



OPEN

## Single trait versus principal component based association analysis for flowering related traits in pigeonpea

Kuldeep Kumar<sup>1,2</sup>, Priyanka Anjoy<sup>3</sup>, Sarika Sahu<sup>3</sup>, Kumar Durgesh<sup>4</sup>, Antara Das<sup>1</sup>, Kishor U. Tribhuvan<sup>1,5</sup>, Amitha Mithra Sevanthi<sup>1</sup>, Rekha Joshi<sup>4</sup>, Pradeep Kumar Jain<sup>1</sup>, Nagendra Kumar Singh<sup>1</sup>, Atmakuri Ramakrishna Rao<sup>3</sup> & Kishor Gaikwad<sup>1</sup>✉

Pigeonpea, a tropical photosensitive crop, harbors significant diversity for days to flowering, but little is known about the genes that govern these differences. Our goal in the current study was to use genome wide association strategy to discover the loci that regulate days to flowering in pigeonpea. A single trait as well as a principal component based association study was conducted on a diverse collection of 142 pigeonpea lines for days to first and fifty percent of flowering over 3 years, besides plant height and number of seeds per pod. The analysis used seven association mapping models (GLM, MLM, MLMM, CMLM, EMLM, FarmCPU and SUPER) and further comparison revealed that FarmCPU is more robust in controlling both false positives and negatives as it incorporates multiple markers as covariates to eliminate confounding between testing marker and kinship. Cumulatively, a set of 22 SNPs were found to be associated with either days to first flowering (DOF), days to fifty percent flowering (DFF) or both, of which 15 were unique to trait based, 4 to PC based GWAS while 3 were shared by both. Because PC1 represents DOF, DFF and plant height (PH), four SNPs found associated to PC1 can be inferred as pleiotropic. A window of  $\pm 2$  kb of associated SNPs was aligned with available transcriptome data generated for transition from vegetative to reproductive phase in pigeonpea. Annotation analysis of these regions revealed presence of genes which might be involved in floral induction like Cytochrome p450 like Tata box binding protein, Auxin response factors, Pin like genes, F box protein, U box domain protein, chromatin remodelling complex protein, RNA methyltransferase. In summary, it appears that auxin responsive genes could be involved in regulating DOF and DFF as majority of the associated loci contained genes which are component of auxin signaling pathways in their vicinity. Overall, our findings indicates that the use of principal component analysis in GWAS is statistically more robust in terms of identifying genes and FarmCPU is a better choice compared to the other aforementioned models in dealing with both false positive and negative associations and thus can be used for traits with complex inheritance.

The United Nations 2nd Sustainable Development Goal (SDG-2) aims to eradicate hunger and malnutrition globally by 2030. The goal has become even more challenging in the current context of the Covid-19 pandemic, which has devastating effect on agricultural sector that by 2030, the number of hungry peoples may exceed 840 million, with the majority (above 381 million) from the Asian (<https://www.un.org/sustainabledevelopment/hunger/>) region. However in order to reach SDG-2 standards and commitments, it is necessary to prioritize nutrition in addition to food security. Pulses are important in combating malnutrition, as in addition to providing a sustainable production system, they are the crucial component of human diet (<http://www.fao.org/resources/infographics/>).

Pigeonpea (*Cajanus cajan* (L.) Millsp.), is a highly nutritious grain legume. Although it is a perennial plant, but primarily cultivated as an annual crop with sowing to flowering duration ranging between 60 to 180 days. While long duration varieties have higher yield potential, lately, a significant shift towards shorter duration

<sup>1</sup>ICAR-National Institute for Plant Biotechnology, New Delhi, India. <sup>2</sup>ICAR-Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India. <sup>3</sup>ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. <sup>4</sup>Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, India. <sup>5</sup>ICAR-Indian Institute of Agricultural Biotechnology, Ranchi, Jharkhand, India. ✉email: kish2012@gmail.com

varieties has occurred so as to accommodate them in diverse cropping systems. Hence, the development of short duration varieties with comparable yield potential is compelling need of the hour. Cultivation of short duration varieties also enables farmers to escape adverse growth conditions, such as drought, severe winter, and disease incidences. The main trait that can be targeted for developing varieties with a specific duration is the number of days to flowering. Thus, biotechnological interventions could be deployed to speed up the development of short-duration varieties<sup>1–4</sup>.

Floral development is a definitive event in the evolution of flowering plants; interestingly no non flowering mutants have been identified to date, and researchers have only been able to alter the days to flowering in plants by modifying a few gene combinations. Floral transition is controlled at both pre and post translation levels<sup>5,6</sup>. Autonomous pathway, vernalization, light dependent floral induction, hormonal control and starch dependent controls are the major floral induction pathways. *FLC*, *SOC1*, *FVE*, *FLD*, *CO*, *FT* are just few of the critical genes involved in this process<sup>7</sup>, while auxin and gibberellic acid serve as the primary hormonal regulators of floral transition<sup>8</sup>. The floral transition phenomenon has been extensively investigated in *Arabidopsis* and a few legumes<sup>9,10</sup>. So far, 306 genes regulating floral development have been characterised in *Arabidopsis*<sup>10</sup>.

To develop shorter duration varieties of pigeonpea, it is essential to understand the mechanisms underlying flowering time variations and its adaptation to different ecologies. Though efforts have been made to characterise the *MADS* box genes, *PEBP* gene family, *CCT* gene family, lncRNAs influencing floral induction and to map the loci governing earliness and domestication related traits, but the genes and markers associated with days to flowering are not yet known in pigeonpea<sup>11–16</sup>. Due to the fact that pigeonpea has four maturity groups<sup>17</sup>, Genome Wide Association Studies (GWAS) can better capture the genetic basis of flowering than bi-parental mapping populations since it uses the natural population that represents all allelic combinations arising out of historical recombinations. The GWAS approach has been shown to be effective in identifying novel genes and QTLs for multiple traits in diverse germplasm of rice and wheat<sup>18,19</sup>. The availability of a 62 K SNP chip and hyper variable markers covering the complete pigeonpea genome, together with low cost sequencing costs, enables efficient GWAS analysis in pigeonpea<sup>20–24</sup>.

Previous reports on GWAS in several legume crops have focused on domestication related loci, resistance against fusarium wilt, and days to flowering in chickpea; days to flowering and maturity in soybean, but a computationally robust analysis is still needed to decipher association and develop markers with high confidence for flowering related traits in pigeonpea<sup>12,25–27</sup>. The current work used an association panel of 142 accessions in order to identify candidate genes and markers for flowering-related traits in pigeonpea.

Principal Component Analysis (PCA) is a powerful dimension reduction and an unsupervised linear transformation technique which aims to extract critical information from phenotypically complex traits while reducing the redundancy in variables and preserving the information parallelly. It reduces a large set of initially correlated variables to a much smaller set of uncorrelated or orthogonal variables termed as PCs. GWAS using PC scores as dependent variables are more reliable and robust than single trait based, and it can reveal possible pleiotropy with increased power<sup>18,28</sup>. As a result, we conducted a PCA based GWAS to discover genetic factors regulating crop architecture with emphasis on flowering. The effectiveness of the approach in identifying significant genes associated with pigeonpea flowering and related traits was further validated through annotation of the flanking regions using transcriptome data of ICPL 20338 accession (PRJNA752250). Thus, the present study was undertaken with the following objectives: (i) Single trait based GWAS using seven association mapping models (GLM, MLM, MLMM, CMLM, EMLM, FarmCPU and SUPER) to identify novel genes to flowering and related traits (ii) PC based GWAS to improve the accuracy and robustness of single trait GWAS and investigate pleiotropy (iii) Genomic prediction and (iv) Annotation of significantly associated SNPs.

## Material and methods

**Plant material, phenotyping and ANOVA.** A collection of 142 accessions representing a global pigeonpea germplasm collection, which includes landraces and breeding lines (mostly of Indian origin) (Table S1), were procured and maintained in the Division of Genetics, Indian Agricultural Research Institute (ICAR), New Delhi, India. These lines were grown using the recommended package of practises for 3 years (2017–18 to 2019–20) in replicates. Data for four quantitative traits, i.e. days to first flowering (DOF), days to fifty percent flowering (DFF), plant height (PH) and average number of seeds/pod (SPP) were taken from each individual line in 2017–18 and 2018–19 and average values from the replicates were used for analysis. In 2019–20, data was collected solely for DOF and DFF traits. Days required to develop one completely opened flower in any plant of a row was noted as DOF, while days required by minimum 50% of plants in a row to have one open flower was noted as DFF. The PH of the plant was noted on maturity considering the last twig as end point, whereas SPP was measured by taking the average of randomly selected 50 pod and rounding off till one decimal point. The basic statistics of the phenotypic data were calculated through the STAR (Statistical Tool for Agricultural Research) tool available at <http://bbi.irri.org/products>. Analysis of variance and broad sense heritability of the data was calculated using Indostat (version 8.1). The whole study complies with relevant institutional, national, international guidelines and legislation.

## Reference based assembly of all lines

The raw sequence data of all 142 lines used in the present study is available at NCBI (<https://www.ncbi.nlm.nih.gov/>). These datasets were downloaded and processed to remove adapters and poor quality reads through Trimmomatic version 0.36<sup>29</sup> using default parameters. High quality reads were mapped to the pigeonpea reference genome of cultivar ICPL 87119 available at NCBI using the BWA tool version 0.7.17<sup>21,30</sup>. Samtools version 1.10 and Freebayes version 1.3.1 were used further for variant calling with minimum 10× depth as basic criteria<sup>31,32</sup>. InDels were removed and SNPs covering minimum of 80% reads were filtered. All 142 .vcf files generated were

Model	Description
General linear model (GLM)	GLM induces the simplest structure for single-locus analysis with population structure (Q) as fixed effect, whereas no random effect component is involved in the model; principal components are used as covariates in such a model to reduce the false positives
Mixed linear model (MLM)	MLM includes the kinship matrix (K) as an additional random effect component; hence it is also called the Q + K model
Multiple loci MLM (MLMM)	MLMM is designed for multiple locus analysis, is an improvement over MLM which incorporates multiple markers simultaneously as covariates in order to partially remove the confounding between testing markers and kinship. Gapit uses forward and backward stepwise linear mixed-model regression to include the markers as covariates
Compressed MLM (CMLM)	In CMLM the similar individuals are assigned into groups through cluster analysis and then groups are used as elements of reduced kinship matrix for random effect structure. The model has improved statistical power compared to regular MLM methods due to grouping or clustering
Enriched CMLM (ECMLM)	ECMLM calculates kinship using different algorithms and then chooses the best combination between kinship algorithms and grouping algorithms
Fixed and random model circulating probability unification (FarmCPU)	This is an iterative approach which iteratively fits both fixed and random effect model to eliminate the models overfitting problem while using stepwise regression in MLMM. To control the false positives, kinship derived from associated markers is used
Settlement of MLM under progressively exclusive relationship (SUPER)	The SUPER model uses the associated genetic markers (pseudo Quantitative Trait Nucleotides) to derive the kinship matrix, instead of all the markers. Whenever a pseudo QTN is correlated with the testing marker, it is excluded from those used to derive kinship. The method has higher statistical power than regular MLM

**Table 1.** Description of the association mapping models.

merged to a single .vcf file through samtools, which was used to develop hapmap through Tassel 5.0<sup>33</sup>. SNPs that showed a Missing Allele Frequency of more than 0.05 were not considered for Hapmap preparation, which ultimately included 168,540 SNPs.

**Population structure analysis.** Population structure is known to affect association studies, but we need to look after its impact. The power to detect population structure is highly dependent on the number of loci utilised<sup>34,35</sup>. Furthermore, increased heterogeneity may lead to false stratification<sup>36,37</sup>. Thus, the combined variant file was filtered for 1 SNPs within a sliding window of 200 kb with maximum depth and minimum number of missing samples and eventually 1229 SNPs out of 168,540 were selected for structure analysis. FastSTRUCTURE v1.0<sup>38</sup> was used to investigate the population structure of all 142 accessions. The number of groups/sub-populations (k) was set from 1 to 10 with the burn-in period, and the number of Markov Chain Monte Carlo (MCMC) replications after burn-in were both set to 100,000 under the "admixture mode". Five independent runs were performed for each k number. Finally, the structure was developed using the STRUCTURE harvester vA.2<sup>39</sup>. The delta K method developed by Evanno et al.<sup>40</sup> was used to determine the optimal value of k.

**Genetic diversity estimation.** The diversity estimation was done using Tassel 5.0<sup>33</sup> based on the nucleotide diversity ( $\pi$ ), Watterson estimator ( $\theta$ ), and Tajima's D index<sup>41–43</sup>.

**Principal component analysis.** PCA was conducted using R software version 4.0.0. Reduced DOF and DFF, but increased SPP, are desirable for the generation of short duration varieties with higher yield potential. Given the intricate inheritance and genetic correlation of quantitative traits, some trade-off is inevitable. As various quantitative phenotypic variables were measured in different units reflecting different types of interpretations, for statistical validity, the original variables were standardized (with mean 0 and variance 1) before attempting PCA. The detailed description about the PCA statistics and their loadings are provided in Table S2.

**Genome wide association study.** GWAS was conducted using the GAPIT package version 3.0 in R software, which employs Bonferroni correction to define statistically significant MTAs. For this study seven association models were implemented namely General Linear Model (GLM), Mixed Linear Model (MLM), Multiple Loci MLM (MLMM), Compressed MLM (CMLM), Enriched CMLM (ECMLM), Fixed and Random Model Circulating Probability Unification (FarmCPU) and Settlement of MLM Under Progressively Exclusive Relationship (SUPER) algorithm<sup>44–50</sup>. These models are distinct in their basic structures and components included as fixed and random effects. Except GLM all the models include mixed effect structures, that is both fixed and random effect components. PC scores were used as covariates or fixed effects in PC based GWAS. CMLM, ECMLM and SUPER generally exhibit higher statistical power as compared to the MLM<sup>51</sup>. Amongst all models, MLMM and FarmCPU algorithms are for multiple loci analysis. FarmCPU model is designed to control both false positives and false negatives as compared to other models<sup>49,51</sup>. Table 1 describes all the association mapping models, also explaining their differences.

**Genomic prediction.** Genomic prediction (GP) modelling was done through the rrBLUP (v4.6) package in R software<sup>52</sup> for ridge-regression based genome-wide regression. The Ridge Regression Best Linear Unbiased Prediction (RRBLUP) model for genome-wide regression assumes the following form,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

Where  $\mathbf{y}$  is the vector of phenotypic values;  $\mathbf{X}$  and  $\mathbf{Z}$  are the design matrices for fixed and random effects respectively;  $\mathbf{b}$  and  $\mathbf{u}$  are the coefficient vectors of fixed and random effects respectively;  $\mathbf{u}$  is assumed to follow normal distribution  $N(0, \mathbf{I}\sigma_u^2)$  and error term  $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$  with  $\mathbf{I}$  being the identity matrix. The random effect coefficient  $\mathbf{u}$  is used to represent the marker effects associated with  $\mathbf{Z}$  being the matrix of genotypes. The variance components  $\sigma_u^2$  and  $\sigma_\varepsilon^2$  are estimated through the Maximum likelihood (ML) or Restricted ML (REML) method.

**Annotation of the associated loci using transcriptome data.** For annotation analysis, a window of  $\pm 2$  kb was used for each associated SNP for annotation analysis. The selected windows were looked at for similarity searches within the assembled *in-house* transcriptome data of early flowering genotypes (ICPL 20338) (PRJNA752250) present in two biological replicates, through the BLASTn program. Initially the raw reads were filtered through trimmomatic version 0.36<sup>29</sup> to remove the poor quality reads with default parameters. Cleaned reads were de novo assembled using Trinity (version 2.1.1)<sup>53</sup> with default parameters. Differentially expressed genes (DEGs) were identified using the edgeR package in the bioconductor environment through R script. These DEGs were filtered on the basis of p value (0.001), FDR < 0.05 and on the basis of fold change of the fragments per kilo-base of transcript per Million fragments mapped (FPKM) value ( $\pm 2$ ). CD-hit web server (<http://weizhong-lab.ucsd.edu/cd-hit/>) was used to remove duplicates. The FPKM values of genes corresponding to vegetative leaves (VL), reproductive leaves (RL), shoot apical meristem (SAM) and reproductive buds (Bud) were retrieved to construct heatmap using Morpheus web server (<https://software.broadinstitute.org/morpheus/>).

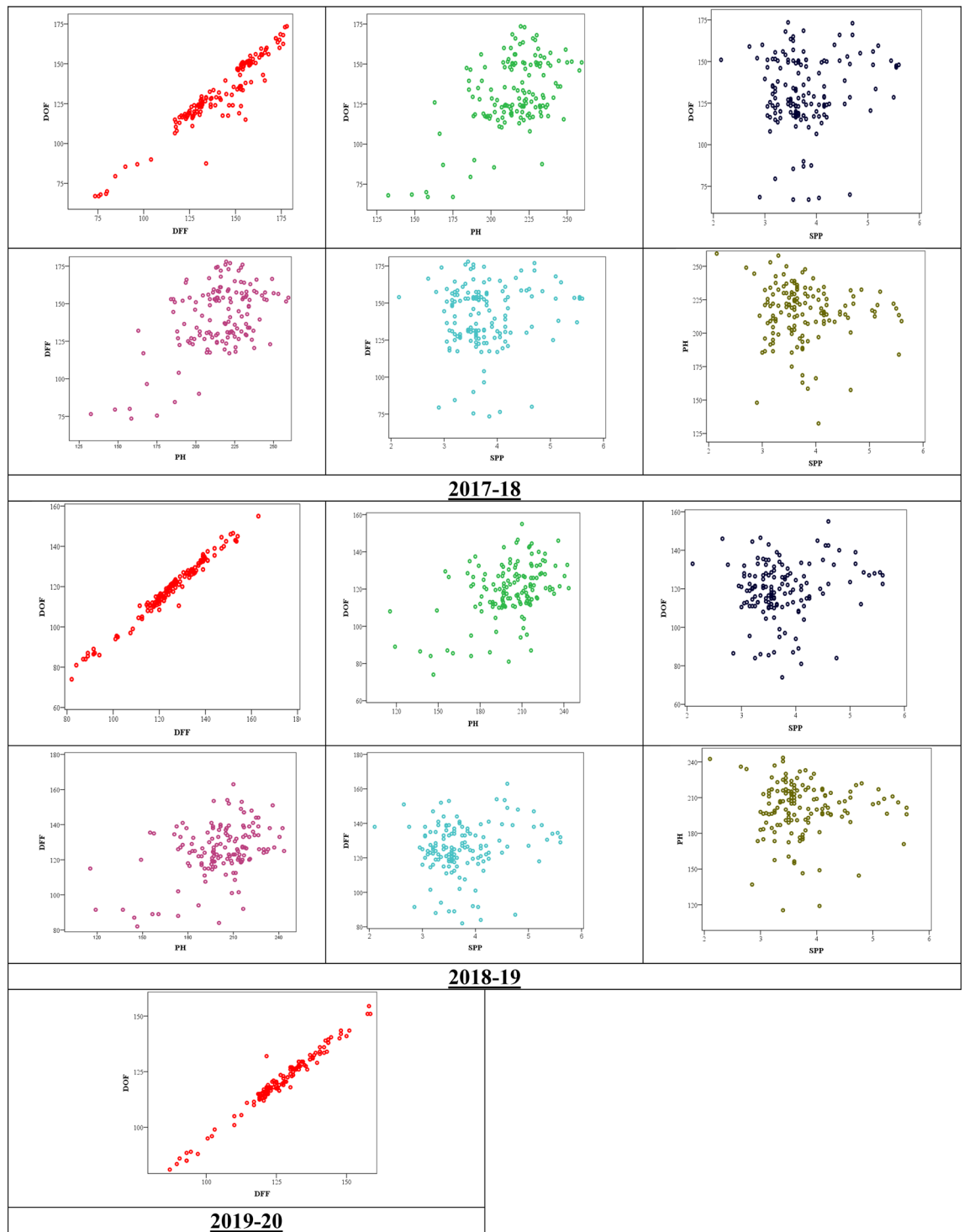
**Ethical approval.** This article does not contain any studies with animals performed by any of the authors.

## Results and discussion

**Descriptive statistics of phenotypes and PCs.** The phenotypic data of quantitative traits, namely DOF and DFF, was taken over 3 years in all 142 lines, whereas data for PH and SPP was taken only for 2 years, i.e. 2017–18 and 2018–19. The cumulative descriptive statistics from the first 2 years of phenotypic data are provided (Tables S3 and S4). The value of DOF was varying from 67 to 174 with an average of 125; the value of DFF was varying from 73 to 178 with a mean of 133; PH in the selected lines were ranging from 118 to 261 with an average of 208; SPP was ranging between 2 to 5 with a mean value of 3. The standard deviations of DOF, DFF, PH and SPP were 0.83, 0.85, 1.01 and 0.02 respectively, while the percentage coefficient of variation (CV) value of DOF, DFF, PH and SPP were 15.4, 15.8, 11.6 and 16.6. Evidently, SPP showed maximum variation and PH manifested minimum variability in terms of percentage CV. For 2017–18 data, the Pearson's linear correlation coefficient between DOF and DFF was 0.95; correlation between DOF and PH was 0.50; DFF and PH had correlation 0.50; whereas, PH and SPP showed negative correlation (Table S4). The same pattern followed in other years too, as depicted in the bi-variate scatter diagram (Fig. 1). Presence of high variability in PH, SPP, DOF and DFF across the year and genotype was observed for all the traits (Table 2). Broad sense heritability ( $h^2$ ) of PH, SPP, DOF and DFF were 0.5449, 0.5897, 0.6593 and 0.7094 respectively (Table 2), which is suggestive that major proportion of the variation is due to difference in genotypes.

The correlation among the original variables is suggestive of using PCA and thereby using PC scores for GWAS analysis. For 2017–18, PC1 explained 58% of the variation in the original data, while PC1 and PC2 together explained 85% of the variation; further adding PC3 explained 98% of the variation. Similarly, in 2018–19, PC1 and PC2 determined 57% and 26% variations respectively. We have taken the first two PCs each year to perform GWAS as they preserved the majority of the variations (> 80%) in the original data (Table S2). The first two traits, namely DOF and DFF, exhibited maximum positive loadings on PC1, followed by PH. Loading on PC2 was higher for SPP trait, whereas plant height showed negative loading on PC2; this means PC2 is mainly representative of SPP. The PCs followed normal distribution, exhibiting non-significant results in Shapiro–Wilk test for Normality (Null hypothesis: Data follows Normal distribution), while amongst the single traits, SPP failed to show normality (SPP had count data). Therefore, PC based GWAS is supposed to improve the statistical power of GWAS analysis.

**Genetic diversity and related analysis.** The nucleotide diversity ( $\pi$ ) is a reflection of genetic diversity which can be used to monitor diversity and genetic variation in crops and related species<sup>54</sup> or to determine evolutionary relationship. In our analysis, the value of  $\pi$  was 0.03573. The Watterson estimator ( $\theta$ ), which is an estimation of the population mutation rate was 0.19898. Usually both  $\pi$  and  $\theta$  ranges in between 0–1, where the inclination toward 0 indicates presence of less diversity. Pigeonpea is an often cross pollinated crop but harbours less diversity, especially in the landraces and cultivated varieties due to progressive bottlenecks during domestication and breeding<sup>12</sup>. Our study was based on only 142 landraces and cultivated varieties mainly of Indian origin, and hence the lesser diversity could be explained. Similar results were reported by other groups also<sup>55,56</sup>. The Tajima's D in our population was  $-2.84379$ . Tajima's D is computed as the difference between two measures of genetic diversity: the mean number of pairwise differences and the number of segregating sites, each scaled so that they are expected to be the same in a neutrally evolving population of constant size. When Tajima's D value is less than 0, it means abundant rare alleles are present, suggesting a possible selective sweep and population expansion.



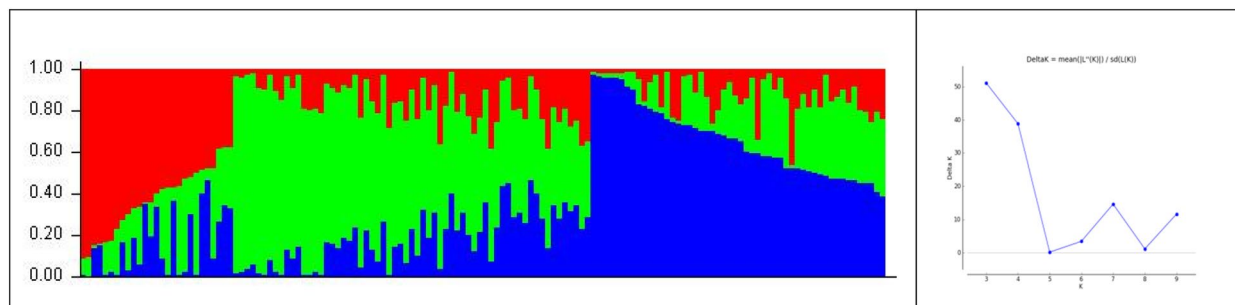
**Figure 1.** Scatter diagrams showing collinearity among the selected phenotypic traits for different years. Upward linear pattern indicates greater extent of positive correlation. Days to first flowering (DOF), days to fifty percent flowering (DFF), plant height (PH) and average number of seeds/pod (SPP).

**Population structure and kinship analysis.** Population structure, kinship analysis, as well as diversity estimates suggested the presence of less diversity among the studied genotypes. The collection was stratified into 3 clusters ( $k = 3$ ) with a substantial level of admixture, probably the result of its pollination behaviour (often cross pollinated), which is in accordance with previous reports<sup>57</sup>. Cluster 1 comprised 28 accessions, while clusters 2 and 3 comprised 63 and 51 lines respectively, in the population structure analysis (Fig. 2). Similar results were obtained with the kinship matrix where the same clustering pattern was observed (Figure S1).



Source	PH		SPP		DOF		DFF	
	d.f.	MSS	d.f.	MSS	d.f.	MSS	d.f.	MSS
Year	1	25,764.88***	1	0.07	2	12,345.08***	2	21,545.25***
Genotype	141	1157.32***	141	0.96***	141	907.85***	141	895.31***
Genotype × Year	282	359.84	282	0.29	282	91.81	282	97.91
Broad Sense heritability ( $h^2$ )	0.5449		0.5897		0.6593		0.7094	

**Table 2.** Pooled Analysis of Variance (Pooled-ANOVA) for four traits evaluated in different environments.



**Figure 2.** Population structure analysis revealed three major clusters in the pigeonpea mini core collection.

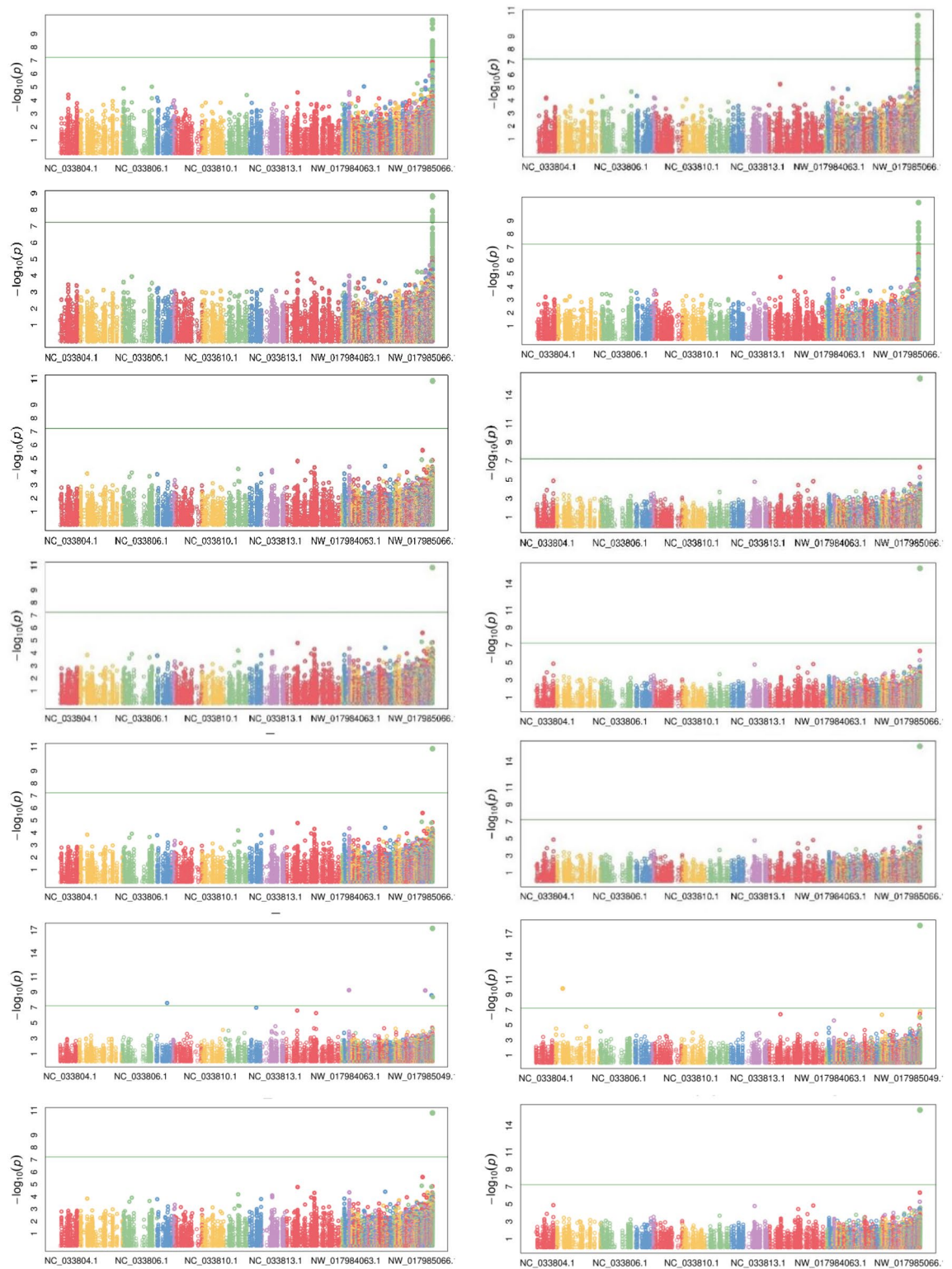
**GWAS results.** In all the aforementioned models, GWAS for quantitative traits like DOF and DFF revealed the association of 19 and 22 non-redundant SNPs distributed on chromosomes 2 and 6 respectively, in the 2017–18 data. However, no significant association was found for PH and SPP. A descriptive summary of GWAS on DOF and DFF in 2017–18 is provided in Tables S5–S6 and Fig. 3. GWAS on 2018–19 data revealed the association of 11 and 9 non-redundant SNPs for DOF and DFF, respectively (Tables S7–S8 and Figure S11). However, no significant association was found for PH and SPP in 2018–19, hence they were excluded from phenotyping in 2019–20. In 2019–20, 13 and 10 non-redundant SNPs were found to be associated with DOF and DFF, by all seven models (Tables S9–S10 and Figure S12). The first two PCs constructed from 2017–18 and 2018–19 data were then analysed for association. PC based GWAS of PC1 in 2017–18 data revealed an association of 18 non-redundant SNPs, while the same in 2018–19 data showed an association of 12 non-redundant SNPs (Tables S11–S12 & Figure S13 and S14). GWAS on PC2 in 2017–18 gave a total of 6 non-redundant SNPs, though no association was found in PC2 2018–19 data (Table S13). As PC2 reflects the loading of SPP, these loci must be regulating SPP traits.

Cumulatively, trait based and PC1 based GWAS on DOF and DFF revealed the association of 22 SNPs with DOF/DFF traits and were selected for further annotation analysis (Table 3). Interestingly, this set of SNPs also included the SNP 812678807:41, which was consistently found to be associated with both the traits as well as the PCs, across all three-year data by each of the seven models employed (Fig. 3 & S11–14). As the distance between the SNPs approached 1.5 Mb, the average  $r^2$  was 0.1<sup>58</sup>. This LD level usually indicates that there is nearly no linkage between the markers after this; thus, we defined SNPs within a 1.5 Mb window as a single locus. So all of the associated SNPs scattered on scaffold NW\_017988637.1 were treated as a single locus for further analysis as it's of only 5 kb in size. Out of 24 SNPs 812678863:261: +, 812678807:41: + and 812679326:250: + were on the same scaffold and hence treated as single loci and the whole scaffold was considered for annotation purposes.

Three SNPs (760222832:55: +, 392479221:11: + and 812679326:250: –) were found to be associated with both PC1 and either DOF/DFF or both, while four SNPs (21256769:324: +, 740074801:308: –, 324910270:94: – and 593701379:271: +) were found to be associated exclusively with PC1 scores, reflecting the ability of PC based GWAS in identifying novel associations. As PC1 is the representative of DOF, DFF and PH, these unique SNPs might be regulating plant height in addition to DOF and DFF. As PH is arrested once flowering starts in the case of determinate lines, the role of these SNPs in regulating plant height can't be ignored. Further, among these novel associations identified by PC, 593701379:271 was annotated as F box protein (Table 4), which is a vital component of auxin signalling playing an important role in vegetative to reproductive phase transition as well as in determining plant height.

Similarly, GWAS on PC2 in 2017–18 gave a total of 6 non-redundant SNPs, though no association was found in PC2 2018–19 data (Table S13). As PC2 reflects the loading of SPP, these loci could be possibly regulating SPP traits. As no marker trait association was found with SPP trait in our analysis and also there is no consistency in PC2 based GWAS, these 6 loci were excluded from the annotation analysis.

**Comparison between the association mapping models.** While association mapping make use of historical recombination to unravel marker trait association, it's difficult to control false positives arising due to linkage disequilibrium (LD), family relatedness and population stratification<sup>51,59</sup>. As a result, choosing an appropriate association mapping model is critical for identifying true marker-trait associations and minimising both false positives and negatives. In essence, an ideal model must have a uniform distribution of expected and



**Figure 3.** Manhattan plots for DOF (Left side) and DFF (Right side) for the year 2017–18. Top to bottom order is GLM, MLM, MLMM, CMLM, ECMLM, FarmCPU and SUPER.

observed p-values. Thus, in this investigation, we examined the Q-Q plots generated by different models in order to identify actual causal maker trait relationships and best suited model. If a Q-Q plot has a straight line close to 1:1, it follows a uniform distribution, indicating that null hypothesis is true (no significant marker trait association is present), whereas deviation depicts the presence of association between testing markers and trait. Upward side deflation of lines represents a false positive association, while a false negative is represented by downward deflation. If the line is close to 1:1 ratio with a sharp upward deviated tail, it indicates that both false positives

S. no.	SNP id	Chromosome	Physical location	Year (trait)
1	812678863:261:+	NW_017988637.1	1117	2017–18 (DOF)
2	392468479:318:+	NW_017984071.1	155,917	2017–18 (DOF)
3	725832748:272:+	NW_017985276.1	22,384	2017–18 (DOF)
4	791831919:74:+	NW_017986933.1	11,488	2017–18 (DOF) and 2019–20 (DOF)
5	142343707:25:+	NC_033807.1	9,366,686	2017–18 (DOF)
6	760222832:55:+	NW_017985856.1	27,685	2018–19 (DOF), 2018–19 (DFF) and 2018–19 (PC1)
7	812678807:41:+	NW_017988637.1	869	2018–19 (DOF) and 2019–20 (DOF), 2018–19 (DFF) and 2019–20 (DFF)
8	376936577:87:-	NW_017984062.1	168,305	2018–19 (DOF)
9	652249420:11:+	NW_017984675.1	23,436	2018–19 (DOF)
10	709017214:7:-	NW_017985090.1	533	2018–19 (DOF)
11	633271872:58:+	NW_017984581.1	74,012	2018–19 (DOF)
12	392479221:11:+	NW_017984071.1	161,167	2019–20 (DOF), 2019–20 (DFF), and 2018–19 (PC1)
13	164755426:8:+	NC_033809.1	6,932,346	2019–20 (DOF) and 2019–20 (DFF)
14	781124881:96:-	NW_017986454.1	6679	2019–20 (DOF)
15	812679326:250:-	NW_017988637.1	863	2017–18 (DFF), 2017–18 (PC1)
16	35373484:284:+	NC_033805.1	6,362,335	2017–18 (DFF)
17	785047004:88:+	NW_017986607.1	3977	2018–19 (DFF)
18	330539130:289:+	NC_033814.1	21,328,862	2019–20 (DFF)
19	21256769:324:+	NC_033804.1	14,401,967	2017–18 (PC1)
20	740074801:308:-	NW_017985477.1	249	2017–18 (PC1)
21	324910270:94:-	NC_033814.1	17,612,083	2018–19 (PC1)
22	593701379:271:+	NW_017984430.1	87,462	2018–19 (PC1)

**Table 3.** List of selected SNPs further used for annotation analysis.

S. no.	SNP id	Putative candidate regulators in 2 kb window of associated SNPs
1.	812678863:261:+	Transcript (TRINITY_DN34349_c0_g1_i9) annotated as cytochrome P450-like TATA box binding protein (cytochrome P450-like TBP)
2.	812679326:250:+	
3.	812678807:41:+	
4.	760222832:55:+	TRINITY_DN35027_c3_g2_i12 was annotated as putative rRNA methyltransferase
5.	785047004:88:+	TRINITY_DN34404_c4_g1_i14 an auxin response factor
6.	633271872:58:+	Genic SNP: <i>in mRNA of pin like transcript variants</i>
7.	593701379:271:+	TRINITY_DN32710_c2_g1_i2 annotated as F-box protein SKIP23
8.	376936577:87:-	TRINITY_DN34296_c0_g1_i10 a serine/threonine protein phosphatase 2A
9.	834373094:36:-	GENIC SNP: ribosomal protein S2
10.	834384838:29:-	GENIC SNP: cytochrome P450 b559 alpha subunit
11.	164755426:80:-	TRINITY_DN34186_c2_g3_i4; annotated as Cytochrome P450 89A2
12.	35373484:284:+	TRINITY_DN33874_c0_g1_i3 annotated as U-box domain-containing protein and TRINITY_DN34453_c0_g3_i10 annotated as chromatin structure remodelling complex protein BSH

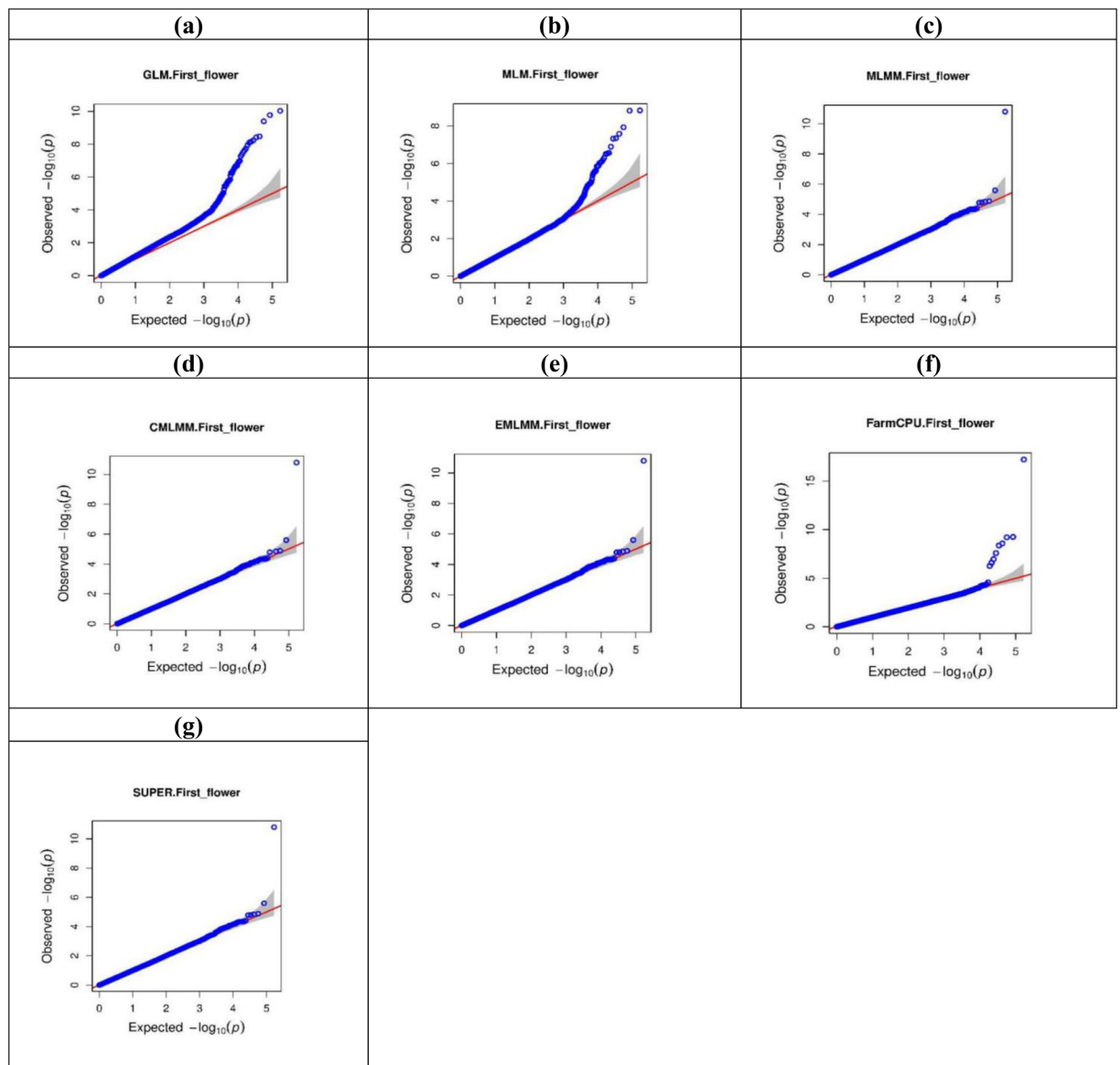
**Table 4.** Annotation of the SNPs showing marker trait association reveals role of auxin pathway genes in flower induction.

and false negatives were controlled and the presence of true associations can be inferred<sup>60,61</sup>. Usually most of the SNPs follow a uniform distribution as they are not in LD with a causal polymorphism, but the few that are in LD with a causal polymorphism will produce significant  $p$  values arising as ‘tail’.

In our analysis, we compared seven models and found four models, viz. CMLM, ECMLM, MLM, and SUPER, showed approximately 1:1 ratio, better than the remaining models, i.e., GLM, MLM, and FarmCPU (Fig. 4 & S2–S10). When the MLM model is used in a genetically diverse panel, its superiority over GLM is lost as the random effect accounted for by the kinship matrix in the former is neutralised by the genetic diversity. Both GLM and MLM are single locus models, i.e. scanning one marker at a time, are computationally demanding and fail to decipher traits which are controlled by multiple loci. From Q-Q plots, it was evident that the GLM model was not able to remove the false positives arising due to LD, and therefore, all SNPs on scaffold NW\_017988637.1 were found to be associated with both DOF and DFF (Fig. 3, S11–S14)<sup>60,61</sup>.

Though the multi-loci models, MLM, CMLM and EMLM are more beneficial in mapping complex traits, they resort to overfitting and give rise to false negatives<sup>49</sup>. In the SUPER model, only the associated SNPs or pseudo quantitative trait nucleotides are used to derive kinship. MLM and its derivative models, which included



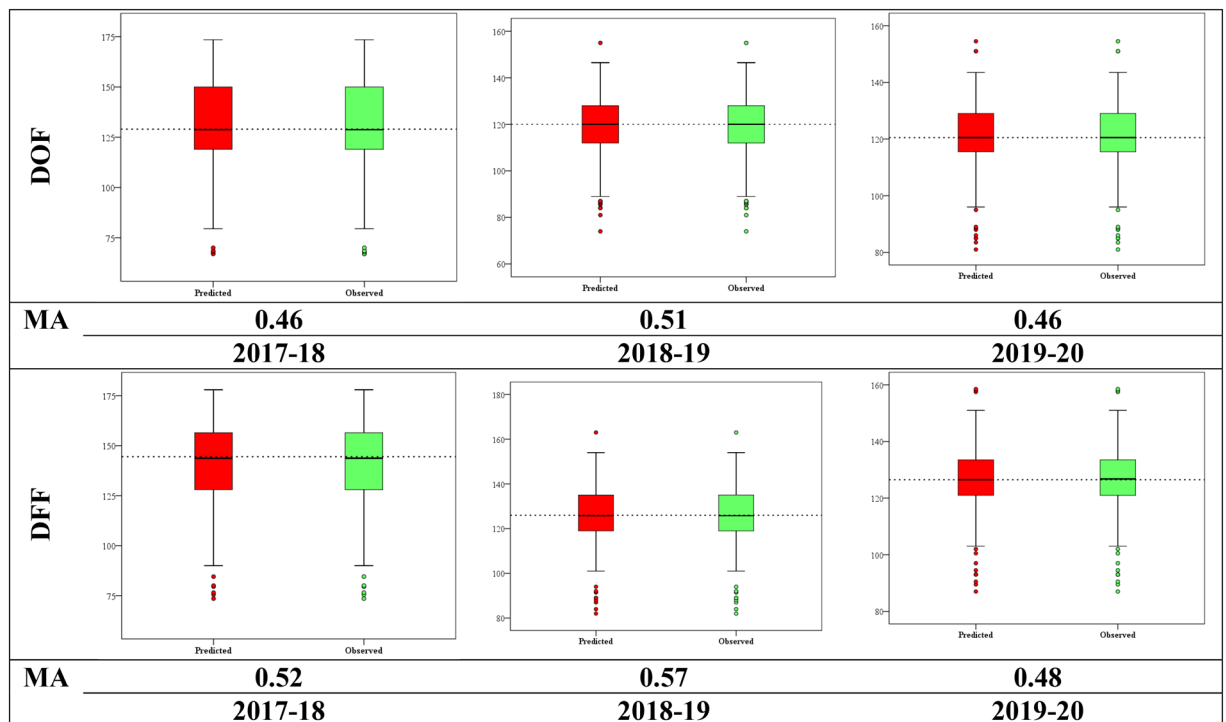


**Figure 4.** Quantile–Quantile (Q–Q) plots based on GWAS results from different association models for DOF in the year 2017–18. Model representations are GLM (a), MLM (b), MLMM (c), CMLMM (d), EMLMM (e), FarmCPU (f) and SUPER (g).  $x$  axis plots expected  $-\log_{10}(p)$  values and  $y$  axis plots observed  $-\log_{10}(p)$  values respectively.

kinship as covariates and perform overfitting of the data leading to the increased  $p$ -value threshold, were superior in controlling the false positives (Table S5–S13), but they also favoured false negatives. Hence, in most cases, these models were able to find only one SNP associated with both DOF and DFF (Fig. 3, S11–S14).

FarmCPU is superior over the other mapping models, as it incorporates multiple markers as covariates to remove confounding between the testing marker and kinship. In our analysis, it was found to be better than the other aforementioned models in dealing with both false positives and negative associations<sup>60,61</sup>. The FarmCPU model overcomes the limitation of false negatives due to overfitting (in the case of CMLM, EMLM, SUPER, and MLMM) and LD based false positive associations (in the case of GLM), as well as being a multi-loci based model, it was appropriate for dissecting the complex traits (Fig. 3 and S11–14).

**Methodological improvement and advantages of PC based GWAS.** Even if we wish to analyse multiple traits, GWAS methodology is essentially based on “*single trait single variant association basis*”. However the phenotypes are not under control of a single locus and there is higher possibility that genetic variants can influence multiple traits or vice-versa. The resolution of understanding complex traits will increase if we study multiple traits simultaneously. PCA based GWAS is one such approach. It is a highly effective method for collecting information from highly correlated, complex and multiple traits through dimension reduction. Many studies have already been done which support the use of PC based GWAS for complex traits<sup>18,62–64</sup>. A compari-

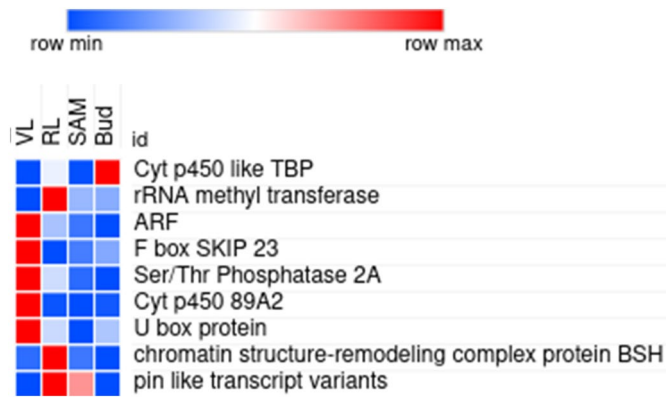


**Figure 5.** Box plot of observed traits vs. predicted flowering days through genomic prediction using RRBLUP method across different year's data. The middle line in each box is the median value. Model accuracy (MA) is provided by setting 80:20 training and testing data sets.

son between the trait and PC based GWAS models suggested the use of PC based GWAS for efficient and high throughput deliverables. This strategy can decrease the likelihood of false positives by avoiding the multiple testing issues<sup>65,66</sup>. As the normal distribution of the phenotype is a must for performing GWAS, PC scores will lead over the single trait as PCA will transform the skewed original variables into an approximate normal distribution, producing reliable GWAS results<sup>67</sup>. Further, GWAS using PC scores may detect genomic regions that could be overlooked by using individual traits, since PC scores represent integrated variables<sup>68</sup>. As many genes contribute to the phenotype of multiple traits of complex architecture, it can be used to describe pleiotropy also. A PC based test has optimal power when the underlying multi-trait signal can be captured by the first PC, and otherwise it will have suboptimal performance<sup>69</sup>. In our analysis, we found some common associations as deciphered by both trait based and PC based GWAS, besides some novel associations identified by PC based GWAS owing to some critical genes during annotation, like locus 593701379:271: + identified exclusively by PC1 based studies was present in the vicinity of TRINITY\_DN32710\_c2\_g1\_i2 annotated as F-box protein SKIP23.

**Genomic prediction.** The genomic estimated breeding value (GEBV) for each line was estimated using all the SNPs and 500 randomly generated train/test sets. The average correlation between the observed DOF and the predicted DOF by GP was 0.46, 0.51 and 0.46 in a model with no significant markers included as fixed effects during 2017–18, 2018–19, and 2019–20, respectively (Fig. 5). Similarly, the correlation between observed DFF and predicted DFF was 0.52, 0.57, and 0.48 during 2017–18, 2018–19 and 2019–20 respectively (Fig. 5). This is comparable to the prediction accuracies (PA) obtained for similar highly heritable traits, days to heading and days to maturity in wheat using a large number of Mexican and Iranian landraces. We observed moderate prediction accuracy in our data. Crossa et al.<sup>70</sup>, found the correlation values for plant height and yield to be 0.87 and 0.49, respectively pertaining to the trait complexity. Though a comparatively smaller number of lines were used in our study, we could still achieve ~50% accuracy, probably owing to the use of the core set. In Brassica, a similar or higher PA were achieved for grain yield related traits<sup>71</sup>. GP will enable high throughput evaluation of germplasm to identify superior one which can then be included in crop breeding programs to perform GP-based progeny selection<sup>70,72</sup>. However, the accuracy of GP models in predicting GEBV in pigeonpea should be increased by including more lines and including more environments for phenotyping to achieve reliable prediction and utilisation of the model.

**Annotation analysis for the significantly associated SNPs showing marker trait association.** All the 22 SNPs representing 12 different loci were considered for gene annotation analysis (Table 3). Four of the 12 loci were found to be present within or in the vicinity of vital flowering related genes (Table 4). The heatmap showing differential expression of nine of these genes in different tissues is presented in Fig. 6. The locus with three SNPs (812678863:261: +, 812679326:250: + and 812678807:41: +) present on the scaffold



**Figure 6.** Expression pattern of the genes found in vicinity of associated SNPs which might have an important role in flowering. Vegetative leaves (VL), reproductive leaves (RL), shoot apical meristem (SAM) and reproductive buds (Bud).

NW\_017988637.1 revealed the presence of cytochrome P450-like TATA box binding protein (cytochrome P450-like TBP) within its vicinity. Several researchers have previously demonstrated the role of plant cytochrome P450s gene family members in various pathways, including hormone biosynthesis<sup>73,74</sup>, which have a bearing on both DOF and DFF, and this observation was strengthened by the expression data with the highest expression in floral bud tissues (Fig. 6).

SNP 785047004:88: + was located adjacent to the Auxin response factor (ARF), which is suggestive of its involvement in SAM to bud transition. Under low auxin concentration, AUX/IAA binds to ARFs, ultimately inhibiting the further downstream genes, whereas when present in higher concentrations, auxin binds to TIR1 (F box protein which is a component of E3 ubiquitin ligase). After auxin binds to TIR1, it gets activated and cleaves AUX/IAA, thereby freeing ARFs and ultimately leading to expression of auxin responsive genes<sup>75–77</sup>. In our analysis, the ARF gene was found to be expressed in VL, which is the phase where auxin is required for the transition from meristem to bud (Fig. 6).

Establishment of a high concentrations of auxin is required for floral induction<sup>78–80</sup>, which is generated by polar auxin transport involving regulators such as the auxin efflux carrier PIN-FORMED1 (PIN1) and the PINOID (PID) kinase, which controls PIN1 activity<sup>81,82</sup>. Interestingly, another SNP (633271872:58: +) was present in the genic region of PIN like transcript variants, which is reported to mediate auxin efflux dependent developmental processes as mutants of these showed defective auxin transport<sup>82,83</sup>. Also, PIN is believed to regulate flowering timing by altering auxin activity in collaboration with ARF. Likewise, a F box protein SKIP23 is a must for induction of downstream auxin responsive genes<sup>84</sup> and was found in the vicinity of SNP (593701379:271: +). As it was found to be highly expressed in VL, it is hypothesised to degrade AUX/IAA and release ARF (found near other SNP i.e. 785047004:88: +).

Similarly, SNP (376936577:87: -) was very close to a serine/threonine protein phosphatase 2A. Although serine/threonine protein phosphatase 2A is known to participate in various stress signals<sup>85</sup>, few reports suggest its role in auxin as well as abscisic acid signalling<sup>86,87</sup>. Although it is not clear how this gene influences flowering, we presume that it regulates flowering by hormonal regulation, mainly by regulating ABA and auxin. Another SNP, 164755426:80: + was found near the cytochrome P450 89A2 subunit. CYP715 (a cyt p450 gene family member) appears to function as a key regulator of flower maturation, synchronising petal expansion and volatile emission<sup>88</sup>. Similarly, this SNP might be involved in the maturation of reproductive buds to flowers.

A transcript (TRINITY\_DN33874\_c0\_g1\_i3) present besides SNP (35373484:284: +) was annotated as a U-box domain-containing protein. The SPIN1 (SPL11-Interacting protein 1) gene has been reported to regulate flowering time in rice and it is ubiquitinated by SPL11 (a U box protein)<sup>89</sup>. The spl11 rice mutants were found to display delayed flowering under long-day conditions. As per a previous report, mutating a U box protein in rice (SPL11) leads to delayed flowering. Transcript (TRINITY\_DN33874\_c0\_g1\_i3) might regulate flowering time as in rice, but interestingly its expression was higher in VL only suggesting that it may regulate flowering through a different mechanism from that of rice.

Another transcript, TRINITY\_DN34453\_c0\_g3\_i10 annotated as chromatin structure remodelling complex protein BSH was found in the vicinity of SNP (35373484:284: +). Several components of chromatin remodelling complexes are evolutionarily conserved in plants, such as the SWI3 subunits<sup>90</sup>, SNF5/BSH subunit<sup>91</sup>, the nuclear actin-related protein ARP4<sup>92</sup>, BRAHMA (BRM), or SPLAYED (SYD). Both of these latter proteins are ATPases of Arabidopsis SWI/SNF complexes and have been shown to participate in the control of flower development and flowering time<sup>93–95</sup>. Likewise, the SWI3B protein interacts with the flowering regulator FCA<sup>90</sup>. Several reports regarding the involvement of epigenetic mechanisms in flower induction regulation are available, and TRINITY\_DN34453\_c0\_g3\_i10 might play a similar role. Interestingly, SNP 760222832:55: + was found in the vicinity of RNA methyltransferase (TRINITY\_DN35027\_c3\_g2\_i12). Many epigenetic regulators have already been reported to regulate flowering timing<sup>96,97</sup>. TRINITY\_DN35027\_c3\_g2\_i12 was found to express constantly in VL, Mer and Bud but the expression increased in RL, suggesting its role in flowering induction.

## Conclusion

In the current study, PC based GWAS was found to be superior over trait based and multi-loci based models for DOF and DFF in Pigeonpea, analysed using 142 accessions and 168,540 SNPs. PC transformation of the traits revealed that PC1 captured 58% of the variation, while PC1 and PC2 cumulatively captured 85% of the variation, suggesting PC1 and PC2 were sufficient enough for GWAS in our datasets. Cumulatively, GWAS revealed the association of 22 SNPs with DOF, DFF or PC1, out of which 15 were solely identified by trait based GWAS, 3 by both trait based as well as PC based GWAS, and 4 SNPs were found to be associated only through PC based GWAS. The 4 SNPs found to be associated with PC1 might be pleiotropic as PC1 also represented PH besides DOF and DFF. One of these 4 SNPs is annotated as F box protein, which plays a vital role in auxin signalling during growth and development, so these 4 SNPs can be inferred as pleiotropic to DOF, DFF and PH. Many of the associated SNPs were in the vicinity of vital genes like Auxin responsive genes like ARF, F box protein, U box protein, PIN like transcripts, chromatin remodelers, RNA methyltransferase the homologs/orthologs, many of which have been previously reported to regulate floral transition in other plant species. A few uncharacterized genes were also found, which are novel and need further characterization in order to decipher their function and role. Associations found in the present study suggest a functional basis of the associations in the regulation of flowering, and hence these genes are excellent candidates for further validation through bi-parental analysis followed by mutagenesis, genome editing, and other approaches. In conclusion, PC based GWAS is effective in deciphering pleiotropy and complex traits over trait based GWAS. Furthermore, the study can be taken forward by combining the PCA and the Multiple Dimension Scaling method to handle both quantitative and qualitative phenotypes as inputs for association mapping models.

## Data availability

All sequencing data used in the current research work are available at (<https://www.ncbi.nlm.nih.gov/>) and the SRA accession numbers to access them are provided in Table S1.

Received: 5 April 2021; Accepted: 18 March 2022

Published online: 21 June 2022

## References

- Kumar, K. *et al.* Climate change mitigation and adaptation through biotechnological interventions. In *Climate Change and Indian Agriculture: Challenges and Adaptation Strategies* (eds Srinivasarao, C. *et al.*) 1–22 (ICAR-National Academy of Agricultural Research Management, 2020).
- Kumar, S. C. *et al.* Mapping QTLs controlling flowering time and important agronomic traits in pearl millet. *Front. Plant Sci.* **8**, 1731 (2017).
- Lu, H. *et al.* QTL-seq identifies an early flowering QTL located near Flowering Locus T in cucumber. *Theor. Appl. Genet.* **127**(7), 1491–1499 (2014).
- Daba, K., Deokar, A., Banniza, S., Warkentin, T. D. & Taran, B. QTL mapping of early flowering and resistance to ascochyta blight in chickpea. *Genome* **59**(6), 413–425 (2016).
- Cho, L. H., Yoon, J. & An, G. The control of flowering time by environmental factors. *Plant J.* **90**(4), 708–719 (2017).
- Putterill, J., Laurie, R. & Macknight, R. It's time to flower: The genetic control of flowering time. *BioEssays* **26**(4), 363–373 (2004).
- Samach, A. Control of flowering. In *Plant Biotechnology and Agriculture Prospects for the 21st Century* (ed. Altman, A.) 387–404 (Academic Press, 2012).
- Wilson, R. N., Heckman, J. W. & Somerville, C. R. Gibberellin is required for flowering in *Arabidopsis thaliana* under short days. *Plant Physiol.* **100**(1), 403–408 (1992).
- Weller, J. L. & Ortega, R. Genetic control of flowering time in legumes. *Front. Plant Sci.* **6**, 207 (2015).
- Bouché, E., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* **44**(D1), D1167–D1171 (2016).
- Kumawat, G. *et al.* Molecular mapping of QTLs for plant type and earliness traits in pigeonpea (*Cajanus cajan* L. Millsp.). *BMC Genet.* **13**(1), 1–11 (2012).
- Varshney, R. K. *et al.* Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* **49**(7), 1082 (2017).
- Kumar, K. *et al.* Identification and characterization of MADS box gene family in pigeonpea for their role during floral transition. *3 Biotech* **11**(2), 1–15. <https://doi.org/10.1007/s13205-020-02605-7> (2021).
- Tribhuvan, K. U. *et al.* Identification and characterization of PEBP family genes reveal CcFT8 a probable candidate for photoperiod insensitivity in *C. cajan*. *3 Biotech* **10**, 1–12 (2020).
- Tribhuvan, K. U. *et al.* Structural and functional analysis of CCT family genes in pigeonpea. *Mol. Biol. Rep.* **49**(1), 217–226 (2022).
- Das, A. *et al.* Non-coding RNAs having strong positive interaction with mRNAs reveal their regulatory nature during flowering in a wild relative of pigeonpea (*Cajanus scarabaeoides*). *Mol. Biol. Rep.* **47**(5), 3305–3317 (2020).
- Upadhyaya, H. D., Reddy, K. N., Gowda, C. L. L. & Singh, S. Phenotypic diversity in the pigeonpea (*Cajanus cajan*) core collection. *Genet. Resour. Crop Evol.* **54**(6), 1167–1184 (2007).
- Yano, K. *et al.* GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl. Acad. Sci.* **116**(42), 21262–21267 (2019).
- Odilbekov, F., Armoniené, R., Koc, A., Svensson, J. & Chawade, A. GWAS assisted genomic prediction to predict resistance to *Septoria tritici* blotch in Nordic winter wheat at seedling stage. *Front. Genet.* **10**, 1224 (2019).
- Singh, N. K. *et al.* The first draft of the pigeonpea genome sequence. *J. Plant Biochem. Biotechnol.* **21**(1), 98–112 (2012).
- Varshney, R. K. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**(1), 83–89 (2012).
- Dutta, S. *et al.* Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *BMC Plant Biol.* **11**(1), 17 (2011).
- Bohra, A. *et al.* New hypervariable SSR markers for diversity analysis, hybrid purity testing and trait mapping in Pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *Front. Plant Sci.* **8**, 377 (2017).
- Singh, S. *et al.* A 62K genic-SNP chip array for genetic studies and breeding applications in pigeonpea (*Cajanus cajan* L. Millsp). *Sci. Rep.* **10**(1), 1–14 (2020).
- Zhang, J. *et al.* Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* **16**(1), 217 (2015).

26. Kamfwa, K., Cichy, K. A. & Kelly, J. D. Genome-wide association study of agronomic traits in common beans. *Plant Genome* **8**(2), 1–12 (2015).
27. Patil, P. G. *et al.* Association mapping to discover significant marker-trait associations for resistance against Fusarium wilt variant 2 in pigeonpea [*Cajanus cajan* (L.) Millspaugh] using SSR markers. *J. Appl. Genet.* **58**(3), 307–319 (2017).
28. Zhang, Y. M., Jia, Z. & Dunwell, J. M. The application of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* **10**, 100 (2019).
29. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120 (2014).
30. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint [arXiv:1303.3997](https://arxiv.org/abs/1303.3997) (2013).
31. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079 (2009).
32. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv preprint [arXiv:1207.3907](https://arxiv.org/abs/1207.3907) (2012).
33. Bradbury, P. J. *et al.* TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
34. Turakulov, R. & Easteal, S. Number of SNPs loci needed to detect population structure. *Hum. Hered.* **55**(1), 37–45 (2003).
35. von Thaden, A. *et al.* Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. *Mol. Ecol. Resour.* **20**(3), 662–680 (2020).
36. Ardlie, K. G., Lunetta, K. L. & Seielstad, M. Testing for population subdivision and association in four case-control studies. *Am. J. Hum. Genet.* **71**(2), 304–311 (2002).
37. Paschou, P. *et al.* PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* **3**(9), e160 (2007).
38. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**(2), 573–589 (2014).
39. Earl, D. A. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**(2), 359–361 (2012).
40. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **14**, 2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x> (2005).
41. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**(10), 5269–5273 (1979).
42. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**(2), 256–276 (1975).
43. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3), 585–595 (1989).
44. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**(8), 904–909 (2006).
45. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**(2), 203–208 (2006).
46. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**(7), 825–830 (2012).
47. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**(4), 355–360 (2010).
48. Li, M. *et al.* Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12**(1), 1–10 (2014).
49. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**(2), e1005767. <https://doi.org/10.1371/journal.pgen.1005767> (2016).
50. Wang, Q., Tian, F., Pan, Y., Buckler, E. S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS ONE* **9**(9), e107684 (2014).
51. Kaler, A. S. & Purcell, L. C. Estimation of a significance threshold for genome-wide association studies. *BMC Genomics* **20**(1), 1–8 (2019).
52. Endelman, J. B. Ridge regression and other kernels for genomic selection with R Package rrBLUP. *Plant Genome* **4**(3), 250–255. <https://doi.org/10.3835/plantgenome2011.08.0024> (2011).
53. Grabherr, M. G. *et al.* Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**(7), 644 (2011).
54. Kilian, B. *et al.* Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during *Triticum monococcum* (einkorn) domestication: Implications for the origin of agriculture. *Mol. Biol. Evol.* **24**(12), 2657–2668 (2007).
55. Kimaro, D., Melis, R., Sibiya, J., Shimelis, H. & Shayanowako, A. Analysis of genetic diversity and population structure of Pigeonpea [*Cajanus cajan* (L.) Millsp.] accessions using SSR markers. *Plants* **9**(12), 1643 (2020).
56. Zavinton, F. *et al.* Genetic diversity and population structure in Beninese pigeonpea [*Cajanus cajan* (L.) Huth] landraces collection revealed by SSR and genome wide SNP markers. *Genet. Resour. Crop Evolut.* **67**(1), 191–208 (2020).
57. Kassa, M. T. *et al.* Genetic patterns of domestication in pigeonpea (*Cajanus cajan* (L.) Millsp.) and wild *Cajanus* relatives. *PLoS ONE* **7**(6), e39563 (2012).
58. Wang, N. *et al.* Genome-wide investigation of genetic changes during modern breeding of *Brassica napus*. *Theor. Appl. Genet.* **127**(8), 1817–1829 (2014).
59. Myles, S. *et al.* Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**(8), 2194–2202 (2009).
60. López-Hernández, F. & Cortés, A. J. Last-generation genome-environment associations reveal the genetic basis of heat tolerance in common bean (*Phaseolus vulgaris* L.). *Front. Genet.* **10**, 954 (2019).
61. Kaler, A. S., Gillman, J. D., Beissinger, T. & Purcell, L. C. Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Front. Plant Sci.* **10**, 1794 (2020).
62. Vujkovic, M., Aplenc, R., Alonzo, T. A., Gamis, A. S. & Li, Y. Comparing analytic methods for longitudinal GWAS and a case-study evaluating chemotherapy course length in pediatric AML. A report from the children's oncology group. *Front. Genet.* **7**, 139 (2016).
63. Rice, B. R., Fernandes, S. B. & Lipka, A. E. Multi-trait genome-wide association studies reveal loci associated with maize inflorescence and leaf architecture. *Plant Cell Physiol.* **61**(8), 1427–1437 (2020).
64. Lin, W. Y. *et al.* Genome-wide association study identifies susceptibility loci for acute myeloid leukemia. *Nat. Commun.* **12**(1), 1–10 (2021).
65. He, L. N. *et al.* Genomewide linkage scan for combined obesity phenotypes using principal component analysis. *Ann. Hum. Genet.* **72**(3), 319–326 (2008).
66. Holberg, C. J. *et al.* Factor analysis of asthma and atopy traits shows 2 major components, one of which is linked to markers on chromosome 5q. *J. Allergy Clin. Immunol.* **108**(5), 772–780 (2001).
67. Boomsma, D. I. & Dolan, C. V. A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. *Behav. Genet.* **28**(5), 329–340 (1998).
68. Goh, L. & Yap, V. B. Effects of normalization on quantitative traits in association test. *BMC Bioinform.* **10**(1), 1–8 (2009).
69. Guo, B. & Wu, B. Integrate multiple traits to detect novel traits-gene association using GWAS summary data with an adaptive test approach. *Bioinformatics* **35**(13), 2251–2257 (2019).



70. Crossa, J. *et al.* Genomic prediction of gene bank wheat landraces. *G3 Genes|Genomes|Genet.* **6**(7), 1819–1834. <https://doi.org/10.1534/g3.116.029637> (2016).
71. Würschum, T., Abel, S. & Zhao, Y. Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. *Plant Breed.* **133**(1), 45–51 (2014).
72. Daetwyler, H. D., Bansal, U. K., Bariana, H. S., Hayden, M. J. & Hayes, B. J. Genomic prediction for rust resistance in diverse wheat landraces. *Theor. Appl. Genet.* **127**(8), 1795–1803. <https://doi.org/10.1007/s00122-014-2341-8> (2014).
73. Miao, Q. *et al.* Genome-wide identification and characterization of microRNAs differentially expressed in fibers in a cotton phytochrome A1 RNAi line. *PLoS ONE* **12**(6), e0179381. <https://doi.org/10.1371/journal.pone.0179381> (2017).
74. Enríquez-Valencia, A. J. *et al.* Differentially expressed genes during the transition from early to late development phases in somatic embryo of banana (*Musa* spp. AAB group, Silk subgroup) cv. Manzano. *Plant Cell Tissue Organ Culture* **136**, 289–302. <https://doi.org/10.1007/s11240-018-1514-6> (2019).
75. Li, W. *et al.* LEAFY controls auxin response pathways in floral primordium formation. *Sci. Signal.* **6**(270), ra23 (2013).
76. Yamaguchi, N. *et al.* A molecular framework for auxin-mediated initiation of flower primordia. *Dev. Cell* **24**(3), 271–282 (2013).
77. Luo, J., Zhou, J. J. & Zhang, J. Z. Aux/IAA gene family in plants: Molecular structure, regulation, and function. *Int. J. Mol. Sci.* **19**(1), 259 (2018).
78. Benková, E. *et al.* Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115**(5), 591–602 (2003).
79. Heisler, M. G. *et al.* Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the Arabidopsis inflorescence meristem. *Curr. Biol.* **15**(21), 1899–1911 (2005).
80. Reinhardt, D., Mandel, T. & Kuhlemeier, C. Auxin regulates the initiation and radial position of plant lateral organs. *Plant Cell* **12**(4), 507–518 (2010).
81. Friml, J. *et al.* A PINOID-dependent binary switch in apical-basal PIN polar targeting directs auxin efflux. *Science* **306**(5697), 862–865 (2004).
82. Gälweiler, L. *et al.* Regulation of polar auxin transport by AtPIN1 in Arabidopsis vascular tissue. *Science* **282**(5397), 2226–2230 (1998).
83. Furutani, M. *et al.* PIN-FORMED1 and PINOID regulate boundary formation and cotyledon development in Arabidopsis embryogenesis. *Development* **131**(20), 5021–5030 (2004).
84. Kipreos, E. T. & Pagano, M. The F-box protein family. *Genome Biol.* **1**(5), 1–7 (2000).
85. Pais, S. M., Téllez-Inón, M. T. & Capiati, D. A. Serine/threonine protein phosphatases type 2A and their roles in stress signaling. *Plant Signal. Behav.* **4**(11), 1013–1015 (2009).
86. Garbers, C., DeLong, A., Deruère, J., Bernasconi, P. & Söll, D. A mutation in protein phosphatase 2A regulatory subunit A affects auxin transport in Arabidopsis. *EMBO J.* **15**(9), 2115–2124. <https://doi.org/10.1002/j.1460-2075.1996.tb00565.x> (1996).
87. Blakeslee, J. J. *et al.* Specificity of RCN1-mediated protein phosphatase 2A regulation in meristem organization and stress response in roots. *Plant Physiol.* **146**(2), 539–553 (2008).
88. Liu, Z. *et al.* A conserved cytochrome P450 evolved in seed plants regulates flower maturation. *Mol. Plant* **8**(12), 1751–1765 (2015).
89. Vega-Sánchez, M. E. *et al.* SPIN1, a K homology domain protein negatively regulated and ubiquitinated by the E3 ubiquitin ligase SPL11, is involved in flowering time control in rice. *Plant Cell* **20**(6), 1456–1469 (2008).
90. Sarnowski, T. J. *et al.* SWI3 subunits of putative SWI/SNF chromatin-remodeling complexes play distinct roles during Arabidopsis development. *Plant Cell* **17**, 2454–2472 (2005).
91. Brzeski, J., Podstolski, W., Olczak, K. & Jerzmanowski, A. Identification and analysis of the *Arabidopsis thaliana* BSH gene, a member of the SNF5 gene family. *Nucleic Acids Res.* **11**, 2393–2399 (1999).
92. Kandasamy, M. K., Deal, R. B., McKinney, E. C. & Meagher, R. B. Silencing the nuclear actin-related protein AtARP4 in Arabidopsis has multiple effects on plant development, including early flowering and delayed floral senescence. *Plant J.* **41**, 845–858 (2005).
93. Wagner, D. & Meyerowitz, E. M. SPLAYED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis. *Curr. Biol.* **12**, 85–94 (2002).
94. Fornara, F., de Montaigu, A. & Coupland, G. SnapShot: Control of flowering in Arabidopsis. *Cell* **141**(3), 550–550 (2010).
95. Wu, J. I. Diverse functions of ATP-dependent chromatin remodeling complexes in development and cancer. *Acta Biochim. Biophys. Sin.* **44**, 54–69 (2012).
96. Yamaguchi, A. & Abe, M. Regulation of reproductive development by non-coding RNA in Arabidopsis: To flower or not to flower. *J. Plant. Res.* **125**(6), 693–704 (2012).
97. Matzke, M. A., Kanno, T. & Matzke, A. J. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annu. Rev. Plant Biol.* **66**, 243–267 (2015).

## Acknowledgements

We acknowledge the support provided by Director, ICAR-NIPB and PG School ICAR-IARI, New Delhi, India. The research work carried out in this manuscript is a major part of Kuldeep Kumar's Ph.D thesis and was funded majorly by ICAR-IARI (NAHEP-CAAST). Authors are grateful and acknowledge the support provided by ICAR-IARI (NAHEP-CAAST) program. Authors also acknowledge the ICAR\_IARI, National Phytotron Facility for providing facilities for controlled experiment on flowering time to generate transcriptome data.

## Author contributions

K.G. and K.K. conceived and designed the study. P.A. and K.K. performed all GWAS analysis and genomic prediction while K.K., S.S., K.T., A.D. and A.R.R. performed all NGS data analysis. K.G., K.D. and K.K. designed and planned the field experiments and data analysis. K.K., P.A., A.D., K.T., K.D., R.J., A.M.S., P.K.J., N.K.S. and K.G. compiled and interpreted the data and wrote the manuscript. All authors have read and approved the final manuscript.

## 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/s41598-022-14568-1>.

**Correspondence** and requests for materials should be addressed to K.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022