

Genome-wide association studies for identification of novel QTLs related to agronomic traits in Horsegram (*Macrotyloma uniflorum*)

Ankita Sharma

CSK HP Agriculture University

Rakesh Kumar Chahota

rkchahota@hillagric.ac.in

CSK HP Agriculture University

Research Article

Keywords: Association mapping, genotype-by-sequencing, marker trait association, population structure, single nucleotide polymorphism, quantitative trait loci's

Posted Date: September 27th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3366692/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

Supplementary File are not included with this version

Version of Record: A version of this preprint was published at Plant Molecular Biology Reporter on February 27th, 2024. See the published version at <https://doi.org/10.1007/s11105-023-01430-5>.

1 **Research article**

2 **Running head**

3 Identification of QTLs in *Macrotyloma uniflorum*

4 **Title**

5 Genome-wide association studies for identification of novel QTLs related to agronomic traits in Horsegram
6 (*Macrotyloma uniflorum*)

7 **Authors**

8 Ankita Sharma¹ and Rakesh Kumar Chahota^{1,*}

9 **Affiliations**

10 ¹ Department of Agricultural Biotechnology, CSK HP Agriculture University, Palampur 176062, India; email:

11 ² Department of Agricultural Biotechnology, CSK HP Agriculture University, Palampur 176062, India; email:
12 rkchahota@hillagric.ac.in

13

14

15 ***Correspondence**

16 Name: Dr. R.K. Chahota

17 Designation: Professor

18 Affiliation: Department of Agricultural Biotechnology

19 e-mail: rkchahota@hillagric.ac.in; Tel.: +919418369800

20

21

Abstract

Horsegram (*Macrotyloma uniflorum*) is an important legume crop well adapted to harsh environments and has significant nutritional value. It is also well recognised due to its nutritional value, health benefits, drought tolerance, and ability to improve soil fertility. It is a versatile crop used in various dishes and is a valuable addition to any diet. However, in-depth molecular analysis must be improved to understand the molecular architecture regulating the essential traits. To achieve this, the study utilized a genome-wide association mapping approach using 88 horsegram accessions genotyped with 55656 single nucleotide polymorphisms (SNPs). A total of 54271 SNPs were mapped to the reference genome. Phenotypic data were collected for four agronomic and yield-related traits Plant height, Days to maturity, days to flowering and number of branches per plant. These traits played a significant role in plant growth, development, protection, and crop yield. The study identified 79 significant SNP-trait associations for the traits under study. The identified SNPs were in 8 chromosomes. The study also identified several novel genomic regions associated with the traits considered, providing new insights into the genetic basis of these essential traits in horse gram. Overall, the study highlights the potential of genome-wide association mapping for identifying genomic regions associated with important agronomic and yield-related traits in horsegram. This could facilitate the development of improved horsegram varieties through marker-assisted breeding programs and overcoming the morphological selection of desirable genotypes in the background of narrow genetic base.

Keywords: Association mapping; genotype-by-sequencing; marker trait association; population structure; single nucleotide polymorphism; quantitative trait loci's

1. INTRODUCTION

Horsegram (*Macrotyloma uniflorum*) is considered one of the most important and least exploited legume crop of the Fabaceae family ($2n=2x=20$). It is native to Southeast Asia and tropical Africa (R. K. Chahota, Sharma, Sharma, Kumar, & Rana, 2013). The center of origin of the cultivated species is reported in southern India ("Britannica, The Editors of Encyclopaedia. "Nikolai Vavilov". Encyclopedia Britannica, 22 Jan. 2023, <https://www.britannica.com/biography/Nikolai-Vavilov>. Accessed 29 July 2023.," ; Guo, Chen, Li, & Cowling, 2014). In India, it is cultivated in Karnataka, Maharashtra, Andhra Pradesh, West Bengal, Himachal Pradesh, Chhattisgarh, Orissa and the foothills of Uttar Pradesh. The total area under horsegram cultivation in India (2012-2015) is 232 thousand hectares, and production is 105 thousand tones Karnataka state contributes 25.71% in production, following Odisha and Chhattisgarh. It is a herbaceous plant with an intermediate growth pattern is adapted to warm climatic conditions and is highly drought tolerant. Horsegram is considered rich source of proteins (18.5 to 28.5% in seeds), dietary fiber, and micronutrients. Besides its nutraceutical importance, horsegram is known to have immense medicinal properties due to the presence of non-nutritive bioactive compounds (Suwarna, Vs, & Shinde, 2019). The United States National Academy of Sciences has recognized this crop as an important food source for future (Tropical Legumes: Resources for the Future, 1979). Unfortunately, in horsegram limited scientific knowledge has been created in comparison to other pulses like *Vigna radiate* and *Cajanus cajan*, however, in the last decade with the advent in the next generation sequencing technology, few efforts have been made to understand important traits in horsegram of agronomic importance (Bhardwaj, et al., 2013; Kaldate, et al., 2017; Mahesh, et al., 2021; Tanin, et al., 2022; Tropical Legumes: Resources for the Future, 1979).

In the last decade, advent in next generation sequencing technologies has ameliorated the modern molecular breeding efforts. Advance technology like Genotype-By-Sequencing (GBS) which is also cheap and cost effective is now used widely for identification of genome-wide high throughput SNPs (Single Nucleotide Polymorphism) which are further utilized for diversity analysis, linkage and association mapping studies (Wang, et al., 2020). Furthermore, Marker Trait Association (MAS) studies are promising strategies and has been widely utilized to understand the underlying genetic architecture of important agronomic traits like grain yield and consensus abiotic stress tolerance in wheat, salinity tolerance in rice, heat stress in tomato, seed quality in Pea, forage quality in auto-tetraploid alfalfa, and biotic stress in pepper (Alsamir & Hameed, 2019; Gali, et al., 2019; Lin, et al., 2020; Liu, et al., 2019; Mulugeta, et al., 2022; Parisi, Alioto, & Tripodi, 2020; Tanin, et al., 2022)-16]. The Genome-Wide Association Studies (GWAS) or the association mapping in the crop plants is very popular as it offers high mapping resolution and evaluates number of alleles rapidly (Korte & Farlow, 2013; Zhu, Gore, Buckler, & Yu, 2008). GWAS have been carried out to identify QTLs related to drought tolerance using 96 diverse Indian accessions, phenological (reproductive period, days to maturity and days to 50% flowering) and morphological (plant height, number of primary and secondary branches) and yield related traits using RILs population (Rakesh Kumar Chahota, et al., 2020; Choudhary, Isobe, & Chahota, 2022; Katoch, Mane, & Chahota, 2021).

Despite one effort with conventional QTL mapping technique, wherein QTLs related to plant height (1) and primary branches (3) and secondary branches (3) has been reported in the horsegram using bi-parental segregating population (RILs). However, no efforts have been made to generate accurate genotypes with diverse panel to characterize and understand their genetic potential and associate them with agronomically important

traits. Moreover, low genetic potential of most of the released variety and scarcity of genomic resources are among the major bottlenecks which hinders the systematic molecular breeding programme in horsegram. Hence, in current study, a diverse panel of 88 accession collected from 2000 Indian horsegram collections maintained at NBPGR (National Bureau of Plant Genetic Resource) Delhi (Choudhary, et al., 2022) were used to identify genetic loci's related to plant height, number of branches per plant, days to flowering and days to maturity using SNPs markers. Furthermore, population structure and Linkage Disequilibrium (LD) was also evaluated for identification of genomic basis of above-mentioned traits. Overall the outcome of the current study will be a first step towards dissecting complex quantitative traits and accelerating varietal improvement in *Macrotyloma uniflorum*.

2. Results

2.1. Genotyping using GWAS panel

Using the Illumina *HiSeqTM* 2500 platform, 88 horsegram accessions were sequenced, which produced 160637255 reads of 151bp with minimum of 666565 and maximum of 8064347 reads per accessions. After deleting low-quality sequences (Quality scores < 20) and filtering out the remainder, 145623560 (90.65%) good-quality reads were left. On average 80% of the reads per accession were mapped to the reference genome (*Medicago truncatula*) using these high-quality reads. Additionally, 55656 high-quality SNPs (localised to the chromosomes) were used for various downstream analyses, such as population structure and marker-trait association analysis in *Macrotyloma uniflorum*, after SNPs were filtered based on MAF (Marker Allele Frequency) of > 0.05 and a minimum depth of reads with 5. A total of 54271 SNPs were mapped to the chromosomes and the highest number of SNPs were found on chromosome 4 (9368), followed by chromosomes 1 (8653), 3 (7752), 8 (6951), 7 (6930), 2 (6792), 5 (5896), and 6 with the fewest SNPs (1929) (Figure 1 & Supplementary file 3). The remaining 1385 SNPs were identified to the respective scaffolds but were not assigned to any of the chromosomes.

The variant calling resulted in transitions (Ts; AG:15277 and CT:15079) were more prevalent among these SNPs than transversions (Tv; GT: 4655, CG:6184, AT:7686, and AC:4742) (Figure 2). Most SNPs were localized to the exons, upstream and downstream regions of the 5' UTR and 3' UTR (untranslated regions). It showed more than 20 % SNPs in exon and downstream regions whereas in upstream regions 16-18% SNPs was found.

2.2. Genetic diversity analysis

STRUCTURE version 2.3.4 was used to evaluate the genetic makeup of 88 accessions of the horsegram, and the model-based maximum likelihood ancestry technique was applied. In the 250Kb window, 1439 high-quality SNPs were chosen based on read depth and the minimal amount of missing data. The most probable clusters (K) were examined between K2 and K10, and K value 3 was ultimately chosen to represent the genetic makeup of 88 horsegram accessions. POP1, POP2, and POP3 were discovered to be the three largest subpopulations in each cluster. Additionally, high-quality filtered SNPs were used for the phylogenetic and PCA analysis. According to the PCA analysis the clustering indicated three groupings based on similarities, genetic variance was assessed for each accession [(PC1; 94.2%) and (PC2; 1.35%)] (Figure 3). The phylogenetic analysis was also done using TASSEL version 4 by neighbour-joining (NJ) method, forming 3 clusters. Two accessions (IC-416486 and TCR-491) were placed in cluster 1, four accessions (TCR 1351, HPKM 317, IC 278828, and HPKM 151) were grouped in cluster 2, and the remaining 82 accessions were grouped in cluster 3.

Based on the SNPs found in the current investigation, the Pearson correlation coefficient (r^2) was calculated between various marker pairs to predict the rate of LD decay. For each chromosome individually and the entire genome, the physical distance of the decay was calculated in Kb, and variance in the LD decay was noted. The LD decay ranges for chromosome 1 were observed between 25 and 220 Kb, indicating that the greatest r^2 value was at the beginning (25 Kb) and steadily decreased. The LD decay ranges were also found 15–225 Kb for chromosome 2, 20–123 Kb for chromosome 3, 18–151 Kb for chromosome 4, 30–120 Kb for chromosome 5, 12–300 Kb for chromosome 6, and 8–121 Kb for chromosome 8. (55 - 260 Kb). Ultimately, it was discovered that the complete genome's LD decay ranged from 48 to 200 Kb. Chromosome 6 (12–300 Kb) showed the highest decay rate, and it was noted that while r^2 values increase at the beginning of a few bases, they gradually decline in some chromosomes as they get farther apart. Also, the distances between the polymorphic loci for each SNP were estimated, and the graphs for all eight chromosomes were displayed based on the colour for the long-range in the base pair (r^2) (Figure 4). Overall, research showed that essentially no linkage was retrieved in the 0.6-0.8 SNPs range, whereas most SNPs identified as connected were in the range of 0.8-1.0.

2.3. Phenotyping and correlation analysis of traits

Multiyear phenotypic data were collected for GWAS panel for the traits namely; plant height, days to flowering, days to maturity, number of branches per plant. Significant variation among all the traits was recorded (Supplementary Table 2 and figure 5). Plant height was recorded maximum (207 cm) and minimum (43 cm) for lines TCR-591) and HPKM-317 respectively. Furthermore, sufficient variation in a range of (26-91) days was recorded for days to flowering. The minimum days for maturity of the plant were recorded 52 and maximum 93 days were observed for days for maturity. A sufficient phenotypic variation was recorded for the traits used in the present study which is required for dissecting association among phenotypic and genotypic traits (Figure 5). Correlation analysis was carried out among the traits to reveal the relationship of one trait with another at $p < 0.05$ (Figure 6). Plant height is found to be positively correlated with days to flowering, days to maturity and number of branches per plant. Days to flowering is positively correlated with plant height and days to maturity. Interestingly, days to flowering is not positively correlated with number of branches per plant but it is in positive correlation with all other traits. Overall we can conclude that all the traits are positively correlated and any variation in one trait directly or indirectly will affect the other. Plant height and number of branches per plant have been reported directly correlated with seed yield previously (Neelam et al. 2014).

2.4. GWAS analysis for traits

In the current study, mixed linear model was used to study the four crucial horsegram agronomic parameters and efforts were made to pinpoint the key loci. The association study's threshold was $1e-04$ for the 88 accessions (Figure 7). Using the MLM that connects for the effects of population structure and genetic relatedness. The total number of marker trait associations (MTAs) for four traits (plant height, Number of branches per plant, days to flowering, days to maturity) were found to be 79. Interestingly the maximum number of associations were found to be 37 for number of branches per plant distributed on *chr1*, *chr2*, *chr4*, *chr5*, *chr7* and *chr8*. The significant marker trait associations found for days to flowering were (20). With respect to days to maturity, total of 7 MTAs was recorded and markers were found distributed on *chr1*, *chr2*. With maximum heritability recorded for plant height the number of significant marker trait associations were 15. (Supplementary Table4).

Discussion

GWAS provides a solid foundation for unlocking plant phenotypes' genetic characteristics at molecular level. The major advances of GWAS study are the use of natural populations as research materials which reduces the research time, high efficiency, high resolution, and simultaneous detection of multiple alleles at the same locus, greatly reducing the costs of gene sequencing. GWAS is a powerful tool for identifying genetic variations contributing to complex plant traits. GWAS allows researchers to identify genetic markers associated with important features in plants, such as yield, disease resistance, and stress tolerance. This can help plant breeders develop new varieties with desirable traits (Atwell, et al., 2010; Choudhary, et al., 2022). These genes can be further studied to understand their function and potentially used for genetic engineering to develop improved varieties of horse gram.

In the current study, GBS analysis was carried out in 88 accessions of horsegram and significant SNPs were identified for regulating four traits such as Plant height, days to flowering, days to maturity and number of branches per plant. The type of maturity plays important role in crop yield. Crops that mature earlier tend to have a greater yield potential than those that mature later, as they can be harvested earlier and are less susceptible to adverse weather conditions or pest and disease pressure. Saxena et al., (2020) also identified the positive contribution of days to 50% flowering, days to 75% maturity, primary branches per plant and secondary branches per plant on yield in pigeon plants (Saxena, et al., 2020).

The GBS study resulted in 54271 high-quality reads after mapping with the reference genome. As GBS is a powerful and advanced tool, it has been used in many crops for genetic diversity characterization of the germplasm, marker-trait associations, marker-assisted breeding, and population structure analysis (Alipour, et al., 2017). After doing GBS in horsegram, 54271 SNPs were mapped and used in the genetic diversity study which helped identify its localization in 8 chromosomes. Several studies have used GBS to identify SNPs in plants, including wheat and rice (Spindel, et al., 2013). Akbari *et al.* (2012) identified 47,000 SNPs using GBS in maize (Akbari, et al., 2006), and Spindel *et al.* (2013) identified over 13 million SNPs using GBS in rice (Spindel, et al., 2013). Comparing the number of identified SNPs between these studies, it is evident that the number of SNPs identified through GBS can vary widely, likely due to differences in the genotyping platform used, sequencing depth, and bioinformatic pipelines employed. Additionally, the number of SNPs identified can be influenced by the genome size, complexity, and ploidy level of the plant species studied. Despite these differences, GBS remains a useful method for identifying large numbers of plant SNPs, which can be used for various downstream applications such as association mapping, marker-assisted selection, and genetic diversity analysis. Transitions are widespread due to the synonymous SNPs and mutations that do not disrupt protein-coding sequences are more prevalent in the Horsegram genome, this suggests that these are tolerated more under natural selection (Delfini, et al., 2021). Also, different chromosomes showed various LD values, with chromosome 6 recording a longer LD (288 Kb). A recent study by Chahota and colleagues (2022) investigated the patterns of LD in a collection of 95 horsegram accessions from diverse geographic regions in India (R. Chahota, 2020). The authors used genotyping-by-sequencing (GBS) to obtain high-density single nucleotide polymorphism (SNP) data for the accessions and then calculated LD measures, such as r^2 , for pairs of SNPs across the genome. They found that LD decayed rapidly with distance, with r^2 values dropping to half of their maximum values within 80-100 kb. LD analysis has been widely used to study the genetic diversity and population structure of various plant species, including rice (Xu, et al., 2011), and grape (Lijavetzky, Cabezas, Ibáñez, Rodríguez, & Martínez-Zapater, 2007). For example, Xu *et al.* (2017) conducted a genome-wide LD analysis in rice and identified significant LD between genes involved in

abiotic stress response and domestication traits, providing insight into the genetic basis of rice adaptation and domestication (Xu, et al., 2011). Zhao *et al.* (2019) performed a population-wide LD analysis on tomatoes. They found that domestication and breeding have reduced LD levels, resulting in the loss of beneficial alleles and potentially limiting future breeding progress. In grape, Lijavetzky *et al.* (2006) analysed LD between simple sequence repeat (SSR) markers and identified genomic regions associated with important traits such as fruit quality and disease resistance (Lijavetzky, et al., 2007). LD analysis is a powerful tool for understanding plant species' genetic diversity and evolutionary history and can be used to guide breeding and conservation efforts.

Additionally, population STRUCTURE analysis indicated three sub-populations (K=3), which have already been observed in another work in horse gram, to comprehend and examine the false positives in the association analysis. Moreover, PCA (Principal Component Analysis) and phylogenetic analysis corroborated these findings (Figure 4). In another study by Sharma *et al.*, in 2020 (Sharma, Sharma, Rana, & Chahota, 2015), amplified fragment length polymorphism (AFLP) markers were used to genotype 91 accessions of horsegram and then calculated several diversity measures, such as expected heterozygosity, observed heterozygosity, and Shannon's diversity index. They found that the accessions exhibited moderate to high levels of genetic diversity, with expected heterozygosity values ranging from 0.192 to 0.293. In their study, the STRUCTURE and PCA analysis grouped the accessions into three genetically distinct clusters, which corresponded to their geographic origins. This suggests that geographic isolation and natural selection have played a role in shaping the genetic structure of horsegram populations. It indicated the importance of population studies to ensure that the observed genetic associations are not simply due to differences in population structure or ancestry (Crawford, et al., 2015). The population study helps to understand that associations are robust and not confounded by population structure or ancestry. They also help identify genetic variants unique to specific populations and can provide insights into the evolutionary history of these variants.

Horsegram is an important legume crop with several morphological traits that play a crucial role in its adaptation and yield potential. A study on horsegram indicated a high degree of genetic variability for most of the traits such as days to 50% flowering, plant height, number of branches per plant. It also revealed a positive and significant correlation between seed yield per plant and plant height ($r=0.409$) (Toppo, et al., 2019). In another study on *Sesamum indicum* L., there was a positive correlation between seed yield and the number of branches per plant ($r=0.53$). They also reported a positive correlation between oil content and the number of branches per plant ($r=0.26$), number of capsules per plant ($r=0.32$), and number of seeds per capsule ($r=0.41$) (Mashilo, Shimelis, & Odindo, 2016). Thus, a correlation study is an important approach to finding out key traits to be focused on for the development of crops. These traits are very important for crop that are positively correlated with yield in horsegram, including plant height, number of branches per plant, number of pods per plant, and number of seeds per pod. This study indicated the traits have a significant correlation and are important for the development of traits. In the present study Presence of high heritability for the traits viz., number of branches per plant, days to flowering, days to maturity indicate that these traits have minor impact of the environment. Therefore studying these traits for marker trait association will augment the available genomic resources this crop.

3. Methods

3.1. Plant Material, phenotyping and statistical analysis

In the current study, 88 diverse accessions of horsegram were selected from 600 available germplasm of from different states of India (Tamil Nadu, Andhra Pradesh, Himachal Pradesh, Madhya Pradesh, Chhattisgarh,

Jharkhand, Bihar, Punjab, Odisha, Rajasthan, Telangana, and Uttarakhand), and one accessions was from Australia (Supplementary Table 1). Majority of the accession are from India and origin of these accessions is mainly from southern India. These accessions were planted in the experimental area of Department of Agriculture, CSK HPKV Palampur, Himachal Pradesh, India (32.1109° N, 76.5363° E) and also at the Hill Agricultural Research and Extension Centre, Bajaura, Himachal Pradesh, India (31.8465° N, 77.1605° E) following the Random Block Design with three replications. In each line of 1.5 meters, 25-30 plants were planted with a space of 40 cms between lines and crop was raised following standard agronomic practices. The morphological data was collected for a period of four years from both the locations (Palampur and Bajaura). The data of five randomly selected plants were measured for four traits viz., plant height, number of branches per plant, days to flowering and days to maturity in each line and the average was calculated before final analysis. Furthermore, total days after the date of sowing to the date till 100% of plants in a single line manifest completely open flowers were recorded for days to flowering and total days after the date of sowing till the complete physiological maturity of the plants (90% of the plants turned brown) were measured for days to maturity trait (Katoch, et al., 2021) (Supplementary Table 2). The Past version 3.25 software was used for the trait distribution study on the basis of kurtosis and skewness statistics. Moreover, the frequency distribution and the Pearson correlation coefficient was also calculated using the same software (Hammer, Harper, & Ryan, 2009). To estimate the presence of variations among location, years, genotypes, genotypes by year, genotype by location, and genotype by environment (genotype by location by year) using R version 4.0.3. The environment was considered as random variables and genotypes as fixed variable.

3.2. DNA extraction, GBS library preparation and Sequencing

Young green leaf tissues from all the accessions were collected, dipped in liquid N₂ to freeze dry and were immediately stored in the -80° C till further use. The DNA was isolated from all the accession using CTAB method with minor modifications (Murray & Thompson, 1980). Total 1gms of the freeze dried leaf tissue was used for the DNA isolation and the purity and concentration of the DNA was measured using Bio-analyzer High Sensitivity DNA Analysis chip (Agilent, USA). Further, high quality 100ng genomic DNA was used for the library preparation and sequencing. For the GBS library preparation the genomic DNA was first digested with the restriction enzyme *ApeKI* (New England Biolabs, Ipswich, MA). Further, the restricted digested DNA (complementary sticky ends) was ligated with the barcoded adapters (P1 and P2) having unique barcodes of 4-8bp and the fragmented DNA with ligated adapters from all the accessions were pooled together in a single falcon tube to make an 88-plex library and size selected. The pooled library was purified with QIAGEN purification kit to remove the excess of adapters and after that PCR amplification was carried out with sequencing primer (amplification of both P1 and P2). Further, the amplified product was again purified to remove any left contamination and the library was quantified using Agilent DNA High Sensitivity Kit (2100 Bio-Analyzer). Total 10 pmole of library was loaded on to the single lane of Illumina *HiSeq*TM 2500 platform using V4 sequencing chemistry (Illumina Inc., San Diego, CA, USA) to perform the high throughput DNA sequencing following 100bp single end protocol.

3.3. Genotyping of GWAS panel

The sequencing data generated by the high throughput sequencing platform was retrieved after base calling using CASAVA software and the generated sequences and its quality score was in FASTQ file. The raw reads generated were de-multiplexed and all the reads were assigned to the 88 accession using unique barcoded

adapters with the help of Fastx Toolkit (v 0.0.6, http://hannonlab.cshl.edu/fastx_toolkit/fastx_toolkit-0.0.6.tar.bz2). All the barcodes were removed from the generated sequences and the sequencing error rate was examined to detect any sites with unusually high rates of incorrectly added bases and finally the high quality reads for each accession was retrieved using Trim Galore (0.6.2, <https://github.com/FelixKrueger/TrimGalore/archive/refs/tags/0.6.2.tar.gz>). After quality filtering, the reads were aligned with the reference genome *Medicago truncatula* (http://ftp.ebi.ac.uk/ensemblgenomes/pub/release-49/plants/fasta/medicago_truncatula/dna/) using BWA (Burrows Wheeler Alignment) tool ver.0.7.5 (Li & Durbin, 2010) to identify the Single Nucleotide Polymorphism (SNPs) between the accessions. These variants were called using GATK version 3.6 (Genome Analyzer Tool Kit) (<https://gatk.broadinstitute.org/hc/en-us>) and were retrieved in the Variant Call Format (VCF). Further, missing data more than 10% or MAF (Minor Allele Frequency) including monomorphic loci's below 5%, indels, non-biallelic, co-efficient of inbreeding (<0.9) were removed using VCF tool version 0.1.15 (Danecek, et al., 2011) and the SNPs were further imputed using Beagle ver. 5 (Browning, Zhou, & Browning, 2018). These high quality SNPs were assigned to the chromosomes of the reference genome used in current study the *Medicago truncatula*. Furthermore, the anchored SNPs were functionally annotated using SnpEff tool (<http://pcingola.github.io/SnpEff/>) a genomic variant annotation and functional effect tool box, which helps to predict and annotate the effects of genetic variations on the genes (amino acid changes) (Cingolani, et al., 2012).

3.4. Genetic diversity, population structure and linkage disequilibrium analysis

Total 54271 high quality SNPs were used for genetic diversity analysis and population structure analysis. Evaluation of these SNPs in the panel of 88 horsegram accessions revealed that less than 1% of the missing data, therefore no genotype was removed from the analysis. The number of subgroups were identified in the accession using PCA (Principal Component Analysis) and a model based clustering method using GAPIT (Genome Association and Integrated Prediction Tool) version 3.0 and STRUCTURE version 2.3.4A (https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html) (Falush, Stephens, & Pritchard, 2007; Pritchard, Stephens, & Donnelly, 2000). The admixture model with MCMC 200000, 10 replications and 50000 burns for number of sub-populations (K =1 to 8) were used and the ΔK value was estimated to predict the number of groups present (Evanno, Regnaut, & Goudet, 2005). The results from the STRUCTURE were visualised with STRUCTURE harvester (Earl & vonHoldt, 2012). Horsegram accession with higher membership coefficient equal to or higher than 0.6 were assigned to the linkage group, however with lower membership coefficient were clustered into the admixture group. Furthermore, SNIPlay a web based tool was used for exploration and large scale analysis for identification of number of transitions, transversions, and to predict the density of SNPs in each sub-population (Dereeper, et al., 2015). The phylogenetic analysis was carried out using TASSEL version 4 and Neighbor Joining tree was made based on distance matrix predicted on the basis of SNPs (Bradbury, et al., 2007). Finally, the *Fst* index, nucleotide diversity (π) and Tajima's D was estimated using VCF tool ver.0.1.15 (Bradbury, et al., 2007; Tajima, 1989; Weir & Cockerham, 1984). Understanding of the pattern of the Linkage Disequilibrium (LD) present within the genomic material is important to determine the requirement of marker density for a genome scan (Siol, et al., 2017). Therefore, in current study the linkage disequilibrium was calculated using TASSEL version 4 (Bradbury, et al., 2007). The pairwise LD for pairs of SNP markers (r^2 , which is squared allele frequency) was estimated using Tomahawk 0.7.0 (<https://github.com/mklarqvist/tomahawk>) and LD was plotted using dplyr and ggplot2 R package. The value of

mean r^2 for each chromosome was calculated and plotted against the physical distance of chromosome's and the distance at which the value of r^2 dropped to half of its average maximum was called the LD decay rate (Huang, et al., 2010).

3.5. Identification of marker-trait association in *Macrotyloma uniflorum*

GWAS analysis was performed for four traits (plant height, number of branches per plant, days to flowering and days to maturity) in *Macrotyloma uniflorum* using 54271 high quality SNPs. The association between the phenotype and genotypes was performed by employing MLM (Mixed Linear Model) and two parameters were considered viz., the Q value (genetic structure of the panel) and K value (kinship coefficient matrix). The estimated K values and co-ordinates (as covariates) were used to run the MLM model and Q-Q plots (quantile-quantile) were drawn between observed and expected P values and on the basis of generated MLM based K and Q values the marker trait association was found.

4. Conclusion

Horsegram is a versatile crop with excellent nutraceutical characteristics that is tolerant of moisture, drought stress, and a wide range of agro-climatic zones. It is a significant legume grown by rural and tribal groups in Asia and Africa for food, fodder, and traditional medicine. In the current work, efforts were made to comprehend the genetic diversity, population structure analysis, and genome-wide association analyses of a panel of 88 accessions of this plant. Four important traits Plant height, number of branches per plant, days to flowering, days to maturity were selected for association study. This study summarized 79 significant SNPs across 8 chromosomes among four traits and the number of branches per plant was associated with a maximum of 37 SNPs followed by 20 significant associations for days to flowering and 15 SNPs for plant height. As discussed the importance of these traits with respect to yield of the crop these SNPs can be used for various genomic studies. Moreover, This study will not only be helpful for marker-assisted genetic improvement of horsegram plant to make it a good nutritional supplement for the fast-growing nutritional demands but can be used in insilco gene prediction.

Declarations

Ethical Approval

“This is a plant based study. Therefore, no ethical approval is required.”

Competing interests

The authors have no financial interest to disclose.

Authors' contributions

RKC conceptualize the idea and provide all the facility to carry out genotypic and phenotypic work and finalized the manuscript, AS performed the experiment, recorded phenotypic data, performed molecular data analysis and written the manuscript.

Funding

The research was funded from our own resources, no external funding was available

Availability of data and materials

All the related. Data will be made available in the journal site, when asked to do so.

Consent to participate

Informed consent was not obtained from any individual participants as the study did not included human re-search.”

Consent to publish

“The authors affirm that there is no data is generated on human participants, therefore no consent is required for publication of the images or data”

References

1. Akbari M, Wenzl P, Caig V et al (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor Appl Genet* 113:1409-1420.
2. Alipour H, Bihamta MR, Mohammadi V et al (2017) Genotyping-by-Sequencing (GBS) Revealed Molecular Genetic Diversity of Iranian Wheat Landraces and Cultivars. 8.
3. Alsamir M, Hameed M (2019) Phenotypic diversity and marker-trait association studies under heat stress in tomato (*Solanum lycopersicum* L.).
4. Atwell S, Huang YS, Vilhjálmsson BJ, et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627-631.
5. Bhardwaj J, Chauhan R, Swarnkar MK et al (2013) Comprehensive transcriptomic study on horse gram (*Macrotyloma uniflorum*): De novo assembly, functional characterization and comparative analysis in relation to drought stress. *BMC Genomics* 14:647.
6. Bradbury PJ, Zhang Z, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635.
7. Britannica, The Editors of Encyclopaedia. "Nikolai Vavilov". Encyclopedia Britannica, 22 Jan. 2023, <https://www.britannica.com/biography/Nikolai-Vavilov>. Accessed 29 July 2023.
8. Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* 103:338-348.
9. Chahota R (2020) Construction of a framework linkage map and genetic dissection of drought- and yield-related QTLs in horsegram (*Macrotyloma uniflorum*). *Euphytica*.
10. Chahota RK, Sharma TR, Sharma SK, et al (2013) 12 - Horsegram. In M. Singh, H. D. Upadhyaya & I. S. Bisht (Eds.), *Genetic and Genomic Resources of Grain Legume Improvement* (pp. 293-305). Oxford: Elsevier.
11. Chahota RK, Sharma V, Rana M, et al (2020). Construction of a framework linkage map and genetic dissection of drought- and yield-related QTLs in horsegram (*Macrotyloma uniflorum*). *Euphytica* 216:61.
12. Choudhary S, Isobe S, Chahota RK (2022) Elucidation of drought tolerance potential of horsegram (*Macrotyloma uniflorum* Var.) germplasm using genome wide association studies. *Gene* 819:146241.

13. Cingolani P, Platts A, Wang le L et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6:80-92.
14. Crawford DC, Goodloe R, Farber-Eger E et al (2015). Leveraging Epidemiologic and Clinical Collections for Genomic Studies of Complex Traits. *Hum Hered* 79:137-146.
15. Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.
16. Delfini J, Moda-Cirino V, dos Santos Neto J et al (2021) Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm. *Scientific reports* 11:2964.
17. Dereeper A, Homa F, Andres G et al (2015) SNIPlay3: a web-based application for exploration and large scale analyses of genomic variations. *Nucleic Acids Res* 43: W295-300.
18. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359-361.
19. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611-2620.
20. Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 7:574-578.
21. Gali KK, Sackville A, Tafesse EG et al (2019) Genome-Wide Association Mapping for Agronomic and Seed Quality Traits of Field Pea (*Pisum sativum* L.). *Front Plant Sci* 10:1538.
22. Guo Y, Chen S, Li Z et al (2014) Center of Origin and Centers of Diversity in an Ancient Crop, Brassica rapa (Turnip Rape). *Journal of Heredity* 105:555-565.
23. Hammer O, Harper D, Ryan P (2009) PAST–Palaeontological statistics, ver. 1.89. *Palaeontologia Electronica* 4.
24. Huang X, Wei X, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961-967.
25. Kaldete R, Rana M, Sharma V et al (2017) Development of genome-wide SSR markers in horsegram and their use for genetic diversity and cross-transferability analysis. *Molecular Breeding* 37:103.
26. Katoch M, Mane RS, Chahota RK (2021) Identification of QTLs Linked to Phenological and Morphological Traits in RILs Population of Horsegram (*Macrotyloma uniflorum*). *Front Genet* 12:762604.
27. Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29.

28. Li H, Durbin, R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589-595.
29. Lijavetzky D, Cabezas JA, Ibáñez A et al (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* 8:424.
30. Lin S, Medina CA, Boge B et al (2020) Identification of genetic loci associated with forage quality in response to water deficit in autotetraploid alfalfa (*Medicago sativa* L.). *BMC Plant Biol* 20:303.
31. Liu C, Chen K, Zhao X et al (2019) Identification of genes for salt tolerance and yield-related traits in rice plants grown hydroponically and under saline field conditions by genome-wide association study. *Rice* 12.
32. Mahesh HB, Prasannakumar MK, Manasa KG et al (2021) Genome, Transcriptome, and Germplasm Sequencing Uncovers Functional Variation in the Warm-Season Grain Legume Horsegram *Macrotyloma uniflorum* (Lam.) Verdc 12.
33. Mashilo J, Shimelis H, Odindo A (2016) Correlation and path coefficient analyses of qualitative and quantitative traits in selected bottle gourd landraces. *Acta Agriculturae Scandinavica, Section B — Soil & Plant Science* 66:558-569.
34. Mulugeta B, Tesfaye K, Ortiz R et al (2022) Marker-trait association analyses revealed major novel QTLs for grain yield and related traits in durum wheat. *Front Plant Sci* 13:1009244.
35. Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* 8:4321-4325.
36. Parisi M, Alioto D, Tripodi P (2020) Overview of Biotic Stresses in Pepper (*Capsicum* spp.): Sources of Genetic Resistance, Molecular Breeding and Genomics. *Int J Mol Sci* 21.
37. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
38. Saxena RK, Kale S, Mir RR et al (2020) Genotyping-by-sequencing and multilocation evaluation of two interspecific backcross populations identify QTLs for yield-related traits in pigeonpea. *Theor Appl Genet* 133:737-749.
39. Sharma V, Sharma T, Rana J et al (2015) Analysis of Genetic Diversity and Population Structure in Horsegram (*Macrotyloma uniflorum*) Using RAPD and ISSR Markers. *Agricultural Research*.
40. Siol M, Jacquin F, Chabert-Martinello M et al (2017) Patterns of Genetic Structure and Linkage Disequilibrium in a Large Collection of Pea Germplasm. *G3 (Bethesda)* 7:2461-2471.
41. Spindel J, Wright M, Chen C et al (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor Appl Genet* 126:2699-2716.

42. Suwarna P, Vs P, Shinde S (2019) Studies on nutritional, chemical and mineral composition of horse gram 7:53-55.
43. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.
44. Tanin MJ, Saini DK, Sandhu KS et al (2022) Consensus genomic regions associated with multiple abiotic stress tolerance in wheat and implications for wheat breeding. *Scientific reports* 12:13680.
45. Toppo NA, Krishi V, Nair S et al (2019) Genetic variability studies on yield and yield contributing traits in blackgram (*Vigna mungo* (L.) Hepper.
46. *Tropical Legumes: Resources for the Future*. (1979). Washington, DC: The National Academies Press.
47. Wang N, Yuan Y, Wang H et al (2020) Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Scientific reports* 10:16308.
48. Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38:1358-1370.
49. Xu X, Liu X, Ge S et al (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 30:105-111.
50. Zhu C, Gore M, Buckler E et al (2008) Status and Prospects of Association Mapping in Plants. *The Plant Genome* 1:5-20.

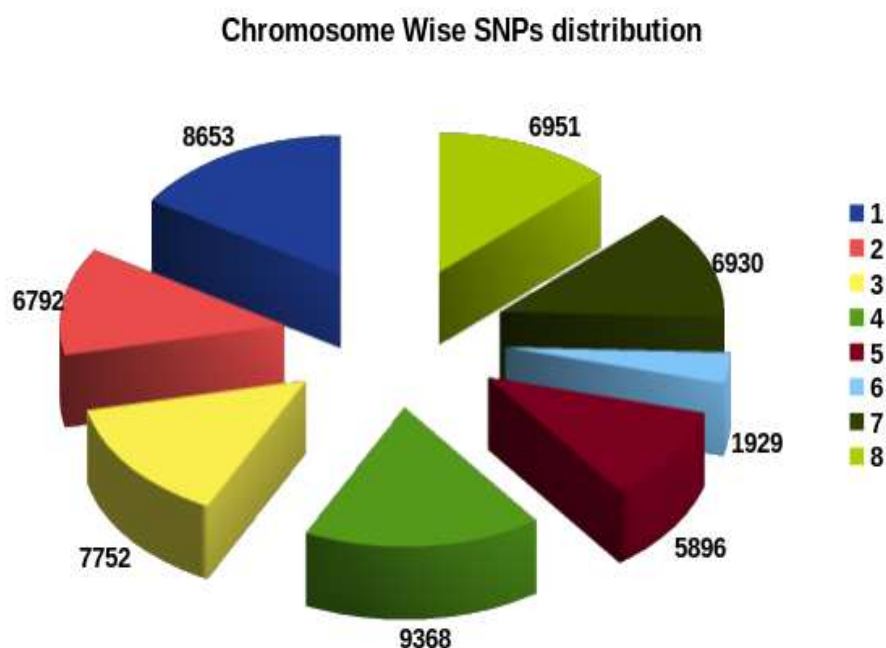
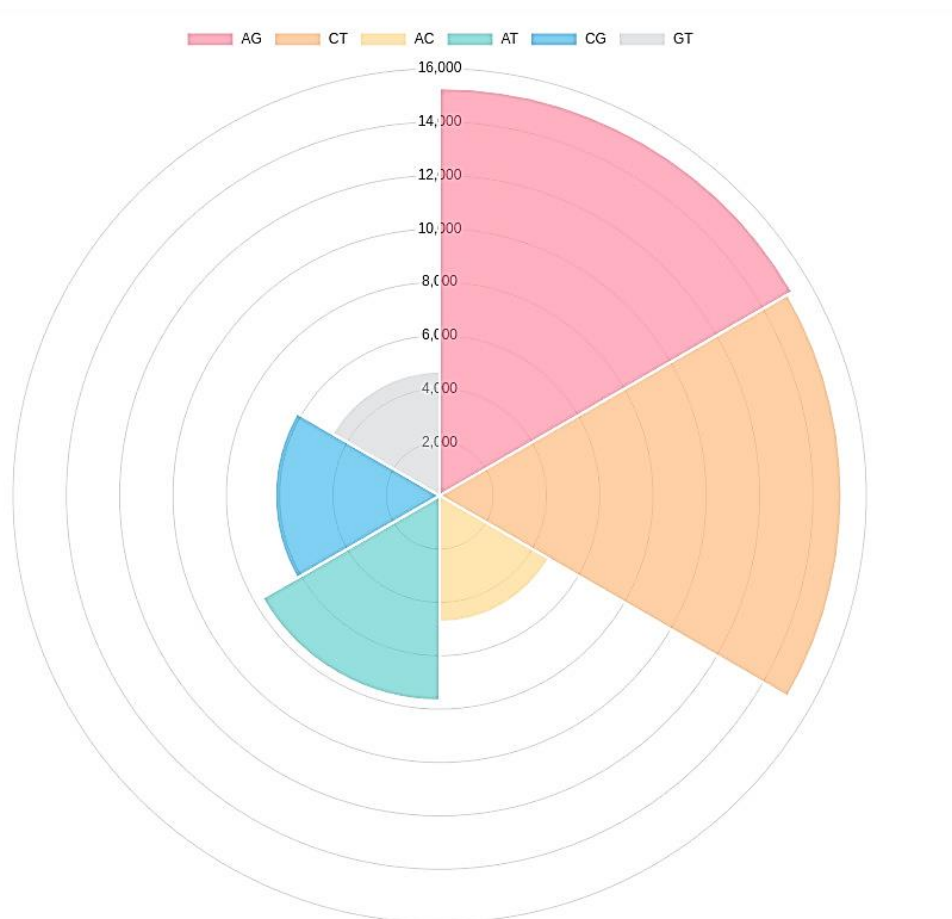


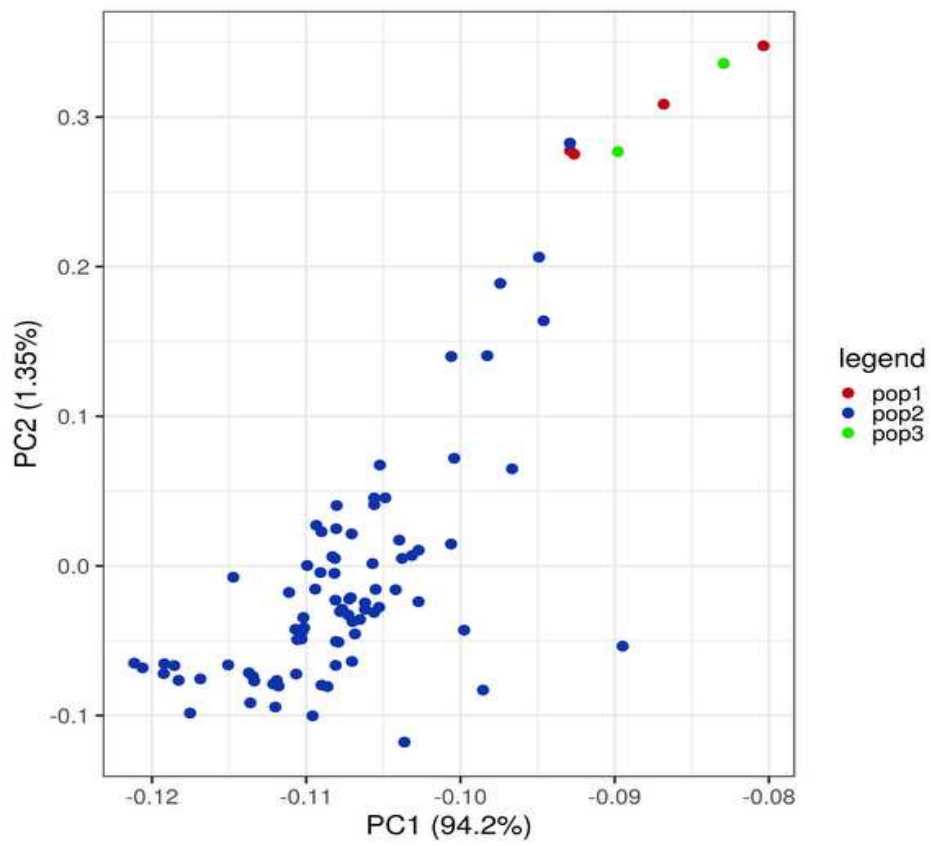
Figure 1: Distribution of SNPs



483

484

Figure 2: Transition transversions



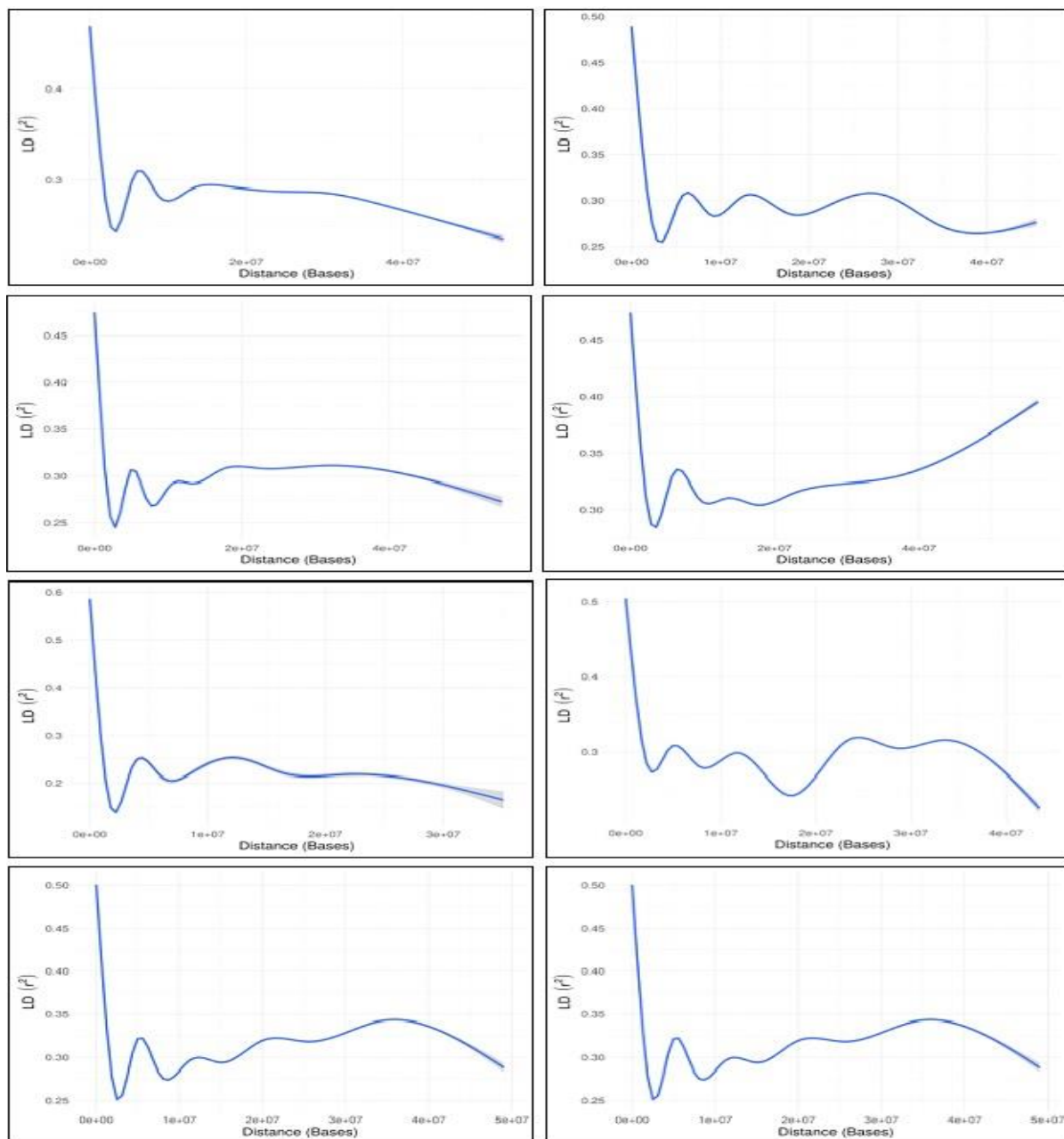
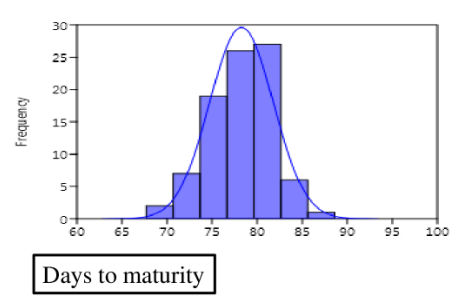
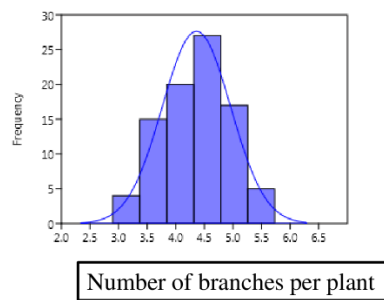
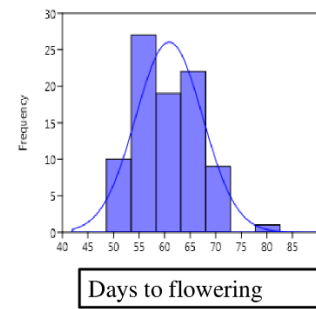
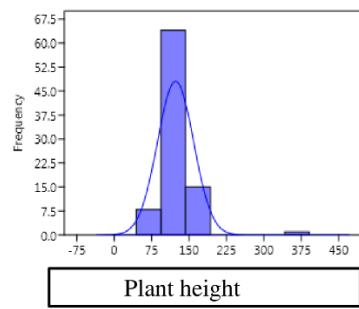


Figure 4: Distances between the polymorphic loci for each SNP



491
 492 **Figure 5:** frequency distribution for phenotypic traits



493

494 **Figure 6:** correlation between the phenotypic traits

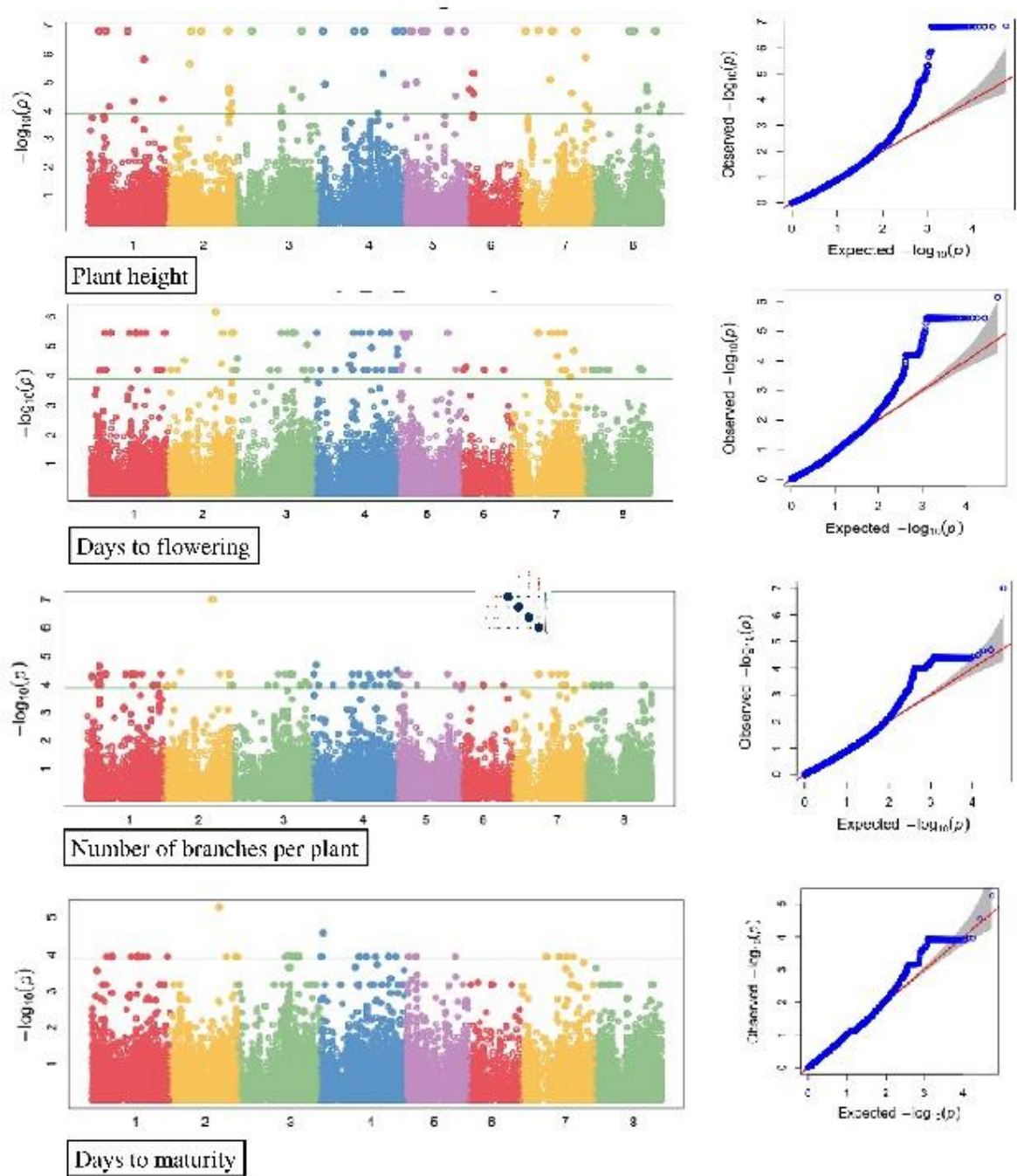


Figure 7: Manhattan and QQ plots