

International Journal of Advanced Biochemistry Research



ISSN Print: 2617-4693
ISSN Online: 2617-4707
IJABR 2024; SP-8(10): 1572-1580
www.biochemjournal.com
Received: 24-09-2024
Accepted: 25-10-2024

Shinde CS

Ph.D. Scholar, Department of
Genetics and Plant Breeding,
MPKV, Rahuri, Maharashtra,
India

Andhale GR

Ph.D. Scholar, Department of
Genetics and Plant Breeding,
MPKV, Rahuri, Maharashtra,
India

Thange VB

Ph.D. Scholar, Department of
Genetics and Plant Breeding,
MPKV, Rahuri, Maharashtra,
India

Bhosale BR

Ph.D. Scholar, Department of
Genetics and Plant Breeding,
MPKV, Rahuri, Maharashtra,
India

Decoding basic GWAS: The beginner's guide

Shinde CS, Andhale GR, Thange VB and Bhosale BR

DOI: <https://doi.org/10.33545/26174693.2024.v8.i10Ss.2783>

Abstract

A genome-wide association study (GWAS) is a research approach used to identify genomic variants that are statistically associated with a particular trait or trait of interest. GWAS study finds associated genes with traits/diseases throughout the genome using SNP's (Single Nucleotide Polymorphism) to determine the location of these genes. SNP can determine the genetic diversity between closely related individuals. GWAS is an efficient approach to study the mechanism behind the complex trait *viz.*, metabolic, or biochemical pathway. The various statistical tools are necessary for computation of association between various genotypes. GWAS mainly uses Minor Allele Frequency (MAF), Linkage Disequilibrium (LD) and P-value to analyze SNPs. On the determination of this value Manhattan Plot was drawn and then association between phenotype and concern genotype was determined. The technique is used for advancement of crops in terms of biotic resistance, abiotic tolerance, improvement in yield and metabolic composition of various crops *viz.*, Maize, Rice, Wheat, Cotton, etc. Any other mapping not as robust as GWAS in terms of discovering the candidate gene, unknown biochemical pathway, finding specific genetic association, etc. Genome-wide association studies (GWAS) have been even more successful in plants than in humans. Mapping approaches can be extended to dissect adaptive genetic variation from structured background variation in an ecological context. These methods, when combined with the increasing capacity and decreasing costs of next-generation sequencing, will allow GWAS in non-model species. As GWAS utilizes a large population, they get a chance to screen a large number of SNPs there by helping to get results more proper.

Keywords: GWAS, association mapping, SNPs, linkage disequilibrium, population structure

Introduction

A genome-wide association study (GWAS) is a research approach used to identify genomic variants that are statistically associated with a particular trait or trait of interest. This method studies the entire set of genomes within a large set of natural populations. If we have the whole genome sequence information of any individual then it becomes easier to identify the function of each gene. So, the next big advancement in the field of genetics after the Human Genome Project (2003) was Genome-Wide Association Studies. As for many studies in the last 40 years the initial applications of GWAS in plants were in Arabidopsis to find the genomic diversity (Borevitz *et al.*, 2007) [2]. Further in cultivated crop Rice was the first to conduct GWAS.

This study finds associated genes with traits/diseases throughout the genome using SNP's (Single Nucleotide Polymorphism) to determine the location of these genes. The GWAS studies required many advancements and assistance. Initially in the beginning era of GWAS it required up to \$250 million of costing in early 5 years. The first application of GWAS was employed in human beings (Klein *et al.*, 2005) [9] and thereafter in plants (Benjamin *et al.*, 2007; Korte and Farlow, 2013) [1, 10]. The genetic sources of phenotypic variation have been a major focus of both plant and animal studies aimed at identifying the causes of disease, improving agriculture and understanding adaptive processes. (Benjamin *et al.*, 2007) [1]. In plants, quantitative trait loci (QTL) were originally mapped in bi-parental crosses, but they were restricted in allelic diversity and in having limited genomic resolution. In such cases the approach GWAS plays a key role to determine allelic diversity. Primary goal of GWAS is to uncover genetic factors that contribute to the risk of developing particular disease or the variation of certain traits. It significantly helps to identify genetic risks of various complex diseases and traits. Also, it gives clues about biological pathways. With the time technique has been advanced with meta- analyses, fine mapping, functional genomics, integration with omics sciences and with machine learning and artificial intelligence (AI) techniques.

Corresponding Author:**Shinde CS**

Ph.D. Scholar, Department of
Genetics and Plant Breeding,
MPKV, Rahuri, Maharashtra,
India

Need

The need of GWAS is varies according to crops and breeding objective but some generalize objectives are depicted here:

- To use diverse genetic resources is important for breeding crop varieties (Glaszman *et al.*, 2010)^[8].
- To exclude crop species with narrow genetic diversity which are susceptible to emerging pathogens or other constraints leading to loss of productivity and this may lead to a serious decline in the areas of adaptation (Dyer *et al.*, 2014)^[6].
- To find genes contributing to variation in phenotypes of interest.
- Non availability of tightly linked flanking markers to the gene of interest. The NGS (Next Generation Sequencing) facilitates development and use of markers like SNP which overcomes this difficulty.

Features

Single nucleotide polymorphism (SNP) markers are desirable for estimation of genetic diversity because of their abundance in the genome.

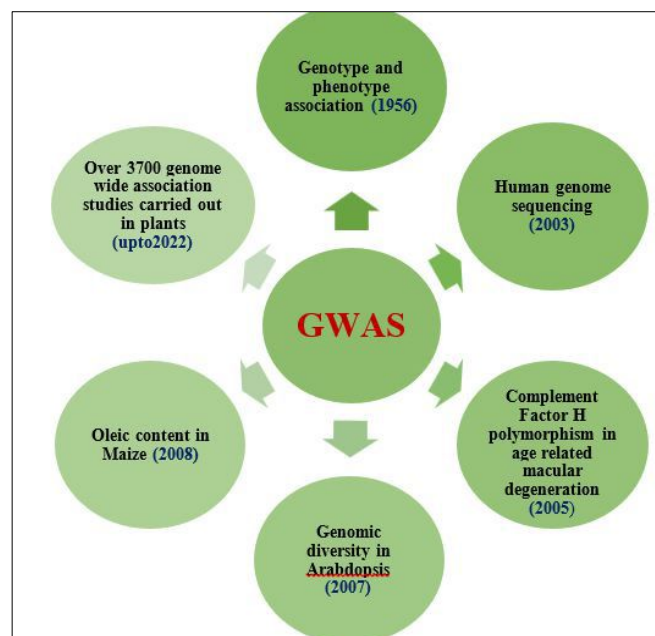
SNPs have the ability to discriminate between closely related individuals at a higher resolution. SNP markers have been developed and used to study genetic diversity and genetic mapping.

Genome-Wide Association Study (GWAS) is an efficient approach to dissect the genetic basis of complex traits using the naturally occurring genetic diversity (Korte and Farlow, 2013)^[10].

GWAS provides higher mapping resolution than classical bi-parental populations to detect associations (Liu *et al.*, 2016; Cui *et al.*, 2017; Xu *et al.*, 2017)^[13, 3].

GWAS requires an assessment of the population structure to determine the genetic relatedness (Kinship) of individuals and minimize detection of false associations (Korte and Farlow, 2013; Sul *et al.*, 2016)^[10].

History



Concepts

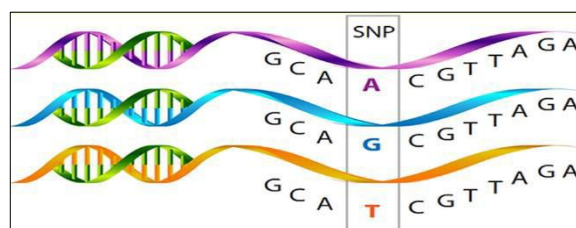
Genetic Variants: Differences in DNA sequence among individuals. SNPs are the most common type of genetic variation.

Genotyping: The process of determining which genetic variants an individual possesses.

Association: Statistical relationship between a specific genetic variant and a trait.

Minor allele frequency (MAF): The frequency of the less common allele of polymorphic locus. Its value lies between 0 to 0.5, and can vary between the population.

Single nucleotide polymorphisms (SNPs) are variations at a single position in a DNA sequence among individuals. Which can identify the single nucleotide at a time and four genetic variants as four different bases. GWAS can identify SNPs that are associated with particular traits or diseases. By examining the correlation between these SNPs and the trait of interest across many individuals, researchers can identify regions of the genome that are likely to contain relevant genes.



³
Source: SciGenom Conferences; NGS Bioinformatic Workshop 2014

Steps to detect SNPs

- **Sample Collection:** Obtain tissue, blood, saliva or other biological materials from individuals in the population. Ensure samples are properly preserved to prevent degradation, typically using freezing or specific preservation solutions.
- **DNA Extraction:** Use extraction kits or protocols to isolate DNA from the collected samples. Assess the quality and quantity of extracted DNA using spectrophotometer or gel electrophoresis.
- **SNP Genotyping:** Choose relevant SNPs for the study, which may involve known markers or genome-wide association studies (GWAS). Use platforms like microarrays, next-generation sequencing (NGS), or specific PCR-based methods to genotype the selected SNPs. Perform quality control to ensure accuracy, such as checking for genotyping errors or low-quality reads.
- **Data Analysis:** Clean the data by filtering out low-quality SNPs and samples. Use tools like STRUCTURE, PCA, or ADMIXTURE to understand the population structure. Further conduct GWAS or other statistical analyses to associate SNPs with traits or diseases. Then go for haplotype analysis by studying the combination of SNPs within a region to understand linkage disequilibrium and evolutionary history.
- **Interpretation:** Investigate the biological significance of associated SNPs, possibly through databases like Ensembl or UCSC Genome Browser. Use tools to analyze whether associated SNPs are enriched in specific biological pathways. Calculate metrics like allele frequency and heterozygosity to understand genetic diversity and differentiation.

- **Reporting:** Visualize result by creating graphs, charts, and maps to visualize SNP distributions and associations. Prepare a comprehensive report or research paper detailing methods, analyses, and findings.
- **Validation:** Validate findings by replicating the study in different cohorts or populations. Conduct laboratory experiments to confirm the biological effects of SNPs.

Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci in a given population. In genome-wide association studies (GWAS), LD plays a crucial role in identifying genetic variants associated with traits or diseases. Linkage equilibrium (LE) is a random association of alleles at different loci and equal product of allele frequencies within haplotypes.

1. **Proximity of SNPs:** SNPs that are physically close to each other on a chromosome tend to be inherited together. This means that if one SNP is associated with a particular trait, nearby SNPs might also show association due to their linkage.
2. **LD Blocks:** The human genome is organized into blocks of high LD, where recombination events are rare, separated by regions with higher recombination rates. Within these LD blocks, SNPs are strongly correlated with each other.
3. **Tagging SNPs:** In GWAS, not every SNP needs to be genotyped. Instead, a subset of SNPs, known as tagging SNPs, can represent the variation in an LD block. This reduces the genotyping effort while still capturing the genetic diversity of the region.
4. **Association Signals:** When a SNP shows a significant association with a trait, it suggests that the causal

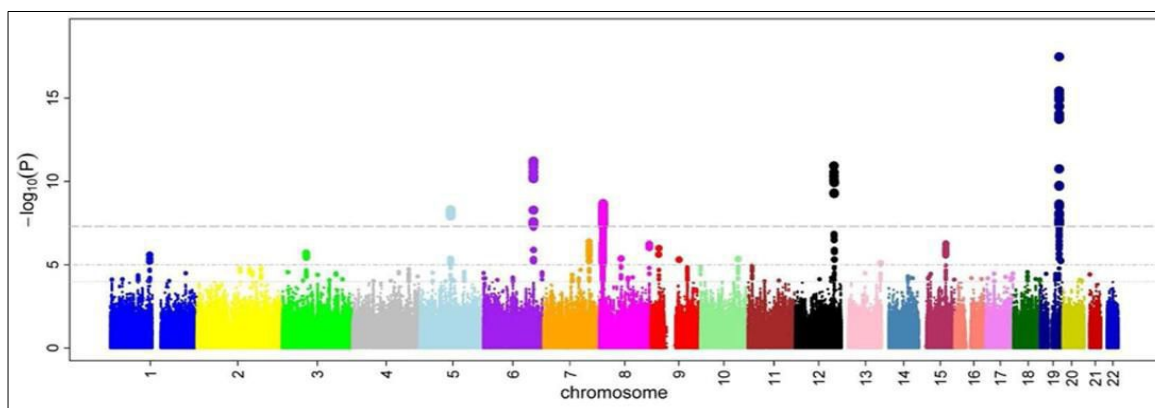
variant (the SNP or gene directly affecting the trait) is nearby, within the same LD block. Researchers then investigate the region around the associated SNP to identify the specific gene or variant responsible.

5. **Haplotype Mapping:** LD patterns are used to construct haplotypes, which are combinations of alleles at different loci that are inherited together. Haplotypes can provide more information about the genetic architecture of a trait than individual SNPs. After identifying regions of interest in a GWAS, fine-mapping techniques are applied to pinpoint the exact causal variants. This involves studying additional SNPs in high LD with the associated SNPs to narrow down the specific genetic changes responsible for the trait.

Importance of LD in GWAS

It reduces the number of SNPs needed by leveraging LD; fewer SNPs need to be genotyped while still capturing the genetic variation across the genome. LD allows researchers to link associated SNPs to nearby genes or regulatory elements, guiding further functional studies. Understanding LD patterns helps in constructing haplotype.

Manhattan plots are a common way to display GWAS results, where each dot/point represents an SNP plotted along the genome. The x-axis typically shows the position of the SNPs ordered by chromosome and base-pair position, while the y-axis shows the $-\log_{10}(p)$ of the association between the SNP and the trait. High Peaks in the Manhattan plot indicate regions with SNPs having low p-values (high significance), making it easy to identify potential genetic loci of interest. Horizontal line marks for the threshold for genome wide significance.



Source: en.m.wikipedia.org

LD-Heatmap

It is a graphical representation in GWAS to visualize correlation between SNPs within a specific genome.

Uses - for visualizing genetic structure and LD regions crucial for identification of haplotype block and understanding genetic linkage. It informs SNPs selection which captures genetic variation of other SNPs in high LD with them, reducing redundancy in genotyping.

Steps: to plot heatmap

Data preparation: Obtain genotype data for SNPs in genes of interest. Calculate LD (r^2 , D') for SNPs in the region of the gene of interest.

Plotting Heat Map: Matrix formation where each cell represents LD between pair of SNPs

Colour coding: Coding gradient represent strength of LD. Darker colour like red indicates low LD. Light colour blue/white indicate high LD. Such a way heat map is plotted with SNPs ordered by their genomic position likewise heat map plotted with SNPs ordered by their genomic position along with both axes.

Interpretation: Heat map shows diagonal pattern SNPs close to each other on chromosomes are often in high LD. The Haplotype block is a region with high LD and it appears dark. Recombination is less frequent and SNPs are inherited together.

Tools: Some software's are used to plot the LD- heat map like haploview, PLINK, R packages etc.

IMPACT	Catastrophic (5)	5	10	15	20	25
	Significant (4)	4	8	12	16	20
	Moderate (3)	3	6	9	12	15
	Low (2)	2	4	6	8	10
	Negligible (1)	1	2	3	4	5
		Improbable (1)	Remote (2)	Occasional (3)	Probable (4)	Frequent (5)
		LIKELIHOOD				

Source: techtarget.com

False Positives

Linkage disequilibrium: LD is non-random association of alleles at different loci, describing the condition of non-equal frequency of haplotype in population. A false positive occurs when a GWAS incorrectly identifies a genetic variant as being associated with a trait when, in fact, there is no real association. This can happen due to random chance, population stratification, or other confounding factors. The consequence of a false positive is that it can lead to pursue non-existent associations, wasting resources and potentially leading to incorrect conclusions.

False Negatives

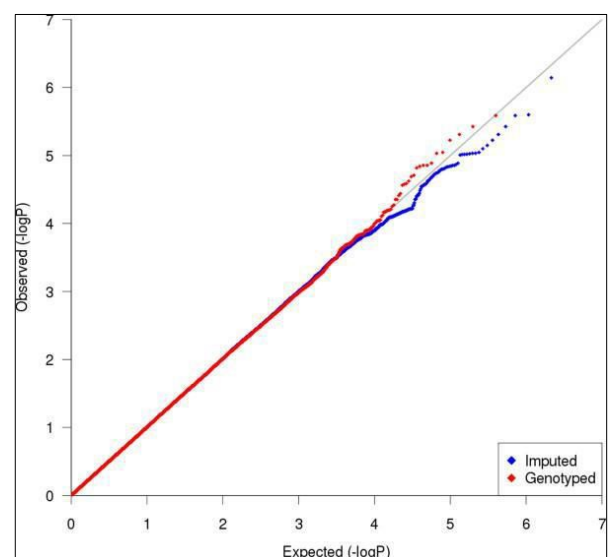
False negative: The declaration of an outcome as statistically non-significant, when effect is genuine. A false negative occurs when a GWAS fails to identify a genetic variant that is truly associated with a trait. This can happen if the study lacks sufficient statistical power, the effect size of the variant is too small to detect, or the variant is not well-represented in the dataset. The consequence of a false negative is that it can prevent the discovery of important genetic associations, leaving gaps in understanding the genetic basis of a trait.

Population Structure

In Genome-Wide Association Studies (GWAS), population structure refers to the presence of systematic differences in allele frequencies between subpopulations within a larger population. This structure can arise due to various factors such as geographical separation, ancestry, or historical migration patterns. If not properly accounted for, population structure can lead to spurious associations between genetic variants and traits, known as false positives, because the allele frequency differences may be confounded with the trait of interest. Population structure can create correlations between genetic variants and traits that are not due to a direct causal relationship but rather due to the shared ancestry within subpopulations. This can lead to the identification of associations that are due to ancestry rather than the trait itself. PCA is commonly used to detect and correct population structure in GWAS. By analyzing the genetic data, PCA identifies principal components that capture the major axes of genetic variation in the study population, which are then included as covariates in the GWAS to control for population structure. Mixed Models is another approach to account for population structure.

These models include a random effect to account for genetic relatedness between individuals, which can also capture and correct for population structure. Population stratification is a form of population structure where the population is divided into subgroups with distinct allele frequencies. If stratification is not accounted for, it can lead to biased results in GWAS. Correcting for population structure is crucial for the accuracy of GWAS results. Failure to do so can lead to incorrect conclusions about the genetic basis of traits and diseases. Therefore, population structure is always a critical consideration in the design and analysis of GWAS.

A Quantile-Quantile (QQ): Plot is a graphical tool used in statistics to compare two probability distributions by plotting their quantiles against each other. In Genome-Wide Association Studies (GWAS), QQ plots are commonly used to assess the presence of population stratification, inflation of test statistics, or other biases in the data. The primary purpose of a QQ plot in GWAS is to visualize the distribution of observed p-values against the expected distribution under the null hypothesis (i.e., no true association between the genetic variants and the trait).



Source: Fosgerau *et al.*, (2022)

Interpretation of QQ Plots

If the points closely follow the diagonal line, it suggests that the test statistics are distributed as expected under the null hypothesis, indicating no major issues like population stratification or inflation. If points deviate significantly from the diagonal line, especially at the upper tail (i.e., smaller p-values), it suggests the presence of true associations or potential confounding factors. Systematic inflation across the plot (with observed p-values consistently higher than expected) could indicate population structure or technical artifacts. Hence the QQ plots are a quick and effective way to assess the quality of GWAS results and check for issues like population stratification, inflation, or technical artifacts before proceeding with further analysis.

Procedure for conducting GWAS

- Step 1: Selection of appropriate population
- Step 2: Phenotyping and Collect samples (DNA)
- Step 3: Genotyping and sequencing
- Step 4: Statistically test each SNP for association
- Step 5: Asses the results
- Step 6: Replication

Step 1: Selection of appropriate population

Identify the trait of interest, which could be related to yield, disease resistance, drought tolerance, or other agronomic characteristics. Choose a suitable population, such as a diverse panel of accessions, landraces, or breeding lines. Population should be large enough which increases repeatability of significant individuals and get proper reproducible results. Population should be homogenous and independent by ancestry. It should be more diverse to have a higher power of GWAS. The increase in sample size may improve detection of causal variants. The ideal size range of population size is 100-500. When population size <100 leads to overestimation of phenotypic variation and gives false association and when Population size beyond 500 will be difficult to handle and manage experimental error.

Step 2: Phenotyping and Collect samples (DNA)

Phenotyping: Accurately measure and record the phenotypic traits of interest. This may involve field trials, greenhouse experiments, or laboratory measurements. Population should be grown as in multilocation trials, different seasons in replicated trials. Phenotyping of population as control and cases eg. Drought and irrigated, Infected and healthy, etc further go for isolation of DNA from both types of cases.

Step 3: Genotyping and sequencing

DNA Extraction: Extract high-quality DNA from plant tissues, such as leaves. **Genotyping:** Use high-throughput genotyping platforms like NGS to identify genetic variants (e.g., SNPs) across the genome. Common methods include SNP arrays or sequencing-based approaches.

Data Cleaning and Preparation

Quality Control: Perform quality control on the genotypic data. Filter out low-quality samples and SNPs with low call rates or minor allele frequencies. Remove samples with too many missing SNP (>95%), Remove SNP with very low MAFs (<0.05%), (Uffelmann *et al.*, 2021) [23]. Remove the genotype that does not follow Hardy-Weinberg equilibrium (i.e. allele and genotype frequencies in population will remain constant from generation to generation in absence of

evolutionary influences.) Imputation Impute missing genotypes using reference panels or software tools to improve the dataset's completeness like selection, mutation, random mating etc.

Step 4: Statistically test each SNP for association

Association Testing: Conduct statistical tests to identify associations between SNPs and the trait. This typically involves linear or mixed models to account for population structure and relatedness. Determine whether a population has sub populations or not. Sub population and main population may not have allelic differences which will influence genetic variation and give false results. Population structure shown by QQ plot (quantile-quantile) to describe experimental groups. Correction for Multiple Testing: Apply methods like Bonferroni correction or False Discovery Rate (FDR) to adjust for multiple comparisons and control false positives.

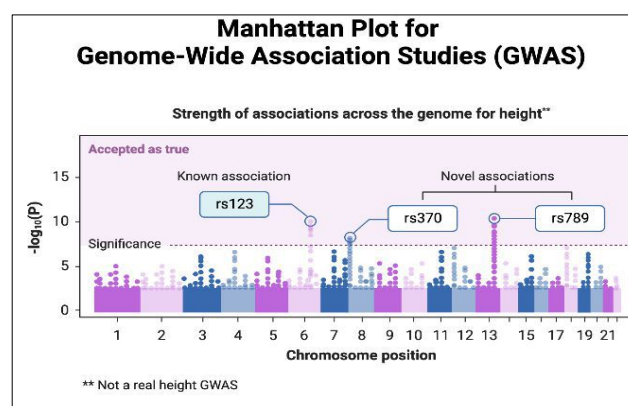
Select genotype having $MAF \geq 0.05$. (Bonferroni correction of 0.05 value for a GWAS, when the GWAS tests about one million independent variants). Calculate LD (difference between expected and observed association between alleles) by different measures like D, r^2 (correlation between a loci), D^2 etc. using software like TASSEL (Trait Analysis by aSSociation Evolution and Linkage). LD shown by heat map. Also we can calculate the genome size. Calculate P value ($-\log_{10} P$) then perform GWAS through appropriate software.

Selection of models

- General linear model (GLM) when no sub population
- Mixed linear model (MLM) when there is sub population structure.
- MLM, Kinship and PCA based correction when sub population and cryptic relatedness.

Step 5: Asses the results

Post-GWAS Analysis through the Manhattan Plot. Generate Manhattan plots to visualize the association results, with SNPs plotted along the genome and their significance levels indicated. The assessment of the result done by Manhattan plot. QMAN package (Turner, 2018) and BigTop is a tool to visualize Manhattan plots in 3D using a virtual reality framework. Manhattan plot for determination of minor variation in population. Where X-axis shows the haplotype of each region of genome tested originally by chromosome shown in different coloured blocks and Y-axis shows relative frequency of association of haplotype in that region for gene of interest or trait.

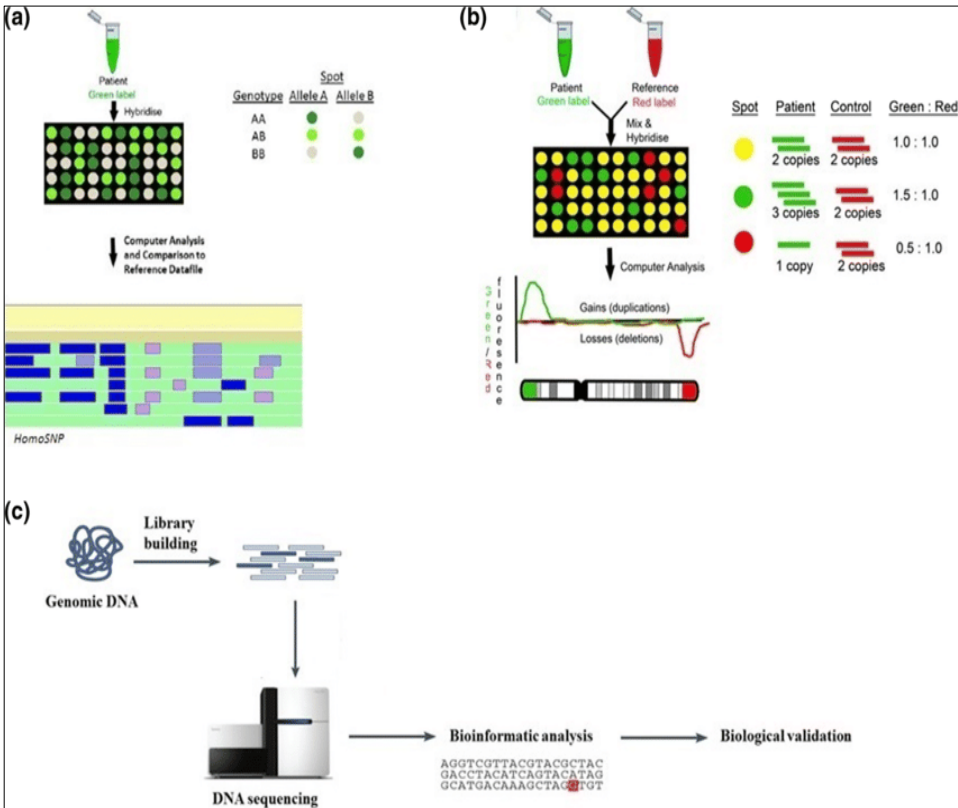


Source: app.biorender.com

Functional Annotation: Annotate significant SNPs to identify potential candidate genes or regulatory elements associated with the trait. **Pathway Analysis:** Perform pathway or network analysis to understand the biological mechanisms underlying the associations.

Step 6 : Replication and Validation

Replication helps ensure that a genotype-phenotype association observed in a genome-wide association study represents a credible association and is not a chance finding or an artifact due to uncontrolled biases. **Functional Validation:** Use experimental methods, such as gene expression studies or gene editing (e.g., CRISPR, TALEN, ZNF), to validate the functional role of candidate genes.



Source: Okutman *et al.*, (2018) ^[16]

Table 1: Difference between QTL and GWAS

QTL	GWAS
Contrasting, crossable parents and multiple generation required	Natural population, unrelated individuals required
Method of mapping linked genes with a particular phenotype with polygenic inheritance	It analyzes complete genetic background behind a particular trait
Uses particular locus	Uses whole genome
Less complex	More complex
No analysis of SNP (specific genomic loci)	Analysis of SNP
Analysis of polygenic inheritance takes place	No analysis of polygenic inheritance
Low resolution power due to limited genetic diversity of parents.	High resolution power due to more possible recombination in the population.
Studies trait with continuous variation eg. Plant height	Studies association between alleles and binary traits.
Linear regression (association between dependent and independent variables)	Logistic regression for binary traits with other independent traits

Some of the common models used in GWAS:
Simple Linear Regression (or Logistic Regression)

This is also known as Single-locus, single trait (SLST) mixed models. This is the most basic model used in initial GWAS for a single SNP-trait association test and single trait at a time. Linear Regression is used when the trait is continuous (e.g., height, blood pressure). Logistic Regression is used when the trait is binary (e.g., presence or absence of a disease).

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Y: Trait (phenotype)

X: Genotype (usually coded as 0, 1, or 2 representing the number of minor alleles)
 β_1 : Effect size of the SNP
 ϵ : Error term

This model is simple and easy to implement but it does not account for population structure or relatedness, which can lead to false positives.

Mixed Linear Models (MLM)

Mixed models are used to account for population structure and relatedness between individuals, reducing false positives due to confounding.

$$Y = \beta_0 + \beta_1 X + Z u + \varepsilon$$

Z: Design matrix for random effects (typically representing relatedness)

u: Random effect (captures genetic relatedness)

Kinship Matrix is a key component of MLMs, representing the genetic relatedness between individuals. This model is better to control population structure and relatedness, and more accurate p-values. The commonly used software includes GCTA, EMMAX, GEMMA, and MLMM etc.

Generalized Linear Mixed Models (GLMM)

It extends mixed linear models to handle non-normal traits (e.g., binary, count, categorical traits). GLMM is flexible in handling various types of traits while accounting for relatedness and population structure. The software like SAIGE is often used for binary traits in large cohorts.

Bayesian Methods

Incorporates prior information into the GWAS analysis, which can be helpful in settings with limited data or when combining GWAS with other types of data. This is generally used for effective fine mapping in candidate regions. It can handle complex models, incorporate prior knowledge, and provide full posterior distributions of effect sizes. Bayesian software for GWAS includes BIMBAM, PIQUE, and FINEMAP etc.

Multivariate GWAS

Analyzes multiple traits simultaneously, which can increase power to detect pleiotropic effects (where a single genetic

variant affects multiple traits). It is more powerful than univariate GWAS when the traits are correlated. Tools like MTAG* (Multi-Trait Analysis of GWAS) and PLINK can be used for this purpose.

Meta-Analysis Models

This model is used to combine results from multiple GWAS studies to increase statistical power and detect associations that might not be evident in individual studies. It increases power, particularly for rare variants, and allows for cross-study comparisons. The common tools include METAL and GWAMA.

Rare Variant Association Models

This focuses on the association of rare variants with traits, often using aggregate tests like burden tests or variance component tests. Burden Tests (e.g., SKAT-O):* Aggregate rare variants within a gene or region. By combining rare variants into a single test statistic called a combine test. The different tools used SKAT, VAT (Variant Association Tools), EPACTS.

Pathway-Based Approaches

It tests for the association of sets of SNPs within biological pathways with the trait, rather than individual SNPs. It can provide insight into the biological mechanisms underlying the trait. And the different tools used are MAGENTA, GSEA (Gene Set Enrichment Analysis) etc.

Table 2: Different software use for GWAS analysis

Sr. No.	Name of software	Application	Reference
1	TASSEL	LD statistics, sequence analysis, association mapping.	http://www.maizegenetics.net
2	Trimmomatic version 0.36	Removing adapters and low-quality bases	Bolger, Lohse, & Usadel, 2014
3	Os-Nipponbare-Reference-IRGSP-1	Sequence alignment	Kawahara <i>et al.</i> , 2013
4	Genome Analysis Toolkit version 3.7-0-gcfedb67 and Picard package version 2.5.0	Recording single nucleotide polymorphism	Van der Auwera <i>et al.</i> , 2014; McKenna <i>et al.</i> , 2009
5	VCFtools version 0.1.13	Identification of biallelic sites	Danecek <i>et al.</i> , 2011
6	Beagle version 4.1	Imputation means statistical inference of SNP	Browning & Browning, 2016
7	PLINK	Package use to calculate P value	https://www.google.com/search?q=name+of+tool+software+use+to+calculate+p+value+in+gwas&source=lmns&bih=625&biw=1366&rlz=1C1SQJL_enIN957IN957&hl=en&sa=X&ved=2ahUKewi3i8aM3J76AhWqi9gFHBokAZgQ_AUoAHoECAEQAA
8	rrBLUP version 4.5	Perform the GWAS	Endelman, 2011

Applications

- Biotic resistance:** Identifying Disease Genes through GWAS can pinpoint genetic factors that contribute to complex diseases. eg. The 2300 bread wheat accessions were used to investigate leaf (Lr) rust, stem (Sr) rust and stripe (Yr) rust diseases (Joukhadar *et al.*, 2020) ^[19] similarly the study carried out for cassava mosaic disease and cassava green mite severity were carried out identifying several novel and previously reported associations (Sonnewald *et al.*, 2020) ^[24].
- Abiotic tolerance:** Identifying Abiotic stress resistant genes through GWAS. Eg (GWAS) was conducted on 298 Iranian bread wheat landraces and cultivars to investigate drought tolerance indices in two cropping

seasons (2018–2019 and 2019–2020) under rainfed and well-watered environments. (Rabieyan *et al.*, 2023) ^[7]. Similarly, Two rice diversity panels, indica and aus, containing a total of 553 genotypes, are phenotyped in waterlogged conditions. (Tamimi *et al.*, 2016) ^[15]

- Yield associated traits:** The GWAS for the yield and yield contributing character to enhance the yield eg. In rice demonstrating subtle links between panicle size and yield performance (Crowell *et al.*, 2015) ^[21]. In upland cotton a GWAS identified two ethylene pathway related genes as associated with increased lint yield. (Fang *et al.*, 2017) ^[11].
- Metabolic composition:** GWAS was used for metabolite profiling and taste panels to characterize the

genetic architecture of tomato fruit taste (Tieman *et al.*, 2017) ^[4]. In maize identified 74 loci associated with kernel oil levels and fatty acid composition. (Hui Li *et al.*, 2019).

Limitations

One of the disadvantages of GWAS is it is unable to give true results in the case of traits controlled by polygenic mechanisms because of the marked influence of the environment on allelic frequency. There is less power of GWAS to study the variation due to natural factors hence there is need to improve it.

Challenges in GWAS

Complex Traits: Many traits are influenced by multiple genes and environmental factors, making it challenging to identify all contributing variants.

Population Stratification: Differences in ancestry among study participants can lead to false associations if not properly controlled.

Data Interpretation: Distinguishing between causative variants and those merely linked to the trait due to their proximity in the genome.

Key Considerations in Plant GWAS

Population Structure: Plants often have complex population structures due to breeding practices and natural selection. Use statistical methods to control for this structure.

Environmental Interactions: Consider the influence of environmental factors on trait expression and ensure phenotyping is consistent across different conditions.

Genomic Resources: Utilize available genomic resources, such as reference genomes and annotated gene databases, to aid in the interpretation of GWAS results.

Conclusion

The larger the population and the larger the number of SNPs the greater the chance of a successful result.

The developments of molecular biology techniques (e.g., CRISPR/Cas9, over-expression, or genetic complementation) have greatly accelerated the biological functions of the causative genes behind the GWAS hits. GWAS is its lack of power to detect rare alleles that are involved in natural variation.

References

- Benjamin B, Borevitz JO, Morris P. Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*. 2007;104(29):12057-62.
- Borevitz JO. Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*. 2007;104(29):12057-62.
- Cui C, Mei H, Liu Y, Zhang H, Zheng Y. Genetic diversity, population structure, and linkage disequilibrium of an association-mapping panel revealed by genome-wide SNP markers in sesame. *Front Plant Sci*. 2017;8:1189.
- Tieman D, Zhu G, Resende Jr MF, Lin T, Nguyen C, Bies D, *et al.* A chemical genetic roadmap to improved tomato flavor. *Science*. 2017;355:391-4.
- Duc G, Agrama HA, Bao S, Berger J, Bourion V, De Ron AM. Breeding annual grain legumes for sustainable agriculture: New methods to approach complex traits and target new cultivar ideotypes. *Crit Rev Plant Sci*. 2015;34:381-411.
- Dyer GA, Lopez-Feldman A, Yunez-Naude A, Taylor JE. Genetic erosion in maize's center of origin. *PNAS*. 2014;111:14094-9.
- Rabieyan E, Bihamta MR, Esmaeilzadeh Moghaddam M, Alipour H, Mohammadi V, *et al.* Analysis of genetic diversity and genome-wide association study for drought tolerance related traits in Iranian bread wheat. *BMC Plant Biol*. 2023;23:431.
- Glaszmann JC, Kilian B, Upadhyaya HD, Varshney RK. Accessing genetic diversity for crop improvement. *Curr Opin Plant Biol*. 2010;13:167-73.
- Klein R. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308(5720):385-9.
- Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. 2013;9:29.
- Fang L, Wang Q, Hu Y, Jia Y, Chen J, Liu B, *et al.* Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat Genet*; c2017 .p. 3886.
- Li H, Wang X, Zhuo Y, Chen S, Lin J, Ma H, *et al.* Molecular characterization and expression analysis of the remorin genes in tomato (*Solanum lycopersicum* L.). *Front Plant Sci*. 2023;14:1175153.
- Liu N, Xue Y, Guo Z, Li W, Tang J. Genome-wide association study identifies candidate genes for starch content regulation in maize kernels. *Front Plant Sci*. 2016;7:1046.
- Fosgerau M, Paulsen M, Kjær Rasmussen T. A perturbed utility route choice model. *Transp Res Part C*. 2022;136:103514.
- Al-Tamimi N, Brien C, Oakey H, Berger B, Saade S, Ho YS, *et al.* Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. *Nat Commun*. 2016;7:13342.
- Okutman O, Rhouma MB, Benkhalifa M, *et al.* Genetic evaluation of patients with non-syndromic male infertility. *J Assist Reprod Genet*. 2018;35:1939-51.
- Kulwal P, Ishikawa G, Benscher D, Feng Z, Yu LX, Jadhav A, *et al.* Association mapping for pre-harvest sprouting resistance in white winter wheat. *Theor Appl Genet*. 2012;125:793-805.
- Gupta PK, Kulwal PL, Jaiswal V. Association mapping in plants in the post-GWAS genomics era. *Adv Genet*. 2018;1-80.
- Joukhadar R, Hollaway G, Shi F, Kant S, Forrest K, Wong D, *et al.* Genome-wide association reveals a complex architecture for rust resistance in 2300 worldwide bread wheat accessions screened under various Australian conditions. *Theor Appl Genet*. 2020;133:3626-9.
- Saleh A, Dimitrina K, Mustafa B, Alisdair RF. Genome-wide association studies: Assessing trait characteristics in model and crop plants. *Cell Mol Life Sci*. 2021;78:5743-54.

21. Crowell S, Korniliev P, Falcaño A, Ismail A, Gregorio G, Mezey J, *et al.* Genome-wide association and high-resolution phenotyping link *Oryza sativa* panicle traits to numerous trait-specific QTL clusters. *Nat Commun.* 2015;6:10527.
22. SciGenom Conferences. NGS Bioinformatic Workshop SciGenom Research Foundation & Institute of Bioinformatics, Bangalore, India; c2014.
23. Uffelmann E, Huang QQ, Nchangwi SM, Jantina V, Yukinori O, Alicia R, *et al.* Genome-wide association studies. *Primer.* 2021;1:59.
24. Sonnewald U, Fernie AR, Gruissem W, Chang SH, Ludewig F, Rascher U, *et al.* The Cassava Source–Sink project: opportunities and challenges for crop improvement by metabolic engineering. *Plant J.* 2020;146:65.
25. Xu Y, Li P, Yang Z, Xu C. Genetic mapping of quantitative trait loci in crops. *Crop J.* 2017;5:175-84.