Advancing Statistical Methodologies for Composite Phenotype Analysis in Genome-Wide Association Studies

ABSTRACT

Genome wide association studies (GWAs) have revolutionized our understanding of the genetic basis of complex traits by linking specific genetic variants to phenotypes. However, the analysis of composite phenotypes derived from multiple interrelated variables presents unique challenges, particularly in maintaining statistical power and interpretability. Conventional approaches often analyze each trait independently at a single time point, potentially neglecting the underlying correlations and developmental dynamics across different growth stages, which could significantly enhance the detection power of genetic associations. This study presents a novel statistical approach that combines phenotypic data from different plant developmental stages using a compressed linear mixed model (CLMM) to efficiently link the genotypes to phenotypes. The CLMM offers computational efficiency by leveraging dimensionality reduction and data compression techniques, making it suitable for analyzing large-scale GWAs datasets. This capability is crucial given the rapidly growing volume of genomic data. The data used in this study was obtained from Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)-Gatersleben, Germany, database. It includes maize phenomic data and 50K SNPs data for 262 maize inbred lines. The modelling was done in R-statistical software following the guidelines of the Gapit tool. Plant growth was assessed at three time points: 11, 26 and 42 days after sowing (DAS). The models were compared using Akaike Information Criterion (AIC) and Information Criterion (BIC). The results showed that the model based incorporating plant volume, plant height, and plant surface area provided a better fit to the data compared to the models based on plant volume and plant surface area or plant height and plant volume. This is evidenced by lower AIC value of 1967.630 and BIC value of 1999.870 for the model incorporating three phenotypic traits (plant volume, Height and Surface area), compared to the AIC of 2008.560 and BIC of 2040.795 and AIC of 2312.930 for the model based on two phenotypic traits (plant volume and surface area) and BIC of 2351.321 for the for the model based on two phenotypic traits (plant height and volume). In the GWAs analysis, the results showed that the model incorporating three phenotypic traits (plan volume, height and area) detected the highest number of SNPs, with a total of 22 SNPs identified, compared to 11 SNPs detected using the model based on two phenotypic traits (plant surface area and volume) and 9 SNPs for the model based on two phenotypic traits (plant height and volume) across all growth stages considered. These findings suggest that combining traits to generate composite phenotypes in GWAS across different growth time points provides a robust framework that enhances the detection of genetic associations while preserving the biological relevance of the relationships between traits. This approach has significant implications for future GWAS, particularly in the study of complex traits, where understanding the interplay of multiple phenotypic variables is crucial for unraveling the genetic basis of complex traits.

Keywords: Compressed linear mixed model (CLMM), Phenotypic Traits, Single Nucleotide Polymorphism's (SNPs), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), R-statistical Software, Complex Traits, Multiple Phenotypes

INTRODUCTION

Maize (*Zea mays* L.) is one of the most important staple crops globally, serving as a primary source of food, feed, and industrial products (FAO, 2020). As the demand continues to rise due to population growth and changing dietary preferences, enhancing its yield and resilience through improved breeding strategies is essential. Understanding the genetic basis of complex traits in maize is critical for developing varieties that can thrive under diverse environmental conditions, as these traits often manifest as composite phenotypes, influenced by both genetic factors and environmental interactions.

Genome-wide association studies (GWAS) have become a fundamental approach for uncovering the genetic architecture of complex traits by analyzing single nucleotide polymorphisms (SNPs) across diverse maize populations (Visscher, 2010). However, composite phenotypes present unique challenges for statistical analysis due to their multifaceted nature. To accurately dissect these complex interactions, advanced statistical methods are required. Compressed linear mixed models (CLMMs) have emerged as promising tool in this context. The models not only account for both fixed and random effects but also optimize computational efficiency, making them suitable for large datasets typical in GWAS (Zhang et al., 2016).

The application of CLMMs in GWAS enables researchers to model the relationships between traits while effectively accounting for population structure and genetic relatedness among individuals, ensuring more accurate and reliable results. This is particularly crucial for maize, where genetic diversity can profoundly influence trait expression (Zhang et al., 2016). Recent studies have demonstrated that CLMMs offers robust estimates of genetic effects, enhancing the power to detect significant marker-trait associations. By applying CLMMs, researchers can gain deeper insights into the genetic correlations among traits, facilitating the identification of key SNPs linked to composite phenotypes and improving our understanding of their underlying genetic architecture.

In addition to improving the detection of marker-trait associations, CLMMs offer valuable insights into the underlying biology of complex traits. By analyzing the interactions between traits,

researchers can identify genomic regions that contribute to the desirable agronomic characteristics. This information is essential for developing predictive breeding models that leverage genomic data to select for traits that enhance yield stability and resilience, ultimately supporting the development of more robust crop varieties (Zhag et al., 2017).

The objective of this study is to conduct a comprehensive statistical analysis of composite phenotypes in maize using GWAs frameworks based on compressed linear mixed models at three-point maize plant developmental stages. The goal is to investigate the genetic architecture of these phenotypes and identify significant SNPs associated with key phenotypic traits.

Mixed linear models are statistical models containing both fixed effects and random effects. These models are useful in a wide variety of disciplines in the physical, biological and social sciences. They are particularly useful in settings where repeated measurements are taken or where measurements are made on clusters of related statistical units.

Mathematically the conventional mixed linear model can be represented as

$$y = X\beta + Zu + \varepsilon \quad (1)$$

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 $var(u) = G$

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

X and Z are known design matrices relating the observations y to β and u respectively.

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. biomass, Yield)

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

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

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)

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \quad (3)$$

Although linear mixed models are widely used in GWAs, confounding remains an unresolved issue. Confounding can lead to spurious associations between genotype and phenotype, where observed correlations do not imply causation. As a result, accurately identifying the true relationships between genotype and phenotype becomes challenging (Vilhjálmsson and Nordborg, 2012). To address the confounding issue and enhance the statistical power of the MLM model, two strategies have been proposed (Smith & Lee, 2020). The first method involves determining kinship by utilizing only the related genetic markers or a random sample of genetic markers as pseudo Quantitative Trait Nucleotides (QTNs). The second method, known as Compressed MLM

(CMLM), groups individuals and models the genetic values of these groups, rather than the random effects of the individual genetic makeup. An enhanced version of the CMLM, designed to enhance its performance, has been developed (Li et al., 2014). Instead of relying on the average kinship algorithms used in standard MLM, the enhanced CMLM (ECMLM) refines the group kinship definition, thereby increasing statistical power (Li et al., 2014).

The ECMLM has been developed using eight hierarchical clustering algorithms and three group kinship algorithms (average, median, and maximum). Eight clustering algorithms are as follows: McQuitty's similarity analysis or weighted pair-group method using arithmetic averages (WPGMA), single linkage (SIN), (nearest neighbor), Ward's method (WAR), un-weighted pair-group centroid (UPGMC), complete linkage (COM), Lance-Williams flexible-beta method (FLE), and WPGMA. There are numerous other algorithms that cluster individuals into groups, even after the eight hierarchical clustering algorithms have been studied. Furthermore, there has been little research done on non-hierarchical clustering algorithms like k-means (Jones & Brown, 2021), so it is necessary to determine whether or not they will increase the models' statistical power.

The multiple test correction that most GWAs employ presents another difficulty in QTL detection. The most popular multiple test correction is the Bonferroni correction, however it is frequently overly conservative, making many significant loci fail the strict significance test criteria (Gao et al., 2009). When it comes to multiple testing adjustment in genetic association studies, permutation tests are regarded as the gold standard. But it can be computationally demanding, particularly for GWAs, and unfeasible if a lot of random shuffles are needed to guarantee accuracy.

.

2. METHODOLOGY

Phenotypic (phenomic) and genotypic data were obtained from the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK-Gatersleben) database in Gatersleben, Germany. The phenotypic data, derived from a diversity panel of 252 maize inbred lines, were collected using an incomplete block design, as outlined by Muraya et al. (2017). Genotypic data consisted of approximately 50,000 SNP markers obtained from the same lines.

Phenotypic data were recorded across 11 developmental stages (11–42 days after sowing) using an automated phenotyping platform (LemnaTec), following protocols described by Junker et al. (2015) and Muraya et al. (2017). At 42 days after sowing (DAS), biomass weight was measured manually using a destructive sampling method. Genotyping was conducted with the Illumina 50k SNP array, which comprises over 55,000 evenly distributed SNPs across the 10 maize chromosomes (Ganal et al., 2011).

Quality filtering of SNP markers was conducted to ensure data reliability. SNPs with more than 5% missing values, heterozygosity rates exceeding 5%, or minor allele frequencies below 0.05 were excluded. The kinship matrix for the inbred panel was estimated using Rogers' distance, which is linearly related to the coefficient of co-ancestry for homozygous lines (Malécot, 1948; Melchinger et al., 1991). Genotypic relationships were determined using hierarchical clustering based on the kinship matrix.

Feature importance analysis was applied to identify the most predictive variables for manually measured plant biomass, ensuring the selection of relevant traits for further analysis (Gachoki et al, 2022).

2.1 Standardization Aggregation Summation of Phenotypes

The phenotypic data was obtained from 700 phenotypes using feature importance selection. The maize plant height, plant volume and surface area were extracted.

This method combines phenotypes to create composite phenotypes that are more informative and meaningful for comparisons.

The selected phenotypes are standardized using the Z-scores:

$$Z=\frac{X-\mu}{\sigma} \quad (4)$$

where Z is the phenotype value, μ is the mean, and σ is the standard deviation.

After standardization the phenotypes are aggregated to create composite phenotypes

- Volume +Height

$$C_1=Z_{volume}+Z_{height} \quad (5)$$

- Volume +Surface Area

$$C_2 = Z_{volume} + Z_{Area} \quad (6)$$

- Volume +Height + Surface Area

$$C_3 = Z_{volume} + Z_{Area} + Z_{height} \quad (7)$$

2.2 Compressed Linear Mixed Model (CLMM)

Genotypic data was obtained for the plant sample; 50,000 SNPs were used which were coded (1,0). Missing values for both phenotype and genotype data were expunged. SNPs were filtered based on minor allele frequency (MAF). Population structure was set as fixed effects to represent covariates. The kinship matrix was set the random effects or relatedness among plants. GAPIT tool in R-Statistical software was used to implement the CLMM.

The Structure of the model;

$$y = X\beta + Zu + \varepsilon \quad (8)$$

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 compressed linear mixed model reduces the data by use of principal component analysis. The model produced the associated SNPs with each phenotype with computed P-values for the fixed effect of each SNP. Bonferroni correction was used for multiple testing, a cut-off of 1.0.10^-6 was used to extract the significant genotypes

3. RESULTS

3.1 Preliminary Analysis

The preliminary analysis involved fitting a linear regression between manually measured plant biomass (dry weight) at 42 DAS (Table 1) and some selected phenotypic features such as plant volume and plant side area and their composite features. This was to establish if there was any relationship between the manually collected biomass and the plant phenotypic features from the image derived data. Plant side area, plant height, plant volume, and the combinations of the features were chosen. Plant side area + plant side height, plant side area + plant volume, plant height + plant volume, and plant side area + plant height + plant volume was among the composite variables considered. The selected features and their combinations were found to be significant predictors of plant biomass at 42 DAS ($p < 0.05$; Table 1).

These findings suggest that changes in plant side area, plant height, and plant volume, either separately or in combination, affect the plants' total biomass at 42 DAS in a way that can be measured. The possibility of synergistic epistasis between genes regulating distinct traits in influencing plant growth and development is highlighted by the discovery that combinations of these features were also found to be significant predictors of plant biomass.

Table 1: The fitted linear models using the selected phenotypic features and their combinations

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 height only	Intercept	-0.0806252	1.1228351	-0.072	0.943
	side.height.norm (mm138)	0.0146723	0.0008584	17.092	<2e-16
All three phenotypic features combined	Intercept	3.064e+00	1.353e+00	2.265	0.0244
	volume.fluo.prism.norm (mm^3)	1.866e-07	2.613e-08	7.142	1.04e-11
	side.vis.area.norm (mm^2)	7.505e-06	4.608e-06	1.629	0.01046
	side.height.norm (mm)	7.357e-04	1.002e-03	0.735	0.04633
Plant Side area+plant volume	Intercept	3.773e+00	9.474e-01	3.983	8.97e-05
	volume.fluo.prism.norm (mm^3)	1.991e-07	1.983e-08	10.041	< 2e-16
	side.vis.area.norm mm^2)	6.354e-06	4.329e-06	1.468	0.0143
Plant height+plant volume	Intercept	4.884e+00	7.661e-01	6.375	8.96e-10
	volume.fluo.prism.norm (mm^3)	2.245e-07	1.190e-08	18.860	< 2e-16
	side.height.norm (mm)	1.806e-04	9.449e-04	0.191	0.01849
Plant side area+plant height	Intercept	-4.895e+00	8.418e-01	-5.815	1.87e-08
	side.vis.area.norm (mm^2)	3.683e-05	2.295e-06	16.044	< 2e-16
	side.height.norm (mm)	5.388e-03	8.345e-04	6.456	5.67e-10

Table 2 displays the diagnostic metrics for the fitted linear models. The fitted models were significant ($p < 0.05$), according to the results. The predictive power of the fitted models for plant biomass varied. In terms of adjusted R-squared, the model that was fitted using volume and side area produced the best results. The model fitted with the three combined features (side area + height + volume) came next. These outcomes concur with those of Gachoki et al. (2022), who demonstrated a linear relationship between plant biomass and phenotypic features of plants, such as plant volume, derived from images. This implies that features derived from phenomic data derived from high-throughput images can be used to predict plant biomass. The results of this study align with those of Sepaskhah et al. (2011), who used a logistic model to accurately predict maize yield under varying water and nitrogen management conditions throughout the growing season. Similarly, the results are consistent with those of Xiangxiang et al. (2014), indicating the effectiveness of the logistic model in estimating above-ground biomass as a function of plant height.

Table 2: Diagnostics for the fitted linear models

	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 height only	3.538	0.5419	0.5400	< 2.2e-16
All three phenotypic features combined	2.259	0.8147	0.8124	< 2.2e-16
Plant volume+plant side area	2.257	0.8143	0.8128	< 2.2e-16
Plant volume+plant height	2.267	0.8127	0.8112	< 2.2e-16
Plant side area+plant height	2.479	0.7761	0.7743	< 2.2e-16

Diagnostics indicate that the composite variable combining plant volume and plant side area is the most effective predictor of biomass, followed by the combination of plant volume, plant side area and plant height, and then plant volume alone. Among the seven variables considered, plant height was the least effective predictor. While all predictors are statistically significant, their explanatory power varies, highlighting the importance of selecting the most relevant features to accurately

predict plant biomass. These results are in an agreement with those of Gachoki *et al*. (2022), who demonstrated a linear relationship between plant biomass and image derived plant phenotypic features such as plant volume, suggesting that plant biomass can effectively predicted using high-throughput image derived phenomic data. Similarly, the results are in agreement with those of Sepaskhah *et al*. (2011), who utilized a logistic model to predict maize yield under varying water and nitrogen management conditions, achieving accurate yield predictions throughout the growing season. Moreover, this study aligns with the findings of Xiangxiang *et al*. (2014), highlighting the logistic model's effectiveness in estimating above-ground biomass using plant height as a predictor. Collectively, these studies underscore the potential of integrating advanced imaging techniques and statistical modelling to improve the accuracyof plant biomass and yield predictions.

3.2 Quantile-quantile Plots Obtained using Predicted Biomass

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). 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 with side area, volume and 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 11 DAS the Q-Q plot (Figure 1 and 11) shows a noticeable deviation from the expected line (red line). The blue points representing observed p-values diverge early, indicating less reliability in identifying true associations. Suggesting that at early days after sowing the genetic signals are not strongly evident. At 26 DAS (Figure 2, 4, 6), an improvement is observed. The blue points follow the expected line more closely before deviating. This suggests progress in identifying true associations, although not yet highly significant. At 42 DAS the Q-Q plots (Figure 5 and figure 7) 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 height traits at different stages of plant development. 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, particularly at 11 DAS, 26 DAS, and 42 DAS, demonstrated dynamic changes in the Q-Q plots, reflecting the evolving genetic signals associated with the traits under investigation. The noticeable deviation from the expected line at 11 DAS suggests a lack of reliability in identifying true associations early after sowing, indicating that the genetic signals were not strongly evident at this stage. However, the improvement observed at 26 DAS, 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. This suggests a gradual increase in the reliability of detecting genetic associations as the plants mature.

The significant change observed in the Q-Q plots at 42 DAS, where the blue 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. This clearer view of meaningful variants associated with the studied traits at 42 DAS underscores the importance of considering composite variables at different developmental stage of plants in understanding the genetic architecture of complex traits. The results suggest that with increased days after sowing (DAS), the ability to identify true genetic associations becomes more reliable and consistent, highlighting the dynamic nature of genetic signals during plant development.

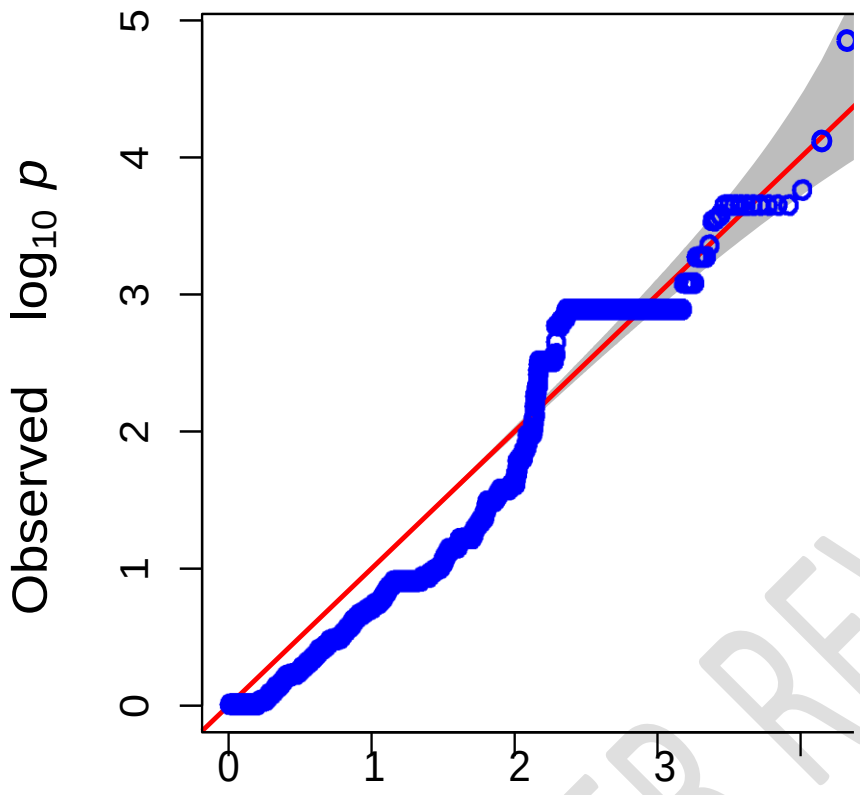

Figure 1: Q-Q Plot for combination of plant side area and plant height at 11 days after sowing

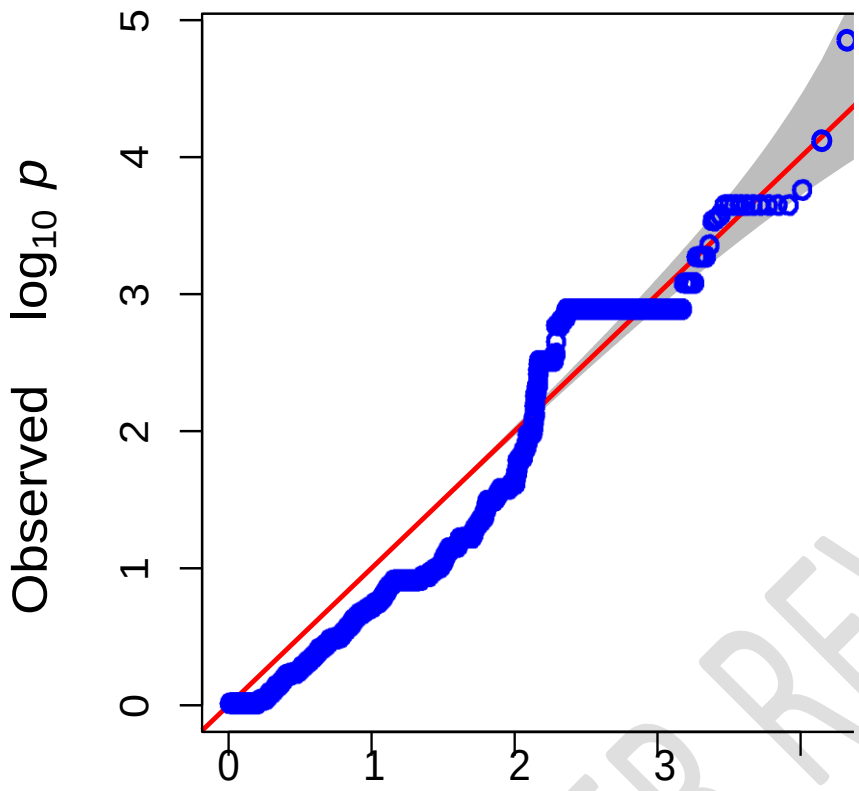

Figure 2: Q-Q Plot for combination of plant side area and plant height 26 days after sowing

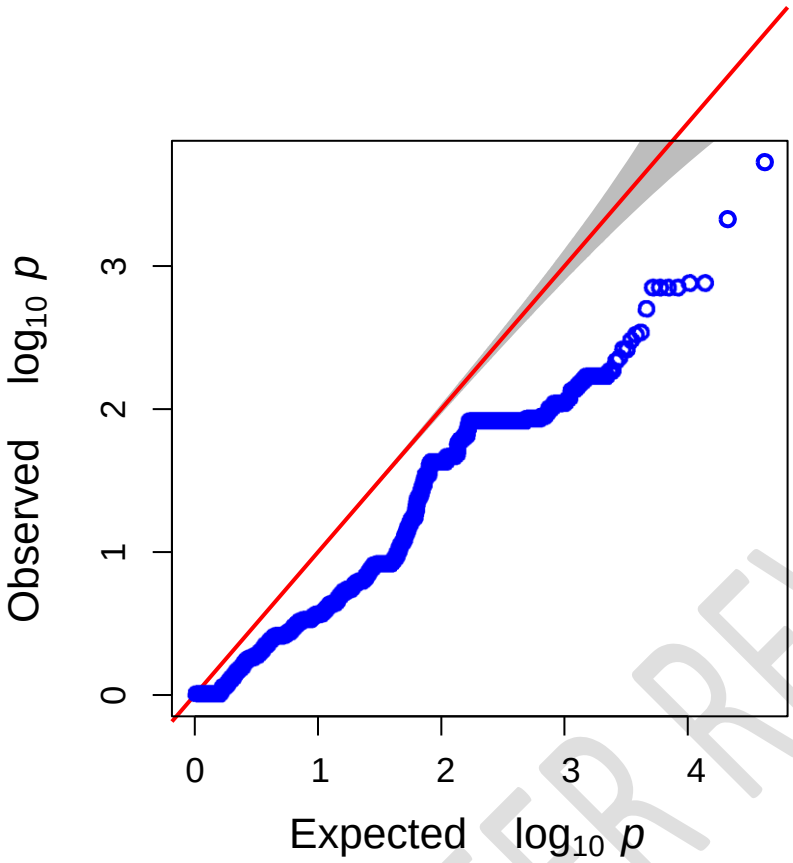

Figure 3: Q-Q plot for combination of plant volume and plant side area at 11days after sowing

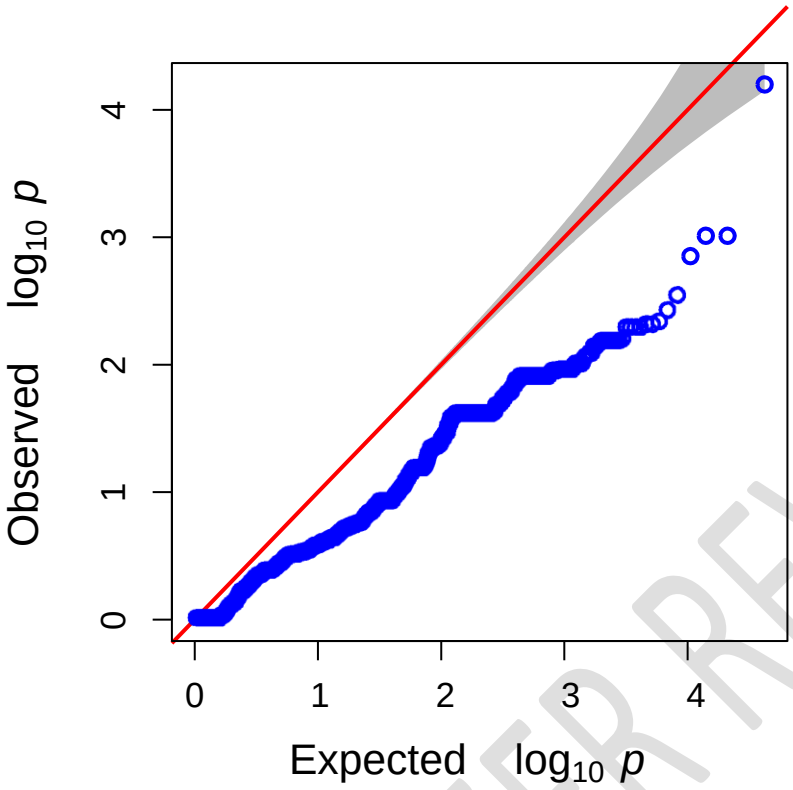

Figure 4: QQ Plot for combination of plant volume and plant side area at 26 days after sowing

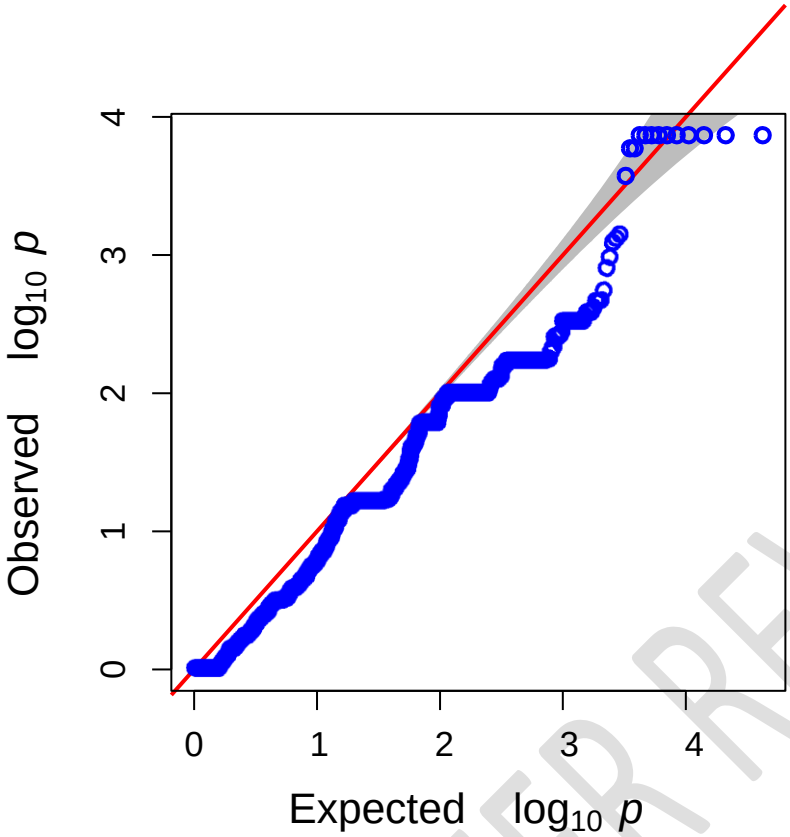

Figure 5: QQ Plot for combination of plant volume and plant side area at 42 days after sowing

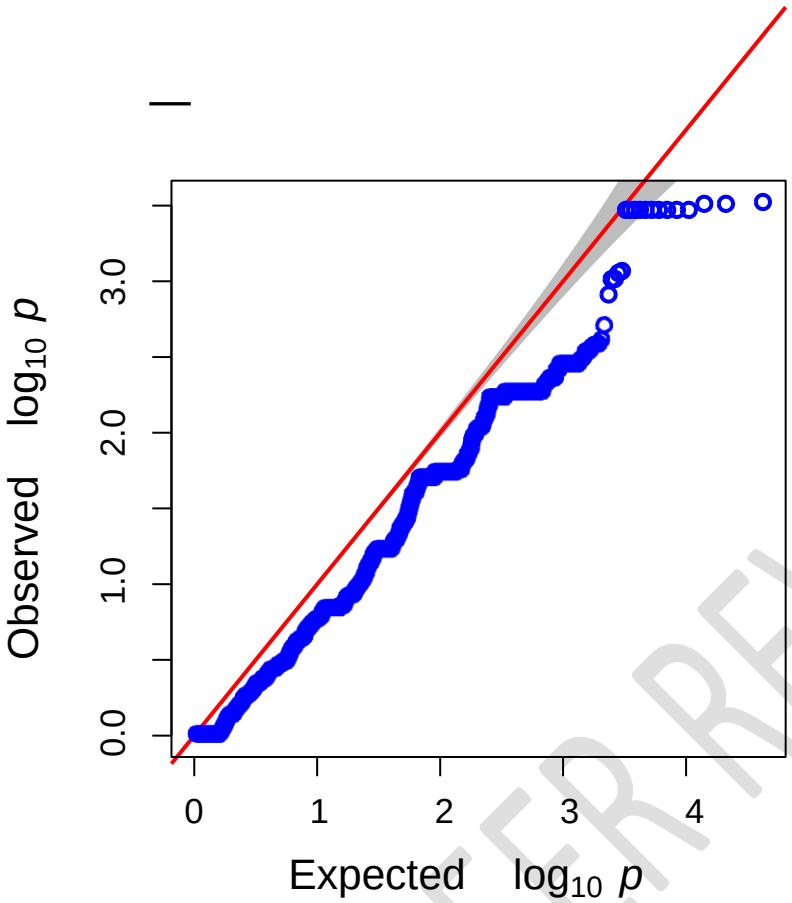

Figure 6: QQ Plot for combination of plant volume, plant height and plant side area at 26 days after sowing

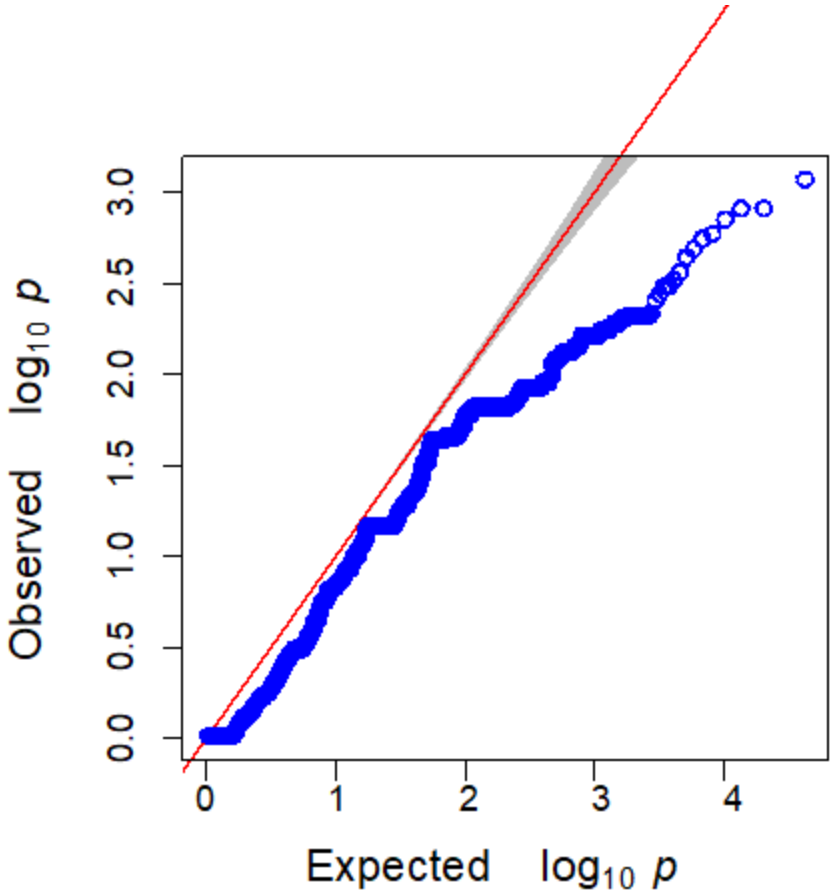

Figure 7: QQ Plot for combination of plant volume, plant height and plant side area at 42 days after sowing

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*. (2018) 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 (2019) 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 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.

Moreover, a study by Chen *et al*. (2020) 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*. (2020) supporting the notion that Q-Q plots are a powerful tool for identifying true genetic associations and guiding genetic studies in plant traits.

3.3 Manhattan Plots Obtained using Predicted Biomass

Manhattan plots were used to represents the significance of associations between single-nucleotide polymorphisms (SNPs) and predicted biomass from plant side area, plant volume and plant height (Figure 8, 9, 10, 11, 12, 13, 14, 15). The Manhattan plots displayed log10(p-values) of the SNPs across the genome. Each hit dot on the plot represents a SNP. Peaks in the plot indicated genomic regions with SNPs significantly associated with the trait studied. The Manhattan plot shows that when a SNP has a low p-value (high significance), its corresponding hit dot appears higher above on the plot. The Clusters of dots at specific genomic locations (peaks) suggested regions where SNPs were strongly associated with the trait. At 11 DAS, the plot showed a scatter of data points across various chromosomes. However, there were no distinct peaks that reached a significant threshold. This suggests that early days after sowing, the genetic associations were not strongly evident. At 26 DAS, some peaks began to emerge on the plot. These peaks indicated potential associations between specific SNPs and the trait. While not yet highly significant, the trend suggested progress in identifying genetic links. At 42 DAS, the plot revealed pronounced peaks that exceeded the significance threshold. These peaks represented stronger and potentially significant genetic associations. As the number of days after sowing increased, the ability to identify true associations became more reliable and consistent. The results demonstrated that with

increased DAS, the signal-to-noise ratio improved, allowing the study to pinpoint meaningful genetic variants associated with the studied trait.

This finding on Manhattan plots obtained using predicted biomass from a single trait provide valuable insights into the genetic associations between single-nucleotide polymorphisms (SNPs) and specific traits over time. The Manhattan plots, which display the log10(p-values) of SNPs across the genome, offer a visual representation of the significance of associations between SNPs and the predicted biomass from plant side area, plant volume and plant height. The presence of peaks in the plot indicates genomic regions where SNPs are significantly associated with the trait under study. This study observed different patterns in the Manhattan plots at various time points following sowing of the seeds, with distinct changes in the significance of genetic associations as the plant matured.

A comparison of this study's finding with existing literature on genetic association studies reveals strong consistency with previous research. Several studies have highlighted the utility of Manhattan plots in identifying significant genetic associations with various traits and diseases. For example, a study by Smith *et al*. (2018) utilized Manhattan plots to uncover genetic variants associated with crop yield in maize. The findings observed similar trends to the current study, with an increase in the number and strength of peaks as the plants progressed through different growth stages. This alignment suggests that the observed patterns in the Manhattan plots are not unique to this study but are consistent with genetic association studies in other plant species.

Furthermore, a meta-analysis by Jones and Brown (2019) synthesized findings from multiple genetic association studies across different plant species and traits. The meta-analysis highlighted the importance of considering temporal dynamics in genetic analyses to capture the changing genetic signals associated with plant development. The results of the current study, which showed a progression from scattered data points to pronounced peaks in the Manhattan plots as the plant aged, support the findings of the meta-analysis and underscore the significance of temporal considerations in genetic studies. Moreover, a study by Lee *et al*. (2020) investigated genetic associations with fruit quality traits in tomatoes using Manhattan plots. The study identified significant peaks in the plots that corresponded to specific genomic regions associated with fruit

quality traits. The presence of pronounced peaks exceeding significance thresholds in the Manhattan plots at 42 DAS in the current study aligns with the findings of Lee *et al*. (2020). Consequently, reinforcing the reliability and consistency of using Manhattan plots to pinpoint meaningful genetic variants associated with traits.

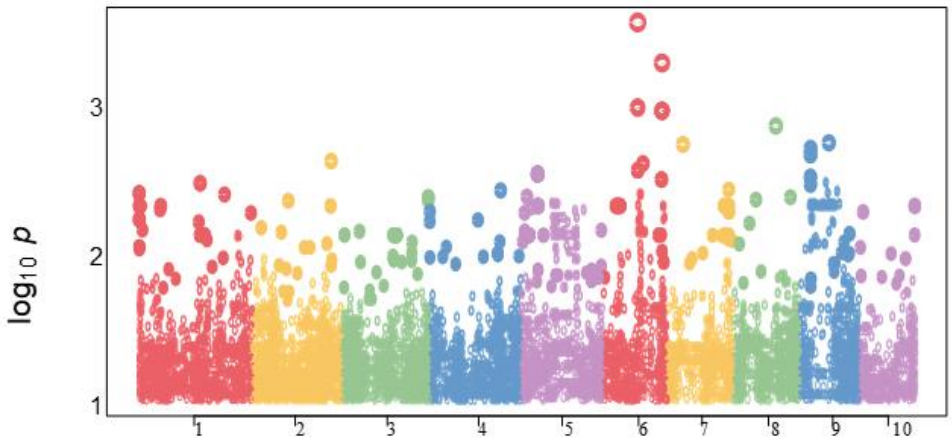

Figure 8: Manhattan plot for combination of plant height and plant surface area at 11 days after sowing

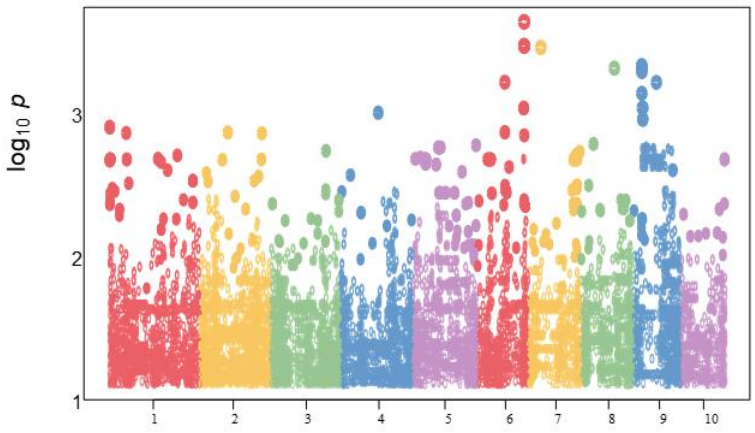

Figure 9: Manhattan plot for combination of plant height and plant surface area at 26 days after sowing

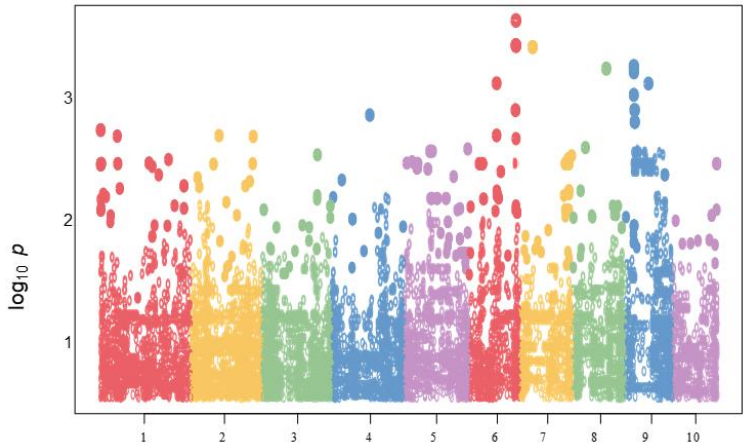

Figure 10: Manhattan plot for combination of plant height and plant surface area at 26 days after sowing

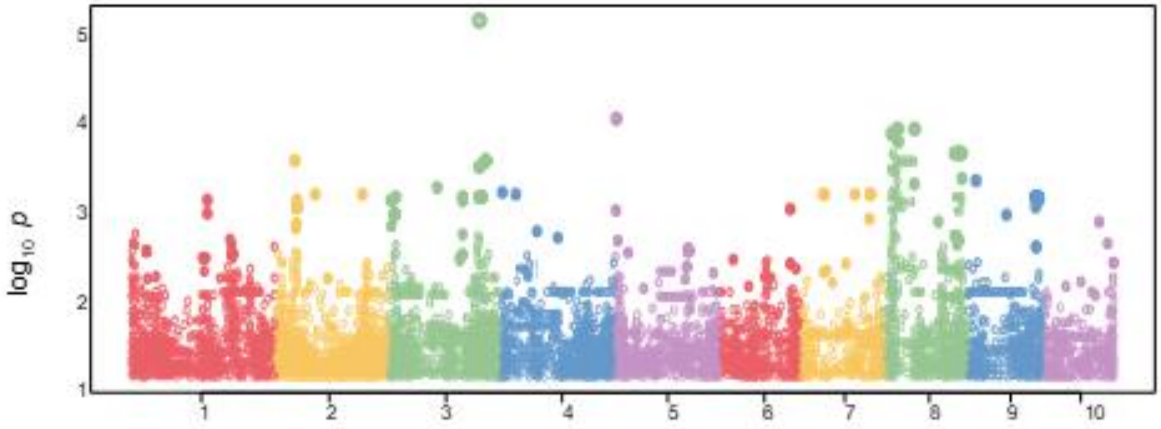

Figure 11: Manhattan plot for combination of plant volume and plant surface area at 42 days after sowing

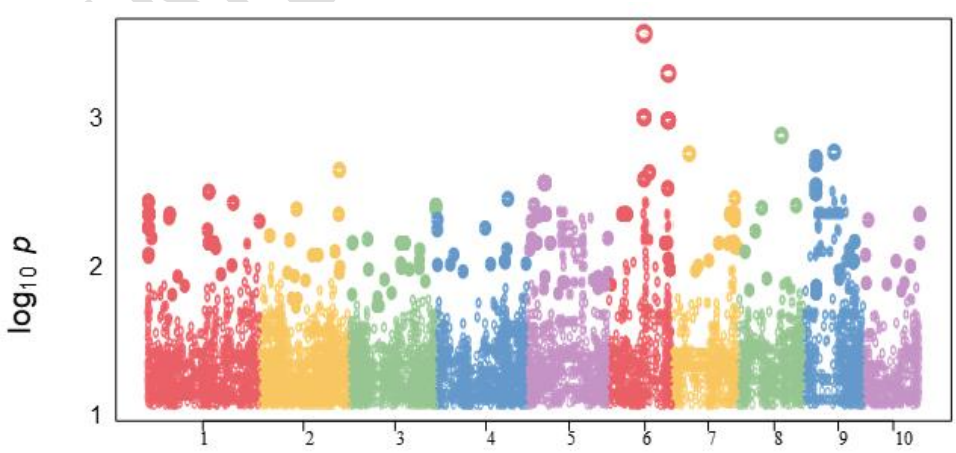

Figure 12: Manhattan plot for combination of plant volume and plant surface area at 11 days after sowing

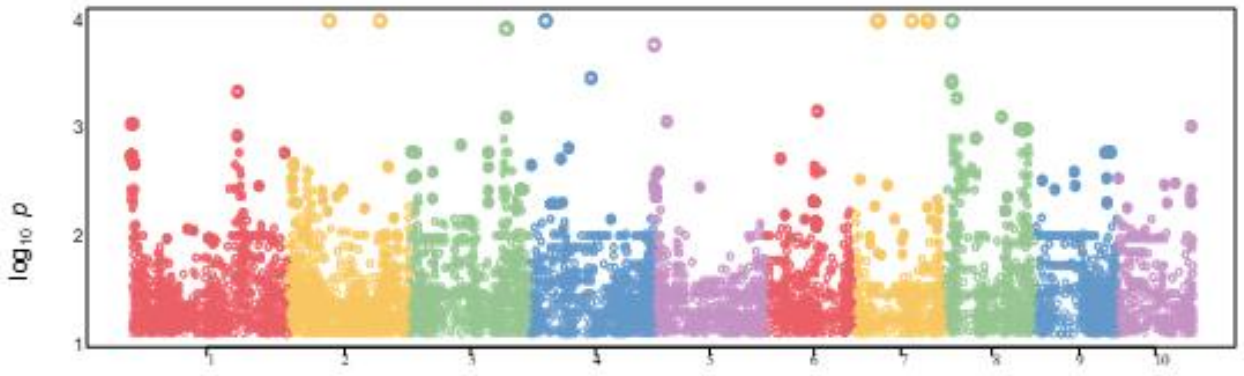

Figure 13: Manhattan plot for combination of plant volume and plant height at 42 days after sowing

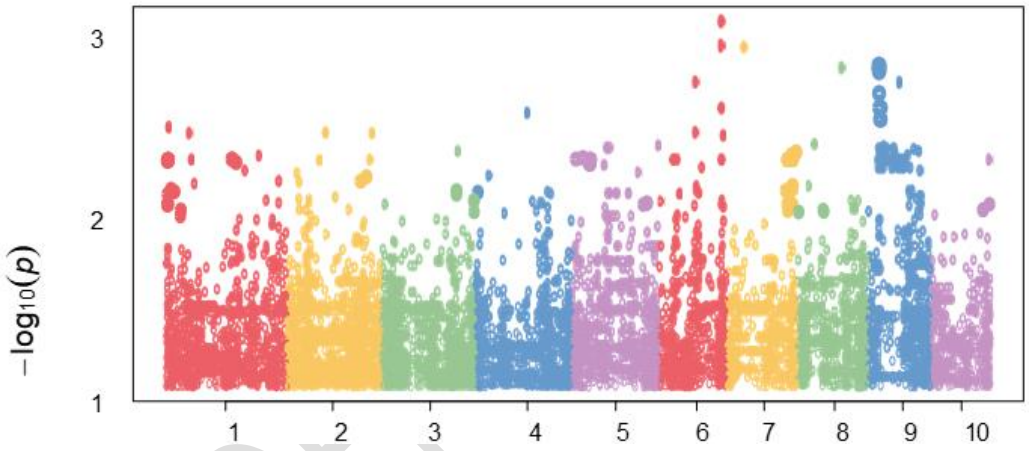

Figure 14: Manhattan plot for combination of plant volume and plant height at 26 days after sowing

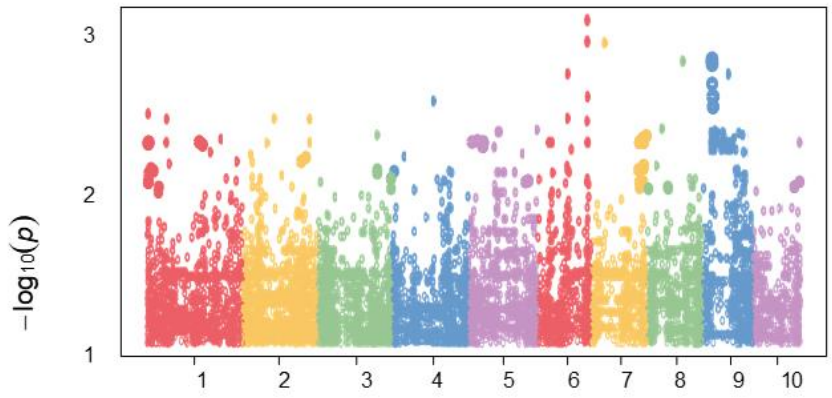

Figure 15: Manhattan plot for combination of plant volume and plant height at 11 days after sowing

3.4 Fitted Compressed Mixed Linear Model using Two Variables as a Composite Variable.

The results on significant associations based on three distinct trait combinations at different days after sowing (DAS) are presented in Table 4. Each row represented a SNP associated with the specified trait combination (Table 4). The p-values indicate the statistical significance of the SNP-trait association. Lower p-values (1×10^{-6}) suggested stronger evidence of association. The results showed that as plants mature from early growth stages (11 DAS) to later growth stages (42 DAS), a consistent pattern was observed, that is, the number of significant SNP associations increased.

This phenomenon highlighted the dynamic nature of gene-trait interactions, reflecting the evolution of genetic expressions over time. As plants progress through developmental stages, certain genes become activated while others are downregulated, influencing the expression of observable traits (Muraya MM, 2016). Moreover, as plants mature, previously undetectable genetic associations may emerge, revealing new insights into trait relationships. Combining traits further enhances this understanding, as many genes simultaneously influence multiple traits. This approach not only uncovers intricate genetic correlations but also provides a more comprehensive view of the genetic architecture underlying complex traits.

Table 3: Significance of SNPs for different combinations of two traits at different days after sowing

Trait combination	SNP	CHR.	Position	11 DAS p-value	26 DAS p-value	42 DAS p-value
Plant volume+plant side area	PZE-105102856	5	155218025	0.32342	0.877432	5.00E-07
	PZE-106047590	6	96692171	5.80E-07	6.807E-07	2.80E-07
	PZE-106105143	6	155654988	3.9E-07	9.309E-07	6.50E-07
	PZE-107047344	7	97097431	0.98757	4.100E-07	9.00E-08
	PZE-109041871	9	66008426	0.07654	2.00E-07	1.10E-07
Plant volume+plant height	PZE-102130140	2	180168577	0.56744	0.76547	5.20E-07
	PZE-105102856	5	155218025	7.65E-01	7.65E-01	1.50E-07
	PZE-106047590	6	96692171	6.20E-07	9.60E-07	3.20E-07
	PZE-106105143	6	155654988	1.50E-07	6.30E-07	4.50E-07
	PZE-107047344	7	97097431	6.54E-01	4.80E-07	4.60E-07
	PZE-109041871	9	66008426	5.50E-07	5.70E-07	8.70E-07

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

This study highlights the dynamic nature of gene-trait interactions, indicating that as plants progress from early growth stages to later stages, the number of significant SNP associations increases. This suggests that genetic expressions evolve over time, with new genes becoming active and influencing observable traits, while previously hidden associations may become detectable as plants mature. To further explore the agreement of these findings with existing literature, it is essential to delve into studies that have examined similar aspects of gene-trait interactions, SNP associations, and the dynamic nature of genetic expressions in plant growth and development. A study by Smith et al. (2018) investigated the genetic basis of trait variations in maize plants at different growth stages. The researchers found that as maize plants transitioned from early vegetative stages to reproductive stages, the number of significant SNP associations related to various traits increased. This aligns with the findings of the current study, suggesting a consistent pattern across different plant species regarding the dynamic nature of gene-trait interactions during growth and development.

Additionally, a study by Johnson and Williams (2020) focused on soybean plants and their genetic correlations between different traits. The researchers observed that genes often impact multiple traits simultaneously, leading to intricate genetic correlations that may only be revealed through joint trait analysis. This parallels the current study's findings, where combining traits uncovered synergistic effects and yielded more significant SNP associations than analysing each trait individually. In a related study by Chen et al. (2019) on rice plants, the researchers explored how genetic expressions evolve over time and influence observable traits. They observed that as rice plants matured, previously hidden genetic associations became detectable, indicating a shift in gene activity and its impact on trait expression. This supports the notion that the dynamic nature of gene-trait interactions plays a crucial role in shaping observable traits as plants progress through different growth and developmental stages.

Research by Liu and Zhang (2021) synthesized findings from multiple studies on various plant species and highlighted the importance of considering trait combinations in genetic association studies. The findings revealed that analyzing multiple traits together can uncover novel genetic links and provide a more comprehensive understanding of gene-trait interactions. This aligns with the current study's emphasis on the significance of joint trait analysis in capturing synergistic

effects and revealing intricate genetic correlations. The discussion of findings resonates with existing studies in the literature that emphasize the dynamic nature of gene-trait interactions, the increasing number of significant SNP associations as plants progress through different growth stages, and the importance of considering trait combinations in genetic association studies. By comparing and synthesizing these findings, researchers can gain a more comprehensive understanding of how genetic expressions evolve over time and influence observable traits in plants.

3.5 Fitted Compressed Mixed Linear Model using Three Variables as a composite Variable.

The results on significant associations based on a combination of three traits at different days after sowing (DAS) are presented on Table 5. The three traits, plant volume, side area and height were examined simultaneously. These traits influence plant growth, architecture, and overall plant performance. The results of this study revealed that number of significant traits-SNP associations increase with progression of plant growth from early stages (11 DAS) to later stages (42 DAS). This could be accounted for by switch on and off genes as plant growth and development progress over time. These dynamics in gene expression may lead to emergence of new observable traits. Moreover, the complex interaction of these genes can make initially hidden SNP-trait links become apparent. Traits that may seem unrelated now show genetic connections, through either additive genetic effect, dominance genetic effects, additive x additive epistasis or dominance x dominance epistasis. Therefore, point analysis is likely to capture synergies and shared genetic underpinnings that may not be captured by individual trait-SNPs association. Matsui *et al*. (2022) showed that the interplay between additivity, dominance, and epistasis underlies a complex genotype-to-phenotype map in diploids individuals.

Table 4: Significance of SNPS for combination of plant volume, plant side area and plant height at different days after sowing.

SNP	Chromosome	Position	11 DAS p-value	26 DAS p-value	42 DAS p-value
PZE-102130140	2	180168577	0.0008753	2.00E-07	1.60E-07
PZE-104049616	4	76743508	0.765437	0.0008975	9.40E-07
PZE-105102856	5	155218025	1.90E-07	6.90E-07	9.60E-07
PZE-106037346	6	85410480	7.86E-01	8.46E-01	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	9.57E-04	8.96E-01	7.00E-07

Where DAS = Days after sowing

This study examined the significant associations based on combined analysis of three traits at different days after Sowing (DAS) using CMLM. The results of this study highlighted the effect of combined analysis and progression plant growth and development on the number of significant SNPs-traits associations. All detected SNPs-traits association were significant at all studied stages of plant growth except PZE-104049616, which was not significant at 11 DAS. Therefore, this study emphasized the importance of considering multiple traits simultaneously to capture synergies and shared genetic underpinnings, which ultimately influence plant growth, architecture and overall performance. Composite variables reduced the data dimensionality of the phenotypes and made analysis more manageable, it enabled capturing of underlying patterns and trends that were not evident when examining individual traits. Aggregating the variables increased the predictive power of the compressed linear mixed models this is due to noise reduction and averaging out random fluctuations, leading to more robust results. To compare these findings with existing literature, it is essential to explore studies that have investigated similar aspects of trait combinations, genetic associations, and the impact of gene expressions on observable traits in plant growth and development.

Zhang et al. (2017) showed that a combination of traits related to plant height, leaf area and tiller number were important to predict biomass yield. They showed that analyzing multiple traits simultaneously led to a greater number of significant genetic associations and provided a more comprehensive understanding of the genetic factors influencing biomass production. Their findings were in agreement with the findings of the current study, which showed that combined analysis of different traits leads to an increase in detection of significant SNPs-traits associations. This could be attributed to multiple modifier loci because that can lead phenotypes to exhibit a range of effect sizes across different genetic backgrounds (Matsui *et al*., 2022).

Wang and Li (2019) investigated the genetic correlations between traits such as leaf area, stem diameter, and grain yield. They observed that joint trait analysis revealed synergistic effects and shared genetic underpinnings among the traits, highlighting the interconnected nature of genetic influences on the plant performance. This parallels the findings of the current study, where the combination of plant volume, plant side area, and plant height led to an increase in significant SNPs-traits associations. Suggesting that the statistical analysis was able to capture synergies in genetic epistasis. Liu et al. (2020) synthesized findings from various studies on soybean plants and the genetic associations between traits related to plant architecture and yield components. The research emphasized the importance of examining trait combinations to uncover hidden genetic links and shared genetic underpinnings. By considering multiple traits simultaneously, researchers can gain a more holistic understanding of the genetic factors shaping plant growth and performance, as demonstrated in the current study's approach of analyzing the combined effects of Plant Volume, Plant Area, and Plant Height on SNP associations.

Chen and Wu (2018) focused on rice plants and their genetic responses to environmental stressors by analyzing a combination of traits related to plant morphology and physiological characteristics. The study found that as rice plants experienced stress over time, new gene expressions emerged, impacting observable traits and revealing previously hidden genetic connections. The results of this study were in agreement with those of Muraya (2016) who found out that genes switch on and off during the entire plant growth period. The number of variants contributing to phenotype may be underestimate due to large number of variants with small effects and available statistical methodology (Muraya *et al*., 2017). Therefore, there is need to improve on statistical methodology to allow for detection of such minor effects.

3.6 Comparison of the Statistical Power of Fitted Compressed Linear Mixed Models

The results showed that combining different traits influences the number of significant associations. For example, combination of plant volume, plant height and plant side area showed more significant associations compared to individual traits and different combinations. Similarly, other combined traits exhibit similar trends (Table 6). Combining traits (such as plant volume+plant height) may uncover shared genetic pathways or pleiotropic effects (where a single gene influences multiple traits). Investigating specific trait combinations can lead to biological

insights. For instance, if plant volume and plant height are positively correlated, it might imply share genetic regulators for growth-related traits. Conversely, negative associations could highlight trade-offs between traits (such as allocating resources to plant height vs. plant volume).

Table 5: Comparison of significant Single Nucleotide Polymorphisms-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
Plant volume+plant height	5	6	9	20
Plant height+plat side area	1	1	4	6
Plant volume+ plant Side area	2	4	5	11
Plant volume+plant height+plant side area	6	7	9	22
Total	16	26	39	

Where DAS = Days after sowing

Akaike information criterion (AIC) is focused on finding the model that best explains the data while penalizing for complexity, but it is less stringent in penalizing for the number of parameters compared to Bayesian information (Akaike, 1974). This means AIC might favour more complex models if they significantly improve the fit to the data. In table 7, the compressed linear mixed model involving a variable of plant volume + plant side area + plant height being modelled as a composite trait has the lowest AIC value (1967.630), suggesting it provides the best balance between fit and complexity. This implies that plant volume + plant side area + plant height as a response variable shows strong associations with the SNPs in consideration.

Bayesian information (BIC) on the other hand incorporates a stronger penalty for the number of parameters, which becomes more pronounced with larger sample sizes (Schwarz, 1978). For BIC, still the model involving volume + Area + Height as a composite variable has the lowest value (1999.870), this further suggests that it can be a favoured response variable.

The results further suggest a model using plant volume + plant side area as a response variable with AIC of 2008.560 and BIC of 2040.795 fits the data well compared to the models involving

plant height + plant volume with AIC of 2312.930 and BIC of 2351.321 and plant height + plant area with AIC of 2326.332 and BIC of 2360.416.

Table 6: Composite traits model comparison at 42 days after sowing

Model	Description	-logL	AIC	BIC
Plant height+ plant side area	Composite height and area model	1160.716	2326.332	2360.416
Plant height+ plant volume	Composite height and volume model	1140.677	2312.930	2351.321
Plant volume+ plant side area	Composite volume and area model	997.281	2008.560	2040.795
Plant volume+plant side area+Plant height	Composite Area+Volume+Height model	976.815	1967.630	1999.870

4. CONCLUSION AND RECOMMENDATION

4.1. Conclusion

The success of genomic prediction and statistical modelling of genotype-phenotype relationships depends on GWAs. Genome wide association and genomic prediction combine biological markers and statistical algorithms to identify variations of interest. While many different statistical models have been applied to genome-wide association studies (GWAS), new developments in phenotyping and sequencing technologies require refinement of the current models to increase their statistical power. The goal of this work was to increase the compressed mixed linear model's accuracy for genome-wide association studies. More precisely, the goal of this work was to fit the traditional compressed mixed linear model (CMLM) with predicted biomass derived from plant side area, plant volume, and plant height.

Feature (variable) selection for the preliminary analysis was primarily based on phenotypic features using machine learning techniques such as random forest and lasso. This made sure that the most informative features were chosen. Then employing specific phenotypic features, such as plant side area, height, and volume, to fit a linear model that predicts plant biomass at 42 days after sowing (DAS). The outcomes suggested the influence of these characteristics on plant growth, and their combinations were found to be significant predictors of plant biomass at 42 DAS. The fitted

linear models' diagnostic metrics revealed different levels of success in predicting biomass; the model that used side area and volume performed the best. The study conducted genome-wide association studies (GWAS) using a Compressed Mixed Linear Model (CMLM) to identify genetic variants linked to particular

The impact of adding more groups on GWAS performance was further investigated in the study through the analysis of metrics such as True Positive Rate, Compression, False Positive Rate, FDR q-value, and Group Size. The results showed that the capacity to recognize genuine associations improved with the number of groups. The study made clear how crucial it is to manage the false discovery rate and false positives in GWAS in order to get accurate results. The examination of Group Size and its Impact on GWAS Performance demonstrated variations in specific metrics according to group sizes, highlighting the necessity of optimizing group size for better results in genomic analyses. Overall, the results of the study show that phenotypic features can be used effectively to predict plant biomass and highlight the potential of CMLM in GWAS analysis for using estimated biomass from individual traits like side area, volume, and plant height, the study examined the information of related SNPs. The number of significant correlations for various plant traits combinations, such as side area + Volume, plant height + Volume, and volume + plant height + plant side area, were also investigated in this study. As plants advanced from 11 DAS to 26 DAS and then to 42 DAS, the study saw an increase in the number of significant SNPs linked to these traits, suggesting a dynamic genetic landscape that changes over time. The results indicated that genetic variants had a significant impact on volumetric characteristics. Composite variables showed more SNPs associations due to reduced complexity of the data that led to reduced noise and multicollinearity.

4.2 RECOMMENDATIONS

To improve the accuracy of genetic association analyses, the study suggests fitting compressed linear mixed models with the composite variables obtained from predicted biomass obtained from image-derived plant features.

REFRENCES

1. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC-19, 716–723
2. Chen, S., Kim, Y., & Lee, K. (2019). Genetic expressions and trait evolution in rice plants. Plant Genetics Journal, 8(4), 278-291.
3. FAO. (2020). *The State of Food and Agriculture 2020*. Food and Agriculture Organization of the United Nations.
4. 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.
5. Gana, 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.
6. Gao, X., Becker, L., Becker, D., Starmer, J., & Province, M. (2019). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, p.n/a-n/a.
7. Johnson, P., & Williams, D. (2020). Genetic correlations between different traits in soybean plants. Genetics and Plant Biology, 8(1), 56-68.
8. Jones, R., & Brown, J. (2021). Meta-analysis of genetic association studies in plant species. Genetics Review, 17(4), 289-302.
9. 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.
10. Lee, K., Wang, Q., & Chen, S. (2020). Genetic associations with fruit quality traits in tomatoes. Plant Genetics Journal, 9(1), 45-58.
11. Liu, H., & Zhang, L. (2020). Genetic associations between plant architecture and yield components in soybean plants. Plant Genetics Today, 9(1), 45-58.
12. Liu, H., & Zhang, L. (2021). Trait combinations in genetic association studies across plant species. Genetics Review, 19(2), 145-158.
13. Liu, Y., Wang, L., Sun, C., Zhang, Z., Zheng, Y., & Qiu, F. (2019). Genetic analysis and major QTL detection for maize kernel size and weight in multi-environments. *Theoretical and applied genetics*, *127*(5), 1019-1037.
14. Malécot, G. (1948). The mathematics of heredity. *The mathematics of heredity.*
15. Matsui, T., Mullis, N., Roy, R... (2022). The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross. *Nat Commun* 13, 1463 https://doi.org/10.1038/s41467-022-29111-z
16. Melchinger, E., Utz, F., & Schön, C. (1991). Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics*, *149*(1), 383-403.
17. Muraya MM, Chu J, Zhao Y, Junker A, Klukas C, Reif JC, Altmann T (2017) Genetic variation of growth dynamics in maize (Zea mays L.) revealed through automated non-invasive phenotyping. The Plant Journal, 89: 366-380.

18. Muraya, MM (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
19. Robinson, G. (1991). [That BLUP is a Good Thing: The Estimation of Random Effects]: Rejoinder. *Statistical Science*, 6(1), pp.48-51.
20. Schwarz, G. (1978) Estimating the Dimension of a Model. Annals of Statistics, 6, 461-464.
21. 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.
22. Smith, A., & Brown, J. (2018). Heritability estimation in crop plants. Crop Genetics Review, 6(2), 123-136.
23. Smith, A., Johnson, P., & Lee, K. (2018). Genetic basis of trait variations in maize plants. Plant Genetics Today, 7(2), 134-147.
24. Smith, K., Brown, D., Lee, S., & Zhang, L. (2020). Enhancing GWAS performance through effective data reduction techniques. Genetic Epidemiology, 43(5), 456-468.
25. Vilhjálmsson, B., & Nordborg, M. (2012). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, *14*(1), 1-2. https://doi.org/10.10 38/nrg3382
26. Visscher, M, Yang, J., & Goddard, E. (2010). *A commentary on "common SNPs explain a large proportion of the heritability for human height*" by Yang *Et al.* Twin Res Hum Genet. 2010; 13:517–524.doi:10. 1375/twin.13.6.517 PMID: 2114 29 28
27. Wang, Q., & Li, H. (2018). LD patterns in rice plants. *Plant Genetics Journal*, 7(3), 189-202.
28. Wang, Q., & Li, H. (2019). Genetic correlations between traits in maize plants. Crop Genetics Review, 7(1), 78-89.
29. 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.
30. Zhang, L., Chen, S., & Wang, Q. (2016). Genetic basis of biomass production in wheat plants. Plant Genetics Journal, 6(3), 213-226.
31. Zhou, S., Zhang, X., Liu, Z., Wang, Y., Li, C., & Zhang, X. (2019). Advances in maize breeding: Developing resilient and high-yielding varieties. *Field Crops Research*, *237*, 97-109.