GWAS

Genome-Wide Association Studies

By

T. Anusha

(2018-11-143)

MSc. Plant Breeding and Genetics

Seminar report

Submitted in partial fulfilment of requirement of the course

PBGN. 591, Seminar (0+1)

Department of Plant Breeding and Genetics

College of Horticulture

Kerala Agricultural University

Vellanikkara - 680656

Thrissur, Kerala

DECLARATION

I, T. Anusha (2018-11-143) hereby declare that the seminar report titled ‘GWAS- Genome-Wide Association Studies’ has been completed by me independently after going through the references cited here and I haven’t copied from any of the fellow students or previous seminar reports.

Vellanikkara	T. Anusha
Date: 24 /1/ 2020	(2018-11-143)

CERTIFICATE

This is to certify that seminar report titled 'GWAS- Genome-Wide Association Studies' for the course PBGN. 591 has been solely prepared by T. Anusha (2018-11-143) under my guidance, and she has not copied from seminar reports of seniors, juniors or fellow students.

Vellanikkara

Date: 24/1/2020

Dr. P. Sindhumole
(Major Advisor)
Assistant Professor
Dept. of Plant Breeding and Genetics
Vellanikara, Thrissur

CERTIFICATE

Certified that the seminar report entitled ‘GWAS- Genome-Wide Association Studies’ is a record of seminar presented by T. Anusha (2018-11-143) on 9th January, 2020 and is submitted for the partial requirement of the course PBGN. 591.

Dr. Anil Kuruvilla
Professor
Dept. of Agricultural Economics
College of Horticulture,
Vellanikkara

Dr. Reshmy Vijayaraghavan
Assistant Professor,
Dept. of Plant Pathology
College of Horticulture,
Vellanikkara

Dr. Sangeetha Kutty M.
Assistant Professor,
Department of Vegetable Science,
College of Horticulture,
Vellanikkara

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PLATES

GWAS

Genome-Wide Association Studies

1. INTRODUCTION

Natural variation is a valuable and sustainable resource of the phenotypic and genetic diversity within plant species worldwide that offer, beneficial traits for plant breeding. The phenotypic variation within-species caused by spontaneously natural genetic mutations that maintained in nature by evolutionary, artificial and natural selection processes (Blanco *et al*, 2009). Natural variation brought great advances to understand crop morphology and their response to biotic and abiotic stresses. The understanding of natural variation in crop plants through thousands of years for domestication e.g. in barley about 10,000 years ago (Badr *et al*, 2007) can be seen in the genetic modification of developmental traits and adaptive features. Natural variation studies in wild species elucidated the molecular basis of phenotypic differences related to domesticated plant adaptation that is important to interpret the maintenance and evolutionary significance of phenotypic variation (Olds *et al*, 2007).

The causal relationship between genetic polymorphism within a species and the phenotypic differences observed between individuals is of fundamental biological interest. The ability to identify variations associated with disease and agronomically important traits like growth rate and yield in plants requires an understanding of both the specific loci that underlie a phenotype and the genetic architecture of a trait. Genome-Wide Association Studies (GWAS) present a powerful tool to reconnect this trait back to its underlying genetics.

1.1 GWAS- Genome-Wide Association Studies

Genome-Wide Association Studies (GWAS) or Whole Genome Association Studies (WGAS) or Common Variant Association Studies (CVAS) investigate a genome-wide set of genetic variants in different varieties to see if any variant is associated with a trait (Manolio, 2010).

The recent advances in DNA sequencing paved the way to genetically improve the important traits (grain quality, biotic and biotic stress tolerance, *etc*. Chip-based microarray technology (Kumar et al., 2015), Illumina (San Diego, California), 90k illumina iselect array (Wang *et al*, 2014), Affymetrix (Distefano and Taverna, 2011), or other next-generation sequencing technologies (Neelapu and Surekha, 2016) *e.g.* genotyping-by-sequencing (GBS), provide thousands of single nucleotide polymorphism (SNPs) covering the most genomic region in chromosomes. Many powerful statistical genetics methods were proposed to identify alleles controlling target traits. Genome-wide association study (GWAS) is one of those useful methods and it is successfully used to identify candidate genes for many important traits and tests the association between the marker type (*e.g* SNP) and the phenotype of a target trait. There are many considerations and recommendations that should be taken into account when geneticists decide to perform GWAS.

2. Working of GWAS

To conduct a GWAS experiment, the first step is to select the population of study with a full consideration of the size of the population (minimum 100 individuals) with preference to increase the number of individuals as much as possible to avoid Beavis effects that lead greatly overestimated of phenotypic variance when the number of individuals are small *e.g.* 100 (Xu, 2003). Then, there are three important stages for performing a successful GWAS experiment.

2.1 Stage I (Phenotyping)

Phenotyping in which all genotypes should be phenotyped for a particular trait or group of traits based on the objectives of the study. Accurate phenotyping is a very critical point to detect genotype-phenotype associations. Phenotyping should be repeated over replications and/or locations and/or years. The broad-sense heritability should be calculated for raw data including all of these factors and considering G x E interaction. High heritability is an indicator that the trait is mostly genetically controlled which is important to detect the association signals. Then, the phenotypic data can be used to estimate the mean *i.e.* BLUE or BLUP. Because the phenotypic data are highly unbalanced in the plants, the estimation of genotypic values is mostly calculated as fixed effects using mixed models.

2.2 Stage II (Genotyping)

Genotyping in which the same set of individuals that were phenotyped should be used for genotyping using DNA molecular markers. Genotyping-by-sequencing is the most frequent method used in genotyping because it generates numerous SNP markers inexpensively that cover the crop genome. The GBS-generated SNPs should be filtered based on missing data, heterozygosity, and minor allele frequency. Before running GWAS, population structure should be tested in order to select the better GWAS model. The general linear model (GLM) and mixed linear model (MLM) are statistical models often proposed for performing GWAS.

The GLM does not take the population structure-related into account. The MLM, on the other hand, considers the population structure in its model (Kinship or kinship+Qmatrix). Finally, the phenotypic and genotypic data are combined using appropriate software by which alleles associated with a particular trait can be detected after the GWAS model was selected

2.3 Stage III (GWAS Analysis)

GWAS can be performed using many software statistical packages (TASSEL, GenStat, PLINK, and R (GAPIT) *e.t.c*. Here, we focused on the most important association analysis software packages that are frequently used.

2.3.1 TASSEL (Trait Analysis by Association, Evolution, and Linkage)

It is the most common software for GWAS in plants. It includes many powerful statistical methods for performing GWAS including GLM and MLM (Bradbury *et al*, 2007). TASSEL can analyse the population structure using kinship and principal component analysis. LD is included also in TASSEL. The software is always used in association analysis *e.g*. The new version of TASSEL (TASSEL 5.0) can analyse genetic diversity and perform SNP calling from GBS data. Interestingly, the software includes many visualizing tools which can be used to present data such as a scatter plot of PCA, linkage disequilibrium, Manhattan plot for GWAS results, the heat map for genetic distance, a phylogenetic tree using archaeoptery in addition to the phenotypic variance explained by markers (R2). The new version also includes some useful data summaries, which provide a quick view for a researcher on genotypes, markers, heterozygous, missing data and number of markers on each chromosome. Old versions of TASSEL such as TASSEL can accept any type of DNA markers (e.g. SNP, SSR, AFLP, RAPD, etc.). The TASSEL accepts only SNP markers. TASSEL is free software and can be downloaded from.

2.3.2 GenStat

GenStat for Windows Edition is another statistical software that can perform marker-trait association analysis in a genetically diverse population using bi-allelic and multi-allelic markers. Using GenStat, GWAS can be done either GLM or MLM models with population structure correction to control genetic relatedness by PCA or Kinship. There is an option to define the threshold of the significance of $–\log_{10}(p)$ of which Bonferroni can be selected. Interestingly, LD decay can be determined and visualized by GenStat software and the effect of each SNP can also be calculated to show the impact of the SNP on the traits. LD decay is important to determine the number of markers required for GWAS. Plots for GWAS profile of the -$\log_{10}$(P) of the test statistics and the map with the location of the detected significant markers, and Q-Q can also be visualized. The GenStat software can be purchased and downloaded from https://www.vsni.co.uk/software/genstat/ (Alqudah *et al*, 2014).

2.3.2 PLINK

Plink allows the study of a large dataset of phenotypes and genotypes (Renteria *et al*, 2013). It is free software that can be downloaded from http://zzz.bwh.harvard.edu/plink/. It provides many characteristics and features of which, PLINK performs analyses for population stratification detection, basic association tests, meta-analyses, and some other tests such as gene-based tests for association and screening for epistasis. Graphical images for Manhattan plot, Q-Q plot, and multidimensional scaling (for population structure) can be illustrated. Also, the results of GWAS and LD among SNP markers can be presented in tables produced by PLINK.

2.3.3 R- (GAPIT) Genome Association and Prediction Integrated Tool

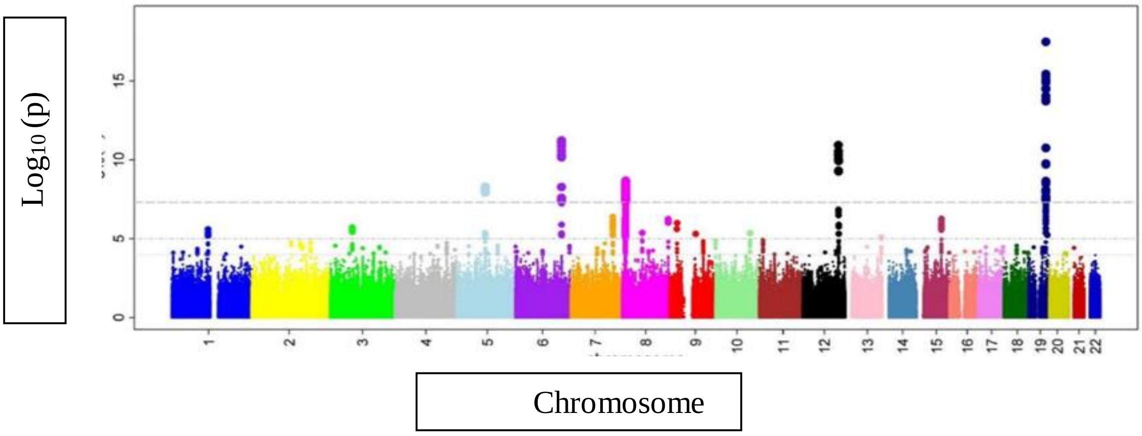
GAPIT is a useful R package that performs GWAS and genomic selection. The main advantages of GAPIT are, it can handle a large amount of data (SNPs and genotypes) and it reduces computational time without compromising statistical power (Lipka et al, 2012). The package includes many statistical methods such as MLM, population parameters previously determined (P3D), and efficient mixed-model association (EMMA). The results of GWAS results can be illustrated by Manhattan plots, Quantile-Quantile plots and a table, including p-value, minor allele frequency, sample size, phenotypic variance explained by markers R2 and adjusted P-value following a false discovery rate. Due to the aforementioned features, GAPIT becomes the most powerful and useful tool for association analysis.

3. The output results of GWAS

Each software program gives slightly different parameters as output results for GWAS. TASSEL software is a good example of producing many parameters that help to dissect the genetic basis of the target trait. These parameters include the p-value of each SNP which is important to determine the significance with the trait, R2 (phenotypic variation explained by marker) that determines if the significant SNP is a minor or major QTL, and allele effects of the significant SNP (increased or decreased the trait).

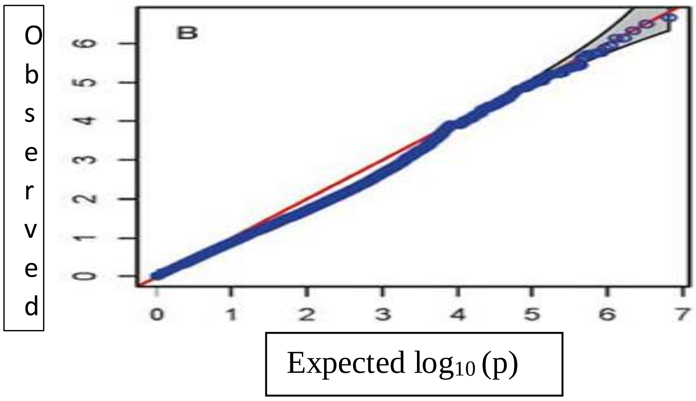
3.1 Manhattan plot

Manhattan plot the x-axis represents the genomic order by chromosome and position on the chromosome, while, the y-axis represents the $–\log_{10}$ of the P-value of each marker (equivalent to the number of zeros after the decimal point plus one). The associated significant SNP (lowest significant p values), representing QTL tend to show up as a strong signal on the Manhattan plot (Fig.3A). The threshold of $–\log_{10}$ (p-value) can be fixed at a confidence value of which $–\log10 \geq 3$ is the most common and reliable value. For further analysis, the threshold can be recalculated using the multiple comparison analysis that makes the p-value of SNP more robust and trustworthy.

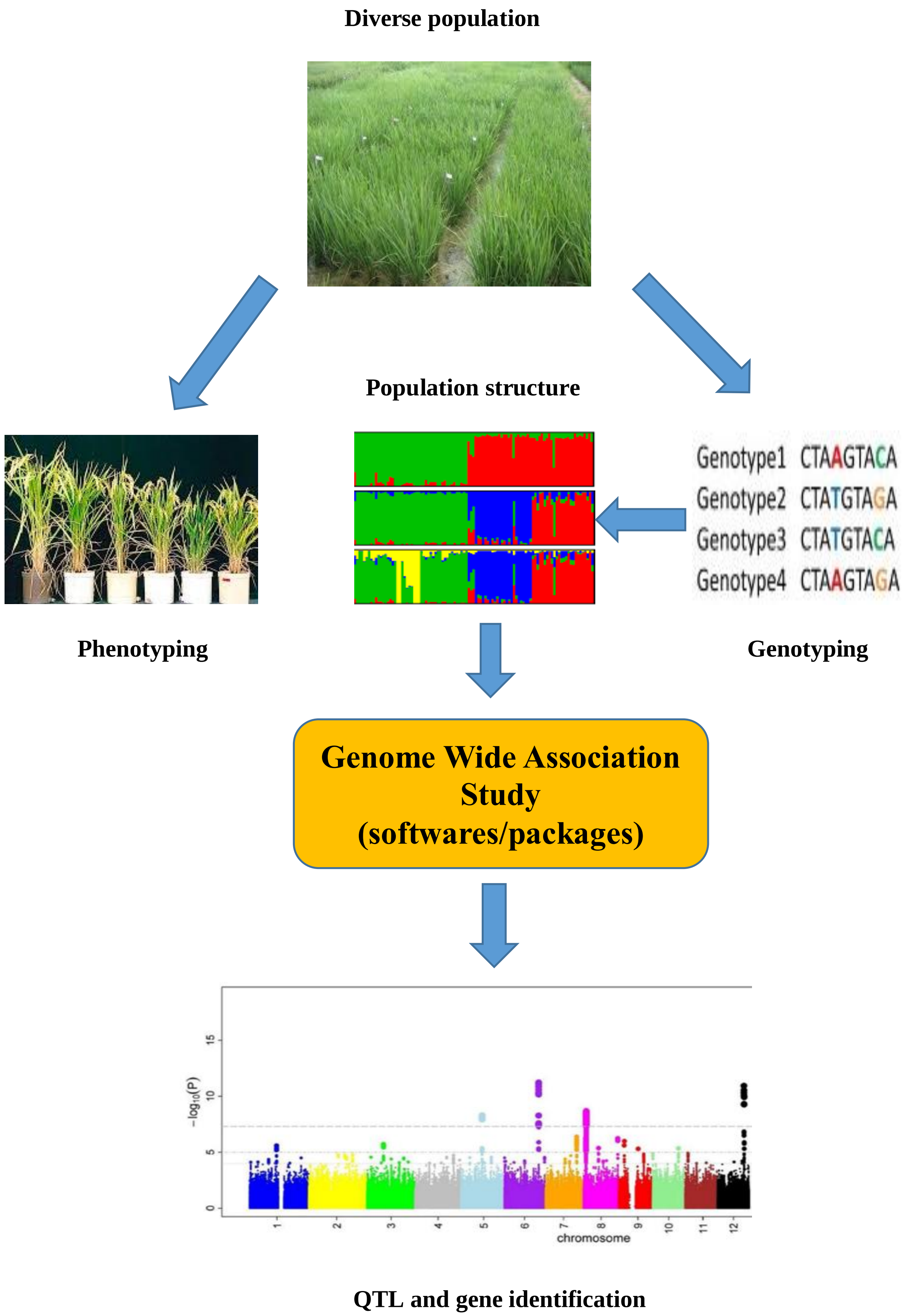


3.2 Quantile-Quantile (Q-Q) plot

Another important graph in GWAS is the Quantile-Quantile (Q-Q) plot which illustrates the relationship between the observed and expected p values. It depicts the deviation of the observed P-value of each SNP from the null hypothesis. The Q-Q plot can be used to compare the observed and expected values among GWAS statically models to show how well the model used in GWAS considering the population structure and familial relatedness and then can be applied.



Overview of working of GWAS



Diverse population

Population structure

Phenotyping

Genotyping

Genome Wide Association Study (softwares/packages)

QTL and gene identification

4. CASE STUDIES

4.1 Genome-Wide Association Analysis and Allelic Mining of Grain Shape-Related Traits in Rice

An experiment was conducted on Genome-Wide Association Analysis and Allelic Mining of Grain Shape-Related Traits in Rice. Based on 16,352 SNPs, 161 natural Indica rice varieties with various grain sizes in Southern China were used for GWAS of grain shape-related traits, referring to grain length (GL), grain width (GW), 1000-grain weight (TGW) and grain length/width (GLW).

Objective

Objective of this study was to detect the SNP loci and determine related candidate genes affecting the rice grain shape to reveal its genetic basis and molecular mechanism which lay foundation for MAS in breeding high yielding varieties

MATERIALS AND METHODS

Rice materials

A total of 161 natural Indica rice varieties with various grain sizes collected from 11 provinces in southern China were used (Supplemental Table 1). All varieties were stored at the China National Rice Research Institute (CNRRI), Hangzhou City, Zhejiang Province, China.

Field trials and phenotypic data collection

All experiments were conducted in the experimental field of CNRRI, Hangzhou, China, and a randomized complete block design was applied. Seeds with uniform germination were directly sown under a spacing of 20 cm × 20 cm (6 line × 6 rows for each variety). When ripening, 10 plants were selected in each variety. After threshing, 30 spikelets were randomly selected from each plant. The grain length and grain width were measured with a ruler, and grain length/width was calculated. Finally, 1000-grain weight was weighed and recorded by an electronic balance.

Statistical analysis

Excel 2014 and SAS 9.4 were employed for data compilation, and the mean, standard deviation and coefficient of variation of each trait were calculated. Correlation analysis was performed for the four grain traits (GL, GW, GLW and TGW).

DNA extraction and SNP genotyping

Genomic DNA from the samples was isolated from three to five leaves of 21-day-old plants per line using the CTAB method and finally diluted to 50 ng/µl. The 60 K SNP chip of of Illumina (Wright *et al*, 2010) was applied in SNP genotyping, and markers with a minimum allele frequency (MAF) less than 0.03 were deleted.

Population structure

Genetic diversity and distance measures were estimated using the PowerMarker (Liu and Muse, 2005). The model-based program Popgene (Glaubitz, 2010) was used to infer population structure and to assign individual varieties into subpopulation. SAS 9.4 was used to perform statistical analysis on the relative kinship between the combinations.

Genome-wide association analysis and allele mining

Association analyses were performed with and without correcting for population structure. General linear model (GLM) approach implemented in TASSEL was used to correlate the grain shape and the corresponding SNP loci, and a Manhattan map was generated by using the R language. Significant marker trait associations were determined based on a threshold of $-\log_{10}(P)$ as 4. Adjacent significant SNP associated with the same trait within a physical distance of 200 kb were regarded as a candidate region. Candidate genes were screened through the Rice Genome Annotation Project Database (http://www.ricedata.cn/gene/), and haplotype analyses of candidate genes were performed in combination with the rice 3K Resource Sequencing Library data (http://www.rmbreeding.cn/Index/).

RESULTS

Statistical analysis of four phenotypes related to grain shape

A total of 161 Indica rice varieties were evaluated for GL, GW, TGW and GLW. The mean values of GL, GW, TGW and GLW were 8.37 mm, 2.97 mm, 26.05 g and 2.87, respectively. CV value ranged from 9.09 per cent to 20.20 per cent, indicating that the grain shape was rich in genetic variation. The distribution pattern of each trait showed a significant normal distribution, and the correlation result revealed that there were a positive correlation between GL with GLW and TGW, and a weak negative correlation with GW. GW was positively correlated with TGW and negatively correlated with GLW.

Table. 1

Trait	Mean	SD	Max	Min	CV%
GL (mm)	8.37	0.83	11.19	7.09	9.92
GW (mm)	2.97	0.27	3.56	2.27	9.09
TGW (g)	26.05	5.26	39.34	16.93	20.20
GLW	2.87	0.47	4.33	2.20	16.38

[GL- Grain Length, GW- Grain Width, TGW- 1000 Grain Weight, GLW- Grain Length/Weight]

Basic statistics of SNP markers

Based on the genomic sequencing results, the set of marker available for GWAS after filtering the minor allele frequency consisted of 16 352 SNPs (4.2 SNP sites per 100 kb on average). Sites are distributed on all 12 chromosomes of rice, with a number of SNP markers per chromosome ranging from 708 to 2 120, PIC values of different chromosome markers ranging from 0.11 to 0.70. The results showed that the selected SNP markers were polymorphic and can perform GWAS analysis covering the entire rice genome.

Analysis of population structure

According to the genetic distance analysis, the neighbor-joining tree identified these varieties into two major groups (cluster I, 45 varieties; cluster II, 116 varieties). The phylogenetic value ranged from 0 to 0.5311 with the mean of 0.2720 and the data fluctuation value was small. The results showed that only 0.3% of the phylogenetic value was greater than 0.50, and 0.2% of the phylogenetic value was smaller than 0.05. Therefore, the relationship between varieties was relatively long and the varieties were suitable for GWAS analysis.

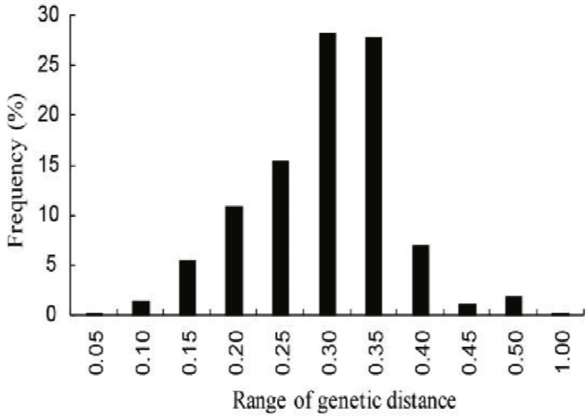


Fig.1 Frequency distribution of kinship between varieties

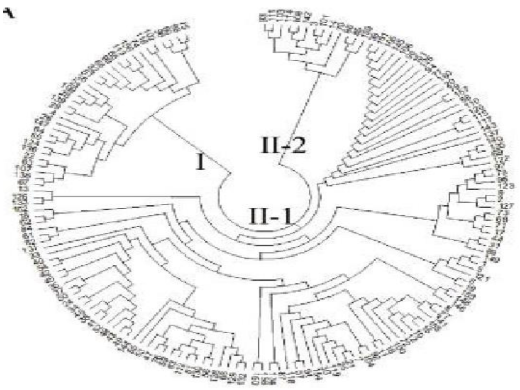
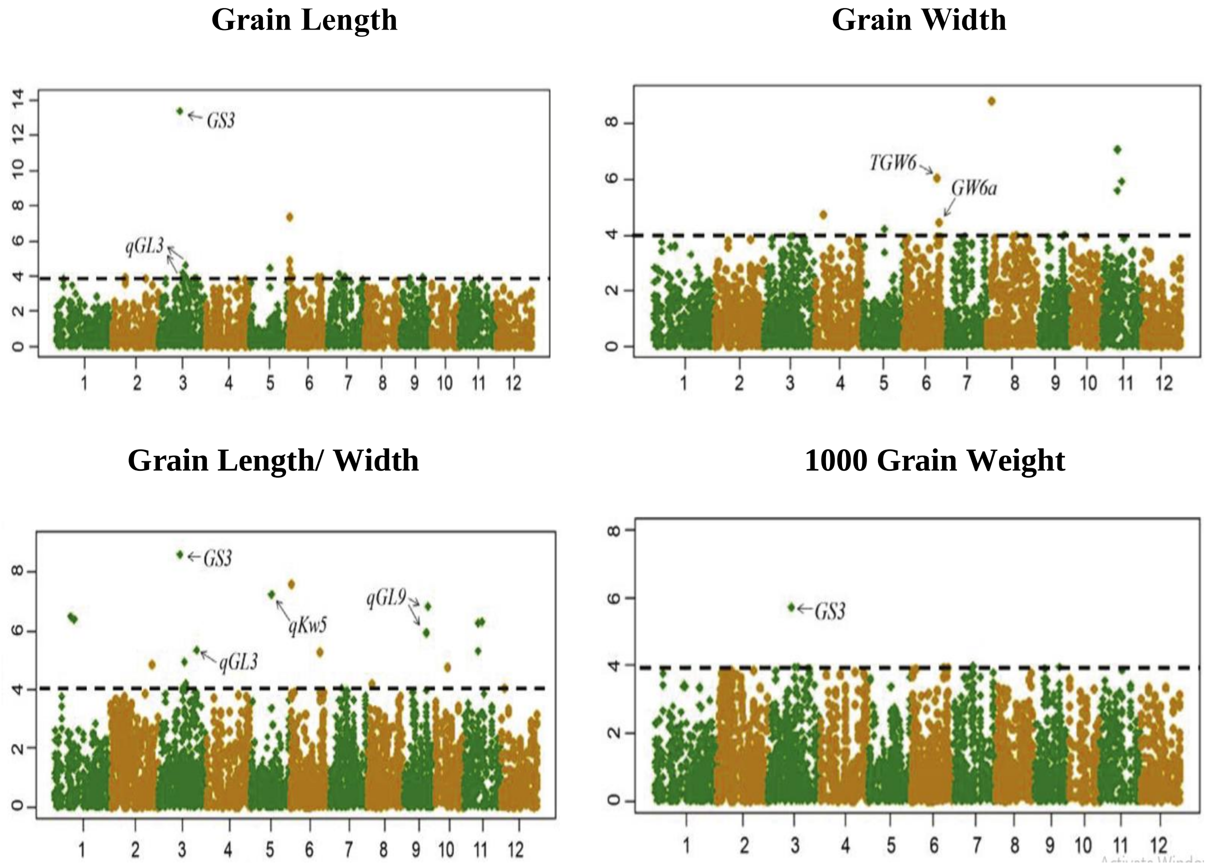


Fig.2 Population genetic relationship evolution analysis.

Genome-wide association analysis

The GLM approach was used to analyze the grain shape of rice materials. The results showed that with $–\log_{10}(P) > 4$ as the screening threshold, 38 significantly associated loci for the four traits were identified and distributed on 12 rice chromosomes. The highest number (9) was distributed on chromosome 3, and the maximum number of $–\log_{10}(P)$ was of 13.38, suggesting that the target candidate genes/QTLs are likely to be present on chromosome 3. There were 8, 9, 20 and 1 SNP loci that were significantly associated with GL, GW, TGW and GLW, respectively, and 11 sites were associated with two or more grain shape traits at the same time, suggesting that there may be a trait correlation or a pleiotropic effect.



Manhattan plots for grain length (A), grain width (B), grain length/width (C), and 1000-grain weight (D) by genome-wide association study. The significance threshold $–\log_{10}(P)$ is 4

Allelic mining and haplotype analysis

Six candidate genes/QTLs were screened out in association with 92 genes/QTLs. The results showed that two grain shape-related QTLs were detected on chromosomes 5 and 9 as qKw5 and qGL9, respectively. Two candidate regions, chr03_17302647 and chr03_19320570 on chromosome 3, were associated with Os03g0407400 and Os03g0646900 genes.

Os03g0407400 negatively regulates grain size and encodes a cysteine-rich domain protein of the TNFR/NGFR family. Os03g0646900 encodes a protein phosphatase that contains a Kelch repeat domain and positively regulates rice grain length. Two candidate regions of chromosome 6, chr06_24170016 and chr06_26025122, were associated with Os06g0623700 and Os06g0650300 genes, respectively, which code for indole-3-acetic acid (IAA)-glucohydrolase and histone acetyltransferase, respectively. A total of 22 rice varieties (overlapped varieties between these two natural populations) in this study were consistent with the 3K rice resource sequencing database (http://www.rmbreeding.cn/ Index/). The haplotype analysis of GS3 (Os03g0407400) and TGW6 (Os06g0623700) genes was performed using the Halpoview software combined with 3K resource sequencing library. Twenty-six SNP loci were detected in five exons of GS3, and 22 varieties were divided into 14 haplotypes. There were differences in the grain traits of different haplotypes. CX145 with grain length longer than 9.6 mm belongs to the dominant haplotype of GS3-11, and the grain size of the haplotype CX79 is also longer than 9 mm. Based on exon differences, 22 varieties of Os06g0623700 were divided into five haplotypes. A total of 11 varieties had excellent haplotype T-G-G, and the mean grain width of these is 3.19 mm. Allelic effect analysis revealed that when the upstream SNP site (chr06_25094225, upstream 0.98 kb) changed from T to G, the TGW6-1 dominant haplotype varieties increased by 23%, and the increase in grain width exceeded 0.4 mm. There was a significant correlation between this locus and grain width traits.

4.2 Genome-wide association study of total starch and its components in common wheat

An experiment was conducted on Genome-wide association study of total starch and its components in common wheat.

Objective

The objectives of this study were to identify markers and candidate genes for loci associated with these traits in order to improve wheat starch quality by breeding.

Materials and methods

Plant material and growth conditions

The association mapping panel of 205 wheat genotypes for GWAS comprised 77 released cultivars, 55 founder parents, and 73 breeding lines from 10 provinces that represent the major winter wheat production regions in China. Two lines from Mexico and France were included as additional founder parents. The panel was grown in the 2013-2014 and 2014-2015 cropping seasons in experimental fields at Shandong Agricultural University, Tai’an (116^{o}360’E,36^{o}570’N) and Dezhou Institute of Agricultural Sciences (116^{o}290’E, 37^{o}450’N). The experimental fields were arranged in randomized block design, with two replicates for each environment. All lines were grown in 2 m plots with 3 rows spaced 25 cm apart, and 70 seeds were evenly spaced in each row. Field management followed local procedures. No serious pest damage or lodging problems occurred during the trials.

Measurement of starch components

Starch, AMS and AMP contents were measured by the double-wave method (Jin *et al*. 2009) with modifications. The main wavelength for determining AMS content was 471 nm, and the comparison wavelength was 632 nm. The main wavelength for determining AMP content was 553 nm, and the comparison wavelength was 740.3 nm. The AMS and AMP contents in each sample were determined according to the extracted dilution factor relationship, and the total starch content (TSC) was taken as the sum of the AMS and AMP contents.

Analysis of phenotypic data

Analysis of variance (ANOVA) and correlations among phenotypic traits were carried out using SPSS version 17.0 (SPSS Inc., Chicago, IL, USA). Heritability (h2) was calculated as hB 2 = rg 2/ (rg 2 ? rge 2 / r ? re 2/re), where rg 2, rge 2 , and re 2 were estimates of genotype, genotype 9 environment and residual error variances, respectively. Estimates of rg 2, rge 2, and re 2 were obtained from the ANOVA, which was performed using the PROC GLM procedure in SAS 8.0 (SAS Institute Inc., Cary, NC, USA).

SNP markers and genotyping

SNP genotyping was performed at the University of California, Davis Genome Center. An Illumina iScan Reader was used to carry out the genotyping assays (Chen *et al,* 2016). The genetic diversity data were reported previously (Chen *et al*, 2016, 2017).

DNA extraction and a composite genetic map

DNA was extracted from the young leaf tissues of each variety. Samples were genotyped using the 90 K iSelect wheat chip, which consists of 81,587 SNP loci distributed across all 21 wheat chromosomes. The total length of the map was 3674.16 cM, with a mean genetic distance of 0.15 cM between markers. Chromosome 1B contained the most markers (n = 2390), followed by 5B (n = 2187), whereas chromosome 4D had the fewest loci (n = 78).

Among the A, B and D genomes, the B genome contained the largest number of loci (n = 12,321) and a total length of 1150.47 cM, followed by the A genome (n = 9523) at 1252.51 cM, and the D genome (n = 2511) at 1271.18 cM (Chen *et al*, 2017).

Population structure

Population structure analysis was performed on genotypic data obtained from unlinked SNP markers in the 205 winter wheat accessions using NJ cluster analysis in STRUCTURE (Chen *et al*, 2017).

Genome-wide association analysis

Significant marker-trait associations (MTAs) were identified using a mixed linear model (MLM) in TASSEL 3.0. Decisions on whether a QTL was associated with a marker was determined by P value. R2 values were used as estimates of the magnitude of MTA effects. SNPs with corrected P values B 0.01 were considered to be significantly associated with phenotypic traits.

Identification of candidate genes

To identify the position of important MTA loci in the physical map and to identify possible candidate genes, a BLAST search was performed on the International Wheat Genome Sequencing Consortium database (IWGSC; http://www.wheatgenome.org/, accessed 27th April, 2018) using the sequences of significant SNP markers identified by GWAS. When a SNP marker sequence from the IWGSC was 100% identical to any wheat contig, the sequence was extended 5 kb using the IWGSC BLAST results. The extended sequence was used to run BLAST searches on the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov, 27th April, 2018) and Ensembl Plants (http://plants. ensembl.org/ *Triticum aestivum*/ Tools/BLAST, 27th April, 2018) to confirm possible candidate genes and putative functions.

Results

Population structure

When DK values were plotted against hypothetical subgroups the highest DK was observed at K = 4, indicating the likelihood of four subgroups in the association panel. Using the maximum membership probability in STRUCTURE, the 205 accessions were segregated into four subpopulations: subgroup 1 (43 accessions), subgroup 2 (32 accessions), subgroup 3

(105 accessions) and subgroup 4 (25 accessions) (Chen *et al*, 2017). The LD values of the different chromosomes were reported in (Chen *et al*, 2016).

Phenotypic data

The phenotypic values for the wheat starch trait in diverse environments. Extensive phenotypic variation for AMS, AMP and TSC among the 205 winter wheat accessions was observed across four environments. The AMS contents ranged from 16.47 to 22.99% in the flour, AMP contents ranged from 38.43 to 61.15%, TSC contents ranged from 55.78 to 82.19%, and AMS/AMP ratio ranged from 33.01 to 52.78%. Broad-sense heritabilities were 89.31, 68.10, 75.36 and 32.45%, respectively, indicating that both genetic and environmental factors influenced the expression of each trait.

Table. 2

Trait	Min (per cent)	Max (per cent)	H^2 (per cent)
AMS	16.47	22.99	89.31
AMP	38.43	61.15	68.10
TSC	55.78	82.19	75.36
AMS/AMP	33.01	52.78	32.45

Thousand kernel weights (TKW) ranged from 26.33 to 60.13 g, and protein contents ranged from 10.30 to 17.98% across environments. Hence, they belonged to typical quantitative traits controlled by multiple loci.

Marker-trait associations and elite allele exploration

A total of 24,355 mapped SNPs was used for MTA analysis. Forty-seven significant MTAs were detected for all four traits across environments. We further analysed MTAs for AMS and AMP by comparing the phenotypic effects of alleles at each locus to identify elite genes for the starch components and AMS: AMP ratio. Nine MTAs were recorded for the two starch traits, and there were 11 MTAs for three traits. These SNPs on eight chromosomes, each accounted for 11.26–23.83% of the phenotypic variance. Eighteen MTAs on chromosomes 1B, 2A, 3B, 3D, 4A, 5B, 6A, 6B and 7B were identified as being related to AMS: AMP ratio, each explaining 5.92-17.2% of the phenotypic variation. Nine MTAs were detected in two environments; seven in E1 and E2, and two in E3 and E4. Fifteen MTAs for AMS were identified on chromosomes 2A, 2B, 3A and 4A explaining 11.8–18.41% of the phenotypic variation. Two MTAs, IAAV4464 (2A_112) and JD_c3742_1130 (2A_112), on chromosome 2A were detected in three environments; these MTAs located at the same position had the highest R2 (18.41%) and smallest P values. Twelve of the 15 MTAs showed significant phenotypic differences among alleles, and the same MTAs exhibited phenotypic differences in environments E2 and E4 (P\0.05). Alleles A and G of marker Kukri_c5615_1214 (3A_93).

Table. 3

Trait	MTA	Chromosomes
AMS	15	3 (2A, 2B, 3A)
AMP	23	8 (2A, 2B, 3A, 3B, 4A, 6A, 6B, 7D)
TSC	22	7 (2A, 2B, 3A, 3B, 4A, 6A, 6B)
AMS/AMP	18	8 (1B, 2A, 3B, 4A, 5B, 6A, 6B,7B)

Putative candidate genes linked to starch-related traits

14 Significant MTAs identified in more than two environments and correlated with more than one trait were selected for candidate gene prediction. For marker IAAV4464 on chromosome2AL there were four candidates but geneTRIAE_CS42_2AL_TGACv1_093900_AA0288950 was related to beta-glucosidase and hydrolysis of O-glycosyl compounds that participate in carbohydrate metabolism.

Table. 4

Marker	Chromosome	Candidate genes
IAAV4464	2AL	4
JD_c3742_1130	2AL	2
wsnp_Ex_c63909_62932437	2AL	1
RFL_Contig4517_1276	2AL	1
RAC875_c6280_292	4AL	1
Tdurum_contig41127_265	4AL	1
BobWhite_c10583_352	4AL	1
Excalibur_c16376_351	6BS	2
CAP11_c1087_327	6BS	1

5. Applications of GWAS

1. GWAS have been very successful in identifying novel variant-trait associations

2. Helps in finding variations among complex traits

3. Relevant for study of low- frequency and rare variants

4. Studies genetic variants other than SNVs like copy no. Variants

5. Data are used for multiple applications beyond gene identification

6. GWAS data generation, management and analysis are straightforward as some software's are used.

6. Limitations of GWAS

1. Phenotypic variation - depends on GXE interaction and heritability. G x E reduces heritability so we should select plants with high heritability.

2. Number of individuals – If number of population increase associations increase so that we can overcome rare-variants. (Kumar *et al,* 2012).

3. Population structure - Not all individuals are equally distantly related to each other at the genetic level (Prichard *et al*, 2000).

4. Allele frequency - Rare allele (<5%) leads to a lack of resolution power so that difficult to identify (Cerda *et al*, 2012).

5. Linkage Disequlibrium (LD) - An indicator to detect the distance between loci, which is important to find the number of required markers for the whole genome scan it is affected by population size and allele frequency (Myles *et al,* 2009).

7. Conclusion

In conclusion, GWAS is a powerful tool for studying multiple traits in response to biotic and abiotic stresses such as drought, salt, temperature, diseases *etc*. and agronomic traits. Through GWAS many novel QTLs and candidate genes were identified. This valuable information can be used for future breeding programmes and in designing better crop varieties.

8. References

Alonso-Blanco, C., Aarts, M. G., Bentsink, L., Keurentjes, J. J., Reymond, M., and Vreugdenhil, D. *et al*. 2009. What has natural variation taught us about plant development, physiology, and adaptation? *The Plant cell* 21(7): 1877-1896.

Alqudah, A. M., Sallam, A., Baenziger, P. S., and Borner, A. 2019. GWAS: Fast-forwarding gene identification in temperate cereals: barley as a case study - a review. *J. Adv. Res.* 3: 3-26.

Alqudah, A. M., Sharma, R., Pasam, R. K., Graner, A., Kilian, B., and Schnurbusch, T. 2014. Genetic dissection of photoperiod response based on GWAS of pre-anthesis phase duration in spring barley. *PloS One*. 9(11): e113120.

Badr, A., Muller, K., Schafer-Pregl, R., El Rabey, H., Effgen, S., and Ibrahim, H. H. *et al*. 2000. On the origin and domestication history of Barley (*Hordeum vulgare*). *Mol. Biol. Evol*. 17(4): 499-510.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics.* 23(19): 2633-2655.

Chen, G. F., Wu, R. G., Li, D. M., Yu, H. X., Deng, Z. Y., and Tian, J. C. 2017. Genome-wide association study for seedling emergence and tiller number using SNP markers in an elite winter wheat population. *J. Genet.* 96: 177-186.

Chen, G. F., Zhang, H., Deng, Z. Y., Wu, R. G., Li, D. M., Wang, M. Y., and Tian, J. C. 2016. Genome-wide association study for kernel weight-related traits using SNPs in a Chinese winter wheat population. *Euphytica* 212: 173-185.

Chen, X., Fang, W., Ji, M., Xu, S., Jiang, Y., Song, S., Chen, G., Tian, J., and Deng, Z. 2019. Genome-Wide Association Study of total starch and its components in common wheat. *Euphytica* 215: 201.

Distefano, J. K. and Taverna, D. M. 2011. Technological issues and experimental design of gene association studies. *Methods Mol. Biol.* 10(3): 7003-7016.

Glaubitz, J. C. 2010. CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Mol. Ecol. Resou.* 4(2): 309-310.

Jin, Y. H., Zhang, K. L., Zhang, X. C., and Du, J. H. 2009. Determination of amylose and amylopectin in wheat and wheat malt by dual -wavelength spectrophotometry. *J. Chinese Cereals Oils Assoc*. 24: 137-140.

Kumar, J., Pratap, A., Solanki, R. K., Gupta, D. S., Goyal, A., and Chaturvedi, S. K. *et al*. 2012. Genomic resources for improving food legume crops. *J. Agric. Sci*. 150(3): 289-318.

Kumar, V., Singh, A., Mithra, S. V., Krishnamurthy, S. L., Parida, S. K., Jain, S., Tiwari, K. K., Kumar, P., Rao, A. R., and Sharma *et al*. 2015. Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Res*. 22 (2): 133-145.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J. Li. M., and Bradbury, P. J. *et al*. GAPIT: Genome Association and Prediction Integrated Tool. *Bioinformatics* 28(18): 2397-2980.

Liu, K. and Muse, S. V. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics* 21(9): 2128-2129.

Manolio, T. A. 2010. Genome-Wide Association Studies and assessment of the risk of disease. *N. Engl. J. Med*. 363(2): 166-176.

Mitchell-Olds, T., Willis, J. H., Goldstein, D. B. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat. Rev. Genet*. 8(11): 84556.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., and Costich, D. E. *et al*. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. *The Plant cell* 21(8): 2194-202.

Neelapu, N. R. R. and Surekha, C. 2016. Next-generation sequencing and metagenomics. In: Wong, K. C. (Ed.), *Computational Biology and Bioinformatics: Gene Regulation*. CRC Press, Boca Raton, pp. 331-351.

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945-59.

Renteria, M. E., Cortes, A., and Medland, S. E. 2013. Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis. In: Gondro C, van der Werf J, Hayes B (eds.). *Genome-Wide Association Studies and Genomic Prediction*. Humana Press, Totowa. pp. 193-213.

Soto-Cerda, B. J. and Cloutier, S. 2012. Genetic diversity in plants. pp. 29-45. Available:www.intechopen.com.[12 Jan 2020].

Wang, S. C., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., and Cattivelli *et al*. 2014. Characterization of polyploid wheat genomic diversity using a high density 90000 single nucleotide polymorphism array. *Plant biotechnol. J.* 12: 787-796.

Wright, M. H., Tung, C. W., Zhao, K., Reynolds, A., Mc Couch, S. R., and Bustamante, C. D. 2010. ALCHEMY: A reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. *Bioinformatics* 26(23): 2952-2960.

Xu, S. Theoretical basis of the beavis effect. *Genet*. 2003. 165(4): 2259-2268.

Yang, L., Yueying, W., Jahan, N., Haitao, H., Ping, C., Lianguang, S., Haiyan, L., Guojun, D., Jiang, H., and Zhenyu, G. *et al.* 2019. Genome-Wide Association analysis and allelic mining of grain shape-related traits in rice. *Rice Sci*. 26(6): 384-392.

9. Discussion – [Questions and Answers]

1. Explain about DNA chip technology?

DNA microarrays or DNA chips consists of thousands of individual DNA sequences arrayed at a high density on a single matrix, usually glass slides or quartz wafers, but sometimes on nylon substrates. Probes with known identity are used to determine complementary binding, thus allowing the analysis of gene expression, DNA sequence variation or protein levels in a parallel format.

2. What is the difference between DNA sequencing and DNA re-sequencing?

In DNA sequencing we will find the exact sequence of a certain length of DNA whereas in re-sequencing we find the variations by comparing with reference genome.

3. Explain about GLM and MLM models?

General Linear Model (GLM) compares how several variables affect different continuous variables and in Mixed Linear Model (MLM) consists of both fixed effects and random effects. They are particularly used where repeated measurements are made on same or related statistical units.

4. Is GWAS cost effective as sequencing cost is comparatively reduced

It is not cost effective even though now a days sequencing cost is reducing as it requires huge investments for carrying out phenotyping and complete genome should be sequenced for each and every individual.

5. Among all the software packages used for GWAS analysis which will be effective?

GAPIT- Genome Association and Prediction Integrated Tool. It can handle large amount of data and most recently developed.

6. What is a haplotype?

A set of SNPs found on the same chromosome are called as haplotype.

7. Which technology is effective and easy among gene chip and illumine?

Illumine is more advanced, it contains large amount of DNA and easy to perform.

8. What is the minimum size of the population to be considered while performing GWAS?

Minimum of 100 individuals should be selected for GWAS. If the size of population is less only less number of associations can be identified and identification of rare alleles becomes difficult

9. How to overcome the limitation of identification of rare alleles in GWAS?

By using Nested Association Mapping (NAM) we can identify rare alleles.

10. If the heritability is less what is to be done?

If heritability is low that individuals should be removed from GWAS analysis.

11. In how many environments the experiment should be replicated?

As G X E interactions will be more it should be replicated in minimum three or four environments to get a standard phenotypic data.

12. What should be done if the size of population is less and less variations observed?

If the size of populations is less the associations observed will not be present above the level of significance and false positives will be more so, such population cannot be used for conducting GWAS.

KERALA AGRICULTURAL UNIVERSITY

COLLEGE OF HORTICULTURE, VELLANIKKARA

Department of Plant Breeding and Genetics

GP. 591: Master's Seminar

Name	: T. Anusha	Venue	: Seminar Hall
Admission No.	: 2018-11-143	Date	: 9-1-2020
Major Advisor	: Dr. P. Sindhumole	Time	: 11.30 am

GWAS - Genome-Wide Association Studies

Abstract

The causal relationship between genetic polymorphism within a species and the phenotypic differences observed between individuals is of fundamental biological interest. The ability to identify variations associated in response to biotic or abiotic stresses, agronomically important traits like growth rate, yield in plants *etc.* requires an understanding of both the genetic architecture of a trait and the specific loci that is underlying a phenotype. Genome-Wide Association Studies (GWAS) present a powerful tool to reconnect this trait back to its underlying genetics.

GWAS focus on capture of Single Nucleotide Polymorphism (SNP) data. SNPs are single base-pair changes (mutations) in the DNA. Millions of SNPs can be captured using genotyping technologies. GWAS or Whole Genome Association Studies (WGAS) or Common Variant Association Studies (CVAS) investigate a genome-wide set of genetic variants in different varieties to see if any variant is associated with a trait (Manolio, 2010).

In a GWAS experiment, initially the population is to be selected with full consideration of the size of the population (minimum 100 individuals). There are three important stages for performing a successful GWAS experiment. Stage I is phenotyping of all genotypes for a particular trait or group of traits based on the objectives of the study, stage II is genotyping using DNA molecular markers, and stage III is GWAS analysis in which phenotypic and genotypic data are combined using appropriate softwares (TASSEL, GenStat, PLINK and R(GAPIT)). Finally, results are visualised by Manhattan and Quantile-Quantile (Q-Q) plots (Alqudah *et al*., 2019).

In an experiment conducted on Genome-Wide Association analysis and allelic mining of 161 natural Indica rice varieties for grain shape-related traits (Grain Length, Grain Width,

1000-Grain Weight and Grain Length/Width) based on 16,352 SNPs, 38 significant loci were identified through general linear model correlation analysis. Additionally, using sequenced 3K-germplasm resources, 22 overlapped varieties, twenty-six SNPs and fourteen haplotypes were identified (Yang *et al*., 2019).

Significant Marker-Trait Associations (MTAs) and candidate genes associated with markers were detected for total starch (TSC), amylose (AMS) and amylopectin (AMP) contents under four environmental regimes by another experiment on GWAS for total starch and its components in a panel of 205 elite winter wheat accessions using SNPs (Chen *et al*., 2019).

GWAS is a powerful tool for studying multiple traits in response to biotic and abiotic stresses such as drought, salt, temperature, diseases *etc*. and agronomic traits. Through GWAS many novel QTLs and candidate genes were identified. This valuable information can be used for future breeding programmes and in designing better crop varieties.

References

Alqudah, A. M., Sallam, A., Baenziger, P. S., and Borner, A. 2019. GWAS: Fast-forwarding gene identification in temperate cereals: barley as a case study - a review. *J. Adv. Res.* 3: 3-26.

Chen, X., Fang, W., Ji, M., Xu, S., Jiang, Y., Song, S., Chen, G., Tian, J., and Deng, Z. 2019. Genome-Wide Association Study of total starch and its components in common wheat. *Euphytica.* 215: 201.

Manolio, T. A. 2010. Genome-Wide Association Studies and assessment of the risk of disease. *N. Engl. J. Med.* 363(2): 166-176.

Yang, L., Yueying, W., Jahan, N., Haitao, H., Ping, C., Lianguang, S., Haiyan, L., Guojun, D., Jiang, H., Zhenyu, G., *et al.* 2019. Genome-Wide Association analysis and allelic mining of grain shape-related traits in rice. *Rice Sci*. 26(6): 384-392.