2020

Statistical Modeling for Genome Wide Association Studies

Alam, Md. Jahangir

University of Rajshahi

STATISTICAL MODELING FOR GENOME WIDE ASSOCIATION STUDIES

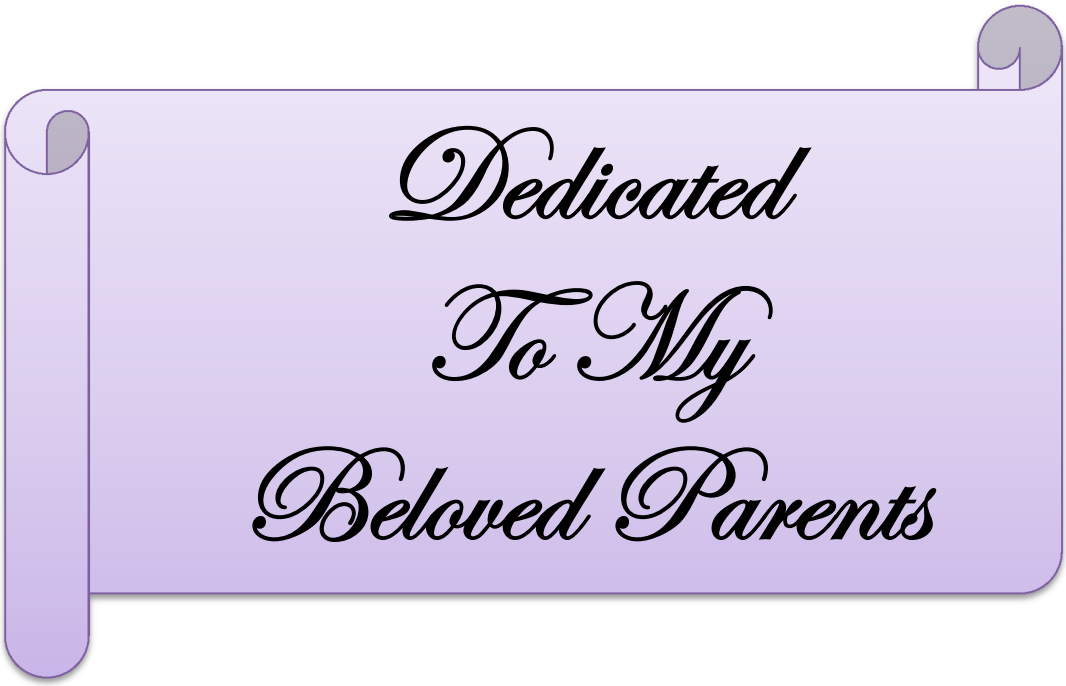
A THESIS SUBMITTED FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN THE
DEPARTMENT OF STATISTICS
UNIVERSITY OF RAJSHAHI
BANGLADESH

BY

MD. JAHANGIR ALAM
B.Sc (Hons.), M.Sc (Statistics)

JUNE 2020

BIOINFORMATICS LAB.
DEPARTMENT OF STATISTICS
UNIVERSITY OF RAJSHAHI
RAJSHAHI 6205
BANGLADESH


Dedicated
To My
Beloved Parents

DECLARATION

I hereby declare that the research work embodied in this thesis entitled **"Statistical Modeling for Genome Wide Association Studies"** has been carried out by me for the degree of Doctor of Philosophy under the supervision of Professor Dr. Md. Nurul Haque Mollah and Professor Dr. Md. Ripter Hossain, Department of Statistics, University of Rajshahi, and Professor Dr. S. M. Shahinul Islam, Institute of Biological Sciences, University of Rajshahi, Rajshahi-6205, Bangladesh. I also declare that the results presented in this dissertation are my own investigation and any part of this thesis has not been submitted elsewhere for any degree/diploma or for the similar purpose.

Md. Jahangir Alam
PhD Fellow
Roll No.: 14225
Session: 2014-2015
Department of Statistics
University of Rajshahi

UNIVERSITY OF RAJSHAHI
RAHSHAHI, BANGLADESH

Certificate of Approval

We are pleased to certify that the thesis entitled **"Statistical Modeling for Genome Wide Association Studies"** is an original work done by **Md. Jahangir Alam**. He has completed the research work under our supervision. As far we know, this thesis has not been previously submitted to any other University or Institute for any kind of degree or diploma.

We also certify that we have perused the thesis and found it satisfactory for the submission to the Department of Statistics, University of Rajshahi for the degree of Doctor of Philosophy (PhD) in Genomics.

---	---	---
Dr. Md. Nurul Haque Mollah	**Dr. Md. Ripter Hossain**	**Dr. S. M. Shahinul Islam**
Principal Supervisor	Co-Supervisor	Co-Supervisor
and	and	and
Professor	Professor	Professor
Department of Statistics	Department of Statistics	Institute of Biological Sciences
University of Rajshahi	University of Rajshahi	University of Rajshahi
Bangladesh	Bangladesh	Bangladesh

CONTENTS

ACKNOWLEDGEMENTS

I would like to express my foremost gratitude to the **Almighty ALLAH** for his endless blessings which make me competent to accomplish my dissertation.

I would like to show my best regards, profound thankfulness and deep appreciation to my honorable principal supervisor **Professor Dr. Md. Nurul Haque Mollah**, Statistics, RU, for his constant supervision, inspiring guidance, enthusiastic encouragement, wise advice and affectionate surveillance throughout the entire period of my research work. Many thanks to my co-supervisors **Professor Dr. Md. Ripter Hossain**, Department of Statistics, RU and **Professor Dr. S. M. Shahinul Islam**, Institute of Biological Science, RU, for their help, continuous encouragement, helpful advices and many discussions during the entire period of my research work that helped me to shape this work.

This work was funded by a Higher Education Quality Enhancement Project (HEQEP) from the University Grant Commission (UGC), Bangladesh. I would like to acknowledge the help of HEQEP Sub-Project (CP-3603). Also I would like to acknowledge the committee members of the HEQEP Sub-Project (CP-3603) for their continuous encouragement and fruitful suggestions in my Ph.D. research period.

I am extremely grateful to **Professor Dr. Md. Golam Hossain**, Chairman, Department of Statistics, University of Rajshahi, who generously helped me by giving permission to use and utilize the equipment as well as all the facilities of the department. I am also very much grateful to all of the teachers of this department for their valuable suggestions and for helping and inspiring me at every stage of this work.

These acknowledgements would not be complete without mentioning my research colleagues: **Md. Mamun Monir**, **Md. Shahjaman**, **Nishith Kumar**, **Md. Masud Rana**, **Md. Nazmol Hasan** and **Zobaer Akond**. It was my great pleasure working with them and I appreciate their ideas, help and good humor.

I also show my appreciation to the other officials of the Dept. of Statistics, RU including **Shah Md. Bodruzzaman** and **Md. Teshem Ali** for their cordial co-operation and administrative support during my research work. I also acknowledge the help of the Bioinformatics Laboratory (Department of Statistics, University of Rajshahi) in facilitating access to this research opportunity. I am also grateful to the administrative and computing staffs of the Department of Statistics (University of Rajshahi) for their support.

My deepest gratefulness belongs to my valued parents **Md. Sadiqul Islam** and **Mrs. Shahazadi Begum**, my wife **Mst. Salma Khatun** and my son **Saif Al-Islam Sabir** for their patience and understanding, who always provided all the supports, encouragements, affections and precious advice when it was necessary.

The Author

SUMMARY

Genomics is one of the most important OMICS research area in Bioinformatics. In **Chapter 1**, we have introduced some basic concepts of genomics. In Genomics, Genome Wide Association Studies (GWAS) have develop gradually over the last ten years into a powerful bioinformatics tool for exploring and investigating the genetic architecture of plant science, animal science and human biology. The advent of new technologies for extracting genetic information from tissue samples has increased the availability of suitable data for finding genes controlling complex traits in plants, animals and humans. Based on the nature of the genomics data GWAS can be divided into major four types: (i) Quantitative trait locus (QTL) mapping based GWAS, (ii) Single nucleotide polymorphism (SNP) based GWAS, (iii) Expression QTL (eQTL) mapping based GWAS and (iv) Sequence based GWAS. Again, based on the number of phenotypes considered in the analysis, GWAS can be of two types: (i) Single-trait GWAS and (ii) Multi-trait GWAS.

Quantitative trait locus (QTL) based GWAS relies on statistical methods to interpret genetic data in the presence of phenotype data and possibly other factors such as environmental factors. Its main goal is to identify the presence of QTLs with significant effects on the trait value(s) as well as to estimate their chromosomal positions on the genome relative to some known markers. In **Chapter 2**, we have proposed a new approach for single-trait QTL analysis based on the properties of bivariate normal distribution. The calculations in our proposed method is very straight forward. Our proposed method shows almost same performance as the existing methods of single-trait QTL analysis. All the classical methods of single-trait QTL analysis, including our proposed method, are very sensitive to outliers and these methods provide misleading results when the phenotypic data are contaminated with outliers. To overcome this problem, we have proposed a robust approach for single-trait QTL analysis by the robust estimation of the parameters of bivariate normal distribution using minimum β–divergence method (**Chapter 2**). Our proposed robust method of single-trait QTL analysis outperforms over the classical method in

presence of phenotypic outliers. Otherwise, the proposed method shows almost equal performance to the existing method of single-trait QTL analysis.

Although single-trait SIM methods can be applied to each trait one-by-one, such approaches do not take into account the pleotropic effects. Therefore, statistical methods for joint analyses of multiple traits are very essentials to identify important QTL locations, which control multiple traits simultaneously. All the existing methods of multi-trait QTL analysis are time consuming and include computation complexity. To overcome this problem, we have developed a new method of multi-trait QTL analysis, called fast multi-trait (FMT) QTL mapping, using the properties of multivariate normal distribution, in which calculations are very straight forward (**Chapter 3**). Our proposed method exhibits almost similar performance to the existing methods of multi-trait QTL analysis. Moreover, our proposed method is very efficient in terms of computation time. Although our proposed method of multi-trait QTL analysis is very fast compared to the existing methods, it is very sensitive to phenotypic outliers and it produces misleading results in case of phenotypic contaminations.

In **Chapter 4**, to overcome the problem of phenotypic outliers, we have developed a robust statistical method for multi-trait QTL analysis by robustifying our proposed FMT QTL mapping approach using minimum β–divergence method. Our proposed robust method of multi-trait QTL analysis outperforms over the classical approaches in presence of phenotypic contaminations. Otherwise, the proposed method shows almost equal performance to the classical methods. This proposed method is also implemented on the eSNPs dataset to explore cis/trans regulatory elements.

Due to the recent advancement in the NGS technologies, SNP data of complete genome can be obtained for GWAS. Nowadays, SNP-based GWAS has been widely used for the genetic study of a variety of species including humans, animals and plants to identify genomic locations/regions responsible for various quantitative traits, which has been made possible by decreasing the cost and time required to obtain sequences of whole genome and genome-wide SNPs. Many methods have been

developed for SNP-based GWAS in the literatures, ranging from simple extensions of single-trait approaches to sophisticated multi-trait approaches designed specifically for multi-trait GWAS. All the existing methods of GWAS are very sensitive to outliers and these methods produce misleading results when the data are contaminated by phenotypic outliers. To overcome this problem, we have developed a robust method for single-trait GWAS by robustifying the simple linear regression model using minimum β–divergence method (**Chapter 5**). Our proposed robust method shows better performance than the existing methods of single-trait GWAS in presence of phenotypic outliers. Otherwise, the proposed method shows almost same performance as the existing methods.

The next step after completing QTL/SNP based GWAS is to integrate the information, and perform structural and functional analysis of the identified associated QTLs/genes/SNPs to investigate the molecular mechanisms of the identified loci/QTLs/genes/SNPs. The entire process is known as sequence matching-based GWAS. In sequence matching based GWAS, a particular sequence of interest (genomic sequence or coding sequence or protein sequence) of a gene is tried to match in the whole genome by searching the similar sequences in the whole genome stored in the databases. If the sequence of interest matches with any portion of the whole genome, then we called that the sequence of interest is associated with that portion of the genome. After this, we select the sequence from the genome that is most associated with the sequence of interest and then we investigate the molecular mechanisms/functions of that selected sequence (i.e., most associated sequence). The whole process described above is performed by using different bioinformatics tools of structural and functional analyses. In **Chapter 6**, we have performed different comparative structural and functional analyses of the 42 finally selected rolling leaf (RL) genes using different bioinformatics techniques including gene structure, conserved domain, phylogenetic, gene ontology (GO), transcription factor (TF), Kyoto Encyclopedia of Genes and Genomes (KEGG), gene network and exploratory gene expression analysis. Exon-intron organization and conserved domain analysis showed diversity in structures and conserved domains of RL genes. Phylogenetic analysis classified the genes into five major groups. GO and TF analyses revealed that regulation-related genes were remarkably enriched in biological process and 10

different TF families were involved in rice leaf rolling. KEGG analysis demonstrated that 14 RL genes were involved in the KEGG pathways, among which 50% were involved in the metabolism pathways. Of the selected RL genes, 55% genes were non-interacting with other RL genes and OsRL9 was the most interacting RL genes. Most of the RL genes exhibited extreme (very high/low) expression at leaf, root and shoot. These results provide important information regarding structures, conserved domains, phylogenetic revolution, protein-protein interactions, gene expression pattern and others genetic basis of RL genes which might be helpful to the researchers for functional analysis of new candidate RL genes to explore their characteristics and molecular mechanisms for high yield rice breeding.

LIST OF TABLES

Table No.	Title	Page

Table No.	Title	Page

LIST OF FIGURES

Figure No.	Title	Page

Figure No.	Title	Page

Figure No.	Title	Page

Figure No.	Title	Page

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
ARF	Auxin Response Factors
BC	Backcross
BP	Biological Process
BND	Bivariate Normal Distribution
CC	Cellular Component
CD	Conserved Domain
CDD	Conserved Domain Database
CDS	Coding Sequence
CHD	Chromo-domain Helicase DNA
Chr	Chromosome
CIM	Composite Interval Mapping
DE	Differential Expression
DESeq	Differential Expression of Sequence data
DB	Database
DH	Double Haploid
DNA	Deoxyribonucleic Acid
DSRM	Double Stranded RNA-binding Motif
EC	Enzyme Commission
edgeR	Empirical analysis of digital gene expression in R
EM	Expectation Maximization
EMMA	Efficient Mixed Model Association
EMMAX	EMMA eXpedited
ESS	Error Sum of Squares
FDR	False Discovery Rate

FPKM	Fragments Per Kilobase Million
GAPIT	Genome Association and Prediction Integrated Tool
GARP	Glycoprotein-A Repetitions Predominant protein
GC	Genomic Control
GE	Gene Expression
GEMMA	Genome-Wide Efficient Mixed Model Analysis
GO	Gene Ontology
GSDS	Gene Structure Display Server
GWA	Genome-Wide Association
GWAS	Genome-Wide Association Studies
HK	Haley and Knott
IBD	Identity by Descent
ICIM	Inclusive Composite Interval Mapping
ID	Identity Document
IM	Interval Mapping
KCS	Ketoacyl-CoA Synthase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthologous
LBD	LOB Domain
LEA	Late Embryogenesis Abundant
LIMMA	Linear Models for Microarrays
LMM	Linear Mixed Model
LMU	Linkage Map Unit
LOD	Log of Odds
LR	Likelihood Ratio
LRT	Likelihood Ratio Test
LS	Least Squares
MDS	Multi-Dimensional Scaling

MF	Molecular Function
MIM	Multiple Interval Mapping
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate
MND	Multivariate Normal Distribution
mRNA	Messenger RNA
MSU	Michigan State University
MVR	Multivariate Regression
NAALAD	N-Acetylated-Alpha-Linked Acidic Dipeptidase
NAC	NAM, ATAF and CUC
NAM	No Apical Meristem
NB	Negative Binomial
NCBI	National Center for Biotechnology Information
NEFA	Non-Esterified Fatty Acids
PA	Protease Associated
PCA	Principal Component
PEBP	Phosphatidyl Ethanolamine Binding Protein
PS	Population Stratification
PSMA	Prostate Specific Membrane Antigen
QTL	Quantitative Trait Locus
RAD	Restriction site Associated DNA
RAP	Rice Annotation Project
RAPD	Random Amplification of Polymorphic DNA
RED	Rice Expression Database
RFLP	Restriction Fragment Length Polymorphism
RFS	Rolled Fine Striped
RGAP	Rice Genome Annotation Project
RI	Recombinant Inbred

RING	Really Interesting New Gene
RL	Rolling Leaf
RNA	Ribonucleic Acid
SA	Structured Association
SAM	Significant Analysis of Microarrays
SANT	Swi3, Ada2, N-Cor, and TFIIIB
SAS	Statistical Analysis System
SD	Standard Deviation
SE	Standard Error
SEA	Singular Enrichment Analysis
SFP	Single Feature Polymorphism
SLL1	SHALLOT-LIKE1
SN	Serial Number
SNP	Single Nucleotide Polymorphism
SIM	Simple Interval Mapping
SS	Sum of Squares
SSLP	Simple sequence length polymorphism
SSR	Simple Sequence Repeat
StAR	Steroidogenic Acute Regulatory Protein
START	StAR-related lipid Transfer
STR	Short Tandem Repeat
TDT	Transmission Disequilibrium Test
TF	Transcription Factor
TFR	Transferrin Receptor
WT	Wild Type

Chapter 1

Introduction

1.1 Introduction to Genome and Genomics

Genomics is the study of the whole genome of any living organism along with its environment and it incorporates different elements from genetics (a branch of biology that generally deals with the heredity). Genomics uses “DNA sequencing methods” to generate sequences of genomes using recombinant DNA, and it utilizes Bioinformatics to assemble the sequences of whole genomes and analyze the structure and function of genomes. It differs from 'classical genetics' in that it considers an organism’s full complement of hereditary material, rather than one gene or one gene product at a time. Moreover, genomics focuses on interactions between loci and allele within the genome and other interactions such as epistasis, pleiotropy and heterosis (Figure 1.1). The availability of complete DNA sequences for entire organisms is made easy by the Genomics. Genomics was made possible by both the pioneering work of Fred Sanger and the more recent next-generation sequencing technology.

Fred Sanger's group established techniques of sequencing, genome mapping, data storage, and bioinformatics analyses in the 1970s and 1980s. This work paved the way for the human genome project in the 1990s (Bentley et al., 2008) an enormous feat of global collaboration that culminated in the publication of the complete human genome sequence in 2003. Nowadays, next-generation sequence technologies have led to remarkable improvements in the speed, capacity and affordability of genome sequencing. Moreover, advances in bioinformatics have enabled hundreds of life-

science databases and projects that provide support for scientific research. Information stored and organized in these databases can easily be searched, compared and analyzed.

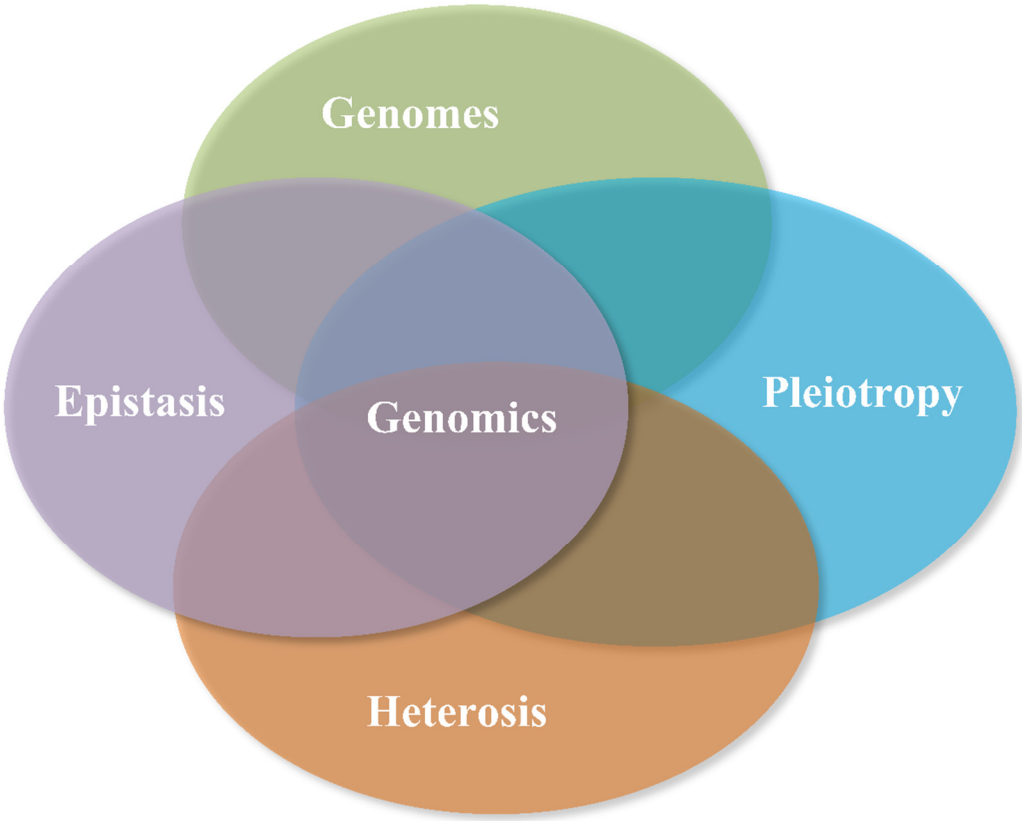


Figure 1.1: The genomics studies of whole organisms and other intragenic interactions.

1.2 Some Important Terminologies Related to Genomics

Cell: The cell is the functional basic unit of life of all living organisms. It was discovered by Robert Hooke in 1665. It is the smallest unit of life that is classified as a living thing, and is often called the building block of life. Some organisms have only a single cell and those organisms are unicellular organisms, e.g., most bacteria. Some organisms have more than one cell and those organisms are called multicellular organisms, such as humans, animals and plants. Humans have about 3.72×10^{13} (i.e., about 37.2 trillion) cells and this number of cells ranges from 10^{12} to 10^{16} (Bianconi et al., 2013). A typical cell size of humans is 10 μm and the mass of a typical cell of

humans is 1 nanogram. The smallest cell varies in size from 0.1 to 0.5 μm (micrometer) which is found in bacteria. On the other hand, the largest cell is of size (170 mm × 130 mm) which is found in the egg of an ostrich. The smallest cell in the human body is the cerebellar granule cell (size: 4 − 4.5 μm) and the largest cell in the human body is the human egg cell (size: 0.12 mm). The cell theory, first developed by Matthias Jakob Schleiden in 1837-1839 and Theodor Schwann in 1838-1839, states that all organisms are composed of one or more cells, all cells come from preexisting cells, vital functions of an organism occur within cells, and all cells contain the hereditary information necessary for regulating cell functions and for transmitting information to the next generation of cells. The pictures of the animal cell and plant cell are given in Figure 1.2.

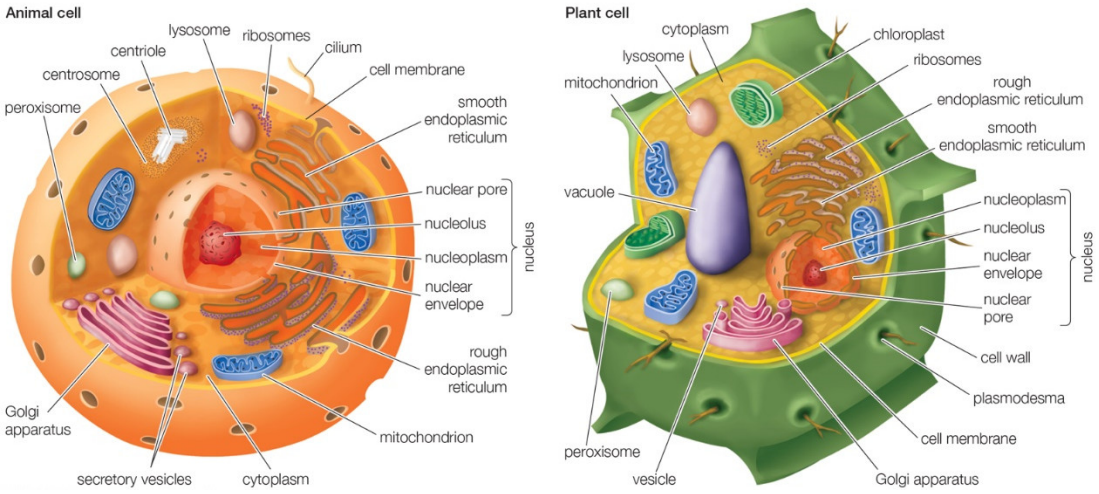


Figure 1.2: The structure of plant cell and animal cell (Image source: Internet).

Chromosome: A chromosome is a thread-like organized structure composed of deoxyribonucleic acid (DNA) molecule and proteins found in the nucleus of a cell. Chromosome is called as the storage unit of DNA and genes. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Chromosome also contains DNA-bound proteins, which serve to package the DNA and control its functions. Chromosomes vary widely between different organisms. The DNA molecule may be circular or linear, and can be composed of 100,000 to 10,000,000,000 nucleotides in a long chain (Harikrishnan and Grace, 2013). Chromosomes are the vital units for cell division and must successfully be

replicated, divided, and passed to their daughter cells in order to ensure the genetic diversity and survival of their progeny. Chromosomes may exist as either duplicated or unduplicated. Unduplicated chromosomes are single linear strands, whereas duplicated chromosomes contain two copies joined by a centromere.

In humans, chromosomes can be divided into two types – autosomes and sex chromosomes. Certain genetic traits are linked to a person's sex and are passed on through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. All act in the same way during cell division. Human cells have 33 pairs of large linear nuclear chromosomes (33 pairs of autosomes and one pair of sex chromosomes), giving a total of 46 per cell. In addition to these, human cells have many hundreds of copies of the mitochondrial genome. Sequencing of the human genome has provided a great deal of information about each of the chromosomes.

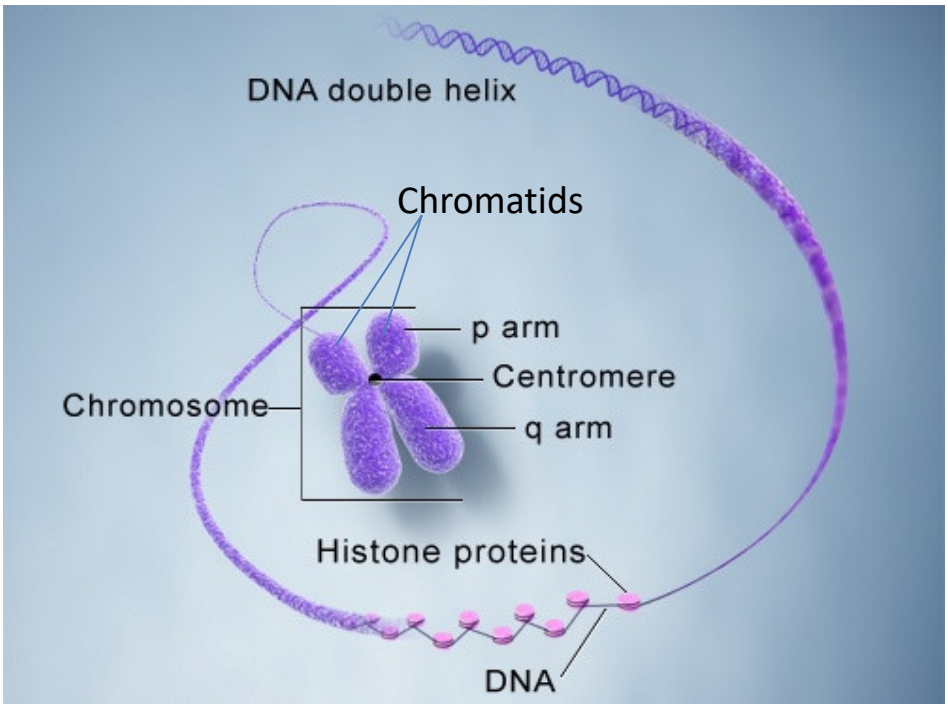


Figure 1.3: Structure of Chromosome (Image source: Internet).

DNA: DNA is a deoxyribonucleic acid that contains the genetic instructions specifying the biological development and functioning of all cellular forms of life. DNA is the hereditary material in humans and almost all other living organisms.

Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA or mtDNA). The information in DNA is stored as a code made up of four nitrogenous chemical bases: (i) adenine (A), (ii) guanine (G), (iii) cytosine (C) and (iv) thymine (T). DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder. DNA acts as the store of genetic information. The sequence of bases along its length is the "language" of the cell and code for all its proteins. DNA is also the molecule of heredity. When a cell or a multi cellular organism reproduced either sexually or asexually, the genetic information stored in the DNA molecules is faithfully copied and exact copies of these DNA molecules passed along from one generation to the next. A DNA double helix is shown in Figure 1.4.

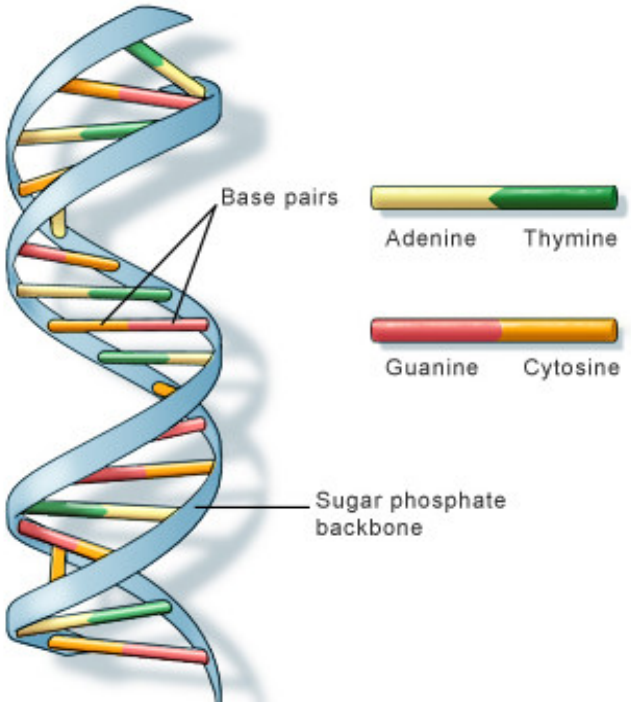


Figure 1.4: Structure of a deoxyribonucleic acid (DNA) double helix (Image source: U.S. National Library of Medicine).

Gene: Genes are discrete units in which biological characteristics are inherited from parents to offspring (Wu et al., 2007). A gene is a unit of heredity in a living organism. A gene consists of a sequence of exons and introns which are transmitted as a whole from generation to generation (Chen, 2016a). Genes are made up of DNA and each chromosome contains many genes. It is a name given to some stretches of DNA and ribonucleic acid (RNA) that code for a type of protein or for an RNA chain that has a function in the organism. Living things depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring, although some organelles (e.g. mitochondria) are self-replicating and are not coded for by the organism's DNA. All organisms have many genes corresponding to various different biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type or increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life. A figure of gene is shown in Figure 1.5.

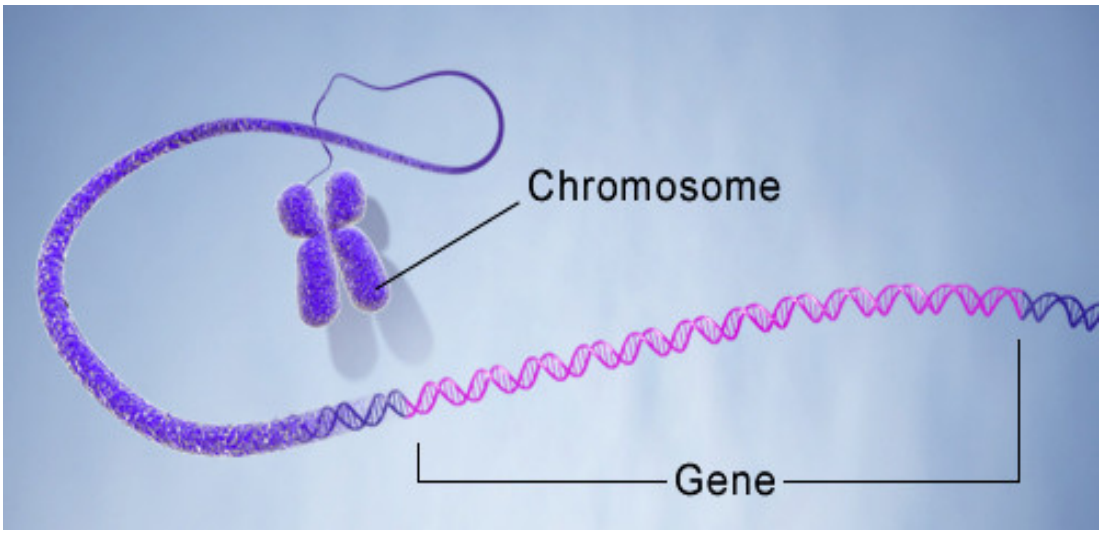


Figure 1.5: Genes: Functional part of DNA (Image source: U.S. National Library of Medicine).

Genome: In board sense, the total amount of DNA including its genes in a single cell (a haploid cell for a diploid organism) of any organism is called genome. In modern molecular biology and genetics, the genome is the entirety of an organism's hereditary information. Each genome of an organism contains all of the information

needed to build and maintain that organism. It is usually encoded in DNA. However, for many types of virus, it is encoded in RNA. The genome includes both the coding sequences (i.e., genes) and the non-coding sequences of DNA or RNA. The study and analysis of genomes is called genomics. Figure 1.6 shows a genome of a living organism

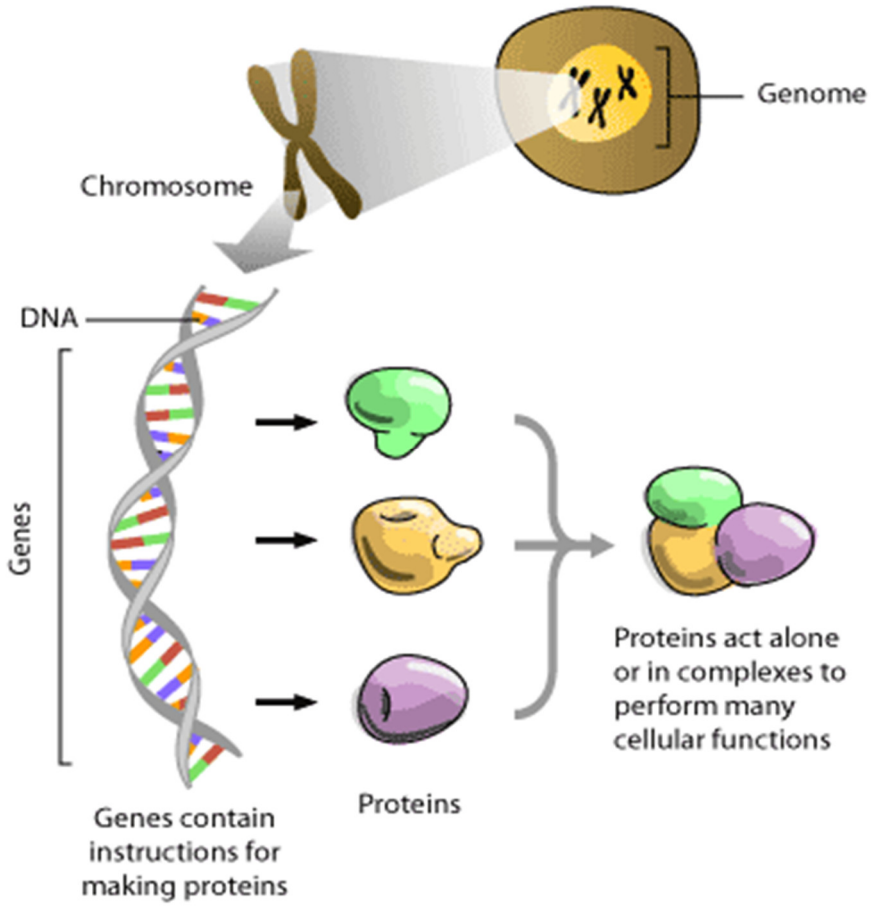


Figure 1.6: Genome of a living organism (Image source: Internet).

Marker: A genetic marker is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed. A genetic marker may be a short DNA sequence, such as a sequence surrounding a single base-pair change (single nucleotide polymorphism, SNP), or a long one, like minisatellites. Figure 1.7 shows a picture of genetic markers. Some commonly used types of genetic markers are as follows:

- RFLP (Restriction fragment length polymorphism)
- SSLP (Simple sequence length polymorphism)
- AFLP (Amplified fragment length polymorphism)
- RAPD (Random amplification of polymorphic DNA)
- VNTR (Variable number tandem repeat)
- Microsatellite polymorphism, SSR (or Simple sequence repeat)
- SNP (Single nucleotide polymorphism)
- STR (Short tandem repeat)
- SFP (Single feature polymorphism)
- DArT (Diversity Arrays Technology)
- RAD markers (or Restriction site associated DNA markers)

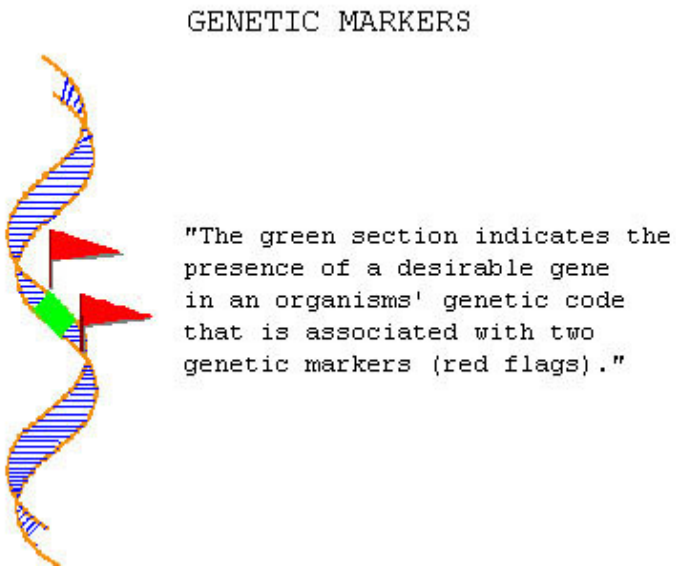


Figure 1.7: Genetic markers are indicated by red flags (Image source: Internet).

Locus: Locus is the location/position of a gene/marker on the chromosome.

Allele: Allele is one variant form of a gene/marker at a particular locus. If we have two alternative genes, say *A* and *a*, there are two types of homozygotes, namely *AA* and *aa*, and one type of heterozygote, namely *Aa*. These alternative genes are called alleles.

Genotypes: At each locus (except for sex chromosomes) there are 2 genes (e.g., *A* and *a*). These constitute the individual's **genotype** at the locus. With a single pair of alleles, there are three different kinds of possible organisms represented by the three genotypes *AA*, *Aa* and *aa*.

Phenotypes: The expression of a genotype is called a **phenotype**. For example, hair color, weight, height, the presence or absence of a disease, etc.

Mendel's Laws:

1. First law or law of segregation
2. Second law or law of independent assortment

Mendel's First Law or Law of Segregation: The *law of segregation* states that characteristics are controlled by pairs of genes that segregate or separate during the formation of the reproductive cells, thus passing into different gametes (Wu et al., 2007). An example of the law of segregation is show in Figure 1.8.

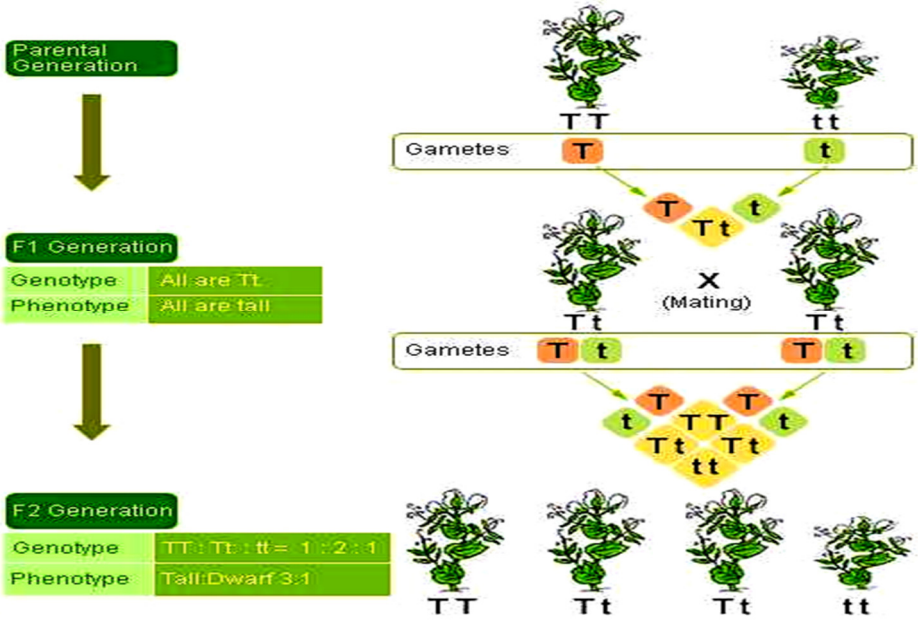


Figure 1.8: Mendel's first law or law of segregation (Image source: Internet).

Mendel's Second Law or Law of Independent Assortment: The *law of Independent Assortment* says that when two or more pairs of genes segregate simultaneously, they do so independently (Wu et al., 2007). An example of the law of independent assortment is show in Figure 1.9.

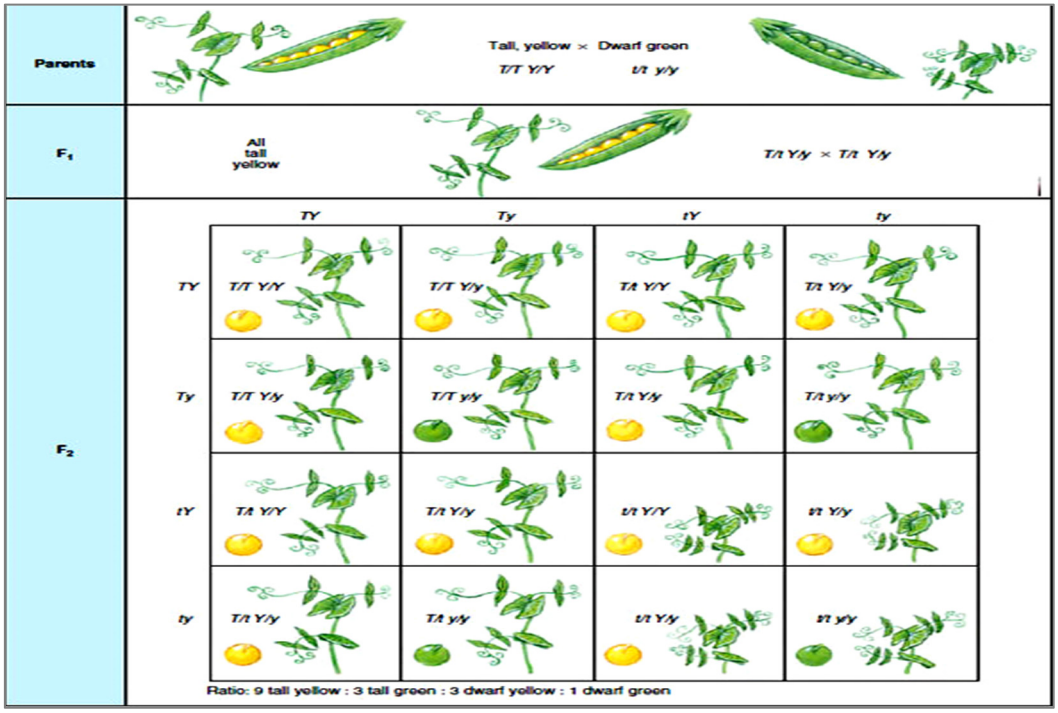


Figure 1.9: Mendel's second law or law of independent assortment (Image source: Internet).

Crossing Over: Crossing over, or recombination, is the process of exchanging the chromosome segments of equal size between two homologous non-sister chromatids of two homologous chromosomes during meiosis. At an early stage of meiosis, the two homologous chromosomes lie side-by-side with corresponding loci aligned (Figure 1.10A). Each of the paired chromosomes is then duplicated to form two sister strands (chromatids) connected to each other at a region called the centromere. The homologous chromosomes form pairs, so that each resulting complex consists of four chromatids known as a tetrad (Figure 1.10B). At this stage, the non-sister chromatids adhere to each other in a semi-random fashion at the regions called chiasmata. Each chiasma represents a point where crossing over between two non-sister chromatids can occur (Figure 1.10C). Crossing over creates new combinations of genes in the

gametes that are not found in either parents, contributing to genetic diversity. The diagram of crossing over is shown in Figure 1.10.

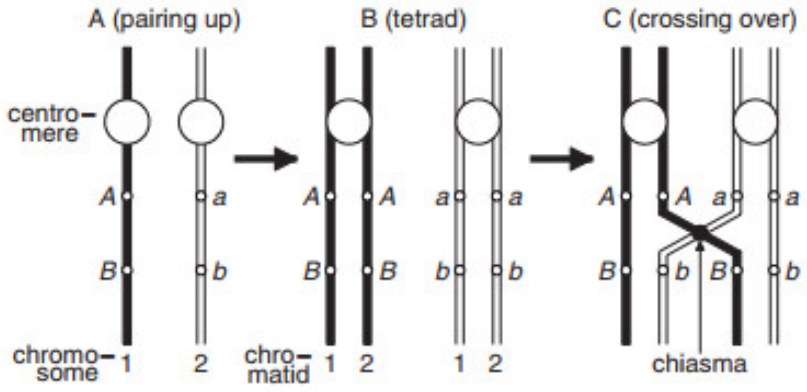


Figure 1.10 (A-B): Crossing over between two linked loci **A** and **B** (Image source: Wu et al. (2007)).

Recombination Frequency or Fraction: The proportion of recombinants in a gametic pool is called the recombination fraction or recombination frequency. Recombination fraction is usually denoted by *r*. This fraction depends on the number of crossovers, although not in a linear fashion. In genetic linkage study, people often use recombination fraction in place of the number of crossovers to measure the distances between loci (Xu, 2013b). Recombination fraction is a measure of genetic linkage and is used to create the genetic linkage map. Recombination fraction is the frequency that a single chromosomal crossover will take place between two genes during meiosis. A centimorgan (cM) is a unit that describes a recombination fraction of 1%. In such way we can measure the genetic distance between two loci, based upon their recombination fraction. This is a good estimate of the real distance. Double crossovers would turn into no recombination. In this case, we cannot tell if crossovers took place. If the loci we are analyzing are very close (< 7 cM) a double crossover is very unlikely. When distances become higher, the likelihood of a double crossover increases. As the likelihood of a double crossover increases we systematically underestimate the genetic distance between two loci.

Genetic Distance: The recombination fraction between two loci depends on how far apart they are in physical terms along the DNA molecule. The genetic distance between two loci is defined as the average number of crossovers. In Genetics, genetic distance is usually denoted by *d*. There are two different measurements for the genetic distance between two loci: the number of crossovers and the recombination fraction. Genetic distance is usually measured in Morgans. The unit of centiMorgan (cM) is also used. A centiMorgan is one hundredth of a Morgan. As a rough guideline, 1% recombination fraction is equivalent to the genetic distance of 1 cM and it corresponds approximately to a physical distance of one million base pairs (i.e., 1 megabase). However, the exact number varies from organism to organism, even from region to region in the genome of the same organism. The relationship between recombination and genetic distance does not remain linear. Because when two loci are in infinite distance apart, the recombination fraction is still only 50%. Recombination fraction is converted into genetic distance using different types of mapping functions, e.g., Haldane map function, Kosambi map functions.

Genetic Linkage: Genetic linkage is the tendency of certain loci or alleles to be inherited together. Genetic loci that are physically close to one another on the same chromosome tend to stay together during meiosis, and are thus genetically linked. The proximity of two or more markers on a chromosome; the closer the markers, the lower the probability that they will be separated during DNA repair or replication processes, and hence the greater the probability that they will be inherited together. Recombination fraction (r) is used as a measure of genetic linkage. The value $r = 0.5$ indicates independent segregation or no linkage. The value $r < 0.5$ indicates there is linkage. The smaller the value of r the stronger the linkage between two loci.

Map Distance: The map distance between any two loci is the average number of points of exchange occurring in the segment (Xu, 2013b). One linkage map unit (LMU) is 1% recombination. Thus, the linkage map distance between two genes is the percentage recombination between those genes.

Example: Suppose we have a total of 300 recombinant offspring out of 2000 total offspring. Map distance is calculated as

Map distance = (# Recombinants)/(Total offspring) ×100.
So our map distance is (300/2000)×100, or 15 LMU.

Map Function: There is a very simple relationship between the recombination fraction and the map distance for a pair of loci. Such a relationship is called a map function. That is, the map function is a mathematical function that converts the recombination fraction (r) between two loci to the genetic distance separating them (d). Widely used map functions: (i) Morgan Map Function, (ii) Haldane Map Function and (iii) Kosambi Map Function.

Morgan Map Function: The Morgan map function is the simplest map function, which assumes that (i) there is at most one crossing-over occurring within the interval of two loci, and (ii) the probability of a crossing-over within an interval is proportional to the map length of the interval (Wu et al., 2007). Under these assumptions, the probability of a chiasma occurring in a distance of d map units is equal to the expected number of crossing-overs per gamete in this distance and therefore to $2d$ which gives

$$r = \frac{1}{2}[1 - \Pr(X = 0)] = \frac{1}{2}[1 - (1 - 2d)] = d \qquad (1.1)$$

This function holds only when $0 \leq d \leq 1/2$ since for $d > 1/2$ it results in recombination fractions of greater than 1/2. It may therefore be used as an approximation for short distances but is not applicable for long segments of chromosomes.

Haldane Map Function: The Haldane map function assumes that crossing-overs occur at random and independently of each other (Haldane, 1919). With this assumption, the occurrence of crossing-overs between two loci on a chromosome can be viewed as a Poisson process so that the number of crossing-overs between the loci can be modelled by a Poisson distribution. Since map distance, d, is defined as the average number of crossing-overs per chromatid within a given interval, the average number of crossing-overs for the tetrad as a whole is $2d$. This assumption of a Poisson

process implies that the probability of no chiasma within the interval, $\Pr(X = 0)$, is e^{-2d}. The Haldane map function is defined as follows:

$$r = \frac{1}{2}[1 - \Pr(X = 0)] = \frac{1}{2}(1 - e^{-2d}) \tag{1.2}$$

whose inverse is

$$d = -\tfrac{1}{2}\ln(1 - 2r).$$

Kosambi Map Function: In 1919, Haldane introduced a differential equation method (Haldane, 1919) that generalized the construction of various map functions. Using this generalization, Kosambi derived a map function, known as Kosambi map function, which is both simple and justifiable in practice (Kosambi, 2016). The Kosambi map function is defined as

$$r = \frac{1}{2}\frac{e^{2d} - e^{-2d}}{e^{2d} + e^{-2d}} \tag{1.3}$$

with inverse

$$d = \frac{1}{4}\ln\frac{1 + 2r}{1 - 2r} \tag{1.4}$$

Qualitative Traits: Phenotypes of organisms can be described in qualitative or quantitative terms. A qualitative trait is a trait that can be assigned to a number of classes, such as round or wrinkled shape of peas. Mendelian traits, controlled by single genes, are a special case of qualitative traits. Much of Mendelian genetics is based on qualitative assessments of phenotypes, where differences in individuals can be identified by their distinct phenotypic values.

Quantitative Traits: A quantitative trait is a trait that is measured numerically, such as body weight, crop yield and the bristle number of a Drosophila. Most characters of economic importance in plants and animals are quantitative traits.

Quantitative trait locus (QTL): A quantitative trait tends to exhibit continuous variation, which is usually a consequence of the combined effects of multiple genes (Bernardo, 2001; Fairbanks and Andersen, 1999; Glazier et al., 2002). A genomic

region that influences a quantitative trait is referred to as a quantitative trait locus (QTL). When a trait is influenced by multiple genes, the inheritance of individual genes follows a Mendelian pattern, but the segregation of each individual gene is obscured by segregation of the rest genes. Continuous variation in quantitative traits may also be influenced by environment. For example, crop yield is not only determined by the genetic composition but also influenced by other factors such as moisture, sunlight and texture of the soil. When a trait is influenced by environment, the segregation of genes underlying the trait may be obscured by environmental effects.

Pleiotropic effect: A QTL is said to have pleiotropic effect if it simultaneously controls several phenotypic traits.

Heritability: There are generally two sources of variation in a quantitative trait: genetic effects and environmental influences. The variance of a quantitative trait can be partitioned into genetic variance, which is induced by genes underlying the quantitative trait, and environmental variance, which is induced by environmental factors, accordingly

$$V_P = V_G + V_E \tag{1.5}$$

where V_P is the phenotypic variance, V_G is the genetic variance and V_E is the environmental variance. Geneticists are often interested in what proportion of the phenotypic variation is genetic, which is conceptually associated with heritability. Generally, there are two types of heritability: (i) Broad-sense heritability and (ii) Narrow-sense heritability.

Broad-sense heritability: The ratio of the genetic variance over the phenotypic variance is defined as broad-sense heritability. That is, it is the proportion of the phenotypic variance caused by genes for the underlying the trait. It is mathematically expressed as follows:

$$H^2 = \frac{V_G}{V_P} \tag{1.6}$$

or equivalently

$$H^2 = \frac{V_G}{V_G + V_E} \tag{1.7}$$

It is easy to see that the broad-sense heritability H^2 falls between 0 and 1. If all the phenotypic variation is due to genetic variance, then $H^2 = 1$ and if all the phenotypic variation is due to environmental variance, $H^2 = 0$.

Narrow-sense heritability: The genetic variance can further be partitioned into the variance for additive effects of genes, the variance for dominant effects of genes and the variance for their interactions (referred to as epistasis):

$$V_G = V_A + V_D + V_I \tag{1.8}$$

where V_A is the variance for additive effects, V_D is the variance for dominant effects and V_I is the variance for epistasis.

The non-additive effect or variance is the summation of dominance effect and epistatic effect or variance. Since the additive effect can be inherited from the parents to offspring whereas the non-additive effect cannot, we use the proportion of the phenotypic variance caused by the additive effects of genes. The ratio of the additive variance over the total phenotypic variance is define as the narrow-sense heritability. Mathematically, it is expressed as follows:

$$h^2 = \frac{V_A}{V_P} \tag{1.9}$$

Narrow-sense heritability quantifies the degree with which the phenotypic value of a quantitative trait is unchanged from one generation to next generation.

The two types of heritability defined in (1.6) and (1.9) are conventionally used to describe the degree of overall genetic control for a trait, including the contributions of all the underlying genes (Lynch and Walsh, 1998). These two types of heritability are now commonly used to describe the contributions of individual genes if these genes can be detected by an approach like genetic mapping.

Backcross (BC): The crossing of first generation (F_1) with one of its parents (P), father/mother, is called backcross (BC). Suppose, two alternative gene *A* and *a*, their zygote in first generation is *Aa*. If we cross this zygote again with *AA* or *aa*, the BC genotype is *AA* are *Aa* in ratio 1:1 and if the population come from BC, then it is called backcross population.

Intercross (F_2): The crossing of first generation (F_1) with the first generation (F_1) is called intercross (F_2). Suppose, two alternative genes are *A* and *a*. Then the first generation is heterozygous of the type *Aa* and their F_2 genotypes are *AA*, *Aa*, and *aa* in ratios 1:2:1, respectively, and if population come from F_2 is called intercross population.

Recombinant Inbred Line: Recombinant inbred line is derived from repeated selfings of F_2 individuals for many generations until all progeny become homozygotes. In animals (except some lower worms), selfing is impossible, and thus, RIL must be obtained by repeated brother–sister matings. For large animals with long generation intervals, RIL cannot be obtained within a reasonable amount of time. Therefore, only small laboratory animals, e.g., fruit flies and mice, are possible to have RIL. An RIL generated via selfing is called RIL1, while an RIL generated via brother–sister mating is called RIL2.

Double Haploid: Double haploid (DH) is obtained by doubling the gametes of first generation (F_1) individuals through some special cytogenetic treatment. DH can be achieved by a single generation of cytogenetic manipulation, just like a BC population. However, a DH individual is homozygous for all loci. Therefore, a DH population contains two possible genotypes, *AA* and *aa*.

Single nucleotide polymorphism (SNP): A single nucleotide polymorphism (SNP) is a site in the genome where the DNA sequences of many individuals differ by a single A, T, C, or G. Nowadays, interest is focusing on the possibilities for using SNPs, especially in association studies. These contain changes in a single base pair at a particular point. The change is either present or absent, so the markers are bi-allelic. The SNPs are extremely numerous, being densely present throughout the whole

genome, and so may offer more potentials for fine-mapping disease genes than microsatellite markers.

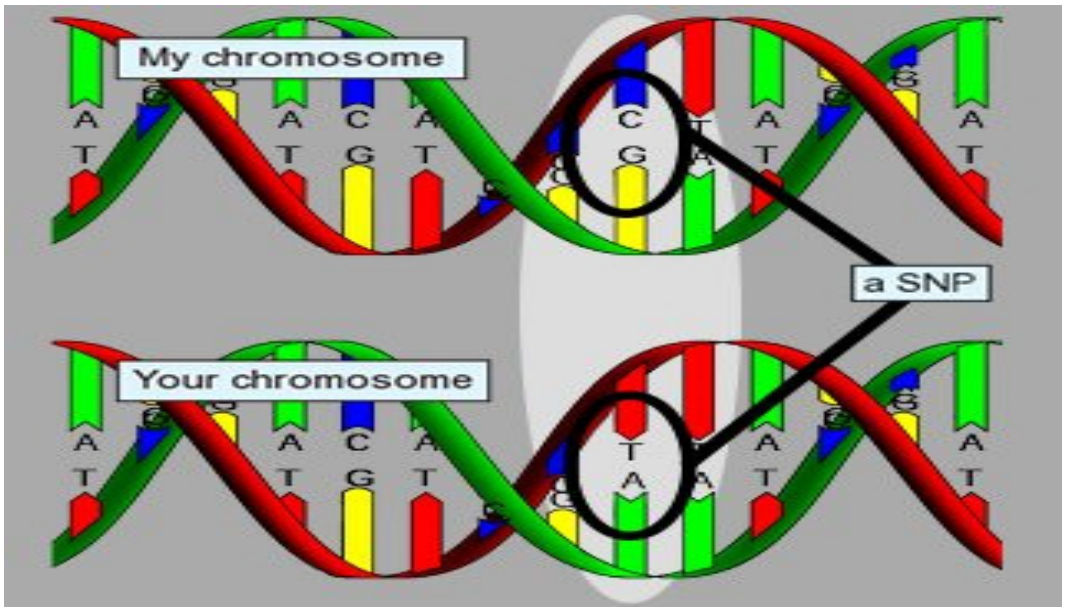


Figure 1.11: Single nucleotide polymorphism (Image source: Internet).

1.3 Genome Wide Association Studies (GWAS)

The human genome project, which was completed in 2003, made it possible for us, for the first time, to read the complete genetic blueprint of human beings. Since then, the researcher started looking into the germline genetics variants which are associated with the heritable diseases and traits among humans, known as genome-wide association studies (GWAS). *A genome-wide association (GWA) study is defined as any study of genetic variation across the whole genome of any organism that is designed to identify genetic associations with observable traits of interest (such as height, weight, blood pressure), or the presence or absence of a particular disease (such as presence or absence of diabetes, or cancer) or any specific condition*. Figure 1.12 shows the pipeline of the GWAS.

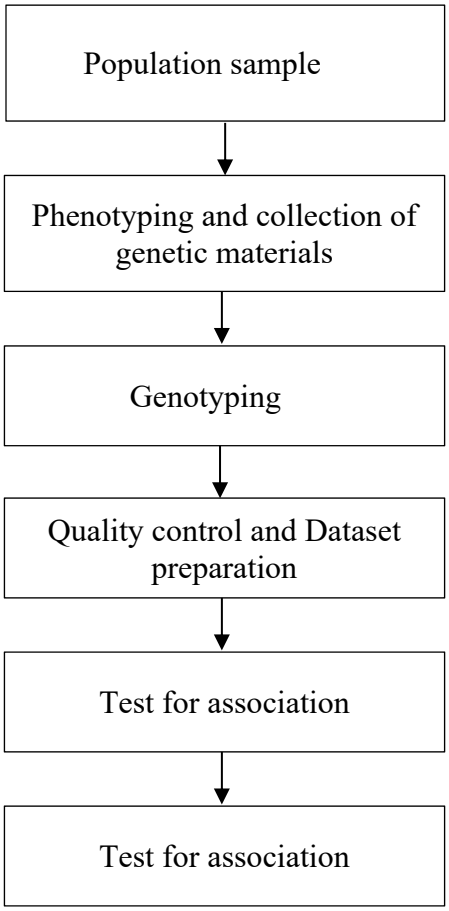


Figure 1.12: Overview of pipeline of the GWAS.

For humans, a typical example of GWAS in machine learning terminology is the association study where the response variable is a disease such as breast cancer, and the predictor variables (or features) are the SNPs (the single positions in the whole genome where the individuals vary by a single nucleotide A or T or G or C). The primary goal in GWAS is to identify all the SNPs that are relevant to the diseases or the observable traits (qualitative or quantitative). GWAS are characterized by high dimension and high-throughput. The human genome has roughly three billion chromosomal positions among which roughly three million positions are SNPs.

GWAS of quantitative traits like human height, growth related traits in plants and animals, or studies of mo1ecu1ar traits like gene expression have identified a large number of loci (Atwell et al., 2010; Bolormaa et al., 2011; Kim et al., 2013; Lee et al.,

2008; Tian et al., 2011; Visscher, 2008). Due to their study design, they have proven especially useful for detecting associations of common variants. Due to their polygeneticity of common human diseases the selective pressure on individual disease-causing mutations is assumed to be reduced, leading to the common diseases, common variant hypothesis (Lander, 1996; Pritchard and Cox, 2002; Reich and Lander, 2001) and the belief that GWAS should be especially useful for these kinds of diseases. The number of loci that have been reliably associated with heritable human diseases is close to nine thousand. Since the publication of the first successful GWAS (Ozaki et al., 2002), the number of publications on human GWAS has constantly been increasing each year (MacArthur et al., 2016; Struck et al., 2018). These GWA studies have detected tens of thousands of genetic variants which are statistically associated with human diseases (MacArthur et al., 2016; Struck et al., 2018).

The main goals of GWAS are -

(i) Discovering genetic markers or SNPs that are associated with a specific trait like disease, height, weight, etc.

(ii) To identify a significant portion of the DNA bases responsible for a disease or trait variability and to aid with disease prediction and prevention.

Broad-sense GWAS: In board sense, GWAS includes the followings areas based on the nature of the genotypic data:

(i) QTL mapping based GWAS

(ii) Transcript based GWAS
   (a) Differential expression (DE) analysis
   (b) Marker based eQTL analysis
   (c) SNP based eQTL analysis

(iii) SNP based GWAS

(iv) Sequence matching based GWAS

Modern-sense GWAS: In modern sense, genome-wide association studies (GWAS) is referred to the SNP-based GWAS.

1.3.1 QTL Mapping Based GWAS

The genome-wide identification of the chromosomal locations of quantitative trait loci (QTLs) along with their effects is very important for biomedical, evolutionary, animal, plant and agricultural genetics. QTL mapping aims to find the association between phenotypes and genotypes (i.e., genetic markers) in one or more chromosomal locations in the whole genome. That is, QTL mapping is the study for locating QTL in the genome by testing the association between phenotypes and genotypes using trait and marker genotype data obtained from certain populations (Chen, 2016c), e.g., backcross (BC) population, intercross (F_2) population, double haploid (DH) population, etc. Figure 1.13 shows the pipeline of the QTL mapping base GWAS.

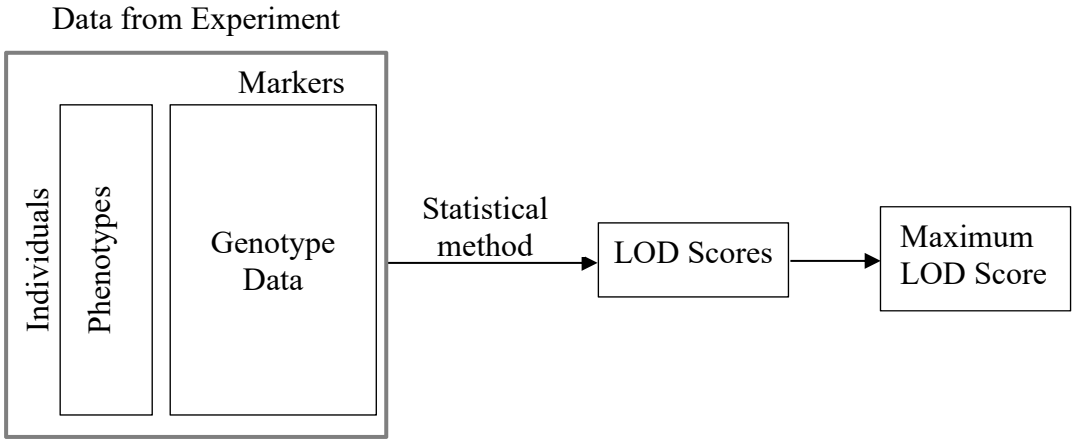


Figure 1.13: Flowchart of pipeline of QTL mapping based GWAS.

A variety of methods have been developed for QTL mapping (Hoeschele et al., 1997; Lynch and Walsh, 1998). These methods of QTL analysis can be classified as follows:

1. Single-marker analysis
   (i) t–test
   (ii) Analysis of variance
   (iii) Linear regression anlaysis
2. Interval mapping (Simple, Composite and Multiple interval mapping)
   (i) Maximum likelihood analysis (ML)

(ii) Least–squares analysis (LS)

(iii) Bayesian analysis.

These methods differ in computational requirements, efficiency in terms of extracting information, flexibility with regard to handling different data structures, and ability to map multiple QTLs. The methods of single-marker analysis are simple and efficient in terms of computational speed but cannot extract all information from the data and is restricted to specific mating designs. The technique of ML based simple interval mapping (SIM) (Lander and Botstein, 1989) is one of the most popular and widely used methods for QTL analysis in controlled crosses or structured pedigrees. However, ML-based interval mapping (IM) is time consuming because it uses the iterative expectation maximization (EM) algorithm. The least squares (LS) based SIM (Haley and Knott, 1992) is very efficient in terms of computation time and can be used for all popular mating designs. The ML and LS based SIM methods have been extended to composite interval mapping (CIM) by Zeng (Zeng, 1993, 1994a, 1994b) and multiple interval mapping (MIM) by Kao (Kao et al., 1999).

In this thesis, we will discuss only the SIM approaches for QTL mapping based GWAS. Based on the number of phenotypes to be considered in the SIM methods of QTL analysis, the SIM methods can be divided into two types:

(i) Single-trait simple interval mapping (i.e., Single-trait QTL anlaysis)

(ii) Multi-trait simple interval mapping (i.e., Multi-trait QTL analysis)

Single-trait Simple Interval Mapping: It searches a QTL within each interval between two adjacent markers on each chromosome, which affects/controls a single phenotypic trait, by performing likelihood ratio test (LRT) or F-test. The most popular and widely used interval mapping approaches are Maximum likelihood (Lander and Botstein, 1989) based SIM and Regression based SIM (Haley and Knott, 1992).

Multi-trait simple interval mapping: In many line crossing experiments of genome-wide QTL mapping studies, measurements are taken on multiple traits along with the marker genotypes. Most often, such traits are correlated with each other and there are

common chromosome regions (chromosomal locations) that affect multiple traits (Chen, 2016b). Although single-trait SIM methods can be applied to each trait one-by-one, such approaches do not take into account the pleotropic effects. A QTL is said to have pleiotropic effect if it simultaneously controls several phenotypic traits. The joint analyses of multiple traits, which include all quantitative traits together in a single model, can increase the power of QTL identification and improve the QTL localization accuracy when multiple traits are correlated genetically in the population (Xu, 2013a). In addition, QTL mapping considering multiple quantitative traits using joint analyses can give insights into the important genetic mechanisms underlying the trait relationships (e.g., genetic linkage versus pleiotropy), which would otherwise be hard to address if multiple traits are analyzed one by one. Therefore, statistical methods are very useful for joint analyses of multiple traits to identify important QTL locations, which control multiple traits simultaneously.

1.3.2 Transcript Based GWAS

In transcript based GWAS, the main activities is to find the differentially expressed genes (DEG) between two or more conditions from gene expression profiles or RNAseq profiles. Gene expression Gene expression is a process by which information from a gene is used in the synthesis of a functional gene product, which may be proteins (Anjum et al., 2016). Figure 1.14 represents the outline of the RNA-seq processing pipeline used to generate data for Expression Atlas. Differential expression (DE) analysis means taking the normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups (e.g., Control group versus Case group). In DE analysis, we use different statistical tests to decide whether, for a given gene, an observed difference in read counts between experimental groups is statistically significant. There are different statistical methods for DE analysis which can be divided into two types: (i) Classical methods and (ii) Bayesian methods. The classical approaches of DE analysis are divided into two types: (i) Parametric test and (ii) Non-parametric test. The most popular parametric classical approaches are t-test, F-test (ANOVA) (Kerr and Churchill, 2001) and likelihood ratio test (LRT) based on normal distribution, and DESeq (differential expression of sequence data) (Anders and Huber, 2010) and

edgeR (empirical analysis of digital gene expression in R) (Robinson et al., 2009) based on negative binomial (NB) distribution. The frequently used non-parametric approaches of DE analysis are Wilcoxon test (Wilcoxon, 1945), Kruskal Wallis test (Kruskal and Wallis, 1952) and significant analysis of microarrays (SAM) test (Tusher et al., 2001). The widely used Bayesian approaches based on a NB model are baySeq (Hardcastle and Kelly, 2010) and EBSeq (Leng et al., 2013), and the empirical Bayes approaches are linear models for microarrays (LIMMA) (Smyth, 2005), EBarrays (Kendziorski et al., 2003) and BRIDGE (Gottardo et al., 2006). It is crucial to consider the design of the experiment when choosing an analysis method for DE analysis. Whereas some of the DE analysis tools can only perform pair-wise comparison, the others such as edgeR, limma-voom (Law et al., 2014), DESeq and maSigPro (Conesa et al., 2006) can perform multiple comparisons.

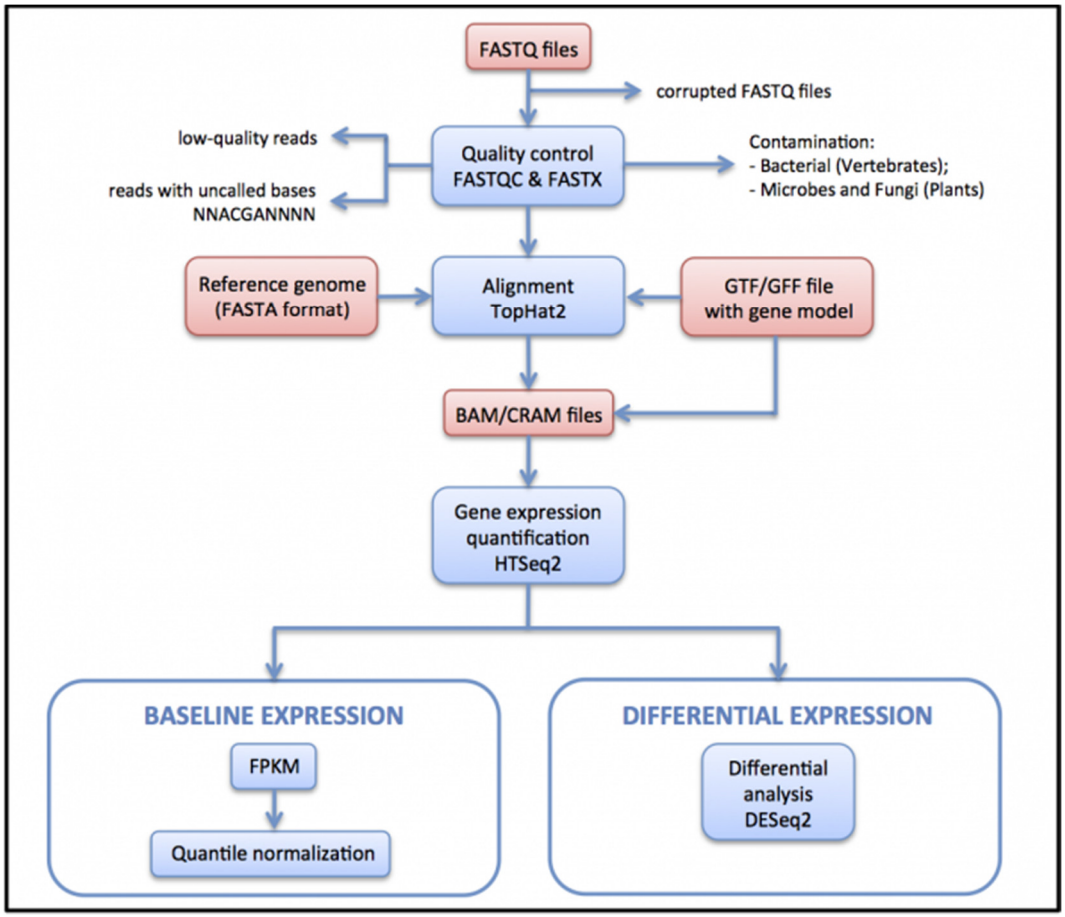


Figure 1.14: Pipeline of the RNA-seq processing used to generate gene expression data.

Marker based eQTL analysis: In maker based eQTL analysis, the methods of QTL analysis are used where the phenotypic data are the expression values of genes and the genotypic data are the marker data. Figure 1.13 shows the pipeline of the marker based eQTL mapping.

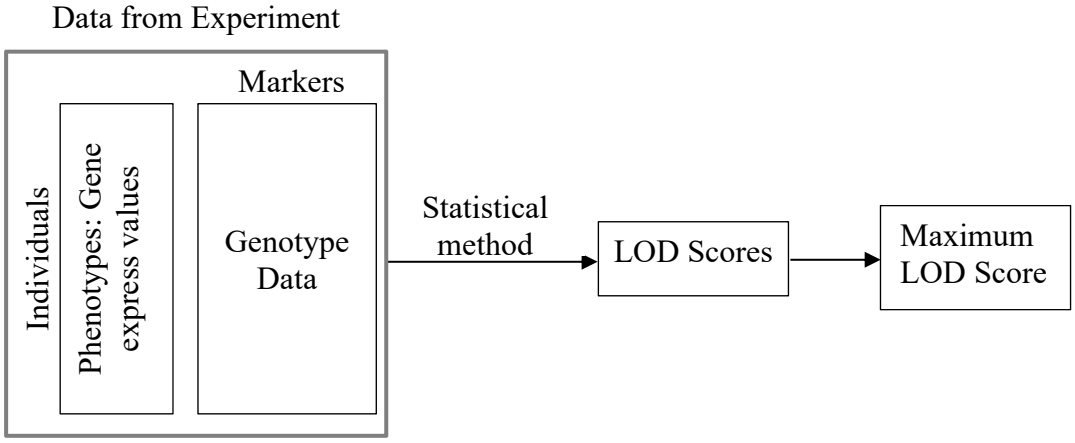


Figure 1.15: Flowchart of pipeline of QTL mapping based GWAS.

SNP based eQTL analysis: In maker based eQTL analysis, the phenotypic data are the expression values of genes and the genotypic data are the SNP data. All the statistical methods of SNP based GWAS with quantitative traits can be used for the SNP based eQTL analysis. Figure 1.14 represents the pipeline of SNP-based GWAS.

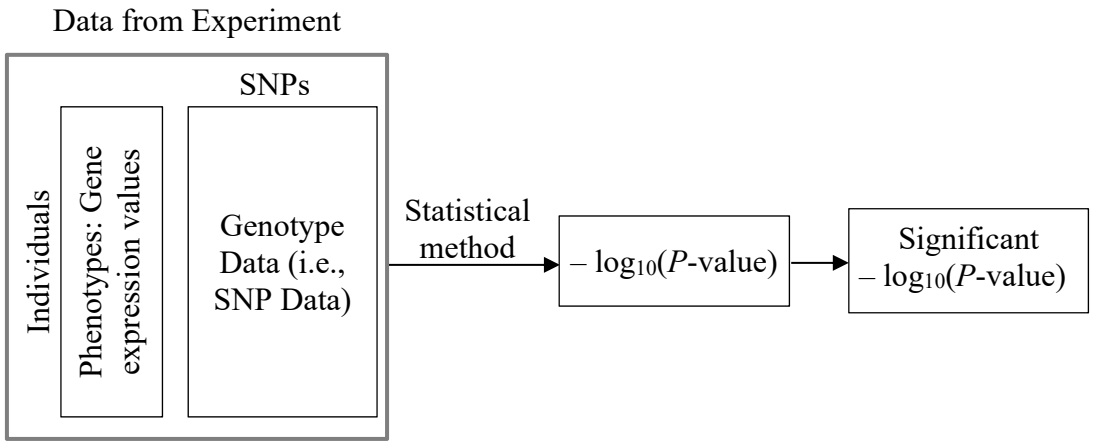


Figure 1.16: Flowchart of pipeline of SNP based GWAS.

In this thesis, we will not discuss any method of GWAS that uses gene expression data. More specifically, we will not discuss any method of DE analysis, marker based eQTL analysis and SNP based eQTL analysis in details in the next chapters of this thesis.

1.3.3 SNP Based GWAS

A SNP-based genome-wide association study (GWAS) is defined as the study of genetic variation across the whole genome of any organism that is designed to identify genetic associations between SNPs and observable traits of interest (such as height, weight, blood pressure, or presence or absence of a particular disease such as diabetes, cancer, or any specific condition). The primary goal in SNP based GWAS is to identify all the SNPs that are relevant to the diseases or the observable traits. Figure 1.17 represents the pipeline of SNP-based GWAS.

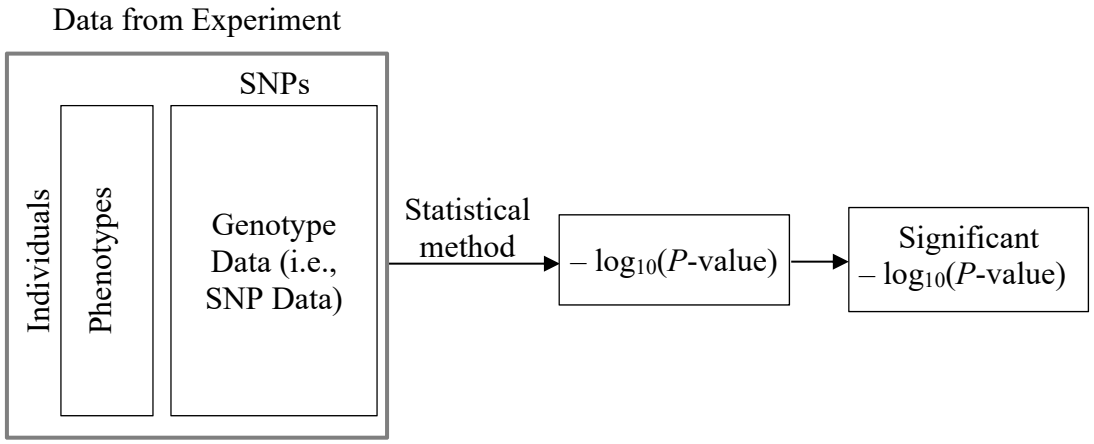


Figure 1.17: Flowchart of pipeline of SNP based GWAS.

The most popular and widely used approaches of SNP based GWAS are as follows:

a. Pearson χ^2 test for association between phenotype and genotype
b. Cochran Armitage test for trend in penetrances
c. The Transmission Disequilibrium Test (TDT) to test association using data from families with at least one affected child
d. SNP based GWAS using linear mixed models, e.g., Efficient Mixed-Model Association (EMMA)

In this thesis, we will discuss SNP based GWAS using simple linear regression

model, which is a special case of linear mixed models.

1.3.4 Sequence Matching Based GWAS

In sequence matching based GWAS, a particular sequence of interest (genomic sequence or coding sequence or protein sequence) of a gene is tried to match in the whole genome by searching the similar sequences in the whole genome stored in the databases. If the sequence of interest matches with any portion of the whole genome, then we called that the sequence is associated with that portion of the genome. In other words, in sequence matching based GWAS, similar sequences to a sequence of interest are searched in the whole genome stored in the databases. Then those similar sequences are said to be associated with the sequence of interest. After this, we select the sequence from the genome that is most associated with the sequence of interest and then we investigate the molecular mechanisms/functions of that selected sequence (i.e., most associated sequence). The molecular mechanisms/functions of that most associated sequence are treated as the molecular mechanisms/functions of the sequence of interest. The molecular mechanisms/functions that we usually interested in include gene structures, conserved domain containing, phylogenetic relationships, protein-protein interaction network, Gene Ontology (GO), transcription factors (TFs), gene-set enrichment, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and gene expression pattern. To explore a specific mechanism/function of a sequence we search the associated (i.e., similar) sequences in the whole genome stored in the database and find out the mechanism/function of the most associated (i.e., similar) sequence. The function of the most associated sequence is treated as the function of the selected sequence.

1.4 Literature Review on GWAS

QTL mapping based GWAS is used to detect the important QTLs and their locations genome wide which control specific phenotypic traits. It is widely used in agricultural and biomedical genetics. QTL mapping aims to find the association between phenotypes and genotypes (i.e., genetic markers) in one or more chromosomal locations in the whole genome. Thoday (1961) first proposed the idea of using two

markers to bracket a region for testing QTLs. Soller et al. (1976) examined the power of experiments at detecting linkage between a quantitative locus and a marker locus. Similar to Thoday (1961), but much improved method called simple interval mapping (SIM) or interval mapping (IM) approach was proposed by Lander and Botstein (1989) which is based on linkage relationships between a QTL and flanking markers. This method uses two adjacent markers to test the existence of a QTL within the interval by performing a likelihood ratio test (LRT) at every position in the interval. The LRT based SIM is time consuming and its calculations are complex. Least squares regression base SIM (Haley and Knott, 1992) is very efficient in terms of computation complexity and time. Liu (1997), Wu et al. (2007) and (Xu, 2013b) have discussed various techniques of QTL mapping based GWAS in their texts.

Often in many line crossing experiments of genome-wide QTL mapping studies, measurements are taken on multiple traits along with the marker genotypes. Very often, such traits are correlated and there are common chromosome regions (chromosomal locations) that affect multiple traits (Chen, 2016b). Several statistical methods for multi-trait QTL analysis have been developed in the literatures, ranging from simple extensions of single-trait QTL mapping approaches to sophisticated multi-trait QTL mapping approaches designed especially for multi-trait QTL analysis.

Several studies showed that some segments of genome (i.e., QTLs) affect different phenotypic traits simultaneously (Doebley and Stec, 1993; Edwards et al., 1992). In 1915, Jiang and Zeng first developed a statistical method for multi-trait QTL with F_2 population analysis, which consider multiple traits simultaneously (Jiang and Zeng, 1995). Korol et al. (1995) demonstrated the advantages of multi-trait QTL mapping within the framework of simple interval mapping where the correlation between trait complexes and the correlation between the QTLs were taken into account. Multi-point QTL analysis using variance-component linkage methods had been developed by Almasy and Blangero (1998) that can be used in the pedigrees of arbitrary complexity and size. They developed a general outline for the probability calculations with multi-point identity by descent (IBD). Mangin et al. (1998) proposed a dimension reduction method for multi-trait QTL analysis which consist of two different steps. First step is to extract the canonical variables that associated with the traits using the estimated

variance-covariance matrix of the traits and second step is to use the single-trait QTL mapping method to each canonical variable to obtain the combined results.

Many other substantial studies has been done in the field of multi-trait QTL mapping (Hackett et al., 2001; Henshall and Goddard, 1999; Knott and Haley, 2000; Korol et al., 2001; Williams et al., 1999). Least squares based multivariate regression (MVR-LS) for multi-trait SIM (Knott and Haley, 2000) and multi-trait SIM using expectation maximization (EM) algorithm (Dempster et al., 1977) based multivariate regression (Xu, 2013a) are two most popular and widely used approaches for multi-trait QTL analysis.

Single nucleotide polymorphism (SNP) based Genome-wide association studies (GWAS) has been widely used for the genetic study of a variety of species including humans, animals and plants to identify genomic locations/regions responsible for various quantitative traits, which has been made possible by decreasing the cost and time required to obtain sequences of whole genome and genome-wide SNPs. Since the publication of the first successful GWAS (Ozaki et al., 2002), the number of publications on human GWAS has constantly been increasing each year (MacArthur et al., 2016; Struck et al., 2018). These GWA studies have detected tens of thousands of genetic variants which are statistically associated with human diseases (MacArthur et al., 2016; Struck et al., 2018). A very large set of SNPs along with a very large number of accessions are simultaneously studied using different GWAS methods to uncover the significant relationship between genomic latent factors and phenotypic variations of interest (Zhao et al., 2011).

Population stratification (PS) is the main concerning issue when extensive genome-wide association analysis with numerous subjects is in consideration (Li and Yu, 2008; Liu et al., 2013; Xu et al., 2009). Some unidentified new population structures are probable to exist due to the large number of subjects that may be liable for systematic differences being selected in SNPs between cases and controls (Liu et al., 2013). Due to higher false discovery rates (FDRs), it is imperative to correct the observed population stratification in GWAS (Campbell et al., 2005; Liu et al., 2013).

The most commonly used statistical methods to avoid the bias of population stratification or genetic relatedness are genomic control (Devlin and Roeder, 1999), structured association (Pritchard et al., 2000), and principal component analysis (Patterson et al., 2006; Price et al., 2006). Genomic control (GC) approach modifies the association statistics by a common factor for all SNPs to correct for PS (Liu et al., 2013). Genomic control suffers from weak power when the effect of population structure is large (Aranzana et al., 2005; Devlin et al., 2001; Price et al., 2006; Yu et al., 2006; Zhao et al., 2007). Structured association analysis technique suggests locating the samples to discrete subpopulation clusters and then collecting evidence of association within each cluster (Pritchard et al., 2000). The SA method is useful for small datasets (http://pritch.bsd.uchicago.edu/software/structure2_1.html) (Liu et al., 2013). Nevertheless, the software package STRUCTURE is computationally intensive and cumbersome for large-scale genome-wide association studies (Price et al., 2006).

Another method based on principal component (PCA) is used for genome-wide association analysis (Price et al., 2006). In this technique, EIGENSTRAT program uses several top principal components (PCs) and applies them as covariates in GWA analysis (Liu et al., 2013). These top PCs are selected using EIGENSTRAT (Price et al., 2006) program based on PCA. Thousands of markers can be analyzed using this PCA method and the adjustment using PCA is definite to a marker's variation in allele frequency across ancestral populations (Liu et al., 2013; Price et al., 2006). PCA approach may however not more appropriate to correct population structure if it arises from the existence of several discrete subpopulations because PCA applies the produced eigenvectors as continuous covariates (Liu et al., 2013). The results obtained from PCA adjustment may be misleading too if there are outliers (Liu et al., 2013). Outlying data were introduced at genotypic level to check the performance of the robust PCA approach (Liu et al., 2013).

Another improved method was proposed to deal with the fact of PS for the presence of hidden population structure for population-based GWAS (Li and Yu, 2008). This method would improve PS by combining the multi-dimensional scaling (MDS) and clustering technique. This approach was however an extension of PCA due to having some similarity matrices between PCA and MDS. It can be applied for both discrete

and continuous population structures and it is well suited for large and small-scale GWA analysis (Li and Yu, 2008). In the recent bioinformatics research, the applications of linear mixed model (LMM) techniques have been popular in different genome-wide linkage analysis for discovery of potential biomarkers from human and agricultural single nucleotide polymorphism (SNP) level data. Nowadays to address the issues of adjustment of population stratification and account for population structure and genetic relatedness (polygenic effects) are effectively overcome by implementing LMM (Endelman, 2011; Kang et al., 2010; Zhang et al., 2010) for large scale GWAS. These approaches have been executed in software programs TASSEL (Bradbury et al., 2007), EMMA (Kang et al., 2008), EMMAX (Kang et al., 2010), rrBLUP (Endelman, 2011), Genome-wide efficient mixed model analysis (GEMMA) (Zhou and Stephens, 2012), GAPIT (Lipka et al., 2012).

1.5 Objectives of the Study

1.5.1 General Objective

The general objective of this study is the statistical modeling for genome wide association studies (GWAS) to identify important biomarker genes which are responsible for one/more particular traits of interest.

1.5.2 Specific Objectives

There are several statistical problems need to be solved for GWAS. To solve some statistical problems for GWAS, our specific objectives in this thesis are as follows:

(1) Regression based single-trait QTL analysis using the properties of bivariate normal distribution.
(2) Regression based fast multi-trait (FMT) QTL analysis using the properties of multivariate normal distribution.
(3) Robustification of regression based fast multi-trait QTL analysis.
(4) Robustification of regression based GWAS for detection of important SNPs.
(5) Sequence matching based GWAS for detection of important genes.

1.6 Layout of the thesis

This thesis contains seven chapters. We have organized the chapters sequentially to discuss different statistical methods that we have developed for GWAS. Figure 1.18 represents the structured layout of this thesis. The chapter wise summary of this thesis is given below:

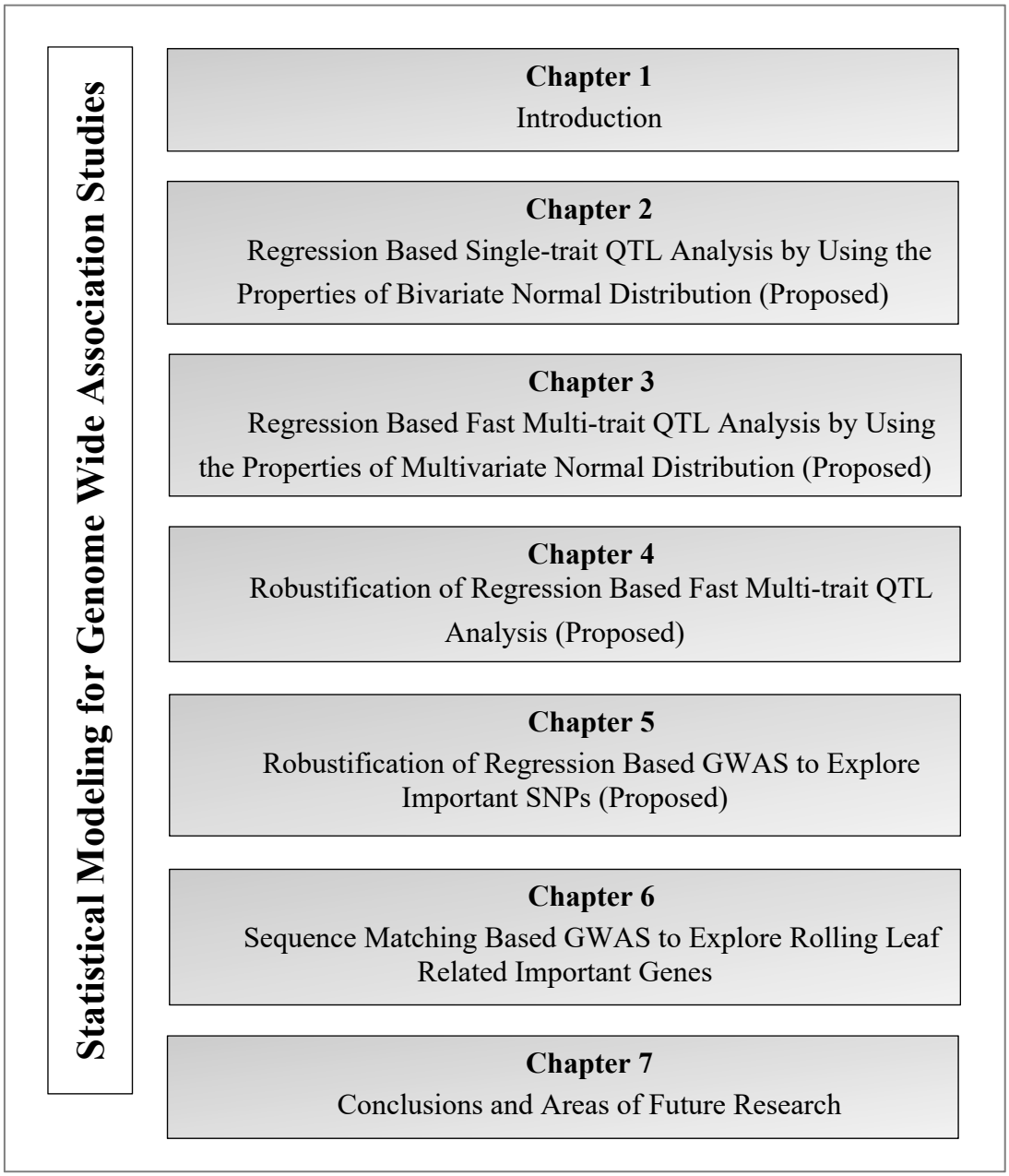


Figure 1.18: Layout of the thesis.

The present chapter (**Chapter 1**) is the introductory chapter which provides the basic concepts and importance of Genomics in plant science, animal science and human biology. This chapter begins with biological background and some important terminologies related to genomics, and at the end this chapter we have provided different objectives of our study along with the layout of this thesis.

In **Chapter 2**, we have introduced a new statistical approach for single-trait QTL analysis using the properties of bivariate normal distribution. The calculation of our new proposed method is very straight forward compared to the existing methods of single-trait QTL analysis. All the methods, including our new proposed method, of single-trait QTL mapping are verity sensitive to phenotypic outliers and these methods provide misleading results when the phenotypic data are contaminated by outliers. To overcome this problem, we have robustified our proposed method by robustifying the parameter of bivariate normal distribution using minimum β–divergence method. The performance of the proposed methods (Proposed1: Classical approach and Proposed2: Robust approach) have been investigated using simulation and real data analysis with the existing methods of single-trait QTL analysis.

In **Chapter 3**, we have discussed our new proposed statistical method for multi-trait QTL analysis, called "regression based fast multi-trait QTL analysis", using the properties of multivariate normal distribution. We have investigated the performance and computation time of the proposed method with the existing methods of multi-trait QTL analysis. Our proposed method is very faster in terms of computation time than the existing methods of multi-trait QTL analysis exhibiting almost the similar performance to the existing methods.

In **Chapter 4**, we have proposed a new robust approach for multi-trait QTL analysis by robustifying the "fast multi-trait QTL mapping" approach (discussed in Chapter 3) with the help of robust estimation of parameters of multivariate normal distribution using the β–divergence method. The performance of the proposed method has been investigated using both simulation and real data analysis in a comparison with the existing methods including the classical "fast multi-trait QTL mapping" approach.

In **Chapter 5**, we have introduced a robust statistical method for SNP-based GWAS by robustifying the least squares method of simple linear regression for SNP-based GWAS using minimum β–divergence method. We evaluated the performance of the proposed method using both simulation and real data analysis in a comparison with the classical least squares method of SNP-based GWAS.

In **Chapter 6**, we have performed sequence matching based genome-wide analysis to investigate the structural and functional mechanism of rolling leaf genes in rice (*Oryza sativa* L.). We have listed almost all the RL genes reported till date throughout several studies and performed different types of comparative and association analyses from different bioinformatics point of view including gene structure and exons/introns pattern analysis, domain analysis, phylogenetic analysis, Gene Ontology (GO) analysis, transcription factor (TF) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis, gene network analysis and gene expression analysis.

In **Chapter 7**, we have summarized all the results and findings throughout the whole study/thesis, and discussed the scope of the future research works.

Chapter 2

Regression Based Single-trait QTL Analysis by Using the Properties of Bivariate Normal Distribution (Proposed)

2.1 Introduction

The rapid advancement in molecular biology has increased the availability of fine scale genetic markers which facilitate the wide use of QTL analysis in the genetic study of quantitative traits in bioinformatics. Liu (1997) and Wu et al. (2007) discussed various techniques of QTL mapping in their texts. Thoday (1961) first proposed the idea of using two markers to bracket a region for testing QTLs. Soller et al. (1976) examined the power of experiments at detecting linkage between a quantitative locus and a marker locus. Similar to Thoday (1961), but much improved method called simple interval mapping (SIM) or interval mapping (IM) approach was proposed by Lander and Botstein (1989) which is based on linkage relationships between a QTL and flanking markers. This method uses two adjacent markers to test the existence of a QTL within the interval by performing a likelihood ratio test (LRT) at every position in the interval. Maximum likelihood (ML) based SIM (Lander and Botstein, 1989) and least squares (LS) regression based SIM (Haley and Knott, 1992) are two most popular and widely used interval mapping approaches. The LS regression based SIM is well known as HK regression based interval mapping to the biologists. The main limitation of ML based SIM is that its calculations are very complex and it is very time consuming because it uses the expectation-maximization

(EM) algorithm. Although LS based SIM takes less time than ML based SIM, its computations are also complex because parameter estimation depends on least squares method, and the calculation of test statistic needs calculation of residuals and residual variance.

In this chapter, we have developed a new approach of single-trait QTL analysis using the properties of bivariate normal distribution (BND). In this approach, the parameter estimation and calculation of test statistic is very straight forward because the calculations depend only the sample means, sample variances and sample covariances of phenotype and the conditional probability of QTL genotype given the flanking marker genotypes. Although our proposed BND based SIM are very useful methods for QTL analysis, it is very sensitive to phenotypic contaminations and provides misleading results when the phenotypic data are contaminated by outliers.

In this work, we have also developed a robust method of single-trait QTL analysis with backcross (BC) population by robustifying our proposed BND based SIM using minimum β–divergence method. We have performed a simulation study to investigate the performance of the proposed methods in comparison with the existing methods of QTL analysis for BC population. Although we have developed our proposed method with BC population, this method can easily be extended for other populations, such as double haploid (DH) and intercross (F_2) population, with some simple modification.

2.2 Methods and Materials

Let us consider no epistatis between two QTLs, no interference in crossing over, and only one QTL in the testing interval. The fixed effect model for Backcross (BC) population, for testing a QTL within a marker interval, is define as

$$y_j = \alpha + \gamma x_{j|i} + \varepsilon_j, i = 1, 2 \text{ and } j = 1, 2, \dots, n \qquad (2.1)$$

where y_j is the phenotypic value of the j-th individual, α is the general mean effect, $x_{j|i} = p_{j|1}$, γ is the QTL additive effect and $\varepsilon_j \sim NID(0, \sigma^2)$ is a random error. Here, $x_{j|i}$ is the conditional probability for QTL genotypes given the flanking marker genotypes. Since conditional expectation is equivalent to conditional probabilities of

QTL genotypes, $x_{j|i}$ is fixed for QTL genotypes given flanking marker genotypes. Since $x_{j|i}$ is fixed, so this model is called fixed effect model.

The conditional probabilities for QTL genotypes QQ and Qq given the flanking marker genotypes are denoted by $p_{j|1}$ and $p_{j|2}$, respectively. The conditional probabilities $p_{j|1}$ and $p_{j|2}$ are shown in Table 2.1 for Backcross population. In Table 2.1, p is defined as $p = r_{MQ}/r_{MN}$ where r_{MQ} is the recombination fraction between the left marker M and the putative QTL and r_{MN} is the recombination fraction between two flanking markers M and N. The possibility of a double recombination event in the interval is ignored.

Table 2.1: Conditional Probabilities of a putative QTL genotype given the flanking marker genotypes for a backcross population

Marker genotypes	Expected frequency	QTL genotypes	
		QQ ($p_{j\|1}$)	Qq ($p_{j\|2}$)
MN/*MN*	$(1-r)/2$	1	0
MN/*Mn*	$r/2$	$(1-p)$	p
MN/*mN*	$r/2$	p	$(1-p)$
MN/*mn*	$(1-r)/2$	0	1

To investigate the existence of a QTL at a given position within a marker interval, we want to test the hypothesis $H_0: \gamma = 0$ (i.e., there is no QTL at a given position) versus H_1: H_0 is not true.

Under null hypothesis (H_0), the model (2.1) reduces to the following model

$$y_j = \alpha + \varepsilon_j, j = 1, 2, \ldots\ldots, n \qquad (2.2)$$

which is called reduced model.

2.2.1 Maximum likelihood (ML) Based Classical Simple IM Approach for Single-trait QTL Analysis

Under the normality assumption of error, the probability density function of the trait value (y) within each QTL genotype class is normal with mean $(\alpha + \gamma x_{j|i})$ and variance σ^2, i.e., $(y_j|x_{j|i}) \sim N(\alpha + \gamma x_{j|i}, \sigma^2)$. Then the likelihood function for the parameters $\boldsymbol{\theta} = (\alpha, \gamma, \sigma^2)$ can be written as follows

$$L(\theta|Y) = \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_j - \alpha - \gamma x_{j|i}}{\sigma}\right)^2\right] \tag{2.3}$$

To test H_0 against H_1, the likelihood ratio test (LRT) statistic is defined as

$$\text{LRT} = -2\ln\left[\frac{\sup_{\Theta_0} L(\boldsymbol{\theta}|Y)}{\sup_{\Theta} L(\boldsymbol{\theta}|Y)}\right] \tag{2.4}$$

where Θ_0 and Θ_0 are the restricted (H_0) and unrestricted (H_1) parameter spaces.

The threshold value to reject the null hypothesis cannot be simply chosen from a chi-square distribution because of the violation of regularity conditions of asymptotic theory under H_0. The number and size of intervals should be considered in determining the threshold value. Since multiple tests are performed in mapping, the hypotheses are usually tested at every position of an interval and for all intervals of the genome to produce a continuous LRT statistic profile. At every position, the position parameter p is predetermined and only α, γ and σ^2 are involved in estimation and testing. If the tests are significant in a chromosomal region, the position with the largest LRT statistic is inferred as the estimate of the QTL position and the maximum likelihood estimates (MLEs) at this position are the estimates of α, γ and σ^2 obtained by iterative way.

An alternative way is to use log of odds (LOD) score (Lander and Botstein, 1989; Ott, 1999; Terwilliger and Ott, 1994; Wu et al., 2007; Xu, 2013d) as a test statistic to test the null hypothesis of no QTL (H_0). The LOD score is the transformation of the LRT statistic, defined as

$$\text{LOD} = \frac{\text{LRT}}{2\times \log(10)} = \frac{\text{LRT}}{4.605} = 0.217\ \text{LRT} \tag{2.5}$$

According Lander and Botstein (1989), the typical threshold of LOD score should be between 2 and 3 to ensure a 5% overall false positive error for identifying a QTL. Terwilliger and Ott (1994), Ott (1999), Wu et al. (2007), and Xu (2013d) suggested a value of LOD = 3 as the critical threshold for declaring the existence of QTL. Thus, the LOD > 3 can be used as a criterion to declare a significant QTL.

The MLEs of the parameters α, γ, and σ^2 are as follows

$$\hat{\alpha} = \bar{y} - \hat{\gamma}\bar{x}, \hat{\gamma} = \frac{\sum_{j=1}^{n}(x_{j|i} - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_{j|i} - \bar{x})^2} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{\alpha} - \hat{\gamma}x_{j|i})^2 \tag{2.6}$$

where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ and $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_{j|i}$, $i = 1, 2$.

Obviously these ML estimates of α, γ and σ^2 are very much sensitive to outliers. Therefore, regression analysis by maximum likelihood estimate (MLE) produces misleading results in presence of contaminated data.

2.2.2 Least Squares (LS) regression Based Classical SIM Approach for Single-trait QTL Analysis

Using (2.1), the error sum of squares (ESS) can be written as

$$\text{ESS} = \sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} (y_j - \alpha - \gamma x_{j|i})^2 \tag{2.7}$$

The least squares estimates of the regression parameters (α and γ) can be obtained by minimizing the ESS with respect to α and γ.

The LS estimates of the regression parameters α and γ are as follows.

$$\hat{\alpha} = \bar{y} - \hat{\gamma}x_{j|i} \text{ and } \hat{\gamma} = \frac{\sum_{j=1}^{n}(x_{j|i} - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_{j|i} - \bar{x})^2} \tag{2.8}$$

where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ and $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_{j|i}$, $i = 1, 2$.

Then the LS estimate of the residual variance is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{j=1}^{n}\left(y_j - \hat{\alpha} - \hat{\gamma}x_{j|i}\right)^2$$

According to Draper and Smith (1998) and Haley and Knott (1992), least squares (LS) regression is equivalent to maximum likelihood when the errors are independently and normally distributed. In this case, the likelihood ratio test can be written in terms of the residual sum of squares under the full model (2.1) (RSS_{full}), the reduced model (2.2) ($RSS_{reduced}$), and the number of observations (n) as follows:

$$\text{LRT} = n \ln\left(\frac{\text{RSS}_{\text{reduced}}}{\text{RSS}_{\text{full}}}\right) \tag{2.9}$$

where RSS_{full} and $RSS_{reduced}$ are the residual sum of squares under the full model (i.e., under H_1) and the reduced model (i.e., under H_0), respectively, and n is the number of observations (Haley and Knott, 1992). This test statistic is approximately distributed as a χ^2-variate with 1 degrees of freedom (Atkin et al., 1989; Haley and Knott, 1992; Haley et al., 1994). The LOD statistic can be obtained by transforming the LRT statistic as (2.4).

2.2.3 Simple Interval Mapping (SIM) approach for Single-trait QTL Analysis Using BND (Proposed1)

In order to estimate the model parameters and the variance of the random error, let us consider that $\boldsymbol{Z} = (Y, X)$ follows a bivariate normal distribution $N\left(\underset{(2\times1)}{\boldsymbol{\mu_Z}}, \underset{(2\times2)}{\boldsymbol{\Sigma_Z}}\right)$ with mean vector $\boldsymbol{\mu_Z}$ and variance-covariance matrix $\boldsymbol{\Sigma_Z}$, where Y and X have been introduced in (2.1).

Then the probability density function for $\boldsymbol{Z} = (Y, X)$ can be written as

$$f(\boldsymbol{Z}) = \frac{1}{(2\pi)|\boldsymbol{\Sigma_Z}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{Z} - \boldsymbol{\mu_Z})^T \boldsymbol{\Sigma_Z^{-1}}(\boldsymbol{Z} - \boldsymbol{\mu_Z})\right] \tag{2.10}$$

We can partition the mean vector $\boldsymbol{\mu_Z}$ as $\boldsymbol{\mu_Z} = [\mu_Y \ \ \mu_X]^T$ and the covariance matrix $\boldsymbol{\Sigma_Z}$ as

$$\boldsymbol{\Sigma_Z} = \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_X^2 \end{bmatrix},$$

where $\sigma_X^2 = E[(X - \mu_X)^2]$, $\sigma_Y^2 = E[(Y - \mu_Y)^2]$ and $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$.

Then the conditional mean of Y given X is obtained as

$$E(Y|X = x) = \mu_Y + \sigma_{YX}\sigma_X^{-2}(X - \mu_X) \tag{2.11}$$

Equation (2.11) can be expressed as

$$\begin{aligned}(Y|X = x) &= \mu_Y + \sigma_{YX}\sigma_X^{-2}X - \sigma_{YX}\sigma_X^{-2}\mu_X \\ &= (\mu_Y - \sigma_{YX}\sigma_X^{-2}\mu_X) + (\sigma_{YX}\sigma_X^{-2})X \\ &= \alpha + \gamma X \end{aligned} \tag{2.12}$$

which is known as simple linear regression surface of Y on X, where $\alpha = (\mu_Y - \sigma_{YX}\sigma_X^{-2}\mu_X)$ is the general mean effect and the (m×1) vector $\gamma = (\sigma_{YX}\sigma_X^{-2})$ is called the regression coefficient. For BC population γ is the additive QTL effects.

Using (2.11), the prediction error can be written as

$$\varepsilon = Y - E(Y|X) = Y - \mu_Y - \sigma_{YX}\sigma_X^{-2}(X - \mu_X) \tag{2.13}$$

Now, the variance of the prediction error is

$$\sigma^2 = V(\varepsilon) = E[\{\varepsilon - E(\varepsilon)\}^2] = E[\varepsilon^2], \text{ since } E(\varepsilon) = 0 \tag{2.14}$$

Using (2.13) in (2.14), we can write

$$\begin{aligned}\sigma^2 &= E[\{Y - \mu_Y - \sigma_{YX}\sigma_X^{-2}(X - \mu_X)\}^2] \\ &= E[(Y - \mu_Y)^2 - 2(Y - \mu_Y)\{\sigma_{YX}\sigma_X^{-2}(X - \mu_X)\} + \{\sigma_{YX}\sigma_X^{-2}(X - \mu_X)\}^2] \\ &= E[(Y - \mu_Y)^2] - 2\sigma_{YX}\sigma_X^{-2}E[(Y - \mu_Y)(X - \mu_X)] + \sigma_{YX}^2\sigma_X^{-4}E[(X - \mu_X)^2] \\ &= \sigma_Y^2 - 2\sigma_{YX}\sigma_X^{-2}\sigma_{YX} + \sigma_{YX}^2\sigma_X^{-4}\sigma_X^2 \\ &= \sigma_Y^2 - 2\sigma_{YX}^2\sigma_X^{-2} + \sigma_{YX}^2\sigma_X^{-2} \\ &= \sigma_Y^2 - \sigma_{YX}^2\sigma_X^{-2} \end{aligned} \tag{2.15}$$

Because $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ are typically unknown, they must be estimated from a random sample in order to construct the multivariate linear predictor and determine expected prediction errors.

Based on a random sample of size *n*, the maximum likelihood estimator of the $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ are given by

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}} = \begin{bmatrix} \hat{\mu}_Y \\ \hat{\mu}_X \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \bar{X} \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\sigma}_{YX} \\ \hat{\sigma}_{XY} & \hat{\sigma}_X^2 \end{bmatrix} = \left(\frac{n-1}{n}\right) \begin{bmatrix} S_Y^2 & S_{YX} \\ S_{XY} & S_X^2 \end{bmatrix} \tag{2.16}$$

where $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j$, $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} y_j$, $S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{Y}\right)^2$, $S_{XY} = S_{YX} = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{Y}\right)\left(x_j - \bar{X}\right)$ and $S_X^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^2$.

Hence, based on a random sample of size *n*, we can get the maximum likelihood estimators of the regression parameters α and γ, and the error variance σ^2.

Using (2.16) into (2.12), we can write

$$\hat{\alpha} = (\hat{\mu}_Y - \hat{\sigma}_{YX}\hat{\sigma}_X^{-2}\hat{\mu}_X) = \bar{Y} - S_{YX}S_X^{-2}\bar{X} \tag{2.17}$$

and

$$\hat{\gamma} = (\hat{\sigma}_{YX}\hat{\sigma}_X^{-2}) = S_{YX}S_X^{-2} \tag{2.18}$$

Therefore, using (2.17) and (2.18) in (2.12), the maximum likelihood estimator of the regression function is

$$\hat{Y} = \hat{\alpha} + \hat{\gamma}X = \bar{Y} - S_{YX}S_X^{-2}\bar{X} + S_{YX}S_X^{-2}X = \bar{Y} + S_{YX}S_X^{-2}(X - \bar{X}) \tag{2.19}$$

Based on a random sample of size *n*, using (2.16) in (2.15), the maximum likelihood estimators of σ^2 under the full model and the reduced model are, respectively,

$$\hat{\sigma}^2 = \hat{\sigma}_Y^2 - \hat{\sigma}_{YX}^2\hat{\sigma}_X^{-2} = \left(\frac{n-1}{n}\right)(S_Y^2 - S_{YX}^2S_X^{-2}) \tag{2.20}$$

and

$$\hat{\sigma}_0^2 = \hat{\sigma}_Y^2 = \left(\frac{n-1}{n}\right)S_Y^2 \tag{2.21}$$

Let $L_1(\alpha, \gamma, \sigma^2)$ is the likelihood function under the full model (2.1) and $L_0(\alpha, \sigma^2)$ is the likelihood function under the reduced model (2.2). To test H_0 against H_1, the likelihood ratio test (LRT) statistic is defined as

$$\text{LRT} = -2 \ln \left[\frac{\max_{\alpha,\sigma^2} L_0(\alpha, \sigma^2)}{\max_{\alpha,\gamma,\sigma^2} L_1(\alpha, \gamma, \sigma^2)} \right]$$

$$= -2 \ln \left[\frac{L_0(\hat{\alpha}_0, \hat{\sigma}_0^2)}{L_1(\hat{\alpha}, \hat{\gamma}, \hat{\sigma}^2)} \right] = -n \ln \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right) \quad (2.22)$$

where $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\sigma}^2$ are the maximum likelihood (ML) estimates of the parameters α, γ and σ^2 under the full model (2.1), and $\hat{\alpha}_0$ and $\hat{\sigma}_0^2$ are the ML estimates of the parameters α and σ^2 under the reduced model (i.e., under H_0).

Under the null hypothesis (H_0), the LRT statistic in (2.22) is expected to have an approximate chi-square distribution with 1 degrees of freedom for a given QTL position in the genome. However, the threshold value to reject the null hypothesis (H_0) cannot be simply chosen from the χ^2 distribution because of the violation of regularity conditions of asymptotic theory under H_0.

An alternative way is to use log of odds (LOD) score (Lander and Botstein, 1989; Ott, 1999; Terwilliger and Ott, 1994; Wu et al., 2007; Xu, 2013d) as a test statistic to test the null hypothesis of no QTL (H_0). The LOD score is the transformation of the LRT statistic, defined as

$$\text{LOD} = \frac{\text{LRT}}{2 \times \log(10)} = \frac{\text{LRT}}{4.605} = 0.217 \text{ LRT} \quad (2.23)$$

According Lander and Botstein (1989), the typical threshold of LOD score should be between 2 and 3 to ensure a 5% overall false positive error for identifying a QTL. Terwilliger and Ott (1994), Ott (1999), Wu et al. (2007), and Xu (2013d) suggested a value of LOD = 3 as the critical threshold for declaring the existence of QTL. Thus, the LOD > 3 can be used as a criterion to declare a significant QTL.

2.2.4 Robust SIM approach for Single-trait QTL Analysis by Robust Estimation of BND (Proposed2)

All the approaches discussed in previous sections are very sensitive to phenotypic outliers and produce misleading results in presence of outliers. So, we need some robust approach which produce similar results in absence of outliers and perform better in presence of outliers being less sensitive to outliers. We observe that the estimates in (2.16) – (2.22) are very sensitive to outliers and give misleading results in presence of outliers. In this section, we have discussed the robustification of the estimates in (2.16) – (2.22) using β–divergence method (Mihoko and Eguchi, 2002; Mollah et al., 2007) to obtain the robust estimates of model parameters and the robust test statistics (LRT and LOD). From (2.16) – (2.22) we observe that if we can robustify the sample means, sample variances and sample covariance, then we can obtain the robust estimates of the model parameters and the test statistics (LRT and LOD).

According to (Mihoko and Eguchi, 2002; Mollah et al., 2007), the β-divergence between two probability density functions $p(\boldsymbol{z})$ and $q(\boldsymbol{z})$ is defined by

$$D_\beta(p,q) = \int \left[\frac{1}{\beta}\{p^\beta(z) - q^\beta(\boldsymbol{z})\}p(\boldsymbol{z}) - \frac{1}{\beta+1}\{p^{\beta+1}(\boldsymbol{z}) - q^{\beta+1}(\boldsymbol{z})\}\right] d\boldsymbol{z},\ for\ \beta > 0 \tag{2.24}$$

which is non-negative, that is $D_\beta(p,q) \geq 0$, equality holds iff $p = q$.

The minimum β-divergence estimators of the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu_Z}, \boldsymbol{\Sigma_Z})$ can be obtained by the iterative solution of the following equations:

$$\boldsymbol{\mu}_{\boldsymbol{z},\,t+1} = \frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)\boldsymbol{z}_j}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)} \tag{2.25}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{Z},\,t+1} = (1+\beta)\frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},\,t})(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},\,t})^T}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)} \tag{2.26}$$

where $w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t), j = 1, 2, \ldots, n$, is called the β-weight function and defined as

$$w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t) = \exp\left[-\frac{\beta}{2}(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},t})^T \boldsymbol{\Sigma}_{\boldsymbol{Z},t}^{-1}(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},t})\right].$$

If $\beta \to 0$, then (2.25) and (2.26) reduces to the classical non-iterative solution.

Let the robust estimates (i.e., β-estimates) of $\boldsymbol{\mu}_{\boldsymbol{Z}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ are denote by $\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)}$. Then we can write

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \hat{\mu}_{Y(\beta)} \\ \hat{\mu}_{X(\beta)} \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \hat{\sigma}^2_{Y(\beta)} & \hat{\sigma}_{YX(\beta)} \\ \hat{\sigma}_{XY(\beta)} & \hat{\sigma}^2_{X(\beta)} \end{bmatrix} \tag{2.27}$$

Then the robust estimates of the regression parameters can be written as

$$\hat{\alpha}_{(\beta)} = \left(\hat{\mu}_{Y(\beta)} - \hat{\sigma}_{YX(\beta)}\hat{\sigma}^{-2}_{X(\beta)}\hat{\mu}_{X(\beta)}\right) \tag{2.28}$$

and

$$\hat{\gamma}_{(\beta)} = \left(\hat{\sigma}_{YX(\beta)}\hat{\sigma}^{-2}_{X(\beta)}\right) \tag{2.29}$$

Now, the robust estimates of σ^2 under the full model and the reduced model are, respectively,

$$\hat{\sigma}^2_{(\beta)} = \hat{\sigma}^2_{Y(\beta)} - \hat{\sigma}^2_{YX(\beta)}\hat{\sigma}^{-2}_{X(\beta)} \tag{2.30}$$

and

$$\hat{\sigma}^2_{0(\beta)} = \hat{\sigma}^2_{Y(\beta)} \tag{2.31}$$

Then we get the robust LRT statistic as follows:

$$\text{LRT}_{(\beta)} = -n \ln\left(\frac{\hat{\sigma}^2_{(\beta)}}{\hat{\sigma}^2_{0(\beta)}}\right) \tag{2.32}$$

The modified LRT statistic has an approximate χ^2-distribution with 1 degrees of freedom. Then the robust LOD statistic can be written as

$$\mathrm{LOD}_{(\beta)} = \frac{\mathrm{LRT}_{(\beta)}}{2\times \log(10)} = \frac{\mathrm{LRT}_{(\beta)}}{4.605} = 0.217\ \mathrm{LRT}_{(\beta)} \tag{2.33}$$

We have develop the proposed method for BC population. However, methods for other mapping populations, such as F_2 and double haploid (DH), are simple extension of that for the BC population with some modifications.

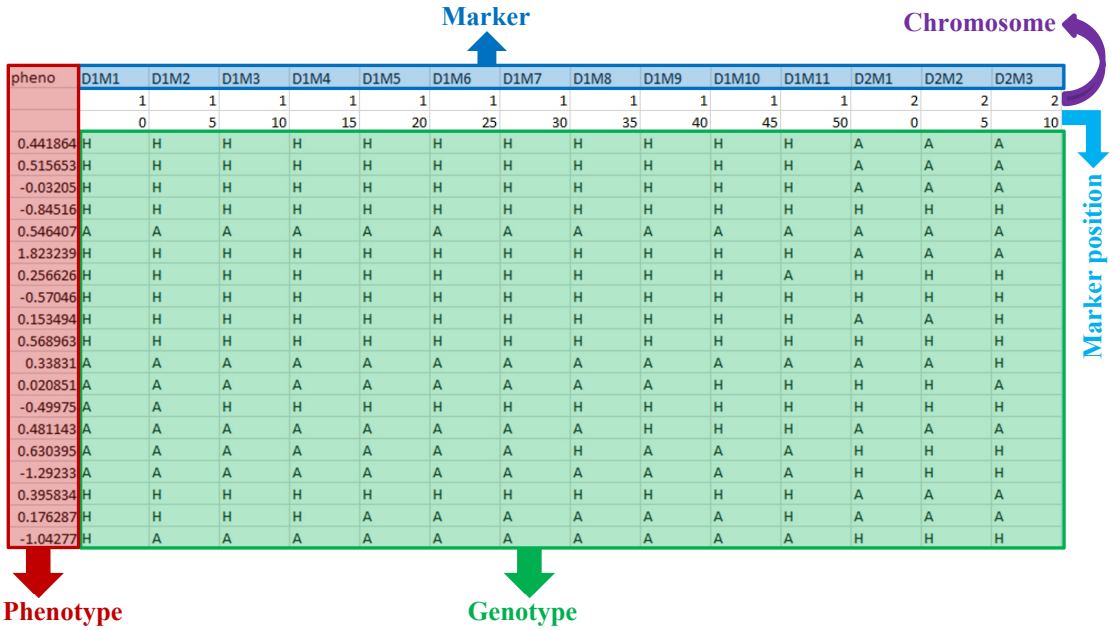
2.3 Results and Discussion

2.3.1 Simulation Results

To measure the performance of the proposed method in comparison of the maximum likelihood (ML), least squares (LS) and bivariate normal distribution (BND) for QTL mapping with Backcross population, we have generated phenotypic and genotypic data with Backcross population using simulation technique. We have considered three unlinked QTLs, total 10 chromosomes and 11 equally spaced markers in each of the 10 chromosomes, where any two successive marker interval size is 5 cM. The true QTL positions are located on chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM). The true values of the parameters in the model are assumed as $\alpha = 0.5$, $\gamma = 0.8$ and $\sigma^2 = 0.25$. We have generated 300 trait values with heritability $h^2 = 0.39$ which means that 39% of the trait variation is controlled by QTL and the remaining 61% is subject to the environmental effects (random error). To investigate the robustness of the proposed method in a comparison of the ML, LS and BND methods, we contaminated 12% of the trait values (i.e., phenotypic values) in this dataset by outliers. Figure 2.1 shows the structure of the dataset obtained from a genome-wide QTL experiment for single-trait QTL analysis. To perform the simulation study we have used R/qtl software (Broman et al. (2003), homepage: http://www.rqtl.org/).

Table 2.2 shows QTL positions (i.e., chromosome, marker and locus position) identified by ML, LS, BND and the proposed method. Figure 2.2(a) and Figure 2.2(b) are representing the scatter plots of 300 trait values in presence and absence of outliers, respectively. Then we computed LOD scores based on ML, LS, BND and the proposed methods for both types of data sets (uncontaminated and contaminated). Figure 2.2(c) and Figure 2.2(d) are showing the LOD scores profile plots for the

uncontaminated and contaminated datasets, respectively. In the LOD scores profile plots the dotted (red colour), two dash (green colour), dot dash (blue colour) and solid (black colour) lines represent the LOD scores at every 1cM position in the chromosomes for ML, LS, BND and the proposed method with $\beta = 0.2$, respectively.



pheno	D1M1	D1M2	D1M3	D1M4	D1M5	D1M6	D1M7	D1M8	D1M9	D1M10	D1M11	D2M1	D2M2	D2M3
	1	1	1	1	1	1	1	1	1	1	1	2	2	2
	0	5	10	15	20	25	30	35	40	45	50	0	5	10
0.441864	H	H	H	H	H	H	H	H	H	H	H	A	A	A
0.515653	H	H	H	H	H	H	H	H	H	H	H	A	A	A
-0.03205	H	H	H	H	H	H	H	H	H	H	H	A	A	A
-0.84516	H	H	H	H	H	H	H	H	H	H	H	H	H	H
0.546407	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1.823239	H	H	H	H	H	H	H	H	H	H	H	A	A	A
0.256626	H	H	H	H	H	H	H	H	H	H	A	H	H	H
-0.57046	H	H	H	H	H	H	H	H	H	H	H	H	H	H
0.153494	H	H	H	H	H	H	H	H	H	H	H	A	A	H
0.568963	H	H	H	H	H	H	H	H	H	H	H	H	H	H
0.33831	A	A	A	A	A	A	A	A	A	A	A	A	A	H
0.020851	A	A	A	A	A	A	A	A	A	H	H	H	H	A
-0.49975	A	A	H	H	H	H	H	H	H	H	H	H	H	H
0.481143	A	A	A	A	A	A	A	A	H	H	H	A	A	A
0.630395	A	A	A	A	A	A	A	H	A	A	A	H	H	H
-1.29233	A	A	A	A	A	A	A	A	A	A	A	H	H	H
0.395834	H	H	H	H	H	H	H	H	H	H	H	A	A	A
0.176287	H	H	H	H	A	A	A	A	A	A	H	A	A	A
-1.04277	H	A	A	A	A	A	A	A	A	A	A	H	H	H

Figure 2.1: Structure of the Dataset obtained from a genome-wide QTL experiment for single-trait QTL analysis.

Table 2.2: QTL positions identified by each method in absence and absence of outliers

Method	**True QTL position**	**Identified QTL position**	
		In absence of outliers	**In presence of outliers**
ML	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	ML method fails to identify any QTL on any chromosome.
LS	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	LS method fails to identify any QTL on any chromosome.

Method	True QTL position	Identified QTL position	
BND	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	(i) On chromosome 3 at marker 8 (locus position 35 cM) (ii) On chromosome 5 at marker 5 (locus position 20 cM) (iii) On chromosome 6 at marker 2 (locus position 5 cM) (iv) On chromosome 8 at marker 3 (locus position 10 cM)
Proposed (Robust BND)	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.

From Table 2.2 and Figure 2.2 it is seen that the highest LOD score peak occurs at the true QTL position on the true chromosome 2, 3 and 5 at marker 5 (locus position 20 cM) for all four methods for the uncontaminated dataset (Figure 2.2(c)). However, in presence of outliers, the highest LOD score peak occurs at the true QTL positions on true chromosomes for the proposed method only (Figure 2.2(d)). That is, from Table 2.2 and Figure 2.2 we observe that all of the four methods (ML, LS, BND and proposed method) identify the true QTL positions correctly in absence of outliers. But in presence of outliers the ML and LS fail to identify any significant QTL position and the BND identify QTLs on chromosomes 3 at marker 8 (locus position 35 cM), on chromosome 5 at marker 5 (locus position 20 cM), on chromosome 6 at marker 2 (locus position 5 cM) and on chromosome 8 at marker 3 (locus position 10 cM). In presence of outliers, inly the position on chromosome 5 at marker 5, identified by BND, is the true QTL position, and all other positions identified by BND are not the true position of QTLs. However, in presence of outliers, the proposed method (robust BND) have identified the QTLs on chromosome 2, 3 and 5 at marker 5 (locus position 20 cM) which are the true QTL positions.

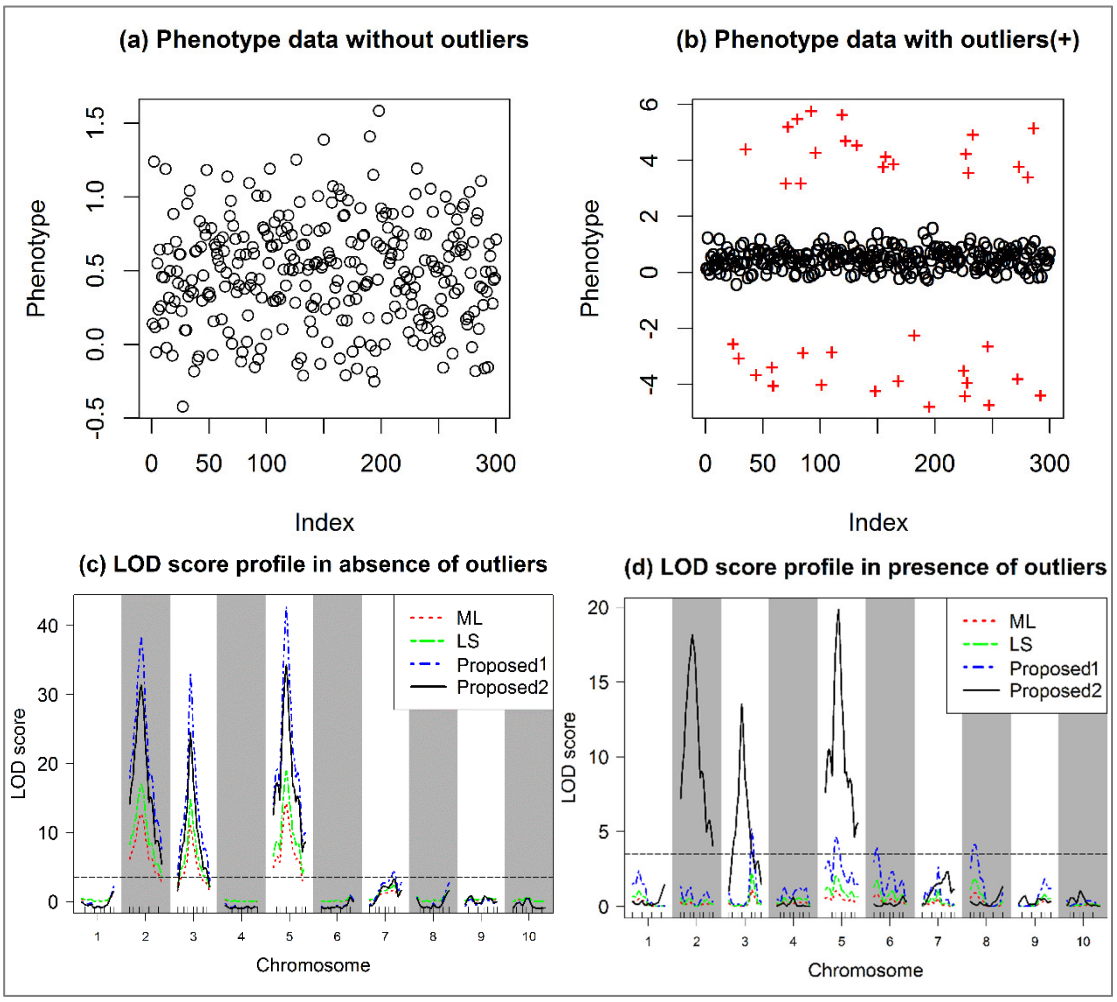


Figure 2.2: Simulated phenotypic observations in (a) absence and (b) presence of 12% outliers, and LOD score profile in (c) absence and (d) in presence of 12% outliers.

Hence, in presence of outliers, the classical methods of SIM (ML, LS and BND) fail to identify the all the true QTL positions whereas the proposed method successfully identifies all the true QTL positions. Also in absence of outliers the proposed method is working as the classical methods.

2.3.2 Real Data Analysis Results

To investigate the performance of the proposed method for real data analysis in a comparison of traditional ML, LS and BND methods, we have considered the hypertension dataset of Sugiyama et al. (2001) which is available in R/qtl package (Broman et al., 2003), homepage: http://www.rqtl.org. A part of the hypertension

dataset have been shown in Figure A2.1 for clear understanding about the data structure. This dataset was analyzed to investigate the genetic control of salt-induced hypertension on male mice from a reciprocal backcross between the salt-sensitive c57BL/6J and the non-salt-sensitive A/J (A) inbred mouse strains.

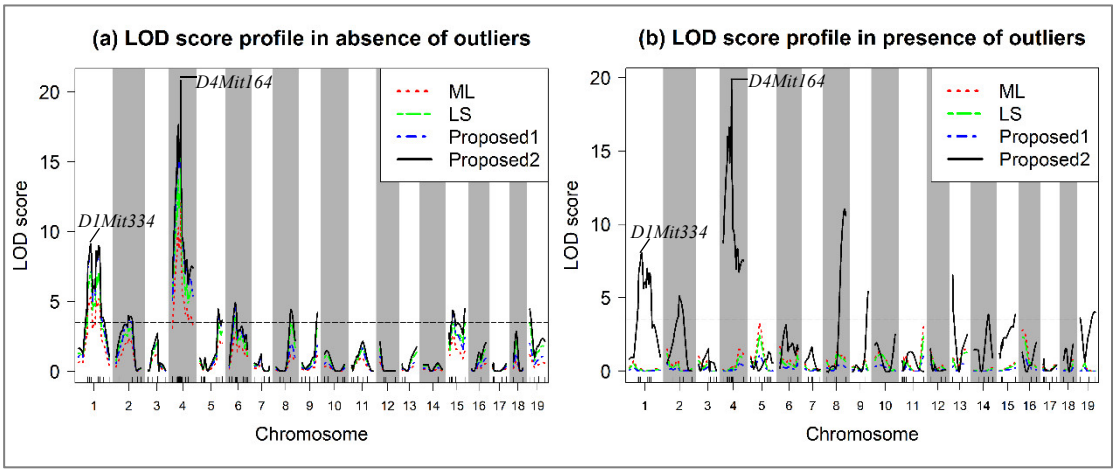


Figure 2.3: LOD score profile plot in absence and in presence of 12% outliers using real data.

Figure 2.3 represents the LOD score profile plots in absence and presence of outliers. Figure 2.3(a) shows the LOD scores profile in absence of outliers, where dotted (red colour), two dash (green colour), dot dash (blue colour) and solid (black colour) lines represents the LOD scores at every 1cM position on the chromosomes for ML, LS, BND and the proposed method, respectively, with $\beta = 0.02$. We select β by cross validation. Figure 2.3(b) shows the LOD scores profile for the contaminated dataset, where dotted (red colour), two dash (green colour), dot dash (blue colour) and solid (black colour) lines represents the LOD scores at every 1cM position on the chromosomes as before for ML, LS, BND and the proposed method, respectively, with $\beta = 0.2$.

Figure 2.3(a) shows that two QTLs on chromosome 1 (QTL/marker: *D1Mit334*) and chromosome 4 (QTL/marker: *D4Mit164*) are statistically significant genome-wide, and one QTL on each of chromosomes 2 (QTL/marker: *D2Mit62*), 6 (QTL/marker: *D6Mit8*), 8 (QTL/marker: *D8Mit271*) and 15 (QTL/marker: *D15Mit152*) are suggestive to be important for controlling blood pressure genome-wide by all four

methods for the uncontaminated real dataset. However, in presence of outliers, almost similar results are obtained by the proposed method only as shown in Figure 2.3(b). Therefore, the proposed method significantly outperforms over the traditional ML, LS and BND methods in presence of outliers. Otherwise, it shows equal performance.

Sugiyama et al. (2001) found that the QTL *D1Mit334* on chromosome 1 and the QTL *D4Mit164* on chromosome 4 were significantly associated with hypertension in mouse which supports our findings by the proposed method. They also suggested the QTLs *D6Mit15* and *D15Mit152* on chromosomes 6 and 15, respectively, as important QTLs for affecting blood pressure which are similar to our suggestive QTLs responsible for hypertension based on our proposed method.

2.4 Conclusion

In this paper, a new robust bivariate normal distribution (BND) based interval mapping approach has been discussed for QTL analysis by maximum β-likelihood estimation with BC population. The value of the tuning parameter β plays a key role on the performance of the proposed method. An appropriate value for the tuning parameter β can be selected by cross validation. The proposed method with tuning parameter $\beta = 0$ reduces to the traditional interval mapping approach. Simulation and real data analysis results show that the proposed method significantly improves the performance over the classical interval mapping approaches in presence of phenotypic outliers.

Chapter 3

Regression Based Fast Multi-trait QTL Analysis by Using the Properties of Multivariate Normal Distribution (Proposed)

3.1 Introduction

Single trait based simple interval mapping (Alam et al., 2018; Alam et al., 2016; Haley and Knott, 1992; Lander and Botstein, 1989) is the most popular and widely used approach to identify quantitative trait locus (QTL) controlling a single trait. However, in many line crossing experiments of genome-wide QTL mapping studies, measurements are taken on multiple traits along with the marker genotypes. Very often, such traits are correlated and there are common chromosome regions (chromosomal locations) that affect multiple traits (Chen, 2016b). Although single-trait simple interval mapping (SIM) methods can be applied to each trait one-by-one, such approaches do not take into account the pleotropic effects. A QTL is said to have pleiotropic effect if it simultaneously controls several phenotypic traits. The joint analyses of multiple traits, which include all quantitative traits together in a single model, can increase the power of QTL identification and improve the QTL localization accuracy when multiple traits are correlated genetically in the population (Xu, 2013a). In addition, QTL mapping considering multiple quantitative traits using joint analyses can give insights into the important genetic mechanisms underlying the trait relationships (e.g., genetic linkage versus pleiotropy), which would otherwise be hard to address if multiple traits are analyzed one-by-one. Therefore, statistical

methods are demanded for joint analyses of multiple traits to identify important QTL locations, which control multiple traits simultaneously.

Many methods for multi-trait QTL mapping have been developed in the literature, ranging from simple extensions of single-trait approaches to sophisticated multi-trait approaches designed specifically for multi-trait QTL mapping. Substantial work has been done in joint mapping for multiple quantitative traits (Almasy and Blangero, 1998; Hackett et al., 2001; Henshall and Goddard, 1999; Jiang and Zeng, 1995; Knott and Haley, 2000; Korol et al., 2001; Korol et al., 1995; Mangin et al., 1998; Williams et al., 1999). Least squares based multivariate regression (MVR-LS) analysis (Knott and Haley, 2000) and maximum likelihood based multivariate regression (MVR-ML) analysis (Xu, 2013a) using expectation maximization (EM) algorithm (Dempster et al., 1977) are two most popular and widely used multi-trait SIM approaches for multi-trait QTL analysis/mapping.

Although MVR-LS and MVR-ML are the most popular approaches for multi-trait QTL mapping, these methods have some limitations. The MVR-ML based multi-trait SIM is very time consuming due to EM algorithm based estimation which is an iterative process. Moreover, in MVR-ML approach, the calculation of likelihood ratio (LR) or log of odds (LOD) statistic is time consuming because it is a five steps process: (i) estimation of regression parameters, (ii) estimation of residuals and its variance-covariance matrix, (iii) estimation of Wilks' lambda statistic, (iv) Chi-square approximation of Wilks's lambda statistic, and (v) calculation of LR or LOD statistic based on approximated Chi-square statistic. Multivariate regression based multi-trait QTL analysis (Knott and Haley, 2000) is a generalization of simple linear regression based interval mapping (Haley and Knott, 1992). Xu (1995) has shown that the regression method can overestimate the residual variance, particularly for large QTL effects or widely spaced markers. Moreover, in MVR-LS approach, the calculation of LR or LOD statistic is also a time consuming five steps process like MVR-ML approach.

The recent advancements of technologies facilitate the generation high-dimensional genotype data of single nucleotide polymorphism (SNP) markers and phenotypic data

on multiple traits with large number of individuals in genome-wide QTL experiments. When the number of phenotypes and individuals are very large, and the markers are very dense (i.e., very large in number) resulting in a very big QTL data set, then computational time is a matter of consideration. In this study, we have proposed a new approach for multi-trait QTL mapping based on the assumption that the phenotypes and the conditional probability of QTL genotype given the marker genotypes follow joint multivariate normal distribution. In this method, the calculation of LR or LOD statistic is very straight forward because it is calculated only based on the sample variance-covariance matrix of the phenotypes and the conditional probability of QTL genotype given the marker genotypes. Our proposed method is able to identify the same QTL positions as identified by the other two existing methods (MVR-ML and MVR-LS). Moreover, it takes comparatively less computation time than the existing methods.

3.2 Materials and Methods

3.2.1 Proposed Method of Multi-trait QTL Analysis

Let us consider no epistasis between two QTLs, no interference in crossing over, and only one QTL in the testing interval for a backcross (BC) population. Let $\boldsymbol{Y}_j = [Y_{j1} \quad Y_{j2} \quad \dots \quad Y_{jm}]$, $j = 1, 2, \dots, n$, be an $(1\times m)$ vector for m phenotypic traits measured from the j^{th} individual of a mapping population. Let $X_j = p_{j|i}$ denote the conditional probability of the putative QTL genotype i $(i = 1, 2)$, given the flanking marker genotypes for the j^{th} individual. Then the linear regression model for BC population with m traits can be written in matrix notation as

$$\underset{(1\times m)}{\boldsymbol{Y}_j} = \underset{(1\times m)}{\boldsymbol{\alpha}} + \underset{(1\times 1)}{X_j}\underset{(1\times m)}{\boldsymbol{\gamma}} + \underset{(1\times m)}{\boldsymbol{\varepsilon}_j}, \; j = 1, 2, \dots, n \tag{3.1}$$

where $\boldsymbol{\alpha}$ is a $(1 \times m)$ vector for general mean effects and $\boldsymbol{\gamma}$ is a $(1 \times m)$ vector for the additive QTL effects, and $\boldsymbol{\varepsilon}_j$ is a $(1 \times m)$ vector for the random errors. The vector of random errors is assumed to be $\boldsymbol{\varepsilon}_j \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $(m \times m)$ variance-covariance matrix of random errors.

Let the flanking markers of the QTL testing interval are denoted by $\mathbf{M}_{\mathrm{L}}$ (left marker) with alleles M_L and m_L, and $\mathbf{M}_{\mathrm{R}}$ (right marker) with alleles M_R and m_R. Suppose that the locus of the unobserved putative QTL located within the testing interval bracketed by the flanking marker $\mathbf{M}_{\mathrm{L}}$ and $\mathbf{M}_{\mathrm{R}}$ is denoted by $\mathbf{Q}$ with alleles Q and q. The conditional probabilities for QTL genotypes QQ and Qq given the flanking marker genotypes are denoted by $p_{j|1}$ and $p_{j|2}$, respectively. The conditional probabilities $p_{j|1}$ and $p_{j|2}$ are shown in Table 3.1 for the BC population. The recombination fraction between the two markers is denoted by r. The possibility of the event of double recombination within the interval of two flanking markers is ignored.

Table 3.1: Conditional Probabilities of a putative QTL genotype given the flanking marker genotypes for a Backcross population.

		QTL Genotypes	
Marker Genotypes	Expected Frequency	$QQ(p_{j\|1})$	$Qq(p_{j\|2})$
M_LM_R/M_LM_R	$(1-r)/2$	1	0
M_LM_R/M_Lm_R	$r/2$	$(1-p^*)$	p
M_LM_R/m_LM_R	$r/2$	p	$(1-p)$
M_LM_R/m_Ln_R	$(1-r)/2$	0	1

$^*p = r_{\mathbf{M_LQ}}/r_{\mathbf{M_LM_R}}$, where $r_{\mathbf{M_LQ}}$ is the recombination fraction between the left marker $\mathbf{M}_{\mathrm{L}}$ and the putative QTL $\mathbf{Q}$, and $r_{\mathbf{M_LM_R}}$ is the recombination fraction between two flanking markers $\mathbf{M}_{\mathrm{L}}$ and $\mathbf{M}_{\mathrm{R}}$.

We want to test the null hypothesis is $H_0: \boldsymbol{\gamma} = \mathbf{0}$ (i.e., there is no QTL at a given position within a marker interval) against $H_1: H_0$ is not true. Under null hypothesis (H_0) the model (3.1) reduces to the following model

$$\underset{(1\times m)}{\boldsymbol{Y}_j} = \underset{(1\times m)}{\boldsymbol{\alpha}} + \underset{(1\times m)}{\boldsymbol{\varepsilon}_j}, \; j = 1, 2, \ldots, n \tag{3.2}$$

Let $L_1(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is the likelihood function under the full model (3.1) and $L_0(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ is the likelihood function under the reduced model (3.2). To test H_0 against H_1, the likelihood ratio test (LRT) statistic is defined as

$$\text{LRT} = -2\ln\left[\frac{\max_{\boldsymbol{\alpha},\boldsymbol{\Sigma}} L_0(\boldsymbol{\alpha},\boldsymbol{\Sigma})}{\max_{\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\Sigma}} L_1(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\Sigma})}\right] = -2\ln\left[\frac{L_0(\hat{\boldsymbol{\alpha}}_0,\hat{\boldsymbol{\Sigma}}_0)}{L_1(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\Sigma}})}\right]$$

$$= -n\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}|}\right) \tag{3.3}$$

where $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Sigma}}$ are the maximum likelihood (ML) estimates of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ under the full model (3.1), and $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\Sigma}}_0$ are the ML estimates of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ under the reduced model (i.e., under H_0).

In order to estimate the model parameter and the variance-covariance matrix of random errors, let us consider that $\boldsymbol{Z} = \left[\underset{(m\times 1)}{\boldsymbol{Y}} \vdots \underset{(1\times 1)}{X}\right]^T$ follows a multivariate normal distribution (MND) $N\left(\underset{(m+1)\times 1}{\boldsymbol{\mu_Z}}, \underset{(m+1)\times(m+1)}{\boldsymbol{\Sigma_Z}}\right)$ with mean vector $\boldsymbol{\mu_Z}$ and variance-covariance matrix $\boldsymbol{\Sigma_Z}$, where $\boldsymbol{Y}$ and X have been introduced in (3.1).

Then the probability density function of $\boldsymbol{Z}$ can be written as

$$f(\boldsymbol{Z}) = \frac{1}{(2\pi)^{(m+1)/2}|\boldsymbol{\Sigma_Z}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{Z}-\boldsymbol{\mu_Z})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu_Z})\right] \tag{3.4}$$

We can partition the mean vector $\boldsymbol{\mu_Z}$ as $\boldsymbol{\mu_Z} = \left[\underset{(m\times 1)}{\boldsymbol{\mu_Y}} \quad \vdots \quad \underset{(1\times 1)}{\mu_X}\right]^T$ and the variance-covariance matrix $\boldsymbol{\Sigma_Z}$ as

$$\boldsymbol{\Sigma_Z} = \begin{bmatrix} \sigma_{Y_1Y_1} & \sigma_{Y_2Y_1} & \cdots & \sigma_{Y_mY_1} & \sigma_{XY_1} \\ \sigma_{Y_1Y_2} & \sigma_{Y_2Y_2} & \cdots & \sigma_{Y_mY_2} & \sigma_{XY_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{Y_1Y_m} & \sigma_{Y_2Y_m} & \ldots & \sigma_{Y_mY_m} & \sigma_{XY_m} \\ \sigma_{Y_1X} & \sigma_{Y_2X} & \cdots & \sigma_{Y_mX} & \sigma_{XX} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma_{YY}} & \boldsymbol{\Sigma_{YX}} \\ \boldsymbol{\Sigma_{XY}} & \sigma_{XX} \end{bmatrix}_{(m+1)\times(m+1)},$$

where $\underset{m\times m}{\boldsymbol{\Sigma_{YY}}} = \begin{bmatrix} \sigma_{Y_1}^2 & \sigma_{Y_2Y_1} & \cdots & \sigma_{Y_mY_1} \\ \sigma_{Y_1Y_2} & \sigma_{Y_2}^2 & \cdots & \sigma_{Y_mY_2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Y_1Y_m} & \sigma_{Y_2Y_m} & \ldots & \sigma_{Y_m}^2 \end{bmatrix}$, $\underset{1\times 1}{\sigma_{XX}} = \sigma_X^2$, $\underset{1\times m}{\boldsymbol{\Sigma_{YX}}} = \left[\sigma_{Y_1X}, \sigma_{Y_2X}, \ldots, \sigma_{Y_mX}\right]$

and $\sigma_{Y_kX} = E\left[(X-\mu_X)\left(Y_k-\mu_{Y_k}\right)\right], k = 1, 2, \ldots, m.$

Then the conditional mean of $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$ given X can be written as

$$E(\boldsymbol{Y}|X = x) = \boldsymbol{\mu_Y} + \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(X-\mu_X)$$

$$= \boldsymbol{\mu_Y} + \boldsymbol{\Sigma_{YX}}\boldsymbol{\Sigma_{XX}^{-1}}X - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\mu_X$$
$$= (\boldsymbol{\mu_Y} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\mu_X) + (\boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1})X$$
$$\Rightarrow E(\boldsymbol{Y}|X = x) = \boldsymbol{\alpha} + \boldsymbol{\gamma}X \quad (3.5)$$

which is known as multivariate multiple linear regression surface of $\boldsymbol{Y}$ on X, where the (m×1) vector $\boldsymbol{\alpha} = (\boldsymbol{\mu_Y} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\mu_X)$ is the vector of general mean effects and the (m×1) vector $\boldsymbol{\gamma} = (\boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1})$ is called the vector of regression coefficients. For BC population $\boldsymbol{\gamma}$ is the additive QTL effects.

The prediction error is

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - E(\boldsymbol{Y}|X) = \boldsymbol{Y} - \boldsymbol{\mu_Y} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(X - \mu_X) \quad (3.6)$$

Now, the variance-covariance matrix of the prediction error is

$$\boldsymbol{\Sigma} = V(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}}]$$
$$= E\left[[\boldsymbol{Y} - \boldsymbol{\mu_Y} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(X - \mu_X)][\boldsymbol{Y} - \boldsymbol{\mu_Y} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(X - \mu_X)]^T\right]$$
$$= \boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(\boldsymbol{\Sigma_{YX}})^T - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\boldsymbol{\Sigma_{XY}} + \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\sigma_{XX}\sigma_{XX}^{-1}(\boldsymbol{\Sigma_{YX}})^T$$
$$= \boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(\boldsymbol{\Sigma_{YX}})^T - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\boldsymbol{\Sigma_{XY}} + \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}(\boldsymbol{\Sigma_{YX}})^T$$
$$= \boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\boldsymbol{\Sigma_{XY}} \quad (3.7)$$

From (3.7), we can write

$$\boldsymbol{\Sigma_{YY}} = \boldsymbol{\Sigma_{YX}}\sigma_{XX}^{-1}\boldsymbol{\Sigma_{XY}} + \boldsymbol{\Sigma}$$
$$\Rightarrow \text{Total SS} = \text{Regression SS} + \text{ Error SS} \quad (3.8)$$

where SS stands for sum of squares.

Since $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ are typically unknown, they must be estimated from a random sample in order to construct the multivariate linear predictor and the test statistics, and to determine expected prediction errors.

Based on a random sample of size n, the maximum likelihood estimator of the $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ are given by

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}} = \begin{bmatrix} \overline{\boldsymbol{Y}} \\ \overline{X} \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} & \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}} \\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} & \hat{\sigma}_{XX} \end{bmatrix} = \left(\frac{n-1}{n}\right)\mathbf{S} = \left(\frac{n-1}{n}\right)\begin{bmatrix} \mathbf{S_{YY}} & \mathbf{S_{YX}} \\ \mathbf{S_{XY}} & \mathrm{S}_{XX} \end{bmatrix} \quad (3.9)$$

where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} = \left[\overline{Y}_1 = \frac{1}{n}\sum_{j=1}^{n} Y_{j1} \quad \overline{Y}_2 = \frac{1}{n}\sum_{j=1}^{n} Y_{j2} \quad \cdots \quad \overline{Y}_m = \frac{1}{n}\sum_{j=1}^{n} Y_{jm}\right]$,

$\hat{\mu}_X = \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} = \boldsymbol{S}_{\boldsymbol{YY}} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}\right)^T$,

$S_{XX} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}} = \boldsymbol{S}_{\boldsymbol{XY}} = \boldsymbol{S}_{\boldsymbol{YX}} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}\right)\left(X_j - \overline{X}\right)^T$ are the maximum likelihood estimates of $\boldsymbol{\mu}_{\boldsymbol{Y}}$, μ_X, $\boldsymbol{\Sigma}_{\boldsymbol{YY}}$, $\boldsymbol{\Sigma}_{\boldsymbol{YX}} = \boldsymbol{\Sigma}_{\boldsymbol{XY}}$ and σ_X^2, respectively.

Then, using (3.9), based on a random sample of size n the maximum likelihood estimators of the vectors regression parameters are

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}\hat{\sigma}_{XX}^{-1}\hat{\mu}_X\right) = \overline{\boldsymbol{Y}} - \mathbf{S}_{\boldsymbol{YX}}\mathrm{S}_{XX}^{-1}\overline{X} \tag{3.10}$$

and

$$\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}\hat{\sigma}_{XX}^{-1}) = \mathbf{S}_{\boldsymbol{YX}}\mathrm{S}_{XX}^{-1} \tag{3.11}$$

Therefore, using (3.10) and (3.11), the maximum likelihood estimator of the regression function can be written as

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\gamma}}X = \overline{\boldsymbol{Y}} - \mathbf{S}_{\boldsymbol{YX}}\mathrm{S}_{XX}^{-1}\overline{X} + \mathbf{S}_{\boldsymbol{YX}}\mathrm{S}_{XX}^{-1}X = \overline{\boldsymbol{Y}} + \mathbf{S}_{\boldsymbol{YX}}\mathrm{S}_{XX}^{-1}(X - \overline{X}) \tag{3.12}$$

Now, based on the maximum likelihood estimates of $\boldsymbol{\mu}_{\boldsymbol{Y}}$, μ_X, $\boldsymbol{\Sigma}_{\boldsymbol{YY}}$, $\boldsymbol{\Sigma}_{\boldsymbol{YX}} = \boldsymbol{\Sigma}_{\boldsymbol{XY}}$ and σ_X^2, the maximum likelihood estimates of $\boldsymbol{\Sigma}$ under the full model (3.1) and the reduced model (3.2) are, respectively,

$$\widehat{\boldsymbol{\Sigma}} = \left(\frac{n-1}{n}\right)(\mathbf{S}_{\boldsymbol{YY}} - \mathbf{S}_{\boldsymbol{YX}}\mathrm{S}_{XX}^{-1}\mathbf{S}_{\boldsymbol{XY}}) \tag{3.13}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}} = \left(\frac{n-1}{n}\right)\mathbf{S}_{\boldsymbol{YY}} \tag{3.14}$$

Now, using (3.13) and (3.14) in (3.3), the LRT statistic to test H_0 can be written as

$$\mathrm{LRT} = -n\ln\left(\frac{\left|\widehat{\boldsymbol{\Sigma}}\right|}{\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}\right|}\right) \tag{3.15}$$

where $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}$ are the estimated variance-covariance matrices of residuals under the full model and the reduced model (i.e., under H_0), respectively.

For large n, the modified LRT statistic is

$$\text{LRT} = -\left[ResDf - \frac{1}{2}(m - TestDf + 1)\right] \ln\left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}|}\right) \tag{3.16}$$

where *ResDf* and *TestDf* are the residual degrees of freedom and test degrees of freedom, respectively.

Equation (3.16) can be written as

$$\text{LRT} = -\left[n - k - 1 - \frac{1}{2}(m - k + 1)\right] \ln\left(\frac{|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}|}{|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}|}\right) \tag{3.17}$$

where *k* is the TestDf which is the number of predictor variables.

Under the null hypothesis (H_0), the modified LRT statistic is expected to have an approximate chi-square distribution with *mk* degrees of freedom for a given QTL position in the genome. However, the threshold value to reject the null hypothesis (H_0) cannot be simply chosen from the χ^2 distribution because of the violation of regularity conditions of asymptotic theory under H_0. An alternative way is to use log of odds (LOD) score (Lander and Botstein, 1989; Ott, 1999; Terwilliger and Ott, 1994; Wu et al., 2007; Xu, 2013d) as a test statistic to test the null hypothesis of no QTL (H_0).

The LOD score is the transformation of the LRT statistic, defined as

$$\text{LOD} = \frac{\text{LRT}}{2\times \log(10)} = \frac{\text{LRT}}{4.605} = 0.217\ \text{LRT} \tag{3.18}$$

According Lander and Botstein (1989), the typical threshold of LOD score should be between 2 and 3 to ensure a 5% overall false positive error for identifying a QTL. Terwilliger and Ott (1994), Ott (1999), Wu et al. (2007), and Xu (2013d) suggested a value of LOD = 3 as the critical threshold for declaring the existence of QTL. Thus, the LOD > 3 can be used as a criterion to declare a significant QTL.

In this study, we have developed the proposed method only for backcross (BC) population. However, the proposed method can be developed for other mapping

populations, such as intercross (F_2), double haploid (DH) and other popular crosses, by some simple modifications of the proposed method for BC population.

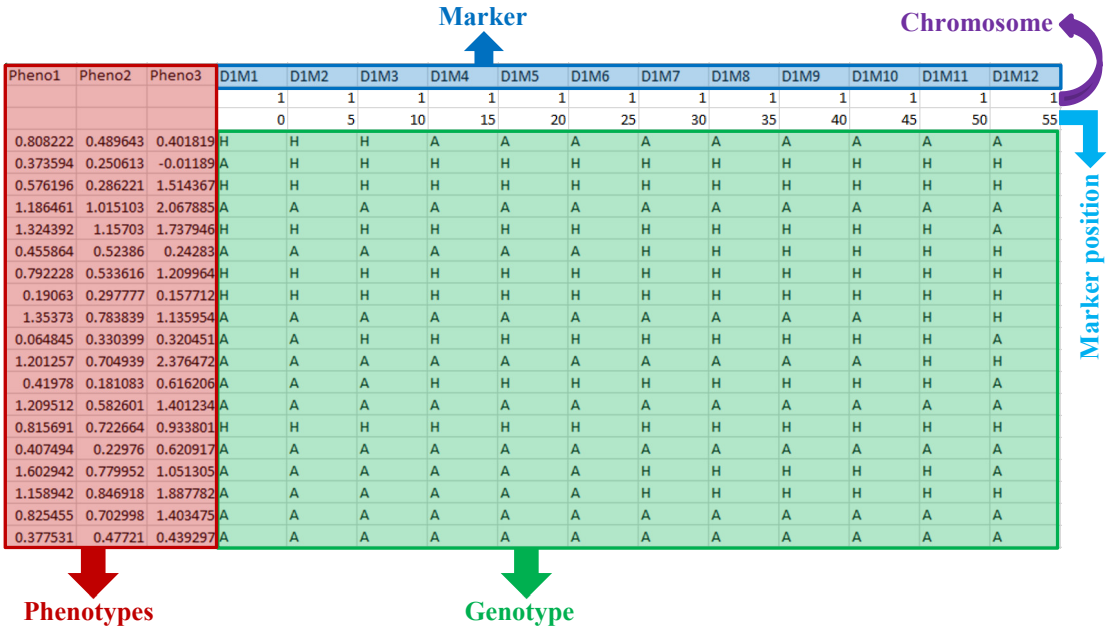
3.2.2 Simulated Datasets

We have used two simulated datasets of BC population to demonstrate the performance of the proposed method in a comparison with the traditional LS regression based multi-trait QTL mapping (Knott and Haley, 2000) and MLE based multi-trait QTL mapping (Xu, 2013a). We have evaluated the performance of the three different methods including the proposed method in terms of the power of QTL detection and computation time. We have generated two simulated datasets (SimData1: Simulated Data 1 and SimData2: Simulated Data 2) using the multivariate regression model (3.1) with BC population.

To generate SimData1 we have considered 3 phenotypes (denoted as Pheno1, Pheno2 and Pheno3), 250 individuals, total 13 chromosomes, and 21 equally spaced (5 cM) markers on each of the 13 chromosomes resulting in a total of 273 equally spaced markers distributed on 13 chromosomes. We have simulated total four QTLs to affect three quantitative traits where the true QTLs are located on chromosomes 2, 4, 6 and 8 at marker 5. Among the simulated QTLs, the QTL on chromosome 4 is a pleiotropic QTL affecting the phenotypes Pheno1 and Pheno2 simultaneously. We consider the parameter values $\boldsymbol{\alpha} = [1.25 \quad 1.75 \quad 1.5]$, $\boldsymbol{\gamma} = [1.50 \quad 1.0 \quad 2.25]$ and $\boldsymbol{\Sigma} = diag\,[0.25 \quad 0.25 \quad 0.25]$. Figure 3.1 shows the structure of a dataset obtained from a genome-wide QTL experiment for multi-trait QTL analysis.

We have generated SimData2 with 5 phenotypes (denoted by Pheno1, Pheno2, Pheno3, Pheno4 and Pheno5), 500 individuals, 13 chromosomes, and 41 equally spaced (5 cM) markers on each of the 13 chromosomes resulting in a total of 533 equally spaced markers distributed on 13 chromosomes. In SimData2, we have considered the true QTL positions on chromosomes 2, 4, 6, 8 and 10 at maker 5 affecting the 5 phenotypic traits. Among the considered QTLs, three QTLs on chromosomes 2, 4 and 6 are considered as pleiotropic QTLs simultaneously affecting the set of phenotypes (Pheno1 and Pheno2), (Pheno2 and Pheno4) and (Pheno1,

Pheno2 and Pheno3), respectively. We consider the parameter values $\boldsymbol{\alpha} = [0.5 \quad 1.0 \quad 1.5 \quad 1.25 \quad 0.75]$, $\boldsymbol{\gamma} = [1.25 \quad 1.75 \quad 1.5 \quad 0.75 \quad 0.5]$ and $\boldsymbol{\Sigma} = diag\,[0.25 \quad 0.25 \quad 0.25 \quad 0.25 \quad 0.25]$.



Pheno1	Pheno2	Pheno3	D1M1	D1M2	D1M3	D1M4	D1M5	D1M6	D1M7	D1M8	D1M9	D1M10	D1M11	D1M12
			1	1	1	1	1	1	1	1	1	1	1	1
			0	5	10	15	20	25	30	35	40	45	50	55
0.808222	0.489643	0.401819	H	H	H	A	A	A	A	A	A	A	A	A
0.373594	0.250613	-0.01189	A	H	H	H	H	H	H	H	H	H	H	H
0.576196	0.286221	1.514367	H	H	H	H	H	H	H	H	H	H	H	H
1.186461	1.015103	2.067885	A	A	A	A	A	A	A	A	A	A	A	A
1.324392	1.15703	1.737946	H	H	H	H	H	H	H	H	H	H	H	A
0.455864	0.52386	0.24283	A	A	A	A	A	A	H	H	H	H	H	H
0.792228	0.533616	1.209964	H	H	H	H	H	H	H	H	H	H	H	H
0.19063	0.297777	0.157712	H	H	H	H	H	H	H	H	H	H	H	H
1.35373	0.783839	1.135954	A	A	A	A	A	A	A	A	A	A	H	H
0.064845	0.330399	0.320451	A	A	H	H	H	H	H	H	H	H	H	A
1.201257	0.704939	2.376472	A	A	A	A	A	A	A	A	A	A	H	H
0.41978	0.181083	0.616206	A	A	A	H	H	H	H	H	H	H	H	A
1.209512	0.582601	1.401234	A	A	A	A	A	A	A	A	A	A	A	A
0.815691	0.722664	0.933801	H	H	H	H	H	H	H	H	H	H	H	H
0.407494	0.22976	0.620917	A	A	A	A	A	A	A	A	A	A	A	A
1.602942	0.779952	1.051305	A	A	A	A	A	A	H	H	H	H	H	A
1.158942	0.846918	1.887782	A	A	A	A	A	A	H	H	H	H	H	H
0.825455	0.702998	1.403475	A	A	A	A	A	A	A	A	A	A	A	A
0.377531	0.47721	0.439297	A	A	A	A	A	A	A	A	A	A	A	A

Figure 3.1: Structure of the dataset obtained from a genome-wide QTL experiment for multi-trait QTL analysis.

3.2.3 Real Datasets

We have also considered two real datasets to investigate the performance of the proposed method in a comparison with the traditional LS regression based multi-trait QTL mapping and MLE based multi-trait QTL mapping. We have employed all the three methods including our proposed method with those two real datasets on QTL experiment, and compared the QTL detection power and the computation time of the proposed method with the traditional methods (MVR-ML and MVR-LS).

3.2.3.1 Barley Data

First, we have compared the three different methods (MVR-ML, MVR-LS and Proposed) of multivariate analysis using a double haploid (DH) population of barley. For DH population, the conditional probability of QTL genotype given marker

genotypes can be calculated using the same formula as that used in the BC population (Xu, 2013c). DH population can be achieved by a single generation of cytogenetic manipulation, just like a BC population and it contains two possible genotypes. The dataset was originally published by (Hayes et al., 1993). We have obtained the barley dataset from the web-database GeneNetwork (http://www.genenetwork.org/). The data consist of 150 double haploid (DH) lines derived from the cross of two spring barley varieties, Steptoe and Morex. A total of eight quantitative traits, including grain yield (YIELD), heading date (HEAD), plant height (HEIGHT), lodging (LODG), grain protein (PROTEIN), alpha amylase (ALPHA), diastatic power (POWER), and malt extract (EXTRACT), were measured from multiple environments with the number of environments ranging from 6 to as many as 16. The average values of trait across the environments were considered as the original phenotypic values for the QTL mapping experiment. QTL by environment (Q×E) interaction is assumed to be absent. The total number of markers was 495 distributed along seven chromosomes of the barley genome. The genotypes of the markers were denoted by A for the Steptoe parent and B for the Morex parent. To get a clear concept about the structure of the barley data, a part of the barley data has been presented in Figure A3.1.

3.2.3.2 Mouse Data

We also investigated the performance of all the three methods (MVR-ML, MVR-LS and Proposed) using another real dataset of BC lines mouse. The data were originally published by Leiter et al. (2009). The data consist of 310 backcross lines of female mice derived from the cross NOD × (NOD × 129.H2^{g7})F1 backcross (N2). The BC mice used in that study Leiter et al. (2009) were measured for different clinical, laboratory and metabolic quantitative/qualitative traits related to type 1 diabetes. In our study, we have considered only the bone mineral and plasma related phenotypes: BMC (Bone Mineral Content, as determined by DXA), AREA (Bone Mineral Area, as determined by DXA), LEPTIN (Leptin in plasma), INSULIN (Insulin in plasma), CHOL (Total cholesterol in plasma), HDLD (HDLD cholesterol in plasma), GLUCOSE (Glucose in plasma), NEFA (Non-Esterified Fatty Acids in plasma) and TG (Triglycerides in plasma). If any of the selected nine phenotypes is missing for

any individual, then that individual has been excluded from the analysis. The total number of SNP markers was 303 distributed along 19 chromosomes of the mouse genome. The genotypes of the markers were denoted by N for the NOD allele and H for the heterozygous allele. A part of the mouse data has been displayed in Figure A3.2.

3.3 Results and discussion

We have compared the performance (i.e., power of QTL detection) and computation times of the proposed method in comparison with other two methods (MVR-ML and MVR-LS) of multi-trait QTL analysis using both simulated and real data. All the three methods, including our proposed method, have been employed with two simulated and two real QTL datasets with BC population to evaluate the performance and computation time of these methods. We used SAS version 9.4 software (SAS Institute Inc., Cary, North Carolina, USA) to implement all the three methods of multi-trait QTL analysis. For the traditional existing methods (MVR-ML and MVR-LS) we have used the PROC QTL procedure of the SAS package "PROC QTL" developed by Hu and Xu (2009). We used same platform/operating system (Windows 8.1 Pro) and environment (Batch Submit with SAS 9.4) to compare the computation time of the three methods (MVR-ML, MVR-LS and Proposed).

3.3.1 Simulated Data Analysis Results

We have compared the performance and computation times of the proposed method with the traditional methods (MVR-ML and MVR-LS) of multi-trait QTL analysis using two simulated data sets: SimData1 and SimData2. To do this we have implemented all the three methods, including our proposed method, with the simulated data sets and evaluated the performance of these methods for true QTL detection. Also we have compared the computation time (in second) of the proposed method with the existing methods.

3.3.1.1 Three-trait QTL Analysis

SimData1 consists of 250 individuals with three phenotypes (Pheno1, Pheno2 and Pheno3) and 273 equally spaced (5 cM) markers distributed on 13 chromosomes. The true QTL positions were considered on chromosomes 2, 4, 6 and 8 at maker 5. Genome-wide QTL search has been performed at each 1 cM flanking marker interval using all the three multi-trait QTL mapping methods (MVR-ML, MVR-LS and Proposed) to identify significant QTLs affecting one or more of the three phenotypes.

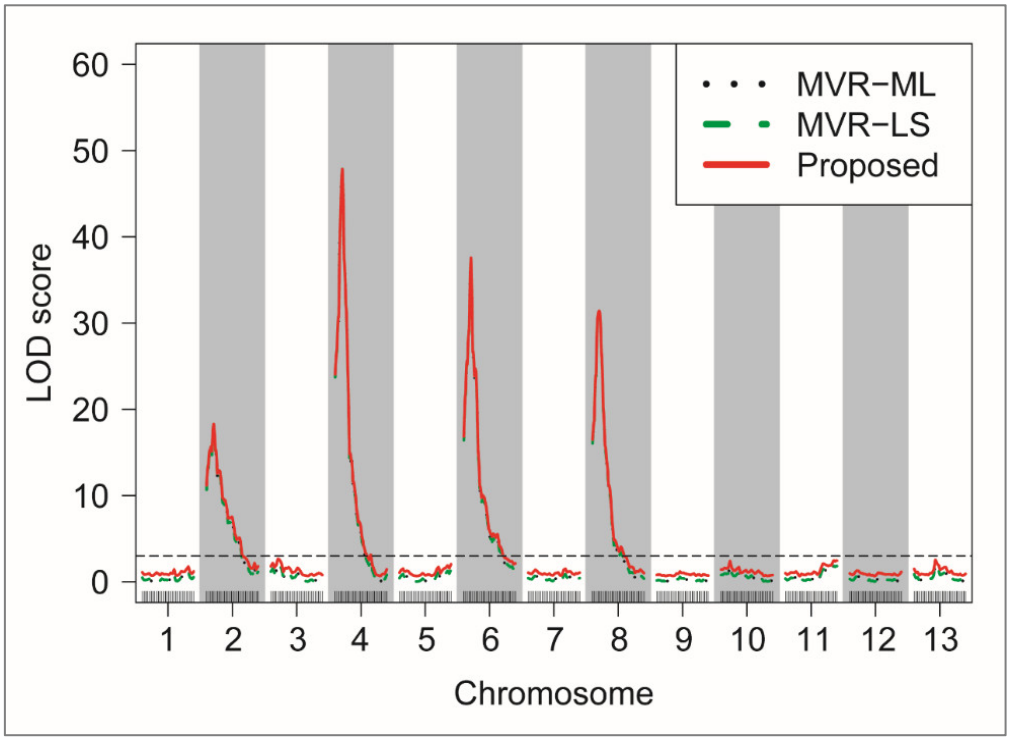


Figure 3.2: LOD score profile plot of multi-trait QTL analysis using the MVR-ML, MVR-LS and proposed methods with SimData1 (Simulated Data 1) with 3 phenotypes. The true QTL positions were considered on chromosomes 2, 4, 6 and 8 at marker 5 (marker position 20 cM).

Figure 3.2 represents the LOD score profile plot of multi-trait QTL analysis with the SimData1 using MVR-ML, MVR-LS and the proposed method. The LOD score plot of multi-trait QTL analysis with SimData1 shows that for all the three methods the highest peaks occur on chromosome 2, 4, 6 and 8 at marker 5 (marker position 20 cM). This indicates that our proposed method identifies the same QTL positions as identified by the existing traditional methods (MVR-ML and MVR-LS). However, the

comparison of computation times of three methods (Table 3.2) reveals that the required computation time of the proposed method (2.18 seconds) is smaller than that of the existing methods MVR-ML (17.75 seconds) and MVR-LS (2.71 seconds).

3.3.1.2 Five-trait QTL Analysis

SimData2 consists of five phenotypes (Pheno1, Pheno2, Pheno3, Pheno4 and Pheno5) of 500 individuals and total 533 equally spaced (5 cM) markers on 13 chromosomes (41 markers on each chromosome) for each individual. The true QTL positions were considered on chromosomes 2, 4, 6, 8 and 10 at maker 5. Genome-wide QTL scanning has been done at each 1 cM flanking marker interval using all the three methods of multi-trait QTL mapping (MVR-ML, MVR-LS and Proposed) to identify significant QTLs affecting one or more of the five phenotypes.

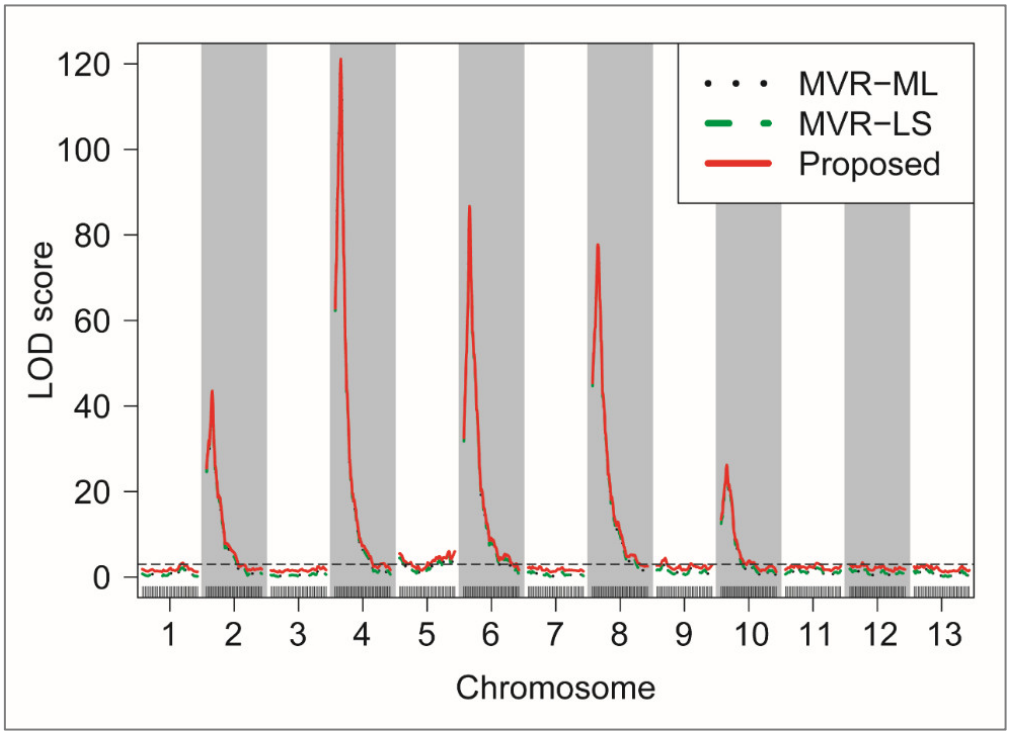


Figure 3.3: LOD score profile plot of multi-trait QTL analysis using the MVR-ML, MVR-LS and proposed methods with SimData2 (Simulated Data 2) with 5 phenotypes (Pheno1, Pheno2, Pheno3, Pheno4 and Pheno5). The true QTL positions were considered on chromosomes 2, 4, 6, 8 and 10.

Figure 3.3 shows the LOD score profile plot of multi-trait QTL analysis with SimData2 using all the three methods including the proposed method. The LOD score plot (Figure 3.3) shows that all the methods exhibit the highest LOD peaks on chromosomes 2, 4, 6, 8 and 10 at marker 5 (marker position 20 cM). This means that the proposed method is able to identify the same QTL positions as identified by the other two tradition methods (MVR-ML and MVR-LS). When we compare the computation times of three methods (Table 3.2), we have found that the proposed method takes less computation time (3.66 sec.) compared to the computation times of the MVR-ML (75.16 seconds) and MVR-LS (9.26 seconds) methods.

Table 3.2: Comparison of computational times of multi-trait QTL analysis among three methods (MVR-ML, MVR-LS and Proposed) with SimData1 and SimData2

Data set	No. of chromosomes	Phenotypic size[a]	Genotypic size[b]	Computation time (in second)		
				MVR-ML[c]	MVR-LS[d]	Proposed[e]
SimData1	13	3×250	250×273	17.75	2.71	**2.18**
SimData2	13	5×500	500×533	75.16	9.26	**3.66**

[a] Phenotypic size indicates (Phenotypes×Indviduals).
[b] Genotypic size indicates (Individuals×Markers).
[c] MVR-ML: Multivariate regression (MVR) using maximum likelihood (ML) method.
[d] MVR-LS: Multivariate regression (MVR) using least squares (LS) method.
[e] Proposed: Multivariate regression (MVR) using the properties multivariate normal distribution.

3.3.1.3 Power Analysis and Comparison of Computation Time

To compare power of QTL detection (percentage of correct QTL identification) and computation times between the proposed and existing methods, we have performed simulation and analyses on 100 replicates of SimData1 and SimData2. Table 3.3 represents the average along with standard deviation (SD) of the locus positions identified in 100 replications by each of the three methods (MVR-ML, MVR-LS and Proposed) with SimData1 and SimData2. We observe that all the methods identify almost the same QTL positions which approximately match with the true QTL positions. This indicates that all the three methods have almost same performance of QTL detection.

Table 3.3: Comparison of descriptive summary of identified QTL positions identified by three different methods (MVR-ML, MVR-LS and Proposed) in 100 replications

QTL	Chromosome	True QTL position (cM)	Identified QTL position		
			MVR-ML (Mean ± SD)	MVR-LS (Mean ± SD)	Proposed (Mean ± SD)
			SimData1		
QTL1	2	20	20.00 ± 0.20	19.96 ± 0.53	19.95 ± 0.52
QTL2	4	20	19.99 ± 0.39	19.84 ± 0.63	19.85 ± 0.66
QTL3	6	20	19.99 ± 0.64	19.92 ± 0.93	19.88 ± 0.92
QTL4	8	20	20.20 ± 0.98	20.11 ± 1.12	20.11 ± 1.12
			SimData2		
QTL1	2	20	19.97 ± 0.39	19.95 ± 0.48	19.95 ± 0.48
QTL2	4	20	20.01 ± 0.39	19.92 ± 0.56	19.92 ± 0.56
QTL3	6	20	19.96 ± 0.47	20.05 ± 0.63	20.04 ± 0.62
QTL4	8	20	19.95 ± 0.44	19.95 ± 0.63	19.95 ± 0.63
QTL5	10	20	20.13 ± 0.56	19.99 ± 0.72	20.00 ± 0.71

Table 3.4 shows the statistical power (percentage of correct identification of QTL positions in 100 replications) of QTL detection and average computation time of the three methods (MVR-ML, MVR-LS and Proposed) of multi-trait QTL mapping from 100 replications of simulation and analyses for SimData1 and SimData2. We find that the statistical powers of the MVR-ML method are 96%, 85%, 76% and 70% to identify true QTLs on chromosomes 2, 4, 6 and 8, respectively, in SimData1 whereas the MVR-LS and Proposed methods exhibit 92%/93%, 78%, 76% and 64% power to identify true QTLs on the same chromosomes. The MVR-ML method shows 85%, 85%, 78%, 81% and 70% powers to identify the true QTL on chromosomes 2, 4, 6, 8, 10, respectively in SimData2 while the MVR-LS and Proposed methods achieve 77%, 78%, 71%, 77% and 65% statistical power to identify true QTLs on the same chromosomes. This means that all the three methods (MVR-ML, MVR-LS and Proposed) of multi-trait QTL analysis have almost the same power of QTL identification with SimData1 and SimData2.

Table 3.4: Observed statistical power (percentage of correct identification of true QTL positions in 100 replications) and average computation time of the three methods (MVR-ML, MVR-LS and Proposed) of multi-trait QTL mapping from 100 replications of simulations

QTL	Chr	True QTL position (cM)	Percentage of correct identification (Power of QTL identification)			Average computation time (in second)		
			MVR-ML	MVR-LS	Proposed	MVR-ML	MVR-LS	Proposed
				SimData1				
QTL1	2	20	96	92	93			
QTL2	4	20	85	78	78			
QTL3	6	20	76	76	76	18.11	2.31	**1.80**
QTL4	8	20	70	64	64			
				SimData2				
QTL1	2	20	85	77	77			
QTL2	4	20	85	78	78			
QTL3	6	20	78	71	72	240.74	22.14	**4.62**
QTL4	8	20	81	77	77			
QTL5	10	20	70	65	66			

Chr: Chromosome, MVR-ML: Maximum likelihood base multivariate regression, MVR-LS: Least squares based multivariate regression.

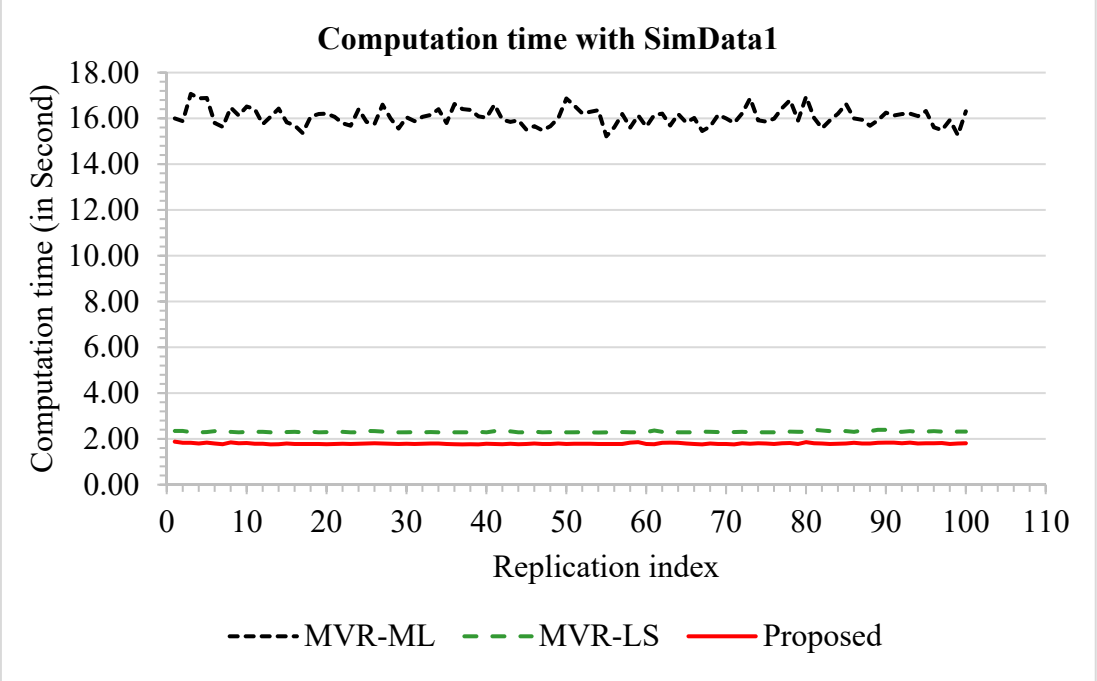


Figure 3.4: Time series plot of computation times (in second) of three different methods (MVR-ML, MVR-LS and Proposed) for SimData1 in 100 replications.

Figure 3.4 and Figure 3.5 represent the time series plots of computation times of three methods for SimData1 and SimData2, respectively. We see that the computation time of the proposed method is always less than that of the other two methods for both SimData1 and SimData2. From Table 3.4, we observe that the average computation times of the MVR-ML, MVR-LS and Proposed methods of multi-trait QTL mapping are 18.11 sec, 2.31 sec and 1.80 sec, respectively, for SimData1. For SimData2, the average computation times required by the MVR-ML, MVR-LS and Proposed methods are 240.74 sec, 22.14 sec and 4.62 sec, respectively. The above results indicates that our proposed method is very less computation time consuming. Hence, we can conclude that our proposed method is very efficient in terms of computation time exhibiting almost the same performance in correct QTL detection (i.e., same statistical power).

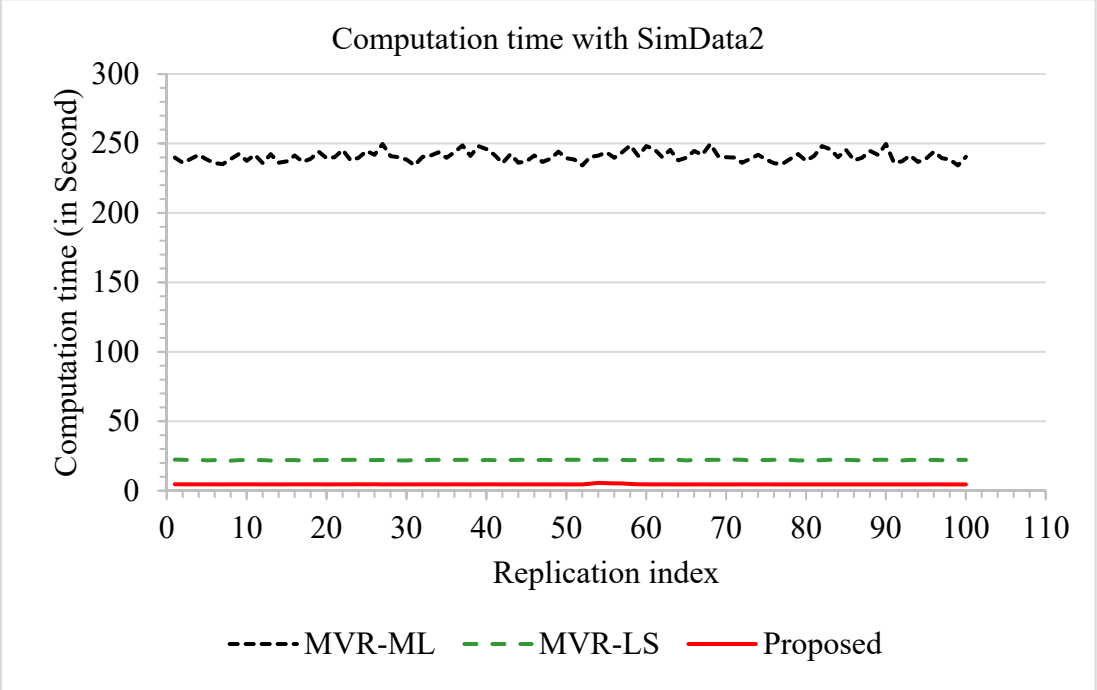


Figure 3.5: Time series plot of computation times (in seconds) of three different methods (MVR-ML, MVR-LS and Proposed) for SimData2 in 100 replications.

3.3.2 Real Data Analysis Results

To investigate the performance of the proposed method in comparison with the existing methods of multi-trait QTL mapping (MVR-ML and MVR-LS), we also employed all the three methods including our proposed method with two real datasets: barley data (Hayes et al., 1993) and mouse data (Leiter et al., 2009).

3.3.2.1 Eight-trait QTL Analysis with Barley Data

We have performed genome-wide QTL scanning at every 1 cM marker interval using all the three methods of multi-trait QTL analysis (MVR-ML, MVR-LS and Proposed) to identify significant QTLs affecting one or more of the eight phenotypes (average yield, loading, height, head, protein, alpha, dias and maltex) in the Barley dataset (Hayes et al., 1993). We have compared the performance and computation time of the proposed method with the other two existing methods (MVR-ML and MVR-LS) for genome-wide QTL searching in the barley data.

Figure 3.6 represents the LOD score profile plot of genome-wide multi-trait QTL mapping using all the three methods with the barley data. The marker or "marker interval" at which maximum LOD occurs on each of the seven chromosomes is presented in Table 3.5 along with the QTL position and maximum LOD score value for each of the three methods. From Figure 3.6 and Table 3.5 we observe that the MVR-ML method identifies the maximum LODs (i.e., highest peak in the LOD score plot) at positions 0.0 (marker: Hor5, max. LOD: 13.53), 36.30 (marker: Tef4, max. LOD: 56.98), 54.40 (marker: Dfr, max. LOD: 31.49), 139.98 (c4.loc112 within marker interval: ksuH11 (139.00 cM) – Tel4M (148.80 cM), max. LOD: 17.49), 78.03 (c5.loc55 within marker interval: snp_0953 (77.10 cM) – snp_0183 (79.90 cM), max. LOD: 18.34), 3.13 (c6.loc3 within marker interval: ABG062 (2.20 cM) – snp_0669 (5.90 cM), max. LOD: 5.10) and 59.2 (c7.loc40 within marker interval: snp_0050 (58.20 cM) – snp_0605 (60.20 cM), max. LOD: 16.50) on chromosomes 1, 2, 3, 4, 5, 6 and 7, respectively. The maximum LODs are identified by the MVR-LS method at the positions 0.00 (marker: Hor5, max. LOD: 13.53), 36.30 (marker: Tef4, max. LOD: 56.14), 54.40 (marker: Dfr, max. LOD: 31.49), 141.94 (c4.loc114 within marker interval: ksuH11 (139.00 cM) – Tel4M (148.80 cM), max. LOD: 17.97),

78.97 (c5.loc56 within marker interval: snp_0953 (77.10 cM) – snp_0183 (79.90 cM), max. LOD: 18.70), 3.13 (c6.loc3 within marker interval: ABG062 (2.20 cM) – snp_0669 (5.90 cM), max. LOD: 5.11) and 57.24 (c7.loc39 within marker interval: ABC156D (53.40 cM) – snp_0050 (58.20 cM), max. LOD: 16.39) on chromosomes 1, 2, 3, 4, 5, 6 and 7, respectively. The proposed method identifies the maximum LODs at the positions 0.00 (marker: Hor5, max. LOD: 14.67), 36.30 (marker: Tef4, max. LOD: 55.58), 54.40 (marker: Dfr, max. LOD: 31.91), 141.94 (c4.loc114 within marker interval: ksuH11 (139.00 cM) – Tel4M (148.80 cM), max. LOD: 18.94), 78.97 (c5.loc56 within marker interval: snp_0953 (77.10 cM) – snp_0183 (79.90 cM), max. LOD: 19.62), 3.13 (c6.loc3 within marker interval: ABG062 (2.20 cM) – snp_0669 (5.90 cM), max. LOD: 6.58) and 57.24 (c7.loc39 within marker interval: ABC156D (53.40 cM) – snp_0050 (58.20 cM), max. LOD: 17.41) on chromosomes 1, 2, 3, 4, 5, 6 and 7, respectively. These results indicate that our proposed method identifies almost the same QTL positions as identified by the other two existing methods (MVR-ML and MVR-LS). This means that the proposed method shows same performance as the traditional methods of multi-trait QTL analysis.

Table 3.5: Position of maximum LOD score on each of the chromosomes identified by the MVR-ML, MVR-LS and Proposed methods in barley data

Method	**Chromosome**	**Position (cM)**	**Marker/marker interval**	**Max. LOD**
MVR-ML	1	0.0	Hor5	13.53
	2	36.30	Tef4	56.98
	3	54.40	Dfr	31.49
	4	139.98	c4.loc112 within marker interval: [ksuH11 (139.00 cM) – Tel4M (148.80 cM)]	17.49
	5	78.03	c5.loc55 within marker interval: [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)]	18.34
	6	3.13	c6.loc3 within marker interval: [ABG062 (2.20 cM) – snp_0669 (5.90 cM)]	5.10
	7	59.2	c7.loc40 within marker interval: [snp_0050 (58.20 cM) – snp_0605 (60.20 cM)]	16.50

Method	**Chromosome**	**Position (cM)**	**Marker/marker interval**	**Max. LOD**
MVR-LS	1	0.00	Hor5	13.53
	2	36.30	Tef4	56.14
	3	54.40	Dfr	31.49
	4	141.94	c4.loc114 within marker interval: [ksuH11 (139.00 cM) – Tel4M (148.80 cM)]	17.97
	5	78.97	c5.loc56 within marker interval: [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)]	18.70
	6	3.13	c6.loc3 within marker interval: [ABG062 (2.20) – snp_0669 (5.90 cM)]	5.11
	7	57.24	c7.loc39 within marker interval: [ABC156D (53.40 cM) – snp_0050 (58.20 cM)]	16.39
Proposed	1	0.00	Hor5	14.67
	2	36.30	Tef4	55.58
	3	54.40	Dfr	31.91
	4	141.94	c4.loc114 within marker interval: [ksuH11 (139.00 cM) – Tel4M (148.80 cM)]	18.94
	5	78.97	c5.loc56 within marker interval: [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)]	19.62
	6	3.13	c6.loc3 within marker interval: [ABG062 (2.20 cM) – snp_0669 (5.90 cM)]	6.58
	7	57.24	c7.loc39 within marker interval: [ABC156D (53.40 cM) – snp_0050 (58.20 cM)]	17.41

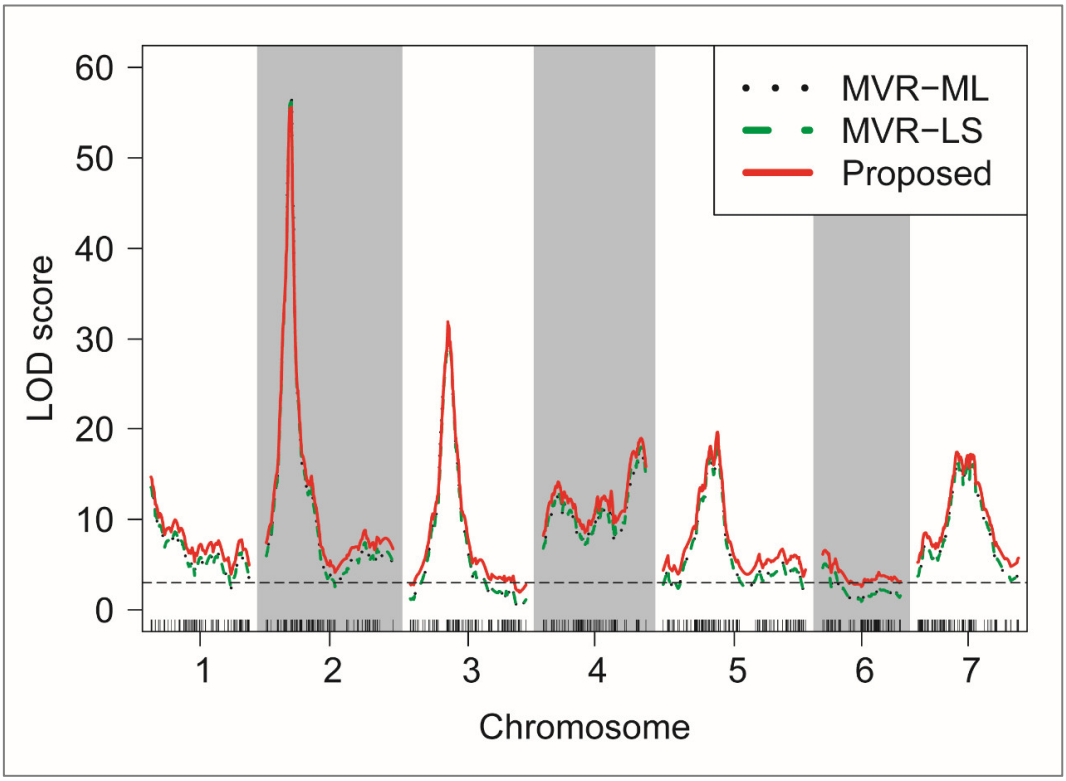


Figure 3.6: LOD score profile plot of genome-wide multi-trait QTL mapping using the MVR-ML (multivariate regression using EM algorithm based maximum likelihood), MVR-LS (multivariate regression using least squares) and Proposed method (multivariate regression using the properties of multivariate normal distribution of phenotypes and conditional probabilities of putative QTL genotype given the marker genotypes) with barley data.

Table 3.7 shows the computation times for all the three methods (MVR-ML, MVR-LS and Proposed). When we compare the computation time of the proposed method with the other two existing methods, it reveals that the required computation time of the proposed method (1.91 seconds) is less than that of the other two methods MVR-ML (48.8 seconds) and MVR-LS (2.62 seconds).

3.3.2.2 Nine-trait QTL Analysis with Mouse Data

We have also investigated the performance of the proposed method in a comparison of the two existing methods (MVR-ML and MVR-LS) by applying these methods with the mouse dataset (Leiter et al., 2009) of BC population. Table 3.6 shows the chromosomal location (i.e., locus position) at which maximum LOD occurs on each

of the 19 chromosomes along with the marker at that locus position or the "marker interval" containing that locus position and the maximum LOD score value for each of the three methods. Figure 3.7 represents the LOD score profile plot of genome-wide QTL analysis of mouse data using all the three different methods (MVR-EM, MVR-LS and Proposed). Multi-trait QTL analysis of mouse data shows that our proposed method identifies almost the same QTL positions as identified by the other two methods (MVR-EM and MVR-LS). However, the comparison of computation times of three methods (Table 3.7) reveals that the required computation time of the proposed method (2.05 sec) is smaller than that of the other two methods MVR-EM (70.47 sec) and MVR-LS (3.20 sec).

Table 3.6: Position of maximum LOD score on each of the chromosomes identified by the MVR-ML, MVR-LS and Proposed methods in mouse data

Method	Chromosome	Position (cM)	Marker/marker interval	Max. LOD
MVR-ML	1	78.03	01_172244784_N	13.60
	2	45.94	02_078062303_M	8.27
	3	26.38	c3.loc23 within marker interval: [03_049202892_M (21.67 cM) – 03_065861493_N (30.15 cM)]	3.52
	4	19.88	04_037053163_M	3.16
	5	8.10	05_015780506_G	5.7913
	6	65.38	06_134747045_G	4.0674
	7	51.86	07_084052305_M	2.4932
	8	72.10	08_123340602_N	2.6327
	9	24.84	09_044591533_N	7.4291
	10	51.98	c10.loc47 within marker interval: [10_091119681_M (45.44 cM) – 10_108166251_N (56.65 cM)]	4.452
	11	5.00	11_008353761_M	2.5189
	12	60.71	12_104882822_M	6.0282
	13	46.14	c13.loc43 within marker interval: [13_056700488_N (30.50 cM) – 13_094920623_M (51.03 cM)]	4.757
	14	21.29	14_031978791_N	7.8529
	15	49.57	c15.loc39 within marker interval: [15_087100507_M (40.68 cM) – 15_097455228_N (52.53 cM)]	3.667

Method	Chromosome	Position (cM)	Marker/marker interval	Max. LOD
	16	39.93	c16.loc36 within marker interval: [16_061226828_N (36.06 cM) – 16_080711577_N (45.73 cM)]	3.481
	17	5.53	17_008378982_M	3.7231
	18	44.19	18_070565016_N	1.5307
	19	3.21	19_007376322_N	3.771
MVR-LS	1	79.01	c1.loc64 within marker interval: [01_172244784_N (78.03 cM) – 01_182577000_P (84.93 cM)]	13.361
	2	45.94	02_078062303_M	8.2675
	3	6.89	c3.loc5 within marker interval: [03_007561998_N (2.02 cM) – 03_027974740_N (11.75 cM)]	3.5684
	4	21.84	c4.loc19 within marker interval: [04_037053163_M (19.88 cM) – 04_055642665_N (31.66 cM)]	3.2201
	5	7.32	c5.loc3 within marker interval: [05_014236855_M (6.54 cM) – 05_015780506_G (8.10 cM)]	5.7913
	6	65.38	06_134747045_G	4.0719
	7	51.86	07_084052305_M	2.5087
	8	72.10	08_123340602_N	2.6342
	9	24.84	09_044591533_N	7.4396
	10	42.83	10_086567143_M	3.7805
	11	5.00	11_008353761_M	2.5292
	12	60.71	12_104882822_M	6.0772
	13	50.05	c13.loc47 within marker interval: [13_056700488_N (30.50 cM) – 13_094920623_M (51.03 cM)]	4.6155
	14	21.29	14_031978791_N	7.8439
	15	49.57	c15.loc39 within marker interval: [15_087100507_M (40.68 cM) – 15_097455228_N (52.53 cM)]	3.715
	16	40.90	c16.loc37 within marker interval: [16_061226828_N (36.06 cM) – 16_080711577_N (45.73 cM)]	3.6396
	17	4.53	17_008378982_M	3.72
	18	38.49	c18.loc33 within marker interval: [18_049823654_M (27.09 cM) – 18_070565016_N (44.19 cM)]	1.6302

Method	Chromosome	Position (cM)	Marker/marker interval	Max. LOD
	19	3.21	19_007376322_N	3.771
Proposed	1	79.01	c1.loc64 within marker interval: [01_172244784_N (78.03 cM) – 01_182577000_P (84.93 cM)]	13.118
	2	45.94	02_078062303_M	8.259
	3	6.89	c3.loc5 within marker interval: [03_007561998_N (2.02 cM) – 03_027974740_N (11.75 cM)]	3.7755
	4	21.84	c4.loc19 within marker interval: [04_037053163_M (19.88 cM) – 04_055642665_N (31.66 cM)]	3.4433
	5	8.10	05_015780506_G	5.8964
	6	65.38	06_134747045_G	4.2556
	7	51.86	07_084052305_M	2.7643
	8	72.10	08_123340602_N	2.8821
	9	24.84	09_044591533_N	7.469
	10	42.83	10_086567143_M	3.9776
	11	5.00	11_008353761_M	2.7836
	12	60.71	12_104882822_M	6.1691
	13	50.05	c13.loc47 within marker interval: [13_056700488_N (30.50 cM) – 13_094920623_M (51.03 cM)]	4.7744
	14	21.29	14_031978791_N	7.8547
	15	49.57	c15.loc39 within marker interval: [15_087100507_M (40.68 cM) – 15_097455228_N (52.53 cM)]	3.9151
	16	40.90	c16.loc37 within marker interval: [16_061226828_N (36.06 cM) – 16_080711577_N (45.73 cM)]	3.8434
	17	4.53	c17.loc1 within marker interval: [17_003335010_M (3.54 cM) – 17_008378982_M (5.53 cM)]	3.9304
	18	38.49	c18.loc33 within marker interval: [18_049823654_M (27.09 cM) – 18_070565016_N (44.19 cM)]	1.9265
	19	3.21	19_007376322_N	3.9686

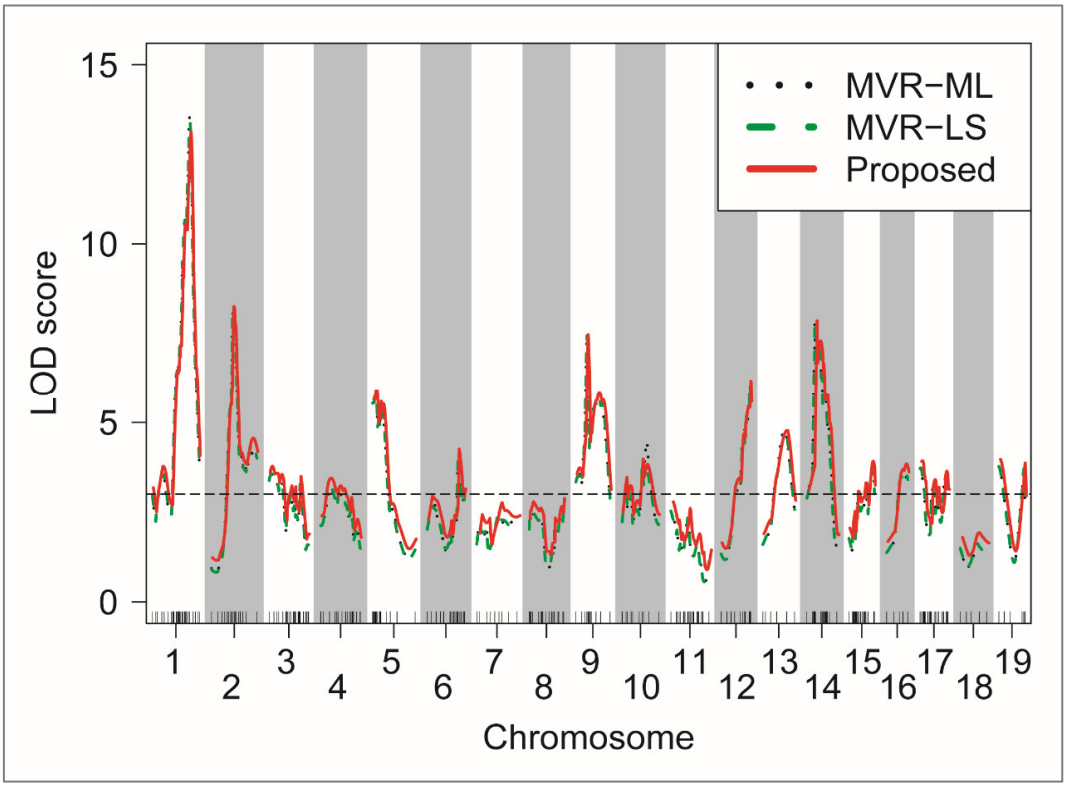


Figure 3.7: LOD score profile plot of genome-wide multi-trait QTL mapping using the MVR-ML (multivariate regression using EM algorithm based maximum likelihood), MVR-LS (multivariate regression using least squares) and Proposed method (multivariate regression using the properties of multivariate normal distribution of phenotypes and conditional probabilities of putative QTL genotype given the marker genotypes) with mouse data of BC population considering 9 phenotypes (BMC, AREA, LEPTIN, INSULIN, CHOL, HDLD, GLUCOSE, NEFA and TG).

Table 3.7: Comparison of computational times of multi-trait QTL analysis among three methods (MVR-ML, MVR-LS and **Proposed**) with Barley and Mouse datasets

Data set	No. of chromosomes	Phenotypic size[a]	Genotypic size[b]	Computation time (in second)		
				MVR-ML[c]	MVR-LS[d]	Proposed[e]
Barley data	7	8×150	150×493	48.8	2.62	1.91
Mouse data	19	9×141	141×303	70.47	3.20	2.05

[a] Phenotypic size indicates (Phenotypes×Indviduals).
[b] Genotypic size indicates (Individuals×Markers).
[c] MVR-ML: Multivariate regression (MVR) using maximum likelihood (ML) method.
[d] MVR-LS: Multivariate regression (MVR) using least squares (LS) method.
[e] Proposed: Multivariate regression (MVR) using the properties multivariate normal distribution.

We found that in all cases our proposed method identified the same QTL positions but it takes less computation compared to the existing methods (MVR-ML and MVR-LS) with both simulated and real data. The MVR-ML approach takes more time because it is an iterative procedure based on EM algorithm. Moreover, in MVR-ML approach, the calculation of likelihood ratio (LR) statistic or LOD statistic is time consuming because it is a five steps process: (i) estimation of regression parameters, (ii) estimation of residuals and its variance-covariance matrix, (iii) estimation of Wilks' lambda statistic, (iv) Chi-square approximation of Wilks's lambda statistic, and (v) calculation of LR or LOD statistic based on Chi-square statistic. Although the least squares based multivariate regression (MVR-LS) is not an iterative process and it requires less computation time than MVR-ML, it takes more time than our proposed method. Because, in MVR-LS approach, the calculation of LR or LOD statistic is also a five steps process like MVR-ML approach. Our proposed method takes comparatively very less time because the calculation of LR or LOD statistic is very straight forward in this method. In our proposed method, the likelihood ratio statistic (or LOD statistic) can be calculated only based on the sample variance-covariance matrix of phenotypes and the conditional probability of QTL genotype given the marker genotypes.

3.4 Conclusion

We have introduced a new and fast approach for multi-trait QTL analysis using the properties of multivariate normal distribution with backcross population. Our proposed method is able to identify the same QTL positions as identified by the existing methods (MVR-ML and MVR-LS) of multi-trait QTL mapping and it takes very less computation time than other two existing methods. This improvement in computation time is very advantageous when the number of phenotypes and individuals are very large, and the markers are very dense resulting in a QTL data set of very big size.

Chapter 4

Robustification of Regression Based Fast Multi-trait QTL Analysis (Proposed)

4.1 Introduction

In many line crossing experiments of genome-wide QTL mapping studies, measurements are taken on multiple traits along with the marker genotypes. Very often, such traits are correlated and there are common chromosome regions (chromosomal locations) that affect multiple traits (Chen, 2016b). Trait-by-trait analysis using single-trait simple interval mapping (SIM) methods (Haley and Knott, 1992; Haley et al., 1994; Lander and Botstein, 1989) cannot take into account the pleotropic effects. The joint analyses of multiple traits, which include all quantitative traits together in a single model, can increase the power of QTL identification and improve the QTL localization accuracy when multiple traits are correlated genetically in the population (Xu, 2013a). In addition, QTL mapping considering multiple quantitative traits using joint analyses can give insights into the important genetic mechanisms underlying the trait relationships (e.g., genetic linkage versus pleiotropy), which would otherwise be hard to address if multiple traits are analyzed one-by-one. Therefore, joint analyses of multiple traits are very essential to identify important QTL locations which control multiple traits simultaneously.

Many methods for multi-trait QTL mapping have been developed in the literature, ranging from simple extensions of single-trait approaches to sophisticated multi-trait approaches designed specifically for multi-trait QTL mapping. Substantial work has

been done in joint mapping for multiple quantitative traits (Almasy and Blangero, 1998; Hackett et al., 2001; Henshall and Goddard, 1999; Jiang and Zeng, 1995; Knott and Haley, 2000; Korol et al., 2001; Korol et al., 1995; Mangin et al., 1998; Williams et al., 1999). Least squares based multivariate regression (MVR-LS) for multi-trait SIM (Knott and Haley, 2000) and multi-trait SIM using maximum likelihood (i.e., using EM algorithm)(Dempster et al., 1977) based multivariate regression (Xu, 2013a) are two most popular and widely used approaches for multi-trait QTL analysis.

Although MVR-LS and maximum likelihood based multivariate regression (MVR-ML) are the most popular approaches for multi-trait QTL mapping, these methods have some limitations. As discussed in Chapter 3, the MVR-ML based multi-trait SIM is very time consuming due to EM algorithm based estimation which is an iterative process. Moreover, in MVR-ML approach, the calculation of likelihood ratio (LR) or log of odds (LOD) statistic is time consuming because it is a five steps process: (i) estimation of regression parameters, (ii) estimation of residuals and its variance-covariance matrix, (iii) estimation of Wilks' lambda statistic, (iv) Chi-square approximation of Wilks's lambda statistic, and (v) calculation of LR or LOD statistic based on Chi-square statistic. MVR-LS (Knott and Haley, 2000) is an alternative approach to MVR-ML to reduce the computation time for multi-trait QTL analysis. However, in Chapter 3, we have discussed the fast multi-trait (FMT) QTL mapping approach which is less time consuming than MVR-LS and MVR-ML approaches and it has almost similar performance to the MVR-LS and MVR-ML approaches. Hence, the FMT QTL mapping is a better approach than all the existing approaches for multi-trait QTL analysis.

Although FMT QTL mapping is a better approach than all the existing approaches of multi-trait QTL analysis, it is very sensitive to outliers and provide misleading results when the data are contaminated by phenotypic outliers. In this study, we have proposed a robust approach for multiple traits QTL mapping by robustifying the FMT QTL mapping approach (treated as classical approach) using minimum β-divergence method (Mihoko and Eguchi, 2002; Mollah et al., 2007). Both the methods (Classical

and Proposed) produce similar results in absence of outliers. However, only our proposed method is able to identify the same QTL positions as identified in absence of outliers.

4.2 Methods and Materials

4.2.1 Classical Fast Multi-trait QTL Analysis

Let us consider no epistasis between two QTLs, no interference in crossing over, and only one QTL in the testing interval. In this study, we have only considered Backcross (BC) population. Methods for other mapping populations, such as F_2 and double haploid (DH), are simple extension of that for the BC population with some modifications. Let $\boldsymbol{Y}_j = [Y_{j1} \quad Y_{j2} \quad \dots \quad Y_{jm}]$, $j = 1, 2, \dots, n$, be an $(1\times m)$ vector for m phenotypic traits measured from the j^{th} individual of a mapping population. Then the linear regression model for BC population with m traits can be formulated in matrix notation as

$$\underset{(1\times m)}{\boldsymbol{Y}_j} = \underset{(1\times m)}{\boldsymbol{\alpha}} + \underset{(1\times 1)}{X_j}\underset{(1\times m)}{\boldsymbol{\gamma}} + \underset{(1\times m)}{\boldsymbol{\varepsilon}_j}, \qquad j = 1, 2, \dots, n \tag{4.1}$$

where $X_j = p_{j|i}$ denote the conditional probability of the putative QTL genotype i $(i = 1, 2)$ given the flanking marker genotypes for the j^{th} individual (Table 3.1), $\boldsymbol{\alpha}$ is a $(1 \times m)$ vector for general mean effects and $\boldsymbol{\gamma}$ is a $(1 \times m)$ vector for the additive QTL effects, and $\boldsymbol{\varepsilon}_j$ is a $(1 \times m)$ vector for the random errors. The vector of random errors is assumed to be $\boldsymbol{\varepsilon}_j \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $(m \times m)$ variance-covariance matrix of random errors.

In chapter 3, we have discussed the regression based FMT QTL mapping approach in details. In this chapter, we have just recall some important formula/equations.

Let us consider that $\boldsymbol{Z} = \left[\underset{(m\times 1)}{\boldsymbol{Y}} \vdots \underset{(1\times 1)}{X}\right]^T$ follows a multivariate normal distribution $N\left(\underset{(m+1)\times 1}{\boldsymbol{\mu}_{\boldsymbol{Z}}}, \underset{(m+1)\times(m+1)}{\boldsymbol{\Sigma}_{\boldsymbol{Z}}}\right)$ with mean vector $\boldsymbol{\mu}_{\boldsymbol{Z}} = \begin{bmatrix}\boldsymbol{\mu}_Y \\ \mu_X\end{bmatrix}$ and variance-covariance

matrix $\boldsymbol{\Sigma_Z} = \begin{bmatrix} \boldsymbol{\Sigma_{YY}} & \boldsymbol{\Sigma_{YX}} \\ \boldsymbol{\Sigma_{XY}} & \sigma_{XX} \end{bmatrix}$, where $\boldsymbol{Y}$ and X have been introduced in (4.1). Then, from equation (3.10) and (3.11), based on a random sample of size n the maximum likelihood estimators of the regression parameters are

$$\widehat{\boldsymbol{\alpha}} = \overline{\boldsymbol{Y}} - \mathbf{S_{YX}} S_{XX}^{-1} \bar{X} \tag{4.2}$$

and

$$\widehat{\boldsymbol{\gamma}} = \mathbf{S_{YX}} S_{XX}^{-1} \tag{4.3}$$

where $\overline{\boldsymbol{Y}} = \left[\bar{Y}_1 = \frac{1}{n}\sum_{j=1}^{n} Y_{j1} \quad \bar{Y}_2 = \frac{1}{n}\sum_{j=1}^{n} Y_{j2} \quad \cdots \quad \bar{Y}_m = \frac{1}{n}\sum_{j=1}^{n} Y_{jm} \right]$

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j ,$$

$$\boldsymbol{S_{YY}} = \frac{1}{n-1}\sum_{j=1}^{n} \left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}\right)^T ,$$

$S_{XX} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2$ and

$$\boldsymbol{S_{XY}} = \boldsymbol{S_{YX}} = \frac{1}{n-1}\sum_{j=1}^{n} \left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}\right)\left(X_j - \bar{X}\right)^T$$

are the maximum likelihood estimates of $\boldsymbol{\mu_Y}$, μ_X, $\boldsymbol{\Sigma_{YY}}$, $\boldsymbol{\Sigma_{YX}} = \boldsymbol{\Sigma_{XY}}$ and σ_X^2, respectively.

From equation (3.12), the maximum likelihood estimator of the regression function is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\gamma}} X = \overline{\boldsymbol{Y}} - \boldsymbol{S_{YX}} S_{XX}^{-1} \bar{X} + \boldsymbol{S_{YX}} S_{XX}^{-1} X = \overline{\boldsymbol{Y}} + \boldsymbol{S_{YX}} S_{XX}^{-1} (X - \bar{X}) \tag{4.4}$$

In multi-trait QTL analysis, we want to test the null hypothesis is $H_0: \boldsymbol{\gamma} = \mathbf{0}$ (i.e., there is no QTL at a given position within a marker interval) against $H_1: H_0$ is not true. From equation (3.15), the likelihood ratio tests (LRT) statistic to test H_0 is as follows

$$\text{LRT} = -n \ln\left(\frac{\left|\widehat{\boldsymbol{\Sigma}}\right|}{\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}\right|}\right) \tag{4.5}$$

where $\widehat{\boldsymbol{\Sigma}} = \left(\frac{n-1}{n}\right)(\boldsymbol{S_{YY}} - \boldsymbol{S_{YX}}S_{XX}^{-1}\boldsymbol{S_{XY}})$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}} = \left(\frac{n-1}{n}\right)\boldsymbol{S_{YY}}$ are the estimated variance-covariance matrices of residuals under the full model and the reduced model (i.e., under H_0), respectively [see equations (3.13)and (3.14)].

For large *n*, from equation (3.17), the modified likelihood ratio test statistic is

$$\text{LRT} = -\left[n - k - 1 - \frac{1}{2}(m - k + 1)\right]\ln\left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}}|}\right) \tag{4.6}$$

where n is the number of individuals (i.e., observations), m is the number of phenotypes, and k is the *"test degrees of freedom"* which is the number of predictor variables.

Under the null hypothesis (H_0), the modified LRT statistic is expected to have an approximate chi-square distribution with *mk* degrees of freedom for a given QTL position in the genome. However, the threshold value to reject the null hypothesis (H_0) cannot be simply chosen from the χ^2 distribution because of the violation of regularity conditions of asymptotic theory under H_0. An alternative way is to use log of odds (LOD) score (Lander and Botstein, 1989; Ott, 1999; Terwilliger and Ott, 1994; Wu et al., 2007; Xu, 2013d) as a test statistic to test the null hypothesis of no QTL (H_0). The LOD score is the transformation of the LRT statistic, defined as

$$\begin{aligned}\text{LOD} &= \frac{\text{LRT}}{2\times\log(10)} = \frac{\text{LRT}}{4.605} = 0.217\ \text{LRT} \\ &= -0.217\left[n - k - 1 - \frac{1}{2}(m - k + 1)\right]\ln\left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}}|}\right)\end{aligned} \tag{4.7}$$

According Lander and Botstein (1989), the typical threshold of LOD score should be between 2 and 3 to ensure a 5% overall false positive error for identifying a QTL. Terwilliger and Ott (1994), Ott (1999), Wu et al. (2007), and Xu (2013d) suggested a value of LOD = 3 as the critical threshold for declaring the existence of QTL. Thus, the LOD > 3 can be used as a criterion to declare a significant QTL.

4.2.2 Robustification of Fast Multi-trait QTL Analysis (Proposed)

One main problem of classical fast multi-trait (FMT) QTL mapping approach (discussed in Chapter 3) is that it is very sensitive to outliers and produce misleading results when the data is contaminated by phenotypic outliers. In most of the QTL experiments, often the phenotypic data are contaminated by some extreme measurement values. So, we need a robust approach for multi-trait QTL analysis to obtain the robust estimates of the model parameters and the robust test statistic which are resistant against phenotypic outliers. In this section, we have discussed the robustification of the multivariate regression based classical FMT QTL mapping approach using minimum β–divergence method (Mihoko and Eguchi, 2002; Mollah et al., 2007) to obtain the robust estimates of model parameters and the robust likelihood ratio test (LRT) statistic.

From equation (4.2) to (4.7), we observe that the classical estimates of the multivariate regression model (4.1), estimates of the residual variance-covariance matrices $\widehat{\boldsymbol{\Sigma}}$ (under full model) and $\widehat{\boldsymbol{\Sigma}}_0$ (under null model), and the calculation of LRT or LOD statistic depend only on the sample mean vector, $\overline{\boldsymbol{Z}} = [\overline{\boldsymbol{Y}} \quad \bar{X}]^T$, and sample variance-variance matrix $\boldsymbol{S_{ZZ}} = \begin{bmatrix} \boldsymbol{S_{YY}} & \boldsymbol{S_{YX}} \\ \boldsymbol{S_{XY}} & S_{XX} \end{bmatrix}$. Hence, if we can robustify the sample mean vector and sample variance-variance matrix, then we can obtain the robust estimates of the regression parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$) and residual variance-covariance matrices ($\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$), and construct robust test statistics (LRT or LOD statistic). In other words, we can robustify the multivariate regression based FMT QTL mapping approach if we can obtain the robust estimates of parameter $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ of multivariate normal distribution.

The minimum β-divergence estimators of the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu_Z}, \boldsymbol{\Sigma_Z})$ can be computed iteratively as follows:

$$\boldsymbol{\mu}_{\boldsymbol{Z},\, t+1} = \frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)\boldsymbol{z}_j}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)} \tag{4.8}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{Z},\,t+1} = (1+\beta)\frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},\,t})(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},\,t})^T}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)}, \tag{4.9}$$

where $w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t), j = 1, 2, \ldots, n$, is called the β-weight function and defined as

$$w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t) = \exp\left[-\frac{\beta}{2}(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},\,t})^T \boldsymbol{\Sigma}_{\boldsymbol{Z},\,t}^{-1}(\boldsymbol{z}_j - \boldsymbol{\mu}_{\boldsymbol{Z},\,t})\right] \tag{4.10}$$

The value of β-weight function ranges from 0 to 1. The tuning parameter β plays an important role to control the performance of the proposed method. The appropriate value of β can be selected by k-fold cross validation. If β = 0, then (4.8) and (4.9) reduces to the classical non-iterative solution and the estimates reduce to classical estimates.

Let the robust estimate (i.e., β-estimate) of the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$ are denote by $\widehat{\boldsymbol{\theta}}_{(\beta)} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)}\right)$. Then we can write

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}(\beta)} \\ \hat{\mu}_{X(\beta)} \end{bmatrix}_{(m+1)\times 1} = \begin{bmatrix} \overline{\boldsymbol{Y}}_{(\beta)} \\ \bar{X}_{(\beta)} \end{bmatrix}_{(m+1)\times 1} \tag{4.11}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} & \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}} \\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} & \hat{\sigma}_{XX} \end{bmatrix}_{(m+1)\times(m+1)} = \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{YY}(\beta)} & \boldsymbol{S}_{\boldsymbol{YX}(\beta)} \\ \boldsymbol{S}_{\boldsymbol{XY}(\beta)} & S_{XX(\beta)} \end{bmatrix} \tag{4.12}$$

Then, using (4.11) and (4.12), in (4.2) and (4.3), the robust estimates of the regression parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$) can be written as

$$\widehat{\boldsymbol{\alpha}}_{(\beta)} = \overline{\boldsymbol{Y}}_{(\beta)} - \mathbf{S}_{\boldsymbol{YX}(\beta)}\mathrm{S}_{XX(\beta)}^{-1}\bar{X}_{(\beta)} \tag{4.13}$$

and

$$\widehat{\boldsymbol{\gamma}}_{(\beta)} = \mathbf{S}_{\boldsymbol{YX}(\beta)}\mathrm{S}_{XX(\beta)}^{-1} \tag{4.14}$$

Using (4.11) and (4.12) in (4.6), the robust LRT statistic can be written as

$$\mathrm{LRT}_{(\beta)} = -\left[n - k - 1 - \frac{1}{2}(m - k + 1)\right]\ln\left(\frac{\left|\widehat{\boldsymbol{\Sigma}}_{(\beta)}\right|}{\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}(\beta)}\right|}\right) \tag{4.15}$$

where n is the number of individuals (i.e., observations), m is the number of phenotypes, and k is the *"test degrees of freedom"* which is the number of predictor variables.

Using (4.11) and (4.12) in (4.7), the robust LOD statistic can be written as

$$\begin{aligned} \mathrm{LOD}_{(\beta)} &= 0.217\ \mathrm{LRT}_{(\beta)} \\ &= -0.217\left[n-k-1-\frac{1}{2}(m-k+1)\right]\ln\left(\frac{\left|\widehat{\boldsymbol{\Sigma}}_{(\beta)}\right|}{\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}(\beta)}\right|}\right) \end{aligned} \tag{4.16}$$

4.2.3 Expression Single Nucleotide Polymorphisms (eSNPs) Mapping by Using the Proposed Robust Multi-trait QTL Mapping Model

The Quantitative Trait Locus (QTL) mapping has been highly successful in determining causative loci underlying several disease phenotypes. Several statistical techniques (i.e. association test, t-test, likelihood or regression approaches) are used to map QTL, with flanking markers to identify the association between the genomic location and the phenotypic traits (Haley and Knott, 1992; Lander and Botstein, 1989). After the advances in gene expression profile datasets, the QTL analysis extended to expression QTL (known as eQTL) analysis which became the more effective approach to identify the biomarker genes associated with the trait variations. However, these QTL and eQTL approaches were not suitable for human populations. Taking the advantages of next generation sequencing (NGS), these technologies have been extended to expression Single Nucleotide Polymorphisms (eSNPs) analysis which is suitable for human population also (Frazer et al., 2007; Gatti et al., 2009; Szatkiewicz et al., 2008). These eSNPs are gene polymorphisms that explain variation in expression levels of mRNAs. An expression trait is an amount of an mRNA transcript or a protein. These are usually the product of a gene with its specific SNPs. This distinguishes expression traits from most complex traits, which are not the product of the expression of that gene. SNPs that explain variance in expression traits are said to be eSNPs located near the gene-of-origin (gene which produces the transcript or protein) are referred to as local eSNPs. By contrast, those located distant from their gene of origin, often on different chromosomes are referred to as distant

eSNPs. Often, these two types of eSNPs are referred to as cis and trans, respectively, but these terms are best reserved for instances when the regulatory mechanism (cis vs. trans) of the underlying sequence has been established. Some cis eQTLs are detected in many tissue types but the majority of trans eQTLs are tissue-dependent (dynamic). eSNPs may act in cis (locally) or trans (at a distance) to a gene. The abundance of a gene transcript is directly modified by SNPs in regulatory elements. Consequently, transcript abundance might be considered as a quantitative trait that can be mapped with considerable power. The combination of whole-genome genetic/SNPs association studies and the measurement of global gene expression allow the systematic identification of eSNPs. By assaying gene expression and SNP variation simultaneously on a genome-wide basis in a large number of individuals, statistical genetic methods can be used to map the genetic factors that underpin individual differences in quantitative levels of expression of many thousands of transcripts.

The calculation for eSNPs creates a computational challenge that can stretch or overwhelm existing tools. These challenges are further compounded by multiple comparison issues arising from the large number of available SNPs and transcripts. Additionally, the majority of identified SNPs in GWAS are located within the non-coding regions (e.g. approximately 88% lie in intergenic or intronic regions) and their causal genetic function remains largely unknown (Zeng et al., 2017). Various methods have been used to address these issues. Multiple comparisons among transcripts has been previously addressed by thresholding transcripts using q-values (Storey and Tibshirani, 2003) obtained from transcript specific testing of association with SNPs using Likelihood Ratio Statistic (LRS) (Chesler et al., 2005) or the mixture over markers method (Kendziorski et al., 2006) or Pearson correlation matrix (Gatti et al., 2009). Thus, there are several methods those are used to overcome the computational problem in the eSNPs based genome-wide association study (GWAS) to identify the biomarker genes and SNPs. However, all methods as early mentioned are sensitive to outlying observations, where outlying observations usually occurs in the gene expression data due to several steps involved in the data generating processes. So these existing methods may produces misleading results. To overcome this problems, therefore, an attempt is made to apply our proposed robust multi-trait QTL mapping approach as an eSNPs approach to identify the biomarker genes and SNPs including

cis and trans regulatory factors. To implement the proposed multi-trait QTL mapping approach to eSNP analysis, detail discussion is given below:

Step 1: Select top g (we have considered $g = 10$ in this study) differentially expressed (DE) genes. To select the top g DE genes we have used idea the DE gene selection method described by Kabir et al. (2016). However, instead of classical sample variance we have used the robust sample variance in the method of top DE gene selection. The robust sample variance can be calculated using (4.12). The top DE gene selection method is composed of the following steps:

Step 1.1: Perform hierarchical clustering with the GE values to stratify the genes into 3 groups/clusters.

Step 1.2: Calculate robust variance of each on the gene expression variables (i.e., each of the genes) in each of the 3 clusters. Calculate average of the variances in each cluster.

Step 1.3: Now remove the cluster with the lowest average variance, which is the cluster of equally expressed genes. So, we get 2 clusters of DE genes.

Step 1.4: Select 5 genes with higher variances, called top 5 DE genes, from each of the 2 clusters of DE genes. So, we get total top 10 DE genes from the 2 clusters of DE genes, which are the resulting (finally selected) top 10 DE genes.

Step 2: Consider the expression values of each of the top 10 finally selected DE genes as the values of a phenotype. So we will have 10 phenotypes that are gene expression of top 10 DE genes.

Step 3: Use the multi-trait SIM approaches with those 10 expression phenotypes and SNP data to calculate LOD scores.

Step 4: Create LOD score plot and identify the positions where the highest peaks occurs in each of the chromosomes. The positions with highest peaks are the important eSNP positions.

The above steps are summarized and shown in the following Figure 4.1.

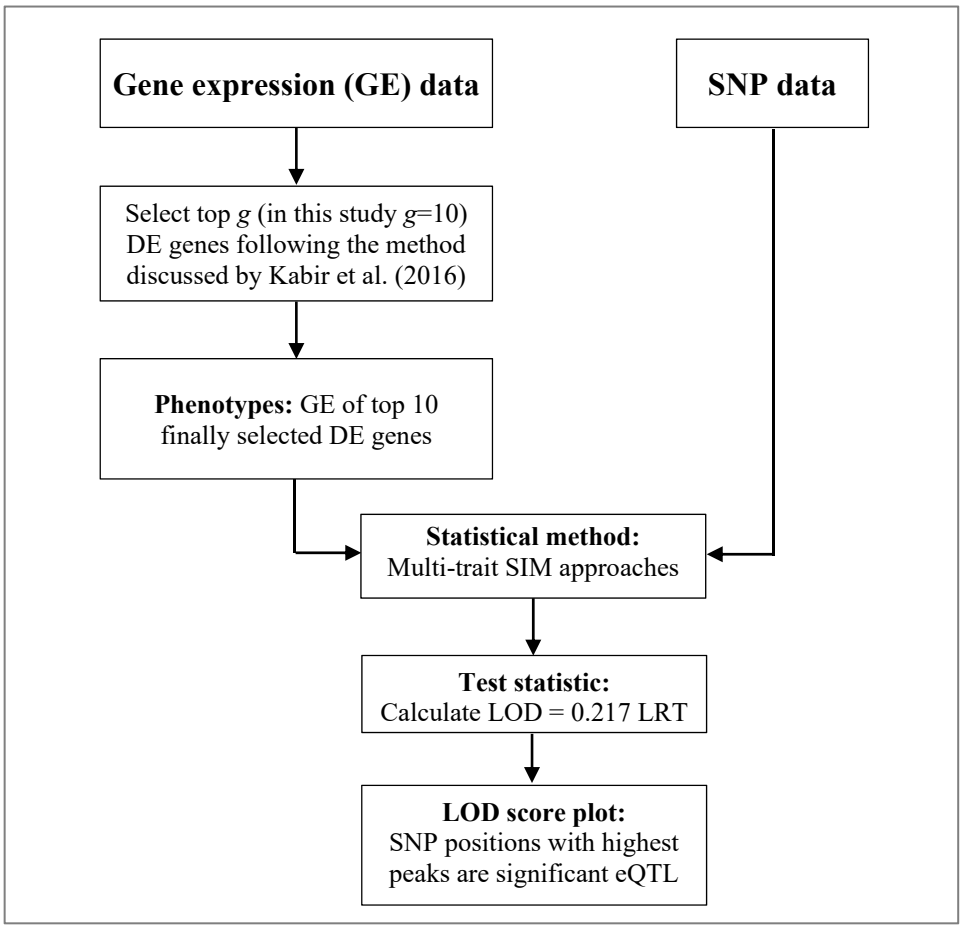


Figure 4.1: Flowchart of multi-trait eQTL analysis.

4.2.4 Simulated Dataset

To investigate the performance of the proposed robust method in comparison with the classical FMT QTL mapping approach, we have generated one artificial/synthetic dataset for BC population based on the multivariate linear regression model (4.1) using simulation technique. To generate the simulated data we have considered 3 phenotypes (denoted as Pheno1, Pheno2 and Pheno3), 250 individuals, total 13 chromosomes, and 15 equally spaced (5 cM) markers on each of the 13 chromosomes resulting in a total of 195 equally spaced markers distributed on 13 chromosomes. We have simulated total four unlinked QTLs to affect three quantitative traits where the true QTLs are located on chromosomes 2, 4, 6 and 8 at marker 5 (locus position 20 cM). Among the simulated QTLs, the QTL on chromosome 6 has been considered as a pleiotropic QTL affecting the phenotypes Pheno1 and Pheno3 simultaneously. To generate the simulated data based on the multivariate linear regression model (4.1),

we have considered the true values of regression parameters as $\boldsymbol{\alpha} = [0.50 \quad 1.00 \quad 1.25]$, $\boldsymbol{\gamma} = [1.50 \quad 1.25 \quad 2.25]$ and true variance-covariance matrix of random errors $\boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & 0.3 & 0.4 \\ 0.3 & 0.5 & 0.035 \\ 0.4 & 0.035 & 0.5 \end{bmatrix}$. We have generated 250 trait values for each of the three phenotypes with the heritability values $h^2 = 0.69, 0.61$ and 0.84 for Pheno1, Pheno2 and Pheno3, respectively. This means that 69%, 61% and 84% of the variation in Pheno1, Pheno2 and Pheno3, respectively, are controlled by QTLs, and the remaining 31%, 39% and 16% variation in Pheno1, Pheno2 and Pheno3, respectively, are subject to the environmental and other effects (random error). To compare the robustness of our proposed method with the FMT QTL mapping method in presence of phenotypic outliers, we have generated contaminated phenotypic data by contaminating 20% of the values of each of the three phenotypes with outliers. To perform the simulation study, we have used R version 3.6.2 software along with the R-package R/qtl (Broman et al. (2003), homepage: http://www.rqtl.org/).

4.2.5 Real QTL Dataset (Barley Data)

We have also investigated the performance of the proposed robust method in comparison with the classical FMT QTL mapping method using one real dataset of barley. The barley dataset was originally published by (Hayes et al., 1993). We have obtained the barley dataset from the web-database GeneNetwork (http://www.genenetwork.org/). We have described the barley dataset in detail in **Chapter 3** (Section 3.2.3.1). To investigate the robustness of our proposed method in comparison with the FMT QTL mapping method, we have employed both the methods with contaminated barley data. We have contaminated 20% values of each of the eight phenotypes (grain yield, heading date, plant height, lodging, grain protein, alpha amylase, diastatic power, and malt extract) with outliers to get the contaminated phenotypic barley data.

4.2.6 Real eSNPs Dataset (Gene Expression and SNP Data of BXD Mouse)

We have also implemented the methods of multi-trait SIM (Classical and Proposed) with gene expression (GE) phenotype and SNP data of mouse as an extended application of the multi-trait SIM approaches in the field of expression SNPs (eSNPs) analysis. The description of the GE in livers and SNP data of 32 recombinant inbred (RI) mouse strains are given below.

BXD gene expression data: The GE dataset of gene expression in Liver of BXD RI mouse strains has been obtained from the FastMap software (Gatti et al., 2009; Gatti et al., 2011) of e-QTL analysis (http://comptox.us/fastmap.php) and has been described by Gatti et al. (2007). Briefly, this GE dataset consists of the expression measurements derived using the Agilent oligonucleotide microarrays for 36182 transcripts in 32 BXD RI strains and the DBA/2J and C57BL/6J parentals of mouse. The GE data provided by FastMap software were normalized using the microarray database of University of North Carolina (UNC). For our convenience to apply the multi-trait SIM approaches as eSNPs analysis, we have considered only the top 10 DE genes/transcripts selected using the DE gene selection method described by Kabir et al. (2016) based on robust sample variances of transcripts. To investigate the performance of the proposed method in comparison with the classical approach in presence of phenotypic outliers, we have contaminated 20% GE values of each of the 10 transcripts (phenotypes) considered in our study. To get a clear concept about the structure of the BXD gene expression data, a part of the BXD gene expression data has been presented in Figure A4.1.

BXD SNP data: We have downloaded the BXD SNP dataset also from the FastMap software (Gatti et al., 2009; Gatti et al., 2011) of eQTL analysis (http://comptox.us/fastmap.php). The BXD SNP dataset consists of 156525 SNPs distributed on 20 chromosomes of mouse. Chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 contain 13615, 11741, 10781, 9782, 9545, 8312, 9091, 8789, 8267, 5467, 8684, 7226, 7138, 8441, 7066, 5226, 5976, 5847, 3880 and 1651 SNPs, respectively. A part of the mouse data has been displayed in Figure A4.2.

4.3 Results and Discussion

4.3.1 Simulated Data Analysis Results

We have employed both the methods (Classical and Proposed) with the simulated data of the BC population and compared their performance of QTL detection. For each of the methods (Classical and Proposed) of multi-trait QTL analysis, we have scanned QTL at each 1 cM marker interval. We have computed LOD scores based on the classical and proposed methods for both types of data sets (non-contaminated and contaminated datasets). Figure 4.2(a) and Figure 4.2(b) are showing the LOD scores profile plots for the non-contaminated and contaminated datasets, respectively. In the LOD scores profile plots, the dot-dash black line and the solid red line represent the LOD scores at every 1cM position in the chromosomes for the classical method (FMT QTL mapping) and the proposed method with $\beta = 0.2$, respectively.

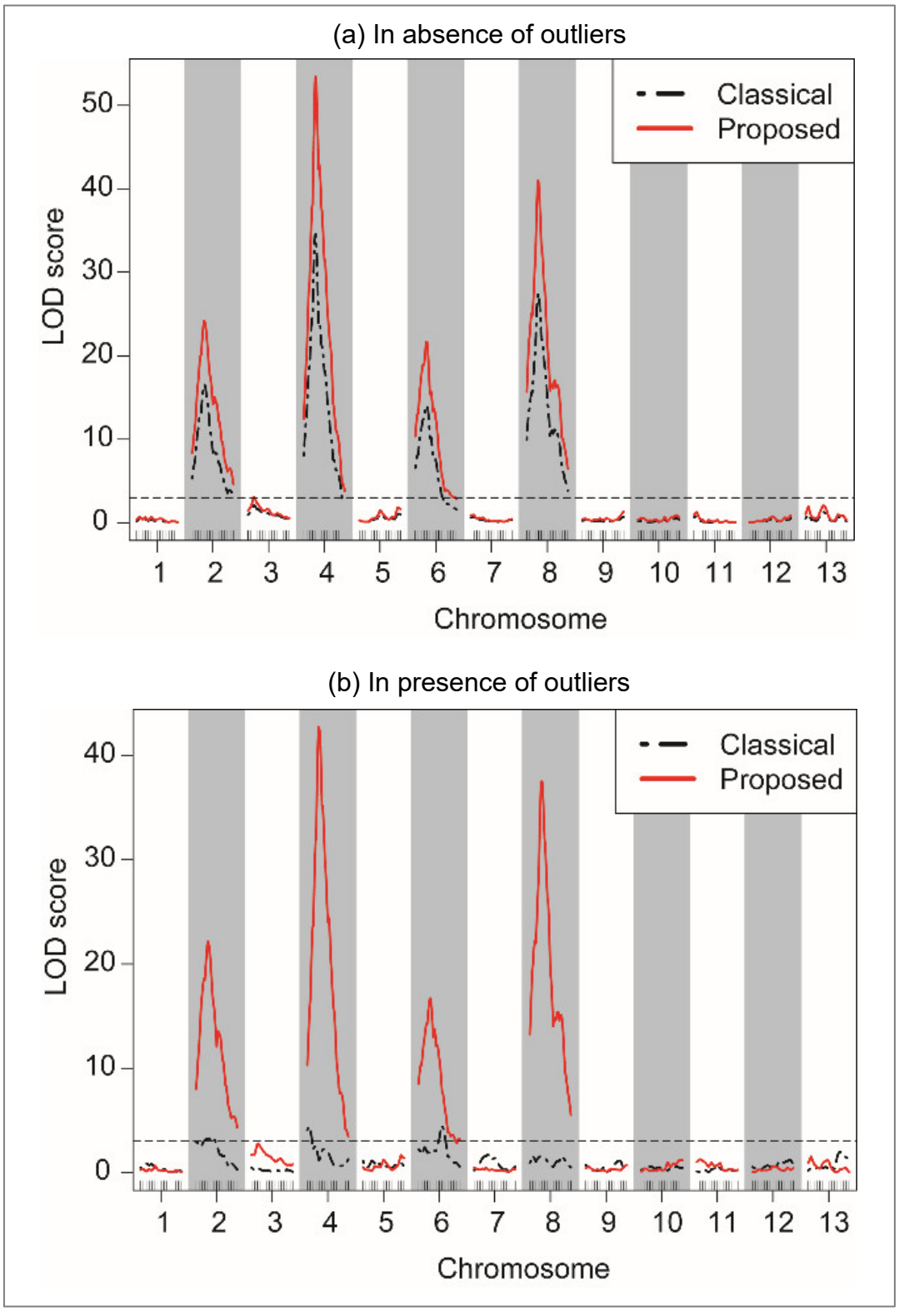


Figure 4.2: LOD score plots with simulated data in (a) absence of outliers and (b) in presence of 20% outliers in each of the phenotypes (Pheno1, Pheno2 and Pheno3).

Table 4.1 represents the significant marker name and its position identified on each of the chromosomes 2, 4, 6 and 8 by each of the methods (classical and proposed) in uncontaminated and contaminated simulated data. From Figure 4.2(a) and Table 4.1, it is seen that the highest LOD score peak occurs at the true QTL position (at marker 5 with locus position 20 cM) of the true chromosome 2, 4, 6 and 8 by the both methods for the uncontaminated dataset. From Figure 4.2(b) and Table 4.1, we observe that, in presence of outliers, the classical method fails to identify any significant QTL on chromosome 2 and 8, and identifies incorrect QTL positions on chromosome 4 (Marker interval: D4M1 - D4M2, Locus position: 3 cM) and chromosome 6 (Marker: D6M9, Locus position: 40 cM).

Table 4.1: Significant QTL positions identified by fast multi-trait (FMT) QTL mapping and Proposed method in simulated data in absence and absence of outliers

Method	**True QTL position**			**Identified QTL position**	
	Chr	**Marker**	**Position (cM)**	**In absence of outliers**	**In presence of outliers**
FMT QTL mapping (Classical)	2	5	20	Marker: D2M5 Position: 20 cM	Marker: D2M5 Position: 20 cM
	4	5	20	Marker: D4M5 Position: 20 cM	Marker: D4M5 Position: 20 cM
	6	5	20	Marker: D6M5 Position: 20 cM	Marker: D6M5 Position: 20 cM
	8	5	20	Marker: D8M5 Position: 20 cM	Marker: D8M5 Position: 20 cM
Proposed	2	5	20	Marker: D2M5 Position: 20 cM	Fails to identify any QTL
	4	5	20	Marker: D4M5 Position: 20 cM	Marker interval: (D4M1 - D4M2) Position: 3 cM
	6	5	20	Marker: D6M5 Position: 20 cM	Marker: D6M9 Position: 40 cM
	8	5	20	Marker: D8M5 Position: 20 cM	Fails to identify any QTL

FMT: Fast multi-trait.

However, in the presence of outliers, the highest LOD score peak occurs at the true QTL positions on true chromosomes by the proposed method only [Figure 4.2(b)]. So, from the simulation study, we can conclude that the proposed method outperforms

over the classical method (FMT QTL mapping) in the presence of outliers. Otherwise, the proposed method shows almost equal performance to the classical method of multi-trait QTL analysis.

4.3.2 Statistical Power of QTL Detection

To compare power of QTL detection (percentage of correct QTL identification) between the proposed and FMT QTL mapping method, we have performed simulation and analyses on 100 replicates of the simulated data. Table 4.2 represents the average along with standard deviation (SD) of the locus positions identified in 100 replications by each of the two methods (FMT QTL mapping and Proposed). From Table 4.2 we observe that in absence of outliers both the methods (FMT QTL mapping and Proposed) identify almost the same QTL positions which approximately match with the true QTL positions. However, in presence of outliers the average QTL positions identified by the FMT QTL mapping does not match with the true QTL positions whereas the average QTL positions identified by the proposed method (robust FMT QTL mapping) approximately match with the true QTL positions. This indicates that the proposed method (robust FMT QTL mapping) outperform over the classical FMT QTL mapping approach in presence of outliers.

Table 4.2: Comparison of descriptive summary of identified QTL positions identified by Fast Multi-trait (FMT) QTL mapping and Proposed method in 100 replications

QTL	Chromosome	True QTL position (cM)	Identified QTL position	
			FMT QTL mapping (Mean ± SD)	Proposed (Mean ± SD)
			In absence of outliers	
QTL1	2	20	20.13 ± 1.58	20.21 ± 1.34
QTL2	4	20	20.00 ± 0.91	20.00 ± 0.83
QTL3	6	20	20.00 ± 1.71	20.20 ± 1.80
QTL4	8	20	19.87 ± 0.72	19.86 ± 0.65
			In presence of outliers	
QTL1	2	20	26.96 ± 18.02	20.20 ± 1.52
QTL2	4	20	24.50 ± 15.04	19.87 ± 0.88
QTL3	6	20	26.68 ± 18.21	20.09 ± 2.18
QTL4	8	20	29.21 ± 20.69	19.90 ± 0.76

Table 4.3 shows the statistical power of QTL detection (percentage of correct identification of QTL positions in 100 replications) of the two methods (FMT QTL mapping and Proposed) of multi-trait QTL analysis from 100 replications of simulation and analyses. In absence of outliers, we find that the statistical powers of the FMT QTL mapping method are 74%, 85%, 77% and 89% to identify true QTLs on chromosome 2, 4, 6 and 8, respectively, whereas the Proposed method (robust FMT QTL mapping) exhibits 77%, 90%, 84% and 95% power to identify true QTLs on the same chromosomes. On the other hand, in presence of outliers the statistical powers of the FMT QTL mapping are 14%, 19%, 13% and 19% to identify true QTLs on chromosome 2, 4, 6 and 8 respectively, while the Proposed method (robust FMT QTL mapping) shows 65%, 85%, 81% and 87% statistical power to identify the true QTL positions on the same chromosomes. This means that our proposed method (robust FMT QTL mapping) shows better performance than the classical FMT QTL mapping method in presence of outliers. Otherwise, the proposed method shows similar performance to the FMT QTL mapping method.

Table 4.3: Observed statistical power (percentage of correct identification of true QTL positions in 100 replications) of the Fast multi-trait (FMT) QTL mapping and proposed method of multi-trait QTL analysis from 100 replications of simulations

QTL	Chr	True QTL position (cM)	% of correct identification in absence of outliers		% of correct identification in presence of outliers	
			FMT QTL mapping	Proposed	FMT QTL mapping	Proposed
QTL1	2	20	74	77	14	65
QTL2	4	20	85	90	19	85
QTL3	6	20	77	84	13	81
QTL4	8	20	89	95	19	87

Chr: Chromosome, FMT: Fast multi-trait, QTL: Quantitative trait locus.

4.3.3 Real Data Analysis Results

4.3.3.1 Barley Data for Multi-trait QTL Analysis

We have also investigated the performance of the proposed method (robust fast multi-trait QTL mapping) in comparison with the classical method (classical fast multi-trait

QTL mapping) using a real multi-trait QTL dataset of barley in absence and presence of phenotypic outliers.

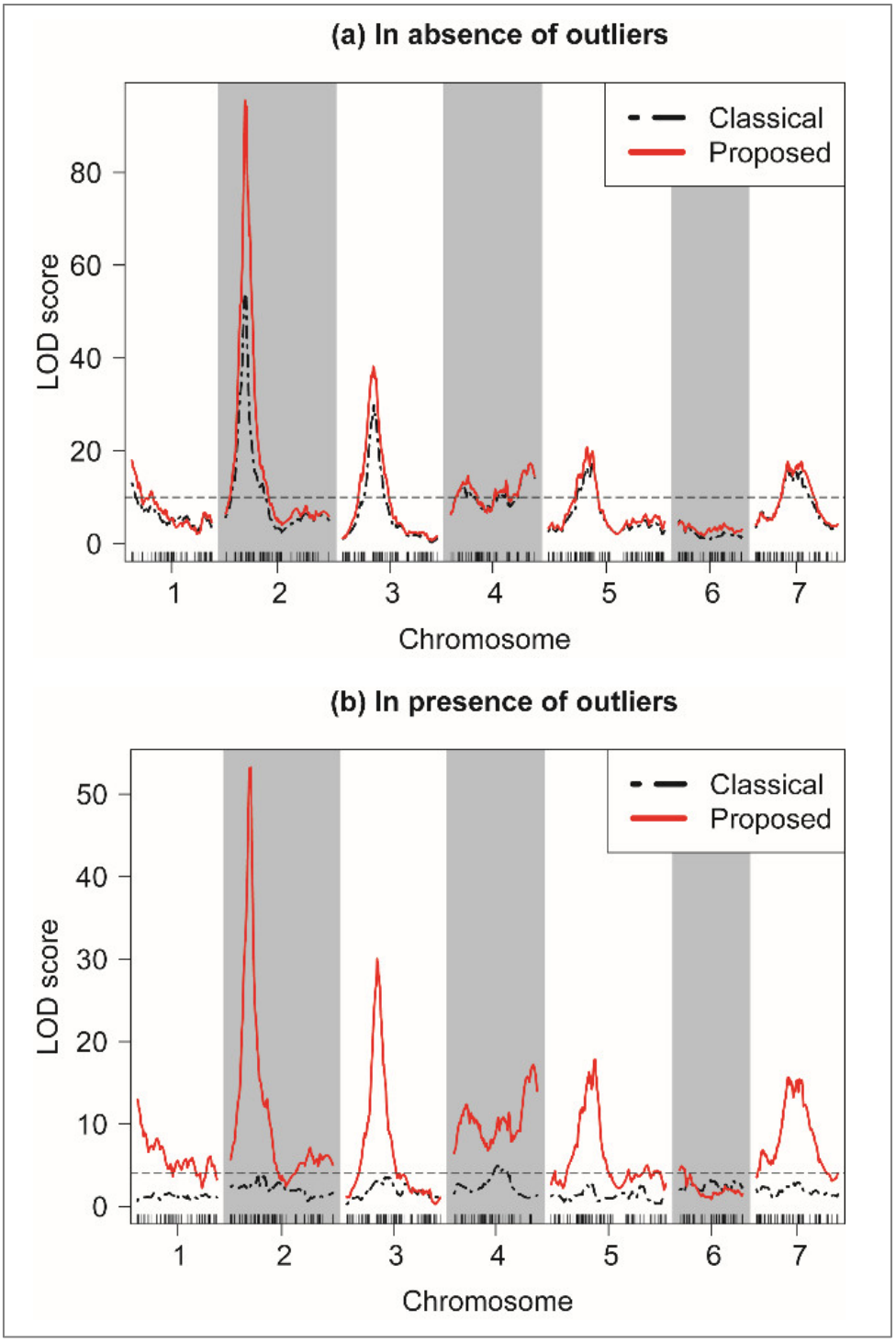


Figure 4.3: LOD score plots with barley data in (a) absence of outliers and (b) in presence of 20% outliers in each of the 8 phenotypes (grain yield, heading date, plant height, lodging, grain protein, alpha amylase, diastatic power and malt extract) considered in the study.

Figure 4.3 represents the LOD score profile plot of multi-trait QTL analysis of barley data with 8 quantitative phenotypes using the classical and proposed method in absence and presence of outliers. In the LOD scores profile plots, the dot-dash line (black color) and solid line (red color) represent the LOD scores at every 1cM position in the chromosomes for the classical method and the proposed method, respectively. Figure 4.3(a) shows the LOD scores profile of multi-trait QTL analysis with uncontaminated (absence of outliers) barley data using the classical method and proposed method with $\beta = 0.02$, respectively. We select β by cross validation. Figure 4.3(b) are representing the LOD score plots of multi-trait QTL analysis with contaminated (presence of outliers) barley data using the classical method and the proposed method with $\beta = 0.2$, respectively. Selection of the β-value was done by cross validation.

Table 4.4 shows the position at which significant maximum LOD occurs on each chromosome by each of the methods in presence and absence of outliers. From Figure 4.3(a) and Table 4.4, we observe that in absence of outliers both the method identified the same QTL positions (Chr1: Marker Hor5 at positon 0.00 cM, Chr 2: Marker Tef4 at position 36.30 cM, Chr3: Marker Dfr at position 54.40 cM, Chr4: At position 141.94 within marker interval [ksuH11 (139.00 cM) – Tel4M (148.80 cM)], Chr5: At position 78.97 cM with in marker interval [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)], Chr6: At position 3.13 within marker interval [ABG062 (2.20 cM) – snp_0669 (5.90 cM)], and Chr7: At position 57.24 within marker interval [ABC156D (53.40 cM) – snp_0050 (58.20 cM)]). However, from Figure 4.3(b) and Table 4.4, we find that in presence of outliers only the proposed method correctly identifies all the QTL positions as identified in absence of outliers. In presence of outliers, the classical method fails to identify any significant QTL position on chromosome 1, 2, 3, 5, 6 and 7, and incorrectly identifies one QTL on chromosome 4 (at position 77.5 cM with in marker interval [BCD453B (76.80 cM) – snp_0523 (78.90)]). Therefore, we can conclude that our proposed method significantly outperforms over the classical method of multi-trait QTL analysis in presence of outliers. Otherwise, the proposed method shows almost equal performance.

Table 4.4: Significant QTL positions identified by each method on each chromosome in barley data in absence and absence of outliers

Method	Chr	Identified QTL position			
		In absence of outliers		In presence of outliers	
		Marker	Position (cM)	Marker	Position (cM)
Classical*	1	Hor5	0.00	Fails to identify the QTL as identified in absence of outliers	-
	2	Tef4	36.30	Fail to identify the QTL as identified in absence of outliers	-
	3	Dfr	54.40	Fail to identify the QTL as identified in absence of outliers	
	4	Marker interval: [ksuH11 (139.00 cM) – Tel4M (148.80 cM)]	141.94	Marker interval: [BCD453B (76.80 cM) – snp_0523 (78.90)]	77.5
	5	Marker interval: [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)]	78.97	Fails to identify the QTL as identified in absence of outliers	-
	6	Marker interval: [ABG062 (2.20 cM) – snp_0669 (5.90 cM)]	3.13	Fails to identify the QTL as identified in absence of outliers	-
	7	Marker interval: [ABC156D (53.40 cM) – snp_0050 (58.20 cM)]	57.24	Fails to identify the QTL as identified in absence of outliers	-
Proposed	1	Hor5	0.00	Hor5	0.00
	2	Tef4	36.30	Tef4	36.30
	3	Dfr	54.40	Dfr	54.40
	4	Marker interval: [ksuH11 (139.00 cM) – Tel4M (148.80 cM)]	141.94	Marker interval: [ksuH11 (139.00 cM) – Tel4M (148.80 cM)]	141.94
	5	Marker interval: [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)]	78.97	Marker interval: [snp_0953 (77.10 cM) – snp_0183 (79.90 cM)]	78.97
	6	Marker interval: [ABG062 (2.20 cM) – snp_0669 (5.90 cM)]	3.13	Marker interval: [ABG062 (2.20 cM) – snp_0669 (5.90 cM)]	3.13
	7	Marker interval: [ABC156D (53.40 cM) – snp_0050 (58.20 cM)]	57.24	Marker interval: [ABC156D (53.40 cM) – snp_0050 (58.20 cM)]	57.24

* Classical method: Fast Multi-trait QTL mapping approach

4.3.3.2 BXD Mouse Data for eSNPs Analysis

We have also implemented the methods of multi-trait SIM (Classical and Proposed) with gene expression (GE) phenotype data along with SNP data of BXD mouse as an extended application of the multi-trait SIM approaches in the field of expression SNPs (eSNPs) analysis. We have considered the only the top 10 DE genes/transcripts as the multiple traits in this study. The top 10 DE transcripts has been selected based on hierarchical clustering method using the DE transcript selection method discussed in section 4.2.3. **Figure 4.4** represents the cluster dendrogram using hierarchical clustering to group the transcripts into 3 groups for selecting top 10 DE transcripts.

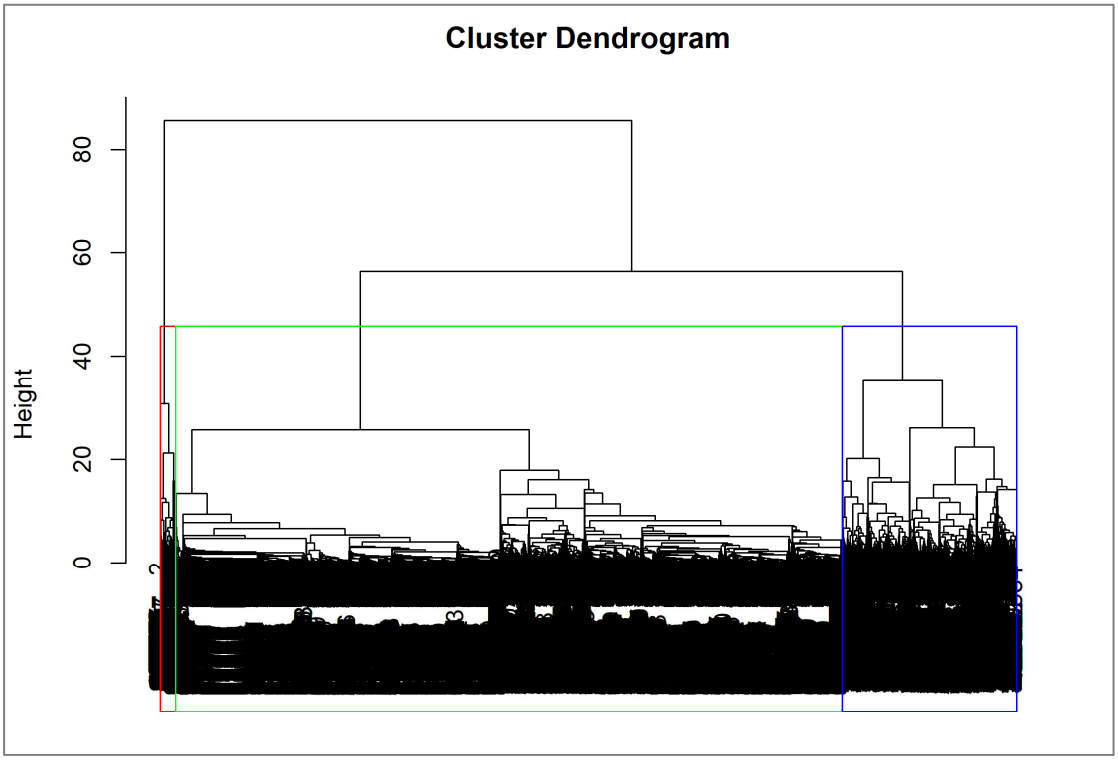


Figure 4.4: Cluster dendrogram using hierarchical clustering method to group the transcripts into 3 groups/clusters for selecting top 10 DE genes.

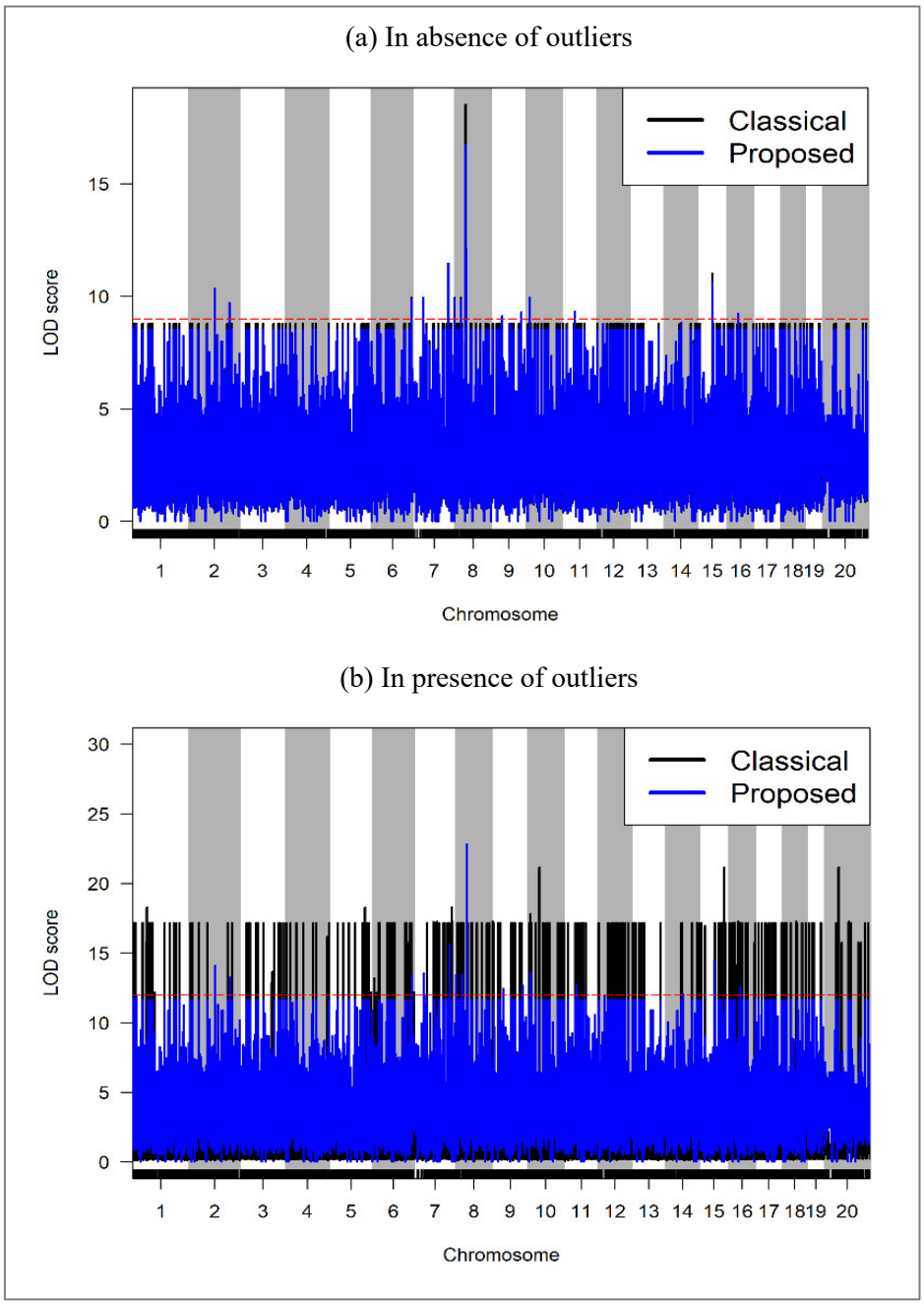


Figure 4.5: LOD score plots with BXD mouse data in (a) absence of outliers and (b) in presence of 20% outliers in each of the top 10 DE genes/transcripts considered in the study.

Figure 4.5 represents the LOD score profile plot of multi-trait eSNP analysis of BXD mouse data with top 10 DE gene expression phenotypes (transcripts) in liver using the classical and proposed method in absence and presence of outliers. In the LOD scores profile plots, the black colored and red colored lines represent the LOD scores at each SNP position in the chromosomes for the classical method and the proposed method,

respectively. Figure 4.5(a) shows the LOD scores profile of eSNPs analysis with uncontaminated (absence of outliers) BXD mouse data using the classical method and proposed method, respectively. Figure 4.5(b) are representing the LOD scores profile of eSNPs analysis with contaminated (presence of outliers) BXD mouse data using the classical method and the proposed method with $\beta = 0.2$, respectively. Selection of the β-value was done by cross validation.

From Figure 4.5(a) we observe that in absence of outliers both the method identified the same eSNP positions on chromosomes Chr2 (position: 145692973 Mb), Chr6 (position: 141509875 Mb), Chr7 (position: 125668666 Mb), Chr8 (position: 43544697 Mb), Chr9 (position: 107479346 Mb), Chr10 (position: 14709884 Mb), Chr11 (position: 47033818 Mb), Chr15 (position: 54192199 Mb) and Chr16 (position: 97748552 Mb). However, from Figure 4.5(b), we find that in presence of outliers only the proposed method correctly identifies all the eSNP positions as identified in absence of outliers. In presence of outliers, the classical method fails to identify eQTL position as identified in absence of outliers on chromosomes 2, 6, 7, 8, 9, 10, 11, 15 and 16, and it incorrectly identifies some eSNP on chromosomes 1, 5, 7, 10, 15 and 20. Thus, we can conclude that the proposed method significantly outperforms over the classical method of multi-trait eSNP analysis in presence of outliers. Otherwise, the proposed method shows almost equal performance.

4.4 Conclusion

In this paper, a new robust regression based fast interval mapping approach has been discussed for multi-trait QTL analysis by maximum β-likelihood estimation with BC population. The value of the tuning parameter β plays a key role on the performance of the proposed method. An appropriate value for the tuning parameter β can be selected by cross validation. We selected the appropriate value of the tuning parameter β using 10-fold cross validation. The proposed method with tuning parameter $\beta = 0$ reduces to the traditional multi-trait QTL interval mapping approach. Simulation and real data analysis results show that the proposed method significantly improves the performance over the classical fast multi-trait QTL mapping approach in presence of phenotypic outliers. Otherwise, the proposed method shows almost same performance as the classical fast multi-trait QTL mapping. We also applied the

proposed multi-trait QTL mapping approach to the eSNPs data analysis to find the biomarker genes and SNPs including cis and trans regulatory factors.

Chapter 5

Robustification of Regression Based GWAS to Explore Important SNPs (Proposed)

5.1 Introduction

One of the main challenges in recent genetic research is the identification of important genetic biomarkers or genes or genetic factors, which are associated with various complex traits of any living organism. Differences in traits in living organisms mainly occur due to molecular genetic variations (i.e., due to the variations in functional parts DNA which are called genes). These variations are mostly observed at the physiological, developmental, and morphological stages. Due to the recent advancements in sequencing technologies and the availability of next-generation sequence data, identification of genetic basis such as causal genetic variants for different phenotypic traits is possible at single nucleotide polymorphism (SNP) level. The statistical methods of exploring the SNP's contribution to phenotypic variation are defined as Genome-Wide Association Studies (GWAS). SNP-based GWAS has widely been used for the genetic study of a variety of species including humans, animals and plants to identify genomic locations/regions responsible for various quantitative traits. GWAS has been made possible by decreasing the cost and time required to obtain sequences of the whole genome and genome-wide SNPs. In GWAS, SNPs are commonly examined for association across the whole genome with the particular trait of interest. GWAS makes it possible to investigate the associations between a very large set of SNPs and the various complex traits of interest (Zhao et

al., 2011) such as complex diseases. The significant SNPs identified by the GWAS can be used for new drug development and prevention of specific complex diseases or complex traits.

The analysis in the GWAS includes three major steps – (i) *Quality control:* Prepressing of the raw data for genotype calls and filtering out particular samples and SNPs based on the specific criteria of quality control; (ii) *Preliminary analyses:* Calculating genotype and allele frequencies, and testing linkage disequilibrium and Hardy-Weinberg equilibrium; and (iii) *Significant SNPs identification:* Identifying the SNPs that are related to the outcome of interest performing the association analysis using SNPs and controlling the false-positive detections by identifying and adjusting population stratification (Liu et al., 2013). Population stratification (PS) is the main concerning issue when extensive genome-wide association analysis with numerous subjects is in consideration (Li and Yu, 2008; Liu et al., 2013; Xu et al., 2009). Some unidentified new population structures are probable to exist due to the large number of subjects that may be liable for systematic differences being selected in SNPs between cases and controls (Liu et al., 2013). Due to higher FDRs, it is imperative to correct the observed population stratification in GWAS (Campbell et al., 2005; Liu et al., 2013). There is however, a number of statistical approaches proposed earlier for genome-wide association mapping to address the effects of population stratification. The most commonly used statistical methods to avoid the bias of population stratification (PS) or genetic relatedness are genomic control (Devlin and Roeder, 1999), structured association (Pritchard et al., 2000), and principal component analysis (Patterson et al., 2006; Price et al., 2006). Genomic control (GC) approach modifies the association statistics by a common factor for all SNPs to correct for PS (Liu et al., 2013). Genomic control suffers from weak power when the effect of population structure is large (Aranzana et al., 2005; Devlin et al., 2001; Price et al., 2006; Yu et al., 2006; Zhao et al., 2007). Structured association analysis (SAA) technique suggests locating the samples to discrete subpopulation clusters and then collecting evidence of association within each cluster (Pritchard et al., 2000). The SAA method is useful for small datasets (Liu et al., 2013). Nevertheless, the software package STRUCTURE is computationally intensive and cumbersome for large-scale genome-wide association studies (Price et al., 2006).

Another method based on principal component (PCA) is used for genome-wide association analysis (Price et al., 2006). In this technique, EIGENSTRAT program uses several top principal components (PCs) and applies them as covariates in GWA analysis (Liu et al., 2013). These top PCs are selected using EIGENSTRAT (Price et al., 2006) program based on PCA. Thousands of markers can be analyzed using this PCA method and the adjustment using PCA is definite to a marker's variation in allele frequency across ancestral populations (Liu et al., 2013; Price et al., 2006). PCA approach may however not more appropriate to correct population structure if it arises from the existence of several discrete subpopulations because PCA applies the produced eigenvectors as continuous covariates (Liu et al., 2013).The results obtained from PCA adjustment may be misleading too if there are outliers (Liu et al., 2013). Outlying data were introduced at genotypic level to check the performance of the robust PCA approach (Liu et al., 2013).

Another improved method was proposed to deal with the fact of PS for the presence of hidden population structure for population-based GWAS (Li and Yu, 2008). This method would improve PS by combining the multi-dimensional scaling (MDS) and clustering technique. This approach was however an extension of PCA due to having some similarity matrices between PCA and MDS. It can be applied for both discrete and continuous population structures and it is well suited for large and small-scale GWA analysis (Li and Yu, 2008).

In the recent bioinformatics research, the applications of linear mixed model (LMM) techniques have been popular in different genome-wide linkage analysis for discovery of potential biomarkers from human and agricultural single nucleotide polymorphism (SNP) level data. Nowadays to address the issues of adjustment of population stratification and account for population structure and genetic relatedness (polygenic effects) are effectively overcome by implementing LMM (Endelman, 2011; Kang et al., 2010; Zhang et al., 2010) for large scale GWAS. These approaches have been executed in software programs TASSEL (Bradbury et al., 2007), EMMA (Kang et al., 2008), EMMAX (Kang et al., 2010), rrBLUP (Endelman, 2011), Genome-wide efficient mixed-model analysis (GEMMA) (Zhou and Stephens, 2012), GAPIT (Lipka et al., 2012).

All the existing methods for GWAS, mentioned above, ae very sensitive to phenotypic outliers and produce misleading results in identifying the important SNPs when the phenotypic data are contaminated by outliers. In this study, we have proposed a robust statistical approach for SNP-based GWAS by robustifying the linear regression based GWAS using minimum β–divergence technique (Mihoko and Eguchi, 2002; Mollah et al., 2007). The performance of the proposed approach has been investigated using both simulated data and real data (SNP data of "grain number per panicle" of rice cultivated in Hangzhou area, China) in terms of power and false discovery rate (FDR) in presence of phenotypic outliers.

5.2 Materials and Methods

5.2.1 Classical Methods of SNP Analysis

Let us consider *m* SNPs with *n* individuals and there are n observations for a phenotype. Then the linear model for GWAS can be written as

$$y_j = \alpha + \gamma x_j + \varepsilon_j,\ \ j = 1,\ 2, \ldots\ldots\ldots, n \tag{5.1}$$

where y_j is the phenotype for the j-th individual, α is the overall mean, γ is the SNP effect and x_j is the SNP value for the j-the individual.

The least square estimates of the parameter of above model is

$$\hat{\gamma} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}} \tag{5.2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\gamma}\bar{x} \tag{5.3}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$,

$$SS_{xy} = \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) \tag{5.4}$$

and $SS_{xx} = \sum_{j=1}^{n}(x_j - \bar{x})^2$ (5.5)

Then the estimated classical model can be written as

$$\hat{y}_j = \hat{\alpha} + \hat{\gamma} x_j \tag{5.6}$$

We want to test the null hypothesis $H_0: \gamma = 0$ (i.e., there is no SNP effect) against $H_1: H_0$ is not true. The test statistic under the null hypothesis $H_0: \gamma = 0$ (i.e., there is no SNP effect) is as follows:

$$t = \frac{\hat{\gamma}}{\mathrm{SE}(\hat{\gamma})} = \frac{\hat{\gamma}}{\sqrt{\frac{MSE}{SS_{xx}}}} \tag{5.7}$$

where SE = Standard error and

$$\text{Mean squared error, } MSE = \frac{1}{(n-k)} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \tag{5.8}$$

The t-statistic in (5.7) follows t-distribution with $(n-k) = (n-2)$ degrees of freedom.

5.2.2 Robust Method of SNP Analysis

One major problem of classical linear regression model based GWAS is that the estimates of regression parameters in (5.2) – (5.3) and the t-statistic defined in (5.7) are very sensitive to outliers and provide misleading results when the data is contaminated with outliers. So, we need a robust approach for SNP analysis to obtain the robust estimates of the model parameters and the robust test statistic which will be less sensitive to outliers. In this section, we have discussed the robustification of the classical linear regression model based GWAS using minimum β–divergence method (Mihoko and Eguchi, 2002; Mollah et al., 2007) to obtain the robust estimates of model parameters and the robust test statistic (i.e., robust t-statistic).

Equation (5.4) and (5.5) can be written, respectively, as

$$SS_{xy} = nS_{xy} \tag{5.9}$$

$$SS_{xx} = nS_x^2 \tag{5.10}$$

where $S_{xy} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$,

$S_x^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$ and

Then the equation (5.2) can be written as

$$\hat{\gamma} = \frac{SS_{xy}}{SS_{xx}} = \frac{nS_{xy}}{nS_x^2} = \frac{S_{xy}}{S_x^2} \tag{5.11}$$

Equation (5.8) can be written as

$$MSE = \frac{1}{(n-k)}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = \frac{SSE}{(n-k)} = \frac{SST - SSR}{(n-k)} \tag{5.12}$$

where

Sum of squares total, $SST = SS_{yy} = \sum_{j=1}^{n}(y_j - \bar{y})^2 = nS_y^2$ (5.13)

$$S_y^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2 \tag{5.14}$$

and

Sum of squares regression, $SSR = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2$

$$= \frac{(SS_{xy})^2}{SS_{xx}} = \hat{\gamma}SS_{xy} \tag{5.15}$$

Using (5.13) and (5.15), equation (5.12) can be written as

$$MSE = \frac{SS_{yy} - \hat{\gamma}SS_{xy}}{(n-k)} = \frac{nS_y^2 - \hat{\gamma}nS_{xy}}{(n-k)} = \frac{n(S_y^2 - \hat{\gamma}S_{xy})}{(n-k)} \tag{5.16}$$

Using (5.10) and (5.16), equation (5.7) can be written as

$$t = \frac{\hat{\gamma}}{\sqrt{\frac{n(S_y^2 - \hat{\gamma}S_{xy})}{(n-k)nS_x^2}}} = \frac{\hat{\gamma}}{\sqrt{\frac{(S_y^2 - \hat{\gamma}S_{xy})}{(n-k)S_x^2}}} \tag{5.17}$$

which follows t-distribution with $(n-k)=(n-2)$ degrees of freedom.

From equations (5.3), (5.11) and (5.17) we observe that the classical estimates of regression parameters ($\hat{\alpha}$ and $\hat{\gamma}$) and the t-statistic depend only the sample means ($\bar{x}$ and $\bar{y}$), sample variances (S_x^2 and S_y^2) and sample covariance (S_{xy}). Hence, we can get the robust estimates of the regression parameters (α and γ) and derive robust t-statistic if we can robustify the sample means ($\bar{x}$ and $\bar{y}$), sample variances (S_x^2 and S_y^2) and sample covariance (S_{xy}).

Let us consider that $\boldsymbol{Z}=(Y,X)$ with population mean vector $\boldsymbol{\mu}_Z=\begin{bmatrix}\mu_Y\\ \mu_X\end{bmatrix}$ and population variance-covariance matrix $\boldsymbol{\Sigma}_Z=\begin{bmatrix}\sigma_Y^2 & \sigma_{YX}\\ \sigma_{XY} & \sigma_X^2\end{bmatrix}$, where Y and X have been introduced in (5.1).

The minimum β-divergence estimators of the parameters $\boldsymbol{\theta}=(\boldsymbol{\mu_Z},\boldsymbol{\Sigma_Z})$ can be obtained by the iterative solution of the following equations:

$$\boldsymbol{\mu}_{\boldsymbol{Z},\,t+1}=\frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)\boldsymbol{z}_j}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)} \tag{5.18}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{Z},\,t+1}=(1+\beta)\frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)(\boldsymbol{z}_j-\boldsymbol{\mu}_{\boldsymbol{Z},\,t})(\boldsymbol{z}_j-\boldsymbol{\mu}_{\boldsymbol{Z},\,t})^T}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)}, \tag{5.19}$$

where $w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t), j=1,2,\dots,n$, is called the β-weight function and defined as

$$w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)=\exp\left[-\frac{\beta}{2}(\boldsymbol{z}_j-\boldsymbol{\mu}_{\boldsymbol{Z},\,t})^T\boldsymbol{\Sigma}_{\boldsymbol{Z},\,t}^{-1}(\boldsymbol{z}_j-\boldsymbol{\mu}_{\boldsymbol{Z},\,t})\right] \tag{5.20}$$

The value of β-weight function ranges from 0 to 1. The tuning parameter β plays an important role to control the performance of the proposed method. The appropriate value of β can be selected by k-fold cross validation. If β = 0, then (5.18) and (5.19)

reduces to the classical non-iterative solution and the estimates reduce to classical estimates.

Let the robust estimate (i.e., β-estimate) of $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ are denote by $\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)}$. Then we can write

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \hat{\mu}_{Y(\beta)} \\ \hat{\mu}_{X(\beta)} \end{bmatrix} = \begin{bmatrix} \bar{y}_{(\beta)} \\ \bar{x}_{(\beta)} \end{bmatrix} \tag{5.21}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \hat{\sigma}^2_{Y(\beta)} & \hat{\sigma}_{YX(\beta)} \\ \hat{\sigma}_{XY(\beta)} & \hat{\sigma}^2_{X(\beta)} \end{bmatrix} = \begin{bmatrix} S^2_{y(\beta)} & S_{yx(\beta)} \\ S_{xy(\beta)} & S^2_{x(\beta)} \end{bmatrix} \tag{5.22}$$

Then, using (5.21) and (5.22) in (5.3) and (5.11), the robust estimates of the regression parameters (α and γ) can be written as

$$\hat{\alpha}_{(\beta)} = \bar{y}_{(\beta)} - \hat{\gamma}_{(\beta)}\bar{x}_{(\beta)} \tag{5.23}$$

$$\hat{\gamma}_{(\beta)} = \frac{S_{xy(\beta)}}{S^2_{x(\beta)}} \tag{5.24}$$

Using (5.22) and (5.24) in (5.17), the robust t-statistic can be written as

$$t_{(\beta)} = \frac{\hat{\gamma}_{(\beta)}}{\sqrt{\frac{(S^2_{y(\beta)} - \hat{\gamma}_{(\beta)}S_{xy(\beta)})}{(n-k)S^2_{x(\beta)}}}} \tag{5.25}$$

which follows t-distribution with $(n-k) = (n-2)$ degrees of freedom.

5.2.3 Simulation Study

To evaluate the performance of the proposed method with the classical method of SNP based GWAS, we have generated synthetic/artificial genotype and phenotype data using the simulation technique based on the linear model described in (5.1). We have generated synthetic data considering one phenotype, 300 individuals, 1000 SNPs on a single chromosome with different values of heritability, h^2 = (0.1, 0.2, 0.3, 0.4,

0.5, 0.6, 0.7, 0.8, 0.9). Five true SNP positions were considered at positions 100, 200, 300, 400 and 500 on a single chromosome. We have used the R package rrBLUP to generate the synthetic/artificial genotype and phenotype data. A part of the simulated phenotype and genotype data has been shown in Figure 5.1. To check the robustness of the proposed method in comparison of the classical method, we have contaminated 1% to 10% of the phenotypic values by outliers in the dataset. We have created the Manhattan plot by plotting SNP positions in X-axis and [$-\log_{10}$(*P*-value)] values in Y-axis. We have computed the threshold value of level of significance using Bonferroni correction (Bonferroni et al., 1936) $\alpha_{\mathrm{BC}} = \frac{\alpha}{\text{Number of SNPs}}$ to identify statistically significant SNPs. So the threshold value for Manhattan plot is $[-\log_{10}(\alpha_{\mathrm{BC}})]$.

We have investigated the performance of the proposed method in comparison with the classical method of SNP based GWAS in terms of statistical power of SNP detection and false discovery rate (FDR). The statistical power of SNPs detection for a method is defined as $\text{Power} = \frac{\text{Number of correctly detected SNPs}}{\text{Total number of true SNPs}}$ and the FDR of a method is defined as $\text{FDR} = \frac{\text{Number of incorrectly detected SNPs}}{\text{Total number of detected SNPs}}$. To compare the power and FDR of the proposed method with the classical over the change of heritability, we have replicated the whole simulation process 1000 times with both the proposed and classical methods for each of the heritability values (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) and calculated the statistical power and false discovery rate (FDR) for each method in absence and presence of 10% outliers. To investigate the power and FDR of the methods (Proposed and Classical) over the change in the % of phenotypic contamination, we have replicated the whole simulation process 1000 times with both the methods (Proposed and Classical) for each of (1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%) contamination at heritability h^2 = 0.5 and calculated the statistical power and false discovery rate (FDR) for each method in absence and presence of outliers using simulation study.

Phenotype Data

line	pheno
1	1.426759818
2	0.981974604
3	-0.771560259
4	-5.650618193
5	1.163927606
6	8.50979028
7	-1.222301064
8	-3.372772776
9	-2.932836648
10	6.058803008
11	2.651224419
12	-3.10292152
13	-0.22339757
14	2.018986607
15	2.672440779
16	0.109511447
17	2.931734942
18	3.272849862
19	3.038239759
20	-1.260983334
21	-1.329489624
22	2.751881059
23	4.233921139
24	5.996717484
25	5.997699933
26	-1.422854488

Genotype Data — **Line/Individual**

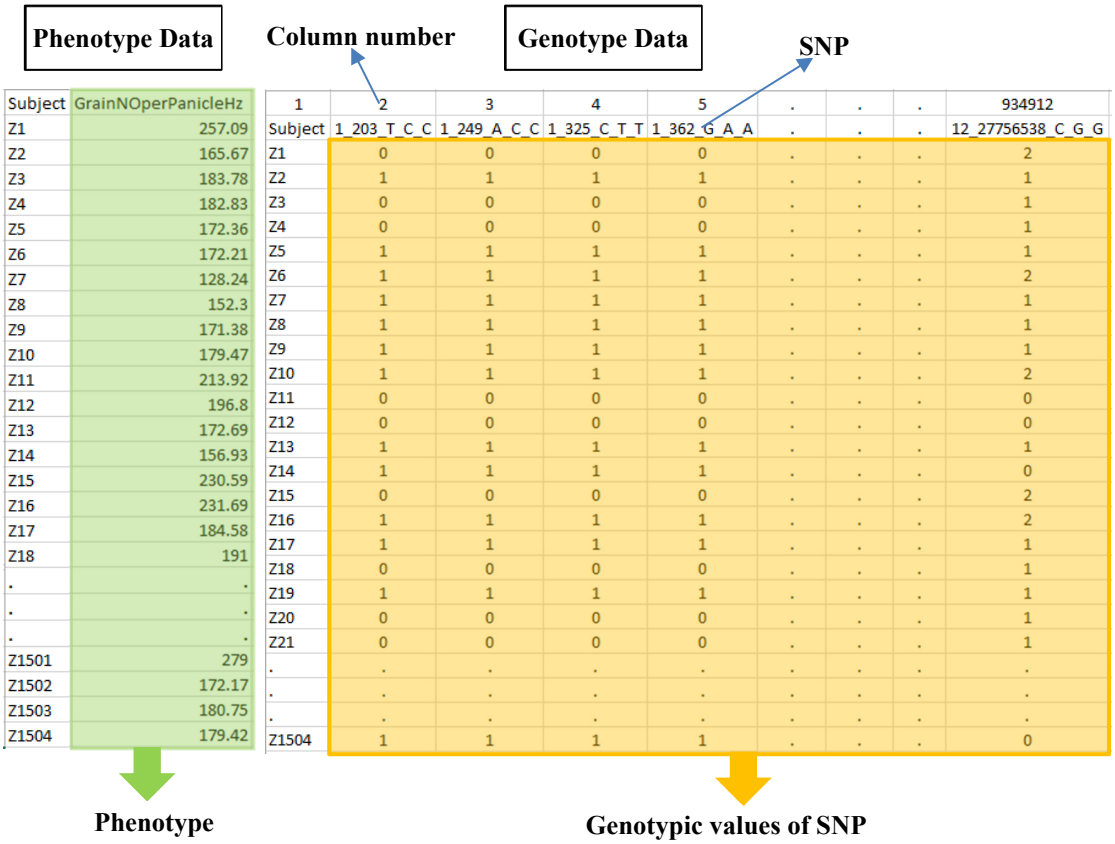
marker	chromosome	position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	1	1	1	-1	1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	-1	-1	1	1	-1	1	1
2	1	2	1	1	1	1	1	1	1	-1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	-1	1	-1	-1	1	1	-1
3	1	3	1	1	-1	-1	1	-1	1	1	-1	1	1	-1	1	1	1	1	-1	-1	-1	-1	-1	1	-1	-1	1	1
4	1	4	1	-1	-1	-1	-1	1	-1	-1	-1	1	-1	1	-1	-1	1	1	-1	-1	-1	1	1	1	1	-1	1	-1
5	1	5	-1	1	1	1	1	1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	-1	-1	1	-1	1	-1
6	1	6	1	-1	-1	1	-1	-1	1	-1	1	1	-1	1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	-1	1
7	1	7	-1	1	1	-1	1	-1	-1	1	1	1	1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	-1
8	1	8	-1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	-1	1	1	1	1	1	-1
9	1	9	1	1	-1	-1	-1	1	1	1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	1	1
10	1	10	1	1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	-1	1	1	1	1	-1	-1
11	1	11	-1	1	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	1	1	1	-1	-1	-1
12	1	12	-1	-1	1	-1	-1	-1	-1	1	-1	1	-1	-1	1	1	1	-1	1	-1	1	-1	-1	1	1	-1	1	1
13	1	13	1	1	-1	-1	1	-1	-1	1	1	1	-1	-1	1	1	-1	-1	-1	1	-1	-1	1	1	1	1	1	-1
14	1	14	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	1	1	1	1	-1	1	1	1	1	1	-1	-1	-1	1
15	1	15	-1	1	1	-1	1	1	-1	1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	1	1
16	1	16	-1	1	-1	1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	-1	-1	1	1	1	-1	1	-1	-1	1
17	1	17	1	1	1	-1	1	1	-1	1	-1	1	1	1	-1	-1	1	1	1	1	1	-1	1	1	1	-1	1	1
18	1	18	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	-1	1	1	1	1	1	-1	1	-1	1	1	-1	-1	-1
19	1	19	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	-1	-1	1	1	1	1	1
20	1	20	-1	1	1	-1	-1	-1	-1	-1	1	-1	1	1	1	1	1	1	1	-1	1	-1	1	-1	1	1	1	1
21	1	21	-1	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1	1	-1	-1	1	-1	1	-1	-1	1	1	-1	1
22	1	22	1	1	1	1	1	1	1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1	-1	-1	1	-1	1	-1	-1
23	1	23	-1	-1	1	1	1	-1	1	-1	1	1	-1	1	1	-1	1	1	-1	1	1	-1	-1	-1	-1	-1	1	1
24	1	24	-1	1	1	-1	1	-1	-1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	1	1	-1	-1	-1	-1	1
25	1	25	1	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	1	1	1	-1	1	-1	-1	1	1	1	-1	-1
26	1	26	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1	1	-1	-1	-1	-1

Figure 5.1: Structure of the simulated Phenotype and Genotype data files used in SNP based GWAS.

5.2.4 Real Data Analysis

We have also validated the performance of the proposed method in comparison with the classical method using a real dataset of rice (*Oryza sativa* L.). The real dataset of rice used in this investigation are obtained from Huang et al. (2015). The dataset contains total 1,495 different varieties of hybrid rice among which 1,439 varieties are from *indica-indica* hybrid crosses (called *indica* hybrid varieties), 38 varieties are from *japonica-japonica* hybrid crosses and 18 varieties are from *indica-japonica* hybrid crosses. For phenotyping, all the varieties of hybrid rice were planted together at two agro-ecologically different locations: Sanya and Hangzhou in China. The phenotypic dataset contains total 12 quantitative traits (i.e., phenotypes) for each hybrid variety: Yield per plant (g), Panicle number, Grain number per panicle, Seed setting rate, Grain weight (g), Heading date (days), Height (cm), Flag leaf length (cm), Flag leaf width (cm), Panicle length (cm), Grain length (mm) and Grain width

(mm). The genotype dataset (i.e., SNP dataset) contains total 934,912 SNPs on 12 chromosomes (Chr 1: 114,948 SNPs, Chr2: 98,502 SNPs, Chr3: 83,352 SNPs, Chr4: 74,265 SNPs, Chr5: 70,404 SNPs, Chr6: 72,144 SNPs, Chr7: 76,618 SNPs, Chr8: 56,123 SNPs, Chr9: 64,298 SNPs, Chr10: 61,832 SNPs, Chr11: 91,095 SNPs, Chr12: 71,331 SNPs) for each of the 1495 varieties of rice. For the convenience of our analysis, we have considered only the 1,439 *indica* hybrids along with only one phenotypic trait **"Grain number per panicle"** for Hangzhou area. In order to measure the performance of the proposed method in comparison with the classical method in presence of outliers we have contaminated 10% of the values of the phenotypic trait (grain number per panicle).



Subject	GrainNOperPanicleHz
Z1	257.09
Z2	165.67
Z3	183.78
Z4	182.83
Z5	172.36
Z6	172.21
Z7	128.24
Z8	152.3
Z9	171.38
Z10	179.47
Z11	213.92
Z12	196.8
Z13	172.69
Z14	156.93
Z15	230.59
Z16	231.69
Z17	184.58
Z18	191
.	.
.	.
.	.
Z1501	279
Z1502	172.17
Z1503	180.75
Z1504	179.42

1	2	3	4	5	.	.	.	934912
Subject	1_203_T_C_C	1_249_A_C_C	1_325_C_T_T	1_362_G_A_A	.	.	.	12_27756538_C_G_G
Z1	0	0	0	0	.	.	.	2
Z2	1	1	1	1	.	.	.	1
Z3	0	0	0	0	.	.	.	1
Z4	0	0	0	0	.	.	.	1
Z5	1	1	1	1	.	.	.	1
Z6	1	1	1	1	.	.	.	2
Z7	1	1	1	1	.	.	.	1
Z8	1	1	1	1	.	.	.	1
Z9	1	1	1	1	.	.	.	1
Z10	1	1	1	1	.	.	.	2
Z11	0	0	0	0	.	.	.	0
Z12	0	0	0	0	.	.	.	0
Z13	1	1	1	1	.	.	.	1
Z14	1	1	1	1	.	.	.	0
Z15	0	0	0	0	.	.	.	2
Z16	1	1	1	1	.	.	.	2
Z17	1	1	1	1	.	.	.	1
Z18	0	0	0	0	.	.	.	1
Z19	1	1	1	1	.	.	.	1
Z20	0	0	0	0	.	.	.	1
Z21	0	0	0	0	.	.	.	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Z1504	1	1	1	1	.	.	.	0

Figure 5.2: Structure of real Phenotype (Grain number per panicle) and Genotype data files of rice in Hangzhou area used in this SNP based GWAS.

5.3 Results and Discussion

5.3.1 Simulated Data Analysis Results

To compare the performance of the proposed method with the classical method of SNP based GWAS, we have implemented both of the two methods (Classical and Proposed) with the simulated dataset and evaluate the performance of these methods for true SNP detection.

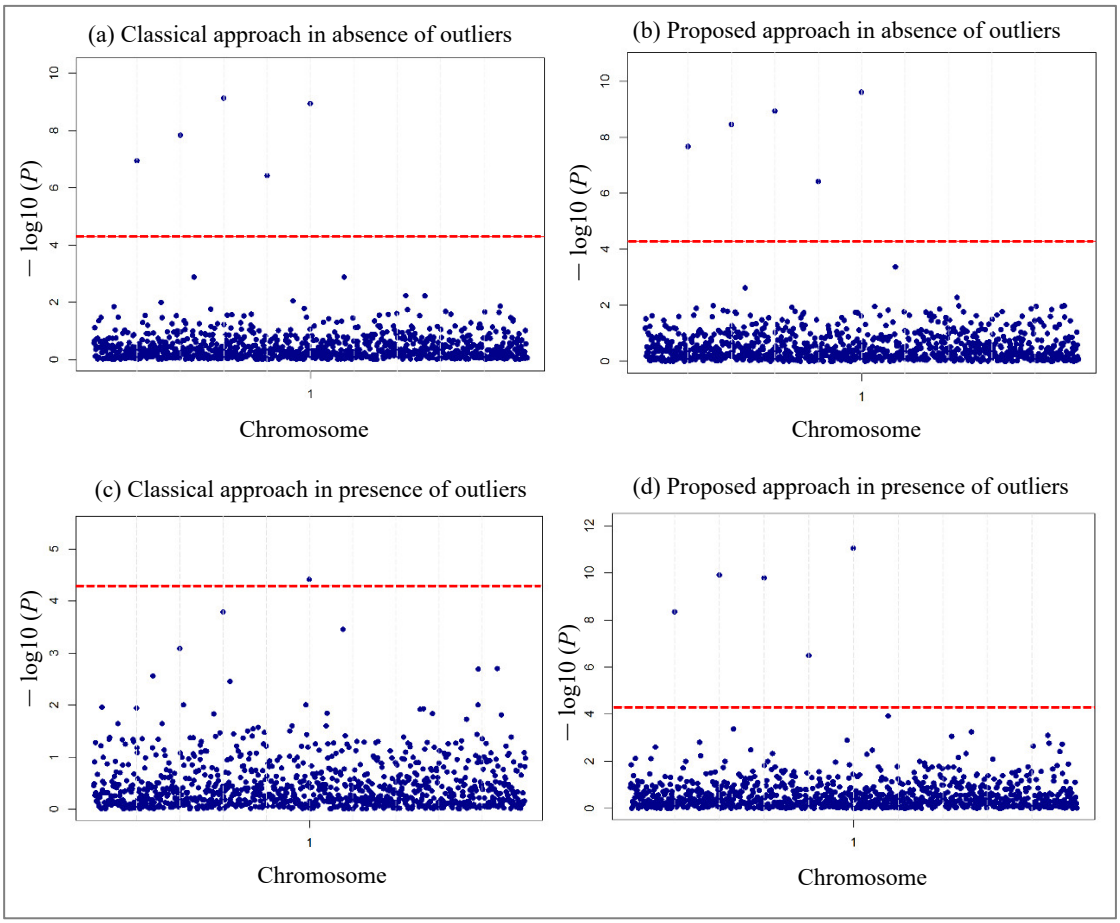


Figure 5.3: Manhattan plot of SNP based GWAS with simulated data using classical approach and proposed approach in absence and presence of outliers. (a) Classical approach in absence of outliers, (b) Proposed approach in absence of outliers, (c) Classical approach in presence of outliers and (d) Proposed approach in presence of outliers.

Figure 5.3 shows the Manhattan plot of SNP based GWAS with simulated data for the classical method and proposed method in absence and presence of 10% outliers.

Manhattan plot represents SNP positions in X-axis and [$-\log_{10}(P)$] values in Y-axis. We created the Manhattan plot for 1000 SNPs with threshold value $-\log_{10}(0.05/1000) = 4.30$ obtained using Bonferroni correction at 5% level of significance. Figure 5.3(a) and Figure 5.3(b) represent the Manhattan plot of GWAS for classical and proposed method, respectively, in absence of outliers. We observe that both the methods identify all the true SNP positions correctly at position 100, 200, 300, 400 and 500. This indicates that the proposed method shows similar performance to the classical method in absence of outliers. Figure 5.3(c) and Figure 5.3(d) show the Manhattan plot of GWAS for classical and proposed method, respectively, in presence of outliers. We find that the classical method fails to identify all the 5 true SNP positions correctly and it identifies only one true SNP position correctly at position 500, whereas the proposed method identifies all the true SNP positions correctly at position 100, 200, 300, 400 and 500. So, our proposed method performs better than the classical method in presence of outliers. Otherwise, it shows almost equal performance.

To compare the performance of the proposed method with the classical method of SNP based GWAS, we have calculated the statistical power and false discovery rate (FDR) for each method in absence and presence of outliers using simulation study. Figure 5.4 shows the prediction power and false discovery rate (FDR) of classical and proposed methods over the change in heritability (h^2) in absence and presence of outliers using simulated study. Figure 5.4(a) and Figure 5.4(b) represent the power of SNP detection and FDR of both methods (Classical and Proposed), respectively, in absence of outliers. We observe that both the methods of SNP based GWAS (Classical and Proposed) have similar power and FDR in absence of outliers. The prediction power increases and FDR decreases with the increase in heritability (h^2) value in absence of outliers. Our findings indicates that both the methods (Classical and Proposed) show almost equal performance in absence of outliers.

Figure 5.4(c) and Figure 5.4(d), respectively, show the prediction power and FDR of both methods in presence of phenotypic outliers. We find that the proposed method exhibits the higher power of correct SNP detection than the classical method in presence of outliers. However, both the methods have almost same FDR in presence

of outliers. The prediction power increases and FDR decreases with the increase in heritability (h^2) value in presence of outliers. These results indicate that the proposed method outperform over the classical method of SNP based GWAS in presence of phenotypic outliers.

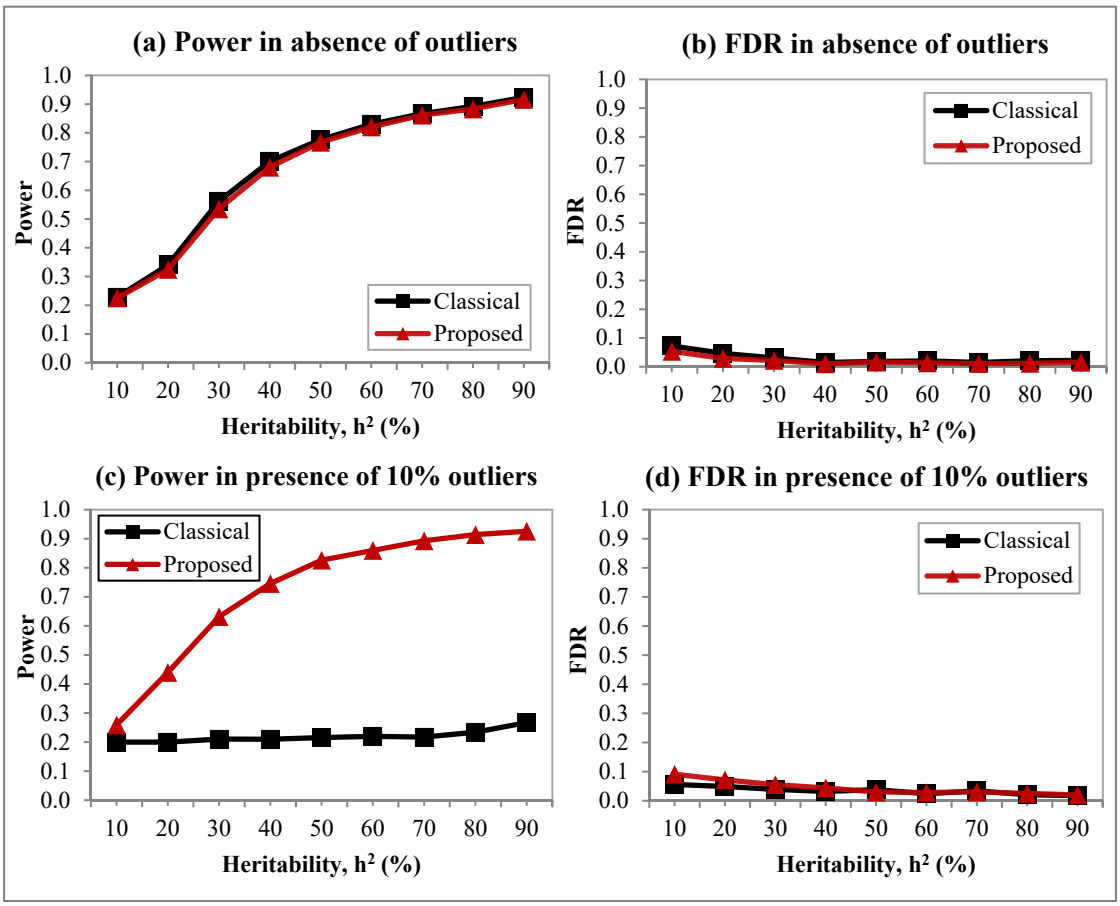


Figure 5.4: Prediction power and false discovery rate (FDR) of classical and proposed methods over the change in heritability (h^2) in absence and presence of outliers using simulated dataset. (a) Prediction power over the change in heritability (h^2) in absence of outliers, (b) FDR over the change in heritability (h^2) in absence of outliers, (c) Prediction power over the change in heritability (h^2) in presence of 10% outliers, and (d) FDR over the change in heritability (h^2) in presence of 10% outliers.

Figure 5.5 represents the statistical power and FDR of classical and proposed methods for different % of phenotypic contaminations (i.e., outliers) with heritability value h^2=0.5. We observe that the power of the proposed method is always higher than that of the classical method. The SNP detection power of the proposed method remain

almost stable with the increase in the % of phenotypic contaminations whereas the power of the classical method decreases with the increase in the % of phenotypic contaminations. The FDR of both the methods is almost similar and it increases very slowly with the increase in the % of phenotypic contaminations. These finding indicate that our proposed method shows better performance than the classical method of SNP based GWAS.

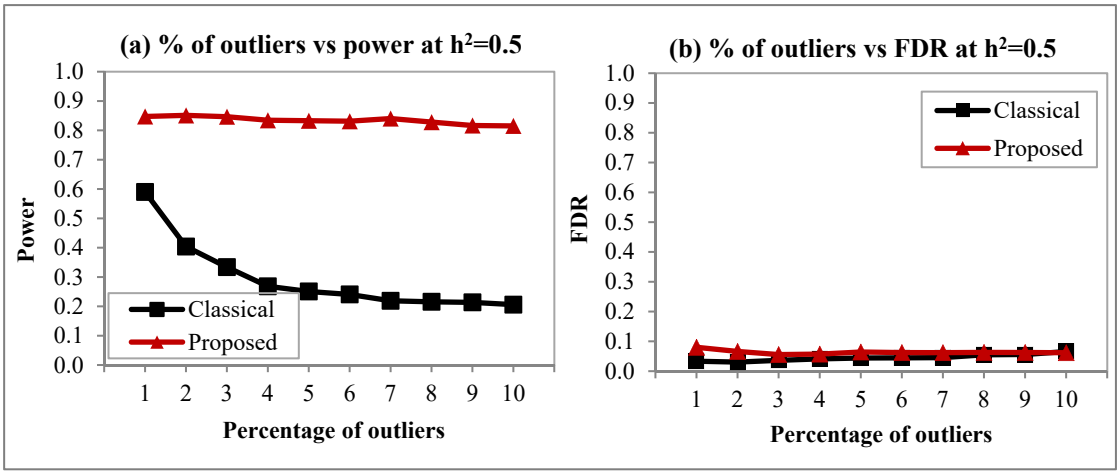


Figure 5.5: Prediction power and false discovery rate (FDR) of classical and proposed methods over the change in percentage (%) of phenotypic contaminations (outliers) using simulated dataset. (a) Percentage of outliers versus prediction power and (b) percentage of outliers versus false discovery rate (FDR).

5.3.2 Real Data Analysis Results

We have evaluated the performance of the proposed method in a comparison with the classical method using a real dataset of 1,439 indica varieties of rice (*Oryza sativa* L.) containing 934,912 SNPs and 12 phenotypes. We have considered the yield-related phenotype "Grain number per panicle" in Hangzhou area as our phenotype of interest. We have performed GWAS with the SNP dataset to identify the significant SNPs controlling the phenotype "Grain number per panicle" using both classical and proposed methods. To investigate the performance of the proposed method in comparison with the classical method in presence of outliers we have contaminated 10% of the values of "Grain number per panicle".

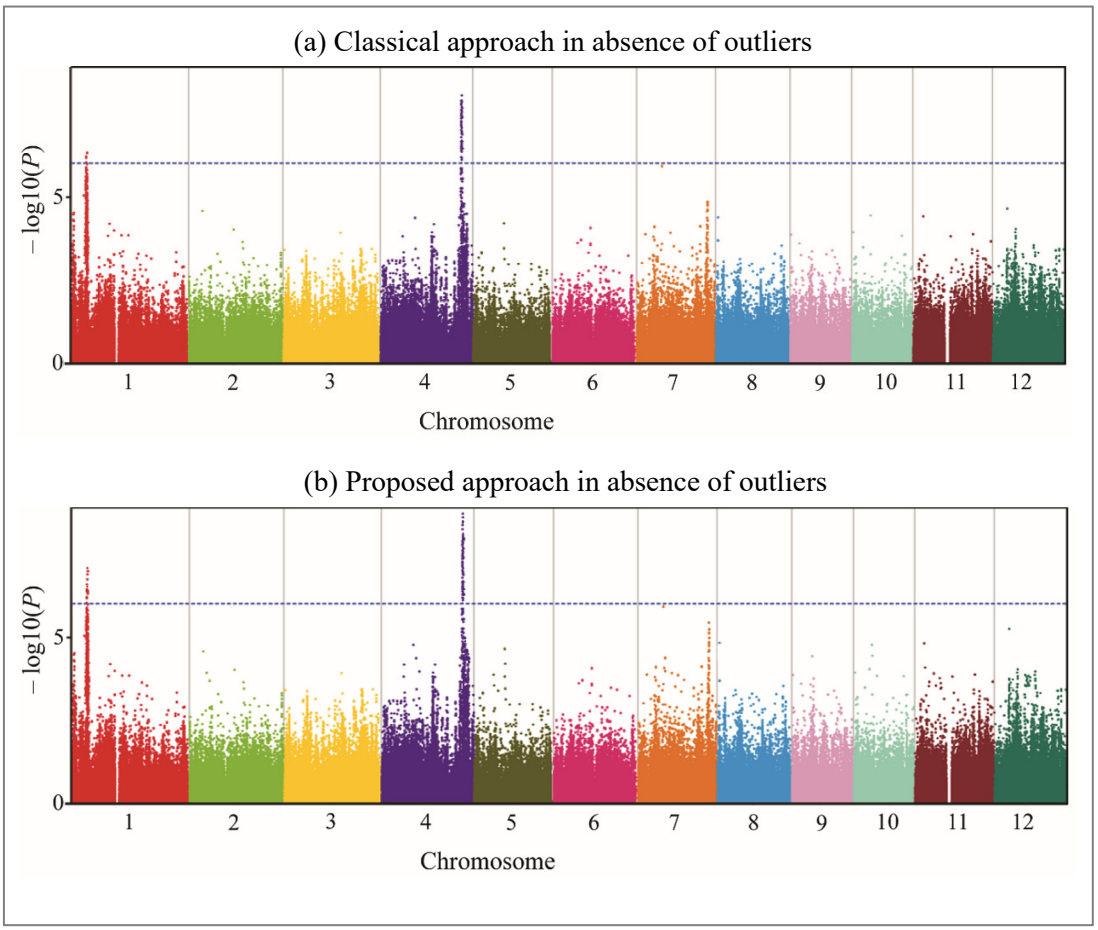


Figure 5.6: Manhattan plot of GWAS to identify important SNPs/QTLs which control the "gain number per panicle" in Hangzhou area in absence of outliers. Manhattan plot has been created plotting $[-\log_{10}(P)]$ values from the linear model in Y-axis and all the SNP positions in X-axis for each of the 12 chromosomes of rice. The horizontal dotted line represent the threshold *P*-value of 10^{-6} to identify the genome-wide significant SNPs using classical and proposed method in absence of phenotypic outliers. (a) Classical approach in absence of outliers and (b) Proposed approach in absence of outliers.

Figure 5.6 represents the Manhattan plot of SNP base GWAS of grain number per panicle using classical and proposed method in absence of outliers. From Figure 5.6(a) we find that classical method identifies significant SNPs on chromosome 1 at locus position 6,346,698 (Table 5.1) and on chromosome 4 locus position 31,493,318 (Table 5.1). From Figure 5.6(b) we observe that the proposed method detects significant SNPs on 1 at locus position 6,346,698 (Table 5.1) and on chromosome 4

locus position 31,493,318 (Table 5.1). So, both the methods identify the same SNP potions. This indicates that the proposed method shows similar performance to the classical method in absence of phenotypic outliers.

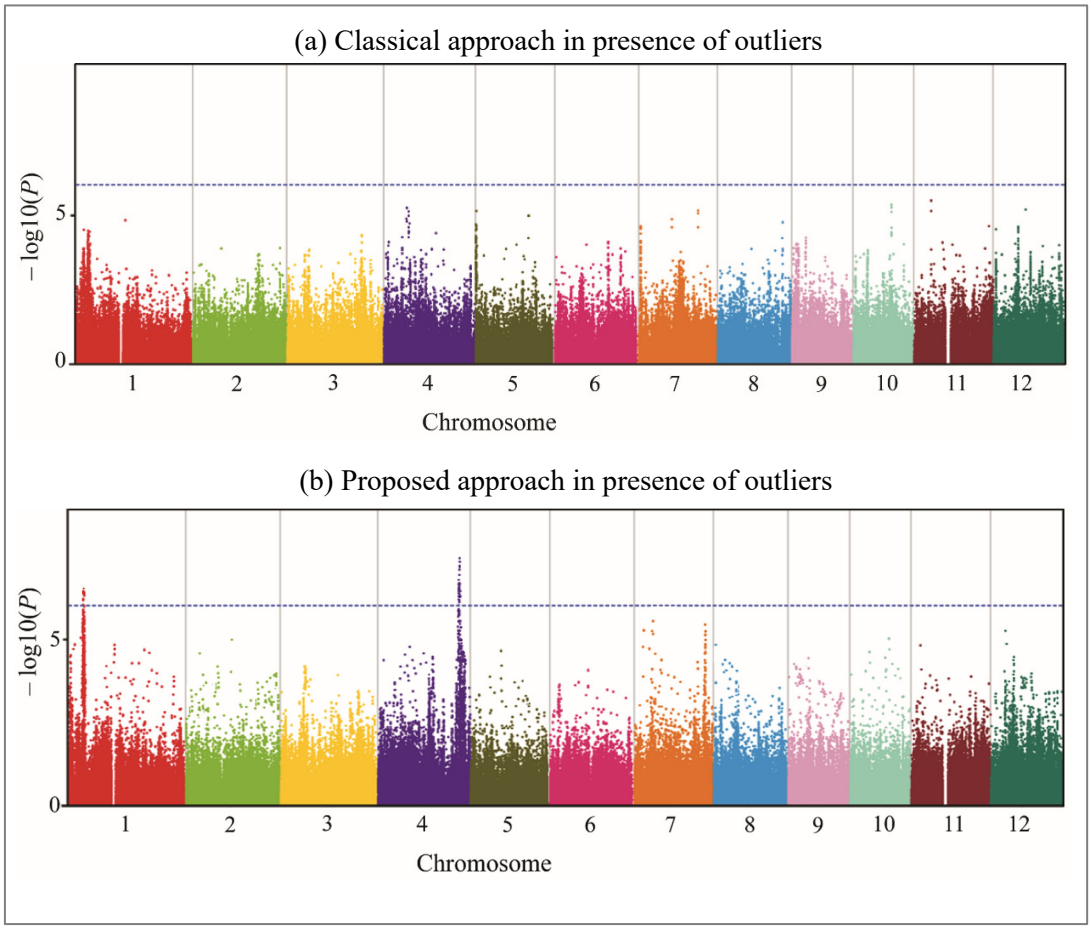


Figure 5.7: Manhattan plot of GWAS to identify important SNPs/QTLs which control the "gain number per panicle" in Hangzhou area in presence of outliers. Manhattan plot has been created plotting $[-\log_{10}(P)]$ values from the linear model in Y-axis and all the SNP positions in X-axis for each of the 12 chromosomes of rice. The horizontal dotted line represent the threshold *P*-value of 10^{-6} to identify the genome-wide significant SNPs using classical and proposed method in absence of phenotypic outliers. (a) Classical approach in presence of outliers and (b) Proposed approach in presence of outliers.

Table 5.1: Significant SNPs (P <10^{-6}) for one yield-related traits "grain number per panicle" in Hangzhou area in absence of outliers.

Chr	Classical method		Proposed method	
	Position	**– log10(*P*)**	**Position**	**– log10(*P*)**
1	6,346,698	6.30	6,346,698	7.09
4	31,493,318	8.05	31,493,318	9.75

Figure 5.7 represents the Manhattan plot of SNP base GWAS of the trait "grain number per panicle" using classical and proposed method in presence of outliers. From Figure 5.7(a) we find that in presence of phenotypic outliers classical method fails to identify the SNPs as identified in absence and even it cannot detect any significant SNP. On the other hand, from Figure 5.7(b), we observe that in presence of phenotypic outliers the proposed method identifies exactly the same SNPs as identified in absence of outlier on chromosome 1 at locus position 6,346,698 (Table 5.2) and on chromosome 4 at locus position 31,493,318 (Table 5.2). This indicates that the proposed method shows better performance than the classical method to identify significant SNPs in presence.

Table 5.2: Significant SNPs (P <10^{-6}) for one yield-related traits "grain number per panicle" in Hangzhou area in presence of outliers

Chr	Classical method		Proposed method	
	Position	**– log10(*P*)**	**Position**	**– log10(*P*)**
1	No SNP identified	Not available	6,346,698	6.69
4	No SNP identified	Not available	31,493,318	8.55

5.4 Conclusion

All the existing classical methods of SNP based single-trait GWAS is very sensitive to outliers and these methods produce misleading results when the phenotypic data are contaminated by outliers. In this study, we have developed a robust approach for SNP based single-trait GWAS analysis. We have investigated the performance of our proposed method in comparison with the classical method of single-trait GWAS in

absence and presence of outliers using simulation study and real data analysis. Simulation studies show that the proposed method produce almost same results as the classical method in absence of outliers. However, the proposed method outperforms over the classical method in presence of outliers. Real data analysis reveals that our proposed method performs better than the classical method in presence of outliers. Otherwise, it shows almost similar performance to the classical method.

Chapter 6

Sequence Matching Based GWAS to Explore Rolling Leaf Related Important Genes

6.1 Introduction

We have discussed the sequence matching based GWAS in section 1.3.4 in details. We have just recalled the idea of sequence matching based GWAS in this chapter. The next step after completing SNP-based GWAS is to integrate the information, and perform structural and functional analysis of the identified associated loci/QTLs/genes to investigate the molecular mechanisms of the identified loci/QTLs/genes. The whole process from SNP identification to functional analysis is known as GWAS, and the sequence searching based process of structural and functional analysis of the genes of interest is called sequence matching based GWAS (Hall, 2019; Han and Huang, 2013; Park et al., 2012). In sequence matching based GWAS, a particular sequence of interest (genomic sequence or coding sequence or protein sequence) of a gene is tried to match in the whole genome by searching the similar sequences in the whole genome stored in the databases. If the sequence of interest matches with any portion of the whole genome, then we called that the sequence is associated with that portion of the genome. In other words, in sequence matching based GWAS, similar sequences to a sequence of interest are searched in the whole genome stored in the databases. Then those similar sequences are said to be associated with the sequence of interest. In this chapter, we have performed a sequence matching based GWAS to explore the structural and functional

characteristics of rolling leaf (RL) related genes in rice (*Oryza sativa* L.). To explore the structural and functional characteristics of rolling leaf (RL) genes, we have conducted different types of GWAS analyses including gene structure analysis, conserved domain (CD) analysis, phylogenetic analysis, protein-protein interaction network construction, Gene Ontology (GO) analysis, transcription factors (TFs) analysis, gene-set enrichment analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analysis and exploratory gene expression analysis.

Rice (*Oryza sativa* L.) is one of the most important staple food crops all over the world, particularly in Asian countries. More than half of world's population consume rice as a primary source of food and in Asia, where approximately 60% of the world's people live, more than 90% of the world's rice is grown and consumed (Khush, 2005). Over 3.5 billion people are solely dependent upon rice for at least 20% of their daily required calories (Khush, 2013). According to Godfray et al. (2010), the food demand of the world population is expected to be 70% to 100% more by the year 2050. Under the condition of limited farmland, the main challenge to meet the food demand is increasing rice yield per unit area which leads to the improvement in overall rice yields. In addition, quality improvement of rice plant is an important factor for increasing the unit yield of rice (Zhu et al., 2017).

Leaf is the major *photosynthetic* organ in rice which is directly related to biomass accumulation and grain yield production (Alamin et al., 2017). It has been proved that appropriate leaf shape improves photosynthesis rate resulting in the higher grain yield in rice (Wu et al., 2010). As a result leaf morphology, such as leaf length, leaf width and the leaf rolling index which are related to plant architecture, is becoming one of the key concerns in the study of high yield plant breeding, especially high yield rice breeding (Li et al., 2014). The flag leaf or uppermost three leaves were proposed to be long, narrow, moderately rolled (V-shaped), erect and thick for the hybrid rice to be super high yielding (Wu, 2009; Yuan, 1997; Zhou et al., 1995). Particularly, a moderate leaf rolling of rice can increase the grain yield by improving photosynthetic efficiency and increasing stress resistance through reduced leaf transpiration during drought resulting the compact tailing and vertical position of leaf (Lang et al., 2004;

Wu et al., 2010; Zhang et al., 2009; Zou et al., 2011). Thus, identifying moderately rolled leaf rice mutants and isolating the responsible genes for controlling leaf rolling will be advantageous for rice breeding with the desired traits (e.g., high yield, stress tolerance) and architecture (Alamin et al., 2017).

To date, at least 103 genes/QTLs for rolling leaf (RL) trait of rice have been identified by several studies (S1 Table). These RL genes/QTLs have been either cloned or mapped throughout the rice genome. *ROLLED LEAF 9* (*RL9*) and *SHALLOT-LIKE1* (*SLL1*) are identical genes which encoded a SHAQKYF class Myb family transcription factor belonging to the KANADI family (Yan et al., 2008; Zhang et al., 2009). *RL9* encodes a GARP protein of KANADI family regulating the completely adaxialized leaves and malformed spikelets (Yan et al., 2008). *SHALLOT-LIKE1* (*SLL1*) mutant have extremely incurved leaves on the abaxial side and encodes a SHAQKYF class MYB family transcription factor belonging to the KANADI family (Zhang et al., 2009). *SHALLOT-LIKE 2* (*SLL2*) showed shrinkage of bulliform cells and this mutant's leaf rolling is caused by a T-DNA insertion (J. J. Zhang et al., 2015). *SEMI-ROLLED LEAF1* (*SRL1*) or *CURLED LEAF AND DWARF 1* (*CLD1*) encodes a putative glycosylphosphatidylinositol anchored protein which modulates leaf rolling (Li et al., 2017; Xiang et al., 2012). The *outcurved laef1* (*oul1*) mutant exhibited abaxial leaf rolling due to the knockout of *RICE OUTERMOST CELL-SPECIFIC GENE 5* (*ROC5*), and interestingly the number and size of bulliform cells decrease due to the over expression of *Roc5* resulting the adaxial leaf rolling (Zou et al., 2011). *RL14* gene encodes a 2OG-Fe (II) oxygenase family protein of unknown function modulating the incurved rice leaves due to the shrinkage of bulliform cells on the adaxial side (Fang et al., 2012).

COW1/NAL7 and *NRL1* encoding a flavin-containing monooxygenase and a cellulose synthase-like protein D4 (OsCslD4), respectively, results in increased number and smaller size of bulliform cells and therefore adaxial leaf rolling (Fujino et al., 2008). *Adaxialized leaf1* (*ADL1*) mutant in rice encodes a plant specific calpain like cysteine proteinase orthologous to maize *DEFECTIVE KERNEL1*, results in number and size change on the adaxial and formation on the abaxial epidermis of the bulliform cells

and caused leaf inward rolling (Hibara et al., 2009). *OsMYB103L* encodes an R2R3-MYB transcription factor and overexpression of *OsMYB103L* results in a rolled leaf phenotype (Yang et al., 2014). *ACL1* (*Abaxially Curled Leaf 1*) encodes a protein of 116 amino acids with unknown conserved functional domains (Li et al., 2010). *OsAG07* gene encodes a protein of 1,048 amino acids including the PAZ and PIWI conserved domains (Shi et al., 2007). Furthermore, *REL1* encodes an unknown protein and plays a positive role in leaf rolling and bending (Chen et al., 2015); and an unknown functional protein which contains DUF630 and DUF632 domains is encoded by *REL2* gene (Yang et al., 2016). Suppression of *OsYABBY6* transcriptional activity results in shrinkage of bulliform cells and adaxially rolled leaves in rice (M. L. Xia et al., 2017).

Isolation of genes controlling leaf rolling is expected to be beneficial for developing crops with the desired architecture (Xu et al., 2014; Zhang et al., 2009). Most of the studies of RL trait are related to the comparison of a rolled leaf mutant with a wild type (WT) and the identification of the genes/QTLs responsible for leaf rolling. Those studies also include the investigation of the molecular functions (e.g., the function of domains; changes in the volume, localization and number of bulliform cells) and mapping/cloning of the identified RL genes/QTLs. Rolling leaf gene isolation, investigation of its molecular functions, and mapping/cloning the identified gene for developing a rice mutant with desired trait and architecture are the time consuming and laborious tasks. So, collective information of the RL genes and their comparative analysis is very essential and helpful before starting the experiment and functional studies of a desired RL rice mutant. Although no fewer than 103 RL genes/QTLs have been cloned or mapped throughout many different studies, there is no collective information on the RL genes; and the comparative analysis of their sequences (i.e., genomic sequence, coding sequence (CDS) and protein sequence) from different bioinformatics point of view is still incomplete. Therefore, this *in silico* study was designed to identify and compare the structures and functions of all RL genes, reported till date through several studies, using various bioinformatics analyses. To our knowledge, this is the first study where we listed almost all the RL genes characterized throughout several studies to date and performed different types of

comparative analyses from different bioinformatics point of view including gene structure and exons/introns pattern analysis, domain analysis, phylogenetic analysis, Gene Ontology (GO) analysis, transcription factor (TF) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis, gene network analysis and gene expression analysis. We found that LR genes are diverse in structure and most of them contain different types of domains. More than 50% of the RL genes of interest are not associated with each other and most of LR genes have some extreme (very high or very low) gene expression values at leaf, root and shoot. Altogether, this study might provide collective information about RL genes and their molecular characteristics till date, and might be helpful for the geneticist as well as rice breeder to develop the rice mutants with rolled leaf trait and desired architecture.

6.2 Materials and Methods

The whole process of this study, from data collection to data analysis, has been illustrated through a schematic diagram in Figure 6.1. The data collection procedures for RL related genes and the bioinformatic analysis techniques used to analyze the data have been discussed in the next subsections.

6.2.1 Data Collection Procedures for RL Related Genes

Identification of RL related genes was done in several ways in this study. First, we searched and collected all the published papers which were related to RL of rice (*Oryza sativa L.*). From these collected papers we extracted all the RL related genes. We also used some most popular and widely used publicly available databases such as Oryzabase (http://shigen.nig.ac.jp/rice/oryzabase), GRAMENE (http://archive.gramene.org/), Rice Genome Annotation Project (RGAP: http://rice.plantbiology.msu.edu/index.shtml), Rice Genome Browser (http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/), Rice Annotation Project (RAP) database (RAP-DB: https://rapdb.dna.affrc.go.jp/), PubMed (https://www.ncbi.nlm.nih.gov/pubmed), and some most popular search engines (e.g., Google Scholar, Google, Bing, etc.) to find out the RL genes of rice along with locus ID. In some papers the MSU locus IDs of RL genes were reported, while some papers mentioned the RAP locus IDs instead of MSU ID for RL related genes. Also in some

papers only the names of the RL genes were reported instead of locus ID. MSU/RAP locus IDs of the RL genes, for which locus IDs were not reported, was found using various web databases of rice including Oryzabase (http://shigen.nig.ac.jp/rice/oryzabase), GRAMENE (http://archive.gramene.org/), Rice Genome Browser (http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/) and Rice Annotation Project (RAP) database (RAP-DB: https://rapdb.dna.affrc.go.jp/). RAP IDs of RL genes were converted to MSU IDs using the Rice Annotation Project (RAP) database (http://rapdb.dna.affrc.go.jp).

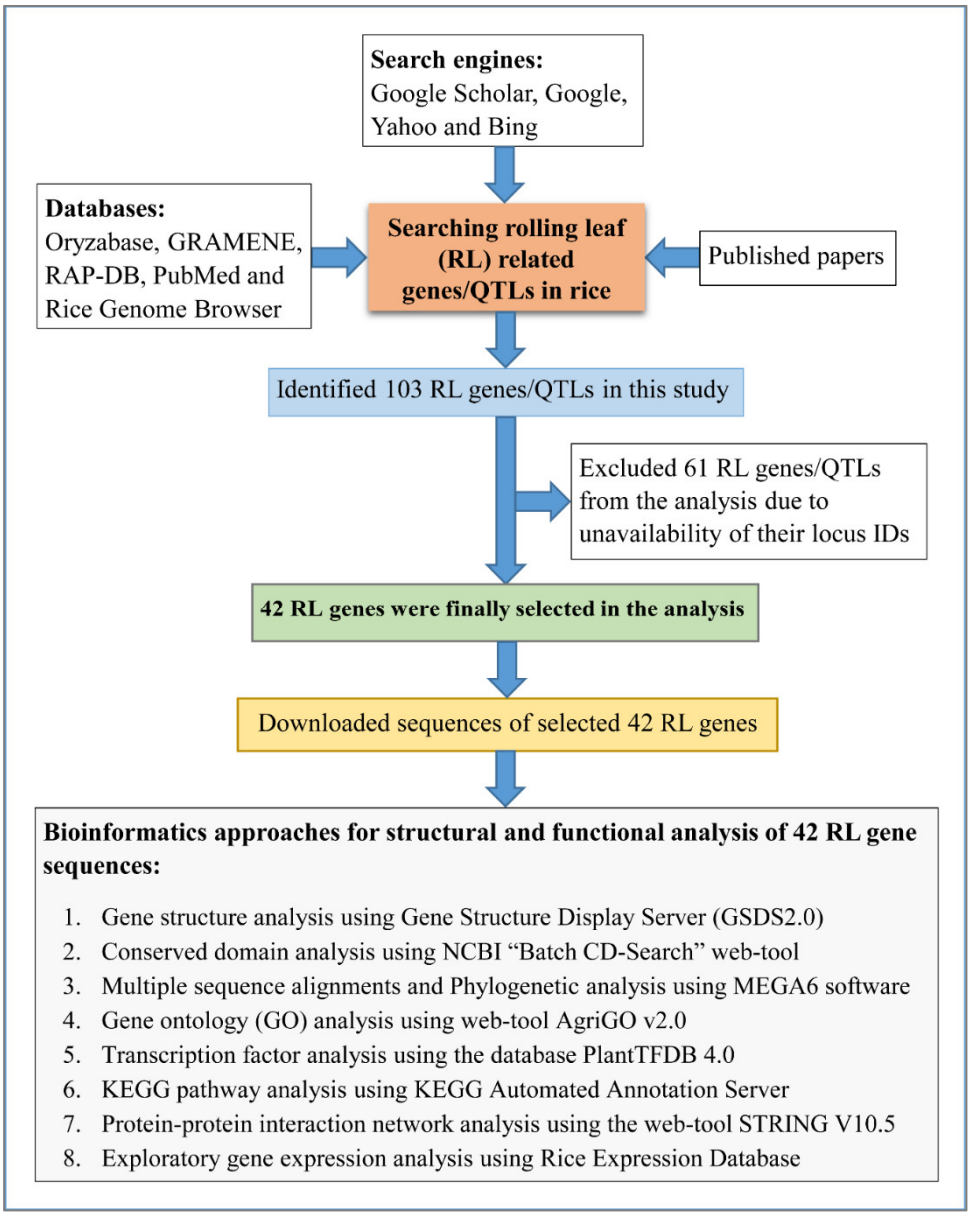


Figure 6.1: Schematic diagram of the study.

Total 103 RL related genes/QTLs reported in different kind of literature and databases were identified in this study. The detail information of the identified 103 RL genes/QTLs are available in Table 6.1. Among the total 103 RL genes/QTLs, 9, 11, 16, 9, 7, 5, 14, 3, 12, 9, 2 and 8 genes/QTLs were identified on chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, respectively. Note that the genes *RL1*, *RL3* and *RL11(t)* have been identified on two different chromosomes (1, 12), (3, 12) and (7, 4), respectively. Among the 103 RL genes, *ND1* (Li et al., 2009), *DNL3* (Shi et al., 2016) and *NRL1* (Hu et al., 2010; Wu et al., 2010) are identical; RL9 (Yan et al., 2008) and *SLL1* (Zhang et al., 2009) are identical; and *COW1* (Woo et al., 2007) and *NAL7* (Fujino et al., 2008) are identical (Table 6.1). So we got a total of 96 distinct RL genes/QTLs after considering one gene from each group of similar genes. Among 96 distinct genes/QTLs, all the QTLs (19 QTLs) were excluded from the analysis since locus ID was not available for any of the QTLs. Then among the remaining 77 genes, we excluded 35 genes for which locus IDs were not found. Finally, we included only 42 RL genes in our analysis for which locus IDs were available. As a result, we confined all of our analyses with only those 42 RL genes. Then we downloaded the genomic sequence, coding sequence (CDS) and protein sequence from the RAP database (http://rice.plantbiology.msu.edu/index.shtml) for each of the 42 RL genes included in our study. For convenience, we defined these 42 finally selected RL genes as “rolling leaf genes of interest” in our study.

Table 6.1: Rolling leaf genes of rice and their references

SN	Gene/QTL	Gene name	Phenotype of rice leaf	Chr	MSU Locus ID	References
1	Gene	*NAL4*	Narrow leaf	4	Unknown	Yen et al. (1968); [http://ejournal.sinica.edu.tw/bbas/content/1968/1/bot091-10.pdf]
2	Gene	*NAL6*	Narrow leaf	3	Unknown	https://shigen.nig.ac.jp/rice/oryzabase/gene/detail/568
3	Gene	*RL1*	Inward rolling	1,12	Unknown	Yoshimura et al. (1997); http://archive.gramene.org/db/genes/search_gene?acc=GR:0060764
4	Gene	*RL2*	Inward rolling	1	Unknown	Yoshimura et al. (1997); http://archive.gramene.org/db/genes/search_gene?acc=GR:0060765
5	Gene	*RL3*	Inward rolling	3,12	Unknown	Yoshimura et al. (1997); http://archive.gramene.org/db/genes/search_gene?acc=GR:0060766
6	Gene	*RL4*	Inward rolling	1	Unknown	Yoshimura et al. (1997); http://archive.gramene.org/db/genes/search_gene?acc=GR:0060767
7	Gene	*RL5*	Inward rolling	3	Unknown	Yoshimura et al. (1997); http://archive.gramene.org/db/genes/search_gene?acc=GR:0060768
8	Gene	*RFS*	Rolled fine striped leaf	7	LOC_Os07g31450	Yoshimura et al. (1997); Cho et al. (2018)
9	Gene	*OsCHR4*	Narrow albino leaf	7	LOC_Os07g31450	Zhao et al. (2012); Xu et al. (2017); Cho et al. (2018)
10	Gene	*CHR729*	Narrow albino leaf	7	LOC_Os07g31450	Hu et al. (2012); Ma et al. (2015); Xu et al. (2017); Cho et al. (2018)
11	Gene	*NAAL1*	Narrow albino leaf	7	LOC_Os07g31450	Xu et al. (2017)
12	Gene	*rl6*	Inward rolling	7	Unknown	http://archive.gramene.org/db/genes/search_gene?acc=GR:0060769
13	Gene	*rl7*	Inward rolling	5	Unknown	Li et al. (1998)

SN	Gene/QTL	Gene name	Phenotype of rice leaf	Chr	MSU Locus ID	References
14	Gene	*rl(t)*	Inward rolling	2	Unknown	Y. Shao et al. (2005); Pan et al. (2011)
15	Gene	*rl8*	Inward rolling	5	Unknown	Y. J. Shao et al. (2005)
16	Gene	*DCL1*	Rolling leaf	3	LOC_Os03g02970	Liu et al. (2005)
17	Gene	*rl9(t)*	Inward rolling	9	Unknown	Yan et al. (2006)
18	Gene	*rl10(t)*	Inward rolling	3	LOC_Os03g06654	Yi et al. (2007)
19	Gene	*RL10*	Inward rolling	9	LOC_Os09g23200	Luo et al. (2007)
20	Gene	*OsAGO7*	Inward rolling	3	LOC_Os03g33650	Shi et al. (2007)
21	Gene	*YABBY1*	Abaxial rolling	7	LOC_Os07g06620	Dai et al. (2007)
22	Gene	*NAL1*	Narrow leaf	4	LOC_Os04g52479	Qi et al. (2008)
23	Gene	*COW1*	Inward rolling	3	LOC_Os03g06654	Woo et al. (2007)
24	Gene	*NAL7*	Inward rolling	3	LOC_Os03g06654	Fujino et al. (2008)
25	Gene	*url1(t)*	Inward rolling	1	Unknown	Yu et al. (2010)
26	Gene	*RL9*	Inward rolling	9	LOC_Os09g23200	Yan et al. (2008)
27	Gene	*SLL1*	Inward rolling	9	LOC_Os09g23200	Zhang et al. (2009)
28	Gene	*ADL1*	Outward rolling	2	LOC_Os02g47970	Hibara et al. (2009)
29	Gene	*rl11(t)*	Inward rolling	7, 4	Unknown	Shi et al. (2009); Zhou et al. (2010)
30	Gene	*OsAS2*	Aberrant twisted leaf	1	LOC_Os01g66590	Ma et al. (2009)
31	Gene	*rl12(t)*	Inward rolling	10	Unknown	Luo et al. (2009)
32	Gene	*ND1*	Narrow leaf and dwarf	12	LOC_Os12g36890	Li et al. (2009)

SN	Gene/QTL	Gene name	Phenotype of rice leaf	Chr	MSU Locus ID	References
33	Gene	*DNL3*	Dwarf and narrow leaf	12	LOC_Os12g36890	Shi et al. (2016)
34	Gene	*nal3(t)*	Inward rolling	12	Unknown	Wang et al. (2009); Feng et al. (2012)
35	Gene	*rl13(t)*	Inward rolling	9	Unknown	Chen et al. (2010)
36	Gene	*ACL1*	Outward rolling	4	LOC_Os04g33860	Li et al. (2010)
37	Gene	*LC2*	Outward rolling	2	LOC_Os02g05840	Zhao et al. (2010)
38	Gene	*NRL1*	Inward rolling	12	LOC_Os12g36890	Hu et al. (2010); Wu et al. (2010)
39	Gene	*nrl2(t)*	Inward rolling	3	Unknown	Wang et al. (2011)
40	Gene	*Roc5*	Outward rolling	2	LOC_Os02g45250	Zou et al. (2011)
41	Gene	*CFL1*	Inward rolling	2	LOC_Os02g31140	Wu et al. (2011)
42	Gene	*rl13*	Inward rolling	6	Unknown	Tian et al. (2012)
43	Gene	*RL14*	Inward rolling	10	LOC_Os10g40960	Fang et al. (2012)
44	Gene	*SRL1*	Inward rolling	7	LOC-Os07g01240	Xiang et al. (2012)
45	Gene	*nul1*	narrow & upper-albino leaf	7	Unknown	F. Wang et al. (2012)
46	Gene	*OsJNBa0003P07*	Rolling leaf	10	Unknown	X. Wang et al. (2012)
47	Gene	*NAL2*	Narrow leaf	11	LOC_Os11g01130	Cho et al. (2013)
48	Gene	*NAL3*	Narrow leaf	12	LOC_Os12g01120	Cho et al. (2013)
49	Gene	*NAL9*	Narrow leaf	3	LOC_Os03g29810	W. Li et al. (2013)
50	Gene	*AGO1a*	Adaxial rolling	2	LOC_Os02g45070	L. Li et al. (2013)
51	Gene	*s1-145*	Adaxial rolling	2	Unknown	Xie et al. (2013)

SN	Gene/QTL	Gene name	Phenotype of rice leaf	Chr	MSU Locus ID	References
52	Gene	*OsZHD1*	Outward rolling	9	LOC_Os09g29130	Xu et al. (2014)
53	Gene	*OsMYB103L*	Upward rolling	8	LOC_Os08g05520	Yang et al. (2014)
54	Gene	*DNAL1*	Narrow leaf	2	Unknown	Sang et al. (2014)
55	Gene	*NAL5*	Narrow leaf	4	Unknown	Cho et al. (2014)
56	Gene	*rl15(t)*	Inward rolling	10	Unknown	Zhang et al. (2014)
57	Gene	*rl28*	Inward rolling	5	Unknown	Feng et al. (2015)
58	Gene	*Nrl3(t)*	Adaxial rolling	2	Unknown	X. H. Zhang et al. (2015)
59	Gene	*SLL2*	Inward rolling	7	LOC_Os07g38664	J. J. Zhang et al. (2015)
60	Gene	*REL1*	Inward rolling	1	LOC_Os01g64380	Chen et al. (2015)
61	Gene	*NAL10*	Narrow leaf	1	Unknown	Fang et al. (2015)
62	Gene	*rl16(t)/RL16*	Rolled leaf	9	LOC_Os09g09360	Liu et al. (2015)
63	Gene	*LRL1*	Late-stage rolled leaf	9	Unknown	Zhao et al. (2015)
64	Gene	*NL(t)*	Narrow leaf	4	Unknown	Pan et al. (2015); Zhang et al. (2016)
65	Gene	*REL2*	Rolling & erect leaf	10	LOC_Os10g41310	Yang et al. (2016)
66	Gene	*SRL2*	Inward rolling	3	LOC_Os03g19520	Liu et al. (2016)
67	Gene	*SCL1*	Semi-curled leaf	2	LOC_Os02g44360	Zhang et al. (2016)
68	Gene	*SRS5*	Leaf rolling	11	LOC_Os11g14220	Segami et al. (2012)
69	Gene	*DTL1*	Twisty leaf	10	Unknown	Zhang et al. (2012)
70	Gene	*OsLBD3-7*	Narrow and adaxially	3	LOC_Os03g57660	Li et al. (2016)

SN	Gene/QTL	Gene name	Phenotype of rice leaf	Chr	MSU Locus ID	References
			rolled leaf			
71	Gene	*NAL11*	Narrow leaf	7	LOC_Os07g09450	Wu et al. (2016);Zhao et al. (2017)
72	Gene	*NRL4*	Narrow and rolling leaf	3	LOC_Os03g19770	Liang et al. (2016)
73	Gene	*OsARVL4*	Abaxially rolled leaves	4	LOC_Os04g33570	Wang et al. (2016)
74	Gene	*OsARF18*	Rolled leaves	6	LOC_Os06g47150	Huang et al. (2016)
75	Gene	*RL15*	Adaxial leaf rolling	1	LOC_Os01g37837	Lee et al. (2016)
76	Gene	*DNL2*	Dwarf and narrow leaf	10	Unknown	Adedze et al. (2017)
77	Gene	*SFL1*	Screw flag leaf	10	LOC_Os10g28060	Alamin et al. (2017)
78	Gene	*OsYABBY6*	Adaxial rolling	12	LOC_Os12g42610	M. L. Xia et al. (2017)
79	Gene	*OsI_14279*	Rolling leaf	3	LOC_Os03g62620	Wang et al. (2017)
80	Gene	*OsRRK1*	Adaxially rolled leaves	6	LOC_Os06g47820	Ma et al. (2017)
81	Gene	*OsHB4*	Adaxially rolled leaves	3	LOC_Os03g43930	Zhang et al. (2018)
82	Gene	*LRRK1*	Adaxially rolled leaves	6	LOC_Os06g07070	Zhou et al. (2018)
83	Gene	*OsSND2*	Rolled leaf	5	LOC_Os05g48850	Ye et al. (2018)
84	Gene	*KAN1*	Upward rolling leaf	9	Unknown	Adedze et al. (2018)
85	QTL	*QFl4*	Inward rolling	4	Unknown	Xu et al. (1999)
86	QTL	*QFl5*	Inward rolling	5	Unknown	Xu et al. (1999)
87	QTL	*QFl7*	Inward rolling	7	Unknown	Xu et al. (1999)
88	QTL	*QFl9*	Inward rolling	9	Unknown	Xu et al. (1999)
89	QTL	*qRL-1*	Outward rolling	1	Unknown	Xu et al. (1999); Guo et al. (2010)

SN	Gene/QTL	Gene name	Phenotype of rice leaf	Chr	MSU Locus ID	References
90	QTL	*qRL3*	Outward rolling	3	Unknown	Xu et al. (1999)
91	QTL	*qRL5*	Outward rolling	5	Unknown	Xu et al. (1999)
92	QTL	*qRL-7*	Outward rolling	7	Unknown	Xu et al. (1999); Guo et al. (2010)
93	QTL	*qRL4-2*	Rolled leaf	4	Unknown	Gao et al. (2007)
94	QTL	*qRL5-9*	Rolled leaf	5	Unknown	Gao et al. (2007)
95	QTL	*qRL5-10*	Rolled leaf	5	Unknown	Gao et al. (2007)
96	QTL	*qRL-2-1b*	Rolled leaf	2	Unknown	Gao et al. (2007)
97	QTL	*qRL-6*	Rolled leaf	6	Unknown	Gao et al. (2007); Guo et al. (2010)
98	QTL	*qRL-8-1*	Rolled leaf	8	Unknown	Gao et al. (2007); Guo et al. (2010)
99	QTL	*qRL-8-2*	Rolled leaf	8	Unknown	Gao et al. (2007); Guo et al. (2010)
100	QTL	*qRL-9*	Rolled leaf	9	Unknown	Gao et al. (2007); Guo et al. (2010)
101	QTL	*qRL-10*	Rolled leaf	10	Unknown	Gao et al. (2007); Guo et al. (2010)
102	QTL	*qRL7b*	Rolled leaf	7	Unknown	Zhang et al. (2016)
103	QTL	*qRL9b*	Rolled leaf	9	Unknown	Zhang et al. (2016)

Chr: Chromosome

6.2.2 Bioinformatics Analysis of RL Genes

We have analyzed the genomic sequences, CDS and protein sequences of the RL genes using various software packages and web-based tools.

6.2.2.1 Gene Structure Analysis

In gene structure analysis, the sequence of gene of interest is searched in the whole genome to match with some portions (which are also sequences) of the genome using different online tools/databases. The sequences that match with the sequence of interest are said to be associated with the sequence of interest. Then the most associated sequence in the genome is selected and its structure is treated as the structure of the gene of interest. In this study, the gene structure analysis has been done based on genomic sequences and CDS using the web-based bioinformatic tool Gene Structure Display Server (GSDS 2.0: http://gsds.cbi.pku.edu.cn) (Hu et al., 2015). Figure 6.2 represents the workflow of the GSDS 2.0. In this analysis, we have investigated the exon-intron structure along with a number of exons and introns of 42 RL genes.

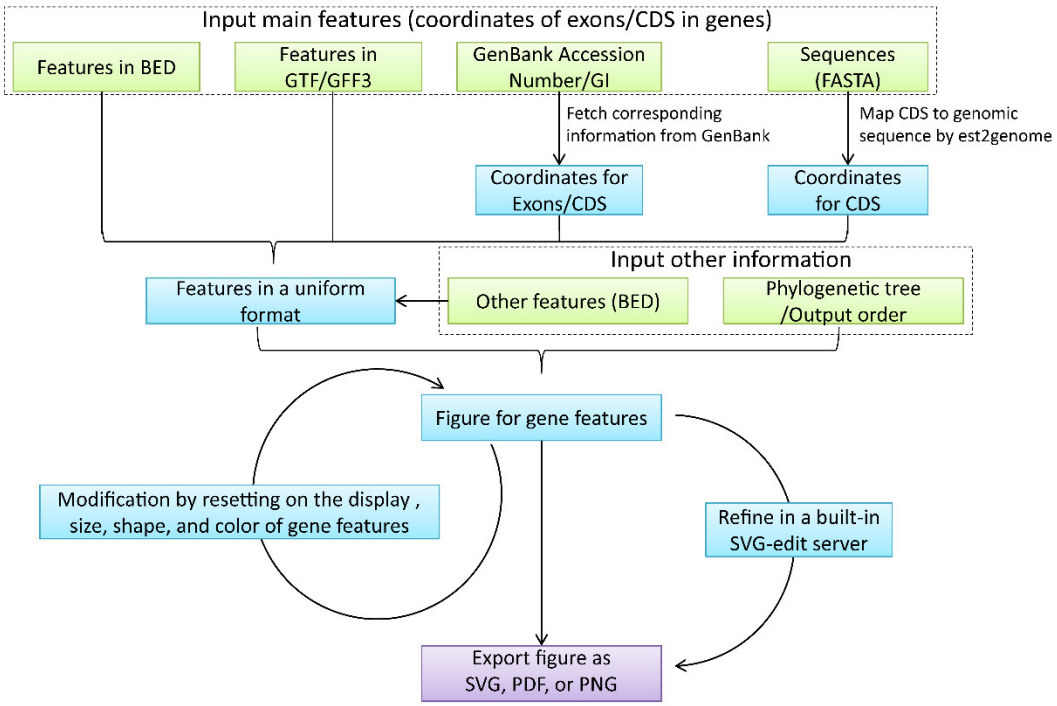


Figure 6.2: Workflow of the gene structure display server (GSDS 2.0).

6.2.2.2 Conserved Domain Analysis

Conserved domain analysis is done using protein sequence of a gene of interest. The identification of a CD might be the only clue towards molecular mechanisms of a gene/protein of interest, as it indicates partial similarity of the query protein to other proteins. Multiple sequence alignments are basis of the conserved domain analysis. The method of multiple sequence alignments has been discussed in appendix A6.1. In conserved domain analysis, multiple sequence alignments is done for different sequences in the whole genome of different species (stored in the database) compared to the query sequence to find the most similar sequence to the query sequence. The obtained sequence is said to be associated with the query sequence of the gene/protein of interest. Then the conserved domains contained by obtained sequence are treated as the conserved domains of the query sequence of the targeted gene/protein. The NCBI "Batch CD-Search" web-based tool (Marchler-Bauer and Bryant, 2004; Marchler-Bauer et al., 2015; Marchler-Bauer et al., 2011) in Conserved Domain Database (CDD: http://www.ncbi.nlm.nih.gov/cdd) has been used to perform conserved domain analysis for the identified RL genes.

6.2.2.3 Phylogenetic Analysis

In our study, MEGA V6 (Tamura et al., 2013) software has been used for the multiple sequence alignments and phylogenetic analysis. Multiple alignments of protein sequences of RL genes are conducted using "ClustalW" method (Larkin et al., 2007; Thompson et al., 1994) and viewed in a compatible printable format using GeneDoc Version 2.6.002 software. The ClustalW method has been discussed in details in appendix A6.1 and Figure A6.1. Maximum likelihood method has been used to construct the phylogenetic tree based on the multiple aligned protein sequences. The maximum likelihood is one of the popular methods for estimating the parameters of a probability distribution (i.e., unknown characteristics/descriptors of a distribution: mean and variance). In phylogenetic analysis there are many parameters including differential transformation costs, rates and the tree. The likelihood function is defined to be the probability of observing the data given the probability distribution (i.e., model), Pr(D|M), where D = Data and M = Model or probability distribution. Hence,

if we have a model (i.e., the parameters and tree) and the data, then we can define the likelihood function as

$$L = \prod_{Data} f(Data|Model) \tag{6.1}$$

which is considered as a function of the model parameters. Then we maximize the likelihood (6.1) with respect to the model parameters to find the maximum likelihood estimates of the parameters of interest (usually the tree and branch lengths). The values of the parameters at which the likelihood function is maximum is called the maximum likelihood estimates of the parameters.

6.2.2.4 Gene Ontology (GO) Analysis

In GO analysis, GO enrichment score is calculated for the protein sequence of the gene of interest and GO term of the sequences in the whole genome stored in the database. If the GO enrichment score for the protein of interest and one GO term is higher, then they are said to be strongly associated with each other. The calculation of the GO score is done using the following statistical method.

Let $G(PT)$ be a set of protein sequences in the genome that have association with the protein sequence of interest *PT*. Given on protein sequence *PT* and one GO term $GO_i, i = 1, 2, \ldots,$ (number of GO terms in the database), the GO enrichment score is calculated as $-\log_{10}(P)$ where *P* is the p-value obtained by hypergeometric test of association between the protein *PT* and GO term GO_i, which can be computed by the following equation.

$$S_{GO}(PT, GO_i) = -\log_{10}\left[\sum_{k=m}^{n} \frac{\binom{M}{n}\binom{N-M}{n-m}}{\binom{N}{n}}\right] \tag{6.2}$$

where $N =$ total number of proteins in the whole genome,

$M =$ number of proteins that are annotated to the GO term GO_i,

$n =$ number of proteins in $G(PT)$ and

$m =$ number of proteins both in $G(PT)$ and annotated to GO term GO_i.

If the GO enrichment score for one protein and one GO term is higher, then they are strongly associated with each other.

The web-based tool AgriGO v2.0 (http://systemsbiology.cau.edu.cn/agriGOv2/) (Tian et al., 2017) has been used for gene ontology (GO) analysis with the option Singular Enrichment Analysis (SEA) to investigate the functional enrichments of the RL genes of interest. *Oryza sativa* japonica was selected for species and MSU7.0 gene ID (TIGR) was selected as a reference background. A GO term was considered significantly enriched among a set of genes if p-value is less than 0.05. We constructed the heatmap of GO term versus Gene along with the bar plot of GO term versus -log10(p-value) for all significant GO terms.

6.2.2.5 Transcription Factor (TF) Analysis

In transcription factor (TF) analysis, the sequence of gene of interest is searched in the whole genome to match with some portions (which are also sequences) of the genome using different online tools/databases. The sequences that match with the sequence of interest are said to be associated with the sequence of interest. Then the most associated sequence in the genome is selected and its TFs are treated as the TFs of the gene of interest. In this study, transcription Factors (TFs) has been identified by using the PlantTFDB 4.0 database (Jin et al., 2017) (http://planttfdb.cbi.pku.edu.cn/).

6.2.2.6 Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Analysis

In KEGG pathway analysis, KEGG enrichment score is calculated for the protein sequence of the gene of interest and KEGG pathway of the sequences in the whole genome stored in the database. If the KEGG enrichment score for the protein of interest and one KEGG pathway is higher, then they are said to be strongly associated with each other. The calculation of the KEGG score is done using the following statistical method.

Let $G(PT)$ be a set of protein sequences in the genome that have association with the protein sequence of interest PT. Given on protein sequence PT and one KEGG

pathway $KEGG_i, i = 1, 2, \ldots,$ (number of KEGG pathways in the database), the KEGG enrichment score is calculated as $-\log_{10}(P)$ where P is the p-value obtained by hypergeometric test of association between the protein PT and KEGG pathway $KEGG_i$, which can be computed by the following equation.

$$S_{KEGG}(PT, KEGG_i) = -\log_{10}\left[\sum_{k=m}^{n} \frac{\binom{M}{n}\binom{N-M}{n-m}}{\binom{N}{n}}\right] \quad (6.2)$$

where $N =$ total number of proteins in the whole genome,

$M =$ number of proteins that are annotated to the KEGG pathway $KEGG_i$,

$n =$ number of proteins in $G(PT)$ and

$m =$ number of proteins both in $G(PT)$ and KEGG pathway $KEGG_i$.

If the KEGG enrichment score between one protein and one KEGG pathway $KEGG_i$ is higher, then they are said to be strongly associated with each other.

The KEGG pathway database is the core of the KEGG resource. This is a vast collection of the pathway maps integrating many entities including the genes, proteins, RNAs, glycans, chemical compounds and chemical reactions, as well as disease genes and drug targets, which are stored as individual entries in the other KEGG databases. In this study, KEGG Automated Annotation Server (http://www.genome.jp/tools/kaas/) (Yoshizawa et al., 2007) has been used for KEGG pathway enrichment analysis to get the summary of gene pathway network.

6.2.2.7 Genome Wide Protein-Protein Association Analysis

In genome wide protein-protein association analysis, association between proteins means that the proteins jointly contribute to a shared function. This does not necessarily mean that the proteins are binding each other. In this genome-wide association analysis, the protein sequence of interest is searched in the whole genome to investigate its interaction with other protein sequences in the genome and then find out the protein sequences with which the query protein sequence interacts to perform some biological functions. The protein sequences that interact with the protein sequence of interest are called associated with the sequence of interest. Functional protein association networks analysis was done using the web-based resource

STRING V10.5 (http://string-db.org/) (Szklarczyk et al., 2017) to examine the protein-protein interactions among the RL genes. The prediction of interactions was done with the default settings of the STRING database.

6.2.2.8 Exploratory Gene Expression Analysis

We also performed the exploratory gene expression analysis of the FPKM (Fragments Per Kilobase Million) expression level data using the Rice Expression Database (http://expression.ic4r.org) (L. Xia et al., 2017) to investigate the gene expression pattern of the RL genes. We have investigated the gene expression pattern using the line charts and to examine the extreme expression (very low or high expression) we have used box plots.

6.3 Results

6.3.1 Summary of Genomic Information of Identified RL Genes

The comprehensive information of 42 RL genes of rice including locus ID, gene location, genomic sequence length, CDS length, number of introns and exons, predicted protein length/size, molecular weight (Mol. Wt.) and isolated point (pI), is shown in Table 6.2. Among the 42 RL genes of our interest 3, 6, 9, 3, 1, 3, 5, 1, 3, 3, 2 and 3 genes were located on chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, respectively. The length of RL genes varied from 315 to 16638 base pairs (or nucleotides). The RL gene *OsREL1* has the minimum genomic length (315 nucleotides) and *OsADL1* has the maximum genomic length (16638 nucleotides). The RL gene *OsSLL2* has the lowest CDS length (225 nucleotides) and protein length (74 amino acids) while the gene *OsRFS* has the highest CDS length (6579 nucleotides) and protein length (2192 amino acids). Gene *OsRFS* has the highest molecular weight (243660 kDa) and *OsSLL2* has the lowest molecular weight (8408.14 kDa). The gene *OsREL1* corresponds to the minimum value of pI (4.2637) whereas the gene OsNRL4 exhibits the maximum value of pI (11.8634).

Table 6.2: Gene name, MSU Locus ID, Gene location, Gene length (nucleotides), CDS length (nucleotides), No. of introns and exons, protein length (no. of amino acids), Mol.Wt. (kDa), Isoelectric point (pI) of rolling leaf genes

SN	Gene name	MSU Locus ID	Gene location (Chr: CDS coordinates 5'-3')	Gene length (nucleotides)	CDS length (nucleotides)	Protein length (amino acids)	No. of introns (Exons)	Mol. Wt. (kDa)	pI
1	*OsACL1*	LOC_Os04g33860	Chr04: 20508616 - 20509616	1001	351	116	0 (1)	12705.1	8.5047
2	*OsADL1*	LOC_Os02g47970	Chr02: 29341254 - 29357891	16638	6489	2162	29 (30)	239859	6.1645
3	*OsAGO1a*	LOC_Os02g45070	Chr02: 27341836 - 27330144	11693	3249	1082	22 (23)	120449	9.7493
4	*OsAGO7*	LOC_Os03g33650	Chr03: 19243920 - 19248524	4605	3147	1048	2 (3)	117498	9.7883
5	*OsARF18*	LOC_Os06g47150	Chr06: 28586445 - 28590358	3914	2103	700	2 (3)	75882.2	7.574
6	*OsARVL4*	LOC_Os04g33570	Chr04: 20321514 - 20319029	2486	522	173	3 (4)	19406.1	9.9632
7	*OsAS2*	LOC_Os01g66590	Chr01: 38676728 - 38679707	2980	810	269	1 (2)	27621.9	8.0637
8	*OsCFL1*	LOC_Os02g31140	Chr02: 18641888 - 18639136	2753	825	274	1 (2)	27695.7	6.9898
9	*OsDCL1*	LOC_Os03g02970	Chr03: 1204839 - 1195075	9765	5652	1883	17 (18)	210203	6.6399
10	*OsHB4*	LOC_Os03g43930	Chr03: 24657650 - 24651800	5851	2589	862	17 (18)	93747.6	6.6202
11	*OsI_14279*	LOC_Os03g62620	Chr03: 35443615 - 35440975	2641	951	316	0 (1)	34798.5	4.6582
12	*OsLBD3-7*	LOC_Os03g57660	Chr03: 32859060 - 32866608	7549	2151	716	9 (10)	77448.9	6.7414
13	*OsLC2*	LOC_Os02g05840	Chr02: 2882177 - 2876553	5625	2250	749	3 (4)	82634.6	7.5743
14	*OsLRRK1*	LOC_Os06g07070	Chr06: 3360524 - 3363523	3000	1104	367	3 (4)	40923.5	7.366
15	*OsMYB103L*	LOC_Os08g05520	Chr08: 2951372 - 2948522	2851	1080	359	2 (3)	39954.7	6.6241
16	*OsNAL1*	LOC_Os04g52479	Chr04: 31203525 - 31214741	11217	1749	582	4 (5)	63252.2	4.8546
17	*OsNAL11*	LOC_Os07g09450	Chr07: 4981656 - 4977967	3690	339	112	1 (2)	12030	10.737
18	*OsNAL2*	LOC_Os11g01130	Chr11: 68344 - 70989	2646	705	234	1 (2)	25921.2	8.8891
19	*OsNAL3*	LOC_Os12g01120	Chr12: 64393 - 65004	612	612	203	0 (1)	22352.1	8.7607
20	*OsNAL7*	LOC_Os03g06654	Chr03: 3355041 - 3360485	5445	1266	421	3 (4)	45720.3	9.1475

SN	Gene name	MSU Locus ID	Gene location (Chr: CDS coordinates 5'-3')	Gene length (nucleotides)	CDS length (nucleotides)	Protein length (amino acids)	No. of introns (Exons)	Mol. Wt. (kDa)	pI
21	*OsNAL9*	LOC_Os03g29810	Chr03: 16994268 - 16997742	3475	780	259	8 (9)	28186.7	7.9423
22	*OsNRL1*	LOC_Os12g36890	Chr12: 22607315 - 22602880	4436	3648	1215	1 (2)	132161	7.9039
23	*OsNRL4*	LOC_Os03g19770	Chr03: 11127352 - 11126732	621	621	206	0 (1)	20837.2	11.8634
24	*OsREL1*	LOC_Os01g64380	Chr01: 37365967 - 37365653	315	315	104	0 (1)	10872.1	4.2637
25	*OsREL2*	LOC_Os10g41310	Chr10: 22202335 - 22198114	4222	2304	767	3 (4)	85721.4	7.1569
26	*OsRFS*	LOC_Os07g31450	Chr07: 18638679 - 18625785	12895	6579	2192	10 (11)	243660	6.8132
27	*OsRL14*	LOC_Os10g40960	Chr10: 21997584 - 21995563	2022	438	145	1 (2)	16325.9	8.6486
28	*OsRL15*	LOC_Os01g37837	Chr01: 21175744 - 21179966	4223	1338	445	8 (9)	50605.7	6.4233
30	*OsRL9*	LOC_Os09g23200	Chr09: 13758215 - 13765052	6838	1599	532	6 (7)	54267.3	9.4739
31	*OsRRK1*	LOC_Os06g47820	Chr06: 28941271 - 28943704	2434	1179	392	5 (6)	43716.1	5.6404
32	*OsRoc5*	LOC_Os02g45250	Chr02: 27494914 - 27487865	7050	2415	804	8 (9)	86047.2	5.4935
33	*OsSCL1*	LOC_Os02g44360	Chr02: 26841585 - 26844331	2747	2130	709	0 (1)	74248.2	5.9887
34	*OsSFL1*	LOC_Os10g28060	Chr10: 14565737 - 14560601	5137	1572	523	1 (2)	56867.7	9.6336
35	*OsSLL2*	LOC_Os07g38664	Chr07: 23217095 - 23220951	3857	225	74	0 (1)	8408.14	10.378
36	*OsSND2*	LOC_Os05g48850	Chr05: 28003165 - 28005001	1837	945	314	2 (3)	34100.4	9.4414
37	*OsSRL1*	LOC_Os07g01240	Chr07: 143578 - 134162	9417	1101	366	4 (5)	39053	6.09
38	*OsSRL2*	LOC_Os03g19520	Chr03: 10979478 - 10970763	8716	2970	989	18 (19)	110441	6.6818
39	*OsSRS5*	LOC_Os11g14220	Chr11: 7960506 - 7963390	2885	1356	451	3 (4)	49737	4.5682
40	*OsYABBY1*	LOC_Os07g06620	Chr07: 3221592 - 3229297	7706	618	205	5 (6)	23325.5	11.2851
41	*OsYABBY6*	LOC_Os12g42610	Chr12: 26477633 - 26487388	9756	624	207	5 (6)	22777.8	9.042
42	*OsZHD1*	LOC_Os09g29130	Chr09: 17705611 - 17703754	1858	840	279	0 (1)	29541.1	7.4822

Chr: Chromosome, Mol.: Molecular, Wt.: Weight, pI: isoelectric point.

6.3.2 Gene Structure Analysis of RL Genes

The gene structure and exon-intron pattern of 42 RL genes of interest were comparatively analyzed and presented in Figure 6.3. The number of introns in the RL genes of interest ranges from 0 to 29. The gene *OsADL1* has the maximum number of exons (30) whereas the genes *OsACL1*, *OsI_14279*, *OsNAL3*, *OsNRL4*, *OsREL1*, *OsSCL1*, *OsSLL2* and *OsZHD1* have the lowest number of the exon (only one exon and zero intron). *OsAGO1a* and *OsSRL2* have the second and third highest number of exons (23 and 19, respectively). Each of the RL genes *OsDCL1*, *OsHB4* and *OsRL16* gene has 18 exons whereas *OsRFS* has 11 exons; and *OsLBD3-7* gene has 10 exons. Also, each *OsNAL9*, *OsRL15* and *OsRoc5* gene, gene *OsRL9*, and each *OsRRK1*, *OsYABBY1* and *OsYABBY6* gene has 9, 7 and 6 exons, respectively in our study. Moreover, each *OsNAL1* and *OsSRL1* gene; each *OsARVL4*, *OsLC2*, *OsLRRK1*, *OsNAL7*, *OsREL2* and *OsSRS5* gene; each *OsAGO7*, *OsARF18*, *OsMYB103L* and *OsSND2* gens; and each *OsAS2*, *OsCFL1*, *OsNAL11*, *OsNAL2*, *OsNRL1*, *OsRL14* and *OsSFL1* gene has 5, 4, 3, 2 exons, respectively. Although each of the genes *OsAS2*, *OsCFL1*, *OsNAL11*, *OsNAL2*, *OsNRL1*, *OsRL14* and *OsSFL1* has two exons and one intron, the length of the exons and introns are different from each other indicating diversity in genes structure. The genes *OsNAL1*, *OsSFL1*, *OsRL9*, *OsSRL1*, *OsYABBY1* and *OsYABBY6* contain an intron with an unusual large size (Figure 6.3). Genes *OsADL1*, *OsNAL1*, *OsNAL7*, *OsSLL2* and *OsSRL2* have long upstream; *OsAS2*, *OsI_14279*, *OsDCL1*, and *OsSRL1* longer downstream; *OsCFL1* and *OsDCL1* do not have upstream; and *OsNAL2*, *OsNAL3*, *OsNRL4*, *OsREL1*, *OsRFS* and *OsRL9* do not have upstream and downstream. The gene structure analysis shows that the genes *OsADL1*, *OsDCL1*, *OsHB4*, *OsRRK1*, and *OsSRL2* are conserved in a sense that they have short length introns (Figure 6.3).

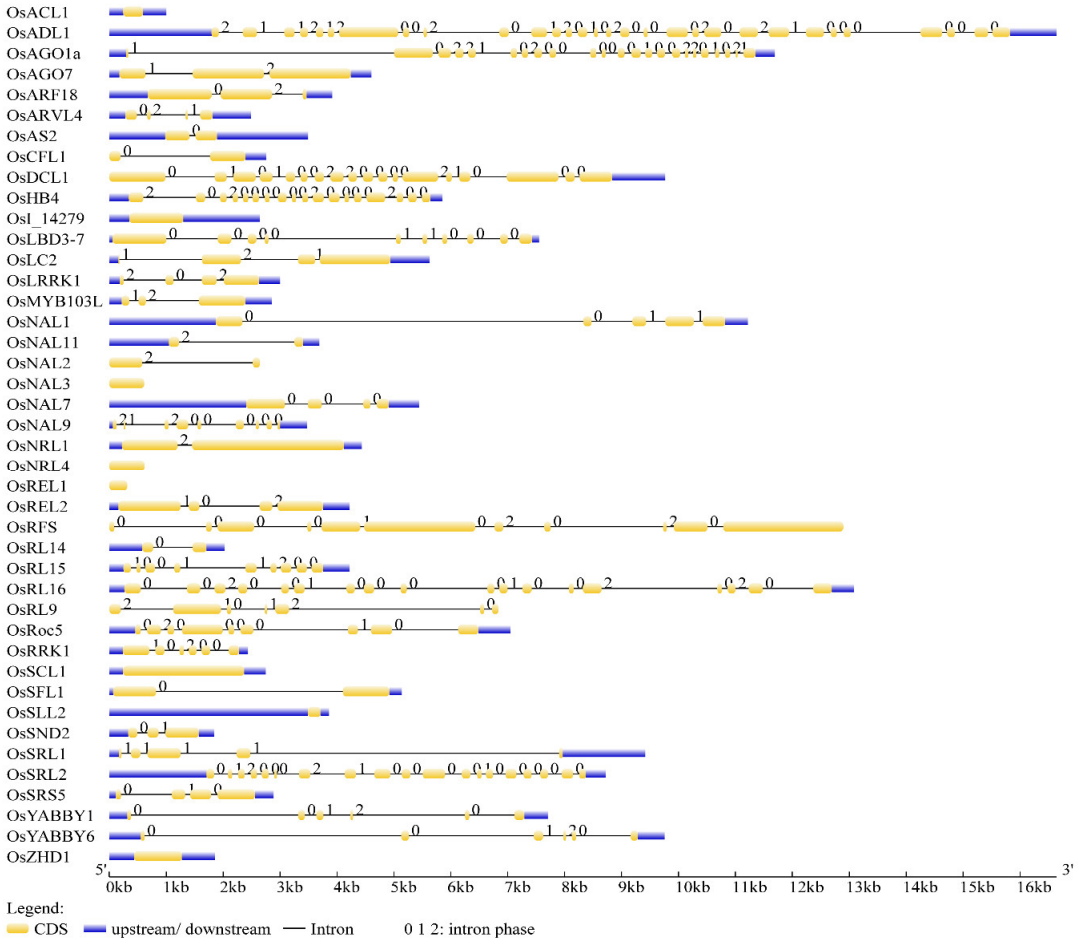


Figure 6.3: Gene structure of 42 rolling leaf genes. The blue color area at the start is representing the upstream, the blue color area at the end is representing the downstream, the yellow color area is representing the exon (CDS) and the black color line is representing the intron of each gene. The intron phase is indicated by the numbers 0, 1 and 2. The exon/intron structure was constructed using Gene Structure Display Server 2.0 (GSDS2.0: http://gsds.cbi.pku.edu.cn).

6.3.3 Domain Analysis of RL Genes

Figure 6.4 and Table A6.1 display the results of the domain analysis of the 42 RL genes identified in this study using the CDD web-tool of NCBI. Conserved domain analysis shows that the genes *OsACL1*, *OsNRL4*, *OsREL1*, *OsSLL2*, *OsSRL1*, *OsSRL2* and *OsNAL1* do not contain any database-recognized domain in spite of having an important role in leaf rolling of rice. Genes *OsARVL4*, *OsAS2*, *OsCFL1*, *OsLRRK1*, *OsNAL11*, *OsNAL2*, *OsNAL3*/*OsWOX3*, *OsNAL9*, *OsRL14*, *OsRL15*, *OsRL16*, *OsRL9*/*OsSLL1*, *OsRRK1*, *OsSCL1*, *OsSFL1*, *OsSND2*, *OsSRS5*, *OsYABBY1* and *OsYABBY6* contain 1 conserved domain; genes *OsAGO1a*, *OsARF18*, *OsI_14279*, *OsLC2*, *OsREL2* and *OsZHD1* contain 2 conserved domains; genes *OsADL1*, *OsLBD3-7*, *OsNAL7*/*OsCOW1*, *OsNRL1* and *OsRoc5* contain 3 conserved domains; *OsHB4* and *OsMYB103L* contain 4 conserved domains; *OsAGO7* contains 5 conserved domains; *OsDCL1* contains 6 conserved domains; and *OsRFS* contains 7 conserved domains. So the gene *OsRFS* contains the maximum number of domains (7 domains), and the genes *OsARVL4*, *OsAS2*, *OsCFL1*, *OsLRRK1*, *OsNAL11*, *OsNAL2*, *OsNAL3*/*OsWOX3*, *OsNAL9*, *OsRL14*, *OsRL15*, *OsRL16*, *OsRL9*/*OsSLL1*, *OsRRK1*, *OsSCL1*, *OsSFL1*, *OsSND2*, *OsSRS5*, *OsYABBY1* and *OsYABBY6* contain the minimum number of domains (1 domain). The genes *OsAGO1a* and *OsAGO7* encode a common conserved domain PIWI. The genes *OsRL9*/*OsSLL1* and *OsMYB103L* contain one similar domain “Myb-like DNA-binding domain”. The genes *OsAGO7* and *OsDCL1* contain a PAZ domain. The genes *OsHB4*, *OsNAL2*, *OsNAL3*/*OsCOW1*, *OsRoc5* and *OsZHD1* have one Homeobox domain. Gene *OsMYB103L* contains two SANT domains and two Myb_DNA_binding domains. *OsRFS* contains two CHROMO domains. Genes *OsRoc5* and *OsHB4* contain one Basic leucine zipper (bZIP) domain. Both the genes *OsYABBY1* and *OsYABBY6* contain one YABBY domain. Genes *OsADL1*, *OsARF18*, *OsARVL4*, *OsAS2*, *OsCFL1*, *OsI_14279*, *OsLBD3-7*, *OsLC2*, *OsLRRK1*, *OsNAL11*, *OsNAL7*, *OsNAL9*, *OsNRL1*, *OsREL2*, *OsRL14*, *OsRL15*, *PGAP1*, *OsRRK1*, *OsSCL1*, *OsSFL1*, *OsSND2* and *OsSRS5* do not have any common domain among them.

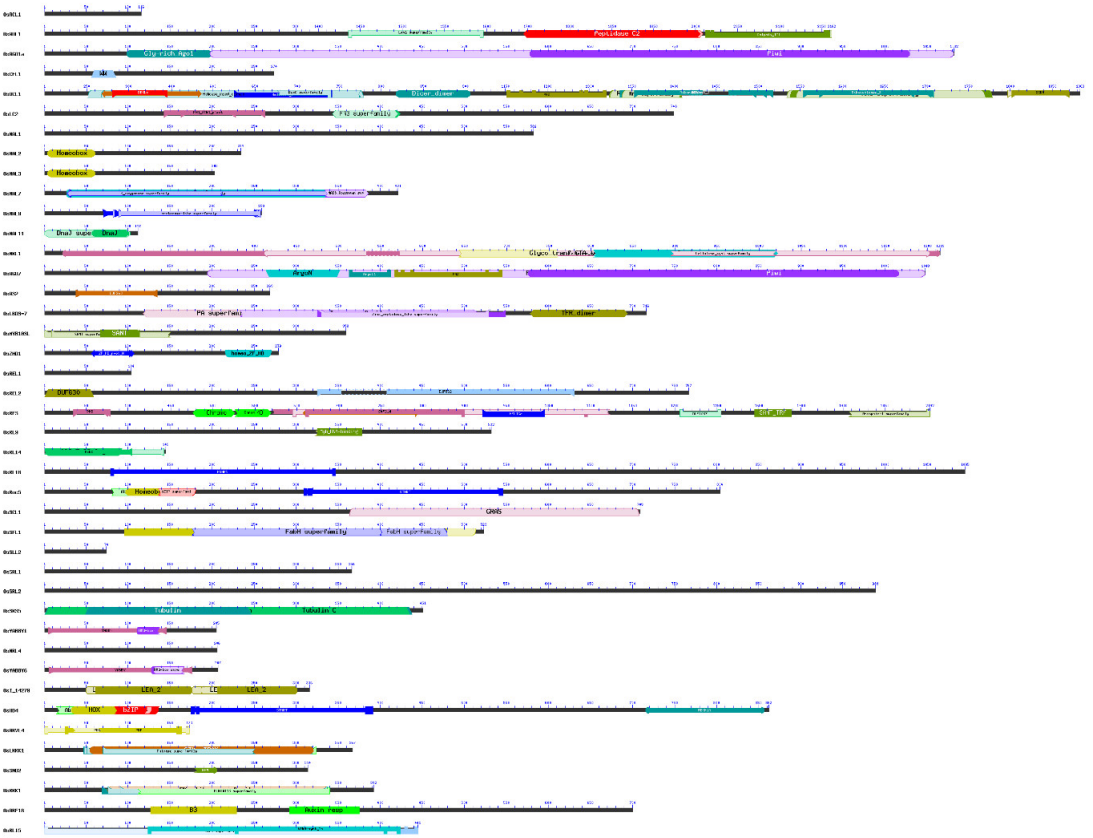


Figure 6.4: Domain organization of the 42 rolling leaf (RL) genes of interest identified in this study. Domains are indicated with different colors except black. Domain analysis of 42 RL genes was done using online Conserved Domain Database (CDD) tool "Batch CD-Search" of National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi).

6.3.4 Phylogenetic Analysis of RL Related Genes

To investigate the evolution of RL genes, we constructed a maximum likelihood (ML) based phylogenetic tree on the alignment of full-length RL proteins (Figure 6.5). The phylogenetic tree showed that the 42 RL proteins clustered into five major groups (I–V) with groups III and IV further divided into two subgroups while groups V separated into four subgroups with robust bootstrap support (Figure 6.5). The RL genes *OsAS2*, *OsNAL1*, *OsSND2*, *OsRL9*, *OsCFL1* and *OsNRL4* gathered into the group I, while the RL genes *OsRRK1*, *OsLRRK1*, *OsSRS5* and *OsACL1* formed the group II. Rolling leaf genes *OsNRL1*, *OsLBD3-7*, *OsRL14*, *OsNAL11* and *OsRL16* constructed the subgroup IIIa while the genes *OsRoc5* and *OsHB4* congregate in

subgroup IIIb. Genes *OsRFS*, *OsRL15*, *OsNAL7*, *OsZHD1*, *OsNAL9*, *OsADL1* and *OsSRL2* grouped into group IVa whereas *OsSFL1* and *OsSRL1* grouped into group IVb. The RL genes *OsREL2*, *OsARVL4*, *OsI_14279*, *OsDCL1* and *OsSLL2*; *OsMYB103L* and *OsSCL1*; *OsAGO1a*, *OsAGO7*, *OsYABBY1* and *OsYABBY6*; and *OsNAL2*, *OsNAL3*, *OsREL1*, *OsLC2* and *OsARF18* appeared in group Va, Vb, Vc and Vd, respectively. These results indicate that the 42 RL genes can be divided into 10 groups namely I, II, IIIa, IIIb, IVa, IVb, Va, Vb, Vc and Vd containing 6, 4, 5, 2, 7, 2, 5, 2, 4 and 5 genes, respectively. These results can be helpful for further evolutionary and functional studies of RL related genes.

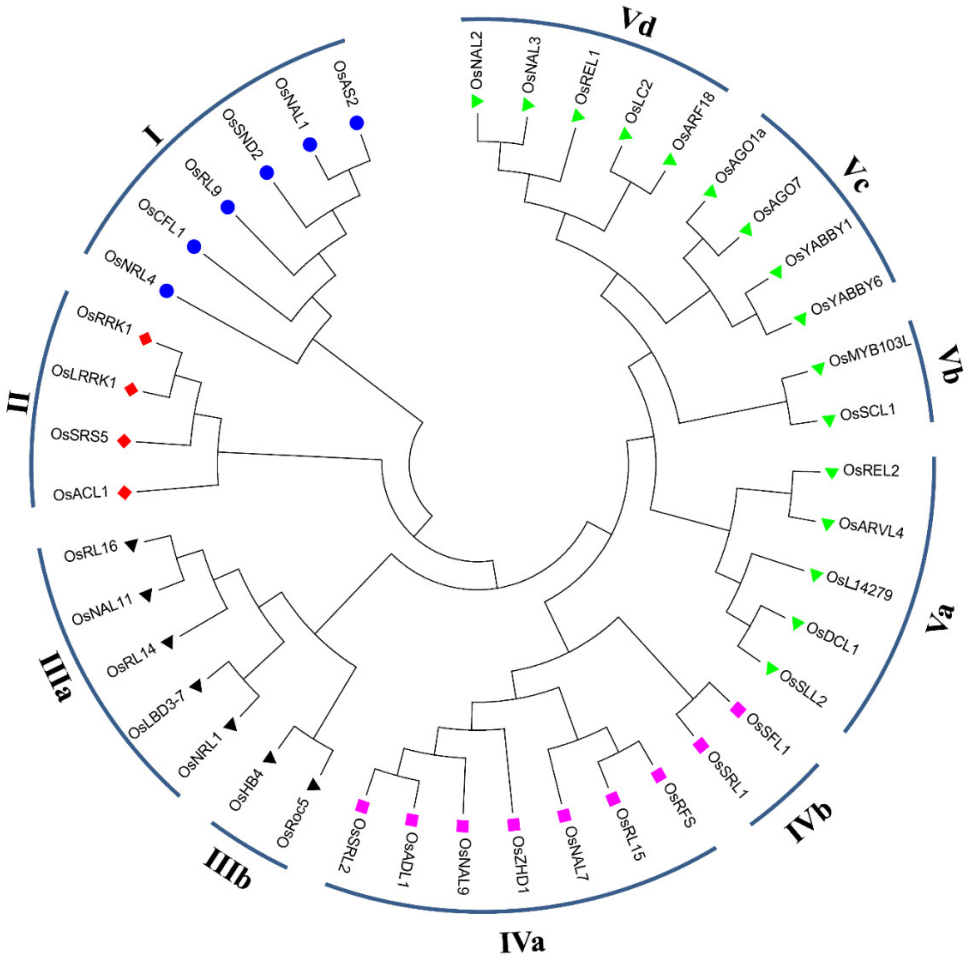


Figure 6.5: Phylogenetic tree of 42 rolling leaf (RL) genes of interest identified in this study. The tree was constructed based on multiple aligned sequences by maximum likelihood (ML) method with bootstrap of 1000 in MEGA6. Multiple sequence alignment was performed using ClustalW program in MEGA6. The colored shapes indicates different clusters of RL proteins. The roman numerals I-Vd indicates groups and subgroups of RL genes.

6.3.5 Gene Ontology, Transcription Factors and KEGG Analysis of RL Genes

In order to better understand the functional involvement and to characterize the selected RL genes, different types of gene enrichment analyses were conducted. The GO analysis was performed to explain the biological importance of RL genes. We used the "GO analysis toolkit" of agriGO v2.0 database (http://systemsbiology.cau.edu.cn/agriGOv2/) with the default setting for performing the GO analysis. Figure 6.6 represents the heatmap of "GO term versus gene" along with the bar plot of –log10(*P*) against all significant GO terms ($P < 0.05$). All of the significant GO terms ($P < 0.05$) of RL genes and their description, the number in the input list, and *P*-value are shown in Table 6.3 and Table A6.2. From the analysis result, it was observed that 38 GO terms were significantly enriched for biological process whereas very few significant overrepresentations were found for cellular component and molecular function (Figure 6.6 and Table A6.2). From Figure 6.6, we found that the 42 RL genes clustered into three groups (Group I, Group II and Group III) based on the enriched GO terms. The genes in three groups were enriched in various biological process. The genes in group I were mostly involved in different types of regulations including regulation of transcription (GO:0045449), regulation of nitrogen compound metabolic process (GO:0051171), regulation of cellular biosynthetic process (GO:0031326), regulation of biosynthetic process (GO:0009889), regulation of primary metabolic process (GO: 0080090), and so on. The genes of this group were also involved in response to abiotic stimulus (GO:0009628) and various biosynthetic process (GO:0044249 - cellular biosynthetic process, GO:0034645 - cellular macromolecule biosynthetic process, GO:0009059 - macromolecule biosynthetic process). This observation indicated that the genes in group I are more effective and biologically more functional. The genes in group II were mainly involved in the biological process (GO:0008150) whereas genes in group III were mostly involved in multicellular organismal development (GO:0007275), cell differentiation (GO:0030154), primary metabolic process (GO:0044238) and macromolecule metabolic process (GO:0043170). The biological process "multicellular organismal development" (GO:0007275; $P < 4.92E\text{-}10$) was the most

representative process that enriched for a large number of genes belonging to group I and group III.

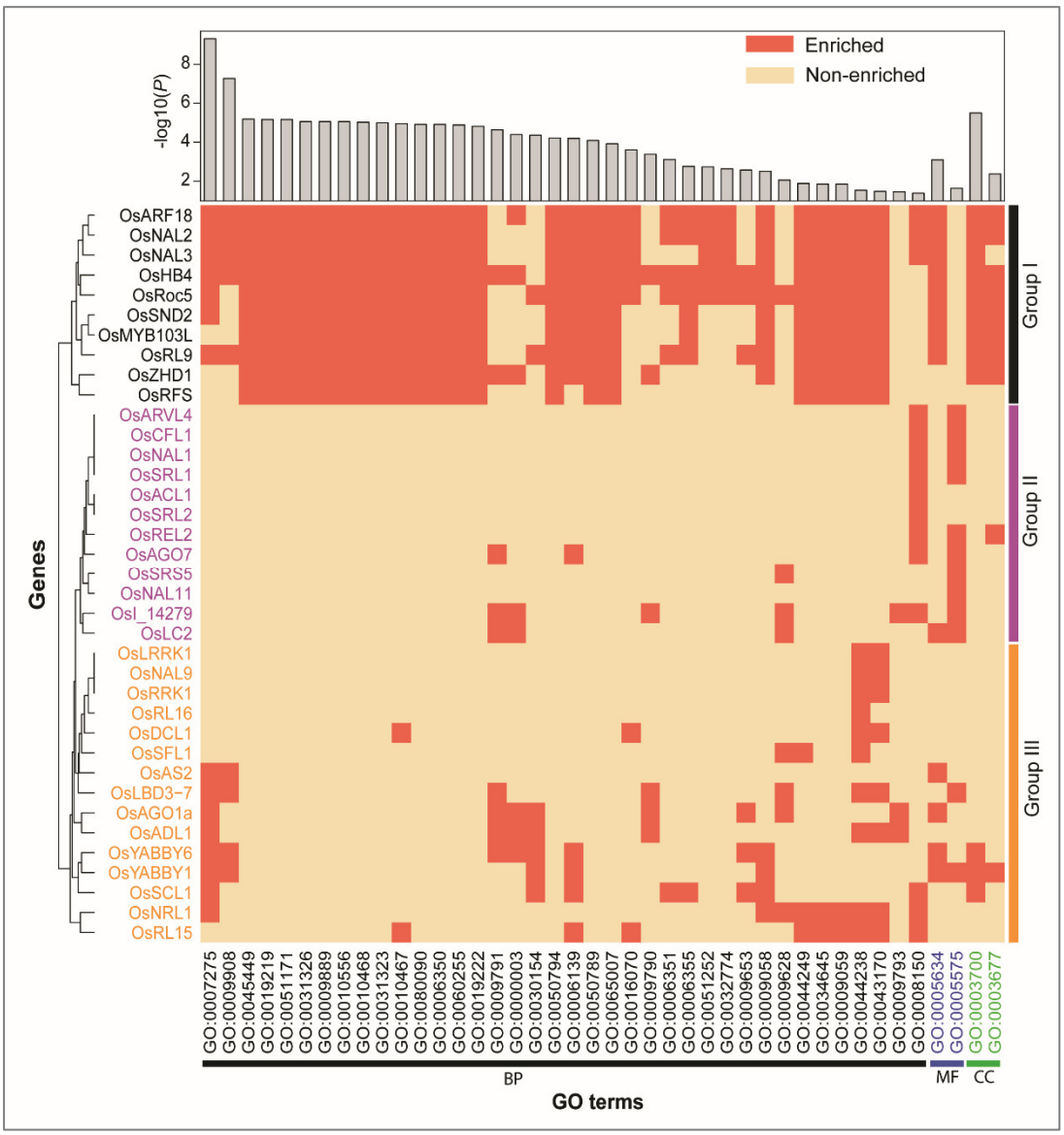


Figure 6.6: The enriched GO terms for all RL genes. The GO terms indicated by BP are involved in the biological process, the GO terms indicated by MF are involved in molecular function and the GO terms indicated by CC are involved in the cellular component.

Table 6.3: The enriched GO terms for all rolling leaf genes identified in this study

GO term	Ontology	Description	Number in input list	P-value
GO:0007275	BP	multicellular organismal development	15	4.92E-10
GO:0009908	BP	flower development	9	5.4E-08
GO:0045449	BP	regulation of transcription	10	6.5E-06
GO:0019219	BP	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	10	6.8E-06
GO:0051171	BP	regulation of nitrogen compound metabolic process	10	6.8E-06
GO:0031326	BP	regulation of cellular biosynthetic process	10	8.7E-06
GO:0009889	BP	regulation of biosynthetic process	10	8.7E-06
GO:0010556	BP	regulation of macromolecule biosynthetic process	10	8.7E-06
GO:0010468	BP	regulation of gene expression	10	9.4E-06
GO:0031323	BP	regulation of cellular metabolic process	10	0.00001
GO:0010467	BP	gene expression	12	1.1E-05
GO:0080090	BP	regulation of primary metabolic process	10	1.2E-05
GO:0006350	BP	transcription	10	1.2E-05
GO:0060255	BP	regulation of macromolecule metabolic process	10	1.3E-05
GO:0019222	BP	regulation of metabolic process	10	1.5E-05
GO:0009791	BP	post-embryonic development	9	2.29E-05
GO:0000003	BP	reproduction	8	4.05E-05
GO:0030154	BP	cell differentiation	7	4.37E-05
GO:0050794	BP	regulation of cellular process	10	6E-05
GO:0006139	BP	nucleobase-containing compound (nucleobase, nucleoside, nucleotide and nucleic acid) metabolic process	14	6.33E-05
GO:0050789	BP	regulation of biological process	10	8.1E-05
GO:0065007	BP	biological regulation	10	0.00012
GO:0016070	BP	RNA metabolic process	7	0.00025
GO:0009790	BP	embryo development	6	0.00041
GO:0006351	BP	transcription, DNA-templated	6	0.00076
GO:0006355	BP	regulation of transcription, DNA-templated	8	0.00172
GO:0051252	BP	regulation of RNA metabolic process	5	0.0018
GO:0032774	BP	RNA biosynthetic process	5	0.0023

GO term	Ontology	Description	Number in input list	P-value
GO:0009653	BP	anatomical structure morphogenesis	6	0.00265
GO:0009058	BP	biosynthetic process	13	0.00309
GO:0009628	BP	response to abiotic stimulus	8	0.00866
GO:0044249	BP	cellular biosynthetic process	13	0.013
GO:0034645	BP	cellular macromolecule biosynthetic process	12	0.014
GO:0009059	BP	macromolecule biosynthetic process	12	0.014
GO:0044238	BP	primary metabolic process	20	0.029
GO:0043170	BP	macromolecule metabolic process	18	0.033
GO:0009793	BP	embryo development ending in seed dormancy	3	0.034
GO:0008150	BP	biological_process	15	0.04107
GO:0005634	CC	nucleus	13	0.0008
GO:0005575	CC	cellular_component	12	0.0225
GO:0003700	MF	sequence-specific DNA binding transcription factor activity	12	3.16E-06
GO:0003677	MF	DNA binding	10	0.00424

GO: Gene ontology, BP: Biological process, CC: Cellular component, MF: molecular function.

Transcription factor analysis showed that a total of 13 genes were related to 10 different TFs families (Table 6.4). Gene sets (*OsRoc5*, *OsHB4*), (*OsNAL3*, *OsNAL2*) and (*OsYABBY1*, *OsYABBY6*) belonged to the TF families HD-ZIP, WOX, and YABBY, respectively. Genes *OsARF18*, *OsRL9*, *OsSCL1*, *OsAS2*, *OsMYB103L*, *OsSND2* and *OsZHD1* were the members of the TF families ARF, G2-like, GRAS, LBD (LOB DOMAIN), MYB, NAC and ZF-HD, respectively.

Table 6.4: Summary of Transcription factors (TFs) identified in this study

Genes Name	Transcription Factors	Description	Features
OsARF18	ARF	Auxin response factors	ARF are transcription factors that regulate the expression of auxin response genes
OsRL9	G2-like	Golden2-like	GLK proteins are members of the newly classified GARP superfamily of Transcription factors

Genes Name	Transcription Factors	Description	Features
OsSCL1	GRAS	GRAS	The GRAS family of putative transcriptional regulators is found throughout the plant kingdom, and these proteins have diverse roles in plant development, including root development, axillary shoot development, and maintenance of the shoot apical meristem
OsRoc5, *OsHB4*	HD-ZIP	Homeodomain leucine Zipper	HD-ZIP gene family are vital regulators of plant development
OsAS2	LBD	LOB DOMAIN	the LBD genes encode a novel class of DNA-binding transcription factors
OsMYB103L	MYB	MYB	The encoded proteins are crucial to the control of proliferation and differentiation in a number of cell types, and share the conserved MYB DNA-binding domain
OsSND2	NAC	NAM, ATAF, and CUC	NAC transcription factors are involved in various aspects of plant development
OsNAL3, *OsNAL2*	WOX	WUS homeobox-containing	WOX family members fulfill specialized functions in key developmental processes in plants, such as embryonic patterning, stem-cell maintenance and organ formation
OsYABBY1, OsYABBY6	YABBY	YABBY	The YABBY gene are expressed in a polar manner in all lateral organs produced by apical and flower meristems
OsZHD1	ZF-HD	Zinc Finger Homeodomain protein	ZF-HD proteins involved in the mesophyll-specific expression of the C4 and C3 plants

The KEGG pathway analysis was carried out using the KEGG database to know the biological process in the RL genes of interest (Figure 6.7). Results showed that only 14 out of 42 RL genes were involved in the KEGG pathways (Figure 6.7b). Of those 14 genes 50% genes were involved in metabolism pathway, followed by cellular processes (36%) and 7% genes were involved in both genetic information processing and organismal system in this study (Figure 6.7a). KEGG pathway enrichment analysis (Figure 6.7b) also showed that three RL genes functioned in the metabolic pathways; two genes functioned in the cell growth and death, cellular community-eukaryotes pathways; and one gene functioned in the pathways: biosynthesis of secondary metabolites, carbohydrate metabolism, lipid metabolism, amino acid metabolism, translation, transport and catabolism, and aging. Table 6.5 represents the KEGG orthologous (KO) of the RL genes of interest along with their description.

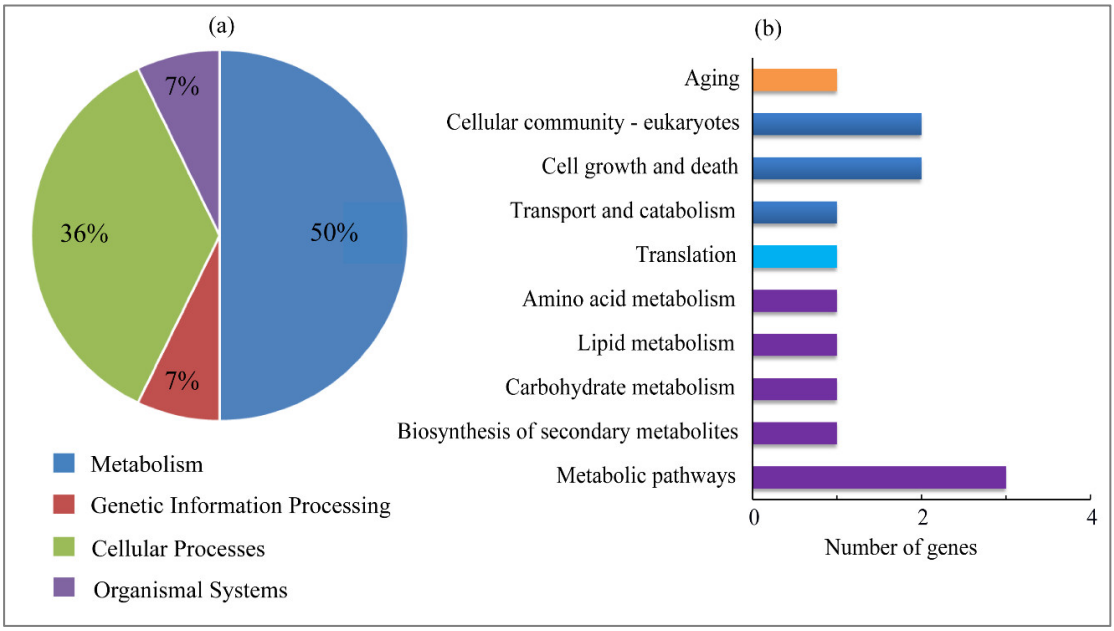


Figure 6.7: KEGG analysis results. (a) Pie chart for all allocated KEGG pathways of all RL genes. (b) Bar chart for all significant KEGG pathways of all RL genes. X-axis represents the number of genes. Y-axis represent second KEGG pathway terms. The second pathway terms are grouped and indicated by different color.

Table 6.5: Identified rolling leaf genes KEGG orthologous (KO) and their description

Gene	KO	Definition
OsAGO1a	K11593	ELF2C; eukaryotic translation initiation factor 2C
OsNAL7	K11816	YUCCA; indole-3-pyruvate monooxygenase [EC:1.14.13.168]
OsNAL9	K01358	clpP; ATP-dependent Clp protease, protease subunit [EC:3.4.21.92]
OsNAL11	K09539	DNAJC19; DnaJ homolog subfamily C member 19
OsNRL1	K00770	E2.4.2.24; 1,4-beta-D-xylan synthase [EC:2.4.2.24]
OsAGO7	K11593	ELF2C; eukaryotic translation initiation factor 2C
OsLBD3-7	K01301	NAALAD; N-acetylated-alpha-linked acidic dipeptidase [EC:3.4.17.21]
OsMYB103L	K09422	MYBP; transcription factor MYB, plant
OsRoc5	K09338	HD-ZIP; homeobox-leucine zipper protein
OsSFL1	K15397	KCS; 3-ketoacyl-CoA synthase [EC:2.3.1.199]
OsSRL2	K21842	EFR3; protein EFR3
OsSRS5	K07374	TUBA; tubulin alpha
OsHB4	K09338	HD-ZIP; homeobox-leucine zipper protein
OsRL15	K01875	SARS; seryl-tRNA synthetase [EC:6.1.1.11]

6.3.6 Network Analysis of RL Related Genes

Network analysis (Figure 6.8) showed that there is no protein-protein interaction for the 23 genes *OsREL2*, *OsNAL2*, *OsYABBY1*, *OsLRRK1*, *OsACL1*, *OsARVL4*, *OsLC2*, *OsREL1*, *OsSLL2*, *OsARF18*, *OsRL16*, *OsNRL4*, *OsNAL3*, *OsRL15*, *OsZHD1*, *OsSRL1*, *OsNAL1*, *OsI_14279*, *OsRFS*, *OsRRK1*, *OsRL14*, *OsNAL11* and *OsSFL1*. The RL gene *OsRL9* is associated with the genes *OsRoc5*, *OsADL1*, *OsNAL7*, *OsAS2*, *OsNRL1*, *OsYABBY6*, *OsHB4* and *OsAGO7*. The RL genes within each of the pairs (*OsMYB103L* and *OsSND2*), (*OsSRS5* and *OsNAL9*), (*OsRoc5* and *OsCFL1*), (*OsRoc5* and *OsADL1*), (*OsADL1* and *OsAGO7*), (*OsLBD3-7* and *OsAGO1a*), (*OsSCL1* and *OsDCL1*), and (*OsSRL2* and *OsYABBY6*) are associated with each other. Genes *OsAGO1a*, *OsAGO7* and *OsDCL1* are associated with each other. Network analysis predicted the association of *OsRL9* with *OsRoc5*, *OsADL1*, *OsNAL7*, *OsAS2*, *OsNRL1*, *OsYABBY6*, *OsHB4* and *OsAGO7*; and association of *OsRoc5* with *OsADL1* by text mining. The interaction between *OsRoc5* and *OsCFL1* was experimentally determined as well as predicted by text mining. The interaction between the RL genes within each of the pairs (*OsMYB103L* and *OsSND2*) and (*OsSRS5* and *OsNAL9*) was predicted by text mining and co-expression analysis. Also, the association between *OsLBD3-7* and *OsAGO1a* was experimentally determined as well as predicted by co-expression analysis. The RL genes *OsAGO1a* and *OsAGO7* exhibited association which was known from curated databases as well as predicted by gene text mining and protein homology. The association of *OsAGO1a* with *OsDCL1* and *OsAGO7* with *OsDCL1* was experimentally determined and known from curated databases, and also predicted by text mining and co-expression analysis.

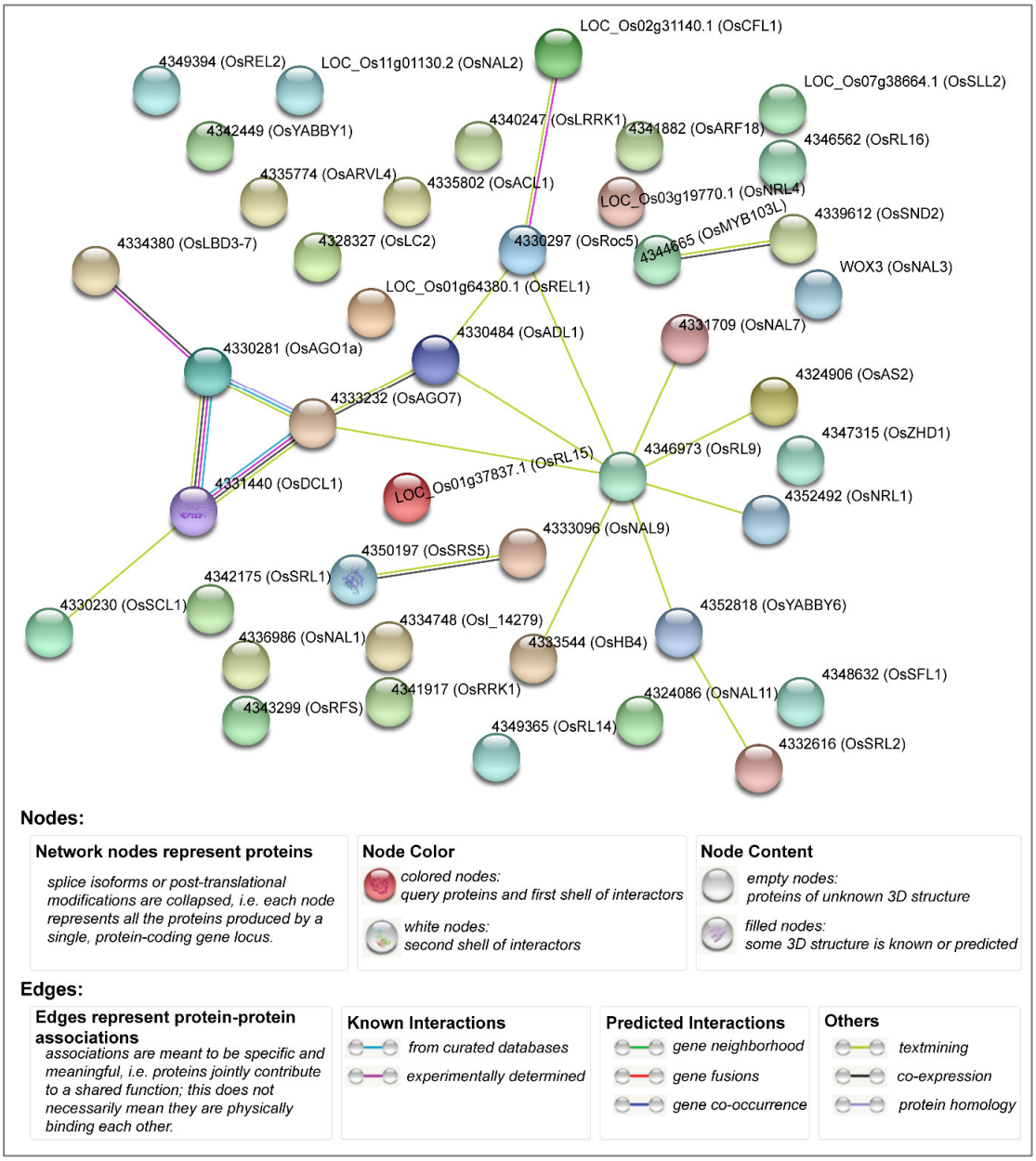


Figure 6.8: Protein-protein interaction network of 42 rolling leaf genes of interest. The different methods of prediction of interactions have been indicated with the different colored connective lines. Protein-protein interaction network was built using the web-based tool STRING V10.5 (http://string-db.org/).

6.3.7 Gene Expression Analysis of RL Related Genes

The analysis of gene expression data of 42 RL genes of interest has been presented in Figure A6.2 – Figure A6.23. The line plots of 42 RL genes at different tissues show that there is no similarity in the pattern of line charts for different genes (Figure A6.2

– Figure A6.11). The box plots of the expression values at different tissues shows that almost all of the RL genes have some extreme (i.e., very high or very low) expression values at all of the three tissues leaf, root and shoot (Figure A6.13 – Figure A6.23) except for *OsAGO1a*, *OsLC2*, *OsNAL1*, *OsNAL9*, *OsNRL4*, *OsRL14*, *OsRRK1*, *OsSFL1* and *OsRL15* (Figure A6.21 – Figure A6.23). Genes *OsAGO1a*, *OsLC2* and *OsNAL9* (Figure A6.21) have some extreme expression values at the root and shoot but not at the leaf. Genes *OsNAL1*, *OsRL15*, *OsRRK1* and *OsSFL1* (Figure A6.22) have some extreme expression values at leaf and shoot but not at the root. Genes *OsNRL4* and *OsRL14* (Figure A6.23) have some extreme expression values at leaf and root but not at the shoot. These results indicate that the gene expression at leaf, shoot and root might be controlling the leaf rolling in rice.

6.4 Discussion

The ultimate goal of rice breeding is the development of rice mutants with super-high yield and stress tolerance traits. In rice, appropriate leaf rolling is considered to be an important agronomic element, particularly, a moderate leaf rolling is regarded as a crucial phenotypic trait of the ideal rice plant as it is important for increasing grain yield (Lang et al., 2004; Wu, 2009; Yuan, 1997; Zhang et al., 2009). However, to the best of our knowledge, there is no study where all the RL genes/QTLs and their genomic information have been put together along with their genome-wide comparative analysis from different bioinformatics point of view. In this study, we listed up 103 RL genes/QTLs along with their genomic information reported in various studies till date and performed various comparative analyses from different bioinformatics stand points with the RL genes of interest.

We selected 42 genes as the genes of interest in our analysis (Table 6.2) among the total 103 identified RL genes/QTLs (Table 6.1) due to the availability of their locus IDs in this study. Maximum number of RL genes were found on chromosome 3 and minimum genes were found on chromosomes 5 and 8 among the 42 RL genes. The RL gene *OsADL1* has the longest genomic length and *OsREL1* has the shortest genomic length. *OsSLL2* has the lowest value of CDS and protein length while the

gene *OsRFS* has the highest value of CDS and protein length. *OsRFS* has the highest molecular weight and *OsSLL2* has the lowest molecular weight. The gene *OsREL1* corresponds to the minimum value of pI whereas the gene *OsNRL4* exhibits the maximum value of pI. The comparative gene structure analysis revealed that some RL genes contained only one exon (i.e., do not contain any intron), some genes did not have upstream, and some genes did not contain any upstream and downstream (Figure 6.3). Most of the genes (60%) contained 0 to 3 introns. Although some genes contained an equal number of exons and introns the length of exons and introns were different from each other indicating structural diversity. From the gene structure analysis (Figure 6.3) it can be inferred that RL genes are diverse in their structures from each other in terms of conserveness and structure of exons and introns.

Domain analysis revealed that very few of the RL genes contain the domains of the same family. Both the RL genes *OsNAL2* and *OsNAL3* contained only one and identical domain named "Homeobox domain" (Figure 6.4 and Table A6.1). According to Cho et al. (2013), *NAL2* and *NAL3* were paralogs and encoded an identical homeobox 3A (*OsWOX3A*) protein which supported our findings. Both the genes *OsYABBY1* and *OsYABBY6* contained only one domain (YABBY domain) which is identical in both genes (Figure 6.4 and Table A6.1). According to Toriba et al. (2007), both *OsYABBY1* and *OsYABBY6* belonged to YABBY family which is in favor of our findings. Most of the RL genes contained the domains of various families indicating the diversity of RL proteins. This study showed that RL genes were diverse in terms of domains they contained and their domain structure. Based on the phylogenetic analysis we grouped the RL genes into five major groups. Again groups III, IV and V were divided into two, two and four subgroups, respectively. We found that both *OsNAL2* and *OsNAL3* belonged to same subgroup Vd, and *OsYABBY1* and *OsYABBY6* belonged to the same subgroup Vc. These findings are supported by Cho et al. (2013) and Toriba et al. (2007).

To explore functional similarity and characterize a set of genes, various enrichment analyses, such as GO and KEGG analyses, were performed along with heatmap of enriched GO terms. The biological process "multicellular organismal development"

(GO:0007275; P < 4.92E-10) was found as the most representative process that enriched for a large number of genes belongs to group I and group III. The GO term GO:0007275 (multicellular organismal development) was found as the most significant. From the heatmap of the GO terms against gene, three distinct clusters of genes were observed and the genes belonged to different clusters were involved in different biological processes. Transcription factors have significant function for controlling plant propagation, maturation, and to react to unfavorable situation condition including drought, chill, salinity, and high temperature (Zhang et al., 2013). For example, myb-protein encoding genes have been reported to function as regulators of cell differentiation (Kirik and Bäumlein, 1996) and NAC TFs involved in various mechanisms including developmental process, flower creation, seed growth, hormone signaling, responses to abiotic stress and aging in plants (Nuruzzaman et al., 2010). Transcription factor analysis showed that 10 different TFs families (such as HD-ZIP, WOX, ARF, G2-like, GRAS, LBD, MYB, NAC and ZF-HD) involved in leaf rolling, where all of these TFs were responsible for different aspect of plant developments (Table 6.4).

Metabolomics plays a significant role in fundamental plant biology and applied biotechnology. Several current studies have found that the analysis of the metabolite composition of mutants can assist the assignment of functions to genes (Schauer and Fernie, 2006). Results of KEGG pathway analysis showed that half of the RL genes were involved in metabolism pathway including biosynthesis of secondary metabolites, carbohydrate metabolism, lipid metabolism, amino acid metabolism and metabolic pathways (Figure 6.7). According to Li et al. (2017), the proteins responsible for leaf rolling in rice are involved in different pathways including biosynthesis of phenylpropanoid, phenylalanine metabolism, and sucrose and starch metabolism. According to Chen et al. (2018) RL proteins in Brassica napus are involved in different pathways including “metabolic pathways”, “biosynthesis of secondary metabolites” and “starch and sucrose metabolism”. So the results of previous studies of rolling leaf support our finding. The results in this study indicate that these GO, TFs and KEGG pathways might be involved in rolling leaf development by controlling transcriptional regulation of a variety of biological

processes related gene expression, transcription and biological regulation, metabolic process and cellular process.

The network analysis was used for investigating the association among the 42 RL genes. Results showed that the RL gene *OsRL9* was associated with a maximum number of genes (9 genes: *OsRoc5*, *OsADL1*, *OsNAL7*, *OsAS2*, *OsNRL1*, *OsYABBY6*, *OsHB4* and *OsAGO7*). We found that the genes *OsAGO1a*, *OsAGO7* and *OsDCL1* were associated with each other. Protein homology and text mining predicted the association between *OsAGO1a* and *OsAGO7* which was also known from the curated databases. All the four methods (experimentally determination, determination from curated databases, prediction using text mining and prediction by co-expression analysis) identified the association between the genes of each of the pairs of RL genes (*OsAGO1a* and *OsDCL1*) and (*OsAGO7* and *OsDCL1*). Gene *OsRoc5* was associated with *OsCFL1* and *OsADL1*; *OsMYB103L* was associated with *OsSND2*; *OsLBD3-7* was associated with *OsAGO1a*; and *OsSRS5* was associated with *OsNAL9*. We found 23 RL genes (*OsREL2*, *OsNAL2*, *OsYABBY1*, *OsLRRK1*, *OsACL1*, *OsARVL4*, *OsLC2*, *OsREL1*, *OsSLL2*, *OsARF18*, *OsRL16*, *OsNRL4*, *OsNAL3*, *OsRL15*, *OsZHD1*, *OsSRL1*, *OsNAL1*, *OsI_14279*, *OsRFS*, *OsRRK1*, *OsRL14*, *OsNAL11* and *OsSFL1*) which were not associated with any other RL genes. This indicates that more than 50% of the RL genes of interest are not associated with each other.

The line charts of the expression values of 42 RL genes at different tissues showed no similarity in the pattern of lines for different genes (Figure A6.2 – Figure A6.11). From the box plots of the expression values of 42 LR genes at different tissues, we found that most of the RL genes have some extreme (very high or low) expression values at leaf, root and shoot (Figure A6.13 – Figure A6.23). This indicates that the gene expression at leaf, root and shoot may be interactively responsible for controlling leaf rolling in rice. It has been demonstrated by several previous studies of leaf mutants with altered leaf morphology that appropriate balance in cells in the shoot apical meristem is very important for normal development of leaf (Luo et al., 2007). This implies that shoot has a crucial role in developing leaf shape which supports our findings.

Altogether, this is the first study where we have studied and put the genetic information of all identified RL genes/QTLs till now and their genomic information along with their genome-wide comparative analysis from different bioinformatics point of view. This study will enable the rice breeders and researchers to get collective information of RL genes/QTLs along with their comparative results. As a result rice breeders will be able to develop super-high-yield rice mutant with moderate leaf rolling and desired architecture. However, further advancement of study is required to explore the complicated process of LR in rice.

6.5 Conclusion

Our present study has identified 103 RL genes/QTLs in the genome of rice plant characterized through several studies to date. Among 103 RL genes/QTLs, 42 genes, for which locus IDs were available, were finally selected in our analysis. This study provides a comparative analysis of the selected 42 RL genes from the various points of bioinformatics view using different bioinformatics techniques including gene structure, conserved domain, phylogenetic, gene expression and protein-protein interaction network analysis. Gene structure analysis shows that the selected RL genes are diverse in structure in terms intron-exon. Domain analysis reveals that the RL genes contain different types of domains except for some genes. Phylogenetic analysis has clustered the RL genes into five major groups with group III, IV and V divided into some subgroups. GO analysis indicates that a total of 42 significant GO terms enriched, TF analysis shows 10 different TFs families and KEGG analysis shows that 14 genes are involved in the KEGG pathways from all RL genes. Protein-protein interaction network shows that 23 RL genes are not associated with any other RL genes and gene *OsRL9* is associated with maximum genes. Gene expression analysis reveals that the expression patterns of RL genes are different and RL genes exhibit some extreme expression at the leaf, shoot and root. Therefore, we may conclude that the RL genes have different types of function for controlling RL in rice. These results might provide important information regarding gene structure, conserved domain information, phylogenetic revolution, gene enrichment, TF families, KEGG pathways, protein-protein interactions, gene expression pattern, and others genetic basic of RL

related rice genes which will be helpful for other researchers to make a quick decision about these genes and to explore new gene's characteristics for rice genetics research.

Chapter 7

Conclusions and Areas of Future Research

7.1 Conclusions

Genomics is one of the most important OMICS research wings for bioinformatics. In Genomics, Genome-wide Association Studies (GWAS) have evolved over the last ten years into a powerful tool for investigating the genetic architecture of plant science, animal science and human biology. The advent of new technologies for extracting genetic information from tissue samples has increased the availability of suitable data for finding genes controlling complex traits in plants, animals and humans. There are different types of GWAS based on the nature of the genomic data such as (i) Quantitative trait locus (QTL) mapping based GWAS, (ii) Single nucleotide polymorphism (SNP) based GWAS, (iii) Expression QTL (eQTL) mapping based GWAS and (iv) Sequence based GWAS. Again, GWAS can be divided into two types based on the number of phenotypes considered in the analysis, GWAS can be divided into two types: (i) Single-trait GWAS and (ii) Multi-trait GWAS. There are various statistical methods for the analysis of GWAS data including maximum likelihood (ML) and least squares (LS) based single-trait and multi-trait GWAS techniques. However, some methods of GWAS are very time consuming and very sensitive to phenotypic contaminations. Thus efficient methods, in terms of computation time and robustness against phenotypic outliers, are demanded for the analysis of GWAS data. In this thesis, we have proposed some less time consuming and robust methods for GWAS that outperform over the existing methods.

In **Chapter 1**, we have provided a brief introduction about the basic concepts of genomics and discussed different important terminologies which are related to genomics, more specifically, GWAS. We have also discussed the genome-wide association studies (GWAS) and its various types. A broad review of the statistical methods for different types of GWAS has also been discussed, along with their advantages and limitations, in **Chapter 1**.

Quantitative trait locus (QTL) analysis relies on statistical methods to interpret genetic data in the presence of phenotype data and possibly other factors such as environmental factors. The goal is to both detect the presence of QTL with significant effects on trait value as well as to estimate their locations on the genome relative to those of known markers. Maximum likelihood (ML) based simple interval mapping (SIM) is the most popular and widely used method for single-trait QTL analysis. However, ML based SIM is very time consuming because it uses expectation maximization (EM) algorithm and also calculations are very complex in this method. Although least squares (LS) regression based SIM overcome the problem of computation time, its calculations are also complex. In **Chapter 2**, to overcome calculation complexity of QTL analysis, we have discussed a regression based new SIM approach for single-trait QTL analysis with BC by estimating the model parameters using the properties of bivariate normal distribution. Our proposed method of single-trait QTL is very straight forward and shows almost same performance as the existing methods (LS and ML based SIM) of single-trait QTL analysis. Simulation study and real data analysis results show that our new proposed approach has almost similar performance to the existing interval mapping approaches. Although our proposed SIM method of single-trait QTL analysis is straight forward in terms of computational complexity, like the existing methods this approach is not robust against phenotypic contamination and produce misleading results when the phenotypic data are contaminated by outliers. To overcome this problem, we have also developed a robust approach for single-trait QTL analysis with BC population by robustifying our newly developed single-trait QTL mapping approach using minimum β–divergence method in this **Chapter 2**. The proposed robust method reduces to the classical approaches when $\beta \to 0$. The tuning parameter β controls the performance of

the proposed robust method. The simulation study and real data analysis results show that our proposed robust method improves performance over the classical SIM approaches (including our new SIM approach) in the case of data contaminations; otherwise, it shows almost the same results as the classical SIM approaches.

In many line crossing experiments of genome-wide QTL mapping studies, measurements are taken on multiple traits along with the marker genotypes. Usually, such traits are correlated with each other and there are common chromosomal regions (or chromosomal locations) that affect multiple traits. Single-trait QTL analysis cannot deal with the pleiotropic problem and trait-trait relationship. This problem can be overcome using multi-trait QTL analysis which consider all the traits of interest simultaneously in the model. Maximum likelihood (ML) and least squares (LS) based multi-trait SIM are two most popular and widely used methods for multi-trait QTL analysis. However, both the methods are time consuming and include computation complexity. In **chapter 3**, we have developed a fast multi-trait QTL mapping method based on the assumption that the phenotypes and the condition probabilities of QTL genotype given flanking marker genotypes have joint multivariate normal distribution. Simulation study and real data analysis results shows that our proposed method performs same as the existing multi-trait QTL mapping methods but it takes very less computation time compared to the other existing methods. So our proposed fast multi-trait QTL mapping approach is very efficient in terms of computation time.

Although our proposed method is faster than the other existing methods of multi-trait QTL analysis, like the existing method this proposed method is very sensitive to phenotypic outliers and produces misleading results when the data are contaminated by outliers. In **chapter 4**, we have discussed a robust technique of multi-trait QTL analysis which is the robustification of fast multi-trait QTL analysis approach (discussed in **Chapter 3**). Simulation study and real data analysis show that only our proposed robust method of multi-trait QTL analysis is able to identify all the QTL positions in presence of outliers as identified in absence of outliers. Otherwise, our proposed method perform almost similar to the traditional methods. So, we can conclude that our proposed robust method of multi-trait QTL analysis outperform

over the classical methods (including the fast multi-trait QTL mapping approach discussed in **Chapter 3**) in presence of outliers. We have also implemented the proposed robust method for eSNPs analysis and found that our proposed robust approach outperform over the existing classical approach. Otherwise, the proposed approach shows similar performance.

Nowadays, due to the recent advancement in the NGS technologies, SNP data of complete genome become very easy obtained for GWAS by decreasing the cost and time required to obtain sequences of whole genome. There are several methods for SNP based GWAS ranging from simple single-trait approaches to complex multi-trait approaches designed specifically for multi-trait GWAS. One main problem in SNP based GWAS is that all the existing methods of SNP based GWAS suffer from the data contamination problems and provide misleading results when the phenotypic data are contaminated by extreme observations. We have developed a regression based robust approach for SNP based GWAS data analysis in **Chapter 5**. Simulation study and real data analysis results show that our proposed method outperforms over the existing methods in presence of outliers. Otherwise, the proposed approach shows similar performance as like as the existing traditional approaches of SNP based GWAS analysis.

In **Chapter 6**, we have discussed sequence based GWAS for all rolling leaf (RL) genes in rice (*Oryza sativa* L.) reported till date in literatures. We found that most of the RL genes are different in structure and contain different conserved domain. We have grouped all the selected 42 RL genes into five major groups through phylogenetic analysis. From gene ontology analysis we found that most of RL genes enriched in biological process. From gene network analysis we observe that most of the genes are not associated with each other. From exploratory gene expression analysis of 42 RL genes, we have found that the expression patterns of RL genes are different and RL genes exhibit some extreme expression at the leaf, shoot and root. Therefore, we may conclude that the RL genes have different types of function for controlling RL in rice. These results might provide important information regarding gene structure, conserved domain information, phylogenetic revolution, gene

enrichment, TF families, KEGG pathways, protein-protein interactions, gene expression pattern, and others genetic basic of RL related rice genes which will be helpful for other researchers to make a quick decision about these genes and to explore new gene's characteristics for rice genetics research.

7.2 Areas of Further Research

Different types of genome wide data are being continuously generated by the bioinformatics and biotechnological researchers. Appropriate computational methods or approaches are needed to properly identify the important genes/QTLs/SNPs genome wide which are responsible for a particular traits (e.g., blood pressure, grain yield, etc.). The following are the future research areas related to GWAS data analysis:

1. Develop fast multi-trait QTL analysis method for composite interval mapping (CIM).

2. Develop robust regression based multi-trait QTL analysis technique for CIM.

3. Extend the SNP based single-trait robust GWAS to multi-trait robust GWAS including other confounding factors in the model.

Bibliography

Adedze, Y.M.N., Feng, B., Shi, L., Sheng, Z., Tang, S., Wei, X., et al. (2018). Further insight into the role of KAN1, a member of KANADI transcription factor family in rice. *Plant Growth Regulation*, 84(2), 237-248. DOI:10.1007/s10725-017-0335-7.

Adedze, Y.M.N., Wei, X.J., Sheng, Z.H., Jiao, G.A., Tang, S.Q., and Hu, P.S. (2017). Characterization of a rice *dwarf and narrow leaf 2* mutant. *Biologia Plantarum*, 61(1), 85-94. DOI:10.1007/s10535-016-0632-4.

Alam, M.J., Alamin, M., Hossain, M.R., Islam, S.S., and Mollah, M.N.H. (2018). Robust linear regression based simple interval mapping for QTL analysis with backcross population. *Journal of Bio-Science*, 24, 75-81. DOI:10.3329/jbs.v24i0.37489.

Alam, M.J., Alamin, M., Sultana, M.H., Amanullah, M., and Mollah, M.N.H. (2016). Regression Based Robust QTL Analysis for F_2 Population. *Rajshahi University Journal of Science and Engineering*, 44, 95-99. DOI:10.3329/rujse.v44i0.30401.

Alamin, M., Zeng, D.-D., Qin, R., Sultana, M.H., Jin, X.-L., and Shi, C.-H. (2017). Characterization and fine mapping of *SFL1*, a gene controlling *screw flag leaf* in rice. *Plant Molecular Biology Reporter*, 35(5), 491-503. DOI:10.1007/s11105-017-1039-x.

Almasy, L., and Blangero, J. (1998). Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees. *The American Journal of Human Genetics*, 62(5), 1198-1211. DOI:10.1086/301844.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*. DOI:10.1038/npre.2010.4282.2.

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A., and Rai, A. (2016). Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. *Journal of Computational Biology*, 23(4), 239-247. DOI:10.1089/cmb.2015.0205.

Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., et al. (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS genetics*, 1(5), e60-e60. DOI:10.1371/journal.pgen.0010060.

Atkin, M., Anderson, D., Francis, B., and Hinde, J. (1989). *Statistical modelling in GLIM*: Oxford University Press, Oxford.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298), 627-631. DOI:10.1038/nature08800.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 53-59. DOI:10.1038/nature07517.

Bernardo, R. (2001). What If We Knew All the Genes for a Quantitative Trait in Hybrid Crops? *Crop Science*, 41(1), 1-4. DOI:10.2135/cropsci2001.4111.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., et al. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6), 463-471. DOI:10.3109/03014460.2013.807878.

Bolormaa, S., Hayes, B.J., Savin, K., Hawken, R., Barendse, W., Arthur, P.F., et al. (2011). Genome-wide association studies for feedlot and growth traits in cattle1. *Journal of Animal Science*, 89(6), 1684-1697. DOI:10.2527/jas.2010-3079.

Bonferroni, C.E., Bonferroni, C., and Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita'. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633-2635. DOI:10.1093/bioinformatics/btm308.

Broman, K.W., Wu, H., Sen, Ś., and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7), 889-890.

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., et al. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, 37(8), 868-872. DOI:10.1038/ng1607.

Chen, Q., Xie, Q., Gao, J., Wang, W., Sun, B., Liu, B., et al. (2015). Characterization of *Rolled and Erect Leaf 1* in regulating leave morphology in rice. *Journal of experimental botany*, 66(19), 6047-6058. DOI:10.1093/jxb/erv319.

Chen, W., Wan, S., Shen, L., Zhou, Y., Huang, C., Chu, P., et al. (2018). Histological, Physiological, and Comparative Proteomic Analyses Provide Insights into Leaf Rolling in *Brassica napus*. *Journal of Proteome Research*, 17(5), 1761-1772. DOI:10.1021/acs.jproteome.7b00744.

Chen, Y., Liu, P., Bai, D., and Li, R. (2010). Genetic analysis and gene mapping of a new rolled leaf mutant in rice (*Oryza sativa* L.). *Guangxi Agricultural Sciences*, 41(5), 403-407.

Chen, Z. (2016a). Biological Background. In *Statistical methods for QTL mapping* (1st ed., pp. 5): Chapman and Hall/CRC.

Chen, Z. (2016b). Multi-trait QTL Mapping and eQTL Mapping. In *Statistical methods for QTL mapping* (1st ed., pp. 219): Chapman and Hall/CRC.

Chen, Z. (2016c). *Statistical methods for QTL mapping*: Chapman and Hall/CRC.

Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37(3), 233-242. DOI:10.1038/ng1518.

Cho, S.H., Lee, C.H., Gi, E., Yim, Y., Koh, H.J., Kang, K., et al. (2018). The Rice Rolled Fine Striped (RFS) CHD3/Mi-2 Chromatin Remodeling Factor Epigenetically Regulates Genes Involved in Oxidative Stress Responses During Leaf Development. *Frontiers in Plant Science*, 9, 364.

Cho, S.H., Yoo, S.C., Zhang, H., Lim, J.H., and Paek, N.C. (2014). Rice *NARROW LEAF1* regulates leaf and adventitious root development. *Plant Molecular Biology Reporter*, 32(1), 270-281.

Cho, S.H., Yoo, S.C., Zhang, H., Pandeya, D., Koh, H.J., Hwang, J.Y., et al. (2013). The rice *narrow leaf2* and *narrow leaf3* loci encode WUSCHEL-related homeobox 3A (OsWOX3A) and function in leaf, spikelet, tiller and lateral root development. *New Phytologist*, 198(4), 1071-1084. DOI:10.1111/nph.12231.

Conesa, A., Nueda, M.J., Ferrer, A., and Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9), 1096-1102. DOI:10.1093/bioinformatics/btl056.

Dai, M., Zhao, Y., Ma, Q., Hu, Y., Hedden, P., Zhang, Q., et al. (2007). The rice YABBY1 gene is involved in the feedback regulation of gibberellin metabolism. *Plant physiology*, 144(1), 121-133.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22. DOI:10.1111/j.2517-6161.1977.tb01600.x.

Devlin, B., and Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics*, 55(4), 997-1004. DOI:10.1111/j.0006-341X.1999.00997.x.

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic Control, a New Approach to Genetic-Based Association Studies. *Theoretical Population Biology*, 60(3), 155-166. DOI:10.1006/tpbi.2001.1542.

Doebley, J., and Stec, A. (1993). Inheritance of the morphological differences between maize and teosinte: comparison of results for two F_2 populations. *Genetics*, 134(2), 559-570.

Draper, N.R., and Smith, H. (1998). *Applied regression analysis* (Vol. 326): John Wiley & Sons.

Edwards, M.D., Helentjaris, T., Wright, S., and Stuber, C.W. (1992). Molecular-marker-facilitated investigations of quantitative trait loci in maize. *Theoretical and Applied Genetics*, 83(6), 765-774. DOI:10.1007/BF00226696.

Endelman, J.B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, 4(3), 250-255. DOI:10.3835/plantgenome2011.08.0024.

Fairbanks, D.J., and Andersen, W.R. (1999). *Genetics : The continuity of life*. Pacific Grove, CA: Brooks/Cole Pub. : Wadsworth Pub.

Fang, L., Zhao, F., Cong, Y., Sang, X., Du, Q., Wang, D., et al. (2012). Rolling-leaf14 is a 2OG-Fe (II) oxygenase family protein that modulates rice leaf rolling by affecting secondary cell wall formation in leaves. *Plant biotechnology journal*, 10(5), 524-532. DOI:10.1111/j.1467-7652.2012.00679.x.

Fang, Y.X., Zhu, L., Pan, J.J., Yu, H.P., Xue, D.W., Rao, Y.C., et al. (2015). Identification and Fine Mapping of a Narrow Leaf Mutant *nal10* in Rice. *Chin J Rice Sci*, 29(6), 587-594. DOI:10.3969/j.issn.1001-7216.2015.06.004.

Feng, G.N., Zhang, C.Q., Tang, M.Y., Zhang, G.Y., Xu, C.W., and Liu, Q.Q. (2012). Genetic analysis and gene mapping of a narrow-leaf mutant (*nal3-t*) in rice

(*Oryza sativa* L.). *Journal of Yangzhou University(Agricultural and Life Science Edition)*, 14(3), 40-43.

Feng, P., Xing, Y.D., Liu, S., Guo, S., Zhu, M.D., Lou, Q.J., et al. (2015). Characterization and Gene Mapping of Rolled Leaf Mutant 28 (*rl28*) in Rice (*Oryza sativa* L.). *Acta Agronomica Sinica*, 41(08), 1164-1171. DOI:10.3724/sp.j.1006.2015.01164.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851-861. DOI:10.1038/nature06258.

Fujino, K., Matsuda, Y., Ozawa, K., Nishimura, T., Koshiba, T., Fraaije, M.W., et al. (2008). *NARROW LEAF 7* controls leaf shape mediated by auxin in rice. *Molecular Genetics and Genomics*, 279(5), 499-507. DOI:10.1007/s00438-008-0328-3.

Gao, Y., Lu, C., Wang, M., Wang, P., Yan, X., and Xie, K. (2007). QTL mapping for rolled leaf gene in rice. *Jiangsu J of Agr Sci*, 23, 5-10.

Gatti, D., Maki, A., Chesler, E.J., Kirova, R., Kosyk, O., Lu, L., et al. (2007). Genome-level analysis of genetic regulation of liver gene expression networks. *Hepatology*, 46(2), 548-557. DOI:10.1002/hep.21682.

Gatti, D.M., Shabalin, A.A., Lam, T.-C., Wright, F.A., Rusyn, I., and Nobel, A.B. (2009). FastMap: Fast eQTL mapping in homozygous populations. *Bioinformatics*, 25(4), 482-489. DOI:10.1093/bioinformatics/btn648.

Gatti, D.M., Shabalin, A.A., Sypa, M., Lam, T.-C., Wright, F.A., Nobel, A.B., et al. (2011). FastMap 2.0: Fast Association Mapping in Heterozygous Populations. *Working paper.*

Glazier, A.M., Nadeau, J.H., and Aitman, T.J. (2002). Finding Genes That Underlie Complex Traits. *Science*, 298(5602), 2345-2349. DOI:10.1126/science.1076641.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., et al. (2010). Food security: The challenge of feeding 9 billion people. *Science*, 1185383. DOI:10.1126/science.1185383.

Gottardo, R., Raftery, A.E., Yee Yeung, K., and Bumgarner, R.E. (2006). Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples. *Biometrics*, 62(1), 10-18. DOI:10.1111/j.1541-0420.2005.00397.x.

Guo, Y., Cheng, B.S., and Hong, D.L. (2010). Construction of SSR linkage map and analysis of QTLs for rolled leaf in japonica rice. *Rice Science*, 17(1), 28-34.

Hackett, C.A., Meyer, R.C., and Thomas, W.T.B. (2001). Multi-trait QTL mapping in barley using multivariate regression. *Genetical Research*, 77(1), 95-106. DOI:10.1017/S0016672300004869.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Gnetics*, 8(29), 299-309.

Haley, C.S., and Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4), 315-324.

Haley, C.S., Knott, S.A., and Elsen, J. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136(3), 1195-1207.

Hall, B. (2019). *In silico identification of novel genetic factors associated with longevity in Drosophila.* Nottingham Trent University,

Han, B., and Huang, X. (2013). Sequencing-based genome-wide association study in rice. *Current Opinion in Plant Biology*, 16(2), 133-138. DOI:10.1016/j.pbi.2013.03.006.

Hardcastle, T.J., and Kelly, K.A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1), 422. DOI:10.1186/1471-2105-11-422.

Harikrishnan, A., and Grace, L. (2013). PVC workers and their cytogenetic effects. *International Journal of Bioassays*, 2(8), 1152-1157.

Hayes, P.M., Liu, B.H., Knapp, S.J., Chen, F., Jones, B., Blake, T., et al. (1993). Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theoretical Applied Genetics*, 87(3), 392-401. DOI:10.1007/BF01184929.

Henshall, J.M., and Goddard, M.E. (1999). Multiple-Trait Mapping of Quantitative Trait Loci After Selective Genotyping Using Logistic Regression. *Genetics*, 151(2), 885-894.

Hibara, K.I., Obara, M., Hayashida, E., Abe, M., Ishimaru, T., Satoh, H., et al. (2009). The *ADAXIALIZED LEAF1* gene functions in leaf and embryonic pattern formation in rice. *Developmental biology*, 334(2), 345-354. DOI:10.1016/j.ydbio.2009.07.042.

Hoeschele, I., Uimari, P., Grignola, F.E., Zhang, Q., and Gage, K.M. (1997). Advances in Statistical Methods to Map Quantitative Trait Loci in Outbred Populations. *Genetics*, 147(3), 1445-1457.

Hooke, R. (1665). Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, with Observations and Inquiries Thereupon. In (pp. 113): Courier Dover Publications.

Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., and Gao, G. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*, 31(8), 1296-1297. DOI:10.1093/bioinformatics/btu817.

Hu, J., Zhu, L., Zeng, D., Gao, Z., Guo, L., Fang, Y., et al. (2010). Identification and characterization of *NARROW AND ROLLED LEAF 1*, a novel gene regulating

leaf morphology and plant architecture in rice. *Plant Molecular Biology*, 73(3), 283-292. DOI:10.1007/s11103-010-9614-7.

Hu, Y., Liu, D., Zhong, X., Zhang, C., Zhang, Q., and Zhou, D.-X. (2012). CHD3 protein recognizes and regulates methylated histone H3 lysines 4 and 27 over a subset of targets in the rice genome. *Proceedings of the National Academy of Sciences*, 109(15), 5773-5778. DOI:10.1073/pnas.1203148109.

Hu, Z., and Xu, S. (2009). PROC QTL—A SAS procedure for mapping quantitative trait loci. *International journal of plant genomics*, 2009, 1-3. DOI:10.1155/2009/141234.

Huang, J., Li, Z., and Zhao, D. (2016). Deregulation of the OsmiR160 Target Gene *OsARF18* Causes Growth and Developmental Defects with an Alteration of Auxin Signaling in Rice. *Scientific Reports*, 6(1), 1-14. DOI:10.1038/srep29938.

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nature Communications*, 6(1), 6258. DOI:10.1038/ncomms7258.

Jiang, C., and Zeng, Z.B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3), 1111-1127.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, 45(D1), D1040-D1045. DOI:10.1093/nar/gkw982.

Kabir, M.H., Ahmed, M.S., and Mollah, M.N.H. (2016). Gene Selection for Patient Clustering by Gaussian Mixture Model. *International Journal of Biometrics and Bioinformatics (IJBB)*, 10(3), 34.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-y., Freimer, N.B., et al. (2010). Variance component model to account for sample structure in

genome-wide association studies. *Nature Genetics*, 42(4), 348-354. DOI:10.1038/ng.548.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., et al. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, 178(3), 1709-1723. DOI:10.1534/genetics.107.080101.

Kao, C.-H., Zeng, Z.-B., and Teasdale, R.D. (1999). Multiple Interval Mapping for Quantitative Trait Loci. *Genetics*, 152(3), 1203-1216.

Kendziorski, C., Chen, M., Yuan, M., Lan, H., and Attie, A.D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, 62(1), 19-27.

Kendziorski, C.M., Newton, M.A., Lan, H., and Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22(24), 3899-3914. DOI:10.1002/sim.1548.

Kerr, M.K., and Churchill, G.A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2(2), 183-201. DOI:10.1093/biostatistics/2.2.183.

Khush, G.S. (2005). What it will take to feed 5.0 billion rice consumers in 2030. *Plant Molecular Biology*, 59(1), 1-6. DOI:10.1007/s11103-005-2159-5.

Khush, G.S. (2013). Strategies for increasing the yield potential of cereals: case of rice as an example. *Plant Breeding*, 132(5), 433-436. DOI:10.1111/pbr.1991.

Kim, Hyun S., Minna, John D., and White, Michael A. (2013). GWAS Meets TCGA to Illuminate Mechanisms of Cancer Predisposition. *Cell*, 152(3), 387-389. DOI:10.1016/j.cell.2013.01.027.

Kirik, V., and Bäumlein, H. (1996). A novel leaf-specific myb-related protein with a single binding repeat. *Gene*, 183(1-2), 109-113.

Knott, S.A., and Haley, C.S. (2000). Multitrait Least Squares for Quantitative Trait Loci Detection. 156(2), 899-911.

Korol, A.B., Ronin, Y.I., Itskovich, A.M., Peng, J., and Nevo, E. (2001). Enhanced Efficiency of Quantitative Trait Loci Mapping Analysis Based on Multivariate Complexes of Quantitative Traits. 157(4), 1789-1803.

Korol, A.B., Ronin, Y.I., and Kirzhner, V.M. (1995). Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics*, 140(3), 1137-1147.

Kosambi, D.D. (2016). The Estimation of Map Distances from Recombination Values. In R. Ramaswamy (Ed.), *D.D. Kosambi: Selected Works in Mathematics and Statistics* (pp. 125-130). New Delhi: Springer India.

Kruskal, W.H., and Wallis, W.A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260), 583-621. DOI:10.1080/01621459.1952.10483441.

Lander, E.S. (1996). The New Genomics: Global Views of Biology. *Science*, 274(5287), 536-539. DOI:10.1126/science.274.5287.536.

Lander, E.S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1), 185-199.

Lang, Y., Zhang, Z., Gu, X., Yang, J., and Zhu, Q. (2004). Physiological and ecological effects of crimpy leaf character in rice (*Oryza sativa* L.) I. Leaf orientation, canopy structure and light distribution. *Zuo wu xue bao*, 30(8), 806-810.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), 2947-2948. DOI:10.1093/bioinformatics/btm404.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), R29. DOI:10.1186/gb-2014-15-2-r29.

Lee, S.H., van der Werf, J.H.J., Hayes, B.J., Goddard, M.E., and Visscher, P.M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS genetics*, 4(10), e1000231-e1000231. DOI:10.1371/journal.pgen.1000231.

Lee, Y.K., Woo, M.O., Lee, D., Lee, G., Kim, B., and Koh, H.J. (2016). Identification of a novel candidate gene for rolled leaf in rice. *Genes & Genomics*, 38(11), 1077-1084. DOI:10.1007/s13258-016-0451-1.

Leiter, E.H., Reifsnyder, P.C., Wallace, R., Li, R., King, B., and Churchill, G.C. (2009). NOD × 129.*H2*^{g7} Backcross Delineates 129S1/SvImJ-Derived Genomic Regions Modulating Type 1 Diabetes Development in Mice. *Diabetes*, 58(7), 1700-1703. DOI:10.2337/db09-0120.

Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M.G., et al. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8), 1035-1043. DOI:10.1093/bioinformatics/btt087.

Li, C., Zou, X., Zhang, C., Shao, Q., Liu, J., Liu, B., et al. (2016). *OsLBD3-7* overexpression induced adaxially rolled leaves in rice. *PloS one*, 11(6), e0156413.

Li, L., Shi, Z.Y., Li, L., Shen, G.Z., Wang, X.Q., An, L.S., et al. (2010). Overexpression of *ACL1* (*abaxially curled leaf 1*) increased bulliform cells and induced abaxial curling of leaf blades in rice. *Molecular plant*, 3(5), 807-817. DOI:10.1093/mp/ssq022.

Li, L., Xue, X., Chen, Z., Zhang, Y., Ma, Y., Pan, C., et al. (2014). Isolation and characterization of rl (t), a gene that controls leaf rolling in rice. *Chinese Science Bulletin*, 59(25), 3142-3152. DOI:10.1007/s11434-014-0357-8.

Li, L., Xue, X., Zuo, S.M., Chen, Z.X., Zhang, Y.F., Li, Q.Q., et al. (2013). Suppressed expression of *OsAGO1a* leads to adaxial leaf rolling in rice. *Chin J Rice Sci*, 27, 223-230.

Li, M., Xiong, G., Li, R., Cui, J., Tang, D., Zhang, B., et al. (2009). Rice cellulose synthase-like D4 is essential for normal cell-wall biosynthesis and plant growth. *The Plant Journal*, 60(6), 1055-1069. DOI:10.1111/j.1365-313X.2009.04022.x.

Li, Q., and Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*, 32(3), 215-226. DOI:10.1002/gepi.20296.

Li, S., Ma, Y., Li, H., Zhou, K., He, P., Chenying, et al. (1998). Genetic analysis and mapping the flag leaf roll in rice (Oryza sativaL.). *Journal of Sichuan Agricultural University*, 16(4), 391-393.

Li, W., Wu, C., Hu, G., Xing, L., Qian, W., Si, H., et al. (2013). Characterization and fine mapping of a novel rice narrow leaf mutant nal9. *Journal of integrative plant biology*, 55(11), 1016-1025.

Li, W.Q., Zhang, M.J., Gan, P.F., Qiao, L., Yang, S.Q., Miao, H., et al. (2017). *CLD1*/*SRL1* modulates leaf rolling by affecting cell wall formation, epidermis integrity and water homeostasis in rice. *The Plant Journal*, 92(5), 904-923. DOI:10.1111/tpj.13728.

Liang, R., Qin, R., Zeng, D.D., Zheng, X., Jin, X.L., and Shi, C.H. (2016). Phenotype Analysis and Gene Mapping of Narrow and Rolling Leaf Mutant *nrl4* in Rice (*Oryza sativa* L.). *Scientia Agricultura Sinica*, 49(20), 3863-3873. DOI:10.3864/j.issn.0578-1752.2016.20.001.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18), 2397-2399. DOI:10.1093/bioinformatics/bts444.

Liu, B., Li, P., Li, X., Liu, C., Cao, S., Chu, C., et al. (2005). Loss of function of OsDCL1 affects microRNA accumulation and causes developmental defects in rice. *Plant physiology*, 139(1), 296-305.

Liu, B.H. (1997). *Statistical Genomics: Linkage, Mapping, and QTL Analysis*: CRC Press.

Liu, C., Kong, W.Y., You, S.M., Zhong, X.J., Jiang, L., Zhao, Z.G., et al. (2015). Genetic Analysis and Fine Mapping of a Novel Rolled Leaf Gene in Rice. *Scientia Agricultura Sinica*, 48(13), 2487-2496. DOI:10.3864/j.issn.0578-1752.2015.13.001.

Liu, L., Zhang, D., Liu, H., and Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics*, 14(1), 132. DOI:10.1186/1471-2105-14-132.

Liu, X., Li, M., Liu, K., Tang, D., Sun, M., Li, Y., et al. (2016). *Semi-Rolled Leaf2* modulates rice leaf rolling by regulating abaxial side cell differentiation. *Journal of experimental botany*, 67(8), 2139-2150.

Luo, Y.Z., Zhao, F.M., Sang, X.C., Ling, Y.H., Yang, Z.L., and He, G.H. (2009). Genetic analysis and gene mapping of a novel rolled leaf mutant *rl12(t)* in rice. *Acta Agronomica Sinica*, 35(11), 1967-1972.

Luo, Z., Yang, Z., Zhong, B., Li, Y., Xie, R., Zhao, F., et al. (2007). Genetic analysis and fine mapping of a dynamic rolled leaf gene, *RL10(t)*, in rice (*Oryza sativa* L.). *Genome*, 50(9), 811-817. DOI:10.1139/G07-064.

Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits* (Vol. 1): Sinauer Sunderland, MA.

Ma, X., Ma, J., Zhai, H., Xin, P., Chu, J., Qiao, Y., et al. (2015). CHR729 is a CHD3 protein that controls seedling development in rice. *PloS one*, 10(9), e0138934.

Ma, Y., Wang, F., Guo, J., and Zhang, X.S. (2009). Rice *OsAS2* gene, a member of LOB domain family, functions in the regulation of shoot differentiation and leaf development. *Journal of Plant Biology*, 52(5), 374-381.

Ma, Y., Zhao, Y., Shangguan, X., Shi, S., Zeng, Y., Wu, Y., et al. (2017). Overexpression of *OsRRK1* Changes Leaf Morphology and Defense to Insect in Rice. *Frontiers in Plant Science*, 8, 1-14. DOI:10.3389/fpls.2017.01783.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1), D896-D901. DOI:10.1093/nar/gkw1133.

Mangin, B., Thoquet, P., and Grimsley, N. (1998). Pleiotropic QTL Analysis. *Biometrics*, 54(1), 88-99. DOI:10.2307/2533998.

Marchler-Bauer, A., and Bryant, S.H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Research*, 32(suppl_2), W327-W331. DOI:10.1093/nar/gkh454.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1), D222-D226. DOI:10.1093/nar/gku1221.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, 39(suppl_1), D225-D229. DOI:10.1093/nar/gkq1189.

Mihoko, M., and Eguchi, S. (2002). Robust blind source separation by beta divergence. *Neural computation*, 14(8), 1859-1886. DOI:10.1162/089976602760128045.

Mollah, M.N.H., Eguchi, S., and Minami, M. (2007). Robust prewhitening for ICA by minimizing *β*-divergence and its application to FastICA. *Neural Processing Letters*, 25(2), 91-110.

Nuruzzaman, M., Manimekalai, R., Sharoni, A.M., Satoh, K., Kondoh, H., Ooka, H., et al. (2010). Genome-wide analysis of NAC transcription factor family in rice. *Gene*, 465(1-2), 30-44. DOI:10.1016/j.gene.2010.06.008.

Ott, J. (1999). *Analysis of human genetic linkage* (3rd ed.). Baltimore, Maryland: Johns Hopkins University Press.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., et al. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4), 650-654. DOI:10.1038/ng1047.

Pan, C.H., Li, L., Chen, Z.X., Xue, X., Zhang, Y.F., Zuo, S.M., et al. (2011). Fine Mapping of a Rolled Leaf Gene rl(t) in Rice. *Chin J Rice Sci*, 25(5), 455-460. DOI:10.3969/j.issn.1001-7216.2011.05.001.

Pan, Y.L., Mao, B.G., Hu, Y.Y., Guo, L.Q., Peng, Y., Shao, Y., et al. (2015). Genetic analysis and gene mapping of a narrow-leaf mutant *nl(t)* in rice (*Oryza sativa* L.). *Journal of Biology*, 32, 92–95.

Park, D.J., Lukens, A.K., Neafsey, D.E., Schaffner, S.F., Chang, H.-H., Valim, C., et al. (2012). Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. *Proceedings of the National Academy of Sciences*, 109(32), 13052-13057.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12), e190-e190. DOI:10.1371/journal.pgen.0020190.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904-909. DOI:10.1038/ng1847.

Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: common disease–common variant…or not? *Human Molecular Genetics*, 11(20), 2417-2423. DOI:10.1093/hmg/11.20.2417.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., and Donnelly, P. (2000). Association Mapping in Structured Populations. *The American Journal of Human Genetics*, 67(1), 170-181. DOI:10.1086/302959.

Qi, J., Qian, Q., Bu, Q., Li, S., Chen, Q., Sun, J., et al. (2008). Mutation of the rice Narrow leaf1 gene, which encodes a novel protein, affects vein patterning and polar auxin transport. *Plant physiology*, 147(4), 1947-1959.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends in Genetics*, 17(9), 502-510. DOI:10.1016/S0168-9525(01)02410-6.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140. DOI:10.1093/bioinformatics/btp616.

Sang, X.C., Lin, T.T., He, P.L., Wang, X.W., Liao, H.X., Zhang, X.B., et al. (2014). Identification and Gene Mapping of a Dominant Narrow Leaf Mutant Dnal1 in Rice (*Oryza sativa*). *Scientia Agricultura Sinica*, 47(9), 1819-1827. DOI:10.3864/j.issn.0578-1752.2014.09.017.

Schauer, N., and Fernie, A.R. (2006). Plant metabolomics: towards biological function and mechanism. *Trends in plant science*, 11(10), 508-516. DOI:10.1016/j.tplants.2006.08.007.

Schleiden, M.J. (1838). Beiträge über Phytogenesis. *Arch. Anat. Physiol. Wiss. Med*, 137–176.

Schwann, T. (1838). Ueber die Analogie in der Structur und dem Wachsthum der Thiere und Pflanzen. *Neue Not Geb Nat Heil*, 1838, 33-36.

Schwann, T. (1839). *Microscopical Researches into the Accordance in the Structure and Growth of Animals and Plants. (English translation by Henry Smith, for the Sydenham Society, 1847)*. Berlin: Sydenham Society.

Segami, S., Kono, I., Ando, T., Yano, M., Kitano, H., Miura, K., et al. (2012). Small and round seed 5 gene encodes alpha-tubulin regulating seed cell elongation in rice. *Rice*, 5(1), 4.

Shao, Y., Pan, C., Chen, Z., Zuo, S., Zhang, Y., and Pan, X. (2005). Fine mapping of an incomplete recessive gene for leaf rolling in rice (*Oryza sativa* L.). *Chinese Science Bulletin*, 50(21), 2466-2472. DOI:10.1007/bf03183637.

Shao, Y.J., Chen, Z.X., Zhang, Y.F., Chen, E.H., Qi, D.C., Miao, J., et al. (2005). One major QTL mapping and physical map construction for rolled leaf in rice. *Yi Chuan Xue Bao*, 32(5), 501-506.

Shi, L., Wei, X., Adedze, Y., Sheng, Z., Tang, S., Hu, P., et al. (2016). Characterization and gene cloning of the rice (*Oryza sativa* L.) dwarf and narrow-leaf mutant *dnl3*. *Genetics and Molecular Research*, 15(3), gmr.15038731. DOI:10.4238/gmr.15038731.

Shi, Y., Chen, J., Liu, W., Huang, Q., Shen, B., Leung, H., et al. (2009). Genetic analysis and gene mapping of a new rolled-leaf mutant in rice (Oryza sativa L.). *Science in China Series C: life sciences*, 52(9), 885-890.

Shi, Z., Wang, J., Wan, X., Shen, G., Wang, X., and Zhang, J. (2007). Over-expression of rice *OsAGO7* gene induces upward curling of the leaf blade that enhanced erect-leaf habit. *Planta*, 226(1), 99-108. DOI:10.1007/s00425-006-0472-0.

Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397-420). New York, NY: Springer New York.

Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 47(1), 35-39. DOI:10.1007/bf00277402.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440-9445. DOI:10.1073/pnas.1530509100.

Struck, T.J., Mannakee, B.K., and Gutenkunst, R.N. (2018). The impact of genome-wide association studies on biomedical research publications. *Human Genomics*, 12(1), 38. DOI:10.1186/s40246-018-0172-4.

Sugiyama, F., Churchill, G.A., Higgins, D.C., Johns, C., Makaritsis, K.P., Gavras, H., et al. (2001). Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics*, 71(1), 70-77.

Szatkiewicz, J.P., Beane, G.L., Ding, Y., Hutchins, L., Pardo-Manuel de Villena, F., and Churchill, G.A. (2008). An imputed genotype resource for the laboratory mouse. *Mammalian Genome*, 19(3), 199-208. DOI:10.1007/s00335-008-9098-9.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1), D362-D368. DOI:10.1093/nar/gkw937.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, 30(12), 2725-2729. DOI:10.1093/molbev /mst197.

Terwilliger, J.D., and Ott, J. (1994). *Handbook of Human Genetic Linkage* (1st ed.). Baltimore, Maryland: Johns Hopkins University Press.

Thoday, J. (1961). Location of polygenes. *Nature*, 191, 368-370.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673-4680. DOI:10.1093/nar /22.22.4673.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics*, 43(2), 159-162. DOI:10.1038/ng.746.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, 45(W1), W122-W129. DOI:10.1093/nar/gkx382.

Tian, X.Q., Sang, X.C., Zhao, F.M., Li, Y.F., Ling, Y.H., Yang, Z.L., et al. (2012). Genetic Analysis and Molecular Mapping of a Rolled Leaf Gene *RL13* in Rice (*Oryza sativa* L.). *Acta Agronomica Sinica*, 38(03), 423-428. DOI:10.3724/sp.j.1006.2012.00423.

Toriba, T., Harada, K., Takamura, A., Nakamura, H., Ichikawa, H., Suzaki, T., et al. (2007). Molecular characterization the *YABBY* gene family in *Oryza sativa* and expression analysis of *OsYABBY1*. *Molecular Genetics and Genomics*, 277(5), 457-468. DOI:10.1007/s00438-006-0202-0.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9), 5116-5121. DOI:10.1073/pnas.091062498.

Visscher, P.M. (2008). Sizing up human height variation. *Nature Genetics*, 40(5), 489-490. DOI:10.1038/ng0508-489.

Wang, D., Liu, H., Li, K., Li, S., and Tao, Y. (2009). Genetic analysis and gene mapping of a narrow leaf mutant in rice (Oryza sativa L.). *Chinese Science Bulletin*, 54(5), 752-758.

Wang, D.Z., Sang, X.C., You, X.Q., Wang, Z., Wang, Q.S., Zhao, F.M., et al. (2011). Genetic Analysis and Gene Mapping of a Novel Narrow and Rolled Leaf Mutant *nrl2*$_{(t)}$ in Rice (*Oryza sativa* L.). *Acta Agronomica Sinica*, 37(7), 1159-1166.

Wang, F., Tang, Y., Miao, R., Xu, F., Lin, T., He, G., et al. (2012). Identification and gene mapping of a narrow and upper-albino leaf mutant in rice (*Oryza sativa* L.). *Chinese Science Bulletin*, 57(28-29), 3798-3803.

Wang, L., Xu, J., Nian, J., Shen, N., Lai, K., Hu, J., et al. (2016). Characterization and fine mapping of the rice gene *OsARVL4* regulating leaf morphology and leaf vein development. *Plant Growth Regulation*, 78(3), 345-356. DOI:10.1007/s10725-015-0097-z.

Wang, X., Gu, F., and Sun, B. (2012). Ds-tagged rice rolling leaf mutant abnormal in bulliform cells. *Journal of Soochow University (Natural Science Edition)*, 28(2), 89-94.

Wang, X., Wang, F., Chen, H., Liang, X., Huang, Y., and Yi, J. (2017). Comparative genomic hybridization and transcriptome sequencing reveal that two genes, *OsI_14279 (LOC_Os03g62620)* and *OsI_10794 (LOC_Os03g14950)* regulate the mutation in the *γ-rl* rice mutant. *Physiology and Molecular Biology of Plants*, 23(4), 745-754.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), 80-83. DOI:10.2307/3001968.

Williams, J.T., Van Eerdewegh, P., Almasy, L., and Blangero, J. (1999). Joint Multipoint Linkage Analysis of Multivariate Qualitative and Quantitative Traits. I. Likelihood Formulation and Simulation Results. *The American Journal of Human Genetics*, 65(4), 1134-1147. DOI:10.1086/302570.

Woo, Y.M., Park, H.J., Su'udi, M., Yang, J.I., Park, J.J., Back, K., et al. (2007). Constitutively wilted 1, a member of the rice YUCCA gene family, is required for maintaining water homeostasis and an appropriate root to shoot ratio. *Plant Molecular Biology*, 65(1-2), 125-136. DOI:10.1007/s11103-007-9203-6.

Wu, C., Fu, Y., Hu, G., Si, H., Cheng, S., and Liu, W. (2010). Isolation and characterization of a rice mutant with narrow and rolled leaves. *Planta*, 232(2), 313-324. DOI:10.1007/s00425-010-1180-3.

Wu, R., Li, S., He, S., Waßmann, F., Yu, C., Qin, G., et al. (2011). CFL1, a WW domain protein, regulates cuticle development by modulating the function of HDG1, a class IV homeodomain transcription factor, in rice and Arabidopsis. *The Plant Cell*, tpc. 111.088625.

Wu, R., Ma, C., and Casella, G. (2007). *Statistical genetics of quantitative traits: linkage, maps and QTL*: Springer Science & Business Media.

Wu, X. (2009). Prospects of Developing Hybrid Rice with Super High Yield. *Agronomy Journal*, 101(3), 688-695. DOI:10.2134/agronj2008.0128f.

Wu, Y., Luo, L., Chen, L., Tao, X., Huang, M., Wang, H., et al. (2016). Chromosome mapping, molecular cloning and expression analysis of a novel gene response for leaf width in rice. *Biochemical and biophysical research communications*, 480(3), 394-401.

Xia, L., Zou, D., Sang, J., Xu, X., Yin, H., Li, M., et al. (2017). Rice Expression Database (RED): An integrated RNA-Seq-derived gene expression database for rice. *Journal of Genetics and Genomics*, 44(5), 235-241. DOI:10.1016/j.jgg.2017.05.003.

Xia, M.L., Tang, D.Y., Yang, Y.Z., Li, Y.X., Wang, W.W., Lu, H., et al. (2017). Preliminary Study on the Rice *OsYABBY6* Gene Involving in the Regulation of Leaf Development. *Life Science Research*, 21(1), 23-30. DOI:10.16605/j.cnki.1007-7847.2017.01.005.

Xiang, J.J., Zhang, G.H., Qian, Q., and Xue, H.W. (2012). *SEMI-ROLLED LEAF1* encodes a putative glycosylphosphatidylinositol-anchored protein and modulates rice leaf rolling by regulating the formation of bulliform cells. *Plant physiology*, 159(4), 1488-1500. DOI:10.1104/pp.112.199968.

Xie, Z.W., Sun, W., Yin, L., Zhao, J.F., Yuan, S.J., Zhang, W.H., et al. (2013). Phenotypic and Genetic Analyses of a Novel Adaxially-rolled Leaf Mutant in Rice. *Acta Agronomica Sinica*, 39(11), 1970-1975. DOI:10.3724/sp.j.1006.2013.01970.

Xu, H., Sarkar, B., and George, V. (2009). A new measure of population structure using multiple single nucleotide polymorphisms and its relationship with F_{ST}. *BMC Research Notes*, 2(1), 21. DOI:10.1186/1756-0500-2-21.

Xu, J., Wang, L., Zhou, M., Zeng, D., Hu, J., Zhu, L., et al. (2017). Narrow albino leaf 1 is allelic to CHR729, regulates leaf morphogenesis and development by affecting auxin metabolism in rice. *Plant Growth Regulation*, 82(1), 175-186.

Xu, J., Zhong, D., Yu, S., Luo, L., and Li, Z. (1999). QTLs affecting leaf rolling and folding in rice. *Rice Genet Newsl*, 16, 51-52.

Xu, S. (1995). A comment on the simple regression method for interval mapping. 141(4), 1657-1659.

Xu, S. (2013a). Mapping QTL for Multiple Traits. In *Principles of Statistical Genomics* (pp. 209-222). New York: Springer.

Xu, S. (2013b). *Principles of statistical genomics*: Springer.

Xu, S. (2013c). QTL Mapping in Other Populations. In *Principles of Statistical Genomics* (pp. 171-185). New York, NY: Springer New York.

Xu, S. (2013d). Recombination Fraction. In *Principles of Statistical Genomics* (pp. 11-22). New York, NY: Springer New York.

Xu, Y., Wang, Y., Long, Q., Huang, J., Wang, Y., Zhou, K., et al. (2014). Overexpression of *OsZHD1*, a zinc finger homeodomain class homeobox transcription factor, induces abaxially curled and drooping leaf in rice. *Planta*, 239(4), 803-816. DOI:10.1007/s00425-013-2009-7.

Yan, C., Yan, S., Zhang, Z., Liang, G., Lu, J., and Gu, M. (2006). Genetic analysis and gene fine mapping for a rice novel mutant (rl9 (t)) with rolling leaf character. *Chinese Science Bulletin*, 51(1), 63-69.

Yan, S., Yan, C.J., Zeng, X.H., Yang, Y.C., Fang, Y.W., Tian, C.Y., et al. (2008). *ROLLED LEAF 9*, encoding a GARP protein, regulates the leaf abaxial cell

fate in rice. *Plant Molecular Biology*, 68(3), 239-250. DOI:10.1007/s11103-008-9365-x.

Yang, C., Li, D., Liu, X., Ji, C., Hao, L., Zhao, X., et al. (2014). OsMYB103L, an R2R3-MYB transcription factor, influences leaf rolling and mechanical strength in rice (*Oryza sativa* L.). *BMC plant biology*, 14(1), 158. DOI:10.1186/1471-2229-14-158.

Yang, S.Q., Li, W.Q., Miao, H., Gan, P.F., Qiao, L., Chang, Y.L., et al. (2016). *REL2*, a gene encoding an unknown function protein which contains DUF630 and DUF632 domains controls leaf rolling in rice. *Rice*, 9(1), 37. DOI:10.1186/s12284-016-0105-6.

Ye, Y., Wu, K., Chen, J., Liu, Q., Wu, Y., Liu, B., et al. (2018). OsSND2, a NAC family transcription factor, is involved in secondary cell wall biosynthesis through regulating MYBs expression in rice. *Rice*, 11(1), 36. DOI:10.1186/s12284-018-0228-z.

Yen, S.T., Lin, M.M., and Hsieh, S.C. (1968). Linkage relations of another induced dwarfness gene $d_{31}^{(1)}$ genic analysis in rice IX. *Bot. Bull. Acad. Sin.*, 9(1), 69-74.

Yi, J.C., Zhuang, C.X., Wang, X.J., Cao, Y.P., Liu, Y.G., and Mei, M.T. (2007). Genetic analysis and molecular mapping of a rolling leaf mutation gene in rice. *Journal of integrative plant biology*, 49(12), 1746-1753.

Yoshimura, A., Ideta, O., and Iwata, N. (1997). Linkage map of phenotype and RFLP markers in rice. *Plant Molecular Biology*, 35(1), 49-60. DOI:10.1023/a:1005764026871.

Yoshizawa, A.C., Itoh, M., Okuda, S., Moriya, Y., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35(suppl_2), W182-W185. DOI:10.1093/nar /gkm321.

Yu, D., Wu, H., Yang, W., Gong, P., Li, Y., and Zhao, D. (2010). Genetic Analysis and Mapping of the Unilateral Rolled Leaf Trait of Rice Mutant B157. *Plant Gene and Trait*, 1, 220-226.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2), 203-208. DOI:10.1038/ng1702.

Yuan, L. (1997). Hybrid rice breeding for super high yield. *Hybrid Rice*, 12(6), 1-6.

Zeng, P., Wang, T., and Huang, S. (2017). Cis-SNPs Set Testing and PrediXcan Analysis for Gene Expression Data using Linear Mixed Models. *Scientific Reports*, 7(1), 15237. DOI:10.1038/s41598-017-15055-8.

Zeng, Z.B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences*, 90(23), 10972-10976. DOI:10.1073/pnas.90.23.10972.

Zeng, Z.B. (1994a). A composite interval mapping method for locating multiple QTLs. *Paper presented at the Proceedings, 5th World Congress on Genetics Applied to Livestock Production, University of Guelph, Guelph, Ontario, Canada*.

Zeng, Z.B. (1994b). Precision mapping of quantitative trait loci. *Genetics*, 136(4), 1457-1468.

Zhang, F.T., Fang, J., Sun, C.H., Li, R.B., Luo, X.D., Xie, J.K., et al. (2012). Characterisation of a rice dwarf and twist leaf 1 (dtl1) mutant and fine mapping of *DTL1* gene. *Hereditas (Beijing)*, 34(1), 79-86. DOI:10.3724/sp.j.1005.2012.00079.

Zhang, G.H., Xu, Q., Zhu, X.D., Qian, Q., and Xue, H.W. (2009). SHALLOT-LIKE1 is a KANADI transcription factor that modulates rice leaf rolling by regulating leaf abaxial cell development. *The Plant Cell*, 21(3), 719-735. DOI:10.1105/tpc.108.061457.

Zhang, J., Zhang, H., Srivastava, A.K., Pan, Y., Bai, J., Fang, J., et al. (2018). Knockdown of Rice MicroRNA166 Confers Drought Resistance by Causing Leaf Rolling and Altering Stem Xylem Development. *Plant physiology*, 176(3), 2082-2094. DOI:10.1104/pp.17.01432.

Zhang, J.J., Wu, S.Y., Jiang, L., Wang, J.L., Zhang, X., Guo, X.P., et al. (2015). A detailed analysis of the leaf rolling mutant *sll2* reveals complex nature in regulation of bulliform cell development in rice (*Oryza sativa* L.). *Plant Biology*, 17(2), 437-448. DOI:10.1111/plb.12255.

Zhang, L.X., Liu, H.Q., Yu, X., Wang, L.Y., Fan, H.H., Jin, Q.S., et al. (2014). Molecular Mapping and Physiological Characterization of a Novel Mutant rl15(t) in Rice. *Scientia Agricultura Sinica*, 47(14), 2881-2888. DOI:10.3864/j.issn.0578-1752.2014.14.018.

Zhang, Q., Zheng, T., Hoang, L., Wang, C., Joseph, C., Zhang, W., et al. (2016). Joint mapping and allele mining of the rolled leaf trait in rice (*Oryza sativa* L.). *PloS one*, 11(7), e0158246.

Zhang, X., Rerksiri, W., Liu, A., Zhou, X., Xiong, H., Xiang, J., et al. (2013). Transcriptome profile reveals heat response mechanism at molecular and metabolic levels in rice flag leaf. *Gene*, 530(2), 185-192. DOI:10.1016/j.gene.2013.08.048.

Zhang, X.H., Qin, Y.Z., Zhang, Y.X., Zhan, X.D., Zhang, Z.H., Shen, X.H., et al. (2015). Gene Mapping of a Narrow and Rolled Leaf Mutant *Nrl3(t)* in Rice. *Chin J Rice Sci*, 29(6), 595-600. DOI:10.3969/j.issn.1001-7216.2015.06.005.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4), 355-360. DOI:10.1038/ng.546.

Zhao, C., Xu, J., Chen, Y., Mao, C., Zhang, S., Bai, Y., et al. (2012). Molecular cloning and characterization of OsCHR4, a rice chromatin-remodeling factor

required for early chloroplast development in adaxial mesophyll. *Planta*, 236(4), 1165-1176. DOI:10.1007/s00425-012-1667-1.

Zhao, F.M., Wei, X., Ma, L., Sang, X.C., WANG, N., Zhang, C.W., et al. (2015). Identification, gene mapping and candidate gene prediction of a late-stage rolled leaf mutant *lrl1* in rice (*Oryza sativa* L.). *Chinese Science Bulletin*, 60(32), 3133-3143.

Zhao, J., Luo, H., Jiang, Y., Yang, X., and Zha, R. (2017). Gene mapping for rice narrow leaf mutant *Narrow leaf 11* (*nal11*). *Journal of Southern Agriculture*, 48(7), 1133-1138.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., et al. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS genetics*, 3(1), e4-e4. DOI:10.1371/journal.pgen.0030004.

Zhao, K., Tung, C.W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communications*, 2(1), 467. DOI:10.1038/ncomms1467.

Zhao, S.Q., Hu, J., Guo, L.B., Qian, Q., and Xue, H.W. (2010). Rice leaf inclination2, a VIN3-like protein, regulates leaf angle through modulating cell division of the collar. *Cell research*, 20(8), 935-947.

Zhou, K., Ma, Y., Liu, T., Shen, M., and Pan, S. (1995). The breeding of subspecific heavy ear hybrid rice-exploration about super-high yield breeding of hybrid rice. *Journal of Sichuan Agricultural University*, 13(4), 403-407.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7), 821–824. DOI:10.1038/ng.2310.

Zhou, Y., Fang, Y., Zhu, J., Li, S., Gu, F., Gu, M., et al. (2010). Genetic analysis and gene fine mapping of a rolling leaf mutant ($rl_{11(t)}$) in rice (*Oryza sativa* L.). *Chinese Science Bulletin*, 55(17), 1763-1769.

Zhou, Y., Wang, D., Wu, T., Yang, Y., Liu, C., Yan, L., et al. (2018). LRRK1, a receptor-like cytoplasmic kinase, regulates leaf rolling through modulating bulliform cell development in rice. *Molecular Breeding*, 38(5), 48.

Zhu, Q., Yu, S., Chen, G., Ke, L., and Pan, D. (2017). Analysis of the differential gene and protein expression profile of the rolled leaf mutant of transgenic rice (*Oryza sativa* L.). *PloS one*, 12(7), e0181378. DOI:10.1371/journal.pone.0181378.

Zou, L.P., Sun, X.H., Zhang, Z.G., Liu, P., Wu, J.X., Tian, C.J., et al. (2011). Leaf rolling controlled by the homeodomain leucine zipper class IV gene *Roc5* in rice. *Plant physiology*, pp. 111.176016. DOI:10.1104/pp.111.176016.

Appendix

A2.1 Supplementary Figures of Chapter 2

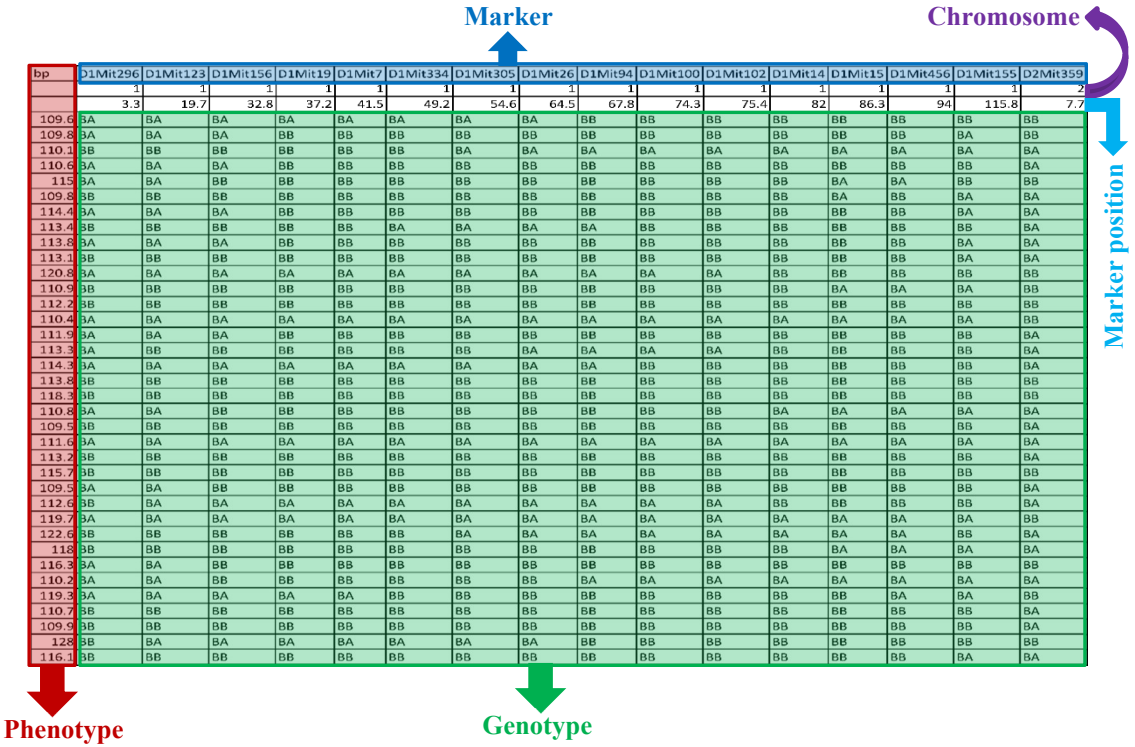
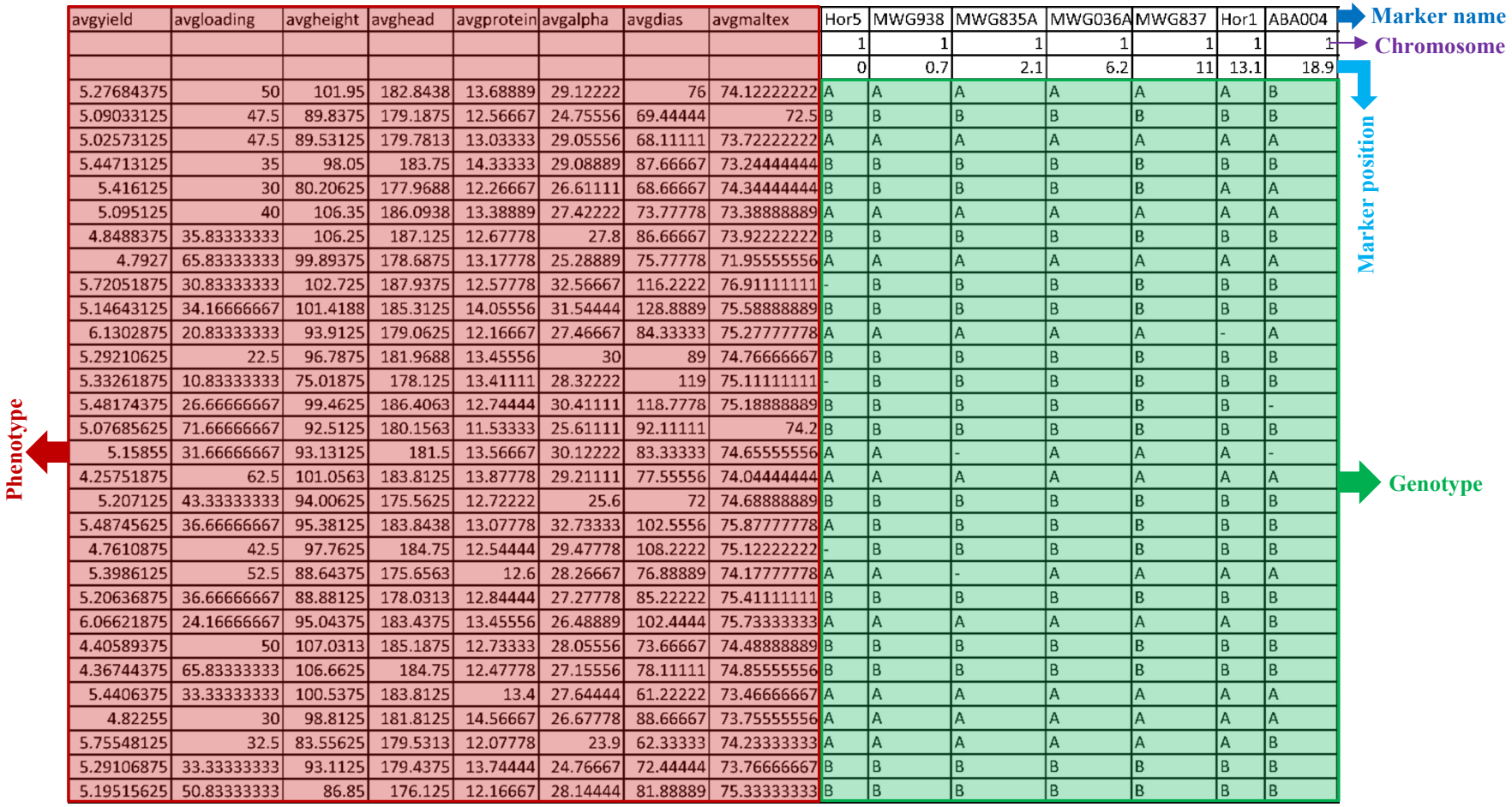


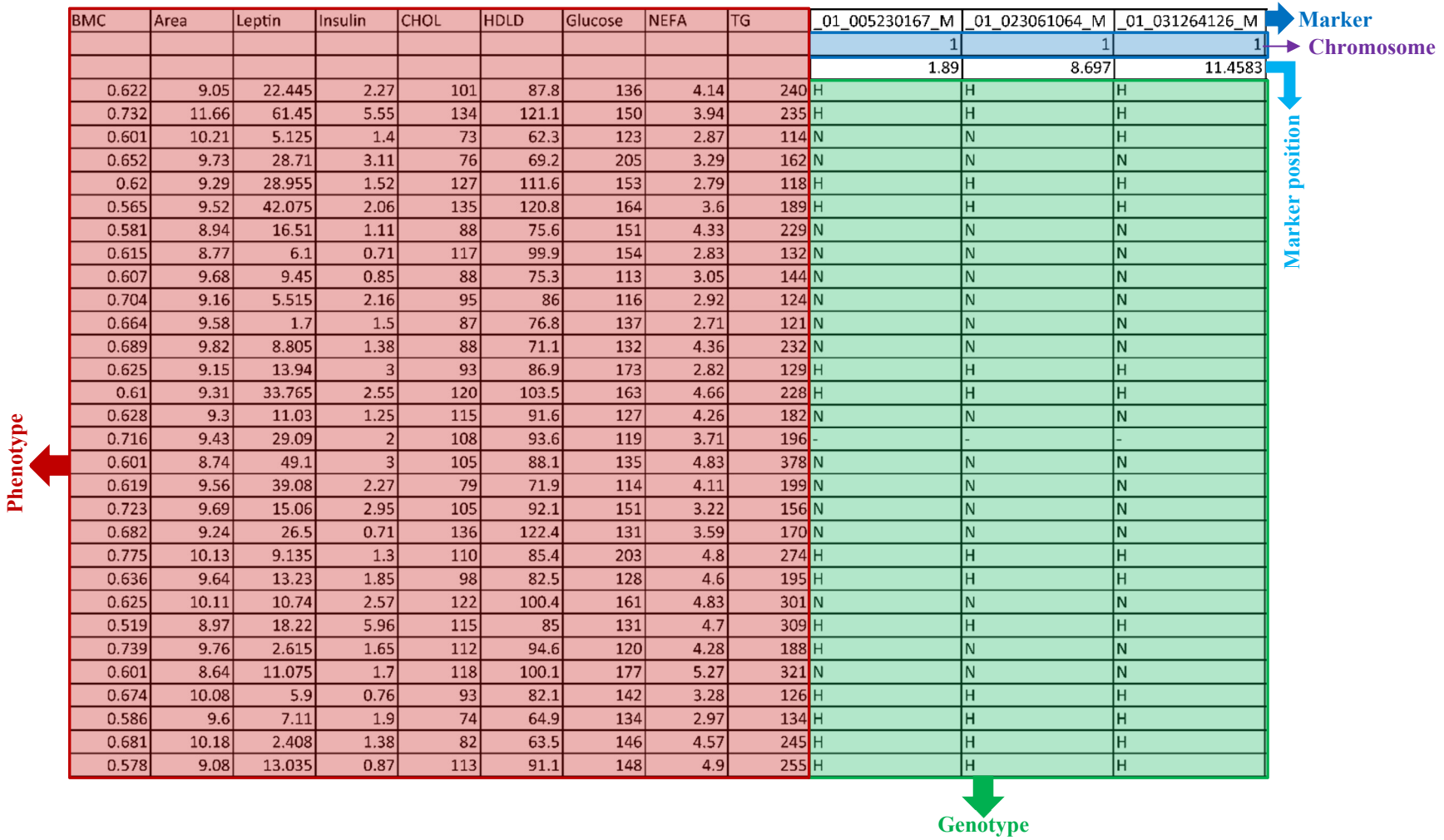
Figure A2.1: Structure of the "salt-induced hypertension" dataset obtained from a QTL experiment on male mice from a reciprocal backcross between the salt-sensitive c57BL/6J and the non-salt-sensitive A/J (A) inbred mouse strains.

A3.1 Supplementary Figures of Chapter 3



avgyield	avgloading	avgheight	avghead	avgprotein	avgalpha	avgdias	avgmaltex	Hor5	MWG938	MWG835A	MWG036A	MWG837	Hor1	ABA004
								1	1	1	1	1	1	1
								0	0.7	2.1	6.2	11	13.1	18.9
5.27684375	50	101.95	182.8438	13.68889	29.12222	76	74.12222222	A	A	A	A	A	A	B
5.09033125	47.5	89.8375	179.1875	12.56667	24.75556	69.44444	72.5	B	B	B	B	B	B	B
5.02573125	47.5	89.53125	179.7813	13.03333	29.05556	68.11111	73.72222222	A	A	A	A	A	A	A
5.44713125	35	98.05	183.75	14.33333	29.08889	87.66667	73.24444444	B	B	B	B	B	B	B
5.416125	30	80.20625	177.9688	12.26667	26.61111	68.66667	74.34444444	B	B	B	B	B	A	A
5.095125	40	106.35	186.0938	13.38889	27.42222	73.77778	73.38888889	A	A	A	A	A	A	A
4.8488375	35.83333333	106.25	187.125	12.67778	27.8	86.66667	73.92222222	B	B	B	B	B	B	B
4.7927	65.83333333	99.89375	178.6875	13.17778	25.28889	75.77778	71.95555556	A	A	A	A	A	A	A
5.72051875	30.83333333	102.725	187.9375	12.57778	32.56667	116.2222	76.91111111	-	B	B	B	B	B	B
5.14643125	34.16666667	101.4188	185.3125	14.05556	31.54444	128.8889	75.58888889	B	B	B	B	B	B	B
6.1302875	20.83333333	93.9125	179.0625	12.16667	27.46667	84.33333	75.27777778	A	A	A	A	A	-	A
5.29210625	22.5	96.7875	181.9688	13.45556	30	89	74.76666667	B	B	B	B	B	B	B
5.33261875	10.83333333	75.01875	178.125	13.41111	28.32222	119	75.11111111	-	B	B	B	B	B	B
5.48174375	26.66666667	99.4625	186.4063	12.74444	30.41111	118.7778	75.18888889	B	B	B	B	B	B	-
5.07685625	71.66666667	92.5125	180.1563	11.53333	25.61111	92.11111	74.2	B	B	B	B	B	B	B
5.15855	31.66666667	93.13125	181.5	13.56667	30.12222	83.33333	74.65555556	A	A	-	A	A	A	-
4.25751875	62.5	101.0563	183.8125	13.87778	29.21111	77.55556	74.04444444	A	A	A	A	A	A	A
5.207125	43.33333333	94.00625	175.5625	12.72222	25.6	72	74.68888889	B	B	B	B	B	B	B
5.48745625	36.66666667	95.38125	183.8438	13.07778	32.73333	102.5556	75.87777778	A	B	B	B	B	B	B
4.7610875	42.5	97.7625	184.75	12.54444	29.47778	108.2222	75.12222222	-	B	B	B	B	B	B
5.3986125	52.5	88.64375	175.6563	12.6	28.26667	76.88889	74.17777778	A	A	-	A	A	A	A
5.20636875	36.66666667	88.88125	178.0313	12.84444	27.27778	85.22222	75.41111111	B	B	B	B	B	B	B
6.06621875	24.16666667	95.04375	183.4375	13.45556	26.48889	102.4444	75.73333333	A	A	A	A	A	A	B
4.40589375	50	107.0313	185.1875	12.73333	28.05556	73.66667	74.48888889	B	B	B	B	B	B	B
4.36744375	65.83333333	106.6625	184.75	12.47778	27.15556	78.11111	74.85555556	B	B	B	B	B	B	B
5.4406375	33.33333333	100.5375	183.8125	13.4	27.64444	61.22222	73.46666667	A	A	A	A	A	A	A
4.82255	30	98.8125	181.8125	14.56667	26.67778	88.66667	73.75555556	A	A	A	A	A	A	A
5.75548125	32.5	83.55625	179.5313	12.07778	23.9	62.33333	74.23333333	A	A	A	A	A	A	B
5.29106875	33.33333333	93.1125	179.4375	13.74444	24.76667	72.44444	73.76666667	B	B	B	B	B	B	B
5.19515625	50.83333333	86.85	176.125	12.16667	28.14444	81.88889	75.33333333	B	B	B	B	B	B	B

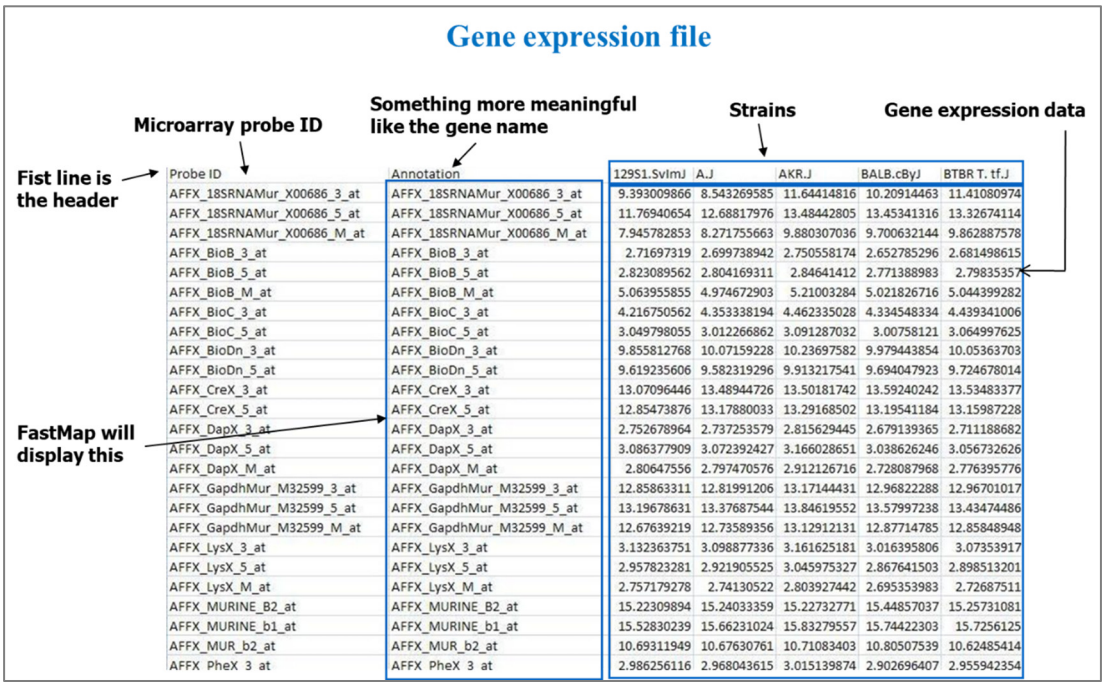
Figure A3.1: Structure of the barley dataset obtained from a QTL experiment on double haploid (DH) population of barley.



BMC	Area	Leptin	Insulin	CHOL	HDLD	Glucose	NEFA	TG	_01_005230167_M	_01_023061064_M	_01_031264126_M
									1	1	1
									1.89	8.697	11.4583
0.622	9.05	22.445	2.27	101	87.8	136	4.14	240	H	H	H
0.732	11.66	61.45	5.55	134	121.1	150	3.94	235	H	H	H
0.601	10.21	5.125	1.4	73	62.3	123	2.87	114	N	N	H
0.652	9.73	28.71	3.11	76	69.2	205	3.29	162	N	N	N
0.62	9.29	28.955	1.52	127	111.6	153	2.79	118	H	H	H
0.565	9.52	42.075	2.06	135	120.8	164	3.6	189	H	H	H
0.581	8.94	16.51	1.11	88	75.6	151	4.33	229	N	N	N
0.615	8.77	6.1	0.71	117	99.9	154	2.83	132	N	N	N
0.607	9.68	9.45	0.85	88	75.3	113	3.05	144	N	N	N
0.704	9.16	5.515	2.16	95	86	116	2.92	124	N	N	N
0.664	9.58	1.7	1.5	87	76.8	137	2.71	121	N	N	N
0.689	9.82	8.805	1.38	88	71.1	132	4.36	232	N	N	N
0.625	9.15	13.94	3	93	86.9	173	2.82	129	H	H	H
0.61	9.31	33.765	2.55	120	103.5	163	4.66	228	H	H	H
0.628	9.3	11.03	1.25	115	91.6	127	4.26	182	N	N	N
0.716	9.43	29.09	2	108	93.6	119	3.71	196	-	-	-
0.601	8.74	49.1	3	105	88.1	135	4.83	378	N	N	N
0.619	9.56	39.08	2.27	79	71.9	114	4.11	199	N	N	N
0.723	9.69	15.06	2.95	105	92.1	151	3.22	156	N	N	N
0.682	9.24	26.5	0.71	136	122.4	131	3.59	170	N	N	N
0.775	10.13	9.135	1.3	110	85.4	203	4.8	274	H	H	H
0.636	9.64	13.23	1.85	98	82.5	128	4.6	195	H	H	H
0.625	10.11	10.74	2.57	122	100.4	161	4.83	301	N	N	N
0.519	8.97	18.22	5.96	115	85	131	4.7	309	H	H	H
0.739	9.76	2.615	1.65	112	94.6	120	4.28	188	H	N	N
0.601	8.64	11.075	1.7	118	100.1	177	5.27	321	N	N	N
0.674	10.08	5.9	0.76	93	82.1	142	3.28	126	H	H	H
0.586	9.6	7.11	1.9	74	64.9	134	2.97	134	H	H	H
0.681	10.18	2.408	1.38	82	63.5	146	4.57	245	H	H	H
0.578	9.08	13.035	0.87	113	91.1	148	4.9	255	H	H	H

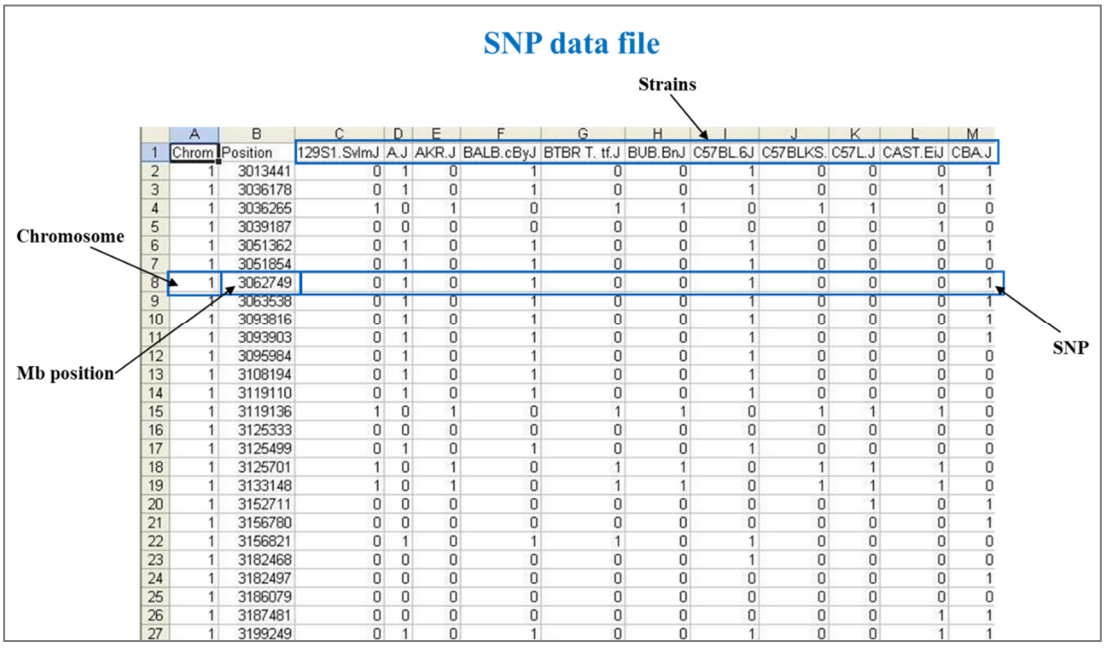
Figure A3.2: Structure of the mouse dataset obtained from a QTL experiment on backcross (BC) lines of mouse.

A4.1 Supplementary Figures of Chapter 4



Probe ID	Annotation	129S1.SvImJ	A.J	AKR.J	BALB.cByJ	BTBR T. tf.J
AFFX_18SRNAMur_X00686_3_at	AFFX_18SRNAMur_X00686_3_at	9.393009866	8.543269585	11.64414816	10.20914463	11.41080974
AFFX_18SRNAMur_X00686_5_at	AFFX_18SRNAMur_X00686_5_at	11.76940654	12.68817976	13.48442805	13.45341316	13.32674114
AFFX_18SRNAMur_X00686_M_at	AFFX_18SRNAMur_X00686_M_at	7.945782853	8.271755663	9.880307036	9.700632144	9.862887578
AFFX_BioB_3_at	AFFX_BioB_3_at	2.71697319	2.699738942	2.750558174	2.652785296	2.681498615
AFFX_BioB_5_at	AFFX_BioB_5_at	2.823089562	2.804169311	2.84641412	2.771388983	2.79835357
AFFX_BioB_M_at	AFFX_BioB_M_at	5.063955855	4.974672903	5.21003284	5.021826716	5.044399282
AFFX_BioC_3_at	AFFX_BioC_3_at	4.216750562	4.353338194	4.462335028	4.334548334	4.439341006
AFFX_BioC_5_at	AFFX_BioC_5_at	3.049798055	3.012266862	3.091287032	3.00758121	3.064997625
AFFX_BioDn_3_at	AFFX_BioDn_3_at	9.855812768	10.07159228	10.23697582	9.979443854	10.05363703
AFFX_BioDn_5_at	AFFX_BioDn_5_at	9.619235606	9.582319296	9.913217541	9.694047923	9.724678014
AFFX_CreX_3_at	AFFX_CreX_3_at	13.07096446	13.48944726	13.50181742	13.59240242	13.53483377
AFFX_CreX_5_at	AFFX_CreX_5_at	12.85473876	13.17880033	13.29168502	13.19541184	13.15987228
AFFX_DapX_3_at	AFFX_DapX_3_at	2.752678964	2.737253579	2.815629445	2.679139365	2.711188682
AFFX_DapX_5_at	AFFX_DapX_5_at	3.086377909	3.072392427	3.166028651	3.038626246	3.056732626
AFFX_DapX_M_at	AFFX_DapX_M_at	2.80647556	2.797470576	2.912126716	2.728087968	2.776395776
AFFX_GapdhMur_M32599_3_at	AFFX_GapdhMur_M32599_3_at	12.85863311	12.81991206	13.17144431	12.96822288	12.96701017
AFFX_GapdhMur_M32599_5_at	AFFX_GapdhMur_M32599_5_at	13.19678631	13.37687544	13.84619552	13.57997238	13.43474486
AFFX_GapdhMur_M32599_M_at	AFFX_GapdhMur_M32599_M_at	12.67639219	12.73589356	13.12912131	12.87714785	12.85848948
AFFX_LysX_3_at	AFFX_LysX_3_at	3.132363751	3.098877336	3.161625181	3.016395806	3.07353917
AFFX_LysX_5_at	AFFX_LysX_5_at	2.957823281	2.921905525	3.045975327	2.867641503	2.898513201
AFFX_LysX_M_at	AFFX_LysX_M_at	2.757179278	2.74130522	2.803927442	2.695353983	2.72687511
AFFX_MURINE_B2_at	AFFX_MURINE_B2_at	15.22309894	15.24033359	15.22732771	15.44857037	15.25731081
AFFX_MURINE_b1_at	AFFX_MURINE_b1_at	15.52830239	15.66231024	15.83279557	15.74422303	15.7256125
AFFX_MUR_b2_at	AFFX_MUR_b2_at	10.69311949	10.67630761	10.71083403	10.80507539	10.62485414
AFFX_PheX_3_at	AFFX_PheX_3_at	2.986256116	2.968043615	3.015139874	2.902696407	2.955942354

Figure A4.1: Structure of the gene expression dataset obtained from the gene expression profile in liver of 32 BXD mouse strains.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Chrom	Position	129S1.SvImJ	A.J	AKR.J	BALB.cByJ	BTBR T. tf.J	BUB.BnJ	C57BL.6J	C57BLKS.	C57L.J	CAST.EiJ	CBA.J
2	1	3013441	0	1	0	1	0	0	1	0	0	0	1
3	1	3036178	0	1	0	1	0	0	1	0	0	1	1
4	1	3036265	1	0	1	0	1	1	0	1	1	0	0
5	1	3039187	0	0	0	0	0	0	0	0	0	1	0
6	1	3051362	0	1	0	1	0	0	1	0	0	0	1
7	1	3051854	0	1	0	1	0	0	1	0	0	0	0
8	1	3062749	0	1	0	1	0	0	1	0	0	0	1
9	1	3063538	0	1	0	1	0	0	1	0	0	0	1
10	1	3093816	0	1	0	1	0	0	1	0	0	0	1
11	1	3093903	0	1	0	1	0	0	1	0	0	0	1
12	1	3095984	0	1	0	1	0	0	1	0	0	0	0
13	1	3108194	0	1	0	1	0	0	1	0	0	0	0
14	1	3119110	0	1	0	1	0	0	1	0	0	0	0
15	1	3119136	1	0	1	0	1	1	0	1	1	1	0
16	1	3125333	0	0	0	0	0	0	0	0	0	0	0
17	1	3125499	0	1	0	1	0	0	1	0	0	0	0
18	1	3125701	1	0	1	0	1	1	0	1	1	1	0
19	1	3133148	1	0	1	0	1	1	0	1	1	1	0
20	1	3152711	0	0	0	0	0	0	0	0	1	0	1
21	1	3156780	0	0	0	0	0	0	0	0	0	0	1
22	1	3156821	0	1	0	1	1	0	1	0	0	0	0
23	1	3182468	0	0	0	0	0	0	1	0	0	0	0
24	1	3182497	0	0	0	0	0	0	0	0	0	0	1
25	1	3186079	0	0	0	0	0	0	0	0	0	0	0
26	1	3187481	0	0	0	0	0	0	0	0	0	1	1
27	1	3199249	0	1	0	1	0	0	1	0	0	1	1

Figure A4.2: Structure of the SNP dataset of 32 BXD mouse strains.

A6.1 ClustalW Method

The ClustalW method is a most popular, accurate and practical method in the category of hierarchical methods. In our study, the multiple alignment of amino acid sequences has been conducted using MEGA V6 software (Tamura et al., 2013) by ClustalW method (Larkin et al., 2007; Thompson et al., 1994). The basic multiple sequences alignment algorithm of ClustalW method consists of the following steps:

Step I: Calculate a distance matrix giving the divergence of each of all possible pairs of sequences.

Step II: Construct an unrooted neighbor-joining tree from the distance matrix.

Step III: Construct a rooted neighbor-joining tree (guide tree) and calculate sequence weights.

Step IV: Align the sequences progressively according to the branching order in the guide tree.

The above steps of multiple sequences alignment algorithm of ClustalW method are shown in Figure A6.1 in a flowchart. The calculations of multiple alignment using ClustalW algorithm has been shown beside the flowchart for 7 arbitrary protein sequences as an example.

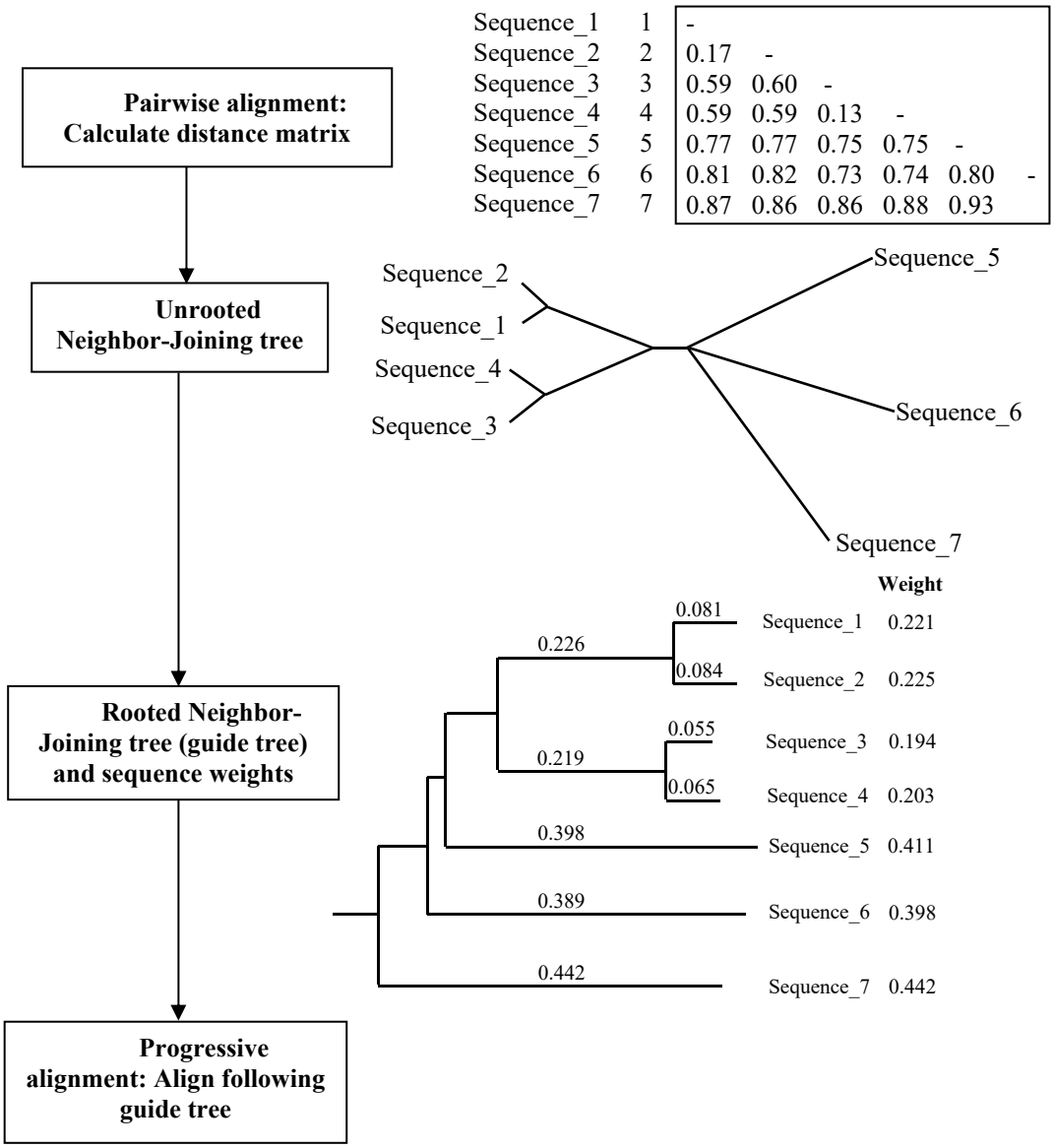


Figure A6.1: Basic procedure of multiple alignment of protein sequences using ClustalW method.

A6.2 Supplementary Figures of Chapter 6

A6.2.1 Investigation of gene expression pattern at different tissues of rolling leaf (RL) genes using line charts

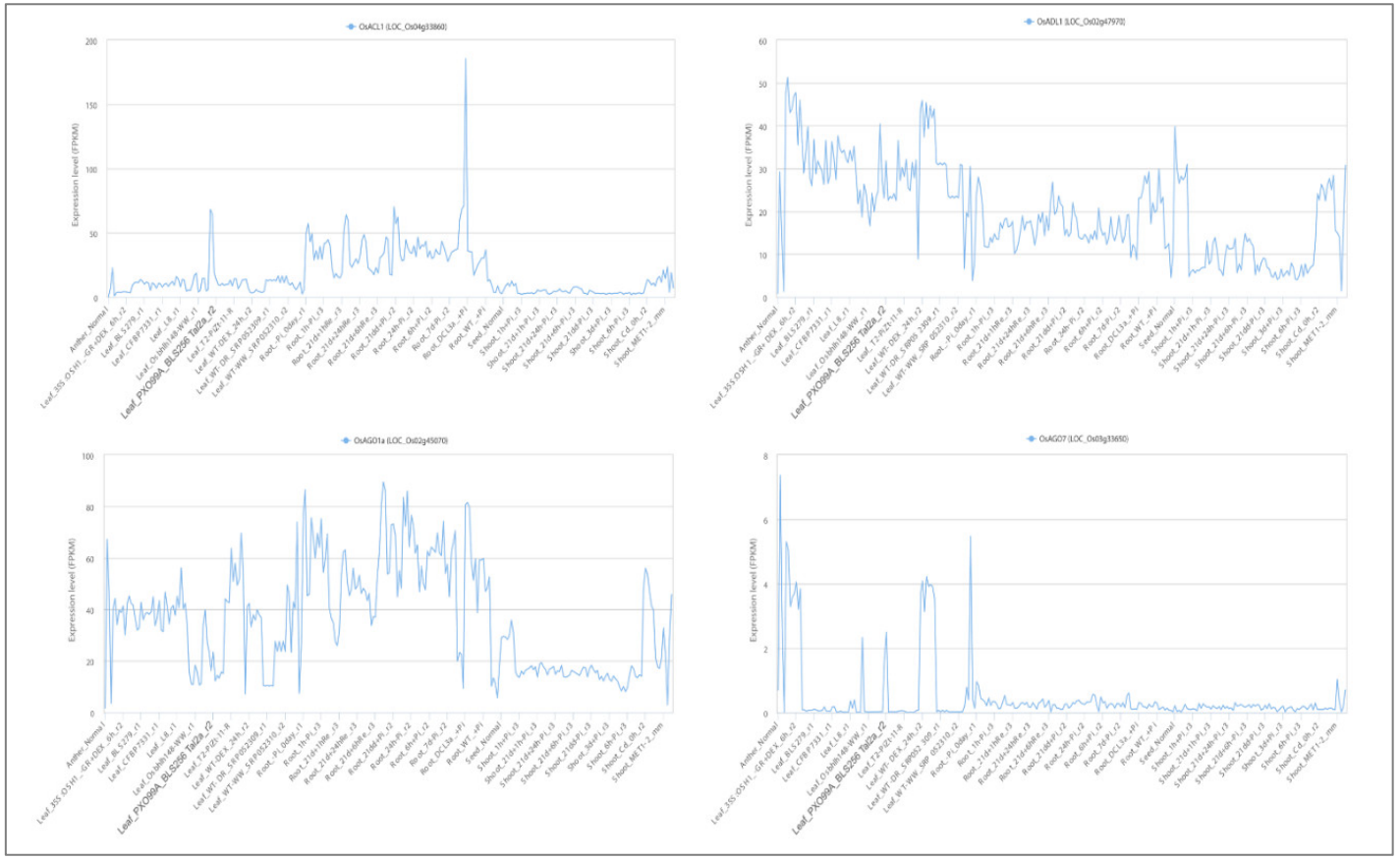


Figure A6.2: Line charts of gene expression at different tissues for genes *OsACL1*, *OsADL1*, *OsAGO1a*, and *OsAGO7*.

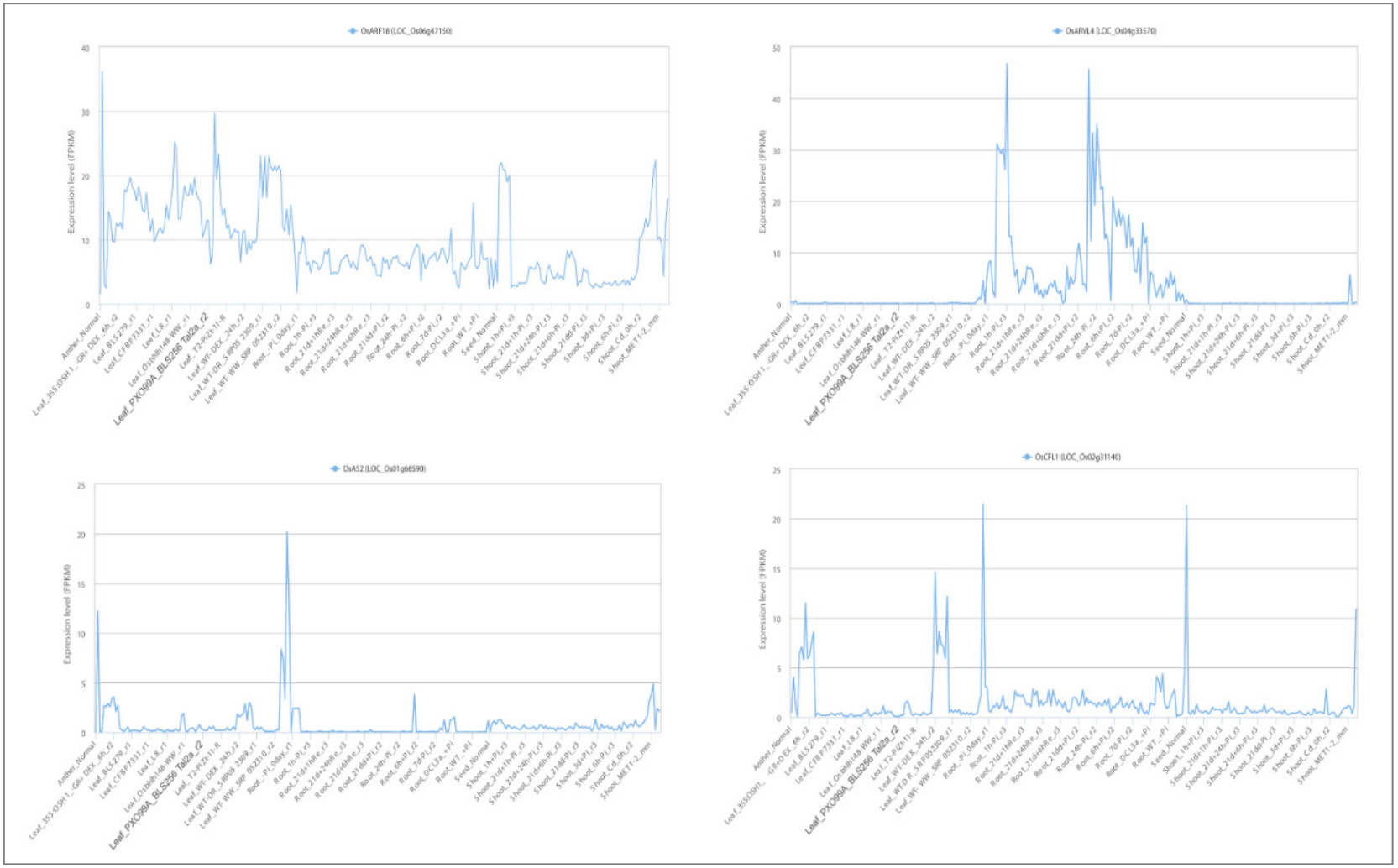


Figure A6.3: Line charts of gene expression at different tissues for genes *OsARF18*, *OsARVL4*, *OsAS2* and *OsCFL1*.

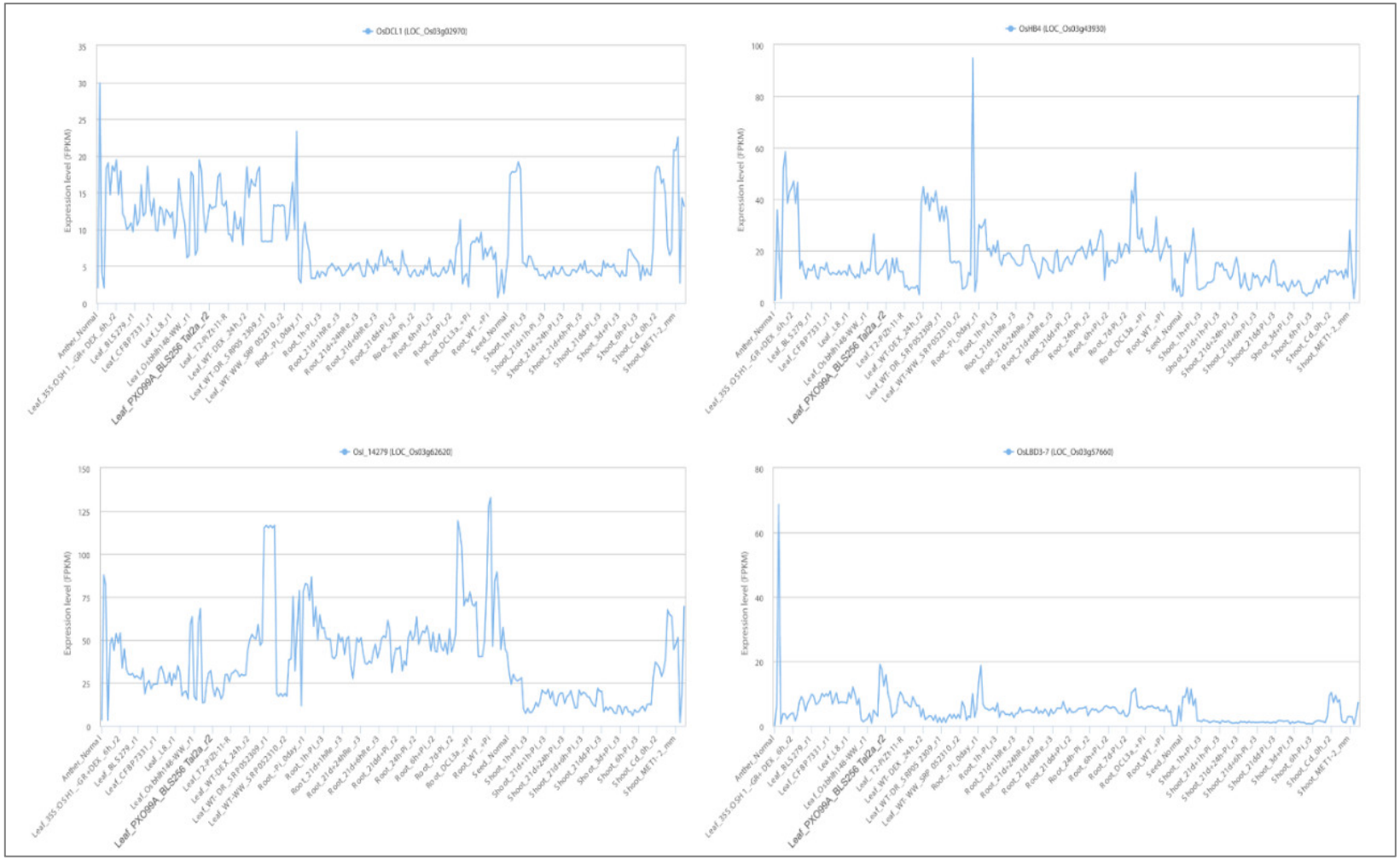


Figure A6.4: Line charts of gene expression at different tissues for genes *OsDCL1*, *OsHB4*, *OsI_14279* and *OsLBD3-7*.

Figure A6.5: Line charts of gene expression at different tissues for genes *OsLC2*, *OsLRRK1*, *OsMYB103L* and *OsNAL1*.

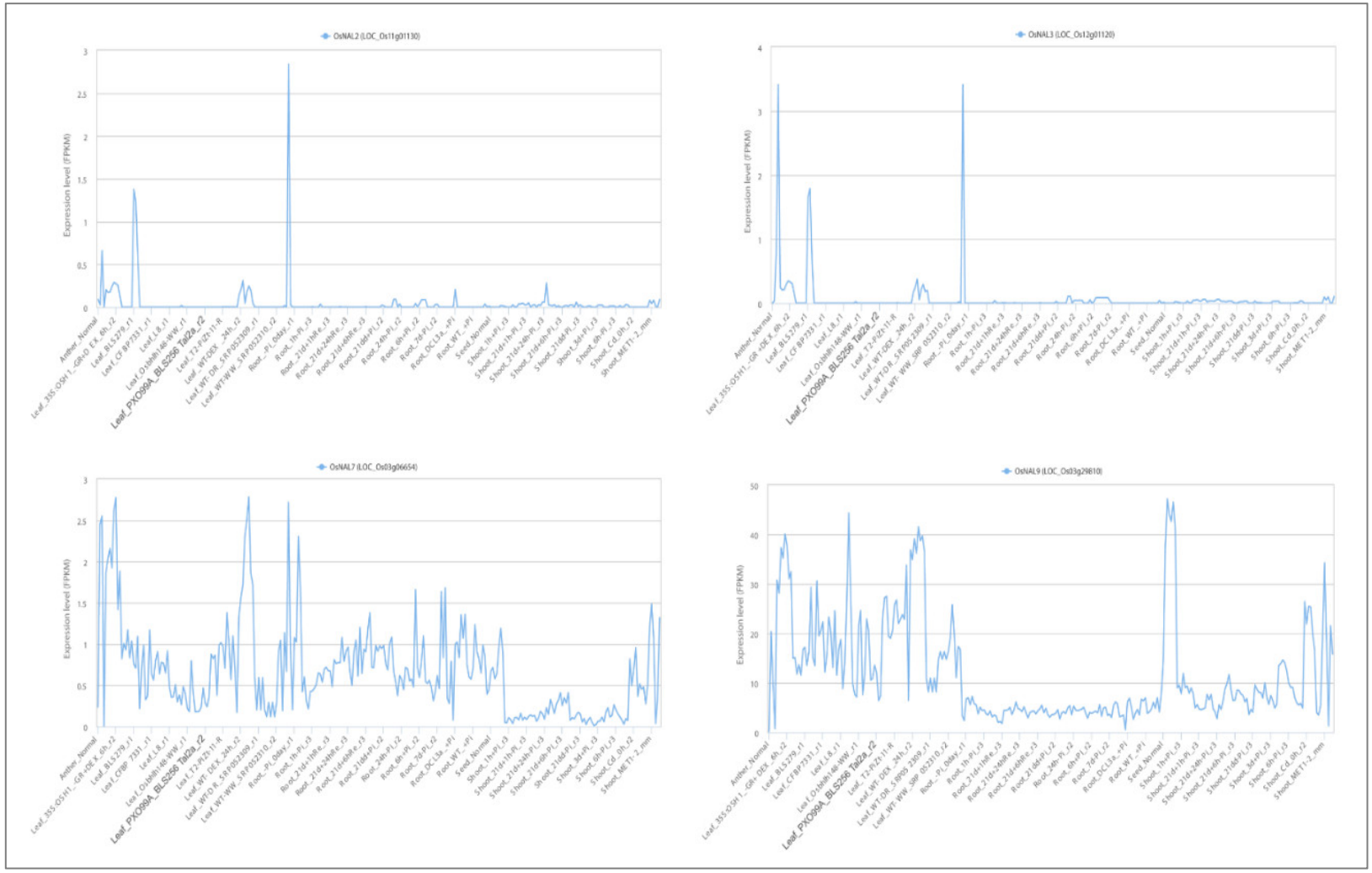


Figure A6.6: Line charts of gene expression at different tissues for genes *OsNAL2*, *OsNAL3*, *OsNAL7* and *OsNAL9*.

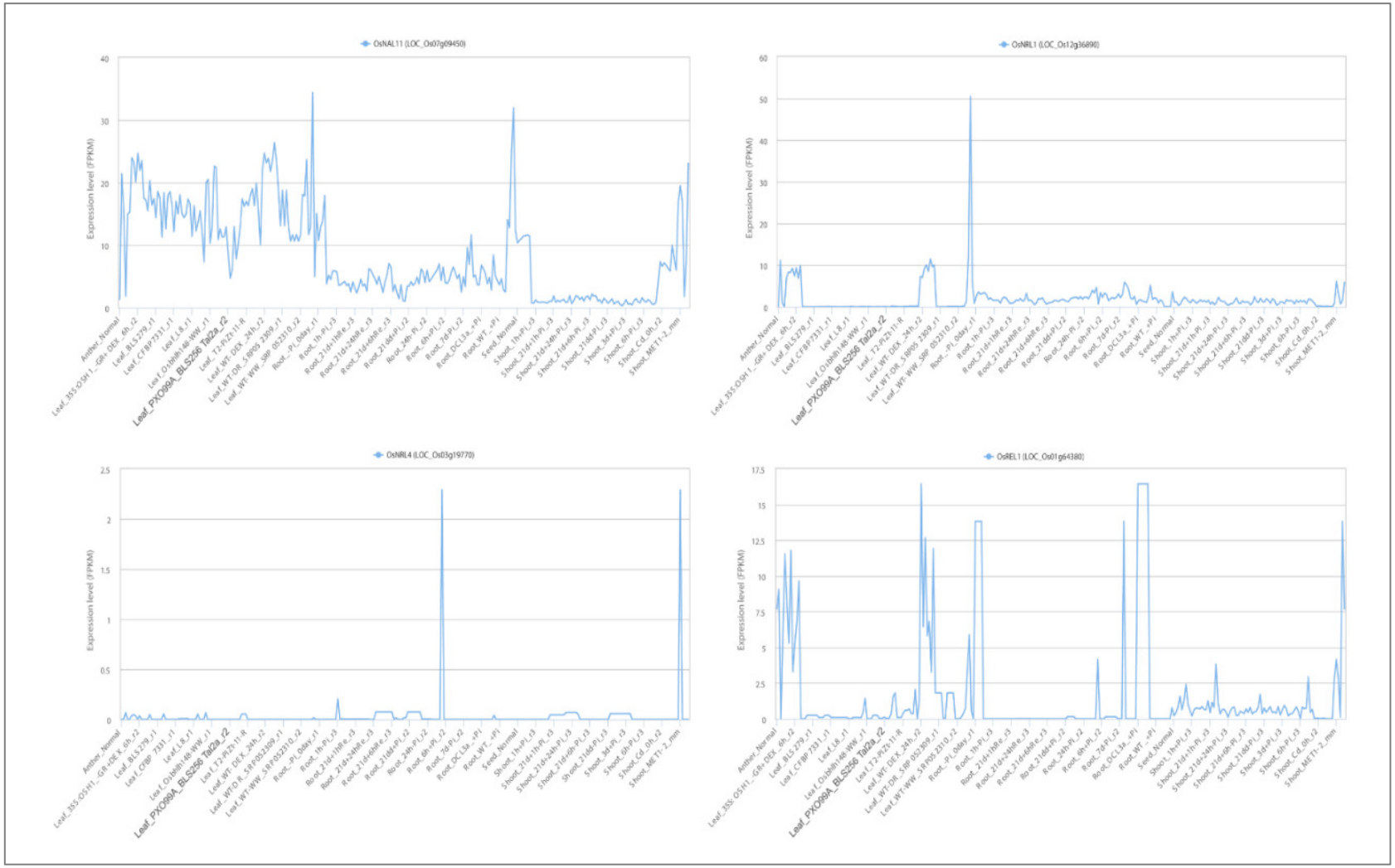


Figure A6.7: Line charts of gene expression at different tissues for genes *OsNAL11*, *OsNRL1*, *OsNRL4* and *OsREL1*.

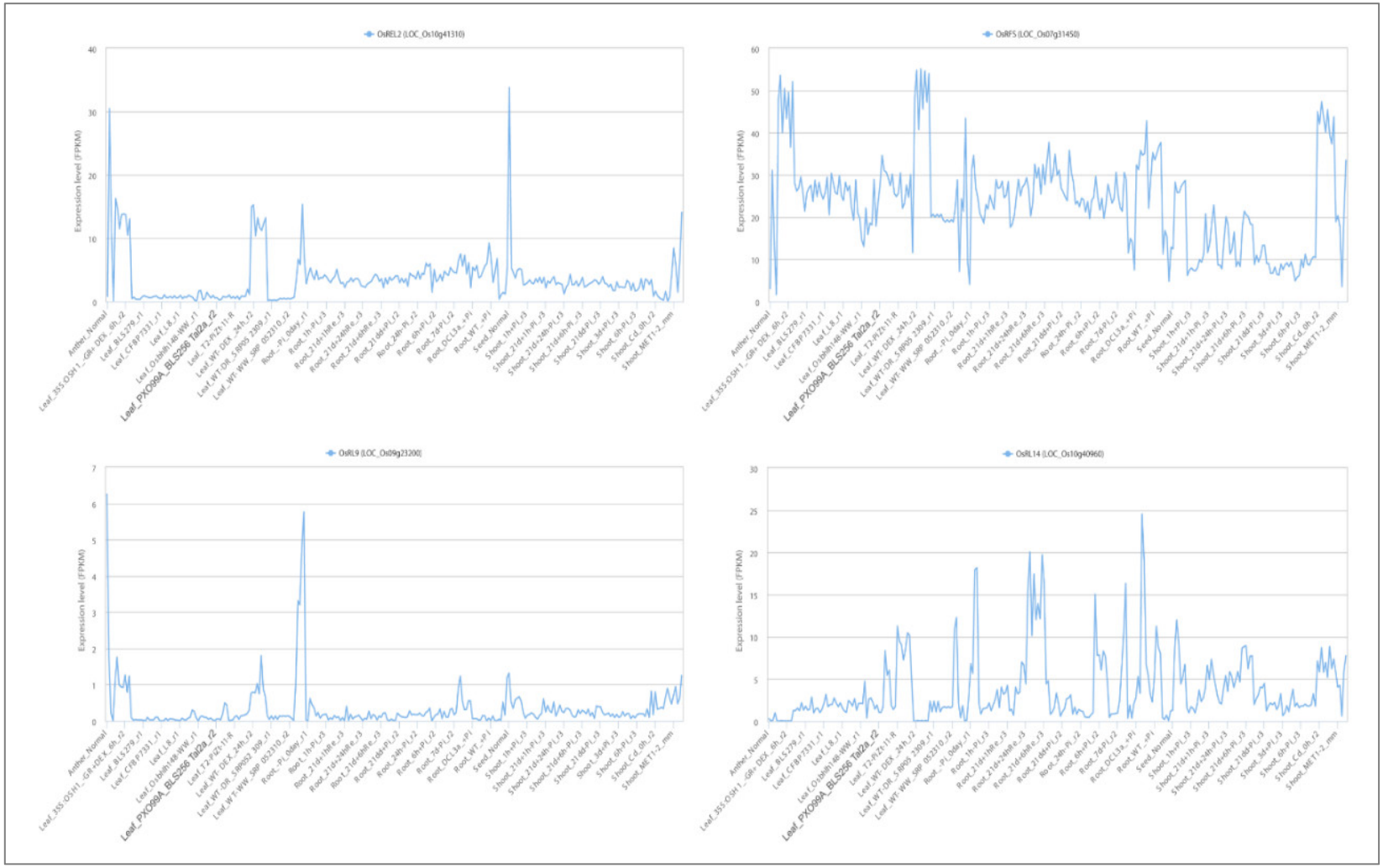


Figure A6.8: Line charts of gene expression at different tissues for genes *OsREL2*, *OsRFS*, *OsRL9* and *OsRL14*.

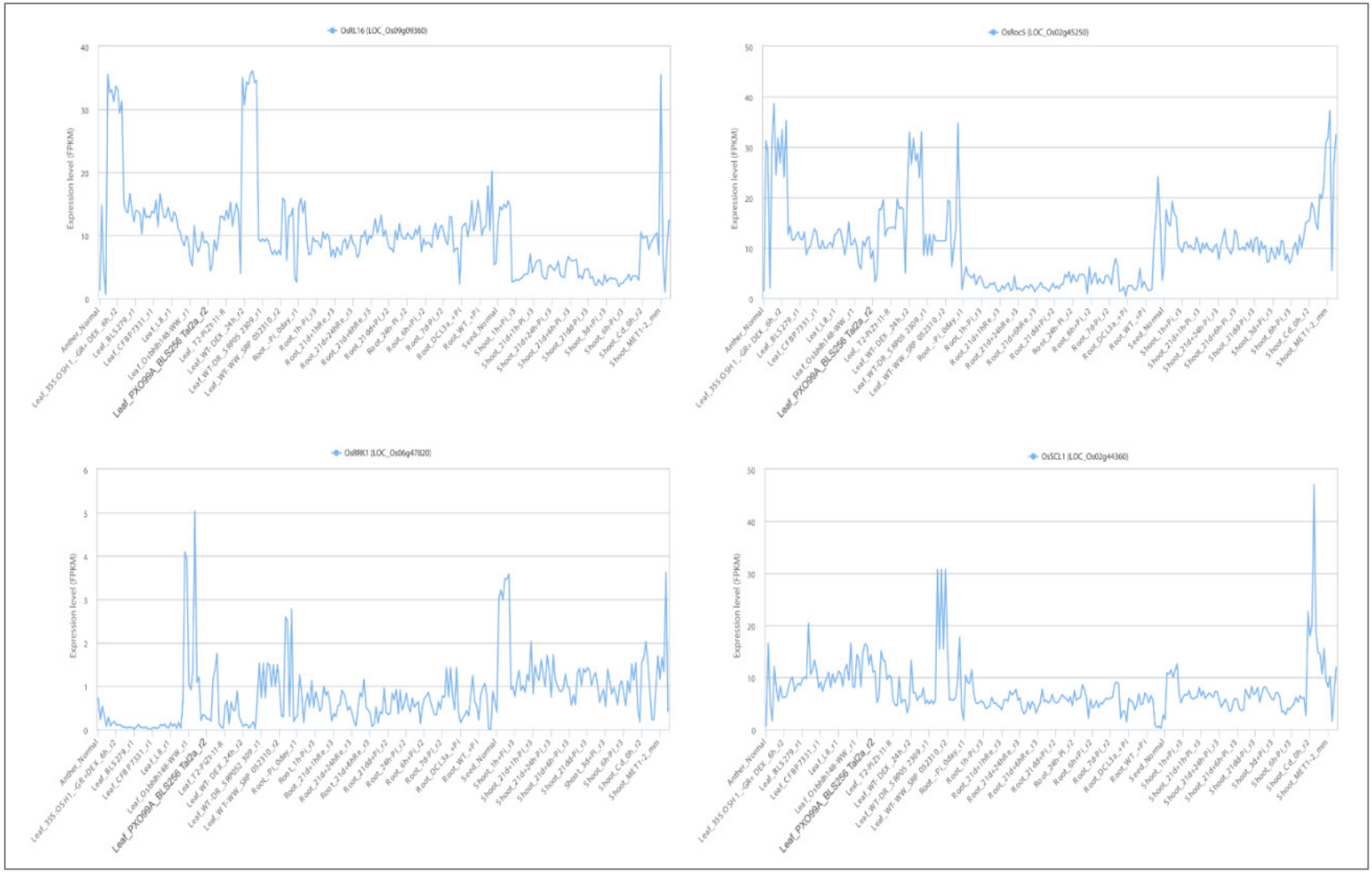


Figure A6.9: Line charts of gene expression at different tissues for genes *OsRL16*, *OsRoc5*, *OsRRK1* and *OsSCL1*.

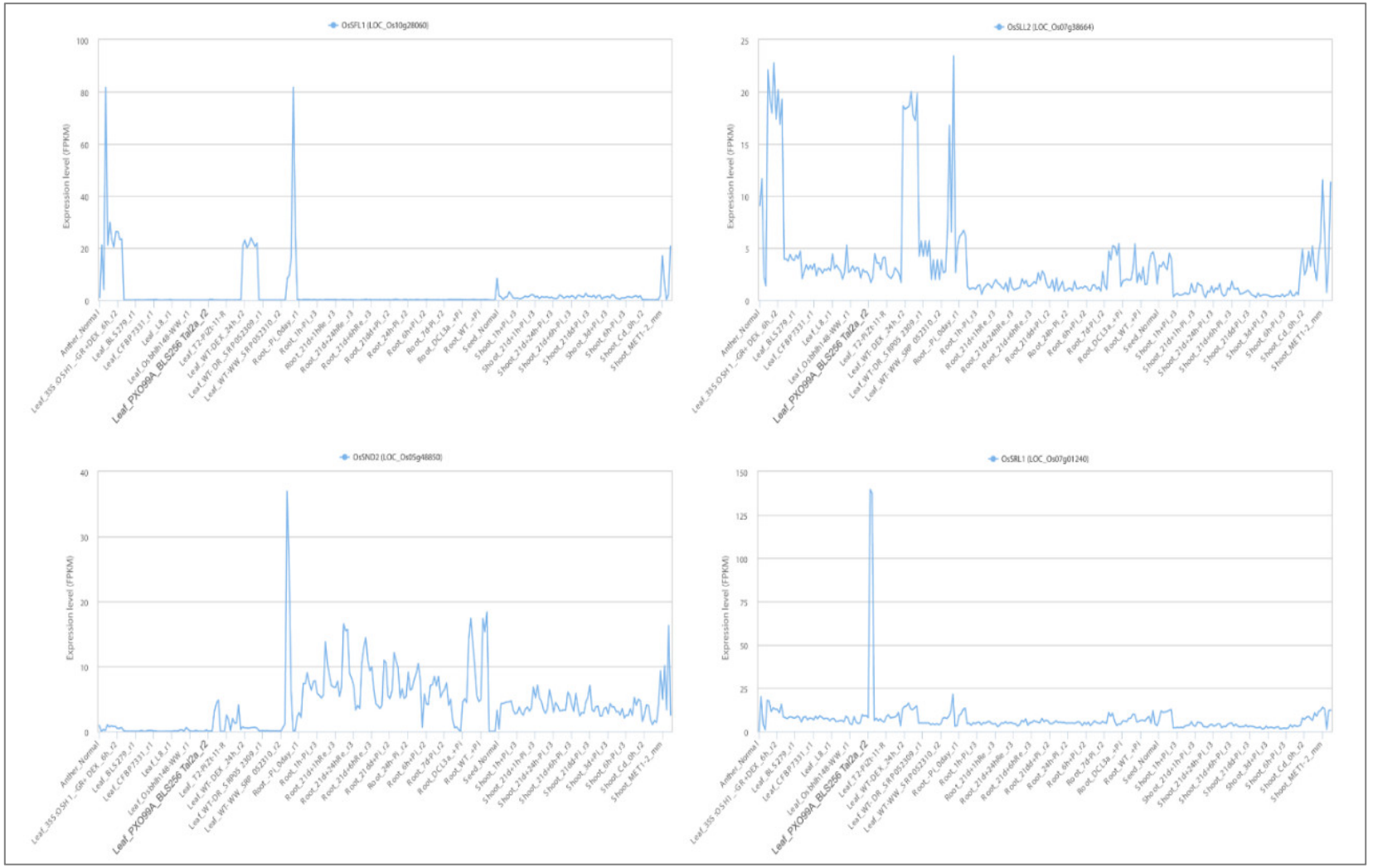


Figure A6.10: Line charts of gene expression at different tissues for genes *OsSFL1*, *OsSLL2*, *OsSND2* and *OsSRL1*.

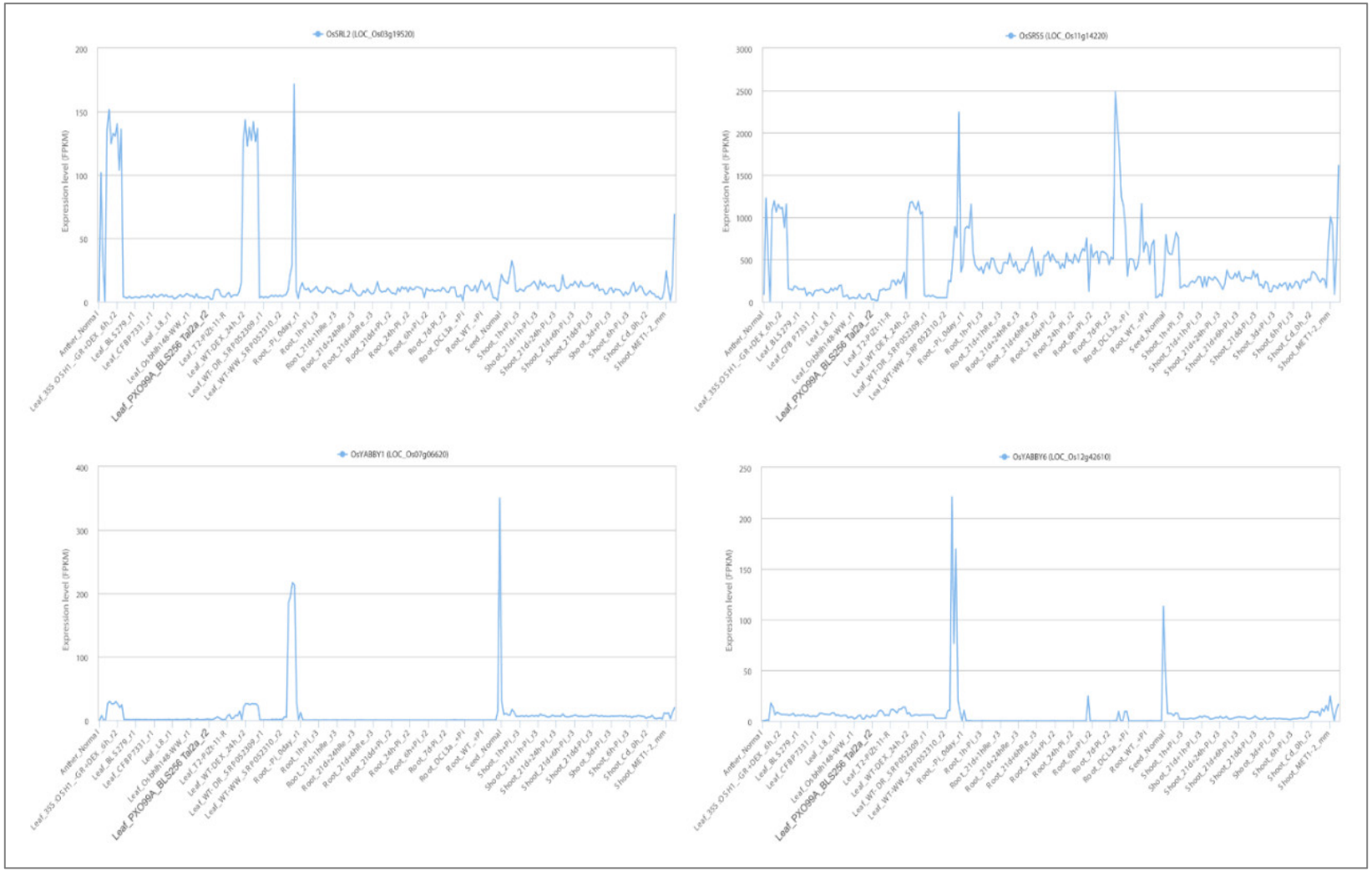


Figure A6.11: Line charts of gene expression at different tissues for genes *OsSRL2*, *OsSRS5*, *OsYABBY1* and *OsYABBY6*.

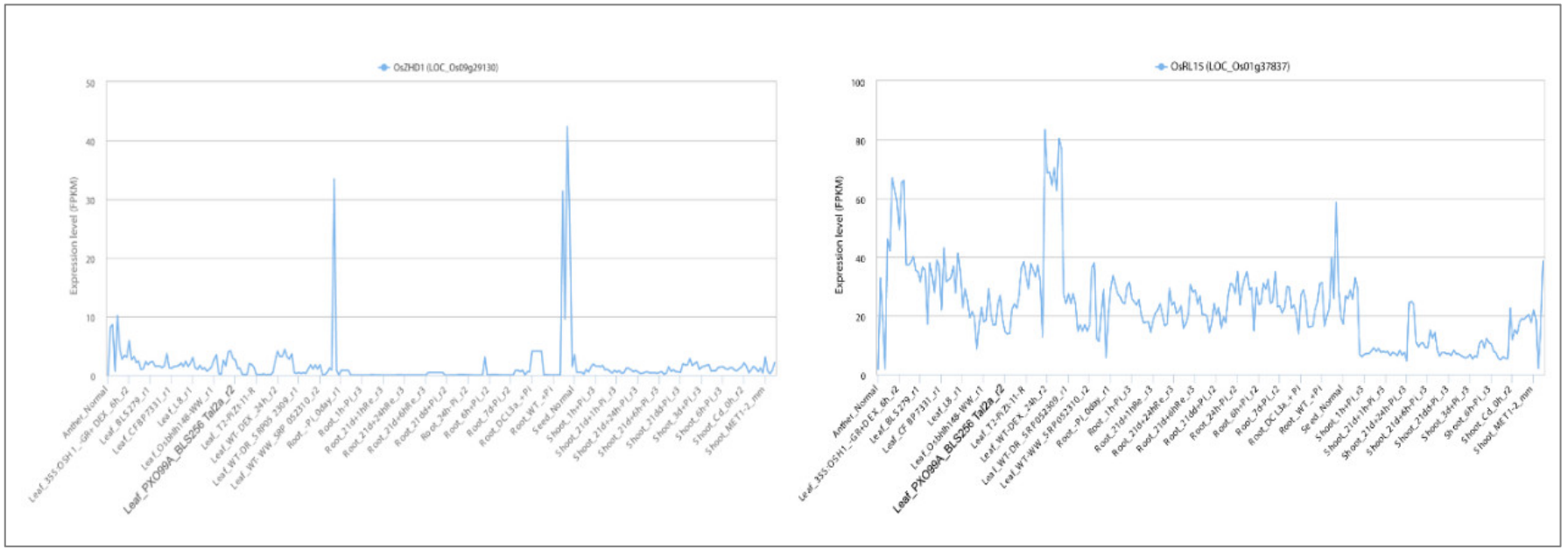


Figure A6.12: Line charts of gene expression at different tissues for genes *OsZHD1* and *OsRL15*.

A6.2.2 Identification of extreme (very high/low) gene expression at different tissues of rolling leaf (RL) genes using box plots

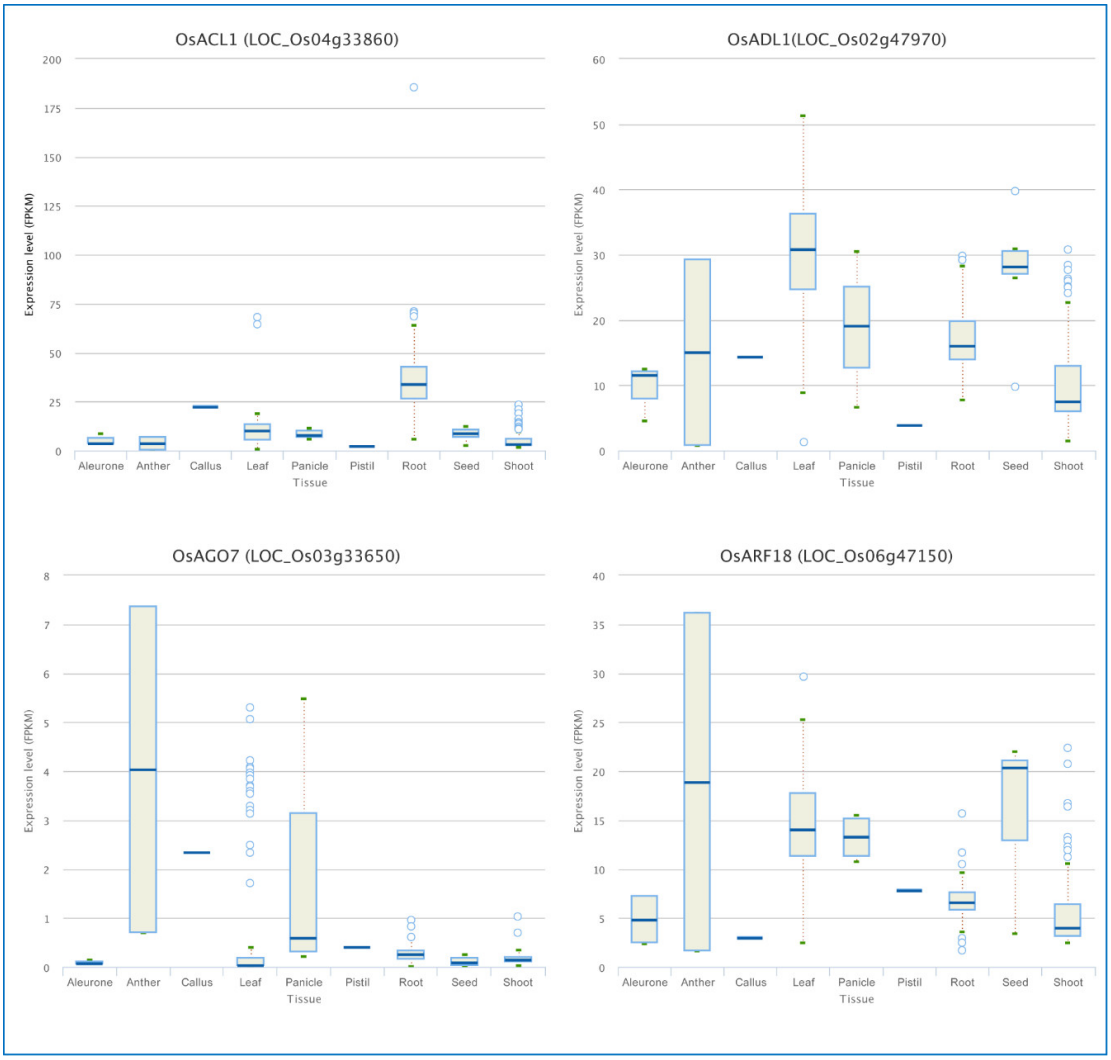


Figure A6.13: Box plot of gene expression at different tissues for genes *OsACL1*, *OsADL1*, *OsAGO7* and *OsARF18*.

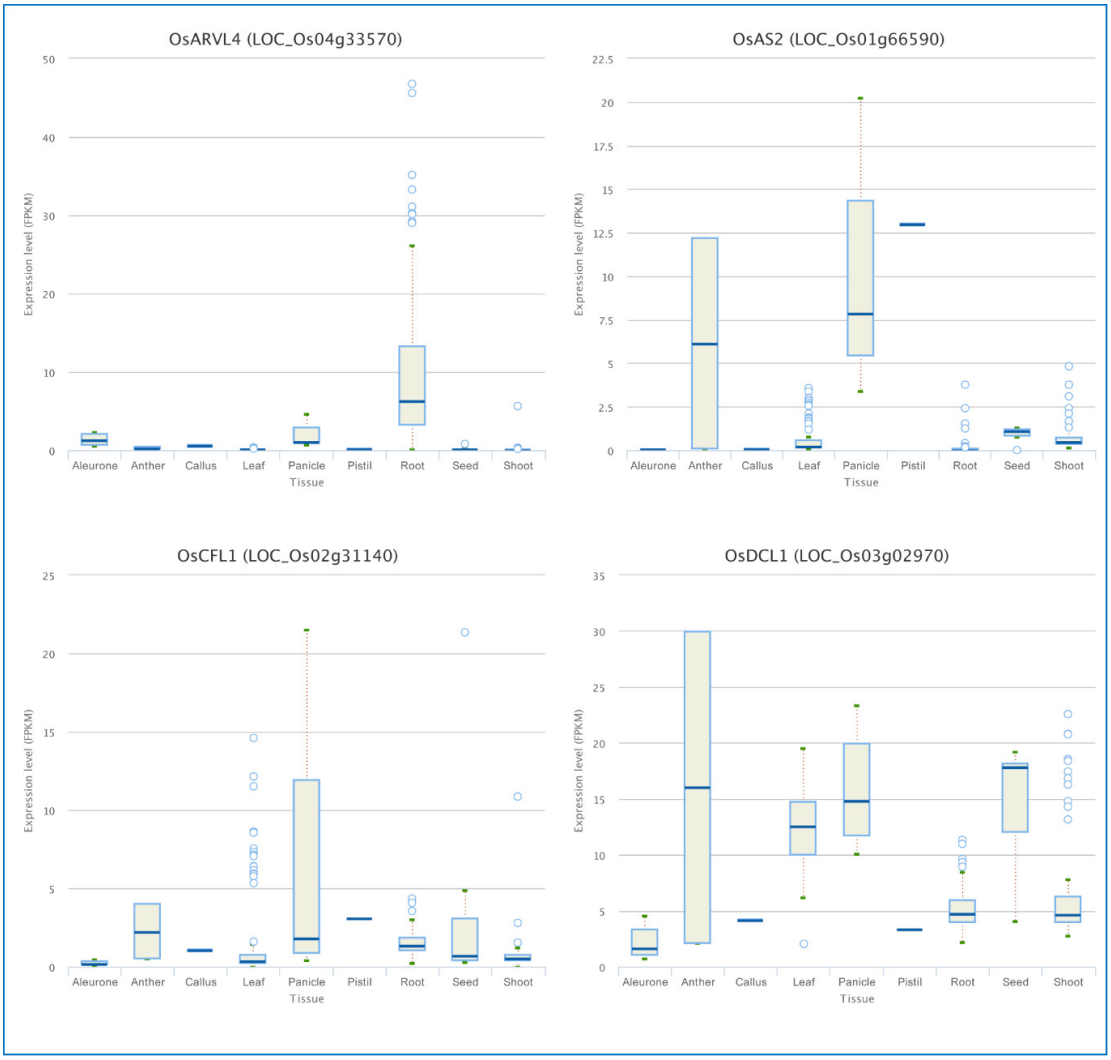


Figure A6.14: Box plot of gene expression at different tissues for gene *OsARVL4*, *OsAS2*, *OsCFL1* and *OsDCL1*.

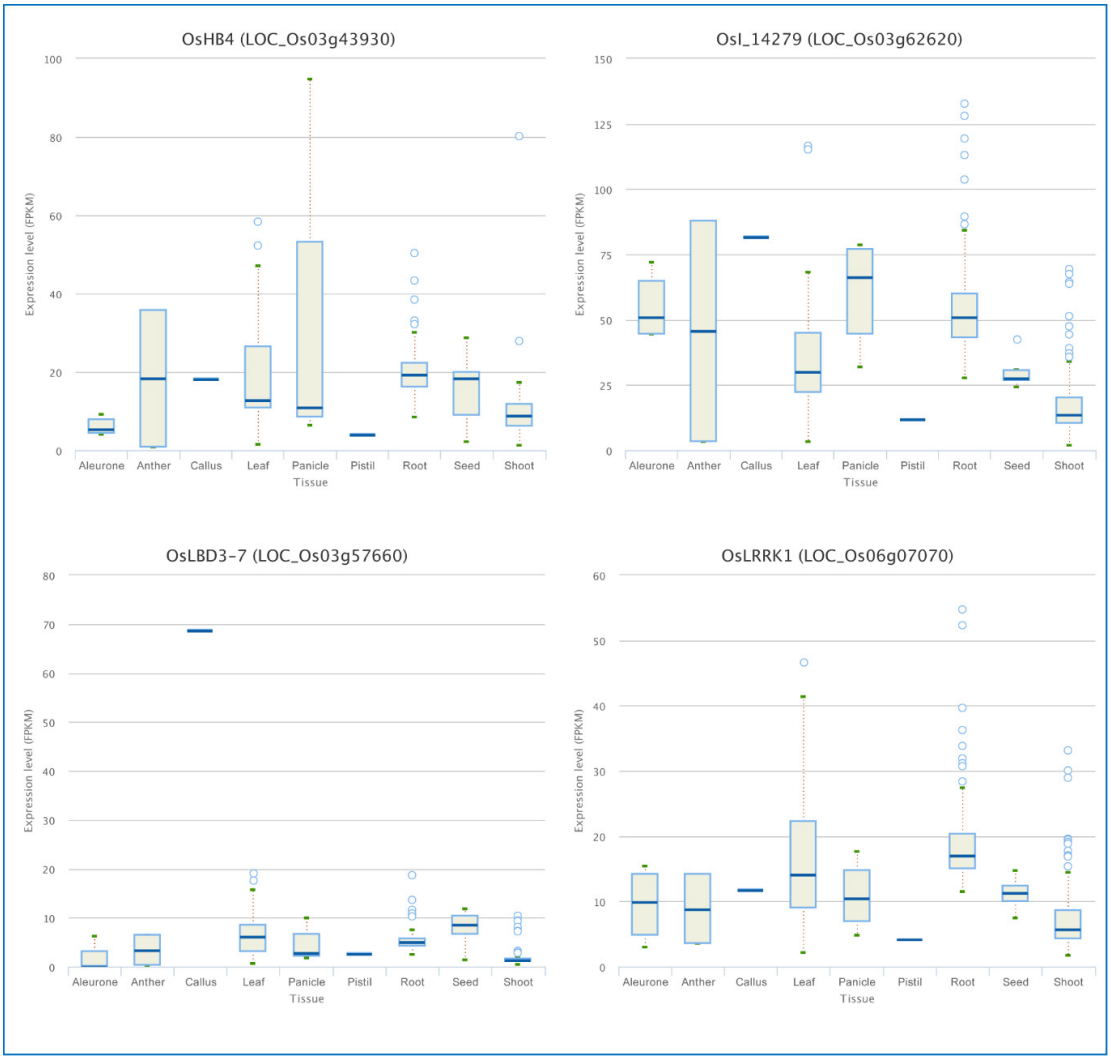


Figure A6.15: Box plot of gene expression at different tissues for genes *OsHB4*, *OsI_14279*, *OsLBD3-7* and *OsLRRK1*.

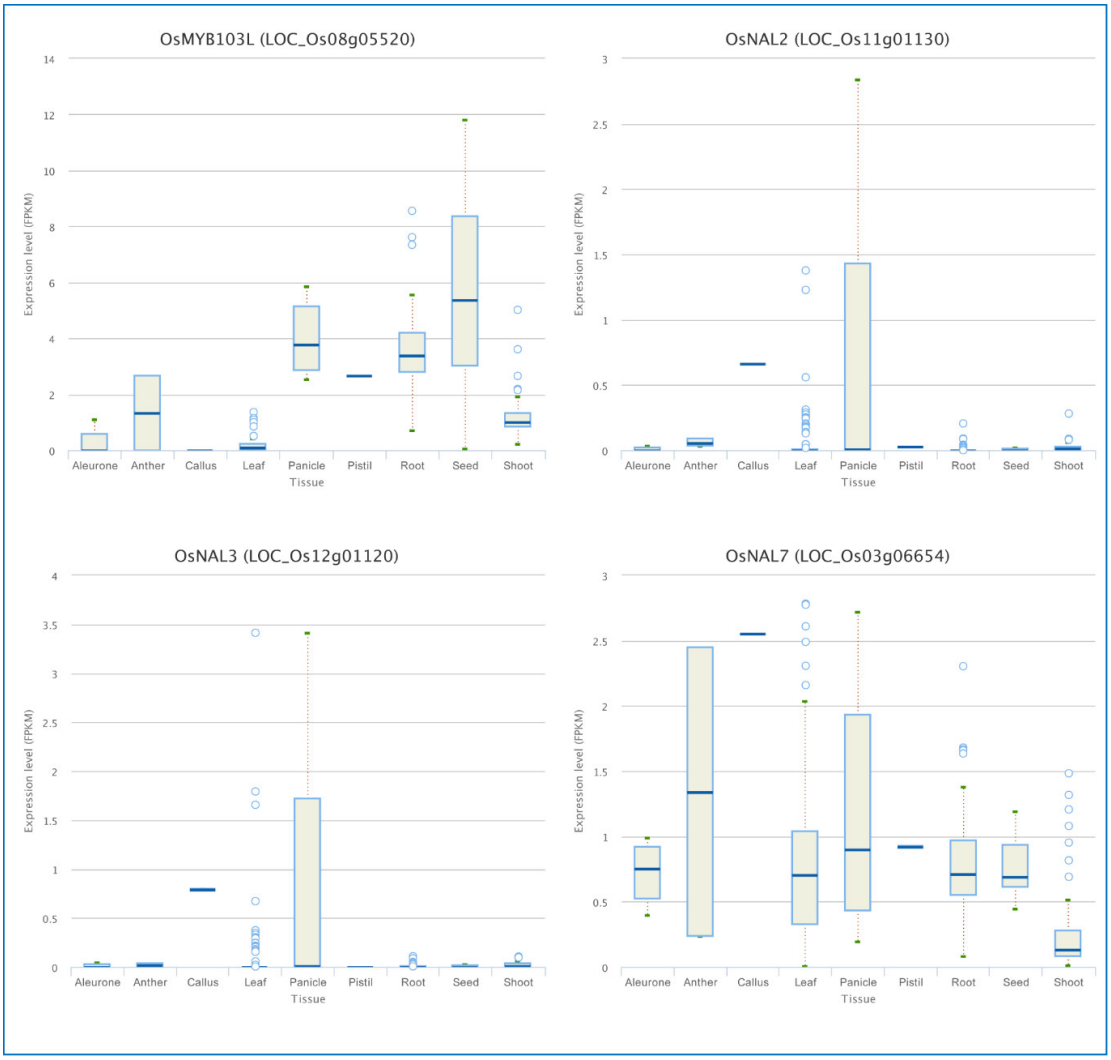


Figure A6.16: Box plot of gene expression at different tissues for gene *OsMYB103L*, *OsNAL2*, *OsNAL3* and *OsNAL7*.

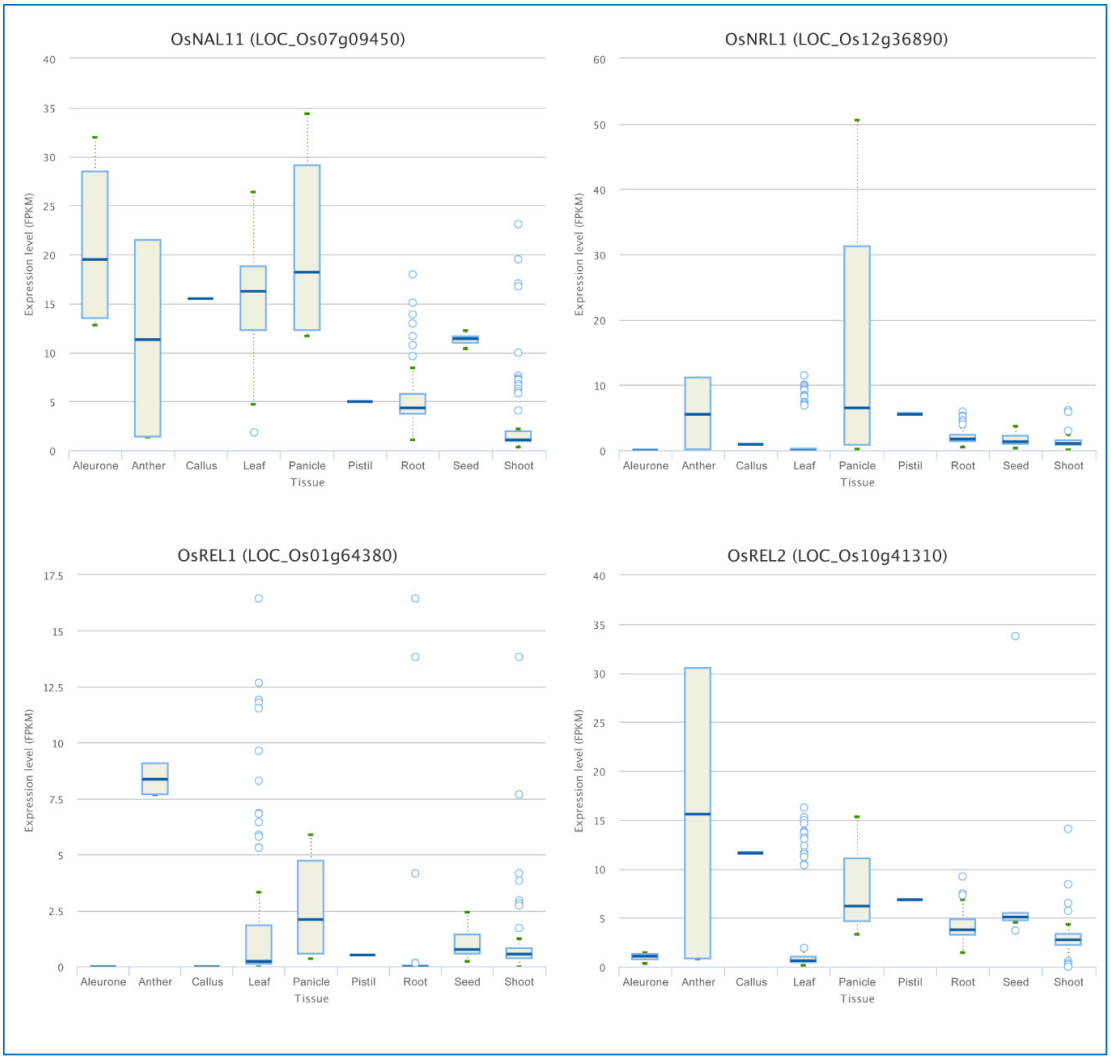


Figure A6.17: Box plot of gene expression at different tissues for genes *OsNAL11*, *OsNRL1*, *OsREL1* and *OsREL2*.

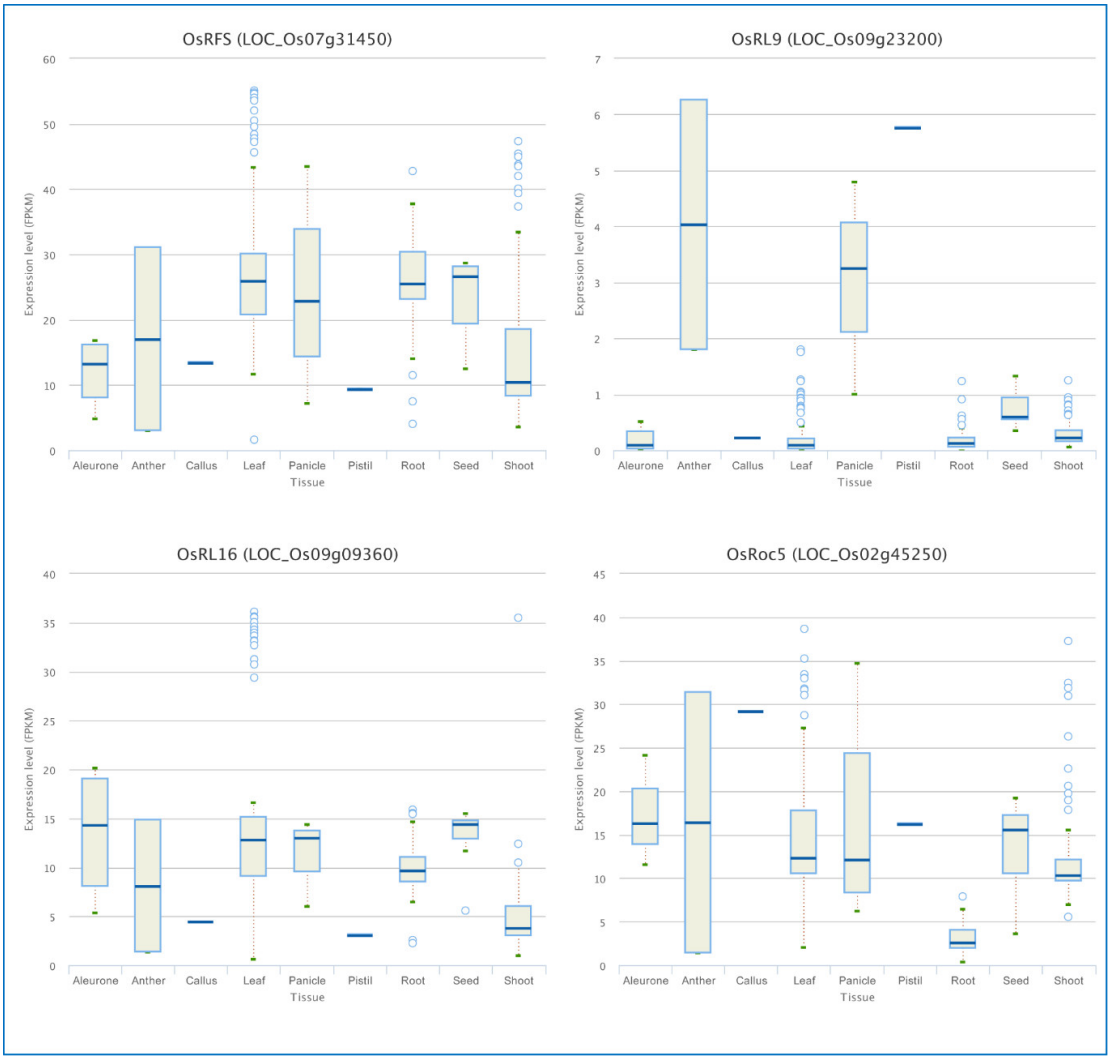


Figure A6.18: Box plot of gene expression at different tissues for genes *OsRFS*, *OsRL9*, *OsRL16* and *OsRoc5*.

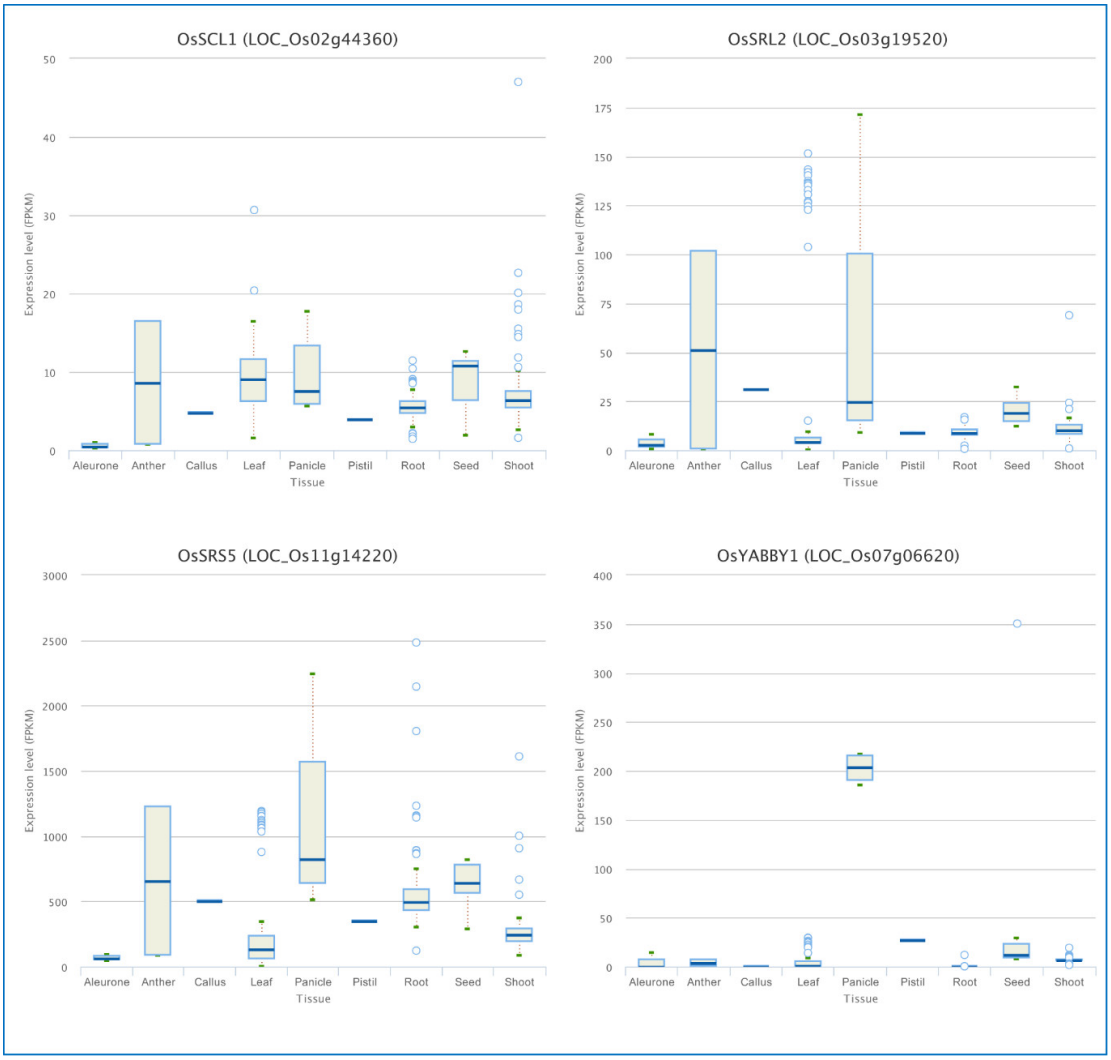


Figure A6.19: Box plot of gene expression at different tissues for genes *OsSCL1*, *OsSRL2*, *OsSRS5* and *OsYABBY1*.

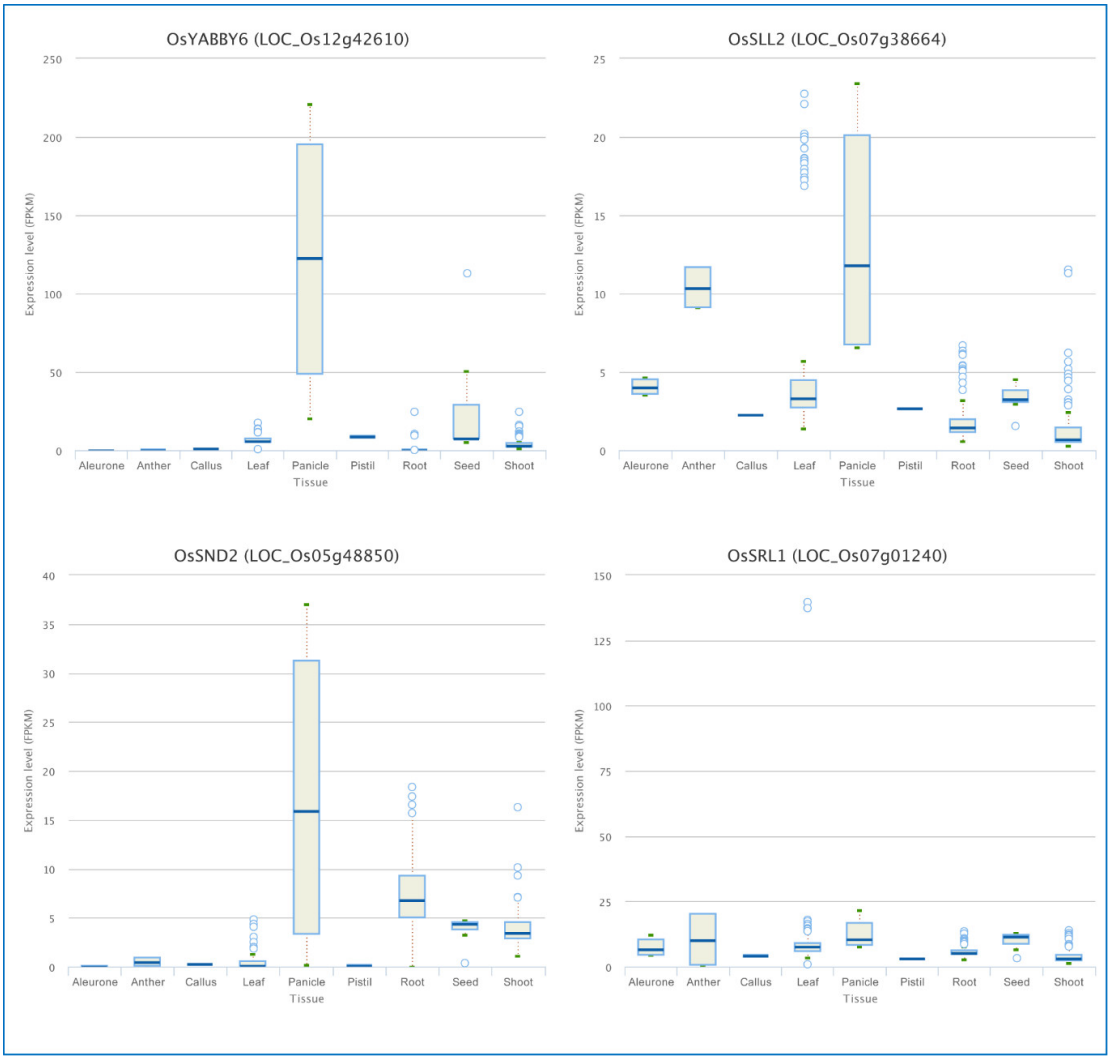


Figure A6.20: Box plot of gene expression at different tissues for genes *OsYABBY6*, *OsSLL2*, *OsSND2* and *OsSRL1*.

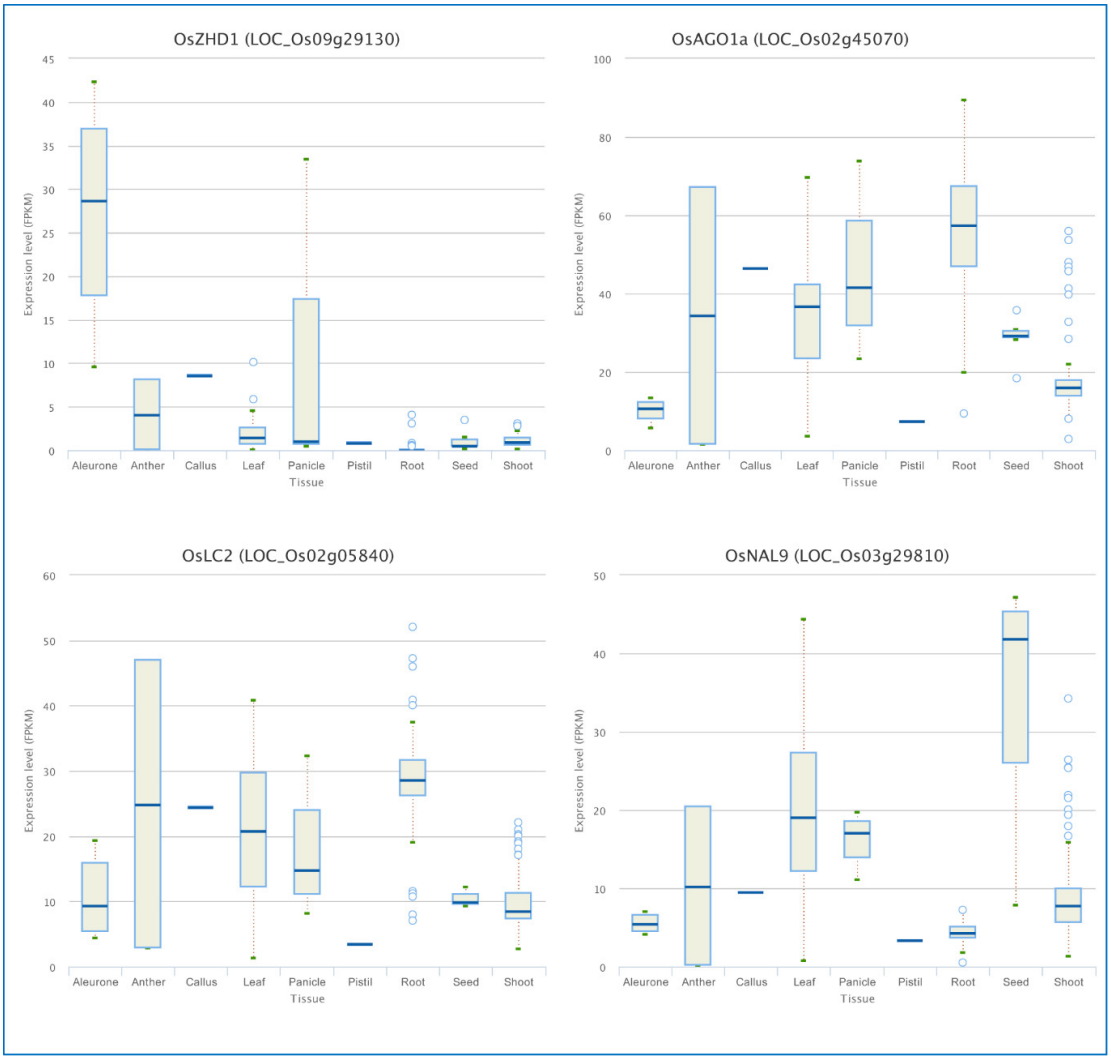


Figure A6.21: Box plot of gene expression at different tissues for genes *OsZHD1*, *OsAGO1a*, *OsLC2*, and *OsNAL9*.

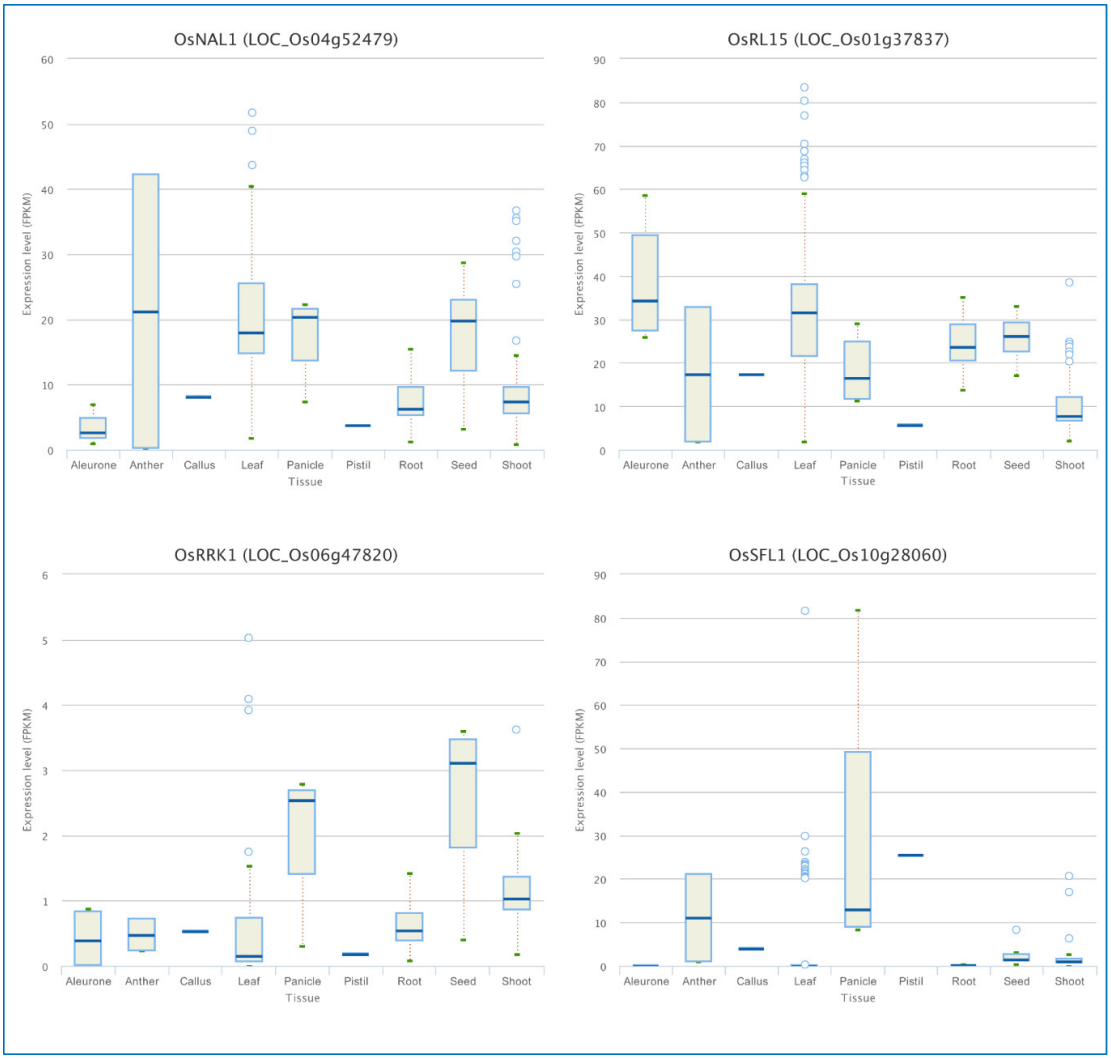


Figure A6.22: Box plot of gene expression at different tissues for gene *OsNAL1*, *OsRL15*, *OsRRK1* and *OsSFL1*.

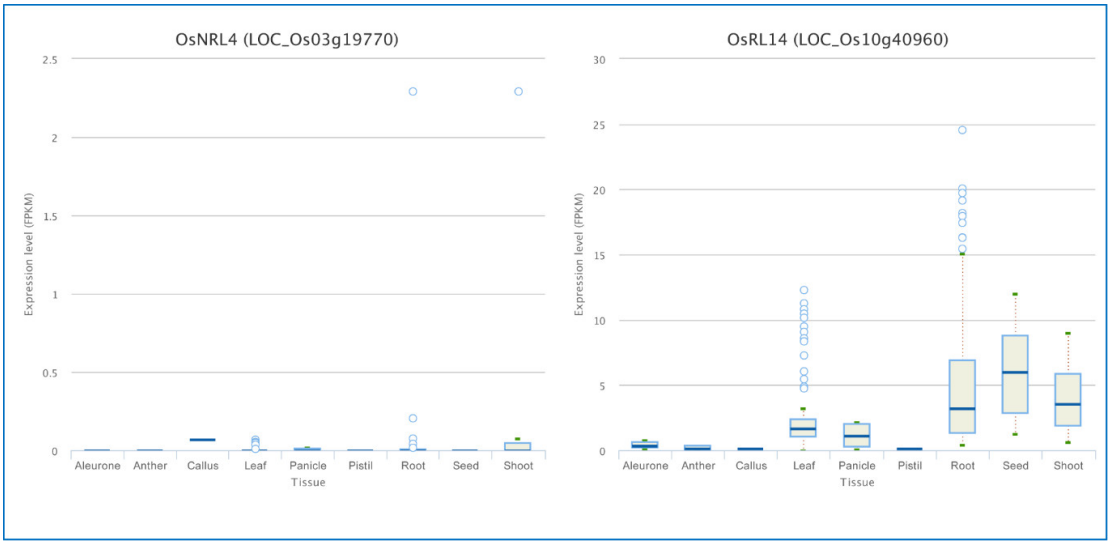


Figure A6.23: Box plots of gene expression at different tissues for gene *OsNRL4* and *OsRL14*.

A6.3 Supplementary Tables of Chapter 6

Table A6.1: Conserved domain analysis of the identified 42 rolling leaf genes of interest using NCBI

Gene Name	Domain Name	Accession	Description	Interval	E-value
OsACL1	No conserved domain found		No conserved domain have been identified for this query sequence		
OsADL1	Peptidase_C2	pfam00648	Calpain family cysteine protease	1707-2006	8.65e-120
	Calpain_III super family	cl00165	Calpain, subdomain III	2012-2162	2.78e-35
	LamG super family	cl22861	Laminin G domain	1436-1596	9.49e-06
OsAGO1a	Piwi-like super family	cl00628	Piwi-like: PIWI domain	197-1082	0e+00
	Gly-rich_Ago1	pfam12764	Glycine-rich region of argonaut	98-200	4.94e-44
OsAGO7	Piwi_ago-like	cd04657	Piwi_ago-like: PIWI domain, Argonaute-like subfamily.	575-1016	0e+00
	ArgoN	pfam16486	N-terminal domain of argonaute	197-352	1.52e-31
	PAZ	pfam02170	PAZ domain; This domain is named PAZ after the proteins Piwi Argonaut and Zwille.	438-544	2.96e-29
	ArgoL1	pfam08699	Argonaute linker 1 domain. ArgoL1 is a region found in argonaute proteins.	363-412	1.38e-11
	PLN03202	PLN03202	protein argonaute; Provisional	194-1048	6.95e-151
OsARF18	Auxin_resp	pfam06507	Auxin response factor. A conserved region of auxin-responsive transcription factors.	292-375	4.85e-45
	B3	pfam02362	B3 DNA binding domain. This is a family of plant transcription factors with various roles in development.	128-229	5.07e-28
OsARVL4	PEBP super family	cl00227	Phosphatidyl Ethanolamine-Binding Protein (PEBP) domain. Phosphatidyl Ethanolamine-Binding Proteins (PEBPs) are represented in all three major phylogenetic divisions (eukaryotes, bacteria, archaea). A number of biological roles for members of the PEBP family include serine protease inhibition, membrane biogenesis, regulation of flowering plant stem architecture, and Raf-1 kinase inhibition.	1-173	9.05e-82

Gene Name	Domain Name	Accession	Description	Interval	E-value
OsAS2	DUF260	pfam03195	Protein of unknown function DUF260.	38-135	4.04e-63
OsCFL1	WW	pfam00397	WW domain. The WW domain is a protein module with two highly conserved tryptophans that binds proline-rich peptide motifs in vitro.	56-86	1.37e-03
OsDCL1	PAZ_CAF_like	cd02844	PAZ domain, CAF_like subfamily.	1152-1298	1.04e-61
	Rnc	COG0571	dsRNA-specific ribonuclease [Transcription]	1538-1779	6.20e-46
	helicase_insert_domain super family	cl17041	helicase_insert_domain. This helical domain can be found inserted in a subset of SF2-type DEAD-box related helicases, like archaeal Hef helicase, MDA5-like helicases and FancM-like helicases.	251-779	6.79e-42
	Dicer_dimer	pfam03368	Dicer dimerisation domain.	817-906	1.45e-30
	RIBOc	cd00593	RIBOc. Ribonuclease III C terminal domain.	1337-1518	1.97e-30
	DSRM super family	cl00054	Double-stranded RNA binding motif.	1797-1870	3.44e-15
OsHB4	START_ArGLABRA2_like	cd08875	C-terminal lipid-binding START domain of the Arabidopsis homeobox protein GLABRA 2 and related proteins; This subfamily includes the steroidogenic acute regulatory protein (StAR)-related lipid transfer (START) domains of the Arabidopsis homeobox protein GLABRA 2 and related proteins.	175-391	1.47e-69
	MEKHLA	pfam08670	MEKHLA domain; The MEKHLA domain shares similarity with the PAS domain and is found in the 3' end of plant HD-ZIP III homeobox genes, and bacterial proteins.	715-857	2.34e-65
	Homeobox	pfam00046	Homeobox domain	32-89	4.52e-17
	bZIP	cd14686	Basic leucine zipper (bZIP) domain of bZIP transcription factors: a DNA-binding and dimerization domain; Basic leucine zipper (bZIP) factors comprise one of the most important classes of enhancer-type transcription factors.	84-123	2.04e-06
OsI_14279	WHy	smart00769	Water Stress and Hypersensitive response;	61-155	2.23e-23
	LEA_2	pfam03168	Late embryogenesis abundant protein; Different types of LEA proteins are expressed at different stages of late embryogenesis in higher plant seed embryos and under conditions of dehydration stress. The function of these proteins is unknown.	206-301	6.86e-18
OsLBD3-7	M28_PSMA_like	cd08022	M28 Zn-peptidase prostate-specific membrane antigen; Peptidase M28	325-549	3.34e-116

Gene Name	Domain Name	Accession	Description	Interval	E-value
			family; prostate-specific membrane antigen (PSMA, also called glutamate carboxypeptidase II or GCP-II)-like subfamily.		
	PA super family	cl28883	PA: Protease-associated (PA) domain.	118-310	1.80e-34
	TFR_dimer	pfam04253	Transferrin receptor-like dimerisation domain; This domain is involved in dimerisation of the transferrin receptor as shown in its crystal structure.	578-694	1.55e-17
OsLC2	PHD_Oberon	pfam07227	PHD - plant homeodomain finger protein; PHD_oberon is a plant homeodomain finger domain of Oberon proteins from plants.	143-263	1.18e-55
	FN3	cd00063	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein fibronectin.	343-424	2.12e-03
OsLRRK1	STKc_IRAK	cd14066	Catalytic domain of the Serine/Threonine kinases, Interleukin-1 Receptor Associated Kinases and related STKs; STKs catalyze the transfer of the gamma-phosphoryl group from ATP to serine/threonine residues on protein substrates.	54-321	1.87e-92
OsMYB103L	Myb_DNA-binding	pfam00249	Myb-like DNA-binding domain. This family contains the DNA binding domains from Myb proteins, as well as the SANT domain family.	67-112	6.27e-16
	Myb_DNA-binding	pfam00249	Myb-like DNA-binding domain. This family contains the DNA binding domains from Myb proteins, as well as the SANT domain family.	14-61	1.29e-15
	SANT	smart00717	SANT SWI3, ADA2, N-CoR and TFIIIB" DNA-binding domains	67-114	5.92e-15
	SANT	smart00717	SANT SWI3, ADA2, N-CoR and TFIIIB" DNA-binding domains	14-63	1.69e-12
OsNAL1	No conserved domain found		No conserved domain have been identified for this query sequence		
OsNAL11	DnaJ super family	cl02542	DnaJ domain or J-domain. DnaJ/Hsp40 (heat shock protein 40) proteins are highly conserved and play crucial roles in protein translation, folding, unfolding, translocation, and degradation.	1-103	1.19e-22
OsNAL2	Homeobox	pfam00046	Homeobox domain	7-61	7.68e-13
OsNAL3/ OsWOX3	Homeobox	pfam00046	Homeobox domain	7-61	1.39e-12
OsNAL7/	Pyr_redox_3	pfam13738	Pyridine nucleotide-disulphide oxidoreductase	27-220	6.31e-17

Gene Name	Domain Name	Accession	Description	Interval	E-value
OsCOW1	NADB_Rossmann super family	cl21454	Rossmann-fold NAD(P)(+)-binding proteins; A large family of proteins that share a Rossmann-fold NAD(P)H/NAD(P)(+) binding (NADB) domain.	334-386	1.86e-04
	CzcO	COG2072	Predicted flavoprotein CzcO associated with the cation diffusion facilitator CzcD [Inorganic ion transport and metabolism]	27-351	2.86e-50
OsNAL9	S14_ClpP_2	cd07017	Caseinolytic protease (ClpP) is an ATP-dependent, highly conserved serine protease.	84-252	4.11e-78
OsNRL1	Glyco_tranf_GTA_type super family	cl11394	Glycosyltransferase family A (GT-A) includes diverse families of glycosyl transferases with a common GT-A type structural fold.	643-956	7.14e-19
	RING super family	cl17238	RING-finger (Really Interesting New Gene) domain.	175-202	1.72e-06
	PLN02248	PLN02248	Cellulose synthase-like protein D4 (CLSD4)	22-1215	0e+00
OsNRL4	No conserved domain found		No conserved domain have been identified for this query sequence		
OsREL1	No conserved domain found		No conserved domain have been identified for this query sequence		
OsREL2	DUF632	pfam04782	Protein of unknown function (DUF632).	325-631	2.07e-113
	DUF630	pfam04783	Protein of unknown function (DUF630).	1-59	2.74e-27
OsRFS	SNF2_N super family	cl26465	SNF2 family N-terminal domain.	595-1123	1.33e-136
	PHD2_CHD_II	cd15532	PHD finger 2 found in class II Chromo domain-Helicase-DNA binding (CHD) proteins.	35-76	3.70e-22
	SANT_TRF	cd11660	Telomere repeat binding factor-like DNA-binding domains of the SANT/myb-like family.	1645-1689	2.10e-08
	DUF1087 super family	cl05792	Domain of Unknown Function (DUF1087).	1206-1255	5.49e-08
	CHROMO	cd00024	Chromatin organization modifier (chromo) domain.	476-526	8.88e-08
	Atrophin-1 super family	cl26464	Atrophin-1 family.	1907-2192	7.00e-06
	Chromo	pfam00385	Chromo (CHRromatin Organisation MOdifier) domain	533-554	7.16e-04
OsRL14	2OG-FeII_Oxy	pfam03171	2OG-Fe(II) oxygenase superfamily	1-93	3.84e-32
OsRL15	PLN02678	PLN02678	seryl-tRNA synthetase	1-445	0e+00
OsRL16	PGAP1	pfam07819	PGAP1-like protein; The sequences found in this family are similar to PGAP1.	79-347	2.23e-101

Gene Name	Domain Name	Accession	Description	Interval	E-value
OsRL9/OsSLL1	myb_SHAQKYF	TIGR01557	myb-like DNA-binding domain, SHAQKYF class.	325-378	1.58e-21
OsRoc5	START_ArGLABRA2_like	cd08875	C-terminal lipid-binding START domain	309-545	8.42e-115
	Homeobox	pfam00046	Homeobox domain	101-154	2.19e-22
	bZIP super family	cl21462	Basic leucine zipper (bZIP) domain of bZIP transcription factors: a DNA-binding and dimerization domain	136-181	4.90e-04
OsRRK1	PKc_like super family	cl21453	Protein Kinases, catalytic domain; The protein kinase superfamily is mainly composed of the catalytic domains of serine/threonine-specific and tyrosine-specific protein kinases. It also includes RIO kinases, which are atypical serine protein kinases, aminoglycoside phosphotransferases, and choline kinases.	75-339	9.36e-87
OsSCL1	GRAS super family	cl15987	GRAS domain family	363-708	6.47e-92
OsSFL1	PLN03169 super family	cl28398	Chalcone synthase family protein; Provisional	96-514	0e+00
OsSLL2	No conserved domain found		No conserved domain have been identified for this query sequence		
OsSND2	NAM	pfam02365	No apical meristem (NAM) protein. This is a family of no apical meristem (NAM) proteins these are plant development proteins.	179-206	3.09e-03
OsSRL1	No conserved domain found		No conserved domain have been identified for this query sequence		
OsSRL2	No conserved domain found		No conserved domain have been identified for this query sequence		
OsSRS5	PLN00221	PLN00221	Tubulin alpha chain; Provisional	1-438	0e+00
OsYABBY1	YABBY	pfam04690	YABBY protein; YABBY proteins are a group of plant-specific transcription involved in the specification of abaxial polarity in lateral organs.	5-146	3.37e-76
OsYABBY6	YABBY	pfam04690	YABBY protein; YABBY proteins are a group of plant-specific transcription involved in the specification of abaxial polarity in lateral organs.	6-176	1.39e-95
OsZHD1	ZF-HD_dimer	pfam04770	ZF-HD protein dimerization region.	56-108	3.03e-35
	homeo_ZF_HD	TIGR01565	homeobox domain, ZF-HD class.	215-271	1.97e-26

Table A6.2: The enriched GO terms for all rolling leaf genes identified in this study

GO term	Ontology	Description	Genes	Number in input list	*P*-value
GO:0007275	BP	multicellular organismal development	*OsAS2, OsSCL1, OsAGO1a, OsRoc5, OsADL1, OsHB4, OsLBD3-7, OsSND2, OsARF18, OsYABBY1, OsRL9, OsNAL2, OsNAL3, OsNRL1, OsYABBY6*	15	4.92E-10
GO:0009908	BP	flower development	*OsAS2, OsHB4, OsLBD3-7, OsARF18, OsYABBY1, OsRL9, OsNAL2, OsNAL3, OsYABBY6*	9	5.4E-08
GO:0045449	BP	regulation of transcription	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	6.5E-06
GO:0019219	BP	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	6.8E-06
GO:0051171	BP	regulation of nitrogen compound metabolic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	6.8E-06
GO:0031326	BP	regulation of cellular biosynthetic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	8.7E-06
GO:0009889	BP	regulation of biosynthetic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	8.7E-06
GO:0010556	BP	regulation of macromolecule biosynthetic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	8.7E-06
GO:0010468	BP	regulation of gene expression	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	9.4E-06
GO:0031323	BP	regulation of cellular metabolic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	0.00001
GO:0010467	BP	gene expression	*OsZHD1, OsHB4, OsDCL1, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL15, OsRL9, OsNAL3, OsARF18*	12	1.1E-05

GO term	Ontology	Description	Genes	Number in input list	*P*-value
GO:0080090	BP	regulation of primary metabolic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	1.2E-05
GO:0006350	BP	transcription	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	1.2E-05
GO:0060255	BP	regulation of macromolecule metabolic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	1.3E-05
GO:0019222	BP	regulation of metabolic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	1.5E-05
GO:0009791	BP	post-embryonic development	*OsLC2, OsAGO1a, OsADL1, OsAGO7, OsHB4, OsLBD3-7, OsI_14279, OsZHD1, OsYABBY6*	9	2.29E-05
GO:0000003	BP	reproduction	*OsLC2, OsAGO1a, OsADL1, OsHB4, OsI_14279, OsARF18, OsZHD1, OsYABBY6*	8	4.05E-05
GO:0030154	BP	cell differentiation	*OsSCL1, OsAGO1a, OsRoc5, OsADL1, OsYABBY1, OsRL9, OsYABBY6*	7	4.37E-05
GO:0050794	BP	regulation of cellular process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	6E-05
GO:0006139	BP	nucleobase-containing compound (nucleobase, nucleoside, nucleotide and nucleic acid) metabolic process	*OsRL15, OsSCL1, OsRoc5, OsAGO7, OsHB4, OsSND2, OsARF18, OsYABBY1, OsMYB103L, OsRL9, OsZHD1, OsNAL2, OsNAL3, OsYABBY6*	14	6.33E-05
GO:0050789	BP	regulation of biological process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	8.1E-05
GO:0065007	BP	biological regulation	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsRL9, OsNAL3, OsARF18*	10	0.00012
GO:0016070	BP	RNA metabolic process	*OsHB4, OsDCL1, OsNAL2, OsRoc5, OsRL15, OsNAL3, OsARF18*	7	0.00025
GO:0009790	BP	embryo development	*OsAGO1a, OsADL1, OsHB4, OsLBD3-7, OsI_14279, OsZHD1*	6	0.00041

GO term	Ontology	Description	Genes	Number in input list	*P*-value
GO:0006351	BP	transcription, DNA-templated	*OsSCL1, OsRoc5, OsHB4, OsARF18, OsRL9, OsNAL2*	6	0.00076
GO:0006355	BP	regulation of transcription, DNA-templated	*OsSCL1, OsRoc5, OsHB4, OsSND2, OsARF18, OsMYB103L, OsRL9, OsNAL2*	8	0.00172
GO:0051252	BP	regulation of RNA metabolic process	*OsNAL3, OsHB4, OsRoc5, OsARF18, OsNAL2*	5	0.0018
GO:0032774	BP	RNA biosynthetic process	*OsNAL3, OsHB4, OsRoc5, OsARF18, OsNAL2*	5	0.0023
GO:0009653	BP	anatomical structure morphogenesis	*OsSCL1, OsAGO1a, OsRoc5, OsHB4, OsRL9, OsYABBY6*	6	0.00265
GO:0009058	BP	biosynthetic process	*OsSCL1, OsRoc5, OsHB4, OsSND2, OsARF18, OsYABBY1, OsMYB103L, OsRL9, OsZHD1, OsNAL2, OsNAL3, OsNRL1, OsYABBY6*	13	0.00309
GO:0009628	BP	response to abiotic stimulus	*OsLC2, OsAGO1a, OsRoc5, OsLBD3-7, OsI_14279, OsSFL1, OsSRS5, OsNRL1*	8	0.00866
GO:0044249	BP	cellular biosynthetic process	*OsZHD1, OsHB4, OsSFL1, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsNRL1, OsRL9, OsNAL3, OsARF18, OsRL15*	13	0.013
GO:0034645	BP	cellular macromolecule biosynthetic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsNRL1, OsRL9, OsNAL3, OsARF18, OsRL15*	12	0.014
GO:0009059	BP	macromolecule biosynthetic process	*OsZHD1, OsHB4, OsSND2, OsRFS, OsNAL2, OsRoc5, OsMYB103L, OsNRL1, OsRL9, OsNAL3, OsARF18, OsRL15*	12	0.014
GO:0044238	BP	primary metabolic process	*OsZHD1, OsMYB103L, OsHB4, OsSFL1, OsNAL9, OsSND2, OsRL16, OsRFS, OsNAL2, OsLBD3-7, OsRoc5, OsADL1, OsNAL3, OsNRL1, OsRL9, OsRRK1, OsLRRK1, OsARF18, OsRL15, OsDCL1*	20	0.029
GO:0043170	BP	macromolecule metabolic process	*OsZHD1, OsHB4, OsADL1, OsNAL9, OsSND2,*	18	0.033

GO term	Ontology	Description	Genes	Number in input list	*P*-value
			OsRFS, OsNAL2, OsLBD3-7, OsRoc5, OsMYB103L, OsNAL3, OsNRL1, OsRL9, OsRRK1, OsLRRK1, OsARF18, OsRL15, OsDCL1		
GO:0009793	BP	embryo development ending in seed dormancy	*OsAGO1a, OsADL1, OsI_14279*	3	0.034
GO:0008150	BP	biological_process	*OsRL15, OsCFL1, OsSCL1, OsSRL2, OsAGO7, OsI_14279, OsARVL4, OsACL1, OsNAL1, OsARF18, OsSRL1, OsREL2, OsNAL2, OsNAL3, OsNRL1*	15	0.04107
GO:0005634	CC	nucleus	*OsAS2, OsLC2, OsAGO1a, OsRoc5, OsHB4, OsSND2, OsARF18, OsYABBY1, OsMYB103L, OsRL9, OsNAL2, OsNAL3, OsYABBY6*	13	0.0008
GO:0005575	CC	cellular_component	*OsLC2, OsCFL1, OsAGO7, OsLBD3-7, OsI_14279, OsARVL4, OsNAL1, OsSRL1, OsYABBY1, OsNAL11, OsREL2, OsSRS5*	12	0.0225
GO:0003700	MF	sequence-specific DNA binding transcription factor activity	*OsSCL1, OsRoc5, OsHB4, OsSND2, OsARF18, OsYABBY1, OsMYB103L, OsRL9, OsZHD1, OsNAL2, OsNAL3, OsYABBY6*	12	3.16E-06
GO:0003677	MF	DNA binding	*OsRoc5, OsHB4, OsSND2, OsARF18, OsYABBY1, OsMYB103L, OsRL9, OsZHD1, OsREL2, OsNAL2*	10	0.00424

BP: Biological process, CC: Cellular component and MF: Molecular function.

List of Publications of Md. Jahangir Alam

A. Journal publications (06)

1. **Md. Jahangir Alam**, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Robust QTL Analysis Based on Robust Estimation of Bivariate Normal Distribution with Backcross population. *International Journal of Statistical Sciences*, 18, 2019. (accepted for publication)

2. **Md. Jahangir Alam**, Md. Alamin, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Robust linear regression based simple interval mapping for QTL analysis with backcross population. *Journal of Bio-Science*, 24, 75-81, 2018.

3. **Md. Jahangir Alam**, Md. Alamin, Most. Humaira Sultana, Md. Amanullah and Md. Nurul Haque Mollah. Regression Based Robust QTL Analysis for F_2 Population. *Rajshahi University Journal of Science and Engineering*, 44, 95-99, 2016.

4. Zobaer Akond, **Md. Jahangir Alam**, Md. Nazmol Hasan, Md. Selim Uddin, Monirul Alam and Md. Nurul Haque Mollah. A Comparison on Some Interval Mapping Approaches for QTL Detection, *Bioinformation*, 15(2), 90-94, 2019. DOI:10.6026/97320630015090.

5. Zobaer Akond, Md. Nazmol Hasan, **Md. Jahangir Alam**, Monirul Alam and Md. Nurul Haque Mollah. Classification of Functional Metagenomes Recovered from Different Environmental Samples. *Bioinformation*, 5(1), 26-31, 2019. DOI:10.6026/97320630015026.

6. Md. Nazmol Hasan, Zobaer Akond, **Md. Jahangir Alam**, Anjuman Ara Begum, Moizur Rahman and Md. Nurul Haque Mollah. Toxic Dose prediction of Chemical Compounds to Biomarkers using an ANOVA based Gene Expression Analysis. *Bioinformation*, 14(7), 369–377, 2018. DOI: 10.6026/97320630014369.

B. Conference proceedings (16)

1. **Md. Jahangir Alam**, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Regression Based Fast Multi-trait QTL Analysis. *2nd International Conference on Applied Statistics 2019*, Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000, Bangladesh, 2019.

2. **Md. Jahangir Alam**, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Robust QTL Analysis Based on Robust Estimation of Bivariate Normal Distribution. *7th International Conference on Data Science and SDGs: Challenges, Opportunities, and Realities*, Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh, 2019.

3. **Md. Jahangir Alam**, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Multivariate QTL Mapping: A Comparative Study. *International Conference On New Paradigms in Statistics for Scientific and Industrial Research 2018*, Central Glass & Ceramic Research Institute, Jadavpur, Kolkata, 2018.

4. **Md. Jahangir Alam**, Md. Alamin, Hafizur Rahman, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Structural and Phylogenetic Analysis of Rolling Leaf Related Genes in Rice (*Oryza sativa* L.). *International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment – 2017*, Department of Statistics, University of Rajshahi, Bangladesh, 2017.

5. **Md. Jahangir Alam**, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. A Comparison on the Computational Tools of QTL Analysis. *International Conference of Biotechnology on Health and Agriculture 2017*, University of Dhaka, Bangladesh, 2017.

6. Md. Amanullah, Md. Mamun Monir, **Md. Jahangir Alam**, Md. Mamunur Rashid and Md. Nurul Haque Mollah. Robust ICIM for QTL Analysis by Minimum β-

Divergence Method for Backcross Population. *International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment – 2017*, Department of Statistics, University of Rajshahi, Bangladesh, 2017.

7. Atul Chandra Singha, Arafat Rahman, **Md. Jahangir Alam** and Md. Nurul Haque Mollah. Comparative Study on Genome-Wide Association Studies Using Canonical Correlation Analysis. *International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment – 2017*, Department of Statistics, University of Rajshahi, Bangladesh, 2017.

8. Most Humaira Sultana, Md. Alamin, **Md. Jahangir Alam**, Longjiang Fan and Md. Nurul Haque Mollah. SNP Based Robust Fast-eQTLMapping for Identification of Important Genes. *International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment – 2017*, Department of Statistics, University of Rajshahi, Bangladesh, 2017.

9. Md. Mamunur Rashid, Fahima Farhana Anni, **Md. Jahangir Alam**, Md. Amanullah, Md. Mamun Monir and Md. Nurul Haque Mollah. Dimension Reduction Approach for Multivariate QTL Analysis. *International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment – 2017*, Department of Statistics, University of Rajshahi, Bangladesh, 2017.

10. Md. Motiur Rahman, Asmaul Husna, Mamun Ur Rashid, **Md. Jahangir Alam**, Md. Asif Ahsan and Md. Nurul Haque Mollah. A Comparative Study on Some Statistical Algorithms for Genome Wide Association Studies (GWAS) using R-Packages. *International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment – 2017*, Department of Statistics, University of Rajshahi, Bangladesh, 2017.

11. **Md. Jahangir Alam**, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Statistical Tools for Genome Analysis in Bioinformatics. *International Conference on Analysis of Repeated Measures Data*, Department of Applied Statistics, East West University, Bangladesh, 2016.

12. **Md. Jahangir Alam**, Atul Chandra Singha, Md. Manir Hossain Mollah, Md. Ripter Hossain, S.M. Shahinul Islam and Md. Nurul Haque Mollah. A comparative Study on R Statistical Packages for Genome-Wide Association Analysis. *The 2nd International Conference on Theory and Application of Statistics 2015*, Department of Statistics, University of Dhaka, 2015.

13. **Md. Jahangir Alam**, Md. Alamin, Most. Humaira Sultana, Md. Amanullah and Md. Nurul Haque Mollah. Regression Based Robust QTL Analysis using Flanking Marker with Intercross (F_2) Population. *International Conference on Materials, Electronics & Information Engineering ICMEIE-2015*, University of Rajshahi, P125, 2015. Website: http://www.ru.ac.bd/icmeie2015/proceedings/pdfs/125.pdf.

14. **Md. Jahangir Alam**, Md. Alamin, Most. Humaira Sultana, Md. Amanullah and Md. Nurul Haque Mollah. Robust Regression Based Interval Mapping Approach for QTL Analysis. *International Conference on Applied Statistics (ICAS) 2014*, Institute of Statistical Research and Training (ISRT), University of Dhaka, P112, 2014.

15. Md. Amanullah, Md. Mamun Monir, Most. Humaira Sultana, **Md. Jahangir Alam** and Md. Nurul Haque Mollah. A Review Study for Quantitative Trait Locus Mapping and Whole Genome Association Studies Tools in Statistical Genomics and Bioinformatics. *International Conference on Applied Statistics (ICAS) 2014*, ISRT, University of Dhaka, P109, 2014.

16. Most. Humaira Sultana, Md. Amanullah, Md. Masud Rana, **Md. Jahangir Alam** and Md. Nurul Haque Mollah. Robustification of Fast Map for eQTL Analysis to Find Genomic Locations Influencing Gene-Expression. *International Conference on Applied Statistics (ICAS) 2014*, ISRT, University of Dhaka, P104, 2014.

C. Submitted to the journal or manuscript is ready for submission (04)

1. **Md. Jahangir Alam**, Md. Alamin, Most. Humira Sultana, Md. Asif Ahsan, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. *In silico* studies on structures and functions of rolling leaf related genes in rice (*Oryza sativa* L.). (Submitted for publication in Plant Genetic Resources).

2. **Md. Jahangir Alam**, Md. Mamun Monir, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Regression Based Fast Multi-trait Genome-Wide QTL Analysis. (Submitted for publication in Journal of Bioinformatics and Computational Biology).

3. **Md. Jahangir Alam**, Md. Mamun Monir, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Robustification of Regression Based Fast Multi-trait QTL Analysis. (Manuscript is ready to submit for publication in a well-reputed International Journal).

4. **Md. Jahangir Alam**, Md. Mamun Monir, Md. Ripter Hossain, S. M. Shahinul Islam and Md. Nurul Haque Mollah. Robustification of Regression Based Genome Wide Association Studies for SNP Analysis. (Manuscript is ready to submit for publication in a well-reputed International Journal).