GENOME-WIDE ASSOCIATION STUDIES OF EAR TRAITS IN MAIZE

A Thesis

by

GALI BAI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Hongbin Zhang
Co-Chair of Committee,	Sing-Hoi Sze
Committee Member,	Steve Hague
Head of Department,	David D. Baltensperger

August 2020

Major Subject: Molecular and Environmental Plant Sciences

ABSTRACT

This study was aimed at finding genetic markers or genes that are associated with ear traits important to maize by genome-wide association study (GWAS) based on the USDA maize GWAS panel consisting of diverse inbred lines that represent the genetic variation and diversity of maize present at the world-wide maize public breeding programs. This study phenotyped 263 diverse inbred lines of the GWAS panel for seven ear traits: grain weight per plant (GW), ear weight per plant (EW), ear length (EL), kernel-row length (KL), ear diameter (ED), number of kernel rows (KR), and cob diameter (CD) in College Station, Texas from 2017 to 2019, and in Lubbock, Texas in 2018. These 263 inbred lines were genotyped using the whole genome shotgun sequencing reads, having an average coverage of 7.0 x, with high performance computing clusters and the maize inbred line B73 genome as the reference genome. A total of 1,553,207 quality single nucleotide polymorphism (SNP markers were identified after filtering with a base-calling quality score of ≤ Q30, missing rate of ≥0.1, and minor allele frequency (MAF) of ≤0.1. Population structure was stratified by population structure analysis, principle component analysis, and phylogenetic tree, respectively, together indicating that the USDA GWAS panel consists of six subpopulations: stiff stalk (SS), non-stiff stalk (NSS), tropical-subtropical (TS), popcorn, sweet corn, and mixed.

By constructing a general linear model (GLM) with kinship matrix as covariate, three SNP markers were identified that were associated with ear traits at a significance level of $-\log_{10}(P) = 7.0$. Three candidate genes and three SNPs that were previously

characterized were identified near these SNPs. These results have provided the candidate genes controlling the traits and SNP markers necessary for enhanced breeding for these traits through marker-based breeding.

DEDICATION

This thesis is dedicated to my parents, who have raised me, educated me to the person I am today. They have always supported me, encouraged me whenever I am encountered with hard choices. With their love and inspiration, I am never afraid to take any challenges in my life.

To my grandparents, who have been always devoted their unselfish love.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to the advisory committee Dr. Hongbin Zhang, Dr. Sing-Hoi Sze, and Dr. Steve Hague. I want to thank my academic advisor Dr. Hongbin who have helped me with all the aspects of my research and life. It is not only the broad knowledge that I have learned from his guidance, but also his passion to pursue the true meaning of academic career inspired me to be more professional in my major fields. I would also like to thank my co-chair Dr. Sing-Hoi who helped me a lot with the computational work. I am so honored to have this opportunity to improve my computational skills in biological field. I want to thank Dr. Steve for advising and supporting me throughout the research, and continuously providing me new ideas of developing thesis. Thank you for your insightful advice that made my research more comprehensive.

I also would like to thank Dr. Seth Merry who have helped me with planting and managing the experimental fields, lab technician Chantel Scheuring who helped me with collecting phenotypic data. I must thank my lab mates Mustafa Cilkiz and Mehmet Dogan, who are always eager to offering supports.

I couldn't have been able to complete this research without any of you all. Thank you everyone for giving me warmth and help.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Hongbin Zhang (Advisor), Steve Hague of the Department of Soil and Crop Sciences and Professor Sing-Hoi Sze (Co-Advisor) of the Department of Computer Science and Engineering.

The data analyzed in Section 4 was partly provided by Professor Hongbin Zhang.

Funding Sources

This research was supported in part by grants from the Texas Corn Producer Board (TCPB) (408116-85360) and the Texas A&M AgriLife Research Crop Improvement Program (06-124329-85360 and 06-124215-85360).

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. Importance of Global Maize Production

Maize, *Zea mays* L., is the largest crop in the world for human food, animal feed, and biofuel. In 2018, global maize production yielded over 1,400M tons of maize grains, harvested from over 230M hectares across 169 countries (1). As one of the most important plant nutrition sources, maize, together with wheat and rice, accounts for 30% of the total food calories, feeding over 4.5 billion people in the world (2). Not only the top maize producing countries, such as the United States, China, and Brazil that produce 63% of the global maize, the but also most middle- and lower-income countries heavily rely on maize (3). For instance, in Ethiopia, agriculture is still the most important part of its economy, to which maize contributed more than 7.3 M tons in 2018 (4). Due to the increasing population and low productivity, the Ethiopian farmers have been struggling with the situation of food insecurity for many years. In addition to population growth, the global climate change, such as elevated temperature and increased drought frequency, is also negatively impacting maize productivity. In the United States, the world's largest maize production belt, including Nebraska, Minnesota, Iowa, and Illinois, is continuously subjected to drought stress (5). Moreover, as the world's population will exceed 9 billion by 2050, the demand for maize production will double (6). Therefore, given the projected global population growth and the adverse impacts of the observed global climate change on crop production, it is apparently necessary to continuously improve maize production and farming efficiency to help feed the world.

Furthermore, the usage of maize in livestock feeds and fuel production is also significantly increasing. It was estimated that approximately 63% of maize is globally consumed as livestock feed (2). This proportion can be higher in most developed countries where the percentage stands around 70%. Another primary use of maize is to generate ethanol fuel in industries. In the past decade, maize ethanol production used around 40% of the total maize production in the United States (7). The uses of maize in food, feed, and biofuel have had the global maize price fluctuate, showing an increasing trend. If the global maize supplies cannot match the increasing demand, global maize prices will be unaffordable for millions of consumers.

1.2. Structure and Architecture of the Maize Genome

Maize is important not only as a crop species, but also as a model species system for genetic and genomic research, due to its high level of genetic diversity. Millions of sequence polymorphisms have been identified in the low-copy genomic region of 27 diverse maize inbred lines (8). The total number of 3.3 million single-nucleotide polymorphisms (SNPs) and nucleotide insertions/deletions (InDels) indicated that there will be 1 polymorphism detected in every 44 bp in the maize genome. Such abundant nucleotide sequence variation in in the maize genome makes maize possible to select nucleotide sequence polymorphisms associated with the variations of quantitative traits. However, the association mapping greatly relies on the linkage disequilibrium (LD), implying that only those tightly linked SNPs and/or InDels may significantly associate with phenotypic traits in a random population (9).

LD is the core of association studies. In a natural maize population, if there were no genetic recombination, mutation, and selection, the alleles of genes at different loci would present in linkage (10). In contrast, because of genome evolution and genetic recombination, the LD level varies among populations, depending on the level of sequence polymorphisms and genetic recombination frequency. The higher correlation means the closer neighboring of two polymorphisms. The LD decay determined with102 maize inbred lines showed that the predicted R^2 values declined rapidly within 2,000 bp(11). Compared to the LD decay determined in the human genome that exceeded 50 kb, the maize genome has a higher resolution for LD-based association studies (12). Therefore, many more DNA markers are needed for genome-wide association study (GWAS) in maize. It is necessary to calculate the LD decay to determine the density of markers for GWAS in a species (13).

1.3. Development of Genetic Maps in Maize

Most of that traits important to agriculture are quantitative traits that exhibit normal distribution, which is in contrast to qualitative traits that are distributed in a discrete manner (14). The studies of quantitative traits have resulted in different molecular tools for enhanced plant genetic improvement , among which marker-assisted selection (MAS) became the most popular approach (15).

1.3.1. DNA marker technology

Molecular genetic mapping is based on various DNA polymorphisms and their associations with the gene loci controlling morphological traits. The earliest maize genetic mapping was conducted in 1980, in which the researchers used co-dominant hetero-multimeric isozymes to position the *alcohol dehydrogenase*-I (*Adh-I*) locus (16). In 1986, RFLP (restriction fragment length polymorphism) markers were developed and used to build a molecular genetic map in maize (17). Nearly 900 loci were mapped with RFLPs in maize (18). RAPD (randomly amplified polymorphic DNA) is PCR-based DNA markers developed following RFLPs. By using single primers of 10- nucleotide arbitrary sequences to amplify random DNA segments with PCR, RAPD markers have the potential to rapidly and readily detect nucleotide sequence polymorphisms (19). SSR (simple sequence repeat) markers were then developed in maize from SSR enriched libraries and by identification of SSR-containing sequences in public and private databases (20). A total of 1,051 SRR markers were identified and 978 of them were used to construct a high-resolution map. Since most SSRs are co-dominant markers, they can distinguish homozygotes from heterozygotes, thus providing more genotypic information for genetic mapping (21). AFLPs (amplified fragment length polymorphisms) are also PCR-based DNA markers that had been widely used for maize DNA fingerprinting. Using AFLPs, restriction fragments can be detected without the previous knowledge of the DNA sequences (22). In comparison among these DNA markers, SRRs have the largest number of alleles per locus while AFLPs have the highest assay efficiency index (23).

The abundance of Single nucleotide polymorphisms (SNPs) and InDels was initially revealed in the human genome. It was discovered that about 90% of polymorphisms in the human genome are single nucleotide variations and they were most related to functional differences (24). In maize, the development of SNP markers was mainly led by construction of the three-generation haplotype map (25, 26). It was found that two randomly chosen maize lines have an average of one SNP per 100 bp (27). Compared to SSR markers, SNPs are much more abundant and have higher heterozygosity, lower missing data rates, and higher repeatability (28). Thus, by leveraging the SNP technology, researchers could obtain increased marker quality and quantity for genetic mapping. Analysis of 632 maize inbred lines using 1,536 SNPs from 582 loci revealed that the LD decay ranged from 1 to 10 kb (29). From the 632 inbred lines, 60 core lines were identified that cover approximately 90% of the total variation. Therefore, broad and deep research of the maize SNPs would offer a great opportunity to improve the accuracy and efficiency of discovering new genes to enhance maize breeding.

1.3.2. QTL Mapping

Quantitative trait loci (QTLs) are the genomic regions that are tightly associated with the variation of quantitative traits. The rapid development of the DNA marker technology has broken the bottleneck of traditional genetic mapping, thus allowing the construction of a high-density genetic map for QTL mapping (30). The first QTL mapping study in crop species was completed in tomato using RFLPs to narrow the mapping resolution within 3 cM (31). RFLP-based QTL mapping in maize revealed that QTL

mapping with DNA markers had several advantages over the traditional QTL mapping, such as improved accuracy of QTL localization, increased mapping resolution, and detection of new QTLs (32). However, QTL mapping heavily relies on manpower and resources. The population size is emphasized in most QTL studies since either thousands of markers or thousands of experimental individuals are needed for QTL mapping (33). In this case, association mapping or QTL mapping with SNP markers provides a substantially improved method that can deal with this resource- and time-consuming processes.

1.3.3. Genome-wide Association Study (GWAS)

Genome-wide association study (GWAS) based on LD is a powerful approach to explore the genetic variation in a large population with higher mapping resolution (13). GWAS has been widely used in the genetic dissection of human diseases (34). The successful application of GWAS in humans has provided a great opportunity of using GWAS for genetic dissection of agronomic traits in crop species. A major obstacle of GWAS in plants probably is the structural population in plants that may lead to spurious associations (35). By estimating the population structure using SSRs, the first GWAS in maize successfully identified a set of polymorphisms in the *Dwarf8* sequence that was associated with flowering time (36). GWAS in maize has recently proliferated, due to the release of the maize B73 reference genome and the development of NGS (next-generation sequencing) technology (37). A variety of agronomic and economic traits have been studied with GWAS in maize and these advances suggested that GWAS is an efficient and reliable approach to explore QTLs controlling different agronomic traits (38).

2. MATERILS AND METHODS

2.1. Plant Materials

Two hundred eighty-two diverse inbred lines of the USDA maize GWAS panel were used for this study. This panel of inbred lines was constructed from all six subpopulations of maize, including stiff stalk (SS), non-stiff stalk (NSS), tropical-subtropical (TS), popcorn, sweet corn, and mixed, and represent the genetic variation and diversity of maize inbred lines present in the public maize breeding programs around the world (39).

2.2. Genotyping

The whole genomes of the 282 inbred lines of the USDA maize GWAS panel have been previously re-sequenced, with an average depth of 7x, using the Illumina HiSeq 2000platform. The original reads are deposited at NCBI under BioProject accession number PRJNA389800 in the BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/389800). We downloaded the sequence read archive (SRA) data from the database site and split them into the paired-end fasta format using the SRAtoolkit (40). The BWA aligner was then used to map the sequencing reads to the maize B73 reference genome (AGPv3) (41). Genome-wide SNPs were scanned with the BCFtools and sorted by chromosome positions (42). Among the 282 inbred lines analyzed, 263 were genotyped. The SNPs were filtered with minor allele frequency (MAF) ≤ 0.1, missing data ≥ 0.1, and base calling Phred quality score < Q30.

2.3. Phenotyping

The 263 maize inbred lines of the USDA maize GWAS panel were grown at the Experimental Farms of Texas A&M AgriLife Research near College Station, Texas from 2017 to 2019 and in Lubbock, Texas in 2018 for phenotyping. The Field trials were designed and carried out as described by Zhang(43). Randomized complete block design (RCBD) was employed with two replicates. Each plot contained one or two rows of 6.1 m spacing by 76.2 cm, with each row having 35 plants, making 33,500 plants per acre. The field practices, including weed control, irrigation, and fertilization, followed those used locally for maize field trials and production. When the ears ripened, the number of plants and number of ears were counted per plot, a section of five plants was randomly selected from the middle of each plot. Therefore, a total of ten plants were sampled from each inbred line for phenotyping. The ears of the selected plants were hand-harvested and naturally dried in a seed drying house. Seven ear traits were measured or counted, including Grain Weight per plant (GW), Ear Weight per plant (EW), Ear Length (EL), Kernel Row Length (KL), Ear Diameter (ED), number of Kernel Rows (KR), and Cob Diameter (CD). GW and EW were measured with an electronic balance; EL and KL were measured with an electronic ruler reader; ED and CD were measured with an electronic caliper, and KR was counted. The mean of each trait over two replicates was used for this study.

2.4. Phenotypic Data Analysis

Heritability (H^2) for each trait was calculated using a linear random effect model. Best linear unbiased prediction (BLUP) has been widely used for selection in plant and animal breeding to predict the average performance of a trait across multiple environments(44, 45). The value of a trait predicted by BLUP has also been used for QTL mapping and GWAS recently(46). Therefore, the value of each of the ear traits predicted by BLUP was used for GWAS in this study. Lines, years, and locations were defined as random effects to construct the linear model in R package "lme4" function (R 3.6.1):

$$\mathrm{Y} = (1 \mid \mathrm{Line}) + (1 \mid \mathrm{Loc}) + (1 \mid \mathrm{Year}) + (1 \mid \mathrm{Line{:}Year}) + (1 \mid \mathrm{Line{:}Loc})$$

where Y is the normalized mean value of the targeted trait; Line represents the 263 inbred lines; Loc is the two locations where the field trials were performed; Year is for the three planting years of the field trials from 2017 to 2018; and ":" indicates the interaction between two random effects. Based on the fitted linear model, the broad-sense heritability (H^2) was calculated by:

$$\mathrm{H}^2 = \mathrm{V_G}/(\mathrm{V_G} + \mathrm{V_{GL}}/\mathrm{L} + \mathrm{V_{GY}}/\mathrm{Y} + \mathrm{V_e}/\mathrm{YL})$$

The correlation between two traits was calculated with the R package (R 3.6.1).

2.5. Population Structure Analysis

SNPs were further pruned with the variance inflation factor (VIF) 1.5 that was calculated on PLINK v1.07(47). After pruning, 720,651 SNPs with VIF larger than 1.5 remained, since a larger VIF indicates a higher level of LD (48). The population Q matrix was calculated based on this subset of 720,651 SNPs using ADMIXTURE (49). The

mixed, popcorn, and sweet corn groups were extracted before determining the number of K because previous studies revealed the estimation bias from the mixed population (50). The structure plot is drawn in R packages (R3.6.1). Principle component analysis (PCA) was calculated with GCTA (51). Phylogenetic tree was constructed by MEGAX. Kinship matrix was constructed with the GAPIT software to stratify the population pedigree(52).

2.6. Association Analysis

Both general linear model (GLM) and mixed linear model (MLM) were used for GWAS. We used both GAPIT and GEMMA to model MLM and only used GAPIT for GLM(53-55). Markers (S) and population structure (PCs) were set to fixed effects:

$$y = \mu + \alpha S + \nu PCs + \epsilon$$

In MLM, markers (S) and first 3 principle components (PCs) were served as the fixed effects, while kinship matrix (K) was set to random effect:

$$y = \mu + \alpha S + \nu PCs + K + \epsilon$$

The Q-Q plot was generated in R to evaluate the fitness of the linear models. The best performing model was used for GWAS. Bonferroni correction methods are used for correcting false discoveries in multiple hypothesis testing(56).

3. OBJECTIVES

The objectives of this study were four-fold:

1. Genotype the 263 diverse inbred lines of the USDA maize GWAS panel with genome-wide SNPs using the reads of their re-sequenced genomes (7x);
2. Phenotype the 263 inbred lines of the USDA maize GWAS panel for seven ear traits through field trials across multiple years and locations in Texas;
3. Estimate the population structure and re-construct the phylogenetic tree of the maize GWAS panel; and
4. Dissect the genetic basis of maize ear traits by GWAS using genome-wide high-density SNPs.

4. RESULTS

4.1. Analysis of Phenotypic Data

The 263 inbred lines of the USDA maize GWAS panel were phenotyped across four environments for seven ear traits, Grain Weight per plant (GW), Ear Weight per plant (EW), Ear Length (EL), Kernel-row Length (KL), Ear Diameter (ED), number of Kernel Rows (KR), and Cob Diameter (CD).

The phenotypic data of the ear traits collected from the four environments were normalized to their average values with BLUP. The prediction accuracy of the traits with BLUP R^2, the square of correlation coefficient between the means of the observed traits across environments and the estimated values of the traits with BLUP ranged from 0.8512 to 0.9875 (**Fig. 1**), indicating that BLUP model properly predicted the average values of the traits across the four environments.

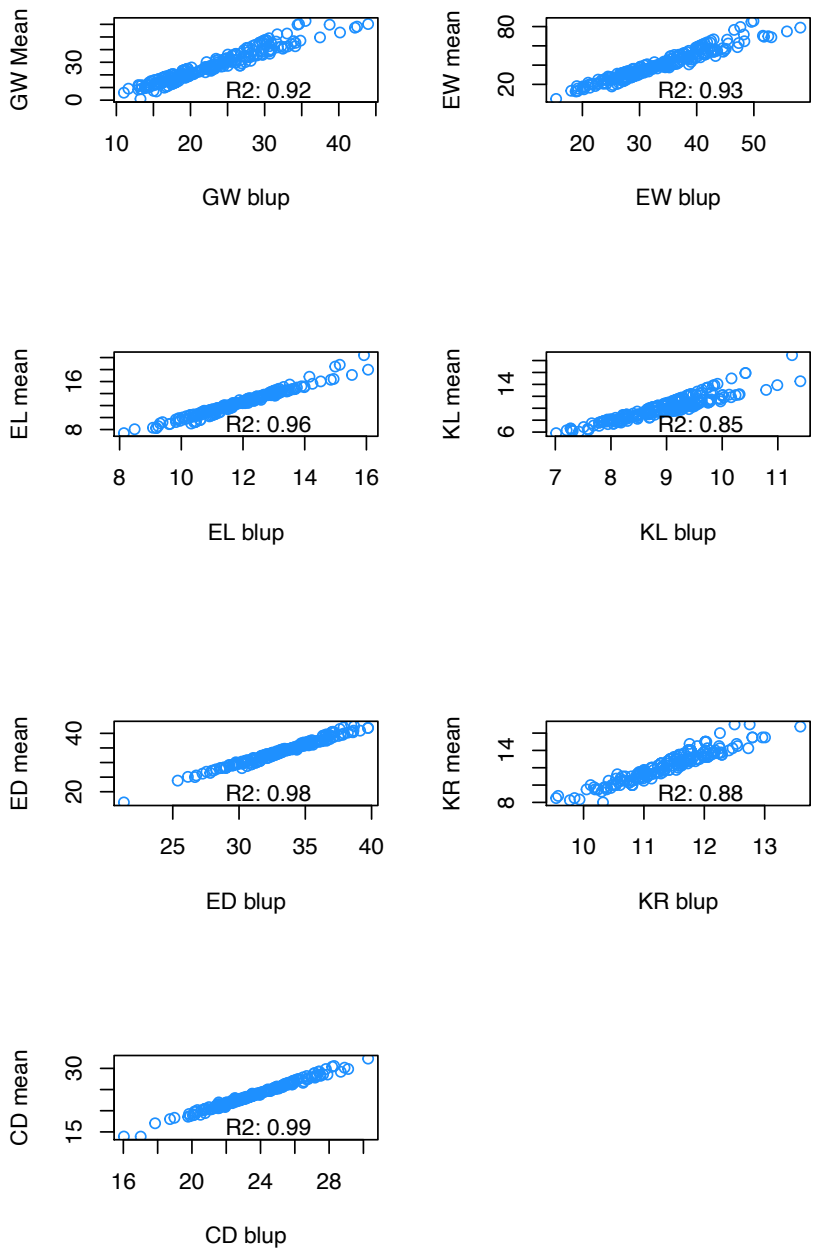

Figure 1 Prediction accuracy of the ear traits with BLUP. GW, grain weight per plant (g); EW, ear weight per plant (g); EL, ear length (cm); KL, kernel-row length (cm); ED, ear diameter (mm); KR, number of kernel rows (n); CD, cob diameter (mm).

The average grain weight per plant (GW) was 23.04 g, with a range from 11.02 g to 43.95g and the largest CV of 28.19% = 6.496, while the mean number of kernel rows (KR) was 11.40, with a range from 9.50 to 13.60 and the smallest CV of 6.14% (**Table 1**). The broad-sense heritability (H^2) was calculated based on the BLUP values of the traits. It ranged from 0.50 for the number of kernel rows (KR) to 0.87 for cob diameter (CD). In comparison, ear length (EL), ear diameter (ED), and cob diameter (CD) had a higher heritability ($H^2 > 0.8$), while kernel-row length (KL), and the number of kernel rows (KR) had a lower heritability ($H^2 < 0.6$).

Table 1 Ear traits predicted with BLUP and their heritability. GW, grain weight per plant (g); EW, ear weight per plant (g); EL, ear length (cm); KL, kernel-row length (cm); ED, ear diameter (mm); KR, number of kernel rows (n); CD, cob diameter (mm).

Trait	**Unit**	**Min**	**Mean**	**Max**	**SD**	**CV (%)**	**Heritability (H^2)**
GW	g	11.02	23.04	43.95	6.50	28.19	0.66
EW	g	15.40	33.32	58.14	8.06	24.19	0.65
EL	cm	8.14	11.87	16.06	1.31	11.04	0.81
KL	cm	7.01	8.94	11.40	0.76	8.50	0.54
ED	mm	21.31	33.16	39.74	2.91	8.76	0.84
KR	n	9.50	11.40	13.60	0.70	6.14	0.50
CD	mm	16.03	23.67	30.28	2.25	9.50	0.87

The seven ear traits studied, with a total of 21 pairs of the traits, were all significantly and positively correlated ($P \leq 0.05$), except for the trait pairs between KL and KR, and between KL and CD. Of the 19 trait pairs that were significantly correlated, GW and EW were

most correlated, with $R^2 = 0.930$ (**Table 2**). The second group of ear trait pairs that were most correlated were between ED and CD, EW and ED, EL and KL, GW and ED, and EW and KL, with $r = 0.7 - 0.8$. The remaining 13 trait pairs had a significant correlation ranging from $r = 0.680$ down to $r = 0.141$.

Table 2 Pairwise correlations between the seven ear traits. GW, grain weight per plant (g); EW, ear weight per plant (g); EL, ear length (cm); KL, kernel-row length (cm); ED, ear diameter (mm); KR, number of kernel rows (n); CD, cob diameter (mm).

	GW	EW	EL	KL	ED	KR	CD
GW	1.000**	0.930**	0.358**	0.680**	0.726**	0.510**	0.383**
EW	0.930**	1.000**	0.454**	0.702**	0.774**	0.562**	0.485**
EL	0.358**	0.454**	1.000**	0.766**	0.154*	0.107	0.108
KL	0.680**	0.702**	0.766**	1.000**	0.365**	0.342**	0.141*
ED	0.726**	0.774**	0.154*	0.365**	1.000**	0.629**	0.784**
KR	0.510**	0.562**	0.107	0.342**	0.629**	1.000**	0.458**
CD	0.383**	0.485**	0.108	0.141*	0.784**	0.458**	1.000**

** P-value < 0.0001
* P-value < 0.05

Due to the existence of missing data for some inbred lines in one or more of the four environments, some distribution skewness of the phenotypic BLUP data were observed for the ear traits. However, the histogram of the traits showed that the normality assumptions could be satisfied for GWAS (**Fig. 2**).

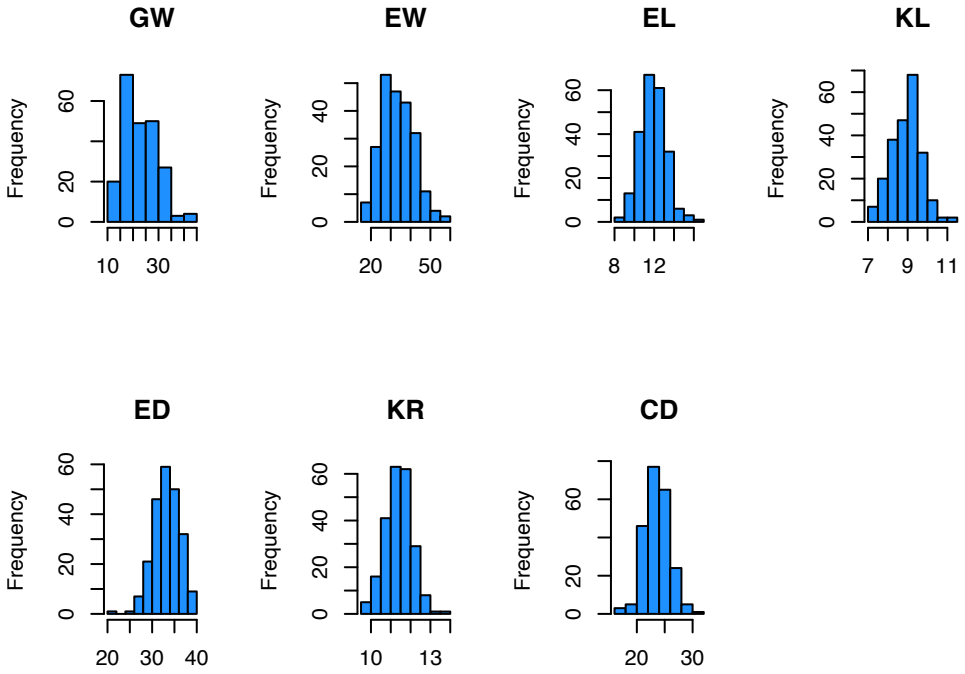

Figure 2 Distribution histogram of the ear traits predicted with BLUP.

4.2. Genotypic Data Quality control and filtering

The total of 7,518,820 SNPs was called for the 263 inbred lines studied for this study, after their sequencing reads were mapped to the maize B73 reference genome. Then, the SNPs were filtered by Phred quality score, missing rate per location, missing rate per individual, and minor allele frequency (MAF), As a result, a total number of 1,537,073 quality SNPs were identified, which well covered all 10 maize chromosomes (**Fig. 3**). The majority of SNPs were filtered out because of high missing rates, which could result from the low read mapping rate of different inbred lines to the B73 reference genome.

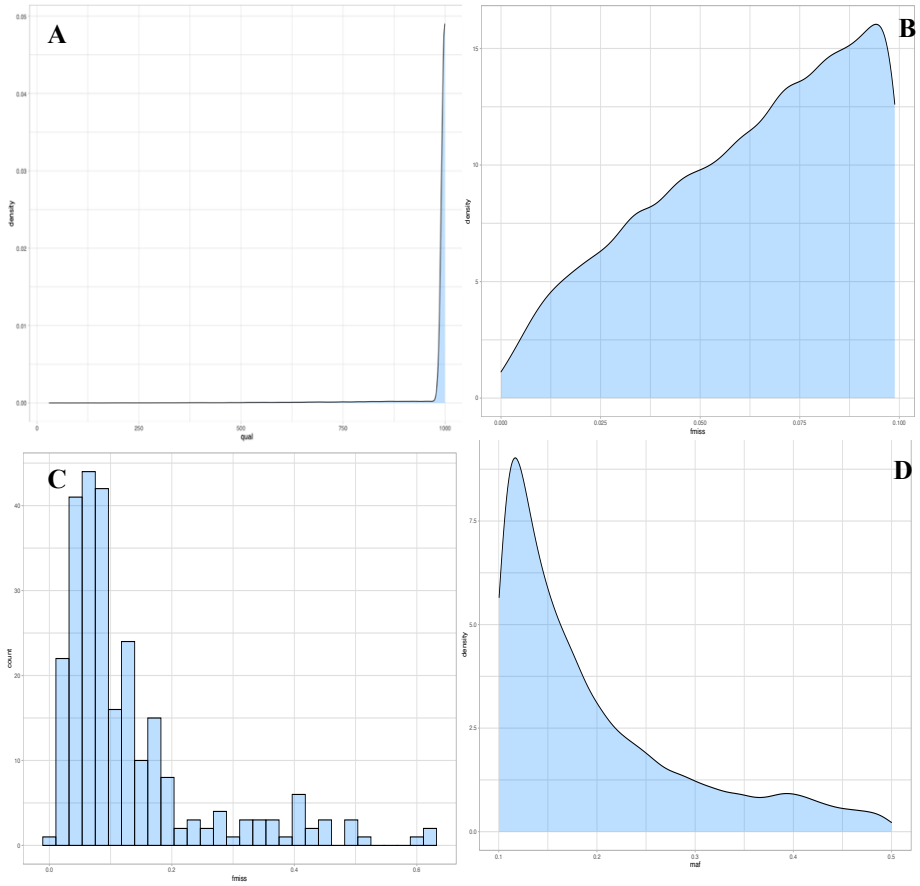

Figure 3 (A) Phred quality score after filtering through Q30; (B-C) Missingness after filtering through missing rate ⩾ 0.1; (D) Minor allele frequency (MAF) after filtering through MAF ⩽ 0.1.

4.3. Population Structure

We examined the population structure and phylogeny of the 263 inbred lines used in this study based on the 1,537,073 quality SNPs using different methods. When the SNP data were used in population structure analysis of the inbred lines and the popcorn, sweet corn, and mixed inbred lines were excluded from the analysis, the number of subpopulations, K = 3, was had the smallest cross-validation error rate 0.45276 (**Fig. 4A**), suggesting that the remaining inbred lines consisted of three subpopulations. Principle component analysis (PCA)

showed a clustering pattern of six subpopulations. According to these analysis results, we initially concluded that the 263 maize inbred lines consisted of subpopulations: non-stiff stalk (NSS), stiff stalk (SS), tropical/subtropical (TS), popcorn, sweet corn, and mixed (**Fig. 4B**).

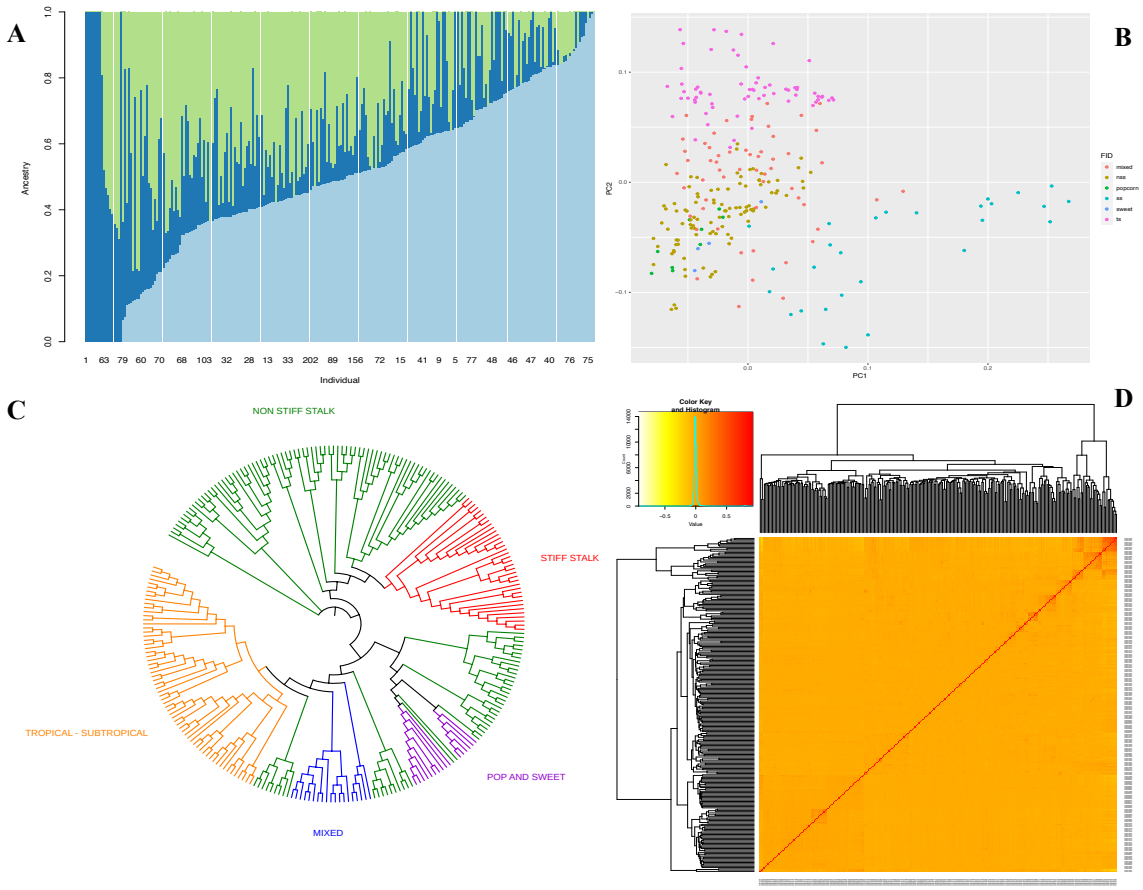

Figure 4 (A) Population structure plot of the 263 inbred lines with k = 3; (B) PCA with six subpopulations; (C) Phylogenetic tree of the six subpopulations; (D) Kinship heat map.

Furthermore, we re-constructed the phylogenetic tree of the 263 inbred lines. The result validated the population structure of the inbred lines achieved above that the 263 maize inbred lines consisted of six subpopulations (**Fig. 4C**). Finally, the heat map of the kinship

matrix suggested that the genetic backgrounds of the inbred lines were not uniformly distributed (**Fig. 4D**). Therefore, it is included that the population structure and kinship matrix should be included in the GWAS to false positives.

4.4. Genome-wide Association Study (GWAS)

We first estimated the LD decay. According to the coefficient square of correlation (R^2) between each marker pairs, we found that the density of SNP markers was highly enriched within 100,000 bp (**Fig. 5 Upper**) and that the genome-wide LD decayed rapidly within 5,000 base pairs (**Fig. 5 Lower**). The LD-decay distances are usually affected by the domestication level and population structure of plant lines used. Since the inbred population we used are highly diverse, it is possible that the LD decay distance is relatively small. Maize has an average genome size of 2.5 Gb. As maize has an average genome size of 2.5 Gb, its genome likely contains approximately 500,000 LD decay segments (2.5 Gb/5kb = 500,000). Thus, the 1,553,207 SNP marker density should be enough to capture genes or loci controlling the ear traits.

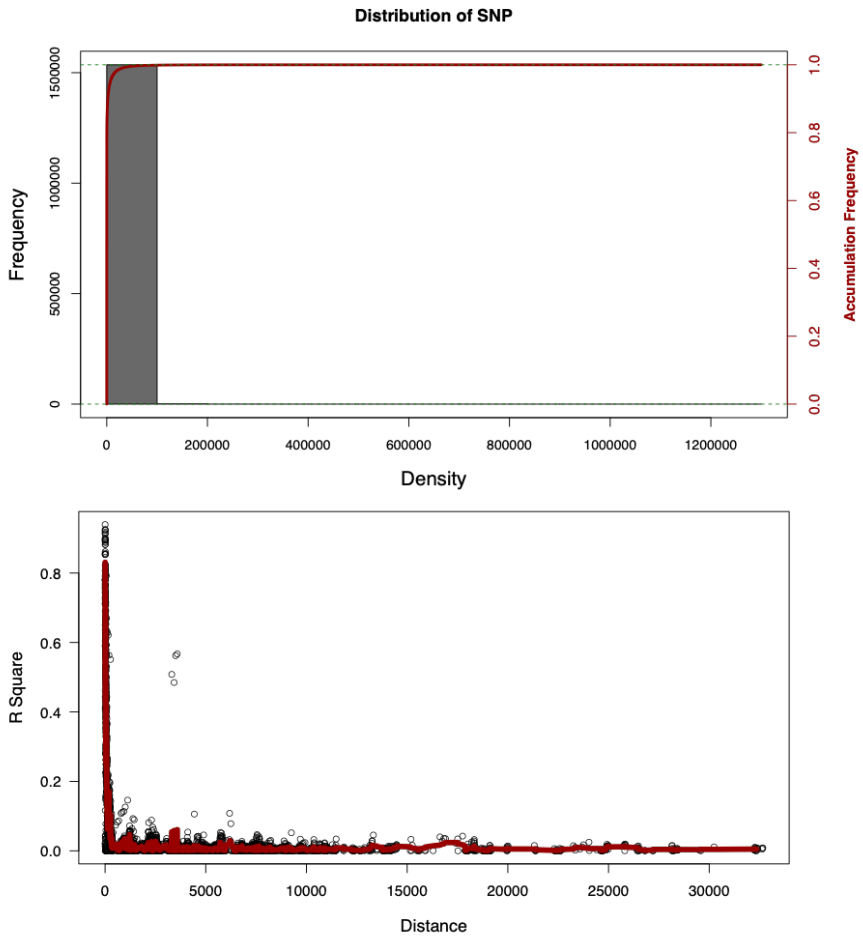

Figure 5 (A) Marker Density Plot; (B) Linkage Disequilibrium decay plot.

The GLM are constructed in GAPIT. The MLM are generated both in GAPIT and GEMMA. We further evaluated the fitness of each model based on Q-Q plot. The fitness of GLM shows a quite great pattern with observed P-value deviate above the expected values (**Fig. 6**). However, in MLM, the observed P-values are distributed below our expectations (**Fig. 7**) According to that, the MLM is suspected to be over-corrected and thus not suitable for our data. To see if this pattern only exists in GAPIT, MLM was also constructed using

GEMMA. Same over-corrected patterns were observed from the new MLM (**Fig. 8**). Since GLM shows a better fit, QTLs are screened based on this model.

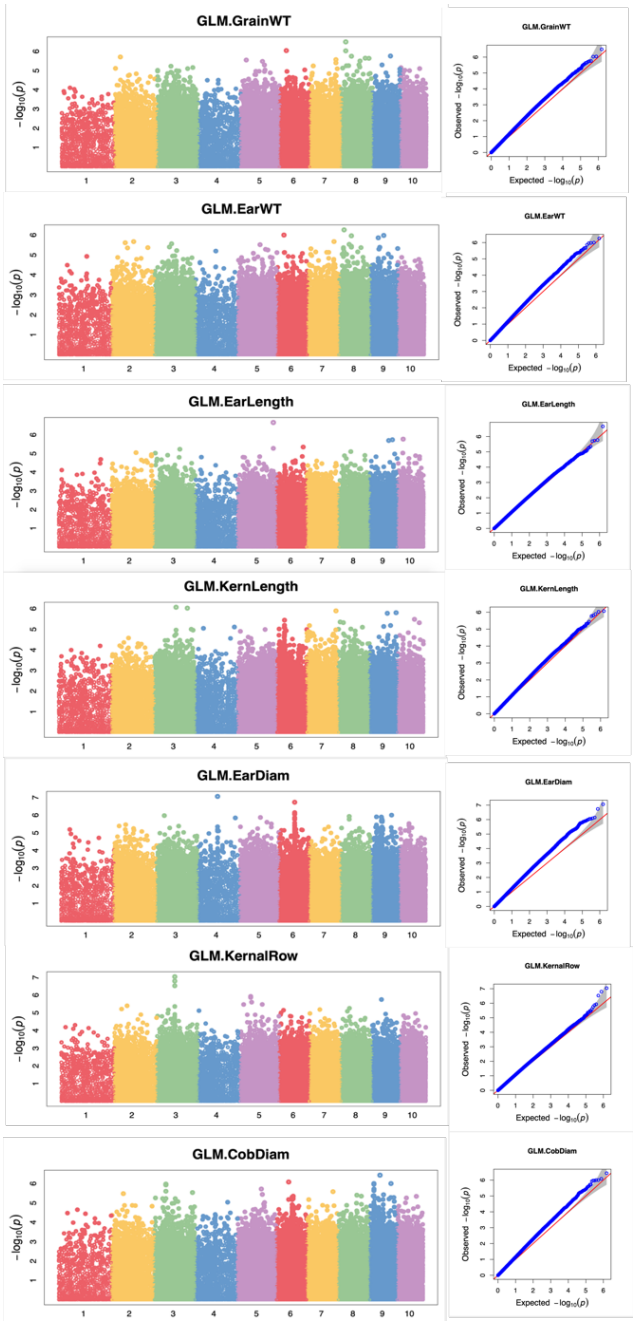

Figure 6 Manhattan plots and Q-Q plots of GWAS GLM model generated by GAPIT.

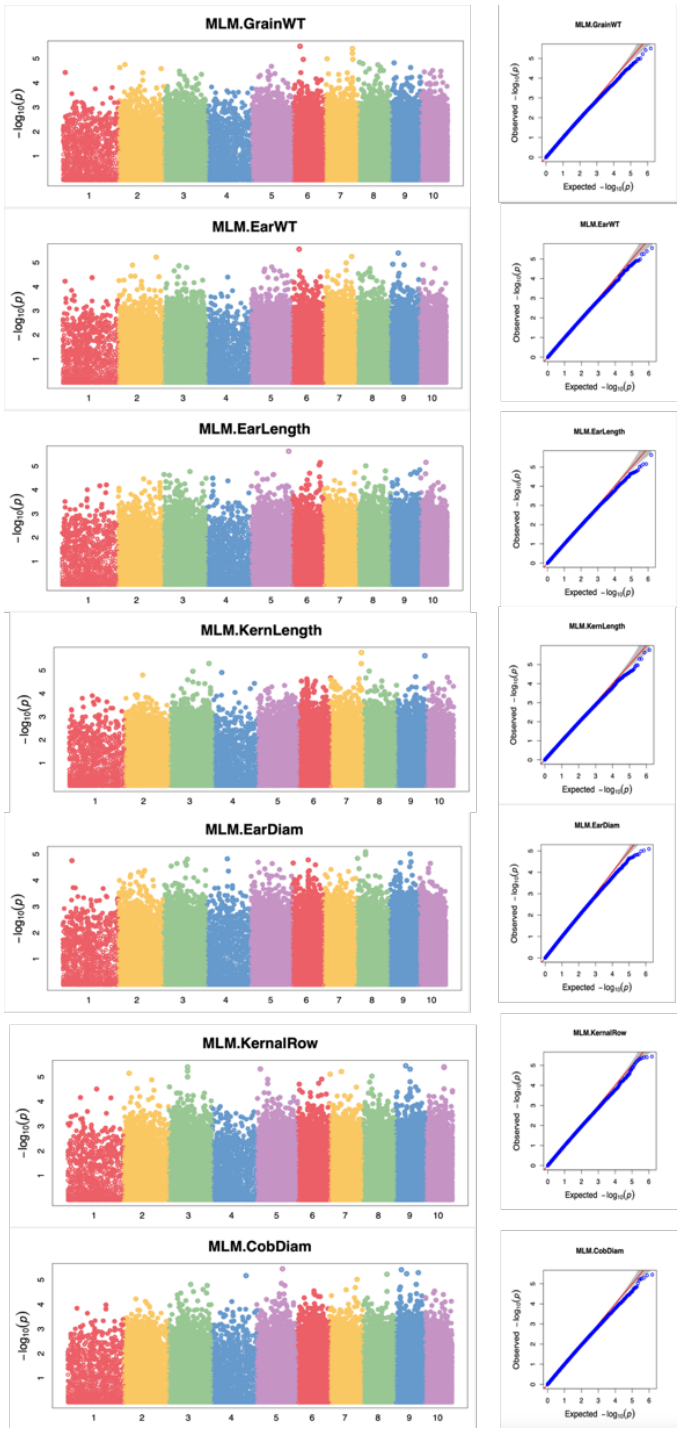

Figure 7 Manhattan plots and Q-Q plots of GWAS MLM model generated by GAPIT.

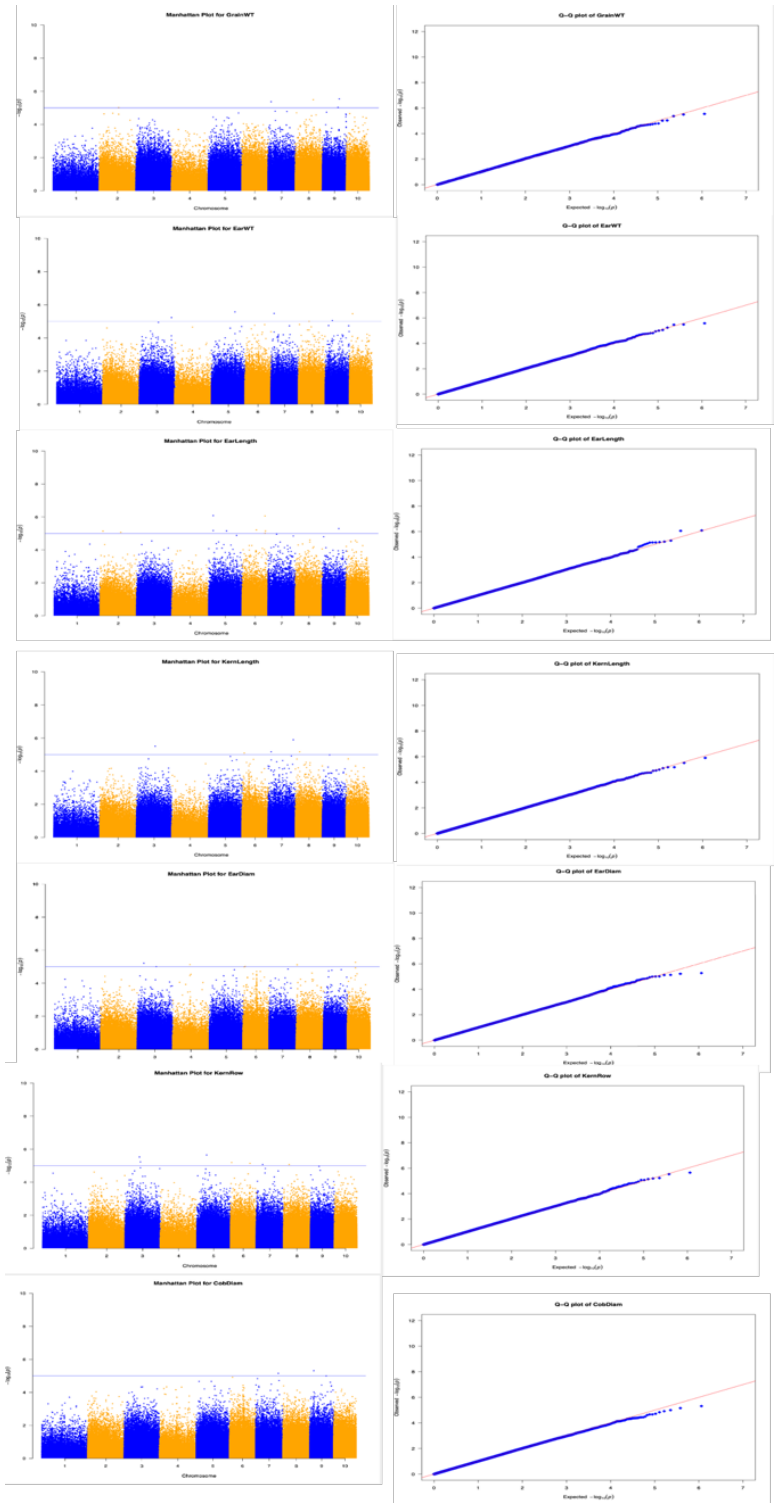

Figure 8 Manhattan plots and Q-Q plots of GWAS MLM model generated by GEMMA.

With the GLM, two SNPs were identified to be associated with KR at -log(*P*-value) = 7.04 and 6.80, respectively, or FDR-adjusted *P*-value = 0.123 (**Table 3**). One SNP was identified to be associated with ED at -log(*P*-value) = 7.06 or FDR-adjusted *P*-value = 0.134. The two SNPs associated with KR were located at Positions 103416743 and 103416737 of Chromosome 3 (**Fig. 9**). The SNP associated with ED was located at Position 116703054 of Chromosome 4.

Table 3 Summary of SNPs that are associated with ear traits.

SNP	Chr	Position	P-value	maf	nobs	FDR P-values	Trait
3.103416743	3	103416743	9.16E-08	0.14	225	0.123096369	KR
3.103416737	3	103416737	1.60E-07	0.14	225	0.123096369	KR
4.116703054	4	116703054	8.71E-08	0.46	225	0.133687704	ED

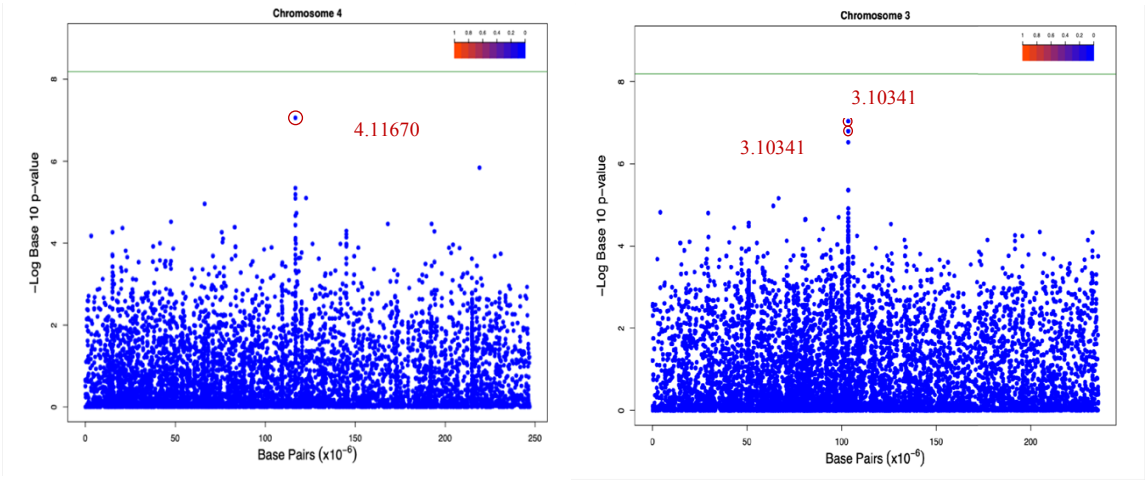

Figure 9 Manhattan plot on chromosome 4 and chromosome3. Red circled showing significant SNPs.

Near the SNP 3.103416743 and 3.103416737, two unannotated genes were detected within 20kb (**Table 4**). One SNP related to Plant Height were found to be previously observed in the same population. 3 unannotated genes were identified near the SNP 4.116703054. Two SNPs were also called near this SNP in the same population.

Table 4 Genes and SNPs that are located near the QTLs.

SNP	Name	Type	Position	Distance
3.103416743	Zm00001e017879	gene	chr3:103383572..103400198	16545
3.103416743	Zm00001e017880	gene	chr3:103617774..103618178	201031
3.103416743	Height_per_day	SNP	chr3:103387298..103387397	29346
4.116703054	Zm00001e022679	gene	chr4:116640546..116642544	62508
4.116703054	Zm00001e022680	gene	chr4:116640546..116642544	62508
4.116703054	Zm00001e022681	gene	chr4:116854247..116856347	151193
4.116703054	Nodes_to_ear	SNP	chr4:116792354..116792453	89300
4.116703054	Sucrose	SNP	chr4:116967543..116967642	264489

5. DISCUSSION

5.1. Phenotypic Data

Among seven ear traits, ear length, ear diameter and cob diameter have a relatively high broad-sense heritability as 0.81, 0.84 and 0.87 respectively. The results are in accordance with the previous studies where broad-sense heritability (H_m^2) was calculated as 0.87, 0.78, and 0.82(39). It further proved that the performance of these three traits are quite stable and mainly affected by genotypic variance. In contrast, the other traits with low heritability are largely affected by the environmental factors as well as the measurement approaches. In this study, two environmental factors are counted toward the total variance of phenotypic data. In the BLUP model, we observed that the effect of multiple locations is much bigger than that of the years. In grain weight data, variation explained by the locations are about 8-fold larger than the variation explained by years (**Table 5**). Similar fractions are also observed in other traits. It should have implied that our two experimental fields College Station and Lubbock have quite large environmental differences that caused the large variation through different locations.

Table 5 Variance of random effects in GW BLUP model.

Groups	Variance	SD
SRR:Year	43.28	6.578
SRR:Loc	31.4	5.604
SRR	79.89	8.938
Year	20.74	4.555
Loc	114.65	10.707
Residual	66.90	8.179

5.2. Population Structure

The phylogenetic tree of 263 maize inbred lines showed similar patterns with previous studies(50). The figure was derived from DNA microsatellites or simple sequence repeats (SSRs) on 260 inbred lines. Non-Stiff Stalk, Stiff Stalk and Tropical-Subtropical constitute the majority subtypes of the whole population, in which Non-Stiff Stalk has a mixed pattern with popcorn and sweetcorn. Besides that, the PCA and Structure results also indicate the above clustering subtypes. Such good agreement of pedigree information of maize inbred lines provides the solid foundation for downstream association analysis.

5.3. Mapping Quality

The missing data rate in the mapped sequences are quite high. This can be due to the low coverage of the sequencing depth or to the low representative of mono-reference. A definition of pangenome was previously put out that all strains of a species only share a certain

amount of genome(57). The genes appeared in all strains are considered as core genome, while those only exist in some strains are called dispensable genome. The feature of pangenome has been revealed in various crops including maize, rice, soybean and wheat. In plants, the core genome represents only 40 to 80% of the total pangenome. Evidence from maize further showed that only half of the genomic structure of B73 and Mo17 was conserved between two individuals(58). Single reference may neglect many information that can be captured by dispensable genome(59). Thus, the B73 assembly reference we used in this study may be insufficient to cover all the genes shared with multiple subtypes, thereby producing high missingness after mapping. It can be optimized if we get access to the de novo assembly pangenome reference to identify variations across different strains.

5.4. Selection of Candidate Genes

The FDR and Bonferroni corrections are the two widely used approaches for correcting false positive rates in the multiple hypothesis testing. However, relevant studies have shown that the thresholds for GWAS should be flexible and depend on many factors such as LD and MAF (60). The traditional Bonferroni correction methods 0.05/n are considered to be too conservative for GWAS and may abandon true positive results by assuming all the individuals are independent(61). In the cases of having large data set, many people prefer to use FDR for correcting multiple hypothesis testing. In this study, we used a threshold of $-\log_{10}$(P-value) = 7 to identify the top associated genes. This value was set according to a looser Bonferroni Threshold 0.1/n (n=1,553,207).

In the above threshold, we captured 5 candidate genes that are in the LD distances of QTLs. However, the functions of those genes haven't been characterized yet. By using sequence information to conduct nucleotide blast, I still haven't been able to find homolog genes. Thus, particular experimental analysis is further required to identify the function of the candidate genes, and validate their associations.

6. CONCLUSION

In this study, we phenotyped seven ear traits in 225 maize core association panels across three years and two locations. Set the B73 assembly as single reference, we successfully genotyped 263 inbred lines. After filtering, we totally got 1,553,207 SNPs. Population structure analysis based on the SNP data indicated a clear clustering pedigree information of U.S. core maize germplasms with six subtypes: NSS, SS, TS, Sweet corn, popcorn and mixed. Using a General Linear Model with kinship matrix as covariates, we identified 3 associated QTLs at significance level $-\log_{10}$(P-value) = 7: two associated with KR and one with ED. 5 candidate genes were found near the three QTLs. The function of specific genes can be further validated by experimental approaches. These genetic markers in the core maize inbred lines can be further utilized in genomic prediction and molecular breeding, thereby facilitate the breeding process.

REFERENCES

1. Faostat F. Statistical databases. Food and Agriculture Organization of the United Nations. 2009.

2. Shiferaw B, Prasanna BM, Hellin J, Bänziger M. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. Food Security. 2011;3(3):307.

3. Ranum P, Peña‐Rosas JP, Garcia‐Casal MN. Global maize production, utilization, and consumption. Annals of the New York Academy of Sciences. 2014;1312(1):105-12.

4. Degefa K, Jaleta M, Legesse B. Economic efficiency of smallholder farmers in maize production in Bako Tibe district, Ethiopia. Developing Country Studie. 2017;7(2):80-6.

5. Kimm H, Guan K, Gentine P, Wu J, Bernacchi CJ, Sulman BN, et al. Redefining droughts for the U.S. Corn Belt: The dominant role of atmospheric vapor pressure deficit over soil moisture in regulating stomatal behavior of Maize and Soybean. Agricultural and Forest Meteorology. 2020;287:107930.

6. Rosegrant M, Ringler C, Sulser TB, Ewing M, Palazzo A, Zhu T, et al. Agriculture and food security under global change: Prospects for 2025/2050. International Food Policy Research Institute, Washington, DC. 2009:145-78.

7. Wallington TJ, Anderson JE, Mueller SA, Kolinski Morris E, Winkler SL, Ginder JM, et al. Corn Ethanol Production, Food Exports, and Indirect Land Use Change. Environmental Science & Technology. 2012;46(11):6379-84.

8. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. Science. 2009;326(5956):1115-7.

9. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 2001;98(20):11479-84.

10. Flint-Garcia SA, Thornsberry JM, Buckler IV ES. Structure of linkage disequilibrium in plants. Annual review of plant biology. 2003;54(1):357-74.

11. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences. 2001;98(20):11479-84.

12. Koch HG, McClay J, Loh EW, Higuchi S, Zhao JH, Sham P, et al. Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. Human Molecular Genetics. 2000;9(20):2993-9.

13. Yu J, Buckler ES. Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology. 2006;17(2):155-60.

14. Mackay TFC. The Genetic Architecture of Quantitative Traits. Annual Review of Genetics. 2001;35(1):303-39.

15. Collard BC, Jahufer M, Brouwer J, Pang E. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica. 2005;142(1-2):169-96.

16. Birchler JA. The cytogenetic localization of the alcohol dehydrogenase-1 locus in maize. Genetics. 1980;94(3):687-700.

17. Weber D, Helentjaris T. Mapping RFLP loci in maize using BA translocations. Genetics. 1989;121(3):583-90.

18. Hoisington DA, Coe EH. Mapping in Maize Using RFLPs. In: Gustafson JP, editor. Gene Manipulation in Plant Improvement II: 19th Stadler Genetics Symposium. Boston, MA: Springer US; 1990. p. 331-52.

19. Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research. 1990;18(22):6531-5.

20. Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, et al. Development and mapping of SSR markers for maize. Plant molecular biology. 2002;48(5-6):463-81.

21. Forster B, Russell J, Ellis R, Handley L, Robinson D, Hackett C, et al. Locating genotypes and genes for abiotic stress tolerance in barley: a strategy using maps, markers and the wild species. The New Phytologist. 1997;137(1):141-7.

22. Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, et al. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research. 1995;23(21):4407-14.

23. Pejic I, Ajmone-Marsan P, Morgante M, Kozumplick V, Castiglioni P, Taramino G, et al. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. Theoretical and Applied genetics. 1998;97(8):1248-55.

24. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome research. 1998;8(12):1229-31.

25. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics. 2012;44(7):803-7.

26. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third-generation Zea mays haplotype map. Gigascience. 2018;7(4):1-12.

27. Zhao W, Canaran P, Jurkuta R, Fulton T, Glaubitz J, Buckler E, et al. Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic acids research. 2006;34(suppl_1):D752-D7.

28. Jones E, Sullivan H, Bhattramakki D, Smith J. A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.). Theoretical and Applied Genetics. 2007;115(3):361-71.

29. Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One. 2009;4(12):e8451-e.

30. Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica. 2005;142(1):169-96.

31. Paterson AH, DeVerna JW, Lanini B, Tanksley SD. Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. Genetics. 1990;124(3):735-42.

32. Edwards MD, Helentjaris T, Wright S, Stuber CW. Molecular-marker-facilitated investigations of quantitative trait loci in maize. Theoretical and Applied Genetics. 1992;83(6):765-74.

33. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 1990;124(3):743-56.

34. Alghamdi J, Padmanabhan S. Chapter 12 - Fundamentals of Complex Trait Genetics and Association Studies. In: Padmanabhan S, editor. Handbook of Pharmacogenomics and Stratified Medicine. San Diego: Academic Press; 2014. p. 235-57.

35. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. The American Journal of Human Genetics. 2000;67(1):170-81.

36. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics. 2001;28(3):286-9.

37. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science. 2009;326(5956):1112-5.

38. Xiao Y, Liu H, Wu L, Warburton M, Yan J. Genome-wide Association Studies in Maize: Praise and Stargaze. Molecular Plant. 2017;10(3):359-74.

39. Flint‐Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high‐resolution platform for quantitative trait locus dissection. The Plant Journal. 2005;44(6):1054-64.

40. Staff S. Using the sra toolkit to convert. sra files into other formats. National Center for Biotechnology Information (US). 2011.

41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

42. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

43. Zhang M, Cui Y, Liu Y-H, Xu W, Sze S-H, Murray SC, et al. Accurate prediction of maize grain yield using its contributing genes for gene-based breeding. Genomics. 2020;112(1):225-36.

44. Robinson GK. That BLUP is a good thing: the estimation of random effects. Statistical science. 1991;6(1):15-32.

45. Piepho H, Möhring J, Melchinger A, Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. 2008;161(1-2):209-28.

46. Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang G, et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nature Genetics. 2018;50(6):803-13.

47. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

48. VanLiere JM, Rosenberg NA. Mathematical properties of the r2 measure of linkage disequilibrium. Theor Popul Biol. 2008;74(1):130-7.

49. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19(9):1655-64.

50. Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J. Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites. Genetics. 2003;165(4):2117.

51. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76-82.

52. VanRaden PM. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science. 2008;91(11):4414-23.

53. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics. 2006;38(2):203-8.

54. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics. 2010;42(4):355-60.

55. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.

56. Weisstein EW. Bonferroni correction. https://mathworld wolfram com/. 2004.

57. Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D. Plant pangenomics: approaches, applications and advancements. Current Opinion in Plant Biology. 2020;54:18-25.

58. Zapparoli E. Identification of structural variation in Zea mays: use of paired-end mapping and development of a novel algorithm based on split reads 2017.

59. Hurgobin B, Edwards D. SNP discovery using a pangenome: has the single reference approach become obsolete? Biology. 2017;6(1):21.

60. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

61. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association. 1967;62(318):626-33.