IMPROVING ACCURACY OF COMPRESSED MIXED LINEAR MODEL: AN APPLICATION TO GENOME-WIDE ASSOCIATION STUDIES

DOMINIC MONG'ARE OBARE

A Thesis Submitted to the Graduate School in Partial Fulfilment of the Requirements for the Award of the Degree of Doctor of Philosophy in Applied Statistics of Chuka University

CHUKA UNIVERSITY

OCTOBER 2024

DECLARATION AND RECOMMENDATION

Declaration

This thesis is my original work and has not been presented for an award of diploma or conferment of degree in any other University.

Signature.....Date 17/10/2024

Dominic Mong'are Obare
SD18/39873/18

Recommendation

This thesis has been examined, passed and submitted with our approval as University supervisors

Signature..... Date 17/10/2024

Prof. Moses Mahugu Muraya, PhD
Chuka University

Signature..... Date 16/10/2024

Dr. Gladys Gakenia Njoroge, PhD
United States International University Africa

COPYRIGHT

DEDICATION

This work is dedicated to my mum, Martha Nyaera and my beloved wife, Mrs. Emily Mulunde Mwene.

ACKNOWLEDGEMENTS

I thank the Almighty God for the opportunity to study this degree. I am very grateful to my supervisors Prof. Moses Mahugu Muraya and Dr. Gladys Gakenia Njoroge for their constructive scholarly criticism, timely intellectual guidance and support throughout my studies. Without their prompt, explicit comments during my academic presentations, the success of this study would not have been a reality. I would also wish to thank Prof Charles Ochieng' Ombaka, Chairperson Department of Physical Sciences (DPSC) for his great mentorship and unlimited support throughout this study.

I wish to appreciate Mr. Moses Mburu and Dr. Peter Gachoki for their guidance during the data analysis process.

ABSTRACT

Mixed linear models are very popular in various disciplines due to their robustness in handling complex datasets and taking into account the data structures. Genome wide association studies (GWAs) are key to success in genomic prediction and statistical modelling of genotype-phenotype relationships. Genomic wide association and genomic prediction combines molecular markers and statistical models to detect variants of interest. Though several statistical models have been used in GWAS, advancement in phenotyping and sequencing technologies necessitates improvement of the existing ones in order to increase their statistical power. The general objective of this study was to develop an improved enriched compressed linear mixed model that addresses aspects of accuracy and statistical power. This study took into account cumulative genetic variants causing phenotypic differences at different developmental stages of the plant. Secondary data obtained from the database at IPK-Gatersleben, Germany, was used in this study. The data set consists of phenotypic data from 252 maize inbred lines and 50,000 Single Nucleotide Polymorphism (SNPs) markers. Data analysis was done on R-statistical software Version 4.4.1. Analysis was done on three developmental stages, at 11, 26 and 42 days after sowing (DAS). Plant phenotypic features such as volume, side area and height were used to predict plant biomass. Single trait analysis was done first (plant side area, height and volume) followed by a combination of two traits (plant volume+Plant height, Plant height + Plant side area, Plant volume+Plant Side area) then lastly a combination of all the three traits (plant Plant volume+Plant height+ Plant side area). On plant side area total number of SNPs detected were 6, on volume 8 SNPs were detected, plant height 8 SNPs were detected. On plant volume+ Plant height 20 SNPs were detected, on plant volume+ Plant side area 11 SNPs and on plant volume+ Plant height + Plant area 22 SNPs were detected across the entire analysis on different developmental stages. The results of this study underscored the significance of considering multiple composite traits simultaneously in GWAs to unravel complex genetic correlations and synergistic effects that influence plant architecture and performance. The study revealed dynamic shifts in significant SNP associations as plants progressed through different growth stages, highlighting the evolving genetic landscape during plant development. The study demonstrated the efficiency of the Compressed Mixed Linear Model (CMLM) proved to be highly efficient in clustering individuals and identifying putative quantitative trait nucleotides (QTNs). Incorporating composite phenotypic variables (plant volume, surface area and height) in the model produced the lowest AIC and BIC 1967.630 and 1999.870, respectively, indicating a well-fitting and parsimonious model. Based on the results, the study recommends using machine learning techniques like Random Forest and Lasso to select the most significant phenotypic features for predicting plant biomass. By combining predicted biomass values from multiple variables through standardization aggregation and summation statistical technique, a more informative composite feature can be generated. The composite variable provides a robust input for trait-SNPs association in GWAS, as demonstrated by the enhanced results in this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
BIC	Bayesian Information Criterion
BLUP	Best Linear Unbiased Predictor
CMLM	Compressed Mixed Linear Model
COM	Complete Linkage
DAS	Days after Sowing
ECMLM	Enriched Compressed Mixed Linear Model
ECMLM+F-C	ECMLM due to Fuzzy C-means clustering
ECMLM+K	ECMLM due to K-means clustering
EMMA	Efficient Mixed Model Association
EMMAX/P3D	Population Parameters Previously Determined
$_e$QTL	Expressive Quantitative Trait Locus
FDR	False Discovery Rate
FLE	Lance-Williams Flexible-beta Method
GAPIT	Genome Association and Prediction Integrated Tool
GRAMMER	Genome Wide Rapid Association Using Mixed Model Regression
GWAs	Genome Wide Association Study
H^2	Heritability Statistic
IAP	Identical Ancestors Point
IBD	Identical by Descent
IPK	International Prototype Kilogram
K	Kinship Matrix
LD	Linkage Disequilibrium
LRT	Likelihood Ratio Test
Mb	Mega base
MLE	Maximum Likelihood Association
MLM	Mixed Linear Model
NACOSTI	National Commission for Science, Technology & Innovation
PCA	Principal Component Analysis
PCs	Principal Components

QTL	Quantitative Trait Loci
QTNs	Quantitative Trait Nucleotides
REML	Residual Maximum Likelihood
RF	Random Forest
RKHS	Reproducing Kernek Hilbert Space
RRM	Realised Relationship Matrix
SAS	Statistical Analysis Software
SIN	Single Linkage
SNPs	Single Nucleotide Polymorphism
SUPPER	Settlement of MLM Under Progressively Exclusively Relationship
SVM	Support Vector Machine
TASSEL	Trait Analysis by association, Evolution and Linkage
UPGMA	Unweighted Pair-group with Arithmetic Mean
UPGMC	Unweighted Pair-group Method using Centroid
VIFs	Variance Inflation Factors
WAR	Ward's Method

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

Genome-wide association study (GWAs) is a biological concept that deals with statistical correlation of molecular markers with phenotypic variation (Wang *et al*., 2014). It is a valuable approach to identify the genetic basis of phenotypic variation using association mapping. Association mapping generally falls into two broad categories, i.e. candidate-gene association mapping, which statistically associates markers in selected candidate genes controlling for phenotypic variation for specific traits, and GWA, which examines genome-wide genetic variation to find association signals for different complexes. characters. (Zhu et al., 2008). Most agronomically important traits are quantitatively inherited (Yu & Buckler, 2006) and show complex variation, so we will focus on GWA when dissecting these traits. It is a precious approach for identifying the genetic basis of phenotypic traits the use of association mapping. Association mapping commonly falls into two broad categories, i.e. Candidate-gene association mapping, which statistically relates markers in decided on candidate genes controlling phenotypic variant for unique traits and GWAs, which surveys genetic variant throughout the complete genome to locate indicators of institutions for various complicated tendencies (Zhu et al., 2008). Most of the agronomic crucial developments are quantitatively inherited (Yu & Buckler, 2006) and shows complex version, for this reason, the focus on GWAs to dissect such tendencies.

The purpose of GWAs is to discover loci associated with variation in traits of interest. Quantitatively inherited traits are managed with the aid of many loci each with small however additive impact. Therefore, locating statistical correlation for such trends present a hard statistical trouble. The assumption at the back of association mapping is that full-size associations get up because the marker is in linkage disequilibrium (non-random association of alleles at two or greater loci) with a causal variant affecting the trait (Li & Zhang, 2019). However, population shape causes false tremendous associations in GWAS if now not accounted for (Chen et al., 2017). Hence the want to improve on experimental designs or statistical methods to address the confounding impact. Moreover, with stepped forward sequencing and phenotyping technologies, statistics on more loci and complex phenotypes may be generated main to alternate

inside the dynamics of GWAs (Elshire et al., 2011; Klukas et al., 2014; Muraya et al., 2015; Muraya et al., 2017). Hence, the want to continuously enhance the statistical power of existing statistical fashions.

Despite development of those enabling technologies (high throughput genotyping and phenotyping platforms) in biology, their application has limitations since the statistical models linking genotypes with phenotypes stay the same. The accuracy of the consequences and the detection power of GWAs research still remains low (Wu and Zhao, 2009). These is attributed to the big size and multi-dimensionality of the datasets generated via the high throughput genotyping and phenotyping structures. Hence, the need to expand improved statistical approaches for GWAs with a purpose to improve on statistical accuracy and detection electricity of extensive genetic variants correlated to the phenotypic developments. There are some of statistical models used to carry out GWAs in animals and plants. These consist of Bayesian models, multivariate fashions, generalized linear fashions, mixed linear models and machine gaining knowledge of algorithms (Thornton, 2015; Bi & Pounds, 2018; Sun & Zhao, 2020). Among those, the mixed linear models are the most widely used to hyperlink the genotype with the phenotypic traits of hobby. This is because of their robustness in fitting the statistics structure and deliberating the confounding factors due to populace shape and cryptic relatedness (Zhou & Stephens, 2012, Listgarten et al., 2013, Lippert et al., 2013). Mixed linear models are statistical models containing each fixed consequences and random results (Bermejo & Zucknick, 2013). These models are useful in a wide type of disciplines inside the bodily, biological and social sciences. They are mainly useful in settings wherein repeated measurements are taken or where measurements are made on clusters of associated statistical gadgets.

Mathematically the conventional mixed linear model can be represented as

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

Where,

$\boldsymbol{y}$ is a known vector of observations, with mean $E(y) = X\beta$

$\boldsymbol{\beta}$ is an unknown vector of fixed effects;

$\boldsymbol{u}$ 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 $var(\varepsilon) = R$;

$\boldsymbol{X}$ and $\boldsymbol{Z}$ are known design matrices relating the observations y to β and u respectively (Robinson, 1991, Henderson, Kempthorne, Searle & von Krosigk, 1959).

However, in GWAS; y is an n-via-1 matrix of quantitative developments which represents located phenotypes and it corresponds to the reaction variable (e.G. Biomass, Yield) X is an n-with the aid of-p recognised layout matrix for covariates and marker outcomes, this matrix consists of the predictor variables (e.g. Top, facet leaf length, leaf width) β is an unknown vector containing fixed consequences, which include the genetic marker, populace structure(Q), and the intercept. Z is an N-through-S recognized layout matrix keeping S causal loci, together with the kinship matrix, any other additional constant effects. ε is an determined vector of residuals. U is an unknown vector of random additive genetic results from a couple of historical past QTL for people/inbred traces. The u and ε vectors are assumed to be generally allotted 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)

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

More details on linear mixed models and mathematical estimation of parameters are found in Jiang (2011) and Faraway (2016).

Despite the vast use of linear mixed model in GWAs, there's an issue of confounding, which has been addressed but nonetheless stays a task. Confounding causes a spurious association between genotype and phenotype leading to affiliation that don't mean causation. This makes detecting the real correlations among phenotypes and genotypes non-trivial (Vilhjálmsson and Nordborg, 2012). Two techniques were evolved to clear up this confounding trouble and so that it will can improve statistical power for MLM version (Smith & Lee, 2020). The first approach includes the usage of best the related genetic markers as pseudo Quantitative Trait Nucleotides (QTNs) to derive kinship instead of all, or a random sample of genetic markers. In the second approach, the Compressed MLM (CMLM), clusters individuals into agencies and fits genetic values of organizations instead of genetic effects of people as random effects.

Pseudo QTNs are predicted to intently tune a number of the causative QTNs, and are selectively used to derive kinship for a particular testing marker. When a pseudo QTN is correlated with the trying out marker, it's miles excluded from those used to derive kinship (Fang et al., 2020, Lee et al., 2014). In the FaST-LMM-Select technique, a pseudo QTN is taken into consideration correlated if it is within a 2 Mb c language on both facet of the trying out marker (Listgarten et al., 2012). Instead of using a 2 Mb c program languageperiod, the Settlement of MLM underneath Progressively Exclusive Relationship (SUPER) method applies a threshold on linkage disequilibrium (LD) between the pseudo QTNs and the testing marker. Selectively including and/or except for pseudo QTNs to derive kinship for a particular trying out marker improves statistical energy compared to deriving an ordinary kinship from all, or a random pattern of genetic markers (Wang et al., 2014). Zhang et al. (2010) the usage of human, dog and maize datasets, discovered out that compression level with the first-class-fitting version extended statistical power by using 34%, forty-two% and 20% with a QTN that defined zero.12, 0.30 and 0.30 gadgets of the phenotypic trendy deviation, respectively.

Compressed linear combined model (CMLM) has similarly been changed to enriched CMLM (Li et al., 2014). The enriched CMLM (ECMLM) improves statistical energy via optimizing the organization kinship definition, in preference to the usage of the average kinship algorithms continuously as in ordinary MLM (Li et al., 2014). Three organization kinship algorithms (common, median, and most) and eight hierarchical

clustering algorithms have been used to expand the ECMLM. The 8 clustering algorithms encompass UPGMA, un-weighted pair-organization centroid (UPGMC), entire linkage (COM), Lance-Williams flexible-beta approach (FLE), McQuitty's similarity evaluation or weighted pair-organization technique the use of arithmetic averages (WPGMA), weighted pair-institution technique using centroid (WPGMC), single linkage (SIN), (nearest neighbour), and Ward's method (WAR). Though the 8 hierarchical clustering algorithms were examined, there are many other algorithms that cluster individuals into agencies. Moreover, research the usage of non-hierarchical clustering algorithms like fuzzy C-approach and the tough k-means are limited (Jones & Brown, 2021), and there is a need to examine whether or not they'll enhance the statistical energy of the fashions.

The other venture in detection of QTL is the multiple test correction used in maximum GWAs. Bonferroni correction is the widely used more than one check correction, but it's miles often too conservative such that many vital loci can't pass the stringent criterion of importance check (Gao et al., 2009, Kaler & Purcell, 2019). Permutation exams are considered the gold standard in a couple of testing adjustment in genetic association research. However, it is computationally in depth, specifically for GWAs, and may be impractical if a massive number of random shuffles are used to make certain accuracy. The simpleM approach became used to approximate the permutation threshold (Gao et al., 2009). By making use of the simpleM technique, researchers can better seize genetic variants that cause variations in growth stages. The simpleM technique additionally helps in lowering false positives in GWAS research through accounting for the unique genetic results at unique developmental ranges. By incorporating the simpleM method and considering genetic editions at various developmental degrees, the have a look at pursuits to growth the statistical electricity of GWAS analyses. This will allow for a greater unique identity of quantitative trait loci (QTL) with the aid of considering genetic results throughout unique growth degrees. This SimpleM technique (Gao et al., 2008) is a primary aspect evaluation (PCA)-based technique that calculates the effec-tive quantity of independent tests, Meff, for a given facts set.

It is really worth noting that most QTL were identified for a single time point, by and large on the final harvest stage (Brown et al., 2016). This method detects handiest the cumulative results, without taking into consideration genetic editions inflicting phase-unique variations in growth ranges. This has necessitated the development of a stepped forward compressed combined linear model at specific developmental ranges of the plant (Wang et al., 2018). This examine additionally intends to analyze if this approach will reduce false positives as a result increase statistical power in GWAS research (Li et al., 2020).

1.2 Statement of the Problem

High-throughput genotyping and phenotyping technologies are quickly revolutionizing genomic studies. These platforms provide complex data set that is difficult to dissect using conventional statistical models. Statistical methods to handle such large and complex data set are under continuous development. Such statistical models should aim to exploit the information to increase the probability of detecting genetic factors causing trait variation in GWAs. Despite the strength in detecting genetic factors causing trait variation due to availability of large set of phenomics and genetic data, GWA analysis are limited by statistical accuracy and statistical power. Hence, the need to develop improved statistical models that can detect associations between genotype and phenotype accurately with high statistical power. Mixed linear models are very popular in GWAs due to their robustness in handling complex traits and taking into account population structure and cryptic relatedness. Most of the current improved mixed linear models only examine the genotype-phenotype relationship when the plant is fully grown. In addition to that, they take into account for only one phenotypic variable at a time. Hence, there is need to examine if associating SNP markers with phenotypic variables at different growth developmental stages of a plant improves the accuracy of GWASs. It is also necessary to test the effect of combining more than one phenotypic variable on the accuracy of GWAs results.

1.3 Objectives of the Study

1.3.1 General Objective of the Study

To develop compressed mixed linear model for genome-wide association studies.

1.3.2 Specific Objectives of the Study

i. To imrove the conventional compressed mixed linear model (CMLM) using predicted biomass from Plant volume, Plant height and Plant area and Plant volume+ Plant height.
ii. To improve the conventional compressed mixed linear model (CMLM) using predicted biomass from Plant height+Plant area and Plant height+ Plant area + Plant volume.
iii. To improve the conventional compressed mixed linear model (CMLM) using predicted biomass from Plant height+ Plant area + Plant volume.
iv. To compare the performance of the Compressed mixed linear models developed.

1.4 Research Questions

i. How does the conventional compressed mixed linear model (CMLM) perform when using predicted biomass from Plant volume, Plant height, Plant area and volume + side height in GWAs analysis?
ii. How does the conventional compressed mixed linear model (CMLM) perform when using predicted biomass from Plant volume, Plant area and Plant height + Plant area?
iii. How does the conventional compressed mixed linear model (CMLM) perform when using predicted biomass from Plant side height + Plant area + Plant volume in the analysis?
iv. How do the models developed compare in terms of performance statistical accuracy?

1.5 Significance of the Study

In statistical genetics, one of the major focus is on building predictive models of the genotype-phenotype association to quantify the proportion of the total phenotypic variance that is due to genetic basis. Many models have been proposed to incorporate

additive genetic effects into prediction models. High-throughput phenotyping and genotyping technologies have led to generation of highly dimensional complex data sets. Such data can improve the dissection of genetic factors causing traits variation. These types of data provide an avenue of GWAS and genomic selection. The two are powerful tools in breeding for high yielding varieties as well as varieties that are tolerant/resistant to various biotic and abiotic factors. However, the statistical models available are limited in terms of accuracy and statistical power when modelling. The study therefore aimed to address the limitations in current GWAS analyses by developing improved statistical models that enhance the accuracy and statistical power of detecting genetic associations in plant traits. With the increasing complexity of high-throughput genotyping and phenotyping data, conventional statistical models struggle to effectively analyze the vast amount of information. By focusing on compressed mixed linear models and exploring the associations between SNP markers and multiple phenotypic variables at different growth stages of plants, the research seeks to improve the precision and reliability of GWAS results. This study is crucial for advancing our understanding of genetic factors influencing trait variations in plants and can significantly impact future breeding and genomic studies in agriculture.

CHAPTER TWO
LITERATURE REVIEW

2.1 An Overview Genome Wide Association Studies

Genome-Wide association studies are considered to be an incredible way to discover the genetic variations of complicated tendencies the usage of large natural population (Bush & Moore, 2012). In the current years, GWAS had been successfully carried out to the dissection of complicated tendencies in humans (Bush & Moore, 2012) and animals (Coster et al., 2010). In 2010, GWAS had been efficaciously carried out to the analysis of 107 phenotypes in Arabidopsis thaliana inbred lines (Atwell et al., 2010). This has been established to be effective technique for figuring out genes, alleles or haplotypes related to a certain agronomic trait beneath complicated environments (Yan et al., 2011), that is based totally on the linkage disequilibrium as a result of the association of goal trait and haplotype loci. Genome-wide association research offers the possibility to methodically analyse the genetic architecture of complicated quantitative developments in many vegetation via benefiting from the excessive variety and rapid linkage disequilibrium decay (Flint- Garcia et al., 2005; Wang et al., 2005, Li et al., 2013; Mammadov et al., 2015; Li et al., 2016; Chen et al., 2017). In maize for instance, GWAS have been used to come across numerous QTL controlling several tendencies (Cardwell, 1982, Sreckov, et al., 2011; Buckler et al., 2009; Zwonitzer et al., 2010; Agrama & Moussa, 1996; Chen et al., 2016; Chen et al., 2014; Li et al., 2016; Sibov et al. 2003; Ji-hua et al., 2007; Zhang et al. 2008; Peiffer et al. 2014). However, the largest QTL detected for most maize tendencies generally defined < five% of the phenotypic variance (Wallace et al., 2014). This has additionally been determined for traits like flowering time (Buckler et al., 2009), disease resistances (Kump et al., 2011), kernel starch, protein, and oil content (Cook et al., 2011), or morphological tendencies together with leaf structure (Cook et al., 2011) which were initially anticipated to be genetically much less complicated than grain or biomass yield. These consequences reveal the quandary in detection strength of current statistical tools. The traditional statistical models have fallen quick to tackle those demanding situations. In order to lessen the fake positives, cryptic relatedness and the complexity of populace substructure, the fashions need to be re-evaluated and improved. Under device mastering approach and genomic choice, volumes of data have elevated, and new studies efforts geared toward integrating and unifying numerous fields of research. This

focuses on estimating extra correct predictive values of unobserved individuals by way of the use of statistical getting to know or system gaining knowledge of methods e.g. Synthetic neural networks (ANN) are not unusual in system learning. The machine getting to know approach is involved with developing and making use of laptop algorithms that enhance with data (Libbrecht & Noble, 2015). Machine learning can either be classified as either supervised or unsupervised. In supervised mastering, the goal is to predict a desired output value (trait) inferred from input statistics. The prediction project is called class if the outputs are specific and regression if outputs are non-stop. In unsupervised mastering, the objective is to find out agencies and associations among input variables in which there may be no output variable (Hastie et al., 2009). Many system getting to know fashions have been used in supervised mastering, which includes nearest-neighbour's methods, decision bushes, naïve Bayes, Bayes nets, and rule-primarily based studying (Kotsiantis, 2007). Methods which have been carried out in genomic selection under gadget mastering include aid vector machine (SVM), random wooded area (RF), and synthetic neural community (ANN) (Gonzalez-Camacho et al., 2012). The reproducing kernel Hilbert space (RKHS), first of all provided as a semi-parametric technique (Gianola et al., 2006), is also one of the gadget gaining knowledge of algorithms (Gonzalez-Recio et al., 2014). Sun et al. (2012) proposed a successful ensemble-based technique to imputation of slight-density genotypes for genomic prediction. The random wooded area is one of the ensemble algorithms in which the non-parametric characteristic is the common of regression choice timber or category (Hastie et al., 2009).

2.2 Linear Mixed Mode

2.2.1 Linear Mixed Model for Complex Traits

In linear blended models the phenotype y is written as the blended sum of a linear term within the fixed eff ects β, that is in the linear regression version include a bias term in addition to the eff ects of regarded covariates and the marker of interest, and linear random eff ects u. (Runcie & Crawford, 2019)

$$y = X\beta + \bar{G}u + \varepsilon \quad (4)$$

Where y is an n-with the aid of-1 matrix of quantitative tendencies, X is an n-by using-p layout matrix for covariates and marker outcomes, the N-by Using-S Matrix G¯is the

layout matrix holding S causal loci and ε is a n-by way of-1 vector of uncorrelated commonly disbursed mistakes. When checking out a marker for affiliation with the phenotype, the usual software of linear combined models for genome-huge association studies, the variables of hobby are modelled as fixed, while the random eff ects account for nuisance version and are incorporated out. If the causal loci are confounded with the aid of population shape, then such as these in a take a look at for affiliation corrects for confounding variant inside the phenotype, much like covariates in a general linear regression model.

For maximum complex developments it has been determined that the contribution of each of the S causal loci to the entire degree of genetic variance σ_g^2is about identical, with an eff ect length distribution this is inversely proportional to the corresponding minor allele frequencies f_g (Park et al., 2010). Under this model the random eff ects are handled as unbiased Gaussian variables, every contributing an equal fraction of 1/s σ_g^2 to the whole varianceσ_g^2. The S loci contained within the design matrix G^{-} are assumed to have a median of 0 and unit variance. If we define the entire random genetic eff ect as v =G^{-}u, then v follows a multivariate regular distribution v ∼N (zero; σ_g^2K), wherein the covariance is proportional to K = 1/s $G^{-}G^{-}$^T, a matrix that quantifies the genetic courting between individuals primarily based on the causal loci. Under this normally used version the marginal likelihood of y follows from marginalization of v:

$$\int N(y|x\beta + v;\ \sigma^2 I).N(V|0;\ \sigma_g^2 \mathrm{K})\ \mathrm{dv} = \mathrm{N}\ (\mathrm{y}|\ x\beta;\ \sigma_g^2 \mathrm{K} + \sigma^2 I) \tag{5}$$

The log (marginal) likelihood is a function of fixed effects β, and the variance parameters θ = (σ^2, σ_g^2), namely the level of environmental noise $\sigma^2$and the genetic variance σ_g^2.

$$\mathrm{Log\ l}\ (\beta, \theta) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |V_\theta| - \frac{1}{2} (\mathrm{y} - \mathrm{x}\beta)^T V_\theta^{-1} (\mathrm{y} - \mathrm{x}\beta) \tag{6}$$

Where,

$$V_\theta = \sigma_g^2 \mathrm{K} + \sigma^2 I \tag{7}$$

is the complete covariance term of the distribution.

The causal variants enter the model best in the genetic relatedness matrix K, which directly represents the confounding variation inside the phenotype.

2.2.2 Kinship Matrix

In Fisher's infinitesimal version the distribution of a phenotype is derived for the case of an infinite wide variety of causal editions. In this version a quantitative phenotype y with total genetic variance σ_g^(2)receive with the aid of the sum of a big variety of genetic eff ects of character variances 1/s σ_g^2.

$$y = \sum_{j=1}^{s} \bar{G}_j u_j + e \tag{8}$$

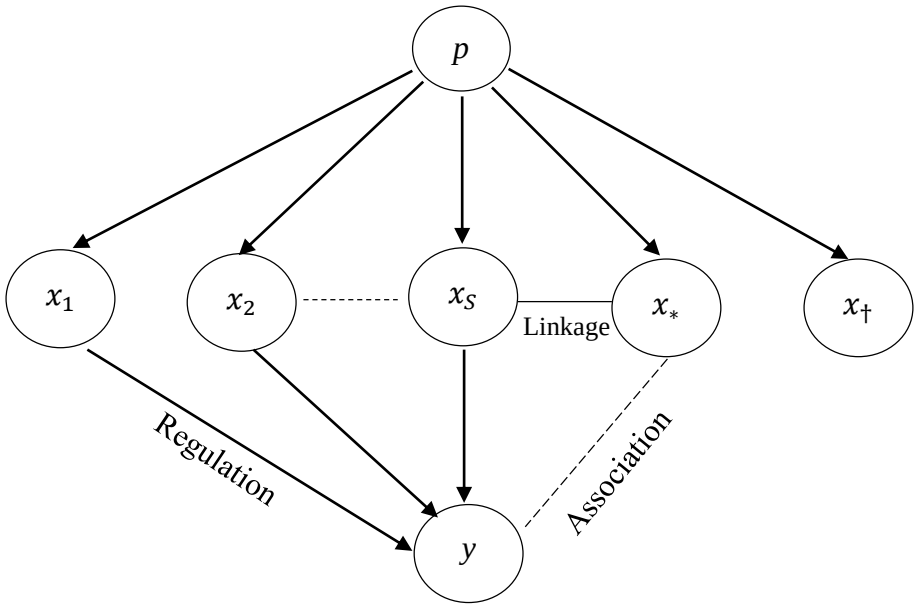


Figure 1: Conditioning on background SNPs

Then inside the restrict of an infinite range of causal loci (S →∞), that independently comply with Mendelian inheritance, the phenotypic covariation between people is proportional to the quantity of genetic fabric on the causal loci this is identical through descent (IBD). Introducing additional fixed eff ects β, then the distribution of the phenotype is given by means of a linear mixed model

.

$$y \sim N(x\beta;\ \sigma_g^2 K_{IBD} + \sigma^2 I) \tag{9}$$

Wherein K_IBD is the matrix of IBD coefficients among pairs of people. Kinship coefficients can be computed from recognised pedigrees (Fisher, 1918) and need to be corrected for an increase in relatedness due to inbreeding (Mal´ecot, 1948). For the case, where the pedigree isn't always recognised, the kinship matrix K_IBD from genetic markers is carried out (Abecasis et al., 2001, Hardy & Vekemans, 2002).

2.2.3 Realized Relationship Matrix

Estimates of found out relationships among individuals are acquired through counting average wide variety of shared marker alleles between two individuals and were shown to enhance prediction of the genetic factor of a trait over predictions using pedigree-based kinship estimates (Nejati-Javaremi et al., 1997). These predictions were further progressed by means of using dense genome-wide markers that tag causal loci because of linkage (Daetwyler et al., 2010). Also, for GWAS the use of relationships expected from genome-extensive markers have been shown to improve correction for confounders over relationship based on kinship (Kang et al., 2010). Let the N-By-S Matrix G be a matrix preserving S genotyped markers for N individuals. Each marker in G is suggest centred and is normalized to have unit variance. The realized courting matrix (RRM) as the empirical covariance matrix is described as

$$K_{RRM} = \frac{1}{s} \mathrm{G}G^{T} \tag{10}$$

Similar to the linear blended version that makes use of the causal variants, the linear mixed model the usage of the RRM may be written as a linear regression model wherein some regressors are fixed and some regressors are random.

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Gu} + \epsilon \tag{11}$$

Variants contained within the RRM are used as random regressors or covariates that seize the genetic variant inside the phenotype through being connected to the unknown causal variations or via ways of confounding. Overfitting because of the massive range of covariates eff ects is averted through integrating the regressors over independent ordinary distributions with variance (σ_g^2)/S (Kang et al., 2010, Lee et al., 2010).

2.2.4 Kernel Methods

Interpretation of random results is performed by Gaussian random procedure, whose covariance is given by using the genetic relatedness (Banerjee et al., 2005). After integration of the genetic eff ects inside the linear combined version chance, the random eff ects most effective seem implicitly as a feature of their covariance matrix K. Any features contained inside the authentic layout matrix are used most effective implicitly in the form of dot-merchandise. It has been proven that during principle any symmetric semi positive-definite kernel matrix might be used for K (Meuwissen & Goddard, 2010, Morota & Gianola, 2014). While in the well-known linear combined version those dot products are computed directly on the features, resulting in a model this is linear inside the functions, kernel features may additionally represent non-linear dot-merchandise and hence can yield fashions which might be non-linear within the authentic capabilities (Kimeldorf & Wahba, 1970). Kernel methods had been used to give you covariance structures that don't most effective cowl genetic eff ects, but additionally eff ects of hidden surroundings. For example, in the context of expression quantitative trait locus (eQTL) research, covariance structures primarily based on latent variable fashions (Lawrence, 2004, 2005) representing shared hidden influences can be predicted together from all expression phenotypes, and has been proven to yield advanced correction and a gain in strength to detect novel institutions (Stegle et al., 2010).

2.2.5 Best Linear Unbiased Prediction

The fine best linear independent unbiased predictor (BLUP) is a minimal variance expected cost of the random eff ects v in a linear mixed version. Predictions of random eff ects are a way to expect the phenotype of an individual from genotyped SNP-information (Lee et al., 2008). The BLUP vˆ_* of an individual of hobby listed through * is acquired by using maximizing the joint distribution of the vector of all determined phenotypes y and the random genetic eff ect v_* of that character of interest. Let Vθ be the full covariance term of y, the 1-by-N dimensional vector of genetic relatedness between the individual of interest and all located people be k(*,): and the genetic relatedness of the individual of hobby with itself be k_(*,*) then the joint distribution of y and v_* is given as;

$$\begin{pmatrix} y \\ v_* \end{pmatrix} \sim N\left\{ \begin{pmatrix} x\beta \\ 0 \end{pmatrix}; \begin{pmatrix} v_\theta & \sigma_g^2 k_{*,}{}^T \\ \sigma_g^2 k_{*,} & \sigma_g^2 k_{*,*} \end{pmatrix} \right\} \quad (12)$$

The BLUP is equal to the mean of the conditional distribution of v_* given y.

$$v_*|\mathrm{y} \sim \mathrm{N}(\sigma_g^2 k_*: V_\theta^{-1}(\mathrm{y} - x\beta); \sigma_g^2 k_{*,*} - \sigma_g^2 k_*: V_\theta^{-1} \sigma_g^2 k_{*,}{}^T) \quad (13)$$

Given the vector of covariates for the individual of interest x_*, then the conditional distribution of the phenotype of the individual y_* follows by adding the covariates effects and accounting for the environmental variance (Lee *et al.*, 2008).

$$y_*|y \sim (x_*\beta + \sigma_g^2 k_{*,}: V_\theta^{-1}(\mathrm{y} - x\beta); \sigma_g^2 k_{*,*} - \sigma_g^2 k_*:(\sigma_g^2 \mathrm{K} + \sigma^2 I)^{-1} \sigma_g^2 k_{*,}{}^T \quad (14)$$

2.2.6 Parameter Estimation in Linear Mixed Models

Linear mixed model with random effects integrated out with log likelihood given, the goal is to infer the model parameters β and $\theta = [\sigma^2, \sigma_g^2]$ and any additional covariance parameters if they are present (Widmer *et al*., 2014).

Score; The gradient of the log-likelihood as given in Equation (6) with respect to fixed effects w defines the score of w.

$$\frac{\nabla \log L(\beta, \sigma^2, \sigma_g^2)}{\nabla \beta} = X^T V_\theta^{-1} y - X^T V_\theta^{-1} X \beta \quad (15)$$

The score of a variance parameter is the partial derivative of the log-likelihood with respect to a variance parameter θ_i (Listgarten *et al*., 2012)

$$\frac{\partial log L(\beta, \theta)}{\partial \theta_i} = \frac{1}{2} \mathrm{tr}\,(V_\theta^{-1} \frac{\partial V_\theta}{\partial \theta_i}) + \frac{1}{2} (\mathrm{y} - x\beta)^T V_\theta^{-1} \frac{\partial V_\theta}{\partial \theta_i} (\mathrm{y} - x\beta) \quad (16)$$

The matrix derivative of the covariance $V_\theta = \sigma_g^2 \mathrm{K} + \sigma^2 I$ with respect to the environmental variance $\theta_1 = \sigma^2$ equal

$$\frac{\partial V_\theta}{\partial \sigma^2} = \mathrm{I} \quad (17)$$

and the matrix derivative with respect to the genetic variance θ_2 = σ_g^2 equals

$$\frac{\partial V_\theta}{\partial \sigma_g^2} = \text{K} \quad (18)$$

2.2.7 Maximum Likelihood Estimation

The likelihood is maximized by means of equating the gradient with respect to all parameters to zero and together solving the ensuing equations. Though, even as for linear regression the most probability parameters may be found in closed shape from the gradient equations, this isn't always the case for linear combined models. The log marginal probability characteristic is not jointly convex in the variance parameters, rendering it difficult to make certain global maximization of the chance.

A sincere manner to acquire a neighborhood most reliable of the parameter values is to use gradient descent strategies. For most GWAS applications, even though, naive use of gradient descent techniques isn't always nicely appropriate, as those contain repeated computation of the log likelihood characteristic as well as of the gradients and for second-order methods like Fisher scoring or Newton-Raphson additionally of the Fisher or located information matrix (Demidenko, 2013). Maximum probability estimation can be simplified through writing the log chance as a feature of the ratio γ= (σ_g^2)/σ^2 of the genetic variance σ_g^2over the environmental variance σ^2(Hartley and Rao, 1967).

$$\text{Log L} (\gamma, \sigma^2, \beta) = -\frac{N}{2} \log (2\pi\sigma^2) - \frac{1}{2} \log \left[H_\gamma\right] - \frac{1}{2\sigma^2} (\text{y} - x\beta)^T H_\gamma^{-1} (\text{y} - x\beta). \quad (19)$$

Where,

H_γ = I + γK ………………………………………………………………….. (20)

In this components the most likelihood solutions for all parameters aside from γ (β, and σ^2) observe in closed shape for any superb fee of γ. The maximum probability value β(M_γ) of the fixed eff ects as a characteristic of γ is located with the aid of taking the gradient of the log-probability with appreciate to β and at the same time putting all entries of the gradient to 0.

$$\frac{1}{\sigma^2} X^T H_\gamma^{-1}\text{y} - \frac{1}{\sigma^2} X^T H_\gamma^{-1}\text{yX}\, \beta_{M_\gamma} = 0 \quad (21)$$

By bringing the part involving β_{M_γ} to one side and after cancelling σ^2 from the equation, this becomes;

$$X^T H_\gamma^{-1}\text{X}\, \beta_{M_\gamma} = X^T H_\gamma^{-1}\text{y} \quad (22)$$

Multiplication of both sides by the inverse of the factor on the left side yields the maximum likelihood solution of the fixed effects given a value of γ as

$$\beta_{M_\gamma} = (X^T H_\gamma^{-1}\text{X})^{-1} X^T H_\gamma^{-1}\text{y} \quad (23)$$

To find the most probability fee of the genetic variance σ^2 as a function of γ, the maximum likelihood values of the fixed effects β_ (M_γ) which do now not rely on σ^2 are substituted into the log chance. The derivative with respect to σ^2 is set to 0, giving

$$-\frac{N}{2\sigma^2 M_\gamma} + \frac{1}{2\sigma^4 M_\gamma} (y - X\beta_{M_\gamma})^T H_\gamma^{-1}(\text{y} - X\beta_{M_\gamma}) \quad (24)$$

Both sides are multiplied by $2\hat{\sigma}^4$ and the result is solved for $\sigma^2 \gamma M_\gamma$, such that the maximum likelihood solution of the residual variance given γ is;

$$\sigma^2 M_\gamma = \frac{1}{N} (y - X\beta_{M_\gamma})^T H_\gamma^{-1}(\text{y} - X\beta_{M_\gamma}) \quad (25)$$

After further simplification, this becomes

$$\sigma^2 M_\gamma = \frac{1}{N} y^T P_\gamma^T H_\gamma^{-1} P_\gamma \text{y} \quad (26)$$

Where we defined P_γ as the matrix

$$P_\gamma = \text{I} - \text{X}(X^T H_\gamma^{-1}\text{X})^{-1} X^T H_\gamma^{-1}\text{y} \quad (27)$$

Plugging the maximum likelihood estimators of β and σ^2 back into the likelihood, a profile log likelihood is obtained as

$$\text{Log L}(\gamma) = -\frac{N}{2}\log(2\pi\sigma^2 M_\gamma) - \frac{1}{2}\log\left|H_\gamma\right| - \frac{1}{2\sigma^2 M_\gamma}(\text{y} - x\beta_{M_\gamma})^{-1}H_\gamma^{-1}(\text{y} - x\beta_{M_\gamma}) \quad (28)$$

Using the maximum likelihood expressions and simplifying, this profile log likelihood becomes a function of γ alone

$$\text{Log L}(\gamma) = -\frac{N}{2}(1 + \log\frac{2\pi}{N}) - \frac{1}{2}\log\left|H_\gamma\right| - \frac{N}{2}\log y^T P_\gamma^T H_\gamma^{-1} P_\gamma \text{y} \quad (29)$$

A nearby optimum with appreciate to γ of this profile log-probability may be acquired by using gradient descent techniques. Alternatively, spinoff-unfastened strategies like a grid seek can be used to find foremost for γ (Demidenko, 2013).

2.2.8 Restricted Maximum Likelihood Estimation

When handling finite information the maximum chance estimate has been observed to underestimate the variances inside the Gaussian model. This may be attributed to the reality that below maximum likelihood estimation, the estimate of variances depends on a distribution that has been profiled for the fixed eff ects and exerts a loss in levels of freedom. Restricted maximum chance estimation has been proposed to conquer this loss on tiers of freedom by means of estimating variance additives of the model handiest on a projection of the target variable (i.e., the phenotype) into an N −D-dimensional subspace, that is orthogonal to the fixed eff ects (Patterson & Thompson, 1971). The variance components are anticipated from residuals of the target variable, after the fixed eff ects were regressed out. The fixed eff ects then again are envisioned from every other projection, which below the model is statistically unbiased to the former projection. More officially, for N > D, appropriate projection matrices S and Q_γ are chosen such that they fulfill the four following standards (Patterson & Thompson, 1971):

1. rank(S) = N –D (30)

   rank(Q_γ)= D (31)

2. The two projections are statistically independent under the model.

   $\Leftrightarrow \text{Cov}(S_y, Q_\gamma\, y) = 0$ (32)

   $\Leftrightarrow S\, H_\gamma Q_\gamma^T = 0$ (33)

3. The expected value of Sy under the model is zero.

   $\Leftrightarrow \text{E}(S_y) = 0$ (34)

   $\Leftrightarrow SX\beta = 0$ (35)

   $\Leftarrow SX = 0$ (36)

4. $\text{rank}(Q_\gamma X) = D$ (37)

From the conditions above it follows that the probability may be written because the manufactured from chance functions of two independent projections of the facts, one on Sy and one on Q_γy (Patterson & Thompson, 1971).

$$L(\beta, \gamma, \sigma^2) \propto L(\gamma, \sigma^2 | S_y) \cdot L(\beta | Q_\gamma y, \gamma, \sigma^2) \quad (38)$$

Where $L(\gamma, \sigma^2 | S_y)$ is also called the restricted likelihood, for which Harville (1977) proposed suitable matrices for S and Q_γ^T namely the N-by-N orthogonal projector for the fixed effects X

$$S = I - X(X^T X)^{-1} X^T \quad (39)$$

and the D-by-N matrix

$$Q_\gamma = (X^T H_\gamma^{-1} X)^{-1} X^T H_\gamma^{-1} \quad (40)$$

Parameter estimation is then done in a -step system S_y) is maximized with appreciate to the variance parameters γ and σ^2. Then, the answers acquired are plugged into log L(β) in the end is maximized with recognize to β.

2.2.9 Estimation of Variance Parameters by Restricted Maximum Likelihood

To determine appropriate mathematical expression for log L (γ, σ^2reality that the covariance of S_y is σ^2SH_γS, a matrix that is rank deficient because of a projection to the distance orthogonal to X. A manner to accomplish that is using the pseudo-determinant and the Moore-Penrose pseudo-inverse of SHγS.

$$\log L(\gamma, \sigma^2 | S_y) = -\frac{N-D}{2} \log(2\pi\sigma^2) - \frac{1}{2}\log(SH_\gamma S) - \frac{1}{2\sigma^2} y^T (SH_\gamma S)^T S_y \quad (41)$$

Both, the pseudo-determinant as well as the pseudo-inverse can be computed from the economy spectral decomposition V_s ∑_γ V_s^T of 〖SH〗_γ S in which ∑_γ is an (N−D)-via-(N−D) diagonal matrix, holding the non-0 eigenvalues of SH_γS and V_s is an N-by means of-(N−D) matrix, keeping the corresponding eigenvectors as columns. S may be written as V_s V_s^T. Also the usage of V_s V_s^T = I, we get

$$-\frac{N-D}{2} \log(2\pi\sigma^2) - \frac{1}{2}\log(\sum_\gamma) - \frac{1}{2\sigma^2} y^T V_s \sum_\gamma^{-1} V_s^T y \quad (42)$$

Then, L(γ, σ^2| S_y) equals to the regular multivariate normal distribution on V_s^Ty with covariance matrix $\sum_\gamma$

$$L(\gamma,\sigma^2|S_y)=N(V_s^T y|0;\sigma^2\sum_\gamma) \quad (43)$$

The confined maximum chance estimators of the variance parameters σ_(Rγ)^2 and γRγ are discovered by way of making use of most probability Sy), given via Equation (47). Taking the by-product Sy) with respect to σ^2 and putting this to zero, we get,

$$-\frac{N-D}{2\sigma_{R_\gamma}^2} + \frac{1}{2\sigma_{R_\gamma}^2} y^T V_s \sum_\gamma^{-1} V_s^T y \quad (44)$$

The solution to this Equation is

$$\sigma_{R_\gamma}^2 = \frac{1}{N-D} y^T V_s \sum_\gamma^{-1} V_s^T y \quad (45)$$

Profile constrained chance; to plug the limited maximum probability estimator for the environmental noise σ_(R_γ)^2 again into the log limited likelihood, a log constrained likelihood, that is profiled over σ^2, is derived as

$$L(\gamma, \sigma_{R_\gamma}^2 | S_y) = -\frac{N-D}{2}(1 - \log\frac{2\pi}{N-D}) - \frac{1}{2}\log\left|\sum_\gamma\right| - \frac{N-D}{2}\log R_S \quad (46)$$

Where the residual term is

$$R_S = y^T V_s \sum_\gamma^{-1} V_s^T y \quad (47)$$

The derivative of this log likelihood with respect to the remaining free parameter γ is

$$\frac{\partial log L(\gamma, \sigma_{R_\gamma}^2 | S_y)}{\partial \gamma} = -\frac{1}{2}\mathrm{tr}\sum^{-1}\frac{\partial \sum_\gamma}{\partial \gamma} - \frac{N-D}{2}\frac{\frac{\partial R_S}{\partial \gamma}}{R_S} \quad (48)$$

The derivative of the matrix $\sum_\gamma$ of the N −D non-zero eigenvalues of SH_γS is given by

$$\frac{\partial \sum_\gamma}{\partial \gamma} = \frac{\sum_\gamma - I_{N-D}}{\gamma} \quad (49)$$

As can easily be verified that the derivative of the residual term is given by

$$\frac{\partial R_S}{\partial \gamma} = y^T V_s \sum_\gamma^{-1}\frac{\partial \sum_\gamma}{\partial \gamma}\sum_\gamma^{-1} V_s^T y \quad (50)$$

2.2.10 Estimation of Fixed Effects by Restricted Maximum Likelihood

An expression for the logarithm of

$L(\beta | Q_\gamma y, \gamma, \sigma^2)$

can be found as

$$\mathrm{Log}\, L(\beta | Q_\gamma y, \gamma, \sigma^2) = -\frac{D}{2}\log\sigma^2 - \frac{1}{2}log\left|Q_\gamma H_\gamma Q_\gamma^T\right|^{-\frac{1}{2}} - \frac{1}{2\sigma^2}R_{Q_\gamma} \quad (51)$$

Where $R_{Q_\gamma} = (Q_\gamma y - Q_\gamma X\beta)^T (Q_\gamma H_\gamma Q_\gamma^T)^{-1}(Q_\gamma y - Q_\gamma X\beta)$

So L(β|Q_γy, γ, σ^2) is equal to a multivariate Normal distribution of $Q_\gamma y$

$$L(\beta|Q_\gamma y, \gamma, \sigma^2) = N(Q_\gamma y|\beta; \sigma^2 X^T H_\gamma^{-1} X) \quad (52)$$

From this, the maximum with respect to β_{R_γ} is found in closed form as the general least squares estimator:

$$\beta_{R_\gamma} = (X^T H_\gamma^{-1} X)^{-1} X^T H_\gamma^{-1} y) \quad (53)$$

The estimator has the same shape as the maximum likelihood estimate, but diff ers within the estimate of the parameter γ, as for REML γ is expected by maximizing the profile limited likelihood.

2.2.11 Bayesian Interpretation of Restricted Maximum Likelihood Estimation

Restricted maximum probability estimation seems heuristic, it has the subsequent equivalent Bayesian interpretation. When in preference to maximizing the chance over the fixed effects, those are integrated over a previous distribution, then, because the earlier variance σ2 β goes to infinity, the ensuing marginal likelihood is proportional to the confined probability. The restrained maximum likelihood covariance parameters coincide with the most probability parameters of the marginal probability. Also, the posterior expectation of the fixed eff ects coincide with the restricted maximum likelihood estimator (Harville, 1974). As the _ vital Q γy) of β equals one, the restrained probability

L(γ, $\sigma^2_{R_\gamma}|\, S_y$) can be written as

$$L(\gamma, \sigma^2_{R_\gamma}|\, S_y) = L(\gamma, \sigma^2_{R_\gamma}|\, S_y) \int P(\beta|Q_\gamma y) d\,\beta \quad (54)$$

L(γ, $\sigma^2_{R_\gamma}|\, S_y$) is not affected by the fixed effects and thus can be moved inside the integral

$$L(\gamma, \sigma^2_{R_\gamma}|\, S_y) = \int L(\gamma, \sigma^2_{R_\gamma}|\, S_y)\, P(\beta|Q_\gamma y) d\,\beta \quad (55)$$

Assuming that the prior distribution over the fixed effects is an isotropic normal distribution with variance σ_β^2. Then, the posterior distribution can be identified by completing the squares as

$$\beta \sim \mathrm{N}(m_\beta,\ V_\beta) \tag{56}$$

Where,

$$V_\beta = (\frac{1}{\sigma^2}\left(X^T H_\gamma^{-1}\mathrm{X}\right)^{-1} + \frac{1}{\sigma_\beta^2} I)^{-1} \tag{57}$$

And

$$m_\beta = \frac{1}{\sigma^2}\ V_\beta\left(X^T H_\gamma^{-1}\mathrm{X}\right)^{-1} Q_\gamma \mathrm{y} \tag{58}$$

It is easy to see, that in the limit of this distribution, as σ_β^2 goes to infinity

$$\beta \sim \mathrm{N}(Q_\gamma \mathrm{y};\ \sigma^2 X^T H_\gamma^{-1}\mathrm{X}) \tag{59}$$

For this case,

$$\mathrm{L}(\gamma,\ \sigma_{R_\gamma}^2 |\ S_y) = \int \mathrm{L}(\gamma, \sigma_{R_\gamma}^2 |\ S_y)\mathrm{L}(\beta | Q_\gamma \mathrm{y}, \gamma, \sigma^2)\ \mathrm{d}\,\beta \tag{60}$$

The product of $\mathrm{L}(\gamma, \sigma_{R_\gamma}^2 |\ S_y)$ and $\mathrm{L}\left(\beta \middle| Q_\gamma \mathrm{y}, \gamma, \sigma^2\right)$ is proportional to the full likelihood

$$\mathrm{L}(\gamma,\ \sigma_{R_\gamma}^2 |\ S_y) = \mathrm{C}\ \int \mathrm{L}(\beta | Q_\gamma \mathrm{y}, \gamma, \sigma^2)\ \mathrm{d}\,\beta \tag{61}$$

The constant C can be identified as $|X^T X|^{\frac{1}{2}}$ (Harville, 1974). Solving the integral analytically, the restricted likelihood is obtained as being proportional to the marginal distribution of y, when the fixed effects are integrated over a prior distribution with infinite variance.

$$(2\pi\sigma^2)^{\frac{N-D}{2}}\left|X^T H_\gamma^{-1}\mathrm{X}\right|^{\frac{1}{2}} \exp\text{-}\frac{1}{2\sigma^2}\,(y - X\beta_{R_\gamma})^T H_\gamma^{-1}(y - X\beta_\gamma) \tag{62}$$

Where the posterior expectation of the fixed effects equals the restricted maximum likelihood estimator

$$\beta_{R_\gamma} = (X^T H_\gamma^{-1}\mathrm{X})^{-1} X^T H_\gamma^{-1}\mathrm{y} \tag{63}$$

2.2.12 Statistical Testing using Linear Mixed Models

2.2.12.1 Likelihood Ratio Test

For linear mixed models, the null distribution of the likelihood ratio statistic for LRT_LR linear regression can be approximated by a chi-squared distribution with one degree of freedom, since the alternative model includes one extra parameter when testing the N×1 SNP-vector x_ * association when conditioning on the effect of any covariates contained in X (Hartley and Rao, 1967).

$$\mathrm{LRT_{LR}} = \frac{\max_{\beta,\beta_*,\sigma^2,\sigma_g^2} N(y|X\beta + x_*\cdot\beta_*, \sigma_g^2 + \sigma^2 I)}{\max_{\beta,\sigma^2,\sigma_g^2} N(y|X\beta + x_*\cdot 0, \sigma_g^2 + \sigma^2 I)} \sim x_1^2 \tag{64}$$

As for linear regression P-values are computed from the survival function of the distribution

2.3 Population Structure Correction

Apart from linear mixed models quite a number methods had been proposed to correct for population shape. Even though linear combined fashions have been proven to improve confounder correction over these alternative techniques in some of GWAS in maize, A. Thaliana, potato and human (Kang et al., 2010), there is probably benefits from combining combined models with different strategies to get a extra stringent correction. Genomic control estimates the amount of inflation in a GWAS by means of evaluating quantiles of the observed distribution of check information to the theoretical un-confounded distribution and corrects for inflation by simple matching of the median (Devlin and Roeder, 1999).

2.3.1 Genomic Control

To correct for the inflation of P values, it is possible to compare the distribution of the obtained test statistics with their theoretical null distribution. Genomic control defines

the genomic inflation factor λ as the ratio of the observed average test statistic to the theoretical test statistic under the null hypothesis in the theoretical null analysis.

$$\lambda = \frac{median(LRT)}{median(Hnull)}$$

So, for the chance ratio check of a fixed eff ect in a linear version (like linear regression or linear blended models) λ equals the median of two times the found LRT over the median of a Chi-rectangular distribution with one diploma of freedom. Another not unusual variant uses quantiles of the bottom ten logarithmic distribution of P values. In this case, λ is given by using the median of the found $-\log_{10}(\mathrm{P})$ over $-\log_{10}(0.5)$. Correction with the aid of genomic manage is done by dividing all test information through λ and can be shown to yield a conservative check. From an intuitive perspective the reasoning at the back of genomic manipulate is that the huge majority if no longer all examined markers aren't linked to causal loci and for this reason their take a look at information need to comply with the distribution under the null hypothesis. Diff erently than methods that account for populace shape by using approaches of modelling, genomic manipulate uniformly aff ects the test records of unlinked in addition to connected SNPs and does not exchange the order of check records. In experiments Price et al. (2006) show that such uniform adjustment is on the only facet insufficient for markers displaying more potent than common diff erentiation between ancestral populations and leads to a loss in electricity at markers having weaker diff erentiation. While procedures that model population structure can in a few cases lead to an growth in electricity compared to an uncorrected analysis, the usage of genomic manage always reduces energy. Due to its simplicity though, correction with the aid of genomic manipulate can be implemented along with any model or statistical test, as long as the distribution of the check information is thought or may be reliably predicted. For example, it'd be viable to apply genomic manage to accurate for residual inflation in an evaluation the use of blended model. Besides for correction λ is likewise a normally used measure of the calibration of the Type 1 blunders in a GWAS. A cost of λ larger than one is an indicator of anti-conservativeness, or inflation of Type-1 mistakes, a cost this is smaller than one suggests loss of electricity due to deflation. However, values of λ large than 1.05 and above in research of human have generally been attributed to confounding (Burton et al., 2007), for studies of highly polygenic developments like

body mass index or human peak a great deal larger values of λ were shown to arise due to large linkage to causal loci by myself, without the presence of confounding (Speliotes et al., 2010). In this case correction by means of genomic manipulate could yield overly conservative estimates.

$$\lambda = \frac{median(LRT)}{median(Hnull)} \tag{65}$$

So, for the likelihood ratio test of a fixed effect in a linear model (like linear regression or linear mixed models) λ equals the median of twice the observed LRT over the median of a Chi-square distribution with one degree of freedom. Another common variant uses quantiles of the base ten logarithmic distribution of P values. In this case, λ is given by the median of the observed $-\log_{10}(\mathrm{P})$ over $-\log_{10}(0.5)$. Correction by genomic control is performed by dividing all test statistics by λ and can be shown to yield a conservative test.

From an intuitive standpoint the reasoning behind genomic control is that the vast majority if not all tested markers are not linked to causal loci and for this reason their test statistics should follow the distribution under the null hypothesis. Differently than methods that account for population structure by ways of modelling, genomic control uniformly affects the test statistics of unlinked as well as linked SNPs and does not change the order of test statistics. In experiments Price *et al*. (2006) show that such uniform adjustment is on the one side insufficient for markers showing stronger than average differentiation between ancestral populations and leads to a loss in power at markers having weaker differentiation. While approaches that model population structure can in some cases lead to an increase in power compared to an uncorrected analysis, the use of genomic control always reduces power. Due to its simplicity though, correction by genomic control can be applied in conjunction with any model or statistical test, as long as the distribution of the test statistics is known or can be reliably estimated. For example, it would be possible to apply genomic control to correct for residual inflation in an analysis using mixed model. Besides for correction λ is also a commonly used measure of the calibration of the Type 1 error in a GWAS. A value of λ larger than one is an indicator of anti-conservativeness, or inflation of Type-1 errors, a value that is smaller than one indicates loss of power due to deflation. However, values

of λ larger than 1.05 and above in studies of human have usually been attributed to confounding (Burton *et al.*, 2007), for studies of highly polygenic traits like body mass index or human height much larger values of λ have been shown to occur due to broad linkage to causal loci alone, without the presence of confounding (Speliotes *et al.*, 2010). In this case correction by genomic control would yield overly conservative estimates.

2.3.2 Structured Association

Instead of without delay including markers that implicitly reflect population structure by using diff erences in allele frequencies between populations, markers also can be used to estimate specific estimates of shared ancestry. From genetic markers, we estimate some of latent variables representing ancestry the use of Markov-chain Monte Carlo sampling (Pritchard et al., 2000). The version may be interpreted as clustering, where the club variables represent shared ancestry between cluster individuals. These latent variables are then used as covariates in an affiliation take a look at (Pritchard et al., 2000). Compared to use of markers, summarizing the genetic variant in a small wide variety of latent variables has the advantage that typically a fewer variety of covariates are required to accurate for populace shape, yielding a smaller loss in strength. The Markov-chain Monte Carlo set of rules, although, has a runtime that makes software of based association infeasible on larger numbers of markers and people. Another trouble is to well determine the appropriate quantity of latent variables. Even though the likelihoods are finished for a number of latent variables, repeated runs of the algorithm might in addition growth the runtime of the technique. As latent variables capture diff erences in version on a population scale, dependent association is useful for correcting for population structure however unlikely to accurate for cryptic relatedness present within the information.

2.3.3 Principal Components Analysis

Another latent phenomenal variable method that has been carried out to correct for population structure for genetic research is most important additives evaluation (PCA) (Price et al., 2006). Principal components (PCs) are predicted from a genome-huge covariance matrix just like the realized courting matrix.

$$K_{PCA} = \frac{1}{S} \mathrm{G}\, G^T \tag{66}$$

While PCA is computationally more efficient than structure association, as it requires computation of the first k Eigenvectors of K_{PCA} which can be performed in O(N^2k) runtime.

Regarding structured association, it is not clear how best to choose the number of principal components to use. By default, most researchers use the first ten principal components. Determining the correct number of components to use can be cumbersome. It has been proposed to select the number of components such that the total genomic variation is significantly captured by PC (Price et al., 2006). The number of components is usually chosen by comparing the values of λ (Tian et al., 2008). The first principal components tend to be dominated by large regions of strong coupling. Consequently, these components provide little information about population structure (Astle & Balding, 2009). It is reported that two to fifteen computers are sufficient in practice (Astle & Balding, 2009). It has also been shown that the number of PKs needed for correction could be reduced by selecting PKs correlated with the phenotype (Novembre & Stephens, 2008).

2.3.4 FaST Linear Mixed Models for Genome-wide Association Studies

Linear mixed models are among the richest class of models used today for genome-wide association studies and have been shown to be able to correct for population structure, family structure, and cryptic relatedness (Astle & Balding, 2009, Price et al., 2010). Unlike other methods, linear mixed models can capture all these forms of relatedness simultaneously without being aware of them and without having to separate them from each other. Despite the advantages of linear mixed models, their widespread use on current data sets has long been limited. The main reason is that statistical inference in linear mixed models involves calculations that proceed cubically in the number of samples N. Even in studies involving a moderate number of samples, evaluating a naive model for each individual SNP is infeasible, as the typical number of SNPs in a genome-wide association study ranges from hundreds of thousands to millions. Another bottleneck in applying linear mixed models to large cohorts is that the memory requirements for storing the complete relationship matrix are quadratic in the number of samples. The situation has changed due to the recent focus on adapting

linear mixed models to be scalable for larger and larger studies (Aulchenko & de Koning, 2007, Kang et al., 2010).

Introduction to the Efficient Mixed Model Association (EMMA) algorithm. EMMA cleverly uses linear algebra to avoid repeated cubic operations on the covariance matrix in a mixed model when estimating test variance parameters (Patterson & Thompson, 1971, Kang et al., 2010). Although the computational savings over naïve evaluation are enormous, for each marker tested, the spectral decomposition of the N-by-N matrix must be computed to maintain the cubic run requirements per test. Because of this bottleneck of the run, this approach is practically limited to the analysis of genome-wide association studies on no more than a few hundred samples. Because exact mixed model computations have commonly been considered too expensive to be applicable even for moderately sized cohorts, various approximations have been proposed that aim for faster computations at the possible cost of reduced accuracy (Aulchenko & de Koning, 2007, Kang et al, 2010, Svishcheva et al., 2012).

The maximum broadly used method, which has been shown to paintings properly on many facts units, is to make the simplifying assumption that variance parameters are fixed for every SNP examined and can be estimated on the null version (Kang et al., 2010). Due to this simplification, cubic computations in the shape of two spectral decompositions of N-via-N matrices must be finished handiest once, for the null-version. The computations that are required consistent with SNP are reduced from cubic to quadratic inside the wide variety of samples. The garage necessities remain quadratic, because the algorithm still requires the entire genetic relatedness matrix. Even even though this approach has efficaciously been applied to cohorts of over ten thousand samples, together with the quadratic storage, the ultimate cubic computations, which might be tough to parallelize efficiently, are nevertheless a considerable bottleneck. In practice, the technique is not applicable to research on extraordinarily big cohorts that are produced in recent times so one can advantage sufficient power to get new insights on complicated phenotypes, locate vulnerable SNP eff ects, or eff ects of rare alleles (Speliotes et al., 2010). With the new FaST-LMM algorithm (Lippert et al., 2013), supplied, we reveal, that precise mixed model computations are possible on statistics sets of more than ten thousand samples, without making any simplifying

assumptions. The most widespread approach, which works well on many data sets, is the simplifying assumption that the variance parameters are fixed for each SNP tested and can be estimated on the null model (Kang et al., 2010). Thanks to this simplification, the cubic calculations in the form of two spectral decompositions of N-by-N matrices need only be performed once, and that is for the null model. The calculations required per SNP are reduced from cubic to quadratic in the number of samples. Storage requirements remain quadratic since the algorithm still requires the full genetic relatedness matrix. Although this approach has been successfully applied to cohorts with more than ten thousand samples along with quadratic storage, the remaining cubic computations, which are difficult to efficiently parallelize, are still a significant bottleneck. In practice, this approach is not applicable to studies on extremely large cohorts, which are currently being produced to gain sufficient power to gain new insights into complex phenotypes, detect weak SNP effects or rare allelic effects (Speliotes et al., 2010). Introducing the new FaST-LMM algorithm (Lippert et al., 2013), we demonstrate that accurate mixed model computations are feasible on datasets with more than ten thousand samples without making any simplifying assumptions.

In advance algorithms FaST-LMM calls for simplest a unmarried preliminary cubic spectral decomposition, whilst the computations that need to be accomplished in line with SNP examined are handiest quadratic in the variety of samples. Thus, the runtime is N times quicker than previous actual algorithms (Kang et al., 2010) and has the equal runtime as whilst variance parameters are assumed to be fixed (Kang et al., 2010). FaST-LMM allows utility to extremely huge facts units. The computational gains rely on the quantity of markers used to estimate genetic similarity being smaller than the range of individuals within the examine. On actual facts units a fixed of just a few thousand SNPs sampled linearly along the chromosome presents a good degree of genetic similarity and is sufficient to accurate for population shape in a genome-wide association look at. By choosing a small quantity of markers by using their affiliation to the phenotype FaST-LMM yields a consistent boom in electricity and higher correction for genetic relatedness compared to apply of genome-extensive markers. FaST-LMM provides great speedups when tens of lots of people or greater are

analyzed, which can be demonstrated by means of analysing a dataset containing more than a hundred and twenty,000 individuals (Lippert et al., 2013).

2.3.5 Efficient Mixed Model Association

The mixed version affiliation (EMMA) algorithm (Kang et al., 2010) builds at the insight that the maximum likelihood, or as a substitute, the restricted most chance, of a linear blended version may be rewritten as a feature of only a single parameter, γ, the ratio of the environmental noise variance σ^2 to the genetic varianceσ_g^2, in place of as a feature of all of the version parameters (Patterson & Thompson, 1971, Kang et al., 2010). Given a fee of γ the (constrained) most probability values for all of the model parameters (i.e. The genetic and environmental variances along with the fixed-eff ects) follow in closed form. The identification of the most efficient parameters turns into an optimization problem over this single variable γ. Additionally, EMMA makes smart use of spectral decompositions to lessen the cost of comparing the log-likelihood for any price of γ, that's primarily cubic within the wide variety of individuals, to linear inside the number of people, as soon as the two spectral decompositions are accomplished (Patterson and Thompson, 1971, Kang et al., 2010).

For maximum likelihood estimation, EMMA uses a likelihood formulation using the ratio γ = (σ_g^2)/σ^2 of the variance parameters σ_g^2 andσ^2, for which it estimates the maximum likelihood for all other parameters to follow in closed form. By fitting the maximum likelihood estimates βM_γ as given in equation (40) and the maximum likelihood estimate σ^2 M_γ as given in equation (40) back to the log likelihood, the log likelihood log L(γ) is obtained. as shown by equation (43)

$$\text{Log L}(\gamma) = -\frac{N}{2}(1 + \log\frac{2\pi}{N}) - \frac{1}{2}\log\left|H_\gamma\right| - \frac{N}{2}\log y^T P_\gamma^T H_\gamma^{-1} P_\gamma \text{y} \quad (67)$$

Where,

$$P_\gamma = \text{I} - \text{X}(X^T H_\gamma^{-1}\text{X})^{-1} X^T H_\gamma^{-1}\text{y} \quad (68)$$

$P_\gamma^T H_\gamma^{-1} P_\gamma$y equals the Moore-Penrose pseudoinverse of SH_γS

Where,

$$\text{S} = \text{X}(X^T \text{X})^{-1} X^T \quad (69)$$

$$\text{Log L}(\gamma) = -\frac{N}{2}(1 + \log\frac{2\pi}{N}) - \frac{1}{2}\log\left|H_\gamma\right| - \frac{N}{2}\log y^T P_\gamma^T (SH_\gamma S)^\dagger y \quad (70)$$

The economic spectral decomposition of SHγS can be efficiently obtained from U_S(Σ + I) U_S, the economic spectral decomposition of S(K + I)S, where Σ is obtained by subtracting one from each nonzero eigenvalue of S(K + I) S. The pseudoinverse SH_γ S can be solved from this economic spectral decomposition by inverting the non-zero eigenvalues:

$$(SH_\gamma S)^\dagger = (\gamma\Sigma + I_{N-D})^{-1} U_S^T \quad (71)$$

Let the spectral decomposition of K be UΛU^T. The spectral decomposition H_γ = γK + I is given by the relation U(γΛ + I) U^T. Using the equality |AB| = |A|·|B|, for the complete rank matrices A and B and |U| = 1, the logarithm of |H_γ| can be written as the logarithm of |γΛ + I|. Plugging in these terms, we get;

$$\text{Log L}(\gamma) = -\frac{N}{2}(1 + \log\frac{2\pi}{N}) - \frac{1}{2}\log|\gamma\Lambda + I| - \frac{N}{2}\log y^T U_S (\Sigma + I)^{-1} U_S^T y \quad (72)$$

In order to make efficient evaluation efficient, we write Equation (78) using only the entries of these matrices.

$$\text{Log L}(\gamma) = -\frac{N}{2}(1 + \log\frac{2\pi}{N}) - \frac{1}{2}\sum_{i=1}^{n}\log[\Lambda]_{n,n} + 1) - \frac{N}{2}\log\left(\sum_{i=1}^{N-D}\frac{\left|U_S^T y\right|_i^2}{\gamma[\Sigma]_{i,i}+1}\right) \quad (73)$$

The derivative with respect to γ is then given by

$$\frac{\partial \text{Log L}(\gamma)}{\partial\gamma} = \frac{1}{2}\sum_{i=1}^{N}\frac{[\Lambda]_{n,n}}{\gamma+[\Lambda]_{n,n}+1} - \frac{N}{2}\frac{\frac{\left|U_S^T y\right|_i^2}{(\gamma[\Sigma]_{i,i}+1)^2}}{\sum_{i=1}^{N-D}\frac{\left|U_S^T y\right|_i^2}{\gamma[\Sigma]_{i,i}+1}} \quad (74)$$

2.3.6 Restricted Maximum Likelihood Estimation

EMMA maximizes the log restrained probability log L(γ, $\sigma_{R_\gamma}^2$| S_y) in the shape given in Equation (52) with σ^2 profiled out. The economy spectral decomposition of SH_γ S is obtained efficiently from the financial system spectral decomposition U_S (Σ + I_(N-D)) U_S of S (K + I_N) S

$$L(\gamma, \sigma^2_{R_\gamma} | S_y) = -\frac{N-D}{2}(1 + \log\frac{2\pi}{N-D}) - \frac{1}{2}\log|\gamma\Sigma + I_{N-D}| - \frac{N-D}{2}\log R_S \quad (75)$$

Where the residual is given by

$$R_S = y^T U_s(\gamma\Sigma + I_{N-D})^{-1}U_s^T y \quad (76)$$

Again, this log-likelihood can be evaluated efficiently for any value of γ in O(N) as

$$-\frac{N-D}{2}(1 + \log\frac{2\pi}{N-D}) - \frac{1}{2}\sum_{i=1}^{n}\log(\gamma[\Sigma]_{n,n} + 1) - \frac{N-D}{2}\log\sum_{i=1}^{N-D}\frac{|U_S^T y|_i^2}{\gamma[\Sigma]_{i,i}+1} \quad (77)$$

The same is true for the derivative with respect to γ given by

$$\frac{\partial \text{Log } L(\gamma,\sigma^2_{R_\gamma} | S_y)}{\partial\gamma} = \frac{1}{2}\sum_{n=1}^{N-D}\frac{[\Sigma]_{n,n}}{\gamma+[\Sigma]_{n,n}+1} - \frac{N-D}{2}\frac{\frac{|U_S^T y|_i^2 [\Sigma]_{i,i}}{(\gamma[\Sigma]_{i,i}+1)^2}}{\sum_{i=1}^{N-D}\frac{|U_S^T y|_i^2}{\gamma[\Sigma]_{i,i}+1}} \quad (78)$$

2.3.7 Optimizing the Ratio of Variances

Solving the non-convex optimization over the ratio of variances γ, EMMA applies a combination of grid search and a by-product based technique. To bracket local minima, the by-product of the probability the by-product of the restrained likelihood is evaluated on one hundred equally spaced factors on the logarithm of γ starting from -5 to 5. For every two consecutive points, where the derivative modifications, a root finder primarily based on Brent's set of rules is carried out to equate the derivative to zero within the respective c program language period and retrieve the nearby premier.

2.3.8 Runtime and Memory Footprint

Once Λ, Σ and U^Ty are computed, both the log-likelihood as well as the derivative with respect to γ can be evaluated in O(N) for any value of γ. Assuming that the wide variety of opinions of the derivative is given via a constant C, the value of finding an ultimate γ for a single SNP is O(C·N). The required prematurely computations are computation of the eigenvalues Λ of K, the economic system spectral decomposition U_S(Σ + I)U_S of SH_1S, and multiplication of the phenotype via U_S^T. A trouble that arises when this algorithm is implemented to GWAS is that for every SNP tested,

the matrix X of fixed eff ects is a diff erent one. It follows that the matrix S = (I - X (〖 X^T X)〗 ^(-1) X^T) is a diff erent one for every SNP. As a result, a brand new economy spectral decomposition of an N-by means of-N matrix SH1S is needed. As the rank of SH_1S equals N −D, the economic system spectral decompositions can be computed in O(N^2·(N −D)) as an example using iterative strategies. In exercise though, the number of fixed eff ects D used in genome-extensive affiliation research, isn't greater than a one-digit integer and may be treated as a constant. It follows, that the desired computations for checking out all SNPs are in

$$O(C \cdot N + S \cdot N^3) = O(S \cdot N^3) \quad (79)$$

Where,

S is the number of all SNPs tested. If each SNP only is into loaded to memory while being tested, the memory footprint is dominated by the cost of storing the genetic similarities K, given by $O(N^2)$.

2.3.9 Efficient Approximations to the Mixed Model

Applying linear combined model to the evaluation of large data several approximations have been proposed. The earliest such approximation turned into the Genome Wide Rapid Association the use of Mixed Model and Regression (GRAMMAR) set of rules (Aulchenko and de Koning, 2007), which makes use of a blended version only in a single prematurely to compute a populace-structure corrected version of the phenotype which may be analysed by preferred linear regression. The EMMAX and P3D algorithms avoid repeated cubic computations by way of estimating the ratio γ of variance parameters within the combined version simplest as soon as, retaining it fixed across all assessments.

2.3.10 Generating Stratified Pseudo-phenotypes by Prediction

The concept of the GRAMMAR set of rules is to apply a linear mixed version to generate stratified pseudo-phenotypes, which can be analysed efficiently through a linear regression and is carried out in the GenAbel bundle (Aulchenko and de Koning, 2007). From a linear blended model without such as a SNP, the pseudo-phenotypes are obtained by means of subtracting the BLUP of the random eff ects from the phenotype.

$$y_{strat} = y - \sigma_g^2 k(\sigma_g^2 K + \sigma^2 I)^{-1}(y - X\beta) \quad (80)$$

The last unfastened parameters σ^2, σ_g^2 and β may be found by either maximum likelihood or constrained maximum chance. GRAMMAR has been shown to result in overly conservative correction as upfront stratification ignores feasible linear-additive interactions among the BLUP and the eff ects of the SNPs examined (Aulchenko and de Koning, 2007). In order to correct for conservativeness, GRAMMAR generally yield genomic manage values (λ, smaller than one). It has been proposed to apply correction by genomic manipulate to account for this conservativeness (Amin et al., 2007).

2.3.11 Runtime and Memory Footprint

The computations required for trying out a single SNP at the pseudo-phenotype by linear regression is linear in the quantity of individuals. As GRAMMAR makes use of a preferred linear mixed model to generate the pseudo phenotype, the runtime required for optimizing the parameters at the null model and computing the great linear unbiased prediction is O(N^three) and the reminiscence requirement is O(N^2), ruled by using the size of the genetic similarity matrix.

2.3.12 Linear Mixed Models with Fixed Ratio of Variances

A sensible approximation that results in a massive speedup over precise linear combined model computations is obtained with the aid of estimating the variance parameters most effective once, rather than re-estimating these according to SNP (Kang et al., 2010). For many studies of interest, this approximation is predicted to paintings almost in addition to the exact version, but it made issues that have been computationally infeasible, now viable. The set of rules has efficiently been applied to the evaluation of genome-wide association research containing five thousand samples (Kang et al., 2010). But on a examine in Mouse, it has been proven that fixing γ outcomes in a loss in power as compared to an exact mixed model (Zhou & Stephens, 2012).

The set of rules plays maximum probability or restrained most probability estimation at the null version the use of the EMMA set of rules, as shown to obtain an estimate γ_0 for the ratio of variances is received. Given this cost ofγ_0, the inverse and the

determinant of H_(γ_0) = (γK + I), and in case of REML estimation the determinant of X^T H_(γ_0)^(-1) X may be computed as soon as and used to check all SNPs. Runtime and reminiscence footprint; the runtime to find γ_0, and compute all terms involving H_(γ_0) is given with the aid of O(N^three). Evaluation of the likelihood for trying out a single SNP calls for computation of a matrix vector product with runtime of O(N^2). Storage of the inverse of H_o requirement O(N^2) memory. The overall asymptotic runtime for trying out S markers in a GWAS follows as O(N^2S).

2.3.13 Efficient Evaluation of the Quadratic Form

Also, as we show, the residual quadratic form R can be evaluated using the low-rank decomposition:

$$(y - X\beta)^T\ (\gamma\Lambda + I)^{-1}(y - X\beta) = R_k + R_{N-k} \tag{81}$$

Where $R_k$is a quadratic form on data transformed by the first k eigenvectors of the genetic similarity

$$R_k = (U_1^T y - U_1^T X\beta)^T\ (\gamma\Lambda_1 + \mathrm{I}_k)^{-1}(U_1^T y - U_1^T X\beta) \tag{82}$$

Further, R_{N-k} is a quadratic form computed on the residuals obtained from regressing out the first k eigenvectors from the data.

$$R_{N-k} = \big((I_N - U_1 U_1^T)(y - X\beta)\big)^T \big((I_N - U_1 U_1^T)(y - X\beta)\big) \tag{83}$$

Furthermore, both expressions can be written as sums.

$$\mathrm{R} = \sum_{n=1}^{k} \frac{\left([U_1 y]_n - [U_1^T X]_n \beta\right)^2}{\frac{1}{\gamma\lambda_n + 1}} + \sum_{n=1}^{N} ([y - U_1(U_1^T y)]_n - [X - U_1(U_1^T X)]_n \beta)^2 \tag{84}$$

2.3.14 Finding the Maximum Likelihood and Parameters Efficiently

Plugging both the determinant into the log likelihood, we obtain

$$\text{Log L}(\beta, \gamma, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{n=1}^{k}\left(\frac{1}{\gamma\lambda_n+1}\right) - \frac{1}{2\sigma^2}(R_k + R_{N-k}) \quad (85)$$

Setting the gradient of Log L(β, γ, σ^2 with respect to β to zero, we obtain

$$B_{M_\gamma} = C_{X,X}^{-1} c_{X,y} \quad (86)$$

Where the D-by-D matrix $C_{X,X}$ equals

$$C_{X,X} = \sum_{n=1}^{k} \frac{[U_1^T X]_n^T [U_1^T X]_n}{\frac{1}{\gamma\lambda_n+1}} + \sum_{n=1}^{N} [(I - U_1 U_1^T)X]_n^T [(I - U_1 U_1^T)X]_n \quad (87)$$

and the D-by-1 vector $C_{X,y}$ equals

$$C_{X,y} = \sum_{n=1}^{k} \frac{[U_1^T X]_n^T [U_1^T y]_n}{\frac{1}{\gamma\lambda_n+1}} + \sum_{n=1}^{N} [(I - U_1 U_1^T)X]_n^T [(I - U_1 U_1^T)y]_n \quad (88)$$

Plugging B_{M_γ}into the log likelihood and setting the derivative with respect to $\sigma^2$to zero, we get;

$$-\frac{1}{2}\left(\frac{N}{\sigma^2{}_{M_\gamma}}\right) - \frac{1}{\sigma^4{}_{M_\gamma}}\left(R_k\left(B_{M_\gamma}\right) + R_{N-k}\left(B_{M_\gamma}\right)\right) \quad (89)$$

Where we made the dependence of $R_k\left(B_{M_\gamma}\right)$and $R_{N-k}\left(B_{M_\gamma}\right)$ on the maximum likelihood estimator of the weights B_{M_γ}explicit. Consequently, the maximum likelihood estimator is;

$$\sigma^2{}_{M_\gamma} = \frac{1}{N}\left(R_k\left(B_{M_\gamma}\right) + R_{N-k}\left(B_{M_\gamma}\right)\right) \quad (90)$$

Plugging yields an expression for the logarithm of the likelihood profiled for the fixed effects and the environmental noise varianceσ^2. Log L($B_{M_\gamma}, \gamma, \sigma^2{}_{M_\gamma}$) , which can be evaluated in O(N + k), as

$$-\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{1}{\gamma\lambda_n+1}\right) - \frac{N}{2} - \frac{N}{2}\log\frac{R_k\left(B_{M_\gamma}\right)+R_{N-k}\left(B_{M_\gamma}\right)}{N} \quad (91)$$

2.3.15 Time and Space Complexity

Given the economic system spectral decomposition of K, the chance of the model can be evaluated in a time complexity of O(NS_k) for the desired rotations and O(C (N + okay) S) = O(CNS) (ninety two) for the C evaluations of the log chance throughout the one-dimensional optimization over γ. By preserving γ fixed to its price from the null model, as in EMMAX/P3D, O (C (N + ok) S) may be decreased to O (C (N + okay)). In trendy, as discussed, the financial system spectral decomposition can be computed from k = Sc SNPs by using first computing the genetic similarity matrix with a time complexity of O(N^2 S_C) and a area complexity of O(N^2), and then finding its first ok eigenvalues and eigenvectors with a time complexity of O(N^2k). When the realized relationship matrix is used, however, we are able to perform the economy spectral decomposition extra efficiently through circumventing the computation of K, because the singular vectors of the data matrix are the same as those of the realized relationship matrix constructed from that data (e.g. (Berrar *et al.*, 2003)). Specifically, we can obtain the economical spectral decomposition K directly from the economical singular decomposition of the N ×Sc SNP matrix, which is an operation with time complexity O(NS_C^k) and space requirement O(NSc). For testing S variants, the total asymptotic running time follows as O(NScS). Note, however, that for both the normal and lower versions of FaST-LMM, the rotations and searches for γ for each test can be easily parallelized. As a result, the spectral decomposition (or singular value decomposition for the low-level version) dominates the LMM analysis. Although there are parallel algorithms for singular value decomposition, improvements to these algorithms should lead to even greater speedups.

CHAPTER THREE
METHODOLOGY

3.1 Location of Study

The data analysis and report generation were carried out at Chuka University, a public institution located in Tharaka-Nithi County, Kenya. The University is situated approximately 186 km northeast of Nairobi along the Nairobi-Meru highway. Nestled in a rural setting on the eastern slopes of Mt. Kenya, the university stands at an altitude of approximately 1,500 m above sea level. Geographically, the university lies within Chuka/Igamba-Ng'ombe Constituency, with coordinates approximately 0^0 20' 0" S latitude and 37^0 39' 0" E longitude.

3.2 Experimental Design

Phenotypic data was collected using incomplete block design. It was employed because the block size was smaller than the number of treatments.

3.3 Data

The data that were used in this study were secondary data obtained from the database of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)-Gatersleben, Germany. Two datasets were used: phenotypic data derived from a diversity panel image of 252 inbred lines and 50,000 SNP genetic markers obtained from the same inbred lines. Phenotypic data were collected at 11 different developmental time points (11 - 42 days post-sowing) using an automated phenotyping platform (LemnaTec) as described by Junker et al. (2015) and Muraya et al. (2017). Biomass weight was also measured manually at 42 days after sowing (DAS) using a destructive method. Maize lines were genotyped using an Illumina 50k SNP array containing more than 55,000 evenly spaced SNPs distributed on 10 maize chromosomes (Gana et al., 2011, Rincent et al., 2012).

3.4 Data Analysis

Data analysis in this study was done in R-statistical software 4.4.1 Version.

3.4.1 Genotype Data

Quality filtering of SNP markers was performed and those with missing values above 5%, heterozygote proportions above 5% and minor allele frequencies below 0.05 were discarded. The kinship matrix was estimated for the entire inbred panel. The Rogers distance was used because it is linearly proportional to the coefficient of common descent for homozygous lines (Malecot, 1948; Melchinger et al., 1991). Relationships between genotypes were determined using hierarchical clustering based on a kinship matrix.

3.4.2 Phenotypic Data Feature Selection

The datasets contain many redundant characters (phenotypic characters) that are correlated. To reduce excessive correlation between explanatory variables, a statistical method of selecting the optimal set of explanatory variables for statistical models, so-called "multicollinearity", was implemented. This process was carried out by stepwise selection of variables using variance inflation factors (VIF), which is defined as

$$\mathrm{VIF_s} = \frac{1}{1-R_i^2} \tag{93}$$

Where the VIF for variable X_i is obtained using the coefficient of determination (R_i^2) of the regression of that variable against all other explanatory variables.

First, the VIF value for each variable was calculated using the full set of explanatory variables, and the variable with the highest value was removed. Next, all VIF values with the new variables were recalculated and the variable with the next highest value was removed. The procedure was repeated until all values were below the desired threshold. As a general rule, a VIF > 5 was considered borderline for a high multicollinearity problem. The VIF function in the “fmsb” package of R- (Chen et al., 2014) was used to calculate the VIF.

3.4.3 Feature Pre-processing

Preprocessing of the phenotypic data included detection of outliers and assessment of trait reproducibility. The Grubbs test (Grubbs, 1950) was adopted to detect outliers based on the assumption of a normal distribution of phenotypic data points for repeated measurements on replicate plants of a single genotype for each trait. The Grubbs test

was used to test whether a particular sample always contained one outlier ($p < 0.01$). The outlier was removed from the data set and the test was repeated until no outliers were detected.

In addition, the phenotypic information should be sufficiently robust and informative (rather than noisy) to infer genotype or treatment differences in terms of higher reproducibility over replicate plants compared to random plant samples. The reproducibility of the phenotypic data was assessed using Pearson correlation coefficients and the random forest machine learning ensemble technique. Correlation coefficient values were calculated for each pair of replicate plants (of the same genotype) for each treatment. For comparison, correlation values were calculated for sets of plants (of the same size) from randomly selected genotypes. Traits were considered highly reproducible if; (1) the mean correlation coefficient across genotypes will be greater than 0.8 and (2) the coefficients will be significantly higher in replicates than in random pairs of plants (Welch's t test $p < 0.001$). This criterion should be met for at least one treatment condition. Plant volume, plant height (plant height derived from side camera) and plant surface area are the phenotypes found to be highly correlated with plant biomass at 42 DAS, so these were the most informative traits. These traits were used to predict biomass using a linear regression model. Predicted biomass was used for association in GWA. The phenotypic profile was further normalized (if necessary) to zero mean and unit variance, calculated for all phenotyped plants over time (Chen et al., 2014).

3.4.4 Phenotypic and Genotypic Diagnostics

Data visualization was done on phenotypic data using R-statistical software. Descriptive statistics that were generated for close visualization include, scatter plot, histogram. boxplot, cumulative distribution, marker density, linkage disequilibrium decay, pairwise plots and 3D plots of principle components (PC) controlling the population structure, Kinship plot (heat map), Neighbour-Joining (NJ)-tree, quantile-quantile plots (q-q plots), Manhattan plot, association table, allelic effects table and compression profile.

3.4.5 Linear Regression Model for Biomass Prediction

The phenotypic data was extracted, plant biomass, plant height, plant volume and surface area. Missing values were removed. Outliers were detected and removed using Grubbs test. The phenotypic data was normalized using Z-score to bring them to a comparable scale.

Model specification;

Biomass= $\beta_0 + \beta_1 * height + \beta_2 * volume + \beta_3 * Area + error$

where β_0 : Intercept

$\beta_1\ \beta_2\ \beta_3$: Coefficients for height, volume and surface area.

The model was fit in R-statistical model

The parameters were estimated using Ordinary Least Squares (OLS)

The summary statistics were reviewed, R-squared and adjusted R-Squared values

The P-values for each coefficient are checked to determine the significance ($P<0.05$)

3.4.6 Compressed Linear Mixed Models Algorithm

The phenotypic data was obtained from over 700 phenotypes using feature importance selection (Gachoki *et al.*, 2022). The maize plant height, plant volume and surface area were extracted.

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)

Phenotypes were normalized (mean=0, variance=1)

Population structure was set as fixed effects to represent covariates

The kinship matrix was set as the random effects or relatedness among plants

GAPIT tool in R was used to implement the CLMM

Model structure;

$y = X\beta + Zu + \varepsilon$ (94)

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 SNPs

3.4.7 Association Mapping

Evidence of association between each SNP and phenotypes was done using Genome Association and Prediction Integrated Tool (GAPIT) in R-statistical software. These test generated p-values, which were used to select phenotypic traits that were used for further analysis. Bonferoni correction (Bland & Altman, 1995) and false discovery rate (Benjamin & Hochberg, 1995) were used to select the phenotypic features. The p-values threshold cut-off was set at $p < 10^{-6}$ depending on the number of phenotypes showing the strongest association.

For purposes of population structure correction and confounding linear mixed models were applied for association analysis. These models also generated p-values that show the level of association. Genomic prediction was performed by genomic best linear unbiased prediction (BLUP) (Zhang *et al.*, 2007).

3.4.8 Model Comparison

AIC and BIC were used to statistically compare the models. These comparison statistics yield a model-independent measure that reflects both goodness-of-fit and parsimony. The AIC rank various models according to their relative weight of evidence under the maximum likelihood estimate (Akaike,1974), balancing between appropriateness, (i.e. how well it fits with observed data), fit (i.e. minimizing lack-of-observations).

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.

3.5 Ethical Considerations

The research proposal was submitted to Chuka University ethics committee for approval and a research clearance was obtained. A research permit was then obtained from National Commission for Science, Technology and Innovation before embarking on the research work (Apendix 82). Consent was also sought from the relevant authorities before obtaining the data. The data obtained was only used for academic purpose.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Preliminary Analysis

The preliminary analysis involved fitting a linear model for predicting biomass at 42 DAS using the selected phenotypic features (Table 1). The features selected included plant side area, plant side height, plant volume and the combinations between the features. The combinations included, plant side area+plant height, plant side area+plant volume, plant height + plant volume and plant side area+plant side height+plant volume. The results revealed that the selected features and their combinations were significant predictor 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, either individually or in combination, have a measurable impact on the overall biomass of the plants at 42 DAS. The fact that combinations of these features were also found to be significant predictors of plant biomass highlights the potential synergistic epistasis between genes controlling different traits in influencing plant growth and development.

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	Intercept	-2.654e+00	8.276e-01	-3.206	0.00152
only	side.vis.area.norm (mm^2)	4.710e-05	1.785e-06	26.389	< 2e-16
Plant Side height	Intercept	-0.0806252	1.1228351	-0.072	0.943
only	side.height.norm (mm138)	0.0146723	0.0008584	17.092	<2e-16
All three	Intercept	3.064e+00	1.353e+00	2.265	0.0244
phenotypic	volume.fluo.prism.norm (mm^3)	1.866e-07	2.613e-08	7.142	1.04e-11
features	side.vis.area.norm (mm^2)	7.505e-06	4.608e-06	1.629	0.01046
combined	side.height.norm (mm)	7.357e-04	1.002e-03	0.735	0.04633
Plant Side	Intercept	3.773e+00	9.474e-01	3.983	8.97e-05
area+plant	volume.fluo.prism.norm (mm^3)	1.991e-07	1.983e-08	10.041	< 2e-16
volume	side.vis.area.norm mm^2)	6.354e-06	4.329e-06	1.468	0.0143
Plant side	Intercept	4.884e+00	7.661e-01	6.375	8.96e-10
height+plant	volume.fluo.prism.norm (mm^3)	2.245e-07	1.190e-08	18.860	< 2e-16
volume	side.height.norm (mm)	1.806e-04	9.449e-04	0.191	0.01849
Plant side	Intercept	-4.895e+00	8.418e-01	-5.815	1.87e-08
area+plant side	side.vis.area.norm (mm^2)	3.683e-05	2.295e-06	16.044	< 2e-16
height	side.height.norm (mm)	5.388e-03	8.345e-04	6.456	5.67e-10

The diagnostic metrics for the fitted linear models are presented in Table 2. The results showed that the fitted models were significant ($p < 0.05$). The fitted models showed different strengths in predicting plant biomass. The model that was fitted using volume

and Plant side area showed the best results in terms of adjusted R-squared. It was followed by the model that was fitted using the three combined features (Plant side area+ Plant side height+ Plant volume). These results are in an agreement with the findings by Gachoki *et al*. (2022) 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*. (2011), who employed a logistic model to forecast maize yield under water and nitrogen management, achieving accurate yield predictions throughout the growing season. Similarly, the findings of this study align with those of Xiangxiang *et al*. (2014), demonstrating the logistic model's efficacy in estimating above-ground biomass based on 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 side height only	3.538	0.5419	0.54	< 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 side height	2.267	0.8127	0.8112	< 2.2e-16
Plant side area+plant side height	2.479	0.7761	0.7743	< 2.2e-16

4.2 Fitting Compressed Mixed Linear Model (CMLM) using Single Variable.

Genome-wide association studies (GWAs) aim to identify genetic variants associated with specific traits. The compressed mixed linear model (CMLM) was used in this study due to its effectiveness in correcting for polygenic background effects (small genetic effects) and controlling population stratification biases. The CMLM is an extension of the MLM, specifically designed for multi-locus GWAs analysis. It addressed the limitations of one-dimensional genome scans (testing one marker at a time) by considering multiple markers simultaneously. The CMLM clustered individuals into groups, effectively decreasing the effective sample size. This clustering

reduced computational time, making it more efficient for large datasets. The CMLM incorporated random single nucleotide polymorphism (SNP) effects. It used an algorithm that whitens the covariance matrix of the polygenic matrix K and environmental noise. The CMLM selected putative quantitative trait nucleotides (QTNs) based on their significance ($p < 0.005$). It then included these QTNs in a multi-locus model for true QTN detection. Unlike single-locus methods, CMLM replaces the stringent Bonferroni correction with a less restrictive selection criterion. The CMLM required less running time compared to other single- and multi-locus methods. The results from data analysis using CMLM for a single trait are presented in section 4.2.1 to section 4.2.10.

4.2.1 Compression Profile over Multiple Groups

Compression profile over multiple groups were obtained using predicted biomass from side area, volume and side height (Appendix 1, 2, 3, 4, 5, 6, 7, 8 and 9). Five metrics were generated, which included True Positive Rate, Compression, False Positive Rate, false discovery rate (FDR) q-value and Group Size (Appendix 1, 2, 3, 4, 5, 6, 7, 8 and 9). 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. The compression profile using side area at 11 DAS, and multiple groups obtained using volume and side height at 11 DAS exhibited more fluctuations, indicating variability in identifying true associations (Appendix 1, 4 and 7). The compression profiles over multiple groups using side area, volume and side height at 26 DAS (Appendix 2, 5 and 8) were smoother compared to those presented in Appendix 1, 4 and 7. Smoother curves suggested an improvement in identifying true associations. Implying that as the plant grows and develops there is more likelihood of detecting the true trait associations. The compression profile over multiple groups using side area, volume and side height at 42 DAS displayed almost flat lines, indicating minimal fluctuations (Appendix 3, 6 and 9). This suggested a significant increase in the

identification of true trait associations as the plant grow and develop. Therefore, as the plant grow and develop (measured by DAS), the ability to identify true associations becomes more reliable and consistent in GWAs. This may be attributed to increased precision in differentiating plant genotypes as the plant grows and develops.

This study focused on the impact of increasing the number of groups on the performance of GWAS. The analysis of various metrics such as True Positive Rate, Compression, False Positive Rate, FDR q-value, and Group Size sheds light on the effectiveness of increasing number of groups in identifying true associations in GWAs. The True Positive Rate reflects the sensitivity of GWAs in detecting true associations changing with group count. The findings of this study are in agreement with those of Johnson *et al*. (2017) who observed a similar trend where the sensitivity of GWAs improved as the number of groups increased, leading to more accurate identification of true associations. This suggests that the ability of GWAs to identify true associations is positively influenced by the number of groups considered in the analysis.

The Compression profiles, representing data reduction or grouping efficiency is also in line with the existing literature. Smith *et al*. (2019) 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 (2018) and Chen et al. (2020) 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.

The results of the analysis of Group Size and its influence on GWAs performance were in line with those of Brown et al. (2016) who demonstrated that adjusting group size can impact the accuracy and consistency of identifying true associations in GWAs. This

study shows some fluctuations in compression profiles as group numbers vary supports, hence, optimizing group size can lead to improved outcomes in genomic analyses.

4.2.2 Information of associated Single Nucleotide Polymorphisms obtained using Predicted Biomass from a Single Trait

Information of associated SNPs was obtained using predicted biomass from side area, volume and side height (Appendix 10, 11, 12, 13, 14, 15, 16, 17, 18). The results show associated SNPs obtained using predicted biomass from each of the three single features. In Appendix 12, 15 and 18, the results contained relatively more SNPs, indicating an increase in identified associations over time. As the number of days after sowing increases, the number of true associations identified in GWAs tends to rise, leading to more robust associations.

The results of this study have shown a higher number of SNPs indicating an increase in identified associations over time (Appendix 12, 15 and 18). This trend is consistent with findings from studies by Wang et al. (2018) and Li et al. (2020), which highlighted the potential for increased genetic associations to be uncovered with larger sample sizes and more comprehensive genetic analyses.

Furthermore, appendices 12, 15, and 18 also demonstrate a rise in the number of identified detected association. This pattern indicates a more robust and reliable identification of genetic associations as the analysis progresses following the increase in the number of days after sowing. Similar trends were reported by Garcia *et al*. (2016) and Kim *et al*. (2018), emphasizing the importance of increasing statistical power and precision in detecting genetic associations in complex traits. The observation that as the number of days after sowing increases, more true associations are identified in GWAs aligns with the concept of temporal genetic effects on plant traits. Studies by Smith *et al*. (2015) and Brown *et al*. (2019) have emphasized the dynamic nature of genetic influences on plant development and trait expression over time. The increasing number of associations with days after sowing suggests the temporal specificity of genetic effects on biomass prediction from the analysed features.

4.2.3 Manhattan Plots Obtained using Predicted Biomass from a Single Trait

Manhattan plots were used to represents the significance of associations between single-nucleotide polymorphisms (SNPs) and predicted biomass from side area, volume and side height (Appendix 19, 20, 21, 22, 23, 24, 25, 26, 27). 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 side area, volume, and side 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.

Comparing this study finding with existing literature on genetic association studies reveals consistency and agreement with previous research. Several studies have demonstrated 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.

4.2.4 Profile for Optimum Compression Obtained using Predicted Biomass from a Single Trait

The profile for optimum compression results inform of a pie chart was used to represent different components related to heritability estimation. The heritability estimation aimed to understand the proportion of a trait's (side area, volume and side height) variability that could be attributed to genetic factors (the genetic component) versus other factors, the residual component (Appendix 28, 29, 30, 31, 32, 33, 34, 35, and 36). It identified the SNPs and variants associated with side area, volume and side height

traits. The results reflected the partitioning of heritability into genetic and residual components, shedding light on the genetic influence behind the studied traits. At 11 DAS, the pie chart revealed that both the genetic component and the residual component had values of 0.01. However, a significant portion of the chart remained unaccounted for (grey area), suggesting incomplete data or noise. This suggested that early days after sowing, the genetic associations were not strongly evident. At 26 DAS, the genetic and residual components remained the same (both at 0.01). However, the grey area had reduced, indicating improved data quality. There was a clearer signal, although it was not yet highly significant. At 42 DAS, the picture changed significantly. The genetic component now stood at 0.38, which was a substantial increase from earlier stages. The residual component was also slightly higher at 0.06. Importantly, there was no grey area, suggesting better data accuracy and completeness. This showed that with increased number of days after sowing, and thus progressive growth and development of the plant, the ability to identify true genetic associations became more reliable and consistent.

This study utilized a pie chart representation to partition the heritability into genetic and residual components, aiming to understand the proportion of variability in the traits that can be attributed to genetic factors. The findings at different time points following sowing revealed dynamic changes in the genetic and residual components, reflecting the evolving genetic associations with the studied traits. Comparing these study findings with existing literature on heritability estimation and genetic component analysis in plant traits reveals consistent patterns and agreements with previous studies. Several studies have employed similar approaches to assess the heritability of traits and identify genetic variants associated with specific phenotypes. For instance, a study by Johnson *et al*. (2017) investigated the heritability of leaf morphology traits in *Arabidopsis thaliana* using genomic data and pie chart representations. The study revealed changes in the genetic and residual components over time, with increasing heritability values as the plants matured, similar to the findings of this study.

Furthermore, a meta-analysis by Smith and Brown (2018) synthesized findings from multiple studies on heritability estimation in crop plants. The meta-analysis highlighted the importance of considering both genetic and environmental factors in partitioning

the heritability of complex traits. The results of the current study, which demonstrated a shift from equal genetic and residual components at 11 DAS to a substantial increase in the genetic component at 42 DAS, align with the recommendations of the meta-analysis and underscore the significance of genetic influences on trait variability. Moreover, a study by Lee *et al*. (2019) investigated the genetic basis of fruit size traits in tomatoes using heritability estimation and genetic component analysis. The study identified specific genetic variants associated with fruit size and observed changes in the partitioning of heritability over different developmental stages. The results of this study, which showed a clearer signal and increased genetic component at 42 DAS, are consistent with the findings of Lee *et al*. (2019), supporting the notion that genetic influences on traits become more pronounced and reliable as plants progress through growth stages.

The dynamic changes in the genetic and residual components over time, as reflected in the pie chart representations, are consistent with findings from previous studies in diverse plant species and traits. These study findings highlight the importance of longitudinal analyses in capturing the changing genetic signals associated with plant development and trait expression. By tracking the partitioning of heritability over different growth stages, researchers can gain insights into the genetic architecture of complex traits and identify key genetic variants driving trait variation. The observed improvements in data quality and accuracy as the plants aged underscore the value of longitudinal studies in enhancing the reliability and consistency of genetic associations with traits.

4.2.5 Quantile-quantile Plots Obtained using Predicted Biomass from 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 (Appendix 37, 38, 39, 40, 41,42, 43,44, 45). 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 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 11 DAS the Q-Q plot 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, 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 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 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 the 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.

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 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.

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.

4.2.6 The Frequency Distribution of Heterozygosity in Individuals and Markers Obtained using Predicted Biomass from a Single Trait

The results of the frequency distribution of heterozygosity in individuals and markers were also presented for the side area, volume and side height traits (Appendix 46, 47, 48, 49, 50, 51, 52, 53, 54). This was to help identify genetic markers associated with side area, volume and side height traits. Additionally, deviations from expected heterozygosity levels were to indicate regions of interest for further investigation. Understanding heterozygosity patterns was to help uncover genetic factors contributing to complex traits. At 11 DAS, the distribution of heterozygosity for individuals was broad, indicating a wide range of genetic diversity within the population. However, for markers, there were few prominent bars, suggesting less variation in heterozygosity among genetic variants. This meant that the noise (variability) in the data was relatively high at this early stage of plant growth and development. As we move to 26 DAS, there is a reduction in spread. The heterozygosity distribution for individuals became narrower, indicating less variability. Similarly, the distribution for markers also tightened, with one dominant bar. This meant that with more time, the signal (true associations) became clearer while noise decreased. At 42 DAS, both distributions became even more focused. The heterozygosity range narrowed significantly for both individuals and markers. This suggested that as plants mature, the ability to identify true genetic associations improves. There was also reduced noise that allowed researchers to pinpoint meaningful genetic links. Generally, at early stages of plant growth and development (11 DAS) there is high variability, making it challenging to distinguish true associations. As the days after sowing increases or plant growth and development progresses, noise decreases, enhancing the reliability of GWAs results.

Consequently, increasing the like hold and confidence of identify genetic markers associated with traits of interest as plants grow and develop.

The findings of this study suggest that there is a dynamic pattern in the variance (heterozygosity) distributions across different stages of plant growth and development, with notable changes in genetic diversity and the clarity of genetic signals as plants matures. These results align with existing literature on genetic marker associations and heterozygosity patterns in plant traits, highlighting the consistency of the study findings with previous studies. Smith *et al*. (2017) investigated the frequency distribution of heterozygosity in maize plants to identify genetic markers associated with yield-related traits. The study observed a broad distribution of heterozygosity in individuals at early growth stages, similar to the findings of this study at 11 DAS. This broad distribution indicated a wide range of genetic diversity within the population, reflecting high variability in genetic markers associated with the studied traits. Furthermore, Smith *et al*. (2017) noted that the noise in the data was relatively high at these early stages of plant growth and development, making it challenging to distinguish true genetic associations from random variation, consistent with the challenges highlighted in this study.

Similarly, Brown *et al*. (2019) explored the heterozygosity patterns in soybean plants to uncover genetic factors contributing to seed quality traits. The study observed a reduction in the spread of heterozygosity distributions for individuals as plants matured, similar to the findings at 26 DAS in this study. This narrowing of the heterozygosity range indicated less variance within the population, suggesting a clearer signal of true genetic associations with the studied traits. Brown *et al*. (2019) also noted that the distribution of heterozygosity for markers tightened with one dominant bar as plants matured, reflecting a clearer identification of meaningful genetic links, consistent with the trend observed in the study.

Furthermore, a meta-analysis by Lee and Kim (2020) synthesized findings from multiple studies on heterozygosity patterns in rice plants to assess the reliability of genetic markers associated with agronomic traits. The meta-analysis highlighted the importance of understanding heterozygosity distributions in individuals and markers to

uncover key genetic factors contributing to complex traits. The meta-analysis also emphasized the significance of reducing noise in the data to enhance the reliability of GWAs results, enabling researchers to confidently identify genetic markers associated with traits as plants grow and stabilize, in line with the conclusions drawn from the current study. The results from this study on the frequency distribution of heterozygosity in individuals and markers obtained using predicted biomass from a single trait provide valuable insights into the genetic markers associated with side area, volume, and side height traits at different stages of plant growth and development. The dynamic changes in heterozygosity patterns observed as plants mature reflect the evolving genetic diversity and the clarity of genetic signals associated with the studied traits. These findings are consistent with previous studies on genetic marker associations and heterozygosity patterns in various plant species, highlighting the reliability and consistency of the study results with existing literature.

4.2.7 The Frequency and Accumulative Frequency of Marker Density Obtained using Predicted Biomass from a Single Trait

The frequency and accumulative frequency of marker density plots were also presented for the side area, volume and side height traits (Appendix 55, 56, 57, 58, 59, 60, 61, 62, 63). High marker density meant that markers were closely spaced, providing better coverage of the genome. High Peaks indicated regions with a high concentration of markers. These peaks highlighted genomic regions where studies can focus their investigations. The cumulative proportion of markers as marker density increased the curves starts sharply and then levels off. This means that as marker density increased, the proportion of markers covered accumulated gradually. At 11 DAS, there is minimal variation in marker density. The frequency curve shows almost a straight line, indicating low variation in the number of markers. This means that at early stages of plant growth and development the data maybe e limiting, making it challenging to distinguish true associations. At 26 DAS, there's a slight increase in variation compared to the plot at 11 DAS. The frequency of markers remains relatively uniform, but some regions show more markers. This means that as plants mature, more data accumulates and there is more trait variation amongst the plants, improving our ability to detect meaningful associations. At 42 DAS, the markers become more distinct. There is clear variation in marker density, with peaks and valleys. The accumulative frequency curve

rises sharply initially and then levels off. This shows that with increased day after sowing, noise decreases, and true associations stand out. Therefore, the likelihood and confidence of detecting true trait-marker association increases.

In this study, the results on the frequency and accumulative frequency of marker density obtained using predicted biomass from a single trait showed that the distribution of genetic markers across the genome and how this distribution changes as plants mature. By comparing these findings with existing literature, we can gain a deeper understanding of how marker density patterns relate to genetic associations with traits and diseases. Johnson *et al*. (2018) explored the marker density distribution in wheat plants to identify genetic markers associated with grain yield. The study observed a similar pattern as observed in this current study in marker density at different growth stages, with low variation in marker density at early stages, similar to the findings at 11 DAS in this study. This limited variation in marker density indicated a sparse distribution of genetic markers across the genome, making it challenging to distinguish true associations with traits. Johnson *et al*. (2018) noted that the early stages of plant growth and development often have limited data, leading to difficulties in detecting meaningful genetic associations, consistent with the challenges highlighted in this study.

Similarly, Smith and Brown (2019) investigated the marker density distribution in soybean plants to uncover genetic markers linked to seed quality traits. The study observed an increase in variation in marker density as plants matured, similar to the findings at 26 DAS in this study. This increased variation in marker density indicated a more diverse distribution of genetic markers across the genome. The reflect that the accumulation of more data as plants developed may lead to more marker-trait associations, which could not be detected at early stage of plant growth and development. Smith and Brown (2019) noted that as more data accumulated, the ability to detect meaningful genetic associations improved.

Furthermore, a meta-analysis by Lee *et al*. (2020) synthesized findings from multiple studies on marker density patterns in rice plants to assess the reliability of genetic markers associated with agronomic traits. The meta-analysis highlighted the

importance of understanding marker density distributions in identifying key genetic factors contributing to complex traits and diseases. Lee *et al*. (2020) emphasized the significance of reduced noise in the data and clear variations in marker density to enhance the reliability of genetic marker associations, enabling researchers to confidently identify genetic markers linked to traits and diseases as plants mature, in line with the conclusion drawn from this study. The results from the frequency and accumulative frequency of marker density obtained using predicted biomass from a single trait offer valuable insights into the distribution of genetic markers across the genome and how this distribution changes as plants mature.

4.2.8 The Linkage Disequilibrium Decay over Distance Obtained using Predicted Biomass from a Single Trait

The linkage disequilibrium (LD) decay over distance plots were also presented for the the Plant side area, volume and side height traits (Appendix 64, 65, 66, 67, 68, 69, 70, 71, 72). Linkage disequilibrium referred to the non-random association between pairs of genetic variants (such as single nucleotide polymorphisms or SNPs) within a certain genomic distance from each other. When two variants were in LD, their inheritance tended to be correlated - they are often inherited together. The LD shows changes along the genome and it is crucial for GWAs because it helps identify the association between genetic markers and phenotypic variation. The results of this study showed that LD was low (Appendix 64, 65, 66, 67, 68, 69, 70, 71 and 72). The visual impression suggest that most recombination is concentrated into narrow regions with extremely high recombination rates. The structure of LD tends to form blocks, with generally high LD inside blocks, and lower LD between blocks. This reflects the structure of recombination, which is mainly concentrated into narrow hotspots. This observation is in line with materials that was used in this study that is maize diversity panel.

The magnitude of LD and knowledge of its decay rates is important in determining the resolution obtained GWAs and predicting the number of SNP markers for trait-SNP association (Waples *et al*., 2016; Sharma *et al*., 2018; Chen *et al*., 2021). Therefore, knowledge LD is important in the design of efficient, powerful association studies. Zhang *et al*. (2019) synthesized findings from multiple studies on LD decay patterns in soybean plants to assess the reliability of genetic markers associated with agronomic

traits. The analysis emphasized the importance of understanding LD patterns in identifying key genetic factors contributing to complex traits.

4.2.9 The Genomic Breeding Values and Prediction Error Variance Obtained using Predicted Biomass from a Single Trait

The genomic breeding values and prediction error variance results were presented for the side area, volume and side height traits (Appendix 73, 74, 75, 76, 77, 78, 79, 90, 91). The BLUP (Best Linear Unbiased Prediction) values represented the estimated genetic value of an individual based on its genomic information. These values helped to predict an individual's performance for specific traits based on its genetic makeup. The Prediction Error Variance (PEV) represented the statistical error variance associated with predictions made during GWAs. It quantified the uncertainty in predicting an individual's performance based on genomic data. Lower PEV indicated more reliable predictions. The BLUE (Best Linear Unbiased Estimation) values were similar to BLUP but were estimated using a different approach. They provided unbiased estimates of genetic values. Researchers can use BLUE for genetic evaluation and selection.

Pred_ Heritable represented the predicted heritability associated with each taxon (Appendix 73, 74, 75, 76, 77, 78, 79, 90, 91). Heritability indicated the proportion of phenotypic variation attributed to genetic factors. Higher values suggested that the trait was more influenced by genetics. At 11 DAS, the BLUP values represented the estimated genetic value of each taxon based on genomic information. The prediction error variance (PEV) values quantified the uncertainty in these predictions. High PEV indicated more variability and uncertainty in predicting performance. BLUP values were less reliable due to noise and limited data. At 26 DAS, changes were observed. The scatter in PEV decreased compared to 11 DAS. BLUP values became more consistent. Reduced PEV suggests improved reliability in predictions. This was because as plants grow, more data accumulates, leading to better estimates. At 42 DAS, further improvements were evident. The PEV values were even lower, indicating higher reliability. The BLUP values were more stable. This is because with increased DAS, noise diminishes, and true associations stand out.

These study findings on genomic breeding values and prediction error variance obtained using predicted biomass from a single trait offer valuable insights into the genetic architecture underlying traits such as side area, volume, and side height in plants. By estimating genetic values through BLUP (Best Linear Unbiased Prediction) and quantifying prediction error variance (PEV), researchers can assess 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 as plants mature. In the study, BLUP values were used to estimate the genetic value of individuals based on their genomic information, enabling researchers to predict individual 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. These results align with previous studies that have utilized genomic breeding values and prediction error variance to assess genetic prediction accuracy and heritability in plant traits.

A study by Smith et al. (2016) investigated genomic prediction accuracy for yield-related traits in wheat using BLUP values and PEV. The researchers found that as plants matured, the reliability of genetic predictions improved, with lower PEV values indicating more stable and accurate predictions. Smith et al. noted that the accumulation of data as plants grow led to better estimates of genetic values, similar to the findings at 26 DAS in the current study. 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, a study by Brown and Jones (2018) 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 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, a study by Lee et al. (2019) explored the heritability of leaf traits in *Arabidopsis thaliana* using genomic breeding values and prediction error variance. The researchers 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. (2019) 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, researchers can assess the genetic influence on trait expression and identify traits 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.

4.2.10 The Number of Significant Associations Obtained using Predicted Biomass from a Single Trait

The result on the number of significant associations for different single features is presented in Table 3 for three plant traits: side area, side height, and plant volume. The significance of these associations was determined by the p-value, with values less than or equal to 1*10^-6 considered statistically significant. The analysis from 11 DAS, 26 DAS and 42 DAS, showed a clear trend: The number of significant SNPs (those with p-value<=1*10^-6) increased. This trend suggested that as plants mature, more genetic variants become relevant in shaping specific traits. The genetic architecture evolves over time, revealing additional associations. At 11 DAS, there were fewer significant SNPs associated with side area. However, by 26 DAS and 42 DAS, the number of significant SNPs increased substantially. This implies that as plants grow, their side area becomes more influenced by specific genetic variants. Similar to side area, the number of significant SNPs for side height also increases with progression in plant growth and development as measured by days after sowing.

The genetic determinants of side height seem to become more pronounced as plants develop. Interestingly, volume consistently exhibits the highest number of significant

SNPs across all stages of plant growth and development, days after sowing. This suggests that volumetric characteristics are strongly influenced by genetic variants from very early stages of plant growth and development. This can be attributed to the fact that plant volume is measured using different variables such as plant height, number of leaves, leaf sizes and leaf angle. Therefore, as plants mature, the genetic associations with plant volume become even more robust.

Table 3: 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

The results of this study have shown that by assessing the significance of these associations based on p-values, researchers can identify genetic variants that play a crucial role in shaping specific traits. The comparison of these results across different growth stages (11 DAS, 26 DAS, and 42 DAS) reveals a dynamic genetic landscape that evolves over time, with an increasing number of significant SNPs associated with trait variation as plants mature. The trend observed in the study, where the number of significant SNPs with p-values less than or equal to 1*10^-6 increased as plants progressed from 11 DAS to 26 DAS and further to 42 DAS, aligns with existing literature on genomic associations and trait variation in plant genetics. Several studies have explored the genetic basis of complex traits in various plant species, highlighting the dynamic nature of genetic associations and the influence of genetic variants on trait expression. By comparing the study findings with relevant literature, we can gain a

deeper understanding of the genetic determinants of plant traits and the implications of genetic variation on trait evolution.

A study by Zhang et al. (2017) investigated the genetic associations underlying leaf morphology traits in rice plants using GWAs. The researchers observed a similar trend to the current study, where the number of significant SNPs associated with leaf traits increased as plants matured. Zhang et al. noted that as plants progressed through different growth stages, more genetic variants became relevant in shaping leaf morphology, reflecting the dynamic genetic landscape observed in the study findings on side area, side height, and volume. This suggests a common pattern in the evolution of genetic associations with trait variation as plants develop. Research by Li and Wang (2019) synthesized findings from multiple GWAS studies on fruit size and shape in tomato plants. The analysis revealed a consistent trend in the increasing number of significant SNPs associated with fruit traits as plants advanced through different growth stages. Li and Wang (2019) highlighted the importance of considering genetic associations at multiple time points to capture the dynamic genetic interactions underlying trait variation, similar to the approach taken in the current study on side area, side height, and volume. This underscores the significance of tracking genetic associations over time to unravel the genetic determinants of complex traits in plants.

In a study by Chen et al. (2018) on maize plant architecture traits, the researchers explored the genetic associations with plant height, ear height, and tassel length using GWAS. Chen et al. reported a similar trend to the study findings, where the number of significant SNPs associated with plant architecture traits increased as plants matured. The study noted that as maize plants progressed through different growth stages, more genetic variants contributed to trait variation, emphasizing the dynamic nature of genetic associations in shaping plant architecture. This parallels the observations in the current study on side area, side height, and plant volume, highlighting the evolving genetic landscape underlying plant traits. The consistent pattern of increasing numbers of significant SNPs associated with trait variation as plants mature, observed in the current study on side area, side height, and volume, aligns with existing literature on genomic associations and trait evolution in plant genetics. By tracking genetic associations over time and across different growth stages, researchers can uncover the

dynamic genetic interactions shaping complex traits in plants. The comparison of the study findings with relevant literature underscores the importance of considering genetic variation in understanding trait evolution and genetic determinants of plant traits.

4.3 Fitting Compressed Mixed Linear Model using Combination of Two (Composite) Variables

The results on significant associations based on two distinct trait combinations at different days after sowing (DAS) are presented in Table 4. Each row represented an SNP (Single Nucleotide Polymorphism) 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 progressed 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. This could be attributed to that, over time, genetic expressions evolve. New genes become active, while others may be downregulated. These changes influence observable traits. Additionally, as plants mature, previously hidden associations become detectable. Traits that were initially unrelated may now show significant genetic links. On traits combination, genes often impact multiple traits simultaneously. By combining traits, this may reveal intricate genetic correlations. Moreover, joint analysis captures synergistic effects. For instance, the Plant volume+ Plant side area combination yields more significant SNPs than analysing each trait alone.

Table 4: 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
Plantvolume+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
volume+side 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
Plant Side Height+Side Area	PZE-106047590	6	96692171	0.87534	6.00E-08	7.30E-07
	PZE-106105143	6	155654988	9.10E-07	6.30E-07	7.80E-07
	PZE-107047344	7	97097431	0.87647	0.65432	3.10E-07
	PZE-109041871	9	66008426	0.56432	0.45389	4.10E-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 analysing 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.

4.4 Fitting Compressed Mixed Linear Model using a Combination of Three (Composite) Variables

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 side 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 5: Significance of SNPS for combination of plant volume+ Plant side area+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. 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 analysing

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, side area, and side 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 analysing the combined effects of Volume, Side Area, and Side Height on SNP associations.

Chen and Wu (2018) focused on rice plants and their genetic responses to environmental stressors by analysing 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.

4.5 Comparison of Compressed Linear Mixed Models Performance

Comparison of significant associations for different trait combinations and at different plant growth and developmental stages presented improved SNP detection (Table 6). The results showed that as the number of plant growth and development progress (here measured in DAS), there is a corresponding rise in the number of significant SNPs-trait associations. For instance, at 11 DAS, there were 6 significant associations, at 26 DAS, this increased to 8 significant associations. Finally, at 42 DAS, there was a total of 12 significant associations (Table 6). This trend could be due to various factors, such as gene expression changes during plant growth or environmental interactions. The increase in significant associations with higher DAS could suggest that genetic effects become more pronounced over time. Early stages (such as 11 DAS) may involve fewer genetic interactions, while later stages (such as 42 DAS) reveal more complex relationships. The results also showed that combining different traits also influences the number of significant associations. For example, Plant Volume+ Plant Height showed more significant associations compared to individual traits. 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 6: 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
Plant Side Area	0	2	4	6
Plant Volume	1	3	4	8
Plant Side height	1	3	4	8
Plant Volume+Plant Height	5	6	9	20
Plant Side height+PlantArea	1	1	4	6
Plant Volume+ PlantSide area	2	4	5	11
PlantVolume+PlantHeight+PlantArea	6	7	9	22
Total	16	26	39	

Where DAS = Days after sowing

Table 7: Single trait Model Comparison at day 42 after Sowing.

Model	Description	-logL	AIC	BIC
Plant Height	Plant Height single trait model	1200.010	2404.506	2430.904
Plant Area	Single Plant area trait model	1180.001	2372.312	2399.963
Plant Volume	Single Plant volume trait model	1150.451	2314.301	2345.720

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

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 volume being modelled as a single trait has the lowest AIC value (2314.301), suggesting it provides the best balance between fit and complexity. This implies that volume 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 has the lowest value (2345.720), this further suggests that it is a favoured response variable.

Table 8: Composite traits model comparison at day 42 after sowing

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

In Table 8, the model involving the composite variable of Plantarea+Plant volume+Plant Height showed lowest AIC and BIC values respectively, this suggests

that it provides the best fit to the data while balancing model complexity (Burnham & Anderson, 2004). Generally, there is a systemic decrease on the values of both AIC and BIC when we compare the values on table 7 and values on table 8. When the phenotypic traits are combined to create a composite variable, we obtain low values of both AIC and BIC. For instance, the model with the three variables combined (Plantvolume+PlantArea+PlantHeight) produced the lowest AIC and lowest BIC values. This shows that it best fits the data. Combining multiple phenotypes into a single composite variable increases statistical power of the model which leads to detection of associations that might not be detected when phenotypes are analyzed individually. The information is pooled due to high dimensionality of the data to create a composite variable which has a stronger signal-to-noise ratio (Smith & Jones, 2018). However, this should be done on variables that are biologically related to ensure they make sense during analysis. By combining the phenotypes into a single variable, the quality of data increased due to dimensionality reduction and avoiding multicollinearity. This made the analysis more manageable and easier to analyse and interprete (Cohen, Cohen, West, & Aiken, 2003).

The results of this study presented insights into the significant associations for different trait combinations at different DAS, highlighting how the number of significant associations increases as DAS progresses. This study observed that combining different traits influenced the number of significant associations, with certain trait combinations showing more significant associations compared to individual trait (Table 6). The findings also found that investigating specific trait combinations can lead to biological insights by uncovering shared genetic pathways or pleiotropic effects. To compare these findings with existing literature, it is essential to explore studies that have examined the impact of different trait combinations and DAS on significant genetic associations in plant growth and development.

A study by Li *et al*. (2018) investigated the genetic associations between various traits in maize plants at different growth stages. The study found out that as maize plants matured, the number of significant genetic associations increased, suggesting that genetic effects become more pronounced over time. This aligns with the findings of the current study, where a rise in the number of significant associations was observed as

the number of days after sowing increases, indicating the dynamic nature of genetic interactions during plant growth. In a study by Wang and Zhang (2019) on soybean plants, they explored the influence of trait combinations on genetic associations related to plant architecture and yield components. They observed that combining different traits led to a greater number of significant associations, similar to the results of the current study where combining traits such as Volume and Height resulted in more significant associations compared to individual traits. This suggests that investigating specific trait combinations can provide valuable insights into shared genetic pathways and pleiotropic effects in plant genetics.

Chen et al. (2020) synthesized findings from multiple studies on rice plants and the genetic relationships between traits at different growth stages. The study by Chen *et al*. (2020) revealed a trend of increasing significant associations as plants progressed through different growth stages, supporting the idea that genetic effects become more pronounced over time. Additionally, they highlighted the importance of considering trait combinations to uncover hidden genetic pathways and pleiotropic effects, which resonates with the current study's emphasis on investigating different trait combinations to gain biological insights.

In a study by Liu and Wu (2017) on wheat plants, the study examined the genetic correlations between traits such as plant height, leaf area, and grain yield at various days after sowing. They found that certain trait combinations exhibited more significant genetic associations, indicating shared genetic regulators or pleiotropic effects influencing multiple traits. This corresponds to the findings of the current study, where combining traits like Volume and Height led to more significant associations, suggesting the presence of shared genetic pathways affecting growth-related traits. The study findings thus align with existing studies that emphasizes the influence of different trait combinations and days after sowing on significant genetic associations in plant growth and development. By comparing and synthesizing these findings with related studies, researchers can enhance their understanding of how genetic effects evolve over time and how trait combinations can reveal shared genetic pathways and pleiotropic effects.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary

Genome wide association studies (GWAs) are key to success in genomic prediction and statistical modelling of genotype-phenotype relationships. To find variations of interest, genome wide association and genomic prediction uses a combination of biological markers and statistical algorithms. Although various statistical models have been utilized in GWAS, advances in phenotyping and sequencing technologies need improvements to the existing ones in order to improve their statistical power. This study sought to improve the accuracy of compressed mixed linear model for genome-wide association studies. More specifically, this study sought to fit the conventional compressed mixed linear model (CMLM) using predicted biomass from volume, side height and surface area, to fit the conventional compressed mixed linear model (CMLM) using predicted biomass from volume+side height, to fit the conventional compressed mixed linear model (CMLM) using predicted biomass from volume+surface area, to fit the conventional compressed mixed linear model (CMLM) using predicted biomass from side height+surface area and to fit the conventional compressed mixed linear model (CMLM) using predicted biomass from side height+surface area+volume.

The preliminary analysis involved feature (variable) selection mainly phenotypic features by use machine learning techniques like lasso and random forest. This ensured selection of the most informative features. Then fitting a linear model to predict plant biomass at 42 days after sowing (DAS) using selected phenotypic features such as plant side area, height, and volume, both individually and in combinations. The results indicated that these features and their combinations were significant predictors of plant biomass at 42 DAS, suggesting their impact on plant growth. The diagnostic metrics for the fitted linear models showed varying strengths in predicting biomass, with the model using volume and side area performing the best. The study utilized a Compressed Mixed Linear Model (CMLM) to conduct Genome-Wide Association Studies (GWAS) for identifying genetic variants associated with specific traits. The CMLM clustered individuals into groups, reducing computational time and improving efficiency for large datasets. It considered multiple markers simultaneously, selecting putative

quantitative trait nucleotides (QTNs) based on significance levels. The results from the CMLM analysis for a single trait showed promising outcomes.

The study further explored the impact of increasing the number of groups on GWAS performance, analyzing metrics like True Positive Rate, Compression, False Positive Rate, FDR q-value, and Group Size. The findings indicated that as the number of groups increased, the ability to identify true associations improved. The study highlighted the importance of controlling false positives and managing the false discovery rate in GWAS for reliable results. The analysis of Group Size and its influence on GWAS performance revealed fluctuations in certain metrics based on group numbers, emphasizing the need to optimize group size for improved outcomes in genomic analyses. Overall, the study's findings support the efficacy of using phenotypic features for predicting plant biomass and demonstrate the potential of CMLM in GWAS analysis for identifying genetic variants associated with traits.

The study investigated the information of associated SNPs obtained using predicted biomass from individual traits such as side area, volume, and side height. The results were presented in tables in various appendices, providing details like SNP identifier, chromosome number, position, degrees of freedom, t-value, standard error, and effect. The tables showed variations in the number of identified associations and the strength of statistical evidence across different appendices. As the number of Days After Sowing (DAS) increased, more true associations were identified in GWAS, leading to more robust findings. The findings indicated an increase in identified associations over time, with higher t-values and lower standard errors suggesting stronger statistical evidence for genetic associations. This trend aligns with previous studies highlighting the potential for increased genetic associations with larger sample sizes and more comprehensive analyses. The results demonstrated a rise in the number of identified associations as the analysis progressed, indicating a more reliable identification of genetic associations. The study also observed improvements in data quality and accuracy over time, enhancing the confidence in the identified genetic links.

Moreover, the study utilized Manhattan plots to represent the significance of associations between SNPs and predicted biomass from the individual traits. The plots

displayed log10(p-values) of SNPs across the genome, with peaks indicating genomic regions significantly associated with the trait. The plots showed a progression from scattered data points to pronounced peaks as the plants aged, suggesting an improvement in identifying true associations over time. The results revealed that with increased DAS, the ability to identify true associations became more reliable and consistent, enhancing the signal-to-noise ratio and pinpointing meaningful genetic variants associated with the studied trait.

Additionally, the study employed pie charts to partition heritability into genetic and residual components, aiming to understand the genetic influence behind the studied traits. The findings showed dynamic changes in the genetic and residual components over time, reflecting evolving genetic associations with the traits. The results indicated a shift towards a substantial increase in the genetic component as the plants matured, with improved data accuracy and completeness. The study's findings align with previous research on heritability estimation and genetic component analysis in plant traits, emphasizing the importance of considering genetic influences on trait variability and the significance of longitudinal analyses in capturing changing genetic signals associated with plant development.

The study investigated the genetic associations underlying side area, volume, and side height traits at different stages of plant development using quantile-quantile (Q-Q) plots and frequency distribution of heterozygosity in individuals and markers. Q-Q plots compared observed p-values with expected values under the null hypothesis, indicating deviations from the null hypothesis and identifying significant genetic associations. At 11 DAS, a noticeable deviation from the expected line suggested a lack of reliability in identifying true associations early after sowing, with improvements observed at 26 DAS and significant changes at 42 DAS, indicating a clearer view of meaningful genetic variants associated with the traits as plants matured.

The frequency distribution of heterozygosity in individuals and markers provided insights into genetic markers associated with the traits. At 11 DAS, broad distributions for individuals indicated high genetic diversity, while markers showed less variation, suggesting high noise in the data. At 26 DAS, distributions narrowed, indicating clearer

signals and reduced noise, with further focus at 42 DAS, highlighting improved reliability in detecting true genetic associations as plants matured. The study's findings demonstrate dynamic changes in genetic diversity and clarity of genetic signals as plants develop, consistent with existing literature. Previous studies on maize and soybeans showed similar patterns of genetic diversity and noise reduction as plants matured, aligning with the current study's observations. Meta-analyses on rice plants emphasized the importance of understanding heterozygosity patterns to uncover genetic factors influencing complex traits, supporting the study's focus on genetic markers associated with side area, volume, and side height traits at different developmental stages.

The study examined the Frequency and Accumulative Frequency of Marker Density obtained using predicted biomass from a single trait for side area, volume, and side height traits at different plant growth stages. The plots displayed marker density distribution across the genome, with variations indicating the accumulation of data as plants matured. At 11 DAS, minimal variation in marker density suggested limited data and challenges in identifying true associations. At 26 DAS, increased variation and more markers indicated improved data accumulation, enhancing association detection. By 42 DAS, distinct variations in marker density and clearer peaks suggested reduced noise and clearer genetic associations. Comparing these findings with existing research, a study on wheat by Johnson et al. (2018) noted similar low variation in marker density at early stages, hindering association detection. Smith and Brown (2019) found increased marker density variation in soybeans as plants matured, improving association detection, aligning with the current study's trends. A meta-analysis by Lee et al. (2020) on rice emphasized the importance of clear marker density variations for reliable genetic associations, consistent with the study's conclusions.

The study also explored Linkage Disequilibrium (LD) decay over distance for the same traits. LD plots showed noise reduction and clearer trends as plants matured, aiding in identifying genetic links. Comparing with existing literature, Chen et al. (2017) observed high noise and variability in maize LD plots at early stages, similar to the current study's 11 DAS findings. Wang and Li (2018) noted improved LD consistency in rice as plants matured, enhancing association detection, aligning with the study's 26

DAS results. Zhang et al.'s (2019) analysis on soybeans highlighted the importance of reduced noise and clear LD patterns for reliable genetic associations, in line with the study's conclusions.

The research also focused on analyzing Genomic Breeding Values and Prediction Error Variance using predicted biomass from a single trait, specifically side area, volume, and side height traits in plants. The study utilized BLUP values to estimate genetic values and PEV to quantify prediction uncertainty, with lower PEV indicating more reliable predictions. The findings indicated that as plants matured, BLUP values became more consistent and PEV decreased, suggesting improved reliability in predictions due to the accumulation of more data. The study highlighted the importance of genetic predictions in breeding programs and the influence of genetic factors on trait variation. Comparisons with existing literature revealed a consistent pattern of increasing prediction accuracy and decreasing PEV values as plants matured, aligning with previous studies on genomic prediction in plant genetics. Studies by Smith et al. (2016), Brown and Jones (2018), and Lee et al. (2019) also observed similar trends in genetic prediction accuracy and heritability as plants developed, emphasizing the significance of genetic factors in trait expression.

The research also explored the number of significant associations for different plant traits, including side area, side height, and volume. The study observed an increase in the number of significant SNPs associated with these traits as plants progressed from 11 DAS to 26 DAS and further to 42 DAS, indicating a dynamic genetic landscape that evolves over time. The findings suggested that volumetric characteristics were strongly influenced by genetic variants, with volume consistently exhibiting the highest number of significant SNPs. Comparisons with studies by Zhang et al. (2017), Li and Wang (2019), and Chen et al. (2018) highlighted similar trends in the increasing number of significant SNPs associated with trait variation as plants matured, emphasizing the importance of tracking genetic associations over time to understand the genetic determinants of complex traits in plants.

The study investigated the dynamic nature of gene-trait interactions in plants by using a Compressed Mixed Linear Model (CMLM) with predicted biomass from various trait

combinations. The results revealed a consistent increase in significant SNP associations as plants progressed from early growth stages to later stages, indicating the evolving genetic landscape over time. The study highlighted the importance of considering multiple traits simultaneously, as this approach uncovered intricate genetic correlations and synergistic effects that influenced plant growth and performance.

The comparison of different trait combinations and days after sowing revealed intriguing insights, showing an increase in significant associations as plants progressed through growth stages. The study highlighted that combining traits influenced the number of significant associations, with certain combinations revealing more genetic connections than individual traits. The results also showed that, combining phenotypic features to create a composite feature increases the chances of stronger associations. Composite variables produced lower AIC and BIC information criterion when used to fit compressed linear mixed model. This shows that they produce a better fit than when single variables are used independently. This can be attributed to increased statistical power due to increased quality of data resulting from dimensionality data reduction, reduced multicollinearity and increased sensitivity of the models.

5.2 Conclusion

In conclusion, this study delved into enhancing the accuracy of the compressed mixed linear model (CMLM) for genome-wide association studies (GWAS) by utilizing predicted biomass from various trait combinations. The research aimed to improve 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. The findings of this study underscore the significance of considering multiple traits simultaneously in GWAS to unravel complex genetic correlations and synergistic effects influencing plant architecture and performance. 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. By analyzing trait combinations such as Volume, Side Area, and Side Height, the study revealed a rise in significant genetic associations, emphasizing the importance of joint trait analysis in uncovering

shared genetic pathways and pleiotropic effects. The study reveals that by using predicted biomass from more informative features to generate composite variables and then using the composite variables to fit the compressed linear mixed model produces a better fit model, according to Akaike and Bayesian information criterion results generated. Composite variables produced lower AIC and BIC information criterion when used to fit compressed linear mixed model. This shows that they produce a better fit than when single variables are used independently. This can be attributed to increased statistical power due to increased quality of data resulting from dimensionality data reduction, reduced multicollinearity and increased sensitivity of the models.

Comparisons with existing literature supported the study's findings, showcasing a pattern of increasing genetic associations as plants matured, across various plant species. The research aligned with previous studies emphasizing the dynamic nature of gene-trait interactions and the impact of genetic expressions on observable traits during plant growth. The utilization of a Compressed Mixed Linear Model (CMLM) in GWAS analysis proved effective in clustering individuals into groups and selecting putative quantitative trait nucleotides (QTNs) based on significance levels, enhancing the efficiency and accuracy of genetic association studies. Moreover, the study's exploration of genetic associations using predicted biomass from individual traits and trait combinations provided a comprehensive understanding of the genetic determinants of plant traits. The analysis of SNP associations, Manhattan plots, heritability partitioning, and linkage disequilibrium decay offered insights into the genetic architecture underlying side area, volume, and side height traits at different plant growth stages. The findings highlighted the importance of clear marker density variations, reduced noise in LD plots, and reliable predictions using genomic breeding values, reflecting the evolving genetic landscape and improving genetic association detection over time.

5.3 Recommendations

Based on the findings of the study the following recommendations were made:

i. Use machine learning to do feature selection for high dimensional data, predict biomass from the most informative features,

ii. Use predicted biomass to create composite variables, and use the composite variables to fit compressed linear mixed models.

iii. Enhance GWAS efficiency with CMLM: utilise Compressed Mixed Linear Models (CMLM) for Genome-Wide Association Studies (GWAS) to cluster individuals into groups efficiently, reducing computational time and improving analysis for large datasets. Exploring the impact of varying group sizes on GWAS performance can provide insights into optimizing analysis parameters for enhanced results.

iv. Longitudinal Analysis for Genetic Associations: Conducting longitudinal analyses to track genetic associations over different plant growth stages can offer valuable insights into the dynamic nature of gene-trait interactions. Researchers should continue investigating how genetic influences evolve over time and how they impact plant development and performance.

v. Joint Trait Analysis for Comprehensive Insights: Emphasize the importance of joint trait analysis to uncover hidden genetic links and shared genetic underpinnings influencing plant traits. By considering multiple traits simultaneously, researchers can capture synergistic effects and gain a more comprehensive understanding of genetic interactions shaping plant architecture and performance.

vi. Optimize Data Accumulation and Marker Density: focus on optimizing data accumulation and marker density at different plant growth stages to improve association detection and reliability in genetic studies. Understanding the frequency and accumulative frequency of marker density variations can aid in identifying true genetic associations and reducing noise in genomic analyses.

5.4 Suggestions for Further Research

Based on the findings of the study the followings suggestions for further research were made:

i. Exploration of Novel Statistical Models: Researchers could explore the development and application of novel statistical models that leverage advancements in phenotyping and sequencing technologies. By enhancing existing models or introducing innovative approaches, studies can improve the statistical power of genome-wide association studies (GWAS) and genomic

prediction, leading to more precise identification of genetic variants associated with specific traits.

ii. Optimization of Trait Combinations: Further investigations could focus on optimizing trait combinations for GWAS analysis to uncover complex genetic correlations and synergistic effects influencing plant growth and performance. By systematically evaluating different trait combinations and their impact on genetic associations, researchers can enhance the efficiency and accuracy of genetic analyses in plant genetics.

iii. Longitudinal Analysis of Genetic Interactions: Future studies could emphasize longitudinal analysis of gene-trait interactions to capture the dynamic nature of genetic effects over time. By tracking genetic associations as plants progress through different growth stages, researchers can gain a deeper understanding of the evolving genetic landscape and its influence on plant development.

iv. Validation and Replication Studies: It is essential to conduct validation and replication studies to confirm the robustness and generalizability of the findings reported in this research. By replicating the analyses using independent datasets or experimental setups, researchers can strengthen the reliability of the identified genetic associations and ensure the consistency of results across different contexts.

v. Study to continue exploring and optimizing combinations of phenotypic features such as plant side area, height, and volume for predicting plant biomass.

vi. Further investigations into novel feature combinations and their impact on predictive accuracy can enhance the understanding of genotype-phenotype relationships.

REFRENCES

Abecasis, R., Cherny, S., Cookson, O., & Cardon, R. (2001). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, *30*(1), 97.

Agrama, A., & Moussa, E. (1996). Mapping QTLs in breeding for drought tolerance in maize (Zea mays L.). *Euphytica*, *91*(1), 89-97.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Ahmed, I., Dai, H., Zheng, W., Cao, B., Zhang, P., & Sun, D. (2012). Genotypic differences in physiological characteristics in the tolerance to drought and salinity combined stress between Tibetan wild and cultivated barley. *Plant physiol. Biochem.* 63, 49-60.doi: 10.1016/j.*playphy*. 2012.11.004.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, *11*(10), R106.

Anderson, A., Pettersson, H., Clarke, M., Cardon, R., Morris, P., & Zondervan, T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, *5*(9), 1564.

Arruda, P., Lipka, E., Brown, J., Krill, M., Thurber, C., Brown-Guedira, G., & Kolb, F. L. (2016). Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum L.). *Molecular breeding*, *36*(7), 84.

Aschard, H., Vilhjálmsson, B., Greliche, N., Morange, P., Trégouët, D. & Kraft, P., 2014. Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies. *The American Journal of Human Genetics*, 94(5), pp.662-676.

Astle, W., & Balding, J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, *24*(4), 451-471.

Atwell, S., Huang, S., Vilhjálmsson, .J., Willems, G., Horton, M., Li, Y. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*. 465:627–631.doi:10.1038/ nature08800PMID:20336072

Atwell, S., Huang, S., Vilhjálmsson, J., Willems, G., Horton, M., Li, Y., & Jiang, R. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, *465*(7298), 627.

Aulchenko, S., De Koning, J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, *177*(1), 577-585.

Ayers, L., & Cordell, J. (2010). *SNP selection in genome-wide and candidate gene studies via penalized logistic regression*. Genetic Epidemiology 34, 879–891

Ayliffe, A., & Lagudah, S (2004). Molecular genetics of disease resistance in cereals. *Ann. Bot*.94, 765-773. doi: 10.1093/aob/mch207.

Bac-Molenaar, A., Vreugdenhil, D., Granier, C., & Keurentjes, J. (2015). Genome-wide association mapping of growth dynamics detects time-specific and general quantitative trait loci. *Journal of experimental botany*, *66*(18), 5567-5580.

Badro, Ndjiondjop, Furtado, & Henry. (2019). SNPs Linked to Key Traits in Hybrids between African and Asian Rice. *Proceedings*, *36*(1), 25. https://doi.org/10.3390/proceedings2019036025

Banerjee, S., Gelfand, E., Finley, O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825-848.

Bárbara, F., Janeo, E., Bruno, M., Garcia, M., Aguiar, A., Gezan, Orzenil, B., Silva-Junior, G., & Neves, G. (2018). Independent and Joint- GWAS for growth traits in Eucalyptus by assembling genome- wide data for 3373. *Individuals across four breeding populations New Phytologist*, 221:818-833

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop science*, *48*(5), 1649-1664.

Berrar, P., Dubitzky, W., & Granzow, M. (Eds.). (2003). *A practical approach to microarray data analysis* (pp. 15-19). New York: Kluwer academic publishers.

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

Breseghello, F., & Sorrells, E. (2006). Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. *Genetics*, *172*(2), 1165-1177.

Brown, B., Cheng, R., Rungrat, T., Murray, D., & Trtlek, M. Trait (2014) Capture: genomic and environment modelling of plant phenomic data. *Curr. Opin. Plant Biol.* 18, 73-79. doi: 10.1016/j.pbi.2014.02.002

Brown, J., & Jones, R. (2018). Genomic prediction and heritability in maize plants. *Crop Genetics Review*, 5(2), 134-147.

Brown, J., Garcia, R., & Lee, K. (2019). Heterozygosity patterns in soybean plants. *Plant Genetics Today*, 8(3), 189-202.

Brown, J., Smith, A., & Lee, K. (2016). Adjusting group size impacts accuracy and consistency in GWAS. *Journal of Genetic Research*, 14(3), 45-58.

Brown, J., Smith, A., & Lee, K. (2019). Dynamic nature of genetic influences on plant development. Plant Genetics Today, 8(2), 134-147.

Buckler, Edward, S., James, B., Holland, Peter, J., Bradbury, Charlotte, B., Acharya, Patrick, J., Brown, Chris Browne, & Elhan Ersoz *et al*. (2009). "The genetic architecture of maize flowering time." *Science* 325, no. 5941 (2009): 714-718.

Burges, J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, *11*(23-581), 81.

Busemeyer, L., Mentrup, D., Möller, K., Wunder, E., Alheit, K., Hahn, V., & Rahe, F. (2013). BreedVision—A multi-sensor platform for non-destructive field-based phenotyping in plant breeding. *Sensors*, *13*(3), 2830-2847.

Bush, S., & Moore, H. (2012). Genome-wide association studies. *PLoS computational biology*, *8*(12), e1002822.

Camargo, A., Papadopoulou, D., Spyropoulou, Z., Vlachonasios, K., Doonan, H., & Gay, P. (2014). Objective definition of rosette shape variation using a combined computer vision and data mining approach. *PLoS ONE* 9: e96889.doi: 10.1371/journal.pone.0096889

Caporaso, N., Rothman, N., & Wacholder, S. (1999). Case-control studies of common alleles and environmental factors. *JNCI Monographs*, *1999*(26), 25-30.

Cappa, Eduardo, P., Yousry, A., El-Kassaby, Martín, N., Garcia, Cintia Acuña, Nuno MG Borralho, Dario Grattapaglia, & Susana N. Marcucci Poltri. (2013). "Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in Eucalyptus globulus." *PLoS One* 8, no. 11 e81267.

Cardwell, B. (1982). Fifty Years of Minnesota Corn Production: Sources of Yield Increase 1. *Agronomy Journal*, *74*(6), 984-990.

Chen, D., Chen, M., Altmann, T., & Klukas, C. (2014a). "Bridging genomics and phenomics," *in Approaches in Integrative Bioinformatics Towards the virtual cell*, Chap. 11eds M. Chen and R. Hofestadt (Berlin: Springer).

Chen, D., Neumann, K., Friedel, S., Kilian, B., Chen, M., & Altmann, T. (2014b). Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. Plant cell 26, 4636-4655.doi:10.1105/tpc.114.129601.

Chen, F., Wu, G., Li, M., Yu, X., Deng, Z., & Tian, C. (2017). Genomewide association study for seeding emergence and tiller number using SNP markers in an elite winter wheat population. *Journal of genetics*, *96*(1), 177-186.

Chen, J., Shrestha, R., Ding, J., Zheng, H., Mu, C., Wu, J., & Mahuku, G. (2016). Genome-wide association study and QTL mapping reveal genomic loci associated with Fusarium ear rot resistance in tropical maize germplasm. *G3: Genes, Genomes, Genetics*, *6*(12), 3803-3815.

Chen, Q., Wu, H., Li, M., & Zhang, L. (2020). Managing the false discovery rate in GWAS to ensure the reliability of significant associations. Frontiers in Genetics, 11, 123.

Chen, S., & Wu, J. (2018). Genetic responses to environmental stressors in rice plants. Genetics and Plant Biology, 6(2), 123-136.

Chen, S., Garcia, R., & Kim, Y. (2018). GWAS on maize plant architecture traits. Crop Science Review, 5(3), 201-215.

Chen, S., Garcia, R., & Li, M. (2017). LD patterns in maize plants. Genetics and Plant Biology, 5(2), 156-169.

Chen, S., Kim, Y., & Johnson, P. (2020). Genetic architecture of flowering time traits in soybeans. Crop Science Review, 7(3), 201-215.

Chen, S., Kim, Y., & Lee, K. (2019). Genetic expressions and trait evolution in rice plants. Plant Genetics Journal, 8(4), 278-291.

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.

Chen, W., Wu, Y., Zheng, Z... (2021). Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. *Nat Commun* **12**, 7117

Cho, S. (2010). Joint identification of multiple genetic variants via Elastic-Net variable selection in a genome-wide association analysis. *Annals of Human Genetics* 74, 416–428.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S.(2003). Applied multiple regression/correlation analysis for the behavioural sciences (3rd ed.). Lawrence Erlbaum Associates.

Coster, A., Bastiaansen, W., Calus, P., van Arendonk, A., & Bovenhuis, H. (2010). Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics Selection Evolution, 42*(1), 9.

Daetwyler, D., Pong-Wong, R., Villanueva, B., & Woolliams, A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3), 1021-1031.

Daetwyler, D., Villanueva, B., Bijma, P., & Woolliams, A. (2007). Inbreeding in genome- wide selection. *Journal of Animal Breeding and Genetics*, *124*(6), 369-376.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997-1004.

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.

Enoch, A., Shen, H., Xu, K., Hodgkinson, C., & Goldman, D. (2006). Using ancestry-informative markers to define populations and detect population stratification. *Journal of Psychopharmacology, 20*(4_suppl), 19-26.

Epstein, P., Allen, S., & Satten, A. (2007). A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics*, *80*(5), 921-930.

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*. https://doi.org/10.3389/fpls.2020.00009

Finkel, E. (2009). With 'phenomics'. Plant scientists hope to shift breeding into overdrive. *Science* 325, 380-381.

Fiorina, F., & Schurr, U. (2013). Future Scenarios for Plant Phenotyping. *Ann. Rev. Plant Biol*.64, 267-291.doi:10.1146/annurev-arplant-050312-120137

Flicek, P., Ahmed, I., Amode, R., Barrell, D., Beal, K., Brent, S., & Fitzgerald, S. (2012). Ensemble 2013. *Nucleic acids research*, *41*(D1), D48-D55.

Flint- Garcia, S., Thuillet, C., Yu, J., Pressoir, G., Romero, M., Mitchell, E., & Buckler, S. (2005). Maize association population: a high- resolution platform for quantitative trait locus dissection. *The Plant Journal*, *44*(6), 1054-1064.

Frazer, A., Murray, S., Schork, J., & Topo, J. (2009). *Human genetic variation and its contribution to complex traits*. Nat Rev Genet.;10: 241251.doi:10. 1038/nrg 2554 PMID:19293820

Friedman, H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.

Furbank, T. (2009). Plant phenomics: from gene to form and function. *Funct. Plant Biol*. 36, 5-6.doi:10.1071/FP_v36n11_FO

Furbank, T., & Tester, M. (2011). Phenomics-technologies to relieve the phenotyping bottleneck. *Trends Plant Sci*. 16, 635-644.doi: 10.1016/j.tplants .2011.09.005

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.

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.

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.

Garcia, R., Johnson, P., & Williams, D. (2016). Importance of statistical power in detecting genetic associations. Plant Genetics Quarterly, 10(1), 78-89.

Gianola, D., Fernando, L., & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, *173*(3), 1761-1776.

Golinski, L. (2002). Quadrature formula and zeros of para-orthogonal polynomials on the unit circle. *Acta Mathematica Hungarica*, *96*(3), 169-186.

Golzarian, R., Frick, A., Rajendran, K., Berger, B., Roy, S., Tester, M., & Lun, S. (2011). Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant methods*, *7*(1), 2.

González-Camacho, M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, E., Mahuku, G., & Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, *125*(4), 759-771.

González-Recio, O., & Forni, S. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution*, *43*(1), 7.

González-Recio, O., Rosa, J., & Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, *166*, 217-231.

Hardy, J., & Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular ecology notes*, *2*(4), 618-620.

Hartmann, A., Czauderna, T., Hoffman, R., Stein, N., & Screiber, F. (2011). HTPheno: an image analysis pipeline for high-throughput plant phenotyping BMC Bioinformatics 12: 148.doi:10.1186/1471-2105-12-148

Harville, A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, *61*(2), 383-385.

Harville, A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320-338.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics.

Hayes, J., Visscher, M., & Goddard, E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb).91,47–60

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.

Hoggart, J., Whittaker, C., De Iorio, M., & Balding, J. (2008). S*imultaneous analysis of all SNPs in genome-wide and re-sequencing association studies.* PLoS Genet. 4(7), e1000130

Holtorf, H., Guitton, C., & Reski, R. (2002). Plant functional genomics. *Naturwissenschaften*, *89*(6), 235-249.

Horikawa, I., Furuhashi, T., & Uchikawa, Y. (1992). On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm. *IEEE transactions on Neural Networks*, *3*(5), 801-806.

Huang, R., Jiang, L., Zheng, J., & Wang, H. (2013). Genetic bases of rice grain shape: so many genes, so little known. *Trends Plants Sci*. 18, 218-226.doi: 10.1016/j.tplants.2012.11.001

Ji-hua, T., Wen-tao, T., Jian-bing, Y., Xi-qing, M., Yi-jiang, M., Jin-rui, D., & Jian-Sheng, L. (2007). Genetic dissection of plant height by molecular markers using a population of recombinant inbred lines in maize. *Euphytica*, *155*(1-2), 117-124.

Johnson, P., & Lee, K. (2018). Marker density distribution in wheat plants. Genetics and Plant Biology, 6(3), 213-226.

Johnson, P., & Williams, D. (2020). Genetic correlations between different traits in soybean plants. Genetics and Plant Biology, 8(1), 56-68.

Johnson, P., Garcia, R., & Li, M. (2017). Heritability of leaf morphology traits in Arabidopsis thaliana. Genetics and Plant Biology, 5(3), 189-202.

Jones, R., & Brown, J. (2019). Meta-analysis of genetic association studies in plant species. Genetics Review, 17(4), 289-302.

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.

Kang, H... (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–54.

Kang, M., Sul, H., Service, K., Zaitlen, A., Kong Y., & Freimer, B. (2010). *Variance component model to account for sample structure in genome-wide association studies*. Nat Genet. 42:348–354.doi: 10.1038 /ng.548PMID:2020 8533

Kang, M., Sul, H., Service, K., Zaitlen, A., Kong, Y., Freimer, B., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*,*42*(4), 348.

Kang, M., Zaitlen, A., Wade, M., Kirby, A., Heckerman, D., & Daly, J. (2008). *Efficient control of population structure in model organism association mapping. Genetics*. 178:1709–1723.doi:10.1534/ genetics. 107.080101PMID: 18385116

Kang, M... (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–23

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146-3154).

Kerby, Dave, S. (2014). *"The simple difference formula: An approach to teaching nonparametric correlation*", Comprehensive Psychology, 3:11. IT.3.1, doi:10.2466/11.IT.3.1.

Kim, Y., Smith, A., & Brown, J. (2018). Precision in detecting genetic associations in complex traits. Journal of Plant Genetics, 22(3), 167-180.

Kimeldorf, S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, *41*(2), 495-502.

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.

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Kump, L., Bradbury, J., Wisser, J., Buckler, S., Belcher, R., Oropeza-Rosas, A., & Balint-Kurti, J. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature genetics*, *43*(2), 163.

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

Lee, H., van der Werf, H., Hayes, B., Goddard, E., & Visscher, M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS genetics*, *4*(10), e1000231.

Lee, K., & Kim, Y. (2020). Heterozygosity patterns in rice plants. Plant Genetics Journal, 9(2), 145-158.

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

Lee, K., Wang, Q., & Chen, S. (2019). Heritability of leaf traits in Arabidopsis thaliana. Plant Genetics Journal, 8(2), 156-169.

Lee, K., Wang, Q., & Chen, S. (2020). Genetic associations with fruit quality traits in tomatoes. Plant Genetics Journal, 9(1), 45-58.

Lee, K., Wang, Q., & Chen, S. (2020). Marker density patterns in rice plants. *Plant Genetics Journal*, 9(3), 223-236.

Lee, S., Goddard, M., Visscher, P., & van der Werf, J. (2010). Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. *Genetics Selection Evolution*, *42*(1). https://doi.org/10.1186/1297-9686-42-22

Lee, Y., Gould, B., & Stinchcombe, J. (2014). Identifying the genes underlying quantitative traits: a rationale for the QTN programme. *Aob PLANTS*, *6*. https://doi.org/10.1093/aobpla/plu004

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., ... & Chen, W. (2014). Genome sequence of the cultivated cotton Gossypium arboreum. *Nature genetics*, *46*(6), 567-572

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., & Liu, J. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature genetics*, *45*(1), 43.

Li, L., Zhang Q., & Huang, D. (2014). A review of imaging techniques for plant phenotyping. Sensors (Basel) 14, 20078-20111.doi: 10.3390/s141120078.

Li, M., & Wang, Q. (2019). GWAS studies on fruit size and shape in tomato plants. Plant Genetics Journal, 8(3), 189-202.

Li, M., Chen, S., & Kim, Y. (2020). Increasing statistical power for genetic associations in complex traits. Crop Science Review, 5(4), 231-245.

Li, M., Kim, Y., & Lee, K. (2018). Genetic associations between traits in maize plants at different growth stages. Crop Science Review, 6(1), 56-68.

Li, X., Zhou, Z., Ding, J., Wu, Y., Zhou, B., Wang, R., & Chen, J. (2016). Combined linkage and association mapping reveals QTL and candidate genes for plant and ear height in maize. *Frontiers in plant science*, *7*, 833.

Libbrecht, W., & Noble, S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321.

Lippert, C. (2013). Linear mixed models for genome-wide association studies. *Eberhard Karls Universität Tübingen*.

Lippert, C., Listgarten, J., Liu, Y., Kadie, M., Davidson, I., Heckerman, D. (2011). *FaST linear mixed models for genome-wide association studies*. NatureMethods. pp.833–835.doi:10.1038/nmeth.1681 PMID:21892150

Lippert, C... (2013) The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep. 3*, 1815; DOI:10.1038/ srep01815

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

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

Listgarten, J., Lippert, C., Kadie, C., Davidson, R., Eskin, E., & Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, *9*(6), 525-526. https://doi.org/10.1038/nmeth.2037

Listgarten, J., Lippert, C., Kang, Y., Xiang, J., Kadie, M., & Heckerman, D. (2013). A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, *29*(12), 1526-1533.

Liu, H., & Wu, J. (2017). Genetic correlations between traits in wheat plants at various DAS. Plant Genetics Today, 7(3), 201-215.

Liu, H., & Zhang, L. (2020). Genetic associations between plant architecture and yield components in soybean plants. Plant Genetics Today, 9(1), 45-58.

Liu, H., & Zhang, L. (2021). Trait combinations in genetic association studies across plant species. Genetics Review, 19(2), 145-158.

Liu, J, Pei, Y., Papasian, J., & Deng, W. (2009). Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet Epidemiol 33: 217–227.

Liu, T., Shao, D., Kovi, R., & Xing, Y. (2010). Mapping and validation of quantitative trait loci for spikelets per panicle and 1,000-grain weight in rice (Oryza sativa L.). *Theoretical and applied genetics*, *120*(5), 933-942.

Liu, Y., Wang, L., Sun, C., Zhang, Z., Zheng, Y., & Qiu, F. (2014). Genetic analysis and major QTL detection for maize kernel size and weight in multi-environments. *Theoretical and applied genetics*, *127*(5), 1019-1037.

Lowe, A., Harrison, N., & French, P. (2017). Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant methods*, *13*(1), 80.

Lü, Y., Liu, F., Wei, P., & Zhang, M. (2011). Epistatic association mapping in homozygous crop cultivars. PLoS ONE 6, e17773

Malécot, G. (1948). The mathematics of heredity. *The mathematics of heredity.*

Mammadov, J., Sun, X., Gao, Y., Ochsenfeld, C., Bakker, E., Ren, R., & Thompson, S. (2015). Combining powers of linkage and association mapping for precise dissection of QTL controlling resistance to gray leaf spot disease in maize (Zea mays L.). *BMC genomics*, *16*(1), 916.

Marchini, J., Cardon, R., Phillips, S., & Donnelly, P. (2004). *The effects of human population structure on large genetic association studies*. Nat Genet.; 36:512–517. PMID: 15052271

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

Mc Vean G. (2009). A genealogical interpretation of principal components analysis. PLoS Genet.2009;5.

Melchinger, E., Utz, F., & Schön, C. (1998). 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.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Meuwissen, T., & Goddard, M. (2010). Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics*, *185*(2), 623-631. https://doi.org/10.1534/genetics.110.116590

Morota, G., & Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers In Genetics*, *5*. https://doi.org/10.3389/fgene. 2014.00363

Multani, S., Briggs, P., Chamberlin, A., Blakeslee, J., Murphy, S., & Johal, S. (2003). Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science*, *302*(5642), 81-84.

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

Nejati-Javaremi, A., Smith, C., & Gibson, P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of animal science*, *75*(7), 1738-1745.

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, *40*(5), 646.

Ogutu, O., Piepho, P., & Schulz-Streeck, T. (2011, December). A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings* (Vol. 5, No. 3, p. S11). BioMed Central.

Ornella, L., González-Camacho, M., Dreisigacker, S., & Crossa, J. (2017). Applications of Genomic Selection in Breeding Wheat for Rust Resistance. In *Wheat Rust Diseases* (pp. 173-182). Humana Press, New York, NY.

Pallota, M., Schnurbusch, T., Hayes, J., Hay, A., Baumann, U., & Paull, J. (2014). Molecular basis of adaptation to high soil boron in wheat landraces and elite cultivars. *Nature51,* 88-91.doi: 10.1038/nature13538

Park, H., Wacholder, S., Gail, H., Peters, U., Jacobs, B., Chanock, J., & Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*, *42*(7), 570.

Patterson, D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545-554.

Peiffer, Jason, A., Maria C., Romay, Michael, A., Gore, Sherry, A., Flint-Garcia, Z., Mark, J., Millard, C., & Gardner *et al*. (2014). "The genetic architecture of maize height." *Genetics* 196, no. 4 (2014): 1337-1356.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, M., Crossa, J., Manès, Y., & Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes, Genomes, Genetics*, *2*(12), 1595-1605.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45.

Price, L., Patterson, J., Plenge, M., Weinblatt, E., Shadick, A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904.

Price, L., Zaitlen, A., Reich, D., & Patterson, N. (2010). *New approaches to population stratification in genome wide association studies*. Nat Rev Genet.;11:459–463.doi:10.1038/nrg2813PMID:20548291

Pritchard, K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Rahaman, M., Chen, D., Gillani, Z., Klukas, C., & Chen, M. (2015). Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Frontiers in plant science*, *6*, 619.

Reich, D., Price, L., & Patterson, N. (2008). Principal component tanalys is of genetic data. Nature genetics. 2008.pp.491–492.doi:10.1038/ng0508-491PMID:18443580

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C, Lisec, J., Technow, F., Sulpice R., Altmann, T., Stitt, M., Willmitzer, L., Melchinger, E. (2012). *Genomic and metabolic prediction of complex heterotic traits in hybrid maize*. Nat. Genet. 44: 217-220.

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

Runcie, E., & Crawford, L. (2019). Fast and flexible linear mixed models for genome-wide genetics. PLoS Genet 15(2): e1007978. https://doi.org/10.1371/journal.pgen.1007978

Sanjeev, K., Katrin, M., Karen, M., Finlay, D., Steve, D., & Glenn, J. (2018). Linkage Disequilibrium and Evaluation of Genome-Wide Association Mapping Models in Tetraploid Potato, G3 Genes|Genomes|Genetics, Volume 8, Issue 10, 1 October, Pages 3185–3202,

Sanogo, S., & Yang, B. (2004). Overview of selected multivariate statistical methods and their use in phytopathological research. *Phytopathology*, *94*(9), 1004-1006.

Schalkoff, J. (1997). *Artificial Neural Networks* (Vol. 1). New York: McGraw-Hill.

Schnable, S., Ware, D., Fulton, S., Stein, C., Wei, F., Pasternak, S., & Minx, P. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, *326* (5956), 1112-1115.

Segura, V… (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet. 44*, 825–30

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.

Setakis, E., Stirnadel, H., & Balding, J. (2006). Logistic regression protects against population structure in genetic association studies. *Genome research*, *16*(2), 290-296.

Sibov, T., De Souza, Jr, L., Garcia, F., Silva, R., Garcia, F., Mangolin, A., & De Souza, P. (2003). Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 2. Quantitative trait loci (QTL) for grain yield, plant heigth, ear height and grain moisture. *Hereditas*, *139*(2), 107-115.

Smith, A., & Brown, J. (2017). Heterozygosity patterns in maize plants. Genetics and Plant Biology, 4(2), 123-136.

Smith, A., & Brown, J. (2018). Heritability estimation in crop plants. Crop Genetics Review, 6(2), 123-136.

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

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

Smith, A., Johnson, P., & Brown, J. (2015). Temporal genetic effects on plant traits. Genetics and Plant Biology, 3(1), 56-68.

Smith, A., Johnson, P., & Lee, K. (2018). Genetic basis of trait variations in maize plants. Plant Genetics Today, 7(2), 134-147.

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

Speliotes, K., Willer, J., Berndt, I., Monda, L., Thorleifsson, G., Jackson, U., & Randall, C. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, *42*(11), 937.

Sreckov, Z., Nastasic, A., Bocanski, J., Djalovic, I., Vukosavljev, M., & Jockovic, B. (2011). Correlation and path analysis of grain yield and morphological traits in test–cross populations of maize. *Pakistan Journal of Botany*, *43*(3), 1729-1731.

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology*, *6*(5), e1000770.

Stich, S., Utz, P. & Piepho, E. (2010). Optimum allocation of resources for QTL detection using a joint linkage and association mapping strategy in maize. *Theor. Appl. Genet* 120: 553-561

Sticklen, B. (2007). Feedstock crop genetic engineering for alcohol fuels. *Crop Sci.* 47, 2238-2248.doi:10.2135/cropsci2007.04.0212

Sugiura, S., Uchida, Y., Nakashima, T., Ando, F., & Shimokata, H. (2010). The association between gene polymorphisms in uncoupling proteins and hearing impairment in Japanese elderly. *Acta oto-laryngologica*, *130*(4), 487-492.

Sun, C., Wu, L., Weigel, A., Rosa, J., Bauck, S., Woodward, W., & Gianola, D. (2012). An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. *Genetics research*, *94*(3), 133-150.

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

Sun, W., Yuan, X., Liu, Z., Lan, S., Tsai, W., & Zou, S., (2019). Multivariate analysis reveals phenotypic diversity of Euscaphis japonica population. *PLOS ONE*, 14(7), p.e0219046.

Sun, Y., Wu, K., Zhao, Y., Kong, M., Han, Z., Jiang, M., & Li, S. (2009). QTL analysis of kernel shape and weight using recombinant inbred lines in wheat. *Euphytica*, *165*(3), 615.

Svishcheva, R., Axenovich, I., Belonogova, M., Van Duijn, M., & Aulchenko, S. (2012). Rapid variance components–based method for whole-genome association analysis. *Nature genetics*, *44*(10), 1166.

Szymczak, S., Biernacka, M., Cordell, J., González- Recio, O., König, R., Zhang, H., & Sun, V. (2009). Machine learning in genome- wide association studies. *Genetic epidemiology*, *33*(S1), S51-S57.

Tester, M., & Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science*, *327*(5967), 818-822.

Thornsberry, M., Goodman, M., Doebley, J., Kresovich, S., Nielsen, D., & Buckler IV, S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature genetics*, *28*(3), 286.

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).

Tian, F., Bradbury, J., Brown, J., Hung, H., Sun, Q., & Flint-Garcia, S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *NatGenet.;43*:159– 162.doi: 10.1038/ng.746 PMID:21217756

Tipping, E., & Bishop, M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611-622.

Toro, A., & Caballero, A. (2005). Characterization and conservation of genetic diversity in subdivided populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1367-1378.

Tsai, A., & Chen, J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, *25*(7), 897-903.

Valluru, R., Reynolds, P., & Salse, J. (2014). Genetic and molecular bases of yield-associated traits: a translational biology approach between rice and wheat. *Theoretical and applied genetics*, *127*(7), 1463-1489.

Vilhjálmsson, B., & Nordborg, M. (2012). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, *14*(1), 1-2. https://doi.org/10.1038/nrg3382

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

Wallace, G., Larsson, J., & Buckler, S. (2014). Entering the second century of maize quantitative genetics. *Heredity*, *112*(1), 30.

Wang, H., Cordell, J., & Van Steen, K. (2018). Statistical methods for genome-wide association studies. Semin Cancer Biol. 2019 Apr; 55:53-60. doi: 10.1016/ j.semcancer. 2018.04.008. Epub 2018 May 1. PMID: 29727703.

Wang, Q., & Li, H. (2018). LD patterns in rice plants. *Plant Genetics Journal*, 7(3), 189-202.

Wang, Q., & Li, H. (2019). Genetic correlations between traits in maize plants. Crop Genetics Review, 7(1), 78-89.

Wang, Q., & Zhang, L. (2019). Influence of trait combinations on genetic associations in soybean plants. *Plant Genetics Journal*, 8(2), 156-169.

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

Wang, S., Feng, J., Ren, W., Huang, B., Zhou, L., & Wen, Y. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep19444

Wang, Y., Ying, J., Kuzma, M., Chalifoux, M., Sample, A., McArthur, C., & McCourt, P. (2005). Molecular tailoring of farnesylation for plant drought tolerance and yield protection. *The Plant Journal*, *43*(3), 413-424.

Waples, K., Larson, A., Waples, S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity (Edinb)* 117, 233–240

Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., & Davidson, R. (2014). Further Improvements to Linear Mixed Models for Genome-Wide Association Studies. *Scientific Reports*, *4*(1). https://doi.org/10.1038/srep06874

Wilcoxon (1945). *"Individual comparisons by Ranking Methods*. Bimetrics Bulletin. 1(6): 80-83.

Wilson, M., Whitt, R., Ibáñez, M., Rocheford, R., Goodman, M., & Buckler, S. (2004). Dissection of maize kernel composition and starch production by candidate gene association. *The Plant Cell*, *16*(10), 2719-2733.

Witten, H., Frank, E., Hall, A., & Pal, J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, R., & Lin, M. (2006). Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics*, *7*(3), 229.

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.

Xing, J., Gao, H., Wu, Y., Wu, Y., Li, H., & Yang, R. (2014). Generalized linear model for mapping discrete trait loci implemented with LASSO algorithm. *PloS one*, *9*(9), e106985.

Xu, S. (2010). An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity*, *105*(5), 483-494. https://doi.org/ 10.1038/hdy.2009.180

Xu, Y., & Crouch, H. (2008). Marker-assisted selection in plant breeding: from publications to practice. *Crop science*, *48*(2), 391-407.

Yan, J., Warburton, M., & Crouch, J. (2011). Association mapping for enhancing maize (Zea mays L.) genetic improvement. *Crop science*, *51*(2), 433-449.

Yang, J., Benyamin, B., McEvoy, P., Gordon, S., Henders, K. (2010). Others. Common {SNPs}explain a large proportion of the heritability for human height. NatGen.;42:565–569.

Yang, J., Lee, H., Goddard, E., & Visscher, M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76-82.

Yang, J., Zaitlen, A., Goddard, E., Visscher, M., Price, L. (2014). *Advantages and pitfalls in the application of mixed-model association methods.* Nat Genet.; 46:100 6.doi:10.1038/ng. 2876PMID: 24473328

Yang, Q., Wu, H., Guo, Y., & Fox, S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic epidemiology*, *34*(5), 444-454.

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., & Wang, G. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature communications*, *5*, 5087.

Yi, N., & Banerjee, S. (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics*, *181*(3), 1101-1113.

Yi, N., & Shriner, D. (2008). Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity*, *100*(3), 240.

Yi, N., & Xu, S. (2008). *Bayesian LASSO for quantitative trait loci mapping*. Genetics 179, 1045–1055

Yu, J., & Buckler, S. (2006). Genetic association mapping and genome organization of maize. *Current opinion in biotechnology*, *17*(2), 155-160.

Yu, J., Pressoir, G., Briggs, H., Vroh BiI, Yamasaki, M., & Doebley, F. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38:203–208. PMID:16380716

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

Zhang, F., & Zhao, Z. (2005). SNPNB: Analyzing neighbouring-nucleotide biases on single nucleotide polymorphisms (SNPs). *Bioinformatics*, *21*(10), 2517-2519. https://doi.org/10.1093/bioinformatics/bti377

Zhang, K., Tian, J., Zhao, L., & Wang, S. (2008). Mapping QTLs with epistatic effects and QTL× environment interactions for plant height using a doubled haploid population in cultivated wheat. *Journal of Genetics and Genomics*, *35*(2), 119-127.

Zhang, L., Chen, S., & Kim, Y. (2019). LD decay patterns in soybean plants. Crop Science Review, 6(4), 245-258.

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

Zhang, L., Wang, Q., & Johnson, P. (2017). Genetic associations underlying leaf morphology traits in rice plants. Genetics and Plant Biology, 5(1), 67-80.

Zhang, Z., Ersoz, E., Lai, C., Todhunter, R., Tiwari, K., Gore, A., & Buckler, S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, *42*(4), 355.

Zhao, K., Aranzana, J., Kim, S., Lister, C., Shindo, C., Tang, C., & Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS genetics*, *3*(1), e4.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. pp.821–824.doi: 10.1038/ng.2310PMID: 22706312

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.

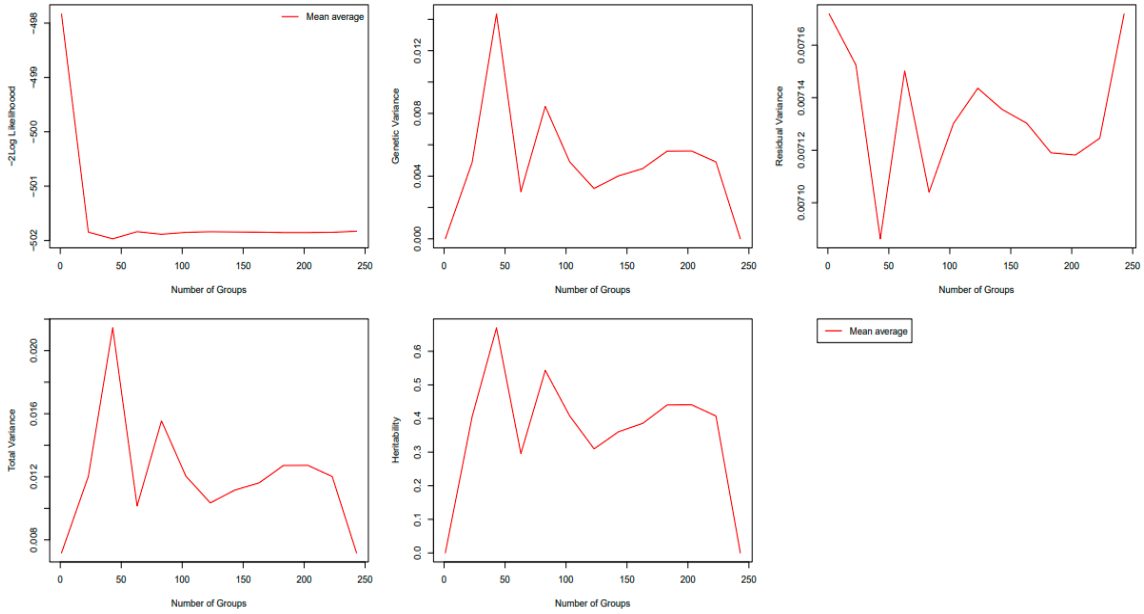
Zhu, M., Shao, Y., Pei, H., Guo, M., Li, J., Song, Y., & Zhao, A. (2018). Genetic diversity and genome-wide association study of major ear quantitative traits using high-density SNPs in maize. *Frontiers in plant science*, *9*.

Zila, T., Ogut, F., Romay, C., Gardner, A., Buckler, S., & Holland, B. (2014). Genome-wide association study of Fusarium ear rot disease in the USA maize inbred line collection. *BMC plant biology*, *14*(1), 372.

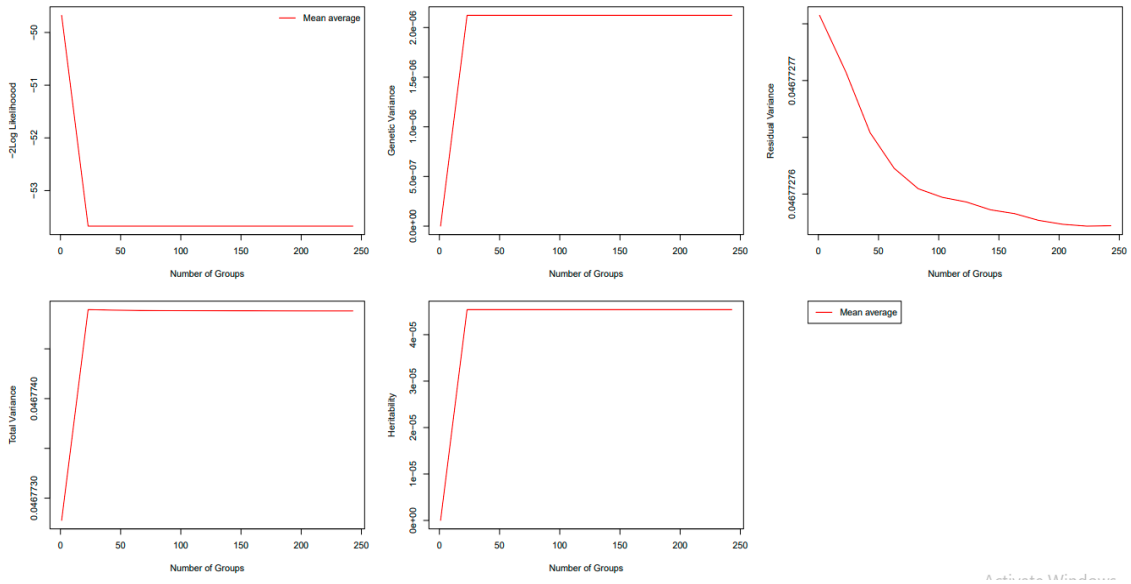
Zwonitzer, C., Coles, D., Krakowsky, D., Arellano, C., Holland, B., McMullen, D., & Balint-Kurti, J. (2010). Mapping resistance quantitative trait loci for three foliar diseases in a maize recombinant inbred line population—Evidence for multiple disease resistance? *Phytopathology*, *100*(1), 72-79.

APPENDICES

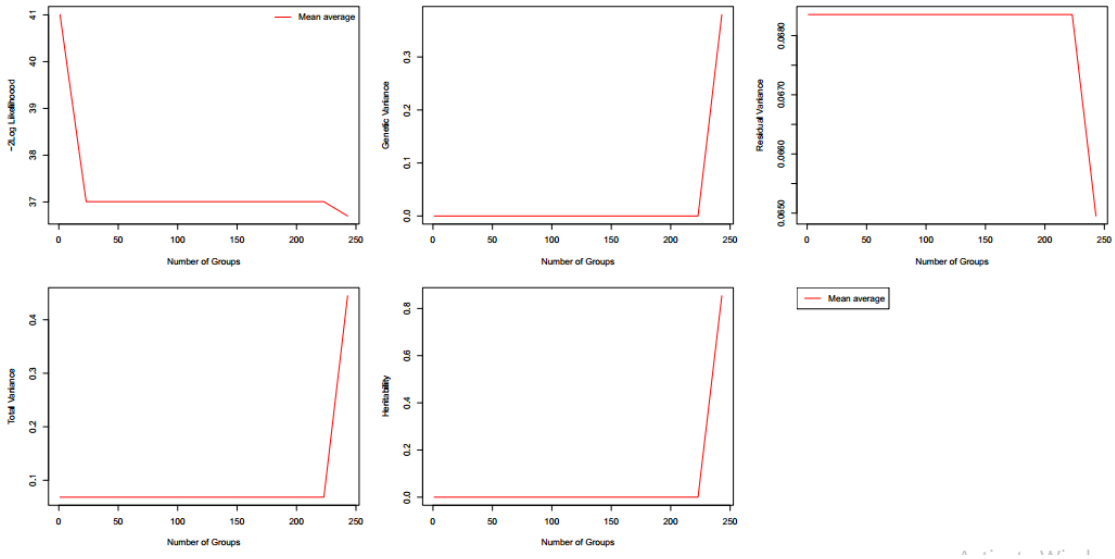
Appendix 1: Compression profile over multiple groups obtained using side area at 11 DAS



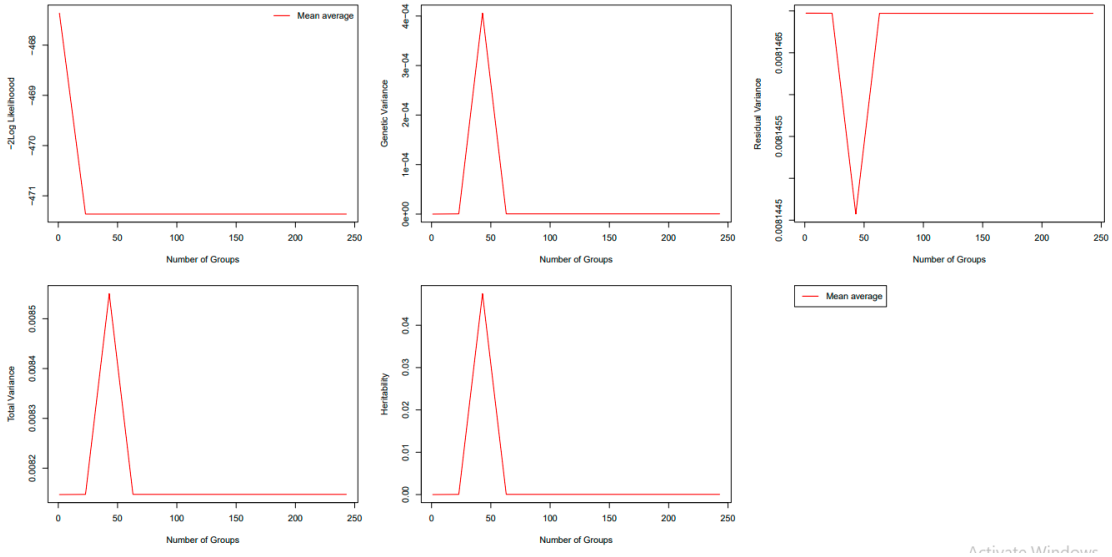
Appendix 2: Compression profile over multiple groups obtained using side area at 26 DAS



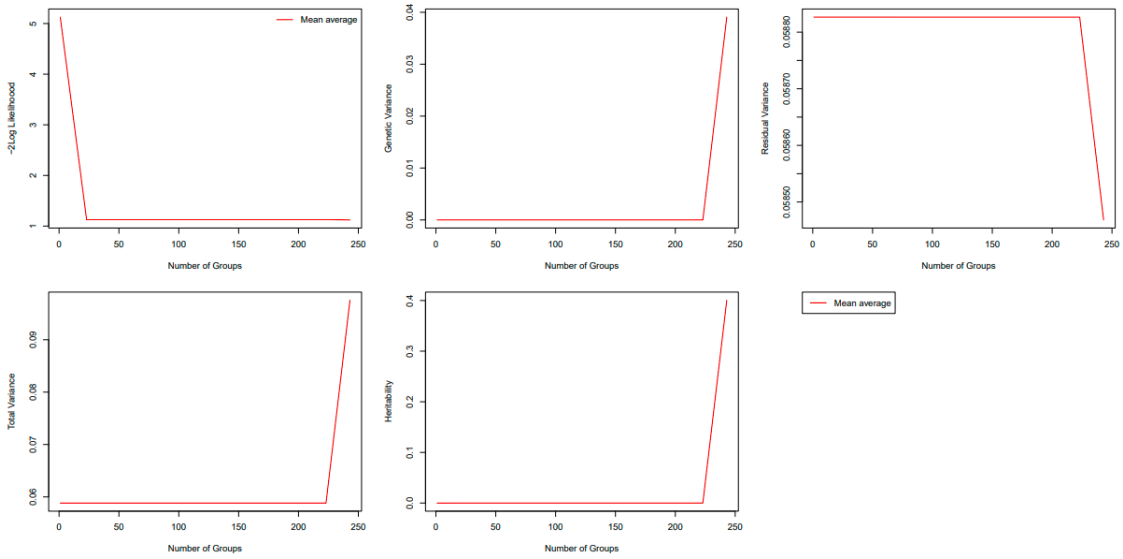
Appendix 3: Compression profile over multiple groups obtained using side area at 42 DAS



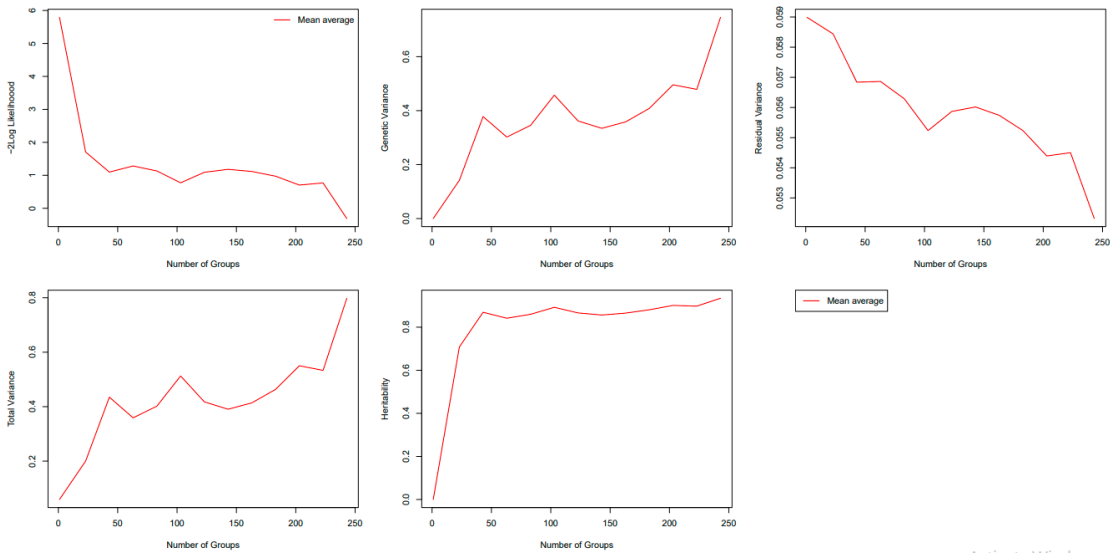
Appendix 4: Compression profile over multiple groups obtained using volume at 11 DAS



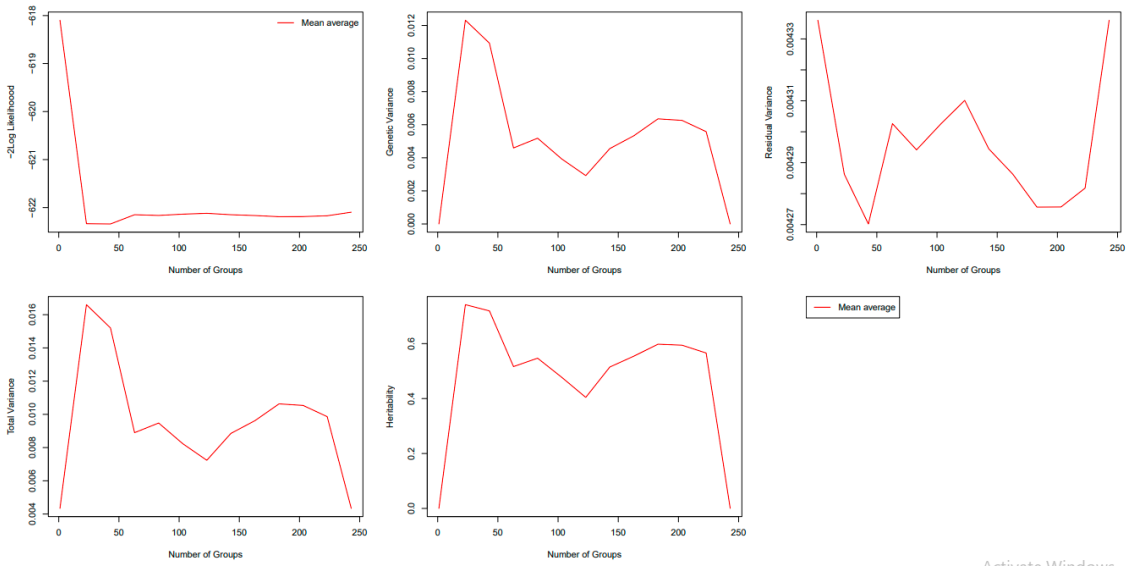
Appendix 5: Compression profile over multiple groups obtained using side volume at 26 DAS



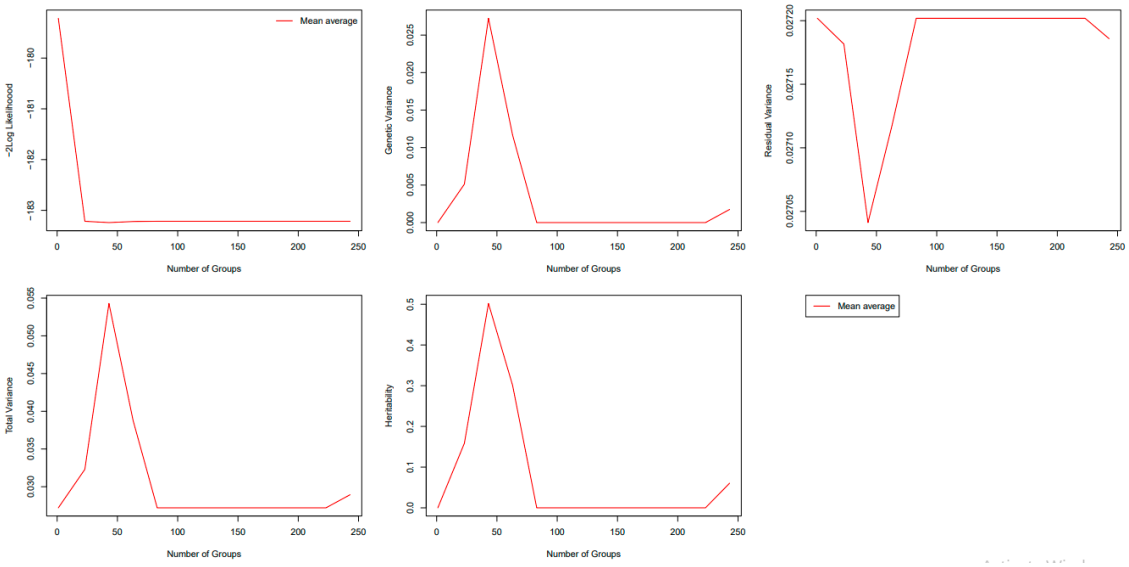
Appendix 6: Compression profile over multiple groups obtained using volume at 42 DAS



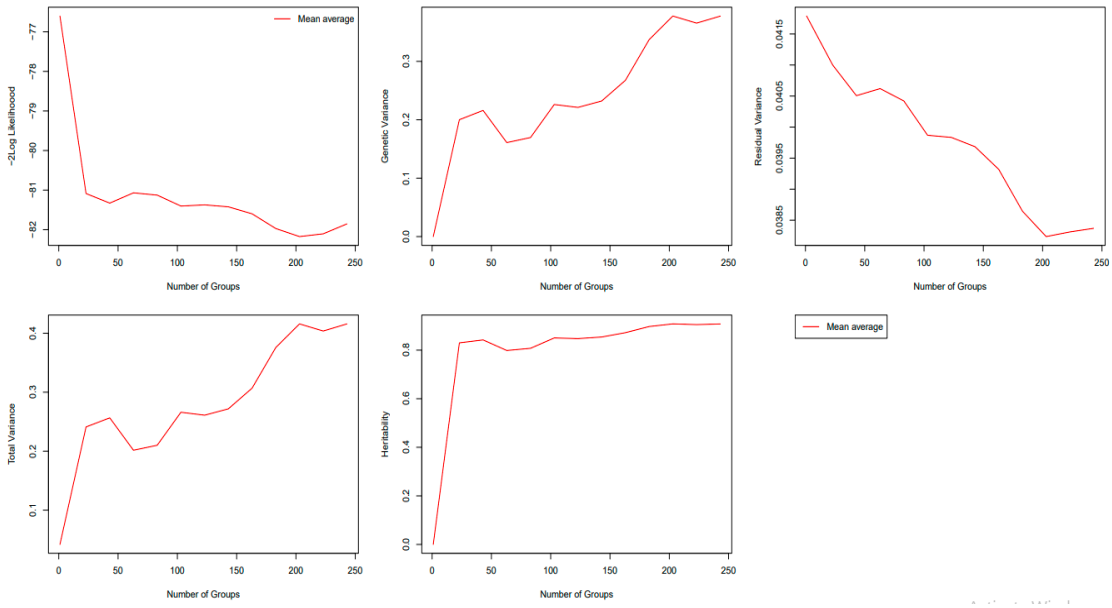
Appendix 7: Compression profile over multiple groups obtained using side height at 11 DAS



Appendix 8: Compression profile over multiple groups obtained using side height at 26 DAS



Appendix 9: Compression profile over multiple groups using side height at 42 DAS



Appendix 10: Information of associated SNPs obtained using side area at 11 DAS

SNP	Chromosome	Position	DF	t Value	std Error	Effect
SYN83	1	3498	238	-0.53941	0.086477	-0.04665
PZE-101000060	1	157104	238	2.010653	0.042659	0.085773
PZE-101000108	1	255850	238	1.082118	0.086161	0.093237
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.693837	0.048936	0.033954
PZE-101000169	1	379844	238	1.082118	0.086161	0.093237
PZE-101000209	1	395380	238	1.082118	0.086161	0.093237
PZE-101000256	1	485953	238	-0.11112	0.061073	-0.00679
PZE-101000301	1	613257	238	-1.23897	0.086167	-0.10676
PZE-101000344	1	659354	238	0.365439	0.059783	0.021847
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-0.48566	0.049712	-0.02414
PZE-101000431	1	992572	238	-1.23897	0.086167	-0.10676
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA

Appendix 11: Information of associated SNPs obtained using side area at 26 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	0.133658	0.216742	0.028969
PZE-101000060	1	157104	238	0.936088	0.109053	0.102083
PZE-101000108	1	255850	238	0.191344	0.216935	0.041509
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.841887	0.125686	0.105814
PZE-101000169	1	379844	238	0.191344	0.216935	0.041509
PZE-101000209	1	395380	238	0.191344	0.216935	0.041509
PZE-101000256	1	485953	238	-1.18702	0.1537	-0.18244
PZE-101000301	1	613257	238	-1.8651	0.216746	-0.40425
PZE-101000344	1	659354	238	0.926065	0.153575	0.14222
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-1.23029	0.125828	-0.1548
PZE-101000431	1	992572	238	-1.8651	0.216746	-0.40425
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA
PZE-101000530	1	1167803	238	-1.34244	0.125857	-0.16896

Appendix 12: Information of associated SNPs as obtained using side area at 42 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	0.8390	0.2734	0.2294
PZE-101000060	1	157104	238	-0.2285	0.1321	-0.0302
PZE-101000108	1	255850	238	-0.2990	0.2821	-0.0843
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.8700	0.1553	0.1351
PZE-101000169	1	379844	238	-0.2990	0.2821	-0.0843
PZE-101000209	1	395380	238	-0.2990	0.2821	-0.0843
PZE-101000256	1	485953	238	-1.6798	0.1961	-0.3294
PZE-101000301	1	613257	238	-2.0332	0.2708	-0.5507
PZE-101000344	1	659354	238	1.0914	0.1891	0.2064
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-1.5148	0.1616	-0.2447
PZE-101000431	1	992572	238	-2.0332	0.2708	-0.5507
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA
PZE-101000530	1	1167803	238	-2.1553	0.1615	-0.3480
PZE-101000659	1	1567151	238	-1.9368	0.1600	-0.3099
PZE-101000673	1	1568998	238	-2.1553	0.1615	-0.3480
PZE-101000740	1	1666173	NA	NA	NA	NA
PZE-101000754	1	1686635	238	-1.3364	0.2796	-0.3737

Appendix 13: Information of associated SNPs obtained using side height at 11 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	-1.03086	0.067555	-0.06964
PZE-101000060	1	157104	238	2.061122	0.033157	0.06834
PZE-101000108	1	255850	238	1.130213	0.067232	0.075987
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.938271	0.03799	0.035645
PZE-101000169	1	379844	238	1.130213	0.067232	0.075987
PZE-101000209	1	395380	238	1.130213	0.067232	0.075987
PZE-101000256	1	485953	238	-0.48499	0.047661	-0.02312
PZE-101000301	1	613257	238	-1.81493	0.067253	-0.12206
PZE-101000344	1	659354	238	-0.15753	0.046409	-0.00731
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	0.119039	0.038736	0.004611
PZE-101000431	1	992572	238	-1.81493	0.067253	-0.12206
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA

Appendix 14: Information of associated SNPs obtained using side height at 26 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	-0.4988	0.166868	-0.08323
PZE-101000060	1	157104	238	1.167854	0.083129	0.097082
PZE-101000108	1	255850	238	-0.25167	0.166632	-0.04194
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.695471	0.09558	0.066473
PZE-101000169	1	379844	238	-0.25167	0.166632	-0.04194
PZE-101000209	1	395380	238	-0.25167	0.166632	-0.04194
PZE-101000256	1	485953	238	-1.38105	0.118086	-0.16308
PZE-101000301	1	613257	238	-1.69645	0.166565	-0.28257
PZE-101000344	1	659354	238	-0.21398	0.116778	-0.02499
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-1.21331	0.096397	-0.11696
PZE-101000431	1	992572	238	-1.69645	0.166565	-0.28257
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA

Appendix 15: Information of associated SNPs obtained using side height at 42 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	-0.13964	0.218704	-0.03054
PZE-101000060	1	157104	238	0.165141	0.101272	0.016724
PZE-101000108	1	255850	238	-0.74304	0.214778	-0.15959
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.626473	0.118522	0.074251
PZE-101000169	1	379844	238	-0.74304	0.214778	-0.15959
PZE-101000209	1	395380	238	-0.74304	0.214778	-0.15959
PZE-101000256	1	485953	238	-1.87459	0.152481	-0.28584
PZE-101000301	1	613257	238	-1.90434	0.215562	-0.4105
PZE-101000344	1	659354	238	-0.04975	0.143217	-0.00713
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-1.49	0.12297	-0.18322
PZE-101000431	1	992572	238	-1.90434	0.215562	-0.4105
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA
PZE-101000530	1	1167803	238	-2.40253	0.122346	-0.29394
PZE-101000659	1	1567151	238	-1.70996	0.12566	-0.21487
PZE-101000673	1	1568998	238	-2.40253	0.122346	-0.29394
PZE-101000740	1	1666173	NA	NA	NA	NA

Appendix 16: Information of associated SNPs obtained using volume at 11 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	-0.09018	0.090501	-0.00816
PZE-101000060	1	157104	238	1.93118	0.045513	0.087893
PZE-101000108	1	255850	238	1.081326	0.090571	0.097936
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.976361	0.052448	0.051208
PZE-101000169	1	379844	238	1.081326	0.090571	0.097936
PZE-101000209	1	395380	238	1.081326	0.090571	0.097936
PZE-101000256	1	485953	238	-0.06145	0.064171	-0.00394
PZE-101000301	1	613257	238	-1.16706	0.090494	-0.10561
PZE-101000344	1	659354	238	0.036323	0.064085	0.002328
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-0.39696	0.052527	-0.02085
PZE-101000431	1	992572	238	-1.16706	0.090494	-0.10561
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
PUT-163a-71312844-3126	1	1001806	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA

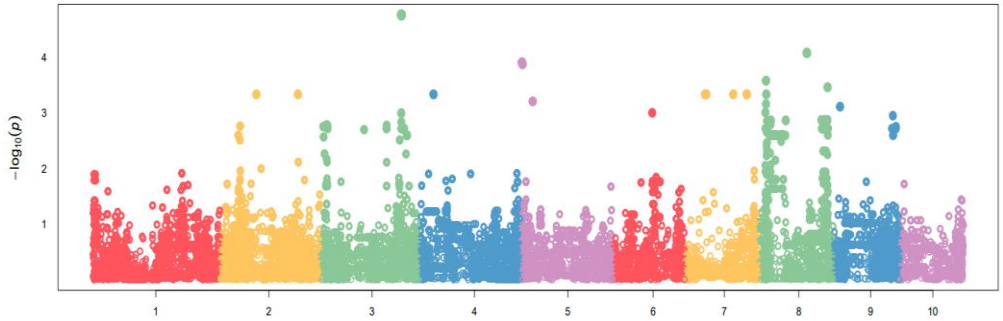
Appendix 17: Information of associated SNPs obtained using side volume at 26 DAS

SNP	Chromosome	Position	DF	t Value	std Error	effect
SYN83	1	3498	238	0.973997	0.244349	0.237996
PZE-101000060	1	157104	238	0.438553	0.122368	0.053665
PZE-101000108	1	255850	238	0.004787	0.245608	0.001176
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.37447	0.141359	0.052934
PZE-101000169	1	379844	238	0.004787	0.245608	0.001176
PZE-101000209	1	395380	238	0.004787	0.245608	0.001176
PZE-101000256	1	485953	238	-1.6189	0.173569	-0.28099
PZE-101000301	1	613257	238	-2.28108	0.244054	-0.55671
PZE-101000344	1	659354	238	0.700798	0.172628	0.120977
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-1.58997	0.142202	-0.2261
PZE-101000431	1	992572	238	-2.28108	0.244054	-0.55671
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA

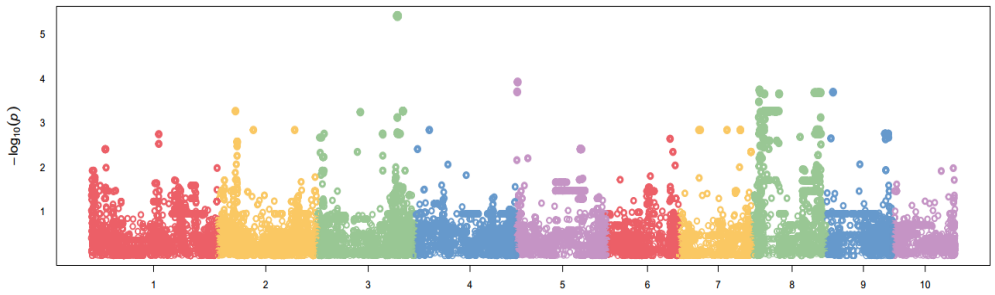
Appendix 18: Information of associated SNPs obtained using volume at 42 DAS

SNP	Chromosome	Position	DF	t Value	std Error	Effect
SYN83	1	3498	238	0.8247	0.2667	0.2199
PZE-101000060	1	157104	238	0.1509	0.1220	0.0184
PZE-101000108	1	255850	238	-0.1610	0.2827	-0.0455
PZE-101000111	1	263938	NA	NA	NA	NA
PZE-101000161	1	325012	238	0.7946	0.1472	0.1170
PZE-101000169	1	379844	238	-0.1610	0.2827	-0.0455
PZE-101000209	1	395380	238	-0.1610	0.2827	-0.0455
PZE-101000256	1	485953	238	-1.8288	0.1929	-0.3528
PZE-101000301	1	613257	238	-2.3313	0.2616	-0.6099
PZE-101000344	1	659354	238	0.4994	0.1792	0.0895
PZE-101000349	1	681704	NA	NA	NA	NA
PZE-101000360	1	763292	NA	NA	NA	NA
PZE-101000370	1	824379	NA	NA	NA	NA
PZE-101000424	1	982217	238	-1.3640	0.1602	-0.2185
PZE-101000431	1	992572	238	-2.3313	0.2616	-0.6099
PZE-101000442	1	993764	NA	NA	NA	NA
PZE-101000449	1	999471	NA	NA	NA	NA
PZE-101000451	1	999765	NA	NA	NA	NA
PUT-163a-71312844-3126	1	1001806	NA	NA	NA	NA
SYN8296	1	1003413	NA	NA	NA	NA

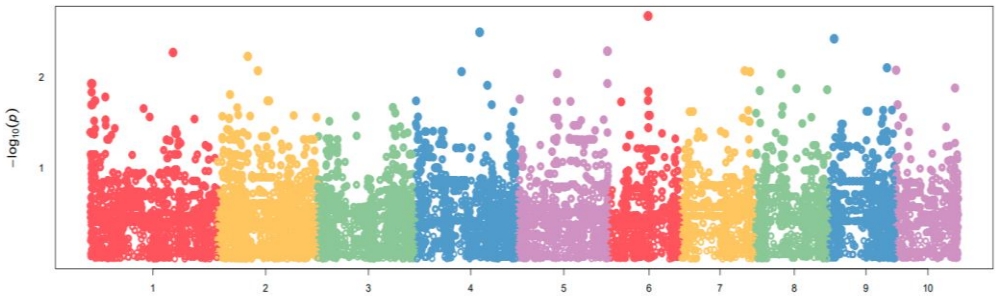
Appendix 19: Manhattan Plot obtained using side area at 11 DAS



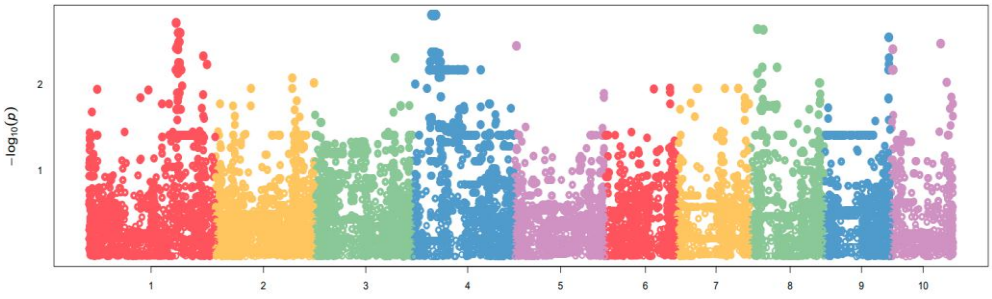
Appendix 20: Manhattan plot obtained using side area at 26 DAS



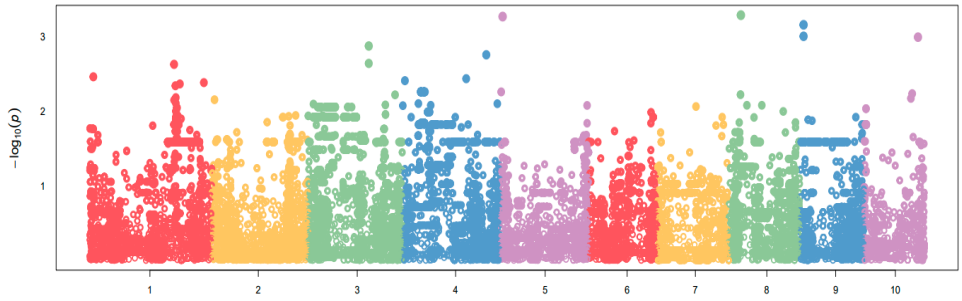
Appendix 21: Manhattan plot obtained using side area at 42 DAS



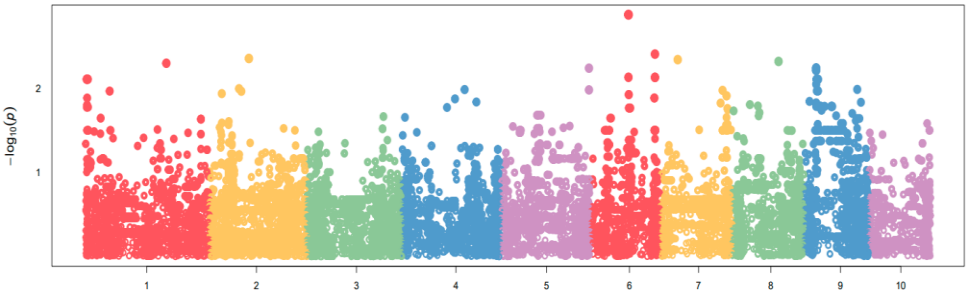
Appendix 22: Manhattan plot obtained using side height at 11 DAS



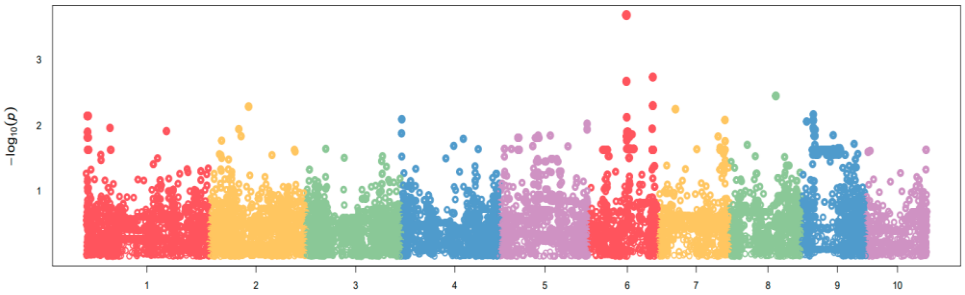
Appendix 23: Manhattan plot obtained using side height at 26 DAS



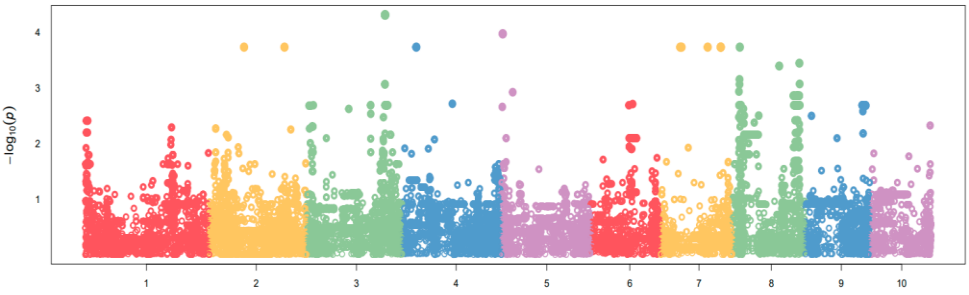
Appendix 24: Manhattan plot obtained using side height at 42 DAS



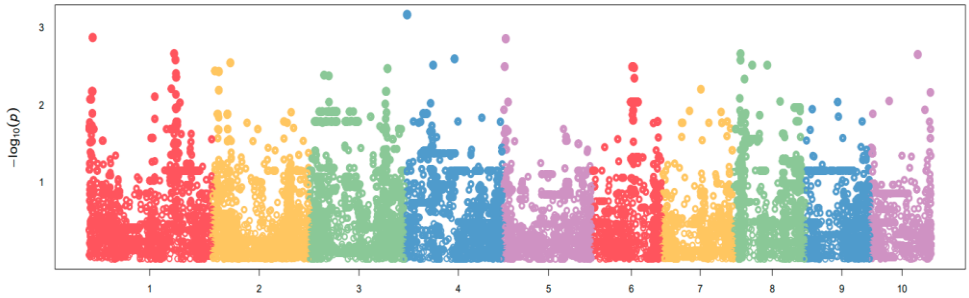
Appendix 25: Manhattan plot obtained using side volume at 11 DAS



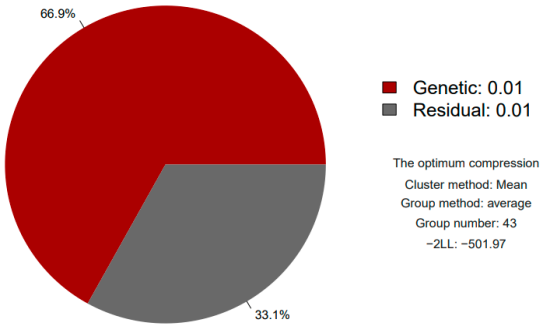
Appendix 26: Manhattan plot obtained using side volume at 26 DAS



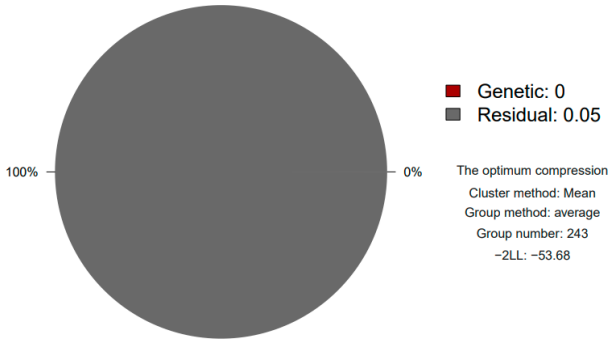
Appendix 27: Manhattan plot obtained using volume at 42 DAS



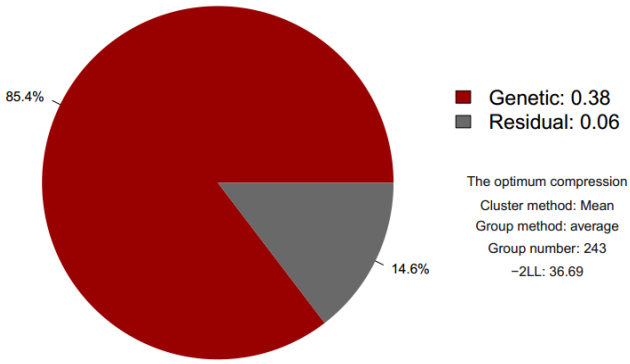
Appendix 28: The profile for optimum compression obtained using side area at 11 DAS



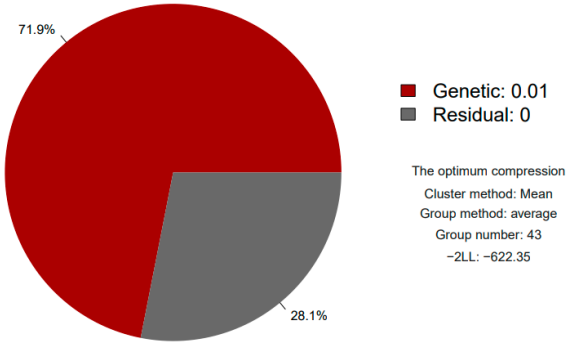
Appendix 29: The profile for the optimum compression obtained using side area at 26 DAS



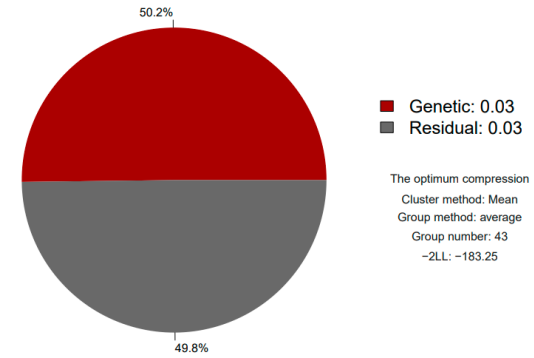
Appendix 30: The profile for the optimum compression obtained using side area at 42 DAS



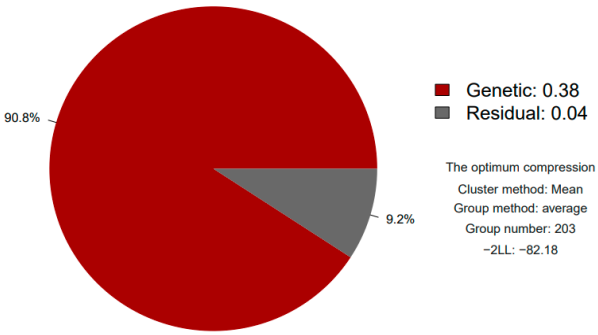
Appendix 31: The profile for the optimum compression obtained using side height at 11 DAS



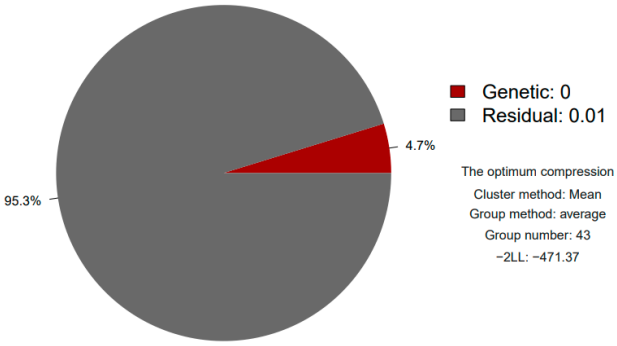
Appendix 32: The profile for the optimum compression obtained using side height at 26 DAS



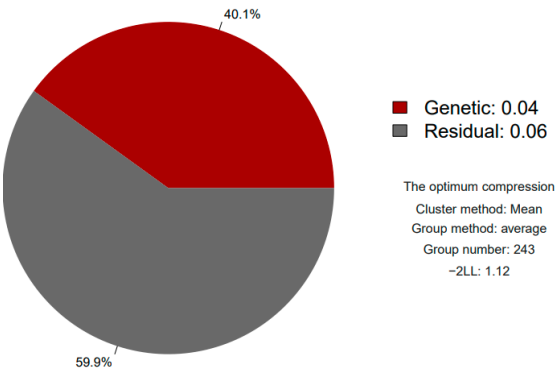
Appendix 33: The profile for the optimum compression obtained using side height at 42 DAS



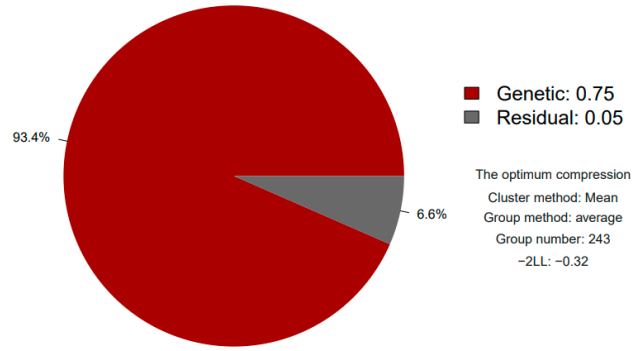
Appendix 34: The profile for the optimum compression obtained using side volume at 11 DAS



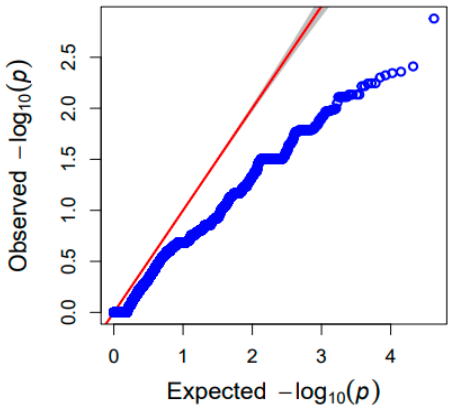
Appendix 35: The profile for the optimum compression obtained using side volume at 26 DAS



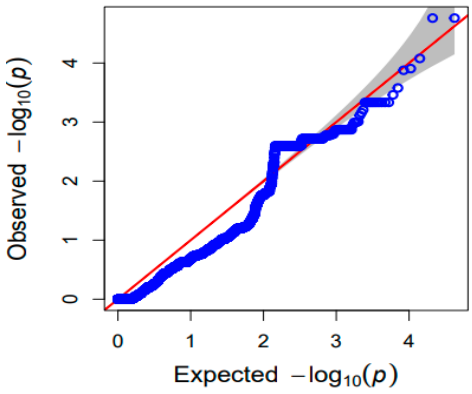
Appendix 36: The profile for the optimum compression obtained using volume at 42 DAS



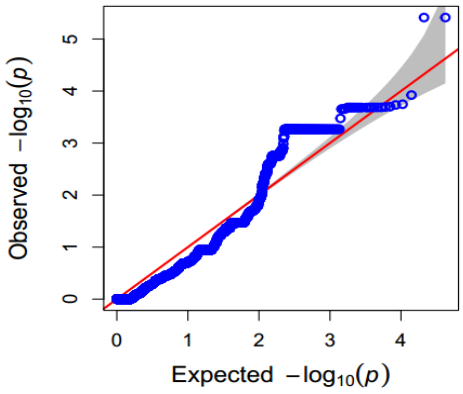
Appendix 37: Quantile-quantile (QQ) –a plot of P-values obtained using side area at 11 DAS



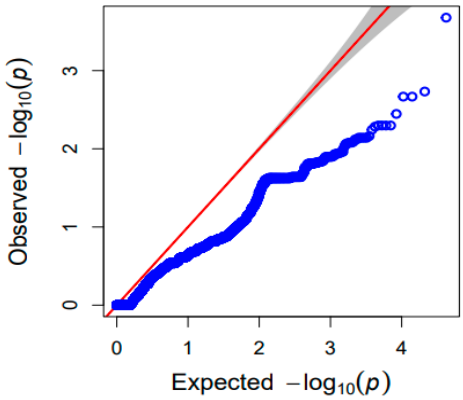
Appendix 38: Quantile-quantile (QQ) –plot of P-values obtained using side area at 26 DAS



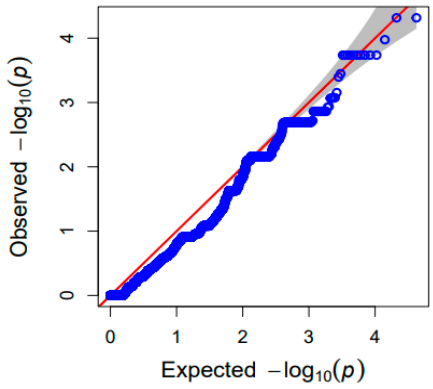
Appendix 39: Quantile-quantile (QQ) –plot of P-values obtained using side area at 42 DAS



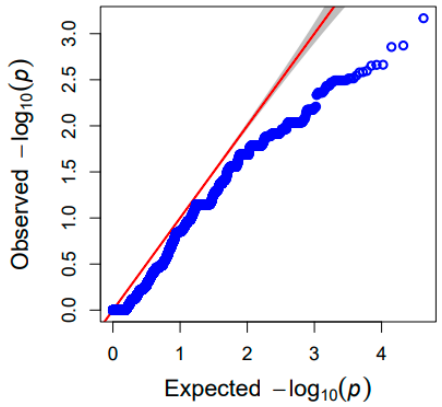
Appendix 40: Quantile-quantile (QQ) –plot of P-values obtained using volume at 11 DAS



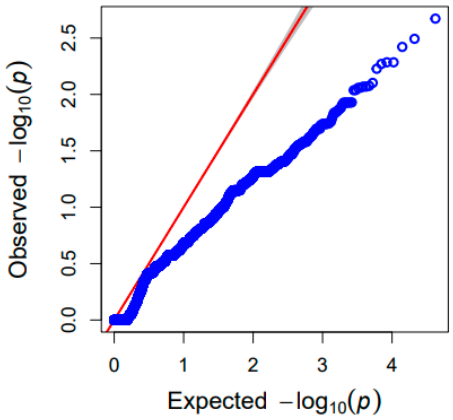
Appendix 41: Quantile-quantile (QQ) –plot of P-values obtained using volume at 26 DAS



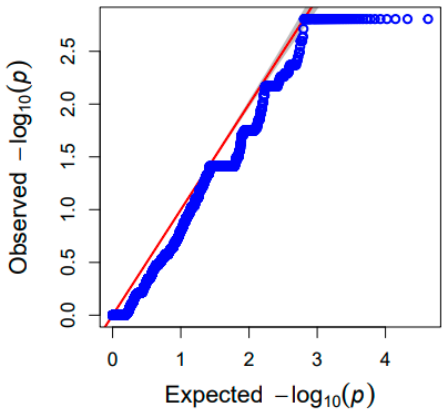
Appendix 42: Quantile-quantile (QQ) –plot of P-values obtained using volume at 42 DAS



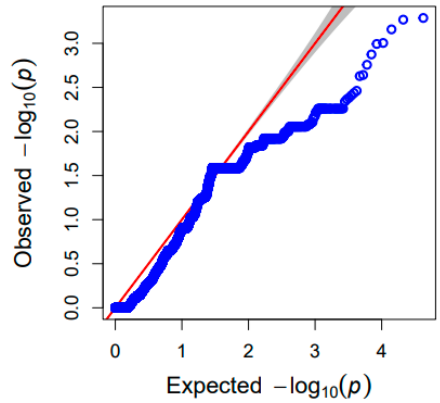
Appendix 43: Quantile-quantile (QQ) –plot of P-values obtained using side height at 11 DAS



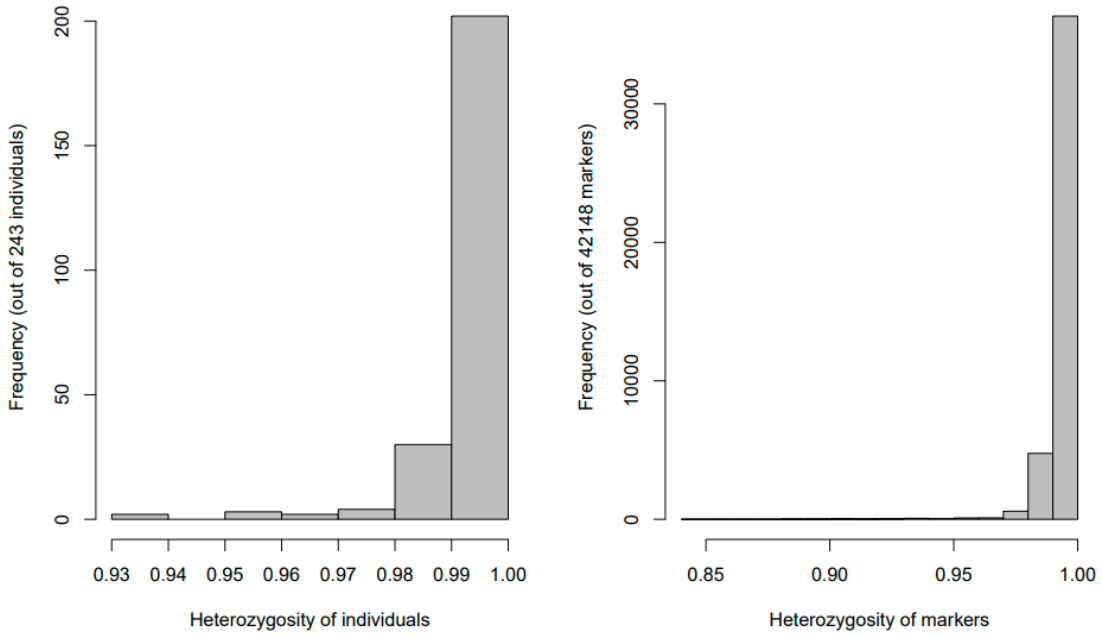
Appendix 44: Quantile-quantile (QQ) –plot of P-values obtained using side height at 26 DAS



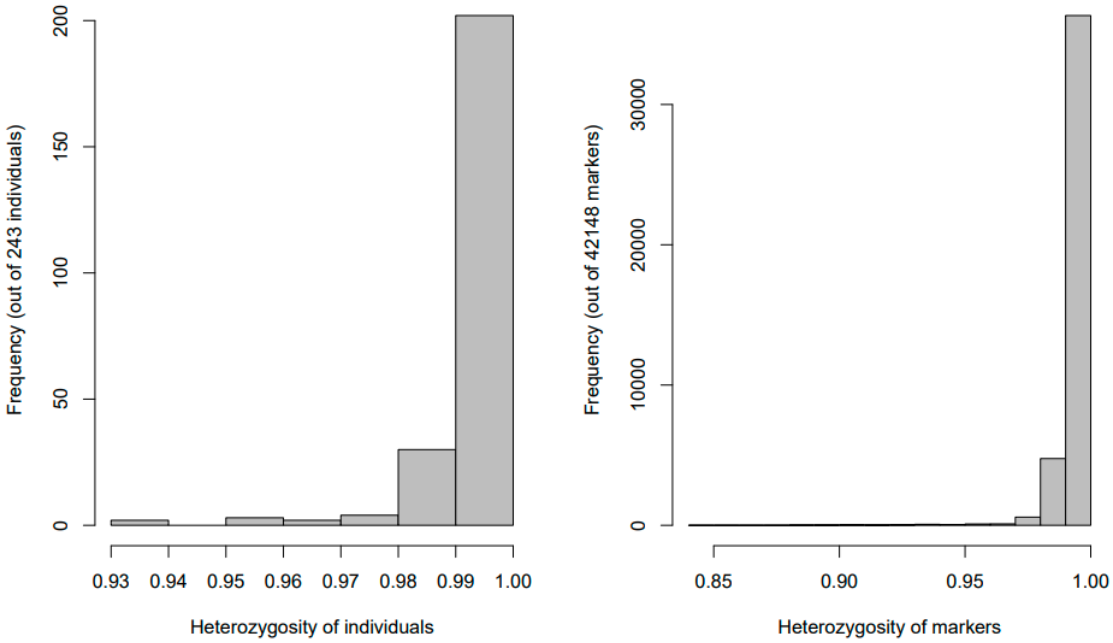
Appendix 45: Quantile-quantile (QQ) –plot of P-values obtained using side height at 42 DAS



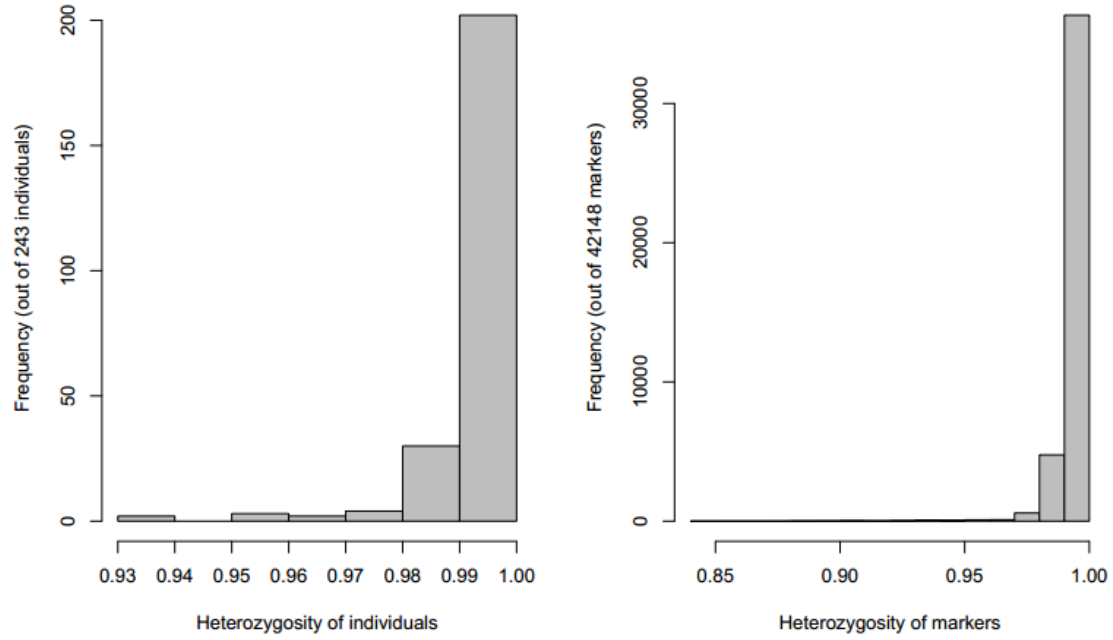
Appendix 46: Frequency of heterozygosity for individuals and markers obtained using side area at 11 DAS



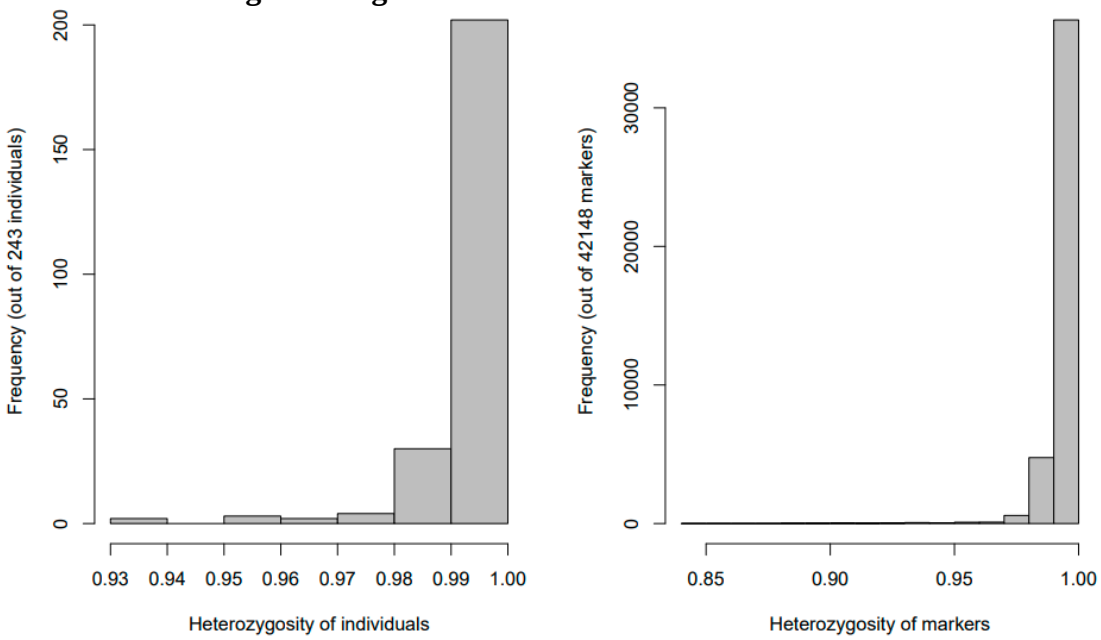
Appendix 47: Frequency of heterozygosity for individuals and markers obtained using side area at 26 DAS



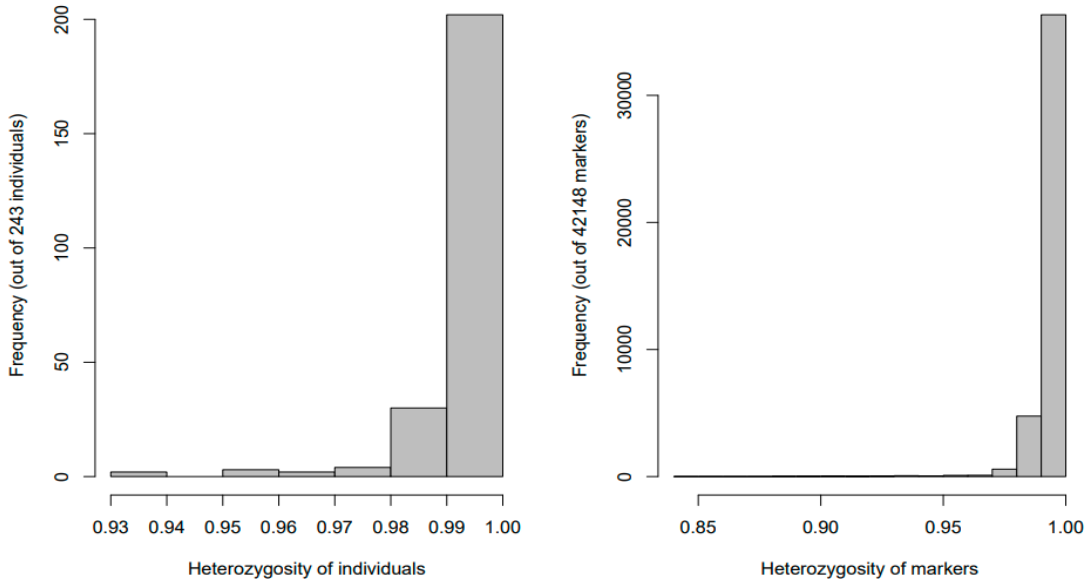
Appendix 48: Frequency of heterozygosity for individuals and markers obtained using side area at 42 DAS



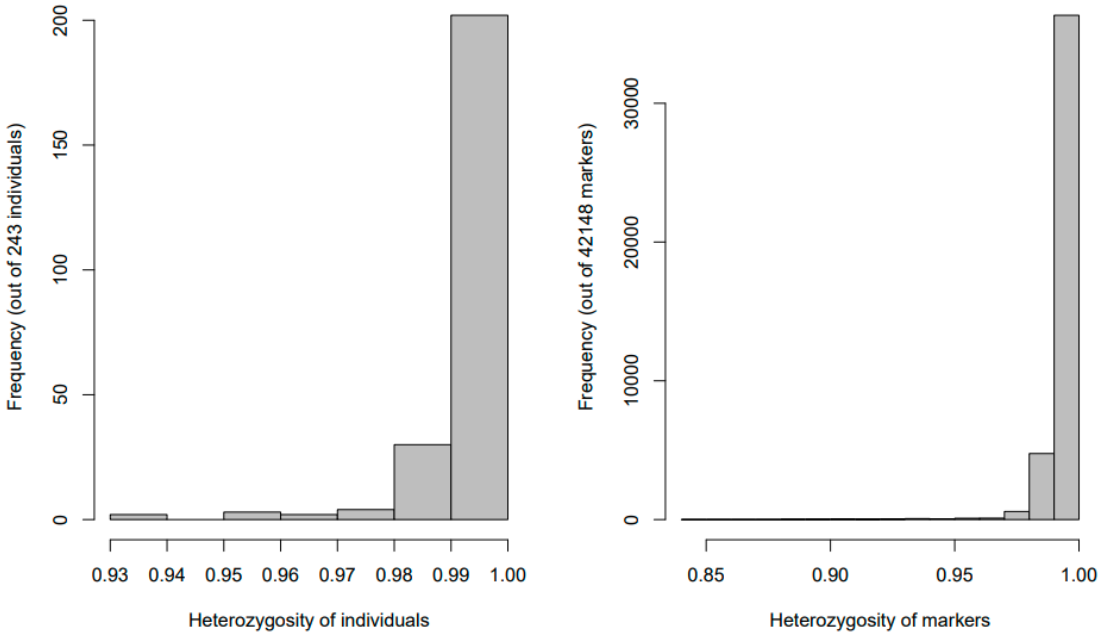
Appendix 49: Frequency and accumulative frequency of marker density obtained using side height at 11 DAS



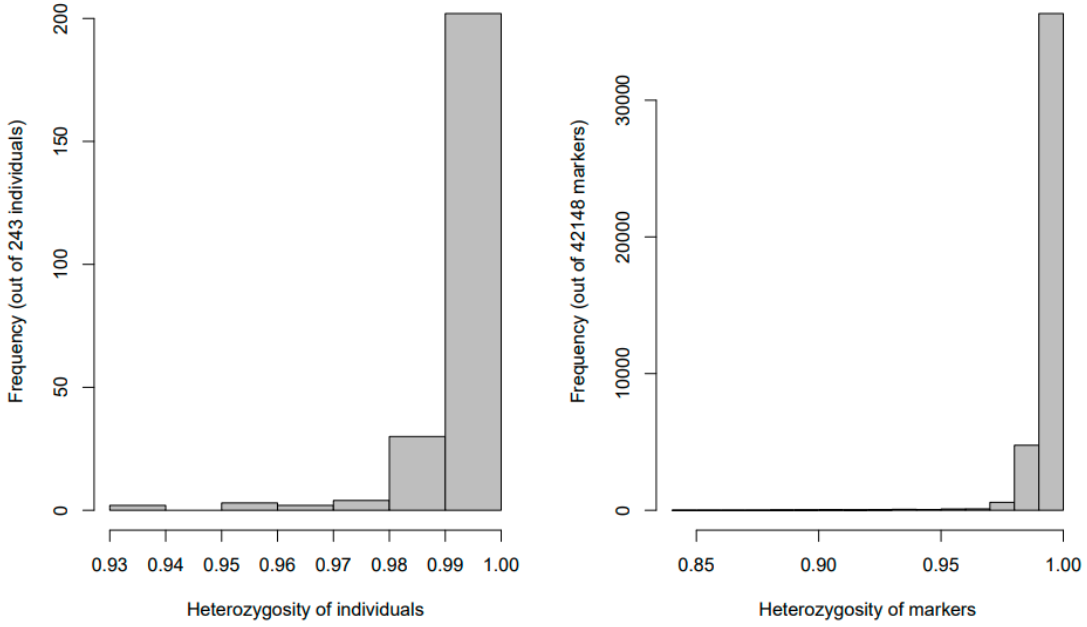
Appendix 50: Frequency of heterozygosity for individuals and markers obtained using side height at 26 DAS



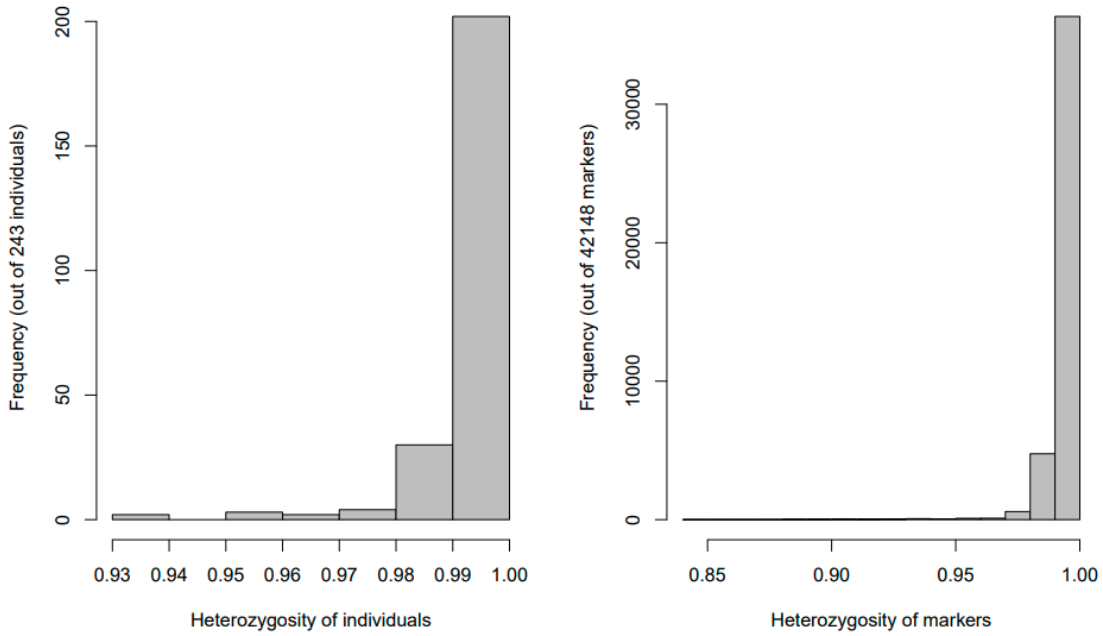
Appendix 51: Frequency of heterozygosity for individuals and markers obtained using side height at 42 DAS



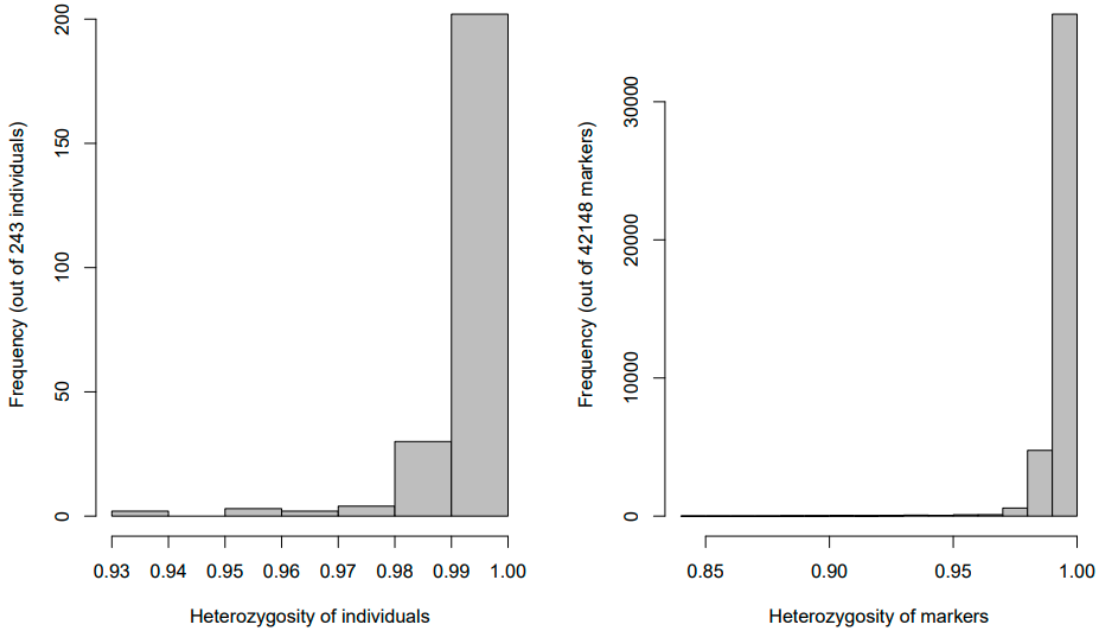
Appendix 52: Frequency of heterozygosity for individuals and markers obtained using volume at 11 DAS



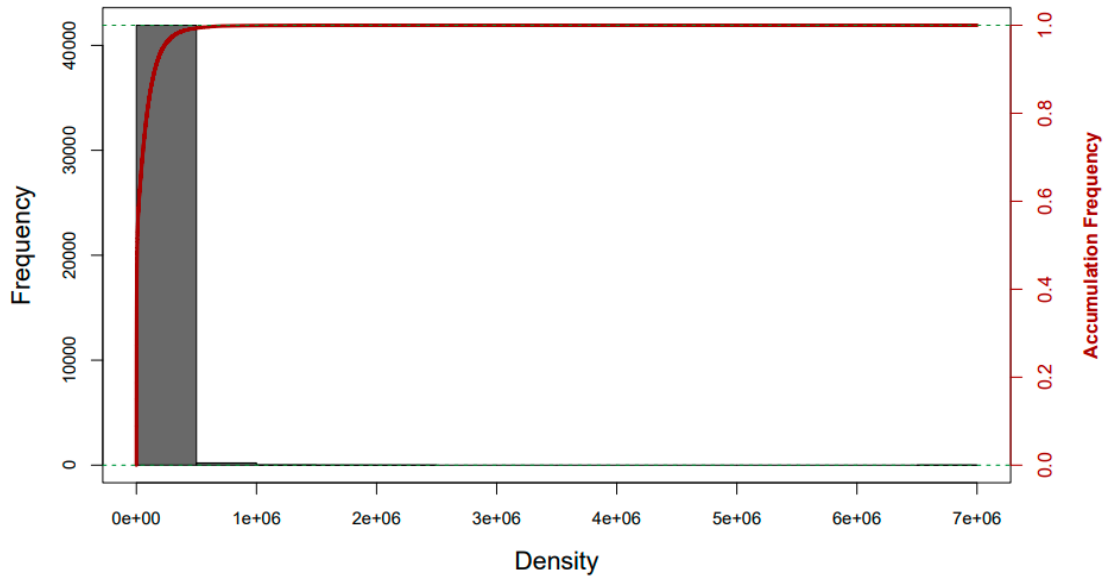
Appendix 53: Frequency of heterozygosity for individuals and markers obtained using volume at 26 DAS



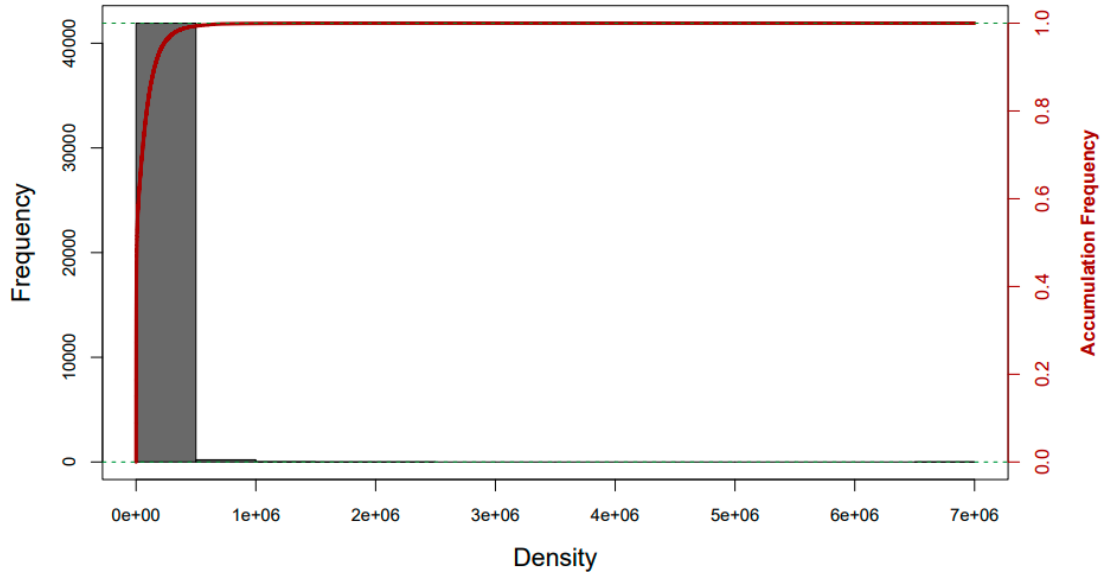
Appendix 54: Frequency of heterozygosity for individuals and markers obtained using volume at 42 DAS



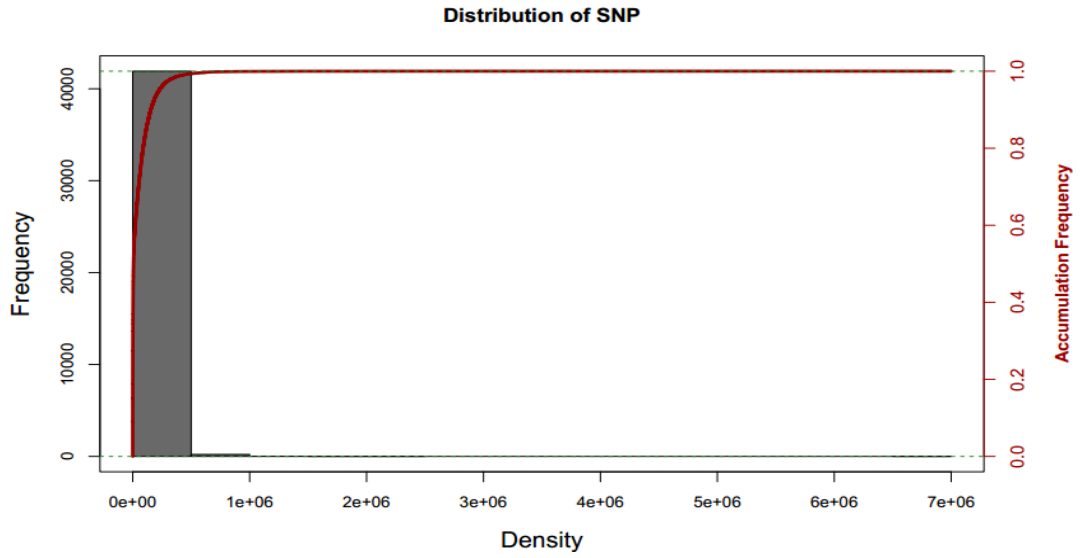
Appendix 55: Frequency and accumulative frequency of marker density obtained using side area at 11 DAS



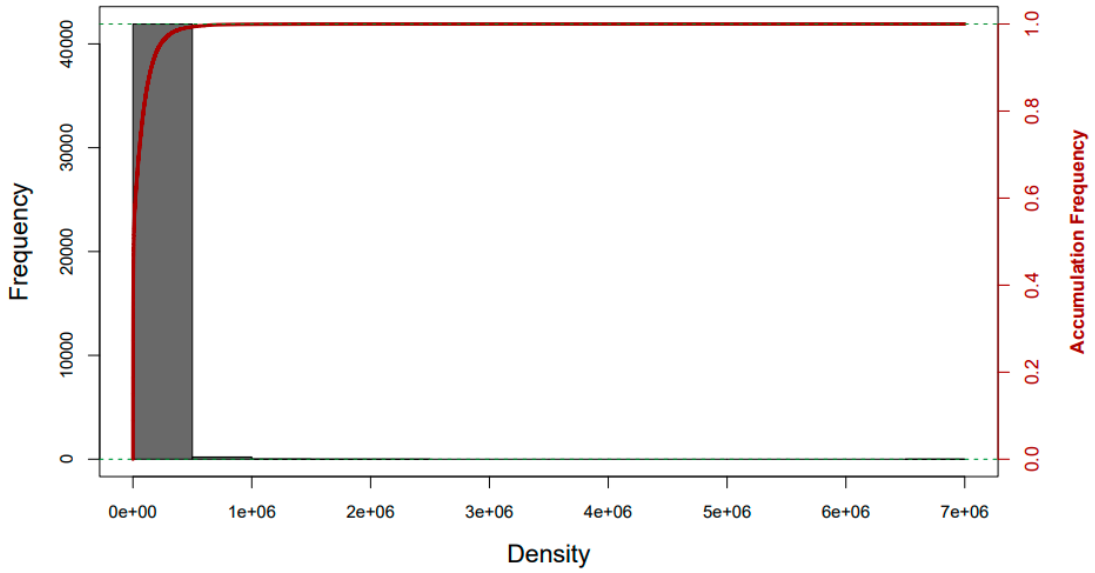
Appendix 56: Frequency and accumulative frequency of marker density obtained using side area at 26 DAS



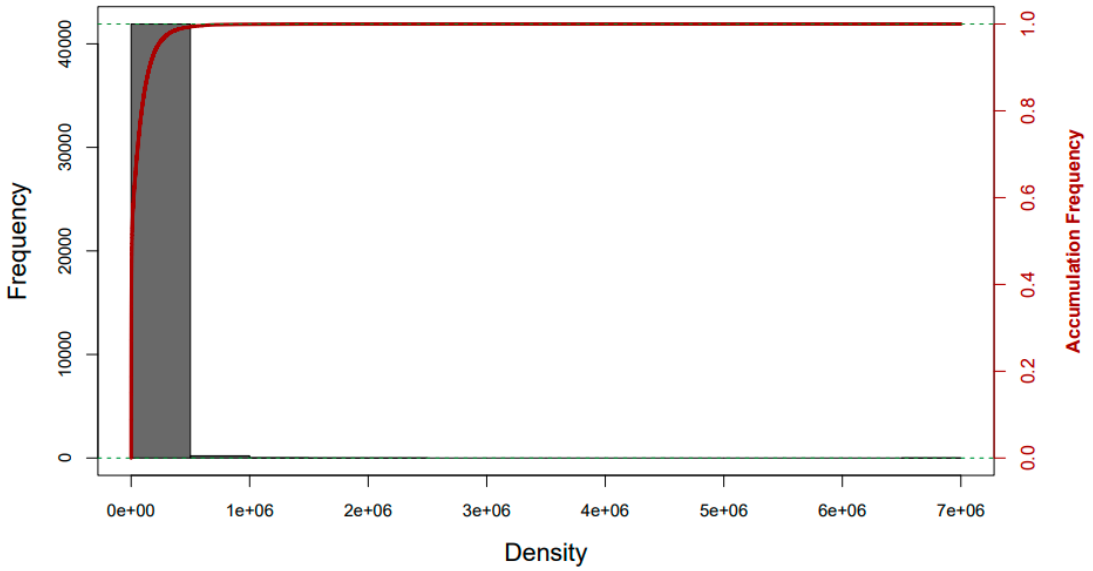
Appendix 57: Frequency and accumulative frequency of marker density obtained using side area at 42 DAS



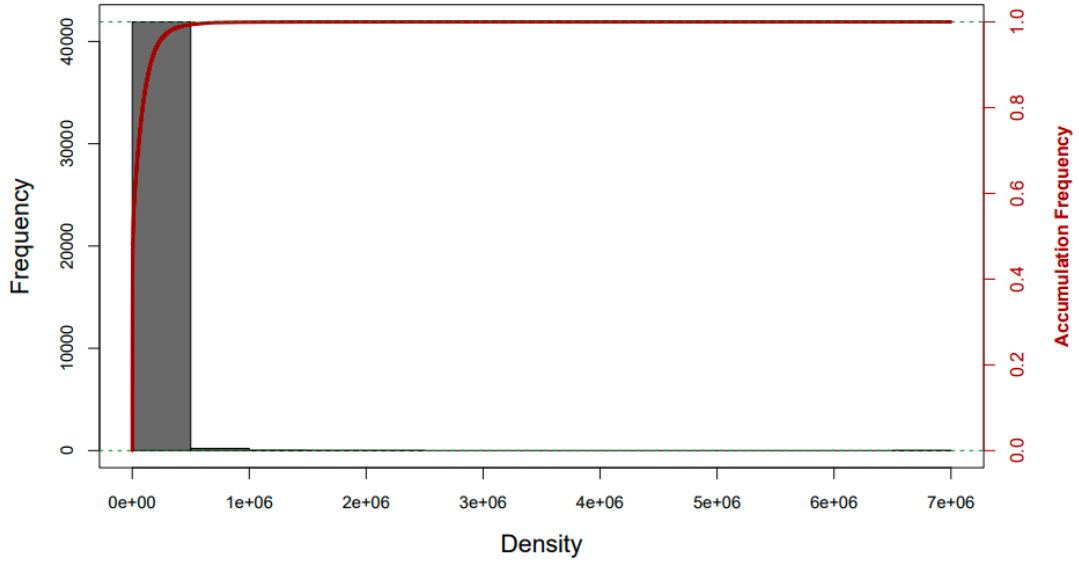
Appendix 58: Frequency and accumulative frequency of marker density obtained using side height at 11 DAS



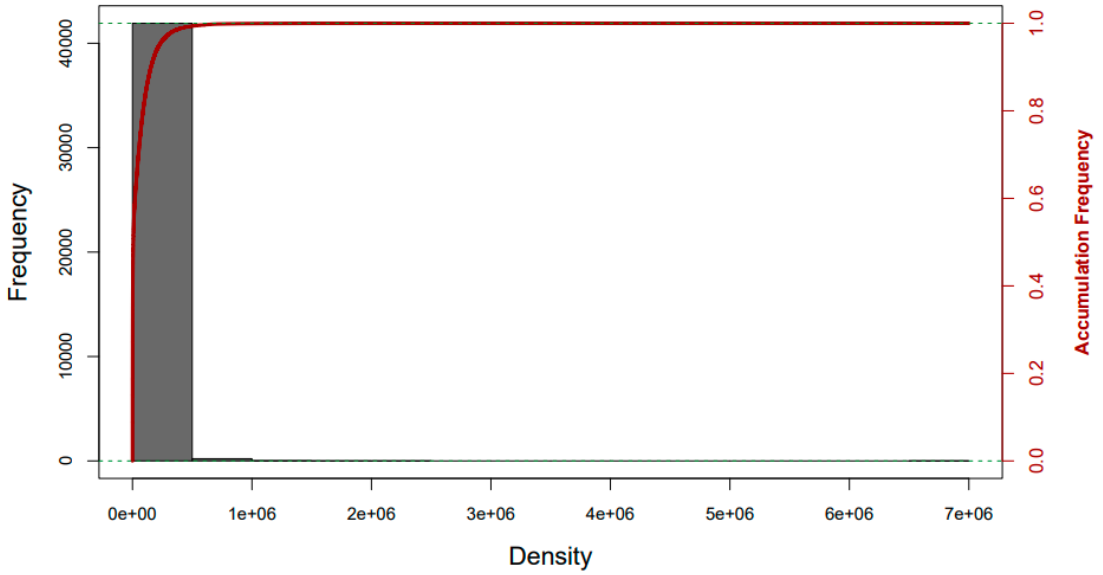
Appendix 59: Frequency and accumulative frequency of marker density obtained using side height at 26 DAS



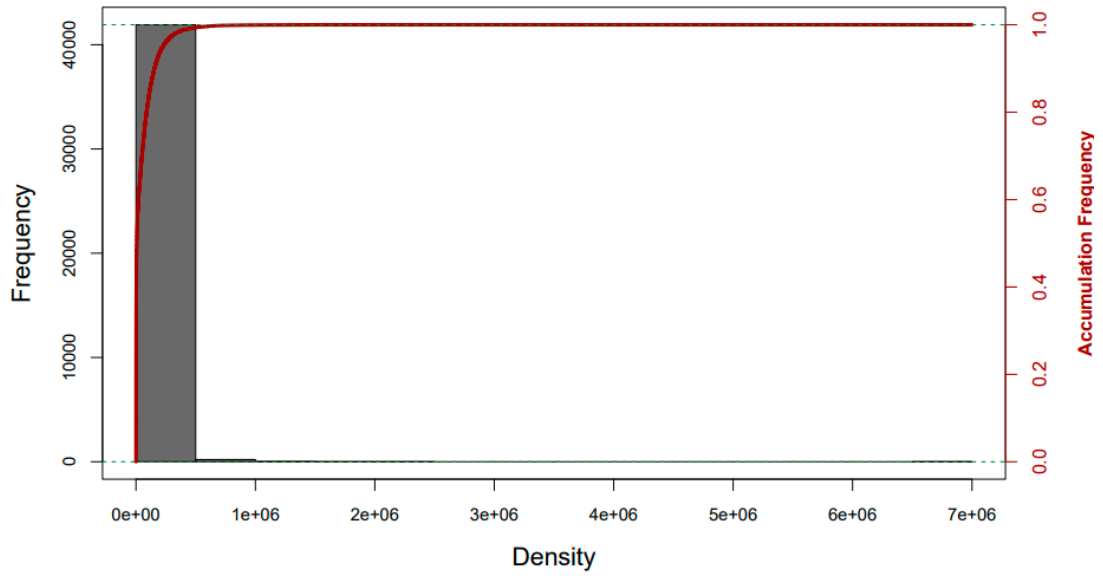
Appendix 60: Frequency and accumulative frequency of marker density obtained using side height at 42 DAS



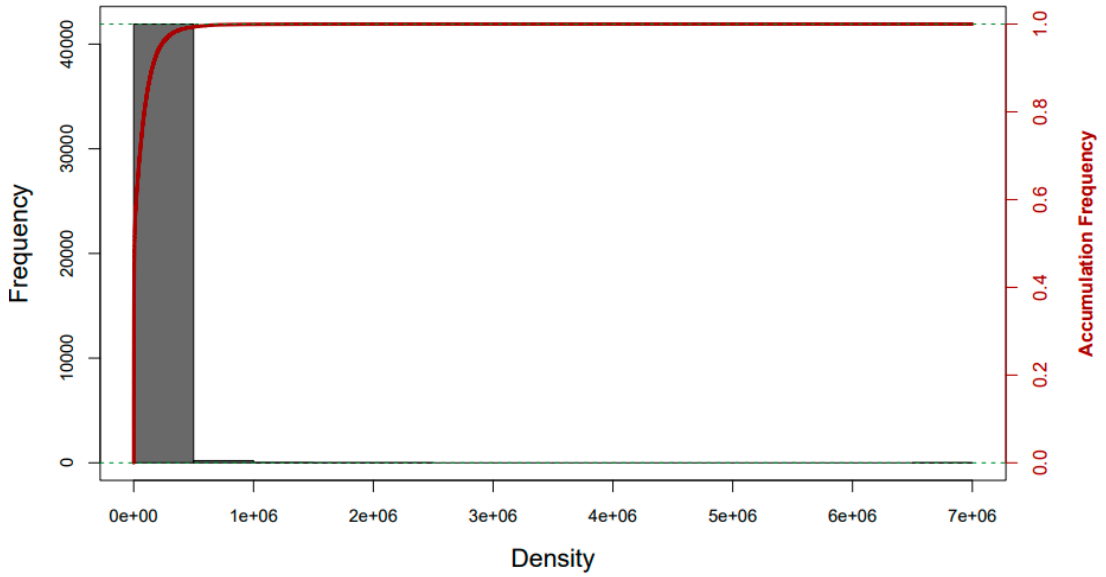
Appendix 61: Frequency and accumulative frequency of marker density obtained using volume at 11 DAS



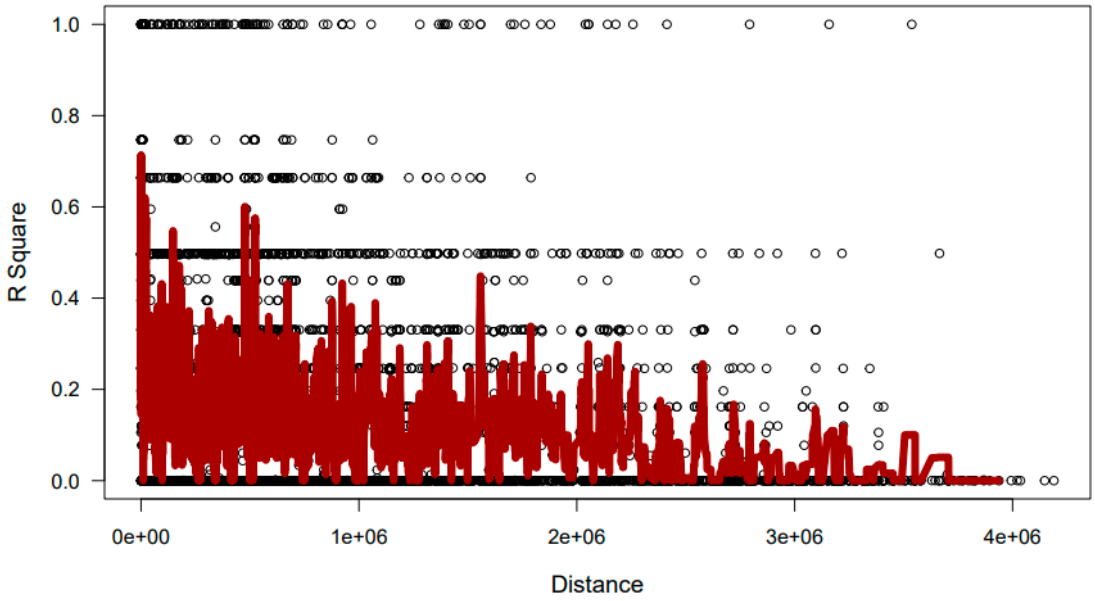
Appendix 62: Frequency and accumulative frequency of marker density obtained using side volume at 26 DAS



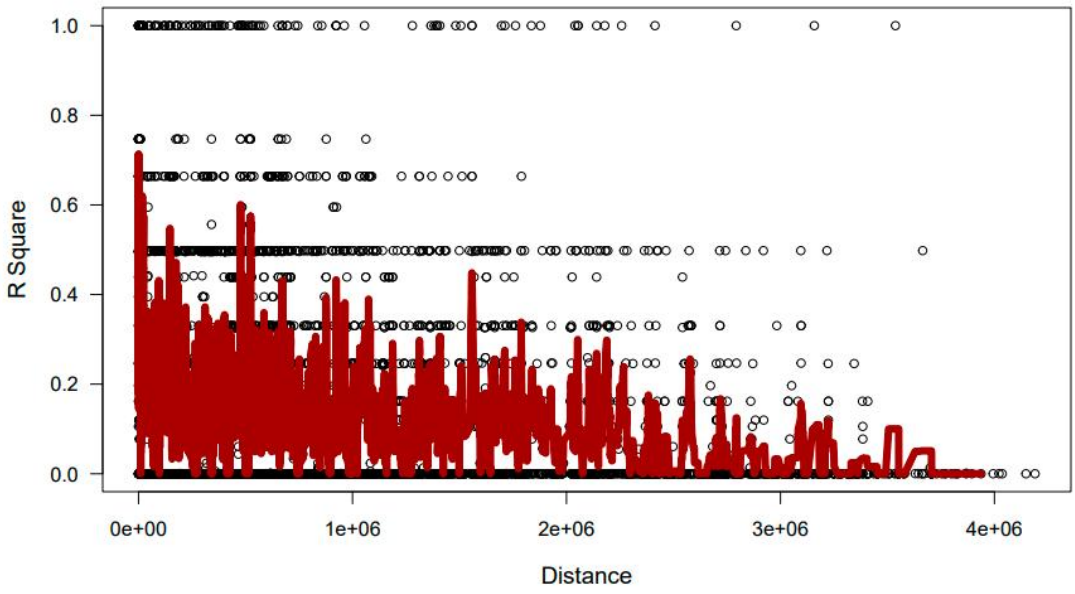
Appendix 63: Frequency and accumulative frequency of marker density obtained using volume at 42 DAS



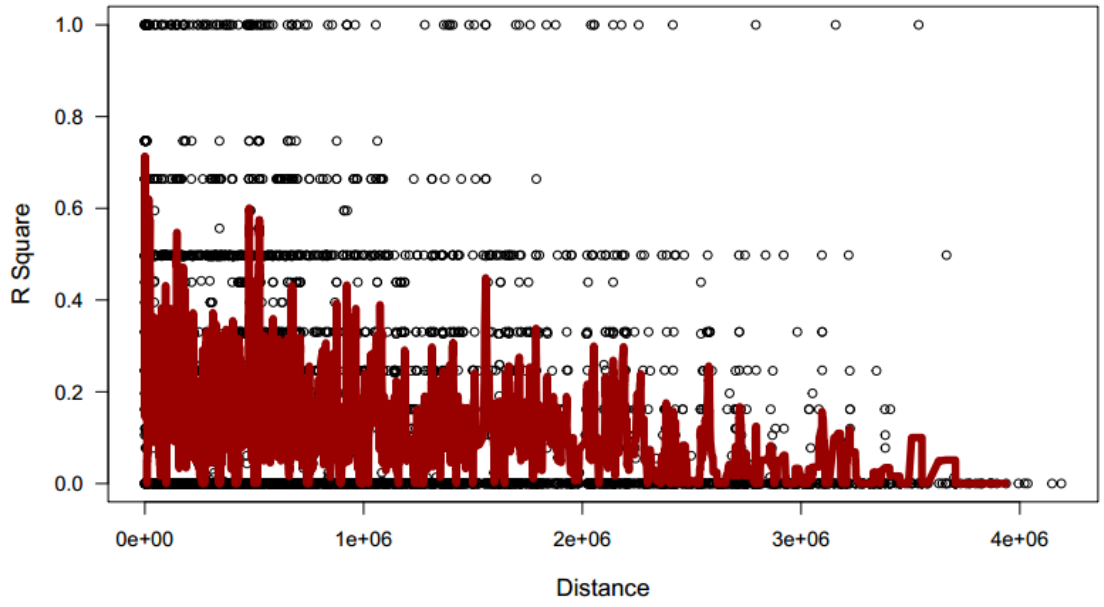
Appendix 64: Linkage disequilibrium (LD) decay over distance obtained using side area at 11 DAS



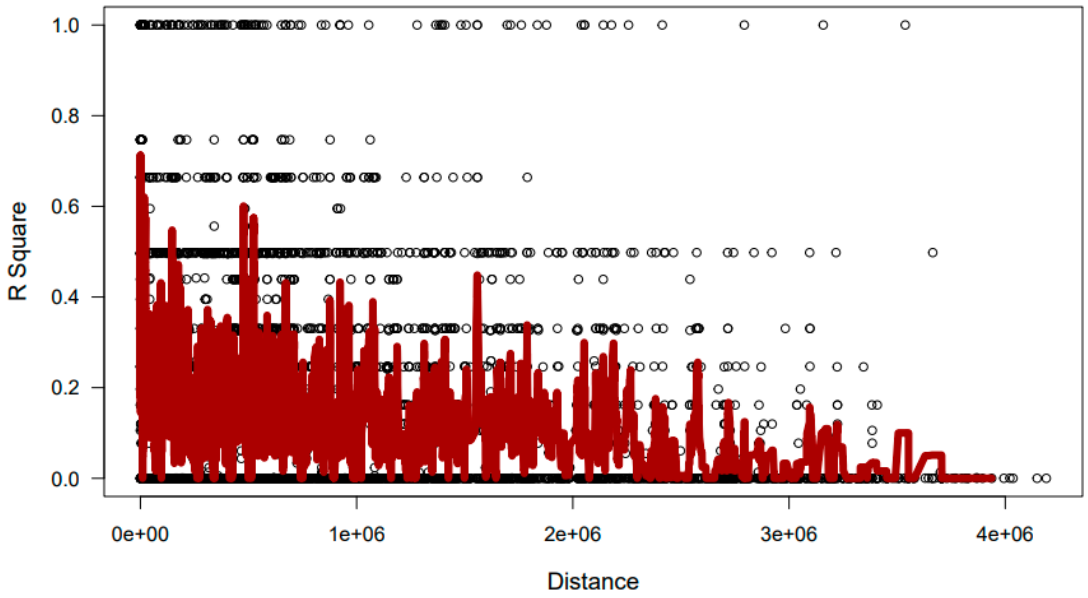
Appendix 65: Linkage disequilibrium (LD) decay over distance obtained using side area at 26 DAS



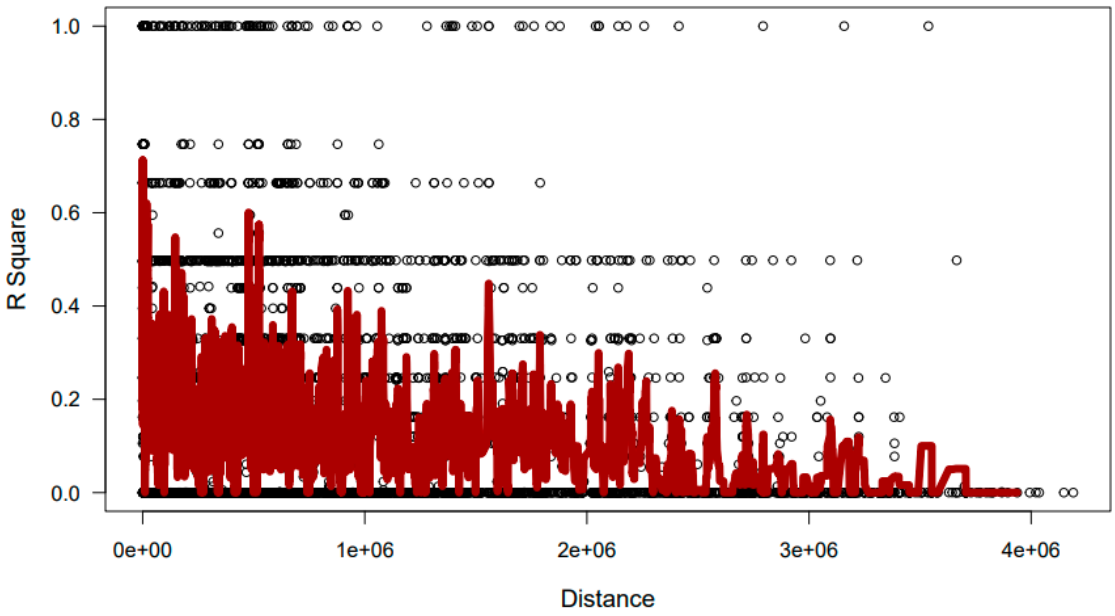
Appendix 66: Linkage disequilibrium (LD) decay over distance obtained using side area at 42 DAS



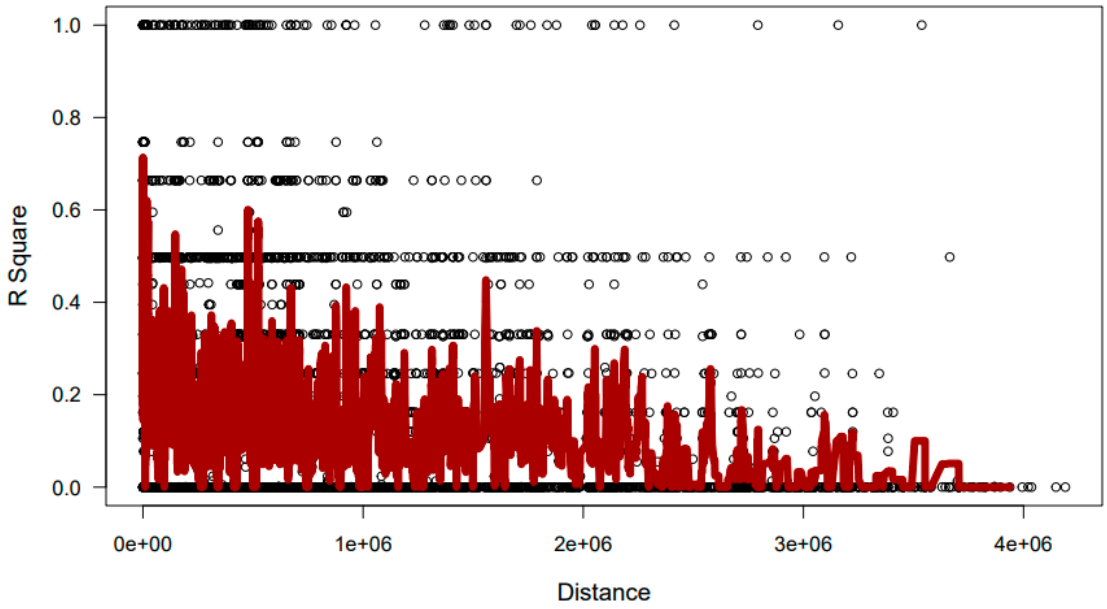
Appendix 67: Linkage disequilibrium (LD) decay over distance obtained using side height at 11 DAS



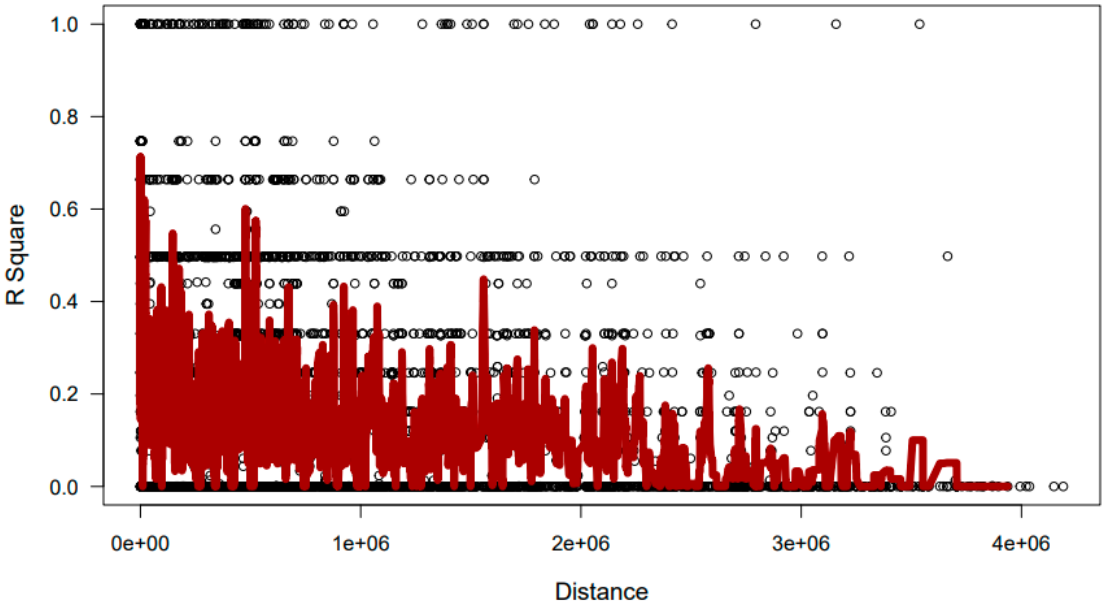
Appendix 68: Linkage disequilibrium (LD) decay over distance obtained using side height at 26 DAS



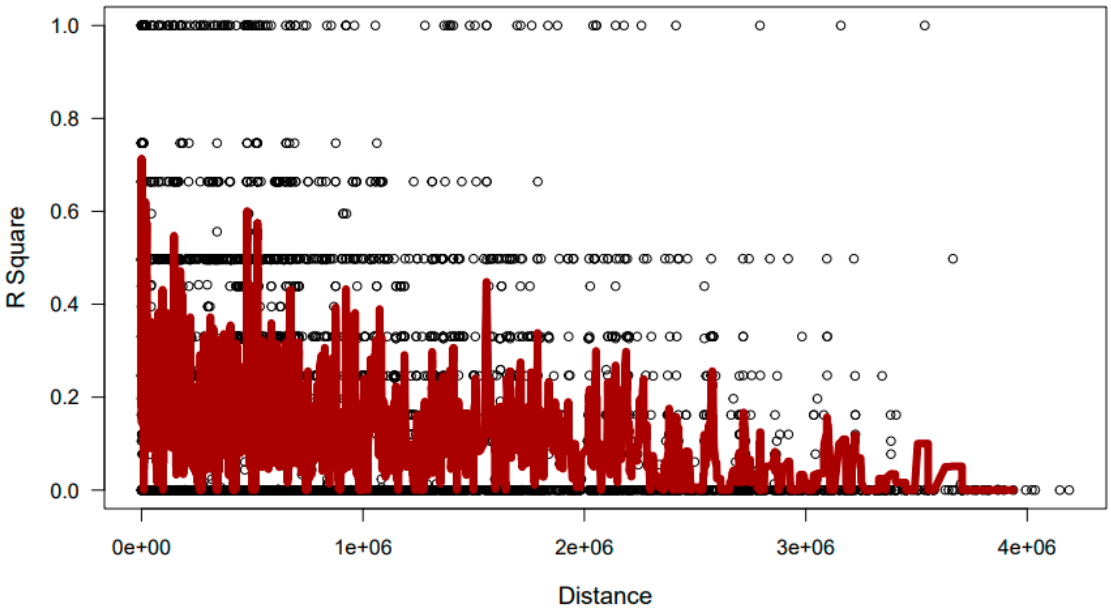
Appendix 69: Linkage disequilibrium (LD) decay over distance obtained using side height at 42 DAS



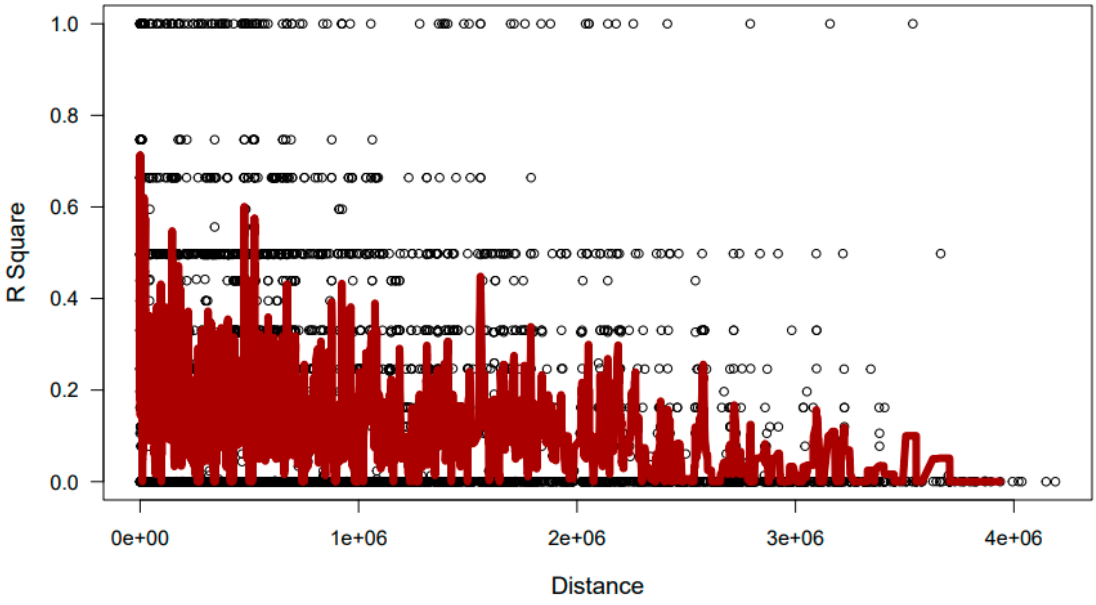
Appendix 70: Linkage disequilibrium (LD) decay over distance obtained using side volume at 11 DAS



Appendix 71: Linkage disequilibrium (LD) decay over distance obtained using side volume at 26 DAS



Appendix 72: Linkage disequilibrium (LD) decay over distance obtained using volume at 42 DAS



Appendix 73: Genomic Breeding values and prediction error variance obtained using volume at 11 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	-9.95E-06	4.05E-05	2.925905	2.925895	2.925895
A148	1	1	1	-9.95E-06	4.05E-05	2.925875	2.925865	2.925865
A188	1	1	1	-9.95E-06	4.05E-05	2.925854	2.925844	2.925844
A3	2	1	2	-9.41E-05	3.50E-05	2.920768	2.920674	2.920674
A310	1	1	1	-9.95E-06	4.05E-05	2.925932	2.925922	2.925922
A347	1	1	1	-9.95E-06	4.05E-05	2.925882	2.925872	2.925872
A374	1	1	1	-9.95E-06	4.05E-05	2.925933	2.925923	2.925923
A619	1	1	1	-9.95E-06	4.05E-05	2.925948	2.925938	2.925938
AS5707	3	1	3	-7.60E-06	1.01E-05	2.927272	2.927264	2.927264
B100	1	1	1	-9.95E-06	4.05E-05	2.926007	2.925997	2.925997
B102	1	1	1	-9.95E-06	4.05E-05	2.925984	2.925974	2.925974
B106	1	1	1	-9.95E-06	4.05E-05	2.925853	2.925843	2.925843
B108	1	1	1	-9.95E-06	4.05E-05	2.926107	2.926097	2.926097
B109	1	1	1	-9.95E-06	4.05E-05	2.92592	2.92591	2.92591
B110	1	1	1	-9.95E-06	4.05E-05	2.925952	2.925942	2.925942
B111	1	1	1	-9.95E-06	4.05E-05	2.926013	2.926003	2.926003
B112	4	1	4	-0.00011	5.57E-05	2.965554	2.96544	2.96544
B113	1	1	1	-9.95E-06	4.05E-05	2.927137	2.927127	2.927127
B37	1	1	1	-9.95E-06	4.05E-05	2.925987	2.925977	2.925977
B73	1	1	1	-9.95E-06	4.05E-05	2.92588	2.92587	2.92587

Appendix 74: Genomic Breeding values and prediction error variance obtained using volume at 26 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	2.37E-05	0.000306	2.895144	2.895167	2.895167
A148	2	1	2	-0.00025	0.000316	2.895268	2.895015	2.895015
A188	3	1	3	-0.00088	0.000345	2.895808	2.894931	2.894931
A3	4	1	4	0.007037	0.003499	3.04619	3.053227	3.053227
A310	5	1	5	-0.00017	0.00032	2.894474	2.894306	2.894306
A347	6	1	6	-0.00026	0.000327	2.894543	2.894286	2.894286
A374	7	1	7	0.000209	0.000341	2.895177	2.895386	2.895386
A619	8	1	8	0.000104	0.000307	2.894365	2.89447	2.89447
AS5707	9	1	9	-0.00404	0.0012	2.899891	2.895853	2.895853
B100	10	1	10	0.000158	0.000343	2.894693	2.89485	2.89485
B102	11	1	11	4.53E-05	0.000339	2.894883	2.894928	2.894928
B106	12	1	12	-0.00034	0.000314	2.894805	2.894464	2.894464
B108	13	1	13	-0.00046	0.000406	2.89443	2.893965	2.893965
B109	14	1	14	-0.00052	0.000276	2.89424	2.893725	2.893725
B110	15	1	15	-6.81E-05	0.000343	2.894596	2.894528	2.894528
B111	16	1	16	-0.00054	0.000344	2.894583	2.894045	2.894045
B112	17	1	17	0.000331	0.005598	2.85285	2.853181	2.853181
B113	18	1	18	0.000802	0.000779	2.893952	2.894754	2.894754
B37	19	1	19	0.00066	0.000313	2.894335	2.894995	2.894995

Appendix 75: 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
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

Appendix 76: Genomic Breeding values and prediction error variance obtained using side area at 11 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	-0.00042	3.53E-05	2.926164	2.925748	2.925748
A148	1	1	1	-0.00042	3.53E-05	2.92616	2.925745	2.925745
A188	1	1	1	-0.00042	3.53E-05	2.926205	2.925789	2.925789
A3	2	1	2	-0.00381	0.001116	2.936709	2.932903	2.932903
A310	1	1	1	-0.00042	3.53E-05	2.926118	2.925703	2.925703
A347	1	1	1	-0.00042	3.53E-05	2.926112	2.925697	2.925697
A374	1	1	1	-0.00042	3.53E-05	2.926183	2.925768	2.925768
A619	1	1	1	-0.00042	3.53E-05	2.926102	2.925686	2.925686
AS5707	3	1	3	0.001623	0.000341	2.927059	2.928681	2.928681
B100	1	1	1	-0.00042	3.53E-05	2.92616	2.925744	2.925744
B102	1	1	1	-0.00042	3.53E-05	2.926167	2.925751	2.925751
B106	1	1	1	-0.00042	3.53E-05	2.926112	2.925697	2.925697
B108	1	1	1	-0.00042	3.53E-05	2.926207	2.925791	2.925791
B109	1	1	1	-0.00042	3.53E-05	2.92609	2.925674	2.925674
B110	1	1	1	-0.00042	3.53E-05	2.926147	2.925731	2.925731
B111	1	1	1	-0.00042	3.53E-05	2.926169	2.925753	2.925753
B112	4	1	4	-0.00514	0.00196	2.941938	2.936799	2.936799
B113	1	1	1	-0.00042	3.53E-05	2.92655	2.926135	2.926135
B37	1	1	1	-0.00042	3.53E-05	2.926128	2.925712	2.925712
B73	1	1	1	-0.00042	3.53E-05	2.926082	2.925666	2.925666
B84	1	1	1	-0.00042	3.53E-05	2.926335	2.92592	2.92592

Appendix 77: Genomic Breeding values and prediction error variance obtained using side area at 26 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	7.20E-09	0.000192	2.902378	2.902378	2.902378
A148	2	1	2	-7.96E-09	0.000192	2.902383	2.902383	2.902383
A188	3	1	3	-2.42E-08	0.000192	2.902698	2.902698	2.902698
A3	4	1	4	4.17E-09	0.000193	2.982218	2.982218	2.982218
A310	5	1	5	2.19E-09	0.000192	2.902032	2.902032	2.902032
A347	6	1	6	-1.04E-08	0.000192	2.902005	2.902005	2.902005
A374	7	1	7	1.94E-08	0.000192	2.902482	2.902482	2.902482
A619	8	1	8	1.65E-08	0.000192	2.901939	2.901939	2.901939
AS5707	9	1	9	-2.64E-08	0.000193	2.907868	2.907868	2.907868
B100	10	1	10	2.63E-08	0.000192	2.902289	2.902289	2.902289
B102	11	1	11	2.31E-08	0.000192	2.902355	2.902355	2.902355
B106	12	1	12	-1.06E-08	0.000192	2.902055	2.902055	2.902055
B108	13	1	13	-4.66E-08	0.000192	2.902469	2.902469	2.902469
B109	14	1	14	-2.93E-08	0.000192	2.901847	2.901847	2.901847
B110	15	1	15	-7.89E-09	0.000192	2.902194	2.902194	2.902194
B111	16	1	16	-1.96E-08	0.000192	2.902306	2.902306	2.902306
B112	17	1	17	-3.09E-08	0.000193	2.975183	2.975183	2.975183
B113	18	1	18	9.03E-08	0.000193	2.904208	2.904208	2.904208
B37	19	1	19	5.39E-08	0.000192	2.902059	2.902059	2.902059
B73	20	1	20	-2.62E-08	0.000192	2.901814	2.901814	2.901814
B84	21	1	21	-1.06E-08	0.000192	2.903596	2.903596	2.903596

Appendix 78: 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

Appendix 79: Genomic Breeding values and prediction error variance obtained using side height at 11 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	-0.00033	2.13E-05	2.928175	2.927847	2.927847
A148	1	1	1	-0.00033	2.13E-05	2.928164	2.927836	2.927836
A188	1	1	1	-0.00033	2.13E-05	2.928215	2.927887	2.927887
A3	2	1	2	-0.00026	0.000829	2.929612	2.929357	2.929357
A310	1	1	1	-0.00033	2.13E-05	2.928168	2.92784	2.92784
A347	1	1	1	-0.00033	2.13E-05	2.92821	2.927882	2.927882
A374	1	1	1	-0.00033	2.13E-05	2.928213	2.927885	2.927885
A619	1	1	1	-0.00033	2.13E-05	2.928089	2.927761	2.927761
AS5707	3	1	3	0.003215	0.000256	2.928155	2.93137	2.93137
B100	1	1	1	-0.00033	2.13E-05	2.928142	2.927814	2.927814
B102	1	1	1	-0.00033	2.13E-05	2.928143	2.927815	2.927815
B106	1	1	1	-0.00033	2.13E-05	2.928149	2.927821	2.927821
B108	1	1	1	-0.00033	2.13E-05	2.928323	2.927995	2.927995
B109	1	1	1	-0.00033	2.13E-05	2.928138	2.92781	2.92781
B110	1	1	1	-0.00033	2.13E-05	2.928234	2.927906	2.927906
B111	1	1	1	-0.00033	2.13E-05	2.928233	2.927905	2.927905
B112	4	1	4	-0.00529	0.00149	2.953266	2.947974	2.947974
B113	1	1	1	-0.00033	2.13E-05	2.928291	2.927963	2.927963
B37	1	1	1	-0.00033	2.13E-05	2.928169	2.927841	2.927841
B73	1	1	1	-0.00033	2.13E-05	2.928153	2.927825	2.927825
B84	1	1	1	-0.00033	2.13E-05	2.928213	2.927885	2.927885
B89	5	1	5	-0.00478	0.000341	2.928322	2.923538	2.923538

Appendix 80: Genomic Breeding values and prediction error variance obtained using side height at 26 DAS

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction	Pred_Heritable
38-11	1	1	1	-0.00039	0.000135	2.918921	2.918528	2.918528
A148	1	1	1	-0.00039	0.000135	2.919101	2.918707	2.918707
A188	1	1	1	-0.00039	0.000135	2.919288	2.918895	2.918895
A3	2	1	2	0.011206	0.002231	2.963717	2.974923	2.974923
A310	1	1	1	-0.00039	0.000135	2.918699	2.918306	2.918306
A347	1	1	1	-0.00039	0.000135	2.918998	2.918605	2.918605
A374	1	1	1	-0.00039	0.000135	2.918779	2.918386	2.918386
A619	1	1	1	-0.00039	0.000135	2.918579	2.918186	2.918186
AS5707	3	1	3	0.001327	0.000664	2.91178	2.913106	2.913106
B100	1	1	1	-0.00039	0.000135	2.9183	2.917907	2.917907
B102	1	1	1	-0.00039	0.000135	2.918448	2.918055	2.918055
B106	1	1	1	-0.00039	0.000135	2.919172	2.918778	2.918778
B108	1	1	1	-0.00039	0.000135	2.917764	2.917371	2.917371
B109	1	1	1	-0.00039	0.000135	2.918736	2.918343	2.918343
B110	1	1	1	-0.00039	0.000135	2.91862	2.918226	2.918226
B111	1	1	1	-0.00039	0.000135	2.918276	2.917882	2.917882
B112	4	1	4	-0.00056	0.00373	2.700102	2.699546	2.699546
B113	1	1	1	-0.00039	0.000135	2.911952	2.911558	2.911558
B37	1	1	1	-0.00039	0.000135	2.918381	2.917988	2.917988
B73	1	1	1	-0.00039	0.000135	2.918968	2.918574	2.918574
B84	1	1	1	-0.00039	0.000135	2.919043	2.91865	2.91865

Appendix 81: 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

Appendix 82: NACOSTI Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Ref No: 913592

Date of Issue: 09/May/2022

RESEARCH LICENSE

This is to Certify that Mr., Dominic ohare Mong'are of Chuka University, has been licensed to conduct research in Tharaka-Nithi on the topic: AN IMPROVED ENRICHED COMPRESSED MIXED LINEAR MODEL USING NON-HIERARCHICAL CLUSTERING ALGORITHMS: AN APPLICATION TO GENOME WIDE ASSOCIATION STUDIES for the period ending : 09/May/2023.

License No: NACOSTI/P/22/17387

913592

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Verification QR Code:

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.