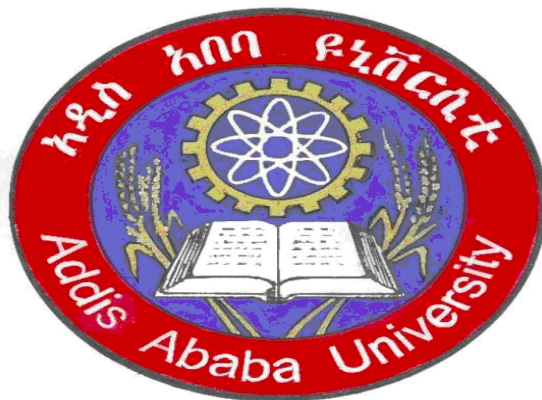




**ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS**

**QTL ANALYSIS OF ETHIOPIAN MAIZE DATA USING MOLECULAR
MARKERS:**

INTERVAL MAPPING AND LINEAR MIXED MODEL APPROACHES

BY

Worku Biyadgie

**A Thesis submitted to the Office of Graduate Programs of Addis Ababa
University in Partial fulfillment of the requirements for the Degree of Master
Science in Statistics**

June, 2014
Addis Ababa
Ethiopia

**ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS**

**QTL ANALYSIS OF ETHIOPIAN MAIZE DATA USING MOLECULAR
MARKERS:**

INTERVAL MAPPING AND LINEAR MIXED MODEL APPROACHES

BY

Worku Biyadgie

**A Thesis submitted to the Office of Graduate Programs of Addis Ababa
University in Partial fulfillment of the requirements for the Degree of Master
Science in Statistics**

Advisor: Dr. Girma Taye

June, 2014
Addis Ababa
Ethiopia

School of Graduate Studies

This is to certify that the thesis prepared by Worku Biyadgie, entitled: **QTL ANALYSIS OF ETHIOPIAN MAIZE DATA USING MOLECULAR MARKERS: INTERVAL MAPPING AND LINEAR MIXED MODEL APPROACHES** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Statistics complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner _____ Signature _____ Date _____

Examiner _____ Signature _____ Date _____

Advisor _____ Signature _____ Date _____

Chair of Department or Graduate Program Coordinator

Acknowledgments

I would like to express my special gratitude to my advisor Dr. Girma Taye for his important directions, preparing the available data, encouragement and support from the initial to final level during my study. My special thanks goes to Addis Ababa University, especially department of statistics for giving me a chance for completing my MSc program and financial support during my study.

I would like to express my sincere thank to Ato Yabebal Ayalew (Lecturer) in department of Statistics, AAU for helping me in developing R software program in the analyzing data for this study and giving me a lot of helps in my research work and my life while I stayed at Addis Ababa University.

Last but not the least my appreciation goes to my beloved parents and grandfather, for them endless support.

Table of Content

Acknowledgments.....	I
Acronyms and Symbols	IV
Abstract	VI
1. INTRODUCTION	1
1.1 Background of the Study	1
1.2 Statement of the problem	3
1.3 Objective of the Study	4
1.4 Significance of the Study	5
2. LITERATURE REVIEW	6
2.1 Basic Concepts and Definitions	6
2.1.1 Genetic markers	6
2.1.2 Basic principles of gene mapping	8
2.1.3 Generations and populations used for QTL mapping	9
2.1.4 Linkage Maps.....	10
2.2 History of the QTL mapping work	13
2.3 Principle of QTL analysis	15
2.4 Review of Literatures Specific to the Study	17
3. DATA AND METHODOLOGY	22
3.1 Sources of Data	22
3.2 Assumptions.....	23
3.3 Estimating Recombination between Two Genetic Markers.....	24
3.4 Conditional probabilities of putative QTL genotypes given markers	25
3.5 QTL Detection	29
3.5.1 Standard Interval Mapping (SIM).....	29
3.5.2 Linear Mixed Model Approach for Mapping QTLs	36
4. RESULTS AND DISCUSSION	49
4.1 Diagnostics for Interval Mapping	49
4.2 QTL detection and estimating their effects using interval mapping	50
4.3 Markers effect illustration.....	53
4.4 QTL localization and mapping using linear mixed model.....	54

4.5 Leveling of Variables as Random and Fixed	56
4.6 Mixed Model Fitting and Diagnostic Checking.....	57
4.6.1 Fixed effect estimates for QTL one (model one).....	57
4.6.2 Fixed effect estimates for QTL two (model two).....	58
4.6.3 Fixed effect estimates for QTL three (model three).....	59
4.6.4 Fixed effect estimates for QTL four (model four)	60
4.6.5 Fixed effect estimates for QTL five (model five)	61
4.6.6 Fixed effect estimates for QTL six (model six)	62
4.6.7 Fixed effect estimates for QTL seven (model seven)	63
4.6.8 Fixed effect tests for the detected QTLs	64
4.6.9 Gene action for the detected QTLs	64
4.7 Genetic variation explained by detected QTL	66
4.8 DISCUSSION	66
5. CONCLUSION AND RECOMMENDATION.....	69
5.1 CONCLUSION.....	69
5.2 RECOMMENDATION	70
REFERNCES	72
APPENDICES	78

Acronyms and Symbols

A	Adenine
AAU	Addis Ababa University
AFLP	Amplified Fragment Length Polymorphism
AIC	Akaike Information Criteria
ANOVA	Analysis of Variance
BC	Backcross
BLUP	Best Linear Unbiased Predictor
BLUE	Best Linear Unbiased Estimator
C	Cytosine
CAPs	Cleavage Amplification Polymorphisms
CGIAR	Consultative Group on International Agricultural Research
CIMMYT	International Maize and Wheat Improvement Center
CIM	Composite Interval Mapping
cM	center Morgan
DF	Degree of Freedom
DNA	Deoxyribonucleic Acid
Dn DF	Denominator Degree of Freedom
E	Environment
ECSA	Ethiopian Central Statistical Agency
EIAR	Ethiopian Institute of Agricultural Research
EM	Expectation and Maximization
ESA	Eastern Southern Africa
FAO	Food and Agriculture Organization
G	Guanine
GY	Grain Yield
LMM	Linear Mixed Model
LOD	Logarithm of Odd
LRT	Likelihood Ratio Test
MAS	Marker Assisted Selection

MLE	Maximum Likelihood Estimates
MSc	Master of Science
N	Normal distribution
Num DF	Numerator Degree of Freedom
OD	Overdominance
PD	Partial Dominance
PCR	Polymerase Chain Reactions
QTL	Quantitative Trait Loci (Locus)
RAPD	Random Amplified Polymorphic
RFLP	Restriction Fragment Length Polymorphism
REML	Restricted Maximum Likelihood
RIL	Recombinant Inbred Line
RNA	Ribonucleic Acid
SIM	Standard Interval Mapping
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
T	Thymine
t/ha	tons of Grain Yield per hectare

Abstract

*The term QTL mapping has been used in a number of scientific disciplines. As the fast advance in molecular genetics, it is much easy to get well-distributed genetic markers in almost every organism nowadays. Quantitative trait loci (QTL) mapping can provide useful information for breeding programs, since it allows the estimation of genomic locations and genetic effects of chromosomal regions related to the expression of quantitative traits. Therefore, as the major direction of quantitative genetics, various statistical methods have been developed to detect or map quantitative trait loci (QTL) by using the genetic marker information. In Ethiopia, till date, researchers have often been challenged by such type of data. Thus, this study is aimed to popularize the principles and methods for QTL mapping in this country. Interval mapping (IM) and mixed model based QTL mapping were included in this study. To realize the genetic basis of grain yield of maize (*Zea mays* L.), an F_2 intercross population with SNP markers covering 201.6cM, which was obtained from EIAR, were used to detect the QTLs for grain yield in maize. As a result, by analyzing the LOD profile in interval mapping, three QTLs associated with grain yield were identified on chromosomes 1, 3 and 4; and these QTLs explain 35.15% of phenotypic variance for grain yield in maize. Whereas, mixed model approach of QTL mapping method by considering localization and detection stages, seven QTLs were detected on chromosomes 1 (two), 2 (two), 3 (two) and 4 (one) and six of them have statistical significant effects on the maize grain yield. The six significant QTLs explained 64.7% of the phenotypic variance. The prevailing mode of gene action revealed overdominance effect in both statistical methods. On the basis of the findings, we conclude that mixed model approach can detect more number of QTLs than interval mapping. Finally, we recommend that statisticians/biometricians in this country need to collaborate with molecular geneticists to promote applications of statistical methods on QTL mapping, molecular data analysis and related areas using recent methods and neutral markers.*

Key words: Marker, QTL, QTL mapping, interval mapping, mixed linear model, maize grain yield.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Maize (*Zea mays* L.) was originated in Central America and introduced to West Africa in the early 16th century (FAO 1992) and to Ethiopia between the 16th and the 17th century. Maize (*Zea mays* L.) is a very important crop for feed, energy, and food production throughout the world. It is a monoecious species with separated male (tassel) and female (ear) organs. Global production exceeds 600 metric tonnes (McDonald and Nicol 2005), with about 60% produced in the developed countries, particularly by the United States of America, China produces 27% of the world's maize and the rest is grown in countries of Africa, Latin America, and southern Asia with a large proportion being produced in the tropics and subtropics.

Maize is one of the most important cereals cultivated in Ethiopia. It ranks second after teff in area coverage and first in total production. The results of the year 2011/12, Meher season post-harvest crop production survey indicate that total land areas of about 12,086,603.89 hectares were covered by grain crops. Out of the total grain crop areas, 79.34% (9,588,923.71 hectares) was under cereals. Of this maize covered 17% (about 2,054,723.69 hectares) and gave 6069413 tonnes of grain yields (ECSA, 2012).

Grain yield is one of the most important traits for maize breeding programs, but its evaluation and improvement are difficult and expensive to assess due to its complex biology, environmental interactions, and low heritability (Hallauer *et al.* 1988). These characteristics have made grain yield the primary trait of interest for QTL mapping studies for marker assisted selection (Tanksley 1993).

Among the various traits normally considered in breeding research, grain yield is generally the most important one (Sabadin *et al.* 2008). But at present, available resources are very limited and conventional breeding method is laborious and time consuming (Ribaut *et al.* 1997); this reduced the progress of yield breeding to some extent. To resolve this problem, an effective approach is

to utilize high yield genes in maize to improve the trait, but, this must depend on understanding of genetic basis.

In Ethiopia, maize grain yield is mainly used for human consumption as a variety of food products and also for local beverages. The use of maize for industrial processing is low but increasing. Maize is mainly processed at the household level with 76% used for home consumption, 9% for seed, about 10% marketed and the remaining 4% being used for different purposes (ECSA 2003).

The total variation i.e., phenotypic variation, present in a population arises due to genotypic and environmental effects. Phenotypic variability is the observable variation present in a character in a population: it includes both genotypic and environmental components of variation and as a result, its magnitude differs under different environmental conditions. Genotypic variation, on the other hand, is the component of variation, which is due to the genotypic differences among individuals within a population and is the main concern of plant breeders; because it is heritable and can therefore, in principle, be manipulated in a breeding programme.

The Highland Maize Germplasm Collection Mission was launched throughout the highlands of Ethiopia in 1998 in collaboration with CIMMYT (Twumasi-Afriyie *et al.* 2001). As part of this project, 287 maize accessions were collected from farmers' fields throughout the highland regions of Ethiopia. A field study of 180 of 287 maize accessions revealed that these accessions are highly variable for morphological and agronomic characteristics (Beyene *et al.* 2005) and thereby generated baseline data for future breeding and molecular studies. However, morphological variation does not always reflect real genetic variation because of genotype X environment interaction and the largely unknown genetic control of polygenic morphological and agronomic traits (Smith and Smith 1992).

Conventional breeding is time consuming and depends on environmental conditions. 'Marker-assisted selection' (also 'marker-assisted breeding' or 'marker-aided selection') may greatly increase the efficiency and effectiveness in plant breeding compared to conventional breeding methods. Once markers that are tightly linked to genes or QTLs of interest have been identified, prior to field evaluation of large numbers of plants, breeders may use specific DNA marker

alleles as a diagnostic tool to identify plants carrying the genes or QTLs (Michelmore 1995; Ribaut *et al.* 1997; Young 1996).

The advantages of MAS include: (i) Time saving from the substitution of complex field trials (that need to be conducted at particular times of year or at specific locations, or are technically complicated) with molecular tests; (ii) Elimination of unreliable phenotypic evaluation associated with field trials due to environmental effects; (iii) Selection of genotypes at seedling stage; (iv) Gene ‘pyramiding’ or combining multiple genes simultaneously; (v) avoid the transfer of undesirable or deleterious genes (‘linkage drag’; this is of particular relevance when the introgression of genes from wild species is involved); (vi) Selecting for traits with low heritability, and (vii) testing for specific traits where phenotypic evaluation is not feasible (Manikanda 2012).

The genetics of quantitative characters is studied in general terms using statistical techniques using population means, variances, co-variances of relatives and heritability etc. rather than individual gene effects. Thus the exact number, mode of action and location of the genes is difficult to be ascertained through Mendelian analyses. But the QTL mapping provides scope for studying this variation in terms of individual gene effects (Asins 2002). However, due to limitations of getting the available marker data in all organisms, the application of statistics for QTL mapping could not be familiarized in this country.

1.2 Statement of the problem

With the arable land gradually decreasing, it is likely that increased planting density will continue to be a target for selection in the future. Hence, breeding of maize with optimized plant architecture is regarded as one of the most important goals in maize breeding projects. To obtain plant varieties possessing the desired architecture, conventional breeding methods are useful; however, these methods are time-consuming due to the selection of superior individuals among segregating progenies resulting from hybridization; and it also depends on environmental conditions. An alternative solution is to utilize marker-assisted selection (MAS). Combined with traditional selection techniques, MAS has emerged as a valuable tool for selecting organisms with interesting traits during crop breeding.

To perform MAS, an understanding of the genetic basis of traits is necessary, and quantitative trait locus (QTL) mapping for traits can help us achieve this purpose.

So far, several studies have been conducted on the analyses of quantitative trait loci using morphological markers, molecular markers and pedigree information. Moreover, statistical methods for mapping quantitative trait loci (QTL) using single marker analysis, interval mapping and composite interval mapping are applied through different researchers for grain yield in maize. However, the putative QTL genotypic means and QTL positions are confounded in single marker analysis, and thus it causes biased estimation of QTL effects and low power in detection. Interval mapping on the other hand considers one QTL at a time in the model for QTL mapping and hence it is biased in estimation of QTL when multiple QTL are located in the same linkage group. Whereas using composite interval mapping, inclusion of too many cofactors reduced the power to identify QTL and QTL effects and marker effects are confounded.

Thus, in order to identify more than one QTL in linkage group and to separate QTL effects from their linked marker effects the recent and important method is linear mixed model QTL mapping approach. However, little is known about mixed model approach for mapping QTL using genetic markers. Moreover, statistical analysis of marker data is not well taken up in this country due to lack of data availability and limited resources, especially on molecular marker data, and bioinformatics expertise. To be honest for our knowledge no one has attempted to do on mixed model approach of QTL mapping for maize grain yield, very little has been done elsewhere for some traits. This is why we are committed to apply and popularized the recently emerging and interesting methods using Ethiopian maize data. Thus, this study considers the grain yield and genotype marker data in maize to identify the putative QTLs.

1.3 Objective of the Study

The general objective of this study is to identify the putative quantitative trait loci (QTL) using two statistical methods (interval mapping and mixed linear model approach) regarding maize grain yield and molecular marker data.

Moreover, this paper has the following specific objectives:

- To detect the regions of the genome that affects maize grain yield.

- To describe the effect of the QTL on the trait: the variation for the trait caused by a specific region, the gene action associated with the QTL- additive effect, dominant effect.
- To test the significance of the effect of QTL on maize grain yield.
- To understand the performances of the two QTL mapping techniques under different circumstances.

1.4 Significance of the Study

The QTLs information such as number, locations, and effects can help geneticist and breeders to improve the quality of plant breeding. However, the fundamental of the QTL mapping methods uses statistics principles. It is thus important to understand the statistical principles governing quantification of QTL with trait of interest and significance of the QTL itself before implementing QTL mapping.

Moreover, it is also useful by comparing different QTL mapping methods to understand the performances of the various methods under different circumstance.

The bioinformatics method in QTL mapping is relatively new area, even in the biological sciences, and it is important to popularize the method so that the methodologies are know better among the statisticians in the field and subject matter experts.

Furthermore, the availability of molecular markers linked to QTLs controlling variation for grain yield will allow for the implementation of marker-assisted selection to improve productivity.

In addition, the bioinformatics method is the integration of genetics, statistical methods and computing (mainly related to managing huge data set often created by genotyping) as well as plant/animal science. For this reason, this work will help initiate the need for team approach among geneticists, statisticians, and plant/animal scientists to genomic research. 0700724, 15/16

CHAPTER TWO

LITERATURE REVIEW

2.1 Basic Concepts and Definitions

2.1.1 Genetic markers

Markers contain information about segregation at various positions of a genome in a population. Genetic markers represent genetic differences between individual organisms or species. Generally, they do not represent the target genes themselves but act as ‘signs’ or ‘flags’. Those genetic markers located in close proximity to genes (i.e. tightly linked) may be referred to as gene ‘tags’. Such markers themselves do not affect the phenotype of the trait of interest because they are located only near or ‘linked’ to genes controlling the trait. All genetic markers occupy specific genomic positions within chromosomes (like genes) called ‘loci’ (singular ‘locus’).

There are three major types of genetic markers: (1) morphological (also ‘classical’ or ‘visible’) markers which themselves are phenotypic traits or characters; (2) biochemical markers, which include allelic variants of enzymes called isozymes; and (3) DNA (or molecular) markers, which reveal sites of variation in DNA (Jones *et al.* 1997; Winter and Kahl 1995).

Morphological markers are usually visual characterized phenotypic characters such as flower colour, seed shape, growth habits or pigmentation. Isozymes markers are differences in enzymes that are detected by electrophoresis and specific staining. The major disadvantages of morphological and biochemical markers are that they may be limited in number and are influenced by environmental factors or the developmental stage of the plant (Winter and Kahl 1995). However, despite these limitations, morphological and biochemical markers have been extremely useful to plant breeders (Eagles *et al.* 2001; Weeden *et al.* 1994).

Molecular markers

DNA markers are the most widely used type of marker predominantly due to their abundance. They arise from different classes of DNA mutations such as substitution mutations (point mutations), rearrangements (insertions or deletions) or errors in replication of tandem repeated

DNA (Paterson 1996a). These markers are selectively neutral because they are usually located in non coding regions of DNA. DNA is the genetic material of organisms and genetic differences between individuals will be reflected directly by the nucleotide sequences of DNA molecules.

Molecular markers offer numerous advantages over conventional morphological markers and isoenzymes. They are stable and detectable in all tissues regardless of growth, differentiation, development and status of the cell. Further, DNA markers are practically unlimited in number and are not affected or confounded by environmental factors, pleiotropic and Epistatic effects (Winter and Kahl 1995). Apart from the use of DNA markers in the construction of linkage maps, they have numerous applications in plant breeding such as assessing the level of genetic diversity within germplasm and cultivar identity (Baird 1997).

DNA markers are particularly useful if they reveal differences between individuals of the same or different species. These markers are called polymorphic markers, whereas markers that do not discriminate between genotypes are called monomorphic markers. Polymorphic markers may also be described as codominant or dominant. This description is based on whether markers can discriminate between homozygotes and heterozygotes. Codominant markers indicate differences in size whereas dominant markers are either present or absent. Strictly speaking, the different forms of a DNA marker (e.g. different sized bands on gels) are called marker 'alleles'. Codominant markers may have many different alleles whereas a dominant marker only has two alleles (Figure 2.1).

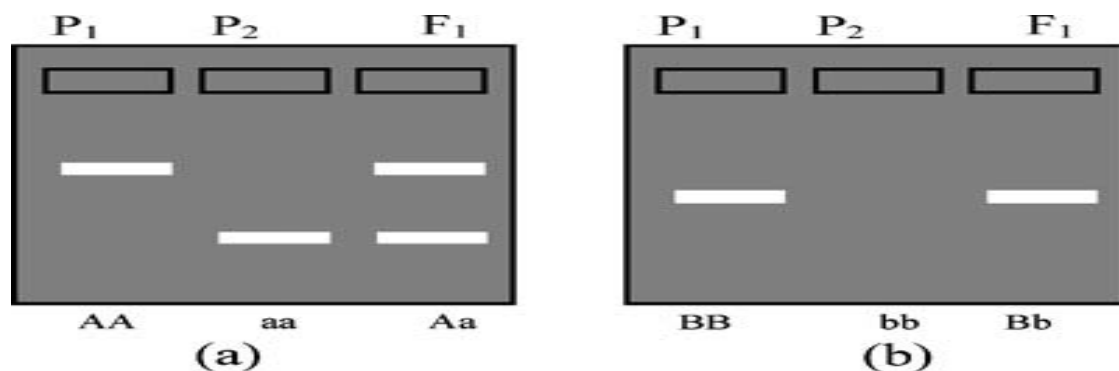


Figure 2.1: Comparison between (a) codominant and (b) dominant markers. Codominant markers can clearly discriminate between homozygotes and heterozygotes whereas dominant markers do not.

A wide variety of techniques can be used to measure DNA variation. Direct sequencing of DNA provides the ultimate measure of genetic variation, but much quicker scoring of variation is sufficient for most purposes. These methods include Restriction Fragment Length Polymorphisms (RFLPs), Polymerase Chain Reaction (PCR), Randomly Amplified Polymorphic DNAs (RAPDs), SNP and microsatellite DNAs.

Single Nucleotide Polymorphism (SNP)

Variations at single-nucleotide level in genome sequence of individuals of a population are known as SNPs (Jordan and Humphries 1994). SNP markers are DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species or between paired chromosomes in an individual. For example, at a certain position in a DNA sequence there may be a C (cytosine) present in some individuals but a T (thymine) present in others.

Seq 1 ATTCAATCCA

Seq 2 ATTTAATCCA

SNPs represent the most basic form of genetic polymorphism. There are tens of millions of SNPs present in the genome of a typical organism. However, usually only a very small subset of these will be developed into genetic markers.

SNPs are most abundant type of DNA polymorphism in animal and plant genome and are new source of markers useful in genetic mapping, QTL mapping, map-based positional cloning and assessment of genetic distances between individuals (Lopez *et al.* 2005). SNPs are useful in the analysis of quantitative trait loci (QTL), and in association studies, because they are conserved during evolution of plants and animals. SNPs have been used in detection of gene mutation using gene sequence information for primer construction (Kumar *et al.* 2009).

2.1.2 Basic principles of gene mapping

Mapping is based on the simple genetic principles, namely, linkage and recombination. Let there are two individuals, homozygous for two alleles of a two loci, A and B. The genotype of one individual is *AABB* and the other is *aabb*. Where *A* and *B*, *a* and *b* are alleles for individual one and two, respectively. They each produce only one type of male and female gamete, *AB* and *ab*.

Crossing between them result in a F_1 progeny of constitution $AaBb$. The F_1 can be selfed or intercrossed to produce F_2 generation. F_1 being heterozygous, throws out segregates in F_2 , based on the random combination obtained between male and female gametes produced either as AB, ab, Ab or aB . Of these four types of gametes, the first two resembling the gametes produced by the original homozygotic parents, are called parental types. The other two, must have been resulted due to a crossing over between the locus A and B in F_1 heterozygote. They are called recombinant types or recombinants.

Recombination is the process by which new combination of parental genes occur by exchanging the alleles of different loci on the chromosomal segments between homologous chromosomes carrying them. In simple terms, thus the recombination value can be calculated as,

$$\text{Recombination frequency (\%)} = \frac{\text{No. of recombinants} \times 100\%}{\text{Total no. of progeny}}.$$

One percentage of recombination is equivalent to one arbitrary map unit called as centiMorgans or cM. For example, if the recombination frequency between two loci A and B is 5% and the same between B and C is 23%, and that between A and C is 26%, using this values, we can order these loci along a chromosome as follows.

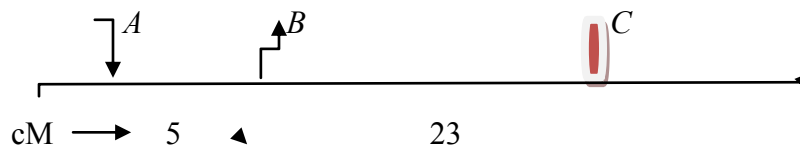
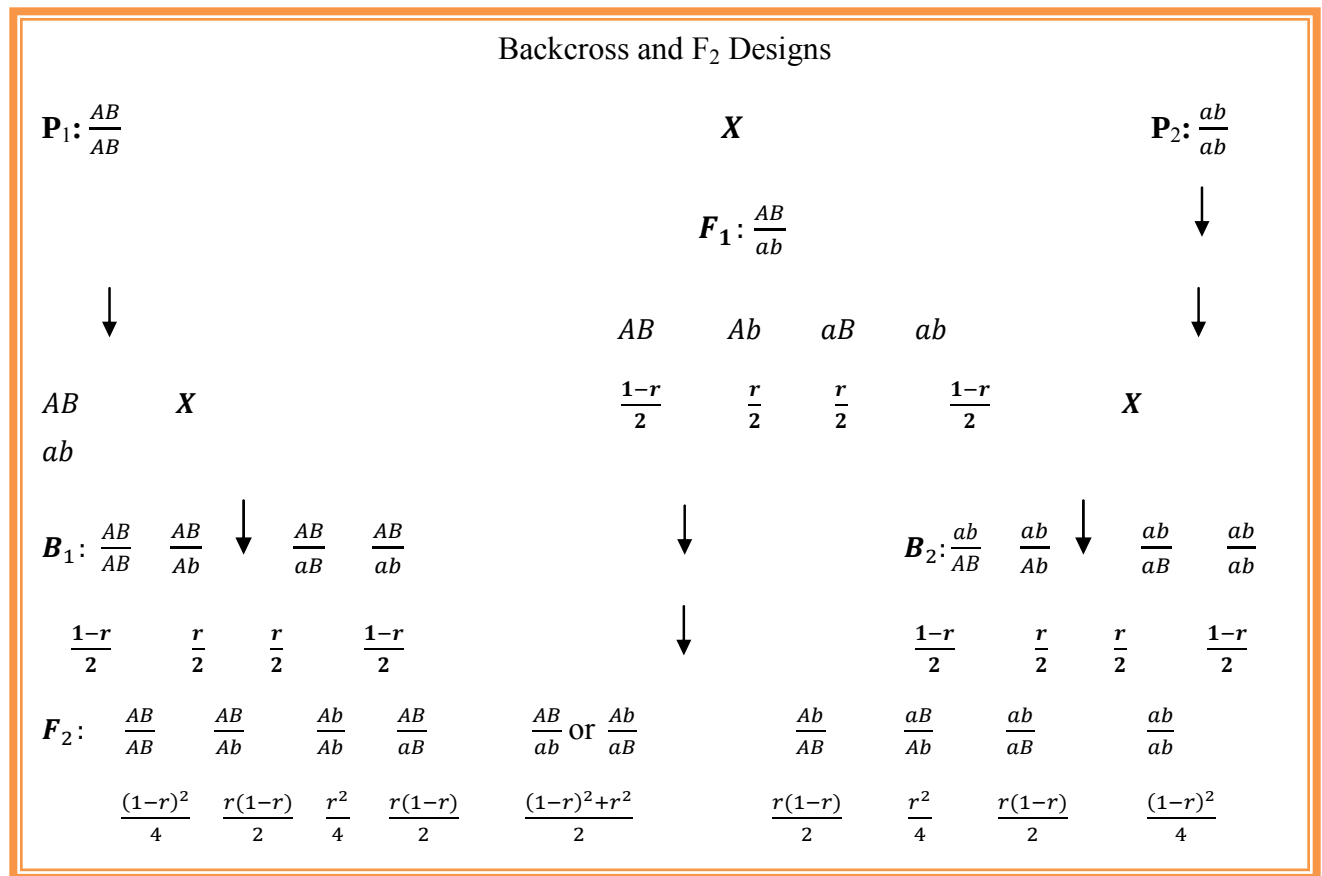


Figure 2.2: diagrammatic representation of markers A , B and C on the molecular marker linkage map.

2.1.3 Generations and populations used for QTL mapping

The procedures of QTL mapping are mainly focused on populations developed from crosses between two inbred lines. Traditional experimental designs for locating QTL start with two parental inbred lines, P_1 and P_2 ; differing both in trait values and in the marker variants they carry. In practice, markers are sought that have different variants in parents. There are two different types of single-cross populations: F_2 and F_2 -derived populations and recombinant inbred lines (RIL) populations (Zeng 1994). Other types of populations, such as backcross or

random mating populations can be used for QTL mapping, but the ability to detect QTL or the information contained in such populations is generally lower compared to F_2 or RIL populations (Zeng 1994). Random mating populations are more difficult for QTL mapping, because, the linkage disequilibrium is a key to detecting QTL with markers. Population generations for backcross, F_1 and F_2 intercross are summarized below.



2.1.4 Linkage Maps

A genotype or the genetic makeup of an organism is defined (Falconer and Mackay 1996) as the combination of alleles at a single autosomal locus in a diploid organism. The physical or visible characteristics resulting from the interaction between the genetic makeup and the environments are referred to as phenotype. Phenotype can be observed, measured, classified or counted. Organisms are determined neither by their genes nor by their environment; they are the consequence of the interaction of genes and environment (Suzuki *et al.* 1981).

Genotype describes the complete set of genes inherited by an individual that is important for the expression of a trait under investigation. Phenotype describes all aspects of the individual's morphology, physiology and ecological relationships. The genotype is essentially a fixed character of the organism; it remains constant throughout life and is unchanged by environmental effects. The phenotype changes continually and the direction of that change is a function of the sequence of environments that the individual experiences (Suzuki *et al.* 1981).

A linkage map may be thought of as a 'road map' of the chromosomes derived from two different parents (Paterson 1996a). Linkage maps indicate the position and relative genetic distances between markers along chromosomes, which is analogous to signs or landmarks along a highway. The most important use for linkage maps is to identify chromosomal locations containing genes and QTLs associated with traits of interest; such maps may then be referred to as 'QTL' (or 'genetic') maps. Thus, a QTL (Quantitative Trait Locus) is a region of the genome that is associated with an effect on a quantitative trait of interest. Conceptually, a QTL can be a single gene, or it may be a cluster of linked genes that affect the trait. The procedures for finding and locating the QTLs are called QTL mapping (Liu, B. 1998).

Many agriculturally important traits such as yield, quality and some forms of disease resistance are controlled by many genes and are known as quantitative traits (also 'polygenic,' 'multifactorial' or 'complex' traits). A gene is a discrete segment of the DNA molecule, located at a specific site (its locus) on a specific chromosome pair. It is the basic physical unit of heredity. Chromosome is the self-replicating threadlike structure found in cells. Chromosomes, which at certain stages of meiosis and mitosis consist of two identical sister chromatids, joined at the centromere, and carry the genetic information encoded in the DNA sequence.

In a segregating population, there is a mixture of parental and recombinant genotypes. The frequency of recombinant genotypes can be used to calculate recombination fractions, which may be used to infer the genetic distance between markers. By analyzing the segregation of markers, the relative order and distances between markers can be determined. The lower the frequency of recombination between two markers, the closer they are situated on a chromosome (conversely, the higher the frequency of recombination between two markers, the further away they are situated on a chromosome).

Markers that have a recombination frequency of 50% are described as ‘unlinked’ and assumed to be located far apart on the same chromosome or on different chromosomes. A mapping function gives the relationship between the distance between two chromosomal locations on the genetic map (in cM) and their recombination frequency.

Linkage maps are constructed from the analysis of many segregating markers. The three main steps of linkage map construction are: (1) production of a mapping population; (2) identification of polymorphism and (3) linkage analysis of markers.

Mapping populations

The construction of a linkage map requires a segregating plant population (i.e. a population derived from sexual reproduction). The parents selected for the mapping population will differ for one or more traits of interest. If the map will be used for QTL studies (which is usually the case), then an important point to note is that the mapping population must be phenotypically evaluated (i.e. trait data must be collected) before subsequent QTL mapping (Collard *et al* 2005).

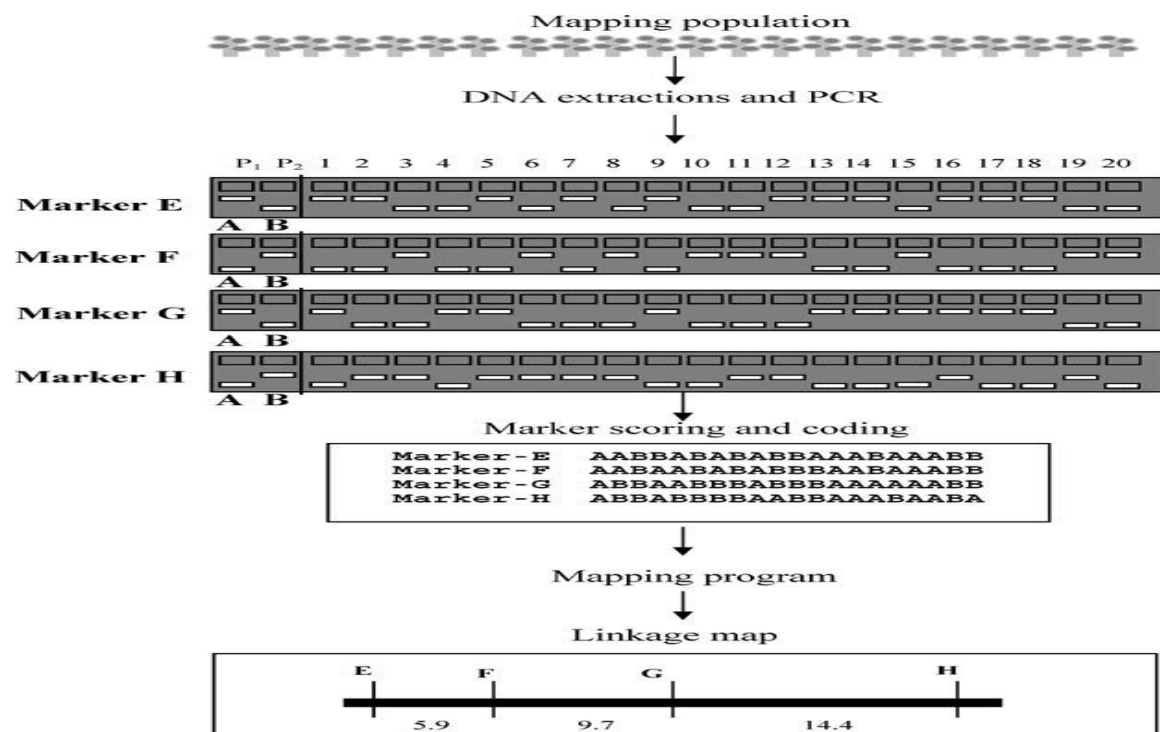


Figure 2.3: Construction of a linkage map based on a small recombinant inbred population (20 individuals). The first parent (P₁) is scored as an ‘A’ whereas the second parent (P₂) is scored as a ‘B’. Coding of marker data varies depending on the type of population used.

Identification of Polymorphism

The second step in the construction of a linkage map is to identify DNA markers that reveal differences between parents (i.e. polymorphic markers). It is critical that sufficient polymorphism exists between parents in order to construct a linkage map (Young 1994). The choice of DNA markers used for mapping may depend on the availability of characterized markers or the appropriateness of particular markers for a particular species (Collard *et al* 2005).

Linkage Analysis of Markers

The final step of the construction of a linkage map involves coding data for each DNA marker on each individual of a population and conducting linkage analysis using computer programs (Collard *et al* 2005).

2.2 History of the QTL mapping work

It is believed that the rediscovery of Mendelian genetics in 1900 was beginning of the modern genetics. Through the demonstration on the inheritance of discrete characters, such as purple vs. white flower, smooth vs. wrinkled seeds, it is clear that the traits are controlled by genetics factors or genes, which will, inherited from generation to generation. Later on, great efforts have been made on understanding how the genes effecting the discrete characters or qualitative traits, especially the nature of the genes to transmit from the parents to their offspring.

However, most economically as well as biologically important traits are not qualitative, but quantitative in nature. Here the quantitative means that the trait's value cannot be divided into several categories and the distribution of these values is continuously over a range in a population. The examples of the quantitative trait are of crop yield, plant height, resistance to diseases, and weight gain in mice and egg or milk production in animals. Due to the complexity nature of the quantitative inheritance, the progress of quantitative genetics is far behind the Mendelian genetics.

The questions how the genes contribute to the quantitative trait values and why the trait values are continuously distributed may be answered partially by polygene theory (East 1916). In this theory, a quantitative trait is controlled by many genes with small effects, and at the same time is also influenced easily by environment effects. However, it is very difficult to dissect the

individual genes that controlling a quantitative trait by classical quantitative genetic means. Therefore, Breeders usually have no idea about the number, location and effect of the individual genes involved in the inheritance of target quantitative traits (Comstock 1978). These genes are also called quantitative trait loci (QTLs). It is impossible to manipulate the QTLs using genetic engineering method and through that to improve the organism's traits without obtaining the QTLs information, such as number, locations, and effects.

The history of QTLs mapping can be traced back to 1920's. Sax (1923) used the morphological markers to demonstrate an association between seed weight and seed coat colour in beans. Thoday (1961) used multiple genetic markers to systematically map the individual polygenes, which control a quantitative trait. He notices: "The main practical limitation of the technique seems to be the availability of suitable markers". It is obvious that the numbers of the morphological or protein markers are very limited. Therefore, genetic markers are the nature choice for detecting or mapping QTLs.

Nowadays, it is much easy to get well-distributed genetic markers in almost every organism, because the fast advance of molecular genetics technology. Varying statistical methods have been developed to detect or map QTLs by using genetic markers information. Lander and Botstein (1989) proposed the interval mapping method (IM), which use two adjacent markers to bracket a region for testing the existence of a QTL by performing a likelihood ratio test at every position in the region. The method has been proven more powerful and requiring fewer progeny than one-marker methods. However, interval-mapping method has some drawback. Because it is a one QTL model, the mapping position of QTLs will be seriously biased when more than one QTL located at same chromosome (Knott and Haley 1992).

Later on, several attempts have been made to solve this problem. Zeng (1993) proved an important property of multiple regression analysis in relation to QTL mapping: "If there is no epistasis, the partial regression coefficient of a trait on a marker depends only on those QTLs that are in the interval bracketed by the two neighboring markers and is independent of QTLs located in other intervals". Zeng (1994) proposed an improved method called composite interval mapping (CIM) by combining interval mapping with multiple regression analysis. Jansen (1993) has also proposed a similar strategy. Composite interval mapping has proved having a better

performance than interval mapping in multiple linked QTLs case. Recently an extended method called multiple intervals mapping (MIM) has been proposed (Kao, Zeng and Teasdale 1999). This method fits all QTLs into the model altogether and has the ability for analyzing QTL epistasis and the associated statistical issues.

A new methodology was also proposed (Zhu 1998, 1999; Zhu and Weir 1998) for systematically mapping QTLs based on the mixed linear model approaches (MCIM). The MCIM method has very similar performance with Zeng's CIM method. However, MCIM method does not have the problem of selecting the background control markers and setting the mapping windows size as CIM method does. MICM method also has the advantage that is very easy to extend for more complicated QTLs mapping situations such as QTL epistasis and QTL by environmental interaction etc.

2.3 Principle of QTL analysis

Identifying a gene or QTL within a plant genome is like finding the proverbial needle in a haystack (Collard *et al* 2005). However, QTL analysis can be used to divide the haystack in manageable piles and systematically search them. QTL are identified via statistical procedures that integrate genotypic and phenotypic data. In simple terms, QTL analysis is based on the principle of detecting an association between phenotype and the genotype of markers. Markers are used to partition the mapping population into different genotypic groups based on the presence or absence of a particular marker locus and to determine whether significant differences exist between groups with respect to the trait being measured (Tanksley 1993; Young 1996). A significant difference between phenotypic means of the groups (either 2 or 3), depending on the marker system and type of population, indicates that the marker locus being used to partition the mapping population is linked to a QTL controlling the trait.

The closer a marker is from a QTL, the lower the chance of recombination occurring between marker and QTL. Therefore, the QTL and marker will be usually being inherited together in the progeny, and the mean of the group with the tightly linked marker will be significantly different to the mean of the group without the marker. When a marker is loosely linked or unlinked to a QTL, there is independent segregation of the marker and QTL. In this situation, there will be no significant difference between means of the genotype groups based on the presence or absence of

the loosely linked marker. Unlinked markers located far apart or on different chromosomes to the QTL are randomly inherited with the QTL; therefore, no significant differences between means of the genotype groups will be detected.

Methods to Detect QTLs

Three widely-used methods for detecting QTLs are single-marker analysis, simple interval mapping and composite interval mapping (Liu 1998; Tanksley 1993). Single-marker analysis (also single point analyses) is the simplest method for detecting QTLs associated with single markers. The statistical methods used for single marker analysis include t-tests, analysis of variance (ANOVA) and linear regression. Linear regression is most commonly used because the coefficient of determination (R^2) from the marker explains the phenotypic variation arising from the QTL linked to the marker. This method does not require a complete linkage map and can be performed with basic statistical software programs. However, the major disadvantage with this method is that the further a QTL is from a marker, the less likely it will be detected. This is because recombination may occur between the marker and the QTL. This causes the magnitude of the effect of a QTL to be underestimated (Tanksley 1993).

The standard interval mapping (SIM) method makes use of linkage maps and analyses intervals between adjacent pairs of linked markers along chromosomes simultaneously, instead of analyzing single markers (Lander and Botstein 1989). The use of linked markers for analysis compensates for recombination between the markers and the QTL, and is considered statistically more powerful compared to single point analysis (Lander and Botstein 1989; Liu 1998).

More recently, composite interval mapping (CIM) has become popular for mapping QTLs. This method combines interval mapping with linear regression and includes additional genetic markers in the statistical model in addition to an adjacent pair of linked markers for interval mapping (Jansen 1993). The method of composite interval mapping that Zeng (1993, 1994) developed includes marker information for controlling background noise while searching for the QTL. The marker effects, as well as the QTL effects, in this model are treated as fixed effects. Therefore, the estimated QTL effects could be affected by the markers included in the model.

A new method was proposed (Zhu and Weir 1998) that used a mixed model approach for composite interval mapping of QTLs. In their mixed model approach, QTLs are fixed variables

while molecular markers as random variables. Thus, the estimates of QTLs do not depend upon a particular fixed set of markers being in the model. This model also provides important estimates of additive and dominance effects of QTLs.

Additive and Dominance allelic effect

The additive allele effect is an estimate of the change in the average phenotype that would be produced by substituting a single allele of one type with that of another type (e.g., a replaced by A) in a population. In a standard F₂ intercross between two inbred parental lines there are two alleles at every polymorphic locus that are often referred to as the recessive " a " allele and dominant " A " allele. F₂ progeny inherit the a/a , a/A , or A/A genotypes at every genetic locus in a ratio close to 1:2:1. The additive effect is half of the difference between the mean of all cases that are homozygous for one parental allele (aa) compared to the mean of all cases that are homozygous for the other parental allele (AA).

The term dominance indicates that the phenotype of intercross progeny closely resemble one of the two parental lines, rather than having an intermediate phenotype. Geneticists commonly refer to an allele as having a dominance effect or dominance deviation on a phenotype. Dominance deviation at a particular marker are calculated as the difference between the average phenotype of all cases that have the Aa genotype at that marker and the expected value half way between the all cases that have the aa genotype and the AA genotype.

2.4 Review of Literatures Specific to the Study

The development of high density linkage maps based on molecular markers in maize and other major crops have allowed the investigations on the genetic basis of quantitative traits, commonly referred to as quantitative trait loci (QTL). Coupled with statistical procedures, along with the availability of computer software (Lander and Botstein 1989; Zeng 1994; Basten *et al.* 1994), these maps provide information on the location and genetic effects of genomic regions affecting economically important quantitative traits. In maize, studies have reported several chromosome regions for QTLs, such as: resistance to biotic and abiotic stresses, morphological characters and grain yield (Stuber *et al.* 1992). DNA markers are widely accepted as potentially valuable tools for crop improvement in maize (Stuber *et al.* 1987).

Stuber *et al* (1992) have mapped QTLs associated with seven major traits, including grain yield, in cross between two widely used elite maize inbred lines B73 and Mo17. The analysis has been carried out to explore heterosis or hybrid vigor and genotype by environment interactions, which are two important phenomena in maize genetics. Seventy six markers, representing 90-95% of the maize genome have been analyzed using single marker analysis and interval mapping methods. Both methods in their study have shown virtually the same results in detecting QTLs affecting grain yield throughout the genome, except on chromosome 6 (Stuber *et al*, 1992). The prevailing mode of action was overdominance. They also conclude that the heterozygote had a higher phenotype than the respective homozygote (with only one exception) suggesting not only overdominance but also that these detected QTLs play a significant role in heterosis.

Sixteen QTL were mapped for grain yield on seven chromosomes Milena *et al.* (2005); QTL were not mapped on chromosomes 8, 9, and 10. The phenotypic variance explained by each QTL in their study was ranged from 1.02 to 4.66%, and all QTL explained 36.28% of the phenotypic variance. Half of the QTL mapped displayed positive additive effects, and half negative additive effects; thus, both parental inbreds had eight QTL with alleles that contributed to increase the trait. Twelve QTL (75.0%) displayed positive dominance effects, and they conclude that there is a tendency for positive unidirectional dominance effects of the QTL and could explain the high inbreeding depression for this trait. In addition, the level of dominance ranged from additive ($LD = 0.01$) to overdominance ($LD = 6.50$), and overdominance ($ALD = 1.26$) was the average level of dominance. Although overdominance was the average level of dominance, only five (31.25%) of the 16 QTL displayed overdominance, and 11 (68.75%) QTL displayed additive, partial dominance, or dominance gene action. Hence, as the ALD value is near the lower bound of the overdominance scale level (1.21) and there is an inherent error associated with it, this result only suggests that overdominance could be the average level of dominance for grain yield.

Overdominance in QTL mapping studies using interval mapping has been reported in both tropical and temperate maize germplasm (Stuber *et al.* 1992; Lu *et al.* 2003), although Sibov *et al.* (2003b) reported that none of the QTL mapped displayed overdominance, and partial dominance was the average level of dominance in the tropical maize population evaluated.

In other study also worked (Austin *et al.* 1996) for grain yield and yield components in $F_{2:3}$ and $F_{6:7}$ recombinant inbred (RI) lines from the same population was conducted to compare maize quantitative trait loci (QTL) detection. Single factor analysis of variance was conducted for each locus trait combination to identify QTL. For grain yield, six QTLs were detected accounting for 22% of the phenotypic variation, and a single major QTL was detected in the $F_{2:3}$ explaining 35% of the phenotypic variation, while all 6 $F_{6:7}$ QTL explaining only 21.8% of the phenotypic variation.

Nagabhushana K. *et al.* (2006) studied on QTL mapping for grain yield and plant height in rice using SIM and CIM. A comparison of QTLs detected by two different models simple interval mapping (SIM) using MAPMAKER/QTL and composite interval mapping (CIM) using QTL Cartographer was made, and common QTLs across the models were identified. They found an economically important trait associated with two significant QTLs for grain yield in their SIM method and three QTLs were detected using CIM method which explained 18.5 % and 26% variation for grain yield, respectively.

In the study of QTL mapping on maize grain yield containing 117 microsatellite loci and 400 F_2 maize plants in Brazil (Sibov *et al.* 2003a, 2003b) detected four QTLs using composite interval mapping. Sibov and his colleagues showed that the detected QTLs explained 32.73% of the phenotypic variance of grain yield; and dominance effects reached significance in three QTLs, whereas only 1 QTL displayed additive gene action.

In the other study, Zhang *et al.* 2010 conducted QTL mapping to realize the genetic basis of grain yield of maize (*Zea mays* L.), a recombinant inbred line population were used to detect the QTLs for grain yield in maize, and as a result, a total of six QTLs associated with grain yield was identified on four chromosomes. On their study, the six QTLs explained about 39.03% of the phenotypic variance; and conclude that the three QTLs could increase grain yield due to positive additive effects and three QTLs could decrease GY due to negative additive effects.

A number of models have been conducted to describe the application and importance of QTL mapping on grain yield and other related quantitative traits. At present, many QTLs associated with grain yield have been identified (Nagaraja *et al.* 2013; Sabadin *et al.* 2008), but different parental lines, segregation population or genetic map can bring about different results in QTL

number, position or effect, variation explained by QTL, and thus, it is necessary and significant that different parents be selected to detect the QTLs for grain yield (Christov *et al.* 2004; Tomlekova *et al.* 2010).

Nagaraja *et al.* (2013) analyzed QTL mapping for grain yield and other agronomic traits in sorghum by performing interval mapping. Thus, six QTL was found for grain yield, and out of the six QTL five QTL were identified for average performance over the three seasons and three of them were detected in multi environment QTL analysis. These QTL was individually explained 4.0-11.4% of phenotypic variance and together accounted 39% of the grain yield variation in the population.

In rice, Xiao, J. *et al.* (1995) have analyzed QTLs for yield related traits by single point analysis and interval mapping using RFLP markers. A total of 37 significant QTLs ($LOD \geq 2$) have been detected with 141 RFLP markers on backcrossing a set of F_7 lines from a subspecific cross showing strong F_1 heterosis to the two parents. Only two significant QTLs have been detected for grain yield in their study and concluded that dominance is the major causal factor of heterosis. No overdominance and epistasis was detected. These results are in disagreement with a series of studies on heterosis in rice (Yu *et al.* (1997); Luo *et al.* (2001), and. It could be because of the type and number of segregation population, the type and number of markers and chromosomes included in the study (Christov *et al.* 2004).

Although no one has attempted to work on mixed model for QTL mapping for grain yield, Shappley *et al.* (1998) worked on the mixed model approach of QTL mapping for agronomic and fiber traits in upland cotton (*Gossypium hirsutum* L.). They used the mixed model approach of Zhu and Weir (1998) to analyze for QTLs associated with 19 agronomic and fiber traits across 96 F_2 derived families from the cross of two cotton lines by applying molecular markers as random and QTLs as fixed variables. They determined the locations of 100 QTLs which mapped to 60 different maximum likelihood locations in 24 linkage groups. They also estimated the additive and dominance effects of the QTLs with some QTLs have both additive and dominance effects for the given traits. Thus, they conclude that negative genetic effect values decreases the phenotypic value of the trait, and positive values indicate an increase in the phenotype.

Currently, QTL mapping using molecular marker data is newly developing in this country by collaborating with other international research centers. Moreover, sequencing or genotyping marker for the given traits is quite expensive and currently being done elsewhere, mainly at Cornell University. Furthermore, statistical analysis of marker data is not well taken up in this country due to lack of data availability and limited resources. Hence information to be reviewed is not widely available; especially literature for mixed model QTL mapping.

CHAPTER THREE

DATA AND METHODOLOGY

3.1 Sources of Data

Data for maize grain yield and markers was obtained from Ethiopian Institute of Agricultural Research (EIAR) in Ethiopia. The data has been obtained from projects being implementing in collaboration with other research institutions. The maize markers are genotyping results from varieties grown in Ethiopia, and evaluated for maize yield (tones of grain yield per hectare, t/ha).

The study is based on grain yield data of 96 elite maize hybrids collected in Wolaita and Gamo Gofa zones of Southern Ethiopia. Between 2001 and 2009, 513 trials were conducted by national agricultural research programs and private seed companies in collaboration with CGIAR centers to identify elite hybrids for the region. CGIAR is an international organization which funds and coordinates research into agricultural crop breeding with the goal of "reducing rural poverty, increasing food security, improving human health and nutrition, and ensuring more sustainable management of natural resources.

A total of 96 tropical maize populations which is an F_2 intercross family obtained from a population derived from hybrid 5003xp118 were used in this study. The F_1 and F_2 population were derived from the cross between the two parental lines 5003 and p118. A total of 48 pairs of SNP markers were selected based on data eligibility.

Among 96 elite maize hybrids 20 of them have missing values (some of the markers are not detected) and hence, this study only includes 76 individuals with grain yield and 48 SNP markers and covering 201.5 cM of mapping distance on four chromosomes. Thus, grain yield in maize is taken as response variable while detected QTLs could be used as fixed and flanking markers as random variables in the given linear mixed model.

Experimental population

The maize populations are derived from a cross between two parental inbred lines P_1 and P_2 , differing substantially in a quantitative trait of interest. Here we call the two different parental

inbred lines (P_1 and P_2). A cross between two parents P_1 and P_2 is performed to produce an F_1 population. The F_1 progeny are all heterozygotes with the same genotype AB/ab . If the F_1 individuals are selfed or intermated, it produces an F_2 population. The F_2 individuals receive two sets of chromosomes from the F_1 generation, each of which will be a combination of parental chromosomes. Thus, at each locus, the F_2 individuals will have the genotypes of AA , AB or BB . F_2 design has three genotypes at a locus, and then can estimate both additive and dominance effects. There are nine observable marker genotypes in an F_2 population. The F_2 population provides the most of genetic information among different types of mapping populations (Lander *et al.* 1987), and is relatively easy to be obtained.

Let loci M_{i-} , with alleles A and a , and M_{i+} , with alleles B and b , denote two flanking markers for an interval where a putative QTL is being tested. Let the unobserved QTL locus Q with alleles Q and q be located in the interval flanked by markers M_{i-} and M_{i+} .

The distribution of unobserved QTL genotypes can be inferred from the observed flanking marker genotypes according to the recombination frequencies between them. The conditional probabilities of the QTL genotypes given marker genotypes are given in Table (3.1) for the F_2 population.

3.2 Assumptions

Throughout this thesis, two unlinked quantitative trait loci (QTL) on different chromosomes, QTL 1 and QTL 2, are assumed. Let A and B denote codominant marker loci linked to QTL 1 and QTL 2, respectively, and let r_1 , and r_2 denote the recombination fraction between marker A and QTL 1, and between marker B and QTL 2, respectively. Then, the marker-QTL orders are $A-r_1$ -QTL1 and $B-r_2$ -QTL2.

The first thing to do is to estimate, for each pair of markers, the recombination fraction between them, r , and calculate a LOD score for the test of $r = 1/2$. Markers on different chromosomes should not appear linked, and for markers on the same chromosome, the estimated recombination fraction should be smaller for more closely linked markers.

3.3 Estimating Recombination between Two Genetic Markers

Recombination or crossover occurred during prophase I stage of meiosis is the reason for individuals with same parents may have different marker genotypes. That is during the production of gametes, an exchange of material between pairs of chromosomes may occur. People can detect and record the variation or recombinants by using laboratory techniques as marker genotype for each individual. There are several facts about the marker genotype:

- The closer of the two markers, the less likely a recombination event is to occur.
- Markers that reside on different chromosome are unlinked.
- Two markers that never experience a recombinant event between them are called completely linked. They travel together during the meiosis process.
- If an even number of crossing over events occurs between two genetic markers, this event is undetectable.

For more than two markers in a chromosome, the simplifying assumption is that recombination between any two of them is independent from others' recombination events. This assumption is called no interference and the phenomenon of a single crossing over between DNA strands can be considered as a Poisson-process. Therefore, Haldane's mapping function (Haldane 1919) can be used for describing the relationship between recombination fraction r and genetic distance θ .

The number of (odd) crossovers (k) in an interval defined by two genetic markers has a Poisson distribution with mean θ .

$$\Pr(\text{recombination}) = \sum_k \frac{\theta^k e^{-\theta}}{k!} = e^{-\theta} \left\{ \frac{\theta}{1!} + \frac{\theta^3}{3!} + \dots \right\} = \frac{e^{-\theta}(e^{\theta} - e^{-\theta})}{2} = \frac{1}{2}(1 - e^{-2\theta})$$

We call this probability r with limits $0 \leq r \leq 0.5$, and θ is the average number of map units (cM) between two markers.

Solving the above equation for θ gives Haldane's map function:

$$\theta = -\frac{1}{2} \ln (1 - 2r) \quad (3.1)$$

When $r \rightarrow 0$, $\theta = 0$ we call it completely linked, on the other hand, $r \rightarrow 0.5$, $\theta = \infty$ (markers are unlinked). The reason why r is set to max of 0.5 is because only half arm of the genes are involved in process.

3.4 Conditional probabilities of putative QTL genotypes given markers

In the study of QTL mapping, the locations and genotypes of QTL are unknown. We can investigate any loci to see whether or not they are QTL. Such loci are referred to as putative QTL. In QTL mapping experiments, only markers are genotyped and the genotypes of putative QTL are generally unknown. However, with experimental crosses, the conditional probabilities of the genotypes of the putative QTL given markers can be obtained. These conditional probabilities play a crucial role in mapping methodologies.

Let Q denote the putative QTL, M_{i-} and M_{i+} denote the left and right flanking markers. Denote the alleles of Q , M_{i-} and M_{i+} by Q, q, A, a and B, b , respectively. The genotypes are indexed as follows. A homozygous genotype with two capital letter alleles is indexed by 2, a heterozygous genotype is indexed by 1, and a homozygous genotype with two small letter alleles is indexed by 0. Let p_{ij} , $i, j = 0, 1$, denote the joint probability that there is i recombination between M_{i-} and Q and j recombination between Q and M_{i+} on a single chromosome.

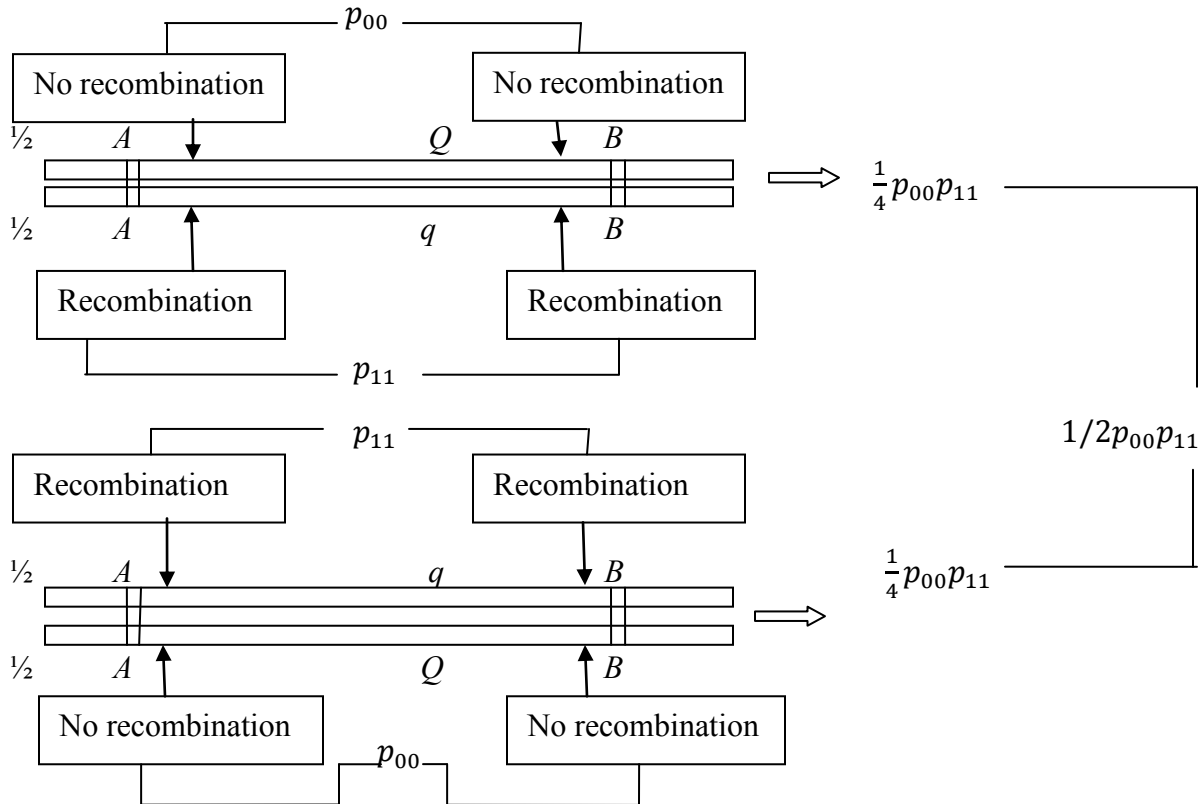


Figure (3.1): Illustration of the derivation of joint genotype probabilities of a putative QTL and its two flanking markers.

As an illustration, the entries of the first row of Table (3.1) are derived as follows. For convenience, we use the indices introduced at the beginning of this section to denote the joint genotypes and $M_{i-} = A, M_{i+} = B$. Thus, 222, 212 and 202 stand, respectively, for the joint genotypes $AAQQBB$, $AAQqBB$ and $AAqqBB$. First, we derive the joint probabilities of 222, 212 and 202. For 222, the probability of having the homozygous genotype of A is $1/4$. On both chromosomes, there is no recombination both between A and Q and between Q and B . The probability for this to happen is p_{00}^2 since the two chromosomes are independent. Thus the joint probability of 222 is $\frac{1}{4}p_{00}^2$.

For 212, there are two possible phases because of the heterozygous putative QTL. The probabilities of the two phases are the same. For each phase, the probability is derived in the same way as for 222. First, the probability for A to be homozygous is $1/4$. On one of the chromosomes, there is no recombination both between A and Q and Q and B , its probability is p_{00} . On the other chromosome, there is recombinations both between A and Q and between Q and B . Its probability is p_{11} . Therefore, the joint probability of 212 is $\frac{1}{2}p_{00}p_{11}$. The joint probability of 202, which is similarly derived, is $\frac{1}{4}p_{11}^2$. The derivation of the joint probability of 212 is demonstrated in Figure (3.1). Dividing these joint probabilities by the joint probability of the two flanking markers given in the second column of Table 3.1 yields the conditional probabilities.

We now relate the joint probabilities p_{11} , p_{10} , p_{01} and p_{00} to recombination fractions. Let r_1, r_2 and r denote, respectively, the recombination fractions between A and Q , between Q and B and between A and B . Note that a recombination fraction is the marginal probability of recombination event occurring between two loci. Let E_1, E_2 and E_3 denote, respectively, the events that there is recombination between A and Q , between Q and B and between A and B . In fact, $r_1 = P(E_1)$, $r_2 = P(E_2)$ and $r = P(E_3)$. Let E_j^c denote the complement of E_j , $j = 1, 2, 3$.

Note that

$$p_{11} = P(E_1 \cap E_2), p_{00} = P(E_1^c \cap E_2^c),$$

$$p_{10} = P(E_1 \cap E_2^c), p_{01} = P(E_1^c \cap E_2),$$

$$p_{10} + p_{01} = P(E_3), p_{11} + p_{00} = P(E_3^c).$$

Furthermore,

$$\begin{aligned} p_{00} &= P(E_1^c \cap E_2^c) \\ &= 1 - P(E_1 \cup E_2) \\ &= 1 - P(E_1) - P(E_2) + P(E_1 \cap E_2), \\ &= 1 - P(E_1) - P(E_2) + p_{11}. \end{aligned}$$

$$\begin{aligned} p_{10} &= P(E_1 \cap E_2^c) \\ &= 1 - P(E_1^c \cup E_2) \\ &= 1 - P(E_1^c) - P(E_2) + P(E_1^c \cap E_2), \\ &= P(E_1) - P(E_2) + p_{01}. \end{aligned}$$

Therefore,

$$\begin{aligned} p_{11} &= 1 - P(E_3) - p_{00} \\ &= 1 - P(E_3) - [1 - P(E_1) - P(E_2) + p_{11}] \\ &= P(E_1) + P(E_2) - P(E_3) - p_{11}, \\ p_{01} &= P(E_3) - p_{10} \\ &= P(E_3) - P(E_1) + P(E_2) - p_{01}. \end{aligned}$$

Thus, it is obtained that

$$p_{11} = \frac{1}{2} [P(E_1) + P(E_2) - P(E_3)] = \frac{1}{2} [r_1 + r_2 - r] = r_1 r_2,$$

$$p_{01} = \frac{1}{2} [P(E_2) - P(E_1) + P(E_3)] = \frac{1}{2} [r_2 - r_1 + r] = r_1(1 - r_2),$$

$$p_{10} = \frac{1}{2} [P(E_1) - P(E_2) + P(E_3)] = \frac{1}{2} [r_1 - r_2 + r] = (1 - r_1)r_2,$$

$$p_{00} = 1 - \frac{1}{2} [P(E_1) + P(E_2) + P(E_3)] = 1 - \frac{1}{2} [r + r_1 + r_2]$$

$$= (1 - r_1)(1 - r_2).$$

In the data of QTL mapping experiments, the genetic linkage map is given, that is, the genetic distance between any two markers is known. Thus, the recombination fractions r , r_1 and r_2 are obtained from the genetic distances by Haldane mapping function given in (3.1).

Table (3.1): conditional probabilities of putative QTL genotypes given two flanking markers in intercross population with no crossover interference.

Marker genotype	Genotype frequency	QTL genotype		
		2	1	0
22	$\frac{(1-r)^2}{4}$	$\frac{(1-r_1)^2(1-r_2)^2}{(1-r)^2}$	$\frac{2r_1r_2(1-r_1)(1-r_2)}{(1-r)^2}$	$\frac{r_1^2r_2^2}{(1-r)^2}$
21	$\frac{r(1-r)}{2}$	$\frac{(1-r_1)^2r_2(1-r_2)}{r(1-r)}$	$\frac{r_1(1-r_1)(1-2r_2+2r_2^2)}{r(1-r)}$	$\frac{r_1^2r_2(1-r_2)}{r(1-r)}$
20	$\frac{r^2}{4}$	$\frac{(1-r_1)^2r_2^2}{r^2}$	$\frac{2r_1r_2(1-r_1)(1-r_2)}{r^2}$	$\frac{r_1^2(1-r_2)^2}{r^2}$
12	$\frac{r(1-r)}{2}$	$\frac{(1-r_2)^2r_1(1-r_1)}{r(1-r)}$	$\frac{r_2(1-r_2)(1-2r_1+2r_1^2)}{r(1-r)}$	$\frac{r_2^2r_1(1-r_1)}{r(1-r)}$
11	$\frac{(1-r)^2 + r^2}{2}$	$\frac{2r_1r_2(1-r_1)(1-r_2)}{r^2 + (1-r)^2}$	$\frac{(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)}{r^2 + (1-r)^2}$	$\frac{2r_1r_2(1-r_1)(1-r_2)}{r^2 + (1-r)^2}$
10	$\frac{r(1-r)}{2}$	$\frac{r_2^2r_1(1-r_1)}{r(1-r)}$	$\frac{r_2(1-r_2)(1-2r_1+2r_1^2)}{r(1-r)}$	$\frac{(1-r_2)^2r_1(1-r_1)}{r(1-r)}$
02	$\frac{r^2}{4}$	$\frac{(1-r_2)^2r_1^2}{r^2}$	$\frac{2r_1r_2(1-r_1)(1-r_2)}{r^2}$	$\frac{(1-r_1)^2r_2^2}{r^2}$
01	$\frac{r(1-r)}{2}$	$\frac{r_1^2r_2(1-r_2)}{r(1-r)}$	$\frac{r_1(1-r_1)(1-2r_2+2r_2^2)}{r(1-r)}$	$\frac{(1-r_1)^2r_2(1-r_2)}{r(1-r)}$

$$\frac{(1-r)^2}{4}$$

$$\frac{r_1^2 r_2^2}{(1-r)^2}$$

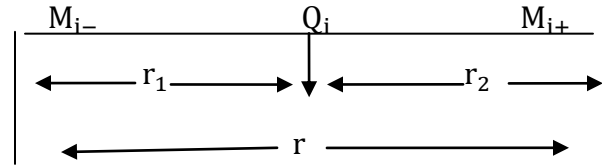
$$\frac{2r_1 r_2 (1-r_1)(1-r_2)}{(1-r)^2}$$

$$\frac{(1-r_1)^2 (1-r_2)^2}{(1-r)^2}$$

3.5 QTL Detection

3.5.1 Standard Interval Mapping (SIM)

Once phenotypic and genotypic data have been checked, one must decide what statistical procedure to use to associate phenotype with genotype. The most commonly used method for QTL analysis is interval mapping, in which one posits the presence of a single QTL and considers each point on a dense grid across the genome, one at a time, as the location of the putative QTL (Karl, W. 2009). Standard interval mapping uses MLE under a mixture model by considering a grid of positions along the genome as the possible locations for the QTL.



where r_1 and r_2 denote the recombination fraction between QTL and left marker M_{i-} and the right marker M_{i+} , respectively. Where r is the recombination fraction between the two flanking markers M_{i-} and M_{i+} . Assuming no crossover interference (or $r = r_1 + r_2 - 2r_1 r_2$), and if an individual is typed at both markers, the QTL genotype probabilities given the marker genotypes are as in Table (3.1).

If we assume that quantitative trait values are affected by genetic factors and environmental effects, the genetic model between trait values and genotypes is shown as

$$P = G + E + \varepsilon. \quad (3.2)$$

here, P represents phenotype values of the quantitative trait, G represents the genotype values, and E is the environmental effect. But our aim is only considering the genetic effect in this study, and the above model can be rewritten if we want to take additive effects (effects when the genotype is the homozygote) and dominant effects (effects when the genotype is the heterozygote) into account as follows:

$$\begin{aligned} \text{Phenotype} &= \text{mean} + \text{additive effect} + \text{dominance effect} + \text{error} \\ \text{additive effect} &= \text{mean of superior parent} - \text{mean of inferior parent} \end{aligned}$$

dominance effect = mean of hybrid – the average of the two parents

$$P = u + a + d + \varepsilon \quad (3.3)$$

here, u is a constant, a represent additive effects, d represents dominance effects, and ε a residual term that is assumed to be normally distributed. As for additive and dominant effects, it becomes $u + a$ in case of the homozygote of the maternal line, $u - a$ in case of the homozygote of the paternal origin, and $u + d$ in case of the heterozygote.

The data needed in QTL analysis is a set of values of the target trait and genotypes of marker loci. Since, the maize grain yield with its markers genotyped from F_2 intercross population, there are three genotypes at marker loci in the data sets in our study: a homozygote of the maternal line (AA), a homozygote of the paternal line (BB), and a heterozygote of both lines (AB).

Genotypic values are composed of the overall mean of the trait (u), the additive effect (a) of the gene due to the substitution of alleles from A to B , or the dominance effect (d) due to the interaction effect of different alleles A and B at the gene. If there is no dominance, $d = 0$; if allele A is dominant over B , d is positive; and if allele B is dominant over A , d is negative. Dominance is complete if d is equal to $+a$ or $-a$, and there is overdominance if d is greater than $+a$ or less than $-a$. Levels of dominance were estimated as $LD = |d|/|a|$ ratio for each QTL, and the average level of dominance (ALD) was computed weighing each LD ratio. Following the studies by Sibov *et al.* (2003b), gene action was characterized as follows: additive (A) $0.0 \leq LD \leq 0.20$; partial dominance (PD) $0.21 \leq LD \leq 0.80$; dominance (D) $0.81 \leq LD \leq 1.20$; and overdominance (OD) $LD > 1.21$.

We consider a particular fixed position for the QTL, and let g_i denote the QTL genotype for individual i . We assume that $y_i|g_i \sim N(\mu_{g_i}, \sigma^2)$, but the QTL genotype is generally not known. For each individual, we may calculate $p_{ij} = \Pr(g = j|M_i)$, where $j = 1, 2, 3$; M_i denotes the multipoint marker genotype data for individual i .

Given the marker data, an individual's follows a mixture of normal distributions, with known mixing proportions (the p_{ij}). That is, the density function for the phenotype of individual i is

$$\sum_{j=1}^3 p_{ij} \phi_j \left(y_i; \mu_j, \sigma^2 \right)$$

where, $\mu_1 = \mu_{QQ} = u + a$, $\mu_2 = \mu_{Qq} = u + d$, $\mu_3 = \mu_{qq} = u - a$, and ϕ_j the density of the normal distribution for the j^{th} QTL genotype and the sum is over the possible QTL genotypes.

3.5.1.1 Statistical Model for SIM Approach

A statistical model for SIM approach based on the genetic model (3.2) for testing a QTL in a marker interval is defined as

$$y_i = \mu + ax_i + dz_i + e_i \quad (3.4)$$

where

$$(x_i, z_i) = \begin{cases} \left(1, -\frac{1}{2}\right) \text{ for } QQ \\ \left(0, \frac{1}{2}\right) \text{ for } Qq \\ \left(-1, -\frac{1}{2}\right) \text{ for } qq \end{cases}$$

y_i is the phenotypic value of the i^{th} , μ is the general mean effect; and e_i is a random error. We assume $e_i \sim N(0, \sigma^2)$.

In the SIM model (3.4), each phenotypic observation y_i is assumed to follow a mixture of three possible Gaussian densities with different means and mixing proportion, since observation y_i 's are influenced by three QTL genotypes QQ , Qq and qq . Therefore, the density of each phenotypic observation (y_i) is defined as

$$f(y_i|\theta) = \sum_{j=1}^3 p_{ij} \phi\left(\frac{y_i - \mu_j}{\sigma}\right), \quad (3.5)$$

where $\theta = (\mu, p, a, d, \sigma^2)$, ϕ is a standard normal probability density function, $\mu_1 = \mu + a$, $\mu_2 = \mu + d$ and $\mu_3 = \mu - a$. The mixing proportions p'_{ij} s, which are functions of the QTL position parameter p , are conditional probabilities of QTL genotypes given marker genotypes.

Given the data with N individuals the likelihood function for $\theta = (\mu, p, a, d, \sigma^2)$ is

$$L(\theta|Y) = \prod_{i=1}^N f(y_i|\theta) = \prod_{i=1}^N \sum_{j=1}^3 p_{ij} \phi\left(\frac{y_i - \mu_j}{\sigma}\right). \quad (3.6)$$

The normal mixture model in (3.4) can be treated as an incomplete data problem (Little and Rubin 1987) because the genotypes of QTL are unknown.

3.5.1.2 EM Algorithm

Dempster *et al.* (1977) proposed the EM (Expectation - Maximization) algorithm as a general approach to iterative computation of maximum likelihood estimates for incomplete data problems. The EM algorithm has been used to obtain MLEs in several studies of QTL mapping analyses (Lander and Botstein 1989; Zeng 1994).

Let

$$g_i(x_i, z_i) = \begin{cases} p_{i1} & \text{if } x_i = 1 \text{ and } z_i = -1/2 \\ p_{i2} & \text{if } x_i = 0 \text{ and } z_i = 1/2 \\ p_{i3} & \text{if } x_i = -1 \text{ and } z_i = -1/2 \end{cases}$$

be the distribution of QTL genotype specified by x_i and z_i . We treat the unobserved QTL genotypes (x_i and z_i) as missing data, denoted by $y_{(mis, i)}$, and treat trait y_i as observed data, denoted by $y_{(obs, i)}$. Then, the combination of $y_{(obs, i)}$ and $y_{(mis, i)}$ is the complete data, denoted by $y_{(com, i)}$. In this setting we can apply the EM algorithm to obtain the MLEs. The conditional distribution of observed data, given missing data, can be considered as an independent sample from a population such that

$$y_i | (\theta, x_i, z_i) \sim N(\mu + ax_i + dz_i, \sigma^2).$$

The complete data density function is thus

$$f(y_{(com, i)} | \theta) = f(y_{(obs, i)} | \theta, x_i, z_i) g(x_i, z_i),$$

the product of densities of conditional observed and missing data.

At a given position, p (in cM) can be determined, and the EM algorithm is used for obtaining the MLEs of a, d and σ^2 . By definition, the iteration of the (s+1) EM step is as follows:

Expectation (E) step: compute the conditional expected complete data log-likelihood with respect to the conditional distribution of Y_{mis} given Y_{obs} and the current estimated parameter value $\theta^{(s)}$,

$$\begin{aligned}
Q(\theta|\theta^{(s)}) &= \int \log L(\theta|Y_{com}) f(Y_{mis}|Y_{obs}, \theta = \theta^{(s)}) dY_{mis} \\
&= \int \log \left[\prod_{i=1}^N \phi\left(\frac{y_i - \mu_j}{\sigma}\right) g_i(x_i, z_i) \right] \times f(Y_{mis}|Y_{obs}, \theta = \theta^{(s)}) dY_{mis}.
\end{aligned} \tag{3.8}$$

Because each observation is independent, the joint probability of N observations $f(Y_{mis}|Y_{obs}, \theta = \theta^{(s)})$ is the product of N individual probabilities $f(y_{mis}|y_{obs}, \theta = \theta^{(s)})$, $i = 1, 2, \dots, N$.

$$\begin{aligned}
Q(\theta|\theta^{(s)}) &= \sum_{i=1}^N \int \log \left[\phi\left(\frac{y_i - \mu_j}{\sigma}\right) g_i(x_i, z_i) \right] \times f(y_{mis}|y_{obs}, \theta = \theta^{(s)}) dy_{mis} \\
&= \sum_{i=1}^N \sum_{j=1}^3 \log \left[\phi\left(\frac{y_i - \mu_j}{\sigma}\right) p_{ij} \right] \times \frac{p_{ij} \phi\left(\frac{y_i - \mu_j^{(s)}}{\sigma^{(s)}}\right)}{\sum_{j=1}^3 p_{ij} \phi\left(\frac{y_i - \mu_j^{(s)}}{\sigma^{(s)}}\right)} \\
&= \sum_{i=1}^N \sum_{j=1}^3 \log \left[\phi\left(\frac{y_i - \mu_j}{\sigma}\right) p_{ij} \right] \times \pi_{ij}^{(s)}.
\end{aligned} \tag{3.9}$$

Where, $\pi_{ij}^{(s)} = \frac{p_{ij} \phi\left(\frac{y_i - \mu_j^{(s)}}{\sigma^{(s)}}\right)}{\sum_{j=1}^3 p_{ij} \phi\left(\frac{y_i - \mu_j^{(s)}}{\sigma^{(s)}}\right)}$, $j = 1, 2, 3$ and $i = 1, 2, \dots, n$.

Maximization (M)-step: Find $\theta^{(s+1)}$ to maximize the conditional expected log-likelihood $Q(\theta|\theta^{(s)})$: by taking the derivatives of $Q(\theta|\theta^{(s)})$ with respect to each parameter, the solutions of parameters in closed form are as follows. For a and d ,

$$\begin{aligned}
a^{(s+1)} &= \frac{\sum_{i=1}^N \left[\left(\pi_{i1}^{(s)} - \pi_{i3}^{(s)} \right) y_i - \frac{1}{2} \left(\pi_{i3}^{(s)} - \pi_{i1}^{(s)} \right) d^{(s)} \right]}{\sum_{i=1}^N \left(\pi_{i1}^{(s)} + \pi_{i3}^{(s)} \right)} \\
d^{(s+1)} &= \frac{\sum_{i=1}^N \frac{1}{2} \left[\left(-\pi_{i1}^{(s)} + \pi_{i2}^{(s)} - \pi_{i3}^{(s)} \right) y_i - \left(\pi_{i3}^{(s)} - \pi_{i1}^{(s)} \right) a^{(s+1)} \right]}{\frac{1}{4} \sum_{i=1}^N \left(\pi_{i1}^{(s)} + \pi_{i2}^{(s)} + \pi_{i3}^{(s)} \right)}
\end{aligned}$$

$$\hat{\sigma}^{2(s+1)} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^3 \pi_{ij}^{(s)} \left(y_i - \hat{\mu}_j^{(s)} \right)^2.$$

The EM algorithm is then initiated by taking

$$\hat{\mu}_j^{(0)} = \frac{\sum_{i=1}^N \pi_{ij}^{(s)} y_i}{\sum_{i=1}^N \pi_{ij}^{(s)}} \quad \text{and} \quad \hat{\sigma}^{2(0)} = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2$$

until the estimates converge, where $\bar{y}$ is the empirical mean of observations.

Hypothesis Testing

In QTL mapping, we test whether there is a QTL at a given position within a marker interval.

The hypotheses are

$H_0: a = 0 \text{ and } d = 0$ (there is no QTL at that position).

$H_1: \text{at least one of them is not zero}$ (there is a QTL at that position).

The corresponding likelihood function is

$$L_0(\theta_0) = \prod_{i=1}^N f(y_i, \mu_0, \sigma_0^2) \quad (3.10)$$

where μ_0 and σ_0^2 are the mean and variance of the overall population, respectively, and $\theta_0(\mu_0, \sigma_0^2)$ is the parameter vector.

Under the null model, the maximum likelihood of parameters can be directly obtained as

$$\hat{\mu}_0 = \bar{y} \text{ and } \hat{\sigma}_0^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2$$

The test statistic for the above hypothesis can be expressed as the log-likelihood ratio of the full model over the null model:

$$LR = 2 \ln \left[\frac{L(\hat{\theta})}{L_0(\hat{\theta}_0)} \right] \quad (3.11)$$

where $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\sigma}^2)$ and $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}_0^2)$ are two vectors of the maximum likelihood estimates under the full model and null model, respectively. If a high peak of the LR profile exceeds a critical threshold then a QTL that controls the trait is asserted to exist in a marker interval.

In addition, we can relate LR and LOD score as follows:

$$LOD = \log_{10} \frac{L(\hat{\theta})}{L_0(\hat{\theta}_0)} \quad (3.12)$$

$$= \frac{1}{2} \log_{10}(e) LR = 0.127 LR$$

$$LR = 2 \log(10) LOD = 4.605 LOD$$

so that the denominator of the LOD score is the likelihood under the null hypothesis that there is no QTL anywhere in the genome.

First calculate a LOD score at each increment (walking step) in the interval and finally to get the profile of LOD score for whole genome. When a peak has exceeded the threshold value, we declare that a QTL have been found at that location.

LOD scores corresponding to each position are then plotted as a function of map position. The position with the highest LOD score is a position that maximizes the likelihood function. If the LOD score at a region exceeds a pre-defined critical threshold, a QTL is indicated at the neighborhood of the maximum of the LOD score with the width of the neighborhood defined by one or two LOD support interval (Lander and Botstein 1989). Lander and Botstein (1989) suggested that a typical LOD score threshold should be between 2 and 3 to ensure a 5% overall false positive error for detecting QTL.

The basic idea for interval mapping is simple. We first consider an interval between two observable markers M_{i-} and M_{i+} , each having two possible alleles for the given data. The genetic distance or recombination frequency between the two markers has been previously estimated. A map function (Haldane) is utilized to translate from recombination frequency to distance or vice visa. To calculate a LOD score at each increment (walking step) in the interval and finally to get the profile of LOD score for whole genome. When a peak has exceeded the threshold value, we declare that a QTL have been found at that location. We also tried to show the effects of closest markers (linked to QTL) using simple plots of grain yield against marker genotypes.

Variance explained by QTL

The QTL effect is often characterized by the proportion of the phenotypic variance explained by the QTL (also called the heritability due to the QTL (Karl W. 2009). Specifically, we consider F_2 intercross maize marker data

$$h^2 = \frac{\text{var}\{E(y|g)\}}{\text{var}(y)}, \quad (3.14)$$

where y is the phenotype and g is the QTL genotype. For an intercross marker data, the heritability due to the QTL is given as

$$h^2 = \frac{(2\hat{a}^2 + \hat{d}^2)}{(2\hat{a}^2 + \hat{d}^2 + 4\hat{\sigma}^2)}. \quad (3.15)$$

where $\hat{a}$, $\hat{d}$ be the estimated value for additive and dominance effects, respectively, where $\hat{\sigma}^2$ is the residual variance.

3.5.1.3 Problems with Interval Mapping

The number of QTLs cannot be determined, locations of QTLs cannot be well-resolved, and there is low statistical power. Even when there is no QTL within an interval, the likelihood profile for the interval can still exceed the significance threshold if there is a QTL at some nearby location on the chromosome. It is generally agreed that the mapping power of IM is low due to lack of background control, and linked QTL cannot be properly separated (Zeng 1994). Due to these problems of interval mapping we are going to show the most important and recent method, mixed model QTL analysis.

3.5.2 Linear Mixed Model Approach for Mapping QTLs

Quantitative trait loci (QTLs) have been researched by several methods, such as interval mapping method (Lander and Botstein, 1989) and composite interval mapping method (Zeng, 1994), which are based on the regression approaches. In this thesis we proposed a method for mapping QTLs based on the mixed model approaches, which has not been applied widely. When a genetic analysis is conducted to search QTLs using information of genetic markers, only QTL position and effects are especially interested and markers can be treated as a random sample from all potential markers.

By employing a mixed model approach where the detected QTLs are considered as fixed and molecular markers are random (Zhu and Weir, 1998) the phenotypic value of a quantitative trait measured on the i^{th} individual can be expressed as a mixed linear equation as follows.

Let a putative QTL is within two flanking codominant genetic markers M_{i-} and M_{i+} , the phenotypic value of quantitative trait measured on the i^{th} individual can be expressed as a mixed linear model,

$$\begin{aligned} y_i &= \mu + ax_{A_i} + dx_{D_i} + \sum_{k \neq M_{i-}, M_{i+}} u_{M_k} z_{M_{ki}} + \varepsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_{M_i}^T \mathbf{u}_M + \varepsilon_i \end{aligned} \quad (3.16)$$

where μ is the population mean; a and d are the additive and dominance effects for the QTL searched; x_{A_i} and x_{D_i} are coefficients for genetic effects; u_{M_k} is the random effect for the k^{th} marker genotype with its coefficients $z_{M_{ki}}$ taking the values of 1 for $M_{k1}M_{k1}$, 0 for $M_{k1}M_{k2}$, or -1 for $M_{k2}M_{k2}$; and ε_i is the random residual effect; $\boldsymbol{\beta}$ is a vector of fixed parameters including μ , a and d ; $\mathbf{u}_M$ is a random vector of marker effects; $\mathbf{z}_{M_i}^T$ is a row vector of the coefficients for $\mathbf{u}_M$ of the i^{th} individual; $\mathbf{x}_i^T$ is a row vector of the coefficients for $\boldsymbol{\beta}$ of the i^{th} individual.

Since QTL genotype of the i^{th} individual is unknown and model (3.16) is a mixture model, the second and third elements of $\mathbf{x}_i^T = [1, x_{A_i}, x_{D_i}]$ can only inferred by the probability of QTL genotype given observed flanking marker genotype, Table (3.1).

Model (3.16) can be expressed as a matrix form of general mixed linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_M \mathbf{U}_M + \boldsymbol{\varepsilon} \quad (3.17)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of phenotypic values of quantitative trait studied (grain yield); $\boldsymbol{\beta}$ is a vector of the fixed effects; $\mathbf{X}$ is the coefficient matrix with row vectors $\mathbf{x}_i^T$; $\mathbf{U}_M \sim N(0, \mathbf{G})$ is a constant matrix described the relationship of $\mathbf{U}_M$; $\mathbf{Z}_M$ is the known coefficient matrix for $\mathbf{U}_M$.

$\boldsymbol{\varepsilon}$ is a vector of residuals with mean $\mathbf{0}$ and variance $\mathbf{R}$ and uncorrelated with $\mathbf{u}$. We can include in $\boldsymbol{\beta}$ and $\mathbf{u}$ any fixed and random terms needed to explain non-genetic variation in the data. Fixed effects $\boldsymbol{\beta}$, include the mean and any putative QTL effects added, while $\mathbf{X}$ contains the

corresponding single column of 1s and the putative QTL covariates, and $\mathbf{Z}$ contains the coded marker data.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{pmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1q} \\ z_{21} & z_{22} & \dots & z_{2q} \\ \vdots & \vdots & \dots & \vdots \\ z_{N1} & z_{N2} & \dots & z_{Nq} \end{bmatrix}$$

For the (co) variance structures of $\mathbf{y}$ the assumptions are $\text{var}(\mathbf{u}) = \mathbf{G}$, $\text{var}(\boldsymbol{\varepsilon}) = \mathbf{R}$, $\text{cov}(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{cov}(\mathbf{y}, \mathbf{u}) = \mathbf{C} = \mathbf{ZG}$, and $\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$.

The simplest covariance matrix for $\mathbf{R}$ is the diagonal structure, in which the residuals associated with observations on the same subject are assumed to be uncorrelated and to have equal variance ($\mathbf{R} = \sigma_e^2 \mathbf{I}$). And $\mathbf{G}$ is the genetic variance-covariance matrix. In this case all genotypes are unrelated ($\mathbf{G} = \sigma_a^2 \mathbf{I}$).

Estimation techniques for mixed model

For the given data vector $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, the likelihood functions which is a functions of $\boldsymbol{\beta}$ and $\mathbf{V}$ is

$$\mathbf{L} = \mathbf{L}(\boldsymbol{\beta}, \mathbf{V}|\mathbf{y}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}}{(2\pi)^{\frac{1}{2}N}|\mathbf{V}|^{\frac{1}{2}}} \quad (3.18)$$

Maximization of $\mathbf{L}$ can be achieved by maximizing the logarithm of $\mathbf{L}$ of (3.18) which shall be denoted by $\mathbf{l}$:

$$\mathbf{l} = \log \mathbf{L} = -\frac{1}{2}N \log 2\pi - \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (3.19)$$

To maximize $\mathbf{l}$, we differentiate (3.19), with respect of $\boldsymbol{\beta}$

$$\mathbf{l}_{\boldsymbol{\beta}} = \frac{\partial \mathbf{l}}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \quad (3.20)$$

A general principle of maximizing $\mathbf{l}$ with respect to $\boldsymbol{\beta}$ is to equate (3.20) to zero and solve the resulting equation for $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of $\boldsymbol{\beta}$.

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (3.21)$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (3.22)$$

In a manner analogous to estimating the residual variance in linear models by ML, the estimates of the variance parameters are biased. The bias arises because estimation of the fixed effects is not taken into consideration when estimating the variance parameters.

An unbiased method for estimating the variance parameters is obtained if the observed data are replaced by a linear contrast. This idea was introduced by Patterson & Thompson (1971) and is called restricted (or residual) maximum likelihood (REML).

A property of ML estimation is that in estimating variance components it takes no account of the degrees of freedom that are involved in estimating fixed effects. The feature of ML not taking account of the degrees of freedom used for estimating fixed effects when estimating variance components is overcome by what has come to be known as restricted (or, more usually in Europe, residual) maximum likelihood (REML) estimation.

REML is often preferred to ML estimation, because it produces unbiased estimates of covariance parameters by taking into account the loss of degrees of freedom that result from estimating the fixed effects in $\boldsymbol{\beta}$.

REML involves the idea of applying ML to linear functions of $\mathbf{y}$, say $\mathbf{K}'\mathbf{y}$, for which $\mathbf{K}'$ is specifically designed so that $\mathbf{K}'\mathbf{y}$ contains none of the fixed effects which are part of the model for $\mathbf{y}$. So in ML, replace $\mathbf{y}$ with $\mathbf{K}'\mathbf{y}$ and one has REML. Two valuable consequences of using REML are first, that variance components are estimated without being affected by the fixed effects. This means that the variance estimates are invariant to the values of the fixed effects. Second, in estimating variance components with REML, degrees of freedom for the mixed effects are taken into account implicitly, whereas with ML they are not.

REML is based on linear combinations of elements of $\mathbf{y}$, chosen in such a way that those combinations do not contain any fixed effects, no matter what their value. These linear combinations turn out to be equivalent to residuals obtained after fitting the fixed effects. This results from starting with a set of values $\mathbf{K}'\mathbf{y}$ where vectors $\mathbf{K}'$ are chosen so that

$$\mathbf{K}'\mathbf{y} = \mathbf{K}'\mathbf{X}\boldsymbol{\beta} + \mathbf{K}'\mathbf{Z}\mathbf{u} + \mathbf{K}'\boldsymbol{\varepsilon} \quad (3.23)$$

contains no term in $\boldsymbol{\beta}$, i.e.,

$$\mathbf{K}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \quad \forall \boldsymbol{\beta} \quad (3.24)$$

and hence

$$\mathbf{K}'\mathbf{X} = \mathbf{0} \quad (3.25)$$

Therefore, the form of $\mathbf{K}$ must be $\mathbf{K}' = \mathbf{c}'(\mathbf{I} - \mathbf{X}\mathbf{X}^+)$ or

$$\mathbf{K}' = \mathbf{c}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{c}'(\mathbf{I} - \mathbf{X}\mathbf{X}^+) = \mathbf{c}'\mathbf{M} \quad (3.26)$$

for any $\mathbf{c}'$ and where $\mathbf{M}$ is defined as

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{X}\mathbf{X}^+ \quad (3.27)$$

With $\mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ we have, for $\mathbf{K}'\mathbf{X} = \mathbf{0}$, $\mathbf{K}'\mathbf{y} \sim \mathbf{N}(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K})$. The REML equations can therefore be derived from the ML equation of (3.18) replacing $\mathbf{y}$ by $\mathbf{K}'\mathbf{y}$, $\mathbf{X}$ by $\mathbf{K}'\mathbf{X} = \mathbf{0}$, $\mathbf{Z}$ by $\mathbf{K}'\mathbf{Z}$ and $\mathbf{V}$ by $\mathbf{K}'\mathbf{V}\mathbf{K}$.

Then the logarithm of the likelihood function of the REML profile is

$$\mathbf{l}_p = -\frac{1}{2} \log |\mathbf{K}'\mathbf{V}\mathbf{K}| - \frac{1}{2} \mathbf{y}'\mathbf{P}\mathbf{y} - \frac{N - \text{rank}(\mathbf{X})}{2} \log 2\pi \quad (3.28)$$

Where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \quad (3.29)$$

$$\mathbf{P} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{V}^{-1}$$

$$\mathbf{P} = \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'$$

The next step is differentiating $\mathbf{l}_p$ with respect to $\mathbf{V}$. let $\boldsymbol{\phi}$ be the arbitrary element of $\mathbf{V}^{-1}$ which needs to be estimated. Mathematically,

$$\frac{\partial \mathbf{l}_p}{\partial \boldsymbol{\phi}} = \mathbf{0}, \quad \forall \boldsymbol{\phi} \in \mathbf{V} \quad (3.30)$$

An iterative method must be used to find estimates for $\emptyset$. The conventional optimization methods, which require first and second derivatives, may be applied. Thus, in particular, we have used ML and REML iteration in *nlme* and *lme4* R packages.

Best linear unbiased predictor (BLUP) is used in linear mixed model to predict the random effects. By taking model in (3.17), with $E(\mathbf{u}) = \mathbf{0}$ we consider the problem of predicting

$$\mathbf{w} = \mathbf{L}'\boldsymbol{\beta} + \mathbf{u} \quad (3.31)$$

for some known matrix $\mathbf{L}'$, such that $\mathbf{L}'\boldsymbol{\beta}$ is estimable; i.e., $\mathbf{L}' = \mathbf{T}'\mathbf{X}$ for some $\mathbf{T}$. Since $\mathbf{w}$ involves both fixed and random variables, there might be a debate as to whether we should estimate $\mathbf{w}$ or predict $\mathbf{w}$. We will predict $\mathbf{w}$, and will choose $\tilde{\mathbf{w}}$ as a predictor to have three properties:

$$\text{“Best” in a sense that: minimizing } E\{(\tilde{\mathbf{w}} - \mathbf{w})'\mathbf{A}(\tilde{\mathbf{w}} - \mathbf{w})\}, \quad (3.32)$$

where $\mathbf{A}$ is any positive definite symmetric matrix.

Linear in $\mathbf{y}$: $\tilde{\mathbf{w}} = \mathbf{a} + \mathbf{B}\mathbf{y}$, with $\mathbf{a}$ and $\mathbf{B}$ not involving $\boldsymbol{\beta}$.

Unbiased: $E(\tilde{\mathbf{w}}) = E(\mathbf{w}) = \mathbf{L}'\boldsymbol{\beta}$.

The resulting predictor is a best linear unbiased predictor (BLUP). Note that unbiasedness is now a criterion of the prediction procedure and not just a by-product of it. Introducing it as a criterion arises from the presence of $\boldsymbol{\beta}$ in (3.31).

Then,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \end{bmatrix} \sim \left\{ \begin{bmatrix} \mathbf{L}'\boldsymbol{\beta} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{C} \\ \mathbf{C}' & \mathbf{V} \end{bmatrix} \right\} \quad (3.33)$$

The unbiasedness required of $\tilde{\mathbf{w}}$ demands that $\mathbf{a} + \mathbf{B}\mathbf{X}\boldsymbol{\beta} = \mathbf{L}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, and $\mathbf{a}$ is not depend on $\boldsymbol{\beta}$ then $\mathbf{a} = \mathbf{0}$ and $\mathbf{B}\mathbf{X} = \mathbf{L}'$. Consequently, the predictor is $\tilde{\mathbf{w}} = \mathbf{B}\mathbf{y}$ and in $\mathbf{w} = \mathbf{L}'\boldsymbol{\beta} + \mathbf{u}$ the term $\mathbf{L}'\boldsymbol{\beta}$ is estimable function of $\boldsymbol{\beta}$ in the model $\mathbf{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. This limits the form of $\mathbf{L}'$ in $\mathbf{w}$, but it is obviously a reasonable limitation, and the predictor is called BLUP. It is

$$BLUP(\mathbf{w}) = \tilde{\mathbf{w}} = \mathbf{L}'\hat{\boldsymbol{\beta}} + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (3.34)$$

$$BLUE(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (3.35)$$

We know that $\tilde{\mathbf{u}} = \boldsymbol{\mu}_u + \mathbf{CV}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$. when $\mathbf{E}(\mathbf{u}) = \mathbf{0} = \boldsymbol{\mu}_u$ and $\mathbf{E}(\mathbf{y}) = \boldsymbol{\mu}_y = \mathbf{X}\boldsymbol{\beta}$, then $\tilde{\mathbf{u}} = \mathbf{CV}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. When we replace $\mathbf{X}\boldsymbol{\beta}$ by its *BLUE*($\mathbf{X}\boldsymbol{\beta}$) = $\mathbf{X}\hat{\boldsymbol{\beta}}$, we get

$$BLUP(\mathbf{u}) = \tilde{\mathbf{u}} = \mathbf{CV}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (3.36)$$

Finally, we can write the predictor as

$$\tilde{\mathbf{w}} = \mathbf{L}'\hat{\boldsymbol{\beta}} + \tilde{\mathbf{u}} \quad (3.37)$$

and $\tilde{\mathbf{w}}$ is thus the sum of BLUE of $\mathbf{L}'\boldsymbol{\beta}$ and the BLUP of $\mathbf{u}$.

3.5.2.1 Searching (localization) of QTLs

The detection stage determines the number of QTL and effect on each chromosome. We can search QTL within two flanking markers M_{i-} and M_{i+} for the whole genome by setting a prior value for recombination fraction $r_{M_{i-}Q}$ between marker M_{i-} and locus Q. For each prior value $r_{M_{i-}Q}$, the likelihood ratio test statistic (LR) can be calculated by

$$LR = 2l_1(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}}, \hat{r}_{M_{i-}Q}) - 2l_0(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}}, r_{M_{i-}Q} = 0.5), \quad (3.38)$$

where l_1 is the observed log likelihood function evaluated at the MLE of the given parameters for the alternative hypothesis, while l_0 for the null hypothesis. Variance components in $\mathbf{V}$ can be replaced by their unbiased estimates,

$$\hat{\mathbf{V}} = \sum_u \sigma_u^2 \mathbf{Z}_u \mathbf{G}_u \mathbf{Z}_u'$$

LR approximately has a χ^2 distribution with $df = 1$, (Zhu 1999).

When the null presentation of no QTL within two flanking markers M_{i-} and M_{i+} is rejected, $r_{M_{i-}Q}$ infers the position of the QTL while $\boldsymbol{\beta}$ gives the estimates of additive and dominance effects of this QTL. However, we have to fix the number of detected QTLs before we make a hypothesis to test the additive and dominance effects for each QTL.

Construction of genetic predictors

Genetic predictors, or regressors', for the additive and dominance genetic QTL effects were constructed for a grid of evaluation points, q, along the genome. These genetic predictors were

introduced as explanatory variables in the mixed models. Genetic predictors were calculated at all the marker positions and at an additional grid of points with a maximum step size of 1 cM along the genome.

Thus, Genetic predictors or genotypic covariables for model (3.17) can be obtained from conditional QTL genotypic probabilities as

$$x_{iq}^{add} = P(QQ/M_i) - P(qq/M_i) \quad (3.39)$$

$$x_{iq}^{dom} = P(Qq/M_i) - \frac{P(QQ/M_i) + P(qq/M_i)}{2}. \quad (3.40)$$

Then, x_{iq}^{add} and x_{iq}^{dom} are the genetic predictors for additive and dominance QTL effects at position q (in cM) for all individuals (Haley and Knott 1992).

Variation explained by each QTL in mixed model

After identifying statistically significant genetic loci (Kennedy *et al.* 1992), many geneticists would ask how much of the phenotypic variation is explained by each quantitative trait locus (QTL) for the interpretation or comparison purpose. In other words, what is the relative degree of improvement of the model fit to the data that result by including this significant genetic effect?

Two research groups (Cox and Snell 1999) have independently proposed the likelihood-ratio based R^2 (R_{LR}^2); a R^2 -like statistic based on the LR:

$$R_{LR}^2 = 1 - \exp\left(-\frac{2}{N}(\log L_M - \log L_0)\right) \quad (3.41)$$

where $\log L_M$ is the maximum log-likelihood of the model of interest, $\log L_0$ is the maximum log-likelihood of the intercept only model, N is the number of observations. The calculation (or estimation technique) is based on maximum likelihood (ML), not restricted ML (REML).

3.5.2.2 Model-Building Strategies

A primary goal of model selection is to choose the simplest model that provides the best fit to the observed data. There may be several choices concerning which fixed and random effects should be included in an LMM. There are also many possible choices of covariance structures for the $\mathbf{G}$

and $\mathbf{R}$ matrices. All these considerations have an impact on both the estimated marginal mean ($\mathbf{X}\boldsymbol{\beta}$) and the estimated marginal variance-covariance matrix $\mathbf{V}(=\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$ for the observed responses in $\mathbf{y}$ based on the specified model.

The process of building an LMM for a given set of data is an iterative one that requires a series of model-fitting steps and investigations, and selection of appropriate mean and covariance structures for the observed data. Model building typically involves a balance of statistical and subject matter considerations; there is no single strategy that applies to every application (Brady, T. *et al.* 2007).

The Top-Down Strategy

The following broadly defined steps are suggested by Verbeke and Molenberghs (2000) for building an LMM for a given data set, which we refer to as a top-down strategy for model building because it involves starting with a model that includes the maximum number of fixed effects that we wish to consider in a model.

1. Start with a well-specified mean structure for the model: This step typically involves adding the fixed effects of as many covariates (and interactions between the covariates) as possible to the model to make sure that the systematic variation in the responses is well explained before investigating various covariance structures to describe random variation in the data.
2. Select a structure for the random effects in the model: This step involves the selection of a set of random effects to include in the model. The need for including the selected random effects can be tested by performing REML-based likelihood ratio tests for the associated covariance parameters.
3. Select a covariance structure for the residuals in the model: Once fixed effects and random effects have been added to the model, the remaining variation in the observed responses is due to residual error, and an appropriate covariance structure for the residuals should be investigated.
4. Reduce the model: This step involves using appropriate statistical tests to determine whether certain fixed-effect parameters are needed in the model. At this stage we use Akaike information criterion (AIC). That is, AIC may be calculated based on the (ML or REML) log-likelihood of the fitted model.

$$AIC = -2l(\hat{\beta}, \hat{u}) + 2p \quad (3.42)$$

where p represents the total number of parameters being estimated in the model for both the fixed and random effects. Note that the AIC in effect “penalizes” the fit of a model for the number of parameters being estimated by adding $2p$ to the $-2\log$ -likelihood. The model having smallest AIC is best.

Influence Measures

As Littell *et al.* (2006) disclosed, the goal of influence analysis is to determine how observations impact the analysis. Of course, all observations are influential, in the sense that their removal changes the numerical results of the analysis. The goal is to identify those observations that are so influential that their presence or absence from the data changes an important aspect of the analysis, yields qualitatively different inferences, or violates of assumptions of the model. The goal is not to determine the observations for removal from the analysis, but to decide prior to a perturbation analysis what matters (Schabenberger 2004).

The basic procedure for quantifying influence:

1. Fit the model to the data and obtain estimates of all parameters.
2. Remove one or more data points from the analysis and compute updated estimates of model parameters.
3. Based on full- and reduced-data estimates, contrast quantities of interest to determine how the absence of the observations changes the analysis.

We use the subscript (U) to denote quantities obtained without the observations in the set U . For example, $\hat{\beta}_U$ denotes the fixed-effects “leave- U -out” estimates. Note that the set U can contain multiple observations.

Change in parameter estimates

A common way to measure the impact on a vector of parameter estimates is to compute a quadratic form in the difference between the full data and reduced data estimates. These quadratic forms are based on Cook’s D (Cook 1977) and the multivariate DFFITS statistic, MDFFITS (Belsley *et al.* 1980). The primary difference between D and MDFFITS is that the

latter uses an externalized estimate of the variance of the parameter estimates, while Cook's D does not. For the fixed effects, the two statistics are

$$D(\boldsymbol{\beta}) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_U)' \widehat{\text{var}}[\hat{\boldsymbol{\beta}}]^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_U)}{\text{rank}(\mathbf{X})} \quad (3.42)$$

$$D(\mathbf{u}) = (\hat{\mathbf{u}} - \hat{\mathbf{u}}_U)' \widehat{\text{var}}[\hat{\mathbf{u}}]^{-1} (\hat{\mathbf{u}} - \hat{\mathbf{u}}_U) \quad (3.43)$$

$$\text{MDFITS}(\boldsymbol{\beta}) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_U)' \widehat{\text{var}}[\hat{\boldsymbol{\beta}}_U]^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_U)}{\text{rank}(\mathbf{X})} \quad (3.44)$$

For both statistics, we are concerned about large values, indicating that the change in the parameter estimate is large relative to the variability of the estimate.

Change in precision of estimates

The effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's D, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large. The two common methods to do this are functions of the trace and determinants of the variance matrices based on the full-data and the reduced-data estimates:

$$\text{COVTRACE}(\boldsymbol{\beta}) = |\text{trace}\{\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) - \widehat{\text{var}}(\hat{\boldsymbol{\beta}}_U)\} - \text{rank}(\mathbf{X})| \quad (3.45)$$

$$\text{COVRATIO}(\boldsymbol{\beta}) = \frac{|\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_U)|}{|\widehat{\text{var}}(\hat{\boldsymbol{\beta}})|} \quad (3.46)$$

The reason behind the COVTRACE statistic is that $\mathbf{X}$ is a positive semi-definite matrix, and then $\text{trace}(\mathbf{X}^{-1}\mathbf{X})$ is equal the rank of $\mathbf{X}$. If the variance parameter estimates is not affected by the removal of observations, then the trace in the COVTRACE statistic should equal the rank of $\mathbf{X}$. The benchmark for lack of "influence" is thus a value of zero. As a result if COVTRACE statistic increases, so does the influential on the precision of the parameter estimates and linear function hereof, such as tests of fixed effects.

The COVRATIO statistic relates the determinants of the covariance matrices of the full data and reduced data estimates. The yardstick of no influence is a value of one. Values larger than one indicate increased precision in the full-data case, and values smaller than one indicate higher precision for the reduced data estimates. Such an interpretation in terms of increase or decrease in precision is not possible with the COVTRACE statistic. The disadvantage of the COVRATIO

statistic is primarily computational. When the $\mathbf{X}$ matrix is of less than full rank, special manipulations are required to compute the determinants of the non-singular components of the variance matrices.

Once we have checked the influential analysis for the two datasets, the next step is residual analysis to check the assumption that random effects and error term follow normal distribution.

Residual diagnostics

In general, residuals should be assessed for normality, constant variance, and outliers. In the context of LMMs, we consider conditional residuals and their “studentized” versions.

A conditional residual is the difference between the observed value and the conditional predicted value of the dependent variable. The conditional distribution of $\mathbf{y}|\mathbf{u}$ has mean $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, meaning that for the realized $\mathbf{u}$, $\mathbf{E}(\mathbf{y}|\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ is the conditional mean, and variance $\mathbf{R}$. the marginal distribution of $\mathbf{y}$ has mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\mathbf{Z}\mathbf{R}\mathbf{Z}' + \mathbf{R}$. Based on these concepts, we can define the conditional residuals as

$$\mathbf{r}_c = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}} \quad (3.47)$$

that measure the deviations from the conditional mean.

and the marginal residuals as

$$\mathbf{r}_m = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \quad (3.48)$$

that measure the deviations from the overall mean. In a model without random effects $\mathbf{u}$, the two sets of residuals coincide. Residual analyses indicate the outliers, homoscedasticity, normality etc.

3.5.2.3 Significance testing

Significance tests for fixed effects can be carried out using tests based on F or t distributions, as we show below. A test can be defined using a contrast; for example, $\mathbf{C} = \mathbf{L}'\hat{\boldsymbol{\beta}} = \mathbf{0}$ for fixed effects. For simple contrast $\mathbf{L}$ will just have one column.

The F test statistic for testing the null hypothesis that the contrast is zero is calculated from a statistic known as the Wald statistic, which is given by

$$W = (\mathbf{L}'\hat{\boldsymbol{\beta}})' \left(\text{var}(\mathbf{L}'\hat{\boldsymbol{\beta}}) \right)^{-1} (\mathbf{L}'\hat{\boldsymbol{\beta}}) \quad (3.49)$$

$$= (\mathbf{L}'\hat{\boldsymbol{\beta}})' (\mathbf{L}'\text{var}(\hat{\boldsymbol{\beta}})\mathbf{L})^{-1} (\mathbf{L}'\hat{\boldsymbol{\beta}})$$

for fixed effects. Thus the Wald statistic can be thought of as the square of the contrast divided by its variance.

Asymptotically, W follows a chi-squared distribution with degree of freedom (DF) equal to the DF of $\mathbf{L}$ (number of linearly independent rows of $\mathbf{L}$).

A better option is to use the Wald F statistic, which is calculated by

$$F_{DF1,DF2} = W/DF1, \quad (3.50)$$

where $DF1$ is the contrast DF (number of linearly independent rows of $\mathbf{L}$), and $DF2$ is the denominator DF corresponding to the DF of the contrast variance, $\mathbf{L}'\text{var}(\hat{\boldsymbol{\beta}})\mathbf{L}$. This statistic takes account of the fact that $\mathbf{L}'\text{var}(\hat{\boldsymbol{\beta}})\mathbf{L}$ is estimated and not known.

The Wald t statistic for tests of single contrasts is given by

$$t_{DF2} = (F_{1,DF2})^{1/2} = W^{1/2}. \quad (3.51)$$

A well-known approximation for the denominator DF for F tests is given by Satterthwaite (1946):

$$DF = \frac{2(\mathbf{L}'\text{var}(\hat{\boldsymbol{\beta}})\mathbf{L})^2}{\text{var}(\mathbf{L}'\text{var}(\hat{\boldsymbol{\beta}})\mathbf{L})}. \quad (3.52)$$

It is equal to twice the variance of the contrast divided by the variance of the variance of the contrast.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Diagnostics for Interval Mapping

Before proceeding with QTL mapping analyses, exploratory analyses of the phenotype and genotype data were performed to understand the distribution of the data and other characteristics. The distribution of the phenotype data displayed (Appendix A) shows maize grain yield approximately follows normal distribution ($\alpha = 0.05$); and it is free from outliers.

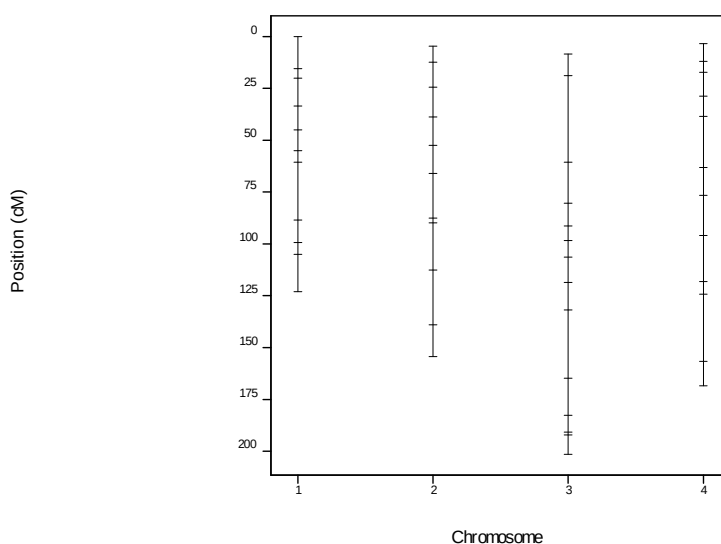


Figure 4.1: The genetic map of the typed markers in the maize data.

The genetic map (Figure 4.1) shows presence of 11 markers in chromosome one and two. Similarly, chromosome three and four has 14 and 12 markers, respectively. As we can see from the genetic map chromosome three has relatively dense markers which reasonably covers the genome. There is no any pattern on missing genotype plot (Appendix A), which indicates that no missing genotype on each chromosome.

Considering the number of markers on each chromosome, it could be tedious to present pairwise recombination fractions and LOD scores for all markers numerically. Thus, the simplest way to describe both recombination fractions and LOD scores would be to use graphical method simultaneously.

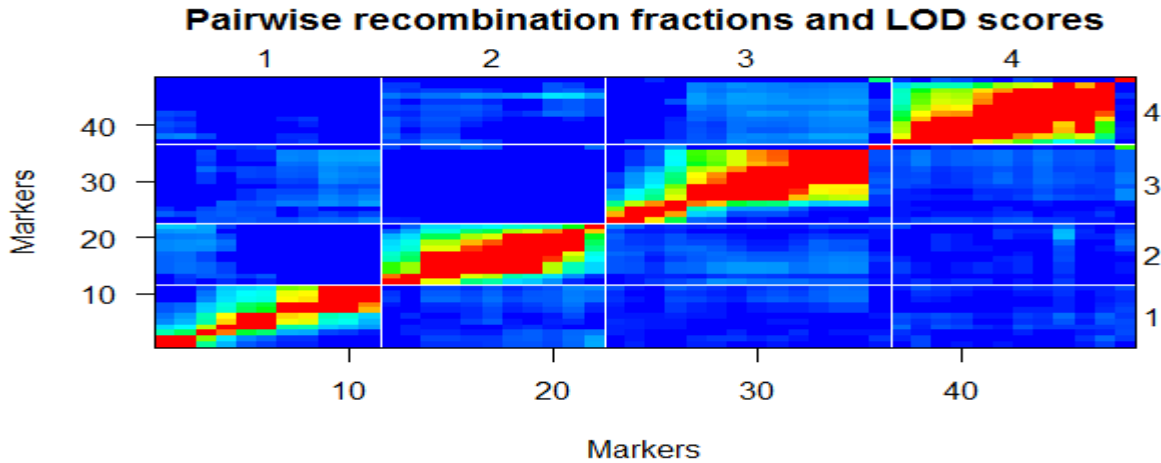


Figure 4.2: Plot of estimated recombination fractions (upper left) and LOD scores for a test of $r = 1/2$ (lower right) for all pairs of markers in the maize grain yield data.

In Figure 4.2, the LOD scores (in the lower right triangle) indicate no evidence for linkage between markers on different chromosomes. Red indicates linkage, blue indicates no linkage.

4.2 QTL detection and estimating their effects using interval mapping

In Figure 4.3, the red dot line shows the threshold LOD score value to detect QTL that affects the grain yield for maize. In chromosomes where the plots exceed the threshold values, indicate presence of significant QTL.

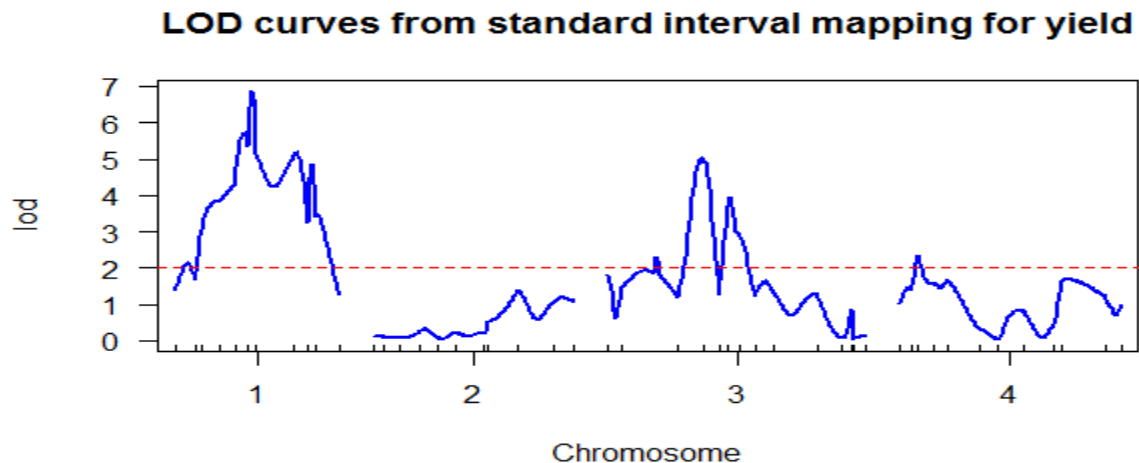


Figure 4.3: Detection of QTLs using LOD curves from a genome scan by EM algorithm for grain yield of maize data.

Thus, we conclude that no evidence for presence of QTLs in chromosome two as the plot does not move above 2.0 LOD value. Although the number of QTLs cannot be determined exactly on each chromosome, three QTLs are detected on chromosome one, three and four.

Both the LOD curve and Table 4.1 indicate that only 3 QTLs have a significant association with grain yield, one on chromosome 1 at position 58.0cM, one on chromosome 3 at 78.4cM, and the other one on chromosome 4 at 17.2cM. The first QTL (Loc58) is located between markers M5 and M7, the second QTL (Loc70) is located between markers M25 and M29, and finally *M39* is located between M37 and M48.

The QTLs identified in this population were tested for acceptance using the likelihood ratio that has approximately the chi-square distribution. Thus, likelihood ratio values of 31.6824, 23.072, and 10.776 represent significant values corresponding to probabilities of 0.05.

Table 4.1: QTL showing significant association with maize grain yield

QTL	Flanking markers	Chrom- ** osome	n.gen	Position (cM)	LR	LOD score	p-value	CI*
<i>loc58</i>	M5-M7	1	3	58.0	31.6824	6.88	0.0001	(4.33, 5.18)
<i>loc70</i>	M25-M29	3	3	78.4	23.072	5.01	0.006	(1.2, 2.93)
<i>M39</i>	M37-M48	4	3	17.2	10.776	2.34	0.043	(1.02, 0.99)

(*) is an approximate confidence interval for QTL location, based on the LOD profile (lower bound LOD value for left closest marker and upper bound LOD value for right closest marker)

(**) n.gen is the number of genotypes in each QTL.

LOD score of *loc58* and *loc70* indicates that the chance of obtaining the observed data, given that there is a QTL at the specified position, is 1,000 times more likely than if there are no QTL. While a LOD score of *M39* indicates that the chance of obtaining the observed data, given that there is a QTL at the specified position, is 100 times more likely than if there are no QTL. Thus, the LOD threshold values used in this study were related with the LOD threshold values of 2, 2.5, or 3 which are commonly used for QTL mapping in maize (Milena *et al.* 2005).

Table 4.2: Estimated QTL effects and LD with percentage of variance they explained.

QTL	Chr.name*	Pos.(cM)**	<i>a</i>	<i>d</i>	LD($ d/a $)	Gene action	variance
<i>loc58</i>	1	58.0	-0.1483	1.3425	9.053	OD ^{1*}	16.91
<i>loc70</i>	3	78.4	0.9669	-0.1543	0.159	A ^{2*}	14.69
<i>M39</i>	4	17.2	- 0.287	0.7614	2.653	OD	3.55

*Chromosomal name; **position in their respective chromosomes; 1* overdominance; 2* additive

Table 4.2 revealed that the additive and dominance effect of *loc58* are -0.1483 and 1.3425, respectively, at position 58.0cM on chromosome one. The additive effect of the gene *Loc58* due to the substitution of alleles from *A* to *B* is -0.1483 at a position of 58.0cM; while the dominance effect due to the interaction effect of different alleles *A* and *B* at the gene *loc58* is 1.3425. In other words, an estimate of the quantitative change in a trait that is associated with substituting one allele (*A*) with that of another allele (*B*) at *loc58* is -0.1483; and the change in grain yield in maize due to the interactions of alleles at *loc58* is 1.3425. On the other hand, the additive and dominance effect of *loc70* at position 78.4 on chromosome 3 are 0.9669 and -0.1543, respectively.

Moreover, the dominance effect for *loc58* and *M39* on chromosome one and four has positive sign which indicates that allele *A* is dominant over *B* at these genes. But, at the locus *Loc70* allele *B* is dominant over *A*, since the dominance effect at this gene has negative value. In addition, the action is overdominance at two quantitative trait loci (*loc58* and *M39*), and additive at *loc70*. This means that the heterozygous combination of the alleles at a locus is superior to either of the two possible homozygous combinations at *loc58* and *M39*.

The variation for grain yield explained by the three QTLs was 16.91% by *loc58*, 14.69% by *loc70* and 3.55% by *M39*. Thus, *loc58* and *Loc70* at their respective positions account for maximum variation among observed yield and have high effect on maize grain yield relative to *M39*.

4.3 Markers effect illustration

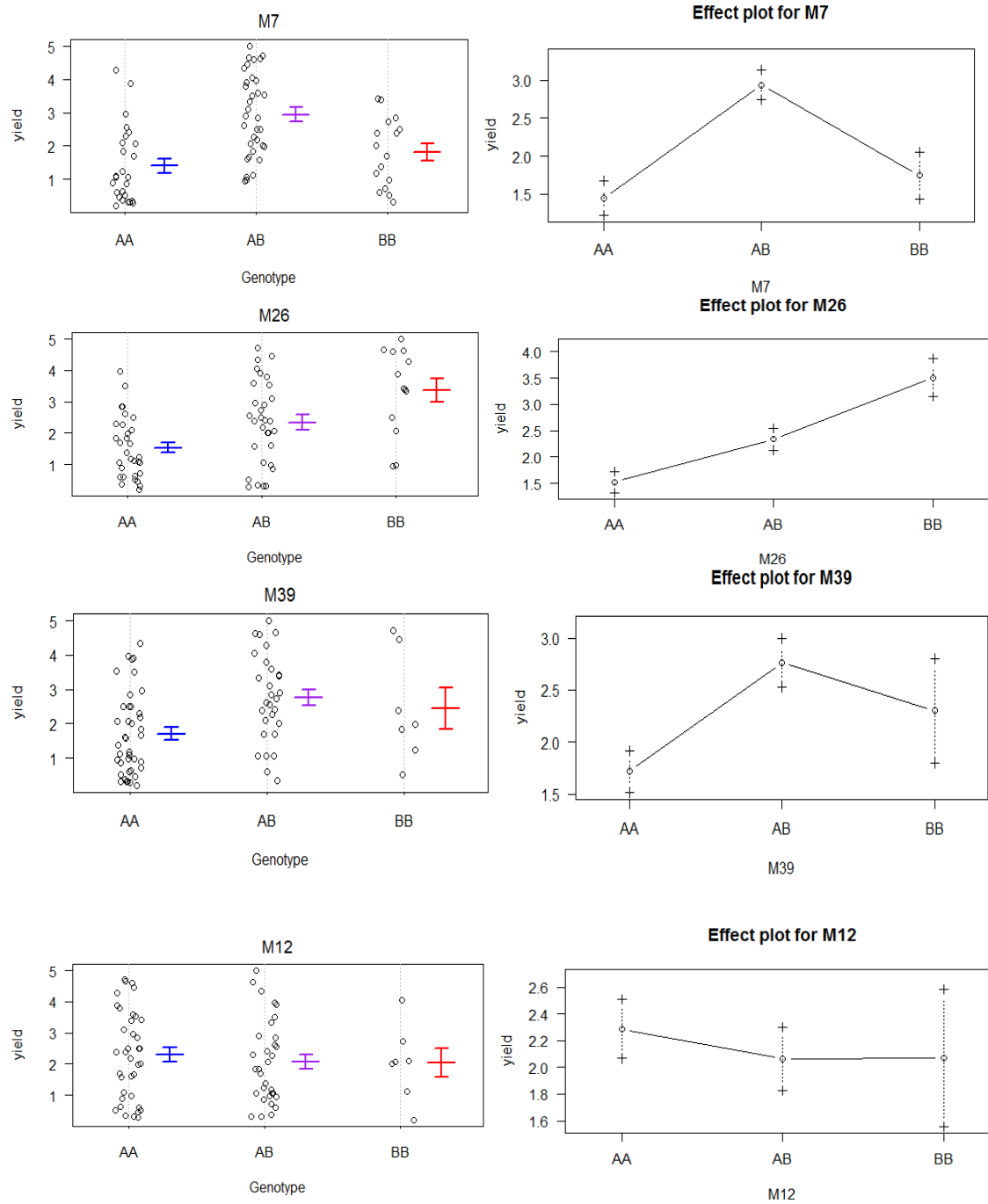


Figure 4.4: Dot plots of the maize yield phenotype against the genotype at four selected markers (left side) and effect plots for closest markers (right side). Confidence intervals for the average phenotype in each genotype group are also shown with different colors.

At marker 7 and marker 39, the heterozygous individuals exhibit a larger average phenotype than the homozygous, indicating that these markers are linked to a QTL (Figure 4.4). In similar manner at marker 26, the two genotype groups (homozygous for parent two and the heterozygous) individuals exhibit a larger phenotype than the homozygous for parent one, and so marker 26 is indicated to be linked to QTL. At *M12* which is located in chromosome two, on the other hand, the three genotype groups show similar phenotypes, and so *M12* is not indicated to be linked to a QTL.

Table 4.3: Grain yield means for three marker genotypes each linked to QTL

genotypes	Markers closest to each QTL		
	M7	M26	M39
AA	1.445	1.519	1.72
AB	2.935	2.336	2.76
BB	1.74	3.51	2.30

From Table 4.3, we see that the heterozygous genotype has highest mean value at *M7* and *M39*, whereas at *M26* the highest mean value contained by the homozygous genotype for parent two and genotype AA has minimum value at *M26*.

4.4 QTL localization and mapping using linear mixed model

Based on the methodology thoroughly discussed in chapter three, the detection/localization of QTLs for grain yield and their effects are presented in the following.

Apart from the QTLs detected by SIM, additional four considerable QTLs were observed in the mixed model approach (Figure 4.5).

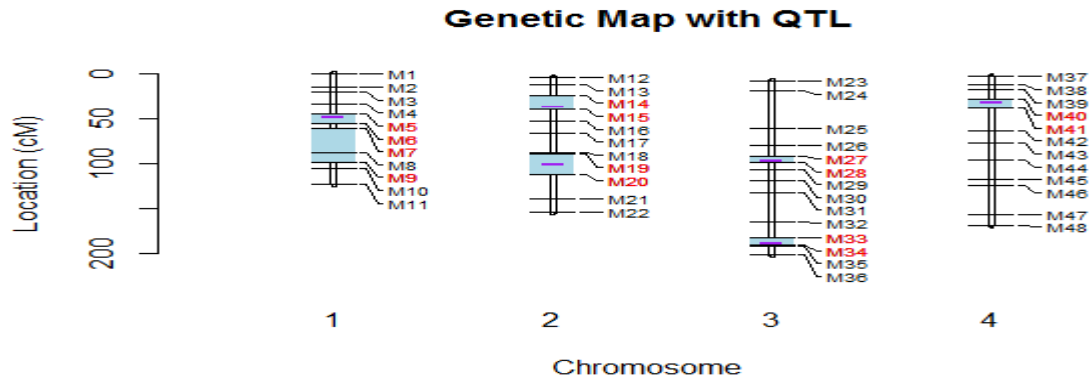


Figure 4.5: Relative positions of QTLs detected by mixed model approach for grain yield in F_2 generation of maize crop.

Referring to Figure 4.5, purple squares indicate the positions estimated for QTL during the localization stage, while flanking markers are indicated in red. Larger rectangles may also be drawn to denote the regions between flanking markers; however these should not be mistaken for QTL support intervals. The above genetic map with its QTLs showed that seven QTLs are localized of which each linkage group has two QTLs, except the fourth chromosome. The genetic map is described more in Table 4.4 below.

Table 4.4: QTL detection using likelihood ratio test

QTLs	Chr.	Pos(cM)	Left M.	Right M.	SE	LR	P-value
<i>loc48</i>	1	48.0	M5	M6	0.17	5.52	<0.0001
<i>M8</i>	1	88.5	M7	M9	0.16	4.63	<0.0001
<i>loc33</i>	2	37.6	M14	M15	0.14	4.49	<0.0001
<i>loc97</i>	2	101.6	M19	M20	0.16	4.2	<0.0001
<i>loc87</i>	3	95.4	M27	M28	0.13	6.85	<0.0001
<i>loc177</i>	3	185.4	M33	M34	0.15	5.77	<0.0001
<i>loc29</i>	4	32.4	M40	M41	0.13	5.81	<0.0001

Referring to Table 4.4, we found two QTLs on the first, second and third chromosomes with the fourth chromosome carrying only one QTL. The two QTLs, *Loc48* and *Marker8* are located on chromosome one at position 48cM and 88.5cM flanked by M5 and M6, and M7 and M9, respectively. In similar manner *loc33* and *loc97*, *loc87* and *loc177* are detected on chromosome two and three flanked by their left and right flanking markers, respectively. Chromosome four has only one QTL in named *loc29* at a position 32.4cM between M40 and M41.

4.5 Leveling of Variables as Random and Fixed

Now, we are seeking to show the effect of each QTL between two flanking markers, and hence the other markers linked with some other QTLs above are likely to have marker effects, which should be considered as random factors in controlling background genetic variation. So, this means that markers on the chromosome with the putative QTL were excluded from the model. The detected QTLs which are listed in the above table (*loc48*, *M8*, *loc33*, *loc97*, *loc87*, *loc177* and *loc29*) are taken as fixed covariates for model 3.17. Thus, the estimates of QTLs do not depend upon a particular fixed set of markers being in the model. Note that, our objective for using mixed model here is to separate the effect of markers being in the model from QTL effects (Zhu and Weir 1998); and to provide important estimates of additive and dominance effects of QTLs. Hence, prediction based on random variables could not be the expected task from our study and do not presented.

Table 4.5: Leveling of markers as random for each QTL fixed covariates.

Flanking Markers	QTL1 (loc48)	QTL2 (M8)	QTL3 (loc33)	QTL4 (loc97)	QTL5 (loc87)	QTL6 (loc177)	QTL7 (loc29)
M5, M6		✓	✓	✓	✓	✓	✓
M7, M9	✓		✓	✓	✓	✓	✓
M14, M15	✓	✓		✓	✓	✓	✓
M19, M20	✓	✓	✓		✓	✓	✓
M27, M28	✓	✓	✓	✓		✓	✓
M33, M34	✓	✓	✓	✓	✓		✓
M40, M41	✓	✓	✓	✓	✓	✓	

Using the above table and model 3.16, we would be able to fit a linear mixed model for each QTL with the corresponding markers as shown below:

4.6 Mixed Model Fitting and Diagnostic Checking

Since we need to fit a linear mixed model for maize datasets, we must first fit the model for each QTL and then check whether they satisfy the basic assumptions of linear mixed model (Appendix B, C and D). Note that, here, we use a new method that used the linear mixed model

General linear mixed model for putative QTLs

$$y_i = \mu + ax_{Ai} + dx_{Di} + \sum_{k \neq M_{i-}, M_{i+}}^m u_{M(k)} z_{M(ki)} + \varepsilon_i$$

$i = 1, 2, \dots, N, k = 1, 2, \dots, m$, where N is the number of observations in the given datasets m is the number of markers included in the model, μ is the general mean, a and d are the additive and dominance effect of the putative QTL with coefficients x_{Ai} and x_{Di} , respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, y_i is the maize grain yield for the i^{th} individual, and ε_i is the random error term. Then, model for each putative QTL is presented as follows.

4.6.1 Fixed effect estimates for QTL one (model one)

Model for QTL one given as

$$\begin{aligned} y_i = & \mu + aloc48_{Ai} + dloc48_{Di} + u_{M(7)} z_{M(7i)} + u_{M(9)} z_{M(9i)} + u_{M(14)} z_{M(14i)} + u_{M(15)} z_{M(15i)} \\ & + u_{M(19)} z_{M(19i)} + u_{M(20)} z_{M(20i)} + u_{M(27)} z_{M(27i)} + u_{M(28)} z_{M(28i)} \\ & + u_{M(33)} z_{M(33i)} + u_{M(34)} z_{M(34i)} + u_{M(40)} z_{M(40i)} + u_{M(41)} z_{M(41i)} + \varepsilon_i \end{aligned}$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given datasets, μ is the general mean, a and d are the additive and dominance effect of the putative QTL *loc48* with coefficients $loc48_{Ai}$ and $loc48_{Di}$, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M5, M6$.

Table 4.6: Additive and Dominance effect estimates for QTL one

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	2.032	0.609	72	3.334	0.0014
<i>loc48_A</i>	-1.05	0.177	72	-5.94	0.0000
<i>loc48_D</i>	0.218	0.169	72	1.29	0.2000

Based on the above table, QTL1 (*loc48*) has additive effect on maize grain yield. Since dominance effect ($p=0.2000$) is not significant at $p=0.05$, QTL one has no dominance effect for maize grain yield. Thus, the average grain yield of maize would be decreased by 1.05 amounts due to substituting allele (*A*) with that of another allele (*B*) at a particular locus *loc48*. Moreover, by holding no dominance and additive effects, the average value of grain yield could be 2.032 tons per hectare at this specific locus.

Additive effect of *loc48* is highly significant ($p<0.0001$) as shown in Table 4.13. Dominance effects, however, were not statistically significant at 5% level of significance.

4.6.2 Fixed effect estimates for QTL two (model two)

We now consider discussion on fixed effect component of model two. We may recall the model as:

$$\begin{aligned}
 y_i = & \mu + aM8_{Ai} + dM8_{Di} + u_{M(5)}Z_{M(5i)} + u_{M(6)}Z_{M(6i)} + u_{M(14)}Z_{M(14i)} + u_{M(15)}Z_{M(15i)} \\
 & + u_{M(19)}Z_{M(19i)} + u_{M(20)}Z_{M(20i)} + u_{M(27)}Z_{M(27i)} + u_{M(28)}Z_{M(28i)} \\
 & + u_{M(33)}Z_{M(33i)} + u_{M(34)}Z_{M(34i)} + u_{M(40)}Z_{M(40i)} + u_{M(41)}Z_{M(41i)} + \varepsilon_i
 \end{aligned}$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given datasets, μ is the general mean, a and d are the additive and dominance effect of the putative QTL *M8* with coefficients $M8_{Ai}$ and $M8_{Di}$, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M7, M9$.

The result is presented in Table 4.7 and discussions follow.

Table 4.7: Additive and Dominance effect estimates for QTL two

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	1.985	0.383	72	5.183	0.000
<i>M8_A</i>	-0.388	0.166	72	-2.34	0.0222
<i>M8_D</i>	-0.594	0.168	72	-3.52	0.0007

As we see in Table 4.7, QTL2 (corresponding to marker 8) has both additive and dominance effects on maize grain yield. Moreover, the average grain yield of maize could be decreased by 0.372 due to the substitution of alleles from one type (*A*) to the other (*B*) at *M8*. In other words, the additive allelic effect (-0.388) is an estimate of the change in the average phenotype that would be produced by substituting a single allele of one type with that of another. On the other hand, dominance effect depicts that the average grain yield is decreased by 0.594 due to the interaction effects of different alleles at the locus *M8*.

The negative dominance effect at locus *M8* on chromosome one indicates that allele *B* is dominant over *A* at this locus. Moreover, both additive and dominance effects at marker 8 are statistically significant at less than 5% level of significances (Table 4.13).

4.6.3 Fixed effect estimates for QTL three (model three)

Model for QTL three

$$\begin{aligned}
 y_i = & \mu + aloc33_{Ai} + dloc33_{Di} + u_{M(5)}Z_{M(5i)} + u_{M(6)}Z_{M(6i)} + u_{M(7)}Z_{M(7i)} + u_{M(9)}Z_{M(9i)} \\
 & + u_{M(19)}Z_{M(19i)} + u_{M(20)}Z_{M(20i)} + u_{M(27)}Z_{M(27i)} + u_{M(28)}Z_{M(28i)} \\
 & + u_{M(33)}Z_{M(33i)} + u_{M(34)}Z_{M(34i)} + u_{M(40)}Z_{M(40i)} + u_{M(41)}Z_{M(41i)} + \varepsilon_i
 \end{aligned}$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given dataset, μ is the general mean, a and d are the additive and dominance effect of the putative QTL *loc33* with coefficients

$loc33_{Ai}$ and $loc33_{Di}$, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M14, M15$.

Table 4.8: Additive and Dominance effect estimates for QTL three

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	2.022	0.280	72	7.21	<0.00001
$loc33_A$	0.0263	0.153	72	0.172	0.9886
$loc33_D$	-0.765	0.140	72	-5.449	<0.00001

Table 4.8 shows that QTL three has significant dominance effect on maize grain yield; and hence the interaction effect of alleles could decrease the average grain yield by the amount of 0.765 tons at $loc33$. Moreover, allele B is dominant once again over allele A at this locus. This QTL has little additive effect which is not considerable at 5% level.

Again, dominance effect of $loc33$ is statistically significant at 1% level of significance for the average change of maize grain yield (Table 4.13).

4.6.4 Fixed effect estimates for QTL four (model four)

We recall that model for QTL four is:

$$\begin{aligned}
y_i = & \mu + aloc97_{Ai} + dloc97_{Di} + u_{M(5)}z_{M(5i)} + u_{M(6)}z_{M(6i)} + u_{M(7)}z_{M(7i)} + u_{M(9)}z_{M(9i)} \\
& + u_{M(14)}z_{M(14i)} + u_{M(15)}z_{M(15i)} + u_{M(27)}z_{M(27i)} + u_{M(28)}z_{M(28i)} \\
& + u_{M(33)}z_{M(33i)} + u_{M(34)}z_{M(34i)} + u_{M(40)}z_{M(40i)} + u_{M(41)}z_{M(41i)} + \varepsilon_i
\end{aligned}$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given datasets, μ is the general mean, a and d are the additive and dominance effect of the putative QTL $loc97$ with coefficients $loc97_{Ai}$ and $loc97_{Di}$, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M19, M20$.

Table 4.9: Additive and Dominance effect estimates for QTL four

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	2.06	0.331	72	6.22	0.0000
<i>loc97_A</i>	0.474	0.209	72	2.269	0.0263
<i>loc97_D</i>	-0.270	0.1997	72	-1.355	0.1796

As we can see in Table 4.9, there is only additive effect at quantitative trait locus *loc97* with an effect of 0.474 changes on average grain yield. The estimated value indicates that, the average grain yield in maize increased by 0.474 tons due to substituting different alleles at *loc97*. Moreover, additive effect of *loc97* is statistically significant at 5% level of significance in (Table 4.13).

4.6.5 Fixed effect estimates for QTL five (model five)

Model for QTL five

$$\begin{aligned}
y_i = & \mu + \mathbf{a}loc87_{Ai} + \mathbf{d}loc87_{Di} + u_{M(5)}z_{M(5i)} + u_{M(6)}z_{M(6i)} + u_{M(7)}z_{M(7i)} + u_{M(9)}z_{M(9i)} \\
& + u_{M(14)}z_{M(14i)} + u_{M(15)}z_{M(15i)} + u_{M(19)}z_{M(19i)} + u_{M(20)}z_{M(20i)} \\
& + u_{M(33)}z_{M(33i)} + u_{M(34)}z_{M(34i)} + u_{M(40)}z_{M(40i)} + u_{M(41)}z_{M(41i)} + \varepsilon_i
\end{aligned}$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given datasets, μ is the general mean, $\mathbf{a}$ and $\mathbf{d}$ are the additive and dominance effect of the putative QTL *loc87* with coefficients *loc87_{Ai}* and *loc87_{Di}*, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M27, M28$.

Table 4.10: Additive and Dominance effect estimates for model five

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	2.045	0.4491	72	4.553	0.000

<i>loc87_A</i>	-0.593	0.1611	72	-3.68	0.0004
<i>loc87_D</i>	-0.0186	0.1614	72	-0.115	0.9084

Table 4.10 shows that, the average grain yield of maize could be decreased by 0.593 tons amounts at *loc87* due to substituting a single allele of one type (*A*) with that of the other (*B*). In other words, the fourth QTL only has additive effect on the grain yield, and is significant at 5 %.

4.6.6 Fixed effect estimates for QTL six (model six)

Model for QTL six is formulated as

$$\begin{aligned}
y_i = & \mu + aloc177_{Ai} + dloc177_{Di} + u_{M(5)}Z_{M(5i)} + u_{M(6)}Z_{M(6i)} + u_{M(7)}Z_{M(7i)} + u_{M(9)}Z_{M(9i)} \\
& + u_{M(14)}Z_{M(14i)} + u_{M(15)}Z_{M(15i)} + u_{M(19)}Z_{M(19i)} + u_{M(20)}Z_{M(20i)} \\
& + u_{M(27)}Z_{M(27i)} + u_{M(28)}Z_{M(28i)} + u_{M(40)}Z_{M(40i)} + u_{M(41)}Z_{M(41i)} + \varepsilon_i
\end{aligned}$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given datasets, μ is the general mean, a and d are the additive and dominance effect of the putative QTL *loc177* with coefficients *loc177_{Ai}* and *loc177_{Di}*, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M33, M34$.

Table 4.11: Additive and Dominance effect estimates for QTL six

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	2.068	0.321	73	6.427	0.000
<i>loc177_A</i>	-0.307	0.2243	73	-1.37	0.1747
<i>loc177_D</i>	-0.436	0.6198	73	-0.704	0.4837

Results in Table 4.11 and Table 4.13 shows that both additive and dominance effects for QTL six did not show significant effect ($p = 0.05$), hence we conclude that *loc177* has no effect on the

grain yield of maize. As a result, we do not have considerable statistical evidence at locus *loc177* in the average changes of the given phenotypic value.

4.6.7 Fixed effect estimates for QTL seven (model seven)

Finally, model for QTL seven is given as:

$$y_i = \mu + aloc29_{Ai} + dloc29_{Di} + u_{M(5)}z_{M(5i)} + u_{M(6)}z_{M(6i)} + u_{M(7)}z_{M(7i)} + u_{M(9)}z_{M(9i)} \\ + u_{M(14)}z_{M(14i)} + u_{M(15)}z_{M(15i)} + u_{M(19)}z_{M(19i)} + u_{M(20)}z_{M(20i)} \\ + u_{M(27)}z_{M(27i)} + u_{M(28)}z_{M(28i)} + u_{M(33)}z_{M(33i)} + u_{M(34)}z_{M(34i)} + \varepsilon_i$$

$i = 1, 2, \dots, N$, where N is the number of observations in the given datasets, μ is the general mean, a and d are the additive and dominance effect of the putative QTL *loc29* with coefficients $loc29_{Ai}$ and $loc29_{Di}$, respectively. $u_{M(k)}$ is the random effect of marker k with its coefficients $z_{M(ki)}$, and $k \neq M40, M41$.

Table 4.12: Additive and Dominance effect estimates for QTL seven

Effects	Estimate	Standard error	DF	t value	Pr> t
Intercept	2.146	0.229	71	9.334	0.0001
<i>loc29_A</i>	-0.813	0.198	71	-4.104	0.0001
<i>loc29_D</i>	0.417	0.187	71	2.225	0.0293

Similarly, QTL seven (*loc29*) which is located in chromosome four has both significant additive and dominance effects on grain yield ($p < 0.05$). By holding the additive and dominance effects zero at *loc29*, the average value of grain yield becomes 2.146 tons per hectare.

When a single allele of one type (*A*) is substituted with that of the second type (*B*), the average grain yield of maize would be decreased by 0.813 tons at *loc29* (Table 4.12). This additive allele effect is an estimate of the change in the average grain yield that would be produced by substituting a single allele of one type with that of another type.

4.6.8 Fixed effect tests for the detected QTLs

Table 4.13: significant test of additive and dominance effects for detected QTLs

QTL	Num DF	Den DF	F value		p-value	
			Additive	Dominance	Additive	Dominance
<i>Loc48</i>	1	72	44.97	1.67	<0.0001	0.20
<i>M8</i>	1	72	34.77	12.41	<0.0001	0.0004
<i>Loc33</i>	1	72	0.06077	29.69	0.806	<0.001
<i>Loc97</i>	1	72	4.37	1.836	0.006	0.1796
<i>Loc87</i>	1	72	13.55	0.0133	0.0004	0.9084
<i>Loc177</i>	1	73	1.8788	0.0896	0.1747	0.285
<i>Loc29</i>	1	71	11.698	4.95	0.0009	0.0293

Table 4.13 depicts that three QTLs (*loc48*, *loc97*, and *loc87*) have significant additive effect effects associated with grain yield in maize, and two loci (*M8* and *loc29*) located on chromosome 1 and 4, respectively, have both significant additive and dominance effects at 5% level of significance. On the other hand, *loc33* which is identified on chromosome two has significant dominance effect for grain yield. On the contrary, we have no considerable statistical evidence at *loc177* on chromosome three for significance of either effect.

Thus, TGYH could decrease at *loc48*, *M8*, *loc87* and *loc29*, due to negative additive effects; and it could be increase at *loc97* due to positive additive effect. Moreover, TGYH could be decreased at *loc33* due to negative dominance effect.

4.6.9 Gene action for the detected QTLs

The degree of dominance for detected QTLs was estimated as $|d/a|$. As a result, we have summarized level of dominance and gene action for each QTL in Table 4.21 below.

Table 4.14: Level of dominance and gene action at detected QTLs

QTL	Chromosomal Name	Position(cM)	<i>a</i>	<i>d</i>	LD	Gene action
<i>Loc48</i>	1	48	-1.05	0.218	0.20	<i>A</i>
<i>M8</i>	1	88.5	-0.388	-0.594	1.53	<i>OD</i>
<i>Loc33</i>	2	37.6	-0.0263	-0.818	29.087	<i>OD</i>
<i>Loc97</i>	2	101.6	0.474	-0.270	0.569	<i>PD</i>
<i>Loc87</i>	3	95.4	-0.593	-0.0186	0.031	<i>A</i>
<i>Loc177</i>	3	185.4	-0.307	-0.436	1.42	<i>OD</i>
<i>Loc29</i>	4	32.4	-0.813	0.4175	0.513	<i>PD</i>

Levels of dominance were estimated as $LD = |d|/|a|$ ratio for each QTL, and the average level of dominance (ALD) was computed weighing each LD ratio. Following the studies (Sibov *et al.* 2003b; and others), gene action was characterized as follows: additive (*A*) $0.0 \leq LD \leq 0.20$; partial dominance (*PD*) $0.21 \leq LD \leq 0.80$; dominance (*D*) $0.81 \leq LD \leq 1.20$; and overdominance (*OD*) $LD > 1.21$. Therefore, Table 4.21 revealed that three QTLs (*M8*, *loc33*, and *loc177*) displayed overdominance effects, and two QTLs (*loc48* and *loc87*) displayed additive effects for decreasing grain yield in maize. On the other hand, the remaining two QTLs in named *loc97* and *loc29* on chromosome 2 and 4 showed partial dominance gene action for increasing and decreasing grain yield in maize, respectively. Hence, as the ALD value is (4.764) this result suggests that overdominance could be the average level of dominance for grain yield in maize.

4.7 Genetic variation explained by detected QTL

Table 4.15: Log likelihood, Likelihood Ratio and variance explained by significance QTLs

Models	LogLik	L. Ratio	R^2
<i>Null model</i>	-135.4216		
<i>QTL 1 (loc48)</i>	-128.9973	12.8486	0.1555
<i>QTL 2 (M8)</i>	-132.8785	5.08632	0.0647
<i>QTL 3 (loc33)</i>	-129.5444	11.7544	0.1432
<i>QTL 4 (loc97)</i>	-132.8040	5.235355	0.0665
<i>QTL 5 (loc87)</i>	-131.3432	8.15686	0.1017
<i>QTL 7 (loc29)</i>	-130.7599	9.3234	0.1154

The proportion of phenotypic variation explained by each of these QTLs varied from 6.65% (*QTL 2*) to 15.55% (*QTL 1*). More than 60.0% of total variation in yield was explained by the QTLs for the given dataset. *QTL 1* and *QTL 3* which are found in chromosome one and two, respectively, explain more variation for the trait of interest relative to the remaining quantitative trait loci (Table 4.15).

The fitted model was checked for possible presence of outliers and influential values and also for homoscedasticity and normality of the residuals. The diagnostic test results for detection of outliers and influential values, test of homoscedasticity and normality are presented in Appendix C and D. Thus, from the goodness of fit test and diagnostic test results presented in the Appendix B, C and D, it is possible to say that our model is adequate.

4.8 DISCUSSION

This study was intended to detect QTLs on the grain yield based on Ethiopian maize data. Accordingly, identification of QTLs was performed by two statistical methods; interval mapping

and mixed model approach techniques were employed. The walking speed chosen for the given trait was 1 cM in both methods.

The statistical model of interval mapping (IM) proposed by Lander and Botstein (1989) is generally a normal mixture model, as the genotypes of the quantitative trait locus (QTL) are not observable and needed to be inferred from its flanking markers. In the parameter estimation of the normal mixture model, the maximum likelihood estimation was implemented to obtain the maximum likelihood estimates (MLE) through the EM algorithm (Dempster *et al.* 1977) by treating the model as an incomplete data problem. Typically, the presence of a QTL, i.e. the null hypothesis of no QTL, were tested over the all possible positions in the whole genome by using likelihood ratio test (LRT) statistics/LOD score and the position with the significantly maximum LRT statistic is regarded as the estimated QTL location. Thus, for interval mapping, the position at which the LOD score curve reaches 2 and above it was used as the estimate of the QTL location. The percentage of phenotypic variance explained by a QTL was estimated as heritability using Karl W. Broman methods (Karl, W. *et al.* 2009).

Under mixed model approach, the identification of QTLs was performed in two stages, localization and detection stages (Zhu and Weir, 1998). Likelihood ratio test statistic using recombination fractions were used for localization stage; and finally the linear mixed model were applied to estimate the additive and dominance effects; and for testing the significances of QTLs. The percentage of phenotypic variance explained by a QTL was estimated using R^2 statistic based on LR (Cox and Snell, 1989; Magee, 1990). Thus, the results obtained from this study are discussed as follows.

Although many studies on QTL identification for grain yield were reported in maize (Stuber *et al.* 1992; Nagabhushana K. *et al.* 2006; Lu *et al.*, 2006; Zhang *et al.* 2010; Sibov *et al.* (2003b); Milena *et al.* 2005), it is not easy to compare previous results with ours, because these differences, including environmental design, genetic maps, markers used and statistical methods, will make comparison difficulties. Therefore, the comparisons here were focused on the selections of segregation population and ecological conditions.

Using interval mapping methods, three QTLs (named *loc58*, *loc70* and *M39*) were identified on chromosome one, three and four associated with grain yield in maize, explaining from 3.55 % to

16.91% of the phenotypic variance. The LOD thresholds values for these QTLs were then 6.88, 5.01, and 2.34, which correspond to LR thresholds of 31.68, 23.07, and 10.776. Thus, the LOD threshold values used in this study were related with the LOD threshold values of 2, 2.5, and 3 commonly used for QTL mapping in maize (Milena *et al.* 2005). In addition, the level of dominance ranged from additive ($LD = 0.159$) to overdominance ($LD = 9.053$), and overdominance ($ALD = 3.955$) was the average level of dominance. Although only three QTLs were identified on three chromosomes for grain yield, the average level of dominance coincides with Milena *et al.* (2005). Parental lines, segregation population or genetic map (including number of chromosomes, type of markers, and number of markers) can bring about different results in QTL number, position or effect, variation explained by QTL. Tropical maize populations were used for both studies but, Milena *et al.* (2005) were used $F_{2:3}$ families evaluated in five environments and a genetic map with 139 microsatellites markers; while in our study F_2 families in one environment and SNP markers are used. On the other study (Zhang *et al.* 2010) six QTLs associated grain yield was detected on four chromosomes, and explained 39.03% of the phenotypic variance.

Even though it was too difficult to find previous studies on grain yield using mixed model mapping approach, we have discussed briefly our findings as follows. Under localization stage, seven QTLs corresponds to LR values ranges from 4.2 (*loc97*) to 6.85 (*loc87*) were detected associated with grain yield in maize; but only six of them have significant effects. Although overdominance could be the average level of dominance for grain yield in maize; our findings revealed that overdominance for three QTLs, additive for two QTLs, and partial dominance for two QTLs. Moreover, the variation in grain yield explained by significant QTL were ranges from 6.47% (*M8*) to 15.55% (*loc48*) which are located on chromosome one.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The study revealed that, interval mapping detected three QTLs located on chromosome 1 (*loc58*), 3 (*loc70*) and 4 (*M39*); and the positions of these QTLs were 58.0cM, 78.4 cM and 17.2cM, respectively. The LR thresholds values for QTL mapping were then 31.68, 23.07 and 10.77, which correspond to LOD thresholds of 6.88, 5.01 and 2.34.

Accordingly, the two QTLs identified on chromosome one and four could explain 20.46% of phenotypic variance, and could decrease TGYH from 0.1483 to 0.287 tons, due to negative additive effects. In other words, when a single allele of parent one is substituted with that of the second parent at loci *loc58* and *M39*, the average grain yield in maize could decrease by 0.1483 and 0.287 tons per hectare, respectively. Whereas, one QTL located on chromosome 3 could account for 14.69 % of phenotypic variance, and due to positive additive effect, it could increase TGYH.

On the other hand, the dominance effects at these detected loci revealed that, the grain yield in maize could be increase due to the interaction between alleles at *loc58* and *M39*; and it could be decreased at *loc70*. In addition, the level of dominance for each locus was 0.159 (*loc70*), 2.635 (*M39*) and 9.05 (*loc58*); which depicts that the gene action were overdominance at *loc58* and *M39* and additive at *loc70*. Moreover, the average level of dominance was 3.948 indicating overdominance gene actions, and 35.15% of variations for grain yield in maize were explained by these detected QTLs.

The findings of this study also revealed that, seven QTLs were identified in the localization stage based on mixed linear model approach. As a result, two QTLs were localized from each chromosome, except chromosome four has only one locus. But, only six of them have significant effects at 5% level of significance for the grain yield in maize.

Four QTLs, *loc48* at 48.0cM, *loc97* at 101.6cM, *loc87* at 95.4cM and *loc29* at 32.4cM positions has significant additive effect on the grain yield at 5% level of significance. The average maize

grain yield could be decreased at locus *loc48* and *loc87* due to negative additive effects; and increase at locus *loc97* due to positive additive effect. On the other hand, one QTL (*loc33*) at 37.6cM position has significance dominance effect for grain yield in maize; and then TGYH could decrease at this locus.

Moreover, *M8* and *loc29* located on chromosome one and 4, respectively, at 88.5cM and 32.4 positions have both additive and dominance significant effects on the maize grain yield. These result revealed that by holding zero dominance effects at *M8* and *loc29*, the average grain yield for maize could be decreased due to negative additive effects. Although the average level of dominance in this model displayed overdominance effect, two QTLs (*loc48* and *loc87*) showed additive, two QTLs (*loc97* and *loc29*) displayed partial dominance effect gene action. The variations in grain yield explained by significant QTL were ranges from 6.47% (*M8*) to 15.55% (*loc48*).

Furthermore, in our study, two QTLs, *M39* and *M8* using interval mapping and mixed model techniques, respectively, were quite linked to their linked markers, with only 0 cM apart, which suggested that the linked markers could probably be co-segregated with the loci controlling grain yield, and thus, could be considered for marker-assisted selection (MAS) in maize breeding project.

In summary, three QTLs and seven QTLs were detected using interval mapping and mixed model approach, respectively. Hence large variations for the given phenotypic trait explained via mixed linear model method. In both methods our findings showed that the gene action were overdominance; which depicts the heterozygote is superior than either homozygote (the performance of heterozygous plants is more extreme than that of both homozygous). The results also suggested that, although the average gene action was displayed overdominance in both methods, mixed model approach can detect more number of QTLs than interval mapping by considering two stages, localization and detection stages.

5.2 RECOMMENDATION

- Recent development in molecular marker technology is expected to enable greater power in detection of QTL using recent techniques for agronomically important traits and

utilization of QTL information for crop improvement. But this is not practical in this country due to lack of available data and limited resources. Thus, molecular geneticists should try to solve this challenge collaborating with molecular breeders of other countries.

- Detecting QTLs for desirable traits is only a first step and much further work is necessary to truly dissect quantitative variation. Fine mapping of the QTL is very important if the QTL information will be used in applied research. Thus, statisticians/biometricians should work with molecular geneticists to show the practical and theoretical applications of statistics on marker data analysis in this country.
- As the result of this study, some of the QTLs were far away from their linked markers, with a range from 1.2 to 27.9 cM apart, for this, other molecular markers should be added to the given regions for mapping these QTLs more finely.
- QTL mapping using mixed model approach is more appropriate than interval mapping to detect more number of QTLs and for testing the significances of QTLs for the given traits.
- For further studies, because of the high genotype by environment interaction in tropical areas, this type of research should focus not only on mapping QTL but, also, on QTL X E and QTL X QTL interactions.

REFERENCES

- Asins, M. (2002), Present and future of quantitative trait locus analysis in plant breeding. *Plant Breed.* **121**: 281-291.
- Austin DF, Lee, M (1996), Comparative mapping in F2:3 and F6:7 generations of quantitative trait loci for grain yield and yield components in maize, **92**:817-826.
- Baird, V., A. Abbott, R. Ballard, B. Sosinski & S. Rajapakse (1997), DNA Diagnostics in Horticulture, p. 111–130: P. Gresshoff (Ed.), *Current Topics in Plant Molecular Biology*.
- Basten, C.J., B.S. Weir & Z.-B. Zeng (1994), Zmap-a QTL cartographer. In: J.S.G.C. Smith, B.J. Benkel, W.F. Chesnais, J.P. Gibson, B.W. Kennedy & E.B. Burnside (Eds.), *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production*.
- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics; Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.
- Beyene Y, Botha AM, Myburg AA. (2005), Phenotypic diversity for morphological and agronomic traits in traditional Ethiopian highland maize accessions.
- Brady, T., West Kathleen, B., Welch Andrzej, T., Galecki (2007), *Linear mixed models, a practical guide using statistical softwares*.
- Christov NK, Todorovska EG, Fasoula DA, Ioannides IM, Atanassov AI, Hristov KN (2004), Molecular characterization of chemical mutagenesis induced diversity in elite maize germplasm. *Genetika* **36(1)**: 47-60.
- Collard, B.C.Y., M.Z.Z. Jahufer, J.B. Brouwer & E.C.K. Pang (2005), An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* **142**: 169–196.
- Comstock, R. E. (1978), Quantitative genetics in maize breeding. pp 191-206. In *maize breeding and genetics*, New York.
- Cook, R.D. (1977), “Detection of Influential Observations in Linear Regression,” *Technometrics*, **19**, 15–18.
- Cox DR, Snell EJ. (1999), *Analysis of Binary Data*, 2nd edn. Chapman and Hall: London.
- Dempster, A. P., N. M. Laird and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39 (1)**: 1–38.

- ECSA (2003), Ethiopian agricultural sample enumeration, 2001/02. Statistical report on socio-economic characteristics of the population in agricultural households, land use, and area and production of crops. Results at country level. Part I. CSA, Addis Ababa, Ethiopia.
- ECSA. (2012), Agricultural Sample survey: report on area and production of major crops (private peasant holdings, Meher season). Statistical Bulletin, (1). Addis Ababa.
- Eagles, H., H. Bariana, F. Ogbonnaya, G. Rebetzke, G. Hollamby, R. Henry, P. Henschke & M. Carter (2001), Implementation of markers in Australian wheat breeding. *Aust J Agric Res* **52**: 1349–1356.
- East, E. M. (1916), Studies on size inheritance in *Nicotiana*. *Genetics* **1**: 164-176
- Falconer, D.S. and Mackay, T.F.C. (1996), Introduction to quantitative genetics. 4th edition, Longman, New York, P.132-133.
- FAO (Food and Agriculture Organization of the United Nations). (1992), Maize in human nutrition. FAO, Rome, Italy.
- Haldane, J.B.S. (1919), “The combination of linkage values, and the calculation of distances between the loci of linked factors.” *Journal of Genetics* **8**, 299–309.
- Hallauer AR, Miranda JBF. (1988), Quantitative Genetics in Maize Breeding 2nd edn. Iowa State University Press, Ames
- Jansen, R. (1993), Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.
- Jones, N., H. Ougham & H. Thomas (1997), “Markers and mapping”: We are all geneticists now. *New Phytol* **137**: 165–177.
- Jordan SA, Humphries P (1994), Single nucleotide polymorphism in exon 2 of the BCP gene on 7q31-q35. *Hum Mol Genet* **3**:1915
- Kao, C.-H., Z.-B. Zeng and R. Teasdale (1999), Multiple intervals mapping for quantitative trait loci. *Genetics* **152**: 1023-1216.
- Karl, W. Broman. Saunak Sen (2009), Statistics for Biology and Health. A guide to QTL mapping with R/qtl.
- Kennedy BW, Quinton M, van Arendonk JA. (1992), Estimation of effects of single genes on quantitative traits.

- Knott, S. A., C. S. Haley (1992), Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genet. Res.* **60**: 139-151.
- Kumar, P., Gupta, V.K., Misra, A.K., Modi, D.R and Pandey, B.K. (2009), Potential molecular markers in plant biotechnology. *Plant Omics Journal*, **2**:141-162.
- Lander, E. & D. Botstein (1989), Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**:185–199.
- Lander, S. P. GREENJ, ABRAHAMSON, A.B ARILOW, M. J. DALY (1987), Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**: 174-181.
- Littell, R.C, Milliken, G.A., Stroup, W.W., Wolfinger, R.D. and Schabenberger, O. (2006), *SAS For Mixed Models*, North Carolina: SAS Institute Inc.
- Little, R. J. A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*. New York: John Wiley.
- Liu, B-H. (1998), *Statistical Genomics: Linkage, Mapping and QTL Analysis* CRC Press, Boca Raton.
- López, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J. and Verdier, V. (2005), Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). *Theor. Appl. Genet.* **110**:425-431.
- Luo, L. J., Z.-K. Li, H. W. Mei, Q. Y. Shu, R. Tabien (2001), Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. *Genetics* **158**: 1755–1771.
- Lu, H., Romero-Severson, J. and Bernardo, R. (2003), Genetic basis of heterosis explored by simple sequence repeat markers in a random-mated maize population. *Theor.ppl. Genet.* **107**: 494 -502.
- Magee, L. (1990), R2 measures based on Wald and likelihood ratio joint significance tests. *Am Stat* **44**: 250–253.
- Manikanda, N. Boopathi (2012), *Genetic Mapping and Marker Assisted Selection Basics, Practice and Benefits*.
- McDonald, AH. Nicol, JM (2005), Nematode Parasites of Cereals. pp. 131-191.In: Luc, M., Sikora, R.A. and Bridge, J. (Eds). *Plant parasitic nematodes in subtropical and tropical agriculture*, 2nd edition. Wallingford, UK, CABI Publishing.

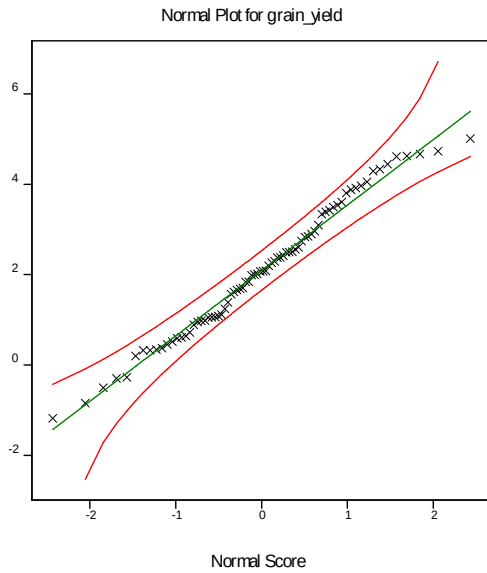
- Michelmore, R. (1995), Molecular approaches to manipulation of disease resistance genes. *Annu Rev Phytopathol* **33**: 393–427.
- Milena de Luna Alves Lima, Claudio Lopes de Souza Jr., Dyeme Antonio Vieira Bento, Anete Pereira de Souza and Luciana Aparecida Carlini-Garcia (2005), Mapping QTL for grain yield and plant traits in a tropical maize population.
- Nagabhushana K., Mane S.P. and Shailaja Hittalmani (2006), Comparative studies on QTL mapping by simple interval mapping and composite interval mapping models for selected growth and yield traits in rice (*Oryza sativa* L.), India.
- Nagaraja Reddy. R., Madhusudhana. S., Murali Mohan. D. V. N., Chakravarthi. S. P. Mehtre. N. Seetharama. J. V. Patil. (2013), Mapping QTL for grain yield and other agronomic traits in post-rainy sorghum [*Sorghum bicolor* (L.) Moench].
- Paterson, A.H. (1996a), Making genetic maps. In: A.H. Paterson (Ed.), *Genome Mapping in Plants*, pp. 23–39. R. G. Landes Company, San Diego, California; Academic Press, Austin, Texas.
- Patterson, H.D. and Thompson, R. (1971), “Recovery of inter-block information when block sizes are unequal,” *Biometrika* **58**, 545-554.
- Ribaut, J.-M., X. Hu, D. Hoisington & D. Gonzalez-De-Leon (1997), Use of STSs and SSRs as rapid and reliable pre selection tools in marker-assisted selection backcross scheme. *Plant Mol Biol Report* **15**: 156–164.
- Sabadin, P. K., Souza, Jr. C. L., Souza, A. P. and Garcia, A. A. F. (2008), QTL mapping for yield components in a tropical maize population using microsatellite markers.- *Hereditas* **145**: 194-203.
- Satterthwaite, F.E. (1946), An approximate distribution of estimates of variance components. *Biometricals Bull.* **2**:110-114
- Sax, K. 1923. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Journal of Theoretical Biology* **117**: 1-10.
- Schabenberger, O. (2004), “Mixed Model Influence Diagnostics,” *SUGI29-Statistics and Data Analysis*, Paper 189-29.
- Shappley, Zachary.W, Johnie N. Jenkins, Jun Zhu, and Jack C. McCarty, Jr. (1998), Quantitative Trait Loci Associated with Agronomic and Fiber Traits of Upland Cotton: *The Journal of Cotton Science* **2**:153-163.

- Sibov, S. T., de Souza Jr, C. L., Garcia, A. A. F. et al. (2003a), Molecular mapping in tropical maize (*Zea mays* L.) using microsatellites markers. 1. Map construction and localization of loci showing distorted segregation. *Hereditas* **139**: 96-106.
- Sibov, T. S., de Souza, Jr, C. L., Garcia, A. A. F. et al. (2003b), Molecular mapping in tropical maize (*Zea mays* L.) using microsatellite markers. 2. Quantitative trait loci (QTL) for grain yield, plant height, ear height and grain moisture. *Hereditas* **139**: 107-115.
- Smith JSC, Smith OS (1992), Fingerprinting crop varieties. *Adv. Agron.* **47**: 85-140.
- Stuber, C. W., M. D. Edwards, J. F. Wendel (1987), Molecular marker facilitated investigations of quantitative trait loci in maize II. Factors influencing yield and its component traits. *Crop Science* **27**: 639-648.
- Stuber, C. W., S. E. Lincoln, D. W. Woff, T. Helentjaris, and E. S. Lander (1992), Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular marker. *Genetics* **132**: 832-839.
- Suzuki, D., Griffiths, A. and Lewontin, R. (1981), An introduction to genetic analysis. W.H. Freeman and Company. San Francisco, *USA*.
- Tanksley, S.D. (1993), Mapping polygenes. *Annu Rev Genet* **27**: 205– 233.
- Thoday, J. M. (1961), Location of polygenes. *Nature* **191**: 368-379.
- Tomlekova NB (2010), Induced mutagenesis for crop improvement in Bulgaria. *Plant Mutation Report* **2(2)**: 4-27.
- Twumasi-Afriyie S, Habatmu Z, Kassa Z, Bayisa Y, Sewagegene T (2001), “Development and improvement of highland maize in Ethiopia. 2nd National maize workshop of Ethiopia”, Addis Ababa, Ethiopia.
- Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, Berlin.
- Weeden, N., G. Timmerman & J. Lu (1994), Identifying and mapping genes of economic significance. *Euphytica* **73**: 191–198.
- Winter, P. and G. Kahl (1995), Molecular marker technologies for plant improvement. *World Journal of Microbiology & Biotechnology* **11**: 438–448.

- Xiao, J., J. Li, L. Yuan and S. D. Tanksley (1995), Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. *Genetics* **140**: 745–754.
- Young, N.D. (1994), Constructing a plant genetic linkage map with DNA markers, p. 39–57, In: I. K.V. Ronald & L. Phillips (Eds.), *DNA-based markers in plants*.
- Young, N.D. (1996), QTL mapping and quantitative disease resistance in plants.
- Yu, S. B., J. X. Li, C. G. Xu, Y. F. Tan, Y. J. (1997), Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. USA* **94**: 9226–9231.
- Zeng, Z.-B. (1993), Theoretical Basis Of Separation Of Multiple Linked Gene Effects On Mapping Quantitative Trait Loci. *Proc. Natl. Acad. Sci. USA* **90**: 10972-10976.
- Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics*. **136**: 1457–1468.
- Zheng Zu-Ping Xiao-Hong Liu¹, Su-Lan He (2010), Identification of the QTLs for grain yield using RIL population under different nitrogen regimes in maize.
- Zhu J. (1998), Mixed model approaches for mapping complex quantitative trait loci, pp. 11-20, *China*.
- Zhu J. (1999), Mixed model approaches of mapping genes for complex quantitative traits. *Journal of Zhejiang University (Natural Science)*, **33(3)**: 327-335
- Zhu J. and B. S. Weir (1998), Mixed model approaches for genetic analysis of quantitative traits. pp. 321-330, *Singapore*.

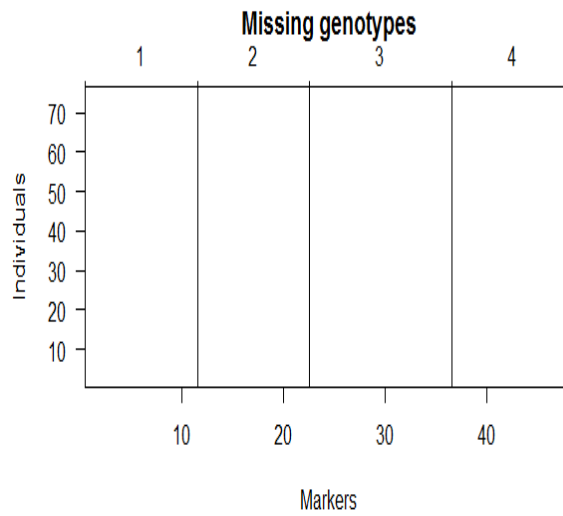
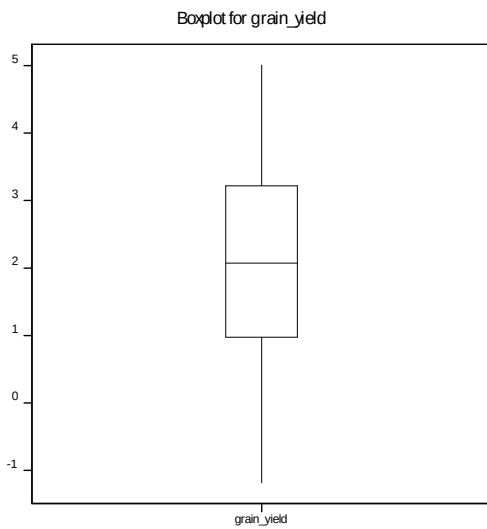
APPENDICES

Appendix A: Diagnostic plots for interval mapping



Shapiro-Wilk test for Normality

Data variate: grain_yield
Test statistic W: 0.9825
Probability: 0.381



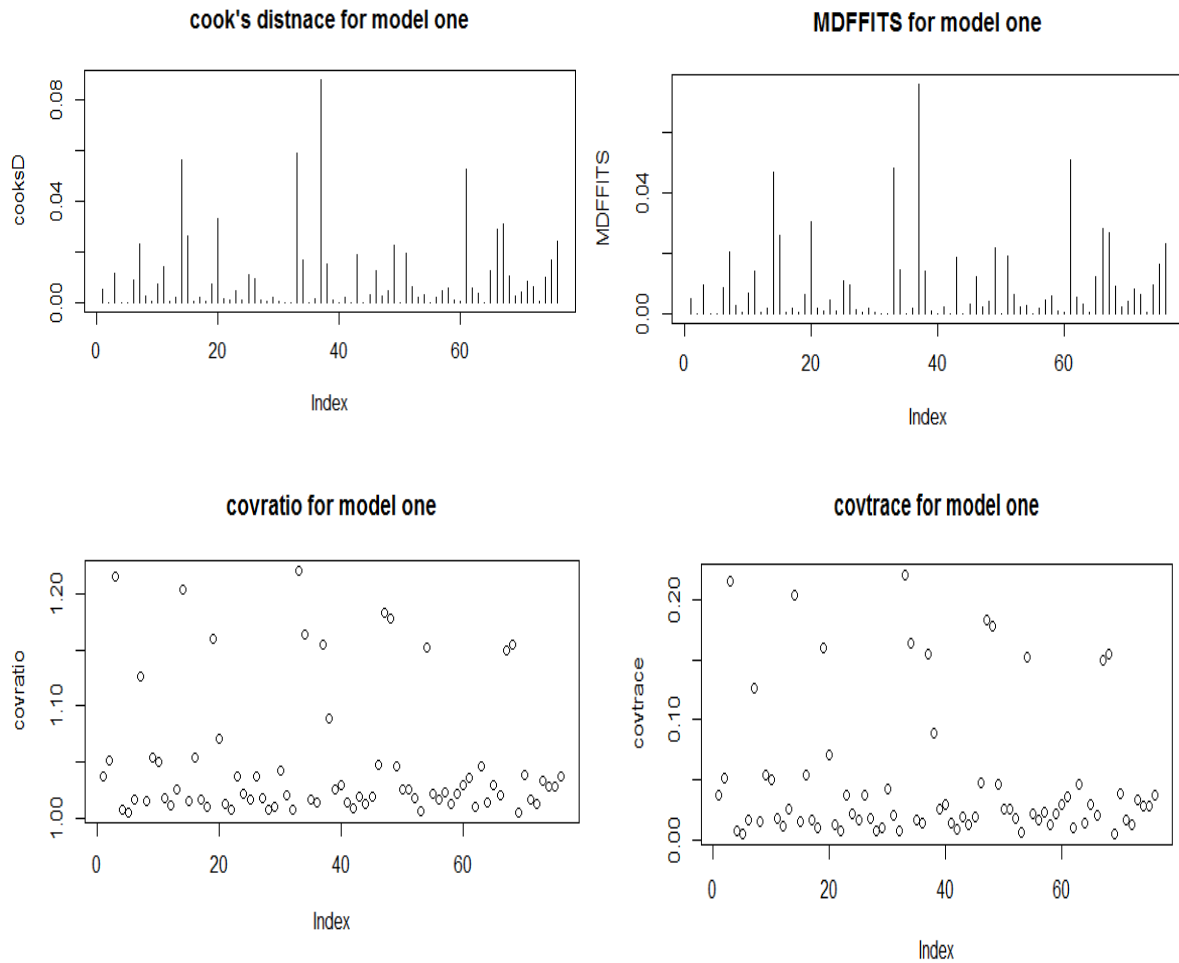
Appendix B: Goodness of fit test for mixed model

Goodness-of-fit test for significance QTLs

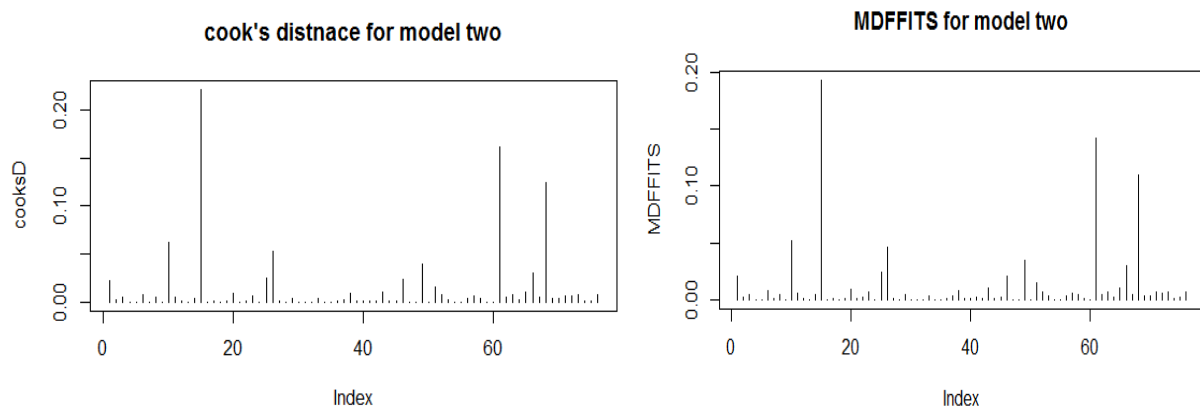
QTLs	Source of variation	DF	Sum of Squares	Mean Square	F value	Pr>F
QTL 1	Model	3	46.910	15.636	9.660	<0.0001
	Error	72	116.54	1.6186		
	Corrected Total	75	163.45	18.753		
QTL 2	Model	3	56.2595	1.488	12.60	<0.0001
	Error	72	107.192			
	Corrected Total	75	163.4516			
QTL 3	Model	3	49.337	16.4458	10.38	<0.0001
	Error	72	114.114	1.5849		
	Corrected Total	75	163.4516			
QTL 4	Model	3	25.4503	8.483	4.43	0.0063
	Error	72	138.001	1.9166		
	Corrected Total	75	163.4516			
QTL 5	Model	3	24.4523	8.1507	4.22	0.0083
	Error	72	138.999	1.9305		
	Corrected Total	75	163.4516			
QTL 7	Model	3	28.0933	9.3644	4.98	0.0034
	Error	72	135.358	1.8799		
	Corrected Total	75	163.4516			

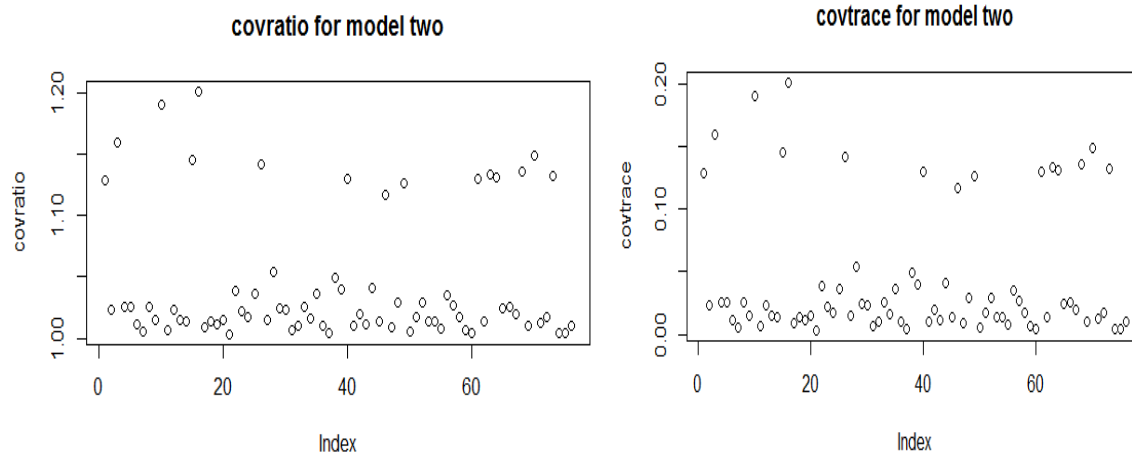
Appendix C: Influence Analysis

C.1 Influence analysis for model one

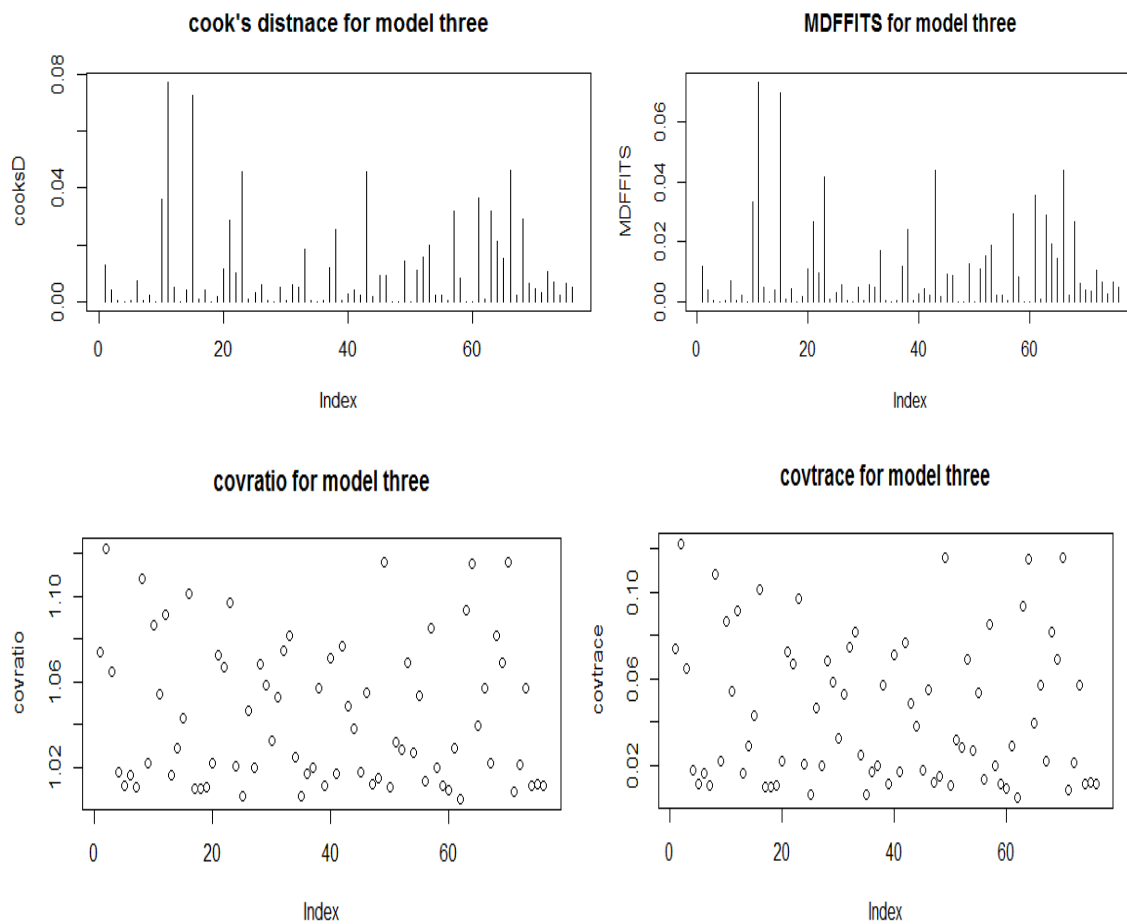


C.2 Influence analysis for model two

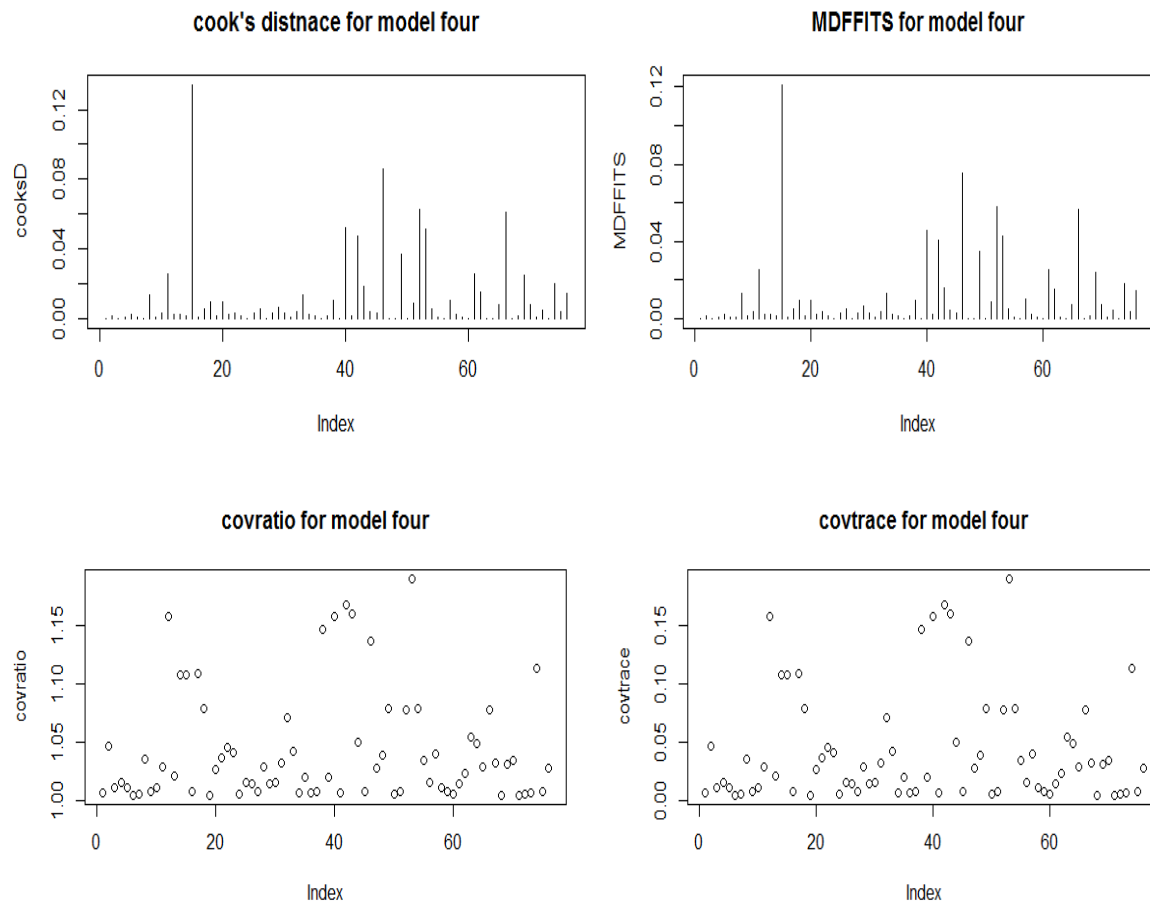




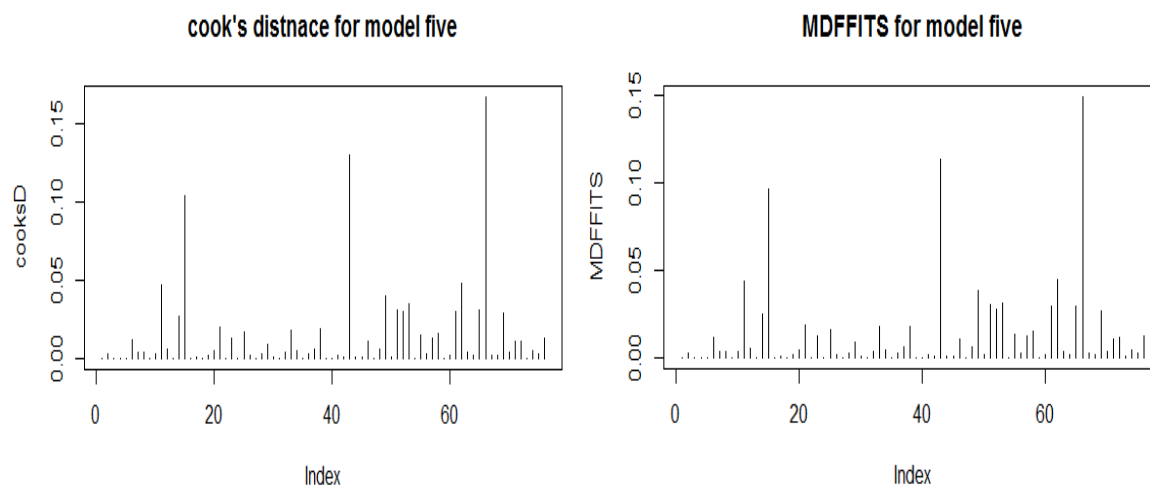
C.3 Influence analysis for model three

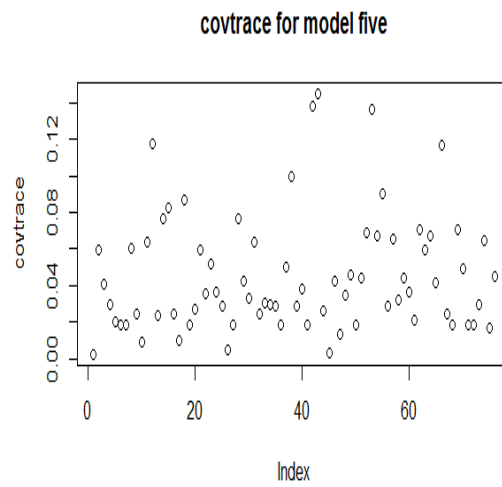
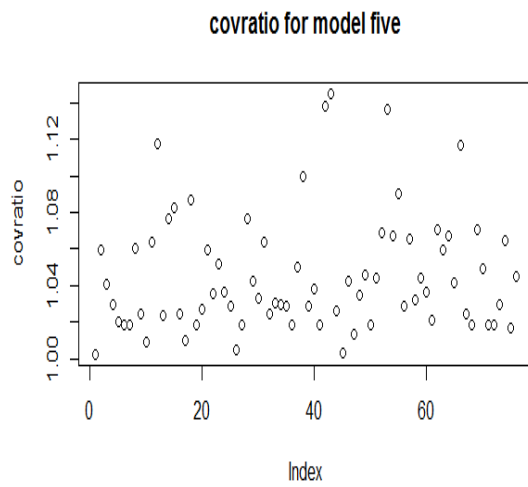


C.4 Influence analysis for model four

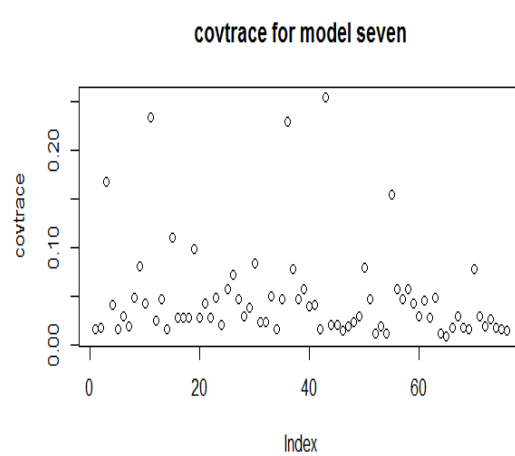
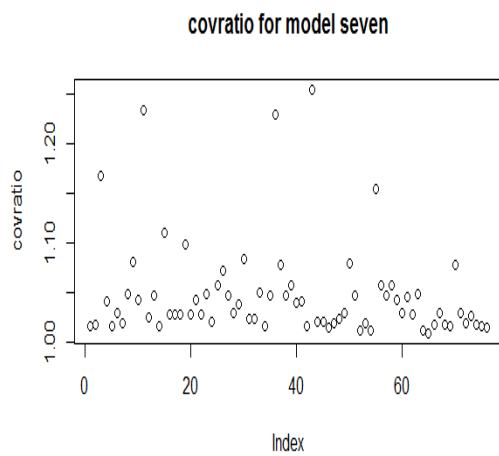
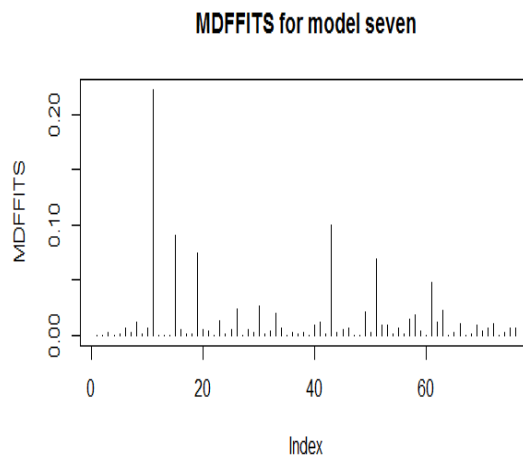


C.5 Influence analysis for model five



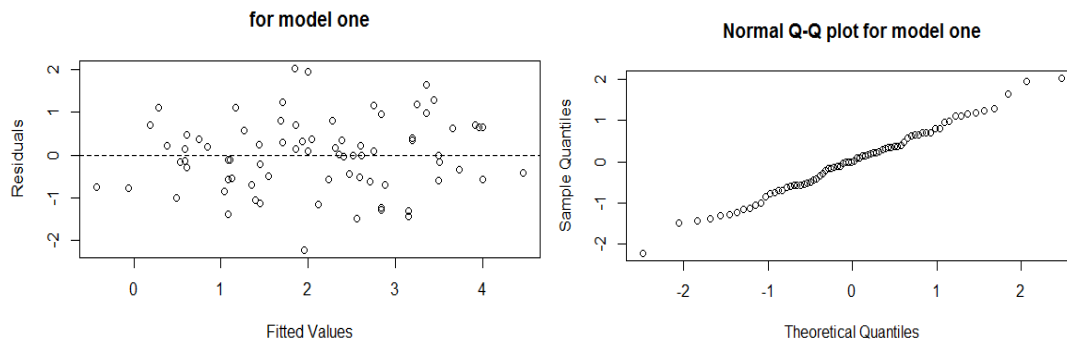


C.6 Influence analysis for model seven

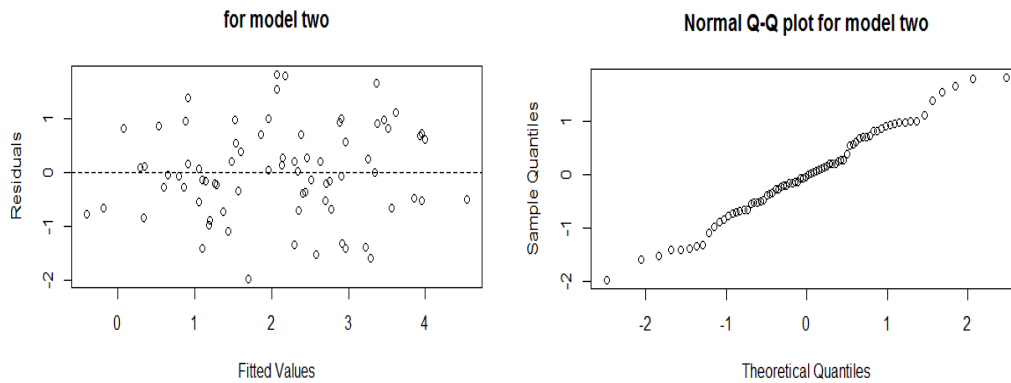


Appendix D: Residual analysis

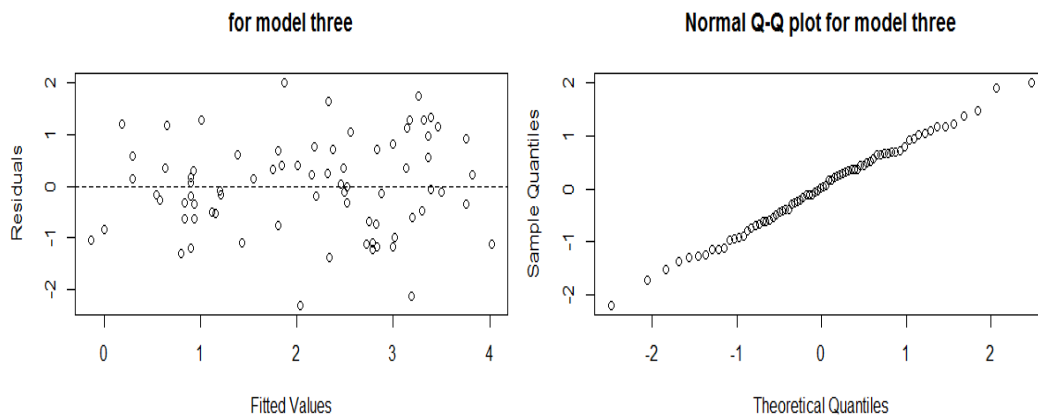
D.1 Residual analysis for model one



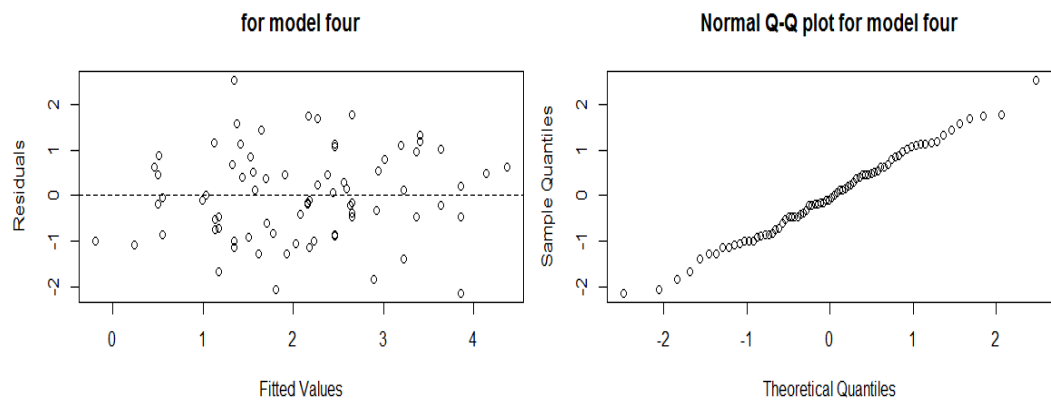
D.2 Residual analysis for model two



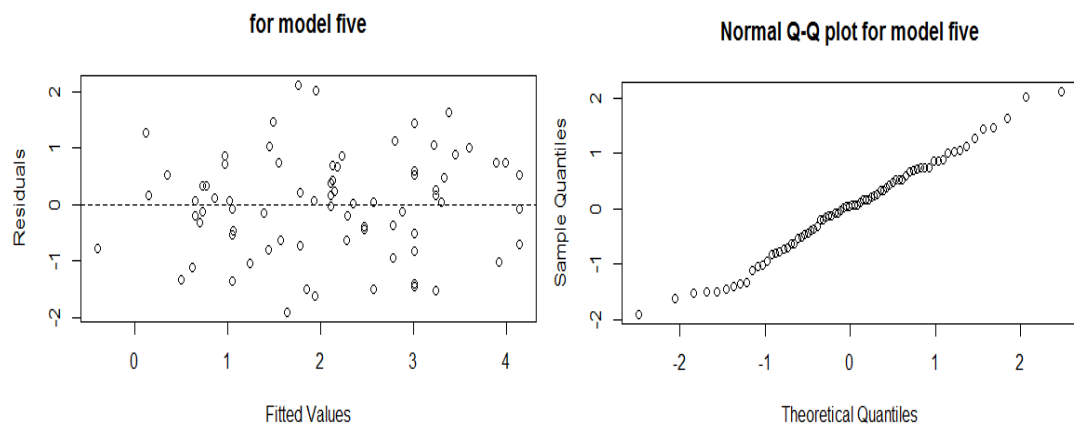
D.3 Residual analysis for model three



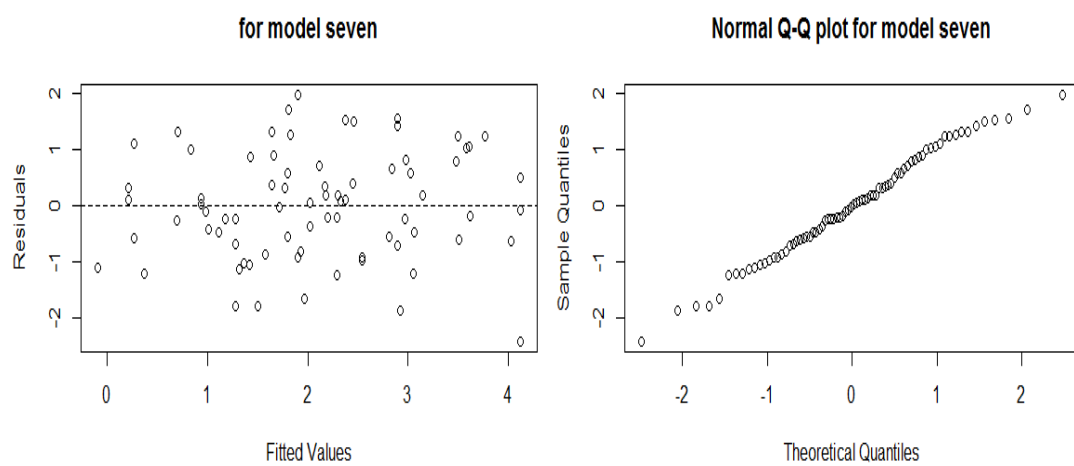
D.4 Residual analysis for model four



D.5 Residual analysis for model five



D.6 Residual analysis for model seven



DECLARATION

I, the undersigned, declare that this thesis is my original work and it has never been presented in any other university. All sources of materials used for this thesis are duly acknowledged.

Name

Signature and date