BMC Plant Biology

RESEARCH

Open Access

Transcriptome-wide association mapping provides insights into the genetic basis and candidate genes governing flowering, maturity and seed weight in rice bean (*Vigna umbellata*)

Tanmaya Kumar Sahu[1,2†], Sachin Kumar Verma[1†], Gayacharan[1†], Nagendra Pratap Singh[3], Dinesh Chandra Joshi[4], D. P. Wankhede[1], Mohar Singh[1], Rakesh Bhardwaj[1], Badal Singh[1], Swarup Kumar Parida[3], Debasis Chattopadhyay[3], Gyanendra Pratap Singh[1] and Amit Kumar Singh[1*]

Abstract

Background Rice bean (*Vigna umbellata*), an underrated legume, adapts to diverse climatic conditions with the potential to support food and nutritional security worldwide. It is used as a vegetable, minor food crop and a fodder crop, being a rich source of proteins, minerals, and essential fatty acids. However, little effort has been made to decipher the genetic and molecular basis of various useful traits in this crop. Therefore, we considered three economically important traits *i.e.,* flowering, maturity and seed weight of rice bean and identified the associated candidate genes employing an associative transcriptomics approach on 100 diverse genotypes out of 1800 evaluated rice bean accessions from the Indian National Genebank.

Results The transcriptomics-based genotyping of one-hundred diverse rice bean cultivars followed by pre-processing of genotypic data resulted in 49,271 filtered markers. The STRUCTURE, PCA and Neighbor-Joining clustering of 100 genotypes revealed three putative sub-populations. The marker-trait association analysis involving various genome-wide association study (GWAS) models revealed significant association of 82 markers on 48 transcripts for flowering, 26 markers on 22 transcripts for maturity and 22 markers on 21 transcripts for seed weight. The transcript annotation provided information on the putative candidate genes for the considered traits. The candidate genes identified for flowering include HSC80, P-II PsbX, phospholipid-transporting-ATPase-9, pectin-acetylesterase-8 and E3-ubiquitin-protein-ligase-RHG1A. Further, the WRKY1 and DEAD-box-RH27 were found to be associated with seed weight. Furthermore, the associations of PIF3 and pentatricopeptide-repeat-containing-gene with maturity and seed weight, and aldo–keto-reductase with flowering and maturity were revealed.

†Tanmaya Kumar Sahu, Sachin Kumar Verma and Gayacharan contributed equally to this work.

*Correspondence:
Amit Kumar Singh
amit_singh79@yahoo.com
Full list of author information is available at the end of the article

Conclusion This study offers insights into the genetic basis of key agronomic traits in rice bean, including flowering, maturity, and seed weight. The identified markers and associated candidate genes provide valuable resources for future exploration and targeted breeding, aiming to enhance the agronomic performance of rice bean cultivars. Notably, this research represents the first transcriptome-wide association study in pulse crop, uncovering the candidate genes for agronomically useful traits.

Keywords GWAS, Rice bean, WRKY1, DEAD box RH27, Phospholipid transporting ATPase-9, Aldo–keto reductase, HSC80, P-II PsbX

Introduction

The rice bean [*Vigna umbellata* (Thunb.) Ohwi & Ohashi] is a relatively short-duration and an annual warm-season leguminous pulse crop. It is reported to be able to adapt to a diverse range of climatic conditions and can be grown in a wide range of soil types, even in the poor quality soil [1]. It is used as a vegetable, a minor food crop, a fodder crop and a green manure as well. Being a leguminous crop, rice bean has an advantageous role in mixed cropping and preventing soil erosion due to its ability to fix nitrogen in nutrient depleted soils [2]. Rice bean is reportedly resistant to a wide range of biotic and abiotic stresses. Especially it is a source of resistance against biotic stresses such as bruchids, yellow mosaic disease and bacterial leaf spots [3]. Among abiotic stresses, it is tolerant to some degree of water logging [4], acid soils [5], drought [6] and high temperatures.

Rice bean is important for the health of the dependent living beings as well as the soil. It has great potential to overcome food and nutritional deficiency across the globe [7]. In favorable climatic conditions, it produces large amounts of healthy animal fodder and superior quality grains. It is a rich source of proteins, minerals, essential fatty acids and amino acids [8]. Its amino acid composition is reportedly well balanced for human consumption [4, 8, 9]. Though it is a beneficial, nutritious and low maintenance crop, little research has been carried out on it. It has also remained as a neglected crop being cultivated on small areas in India, Nepal and parts of Southeast Asia [2, 10, 11]. Due to this negligence, it suffers from wild and disadvantageous traits such as indeterminate growth habit, asynchronous and late maturity, pod shattering and anti-nutritional compounds [7]. Additionally, the level of enzyme inhibitors, anti-nutritive or toxic factors are low in rice bean compared to other legumes [12].

Rice bean is a diploid crop with 11 pairs of chromosomes ($2n=22$) and has an estimated genome size of ~ 525.60 Mb [13]. The availability of inadequate genomic information on rice bean has made the investigations on this crop challenging. A few genetic maps of rice bean have been constructed and utilized to localize genes for quantitative traits such as seed weight and several domestication related traits [3, 14]. However, these maps do not provide high resolution mapping of targeted traits because they were based on a small number of simple sequence repeat (SSR) markers derived from the related legume species [3]. Very recently in September 2022, one genome assembly of rice bean at the chromosomal level has been released by Chinese Academy of Agricultural Sciences, Beijing, China [13] with the genome size of 475.64 Mb that is 90.49% of the estimated genome size. However, another assembly by International Centre for Genetic Engineering and Biotechnology, New Delhi, India is available at scaffold level [15] having the assembled genome size of 414 Mb. They estimated a total of 31276 highly confidential genes from 15521 scaffolds. Availability of these two assemblies opened an avenue for annotating markers and the corresponding candidate genes for important traits revealed through genome or transcriptome wide association analysis.

The cultivars of rice bean are highly photoperiod sensitive. Therefore, when the crop is grown in subtropical areas, vegetative growth continues for a longer duration and the crop flowers very late. Further, the rice bean crop improvement and its cultivation faces a challenge from other *Vigna* group of crops of similar nature but with superior agronomic traits. The yield potential of rice bean is estimated to be 1200 kg/ha [16], which is comparatively lesser than green gram, black gram and cowpea. Similarly, indeterminate growth habit and pod shattering make the crop unsuitable for large scale or mechanized farming primarily grown as intercrop [16]. Owing to such constraints and the competition from other crops, rice bean cultivation area has gradually decreased and is even discontinued from traditional cultivation areas. However, with the recent trend towards food crop diversification amid climate change and health awareness of the society, rice bean crop is recognized as one of the most potential legume crops to meet out the current and future needs. Therefore, we selected the three most important productivity related traits i.e., days to flowering, days to maturity after crop sowing and seed size, to understand their genetics. Understanding the underlying genetic mechanism governing these traits will help in developing rice bean genotypes with early flowering, early maturity,

and bold seeds, which will enhance the large-scale area expansion under rice bean crop.

Genome-wide association studies (GWAS) involving genotypic and phenotypic data has empowered efficient detection of significant markers associated with the trait of interest [17, 18]. If interest lies in the expressed part of the genome, variant calling from transcriptome sequencing can be adopted, which is comparatively easier, cost effective and time efficient than whole genome sequencing [19, 20]. Further, transcriptome-based variant calling is expected to capture the variants in the parts of the transcripts modified during post-transcriptional modifications and RNA editing. Therefore, we attempted here transcriptome-based variant calling along with association analysis to unravel the putative candidate genes for important productivity related traits like flowering, maturity and seed weight of rice bean. The comprehensive associative transcriptome analysis on rice bean explored here is expected to enlighten the stakeholders on the implementation of transcriptome-based variant calling to identify putative candidate genes for economically important traits in other crops, too.

Materials and methods

Plant material

A diverse set of 100 accessions of rice bean was selected based on the phenotypic characteristics of 1800 accessions, which were acquired from National Genebank of India (http://pgrportal.nbpgr.ernet.in/) and characterized for important agro-morphological traits during 2018 and 2019. We also tried to make a selection of these accessions in such a way that each rice bean growing geographical region of India is represented (Fig. 1). To avoid any admixture, a single plant selection representative to the accession population was made. Based on the phenotypic characteristic information generated from the entire set of rice bean accessions, 100 accessions were selected, having a good range of variation for flowering and maturity period and seed size. The selected 100 accessions were further grown during the rainy season of

Fig. 1 Geographical distribution of 100 rice bean accessions. The map used in this figure has been created using the diva-gis (https://www.diva-gis.org/) web based software. The red points on the map are the site of collection for the genotypes under study

years 2020 and 2021 at the Experimental Farm of ICAR-National Bureau of Plant Genetic Resources, Issapur (28° 34′25″N, 76°50′41″E, 215 m msl) and Experimental Farm Hawalbagh, ICAR-Vivekananda Institute of Hill Agriculture, Almora (79.39° E longitude and 25.35° N latitude, mean rainfall-1000 mm and 1250 m above msl) following augmented block design. Each accession was sown in paired rows of 2 m in length and a row-to-row distance of 60 cm, and a plant to plant distance of 30 cm. The standard rice bean growing practices were followed to obtain the ideal phenotypic expression of genotypes. The detailed passport information about these accessions, such as date of collection, site of origin and cultivar name, is given in Table S1. The phenotypic data for 100 accessions, including days to 50% flowering, days to 80% maturity and 100-seed weight were recorded for all the accessions.

Data acquisition and processing

Phenotypic data analysis

The phenotypic data collected from two locations in two consecutive years was analyzed, and significant variations in the datasets were detected through analysis of variance (ANOVA) for each trait in the interface of SPSS version 15. The frequency distributions and trait-wise comparative density plots were generated using the *sm* package [21] of R.

Genotypic data generation

The transcriptome sequencing of the considered genotypes was carried out in the Illumina Hiseq6000 platform. Single-nucleotide polymorphisms (SNPs) were identified using GATK version v4.1 by mapping cleaned short reads (obtained by fastp v0.12) against the rice bean transcriptome reference VRB3 [22] using BWA-mem v0.7 [23]. Duplicate reads were marked by Picard tools (https://broadinstitute.github.io/picard/) and genomic variants were identified using GATK HaplotypeCaller [24]. A joint variant call set was generated using GATK Genotype GVCFs, and subsequently, the SNP variants were selected and filtered to retain high-quality SNPs.

Pre-processing of genotypic data

The genotypic data was pre-processed in the interface of TASSEL software (Trait Analysis by Association, Evolution and Linkage version 5. 2. 85) [25]. Initially, the markers were filtered out based on minor allele frequency (MAF > 5%), missing data (< 20%) and heterozygosity (< 50%). As rice bean is a cross-pollinated crop the heterozygosity filter of 50% was implemented. Further, indels, triallelic markers and markers with minor states were removed. The genotypes were also examined for > 30% missing data and > 50% heterozygosity.

Population structure and diversity

To estimate the number of populations, the genotypic data was subjected to principal component analysis (PCA) and population structure analysis. The PCA was carried out in the graphical user interface of TASSEL on the filtered set of markers, and the first three principal components were plotted to analyze the population structure. To execute the admixture model of STRUCTURE v2.3.4 [26] with Bayesian Markov Chain Monte Carlo model (MCMC) simulation, the markers having high polymorphic information content (PIC) and genetic diversity (GD) were considered. The PIC and GD for each marker were computed based on the following formulae:

$$GD = 1 - \sum_{i=1}^{2} a_i^2$$

and,

$$PIC = GD - 2a_1^2 a_2^2$$

where, a_i ($i = 1, 2$) is the frequency of i^{th} allele in the population. Here, the value of i varies from 1 to 2, because we have considered only the bi-allelic markers. The markers with PIC value >= 0.35 were selected for STRUCTURE analysis with the parameters: K = 2–7, 25000 burnin, 50000 MCMC iterations and ten independent runs. The best value of K was determined using Structure Harvester [27] and the corresponding population membership file was used as the Q-matrix for the association analysis.

Linkage disequilibrium analysis

The linkage disequilibrium (LD) decay analysis was carried out in the interface of TASSEL to examine the genomic distance within which the genomic elements are believed to have strong LD. As we have used transcriptome data, short-distance LD has been examined by plotting the frequency of the squared allele of LD (r^2) (Y-axis) against distance in base pair (X-axis) in the R command line for all pair-wise comparisons.

Association analysis

The association analysis of markers with the considered traits was carried out using two single locus models (with two variations each) and six multi-locus models. Among the single locus models, generalized linear models (GLM) [28] and mixed linear models (MLM) [29, 30] have been used. Two variants each of GLM [GLM (Q) and GLM (PCA)] and MLM [MLM (Q + K) and MLM (PCA + K)] have been used where Q involves the population membership information (Q matrix) as covariates and PCA involves first three principal components (PC) as covariates in the models. The K in MLM refers to the kinship matrix (K) generated based on identity-by-state analysis.

These models were implemented using TASSEL, GAPIT [31] and mrMLM.GUI package [32].

Six multi-locus models viz., multi-locus random-SNP-effect MLM (mrMLM) [33], fast multi-locus random-SNP-effect MLM (FASTmrMLM) [34], fast multi-locus random-SNP-effect EMMA(FASTmrEMMA) [32], Iterative Sure Independence Screening EM-Bayesian LASSO (ISIS EMBLASSO) [35], fixed and random model circulating probability unification (FarmCPU) [36] and Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK) [37] were used in the study to analyze the marker-trait associations. The FarmCPU and BLINK were implemented through GAPIT package whereas all other multi-locus models were executed through mrMLM.GUI package in R.

Marker selection

Markers have been selected based on different parameters for thresholds in different software packages. For the single locus models implemented through TASSEL, the markers were selected based on "–log10 (p)" value > 5.99 after Bonferroni correction [38] *i.e.,* 0.05/total number of markers. In the case of multi-locus models implemented through GAPIT, the markers were also selected based on "–log10(p)" value > 3.69(*i.e.,* 2e-4) [39]. However, in mrMLM a separate marker screening parameter viz, logarithm of the odds (LOD) score was used. Here, markers were considered to be associated with a trait of interest if they had the LOD score > 3 [39]. Though the initial screening of the markers was based on the above defined parameters, the final screening was on the basis of their association predicted by at least two GWAS models.

Candidate gene identification

The transcripts on which the significant markers are located were subjected to a BLAST (Basic Local Alignment Search Tool) search for identifying the corresponding genes. To perform the BLAST search, the transcript sequences were aligned to a local protein sequence database using the blastx program of the offline BLAST. For creating a local database, the protein sequences of *Vigna*, G*lycine max* and *Arabidopsis* genus were collected from the National Center for Biotechnology Information (NCBI), USA. As the latter two plant species are well annotated, and *G. max* is the close relative of rice bean apart from other *Vigna* members, the protein sequences of these two plants were included to develop the local BLAST database. Further, all the transcripts were subjected to BLAST2GO of OmicsBox tool (https://www.biobam.com/omicsbox/) for annotation with gene ontology (GO) terms.

Chromosomal localization of associated markers and transcript synteny

To unravel the putative chromosomal location of markers as well as candidate genes, 50 bp left and 50 bp right flanking to the markers were extracted using a developed R- script and the resulting 101 bp fragments were subjected to the offline blastn program against the recently released fully sequenced genome of *V. umbellata* cultivar FF25 [13]. The exact marker locations were identified from the ungapped alignment of SNP sequences and chromosomal sequences obtained through blastn program. A chromosomal map of associated markers was created based on the newly released fully sequenced genome of *V. umbellata* in the interface of MapChart [40]. Further, full length transcripts were subjected to blastn locally, against the genomes of *V. radiata, V. angularis, V. mungo, V. unguiculata* and *V. umbellata* to identify their corresponding chromosomal locations. The resulting chromosomal coordinates were used to carry out a synteny analysis using ShinyCircos software [41].

Expression analysis of associated transcripts

The expression of genes related to flowering, maturity and seed weight represented by the associated transcripts was checked in transcriptome data of inflorescence and developing seed tissues. The RNA from the two different developing seed stages i.e., 5-days post anthesis (SRR16122607) and 10-days post anthesis (SRR16122602) of rice bean (accession: IC426787) was sequenced on Illumina HIseq4000. Additionally, the RNA sequencing reads of rice bean samples at the young inflorescence stage (SRR5764826) were also downloaded from NCBI. These reads were processed using FastQC (version 0.11.9) [42], Trimomatic(version 0.40) [43], bwa [44] and samtools [45] for quality check, trimming, mapping and obtaining FPKM (fragments per kilobase of exon per million mapped fragments) values, respectively. A heatmap of the expression pattern was generated using an R-script with heatmap() function.

Results

Variation in phenotypic data

The phenotypic data for two years (2020 and 2021) from two locations (Delhi and Almora) for 100 selected accessions for some traits were found to be significantly different from each other either location-wise or year-wise (Fig. 2). The distribution of trait data for all the datasets is shown in Fig. 3. The minimum, maximum and mean values of Almora datasets were found within a short range (days to 50% flowering: 45–101 days, days to 80% maturity: 67–148 days, 100-seed weight: 5.65–9.68 g), whereas in case of Delhi datasets, these values are observed within

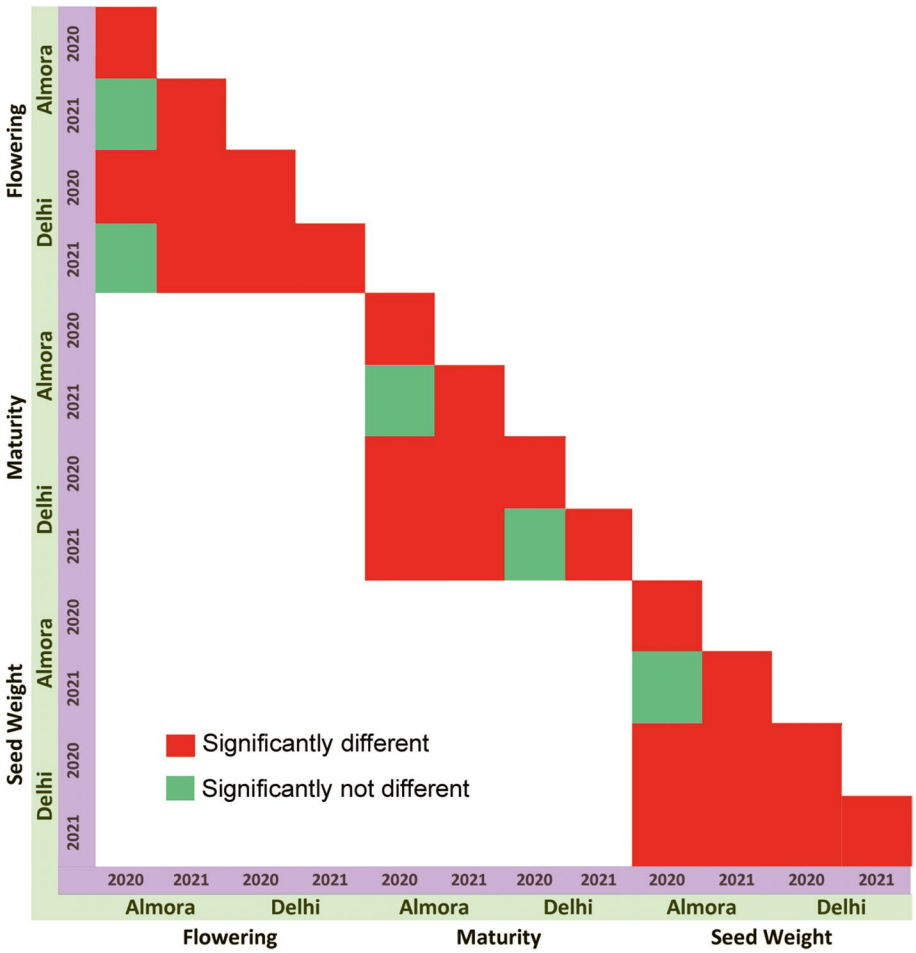

Fig. 2 Analysis of significant difference in the phenotypic data of two years (2020–2021) from two locations (Delhi and Almora) based on ANOVA

a long range (days to 50% flowering: 44–144 days, days to 80% maturity: 52–156 days, 100-seed weight: 0.73–10.79 g) probably due to unexpectedly changing climatic conditions at Delhi. Further, the density plot for days to 50% flowering reveals similar distribution of Delhi datasets, whereas the density plot for 100-seed weight shows similar distribution of Almora datasets. However, for maturity data points of all the datasets were found to be homogeneously distributed. Furthermore, two accessions; EC934417 and IC116118 were identified as early flowering (50 days and 65 days to 50% flowering respectively, averaged over 4 locations). Additionally, these accessions also exhibit early maturation, with EC934417 taking 77 days and IC116118 taking 98 days on an average to reach 80% maturity across the same four locations.

Processed genotypic data

The initial filtration of sequencing reads before variant calling generated cleaned short reads. The number of bases and reads before and after filtration is given in Table S2. The pre-processing of genotypic data with marker filtration parameters like indels, non bi-allelic markers, minor SNP states, minor allele frequency (< 5%), missing genotype (> 20%), heterozygosity (< 50%) resulted in 49,271 markers. Further, the implementation of genotype filtration parameters retained all the genotypes as all have missing data < 30% and heterozygosity < 50%. The final processed genotypic data contained 49,271 markers for 100 genotypes, for which the phenotypic data contained observations for 50% flowering, 80% maturity and 100 seed weight.

Population structure

The calculated GD values for each marker varied from 0.0582 to 0.5000 with an average of 0.2690 whereas the PIC values varied from 0.0565 to 0.3750 with an average of 0.2253 (Figure S1). The markers having the PIC above 0.3500 were found to have the GD above 0.4500. Therefore, the STRUCTRE software was executed with 5416 markers having high PIC and GD that is expected to

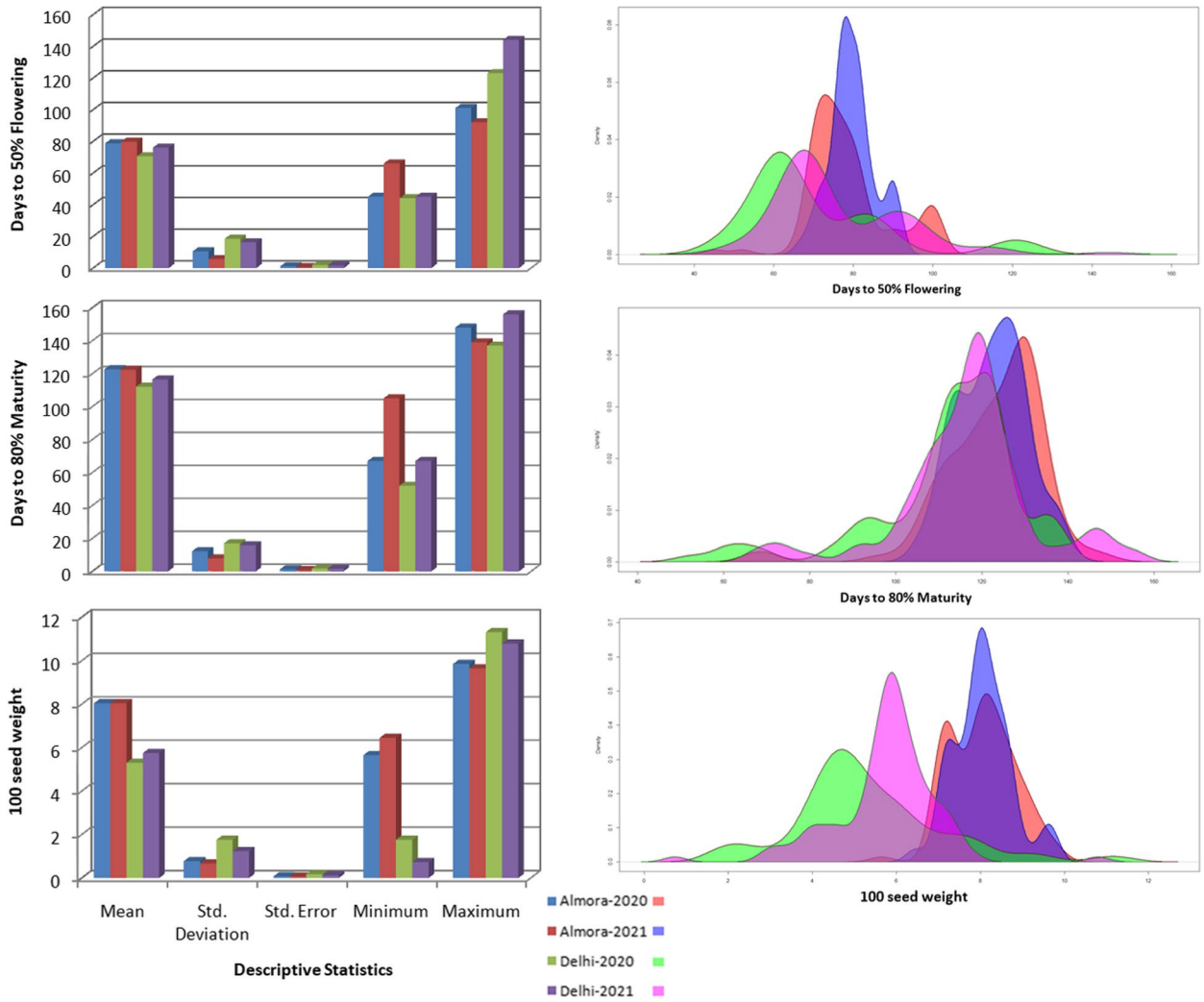

Fig. 3 The distribution of phenotypic data under each trait for all the datasets showing the descriptive statistics and density plots

correctly determine the number of populations. The best value of K with the highest ΔK was determined to be 3 (Figure S2) suggesting that the genotypes are distributed in three putative populations (Fig. 4A). Further, three putative populations have also been revealed through the genotypic cluster (Fig. 4B) by Neighbour Joining method and by plotting the first three principal components (Fig. 4C). Though the genotypic clustering reveals 3 distinct populations split from the root node, STRUCTURE showed a few admixed individuals in three sub-populations. Large number of pure individuals was observed in the largest cluster (sub-population-1 with 53 individuals), followed by a smaller cluster (sub-population-2 with 32 individuals) than sub-population-1 and the smallest cluster (sub-population-3 with 15 individuals). Although sub-population-2 is larger than sub-population-3, the number of pure individuals in these sub-populations are nearly equal. These three putative sub-populations when matched with the passport information of the cultivars (Table S1), sub-population -1 was found to contain mostly the individuals of eastern and north-eastern regions of India and sub-population -2 was observed to have mostly the north Indian cultivars. The sub-population -3 was noticed to contain mixed individuals. However, few exotic cultivars were also noticed to fall within the sub-populations-1 & 2. The distribution of Indian cultivars is given in the Fig. 1.

LD decay

The LD was observed to be decayed at a distance of 1.5 kb at the r^2 cutoff of 2% (Fig. 5). Further, the r^2 was noticed to be on an average of 5% within a distance of 0.5 kb; however, it decayed rapidly from 4 to 2% within a

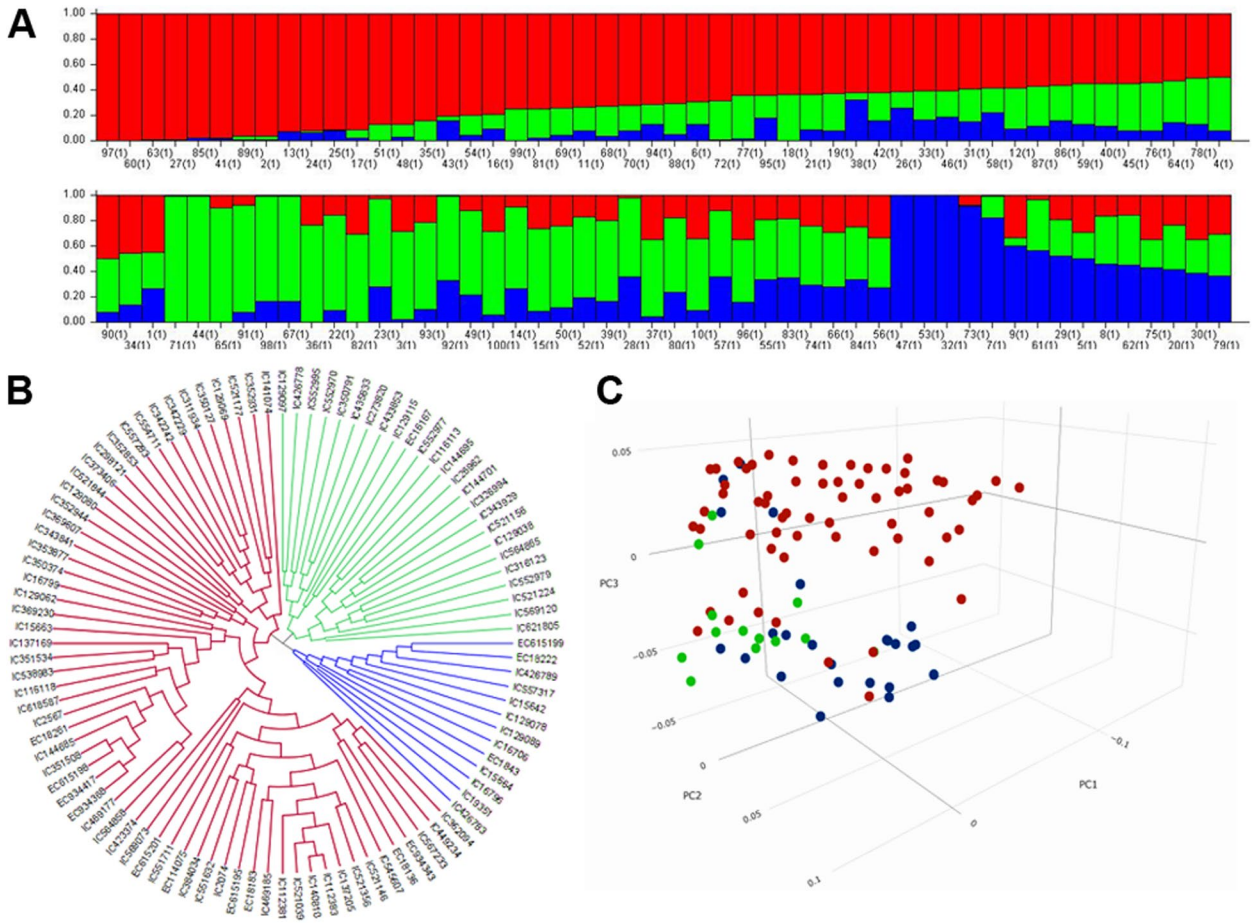

Fig. 4 Three putative sub-populations revealed from (**A**) the STRUCTRE-MCMC simulations, (**B**) the clustering based on Neighbor Joining method and (**C**) the plot of first three principal components

distance of 0.5 kb to 2 kb. After that, the trend line seems to be in equilibrium up to a distance of 6 kb.

Marker-trait association

Marker-trait association with Almora datasets

For Almora 2020, a total of 33 markers have been detected by single and multi-locus models for all the considered traits (Table 1, Fig. 6). The markers considered here are predicted by at least two methods. With this dataset, 3 markers for flowering, 24 markers for maturity and 7 markers for seed weight were detected. However, one marker (SVUTC25856_1400) was found common for both maturity and seed weight traits. The marker, SVUTC06910_1648, predicted for maturity was predicted by 6 out of 10 GWAS models. Further, five markers at different positions (177 bp, 1616 bp, 1648 bp, 1702 bp and 1746 bp) were found significant on one transcript VUTC06910, which are predicted to be associated with maturity trait. The remaining markers in this dataset were predicted on different transcripts.

In the case of phenotypic data collected from the Almora location in the year 2021, a total of 11 markers were detected to be associated with different traits (Table S3, Figure S3). For this dataset, only multi-locus models of mrMLM have detected significant markers which are selected based on the LOD score. In this case, 3 markers (SVUTC21295_283, SVUTC22319_77, SVUTC31327_3358) for flowering, 3 markers (SVUTC21295_283, SVUTC05378_2089, SVUTC21831_6083) for maturity and 6 markers (SVUTC01132_259, SVUTC06850_2182, SVUTC09951_1960, SVUTC23058_1028, SVUTC24042_2561, SVUTC24699_1593) for seed weight traits were predicted where one marker (SVUTC21295_283) was found common for both flowering and maturity traits. Here, all the markers except SVUTC21295_283 were predicted by exactly two different multi-locus models. Though SVUTC21295_283 has been predicted by only one method, it was considered, as it was predicted for two different traits. All the markers

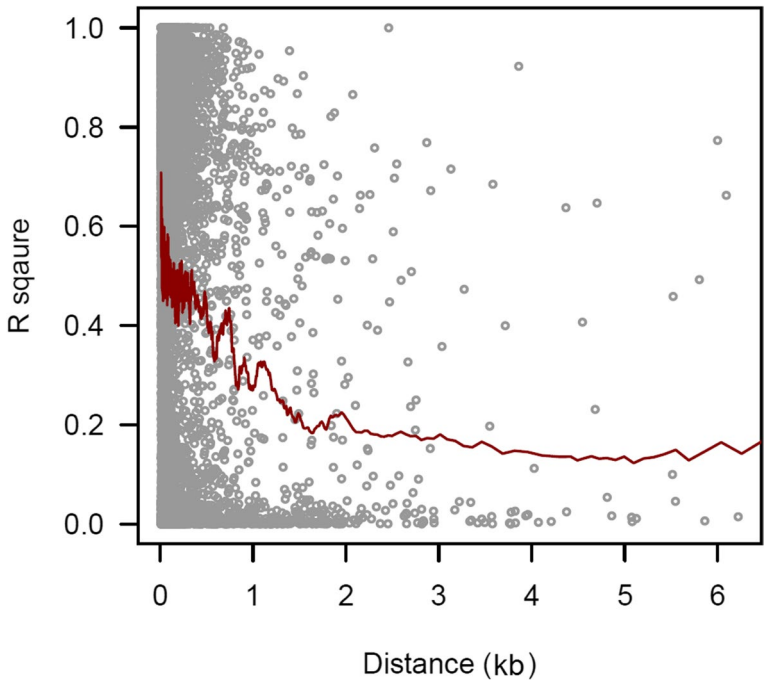

Fig. 5 The decay in LD plotted by estimating the LD between all possible pair-wise markers. The LD decays at a distance of 1.5 kb at the *R*-square cutoff of 2%

were predicted on different transcripts. However, one transcript VUTC22319 was found common in both years with different marker positions (438 in 2020 and 77 in 2021) and was found to be associated with flowering for the datasets of both years.

Marker-trait association with Delhi datasets

The marker-trait association analysis for phenotypic data of Delhi in the year 2020 revealed a total of 58 markers, out of which the majority of the markers (49) were for flowering trait, whereas 4 for maturity and 5 for seed weight were identified (Table S4, Figure S4). Here, none of the markers were found to be associated with more than one trait. The 58 markers are located on 24 transcripts, where seven transcripts contain more than two markers each. The transcript VUTC28154 contains the highest (8) number of markers on it. Out of 49 markers of flowering, 43 markers were predicted only by single locus GWAS models. A marker, SVUTC28154_1646 predicted to be associated with flowering, has been predicted by 8 out of 10 considered GWAS models.

For the phenotypic data collected from Delhi in 2021, the GWAS analysis revealed the association of 49 markers for the three considered traits (Table S5, Figure S5). Out of these 49 markers, 4 for seed weight, 4 were for maturity and 42 were found to be associated with flowering. Interestingly, one marker (SVUTC25248_8323) being predicted by three methods was found to be associated with maturity and seed weight. Nine transcripts associated with flowering were identified by at least two markers on them, where the transcript VUTC28201 contains highest number of markers on it. A total of 15 markers and eight transcripts were found common between two years of data of Delhi, all of which were observed to be associated with the flowering trait.

Associated markers and corresponding transcripts

From the overall analysis, 87 transcripts were found associated with the traits, having 127 markers in total from all the datasets considered. From the datasets of two locations in two consecutive years, neither any marker nor any transcript could be found in common for all. However, 15 markers were noticed to be common between Delhi-2020 & Delhi-2021 datasets (Fig. 7A). Transcripts are concerned, one transcript (VUTC22319: flowering) between Almora-2020 & Almora-2021 datasets and 8 transcripts (VUTC28154, VUTC28165, VUTC28183, VUTC28185, VUTC28192, VUTC28201, VUTC29109, VUTC28186; all associated with the flowering trait) between Delhi-2020 and Delhi-2021 datasets were identified to be common (Fig. 7B). Interestingly, one transcript (VUTC25312: flowering) was also found common between Almora-2020 and

Table 1 Associated markers for flowering, maturity and seed weight predicted using phenotypic data of Almora location in the year 2020

Marker ID[a]	SNP[c]	Transcript ID	Position (in bp)	Models[b]	Trait(s)	-log_{10} (P)/ LOD[c]	MAF[c]	R^2
SVUTC32100_1612	C/T	VUTC32100	1612	6,8,5	Flowering	*LOD*: 4.52 to 7.07	0.260	5.63 to 11.99
SVUTC02300_1585	C/A	VUTC02300	1585	6,7	Flowering	*LOD:* 6.15 to 11.00	0.464	7.10 to 5.38
SVUTC22319_438	C/T	VUTC22319	438	6,7	Flowering	*LOD:* 3.84 to 7.72	0.355	4.72 to 9.66
SVUTC06910_1648	C/T	VUTC06910	1648	4,3,2,1,9,10	Maturity	*-log10(P):* 6.31 to 13.33	0.055	38 to 47
SVUTC06910_1702	G/T	VUTC06910	1702	4, 3,2,1,10	Maturity	*-log_{10}(P):* 6.31 to 13.33	0.055	38 to 47
SVUTC06910_1746	C/G	VUTC06910	1746	4,3,2,1,10	Maturity	*-log10(P):* 6.31 to 13.33	0.055	38 to 47
SVUTC06910_177	C/G	VUTC06910	177	4,3,2,1	Maturity	*-log10(P):* 6.79 to 13.51	0.063	37 to 49
SVUTC02283_1297	G/T	VUTC02283	1297	4,2,1	Maturity	*-log10(P):* 6.07 to 10.96	0.092	28 to 41
SVUTC07229_692	T/C	VUTC07229	692	6,5,8	Maturity	*LOD*: 5.68 to 7.24	0.220	23 to 24
SVUTC08788_874	A/G	VUTC08788	874	3,2,1	Maturity	*-log10(P):* 6.20 to 11.35	0.057	29 to 43
SVUTC21395_469	A/G	VUTC21395	469	5,8,6	Maturity	*LOD:* 3.56 to 6.03	0.200	8.33 to 18.24
SVUTC27187_1237	C/T	VUTC27187	1237	3,2,1	Maturity	*LOD*: 6.15 to 11.00	0.060	30 to 43
SVUTC27441_1674	A/C	VUTC27441	1674	3,2,1	Maturity	*-log10(P):* 5.99 to 10.97	0.060	28 to 41
SVUTC06910_1616	C/G	VUTC06910	1616	7,1	Maturity	*-log10(P):* 6.61 *LOD:* 4.07	0.056	10 to 27
SVUTC16925_1108	T/C	VUTC16925	1108	8,6	Maturity	*LOD*: 6.02 to 7.39	0.262	13.33 to 14.64
SVUTC31321_227	A/C	VUTC31321	227	8,10	Maturity	*log10(P):* 6.30 *LOD:* 5.43	0.100	16.84
SVUTC34182_2284	C/T	VUTC34182	2284	6,10	Maturity	*log10(P):* 6.31 *LOD:* 3.01	0.306	4.60
SVUTC25856_1400	T/G	VUTC25856	1400	5,8,6	Maturity, Seed weight	*LOD:* 3.99 to 5.88	0.214	7.06 to 13.22
SVUTC06074_3949	G/A	VUTC06074	3949	8,6	Seed weight	*LOD:* 3.85 to 6.63	0.146	9.59 to 11.35
SVUTC12822_2877	A/G	VUTC12822	2877	8,6	Seed weight	*LOD:* 4.36 to 5.59	0.191	6.99 to 7.43
SVUTC21543_375	C/T	VUTC21543	375	8,6	Seed weight	*LOD*: 3.5 to 5.46	0.247	8.00 to 11.45
SVUTC25312_5309	G/A	VUTC25312	5309	8,6	Seed weight	*LOD:* 9.18 to 9.86	0.468	6.90 to 12.33
SVUTC25491_6046	T/C	VUTC25491	6046	8,6	Seed weight	*LOD*: 7.91 to 8.72	0.261	3.15 to 4.06
SVUTC30884_3008	T/G	VUTC30884	3008	8,6	Seed weight	*LOD:* 6.03 to 11.58	0.205	10.51 to 26.26

[a] The markers highlighted in yellow are associated with more than one trait

[b] 1. GLM (Q), 2. GLM (PCA)], 3. MLM (Q + K), 4. MLM (PCA + K), 5. mrMLM, 6. FASTmrMLM, 7. FASTmrEMMA, 8. ISIS EMBLASSO, 9. FarmCPU, 10. BLINK

[c] *SNP* single nucleotide polymorphism, *LOD* logarithm of the odds, *MAF* minor allele frequency

Delhi-2021 datasets, even though no marker was identified to be common between these two datasets.

From 3 out of 127 markers and 4 out of 87 transcripts were identified to be associated with more than one trait. One marker (SVUTC21295_283) between flowering and maturity and two markers (SVUTC25856_1400, SVUTC25248_8323) between flowering and seed weight were identified (Fig. 7C). Further, one transcript (VUTC21295) between flowering and maturity, two transcripts (VUTC25856, VUTC25248) between maturity and seed weight and one transcript between flowering and seed weight (VUTC25312) have been identified (Fig. 7D). However, later one was predicted for two different datasets (Almora-2020 and Delhi-2021 with two different markers *i.e.,* SVUTC25312_5309 and SVUTC25312_4994, respectively). These common markers and transcripts are expected to help understand the interrelation between the considered traits in terms of the governing genes.

Annotation of candidate genes using associated transcript sequences

The annotation of transcripts in terms of chromosomal localization (Table S6) and molecular function (Table S7) has revealed many functional proteins related to the associated traits. Based on the BLAST results, majority of transcripts showed top hits against the proteins of *V. umbellata* and *Vigna anguilaris.* BLAST2GO revealed GO terms for each transcript sequence. The major GO terms for biological processes obtained as annotations for 87 transcripts are presented in the form of a pi-chart in Fig. 8. From the annotation of candidate genes, the potential contributors to the process of flowering, such as HSC80, P-II PsbX, phospholipid-transporting-ATPase-9, pectin-acetylesterase-8, and

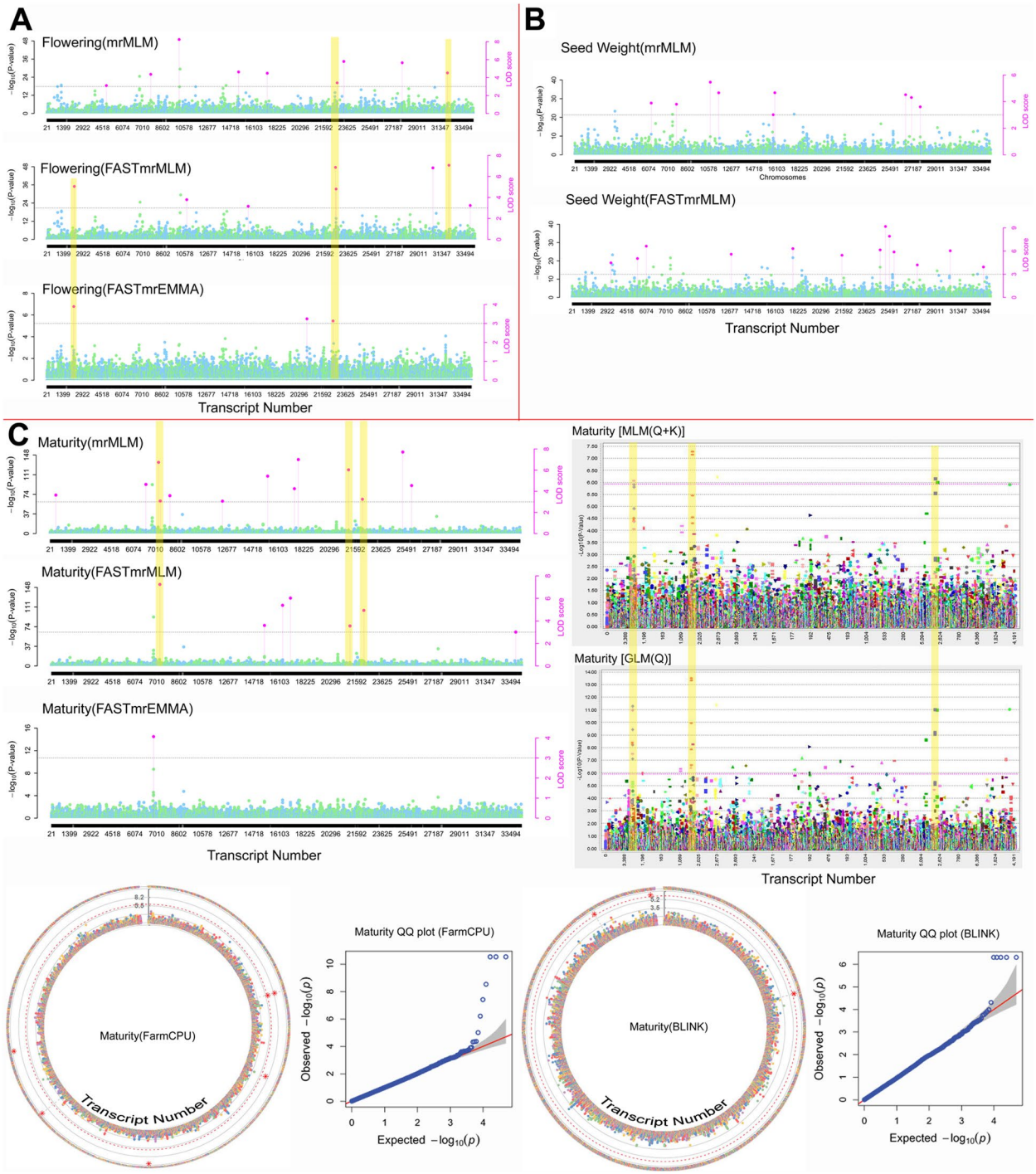

Fig. 6 Significantly associated markers for (**A**) flowering, (**B**) seed weight and (**C**) maturity shown on the Manhattan plots, predicted by various models using the phenotypic dataset from Almora in 2020. The yellow colour highlighted regions show the consistent markers predicted by at least two models. The Manhattan plots are not generated for the markers that are predicted by ISIS EM-BLASSO

E3-ubiquitin-protein-ligase-RHG1A have been identified. Additionally, two genes, WRKY1 and DEAD-box-RH27 were found to be associated with only seed weight, while the association of PIF3 [having Basic Helix-Loop-Helix (bHLH) motif] and pentatricopeptide-repeat-containing-gene with maturity and seed weight and aldo–keto-reductase with flowering and maturity were identified.

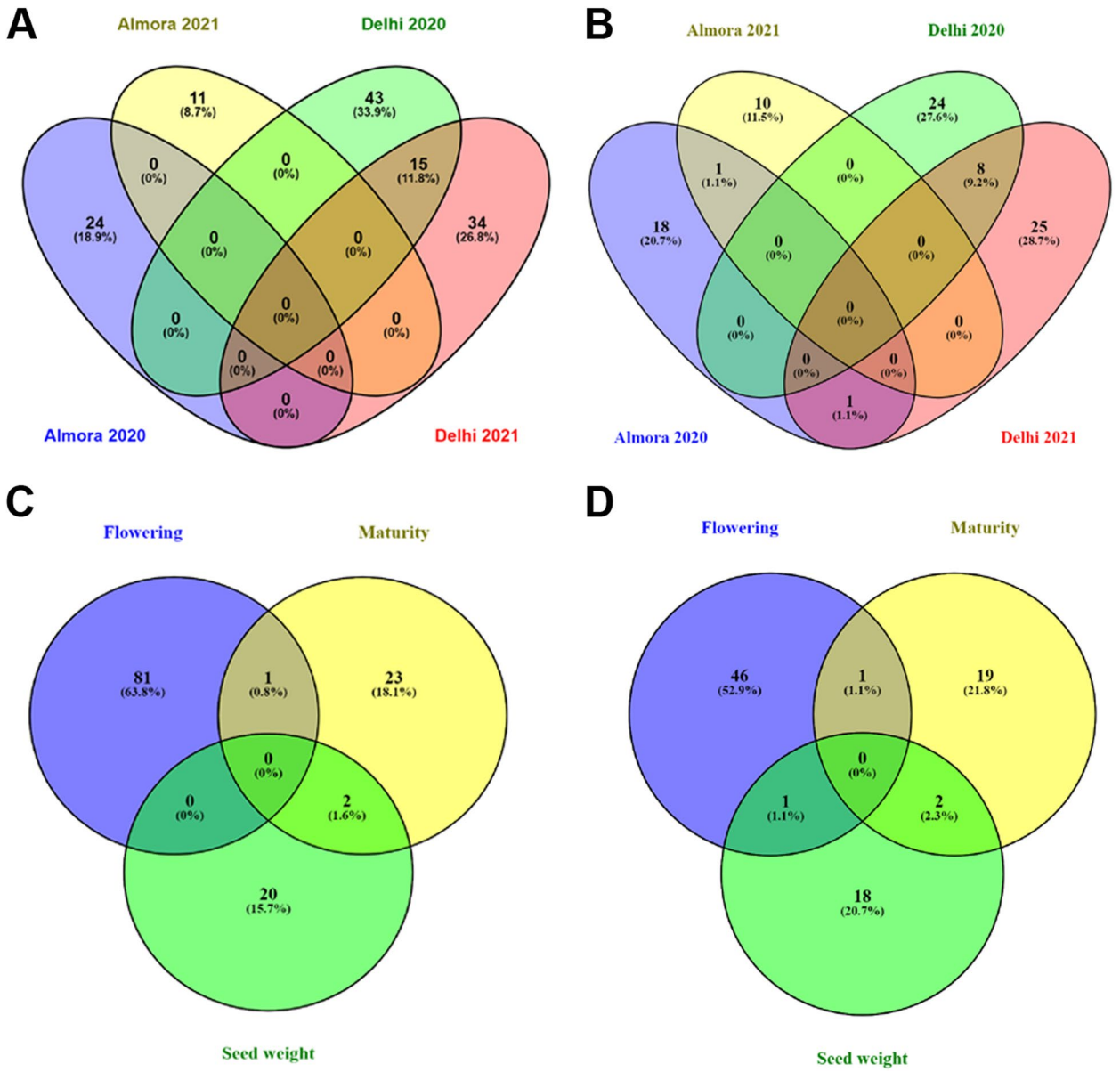

Fig. 7 Common markers and transcripts identified for different traits and different datasets. (**A**) Common markers for datasets, (**B**) common transcripts for datasets, (**C**) common markers for traits, (**D**) common transcripts for traits

Chromosomal localization of markers

The exact chromosomal positions of the associated markers were revealed from the ungapped alignment between marker sequences (101 bp; SNP at 51st position) and chromosomal sequences of the recently released rice bean cultivar FF25. The trait-wise associated markers mapped onto different chromosomes of rice bean is shown in Fig. 9. The associated markers were found distributed over all the chromosomes of rice bean. Most of the markers associated with flowering traits are found on chromosome 1. The highest number of markers for maturity was found on chromosome 11, whereas the highest number of markers for seed weight was found on chromosome 5. On chromosome 9, only two markers for flowering were identified. A set of 35 markers were identified on chromosome 1 within a distance of 80.83 kb at 41.6 Mbp to 42.4 Mbp.

Inter-species synteny based on chromosomal localization of associated transcripts

The associated transcripts, when mapped onto the recently released genome of rice bean [13]; with > = 95%

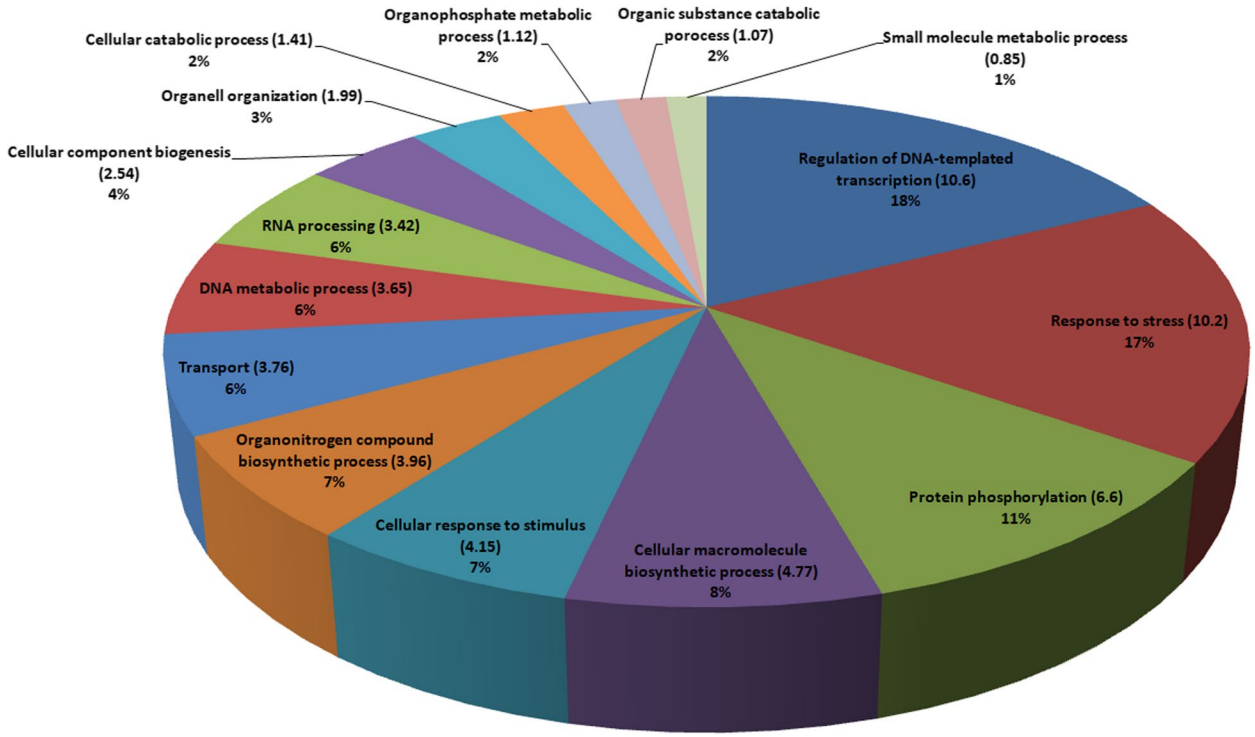

Fig. 8 The major gene ontology terms for biological processes obtained as annotations for 87 transcripts

query coverage, all the associated transcripts were mapped onto the chromosomes (Table S6). 81 out of 87 transcripts were identified with 100% percentage identity and > = 96% query coverage. Further, the mapping of the 87 transcripts onto the chromosomes of different *Vigna* species revealed their inter-species chromosomal locations, which assisted in establishing a synteny between all the considered *Vigna* genomes in terms of the trait-associated transcripts (Fig. 10). The pairwise synteny between *V. umbellata* and other *Vigna* species revealed the inter-species chromosomal syntenic relationship. Based on these pair-wise relationships with *V. umbellata,* the synteny between other *Vigna* species was also established (Fig. 10).

Expression analysis of associated transcripts

The expression pattern of genes represented by 87 transcripts for the stages like inflorescence, 5 days post anthesis and 10 days post anthesis is given in the Fig. 11. Most of the associated transcripts for flowering were found enriched with the reads from samples taken at inflorescence and five days post anthesis stages. However, the few transcripts associated with maturity and seed weight were also observed to be enriched with the reads from all three stages, suggesting their expression during the entire process of flowering to maturity. Among the top 10 expressed transcripts across all three stages, five transcripts associated with flowering were found to encode the proteins such as heat shock cognate protein 80 (HSC80), Photosystem II PsbX (P-II PsbX), plasma membrane ATPase 4, photosystem II stability/assembly factor HCF136, and 40S ribosomal protein S19-1.

Discussion

Rice bean, being an underutilized crop with high nutritional value, bears the potential to contribute to the food and nutritional security across the globe [1, 7, 8]. However, its importance is being recognized, and several research projects have been undertaken for the genetic improvement of this crop. The genomes released by the Chinese Academy of Agricultural Sciences, Beijing, China [13] at the chromosomal level and the International Centre for Genetic Engineering and Biotechnology, New Delhi, India, at the scaffold level [15] have increased the scope of genomic research on rice bean. The phenotypic data for 100 rice bean varieties considered here, collected from two locations in two consecutive years, was found significantly different for most of the traits (Fig. 2). The significant difference in the phenotypic data is expected for different locations as Almora has a hilly topography and receives high rain, whereas Delhi has a plain topography and has much less rain during the crop season. However, the difference in the data during consecutive years in the same location could be due to

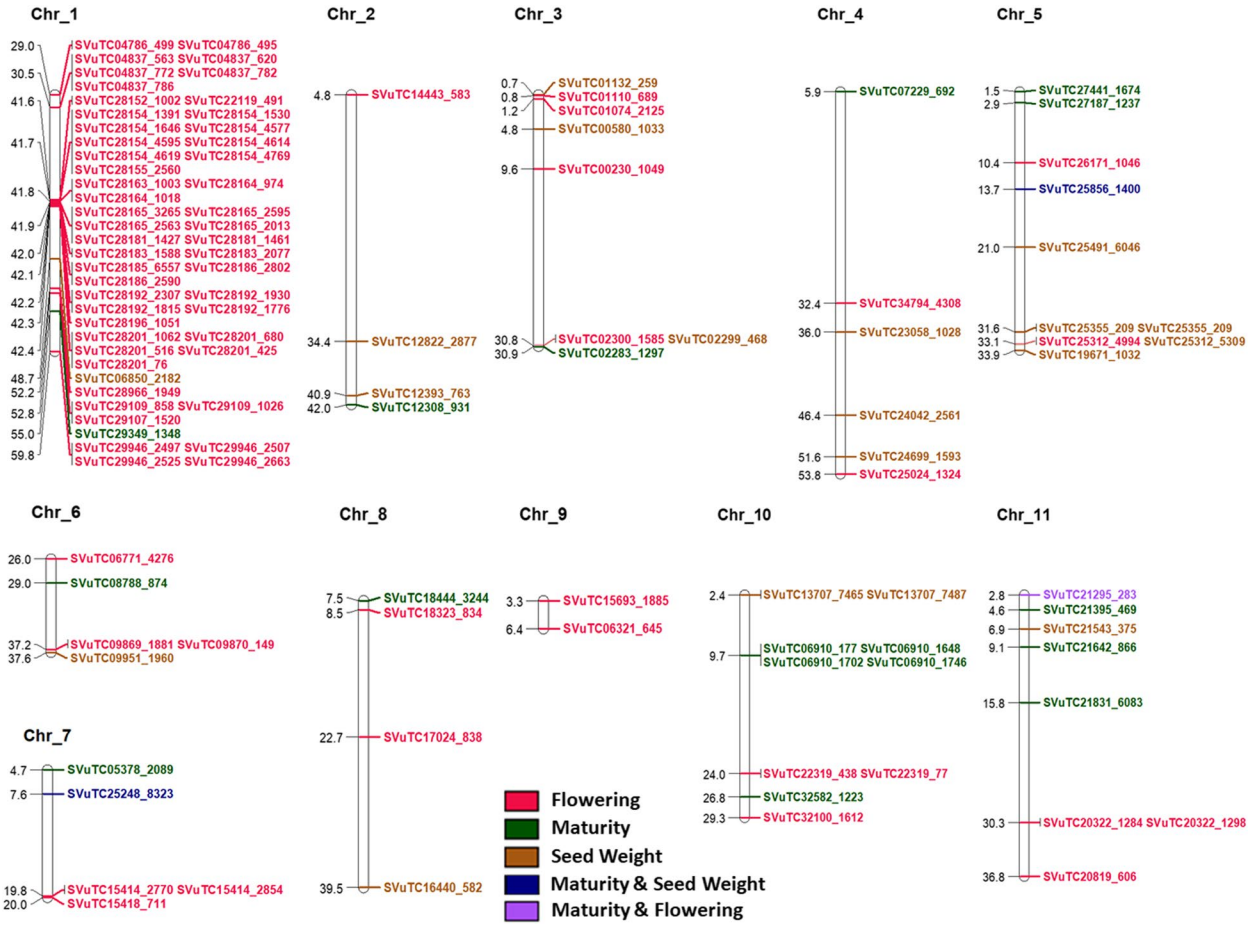

Fig. 9 The trait-wise associated markers mapped onto different chromosomes of the rice bean cultivar FF25

unpredictable climate change at these two locations in consecutive years.

Information on population structure in GWAS analysis is often incorporated to avoid false marker-trait associations. Q-matrix generated through STRUCTURE and PC scores from PCA help infer the population structure from the genotypic data in GWAS analysis [46]. STRUCTURE and PCA indicated the classification of genotypes into three subpopulations (sub-population 1–3) that have also corroborated with the genotypic cluster derived through TASSEL (Fig. 4). Sub-population 1 contains 53 genotypes with 80 days on an average to 50% flowering, 123 days on an average to 80% maturity and an average 100-seed-weight of 6.96 g (Table S8). Further, sub-population 2 was noticed to have 32 genotypes with average values of 74 days, 116 days, and 6.61g for 50% flowering, 80% maturity and 100 seed weight, respectively (Table S8). Furthermore, sub-population 3 was observed to have a group of 15 genotypes with average values of 69 days, 106 days and 6.54g for 50% flowering, 80% maturity and 100-seed-weight, respectively (Table S8). This implies that sub-population 1 contains mostly the late flowering and late maturity genotypes, whereas sub-populations 2 and 3 contain most of the genotypes of early and normal flowering types.

In cross pollinated crops LD decays at short distances as compared to self-pollinated crops [47, 48]. Rice bean is a highly cross-pollinated crop, thus, we have also observed decay of LD at short distances. The LD decay at an r^2 cutoff of 2% [49, 50] was observed to be 1.5 Kb (Fig. 5), suggesting a high genetic diversity in the genotypes taken into consideration. Further, the LD decay at short distances can also be expected with the transcript data as mature mRNA lacks introns. Thus, markers identified through GWAS within a distance of 1.5 kb are expected to be associated with similar or related traits and are expected to be inherited together with a little chance of contemporary recombination. Several GWAS models exist with their own advantages and disadvantages. We implemented ten different models and selected markers on the basis of their threshold scores. The markers were further screened if they are predicted by at least

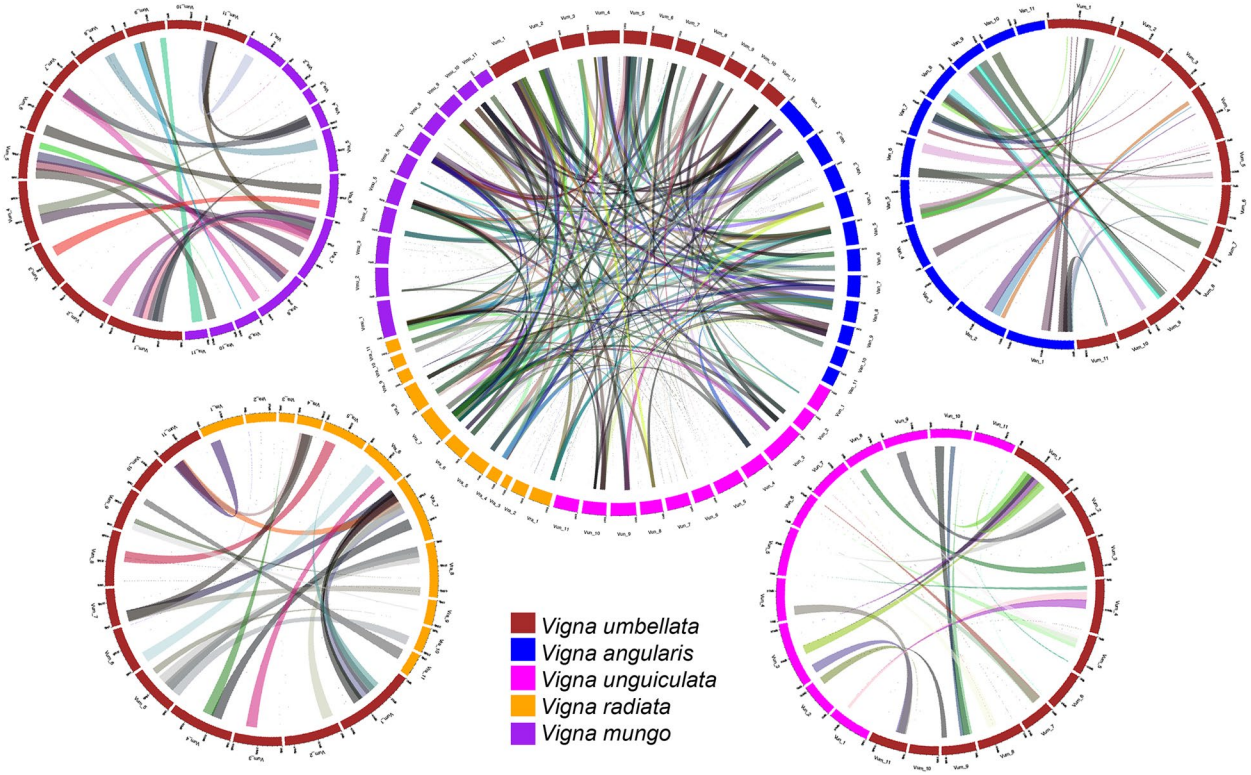

Fig. 10 Synteny between all the considered *Vigna* genomes in terms of the trait-associated transcripts along with pair-wise synteny between *V. umbellata* and other *Vigna* species

two models. Our approach was to minimize the false positive as well as not to discard biologically important markers even if they were not predicted by high number of methods.

In spite of applying several models, we could not find any marker for any trait found consistently with all datasets. However, we observed few markers and corresponding transcripts consistent with the two datasets (Fig. 7). It can be expected as that two locations included in our study, Almora and Delhi differs in their topography, latitude, longitude and climate conditions, causing large differences in the phenology and related traits and thus necessitating the involvement of different set of genes at these two locations. This assertion is supported by the findings of Guan et al. [13] who reported distinct association signals for the different locations. They conducted GWAS analysis for various traits including *i.e.,* flowering and seed weight, which were considered in our study. Two of their reported markers, Chr6: 25.86 Mb and Chr10: 28.15 Mb have shown association with flowering. We have also observed the association of two markers (SVUTC06771_4276: Chr6: 26.03 Mb and SVUTC32100_1612: Chr10: 29.31 Mb, respectively) for flowering within the vicinity (1.5 Mb) of their reported markers. Additionally, we found four markers (SVUTC06850_2182: Chr1: 48.73 Mb, SVUTC25856_1400: Chr5: 13: 66 Mb, SVUTC09951_1960: Chr6: 37.65 Mb, SVUTC25248_8323: Chr7: 7.58 Mb) for seed weight nearby the four markers (Chr1: 48.26 Mb, Chr5: 14.78 Mb, Chr6: 36.35 Mb, Chr7: 6.14 Mb, respectively) reported by Guan et al. [13] for seed weight. However, two of these markers (SVUTC25856_1400: Chr5: 13.66 Mb, SVUTC25248_8323: Chr7: 7.58 Mb) identified in our study were also found to be associated with maturity trait.

The chromosomal localization of the genes revealed the flowering genes on chromosome 1 (Fig. 9), and the role of chromosome 1 in controlling the flowering time has been reported in common bean [51–53]. It is worth mentioning here that chromosome 1 of cow pea corresponds to chromosome 1 of common bean [54] and chromosome 1 of rice bean [13], which indicates the correspondence between the common bean and rice bean genome in terms of chromosome 1. The inter-species syntenic relationship (Fig. 10) based on associated transcripts revealed the higher closeness of rice bean with *V. angularis*, *V. mungo* and *V. radiata* than *V. unguiculata*. The same has also been depicted in Guan et al. [13] and Pattanayak et al. [7].

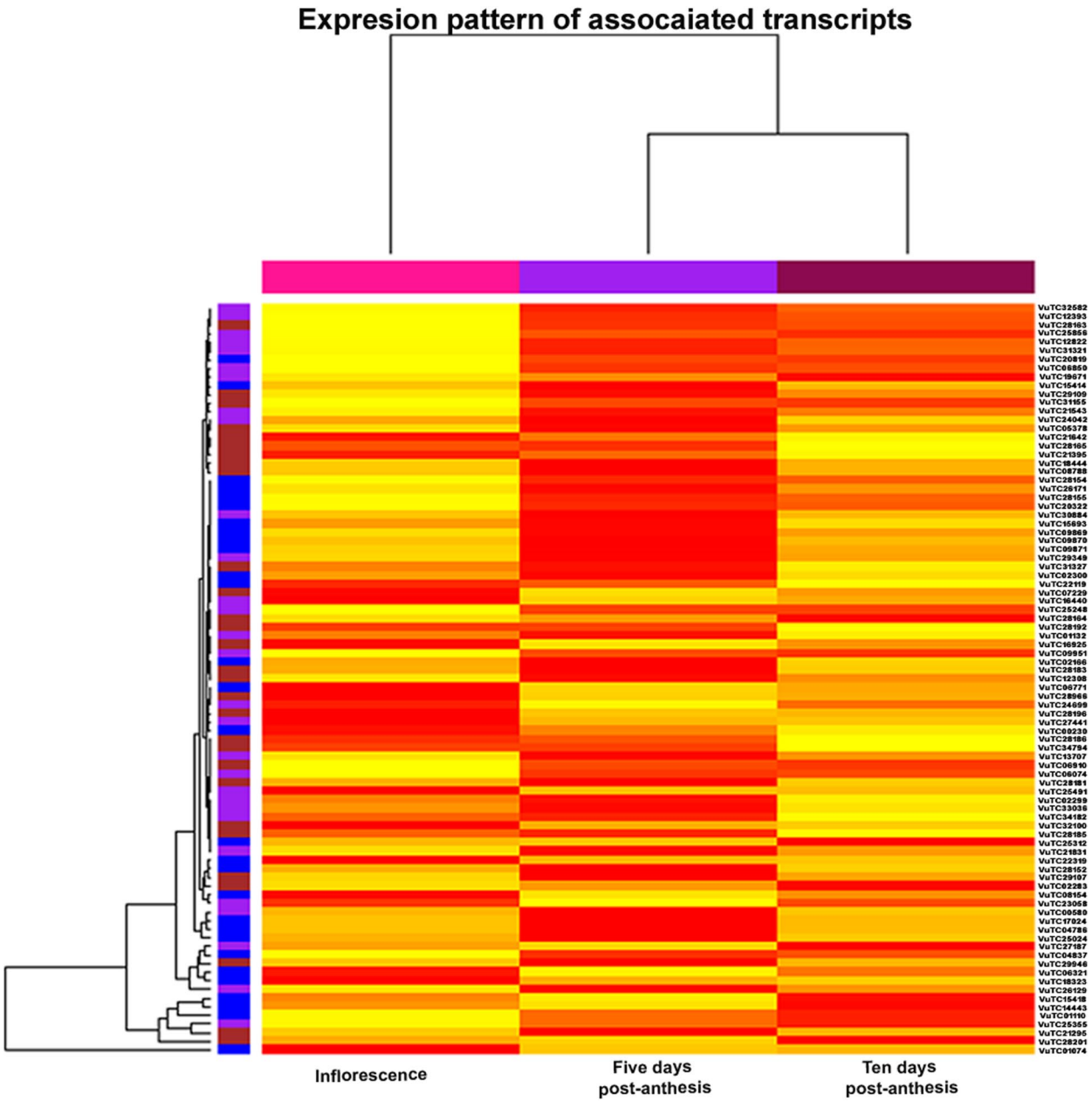

Fig. 11 The expression pattern of genes represented by 87 transcripts for the stages like inflorescence, 5 days post anthesis and 10 days post anthesis

Among the identified markers, SVUTC25856_1400 associated with maturity and seed weight traits located on the transcript (VUTC25856) that has shown similarity with transcription factors bHLH149, bHLH147 and PIF3. The bHLH transcription factors regulate the expression of the genes related to biosynthesis, metabolism and transduction of plant hormones [55]. The phytochrome-interacting factor (PIF3) belongs to bHLH family of proteins that has a specific function in light-induced developmental processes [56]. During seed maturation in light conditions, the amount of abscisic acid in seeds and sensitivity of seeds to abscisic acid is influenced [57] that has an important role in the biosynthesis of storage compounds in the embryo, seed dormancy, and the inhibition of precocious germination [58–60]. Besides, [61] reported that mutants of *Arabidopsis* deficient with abscisic acid produced seeds with increased size, mass, and embryo cell number. Thus, transcript (VUTC25856)

similar to PIF3 with bHLH domain is expected to regulate the abscisic acid-dependent behavior of seeds during light-induced developmental processes.

A transcript VUTC21295 containing the marker SVUTC21295_283 associated with maturity and flowering was found to have aldo–keto reductase function. The aldo–keto reductase enzymes reduce carbonyl substrates like sugar aldehydes, keto-steroids and keto-prostaglandins [62]. The D-GalUA reductase, an aldo–keto reductase, plays an important role in D-Galacturonic acid pathway for ascorbate biosynthesis in plants [63]. Kotchoni et al. [64] reported that an artificial increase in ascorbic acid delayed the flowering. Delaying in flowering affects the days to maturity. Thus, the transcript VUTC21295 having an aldo–keto reductase function might be playing a regulatory role in ascorbate biosynthesis in rice bean, having an effect on its flowering and maturity.

On the transcript VUTC28154, 8 markers were identified where the marker SVUTC28154_1646 was predicted by 8 out of 10 models. The transcript showed similarity with putative phospholipid-transporting ATPase 9, and the GO annotation revealed its involvement in the phospholipid translocation process. Zhou et al. [65] reported that *phospholipid-transporting ATPase* 9, along with other phospholipases, plays a part in accelerating the pollen tube aging in *Pyrus bretschneideri*. Five markers were identified on the transcript VUTC28201 for flowering. The transcript annotation shows its similarity with pectin acetylesterase 8-like protein having a role in cell wall organization and pectin acetylesterase activity. However, the expression of pectin acetylesterase 8 of *Arabodopsis thaliana* has been reported to be highly regulated during plant growth and development. It is expressed in different flowering parts and stages (pollen, inflorescence meristem, stamen, petal differentiation and expansion stage), suggesting a role in the control of the degree of acetylation of pectins [66, 67].

Among the top 10 expressed transcripts in all three stages (inflorescence, 5 days post anthesis and 10 days post anthesis), 8 transcripts are found common. Out of these eight transcripts, associations of 5 transcripts with flowering, one with both flowering and maturity, one with maturity, and one with seed weight have been revealed. The transcripts associated with only flowering were found to encode heat shock cognate protein 80 (HSC80; VuTC01074), Photosystem II PsbX (P-II PsbX; VuTC14443), plasma membrane ATPase 4 (VuTC29946), photosystem II stability/assembly factor HCF136 (VuTC01110) and 40S ribosomal protein S19-1(VuTC15418). The strong expression of HSC80 in floral shoot apices until six days post anthesis has been observed by Koning et al. [68]. Further, the expression of P-II PsbX [69] and plasma membrane ATPase 4 [70] in the flowers of *Spinacia oleracea* and *A. thaliana,* respectively has also been reported. Meurer et al., [71] have demonstrated that the *A. thaliana* plants with mutant HCF136 deficient in PSII activity were failed to produce flowers. Another transcript annotated as 40S ribosomal protein S19-1, was shown to be differentially expressed by Yan et al., [72] while analyzing the differential expression in anther and stigma of *Eruca sativa* during pre-bloom and after flowering stages. One transcript (VUTC21295) associated with both flowering and maturity, which codes for aldo–keto reductase has been discussed earlier regarding its association with the related traits. Its appearance within the top 10 expressed transcripts in all three stages confirms its association. A transcript VuTC26129 annotated as WD repeat-containing protein 48(WDR48) and associated with maturity appeared within 10 expressed transcripts for all three stages. A WDR48 analog in *A. thaliana* has been demonstrated to interact and activate a deubiquitinase UBP3 [73]. Additionally, plant deubiquitinases are reported to control flowering, embryogenesis, pollen and seed development [74–76]. The transcript VuTC25355 associated with seed weight was annotated as 60S ribosomal protein L18a (RPL18a) and the role of RPL18a in embryo development has been established by Yan et al., [77].

Apart from all the above discussed genes, a few major genes like WRKY transcription factor 1(VuTC01132: seed weight), E3 ubiquitin-protein ligase RHG1A (VuTC02166: flowering), DEAD-box ATP-dependent RNA helicase 27(VuTC12393: seed weight), pentatricopeptide repeat-containing protein (VuTC09951: seed weight, VuTC31321: maturity) were found to contain trait associated markers. A *WRKY1* gene of *Solanum chacoense* has been reported to be involved in the process of seed development, having a specific role during the process of embryogenesis, and was also found to be highly expressed in fertilized ovules at the late torpedo stage in wild potato [78]. Shu and yang, [79] demonstrated the role of E3 ubiquitin-protein ligase RHG1A protein containing RING-H2_TTC3 type domain in flowering time control and light response. Further, the role of deadbox RH27 in the process of seed development in *A. thaliana* has been established by Hou et al. [80] through a mutagenesis experiment. They reported that a recessive mutation in the gene produced shrivelled or wrinkled seed.

The associative transcriptomics approach followed in the present investigation revealed a total of 127 markers on 87 transcripts associated with flowering, maturity and seed weight traits. Besides, 81 markers on 46 transcripts for flowering, 23 markers on 19 transcripts for

maturity, 20 markers on 18 transcripts for seed weight, 2 markers on 2 transcripts for both seed weight and maturity, 1 marker on 1 transcript for both flowering and maturity were found associated. However, 1 transcript was found associated with both flowering and seed weight with different markers for both traits. The functional annotation of transcripts revealed the corresponding gene description, domains, super families and GO terms of all the associated transcripts. Further, the role of the identified genes in influencing the corresponding traits has been corroborated with findings in other crops. The association analysis involving SNPs obtained from transcriptome-based variant calling followed in this study is expected to provide insights into genetic mechanisms governing economically important production traits for various crops.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-04976-y.

Supplementary Material 1.

Authors' contributions

T.K.S: Data analysis, writing the original draft and editing; S.K.V, N.P.S: Data analysis; Gayacharan, D.C.J, S.K.V, BS: Generation of phenotyping data; D.P.W, M.S, R.B, S.K.P, D.C.J, DC, G.P.S: writing, review and editing, A.K.S: conceptualization, writing, review and editing.

Funding

This work was supported by the grants received from the Department of Biotechnology under project: BT/Ag/Network/Pulses-1/2017–18. DC acknowledges J.C. Bose Fellowship (JCB/2020/000014) from Science and Engineering Research Board, Department of Science and Technology.

Availability of data and materials

The transcriptome data used in this study has been submitted to NCBI under the Bioproject accession PRJNA916051 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA916051). The RNA sequenced for two different developing seed stages *i.e.,* 5-days post anthesis and 10-days post anthesis of rice bean (accession: IC426787) has been submitted to NCBI under the accessions SRR16122607 (https://www.ncbi.nlm.nih.gov/sra/SRR16122607) and SRR16122602 (https://www.ncbi.nlm.nih.gov/sra/SRR16122602) respectively.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]ICAR-National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi 110012, India. [2]ICAR-Indian Grassland and Fodder Research Institute, Jhansi, Uttar Pradesh, India. [3]National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi, India. [4]ICAR-Vivekananda Parvatiya Krishi Anusandhan Sansthan, Almora, Uttarakhand, India.

Received: 19 June 2023 Accepted: 2 April 2024
Published online: 08 May 2024

References

1. Asha RK, Koundinya AVV, Das A, Chattopadhyay SB. A review on an underutilised multipurpose legume: rice bean. Acta Hort. 2019;1241:57–64.
2. Dahiphale AV, Kumar S, Sharma N. Rice bean-A multipurpose, underutilized, potential nutritive fodder legume - a review. Journal of Pure and Applied Microbiology. 2017;11(1):433–9. https://doi.org/10.22207/JPAM.11.1.57.
3. Isemura T, Kaga A, Tomooka N. The genetics of domestication of rice bean Vigna umbellata. Ann Bot. 2010;106(6):927–44.
4. de Carvalho NM, Vieira RD. Rice bean (Vigna umbellata (Thunb.) Ohwi et Ohasi). In: Nwokolo E, Smartt J, eds. Food and Feed from Legumes and Oilseeds. Boston: Springer, 1996;222–228. https://doi.org/10.1007/978-1-4613-0433-3-25.
5. Dwivedi GK. Tolerance of some crops to soil acidity and response to liming. J Indian Soc Soil Sci. 1996;44:736–41.
6. NAS.Tropical Legumes: Resources for the Future. National Academy of Science, Report of an Ad Hoc Panel of the advisory committee on technology innovation 1979. https://nap.nationalacademies.org/read/19836/chapter/1#iii.
7. Pattanayak A, Roy S, Sood S. Rice bean: a lesser known pulse with well-recognized potential. Planta. 2019;250(3):873–90.
8. Mohan VR, Janardhanan K. Chemical and nutritional evaluation of two germplasms of the tribal pulse, Bauhinia racemosa Lamk. Plant Foods Hum Nutr. 1994;46(4):367–74. https://doi.org/10.1007/BF01088438.
9. Chandel KP, Joshi BS, Arora RK. Rice bean - a new pulse with high potential. Indian Farming. 1978;28:19–22.
10. Dhillon PK, Tanwar B. Rice bean: A healthy and cost-effective alternative for crop and food diversity. Food Security. 2018;10:525–35. https://doi.org/10.1007/s12571-018-0803-6.
11. Tian J, Isemura T, Kaga A. Genetic diversity of the rice bean (Vigna umbellata) genepool as assessed by SSR markers. Genome. 2013;56:717–27. https://doi.org/10.1139/gen-2013-0118.
12. Smil V. Some unorthodox perspectives on agricultural biodiversity. The case of legume cultivation. Agri Ecosyst Environ. 1997;62:135–44.
13. Guan J, Zhang J, Gong D. Genomic analyses of rice bean landraces reveal adaptation and yield related loci to accelerate breeding. Nat Commun. 2022;13:5707. https://doi.org/10.1038/s41467-022-33515-2.
14. Somta P, Kaga A, Tomooka N. Development of an interspecific Vigna linkage map between Vigna umbellata (Thunb.) Ohwi & Ohashi and V. nakashimae (Ohwi) Ohwi & Ohashi and its use in analysis of bruchid resistance and comparative genomics. Plant Breeding. 2006;125:77–84. https://doi.org/10.1111/j.1439-0523.2006.01123.x.
15. Kaul T, Easwaran M, Thangaraj A. De novo genome assembly of rice bean (Vigna umbellata) – A nominated nutritionally rich future crop reveals novel insights into flowering potential, habit, and palatability centric – traits for efficient domestication. Front Plant Sci. 2022;13:739654. https://doi.org/10.3389/fpls.2022.739654.
16. Gautam R, Kumar N, Yadavendra JP. Food security through rice bean research in India and Nepal (FOSRIN). Report 1. Distribution of rice bean in India and Nepal. Local Initiatives for Biodiversity, Research and Development, Pokhara, Nepal and CAZS Natural Resources, College of Natural Sciences, Bangor University, Wales, UK. 2007.
17. Zhao K, Tung CW, Eizenga G. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun. 2011;2:467. https://doi.org/10.1038/ncomms1467.
18. Xiao Y, Liu H, Wu L. Genome-wide association studies in Maize: praise and stargaze. Mol Plant. 2017;10:359–74.
19. Hong SE, Kneissl J, Cho A. Transcriptome-based variant calling and aberrant mRNA discovery enhance diagnostic efficiency for neuromuscular diseases. J Med Genet. 2022;59(11):1075–81. https://doi.org/10.1136/jmedgenet-2021-108307.
20. Jehl F, Degalez F, Bernard M. RNA-Seq data for reliable SNP detection and genotype calling: interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. Front Genet. 2021;12:655707. https://doi.org/10.3389/fgene.2021.655707.

21. Bowman AW, Azzalini A. R package 'sm': nonparametric smoothing methods 2021. (version 2.2–5.7). http://www.stats.gla.ac.uk/~adrian/sm.
22. Francis A, Singh NP, Singh M, Sharma P, Gayacharan, Kumar D, Basu U, Bajaj D, Varshney N, Joshi DC, Semwal DP, Tyagi V, Wankhede D, Bharadwaj R, Singh AK, Parida SK, Chattopadhyay D. The ricebean genome provides insight into Vigna genome evolution and facilitates genetic enhancement. Plant Biotechnol J. 2023;21(8):1522–1524. https://doi.org/10.1111/pbi.14075.
23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arxiv 2013; arXiv:1303.3997v2 [q-bio.GN].
24. Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.
25. Bradbury PJ, Zhang Z, Kroon DE. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
26. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
27. Earl DA, vonHoldt BM. Structure Harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012;4(2):359–436. https://doi.org/10.1007/s12686-011-9548-7.
28. Price A, Patterson N, Plenge R. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
29. Yu J, Pressoir G, Briggs WH. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–8. https://doi.org/10.1038/ng1702.
30. Gupta PK, Kulwalb PL, Jaiswal V. Association mapping in plants in the post-GWAS genomics era. Adv Genet. 2019;104:75–154. https://doi.org/10.1016/bs.adgen.2018.12.00.
31. Lipka AE, Tian F, Wang Q. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28(18):2397–9. https://doi.org/10.1093/bioinformatics/bts444.
32. Wen YJ, Zhang H, Ni YL. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform. 2018;19(4): https://doi.org/10.1093/bib/bbw145.
33. Wang SB, Feng JY, Ren WL. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep. 2016;6:19444. https://doi.org/10.1038/srep19444.
34. Tamba CL, Zhang YM. A fast mrMLM algorithm for multi-locus genome-wide association 2018;studies. bioRxiv. https://doi.org/10.1101/341784.
35. Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. PLoS Comput Biol. 2017;13(1):e1005357. https://doi.org/10.1371/journal.pcbi.1005357.
36. Liu X, Huang M, Fan B. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 2016;12(2):e1005767. https://doi.org/10.1371/journal.pgen.1005767.
37. Huang M, Liu X, Zhou Y. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience. 2019;8(2):giy154. https://doi.org/10.1093/gigascience/giy154.
38. Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. BMJ. 1995;310(6973):170. https://doi.org/10.1136/bmj.310.6973.170.
39. Zhang YM, Jia Z, Dunwell JM. Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. Front Plant Sci. 2019;10:100. https://doi.org/10.3389/fpls.2019.00100.
40. Voorrips RE. 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. J Hered. 2002;93(1):77–8.
41. Yu Y, Ouyang Y, Yao W. shinyCircos: an R/Shiny application for interactive creation of Circos plot. Bioinformatics. 2018;34(7):1229–31. https://doi.org/10.1093/bioinformatics/btx763.
42. Andrews S. FastQC: A quality control tool for high throughput sequence data 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
43. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
44. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
45. Li H, Handsaker B, Wysoker A. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
46. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. PLoS One. 2014;9(4):e93766. https://doi.org/10.1371/journal.pone.0093766.
47. Dinesh A, Patil A, Zaidi PH. Genetic diversity, linkage disequilibrium and population structure among CIMMYT maize inbred lines, selected for heat tolerance study. Maydica. 2016;61(3):1–7.
48. Yu H, Deng Z, Xiang C. Analysis of diversity and linkage disequilibrium mapping of agronomic traits on B-genome of wheat. J Genomics. 2014;2:20–30. https://doi.org/10.7150/jgen.4089.
49. Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, André I, Duarte J, Gauthier V, Lucante N, Marty A, Pauchon M, Pichon J-P, Ribière N, Trotoux G, Blanchard P, Rivière N, Martinant J-P, Pauquet J. High-density SNP-based genetic map development and linkage disequilibrium assessment in Brassica napus L. BMC Genomics. 2013;14:1–18. https://doi.org/10.1186/1471-2164-14-120.
50. Li X, Han Y, Wei Y, Acharya A, Farmer AD, Ho J, Monteros MJ, Brummer EC. Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. PLoS One. 2014;9:e84329. https://doi.org/10.1371/journal.pone.0084329.
51. Gu W, Zhu J, Wallace DH. Analysis of genes controlling photoperiod sensitivity in common bean using DNA markers. Euphytica. 1998;102:125–32. https://doi.org/10.1023/A:1018340514388.
52. Kwak M, Velasco D, Gepts P. Mapping Homologous Sequences for Determinacy and Photoperiod Sensitivity in Common Bean (Phaseolus vulgaris). J Hered. 2008;99(3):283–91. https://doi.org/10.1093/jhered/esn005.
53. González AM, Yuste-Lisbona FJ, Saburido S. Major contribution of flowering time and vegetative growth to plant production in common bean as deduced from a comparative genetic mapping. Front Plant Sci. 2016;7:1940. https://doi.org/10.3389/fpls.2016.01940.
54. Lonardi S, Muñoz-Amatriaín M, Liang Q. The genome of cowpea (Vigna unguiculata [L.] Walp.). Plant Journal. 2019;98(5):767–82. https://doi.org/10.1111/tpj.14349.
55. Hao Y, Zong X, Ren P. Basic Helix-Loop-Helix (bHLH) Transcription Factors Regulate a Wide Range of Functions in Arabidopsis. Int J Mol Sci. 2021;22(13):7152. https://doi.org/10.3390/ijms22137152.
56. Ni M, Tepperman JM, Quail PH. PIF3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein. Cell. 1998;95(5):657–67. https://doi.org/10.1016/s0092-8674(00)81636-0.
57. Contreras S, Bennett MA, Metzger JD. Maternal light environment during seed development affects lettuce seed weight, germinability, and storability. HortScience. 2008;43:845–52.
58. McCarty DR. Genetic control and integration of maturation and germination pathways in seed development. Annu Rev Plant Physiol Plant Mol Biol. 1995;46:71–93.
59. Finkelstein RR, Gampala SS, Rock CD. Abscisic acid signaling in seeds and seedlings. Plant Cell. 2002;14(Suppl):S15-45. https://doi.org/10.1105/tpc.010441.
60. Kanno Y, Jikumaru Y, Hanada A. Comprehensive hormone profiling in developing Arabidopsis seeds: Examination of the site of ABA biosynthesis, ABA transport and hormone interactions. Plant Cell Physiology. 2010;51:1988–2001.
61. Cheng ZJ, Zhao XY, Shao XX. Abscisic acid regulates early seed development in Arabidopsis by ABI5-mediated transcription of SHORT HYPOCOTYL UNDER BLUE1. Plant Cell. 2014;26(3):1053–68. https://doi.org/10.1105/tpc.113.121566.
62. Penning TM. The aldo-keto reductases (AKRs): Overview. Chem Biol Interact. 2015;234:236–46. https://doi.org/10.1016/j.cbi.2014.09.024.
63. Ishikawa T, Maruta T, Yoshimura K. Biosynthesis and regulation of ascorbic acid in plants. In: Gupta D, Palma J, Corpas F, eds. Antioxidants and antioxidant enzymes in higher plants. Cham: Springer, 2018. https://doi.org/10.1007/978-3-319-75088-0_8.
64. Kotchoni SO, Larrimore KE, Mukherjee M. Alterations in the endogenous ascorbic acid content affect flowering time in Arabidopsis. Plant Physiol. 2009;149(2):803–15. https://doi.org/10.1104/pp.108.132324.

65. Zhou H, Yin H, Chen J. Gene-expression profile of developing pollen tube of Pyrus bretschneideri. Gene Expr Patterns. 2016;20(1):11–21. https://doi.org/10.1016/j.gep.2015.10.004.
66. Philippe F, Pelloux J, Rayon C. Plant pectin acetylesterase structure and function: new insights from bioinformatic analysis. BMC Genomics. 2017;18:456. https://doi.org/10.1186/s12864-017-3833-0.
67. Schmid M, Davison T, Henz S. A gene expression map of Arabidopsis thaliana development. Nat Genet. 2005;37:501–6. https://doi.org/10.1038/ng1543.
68. Koning AJ, Rose R, Comai L. Developmental expression of tomato heat-shock cognate protein 80. Plant Physiol. 1992;100(2):801–11. https://doi.org/10.1104/pp.100.2.801.
69. Shi LX, Kim SJ, Marchant A. Characterisation of the PsbX protein from Photosystem II and light regulation of its gene expression in higher plants. Plant Mol Biol. 1999;40(4):737–44. https://doi.org/10.1023/a:1006286706708.
70. Zhang J, Wei J, Li D. The role of the plasma membrane H+-ATPase in plant responses to aluminum toxicity. Front Plant Sci. 2017;8:1757. https://doi.org/10.3389/fpls.2017.01757.
71. Meurer J, Plücken H, Kowallik KV. A nuclear-encoded protein of prokaryotic origin is essential for the stability of photosystem II in Arabidopsis thaliana. EMBO J. 1998;17(18):5286–97. https://doi.org/10.1093/emboj/17.18.5286.
72. Yan F, Wan-cang S, Jun-yan W. Differential display and expression analysis of self-compatibility associated gene in Eruca sativa. Chin J Oil Crop Sci. 2014;36(5):580–5.
73. Baskerville A, Donahue J, Gillaspy G. Identification of a WD-repeat protein that binds and activates the deubiquitinase UBP3 from Arabidopsis thaliana. Bios. 2020;91(2):90–9. https://doi.org/10.1893/BIOS-D-18-00029.
74. Schmitz RJ, Tamada Y, Doyle MR. Histone H2B deubiquitination is required for transcriptional activation of FLOWERING LOCUS C and for proper control of flowering in Arabidopsis. Plant Physiol. 2009;149:1196–204.
75. Doelling JH, Yan N, Kurepa J. The ubiquitin-specific protease UBP14 is essential for early embryo development in Arabidopsis thaliana. Plant J. 2001;27(5):393–405.
76. Doelling JH, Phillips AR, Soyler-Ogretim G. The ubiquitin-specific protease subfamily UBP3/ UBP4 is essential for pollen development and transmission in Arabidopsis. Plant Physiol. 2007;145:801–13.
77. Yan H, Chen D, Wang Y. Ribosomal protein L18aB is required for both male gametophyte function and embryo development in Arabidopsis. Sci Rep. 2016;6:31195. https://doi.org/10.1038/srep31195.
78. Lagacé M, Matton DP. Characterization of a WRKY transcription factor expressed in late torpedo-stage embryos of Solanum chacoense. Planta. 2004;219:185–9. https://doi.org/10.1007/s00425-004-1253-2.
79. Shu K, Yang W. E3 ubiquitin ligases: ubiquitous actors in plant development and abiotic stress responses. Plant Cell Physiology. 2017;58(9):1461–76. https://doi.org/10.1093/pcp/pcx071.
80. Hou XL, Chen WQ, Hou Y. DEAD-BOX RNA HELICASE 27 regulates microRNA biogenesis, zygote division, and stem cell homeostasis. Plant Cell. 2021;33(1):66–84. https://doi.org/10.1093/plcell/koaa001.

Publisher's Note