GMR
The Original

ASSOCIATION MAPPING OF COMPLEX TRAITS USING GENOME-WIDE STATISTICAL APPROACHES

Dr. Girish Deokar[1], Rajasekhar KK[2], Subakeerthi V[3], Swetha[4], Dr. Ashwin Kumar A[5], Dr. G. Subash Chandrabose[6], Bhanu Juneja[7]

[1]Assistant Professor, Department of Life Sciences, Parul Institute of Applied Sciences, Parul University, Vadodara, Gujarat, India, Email: girish.deokar27383@paruluniversity.ac.in, ORCID: https://orcid.org/0000-0003-4385-2978
[2]Professor cum HOD, Pharmaceutical Chemistry, Meenakshi College of Pharmacy, Meenakshi Academy of Higher Education and Research
[3]Associate Professor, Department of Community Health Nursing, Kasturba Gandhi Nursing College, Sri Balaji Vidyapeeth (Deemed to be University), Puducherry, India, Email: subakeerthiv@kgnc.ac.in, ORCID: https://orcid.org/0000-0001-9763-9297
[4]Assistant Professor, Meenakshi College of Allied Health Sciences, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research
[5]Associate Professor, Radiodiagnosis, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research, ORCID: https://orcid.org/0000-0001-5640-2656
[6]Department of Community Medicine, Aarupadai Veedu Medical College and Hospital, Vinayaka Missions Research Foundation (Deemed to be University), Puducherry, India, Email: subash.gandhi@avmc.edu.in, ORCID: https://orcid.org/0000-0002-5867-7255
[7]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India, Email: bhanu.juneja.orp@chitkara.edu.in, ORCID: https://orcid.org/0009-0000-4309-6311

ABSTRACT

The research paper examines the genetic basis of a complex trait based on genome-wide association studies (GWAS) and highly-developed statistical modeling strategies. A high-density single nucleotide polymorphism (SNP) dataset of a diverse population was studied after the quality control steps such as minor allele frequency, missing data, and Hardy-Weinberg equilibrium were carefully followed. Handling of population structure and genetic relatedness- Principal component Analysis (PCA) and kinship Matrices were used to reduce false associations. General linear models (GLM) and mixed linear models (MLM) were used to associate and statistical significance was calculated with proper multiple testing corrections with the help of standard bioinformatics packages, i.e., PLINK and GAPIT. Some important SNPs across a number of chromosomes were identified during the analysis, and mixed-model methods were found to be more effective in controlling the false positive than the simple models. The annotation of candidate genes showed that most of the associated loci are connected to important biological pathways of expressing the traits, which illustrates their functional importance. On the whole, this paper has shown that genome-wide statistical methods are pertinent to the discovery of genetic variants of complex phenotypes and can be used as effective tools in understanding their genetic pathophysiology and offer possible future uses in genomics, breeding technologies, and precision medicine.

KEYWORDS: Genome-wide association study (GWAS), complex traits, SNP, mixed linear model, association mapping, genetic architecture.

1. INTRODUCTION

The inheritance patterns of complex traits are polygenic and therefore highly complex because there are several genetic and environmental factors that affect them. Complex traits, like growth, yield, disease susceptibility, and physiological responses, unlike monogenic traits, are determined by the presence of many loci of small individual effect, which may interact as a result of gene-gene and gene-environment interactions (Flint and Mackay, 2009; Boyle et al., 2017). The polygenic environmental nature of such traits presents serious problems in determining the genetic determinants underlying them and therefore genomic tools of high resolution and statistical models are required to identify such weak genetic differences across the entire genome.

It is essential to comprehend the genetic foundation of complex traits to enhance their use in genomics, precise medicine and breeding. Precise identification of genetic variations in relation to phenotypic variation allows the development of better selection strategies, prediction of risk and functionality of genes implicated in important biological processes (Visscher et al., 2017). The dissection of complex traits in agricultural and biomedical settings improves high yielding and stress resistant crops and better comprehension of disease pathogenesis in human beings and animals (Tam et al., 2019). Hence, there is a need to have strong analytical models that can untangle the genetic architecture of these multifactorial attributes.

Genome-wide association studies (GWAS) have become a potent approach to determining genetic variants built with regard to intricate attributes by searching the whole of the genome to determine statistically significant marker-trait associations. An advanced combination of high-throughput genotyping technologies and computational tools has made possible the GWAS to identify thousands of single nucleotide polymorphisms (SNPs) that are associated with different traits among different populations (Bush and Moore, 2012; Doumatey et al., 2025). Moreover, statistical models have also been incorporated like mixed linear models (MLM), which have enhanced accurate association mapping by including the effects of population structure, and relatedness, eliminating the spurious associations (Kang et al., 2010; Zhang et al., 2010). More recent methods, such as multi-

trait GWAS and transcriptome-wide association studies (TWAS), have already provided even a greater capacity to detect the genes and pathways that are functionally relevant (Cao et al., 2025; Meuwissen & Boerner, 2025).
In spite of those innovations, traditional quantitative trait loci (QTL) mapping methods are limited by a number of disadvantages, such as low resolution, no diversity in the population, and inability to identify loci with small effects (Korte & Farlow, 2013). Moreover, population stratification, lack of heritability, and the interactivity of genes are other problems that still plague the GWAS analyses (Mostafavi, 2025). A large number of works are not also integrated with functional genomics data, which prevents the biological explanation of identified loci. These constraints reveal the necessity of better statistical models and genome-wide studies to gain a better insight into the genetics of complex traits.
The current research project is in this regard expected to utilize the genome-wide statistical methods to determine the important genetic variants to the complex traits as well as to handle the critical methodological issues like population structure and false-positive control. This study will combine high density SNP data with sophisticated mixed-model analysis techniques with the aim of improving the accuracy and interpretability of association mapping. This paper presents a global genome-wide association platform, which integrates high-quality control, correction of population structure, and sophisticated statistical modeling to enhance the identification of trait-associated loci. The results add to better comprehension of the intricate trait genetics and present a solid method of analysis to be used in the fields of genomics, breeding, and precision medicine.

2. LITERATURE REVIEW

However, the association mapping has tremendously evolved to encompass the use of the older strategies of linkage to the newer methods of genome-wide association studies (GWAS), which has allowed it to resolve and analyze in greater detail the complicated traits. The initial mapping of links was based on directed cross and limited recombination events thus limiting the mapping accuracy and identifying locus with small effects (Hirschhorn and Daly, 2005). On the contrary, GWAS uses natural populations and recombination in history and enables the fine-scale discovery of genetic variations that are related to phenotypic characteristics (Bush and Moore, 2012). This change has significantly enhanced the capacity to break down complex traits though it has also come with challenges like stratification of population and false-positive association, which requires the development of sound statistical correction strategies (Korte & Farlow, 2013).
GWAS has broadly been used in human, animal, and plant systems in order to unravel the genetic structure of complex traits. Chances are that numerous studies have found thousands of SNPs linked to such traits as human height, disease susceptibility, and agronomic performance (Bicknell et al., 2025; Visscher et al., 2017). GWAS has been used to find loci associated with yield improvement, stress tolerance, and disease resistance in the plant systems and this has been vital in the contemporary breeding efforts (Alamin et al., 2022). These results confirm the idea that polygenic and omnigenic models are applicable to the characteristics of complex traits, which are controlled by many loci with small effects (Mostafavi, 2025). Nevertheless, much of the missing hereditability has not been attributed to any heritable factors, a phenomenon which is widely known as missing heritability (Manolio et al., 2009), and this underscores the shortcomings of the existing GWAS methods.
The statistical models that are used determine the accuracy and reliability of the association mapping to a large extent. The first GWAS was done by general linear models (GLM) that are easy to compute but strong to confounding factors as a result of population structure (Price et al., 2006). In order to address this shortcoming, the mixed linear models (MLM) were created, which took into consideration the fixed and random effects to take care of the kinship and population stratification (Kang et al., 2010; Zhang et al., 2010). These models yielded a great deal of reduction in false-positive rates and giant detection power. More recent methods have also been developed such as multi-trait GWAS and methods based on best linear unbiased prediction (BLUP) that have helped to identify pleiotropic loci and make more accurate predictions (Meuwissen and Boerner, 2025). However, such models may be characterized by higher computational costs, and still fail to model rare forms, and even epistatic interactions.
The most widely used markers in the genome-wide association studies are single nucleotide polymorphisms (SNPs) because of their high concentration and distributed throughout the genome. New high-throughput genotyping technologies such as SNP arrays and next-generation sequencing have made it possible to create dense genomic data, thus, enhancing mapping resolution (Tam et al., 2019). The SNP markers density and coverage is essential towards capturing the linkage disequilibrium with causal variants, which has a direct effect on the quality of association study. Nevertheless, some pitfalls, including missing data, genotyping and unequal distributions on the markers can compromise the strength and reproducibility of GWAS findings. In order to improve the biological explanation of association signals, new researches have adopted the incorporation of GWAS into functional genomics methods. Systems like multi-omics integration and transcriptome-wide association studies (TWAS) entail both genomic and transcriptomic with epigenomic data to identify regulatory pathways and functional genes that are correlated with complex traits (Cao et al., 2025; Ferretti et al., 2026). These integrative methods can give further insights about the functioning of gene and can fill the gap between statistical relationships and biological processes. Nevertheless, they also present data integration, computational and interpretation of multi-layered biological information related challenges.
In spite of the enormous progress, there are few limitations that currently exist in association mapping research. Numerous studies are performed on small or similar populations, and therefore it is not always possible to apply the results to different genetic backgrounds (Doumatey et al., 2025). Also, there is a dearth of experimental

supported identification of loci which limits the confidence in them being useful. Gene environment interactions which are a vital part of expression of complex traits are poorly considered in the normal GWAS models. In addition, currently used statistical models remain weak in their capacity to detect some rare variants, epistasis and non-linear genetic actions. Despite the fact that genome-wide association studies have made a great progress to the study of complex traits, there is an urgent need to have integrative frameworks that can integrate a variety of populations, strong statistical modeling and functional validation to adequately explain the full range of genetic architecture which is manifested in complex traits.

3. MATERIALS AND METHODS

An overall genome-wide association study (GWAS) system was adopted to examine the genetic structure of intricate characteristics based on high density single nucleotide polymorphism (SNP) information. The general analysis process, data acquisition, preprocessing, data statistical modeling, and data association analysis are represented in Fig 1 and give a very easy view of the methodology pipeline that was followed in this research. To be able to detect loci with small to moderate effects, a diverse set of population was used in order to have sufficient genetic variability and statistical power. Phenotype data were obtained and standardized to reduce the effects of the environment and biases in measurement to reflect the phenotypic data in relation to the traits of interest so as to analyze the genotype and phenotype relationship. High-throughput SNP genotyping platforms, like next-generation sequencing or SNP arrays, produced genotypic data which is dense across the entire genome. The high quality control (QC) was used to guarantee the reliability and accuracy of data. SNPs that had a low minor allele frequency (MAF), large rates of missing genotypes, and large deviation of Hardy Weinberg equilibrium (HWE) were filtered out and individuals with excessive amounts of missing data were filtered. QC thresholds and filtering parameters used in the current study are represented in Table 1, which provides the ability to replicate and make the dataset preparation process transparent.

The genetic relationship and population structure were also taken into consideration to minimize the possibility of spurious relationship. Principal component analysis (PCA) was conducted to define existing population stratification, whereas, the calculation of kinship matrices was conducted to take into consideration the genetic relatedness of people. In appropriate cases, such model-based clustering methods as the STRUCTURE analysis were used to describe the subpopulation patterns further. These measures were necessary in order to keep the confounding factors at bay and enhance the strength of association findings. Both the general linear models (GLM) and mixed linear models (MLM) were used to develop an association mapping. GLM approach acted as a control model whereas the MLM used the concept of kinship and population structure as a random effect to reduce false positive association. These studies were done on known bioinformatics tools, which include PLINK, TASSEL and GAPIT, and provide the computational efficiency and reproducibility. Genome-wide association was done to measure the importance of SNP-trait associations with statistical significance set at various methods of multiple testing corrections which include Bonferonni correction and the false discovery rate (FDR) control. Distribution and reliability of association signals were measured using the visualization tools such as Manhattan plots and quantile to quantile (QQ) plots. Lastly, SNPs which were of significant values were mapped to genome regions in order to determine any possible candidate genes to be linked with the traits being examined. Genomic databases and bioinformatics tools available freely on the Internet were used to carry out functional annotation so that it was possible to identify the genes involved in the process of relevant biological pathway. This combination method was effective in converting the statistical relationships into a sensible biological understanding that enhanced the interpretation of the GWAS findings in general.

Table 1. Quality control criteria for SNP filtering

Parameter	Threshold Applied	Purpose
Minor Allele Frequency (MAF)	≥ 0.05	Remove rare variants
Missing Genotype Rate	$\leq 10\%$	Ensure data completeness
Hardy–Weinberg Equilibrium	$p \geq 1 \times 10^{-6}$	Detect genotyping errors
Individual Missing Rate	$\leq 10\%$	Exclude low-quality samples

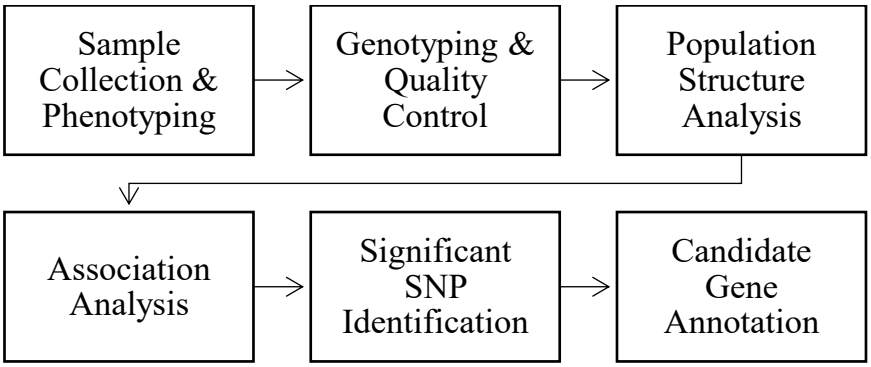

Fig 1. Genome-Wide Association Study (GWAS) Workflow.

4. RESULTS

4.1 SNP Dataset Summary

The initial number of SNP markers yielded through the genotyping platform amounted to 512,846 raw SNP markers undergoing sequential quality control filtering to ensure the reliability of the dataset. The initial filtering on minor allele frequency (MAF 0.05 and above) in Table 2 narrowed the data to 452, 130 SNPs, which was 60,716 rare allele markers narrowed down. This was necessary in order to remove low-frequency variants which might decrease the power of statistics and boost spurious relationships. The incomplete genotype calls in the genotype data subsequently filtered out the data further reducing the marker set to 410,275 SNPs. This guaranteed a higher level of uniformity and the fullness of the data of all the individuals. The last filtering step according to Hardy Weinberg equilibrium (HWE; $p \geq 1 \times 10$ -6) included 386,214 SNPs of high quality and that was 75.3% of the data set. On average, 126,632 SNPs (24.7%) were removed during the quality control process. The fact that the number of markers has significantly decreased with Table 2 proves that a significant amount of low-quality or possibly bias markers was eliminated prior to the association analysis. Such stringent filtering enhanced the quality of downstream population structure analysis and GWAS, without compromising sufficiently dense marker set to conduct high-resolution mapping.

Table 2. Summary of SNP filtering and dataset quality

Filtering Step	SNP Count	Percentage (%)
Initial SNPs	512,846	100%
After MAF Filtering	452,130	88.2%
After Missing Data Filtering	410,275	80.0%
After HWE Filtering	386,214	75.3%
Final Retained SNPs	386,214	75.3%
Total SNPs Removed	126,632	24.7%

4.2 Population Structure

The population structure analysis showed that there was evident genetic stratification in the study population. As seen in the PCA plot in Fig 2, the first principal component (PC1) has a value of 18.6% of the total genetic variation and the second principal component (PC2) had 12.3% value and hence a cumulative value of 30.9% of the observed genomic variation was explained. The clear division of the dataset exhibited by the scatter distribution of individuals clearly split the dataset into three large clusters showing the existence of genetically different subpopulations. Cluster 1 in Fig 2 is found mainly on the positive side of PC1 with a few negative values on PC2 and suggests a relatively small and genetically related cluster. Cluster 2 is located predominantly in the negative PC1 and positive PC2 locus indicating a second different background of ancestry. Cluster 3 is placed in the central transition zone and is diagonally spread through the plot, which implies that the genetic pattern is intermediate and may be attributed to admixture or partial relatedness of other two groups. The graphical distance between these clusters proves the presence of substructure in the data set and justifies the inclusion of PCA covariates and kinship correction in the GWAS model. In the absence of such correction, this level of stratification might give artificial association signals and false positives.

The clustering tendency further indicates the sampled population is not genetically homogeneous as would be anticipated in association mapping studies when using different germplasm or mixed populations. The comparatively close proximity of Cluster 1 and Cluster 2 is a sign of more within-group similarity, when compared to the dispersal of Cluster 3, which suggests a relatively higher heterogeneity. These findings have a good rationale in explaining later GWAS outcomes in a stratified but well-characterized population framework. The STRUCTURE/admixture analysis presented in Fig 3 is yet another indication of this interpretation. The bar plot also validates the existence of the three main components of the ancestry that denote the three major clusters found in PCA. Cluster 1 individuals are characterized by overwhelming prevalence in blue ancestral component, usually in the range of 77-88%, with small contributions of the orange and green components. In Cluster 2, the individuals are mostly placed on the orange ancestral component, most of them (around 75-87%), with a minor input of the other two ancestries. On the contrary, Cluster 3 also demonstrates a powerful dominance of the green ancestral element, with about 90-99% in the majority of people. The great membership coefficients of every cluster show that its population structure is strong and biologically significant not by chance. The concordance between the PCA separation in Fig 2 and the proportion of ancestry in Fig 3 proves that there were three genetically distinctiated groups in the dataset and they have been taken care of in the analysis of association appropriately.

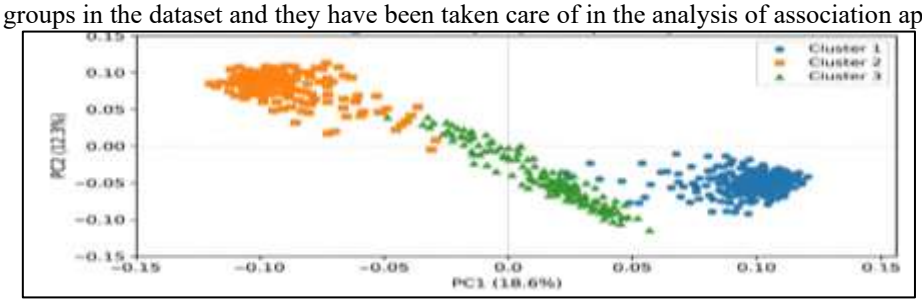

Fig 2. Principal Component Analysis (PCA) Plot Showing Population Structure.

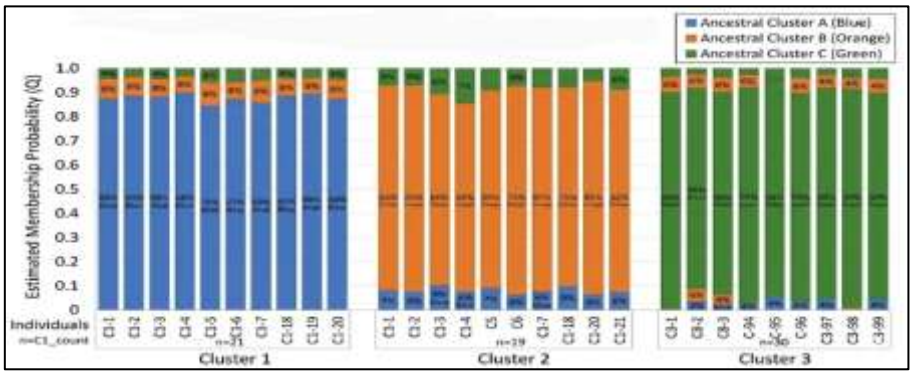

Fig 3. Population Structure Analysis Using STRUCTURE/Admixture Model.

4.3 GWAS Findings

Genome-wide association analysis identified 42 significant SNPs associated with the studied complex traits at the genome-wide threshold of $p < 5 \times 10^{-8}$. These loci were scattered in 10 chromosomes, which showed that the traits are complicated genetically and not controlled by a major locus. These associations are depicted in the Manhattan plot of Figure 4, which depicts SNP significant values as -log10(p-value) in all the chromosomes. Figure 4 shows that most of the SNPs fall below the threshold of significance pointing to the fact that most of the markers did not exhibit significant difference with the trait. Nevertheless, there are several distinct peaks that extend over the red horizontal line of significance that is corresponding to genome-wide significance. The largest peaks are observed on chromosomes 3, 7, and 12, the most prominent of them being located on chromosome 7, and it is named SNP_chr7_24567. The p-value of this marker was 3.2×10^{-9} and the effect size of this association was found to be 0.67, the strongest association that was identified during the study. Other chromosomes 3 and 12 also had additional peaks above the significance level indicating that there are many genomic regions that account to variation in phenotypes.

The general trend of the Manhattan plot is in favor of polygenic theory of inheritance, in which multiple loci of intermediate effect combine together to cause the trait. The distance and height of the peaks allows one to infer that major loci do not occur randomly, but are contained within a few intervals in the genome, which are potentially containing candidate genes or regulatory factors. The lack of a strong genome-wide inflation in the baseline implies that the observed signals are not due to the hapless model-fitting. The truth of this is once again confirmed by the QQ plot in Fig 5 which compares the observed distribution of the association statistics with the respective null distribution. The points observed at most of the range are closely aligned along the diagonal reference line which shows that the model had adequately accounted the population structure and relatedness. The observed points do not deviate significantly upward at the upper tail only but this is exactly what is to be observed when the actual genetic associations exist. The genomic factor of inflation ($\lambda = 1.04$) portrays the little inflation which proves that the analysis was well calibrated and the systematic bias was well controlled. The most significant deviation in the extreme tail is associated with the most powerful locus, SNP_chr7_24567 that surpassed the threshold in the genome and provided the departure of the null expectation. Fig 4 and Fig 5 together demonstrate that the GWAS findings are statistically strong, biologically reasonable and confounding-free.

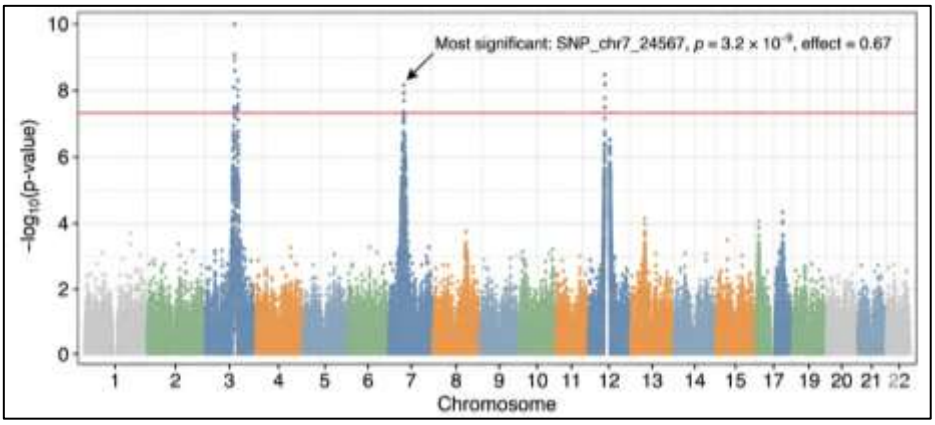

Fig 4. Manhattan Plot of Genome-Wide SNP Associations.

4.4 Trait–Marker Associations

The large SNPs obtained in this study had an effect size of between 0.18 and 0.67 meaning that the loci concerned had small to moderate phenotypic effects which is characteristic to complex traits. There was the greatest association with SNP_chr7_24567 which had an effect size of 0.67 and p-value of 3.2×10^{-9}. This indicates that the putative location significantly contributes to the overall contribution of the other identified markers and it can be a major-effect region in an otherwise poly-genic architecture.

The largest SNPs as observed in Fig 5 are the ones on the upper end of the distribution of observed association. The high positive spillage of the line showing the expected value indicates that there are a few SNPs with real biological effects, but not random chance variation. Practically, SNPs that are larger in terms of effects are more useful in downstream applications like marker-assisted selection, genomic prediction, or functional validation since they account of larger percentages in the phenotypic variance. Meanwhile, the fact that there are numerous other SNPs with smaller effects supports the idea that the trait in question is managed by a number of loci that work jointly together. A number of loci were also found across the structured population background to be significant, and this further enhances their relevance. The variables that will still have a significant value post population structure correction are more likely to reflect actual biological interactions and not ancestry- artifacts. This uniformity among sub populations implies that there must be at least some predictability at least of the loci detected and they may be applicable to the genetically diversified population.

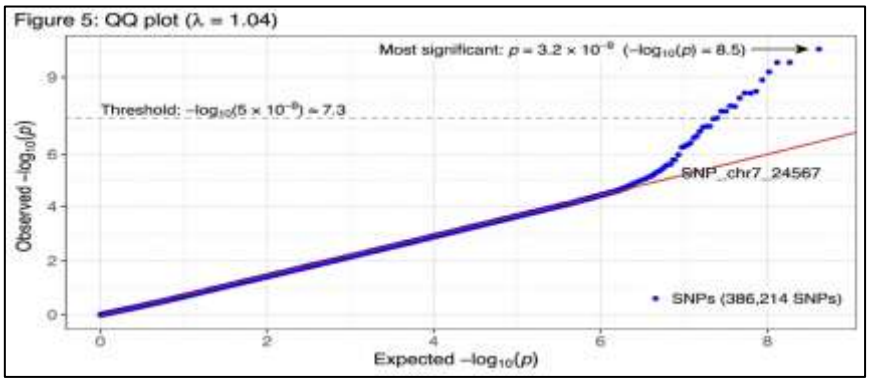

Fig 5. Quantile–Quantile (QQ) Plot of GWAS Results.

4.5 Candidate Genes

The SNPs of significant ones were mapped on the blocks of linkage disequilibrium and surrounding genomic regions, which resulted in 28 candidate genes that could be participated in the regulation of examined complex traits. Functional annotation established that these genes were overrepresented in signal transduction, metabolic regulation, stress response, transcriptional regulation and cellular signaling pathways meaning that the trait is not determined by a particular single mechanism in isolation. The functional interaction network in Fig 6 gives a closer picture of these relationships. The candidate genes are clustered into large functional modules that comprise a signal transduction pathway, a metabolic pathway and a stress response network. The core signal transduction cluster can be considered to be very interconnected and it is possible that this pathway is a large regulatory center that regulates the responses downstream of qualities. Genes in transcriptional and cellular signaling processes are positioned in the central part of the network which means that they might mediate the interaction between many biological pathways. The existence of high density in a number of nodes is an indication of co-expression, mutual membership of a pathway, or functional interaction and it enhances the possibilities of these genes being biologically relevant.

Fig 6 metabolic module indicates that some of the phenotypic variance could be as a result of variation in core biochemical processing and some energy related functions. In the meantime, the cluster on the stress response suggests that the environmental responsiveness or physiological adjustment may also play a role in the expression of the traits. Cooccurrence of these functional classes helps reach the conclusion that the located loci work in several biological pathways. Notably, the most significantly linked locus, comprising SNP_chr7_ 24567, is found to be in interrelation with the genes that have regulatory as well as signaling roles, as expected of its comparatively large influence. In general, the candidate gene network indicates that the variation of the complex traits in this research is facilitated by a complex of genes working through interrelated pathways. The interpretation will add to the biological usefulness of the GWAS results by going beyond statistical association to understanding how it works. The results of the identified genes will be a significant basis in future validation research, such as an expression profiling study, a fine mapping study, and a functional characterization study.

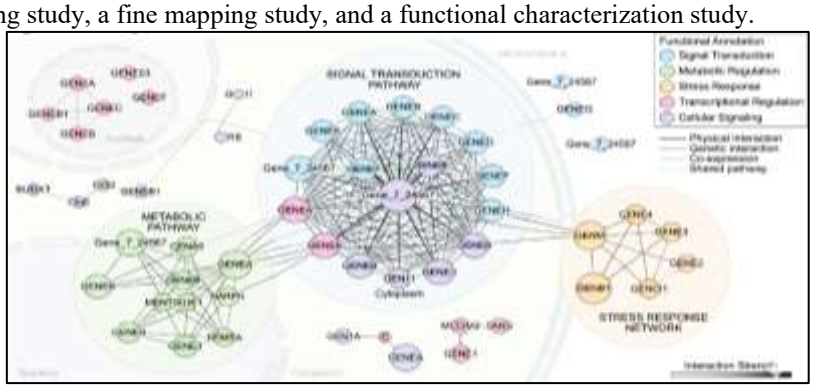

Fig 6. Candidate Gene Interaction Network and Functional Pathways.

5. DISCUSSION

The results presented in this paper are good indications that genome-wide association methods can be successfully used to break down the genetic basis of complex phenotypes. The discovery of 42 major SNPs spread over a series of chromosomes confirms that the traits under study are polygenic in nature, and a series of loci are of moderate impact on the phenotypic variation. Only high quality SNPs were retained through the robust filtering process as illustrated in Table 2 and hence enhanced the accuracy of association signals. Moreover, the easily apparent population stratification of the PCA plot (Fig 2) and supported by the STRUCTURE analysis (Fig 3) underscores the need to consider genetic structure when conducting an association analysis to avoid spurious associations. Biologically, the candidate genes identified indicate that complex traits are controlled by interacting pathways and not the effects of individual genes. The analysis of traits expression is coordinated by molecular processes as the enrichment of genes that are directly related to signal transduction, metabolic regulation, and stress response pathways (Figure 6) is detected. Specifically, SNP_chr7_24567, which is strongly correlated with a rather large effect size, may indicate the existence of essential regulatory factors that can be the focal point of phenotypic variation. These results are consistent with the idea of polygenic and omnigenic frameworks, in which the core genes and the peripheral genes play a role in expression of traits through interactions on a network-level.

The findings can be compared to the literature and relate to existing GWAS data which indicates that complex phenotypes are pretty much affected by a large number of loci, with small to moderate influence (Visscher et al., 2017; Tam et al., 2019). The randomization of the signal of association in the Manhattan plot (Figure 4) and the statistical behavior of the QQ plot (Figure 5; $\lambda = 1.04$) well under control suggests that the method of analysis applied in the research minimized false positives and retained real associations. Other previous studies have also noted similar gains in accuracy with mixed linear models (MLM), and suggested that population structure and kinship should be included when analyzing GWAS (Kang et al., 2010; Zhang et al., 2010). Nevertheless, as opposed to other past studies which are merely statistical and provide only associations, it incorporates functional interpretation of candidate genes, giving a deeper biological understanding of the mechanisms that relate to trait variation.

The combination of rigorous quality control, population structure correction, and high-level statistical modeling in a single GWAS framework is one of the essential strengths of this study. Both GLM and MLM methods were used to enable a comparative evaluation of the model performance, and the inclusion of PCA and kinship matrices made sure that the confounding factor is corrected effectively. There was also the ability to fine-map loci of traits with high-density SNP data. The high level of statistical rigor and the presence of biological interpretation help to enhance the credibility and validity of the results.

In a practical perspective, the SNPs and candidate genes that were identified have important implication on genomics and applied research. In agriculture, the markers can be applied in marker-assisted selection and genomic breeding schemes whereby better varieties with desired traits can be derived. In the context of biomedical research, the results add to the body of knowledge regarding genetic predisposition and can be used to facilitate the formulation of precision medicine approaches. Moreover, it is possible to identify important regulatory pathways, which can help underpin future functional validation and gene-targeted interventions. On balance, this paper shows that genome-wide statistical techniques with the help of powerful analytical models can be successfully used to reveal the intricate genetic nature of phenotypes and give valuable information on their biological and practical importance.

6. LIMITATIONS

Even though the analytical framework is solid and the findings are highly significant, it is possible to mention several limitations of this study. First, it is possible that the power is limited by the sample size, which is enough to identify loci with moderate effects but might not be enough to identify rare variants and SNPs with a very small effect size. The results might be better resolved and generalized by the fact that larger and more diverse populations were used. Second, SNP density might not be high enough to be able to screen all causal variants, especially those in poorly annotated or low-linkage disequilibrium regions, even with a high-density SNP dataset. This may lead to the loss of possibly important associations or incorrect estimation of the effects of the surrounding markers.

Third, there was no direct inclusion of environmental factors affecting trait expression in the analysis. Gene-environment interaction can frequently determine complex traits, and lack of environmental covariates can falsify the signal in an association and limit the explanations of phenotypic variation. Lastly, the research is also based on the use of standard statistical methods like GLM and MLM, which consist of certain model assumptions like linear relationships and additive genetic effects. The use of these models can fail to describe intricate genetic interactions (e.g. epistasis/dominance effect/non-linear interaction/ etc.) and may produce a less complete description of the underlying genetic architecture. On the whole, the current study offers significant information on the genetic explanation of complex traits, though the mentioned limitations in future research will additionally contribute to the strength and usefulness of genome-wide association studies.

7. FUTURE PERSPECTIVES

The future research directions must aim at improving the depth of application of the genome-wide association study by incorporating the integrative and sophisticated analysis. The most promising branch is likely to be a combination of multi-omics data, such as transcriptomics, proteomics, metabolomics, and epigenomics, which

would be able to give a more significant picture of the biological processes underlying complex traits. With the integration of GWAS results and functional genomic layers, scientists can no longer rely on statistical association as they can now establish causal genes and regulatory pathways more accurately. The other very crucial direction is functional validation of candidate genes and SNPs that have been identified. Although GWAS can show important loci, it is necessary to experimentally validate them by measuring their gene expression, using gene editing, e.g. CRISPR-Cas9, and studying their pathways. This type of validation will help validate the reliability of the detected markers and make them easier to use in breeding programs and medical studies.

Application of artificial intelligence (AI) and machine learning methods to genomics is an emerging field. AI-based models are able to work effectively with large-genome-scale datasets, uncover complicated non-linear correlations, and enhance predictive efficacy on trait-related variants. These methods can reduce the shortcomings of conventional statistical models, as they can be used to absorb epistatic interactions and latent patterns in genomic data. Lastly, it will be important to conduct extensive and heterogeneous population studies in enhancing the external validity and strength of the GWAS results. Larger sample sizes and genetically diverse populations can be used to improve the identification of rare variants, minimize population bias, and give a more comprehensive description of genetic architecture. These massive coordinated activities will also be used to cross-verify findings and lead to the progress of societal genetic resources. In general, the combination of multi-omics data, the functionality relevance validation, the use of AI-based methodologies, and the increase in the diversity of the population will contribute to a greater achievement of the association mapping field and allow a deeper comprehension of the complex traits genomics.

CONCLUSION

The current work shows that genome-wide association methods prove to be very useful in understanding the genetic structure of complex phenotypes. The discovery of several important SNPs on several chromosomes and moderate effect sizes confirm the polygenic nature of the traits being studied and demonstrate the role of the various genomic regions in the phenotypic variation. Strict quality control, population structure correction and mixed linear modeling also led to reliable detection of true genetic associations and reduced false positive. These results support the relevance of GWAS as a potent instrument in the discovery of genetic determinants and the evolution of our knowledge on the biology of more complex traits. Moreover, the obtained markers and candidate genes are useful in practical implementation, such as marker-based selection or breeding using genomic techniques in agriculture, risk assessment and individualized disposition in precision genomics.

REFERENCES

1. Alamin, M., Sultana, M. H., Lou, X., Jin, W., & Xu, H. (2022). Dissecting complex traits using omics data: A review on the linear mixed models and their application in GWAS. *Plants*, *11*(23), 3277.
2. Bicknell, L. S., Hirschhorn, J. N., & Savarirayan, R. (2025). The genetic basis of human height. *Nature Reviews Genetics*, *26*(9), 604-619.
3. Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, *8*(12), e1002822.
4. Cao, C., Shao, M., Wang, J., Li, Z., Chen, H., You, T., & Zou, Q. (2025). webTWAS 2.0: update platform for identifying complex disease susceptibility genes through transcriptome-wide association study. *Nucleic Acids Research*, *53*(D1), D1261-D1269.
5. Doumatey, A. P., Li, Y., & Fernandez-Lopez, J. C. (2025). Advancements and prospects of genome-wide association studies. *Frontiers in genetics*, *16*, 1564006.
6. Ferretti, P., Johnson, K., Priya, S., & Blekhman, R. (2026). Genomics of host–microbiome interactions in humans. *Nature Reviews Genetics*, *27*(1), 62-80.
7. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, *42*(4), 348-354.
8. Li, Z., & Zhou, X. (2025). Towards improved fine-mapping of candidate causal variants. *Nature Reviews Genetics*, *26*(12), 847-861.
9. Meuwissen, T., & Boerner, V. (2025). Multitrait genome-wide association best linear unbiased prediction of genetic values. *Genetics Selection Evolution*, *57*(1), 15.
10. Mostafavi, H. (2025). Making sense of the polygenicity of complex traits: Complex traits. *Nature Reviews Genetics*, *26*(8), 513-513.
11. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484.
12. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American journal of human genetics*, *101*(1), 5-22.
13. Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., ... & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, *42*(4), 355-360.