GENOME-WIDE ASSOCIATION STUDIES OF FOUR AGRONOMIC TRAITS IN RICE (*Oryza sativa)*

Dissertation submitted in partial fulfilment of the requirement for the degree of

MASTER OF SCIENCE
IN
BIOTECHNOLOGY

By

Shalini Thakur
Enrollment No. 217809

Under the Supervision of

Dr. Shikha Mittal
(Assistant Professor)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY WAKNAGHAT,
DEPARTMENT OF BIOTECHNOLOGY & BIOINFORMATICS
SOLAN-173234, HIMACHAL PRADESH

DECLARATION

I hereby declare that work reported in the M.Sc. Biotechnology project entitled **"Genome-Wide Association Studies of four agronomics traits in rice (*Oryza sativa)*"** submitted at **Jaypee University of Information Technology, Waknaghat, H.P, India,** is an authentic record of my work carried out over a period from January 2023 to May 2023 under the supervision of **Dr. Shikha Mittal** (Assistant professor)**.** I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my **M.Sc.** Dissertation report.

Shalini Thakur

Enrollment Number: - 217809 **Date:**
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, India – 173234

SUPERVISOR CERTIFICATE

This is to certify that the work reported in the **M.Sc.** Dissertation report **"Genome Wide Association Studies of four agronomics traits in rice (*Oryza sativa)*"** submitted by **Shalini Thakur** at **Jaypee University of Information Technology, Waknaghat, India,** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Supervisor:

Dr. Shikha Mittal **Date:**

Assistant Professor
Department of Biotechnology &Bioinformatics
Jaypee University of Information Technology
Waknaghat, India-173234

ACKNOWLEDGMENT

I would like to express my profound gratitude to my guide **Dr. Shikha Mittal** for his guidance, support and constant encouragement throughout the course of this project work. She has been more than just my project guide; at times a mentor to rescue me out of my doubts. She has always helped me to work hard and also taught me how to implement different ideas to deal with the problem. Moreover, she taught me to not give up and many other valuable lessons.

Furthermore, I would like to acknowledge Vice-Chancellor **Prof. (Dr.) Rajendra Kumar Sharma**, **Prof. (Dr.) Ashok Kumar Gupta,** Dean of academics & research for providing me with an opportunity to be a part of the institute and to complete my Master's Degree.

I also want to mention the HOD of Biotechnology and Bioinformatics **Prof. (Dr.) Sudhir Kumar** has been a source of immense motivation and inspiration both for my academic and personal life. He was never, and I know will never be, more than just a phone call away. He has helped me in almost every aspect I have asked him for.

In addition, I would like to thank all the faculty members of the BT/BI Department of JUIT, who have helped me whenever I needed and also would like to thank all the lab engineers and specially **Ms. Somlata Sharma** for providing me with a workplace and for always motivating me.

I would also like to appreciate the part that my classmates (Pallavi, Gargi, Raj laxmi and swalpana) have played in shaping this project work. They have been my constant support and cheered me up at hard times. They helped me whenever I had any doubts. Thanks a lot!

I would like to thank the almighty God for his grace throughout my life. Last but not the least I would like to thank my Mother and Father who have always supported me through thick and thin and have been a constant source of encouragement and support; also, who has never given up on me and always motivated me.

[Thanks to JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY]

SHALINI THAKUR
M.Sc. Biotechnology
JUIT, Solan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GWAS: - Genome-Wide Association Study

IARI: - Indian Agricultural Research Institute

MLM: - Mixed Linear Model

GLM: - General Linear Model

QTL: - Quantitative Trait Loci

LD: - Linkage Disequilibrium

SNP: - Single Nucleotide Polymorphism

Q-Q: - Quantile-Qunatile

HT: - Height

GN: - Grain Number per Plant

GAPIT:-Genome Association and Prediction Integrated Tool

TASSEL: - Trait Analysis by association, Evolution and Linkage.

NGS: - Next Generation Sequencing

MAF: - Minor Allele Frequency

EMMA: - Efficient Mixed-Model Association

Farm CPU: - Fixed and random model Circulating Probability Unification

MCMC: - Markov Chain Monte Carlo

ABSTRACT

The purpose of this study was to find common single nucleotide polymorphisms (SNPs) associated with rice phenotypic traits. Genome-wide association studies assist scientists in identifying genes associated with a specific trait. A genome wide association study was performed in four rice trait (grain per plant, grain yield, grain yield per meter square, plant height) were selected to examine whether significant phenotypic variances exist in the yield among the 186 rice genotype. GWAS will perform in 186 rice varieties with 50051 high-qualities SNPs. The full phenotypic data of rice core accessions were obtained from IARI. (ICAR- Indian Agricultural Research Institute). Manhattan plot show the significant value above the *p*-value >0.001. In this study, we did a comparative analysis between TASSEL software and GAPIT R package by using two univariate methods MLM and GLM. Phylogenetic analysis revealed the two clusters in the rice diverse collection and population structure analysis show they have a four population were present in the rice diverse collection. Markers traits associations' analysis using different GLM and MLM models revealed the 23 candidate genes those were associated with four different agronomics traits. Further we analysed the linkage disequilibrium with the genotype data. They show the closely linked markers.

Keywords: SNPs, Mixed linear model, General linear model, Genome-wide association study, GAPIT and TASSEL.

CHAPTER 1-

INTRODUCTION

1.1 INTRODUCTION

Genome-wide association studies are just one type of recent method that allows researchers studying the genetics of complex traits in different rice varieties to locate the causal loci (or perhaps the causal genes) that underlie these features. The genetic underpinnings of the agronomic features in agricultural landraces which have evolved to varied agriculture-climatic environments must be discovered in order to ensure global food security. GWAS uses statistical methods to look for links among sequence polymorphisms in the rice genome and phenotypic variation throughout rice varieties. Genome-wide association studies (GWAS) seek to identify genotype-phenotype associations by looking for genetic variant allele frequency differences among individuals who are genetically similar but differ phenotypically. Although in the human genome, GWAS can look at copy-number variants or sequence variations. Single-nucleotide polymorphisms are the most commonly studied genetic variants in GWAS (SNPs). Genomic risk loci are blocks of SNPs that are all statistically significant for the trait of interest in GWAS. There is numerous duplicated genetic risk loci connected to various diseases and phenotypes after 15 years of GWAS [1].

The results of GWAS can be used for a variety of purposes. For instance, trait-associated genetic variations might be employed as a control variable in epidemiological studies to take genetic group differences into account. The outcomes of a person's genetic profile can also be used to forecast their risk for both physical and mental illness [2] +

-*36GWAS has two advantages over conventional biparental populations: The primary lines utilised in the segregate populations are substantially less genetically diverse than the rice types employed in the GWAS populations & the majority of GWAS can generate rather high mapping resolution as a result of the numerous previous recombination events [3]. Some critical factors for a successful GWAS include population generation, genotype, phenotype and a pipeline for software. We as well summarise the key outcomes of current GWAS for rice, as well as functional examinations of GWAS interesting gene. These recent investigations have improved our genetic map capability and knowledge of the genetic influences on many significant rice traits. For future Prospective breeding applications, the future of GWAS, and rapid advances in GWAS follow-up researches [4].

1.2 PROBLEM STATEMENT

In this study, between present straight increases in worldwide food production and anticipated demand, there is a sizable gap. One of the most significant food crops in the world and a model plant for so many purposes is rice (*Oryza sativa*), including the rich natural variation within. Genome-wide association studies have emerged in recent decades (GWAS), have been conducted employing high-throughput sequencing to analyse the genetic makeup of key rice properties. Given its importance to global food security, yield is one of the key characteristics that rice breeders concentrate on yield. Rice is a vital cereal that feeds greater than 50% of the global population, especially in developing countries. Around 760 million tonnes of paddy, or 35% more than the amount of rice produced in 1996, will be needed by 2025 for the world to meet the rising demand. Arable land is, however, primarily exploited, particularly in Asia, where 90% of the world's rice is produced and consumed. Food production has become an extreme challenge due to the rapid growth of the global population, the energy the energy sector, and the widespread use of pesticides and fertilisers in agriculture have resulted in heavy metal contamination in soil, such as the metalloid Arsenic (As)., manganese, nickel (Ni), and cadmium (Cd) Increased levels of heavy metals in soil inhibit crop germination and growth, lowering farm productivity. In the meantime, plants take in poisonous heavy metals from polluted soil and build up the elements in consumable plant tissues, poisoning food [3, 4]. Additionally known as a multigene regulated characteristic, yield is thought to be differentiated by a variety of genes and loci. Additionally, yield is a complex characteristic that is influenced by a wide range of factors, including the number of tillers, plant Ht. no. of grain per plant, grain weight, grain yield, and number of major branches [5]. In this research, we collected rice data from IARI. & we performed GWAS to examine their HT (PLANT HEIGHT), GW (GRAIN WEIGHT), GN (GRAIN NUMBER PER PLANT), and DY (GRAIN YEILD) with the intention of identifying genes with linked SNPs that might explain phenotypic differentiation and are anticipated for use in subsequent breeding programmes.

1.3 OBJECTIVE

- To identify the sub-population in a rice diversity panel.
- Identification of marker trait association.
- Identification of genes associated with significant markers.

CHAPTER 2 – REVIEW OF LITERATURE

2.1 Genome-Wide Association Study (GWAS)

Genome-wide association research permits the identification of the genes associated with specific traits. Genome-wide association studies look for correlations across phenotypes genotypes and by comparing the allele frequencies of genetic variants in people with similar ancestry but diverse characteristics. The more than 5,700 GWAS have examined more than 3,300 characteristics. GWAS sample sizes have increased to well over a million people as a result of efforts to boost statistical power, showing a significant number of linked and reproducible variants for many heritable variables [6, 7].

Genome-wide association research has frequently been utilised to identify genetic variants that influence complex traits, either through comparative analysis or correlation analysis, and have found a large number of SNPs linked to the target characteristics [8]. Studies employing GWAS platforms have successfully analysed the genetic underpinnings of various complex variables in important crops during the past few decades, including flowering time and yield-related parameters [9, 10]. GWAS were conducted to find significant associated loci that were persistently expressed across various contexts. The genetics governing the build-up of elements in rice grain has also been studied, and this has led to the discovery of closely linked loci and possible causal genes [11].

Human genome sequencing and the conclusion of the human genome haplotype mapping project allowed for these research to be conducted (International Hap Map project, 2005), It led to the identification of billions of common significant SNPs, as well as the recording of the alleles' correlation structures or linkage disequilibrium, countless numbers of common significant SNPs were found, and the alleles at those loci were recorded with regard to their relationship structure or linkage disequilibrium. Having chips for testing over 100,000 SNPs made available for a few hundred dollars or less per sample because of this understanding of genomic variation and cutting-edge bioengineering techniques. With the decline in the genome-wide genotyping costs, the number of GWAS research has significantly grown [12].

2.2 Genome-wide association study in rice (*Oryza sativa*)

The "big three" global cereals— wheat, rice and maize—accounted for around 87% of all grains produced globally. Rice has a relatively small genome that has been fully sequenced using the map-based approach [13]. Numerous genetic researches have been conducted in order to categorise the biological functions of hundreds of rice genes, using information from

the rice reference genome sequence. There are also a number of technical platforms for functional genomics studies on rice [14]. The most crucial sources for breeding are the extensive germplasm of rice's cultivated species & wild accessions. These new varieties & natural accessions had already modified for various agricultural climatic environments & continue to have a sizable amount of genetic variation. Rice genetics research has focused substantially on knowing the genetic basis of phenotypic variability among germplasm, particularly agronomical significant traits. Among the most frequently grown worldwide cereal crops, rice is grown in a variety of geographic, ecological, and climatic environments [15, 16].

The small genome size of rice makes it an excellent model crop for functional genomics research. Because of large germplasm sources and low sequencing costs, Gene-wide association research has enabled researchers to analyse the genetic variation underlying agronomic characteristics in rice. Many accessions with significant phenotypic and genotypic variety are available due to the diverse adaptations of rice genotypes [17].Many of these rice accessions, including those from the *japonica, indica* & *javanica* subspecies, have been preserved in international gene banks [18]. This is significant because it represents a potential source of gene reservoirs for crop improvement initiatives [19]. The majority of rice breeding initiatives, however, have only utilised a small percentage of the genetic resources that are accessible for rice, there is a high degree of genetic similarity among most commercial rice cultivars [17]. Additionally, it is essential to diversify the genetic background of rice genotypes by adding genes from wild or close relatives that may result in Quantitative trait loci or new genes for essential agronomic characteristics.

How genetically diverse is a population and the extent to which the desired features are heritable both play a significant role in determining the success of any plant breeding programme [19]. As a result, finding molecular markers or QTLs linked to desirable traits and potentially useful for marker-assisted selection has necessitated association mapping using phenotypic and genotypic data. As a result, it is possible to employ a larger variety of germplasm that provides a wider allelic coverage without necessarily establishing bi-parental mapping populations [20].

Understanding the genetic foundation for agricultural features landraces of crops that have evolved diverse agricultural climates environments is crucial for ensuring the world's food security. Rice (*Oryza sativa*) is a primary source of nutrition for more than half of all humans

[21]. Natural and human selections have led to the development of rice landraces from their wild ancestor, maintaining a significant genetic diversity [22]. Moreover, these grown varieties have a great tolerance for biotic and abiotic stress, which leads to extremely steady yields and an intermediate yield in low-input farming systems. Abiotic stress tolerance genes, on the other hand, may be linked to unfavourable traits like Poor grain quality, an excessive plant height, and a lack of yield potential caused these cultivars to be disqualified from selection during the selection process [23].

In order to breach the yield improvement barrier under drought, advances in molecular biology have given breeders new chances to identify such regions, refine them through precise mapping, and incorporate them into climatic variety. Until a few years ago, these opportunities were not available [24]. Finding the genetic underpinnings of these many types will offer crucial insights for creating exceptional types of crops to promote sustainable agriculture.

GWAS have become a popular method for rapidly discovering the genes driving complicated traits [25]. GWAS have not been widely used to analyse complex features in crop plants, despite their promise [26]. Rice is a superior a potential system for the implementation of the GWAS since it reproduces by itself, has access to phenotyping resources, and has a superior high-quality reference genome sequence. Such a technique ought to enable the discovery of high-quality haplotypes required for precisely connecting molecular markers with phenotypes [27]. A significant obstacle to increasing food production is projected to be the effect of climate change on agricultural output [28].

It has long been of tremendous interest to use the GWAS approach to simultaneously map several agronomic features in various kinds of rice because it is a significant crop. Rice is a selfing crop with substantial germplasm banks, making it an excellent choice for GWAS. In previous study, the functional impact of 18 genes related to starch synthesis on regulating the cooking and eating quality of rice was previously examined in rice by the use of candidate-gene association analysis [13]. Even for those who are unfamiliar with the causal genes and variations, GWAS on agronomic variables gives useful insights and data that may be immediately used to rice breeding [14].

Single nucleotide polymorphisms (SNPs) have replaced simple sequence repeat markers as the preferred marker for use in genotyping studies with GWAS because next-generation sequencing technologies are developing quickly. Comprehensive genome coverage and

certain SNP identification are made possible by the enormous amounts of data that can be gathered with NGS [29].

2.3 Experimental design:

More than ten years ago, the first extensive GWAS was created for humans. As used in the medical genetics, In GWAS analyses Individuals with a disease's genome-wide polymorphisms to those of comparable individuals without the ailment. This method is known as the phenotype-first design because individuals are classified first according to their characteristics then included in the GWAS panel based their phenotypic information [30]. The GWAS in plants often uses genotyping first tactics, compared to the phenotype first approach, where data for the GWAS panel are mostly picked according to their genetic diversity, with no particular emphasis given to one feature [31].

In a typical GWAS for crops, genetic information from the population is used for statistical analysis in order to examine the relationships between genotype and a variety of phenotypes. The population is frequently a source of germplasm with a diverse range of phylogenetic relationships and a large geographic spread. Population genetics is therefore frequently used while choosing the GWAS panel that was with a highly amount of the genetic variation as well as a weak population structure usually serving as the main selection factor. A GWAS may capture more loci linked to higher phenotypic variety when there is a high genetic diversity among the population, while fewer false positive associations occur from a low population structure. In addition, because numerous seed bank accessions lack homozygosity, seed purification is a crucial step in creating a viable GWAS population [32].

Among the most significant crops on the planet, rice is adapted to a variety of ecological and agronomic conditions and has a very wide geographic distribution. As a result, rice has an extremely diverse genetic makeup. Analysis is based on the rice pan-genomic information set, In addition to many more variants in non-coding regions, the typical value for each rice gene of sixteen Coding variations distributed over various haplotypes. Such as the promoter regions, that may have an impact on the control of gene expression [33]. The GWAS greatly benefit from this tremendous diversity. But because it is a self-pollinating plant, rice has a very strong population structure.

For GWAS to have adequate statistical power, sample size is crucial. The mapping power of single rice GWAS perhaps too low if sample size is too small, and if it is a very large number of accessions were collected, the cost may be too high. Hence, the range of sample sizes in

rice GWAS that can produce meaningful results is typically 200 to 3000. While deciding on the sample size, there are a number of things to take into account GWAS of complex composite characteristics, which are traits governed by a number of genes, each with a tiny genetic influence (such as grain yield per plant), require a sizeable sample, whereas GWAS of qualitative qualities governed by a limited number of essential genes only need a small sample size. For GWAS of challenging to interpret complex features for properly examine (such heat tolerance), it is required to increase the number of samples as well as the number of duplicates within each sample. When traits are presumably controlled by genes with low-frequency alleles, the experimental design used for the GWAS must also be improved on the basis of sample size and sample diversity. Unless rice accessions with traits enhancing resistance to blast disease are enriched in the gathered populations, the GWAS population must be sizable [4].

2.4 Whole-genome variant genotyping:

Genotyping work can begin as soon as the rice accessions are required available for a GWAS. With a resolution of roughly 100 kb in indicia and 200 kb in japonica species on average, rice's linkage disequilibrium is relatively mild, necessitating the use of 1000 of segregating markers spread across the entire genome. A snips genotyping array is used for conduct in rice, GWAS prior to the widespread usage of high-throughput sequencing technology. Many common variations for complex features were discovered thanks to the increased genotyping resolution provided by chips [34].

New high-throughput genotyping methods for rice GWAS include second-generation sequencing [8]. There was a significant amount of missing genotypic information each addition of rice, as evidenced reads by the raw sequence coverage of >50% the genome of rice. Because of linkage disequilibrium between local region polymorphisms in rice, Statistical methods could be used to simulate the missing data [35]. The assertions of the genotypic sequence-based data were carried out using the K-nearest-neighbours technique, which performed admirably for the rice accessions. Second-generation sequencing has seen a significant improvement in throughput over the last ten years, and the cost has been falling quickly. The cost of building a library and performing whole-genome sequencing for a single rice accession is currently around $30; making rice GWAS use the sequencing-based genotyping technique as the default strategy [34].

The majority of rice accessions are inbred lines since rice is a self-pollinating plant. Both genotype calling and missing data imputation are considerably aided by the homozygous genotypes. However there are both a lot and a little bit of heterozygous genotypes found in native wild rice accessions and hybrid rice, respectively. By sequencing the inbred parental lines of hybrid rice accessions, the genotype of those accessions may be precisely known, and numerous generations of selfing can "purify" wild rice accessions grown under natural conditions. The processes would be more difficult without these data or trials since, it's possible that the raw genotypes are not the true genotypes, even with excellent coverage. For instance, it is possible that just one allele at a heterozygous location receives coverage from several reads; the other allele has not been sequenced. In order to resolve the ambiguity around heterozygous genotypes as well as sequencing and alignment issues, imputation approaches are required. With deeper sequencing levels and hidden Markov model-based imputation techniques, it is possible to improve genotyping determination considering these heterozygous genomes of wild rice or hybrid [36]. The performance is especially enhanced when reference haplotype maps are made available [37].

To find replicable genome-wide significant connections and the appropriate sample size, GWAS frequently need very high sample sizes. The research issue, the intended sample size, and the presence of existing data, or the simplicity with which new data can be obtained, all have an impact on the GWAS study design and data resource selection. Direct to consumer studies or information from sources like cohorts or bio banks that includes disease- or population-based data enrolment can be used to conduct GWAS. For a complicated characteristic, a well-powered GWAS needs substantial time and financial commitments that are beyond the capabilities of the majority of individual laboratories. However, the majority of GWAS are carried out using a number of great public resources that already exist and offer access to sizable cohorts with data on genotype and phenotype.

2.5 Phenotyping with high throughput:

Grain yield, salt resistance, grain quality, and stress resistance are agronomic traits, that have attracted a lot of attention in the rice molecular genetic studies, and rice GWAS rely heavily on a high-quality phenotypic dataset. Millions of rice accessions with numerous replicates were used to phenotype these agronomic features, which requires a great deal of labour and time, It often takes numerous researchers or farmers a few months or even two to three years to complete, which is substantially slower than the genotyping processes.

The possibility of high-throughput phenotyping in rice has been made possible by the incredibly quick improvements in robotics and remote sensing technology [38]. By obtaining relevant photos for the software processes for the analysis of images have been created to predict a range of phenotypic for each plant line of rice grown plant in greenhouse or in the wild. [39]. In current GWAS of rice's drought tolerance, morphological changes before and after drought shocks were measured using a computerised phenotyping platform [40]. The technology for high-throughput research is getting cheaper and smaller, while plants phenotypic picture recognition using artificial intelligence systems are getting more accurate. As a result, it is extremely possible that additional phenotyping studies for rice GWAS will carried out on systems that are automated and require little to no manual input.

In a broad sense, complex quantitative traits include methylation, metabolite, and gene expression profiling. Over the past few years, a rise in curiosity has been shown in utilising GWAS to examine genetic diversity in gene expression and methylation levels and metabolite content, between rice accessions. As an illustration, the metabolic GWAS of rice grains and leaves made it possible to identify and annotation of large number of candidate genes implicated in metabolic pathways [41]. The presence of metabolites was discovered to have potential physiological and dietary significance. An eGWAS is a GWAS that uses gene expression profiling data as the characteristics [42]. Along with the genes already recognised to play a role in epigenetic, in GWAS also discovered numerous unique sites that control the degree of DNA methylation in the *Arabidopsis* genome. Future investigations into the genetic differences that naturally arise in gene regulation on rice will provide essential resources & genetic insights [43].

2.6 GWAS association method:

A separate model was used to evaluate the relationship between the genotypes of the SNP markers and the associated characteristics. The Mixed Linear Model & General Linear Model was both used by TASSEL software to do association mapping analysis. [44].

TASSEL continues to serve as a tool for examining the link between phenotypes and genotypes, despite the fact that it has undergone significant changes since its original public release in 2001 [44]. TASSEL provides features for association study, assessing evolutionary links, principal component analysis, linkage disequilibrium analysis, missing data imputation (SNPs imputation), cluster analysis (phylogenetic analysis), and data visualisation. TASSEL design and computational optimisations take into account the biology present in many plants

and breeding scenarios because its team with expertise in corn genetics and genomics has been in charge of the development. Comparing crop genetics to human genetics reveals that inbreeding, large families, and entire genome prediction are all more common, and many crops exhibit high levels of nucleotides and structural variation. Due to these biological variations, other biological systems other than crops have benefited from various optimisations.

TASSEL was developed to be accessible to a wide range of users, including those with little prior knowledge regarding statistical genetics or computational science. In a few simple steps, a GWAS can be carried out by "clicking" on the appropriate options on a graphical interface, combining population structure data with cryptic associations using the mixed linear model approach [45, 46]. The entire analysis is completed automatically, including the import of genotypic and phenotypic data, the imputed generation of missing genotypic or phenotypic data, and the conclusion, filtering of markers based on the minor allele frequency (MAF), the creation of kinship and main component matrices to depict population structure and hidden relationships, as well as the optimisation of the compression level and GWAS execution.

The Genomic Association and Prediction Integrated Tool was used to assess the trait-SNP relationships for grain ionomic and agronomic traits in rice was utilised in conjunction with 2 univariate GWAS methods (GLM and MLM) and 2 multivariate GWAS methods (MLMM and FarmCPU) [47]. A package called GAPIT is used with the R programming language. EMMA, the unified mixed model, and P3D /EMMAx & compressed mixed linear model, are only a few of the cutting-edge statistical genetics techniques used in this programme [47].

The general linear model, also referred to as the generic multidimensional regression model, is an efficient way to write several different linear regression models simultaneously. In that regard, it cannot be considered a different statistical linear model. The following are short ways to express the many multiple linear regression models:-

$Y= XB+U,$

Here, Y is a matrix that is contains a number of the multivariate measurements.

X is a matrix of independent variable observations that could be a design matrix.

U is a matrix that contains error, and B is a matrix that contains parameters that must typically be approximated.

The error typically has a multivariate normal distribution and is considered to be uncorrelated across measurements. If the errors don't fit a multivariate normal distribution, generalised linear models can be employed to modify the Y and U assumptions [48].

MLM features both random and fixed effects. An MLM can contain information about inter-person interactions by including people as random effects. This relationship-related information is communicated by the kinship matrix that is used in an MLM (mixed linear model) as the matrix of variance-covariance across the individuals. The "Q+K" approach, which combines population structure and a genetic marker-based kinship matrix, has greater statistical power than the "Q" approach alone [49, 50].

Henderson's matrix notation can be used to describe an MLM is as follows:-

$Y = X\beta + Zu + e.$

Where Y is a vector encoding the traits discovered; β is an unidentified vector with fixed effects, including the genetic marker, population structure. U is an unidentified vector of the additive genetic effects resulting from different background QTLs for individuals. E is the undetected vector of residuals, whereas X & Z are the known design matrices.

In compressed MLM, Even though kinship matrix is obtained from each of the markers, using kinship (k) to test the MLM markers creates confusion among the testing markers and the genetic effects of the people, with the structure of variance represented by the kinship. To avoid confusion, Zhang et al.'s 2010 compressed MLM changed individuals by their corresponding groups [48]. To categorise people who are similar, cluster analysis is employed. In the clustering analysis, similarity metrics are applied to the kinship matrix's components. The lines can be grouped together using a variety of connection criteria. The user determines the number of groups. After the lines have been divided into groups, a reduced kinship matrix is constructed using summarised data on the relationships between and within subgroups. At each compression level, a reduced kinship matrix is produced using this process. The ideal compression setting is chosen by fitting a number of mixed models. Each model's log likelihood function value is calculated, and the level with the highest possible log likelihood functional value from the fitted mixed model is deemed the ideal compression level.

In Farm CPU model, In order to address the problems of false +ve controls & confusion between the testing marker and cofactors at the same time, the iterative technique known as

Fixed and Randomised Model Circulation Probabilities Unification was developed in 2016. To avoid false positives while evaluating the remaining markers, the connected markers discovered during iterations are fitted as cofactors in a fixed-effect model. Stepwise regression's over-model fitted issue is prevented by using a model with random effects as well as the maximum likelihood method to select the relevant markers [51].

2.7 Significant threshold value in GWAS

The rice GWAS's association significance threshold values are essential. The genome-wide significance cut-off value in the majority of human GWAS research is P value $< 5 \times 10{-}8$. With rice GWAS, there are no set or accepted thresholds, especially when multiple population and several marker counts are used. Permutations tests are used to calculate the GWAS the p-value threshold by rearranging phenotypic data & then running a GWAS via the rearranged phenotypes. The threshold 10−7 is appropriate in the majority of rice GWAS scenarios carried out with a linear mixed model. Furthermore, since each linkage disequilibrium block has enough markers and the majority of rice GWAS employ whole-genome sequencing data, it is frequently incorrect for a single SNP to pass the Manhattan plot criterion. In recent years, it has been demonstrated that algorithms, particularly deep learning, are quite effective in a variety of fields, including speech and image recognition. Deep learning didn't significantly outperform linear models, according to certain exploratory investigations in human genetics that compared their performance compared with deep learning's ability to solve challenges based on heredity [52].

2.8 Population structure

On the frequency and distribution of alleles controlling economically significant features, Analysis of population structure & genome-wide association studies carried out on agricultural germplasm collections are very helpful. The usefulness of these analyses is significantly increased when the accession numbers are raised from 1,000 to 10,000 or more. Population structure across landscapes also affects the source-sink relationships that control population survival and the ability to recolonize areas following disruptions. For instance, demes' size and dispersion affect their capacity to sustain gene flow or to diverge into new species. The distribution of demes provides information about the origin and initial genetic composition of colonists coming into a new habitat patch [53].

Genome association studies should take population structure analysis or genetic relatedness into account to prevent misleading associations. The most popular techniques for the

genome-wide association studies take population structure into consideration; however they are only applicable to genotyped people with phenotypes. Phenotypes from ungenotyped relatives can be used in single-step GWAS (ssGWAS), although its capacity to take population structure into account hasn't been thoroughly studied. Here, we evaluate at how single-SNP analyses without population structure adjustment and efficient mixed-model association expedited and genomic best linear unbiased prediction GWAS compare to each other [54].

Using a Bayesian technique, the population structure was examined. STRUCTURE software was used to achieve model-based clustering. K-values or the population's estimated fixed there are between 1 to 12 subpopulations was used to assess population structure. For each K-value, three independent analyses were employed, and the STRUCTURE programme was configured with 100,000 Markov Chain Monte Carlo and 50,000 burn-in iterations replications following burn-in period [55].

2.9 Population use in GWAS study

GWAS that are population-based, GWAS frequently use & phenotypic findings from cohorts based on population data, where participants are thought to have been randomly selected from the population. Testing for associations with genotyped or imputed variations can be done for phenotypes related to continuous or binary dependent variables. Case-control research, where controls & cases are differentiated according to the existence or absence of the particular feature, is a typical GWAS design. The case and control cohorts are frequently chosen on purpose in case-control studies in order to ensure that the average number of the cases does not coincide with the frequency determined by the population. As a result, the statistical analyses should take this into account; for instance, covariate correction would require further thought [56].

Using controls from population cohorts with uncertain illness status can enable an increase cases in people's lives in the control group, even If the population's diseases won't be affected much by this < 1%. As an alternative, controls also are purposefully similar to patients based on ancestry and gender. It has been demonstrated that the latter strategy is cost-effective and has appropriate power when the disease frequency in the population is low (20%). Active case and control recruiting is typically favoured when there are limited financial resources and a need to increase statistical power [57].

If Cases & controls must be genotyped on the same chip at the same time to limit artefacts, then extra care must be given during quality control and following analysis [58]. Noteworthy is the fact that, while sample is considered to be picked at randomly using the population, It's not like that, Participation is prejudiced and out of place & social-demographic characteristics make this assumption false [59].

GWAS was in its early stages, 1st -degree relatives were commonly utilised in association testing, since well-phenotyping identical and other familial cohorts are available [60]. Family-based GWAS require bigger sample numbers to attain the identical statistical strength as unrelated people. Concerns with incorrect grouping in based on populations GWAS have lately revived interest in doing within-family investigations [61]. Within-family approaches often involve variation on the gearbox incompatibility test to study the dispersion of an allele amongst a population of people. PLINK allows for the implementation of many variations of this test, like a test for qualities that are quantitative which includes Organisation both within and between families, albeit, critically, and only within the family component is resistant for stratified populations [62].

The advantage that incorporating family information into GWAS has this benefit it enables to researchers to examine how an allele affects a person's phenotype through the indirect effects on members of their immediate family [63]. Furthermore, it has been demonstrated that employing phenotypic data from non-genotyped family members, or "GWAS by proxy," greatly boosts power for a few features, when researching late-onset disorders for which those have some challenging to gather big data sets. A word warning: Self-reported family history is used in GWAS by proxy, and is not necessarily reliable [64].

Here are several advantages of doing GWAS in communities that have been separated for an extended period of time because of a founder event like physical or cultural barriers, have limited gene exchange with nearby populations, or both. An important advantage is that isolated populations may have more instances of functional variations that are often rare, increasing the power of association's research for these variants [58]. Suppose even only a few people from the isolated group are included in the reference panel, the lengthy linkage disequilibrium distinctive of isolated populations enhances imputation accuracy and power above comparably large non-isolated cohorts [59].

Due of the close kinship among isolated populations; GWAS frequently uses a method based on a linear mixed model. Due to genetic bottlenecks that cause alleles to disappear, isolated

populations have a tendency to have high genetic homogeneity, which by fewer neutral variants can increase the strength of burden testing [60]. If a variant is sufficiently uncommon, discoveries made in remote groups could be difficult to replicate in other populations, even if other variants involving the same genes can provide more evidence. For instance, variations linking APOA5 to myocardial infarction in various European communities may support findings linking triglycerides to the disease in the Sardinian community [61].

In Biobank, Researchers have access to a variety of sizable, public population biobanks. Biobanks contain information from tens of thousands of genotyped people who have undergone extensive phenotyping via lab tests, surveys. These individuals were not chosen for specific illness features. The UK Bio bank is one notable example, it has allowed for well-powered GWAS of hundreds of quantitative variables, such as blood cell features, depressive symptoms, anthropometric traits, metabolites, and brain imaging traits, in addition to increasing sample sizes for GWAS of common disorders. [62].

2.10 Linkage disequilibrium

It is therefore possible to determine whether selection was natural, epigenetic, or due to other mechanisms like as genetic drift or gene flow, or to past genetic conditions and limitations, thanks to the insights into the genetic limitations and situations of the past that are provided. When examining populations, LD may be found throughout the genome, which can reflect patterns in the breeding system, geographic subdivision, and population history. When LD studies genomic regions, it reveals the history of gene conversion, natural selection, mutation, and other factors that either cause or contribute to gene-frequency evolution.

As a result, finding LD does not demonstrate linkage or absence of equilibrium. The local rate of recombination ultimately determines how the aforementioned criteria affect LD between paired loci or within a specific genomic region. Chromosome linkage is the hypothesis that two markers on a chromosome will stay physically connected over the course of multiple generations of a family, is related to this, as is to be expected. However, chromosomal portions within a family will be split apart by recombination events from generation to generation, and the effect is increased over several more generations. Until linkage equilibrium is reached among alleles in a population, recombination events will inevitably separate portions of chromosomes carrying linked alleles. Linkage disequilibrium, put simply, is the coupling of population-level indicators.

Additionally, the number of founding chromosomes within a population, the population's size, the population's age and the number of generations has been around all affect the LD decay rate. Therefore, when comparing distinct human subpopulations, it is not surprising to see a range of LD levels and patterns.

Strong LD exists between tightly related polymorphic SNPs in general. According to research by the International Hap Map project, the human genome comprises haplotype blocks that either contain all or the majority of high LD SNPs. As a result, LD is prevalent on a finer scale in human populations. As a result, it has been believed that SNPs with significant levels of LD correspond to alleles with elevated chances of complicated hereditary disorders.

When GWAS are conducted and a significant number of SNPs are surveyed, it's interesting to note that this has really been analysed for those SNPs that are strongly associated with breast cancer. However, it's crucial to remember that LD in GWAS can result from population stratification that was either missed or was not known to exist.

Linkage disequilibrium is a phrase that unfortunately hides its meaning. Every population genetics educator is aware of the term's limitations rather than benefits. A lack of equilibrium or linkage is not automatically implied by the presence of LD, which is merely a non-random connection between alleles at more than one locus. The term 1st was coined by Lewontin and Kojima in 1960, and it has stuck around ever since because LD was first a topic of interest for population geneticists who didn't care what they called things as long as the mathematics made sense. Because there were initially few data available for LD studies, its importance for the genetics and evolutionary biology in humans was disregarded besides population genetics [63]. However, when extensive analyses of closely related loci became practical and the value of LD for the mapping of genes became apparent, interest in LD grew rapidly in the 1980s. A phrase had already become too well-known by that point to be changed.

LD is significant in evolutionary biology and human genetics because of the numerous factors that influence and are influenced by it. The potential to respond in both natural selection and artificial selection is constrained by LD, which further offers information on past occurrences. LD across the genome reveals the breeding system, history of the population and the pattern of geographic subdivision, while LD in each genomic region reflects the history of factors like gene conversion, natural selection, mutation, and other methods that impact the evolution of gene-frequency. Local recombination rates control how these factors influence the degree of differentiation within a genomic area or a group of loci.

A substantial effort is being made for mapping genes in human beings and other species using the well established population genetics concept LD. [64].

The extent to which an allele of one SNP is inherited or linked with an allele of another SNP in a population is described by a property of SNPs called linkage disequilibrium (LD) on a continuous stretch of genomic sequence. Population geneticists developed the concept of linkage disequilibrium to statistically evaluate modifications to genetic variation over a perio d of time among a population. Two chromosomal markers continue to remain physically linked over the course of a family's generations, which is related to the idea of chromosomal linkage. Chromosome segments are split apart from generation to generation by recombination events within a family. Through successive generations, repeated random recombination events in a fixe sized population that experiences random mating will split up regions of contiguous chromosomes (carrying linked alleles) until finally all of the population's alleles are independent or in linkage equilibrium. This effect is magnified. Therefore, linkage disequilibrium describes the relationship amongst biomarkers on a population scale [55].

How many founding chromosomes there are in the population, the population's size, and the amount of generations the population has been around are all factors that affect the rate of LD decay. As a result, the level & patterns of LD vary among various human subpopulations. The underlying element of all recommended assessments of LD is the discrepancy among the actual frequency of overlap for two alleles and the frequency expected when two markers are distinct [65].

Since each SNP is genotyped separately and the stage or chromosome of genesis for each alle le is unknown, present technology cannot be used for a direct evaluation of haplotype frequen cies from a sample. This is a problem that is occasionally missed with LD measurements.
There are numerous established and well-documented techniques for determining the two-marker haplotype frequencies and inferring haplotype phase, and these techniques typically produce accurate findings [66].

Tag SNPs are SNPs that have been specifically chosen to capture variation at neighbouring genomic sites since each of the the SNPs' alleles tag a nearby stretches of LD. Since LD patterns are population-specific, as was already explained, tag SNPs chosen for one population could not be useful for another. By avoiding genotyping repetitive information-giving SNPs, LD is used to enhance genetic investigations [67].

2.11 Meta-Analysis in GWAS

To do a meta-analysis, the findings of various GWAS research can be combined. Meta-analysis techniques were developed with the initial goal of assessing and enhancing estimations of significant and effect size resulting from numerous studies investigating an identical assumption in the research that has been published. The emergence of huge academic consortiums has made it possible to synthesise data from numerous studies using meta-analysis methods; no protected genotype or clinical data must be disclosed for parties who were not involved in the only statistical information from a study that was approved in the original need be transferred. For instance, a meta-analysis of 46 researches was used to inform a recent study that looked at lipid profiles [68].

The research that was included in the meta-analysis looked at the same hypothesis, which is a key concept. The assessment of clinical characteristics and phenotypes, as well as the methods used to select which SNPs have been included from each site and any covariate adjustments should be similar across numerous sites. The idea that the sample sets used in each study should be independent should always be tested because researchers commonly use the same samples in multiple studies. Making sure that all research publishes results in relation to a shared genomic structure and reference allele is also a very crucial and somewhat annoying logistical issue. Because the effects of the two studies cancel each other out, if results from one study are published in relation to the A allele and those from another study in relation to allele B, a meta-analysis outcome for this SNP could not be significant [69].

It is uncommon to discover several studies that match completely on every criterion when all of these aspects are taken into account. As a result, to determine how distinct studies are from one another, research diversity is typically analytically evaluated through a meta-analysis. In the same way that outlier analysis is used to discover locations with excessive influence. It is crucial to emphasise that these data points should be used as a guide when looking for studies that could test a different core hypothesis from the ones that were examined by the other studies that made up the meta-analysis. However, much as with outliers, a study should only be disregarded if there is a clear justification based on the study's parameters, not just because a statistic implies that the study raises diversity. Otherwise, statistical methods employed in meta-analysis to reduce heterogeneity will produce more erroneous findings [70].

Genome-wide association studies have significantly influenced the field of human genetics. They have helped the genetics community think on a genome-wide scale and revealed novel risk factors that are genetic for a number of common human diseases. The whole genome will soon be sequenced. In the coming years, with the entire 300,000,000 nucleotide genome sequence in place of one million SNPs, inexpensive sequencing technology will be used. The infrastructure and knowledge of computer science and bioinformatics will be put to the test by exponentially more complicated challenges related to storage and processing of data, data analysis and quality control. Combining sequencing data with information from other highly efficient technologies, such as the enormous amounts of data generated by neuro-imaging, will only make it more difficult to comprehend the connection between phenotype and genotype and, ultimately, to improve healthcare. Data manipulation and storing these additional high-throughput techniques assess phenotypes, the environment, the proteome, and the transcriptase. The future of human genetics lies on the integration of these numerous levels of complicated biological data and their coupling with experimental methods [71].

CHAPTER 3 – MATERIAL & METHODS

3.1 MATERIAL & METHODS

GWAS will perform for 186 genotype with 50051 high-quality SNPs. We obtained the phenotypic data for rice core accessions from IARI (Indian Agricultural Research Institute) four yield-related traits (including Height, grain yield, grain no. per meter square, GN_P) were chosen to investigate whether there are any significant phenotypic differences in rice yield among the 186 rice types.

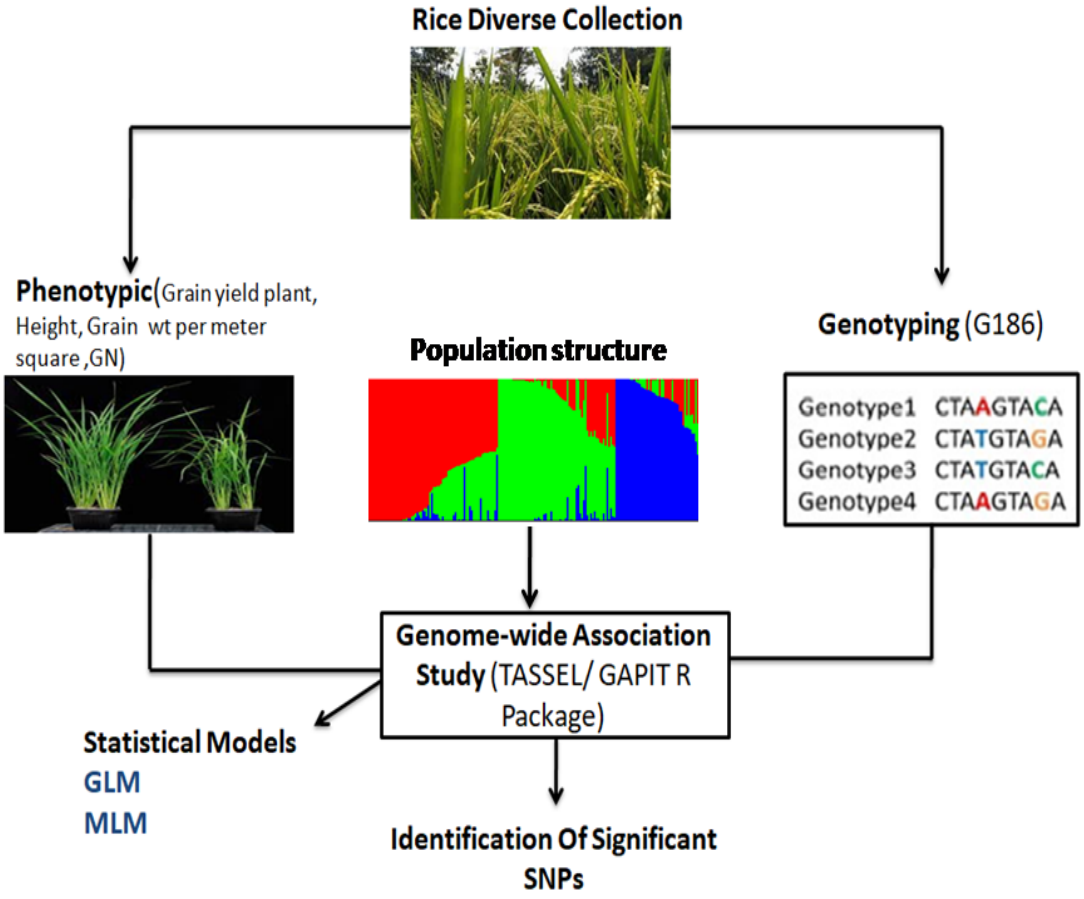


Fig. 1: The overview of the GWAS methodology.

Rice diverse collection

The rice diverse collection was taken from IARI (ICAR-Indian Agricultural Research Institute, New Delhi).

Genotyping

50051 high-quality SNPs were used in GWAS for 186 different rice types. We used hap map format for genotyping data in our study.

Phenotyping

In this study we analyzed a collection of rice varieties to perform association studies with traits of high agronomical interest such as; four agronomics traits (plant height, grain weight per meter square, grain yield plant and grain number per plant).

Genome-wide association studies for 4 agronomics traits

GWAS was carried out across 186 rice accessions descended from IARI (Indian agriculture research institute) with 50051 high-qualities SNPs. Association analyses were performed using two univariate models, including MLM (mixed linear model) and GLM (general linear model) used to assess the relationship between grain agronomic characteristics and SNPs by using the TASSEL software version 5.0 (https://bitbucket.org/tasseladmin/tassel-5-source) [41].

A software programme called TASSEL is used to assess the relationships between characteristics, evolutionary trends, and linkage disequilibrium. TASSEL's design and computational optimisations take into account the biology present in various plants and breeding circumstances because its development has been directed by a team specialising in maize genetics and genomics. Numerous crops exhibit high nucleotide and structural diversity when compared to human genetics.

After this, association analyses were performed using two models, GLM and MLM by using Genomic Association and Prediction Integrated Tool (GAPIT R Package) (http://www.r-project.org & source (http://zzlab.net/GAPIT/gapit_functions.txt) [44].

GAPIT is a Genome Association and Prediction Integrated Tool that is free to the public. It has been periodically updated to include the most recent approaches for the Genomic Selection and Genome Wide Association Study.

GAPIT Following association analysis, GWAS findings were presented as Manhattan plots based on the observed p-values for each SNP-trait relationship that had been negatively (-log10) transformed. To deem an SNP significant, we utilised a threshold of above log 3.

Identification of significant SNPs

The significant SNPs were identified by using different model (GLM &MLM) and different software TASSEL and GAPIT R Package. We did a comparative analysis between GAPIT tool and TASSEL software by using GLM & MLM models. After this, the common SNPs which were identified in GAPIT & TASSEL by using the different model GLM &MLM were considered as significant SNPs.

Population structure analysis & phylogenetic analysis

For population structure analysis, we analysed by using the structure software version 2.3.4. (http://web.stanford.edu/group/pritchardlab/structure.html.) [72]. STRUCTURE software is a freely software tool for investigating population structure using multi-locus genotyping data. Its applications include inferring presence of the separate populations, assigning people to populations, researching hybrid zone, detecting migrants & admixed individuals, & calculating populations' allele frequencies in situations when many individuals are migrants or admixed. It is applicable to the vast majority of frequently used SNPs, including genetic markers.

The STRUCTURE algorithm employs a systematic Bayesian clustering approach with Markov Chain Monte Carlo (MCMC) estimate. In the MCMC process, people are first randomly assigned to a set number of groups, after which the variant frequencies in each group are estimated and new groupings of individuals are made based on those estimates. We ran a series of model with ranging from 1 to 5, in all the loci. We fixed a Length of Burnin Period: - 10,000 and No. of the MCMC Reps after Burnin: - 100000. Find out the highest peak of delta K value by using the structure harvester. The tool offers a quick way to evaluate and show likelihood values over a range of K values and hundreds of iterations, making it simpler to identify how many genetic groups best suit the data.

The phylogenetic tree was created by the TASSEL software version 5.0 (https://bitbucket.org/tasseladmin/tassel-5-source) [41] with the NJ method (Neighbor-joining method). To examine the cluster in the rice subgroup, the tree was plotted using Archaeopteryx.

Analysis of linkage disequilibrium

Linkage disequilibrium analysis was determined by using the TASSEL software.

Identification of candidate gene

In this study, we identified a candidate gene and its annotation by using the rice genome annotation project (http://rice.uga.edu/).we identified the candidate gene by using the locus search.

CHAPTER 4 - RESULTS

4.1 Population structure & phylogenetic analysis:

To comprehend the population structure of the rice diversity collection, we analysed population structure by using the STRUCTURE software based on the admixture- based model. The highest peak value for Δk value was noticed at K = 4 (fig.5). Here four population observed in the rice diversity panel.

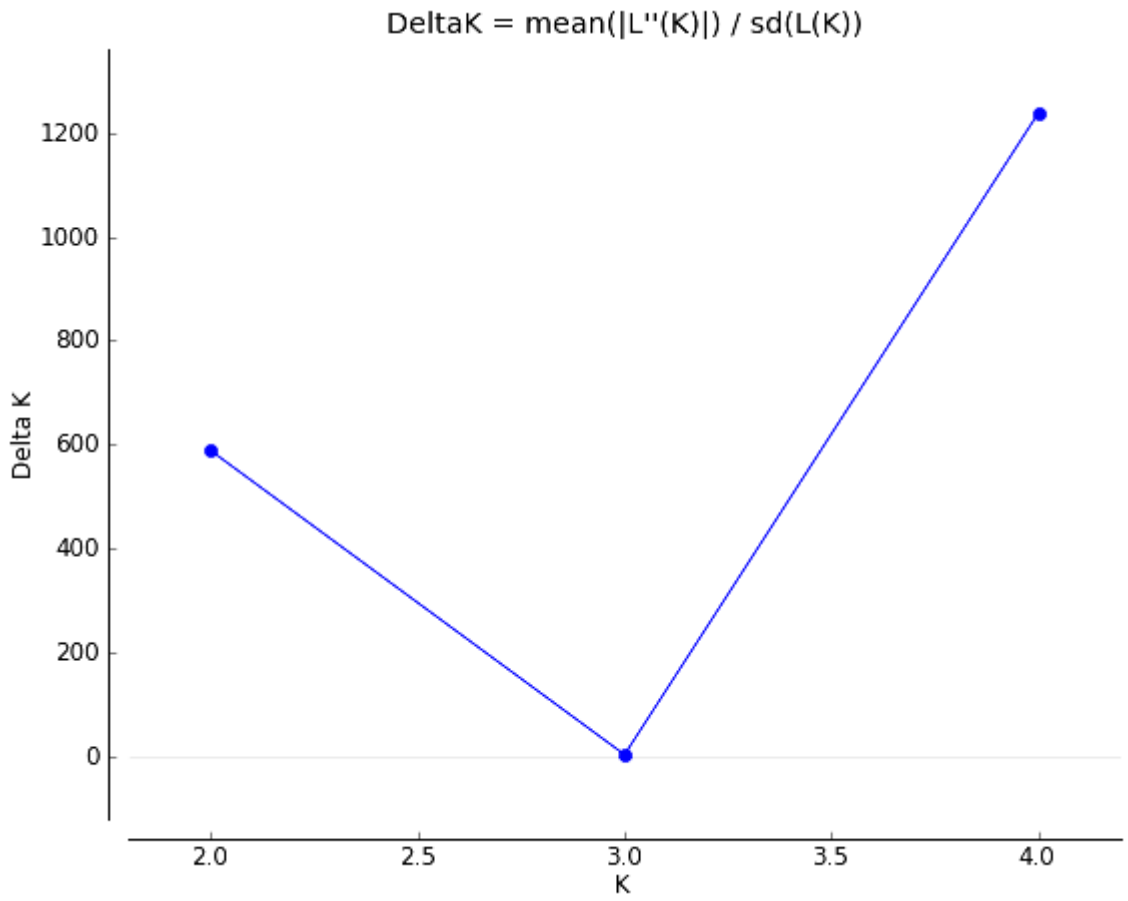
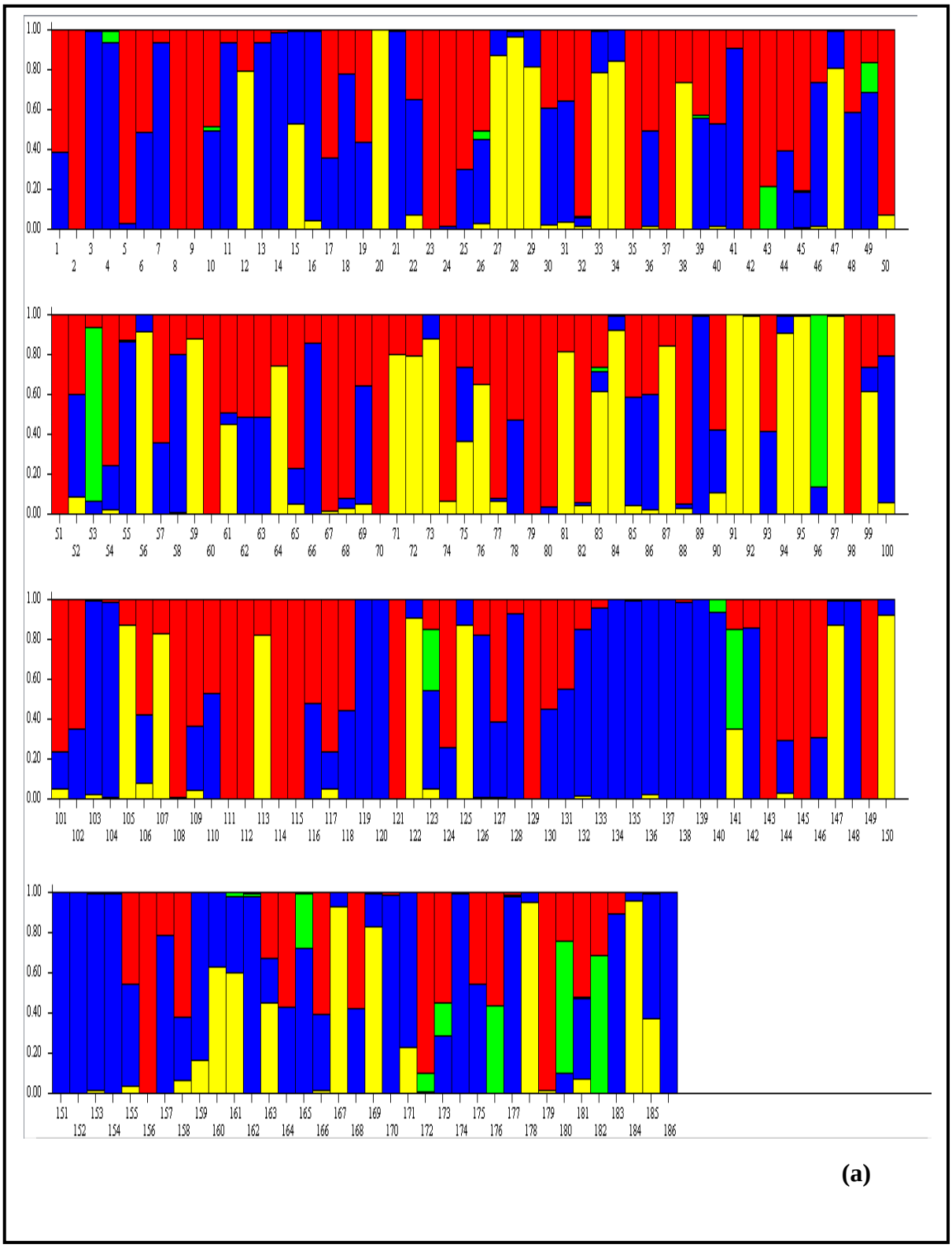


Fig. 2: This plot showing the maximum peak for Δk value. The y-axis shows the Δk value & X-axis displays the k value. The maximum peak correspond to k = 4.


1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186

(a)

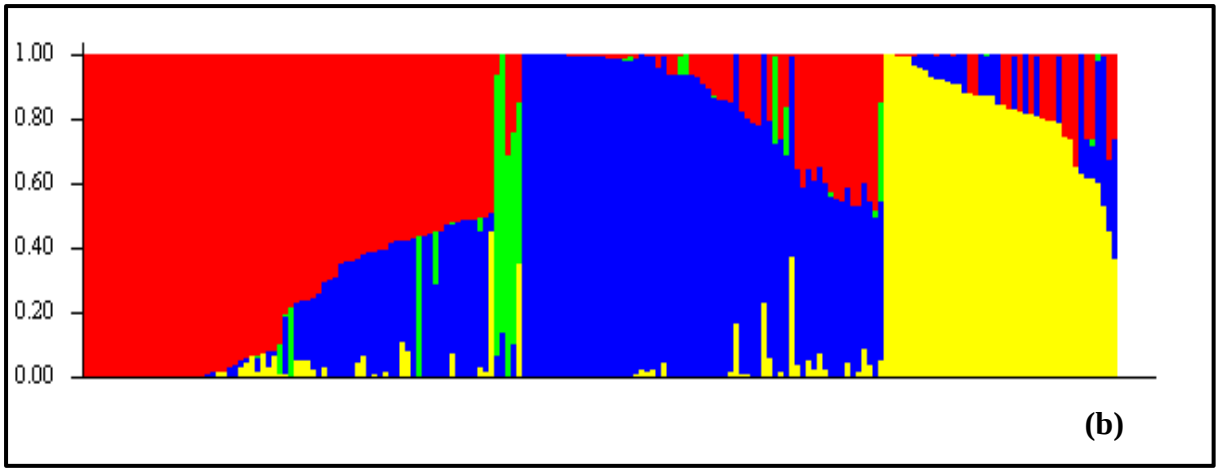
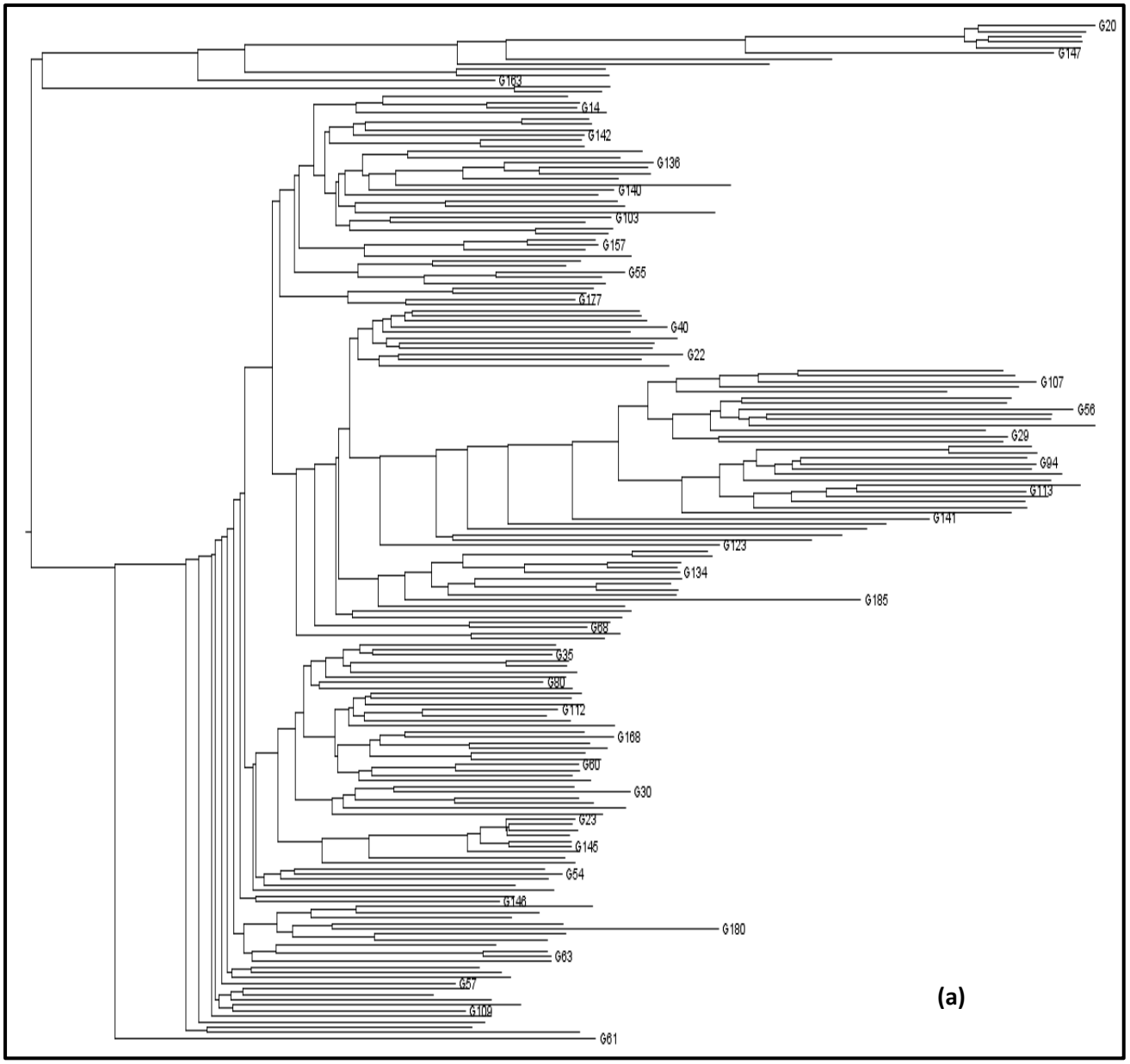


Fig. 3: Population structure of the rice diverse collection showing four subpopulations. **(a)** This is a graph of a multiple lines of different population **(b)** This graph is a Q-short graph of population structure analysis.

Phylogenetic analysis

High-quality SNPs (50051 SNPs) were chosen to create a NJ (neighbour-joining) tree to show how the 186 rice accessions are related phylogenetically. In a phylogenetics analysis, in a total of individuals were presented in two clusters (cluster I and cluster II). The phylogenetic tree of the 186 rice accessions was grouped into two clusters.


G20
G147
G163
G14
G142
G136
G140
G103
G157
G55
G177
G40
G22
G107
G56
G29
G94
G113
G141
G123
G134
G185
G88
G35
G80
G172
G168
G60
G30
G23
G145
G54
G146
G180
G63
G57
G109
G61
(a)

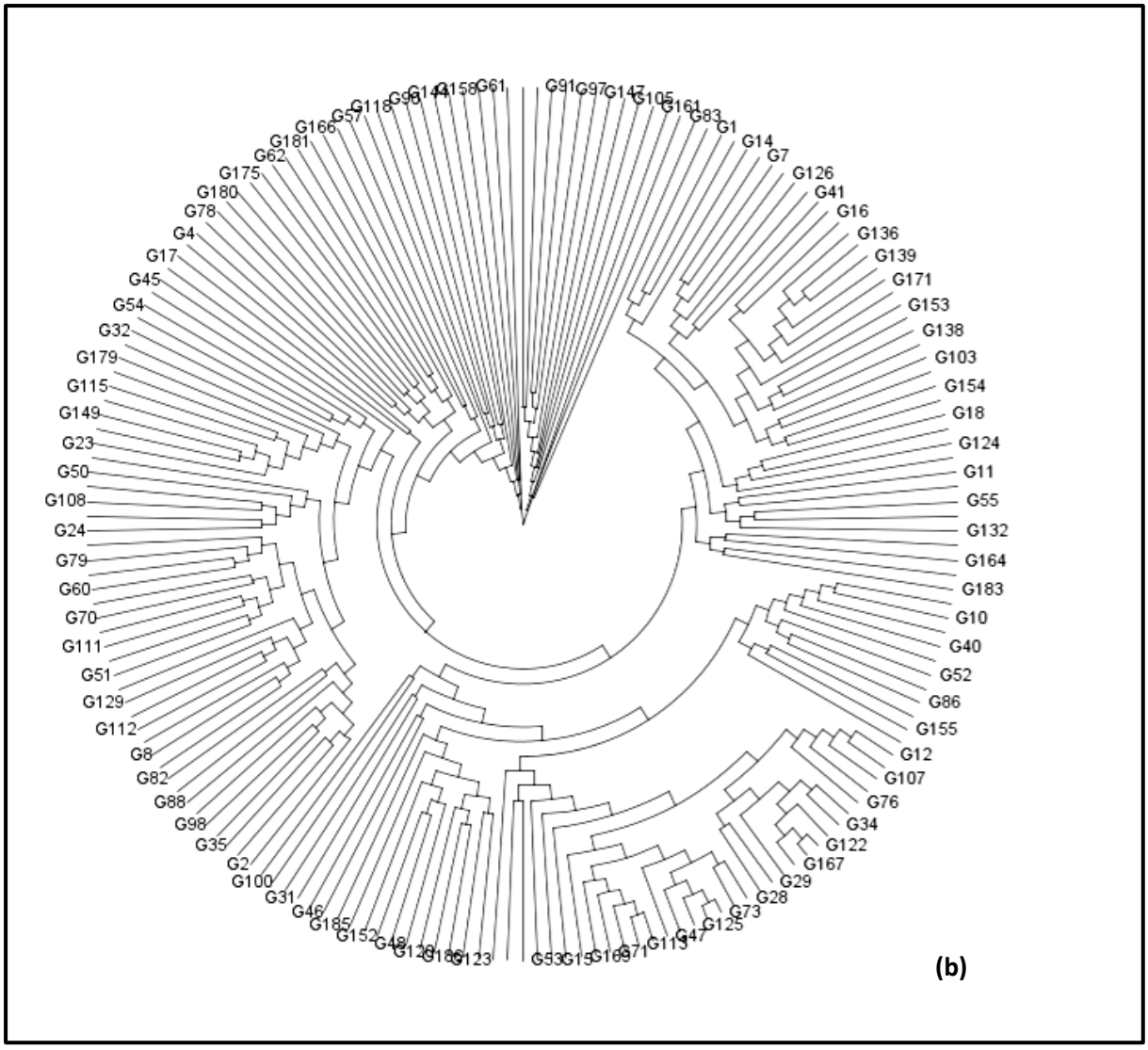


Fig. 4: A Phylogenetic tree analysis of 186 accessions of rice by using the NJ method. **(a)** Phylogenetic tree in a circular (alpha) type. (**b)** phylogenetic tree in a rectangular type.

4.2 Linkage Disequilibrium:

The TASSEL generated the triangle plot of linkage disequilibrium. The comparison of two sets of marker sites is shown in each cell, with colour coding used to indicate the presence of significant LD. The significant threshold values in both diagonals are represented by coloured bar codes. A hypothetical genome fragment's genetic distance scale was manually created.

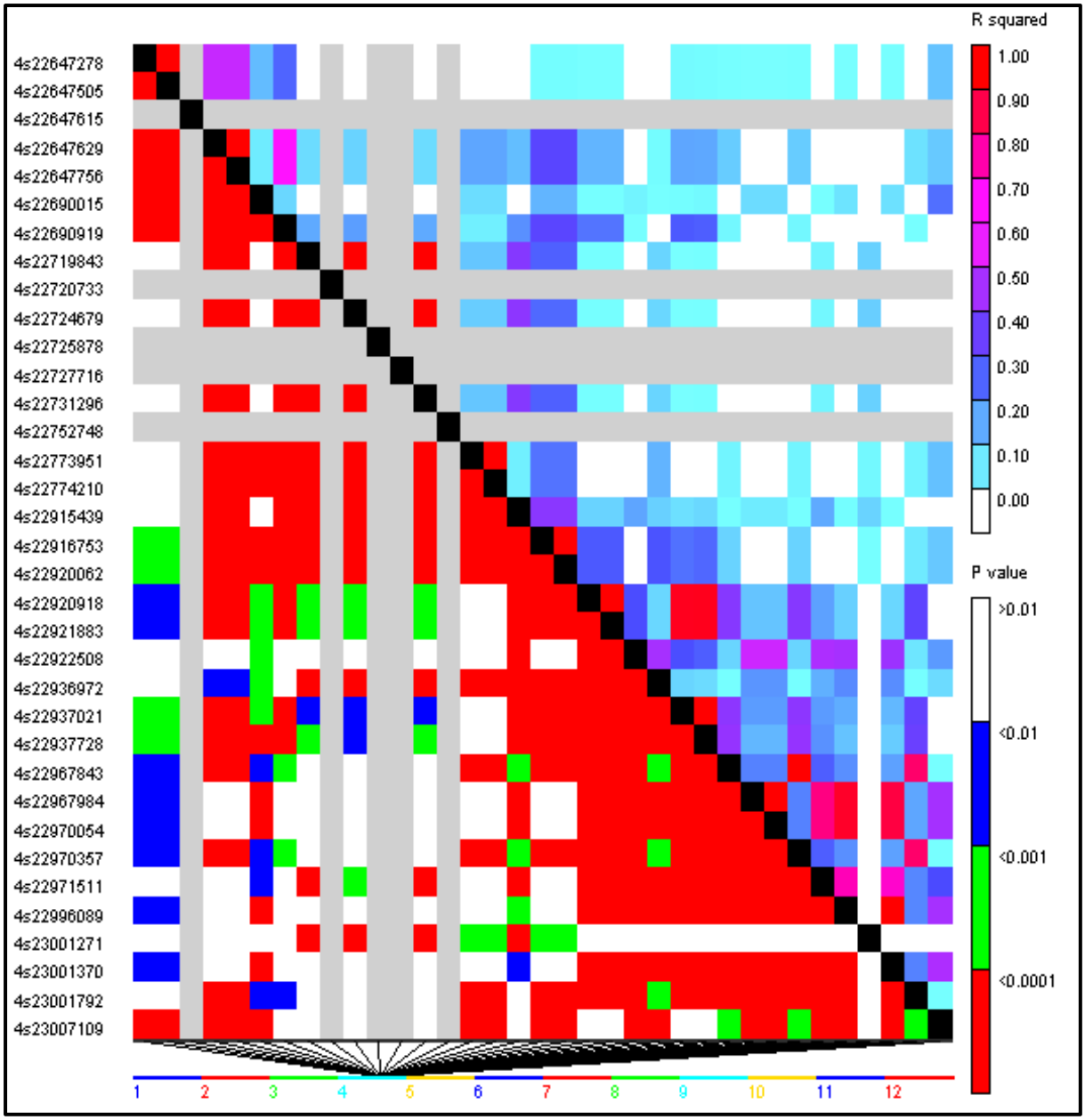


Fig. 5: The triangle plot was created by TASSEL. Above the diagonal shows the r2 values and below the diagonal shows the corresponding p-values.

4.3 Genome-wide association study by TASSEL software with using the GLM & MLM models:

Four agronomics traits (plant height, grain weight per meter square, grain yield plant, and grain number per plant) were examined utilising 2 univariate GWAS (GLM & MLM) methods to identify the significant SNPs.

A total number of 50051 Snips was used for GWAS analysis using two univariate methods using GLM and MLM. Significant SNPs were identified using *p*-value <.001 & 0.0001. Under the *p*-value <0.001, GLM model identified 14, 798, 225 and 1654 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 5, 12, 33, and 77 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively. Under the *p*-value <0.0001, GLM model identified 2, 26, 44, and 187 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 1, 7, 1, and 0 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively.

Table 1: List of total number of SNPs after using the filter of *p*-value >0.001 by TASSEL SOFTWARE (GLM model).

Traits	**Total SNPs**	**After using the filter of *p*-value >0.001**	**MAF <0.5**
Height	50051	1654	34044
Grain yield plant	50051	225	34044
Grain wt. per meter square	50051	798	34044
Grain no. per plant	50051	14	34044

Table 2: List of total number of SNPs after using the filter of *p*-value > 0.001 by TASSEL software (MLM model)

Traits	**Total SNPs**	**After using the filter of *p*-value >0.001**	**MAF <0.5**
Height	50051	77	34044
Grain yield plant	50051	33	34044
Grain wt. per meter square	50051	12	34044
Grain no. per plant	50051	5	34044

4.4 Genome-wide association study by GAPIT R packages with using the GLM & MLM models:

Further, Four agronomics traits (plant height, grain weight per meter square, grain yield plant, and grain number per plant) were analyzed by GAPIT R package with using the two univariate GWAS (GLM & MLM) methods to identify the significant SNPs.

Total number of 50051 SNPs was used for GWAS analysis using two univariate methods using GLM and MLM. Significant SNPs were identified using *p*-value <.001 & 0.0001. Under the *p*-value <0.001, GLM model identified 129, 1873, 1253 and 2756 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 7, 29, 31 and 38 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively. Under the *p*-value <0.0001, GLM model identified 14, 798, 225 and 1654 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 2, 1, 1 and 6 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively.

Table 3: List of total number of SNPs after using the filter of *p*-value by GAPIT R package (GLM model).

Traits	**Total SNPs**	**After using the filter of *p*-value >0.001**	**MAF <0.5**
Height	50051	2756	34044
Grain yield plant	50051	1253	34044
Grain wt. per meter square	50051	1873	34044
Grain no. per plant	50051	129	34044

Table 4: List of total number of SNPs after using the filter of *p*-value by GAPIT R package (MLM model)

Traits	**Total SNPs**	**After using the filter of *p*-value >0.001**	**MAF <0.5**
Height	50051	38	34044
Grain yield plant	50051	31	34044
Grain wt. per meter square	50051	29	34044
Grain no. per plant	50051	31	34044

A Millions of genetic variants can be shown in a single figure using a Manhattan plot, which shows chromosomes on the x-axis and the association statistical significance value as -log10 (*p*-value) on the y-axis. In below, figure 2 and figure 3 shows the Manhattan plot of different traits by using two different models (GLM and MLM). Above the threshold p- value 3, all the SNPs considers as a significant SNPs in this plot.

Quantile-quantile plots, also referred to as Q-Q plots, are a probability graphic that contrasts the quantiles of two probability distributions. In QQ-plot, the y axis displays the observed -log10 while the x axis displays the predicted log10 (*p*-value).

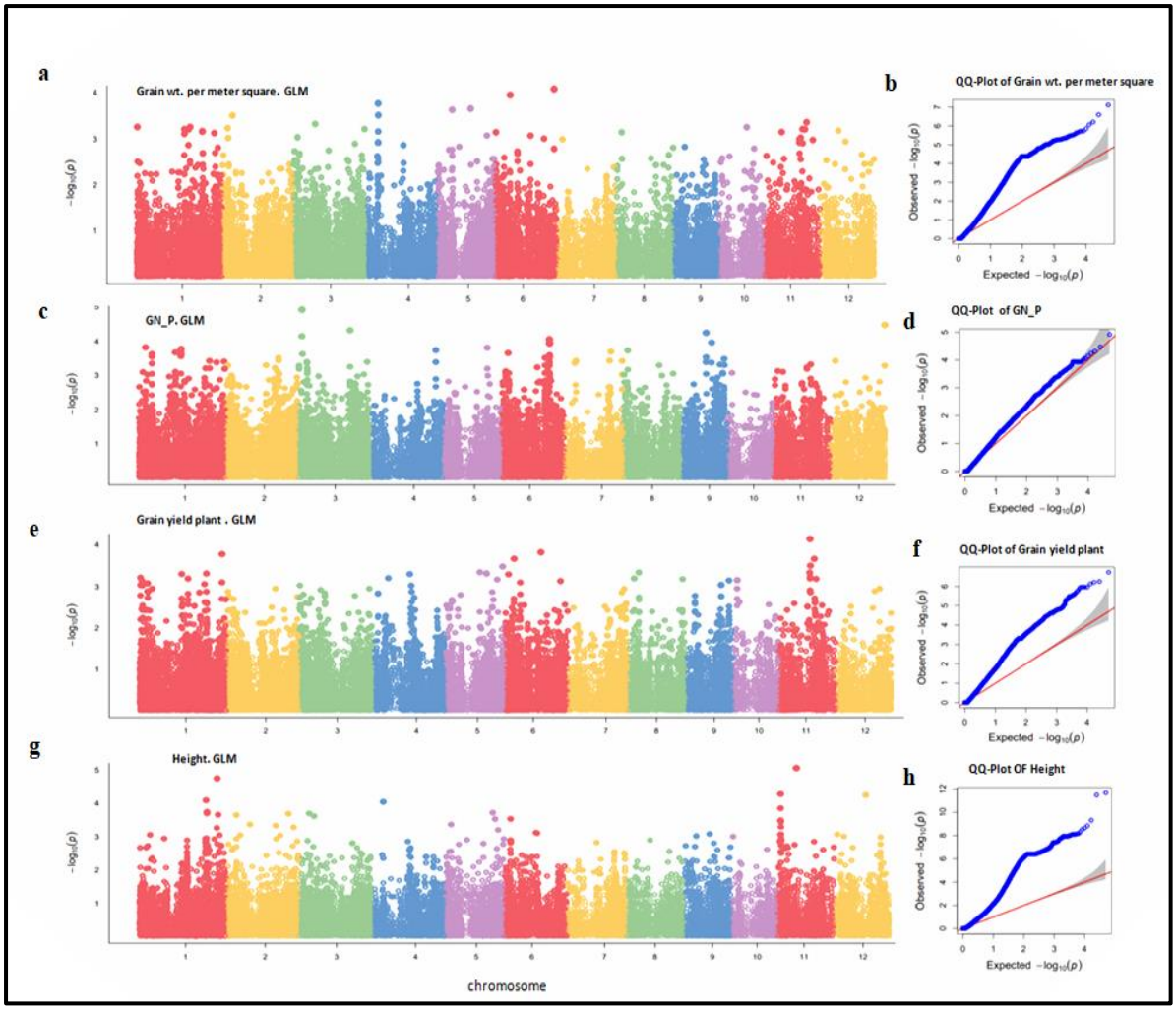


Fig. 6: Genome-wide association studies of four different traits. (a) This is a Manhattan plot for the simple model for grain weight per meter square. Which shows chromosomes on the x-axis and the association statistical significance value as -log10 (*p*-value) on the y-axis. (b) Quantile-quantile (Q-Q) plot of the MLM model for grain weight per meter square. (c) Manhattan plots for the GLM for grain number per plant (GN_P). (d) Q-Q plot of GLM for grain number per plant. (e) Manhattan plots of the GLM model for grain yield plant (f) Q-Q plot of the GLM model for grain yield plant. (g) Manhattan plots of the GLM model for plant height (h) Q-Q plot of GLM for plant height (ht.)

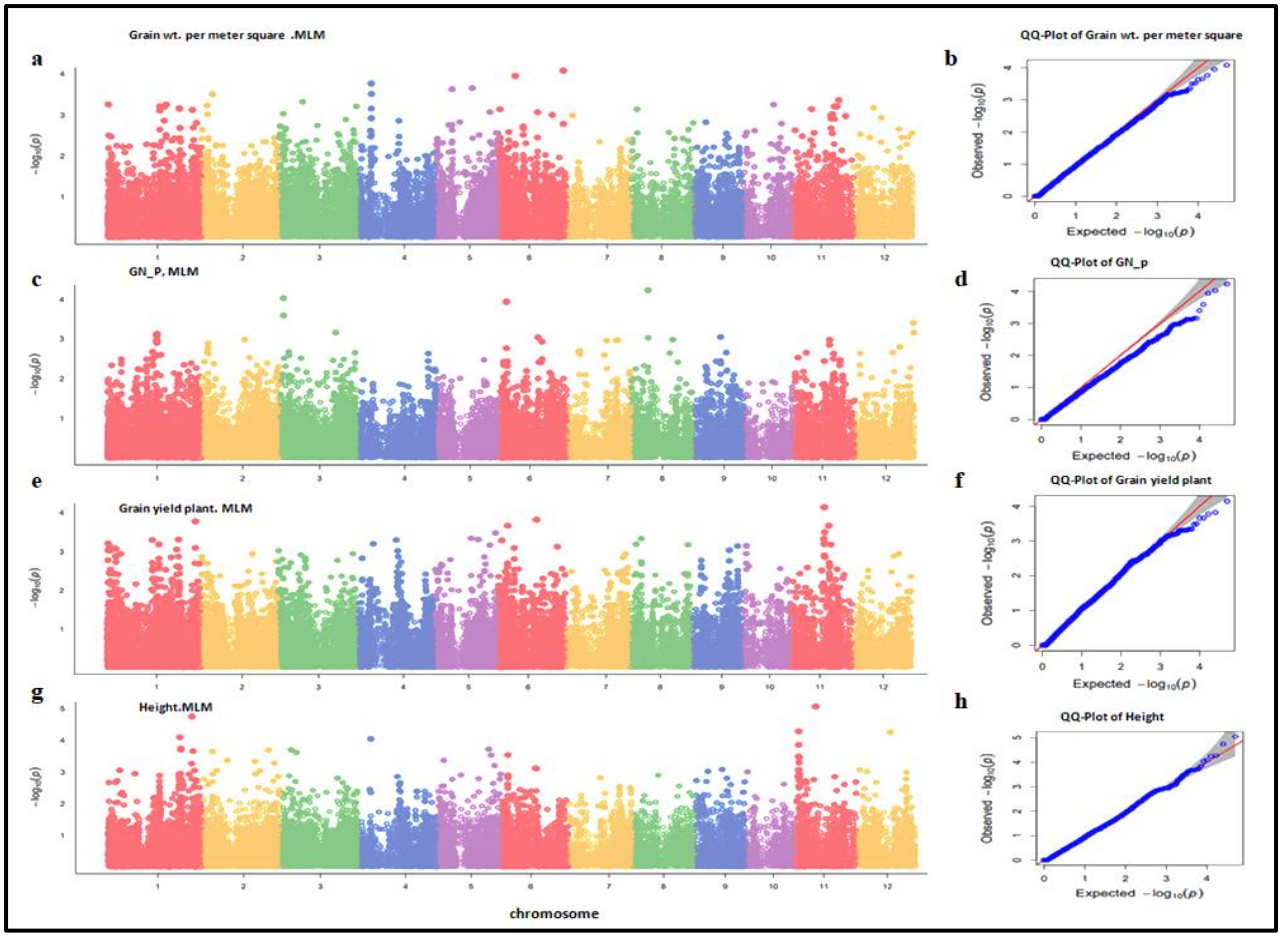


Fig.7: Genome-wide association studies of four different traits. (a) This is a Manhattan plot for the simple model for grain weight per meter square. Which shows chromosomes on the x-axis and the association statistical significance value as -log10 (*p*-value) on the y-axis. (b) Quantile-quantile (Q-Q) plot of the MLM model for grain weight per meter square. (c) Manhattan plots for the MLM for grain number per plant (GN_P). (d) Q-Q plot of MLM for grain number per plant. (e) Manhattan plots of the MLM model for grain yield plant (f) Q-Q plot of the MLM model for grain yield plant. (g) Manhattan plots of the MLM model for plant height (h) Q-Q plot of MLM for plant height (ht.)

4.5 Identification of common significant SNPs by using the comparative analysis between TASSEL software and GAPIT R package:

A total no. of 50051 Snips was using for GWAS analysis by using two univariate models (GLM and MLM) with TASSEL software and GAPIT R package. After doing the comparative analysis, common significant SNPs were identified by using the *p*-value <0.001

& <0.0001. A total of 23 common significant SNPs those are associate with different traits were identified with GAPIT R package and TASSEL software.

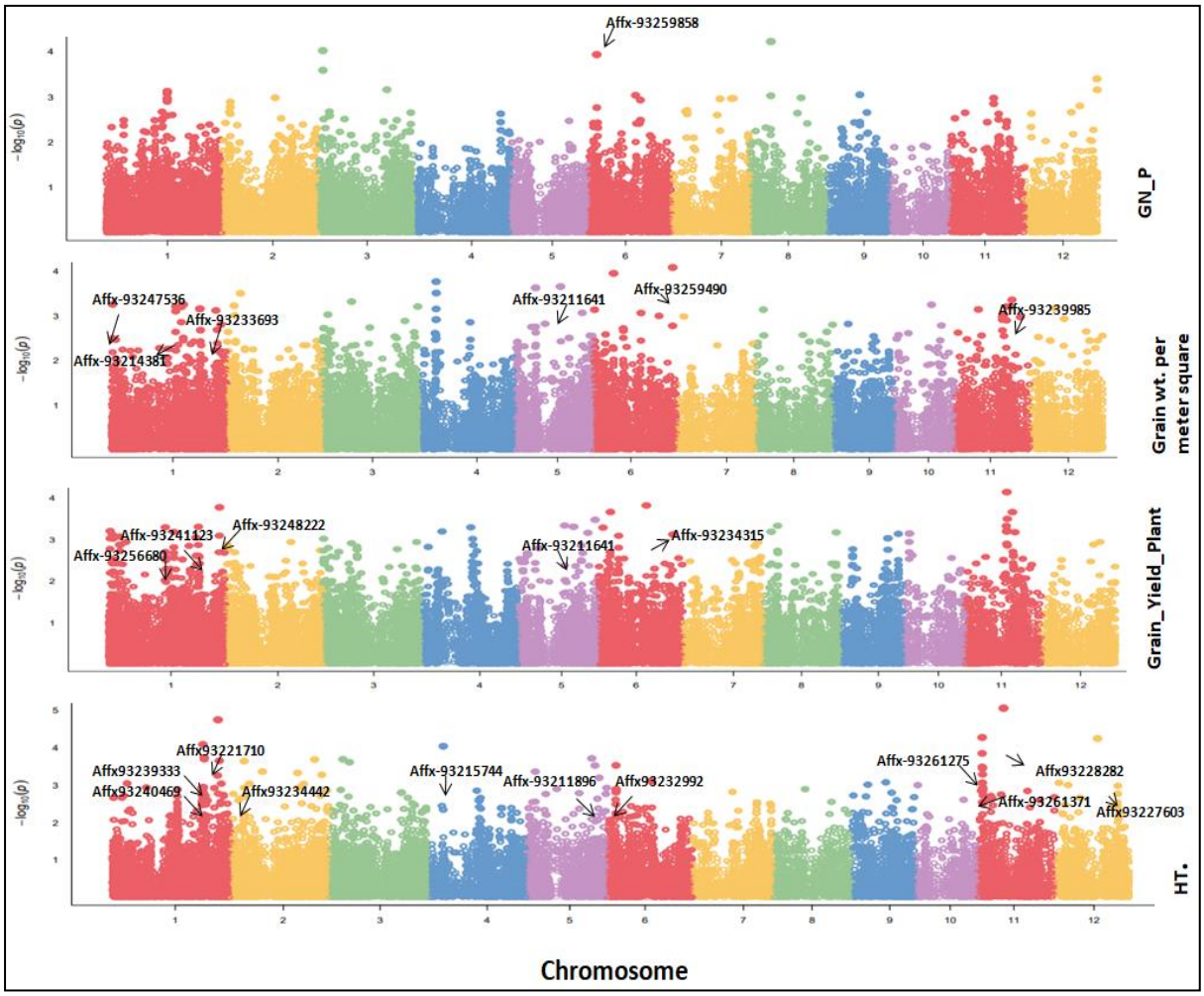


Fig.8: Manhattan plot for shows the common significant SNPs for four different agronomics traits by using the comparative analysis between TASSEL software and GAPIT software package through two univariate methods (GLM and MLM). The common significant SNPs were identified using *p*-value<.001 & 0.0001. Black arrows indicate the common significant SNPs. The X-axis shows the position of the snips on which chromosome and Y- axis shows -log p-value.

After getting the result by TASSEL software and GAPIT R package, we used filter of p-value >0.001 and >0.0001. After the comparative analysis between GAPIT and TASSEL software by using the MLM model. Under the *p*-value <0.001 we identified 2, 4, 6 and 12 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas under the *p*-value <0.0001, we identified 0, 0, 0 and 1 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. The details of SNPs and its chromosomes position shown in below (Table 1).

Table 5: Common significant SNPs by using MLM method (comparative analysis between GAPIT)

Traits	**Significant SNPs**	***p*-value**	**Total no. of significant SNPs**	**Chromosome position**
Grain yield plant	Affx-93211641 Affx-93234315 Affx-93241123 Affx-93248222 Affx-93248879 Affx-93256680	<0.001	4	Chr5_17096019 Chr6_17704662 Chr1_631690 Chr1_1672453 Chr4_17531730 Chr1_33332505
Grain wt. per meter Square	Affx-93247536 Affx-93233693 Affx-93214381 Affx-93211641 Affx-93259490 Affx-93239985	<0.001	6	Chr1_24239743 Chr1_27129151 Chr1_27130473 Chr5_17096019 Chr6_29615718 Chr11_21672937
Height	Affx-93221710 Affx-93239333 Affx-93240469 Affx-93234442 Affx-93232992 Affx-93261796 Affx-93261275 Affx-93261371 Affx-93228282 Affx-93227603 Affx-93211896 Affx-93215744	<0.001	12	Chr1_33320802 Chr1_33805233 Chr1_38865687 Chr2_30752216 Chr6_3221259 Chr11_1935453 Chr11_1936049 Chr11_1936531 Chr11_9781236 Chr12_15840841 Chr5_24021196 Chr4_4758589
Grain no. per plant	Affx-93259858 Affx-93232175	<0.001	2	Chr6_3057131 Chr8_7680994
Grain yield plant	No significant SNPs	<0.0001	0	_
Grain wt. per meter Square	No significant SNPs	<0.0001	0	_
Height	Affx-93221710 Affx-93228282 Affx-93215744	<0.0001	3	Chr1_33320802 Chr11_9781236 Chr4_4758589
Grain no. per plant	Affx-93232175	<0.0001	1	Chr11_7680994

After getting the result by TASSEL software and GAPIT R package, we used filter of *p*-value >0.001 and >0.0001. After the comparative analysis between GAPIT and TASSEL software by using the GLM model. Under the *p*-value <0.001 we identified 7, 102, 134 and 269 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas under the *p*-value <0.0001, we identified 0, 14, 11 and 75 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively.

Table 6: Common significant SNPs by using GLM method (comparative analysis between GAPIT &Tassel)

Traits	***p*- value**	**Total no. of SNPs**
Grain yield plant	<0.001	134
Grain wt. per meter square	<0.001	102
Height	<0.001	269
GN_P	<0.001	7
Grain yield plant	<0.0001	11
Grain wt. per meter square	<0.0001	14
Height	<0.0001	75
Grain no. per plant	<0.0001	No significant SNPs

A total number of SNPs 103 were identified after doing the comparative analysis between GLM and MLM models (TASSEL software). Under the *p*-value <0.001 we identified 3, 8, 17 and 68 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas under the *p*-value <0.0001, we identified 1, 0, 0 and 6 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. The detail description was given in below (Table 3) about the associated SNPs with particular traits.

Table 7: Common Significant SNPs of agronomics traits by using comparative analysis method (GLM & MLM in tassel)

Traits	**SNP id**	***p*- value**	**Total no. Of SNPs**	**Chromosome position**
Grain yield plant	Affx-93211641 Affx-93216822 Affx-93217649 Affx-93227543 Affx-93230744 Affx-93234315 Affx-93235285 Affx-93238376 Affx-93238743 Affx-93241123 Affx-93242457 Affx-93248222 Affx-93252181 Affx-93254542 Affx-93255342 Affx-93256680 Affx-93259084	<0.001	17	Chr5_17096019 Chr1_587230 Chr6_1635848 Chr1_41107928 Chr6_8309276 Chr6_17704662 Chr1_40939528 Chr3_10537869 Chr1_623173 Chr1 _631690 Chr8_2882889 Chr1_1672453 Chr1_41003616 Chr8_5526927 Chr1_41155466 Chr1_33332505 Chr3_290809
Grain wt. per meter square	Affx-93211641 Affx-93214381 Affx-93233693 Affx-93239985 Affx-93242457 Affx-93247536 Affx-93247708 Affx-93259490	<0.001	8	Chr5_17096019 Chr1_27130473 Chr1_27129151 Chr11_21672937 Chr8_2882889 Chr1_24239743 Chr4_17158963 Chr6_29615718
Height	Affx-93211379 Affx-93211713 Affx-93257821 Affx-93212357 Affx-93213301 Affx-93213819 Affx-93215521 Affx-93216212 Affx-93217819 Affx-93218894 Affx-93219343 Affx-93219464 Affx-93219940 Affx-93220999 Affx-93221710 Affx-93222027 Affx-93222177	<0.001	68	Chr_1 39816396 Chr_1 39698544 Chr6_3199146 Chr_1 39694313 Chr_6 3082967 Chr_6 3079615 Chr_6 3082458 Chr6_3214931 Chr_6 3253283 Chr_12 23163715 Chr6_3253397 Chr4_ 4444977 Chr6_ 3217070 Chr4_ 29752183 Chr1_ 33320802 Chr1_ 13783038 Chr2 _12418863

	Affx-93216212			Chr6_3214931
	Affx-93223431			Chr6_ 3083259
	Affx-93224590			Chr2_ 30747337
	Affx-93225562			Chr5_3654820
	Affx-93226246			Chr1_33742361
	Affx-93227603			Chr12_15840841
	Affx-93228186			Chr6_ 872823
	Affx-93228282			Chr11_ 9781236
	Affx-93228861			Chr6_ 3083998
	Affx-93229936			Chr1_39852428
	Affx-93230964			Chr11_ 5894739
	Affx-93231495			Chr6_3082622
	Affx-93231681			Chr6_3082401
	Affx-93232417			Chr6_ 3083782
	Affx-93232861			Chr6_ 3084901
	Affx-93232992			Chr6_ 3221259
	Affx-93234442			Chr2_ 30752216
	Affx-93237242			Chr7_ 23147795
	Affx-93237573			Chr6_ 3080670
	Affx-93238592			Chr6_ 3254415
	Affx-93239247			Chr1_ 34474363
	Affx-93239333			Chr1_33805233
	Affx-93239381			Chr8_25329650
	Affx-93240469			Chr1_38865687
	Affx-93242087			Chr4_21330713
	Affx-93245061			Chr1_39853599
	Affx-93245297			Chr6 _1612901
	Affx-93246424			Chr6_3259043
	Affx-93246702			Chr6_3082480
	Affx-93247629			Chr6_3216262
	Affx-93247702			Chr6_3255072
	Affx-93247751			Chr1 _23909305
	Affx-93248193			Chr4 _29750281
	Affx-93248783			Chr6_ 3079094
	Affx-93248815			Chr6_3217847
	Affx-93249336			Chr6_3083247
	Affx-93251039			Chr6_3146449
	Affx-93251425			Chr6_3079863
	Affx-93251469			Chr6_3082080
	Affx-93251994			Chr9_14540271
	Affx-93252105			Chr6_3257231
	Affx-93252804			Chr12_9361340
	Affx-93253250			Chr6_3084733
	Affx-93255286			Chr2_12465795
	Affx-93257859			Chr6_ 3217504
	Affx-93258500			Chr6_ 3314059

	Affx-93259067 Affx-93260636 Affx-93260941 Affx-93261159 Affx-93261350			Chr6 _ 3083366 Chr1_12574524 Chr6_3083881 Chr6_3079191 Chr5_3651352
Grain no. per plant	Affx-93221120 Affx-93232175 Affx-93259858	<0.001	3	Chr3_3164066 Chr8_7680994 Chr6_3057131
Grain yield plant	No significant SNPs	<0.0001	0	-
Grain wt. per meter square	No significant SNPs	<0.0001		-
Height	Affx-93219343 Affx-93221710 Affx-93239247 Affx-93234442 Affx-93242087 Affx-93228282	<0.0001	6	Chr6_3253397 Chr1_33320802 Chr1_34474363 Chr2_30752216 Chr4_21330713 Chr11_9781236
Grain no. per plant	Affx-93232175	<0.0001	1	Chr8_7680994

A total number of 98 SNPs were identified after doing the comparative analysis between GLM and MLM models (TASSEL software). Under the *p*-value <0.001 we identified 7, 29, 23 and 31 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas under the *p*-value <0.0001, we identified 1, 1, 1 and 5 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. The detail description was given in below (table 4) about the associated SNPs with particular traits.

Table 8: Common Significant SNPs of agronomics traits by using comparative analysis methods (GLM & MLM in GAPIT)

Traits	Significant SNPs	*p*-value	Total no. of Significant SNPs	Chromosome position
Grain yield plant	Affx-93211641	<0.001	29	Chr5_17096019
	Affx-93211996			Chr11_15572855
	Affx-93212388			Chr11_15777699
	Affx-93213333			Chr11_15570055
	Affx-93219830			Chr9_17311319
	Affx-93224741			Chr9_21197218
	Affx-932265			Chr6_4371259
	Affx-93227314			Chr11_15569746
	Affx-93234315			Chr6_17704662
	Affx-93236110			Chr11_15409636
	Affx-93239691			Chr1_24247953
	Affx-93241123			Chr1_631690
	Affx-93245490			Chr5_20353432
	Affx-93245777			Chr1_21111070
	Affx-93248082			Chr11_15640800
	Affx-93248222			Chr1_1672453
	Affx-93250221			Chr6_7401783
	Affx-93250332			Chr6_27428615
	Affx-93251143			Chr9_17310198
	Affx-93252350			Chr5_20353475
	Affx-93252682			Chr5_20354059
	Affx-93252849			Chr1_4141784
	Affx-93254004			Chr3_335220
	Affx-93254700			Chr1_4134887
	Affx-93256063			Chr9_17310389
	Affx-93256178			Chr9_17309981
	Affx-93256680			Chr1_33332505
	Affx-93259041			Chr4_18239785
	Affx-93259970			Chr8_27139226
Grain yield per meter square	Affx-93211641	<0.001	23	Chr5_17096019
	Affx-93214799			Chr11_18679859
	Affx-93217800			Chr1_33285434
	Affx-93217894			Chr3_11004153
	Affx-93219645			Chr3_35930607
	Affx-93221653			Chr1_33289000
	Affx-93224242			Chr1_33289833
	Affx-93225522			Chr1_33335369
	Affx-93232506			Chr2_2908628
	Affx-93233963			Chr2_2623254
	Affx-93235980			Chr12_9353998

	Affx-93239049			Chr5_25277519
	Affx-93239691			Chr1_24247953
	Affx-93239985			Chr11_21672937
	Affx-93241873			Chr1_33329230
	Affx-93244148			Chr6_203398
	Affx-93246152			Chr1_475294
	Affx-93247536			Chr1_24239743
	Affx-93248205			Chr10_14261270
	Affx-93248244			Chr5_7748187
	Affx-93250183			Chr6_24494532
	Affx-93250221			Chr6_7401783
	Affx-93259490			Chr6_29615718
Height	Affx-93211896	<0.001	31	Chr5_24021196
	Affx-93214447			Chr11_2065080
	Affx-93215744			Chr4_4758589
	Affx-93221710			Chr1_33320802
	Affx-93223785			Chr2_11595159
	Affx-93223893			Chr10_1177199
	Affx-93227480			Chr1_38710379
	Affx-93227603			Chr12_15840841
	Affx-93227665			Chr6_15867674
	Affx-93228282			Chr11_9781236
	Affx-93232992			Chr6_3221259
	Affx-93234442			Chr2_30752216
	Affx-93239333			Chr1_33805233
	Affx-93240469			Chr1_38865687
	Affx-93243068			Chr5_3305808
	Affx-93243861			Chr11_2063736
	Affx-93244800			Chr1_33800408
	Affx-93249760			Chr11_2106739
	Affx-93250190			Chr1_5508658
	Affx-93251836			Chr2_26699583
	Affx-93255782			Chr1_39358978
	Affx-93256813			Chr3_5256083
	Affx-93256846			Chr12_1703264
	Affx-93259809			Chr2_33360346
	Affx-93260832			Chr3_7676514
	Affx-93261218			Chr11_2060290
	Affx-93261275			Chr11_1936049
	Affx-93261371			Chr11_1936531
	Affx-93261437			Chr11_1860463
	Affx-93261673			Chr11_1931241
	Affx-93261796			Chr11_1935453
Grain no. per plant	Affx-93215477	<0.001	7	Chr12_27316019

	Affx-93215760 Affx-93259858 Affx-93218948 Affx-93224089 Affx-93242996 Affx-93256329			Chr3_2240618 Chr6_3057131 Chr3_2199340 Chr9_12221443 Chr3_26332851 Chr6_17675280
Grain yield plant	Affx-93213333	<0.0001	1	Chr11_15570055
Grain yield per meter square	Affx-93259490	<0.0001	1	Chr6_29615718
Height	Affx-93240469 Affx-93215744 Affx-93261371 Affx-93228282 Affx-93227603	<0.0001	5	Chr1_38865687 Chr4_4758589 Chr11_1936531 Chr11_9781236 Chr12_15840841
Grain no. per plant	Affx-93213333	<0.0001	1	Chr11_15570055

After doing the comparative analysis between both GLM AND MLM models and TASSEL software and GAPIT package (table 1, 2, 3, and 4), we identified 23 common SNPs those are associated with different traits. Under the *p*-value <0.001 we identified 1, 6, 5 and 11 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas under the *p*-value <0.0001, we identified 0, 1, 8 and 1 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. The details of SNPs and its chromosomes position shown in below (Table 9).

Table 9: Common significant SNPs of agronomic traits by using Tassel and GAPIT R package**.**

Traits	**Significant SNPs**	***p*-value**	**Total no. of significant SNPs**	**Chromosome position**
Grain yield plant	Affx-9321164 Affx-93234315 Affx-93241123 Affx-93248222 Affx-93256680	<0.001	5	Chr5_17096019 Chr6_17704662 Chr1_631690 Chr1_1672453 Chr1_ 3333250
Grain wt. per meter square	Affx-93211641 Affx-93239985 Affx-93247536 Affx-93259490 Affx-93233693 Affx-93214381	<0.001	6	Chr5_17096019 Chr11_21672937 Chr1_24239743 Chr6_29615718 Chr1_27129151 Chr1_27130473
Height	Affx93221710 Affx93227603 Affx93228282 Affx93232992 Affx93234442 Affx93239333 Affx93240469 Affx-93211896 Affx-93215744 Affx-93261275 Affx-93261371	<0.001`	11	Chr1_33320802 Chr12_15840841 Chr11_9781236 Chr6_3221259 Chr2_30752216 Chr1_33805233 Chr1_38865687 Chr5_24021196 Chr4_4758589 Chr11_1936049 Chr11_1936531
Grain no. per plant	Affx-93259858	<0.001	1	Chr6_3057131
Grain yield plant	Affx-93211641 Affx-93211996 Affx-93224741 Affx-93236110 Affx-93245490 Affx-93252350 Affx-93252682 Affx-93256680	<0.0001	8	Chr9_21197218 Chr11_15572855 Chr5_17096019 Chr11_15409636 Chr5_20353432 Chr5_20353475 Chr5_20354059 Chr1_33332505
Grain wt. per meter square	Affx-93259490	<0.0001	1	Chr6_29615718
Height	Affx-93228282	<0.0001	1	chr11_978123

Grain no. per plant	No Significant SNPs	<0..0001	0	……

4.6 Identification of candidate gene and its annotation related with different traits:

The search for candidate genes by using Rice Annotation Project (RAP) database genome browser. According to the results of GWAS, a total of 23 genes were identified, 5 potential genes were discovered to be connected to grain yield plant, 6 genes were found to be related with grain wt. per meter square, 11 genes were found to be related with height, and 1 gene were found to be related with Grain number per plant. These significant candidate genes encoded SNF2 family N-terminal domain containing protein, expressed, Cation ellux family protein, putative, expressed, RING-H2 finger protein, putative, expressed, OsCML6 - Calmodulin-related calcium sensor protein, expressed, heat shock protein DnaJ, putative, expressed, transcription factor, putative, expressed, insulin-degrading enzyme, putative, expressed, PPR repeat domain containing protein, putative, expressed, roothairless 1, putative, expressed, hydrolase, alpha/beta fold family domain containing protein, expressed, helicase domain-containing protein, putative, expressed, Homeobox domain containing protein, expressed, DUF647 domain containing protein, putative, expressed, avr9/Cf-9 rapidly elicited protein 137, putative, expressed, GRAS family transcription factor containing protein, expressed, GDSL-like lipase/acylhydrolase, putative, expressed, RING-H2 finger protein, putative, expressed, tetratricopeptide repeat domain containing protein, expressed, and OsDegp7 - Putative Deg protease homologue, expressed.

Table 10: list of candidate genes and its annotation for four agronomics traits.

Traits	SNP ID	*p*-value	Total SNPs	Chromosome position	Gene ID	Annotation
Grain yield plant	Affx-9321164	<0.001	5	chr5_17096019	LOC_Os05g27340.1	expressed protein
	Affx-93234315		6	chr6_17704662	LOC_Os06g30600.1	expressed protein
	Affx-93241123		1	chr1_631690	LOC_Os01g02170.1	expressed protein
	Affx-93248222		1	chr1_1672453	LOC_Os01g03914.1	cation efflux family protein, putative, expressed
	Affx-93256680		1	chr1_ 3333250	LOC_Os01g57110.1	SNF2 family N-terminal domain containing protein,
Grain wt. per meter square	Affx-93211641	<0.001	5	chr5_17096019	LOC_Os05g29710.1	RING-H2 finger protein, putative, expressed
	Affx-93239985		11	chr11_21672937	LOC_Os11g37550.1	OsCML6 - Calmodulin-related calcium sensor protein, expressed
	Affx-93247536		1	chr1_24239743	LOC_Os01g42190.1	heat shock protein DnaJ, putative, expressed
	Affx-93259490		6	chr6_29615718	LOC_Os06g48920.1	Expressed protein
	Affx-93233693		1	Chr1_27129151	LOC_Os01g46970.1	transcription factor, putative, expressed
	Affx-93214381		1	Chr1_27130473	LOC_Os01g46970.1	transcription factor, putative, expressed
Height	Affx93221710	<0.001	1	chr1_33320802	LOC_Os01g57082.1	insulin-degrading enzyme, putative, expressed
	Affx93227603		11	chr12_15840841	LOC_Os12g27060.1	PPR repeat domain containing protein, putative, expressed
	Affx93228282		12	chr11_9781236	LOC_Os11g17600.3	roothairless 1, putative, expressed
	Affx93232992		6	chr6_3221259	LOC_Os06g06820.1	hydrolase, alpha/beta fold family domain containing , protein, expressed
	Affx93234442		2	chr2_30752216	LOC_Os02g50370.1	helicase domain-containing protein, putative, expressed
	Affx93239333		1	chr1_33805233	LOC_Os01g57890.1	Homeobox domain containing protein, expressed
	Affx93240469		1	chr1_38865687	LOC_Os01g66350.1	DUF647 domain containing protein, putative, expressed
	Affx-93211896		5	Chr5_24021196	LOC_Os05g41150.1	expressed protein

	Affx-93215744		4	Chr4_4758589	LOC_Os04g08764.1	avr9/Cf-9 rapidly elicited protein 137, putative, expressed
	Affx-93261275		11	Chr11_1936049	LOC_Os11g04570.1	GRAS family transcription factor containing protein, expressed
	Affx-93261371		11	Chr11_1936531	LOC_Os11g04570.1	GRAS family transcription factor containing protein, expressed
Grain no. per plant	Affx-93259858	<0.001	6	chr6_3057131	LOC_Os06g06520.1	GDSL-like lipase/acylhydrolase, putative, expressed

CHAPTER 5 - DISCUSSION

5.1 Population structure analysis, phylogenetic analysis and linkage disequilibrium:

Basically, population structure analysis shows a having a diverse population is made possible by population structure. Population structure shows a sub-population between the populations. In our study, to comprehend the population dynamics of the rice diversity panel, we analysed population structure by using STRUCTURE software. The best K value at 4. That's means in rice diverse collections have four populations present.

Phylogenetic analysis shows the evolutionary relationships between different population and sub population. Basically, they show the closely related relationship between different groups. In our study, 2 clusters are present in the phylogenetic tree, those I have mentioned in the previous page (Fig.No.4).

LD is important because of the many variables that affect and are affected by it. The potential responds to both artificial and natural selection is restricted by LD, which also provides historical event information. The TASSEL generated the triangle plot of linkage disequilibrium. The comparison of two sets of marker sites is shown in each cell, with colour coding used to indicate the presence of significant LD. It is customary to visually separate pairings of loci with high levels of LD from those with low levels of LD when considering more than two loci together.

5.2 Comparative analysis by using different method and software:

We selected to conduct GWAS on 186 rice diverse collections landrace populations in this work because the greater size of the sample along with increased genetic diversity enough effectiveness of the association analysis. While numerous loci were all mapped of the two univariate tested GWAS methods by TASSEL software and GAPIT R package, both of the two techniques located a few of the known function loci. We did a comparative analysis between GLM and MLM models with the result of TASSEL software and GAPIT R package. And after this, those common SNPs were identified after the comparative analysis, which is consider as a significant SNPS .In our study, 23 genes were identified by using the comparative analysis those were associated with particular traits. All the result of the comparative study was shows in previous pages (table 5, 6, 7, 8 and 9).

5.3 Application of the potential genes in breeding and future research:

Twenty- three QTLs or genes with crucial agronomic trait agronomic trait control were discovered in our study. We identified we identified 1, 6, 5 and 11 common significant SNPs

for GN_P, grain weight per meter square, grain yield plant and ht respectively. The candidate genes involved in important agronomics traits are crucial tools for comprehending the mechanisms that underlie the optimum yield and agricultural breeding.

However, it is crucial to remember that although though all of these potential genes were found based on the expression, sequence and homology analyses of QTLs, more research must be done to confirm the findings before drawing any definite conclusions.

CONCLUSION

In this study, we did a comparative analysis between TASSEL software and GAPIT R package by using the GLM and MLM models .BY USING THE tassel software, Significant SNPs were identified using *p*-value <.001 & 0.0001. Under the *p*-value <0.001, GLM model identified 14, 798, 225 and 1654 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 5, 12, 33, and 77 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively. Under the *p*-value <0.0001, GLM model identified 2, 26, 44, and 187 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 1, 7, 1, and 0 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively.

By using the GAPIT R package, Under the *p*-value <0.001, GLM model identified 129, 1873, 1253 and 2756 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 7, 29, 31 and 38 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively. Under the *p*-value <0.0001, GLM model identified 14, 798, 225 and 1654 significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas MLM model identified 2, 1, 1 and 6 significant SNPs for GN_P, grain wt. per meter square, grain yield plant and ht respectively.

The common significant SNPs after using the comparative analysis, WAS analysis for four agronomic traits based on 50051 SNPs was performed using two univariate methods (GLM and MLM). Under the *p*-value < 0.001, we identified 1, 6, 5 and 11 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively. Whereas under the *p*-value <0.0001, we identified 0, 1, 8 and 1 common significant SNPs for GN_P, grain weight per meter square, grain yield plant and ht respectively.

The search for candidate genes by using Rice Annotation Project (RAP) database genome browser. According to the results of GWAS, a total of 23 genes were identified, 5 potential genes were discovered to be associated with grain yield plant, 6 genes were found to be related with grain wt. per meter square, 11 genes were found to be related with height, and 1 gene were found to be related with Grain number per plant. The common significant markers which might be further used for genetic studies. These results will serve as a guide for high-yielding rice variety breeding in the near future.

FUTURE PROSPECTIVE

A potent tool for analysing complex phenotypes is the use of genome-wide association studies, which are genetic explorations of the entire genome to identify variants associated with a trait in wild populations. The common significant markers which might be further used for genetic studies. These results will serve as a guide for the immediate future breeding of high-yielding rice cultivars.

REFERENCES

[1] P. M. Visscher *et al.*, "10 years of GWAS discovery: Biology, function, and translation," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, 2017.

[2] D. J. Benjamin *et al.*, "The promises and pitfalls of genoeconomics," *Annu. Rev. Econom.*, vol. 4, no. 1, pp. 627–662, 2012.

[3] S. Takeda and M. Matsuoka, "Genetic approaches to crop improvement: responding to environmental and population changes," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 444–457, 2008.

[4] Q. Wang, J. Tang, B. Han, and X. Huang, "Advances in genome-wide association studies of complex traits in rice," *Züchter Genet. Breed. Res.*, vol. 133, no. 5, pp. 1415–1425, 2020.

[5] M. Yano, "Genetic and molecular dissection of quantitative traits and its application in rice breeding," *Ikushugaku Kenkyu*, vol. 9, no. 4, pp. 135–140, 2007.

[6] J. J. Lee *et al.*, "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals," *Nat. Genet.*, vol. 50, no. 8, pp. 1112–1121, 2018.

[7] P. R. Jansen *et al.*, "Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways," *Nat. Genet.*, vol. 51, no. 3, pp. 394–403, 2019.

[8] X. Huang *et al.*, "Genome-wide association studies of 14 agronomic traits in rice landraces," *Nat. Genet.*, vol. 42, no. 11, pp. 961–967, 2010.

[9] H. Begum *et al.*, "Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa)," *PLoS One*, vol. 10, no. 3, p. e0119873, 2015.

[10] X. Huang *et al.*, "Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm," *Nat. Genet.*, vol. 44, no. 1, pp. 32–39, 2012.

[11] G. I. Descalsota-Empleo *et al.*, "Genetic mapping of QTL for agronomic traits and grain mineral elements in rice," *Crop J.*, vol. 7, no. 4, pp. 560–572, 2019.

[12] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for

genome-wide association studies," *Bioinformatics*, vol. 26, no. 4, pp. 445–455, 2010.

[13] Z. Tian *et al.*, "Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 51, pp. 21760–21765, 2009.

[14] M. Ashikari and M. Matsuoka, "Identification, isolation and pyramiding of quantitative trait loci for rice breeding," *Trends Plant Sci.*, vol. 11, no. 7, pp. 344–350, 2006.

[15] M. Surapaneni, D. Balakrishnan, S. Mesapogu, A. Krishnam Raju, Y. V. Rao, and S. Neelamraju, "Genetic characterization and population structure of Indian rice cultivars and wild genotypes using core set markers," *3 Biotech*, vol. 6, no. 1, 2016.

[16] S. C. A. Akouelamouai, C. D. Yila Moutelet, and P. G. Ondongo, "The impact of improved seed subsidies on cereal yields in Burkina Faso: The case of rice," *International Journal of Business Management and Economic Review*, vol. 04, no. 06, pp. 351–364, 2022.

[17] B. Das *et al.*, "Genetic diversity and population structure of rice landraces from Eastern and North Eastern States of India," *BMC Genet.*, vol. 14, no. 1, 2013.

[18] S. D. Koutroubas, F. Mazzini, B. Pons, and D. A. Ntanos, "Grain quality variation and relationships with morpho-physiological traits in rice (Oryza sativa L.) genetic resources in Europe," *Field Crops Res.*, vol. 86, no. 2–3, pp. 115–130, 2004.

[19] R. Kumar and P. S. Rathiya, "Effects of nutrient management on growth attributes and yield of high yielding rice (Oryza sativa L.) varieties of Chhattisgarh," *Int. J. Chem. Stud.*, vol. 8, no. 1, pp. 1020–1024, 2020.

[20] W. Tadesse *et al.*, "Genome-wide association mapping of yield and grain quality traits in winter wheat genotypes," *PLoS One*, vol. 10, no. 10, p. e0141339, 2015.

[21] T. Izawa and K. Shimamoto, "Becoming a model plant: The importance of rice to plant science," *Trends Plant Sci.*, vol. 1, no. 3, pp. 95–99, 1996.

[22] D. Zhang *et al.*, "Genetic structure and differentiation of Oryza sativa L. in China revealed by microsatellites," *Züchter Genet. Breed. Res.*, vol. 119, no. 6, pp. 1105–1117, 2009.

[23] P. Vikram *et al.*, "Drought susceptibility of modern rice varieties: an effect of linkage of drought tolerance with undesirable traits," *Sci. Rep.*, vol. 5, no. 1, 2015.

[24] A. Kumar, S. Dixit, T. Ram, R. B. Yadaw, K. K. Mishra, and N. P. Mandal, "Breeding high-yielding drought-tolerant rice: genetic variations and conventional and molecular

approaches," *J. Exp. Bot.*, vol. 65, no. 21, pp. 6265–6278, 2014.

[25] E. M. Brown and B. J. Barratt, "The HapMap– A haplotype map of the human genome," in *Bioinformatics for Geneticists*, Chichester, UK: John Wiley & Sons, Ltd, 2007, pp. 33–58.

[26] S. Atwell *et al.*, "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines," *Nature*, vol. 465, no. 7298, pp. 627–631, 2010.

[27] T. Sasaki and B. Burr, "International Rice Genome Sequencing Project: the effort to completely sequence the rice genome," *Curr. Opin. Plant Biol.*, vol. 3, no. 2, pp. 138–141, 2000.

[28] R. J. Henry, "Genomics strategies for germplasm characterization and the development of climate resilient crops," *Front. Plant Sci.*, vol. 5, p. 68, 2014.

[29] E. W. Petersdorf and C. O'hUigin, "The MHC in the era of next-generation sequencing: Implications for bridging structure with function," *Hum. Immunol.*, vol. 80, no. 1, pp. 67–78, 2019.

[30] The Wellcome Trust Case Control Consortium *et al.*, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[31] M. Nordborg and D. Weigel, "Next-generation genetics in plants," *Nature*, vol. 456, no. 7223, pp. 720–723, 2008.

[32] K. Zhao *et al.*, "Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa," *Nat. Commun.*, vol. 2, no. 1, 2011.

[33] D. R. Wang *et al.*, "An imputation platform to enhance integration of rice genetic resources," *Nat. Commun.*, vol. 9, no. 1, 2018.

[34] B. L. Browning and S. R. Browning, "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals," *Am. J. Hum. Genet.*, vol. 84, no. 2, pp. 210–223, 2009.

[35] D. Reynolds, J. Ball, A. Bauer, R. Davey, S. Griffiths, and J. Zhou, "CropSight: a scalable and open-source information management system for distributed plant phenotyping and IoT-based crop management," *Gigascience*, vol. 8, no. 3, 2019.

[36] W. Yang *et al.*, "Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice," *Nat. Commun.*, vol. 5, no. 1, 2014.

[37] Z. Guo *et al.*, “Genome-wide association studies of image traits reveal genetic architecture of drought resistance in rice,” *Mol. Plant*, vol. 11, no. 6, pp. 789–805, 2018.

[38] W. Chen *et al.*, “Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism,” *Nat. Genet.*, vol. 46, no. 7, pp. 714–721, 2014.

[39] K. A. G. Kremling *et al.*, “Dysregulation of expression correlates with rare-allele burden and fitness loss in maize,” *Nature*, vol. 555, no. 7697, pp. 520–523, 2018.

[40] T. Kawakatsu *et al.*, “Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions,” *Cell*, vol. 166, no. 2, pp. 492–505, 2016.

[41] P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler, “TASSEL: software for association mapping of complex traits in diverse samples,” *Bioinformatics*, vol. 23, no. 19, pp. 2633–2635, 2007.

[42] J. M. Thornsberry, M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. S. Buckler IV, “Dwarf8 polymorphisms associate with variation in flowering time,” *Nat. Genet.*, vol. 28, no. 3, pp. 286–289, 2001.

[43] J. Yu *et al.*, “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness,” *Nat. Genet.*, vol. 38, no. 2, pp. 203–208, 2006.

[44] A. E. Lipka *et al.*, “GAPIT: genome association and prediction integrated tool,” *Bioinformatics*, vol. 28, no. 18, pp. 2397–2399, 2012.

[45] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. San Diego, CA: Academic Press, 1979.

[46] J. K. Pritchard, M. Stephens, and P. Donnelly, “Inference of population structure using multilocus genotype data,” *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[47] K. Zhao *et al.*, “An Arabidopsis example of association mapping in structured samples,” *PLoS Genet.*, vol. 3, no. 1, p. e4, 2007.

[48] X. Liu, M. Huang, B. Fan, E. S. Buckler, and Z. Zhang, “Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies,” *PLoS Genet.*, vol. 12, no. 2, p. e1005767, 2016.

[49] P. Bellot, G. de los Campos, and M. Pérez-Enciso, “Can deep learning improve genomic prediction of complex human traits?,” *Genetics*, vol. 210, no. 3, pp. 809–819, 2018.

[50] "Preface to second edition," in *Infectious Diseases*, Elsevier, 1967, p. ix.

[51] E. Mancin, D. Lourenco, M. Bermann, R. Mantovani, and I. Misztal, "Accounting for population structure and phenotypes from relatives in association mapping for farm animals: A simulation study," *Front. Genet.*, vol. 12, 2021.

[52] The Pan African Journal, "Retraction: Analysis of the effect of health insurance on health care utilization in Rwanda: a secondary data analysis of Rwanda integrated living condition survey 2016-2017 (EICV 5) (PAMJ - One Health. 2021;4:10. Doi: 10.11604/pamj-oh.2021.4.10.25256)," *PAMJ One Health*, vol. 5, 2021.

[53] M. Pirinen, P. Donnelly, and C. C. A. Spencer, "Including known covariates can reduce power to detect genetic effects in case-control studies," *Nat. Genet.*, vol. 44, no. 8, pp. 848–851, 2012.

[54] V. Moskvina, P. Holmans, K. M. Schmidt, and N. Craddock, "Design of case-controls studies with unscreened controls," *Ann. Hum. Genet.*, vol. 69, no. 5, pp. 566–576, 2005.

[55] A. Fry *et al.*, "Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population," *Am. J. Epidemiol.*, vol. 186, no. 9, pp. 1026–1034, 2017.

[56] N. Pirastu *et al.*, "Genetic analyses identify widespread sex-differential participation bias," *Nat. Genet.*, vol. 53, no. 5, pp. 663–671, 2021.

[57] B. Brumpton *et al.*, "Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses," *Nat. Commun.*, vol. 11, no. 1, 2020.

[58] G. R. Abecasis, L. R. Cardon, and W. O. C. Cookson, "A general test of association for quantitative traits in nuclear families," *Am. J. Hum. Genet.*, vol. 66, no. 1, pp. 279–292, 2000.

[59] T. C. Bates *et al.*, "The nature of nurture: Using a virtual-parent design to test parenting effects on children's educational attainment in genotyped families," *Twin Res. Hum. Genet.*, vol. 21, no. 2, pp. 73–83, 2018.

[60] Y. Xue *et al.*, "Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations," *Nat. Commun.*, vol. 8, no. 1, 2017.

[61] A. F. Herzig, T. Nutile, M.-C. Babron, M. Ciullo, C. Bellenguez, and A.-L.

Leutenegger, “Strategies for phasing and imputation in a population isolate,” *Genet. Epidemiol.*, vol. 42, no. 2, pp. 201–213, 2018.

[62] E. Zeggini, A. L. Gloyn, and T. Hansen, “Insights into metabolic disease from studying genetics in isolated populations: stories from Greece to Greenland,” *Diabetologia*, vol. 59, no. 5, pp. 938–941, 2016.

[63] R. Do *et al.*, “Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction,” *Nature*, vol. 518, no. 7537, pp. 102–106, 2015.

[64] M. M. Iles, “Genome-wide association studies,” in *Methods in Molecular Biology*, Totowa, NJ: Humana Press, 2011, pp. 89–103.

[65] R. C. Lewontin and K.-I. Kojima, “The evolutionary dynamics of complex polymorphisms,” *Evolution*, vol. 14, no. 4, p. 458, 1960.

[66] M. Slatkin, “Linkage disequilibrium — understanding the evolutionary past and mapping the medical future,” *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–485, 2008.

[67] S.-W. Guo, “Linkage disequilibrium measures for fine-scale mapping: A comparison,” *Hum. Hered.*, vol. 47, no. 6, pp. 301–314, 1997.

[68] D. Fallin and N. J. Schork, “Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data,” *Am. J. Hum. Genet.*, vol. 67, no. 4, pp. 947–959, 2000.

[69] M. Li, C. Li, and W. Guan, “Evaluation of coverage variation of SNP chips for genome-wide association studies,” *Eur. J. Hum. Genet.*, vol. 16, no. 5, pp. 635–643, 2008.

[70] S. Sanna *et al.*, “Common variants in the GDF5-UQCC region are associated with variation in human height,” *Nat. Genet.*, vol. 40, no. 2, pp. 198–203, 2008.

[71] J. P. T. Higgins, “Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified,” *Int. J. Epidemiol.*, vol. 37, no. 5, pp. 1158–1160, 2008.

[72] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, and M. V. Lareu, “An overview of STRUCTURE: applications, parameter settings, and supporting software,” *Front. Genet.*, vol. 4, 2013.