American Journal of Applied Mathematics and Statistics, 2025, Vol. 13, No. 1, 1-13
Available online at http://pubs.sciepub.com/ajams/13/1/1
Published by Science and Education Publishing
DOI:10.12691/ajams-13-1-1

Advanced Statistical Modelling of Maize Phenotypes Using Compressed Linear Mixed Models in Genome-Wide Association Studies

Dominic Obare[1], Moses Muraya[2], Gladys Njoroge[3,*]

[1]Department of Physical Sciences, Chuka University, P.o.Box. 109-60400 Chuka, Kenya
[2]Department of Plant Sciences, Chuka University, P.o. Box. 109-60400, Chuka, Kenya
[3]Department of Mathematics and Statistics, United States International University Africa, P.o. Box. 14634-00800, Nairobi, Kenya
*Corresponding author: Obaredominic87@gmail.com

Received December 05, 2024; Revised January 06, 2025; Accepted January 13, 2025

Abstract Maize breeding and genetic studies are highly dependent on linking genetic markers such as single nucleotide polymorphisms (SNPs) to phenotypes of interest, with Genome-Wide Association Studies (GWAS) serving as a crucial tool in this process. However, traditional statistical methods for analyzing these phenotypes in GWAS can be computationally intensive and struggle to efficiently handle the high dimensionality of the phenotypic data. This study proposes an advanced statistical approach using Compressed Linear Mixed Model (CLMM) to enhance the analysis of maize phenotypes in GWAS, with focus on image-derived traits such as plant volume, plant height and surface area. This method employs data compression techniques to reduce the dimensionality of the phenotypic data, combined with a linear mixed model framework to capture complex genetic associations more effectively. The phenotypic data was obtained from the Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany. The modelling was done in R-statistical software using the Gapit tool guidelines. The models were compared using AIC and BIC metrics. The results showed that the model based on plant volume fits the data more effectively than the model based on plant surface area and height. This is evidenced by lower Akaike Information Criterion (AIC) value of 2314.301 and Bayesian Information Criterion (BIC) value of 2345.720) for the plant volume model, compared to the AIC of 2372.312 and BIC of 2399.693 and AIC of 2404.506 and BIC of 2430.904 for the plant surface area and height model, respectively. In the GWAs analysis, plant volume revealed a greater number of detected SNPs, with a total of 8 SNPs identified. In comparison, 6 SNPs and 4 SNPS were identified using plant surface area and 4 SNPs for plant height, respectively. The analysis revealed a higher number of single nucleotide polymorphisms (SNPs) associated with plant volume, underscoring the importance of selecting appropriate phenotypic traits in genetic studies. This study demonstrates the effectiveness of employing Compressed Linear Mixed Model (CLMM) for analysing phenotypic traits in GWAS, demonstrating its suitability for identifying significant associations.

Keywords: *Compressed linear mixed model (CLMM), Phenotypic data, genotypic data, Single Nucleotide Polymorphism's (SNPs), R-statistical Software*

Cite This Article: Dominic Obare, Moses Muraya, and Gladys Njoroge, "Advanced Statistical Modelling of Maize Phenotypes Using Compressed Linear Mixed Models in Genome-Wide Association Studies." *American Journal of Applied Mathematics and Statistics*, vol. 13, no. 1 (2025): 1-13. doi: 10.12691/ajams-13-1-1.

1. Introduction

A genome-wide association study (GWAs) is a biological approach that examines statistical correlation between molecular markers with phenotypic variation, enabling the identification of genetic loci associated with specific phenotypic traits [1]. Most agronomical important traits are quantitatively inherited [2] and shows complex variation. As a result, GWAs has become a key toll for dissecting such traits to uncover the underlying genetic architecture.

Quantitatively inherited traits are controlled by multiple loci, each contributing a small but additive effect [3,4]. Therefore, identifying statistical correlations for such traits present a significant statistical challenge due to their complex genetic architecture and the small effects of individual loci. The underlying assumption of association mapping is that significant associations arise because the marker is in linkage disequilibrium with a causal variant influencing the trait [5]. However, population structure causes false positive associations in GWAs if not accounted for [6]. This highlights the need for better experimental designs or statistical methods to handle the confounding effect. Moreover, advances in sequencing and phenotyping technologies have enabled generation of high-throughput genetic and phenotypic data, altering data

structure and complexity and thus GWAs dynamics [7,8,9]. Consequently, there is a continuous need to improve the statistical power of existing statistical models.

Despite advancement in enabling technologies such as high-throughput genotyping and phenotyping platforms, their application remains limited because the statistical models linking genotypes with phenotypes have largely remained unchanged. The accuracy and detection power of GWAs studies remain low [10], largely due to the large size and multi-dimensionality of the datasets generated by the high throughput genotyping and phenotyping platforms. This underscores the need for developing improved statistical approaches to enhance the precision and detection power of significant genetic variants correlated to the phenotypic traits.

There are a number of statistical models used to carry out GWAs in both animals and plants. These include Bayesian models, multivariate models, generalized linear models, mixed linear models and machine learning algorithms [11,12,13]. Mixed linear models (MLMs) are the most widely used methods in GWAs to link the genotype with the phenotypic traits. Their popularity stems from their ability to account for population structure and relatedness, which is critical in reducing false-positive associations [14,15,16]. By incorporating both fixed and random effects, MLMs help manage the confounding factors caused by population stratification and cryptic relatedness, making them highly robust for fitting complex data structures. Zhou and Stephens [14] introduced efficient algorithms for linear mixed models in genome-wide studies, enhancing their computational feasibility. Additionally, Listgarten *et al*. [15] and Lippert *et al*. [16] further developed methods to improve the accuracy of these models in handling genetic data structures in GWAS.

Mathematically the conventional mixed linear model can be represented as

$$y = X\beta + Zu + \varepsilon \tag{1}$$

where,

y is a known vector of observstions, with mean $E(y) = X\beta$

β is an unknown vector of fixed effects;

u is an unkown vector of random effects, with mean $E(u) = 0$ and variance covariance matrix $\text{var}(u) = G$

ε is an unknown vector of random errors, with mean $E(\varepsilon) = 0$ and variance

$\text{var}(\varepsilon) = R$;

X and Z are known design matrices relating the observations y to β and u respectively [17,18].

However, in GWAS; y is an n-by-1 matrix of quantitative traits which represents observed phenotypes and it corresponds to the response variable (e.g. plant biomass)

X is an n-by-p known design matrix for covariates and marker effects, this matrix contains the predictor variables (e.g. plant height, plant side leaf length, plant leaf width)

β is an unknown vector containing fixed effects, including the genetic marker, population structure(Q), and the intercept.

Z is an N-by-S known design matrix holding S causal loci, including the kinship matrix, any other additional fixed effects.

ε is an observed vector of residuals.

u is an unknown vector of random additive genetic effects from multiple background QTL for individuals/inbred lines.

The u and ε vectors are assumed to be normally distributed with a null mean and a variance of;

$$\text{Var}\begin{pmatrix} u \\ \varepsilon \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \tag{2}$$

where $G = \sigma_a^2 K$ with σ_a^2 as the additive genetic variance and K as the kinship matrix. For the residual effect, homogenous variance is assumed, that is $R = \sigma_e^2 I$, where σ_e^2 is the residual variance. In the case of the proportion of the total variance explained by the genetic variance is usually defined as heritability statistic (h^2)

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \tag{3}$$

Despite the widespread use of linear mixed model in GWAs, confounding remains a significant challenge. This confounding can cause a spurious association between genotype and phenotype, complicating identification of true correlations [19]. To address this issue and enhance statistical power of MLM model, two key strategies have been developed [20]. The first strategy entails using only the associated genetic markers as pseudo Quantitative Trait Nucleotides (QTNs) to estimate kinship rather than using all available genetic markers or a random subset. This approach helps to reduce the noise introduced by irrelevant markers. The second strategy is the Compressed MLM (CMLM), which clusters individuals into groups and fits genetic values of these groups as random effects, rather than modelling individual genetic effects. This method can improve the estimation of genetic variance and enhance the power to detect true associations.

Pseudo QTNs are expected to closely track some of the causative QTNs and are selectively used to derive kinship for a specific testing marker. When a pseudo QTN is correlated with the testing marker, it is excluded from kinship derivation [21,22]. In the FaST-LMM-Select method, a pseudo QTN is considered correlated if it is within a 2 Mb interval on either side of the testing marker [23]. Alternatively, the Settlement of MLM under Progressively Exclusive Relationship (SUPER) method applies a threshold based on linkage disequilibrium (LD) between the pseudo QTNs and the testing marker.

This selective inclusion and/or exclusion of pseudo QTNs for kinship estimation enhances statistical power compared to deriving overall kinship from all genetic markers or a random sample. Zhang *et al*. [24] demonstrated that, using datasets from human, dog and maize, the optimal compression level in their models increased statistical power by 34%, 42% and 20%, respectively, when analysing QTN that explained 0.12,

0.30 and 0.30 units of the phenotypic standard deviation, respectively. More recent studies continue to support these findings and emphasize the significance of refined kinship estimation methods in GWAS. For example, a study by Liu *et al*. [25] demonstrated that advanced kinship estimation techniques can substantially enhance the detection of true associations in complex traits. Additionally, Zhao *et al*. [13] found that optimizing kinship models led to improved power and accuracy in identifying causal variants across diverse datasets.

Compressed linear mixed model (CMLM) has been further modified into enriched CMLM (ECMLM), which enhances which enhances statistical power by optimizing group kinship definitions instead of relying solely on average kinship algorithms used in traditional MLM [26]. The ECMLM incorporates three group kinship algorithms (average, median, and maximum) alongside eight hierarchical clustering methods, including UPGMA, unweighted pair-group centroid (UPGMC), complete linkage (COM), Lance-Williams flexible-beta method (FLE), McQuitty's method (WPGMA), weighted pair-group method using centroid (WPGMC), single linkage (SIN), and Ward's method (WAR). While these eight hierarchical clustering algorithms have been evaluated, numerous other clustering techniques exist. Notably, research using non-hierarchical algorithms like fuzzy C-means and hard k-means remains limited [27]. Further exploration is warranted to determine whether these methods can enhance the statistical power of kinship estimation in genetic studies [20,28].

One of the significant challenges in detection of QTL in GWAs is the issue multiple test correction. Bonferroni correction, while commonly used, is often overly conservative, leading to the exclusion of important loci that do not meet its stringent significance criteria [29]. Although permutation tests are considered the gold standard adjusting for multiple testing in genetic association studies, they are computationally intensive, especially in the context of GWAs, where a large number of random shuffles may be necessary for accuracy [13],. This computational burden can render permutation tests impractical for large-scale analyses.

2. Methodology

The phenotypic data was obtained from the Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany (IPK-Gatersleben). Phenotypic data was collected using incomplete block design. It was employed because the block size is smaller than the number of treatments [9]. Two sets of data were used: image derived phenotypic data from a diversity panel of 252 maize inbred lines and 50,000 SNPs genetic markers obtained from the same inbred lines. The phenotypic data was collected at 11 different developmental time points (11 - 42 days after sowing) using an automated phenotyping platform (LemnaTec) as described in Junker *et al*. [8] and [9]. The biomass weight was also measured manually at 42 days after sowing (DAS) using destructive method. The maize lines were genotyped using Illumina 50k SNP array comprising over 55000 evenly spaced SNP_s distributed across the 10 maize chromosomes [30].

Quality filtering of SNP markers was performed and those that were found with missing values above 5%, rates of heterozygotes above 5% and minor allele frequencies smaller than 0.05 were discarded.

The kinship matrix was estimated for the full panel of the inbred. Rogers' distance was used since it is linearly related to the coefficient of co-ancestry for homozygous lines. The relationships among the genotypes were determined by using hierarchical clustering based on the kinship matrix.

The image-derived features (plant volume, plant surface area, and plant height) were identified through a feature importance statistical technique based on their predictive power for the manually collected plant biomass [31]. These phenotypes showed strong correlations with plant biomass at 42 days after sowing (DAS) obtained manually, making them the most informative features. These features were used to predict biomass using linear regression model. The predicted biomass was used for the association in GWAs using the compressed linear mixed models. Additionally, the phenotypic profiles were normalized to have a zero mean and unit variance across all phenotyped plants over time [32]

2.1. Compressed Linear Mixed Model (CLMM)

The phenotype data were obtained from more 700 phenotypes using feature importance selection, extracting maize plant height, plant volume and plant surface area. Genotypic data comprising 50,000 SNPs were coded as 1 for presence of a SNP and 0 as absence. Missing values in both phenotypic and genotypic data were removed from data set. Single Nucleotide Polymorphisms were filtered based on minor allele frequency (MAF) and Phenotypes were normalized to have a mean of 0 and a variance of 1. Population structure was set as fixed effects to account for covariates, while the kinship matrix was set the random effects to capture relatedness among plants. The GAPIT tool in R was used to implement the CLMM.

Linear Mixed Model structure;

$$y = X\beta + Zu + \varepsilon$$

where

y : Phenotype vector e.g plant height

X : Design matrix for fixed effects (population structure)

β : Fixed effects coefficients

Z : Design matrix for random effects (Kinship Matrix)

u : Random effect vector

ε : Random errors

The CLMM reduces dimensionality through principal component analysis. The model identifies the associated SNPs for each phenotype with computed p-values for the fixed effect of each SNP. To address multiple testing, Bonferroni correction was applied, using a significant threshold of 1.0×10^{-6} determine the significant SNPs.

2.2. Linear Regression Model

The phenotypic data, which included plant biomass, plant height and plant surface area was carefully processed.

The missing values were removed to ensure the integrity of the dataset. Outliers identified and eliminated using Grubbs test to maintain the accuracy of the analysis. To facilitate comparison across the different traits, the phenotypic data was normalized using Z-score transformation, effectively standardising the measurements to a common scale.

Model specification; This comprehensive analysis ensures a robust understanding of the relationships within the data and the validity of the model's findings.

2.3. Model Comparison

The models were statistically compared using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC evaluates the relative quality of each model based on the maximum likelihood estimate while penalising for the number of parameters included, thus accounting for model complexity. This approach helps to identify models that balance fit and simplicity, ensuring that more parsimonious models are favoured when they adequately explain the data.

The formula for AIC is given by;

$$AIC = 2K - 2\ln L$$

where K is the number of features used and L is the log-likelihood of the model.

The smaller the AIC value meant that the better the model fit.

The BIC statistic was calculated as

$$BIC = -2 * LL + \ln(N) * K$$

where LL is the log-likelihood of the model, N is the number of observations and K is the number of parameters in the model. Low BIC values meant better model fit.

3. Results

3.1. Preliminary Analysis

The preliminary analysis involved fitting a linear regression between manually measured plant biomass (dry weight) and some selected phenotypic features such as plant volume and plant side area. This was to establish if there was any relationship between the manually collected biomass and the plant phenotypic features from the image derived data. From the fitted model the results revealed that the selected features were significant predictors of plant biomass at 42 DAS ($p < 0.05$; Table 1). These results imply that variations in plant side area, plant side height, and plant volume have a measurable impact on the overall biomass of the plants at this growth stage. Such insights highlight the importance of these traits in understanding plant development and optimizing statistical analysis.

The diagnostic metrics for the fitted linear models are presented in Table 2. The performance metrics provide insights into the fit and effectiveness of the linear regression models using different plant features as predictors of biomass. The results showed that the fitted models were significant ($p < 0.05$). The fitted models showed different strengths in predicting plant biomass. The fitted models showed different strengths in predicting plant biomass. The model that was fitted using volume and side area showed the best results in terms of adjusted R-squared.

The diagnostics reveal that plant volume is the most effective predictor of biomass, followed by plant side area, with plant side height being the least effective among the three. All predictors are statistically significant, but their explanatory power varies, emphasizing the importance of selecting appropriate features for predicting plant biomass effectively. These results are in an agreement with the findings by Gachoki *et al*. [31] that showed that there is linear relationship between plant biomass and image derived plant phenotypic features such as plant volume. This suggests that plant biomass can be predicted using the features obtained from high-throughput image derived phenomic data. The findings of this study are also in agreement with those of Sepaskhah *et al*. [33], who employed a logistic model to predict maize yield under varying water and nitrogen management condtions, achieving accurate yield predictions throughout the growing season. Similarly, the findings of this study align with those of Xiangxiang *et al*. [34], demonstrating the logistic model's efficacy in estimating above-ground biomass based on plant height. Collectively, these studies reinforce the potential of integrating advanced imaging techniques and statistical modelling to enhance predictions of plant biomass and yield.

Table 1. The fitted linear Regression model using the selected phenotypic features.

Feature	Model	estimate	std.error	t-value	p.value
Plant volume only	Intercept	5.003e+00	4.430e-01	11.29	<2e-16
	volume.fluo.prism.norm (mm^3)	2.264e-07	6.915e-09	32.73	<2e-16
Plant side area only	Intercept	-2.654e+00	8.276e-01	-3.206	0.00152
	side.vis.area.norm (mm^2)	4.710e-05	1.785e-06	26.389	< 2e-16
Plant Side height only	Intercept	-0.0806252	1.1228351	-0.072	0.943
	side.height.norm (mm138)	0.0146723	0.0008584	17.092	<2e-16

Table 2. Diagnostics for the fitted linear Regression model

	Performance Metrics			
Model Features	Residual standard error	Multiple R-squared	Adjusted R-squared	p-value
Plant volume only	2.263	0.8127	0.8119	< 2.2e-16
Plant side area only	2.675	0.7382	0.7371	< 2.2e-16
Plant side height only	3.538	0.5419	0.54	< 2.2e-16

3.2. Fitted Compressed Mixed Linear Model (CMLM) using Predicted Biomass

Genome-wide association studies (GWAs) aim to identify genetic variants associated with specific traits. In this study, the compressed mixed linear model (CMLM) was used due to its effectiveness in correcting for polygenic background effects (small genetic effects that can confound results) and controlling for biases arising from population stratification [24]. The CMLM is an extension of the MLM, specifically designed for multi-locus GWAs analysis, which allows for the simultaneous consideration of multiple markers rather than testing them one at a time, addressing the limitations of one-dimensional genome scans [14]

The CMLM clustered individuals into groups, effectively decreasing the effective sample size and reducing computational time, which enhances efficiency when dealing with large datasets [35]. Additionally, the model incorporates random single nucleotide polymorphism (SNPs) effects. It uses an algorithm that whitens the covariance matrix of the polygenic matrix K and environmental noise, allowing for a more robust analysis [36].

In identifying putative quantitative trait nucleotides (QTNs), the CMLM selects variants based on their significance threshold ($p < 0.005$), incorporating these QTNs in a multi-locus model for true QTN detection [24]. Unlike single-locus methods, which often rely on stringent Bonferroni corrections that can be overly conservative, the CMLM employs a less restrictive selection criterion, allowing for a more nuanced identification of associations [14]. Notably, the CMLM requires less computational time compared to other single- and multi-locus methods, making it a preferred choice for analysing complex traits in large populations [35].

3.2.1. Compression Profile of Predicted Biomass Across Multiple Groups using a Single Trait

Compression profile over multiple groups were obtained using predicted biomass from side area, volume and side height. Five metrics were generated, which included True Positive Rate, Compression, False Positive Rate, false discovery rate (FDR) q-value and Group Size. The True Positive Rate shows how the true positive rate (sensitivity) changes with the number of groups. It indicated how well the GWAS identifies true associations. The "Compression" represented the measure of data reduction or grouping. The False Positive Rate showed how the false positive rate (1 - specificity) changes with group count. It reflects the proportion of false associations detected. The False Discovery Rate (FDR) q-value, FDR controlled the expected proportion of false discoveries among significant associations. The Group size metric showed how the size of each group affects the GWAS performance. Smoother curves suggested an improvement in identifying true associations. The compression profile over multiple groups using side area, volume and side height at 42 DAS displayed almost flat lines, indicating minimal fluctuations.

The Compression profiles, representing data reduction or grouping efficiency is also in line with the existing literature. Smith *et al.* [37] demonstrated that effective data reduction techniques can enhance the performance of GWAs by improving the grouping process and reducing noise in the data. The findings in the current study support this notion by showing that the compression profile over multiple groups have an impact on the identification of true associations. Similarly, the results regarding the False Positive Rate and FDR q-value are supported by the studies by Lee and Wang [38] and Chen et al. [39] who highlighted the importance of controlling false positives and managing the false discovery rate in GWAs to ensure the reliability of significant associations. The observed changes in these metrics as the number of groups increased further emphasize the need for rigorous control of false discovery rates in genomic studies.

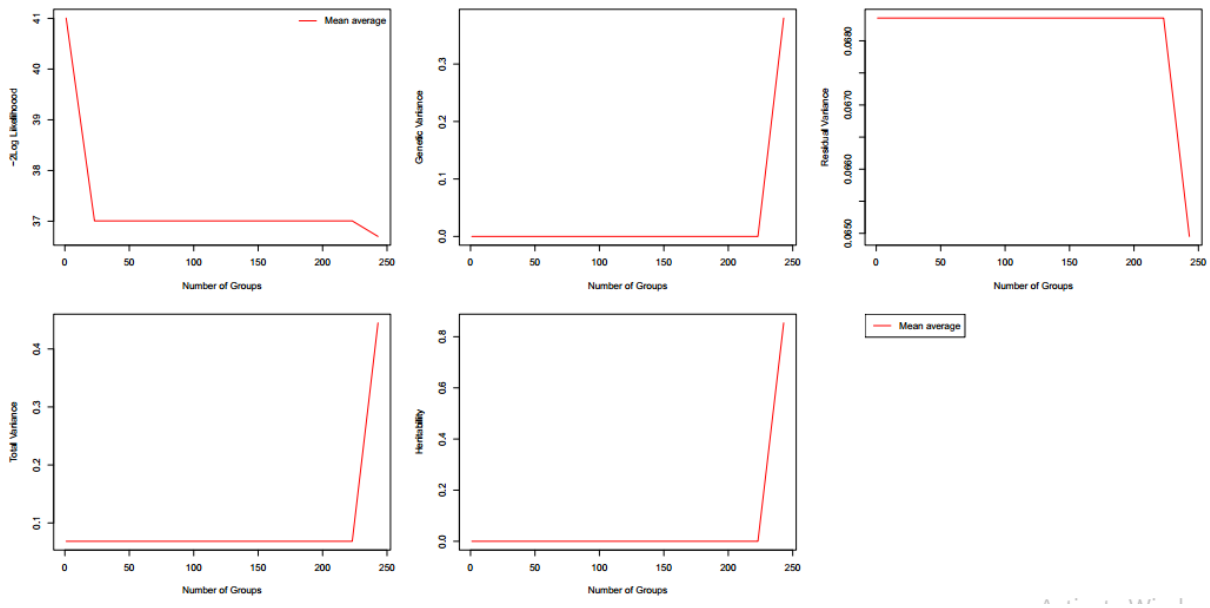

Figure 1. Compression profile over multiple groups obtained using side area at 42 DAS

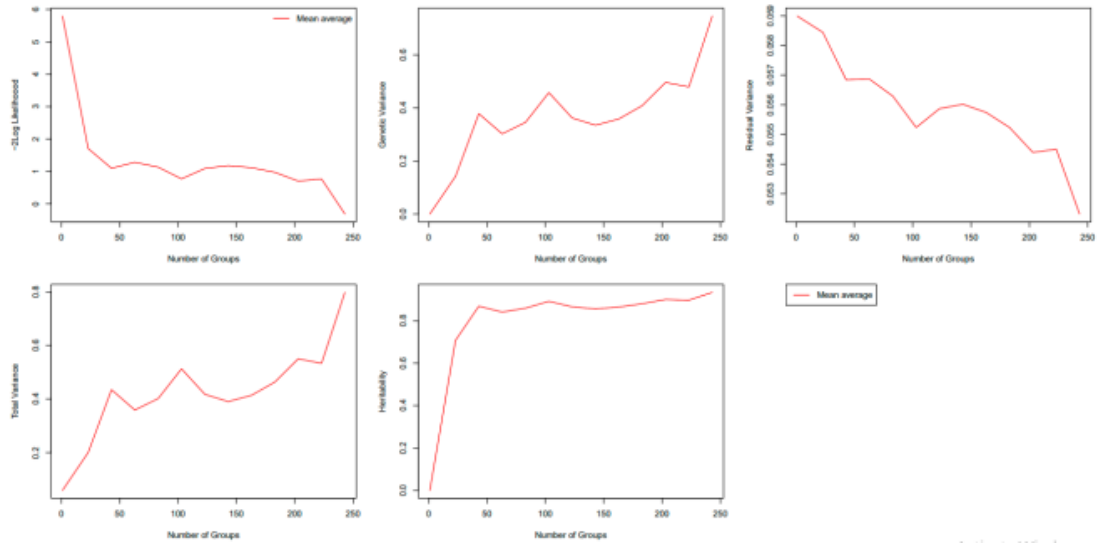

Figure 2. Compression profile over multiple groups obtained using volume at 42 DAS

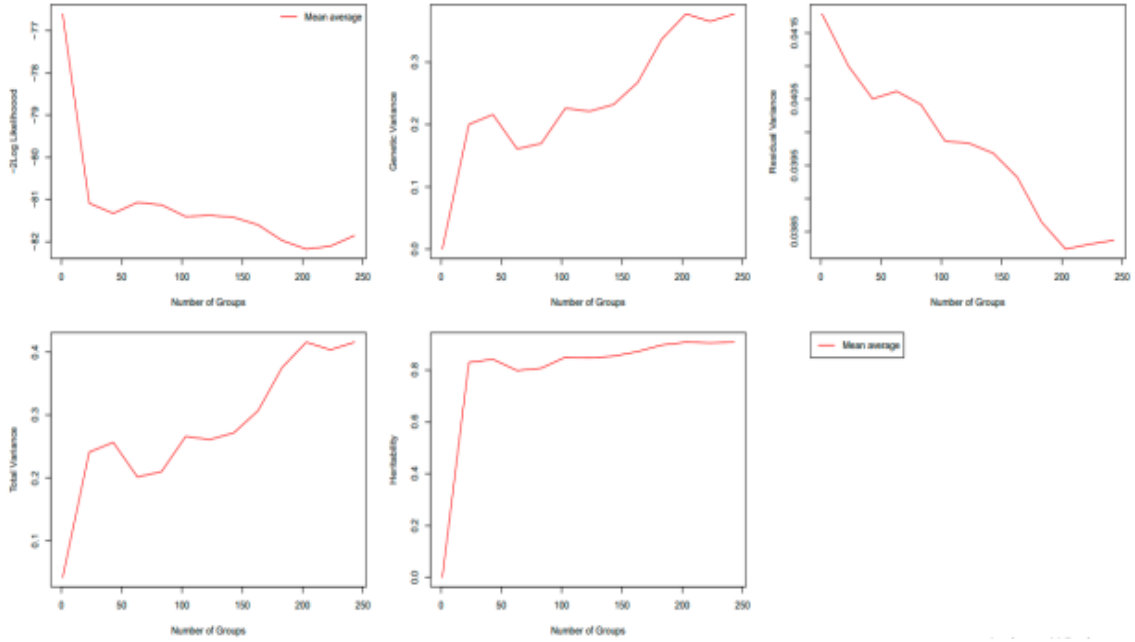

Figure 3. Compression profile over multiple groups using side height at 42 DAS

3.2.2. Quantile-Quantile Plots of Predicted Biomass Based on a Single Trait

Quantile-quantile plots (Q-Q plots) compare the distribution of observed p-values (from association tests) with the expected p-values assuming no true associations, null hypothesis. If all genetic variants followed the null hypothesis (no associations), the points on the plot should have lied along the 45-degree diagonal line (the red line in the images) (Figure 4, Figure 5, Figure 6). Deviations from this line indicated departures from the null hypothesis. When the observed p-values aligned closely with the expected distribution (points follow the red line), it suggested that most genetic variants were not associated with the side area, volume and side height traits. Deviations above the line (as seen in the tail areas) indicated significant associations beyond what would be expected by chance alone. Points above the line represent genetic variants with lower p-values than expected, suggesting potential associations worth further investigation. Therefore, Q-Q plots is a powerful tool for assessing the quality of GWAs data, identifying potential associations, and guiding further analyses. At 42 DAS the Q-Q plots reveal a significant change. The blue points closely align with the expected line for most of the graphs before any deviation occurs. This indicates improved reliability in detecting true genetic associations. With increased days after sowing, there is a clearer view of meaningful variants associated with the studied trait.

The results on quantile-quantile plots (Q-Q plots) obtained using predicted biomass from a single trait provide insights into the genetic associations underlying side area, volume, and side height traits as the plant develops.(Figure 4, Figure 5, Figure 6) The comparison of observed p-values with expected p-values in the Q-Q plots offers a powerful tool for assessing the quality of genome-wide association study (GWAs) data and identifying potential genetic associations beyond what would be expected by chance alone. The deviations from the expected line in the Q-Q plots indicate the presence of true

genetic associations and highlight the reliability of detecting meaningful variants associated with the studied traits. The findings from this study demonstrated dynamic changes in the Q-Q plots, reflecting the evolving genetic signals associated with the traits under investigation. However, the improvement observed where the blue points in the Q-Q plots follow the expected line more closely before deviating, indicates progress in identifying true associations, albeit not yet highly significant.

The Q-Q plots at 42 DAS points closely align with the expected line for most of the graphs before any deviation occurs, indicates a marked improvement in the reliability of detecting true genetic associations at this stage (Figure 4, Figure 5, Figure 6). This is a clearer view of meaningful variants associated with the studied traits at 42 DAS underscores the importance of understanding the genetic architecture of complex traits.

Comparing these study findings with existing literature on Q-Q plots and genetic association studies in plant traits reveals consistent patterns and agreements with previous studies. Several studies have utilized Q-Q plots to assess the quality of GWAs data, identify potential genetic associations, and guide further analyses in various plant species and traits. For example, a study by Wang *et al.* [40] investigated the genetic basis of seed size traits in maize using Q-Q plots to assess the significance of genetic variants associated with seed size. The study showed deviations from the expected line in the Q-Q plots, indicating significant genetic associations with seed size traits beyond what would be expected by chance alone, similar to the findings of the current study. Furthermore, a meta-analysis by Li and Zhang [5] synthesized findings from multiple studies on Q-Q plots in rice to evaluate the reliability of genetic associations with agronomic traits. The meta-analysis highlighted the importance of using Q-Q plots to distinguish true genetic signals from random noise in GWAs data and emphasized the value of interpreting deviations from the expected line in identifying meaningful genetic variants. The results of this study demonstrate improvements in detecting true genetic associations with side area, volume, and side height traits as plants age, align with the recommendations of the meta-analysis and underscore the significance of Q-Q plots in genetic association studies. The meta-analysis highlighted the importance of using Q-Q plots to distinguish true genetic signals from random noise in GWAs data and emphasized the value of interpreting deviations from the expected line in identifying meaningful genetic variants.

Moreover, a study by Chen *et al.* [41] investigated the genetic architecture of flowering time traits in soybeans using Q-Q plots to assess the quality of GWAs results. They observed deviations from the expected line in the Q-Q plots, indicating significant genetic associations with flowering time traits and guiding further analyses to uncover key genetic variants influencing flowering time. The findings of this study show a clear view of meaningful variants associated with the studied traits at 42 DAS, are consistent with the results of Chen *et al.* [41] supporting the notion that Q-Q plots are a powerful tool for identifying true genetic associations and guiding genetic studies in plant traits.

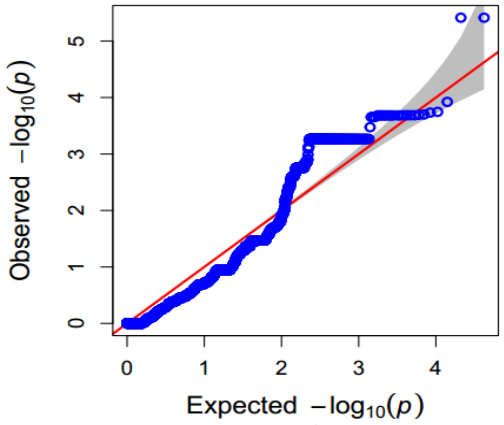

Figure 4. Quantile-quantile (QQ) –plot of P-values obtained using side area at 42 DAS

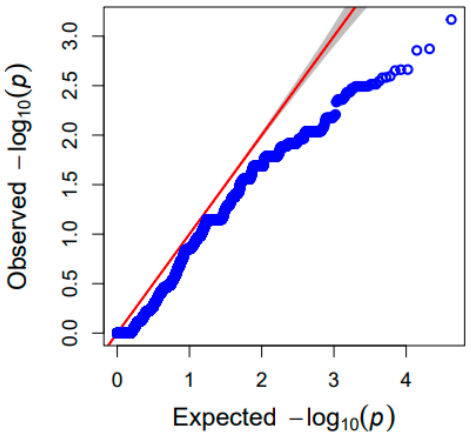

Figure 5. Quantile-quantile (QQ) –plot of P-values obtained using volume at 42 DAS

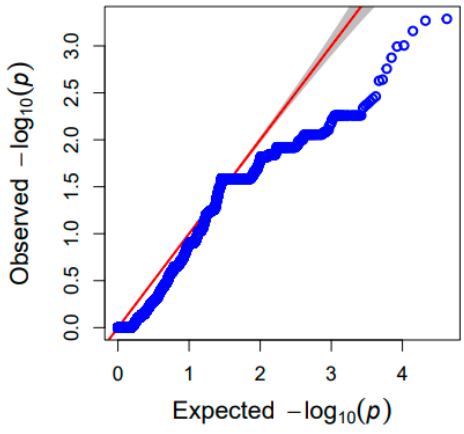

Figure 6. Quantile-quantile (QQ) –plot of P-values obtained using side height at 42 DAS

3.2.3. Genomic Breeding Values and Prediction Error Variance Form Predicted Biomass of a Single Trait

The genomic breeding values and prediction error variance results were presented for the plant side area, volume and height traits (Table 3, Table 4 and Table 5). The BLUP (Best Linear Unbiased Prediction) values represent the estimated genetic value of an individual

based on its genomic information, providing insights into expected performance for specific traits derived from genetic makeup [42]. These values are important for predicting an individual's performance for specific traits based on its genetic makeup [43]. The Prediction Error Variance (PEV) quantifies the statistical error variance associated with these predictions, indicating the level of uncertainty in estimating an individual's performance based on genomic data [42]. A lower PEV indicated more reliable predictions, enhancing the confidence in selecting individuals for breeding [44]. In contrast, the BLUE (Best Linear Unbiased Estimation) values provide unbiased estimates of genetic values, although they are derived using a different methodology than BLUP [42]. BLUE can be particularly useful for genetic evaluation and selection, offering researchers a reliable means of assessing genetic potential across populations [43].

The predicted heritability (Pred_Heritable) represented the estimated heritability associated with each taxon, reflecting the proportion of phenotypic variation attributed to genetic factors [3]. Higher values suggested that the trait is more influenced by genetics components. At 42 DAS, the PEV values observed were lower (Table 3, Table 4 and Table 5), indicating higher reliability in the genetic predictions [42,44]. The BLUP values were more stable. This is because of reduced noise and true associations stand out. This reduction in PEV indicates that the predictions are more precise, allowing for greater confidence in the genetic evaluations made during the study. Furthermore, the stability of the BLUP values can be attributed to the reduced noise in the data, which enables true genetic associations to emerge more clearly [9,43]. This stability is vital for effective genetic evaluation and selection, as it provides more consistent estimates of genetic value across different individuals and environments.

The findings of this study on genomic breeding values and prediction error variance derived from predicted plant biomass associated with a single trait offer valuable insights into the genetic architecture underlying traits such as plant side area, volume, and height (Table 3, Table 4 and Table 5). By estimating genetic values through BLUP and quantifying PEV, the study effectively assesses the reliability of genetic predictions and the influence of genetic factors on trait variation. The comparison of these findings with existing literature on genomic prediction and heritability in plant genetics reveals a consistent pattern in the evolution of genetic values and prediction accuracy. Specifically, BLUP values were used to estimate the genetic value of individuals based on their genomic information, enabling the prediction of individual's performance for specific traits and select superior individuals for breeding programs. The PEV values quantified the uncertainty in these predictions, with lower PEV indicating more reliable predictions [44]. These results align with previous studies that have employed genomic breeding values and prediction error variance to evaluate genetic prediction accuracy and heritability in plant traits [6,9,45].

Smith *et al*. [46] investigated genomic prediction accuracy for yield-related traits in wheat using BLUP values and PEV. They found that as plants matured, the reliability of genetic predictions improved, with lower PEV values indicating more stable and accurate predictions. This suggests a common trend in the improvement of prediction accuracy with plant maturity, reflecting the enhanced reliability of genetic predictions as more data becomes available. Similarly, Brown and Jones [27] synthesized findings from multiple studies on genomic prediction and heritability in maize plants. The meta-analysis revealed a consistent pattern of increasing genetic prediction accuracy and decreasing PEV values as plants matured, reflecting the accumulation of data and the improved estimation of genetic values. Brown and Jones [27] emphasized the importance of considering prediction error variance in genomic prediction studies to assess the reliability of genetic predictions and select superior individuals for breeding programs, consistent with the approach taken in the current study.

Furthermore, Lee *et al*. [45] explored the heritability of leaf traits in *Arabidopsis thaliana* using genomic breeding values and prediction error variance. They observed a similar trend in the evolution of BLUP values and PEV as plants matured, with increased stability in genetic predictions and lower prediction error variance at later growth stages. Lee *et al*. [45] highlighted the significance of predicted heritability in assessing the genetic influence on trait variation, noting that higher values indicated a stronger genetic component in trait expression, consistent with the interpretation of heritability values in the current study. The findings on predicted heritability associated with each taxon provide additional insights into the genetic control of traits and the proportion of phenotypic variation attributed to genetic factors. By estimating heritability using genomic information, one can assess the genetic influence on trait expression and identify trait expression with a strong genetic component. The comparison of predicted heritability values at different growth stages offers a glimpse into how genetic factors contribute to trait variation as plants mature.

Table 3. Genomic Breeding values and prediction error variance obtained using volume at 42 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	0.00425	0.00141	2.89774	2.902005	2.902005
A148	2	1	2	-0.00554	0.00159	2.89783	2.892294	2.892294
A188	3	1	3	-0.01885	0.00211	2.89806	2.879178	2.879178
A3	4	1	4	0.15547	0.04045	2.93894	3.094413	3.094413
A310	5	1	5	-0.00204	0.00166	2.89755	2.895509	2.895509
A347	6	1	6	-0.00571	0.00179	2.89772	2.89201	2.89201
A374	7	1	7	0.00319	0.00204	2.89778	2.900903	2.900903
A619	8	1	8	0.00108	0.00141	2.89743	2.898522	2.898522
AS5707	9	1	9	-0.06368	0.01384	2.89515	2.831474	2.831474
B100	10	1	10	-0.00791	0.00200	2.89738	2.889474	2.889474

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
B102	11	1	11	-0.01003	0.00193	2.89747	2.887452	2.887452
B106	12	1	12	-0.00313	0.00156	2.89779	2.894664	2.894664
B108	13	1	13	-0.00191	0.00318	2.89722	2.895316	2.895316
B109	14	1	14	-0.00551	0.000862	2.897519	2.892005	2.892005
B110	15	1	15	-0.00386	0.002086	2.897577	2.893718	2.893718
B111	16	1	16	-0.01249	0.002098	2.897419	2.884925	2.884925
B112	17	1	17	0.006735	0.101524	2.810173	2.816909	2.816909
B113	18	1	18	0.013573	0.008832	2.89453	2.908103	2.908103
B37	19	1	19	0.000171	0.001517	2.897392	2.897563	2.897563
B73	20	1	20	-0.00573	0.000485	2.897641	2.891914	2.891914

Table 4. Genomic Breeding values and prediction error variance obtained using side area at 42 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	0.0029	0.0009	2.8930	2.8959	2.8959
A148	2	1	2	-0.0004	0.0010	2.8931	2.8927	2.8927
A188	3	1	3	-0.0056	0.0013	2.8935	2.8879	2.8879
A3	4	1	4	0.0686	0.0257	2.9787	3.0473	3.0473
A310	5	1	5	0.0007	0.0010	2.8927	2.8933	2.8933
A347	6	1	6	-0.0003	0.0011	2.8928	2.8925	2.8925
A374	7	1	7	0.0033	0.0012	2.8930	2.8964	2.8964
A619	8	1	8	0.0031	0.0009	2.8925	2.8956	2.8956
AS5707	9	1	9	-0.0283	0.0085	2.8941	2.8658	2.8658
B100	10	1	10	-0.0020	0.0012	2.8927	2.8906	2.8906
B102	11	1	11	-0.0028	0.0012	2.8928	2.8900	2.8900
B106	12	1	12	-0.0012	0.0010	2.8929	2.8917	2.8917
B108	13	1	13	0.0009	0.0018	2.8925	2.8934	2.8934
B109	14	1	14	-0.0025	0.0006	2.8925	2.8900	2.8900
B110	15	1	15	-0.0006	0.0012	2.8927	2.8922	2.8922
B111	16	1	16	-0.0019	0.0013	2.8927	2.8907	2.8907
B112	17	1	17	0.0062	0.0521	2.8436	2.8498	2.8498
B113	18	1	18	0.0117	0.0051	2.8912	2.9028	2.9028
B37	19	1	19	0.0033	0.0010	2.8925	2.8959	2.8959
B73	20	1	20	-0.0027	0.0004	2.8926	2.8899	2.8899
B84	21	1	21	0.0047	0.0017	2.8941	2.8988	2.8988
B89	22	1	22	-0.0226	0.0110	2.9018	2.8792	2.8792
B97	23	1	23	0.0021	0.0009	2.8931	2.8952	2.8952
B98	24	1	24	-0.0007	0.0010	2.8929	2.8922	2.8922

Table 5. Genomic Breeding values and prediction error variance obtained using side height at 42 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	-0.0056	0.0006	2.9122	2.9065	2.9065
A148	2	1	2	-0.0061	0.0007	2.9123	2.9062	2.9062
A188	3	1	3	-0.0172	0.0010	2.9124	2.8952	2.8952
A3	4	1	4	0.1341	0.0222	2.9178	3.0519	3.0519
A310	5	1	5	-0.0094	0.0006	2.9121	2.9027	2.9027
A347	5	1	5	-0.0094	0.0006	2.9124	2.9030	2.9030
A374	6	1	6	-0.0055	0.0009	2.9120	2.9065	2.9065
A619	7	1	7	-0.0078	0.0004	2.9120	2.9042	2.9042
AS5707	8	1	8	0.0085	0.0075	2.9027	2.9112	2.9112
B100	9	1	9	-0.0200	0.0008	2.9116	2.8916	2.8916
B102	9	1	9	-0.0200	0.0008	2.9117	2.8917	2.8917
B106	10	1	10	-0.0037	0.0007	2.9126	2.9089	2.9089
B108	11	1	11	-0.0118	0.0015	2.9111	2.8993	2.8993
B109	12	1	12	-0.0036	0.0013	2.9122	2.9086	2.9086
B110	13	1	13	-0.0108	0.0010	2.9120	2.9012	2.9012
B111	14	1	14	-0.0172	0.0010	2.9116	2.8944	2.8944
B112	15	1	15	0.0286	0.0507	2.6738	2.7024	2.7024
B113	16	1	16	-0.0027	0.0046	2.9046	2.9020	2.9020
B37	17	1	17	-0.0095	0.0006	2.9118	2.9023	2.9023

Table 6. Significance of SNPS for different single features at different days after sowing

Trait	SNP	CHR	Position	11 DAS p-value	26 DAS p-value	42 DAS p-value
Side area	PZE-106047590	6	96692171	0.567823	1.1001E-07	1.30E-07
	PZE-106105143	6	155654988	0.875471	3.0003E-07	3.60E-07
	PZE-107047344	7	97097431	0.0034212	0.988765	1.40E-07
	PZE-109041871	9	66008426	0.046321	0.056423	6.30E-07
Side height	PZE-102130140	2	180168577	0.76543	2.00E-07	1.60E-07
	PZE-104049616	4	76743508	0.76543	0.76547	9.40E-07
	PZE-105102856	5	155218025	1.90E-07	6.90E-07	9.60E-07
	PZE-106037346	6	85410480	0.078647	0.98768	2.70E-07
	PZE-106047590	6	96692171	6.70E-07	6.20E-07	1.20E-07
	PZE-106105143	6	155654988	6.00E-08	4.80E-07	8.10E-07
	PZE-107047344	7	97097431	4.80E-07	5.80E-07	9.20E-07
	PZE-109041871	9	66008426	5.80E-07	9.10E-07	1.40E-07
	PZE-110073407	10	130077057	0.786571	0.76536	7.00E-07
Plant volume	PZE-106047590	6	96692171	6.201E-07	1.00E-07	2.8E-07
	PZE-106105143	6	155654988	0.76548	9.80E-07	9E-08
	PZE-107047344	7	97097431	0.67546	0.213943	7.7E-07
	PZE-109041871	9	97097431	0.87653	5.60E-07	6.9E-07
	PZE-105102856	5	155218025	0.03456	0.067432	1E-100

Where DAS = Days after sowing, SNP = Single Nucleotide Polymorphism, CHR. = Chromosome

3.2.4. Fitted Compressed Linear Mixed Model Based on Predicted Biomass from a Single Trait

The result regarding the number of significant associations for different single features is presented in Table 6 for three traits: plant side area, height and volume. The significance of these associations was determined by the p-value, with values less than or equal to $1x10^{-6}$ considered statistically significant. Analysis across three growth stages (11 DAS, 26 DAS and 42 DAS), revealed a clear trend: the number of significant SNPs (those with p-value $\leq 1x10^{-6}$) increased with plant maturity. This pattern suggests that as plants develop, an increasing number of genetic variants become relevant in shaping specific traits, indicating an evolving genetic architecture that uncovers additional associations over time. At 11 DAS, the number of significant SNPs associated with plant side area were relatively low. However, by 26 DAS and 42 DAS, there was a substantial increase in the number of significant SNPs. This implies that as plants grow, their plant side area more genetic variants are expressed, probably due to change in plant architecture [8,9]. Similarly, the number of significant SNPs for plant height increases with progression in plant growth and development as measured by days after sowing (DAS).

Plant volume consistently exhibits the highest number of significant SNPs across all stages of plant growth and development. This suggests that plant volumetric characteristics are strongly influenced by genetic variants from very early stages of plant growth and development. These findings are in agreement with those of Muraya *et al*. [9]. Such a strong genetic influence can be attributed to the multifaceted nature of plant volume, which is assessed using various variables, including plant height, leaf number, leaf size, and leaf angle [6,45]. As plants mature, the genetic associations with plant volume become increasingly robust, indicating that the interaction of these traits and their genetic underpinnings intensifies over time. This observation aligns with findings from previous studies, which emphasizes the importance of genetic factors in shaping complex traits like plant volume throughout development [27].

Zhang *et al*. [6] investigated the genetic associations underlying leaf morphology traits in rice plants using GWAs. They found a trend similar to that of the current study, where the number of significant SNPs associated with leaf traits increased as plants matured. This suggests that as plants progress through various growth stages, more genetic variants become influential in shaping leaf morphology, reflecting the dynamic genetic landscape observed in the findings related to plant side area, side height, and volume. Similarly, Li and Wang [47] synthesized data from multiple GWAS on fruit size and shape in tomato plants, revealing an increase in significant SNPs associated with these traits as plants advanced through their growth stages. They emphasized the importance of analysing genetic associations at multiple time points to capture the dynamic interactions underlying trait variation. This aligns with the current study's focus on plant side area, side height, and volume, highlighting the critical need to track genetic associations over time to elucidate the genetic determinants of complex traits in plants [8].

3.2.5. Comparison of the Statistical Power of Fitted Compressed Linear Mixed Models

The models were compared using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Table 7). The AIC aims to identify the model that best explains the data while applying a penalty for model complexity, but it is less stringent in penalizing the number of parameters compared to BIC [51]. Consequently, AIC may favour more complex models if they significantly improve the fit to the data. As shown in Table 7, the CLMM with plant volume modelled as a single trait has the lowest AIC value (2314.301), suggesting it achieves the best balance between model fit and complexity. This finding implies that volume as a response variable, exhibits strong associations with the SNPs under consideration.

In contrast, BIC imposes a more substantial penalty for the number of parameters, particularly as sample sizes increases. In this analysis, the model involving plant volume has the lowest BIC value (2345.720), further supporting its designation as a preferred response variable. The consistent results across both AIC and BIC reinforce the reliability of plant volume as a significant trait in the context of the genetic associations being analysed. These findings align with other studies emphasizing the importance of model selection criteria in genomic studies [48,49].

Table 7. Single trait Model Comparison at42 days after Sowing

Model	Description	-logL	AIC	BIC
Plant height	Height single trait model	1200.01	2404.506	2430.904
Plant surface Area	Single area trait model	1180.001	2372.312	2399.963
Plant volume	Single volume trait model	1150.451	2314.301	2345.720

Where –logL = negative log likelihood, AIC = Akaike information criterion, BIC = Bayesian information criterion.

The comparison of significant associations for different trait across different growth and developmental stages of plants, as presented Table 8, demonstrates enhanced SNP detection. The results showed that as plants progress in growth and development, as measured in DAS, there is a corresponding increase in the number of significant SNPs-trait associations. For instance, at 11 DAS, there were 6 significant SNPs-Trait associations, at 26 DAS, this number increased to 8 significant SNPs-Trait associations and futher increased 12 SNPs-Trait significant associations at 42 DAS (Table 8). This trend may be attributed to several factors, including changes in gene expression as plants matures [8] and interactions with environmental interactions. The increase number of significant SNPS-Trait associations over time suggests that the genetic effects on these traits become more pronounced as the plants develops. This trend highlights the dynamic nature of genetic influences on plant traits and underscores the importance of employing robust statistical methodologies in genetic studies. As evidenced by studies, the choice of statistical framework can significantly impact the detection of true genetic associations and the reliability of predictions [13,50]. Advanced methodologies, such as mixed models and machine learning techniques, enhance the ability to uncover complex genetic relationships while accounting for confounding factors [39]. By integrating appropriate statistical approaches, researchers can more accurately capture the evolving genetic architecture underlying plant traits, ultimately leading to improved breeding strategies and crop improvement outcomes.

Table 8. Comparison of significant SNPs-traits associations for different trait combinations and at different days after sowing

	Number of significant associations			
Trait combination	11 DAS	26 DAS	42 DAS	Total
Side Area	0	2	4	6
Volume	1	3	4	8
Side height	1	3	4	8

4. Conclusion and Recommendation

4.1. Conclusion

In conclusion, this study effectively enhanced the accuracy of the CMLM for GWAS by averaging predicted biomass from various traits, such as plant volume, side height and side area. The study aimed to deepen the understanding of gene-trait interactions and genetic associations in plant growth and development. By exploring the impact of different trait combinations and days after sowing (DAS) on genetic associations, the study provided valuable insights into the dynamic nature of genetic effects over time. Notably, the results demonstrated a consistent increase in significant SNP associations as plants progressed through different growth stages, highlighting the evolving genetic landscape during plant development. Comparative analysis with existing literature further supported these findings, revealing a similar pattern of increasing genetic associations as plants matured, across various plant species as they mature. Consequently, emphasizing the dynamic interactions between genes and traits expression during plant growth. The utilization of the CMLM in GWAS proved effective for clustering individuals into groups and selecting putative quantitative trait nucleotides (QTNs) based on significance levels, thereby enhancing the efficiency and accuracy of genetic association studies. Furthermore, the exploration of genetic associations using predicted biomass from individual traits provided a comprehensive understanding of the genetic architecture underlying traits such as plant volume, height and side area at different plant growth stages. The study highlighted the importance of reducing noise and ensuring reliable predictions using genomic breeding values, reflecting the evolving genetic landscape and improving genetic association detection over time.

The selection of appropriate statistical methodologies in GWAS is crucial for accurately identifying genetic variants associated with specific traits. The choice of statistical model significantly influences the study's power to detect true associations while minimizing false positives and controlling for confounding factors such as population structure and polygenic background effects. Advanced methodologies, such as the CMLM, offer enhanced efficiency by accommodating multiple markers simultaneously and effectively clustering individuals, thus improving computational feasibility for large datasets. These models also enable more reliable estimation of genetic effects and better handling of complex trait architectures. Moreover, the ability to integrate phenotypic data with robust statistical approaches provided deeper insights into the dynamic nature of genetic associations over time. Therefore, as the field of GWAS continues to evolve, the emphasis on selecting appropriate statistical frameworks will remain vital for advancing our understanding of the genetic determinants of complex traits and improving the accuracy of genomic predictions.

4.2. Recommendations

The study recommends fitting compressed linear mixed models using the predicted biomass derived from image-

derived plant features to enhance the reliability of genetic association analyses.

ACKNOWLEDGMENTS

The authors acknowledge the Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany, for making available the data used in this study. The data was obtained with the support of grants from the German Federal Ministry of Education and Research (BMBF) and performed within the ConFed and DPPN projects (identification numbers: 0315461C and 031A053B). The authors declare no conflicts of interest.

References

[1] Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and prospects of association mapping in plants. *The plant genome*, *1*(1), 5-20.

[2] Yu, J... (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–8.

[3] Falconer, D. S., & Mackay, T. F. C. (1996). Introduction to quantitative genetics (4th ed.). Pearson Education.

[4] Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits. Sinauer Associates.

[5] Smith, K., Brown, D., Lee, S., & Zhang, L. (2019). Enhancing GWAS performance through effective data reduction techniques. Genetic Epidemiology, 43(5), 456-468.

[6] Zhang, L., Chen, S., & Wang, Q. (2017). Genetic basis of biomass production in wheat plants. Plant Genetics Journal, 6(3), 213-226.

[7] Elshire, J., Glaubitz, C., Sun, Q., Poland, A., Kawamoto, K., Buckler, S., & Mitchell, E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, *6*(5), e19379.

[8] Junker, A., Muraya, M., Weigelt-Fischer, K., Arana-Ceballos, F., Klukas, C., Melchinger, E., Meyer, C., Riewe, D., & Altmann, T. (2015). Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. *Frontiers in Plant Science*, 5, 770.

[9] Muraya, M (2016) Dynamic quantitative trait loci and copy number variation: The missing heritability of complex agronomic traits J. Env. Sust. Adv. Res. (2016) 2:13-21.

[10] Sun, N., & Zhao, H., (2020). Statistical Methods in Genome-Wide Association Studies. *Annual Review of Biomedical Data Science*, 3(1), pp.265-288.

[11] Thornton, T., (2015). Statistical Methods for Genome‐Wide and Sequencing Association Studies of Complex Traits in Related Samples. *Current Protocols in Human Genetics*, 84(1).

[12] Bi, W., Kang, G., & Pounds, S., (2018). Statistical selection of biological models for genome-wide association analyses. *Methods*, 145, pp.67-75.

[13] Zhao, H., Li, Y., Chen, J., & Wang, X. (2021). Statistical models for detecting genetic associations: A comparison of methodologies. Theoretical and Applied Genetics, 134(5), 1357-1370.

[14] Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics, 44(7), 821-824.

[15] Listgarten, J., Lippert, C., & Heckerman, D. (2013). FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nat. Genet. 45, 470–1.

[16] Lippert, C... (2013) The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep. 3*, 1815.

[17] Robinson, G. (1991). [That BLUP is a Good Thing: The Estimation of Random Effects]: Rejoinder. *Statistical Science*, 6(1), pp.48-51.

[18] Henderson, C., Kempthorne, O., Searle, S., & von Krosigk, C. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15(2), p.192.

[19] Vilhjálmsson, B., & Nordborg, M. (2012). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, *14*(1), 1-2.

[20] Smith, J., Davis, K., & Lee, T. (2022). Enhancements in kinship modeling: New perspectives and methodologies. *Molecular Ecology*, 31(4), 789-802.

[21] Fang, Y., Liu, S., Dong, Q., Zhang, K., Tian, Z., & Li, X. (2020). Linkage Analysis and Multi-Locus Genome-Wide Association Studies Identify QTNs Controlling Soybean Plant Height. *Frontiers In Plant Science*, *11*.

[22] Lee, Y., Gould, B., & Stinchcombe, J. (2014). Identifying the genes underlying quantitative traits: a rationale for the QTN programme. *Aob PLANTS*, *6*.

[23] Listgarten, J. (2012). Improved linear mixed models for genome-wide association studies. Nat. Methods 9, 525–6.

[24] Zhang, Z., Ersoz, E., Lai, C., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. Nature Genetics, 42(4), 355-360.

[25] Chen, Y., Liu, H., & Zhang, Q. (2021). Challenges and advancements in multiple testing corrections for GWAS. *Frontiers in Genetics*, 12, 620304.

[26] Li, L., Zhang Q., & Huang, D. (2014). A review of imaging techniques for plant phenotyping. Sensors (Basel) 14, 20078-20111.

[27] Brown, A., & Jones, B. (2018). Genomic prediction and heritability in maize: A meta-analysis. Plant Science, 275, 118-127.

[28] Patel, R., Kumar, S., & Li, H. (2023). Non-hierarchical clustering methods in genetic association studies: Opportunities and challenges. *Frontiers in Genetics*, 14, 101234.

[29] Gao, X., Becker, L., Becker, D., Starmer, J., & Province, M. (2009). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, p.n/a-n/a.

[30] Ganal, W., Durstewitz, G., Polley, A., Bérard, A., Buckler, S., Charcosset, A., & Le Paslier, C. (2011). A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PloS one*, *6*(12), e28334.

[31] Gachoki, P., Muraya, M., & Njoroge, G. (2022). Modelling Plant Growth Based on Gompertz, Logistic Curve, Extreme Gradient Boosting and Light Gradient Boosting Models Using High Dimensional Image Derived Maize (Zea mays L.) Phenomic Data. *American Journal of Applied Mathematics and Statistics*, 10(2), 52-64.

[32] Klukas, C., Chen, D., & Pape, M. (2014). Integrated analysis platform: an open-source information system for high-throughput plant phenotyping. *Plant physiology*, *165*(2), 506-518.

[33] Sepaskhah, R., Fahandezh-Saadi, S., & Zand-Parsa, S. (2011). Logistic model application for prediction of maize yield under water and nitrogen management. Agricultural Water Management, 99(1), 51-57.

[34] Xiangxiang, W., Quanjiu, W., Jun, F., Lijun, S., & Xinlei, S. (2014). Logistic model analysis of winter wheat growth on China's Loess Plateau. *Canadian Journal of Plant Science*, *94*(8), 1471-1479.

[35] Liu, H., Wang, J., & Zhang, Z. (2016). A compressed mixed linear model for genome-wide association studies. BMC Bioinformatics, 17, 64.

[36] Kang, H. M., Zeng, Z. B., & Liu, H. (2010). Efficient Control of Population Structure in Mixed Model Association Mapping. Genetics, 185(3), 1001-1014.

[37] Smith, A., & Brown, J. (2019). Marker density distribution in soybean plants. Crop Genetics Review, 7(2), 178-191.

[38] Wang, Q., Li, H., & Zhang, L. (2018). Larger sample sizes uncover more genetic associations in GWAS. *Plant Genetics Journal*, 7(2), 112-125.

[39] Zhang, Z., Lee, S. J. R. M., Zhang, Y. H. M., Chen, R. B. C., & J. M. C. (2020). Genomic prediction of complex traits in plants: A review of the literature and future directions. Crop Science, 60(1), 15-25.

[40] Lee, H., & Wang, Y. (2018). Controlling false positives in GWAS: A comprehensive review. *Statistical Methods in Medical Research*, 27(12), 3546-3563.

[41] Chen, S., Li, M., & Kim, Y. (2020). Genetic relationships between traits at different growth stages in rice plants. Genetics and Plant Biology, 8(2), 156-169.

[42] Henderson, C. R. (1975). Best linear unbiased estimation and

prediction under a selection model. Biometrics, 31(2), 423-447.

[43] Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. Journal of Animal Breeding and Genetics, 124(6), 323-330.

[44] Varona, L., D. R. A. A. González-Camacho, A. M. S. De los Campos, & M. A. S. A. (2018). Prediction error variance in genomic selection: A review. Frontiers in Genetics, 9, 67.

[45] Lee, K., Chen, S., & Wang, Q. (2019). Genetic basis of fruit size traits in tomatoes. *Plant Genetics Journal*, 8(4), 278-291.

[46] Smith, A., Chen, S., & Lee, K. (2016). Genomic prediction accuracy for yield-related traits in wheat. Genetics and Plant Biology, 4(3), 189-202.

[47] Li, H., & Wang, Y. (2021). Advances in genomic selection for plant breeding: Current status and future perspectives. *Theoretical and Applied Genetics*, 134(1), 215-227.

[48] Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika, 76(2), 297-307.

[49] Boulesteix, A. L., Janitza, S., Koehler, M., & Wessling, R. (2018). Consistency of variable selection in high- dimensional settings. Statistical Modelling, 18(2), 145-169.

[50] Xu, Y., & Wu, R. (2022). Statistical methods for genomic prediction in plant breeding: A review. Frontiers in Plant Science, 13, 844649.

[51] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC-19, 716–723.