

Identification of candidate genes involved in genetic variation in seed weight in *Zea mays* (L.) using Genome-Wide Association Studies

Shubham Sharma, Haroon Rashid Hakla and Mohammad Urfan

Plant Physiology Laboratory, Department of Botany, University of Jammu, Jammu 180 006, India

(Received 2 January, 2024; Accepted 29 February, 2024)

ABSTRACT

Cereal crops are the most important crops globally. They constitute the staple food for most of the people. With the increasing world population, food scarcity is posing a global challenge. So, there is a need to increase the food production with the available land. This increase is only possible with the use of modern scientific techniques. Maize is the third most important cereal crop which is grown globally. To increase its yield, there is an urgent need to find those varieties that have a large seed size and weight. This can be done using Genome-wide association studies (GWAS), which is a better approach than the traditional breeding methods. For the current study, a large population size consisting of 334 maize accessions of the Ames panel was used. The Genome Based Sequence of the accessions is freely available at <https://www.panzea.org>. The genomic data was filtered using Trait Analysis by Association, Evolution, and Linkage (TASSEL) software. The GWAS analysis was done using the BLINK model in the GAPIT library in R-studio. One significant SNPs was found to be associated with seed weight. The findings provided valuable insights into the genetic basis of the seed biomass allocation and will aid in the development of such varieties that produce large-sized seeds and are crucial for food security.

Key words: GWAS, Maize, GAPIT, TASSEL.

Introduction

Cereal crops such as rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and maize (*Zea mays* L.) are staple food for the majority of the people (Farooq *et al.*, 2015). An estimated 50 % increase in the grain yield of these major cereals is required to meet the food requirement for the projected population by 2050 (Godfray *et al.*, 2010). Maize is the third most important cereal crop which is grown all around the globe under a wide range of environmental conditions. *Zea mays* L. (2n=20) with the global area under its cultivation of about 197 million hectares is an established (FAO Stat, 2021) and important food crop in numerous countries. It contributes about 20% of

total calorie intake in humans. Maize is a more versatile and multi-purpose crop compared to wheat and rice. With the dependence on maize for food, fiber, and fuel, there is a requirement of a substantial increase in the yield of maize plants (Sun *et al.*, 2015). Maize seed size plays a crucial role in determining yield and plant vigor. Larger seeds often result in stronger seedlings with better early growth (Steiner *et al.*, 2019). Farmers often select maize varieties based on seed size to optimize their crop performance (Guo *et al.*, 2020). Additionally, seed size can impact the efficiency of planting and seedling emergence. Understanding and managing maize seed size is key to maximizing agricultural productivity.

Seed size is a complex agronomic trait due to the involvement of multiple genetic factors and environmental influences. Maize seeds exhibit substantial variations in size and weight within and among different populations (Li *et al.*, 2019). Grain yield is a highly quantitative trait that is under the control of many genes with small effects which usually escape detection by QTL mapping (Neuweiler *et al.*, 2020). So, there is a need to find the genetic variation involved in the resource allocation in seeds, which can be done by performing a whole genome analysis. Genome-wide association Studies (GWAS) is a computational biology technique that uses large population size to link a trait with the genome and find the genetic variation responsible for that trait. The current study involves the study of large population size (334 maize accessions) on which association analysis was performed for seed weight.

Materials and Methods

Plant material

334 maize accessions from the Ames panel were procured from the United States Department of Agriculture (USDA) USA via the germplasm exchange program of the National Bureau of Plant Genetic Resources (NBPGR), New Delhi, India.

Agronomic traits

100 seeds were randomly selected from each accession and hundred seed weight (HSWT) was measured (in g) using a digital balance.

Genome-wide association studies (GWAS)

The maize accessions from the Ames panel consisted of 334 accessions. The Genotyping by sequencing (GBS) data was retrieved from free data source of <https://www.panzea.org> (<https://datacommons.cyverse.org/browse/iplant/home/shared/panzea/genotypes/GBS/v27>). The genotype data had 9,45574 single nucleotide polymorphisms (SNPs), which were filtered with Trait Analysis by Association, Evolution, and Linkage (TASSEL) software. For the association analysis, SNPs were filtered to 3,30256 with a minor allele frequency (MAF) > 0.05. For the estimation of Linkage disequilibrium (LD), 2,71556 SNPs were filtered from total SNPs with MAF > 0.1. For association analysis Genome Association and Prediction Integrated Tool (GAPIT) library was used in R-Studio.

Prediction of candidate genes

The high-quality SNPs obtained after filtration were used to run GAPIT for HSWT. The SNP hits which lie above the threshold were used for gene mapping. The candidate genes were predicted using the significant SNPs. Each SNP was deeply analyzed and a window of 50,000 nucleotides plus and minus strands of SNP was studied using https://ensembl.gramene.org/Zea_mays/Info/Index. The gene (s) falling within the range of SNP (+ or -, 50,000 bp) were enlisted for each trait. The localization of a gene (s) was performed using the maize eFP browser (utoronto.ca).

Results

The current study was involved in the GWAS statistical approach to study the association of the whole genome with the seed weight. The 334 maize accessions procured from USDA showed phenotypic variations (Figure 2A) in terms of all the seed allometric trait's weight. The highest HSWT was observed to be 38.95 g in accession PI221747 and the lowest to be 8.18g in PI198895.

GWAS analysis

The GWAS analysis was performed on HSWT using BLINK model of GWAS in GAPIT library of R-studio. Kinship was also done which showed the closeness among the accession of maize population (Figure 1B). Linkage disequilibrium (LD) plot showed the LD decay in the accessions, which is the non-random association among the accessions (Figure 1C). The three-dimensional principal component analysis (PCA) provides significant estimation of population structure of the maize accessions for GWAS analysis (Figure 1D).

Using BLINK a single significant SNP hit was obtained in the Manhattan plot above the threshold line viz. S5_163940647 located on the chromosome number 5 (Figure 2B). The quantile-quantile (Q-Q) plot (Figure 2C) showed SNPs, and reliability of the Manhattan plot.

Prediction of candidate gene

For the prediction of the candidate gene 100kb genomic region was scanned keeping the SNP at the center. The SNP S5_163940647 located on chromosome 5 was mapped to *Zm00001d016584* (Table 1) which is a *LITAF-domain-containing protein*.

Table 1. Showing SNP provided by BLINK for HSWT

Trait	SNP	Chromosome	Position	p-value	MAF	Gene
HSWT	S5_163940647	5	167817986	6.29E-09	0.191395	Zm00001d016584

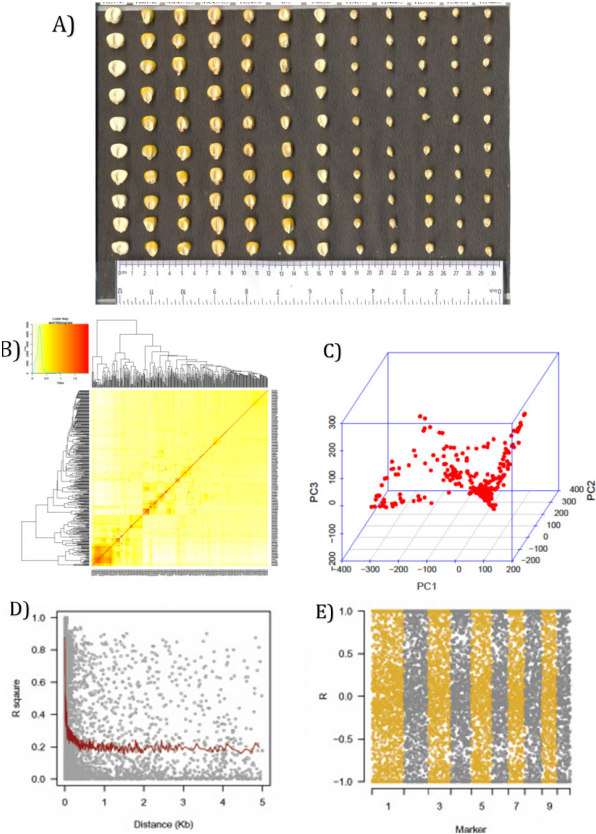


Fig. 1. A) left to right seeds of largest to smallest maize seeds with two references B73 and CML at the center, B) Kinship matrix showing genetic relatedness among maize accessions, C) 3D-principal component analysis illustrating the population structure of the population, D) Linkage disequilibrium (LD) plot showing the LD decay C) marker density.

Bioinformatic analysis of candidate genes

The candidate gene Zm00001d016584 for HSWT showed that it codes for DNA sequence (CDS) that spanned 1149 base pairs, translating into an 184 amino-acid residue sequence. Its transcript was characterized by two exons, annotated with six domains, and correlated with 1315 variant alleles, aligning with 58 oligo probes. Interactions predicted through the STRING database highlighted ten proteins potentially interacting with Zm00001d016584 (Figure 3).

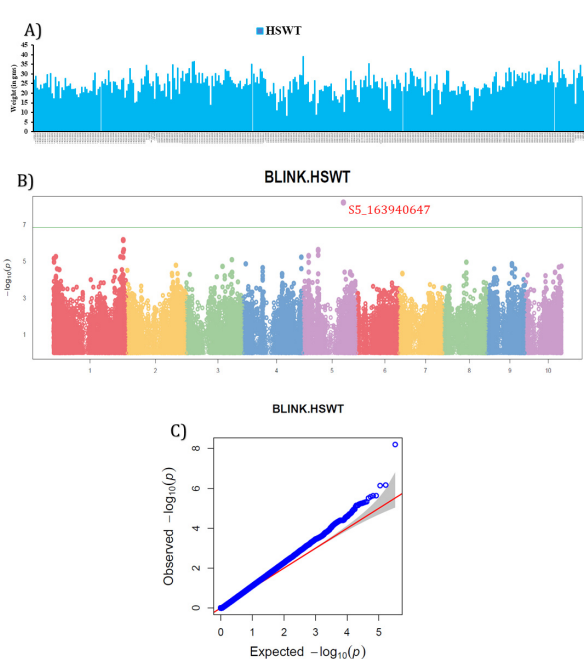


Fig. 2. BLINK for hundred seed weight HSWT) A) Phenotypic variation among 334 accessions B) Manhattan plot showing a single significant SNP above the threshold line, and C) Quantile-quantile (Q-Q) plot.

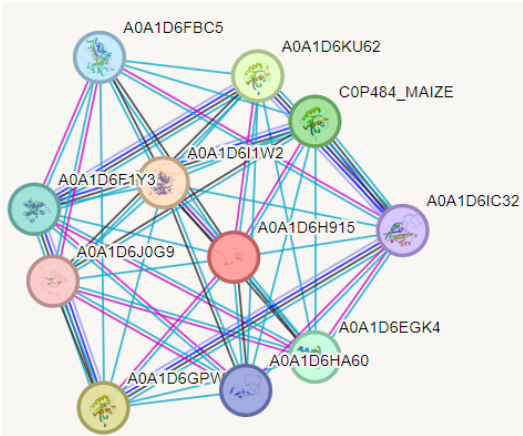


Fig. 3. STRING predicted the interaction of proteins. The genes and their protein IDs (shown in bold) show the interaction of proteins of identified candidate genes using STRING database version 12.0 for hundred seed weight. The red circle represents the identified candidate gene Zm00001d016584 with protein ID A0A1D6H915.

Discussion

Seed weight is controlled by various genes, which makes it necessary to understand the genetic basis of seed weight via the identification of probable candidate QTLs that are associated with the trait (Zhang *et al.*, 2021). Seed weight is the most important parameter determining the yield of a crop (Khan *et al.*, 2019). GWAS on seed weight in mulberry plant showed genetic variation in three genes *Bp02g2123*, *Bp01g3291* and *Bp10g1642* (Hu *et al.*, 2022). In the present study, BLINK was used to study GWAS for hundred seed weight. Association analysis predicted a SNP on chromosome 5 to be associated with HSWT trait. A novel candidate gene *Zm00001d016584* was mapped 53kb away from this SNP and is 291 bp in length. It is a LITAF-domain-containing protein. This finding lays the foundation for marker-based improvement in maize. In depth analysis are needed to validate the functioning of this gene.

Acknowledgements

The authors are thankful to the Department of Botany, University of Jammu-180006 for providing the necessary facilities.

Conflicts of interest

The authors declare no conflict of interest.

References

- FAO Stat, 2021. FAO Stat. FAO, Rome. <http://www.fao.org/faostat>
- Farooq, M., Hussain, M., Wakeel, A. and Siddique, K.H. 2015. Salt stress in maize: effects, resistance mechanisms, and management. A review. *Agronomy for Sustainable Development*. 35: 461-481.
- Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M. and Toulmin, C. 2010. Food security: the challenge of feeding 9 billion people. *Science*. 327(5967): 812-818.
- Guo, Y., Gao, M., Liang, X., Xu, M., Qu, S. and Luan, F. 2020. Quantitative trait loci for seed size variation in cucurbits—a review. *Frontiers in Plant Science*. 11: 494620.
- Hu, Y., Peng, X. and Shen, S. 2022. Identification and investigation of the genetic variations and candidate genes responsible for seed weight via GWAS in paper mulberry. *International Journal of Molecular Sciences*. 23(20): 12520.
- Khan, S.U., Yangmiao, J., Liu, S., Zhang, K., Khan, M.H.U., Zhai, Y., Olalekan, A., Fan, C. and Zhou, Y. 2019. Genome-wide association studies in the genetic dissection of ovule number, seed number, and seed weight in *Brassica napus* L. *Industrial Crops and Products*. 142: 111877.
- Li, L., Li, X., Li, L., Schnable, J., Gu, R. and Wang, J. 2019. QTL identification and epistatic effect analysis of seed size-and weight-related traits in *Zea mays* L. *Molecular breeding*. 39: 1-11.
- Neuweiler, J.E., Maurer, H.P. and Würschum, T. 2020. Longterm trends and genetic architecture of seed characteristics, grain yield and correlated agronomic traits in triticale (× *Triticosecale* Wittmack). *Plant Breeding*. 139(4): 717-729.
- Steiner, F., Zuffo, A.M., Busch, A., Sousa, T.D.O. and Zoz, T. 2019. Does seed size affect the germination rate and seedling growth of peanut under salinity and water stress? *Pesquisa Agropecuária Tropical*. 49.
- Sun, C., Gao, X., Fu, J., Zhou, J. and Wu, X. 2015. Metabolic response of maize (*Zea mays* L.) plants to combined drought and salt stress. *Plant and Soil*. 388: 99-117.
- Zhang, W., Xu, W., Zhang, H., Liu, X., Cui, X., Li, S., Song, L., Zhu, Y., Chen, X. and Chen, H., 2021. Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean. *Theoretical and Applied Genetics*. 134: 1329-1341.