2019

# Statistical Modeling for Genome Data Analysis to Detect Agricultural Biomarkers

Akond, Zobaer

University of Rajshahi, Rajshahi

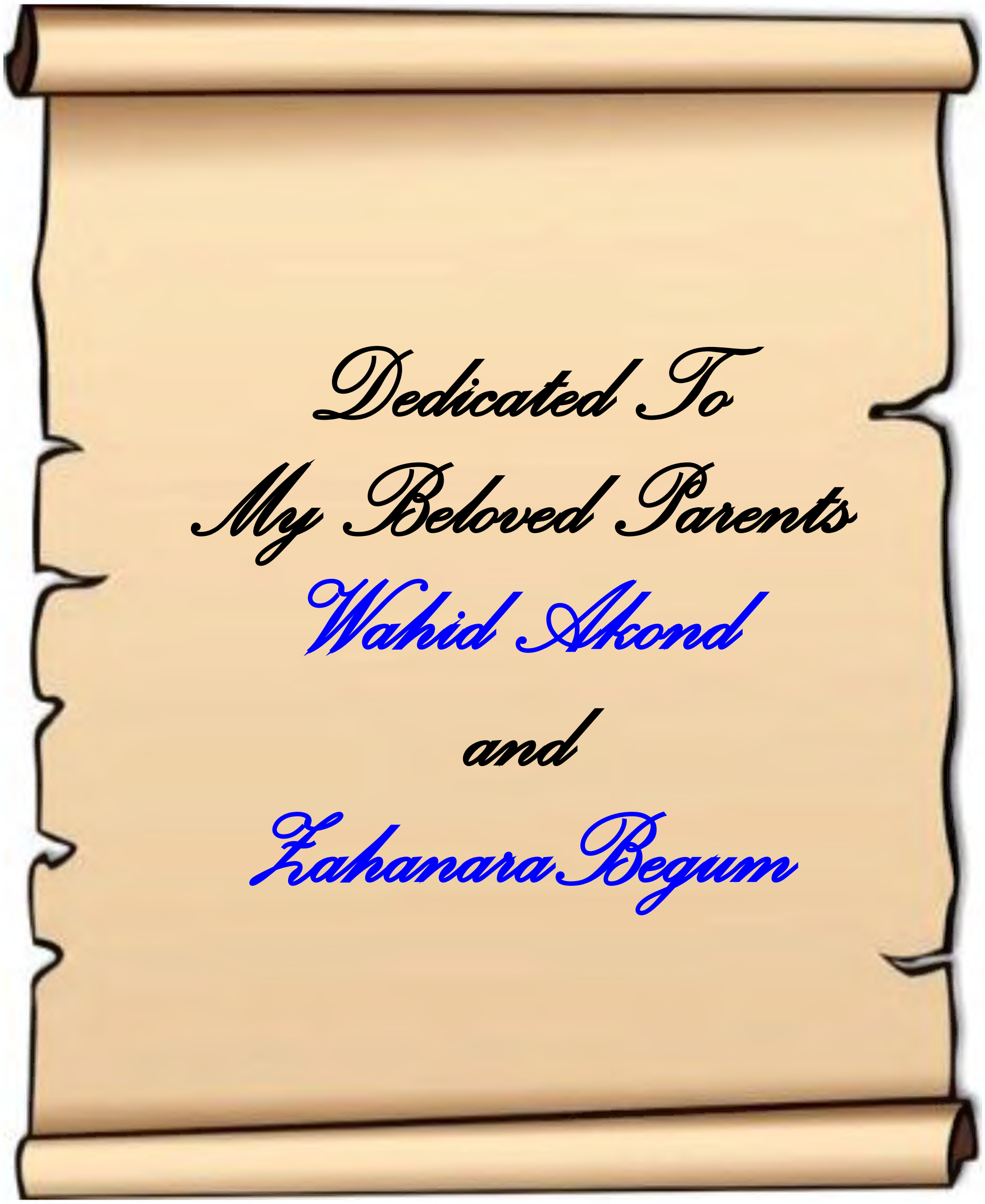
# STATISTICAL MODELING FOR GENOME DATA ANALYSIS TO DETECT AGRICULTURAL BIOMARKERS

**THESIS SUBMITTED FOR THE DEGREE**
**OF**
**DOCTOR OF PHILOSOPHY**
**IN THE**
**INSTITUTE OF ENVIRONMENTAL SCIENCE**
**UNIVERSITY OF RAJSHAHI**
**BANGLADESH**

**BY**
**ZOBAER AKOND**
**B.Sc (Hons.), M.Sc (Statistics), M.Sc (Computer Science)**

**SEPTEMBER , 2019**

**INSTITUTE OF ENVIRONMENTAL SCIENCE**
**UNIVERSITY OF RAJSHAHI**
**RAJSHAHI-6205**
**BANGLADESH**



Dedicated To

My Beloved Parents

Wahid Akond

and

Zahanara Begum

# DECLARATION

I hereby declare that the research work accomplished in this thesis entitled "**Statistical Modeling for Genome Data Analysis to Detect Agricultural Biomarkers"** has been carried out by me for the degree of Doctor of Philosophy. This research work has been accomplished under the guidance of Professor Dr. Md. Nurul Haque Mollah, Department of Statistics, University of Rajshahi and Dr. Munirul Alam, Senior Scientist, Emerging Infections, Infectious Diseases Division, Internal Center for Diarrheal Disease Research, Bangladesh (icddr,b). I also declare that the results presented in this dissertation are my own investigation and any part of this thesis work has not been submitted to elsewhere for any degree/diploma or for similar purposes.

------------------------

**Zobaer Akond**

PhD Fellow
Roll No. 15201
Session: 2015-2016
Institute of Environmental Science (IES)
University of Rajshahi

**UNIVERSITY OF RAJSHAHI**
**RAJSHAHI, BANGLADESH**

# Certificate of Approval

It is our pleasure to certify that the thesis entitled **Statistical Modeling for Genome Data Analysis to Detect Agricultural Biomarkers** is an original research work completed by **Zobaer Akond**. The research work has been carried out under our close supervision. As far as we know that the thesis has not been formerly submitted to any other university/institute for any kind of degree or diploma.

We also certify that we have read the thesis and found it satisfactory for submission to the Institute of Environmental Science (IES), University of Rajshahi for the degree of Doctor of Philosophy (Ph.D.) in **Bioinformatics.**

**Dr. Md. Nurul Haque Mollah**
Principal Supervisor
and
Professor
Department of Statistics
University of Rajshahi

**Dr. Munirul Alam**
Co-Supervisor
and
Senior Scientist
Emerging Infections, Infectious Diseases Division,
International Center for Diarrheal Disease,
Bangladesh(icddr,b)

# CONTENTS

# ACKNOWLEDGEMENTS

**Almighty Allah** is the great creator of this whole universe whatever living and nonliving things exist on this earth. **Allah** observes and maintains things in our imagination and beyond our imagination whatsoever. I am immensely grateful to **Allah** for His blessings for the accomplishment of the PhD dissertation finally.

I would like to express my profound gratitude and appreciation to my principal supervisor **Professor Dr. Md. Nurul Haque Mollah** for his active supervision, motivating guidance, sincere support as well as creating fruitful research environment throughout the whole period of my research work.

I am also very thankful to my co-supervisor **Dr. Munirul Alam**, Senior Scientist, Emerging Infections, Infectious Diseases Division, International Center for Diarrheal Disease, Bangladesh (icddr,b) for his encouragement, helpful advices and discussions at the time of my research work that made me prompt to complete the whole work successfully.

I would like to acknowledge the Bangladesh Agricultural Research Institute (BARI) for providing me the stipend to accomplish my PhD research work.

I would like to acknowledge the Bioinformatics Lab., Department of Statistics, University of Rajshahi for providing the lab facilities during my research work.

I also pay many thanks to the honorable director and respected teachers and staffs of Institute of Environmental Science (IES) for their valuable cooperation to finish my course works and research during the whole period of my PhD program.

I am extremely grateful to the scientists and staffs of Agricultural Statistics and ICT (ASICT) Division and other divisions of Bangladesh Agricultural Research Institute (BARI) for their valuable administrative and others supports during my PhD work.

No doubt, these acknowledgements would be incomplete unless I state the supports and helps belong to a few research personnel – Dr. Hafizur Rahman, Dr. Matiur Rahman, Dr Asif Ahmed and Mohammand Nazmol Hasan(PhD Fellow).

Finally, I am very grateful and indebted to my beloved parents, my dear wife Lutfun Nahar Lata, my loving son Muaaz Abrar Akond (Yaafi) and Yaaz Mashroor Akond for their patience and sacrifices. I am indeed thankful for providing encouragements and affections to me when and where it was essential during my whole period of PhD research. Best wishes for them.

***The Author***

# ABSTRACT

The focuses of this study were to evaluate the performance of different statistical methods from the perspective of various genomic data such as phenotypic-genotypic data, gene expression (microarray/RNA-Seq) data, SNP data and meta-genomic data collected from different environmental samples. We also performed some *in silico* analysis of RNA silencing machinery genes in wheat (*Triticum aestivum*) based on the RNAi genes of arabidopsis thaliana and expression profile analysis of seven TaDCL genes in leaves and roots as well as against drought stress using qRT-PCR.

In **Chapter Two,** we explored better QTL mapping approach by comparative study. We found that Composite Interval Mapping (CIM) performs significantly better than the other four Simple Interval Mapping (SIM) methods in detecting QTL positions in backcross technique both on simulated data and on real rice genome dataset. In the case of real rice genome data analysis for backcross population, the CIM identified some vital positions that were not detected by the traditional SIM approaches.

In **Chapter Three,** we have discussed the transcriptomics data (e.g. microarray, RNA-Seq) analysis from the robustness point of view. Transcriptomics datasets poses different computational challenges by virtue of the large number of transcripts surveyed with small sample sizes. Usually, these types of data analysis require identification of differentially expressed (DE) genes between two or more conditions and classification/clustering of samples (DE genes) based on the DE genes (samples). There are several statistical methods for these types of works. However, most of them are suffering from the small sample size and outlying observations. So transcriptomics data analysis by most of the conventional algorithms might be produced misleading results, since transcriptomics datasets are also often contaminated by outlying observations due to several steps involve in the data generating processes. To overcome these problems, in this chapter, we proposed logistic transformation of transcriptomics data for (i) robust identification of DE genes by SAM approach and (ii) robust classification of samples (DE genes) based on the reduced DE gene-set (samples). Simulation and real transcriptomics data analysis results showed that the proposed procedure outperform over the conventional procedure in presence of outliers, otherwise it keeps almost equal performance. It should be mention here that in the case of real rice genome data (control vs blast fungus disease) analysis, our proposed method detected two additional genes that were significantly associated with the rice blast fungus disease. This report is also supported by the literature review.

DCL, AGO and RDR are known as RNA silencing machinery genes play significant role in the regulation of gene expression through the generation of small RNA (sRNA) molecules in plants. The wheat (*Triticum aestivum*) possesses 7 DCL, 39 AGO and 16 RDR genes. The aim of **Chapter Four** was *in silico* identification and characterization of these genes and analyses of expression pattern of 7 TaDCLs. The phylogenetic analysis provided that all subfamilies of these three gene sets maintain their evolutionary relationships similar to their rice and Arabidopsis counterpart. Conserved domain structure analysis suggested that these genes were also contained consistent domain structure similar to rice and Arabidopsis. Although there was a difference in possessing *cis*-acting elements of the promoter regions of TaDCL genes but the promoters of TaDCL1a, TaDCL3a, and TaDCL4 contained a large number of stress-related *cis*-elements. GO investigation identified different metabolic process, biosynthetic process, molecular binding such as RNA, nucleic acid, protein binding, posttranscriptional, transferase and nuclease activities are significantly connected to RNAi gene regulation in wheat. Cytosol is however found to be the key molecular component that possesses the maximum number of wheat RNA silencing proteins. Expression analyses indicated that TaDCL3 and TaDCL4 genes are likely to play distinct roles in development and drought stress tolerance. This work provides the important indication of DCL, AGO and RDR genes evolutionary resemblances of their rice and Arabidopsis counterpart.

Analysis for SNP biomarkers identification following this approach encompasses two major challenges such as existing embedded population structures or stratification along with polygenic effects or genetic relatedness among population subjects or individuals. To overcome these complexities recently linear mixed model has been widely used for GWAS for large-scale whole genome dataset. It is however observed that this method is sensitive

to outliers and produce higher FDR and lower statistical power. We therefore proposed in **Chapter Five** a robust approach for handling outliers and reducing the differential effects of the population stratification. We introduced a $\beta$-weight function termed as minimum $\beta$-divergence method accompanied by a tuning parameter $\beta$ to overcome the problem of data contamination. The proposed approach performed superior in terms of lower FDR and higher statistical power compared to LRM and LMM at changing outliers for two heritability rates 0.2 and 0.3. We also identified 11 rice SNP makers by the proposed method. It is projected that the gene LOC_Os06g18000 might play functional roles in flower development and in response to stress in rice. The marker genes LOC_Os02g21880, LOC_Os06g18000, LOC_Os02g24134 exhibited larger expression in seedling, vascular cell, root, shoot, and panicle. The gene LOC_Os08g25060 is predicted to express highly in vascular cell, root, and panicle. Our proposed robust method outperforms the existing methods for GWAS in the presence of subject outliers.

Classification of the metagenomic data obtained from different microbial samples is a significant issue in the context of their associated functional metagenomic variables. In **Chapter Six,** Random forest classifier presented the lowest FDR and MER in conjunction with highest TPR in all cases of data compared to Bayes, SVM, KNN, AdaBoost and LogitBoost classifiers based on the feature selection by the proposed beta-*t* statistic. It is therefore concluded that the proposed beta-*t* based random forest classifier is considered the optimum classifier in grouping the metagenomes collected from different environmental community.

# LIST OF TABLE

# LIST OF FIGURE

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ACC | Accuracy |
| AdaBoost | Adaptive Boosting |
| AGO | Argonaut |
| ANOVA | Analysis of Variance |
| AUC | Area Under the Receiver Operating Characteristic Curve |
| bagSVM | Bagging Support Vector Machine |
| BLAST | Basic Local Alignment Search Tool |
| CART | Classification and Regression |
| CC | Cellular Component |
| CDA | Canonical Discriminate Analysis |
| CDS | Coding Sequence |
| cDNA | Complementary Deoxyribonucleic Acid |
| CIM | Composite Interval Mapping |
| CRAN | Comprehensive R Archive Network |
| DAP | Days After Planting |
| DCL | Dicer-like |
| DE | Differentially Expressed |
| DEG | Differentially Expressed Gene |
| DMRT | Duncan Multiple Range Test |
| DSRM | Double Stranded RNA-binding Motif |
| dsRNA | Double Stranded Ribonucleic Acid |
| DUF283 | Domain of Unknown Function283 |
| edgeR | Differential Expression Analysis of Digital Gene Expression Data |
| EE | Equally Expressed |
| eHK | Extended Haley Knott |
| EMMA | Efficient Mixed Model Association |
| EMMAX | Efficient Mixed Model Association Expedited |
| EM | Expectation Maximization |

| | |
|---|---|
| FC | Fold Change |
| FDR | False Discovery Rate |
| FEM | Fixed Effects Model |
| FP | False Positive |
| FN | False Negative |
| FPR | False Positive Rate |
| FPKM | Fragments Per Kilobase Million |
| FNR | False Negative Rate |
| GAPIT | Genomic Association and Prediction Integrated Tool |
| GC | Genomic Control |
| GEO | Gene Expression Omnibus |
| GO | Gene Ontology |
| GSDS | Gene Structure Display Server |
| GWA | Genome Wide Association |
| GWAS | Genome Wide Association Studies |
| HIV | Human Immunodeficiency Virus |
| HK | Haley Knott |
| HT | High Throughput |
| ICM | Independent Contamination Model |
| IM | Interval Mapping |
| IMP | Multiple Imputation |
| IT | Information Technology |
| KNN | K-Nearest Neighbor |
| LOD | Logrithm(10 base) of Odds |
| LRT | Likelihood Ratio Test |
| LRM | Linear Regression Model |
| MAF | Minor Allele Frequency |
| MEGA | Molecular Evolutionary Genetics Analysis |
| MER | Misclassification Error Rate |
| MDS | Multidimensional Scaling |

| | |
|---|---|
| MF | Molecular Function |
| miRNA | Micro Ribonucleic Acid |
| MLE | Maximum Likelihood Estimator |
| MLM | Mixed Linear Model |
| MR | Multiple Regression |
| NPV | Negative Predictive Value |
| NJ | Neighbor Joining |
| NCBI | National Center for Biotechnology Information |
| pAUC | Partial Area Under the Receiver Operating Characteristic Curve |
| PAZ | Piwi/Argonaute/Zwille |
| PC | Principle Component |
| PCA | Principal Component Analysis |
| PPI | Protein-Protein Interaction |
| PPV | Positive Predictive Value |
| PSI | Plant Subcellular localization Investigative Predictor |
| PTM | Post Translational Modification |
| qRT-PCR | Quantitative Real Time Polymerase Chain Reaction |
| qtl/QTL | Quantitative Trait Loci |
| rANOVA | Robust Analysis of Variance |
| rrBLUP | Ridge Regression Best Linear Unbiased Prediction |
| RDR | RNA-dependent RNA-polymerase |
| REM | Random Effects Model |
| RF | Random Forest |
| RGAP | Rice Genome Annotation Project |
| RNA | Ribonucleic Acid |
| RNaseII | Ribonuclease II |
| RNAi | Ribonucleic Acid Interference |
| ROC | Receiver Operating Characteristic |
| SA | Structured Association |
| SAMseq | Significance Analysis of Microarray Sequencing |

| | |
|---|---|
| SCL | Subcellular Localization |
| SE | Standard Error |
| SIM | Simple Interval Mapping |
| siRNA | Short Interfering Ribonucleic Acid |
| SNP | Single Nucleotide Polymorphism |
| STRING | Search Tool for the Retrieval of Interacting Genes/Proteins |
| SVM | Support Vector Machine |
| ssRNA | Single-Stranded Ribonucleic Acid |
| TP | True Positive |
| TN | True Negative |
| TPR | True Positive Rate |
| TNR | True Negative Rate |
| THCM | Tukey-Huber Contamination Model |
| TASSEL | Trait Analysis by aSSociation, Evolution and Linkage |
| voom | Mean Variance Modelling at the Observational level |
| limma | Linear Models for Microarrays |
| WGS | Whole Genome Sequencing |

# CHAPTER ONE
# GENERAL INTRODUCTION

# GENERAL INTRODUCTION

## 1. Introduction

Cells in the living being are the smallest basic structural, functional, and biological units. Each biological phenomenon of a life is continuously occurring and maintaining by the nucleus of a cell. Each cell nucleus contains long chromosome where lies Deoxyribonucleic Acid (DNA) carrying genetic information in the form of genes. Genotypic and phenotypic variation in the species ultimately occurs due to different genes and for their different expressions level. The genetic code or information stored in DNA is interpreted by gene expression and the properties of the expression give rise to the organism’s phenotype. This gene expression is associated to synthesize gene product such as RNA or protein from DNA. There is a profound impact on the functions or actions of the gene in a cell or in a multicellular organism due to the amount of gene expression. Regulation of gene expression in a genotype gives rise to the phenotype that is the observable trait. Variation in phenotype involves the mechanism of protein synthesis that control the organism’s shape or that act as enzymes catalyzing specific metabolic pathways characterizing the organism. Recently, bioinformatics has been evolved as the core research discipline or platform, which combines the shared in-depth knowledge and viewpoint of molecular biology, statistical and computer science. The common focus of the associated researchers in this discipline is to identify differentially expressed gene(s) and their expression levels, gene functions, and biological, molecular, and cellular component pathways, specifically identification of genes even at single nucleotide polymorphism (SNP) level responsible for particular reasons of interest such as for substantial phenotypic variations in terms of healthy and diseased one. Sometimes it is key issue to identify cause of genotypic and phenotypic variations in agricultural crops against different biotic stressors (virus, bacteria, archaea, fungi, insects etc) and abiotic stressors (drought, salt, cold, water logging, gas, air etc) by the bioinformaticians from the whole genome sequencing which drive bioinformaticians to pinpoint the so-called differentially expressed (DE) genes for such reasons. Earlier this genomic information was extracted using quantitative trait loci (qtl) mapping or linkage analysis that is the initial branch of genomics relates to the identification of prospective marker/gene location in the target chromosome(s). This qtl technique is commonly used by the conventional plant breeders to recognize the prospective agricultural biomarker(s)/gene(s) for crop improvement in agricultural practices. Due to rapid technological advancement, the scope of the bioinformatics research has been expanding considerably in all biological research in terms of generating traditional microarray genome data, DNA, RNA, proteomic, metagenomics and SNP level data and their fine tuning assembly and annotations. However, the major areas of bioinformatics research encompass the qtl analysis, transcriptomics (microarrary/RNA-seq), proteomics; metabolomics, metagenomics data analysis and recently started genome-wide association studies (GWAS) at SNP level. Existing among many transcriptome profiling tools, DNA microarrays technology was invented in 1990s. Microarrays were perhaps the technology that allowed biologists to vast amounts of complex digital data. In the late 90's and 2000's, DNA microarray technology progressed rapidly as both new methods of production and fluorescent detection were adapted to the task. Microarrays are simply devices simultaneously to measure the relative concentrations of many different DNA or RNA sequences such as spotted arrays on glass, in-situ synthesized arrays, and self-assembled arrays. The advent of next generation sequencing technologies combined with the rapid decrease in the cost of sequencing made sequencing cost competitive with microarrays for all assays with the possible exception of genotyping. Sequencing is a relatively unbiased approach to measuring which nucleic acids are present in solution. While sample preparation or different enzymes may bias sequencing counts, unlike DNA arrays, sequencing is not dependent on prior knowledge of which nucleic acids may be present. Sequencing is also able to detect closely related gene sequences, novel splice forms or RNA editing that may be missed due to cross hybridization on DNA microarrays. Because of these advantages and the decreasing cost of sequencing, DNA arrays are being rapidly replaced by sequencing for nearly every assay that has been previously performed on microarrays (Wold and Myers, 2008). Transcriptomics is one of the key branches of omics technology  that links to the study of the total set of transcripts in a given organism, or to the specific subset of transcripts present in a particular cell type that are produced from DNA. Transcriptomics is an emerging and continually growing field in biomarker discovery for use in assessing the safety of drugs or chemical assessment as well. It may also be used to infer phylogenetic relationships among individuals. Transcriptomics techniques include DNA microarray and next-generation sequencing technologies called RNA-seq. RNA-Seq can identify disease-associated single nucleotide polymorphisms (SNPs), allele-specific expression, and gene fusions, which contributes to the understanding of disease causal variants(Khurana et al., 2016). RNA-seq data analysis however helps bioinformaticians to identify an organism’s genes that are differentially expressed (DE) between diseased and healthy tissues or different conditions, or at different times, gives information on how genes are regulated and reveals details of an organism’s biology. This technology can infer the functions of previously unannotated genes. For improved detection of DE genes, RNA-seq possesses higher sensitivity with absolute values and lower technical than microarrays. Therefore, the transcriptomic profiling using RNA-seq has been more popular than older microarray technology. Large high-dimensional sequencing data with many biological variables have been available due to introduction of hi-tech computational facility nowadays. These high-dimensional data mainly involves two attributes such as a large number of variables (genes) and small number of samples (patients). Proteomics is another branch of OMICS field for large-scale study of proteomes. Proteins are produced from mRNA through the translation by building chain of triplet nucleotide called amino acids. A proteome is a set of proteins produced in an organism, system, or biological context. We may refer to, for instance, the proteome of a species (for example, wheat) or an organ (for example, the leaf). The proteome is not constant; it differs from cell to cell and changes over time. To some degree, the proteome reflects the underlying transcriptome. However, many factors modulated the protein activity (often assessed by the reaction rate of the processes in which the protein is involved) in addition to the expression level of the relevant gene. Metagenomics is one of the prominent areas of bioinformatics that study the microbial community (virus, bacteria, archaea etc.) obtained from different environmental samples. Metagenomics applies a suite of genomic technologies and bioinformatics tools to access the genetic content of entire communities of organisms directly. With the growing numbers of activities also comes a plethora of methodological knowledge and expertise that should guide future developments in the field(Simon and Daniel, 2011). This key area of bioinformatics research involves similar steps as like as previously mentioned bioinformatics research platforms viz., sample processing, sequencing technology, assembly, binning, annotation, experimental design, statistical analysis, data storage, and data sharing. Finally, the of late started area of bioinformatics research is called Genome wide association studies (GWAS).GWAS are hypothesis free methods to identify associations between genetic regions (loci) and traits (including diseases).It has long been known that genetic variation between individuals can cause differences in phenotypes. These causal variants, and those which are tightly linked to their region of the chromosome, are therefore present at higher frequency in cases (individuals with the trait) than controls (individuals without the trait) (www.ebi.ac.uk).The first successful GWAS was published in 2002 studying myocardial infarction. In genetics, GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major plant diseases, but can equally be applied to any other genetic variants and any other organisms. GWA studies identify SNPs and other variants in DNA associated with a disease, but they cannot on their own specify which genes are causal(Manolio, 2010; Pearson and Manolio, 2008). As of 2017, hundreds or thousands of individuals are tested in a typical GWA study. Over 3,000 human GWA studies have examined over 1,800 diseases and traits, and thousands of SNP associations have been found(Pearson and Manolio, 2008). Bioinformatics research in agriculture help researchers identify DE genes for distinct phenotypic variations in crops, genesis of diseases for diagnosis, prognosis and to predict as well as evaluate the response of certain treatments even at molecular level. Researchers however sometimes experience very challenges for statistical computing and analysis in terms of making inference, estimation, prediction or classification and draw precise conclusion from such voluminous data that definitely leads to consider or develop modern and robust statistical algorithms or methods and bioinformatics tools. The results will be misleading unless the statistical hypotheses and methods are appropriately selected and employed for such high-throughput data that hold different characteristics. The next subsequent subsections will be presented with the basic terminology of molecular biology and several OMICS technologies in bioinformatics research. Next, a brief discussion on different genome data characteristics and their analytical process in terms of statistical methods and bioinformatics tools will be discussed to address the some specific challenges and issues arise from these different genome dataset to draw precise and valid conclusions. Then we will conclude this first chapter with different objectives and outline of our study.

## 1.1 Biological Terminology Related to This Thesis

Cell: A cell is the smallest unit of life. Cells are often called the "building blocks of life". The study of cells is called cell biology or cellular biology. Cells consist of cytoplasm enclosed within a membrane, which contains many biomolecules such as proteins and nucleic acids(Bruce Alberts 2002). Organisms can be classified as unicellular (consisting of a single cell; including bacteria) or multicellular (including plants and animals). Cells are of two types: eukaryotic, which contain a nucleus, and prokaryotic, which do not.

Prokaryotic cells: Prokaryotes are single-celled organisms, while eukaryotes can be either single-celled or multicellular. Prokaryotes include bacteria and archaea, two of the three domains of life. The DNA of a prokaryotic cell consists of a single circular chromosome that is in direct contact with the cytoplasm.

Eukaryotic cells: Plants, animals, fungi, slime molds, protozoa, and algae are all eukaryotic. These cells are about fifteen times wider than a typical prokaryote and can be as much as a thousand times greater in volume. The main distinguishing feature of eukaryotes as compared to prokaryotes is compartmentalization: the presence of membrane-bound organelles (compartments) in which specific activities take place. Most important among these is a cell nucleus (NCBI document: "What Is a Cell?". 30 March 2004). An organelle that houses the cell's DNA. This nucleus gives the eukaryote its name, which means "true kernel (nucleus)".

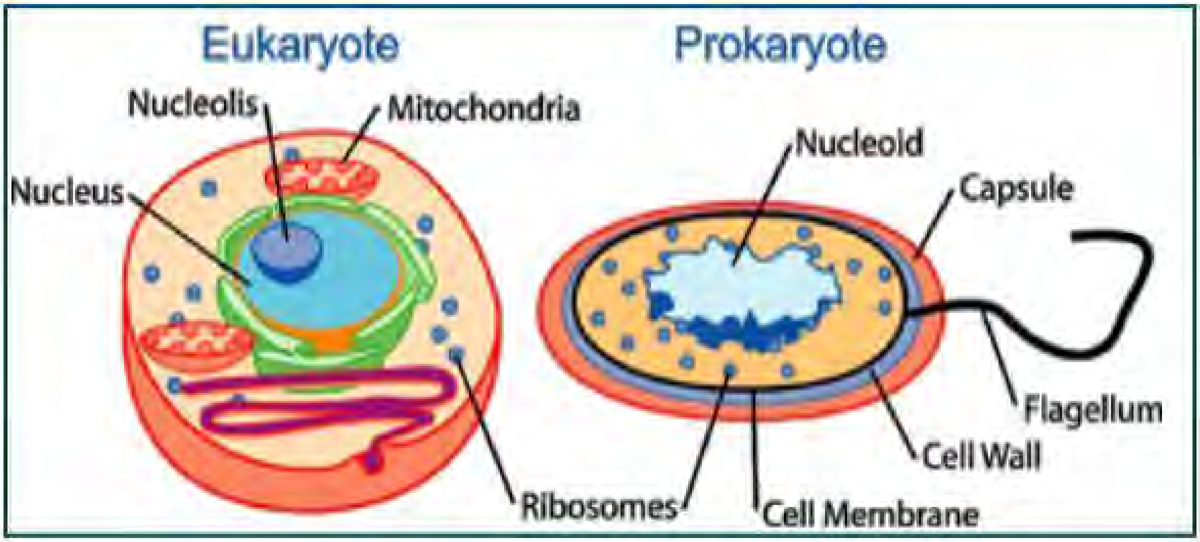


**Fig.1.1** Structure of Typical Eukaryote and Prokaryote cell

**DNA:** Deoxyribonucleic acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth, and reproduction of all known organisms and many viruses. Both strands of double-stranded DNA store the same biological information. This information is replicated as and when the two strands separate. The two DNA strands are also known as polynucleotides as they are composed of simpler monomeric units called nucleotides (Boyle 2008; Purcell A 2017). Each nucleotide is composed of one of four nitrogen-containing nucleobases (cytosine [C], guanine [G], adenine [A] or thymine [T]), a sugar called deoxyribose, and a phosphate group. The nitrogenous bases of the two separate polynucleotide strands are bound together, according to base pairing rules (A with T and C with G), with hydrogen bonds to make double-stranded DNA. The DNA of a prokaryotic cell consists of a single circular chromosome that is in direct contact with the cytoplasm. The eukaryotic DNA is organized in one or more linear molecules, called chromosomes, which are associated with histone proteins. All chromosomal DNA is stored in the cell nucleus, separated from the cytoplasm by a membrane (NCBI document: “What Is a Cell?” 30 March 2004). Some eukaryotic organelles such as mitochondria also contain some DNA.

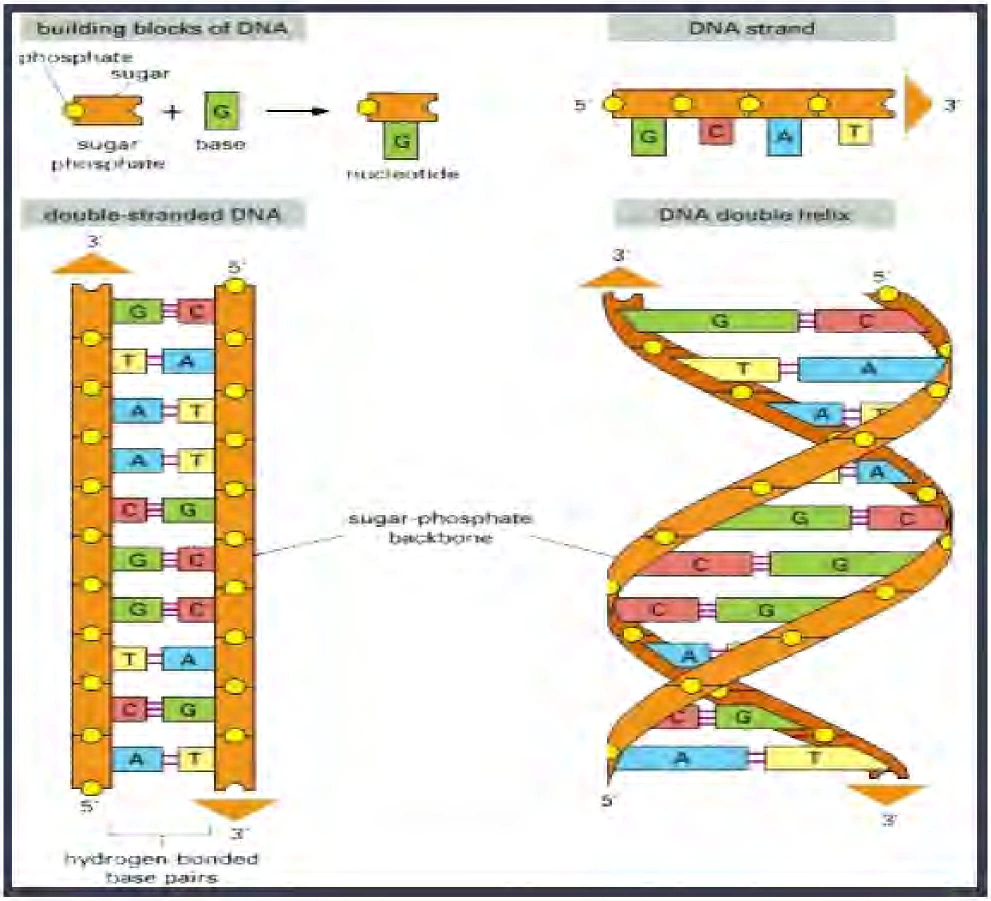


**Fig.1.2** Structure of DNA

**RNA:** Ribonucleic acid (RNA) is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression of genes. RNA and DNA are nucleic acids, and, along with lipids, proteins and carbohydrates constitute the four major macromolecules essential for all known forms of life. Like DNA, RNA is assembled as a chain of nucleotides, but unlike DNA, it is more often found in nature as a single-strand folded onto itself, rather than a paired double-strand. Cellular organisms use messenger RNA (mRNA) to convey genetic information (using the nitrogenous bases of guanine, uracil, adenine, and cytosine, denoted by the letters G, U, A, and C) that directs synthesis of specific proteins.

Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function in which RNA molecules direct the synthesis of proteins on ribosomes. This process uses transfer RNA (tRNA) molecules to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) then links amino acids together to form coded proteins.

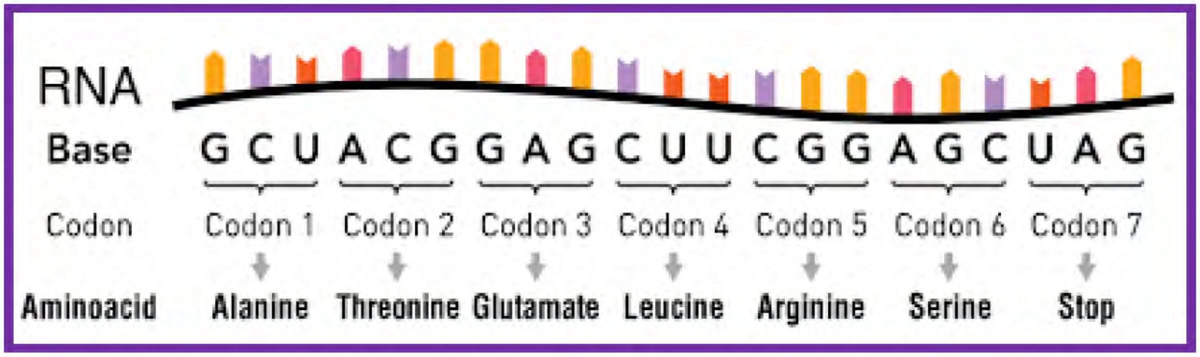


**Fig.1.3** Structure of RNA

**Types and Length of RNA:**

Messenger RNA (mRNA) is the RNA that carries information from DNA to the ribosome, the sites of protein synthesis (translation) in the cell. The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced *(*Cooper and Hausman 2004). However, many RNAs do not code for protein (about 97% of the transcriptional output is non-protein-coding in eukaryotes(Mattick, 2001, 2003; Mattick and Gagen, 2001). These so-called non-coding RNAs ("ncRNA") can be encoded by their own genes (RNA genes), but can also derive from mRNA introns(Wirta, 2006). The most prominent examples of non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation(Berg et al., 2002). There are also non-coding RNAs involved in gene regulation, RNA processing and other roles. Certain RNAs are able to catalyse chemical reactions such as cutting and ligating other RNA molecules, and the catalysis of peptide bond formation in the ribosome; these are known as ribozymes(Nissen et al., 2000; Rossi, 2004).

According to the length of RNA chain, RNA includes small RNA and long RNA(Hannon et al., 2006). Usually, small RNAs are shorter than 200 nt in length, and long RNAs are greater than 200 nt long(Fatica and Bozzoni, 2014). Long RNAs, also called large RNAs, mainly include long non-coding RNA (lncRNA) and mRNA. Small RNAs mainly include

5.8S ribosomal RNA (rRNA), 5S rRNA, transfer RNA (tRNA), microRNA (miRNA), small interfering RNA (siRNA), small nucleolar RNA (snoRNAs), Piwi-interacting RNA (piRNA), tRNA-derived small RNA (tsRNA)(Chen et al., 2016) and small rDNA-derived RNA (srRNA)(Wei et al., 2013).

**Gene:** A gene is a sequence of nucleotides in DNA or RNA that codes for a molecule that has a function. The transmission of genes to an organism's offspring is the basis of the inheritance of phenotypic trait. These genes make up different DNA sequences called genotypes. Most biological traits are under the influence of polygenes (many different genes) as well as gene–environment interactions. Some genetic traits are instantly visible, such as eye color or number of limbs, and some are not, such as blood type, risk for specific diseases, or the thousands of basic biochemical processes that constitute life. A broad, modern working definition of a gene is any discrete locus of heritable, genomic sequence which affect an organism's traits by being expressed as a functional product or by regulation of gene expression(Group, 2006; Pennisi, 2007).The term gene was introduced by Danish botanist, plant physiologist and geneticist Wilhelm Johannsen in 1909.

**Gene Expression:** In all organisms, two steps are required to read the information encoded in a gene's DNA and produce the protein it specifies. First, the gene's DNA is transcribed to messenger RNA (mRNA)(Alberts et al., 2017). Second, that mRNA is translated to protein. RNA-coding genes must still go through the first step, but are not translated into protein(Eddy, 2001). The process of producing a biologically functional molecule of either RNA or protein is called gene expression, and the resulting molecule is called a gene product.

Genes can acquire mutations in their sequence, leading to different variants, known as alleles, in the population. These alleles encode slightly different versions of a protein, which cause different phenotypical traits. Usage of the term "having a gene" (e.g., “good genes”, “hair color gene”) typically refers to containing a different allele of the same, shared gene. Genes evolve due to natural selection / survival of the fittest and genetic drift of the alleles.

**Genome and Genomics:** In the fields of molecular biology and genetics, a genome is the genetic material of an organism. It consists of DNA (or RNA in RNA viruses). The genome includes both the genes (the coding regions) and the noncoding DNA, as well as mitochondrial DNA and chloroplast DNA. The study of the genome is called genomics.

**Biomarker:** A biomarker or biological marker is a measurable indicator of some biological state or condition. Biomarkers are often measured and evaluated to examine normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention(Siderowf et al., 2018). They can be used to create genetic maps of whatever organism is being studied. The term "biological marker" was introduced in 1950s.

**Agricultural Biomarker:** An agricultural biomarker (identified as genetic marker) is a DNA or RNA or protein sequence that causes disease, is associated with susceptibility to disease, or is associated to the variation of particular phenotype.

## 1.2 Central Dogma of Molecular Biology

The *Central Dogma* is the process by which the instructions in DNA are converted into a functional protein. DNA, or deoxyribonucleic acid, is the hereditary material in all living species. Nearly every cell in an organism has the same DNA. Most DNA is located in the cell of a nucleus though a small amount of DNA can also be found in the mitochondria. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine(C), and 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.

An important property of DNA is that it can replicate, or make copies of itself. There is another molecule produced from DNA also formed of the four nucleotides, but instead of the Thymine (T), it has another nucleotide called uracil (U). Unlike DNA, RNA is a single-stranded molecule.

The basic framework for how genetic information flows from a DNA sequence to a protein product inside cells is known as the central dogma of molecular biology (**Fig.1.1(a)**) The protein is synthesized from the DNA into three stages: (1) transcription (2) splicing and (3) translation (**Fig.1.1(b)**).

**(1) Transcription:** Transcription is the first step in the synthesis of proteins from specific gene sequences. In this process, the genetic information in the DNA is copied into a new molecule of messenger RNA (mRNA) or pre-mRNA with the help of an enzyme called RNA polymerase and several transcription factors. There are main three steps are associated in transcription: initiation, elongation, and termination.

**(2) Splicing:** For most eukaryotic genes(and some prokaryotic ones), the initial RNA that is transcribed from a gene's DNA template must be processed before it becomes a mature messenger RNA(mRNA) that can direct the synthesis of protein. One of the steps in this processing, called RNA splicing involves the removal or splicing out of certain sequences referred to as intervening sequences, or introns. The final mRNA thus consists of the remaining sequences, called exons, which are connected to one another through the splicing process. RNA splicing was initially discovered in the 1970s.

**(3) Translation:** Translation is the process of producing proteins by joining amino acids in the order encoded in the mRNA. An amino acid is determined by three adjacent nucleotides (triplets) in the DNA. This is known as the triplet or genetic code. Each triplet is called a codon and codes for one amino acid. As there are 64 codons, 61 resent amino acids, and three are stop signals. For example, the codon CAG represents the amino acid glutamine, ATG triplet the amino acid methionine it is known as start codon and TAA (Ochre), TAG (Amber), TGA (Opal) are three stop codons out of 64 codons. Translation however takes place on ribosome. When proteins are made in a cell by ribosomes, mRNA directs protein synthesis. The mRNA sequence is determined by the sequence of genomic DNA.

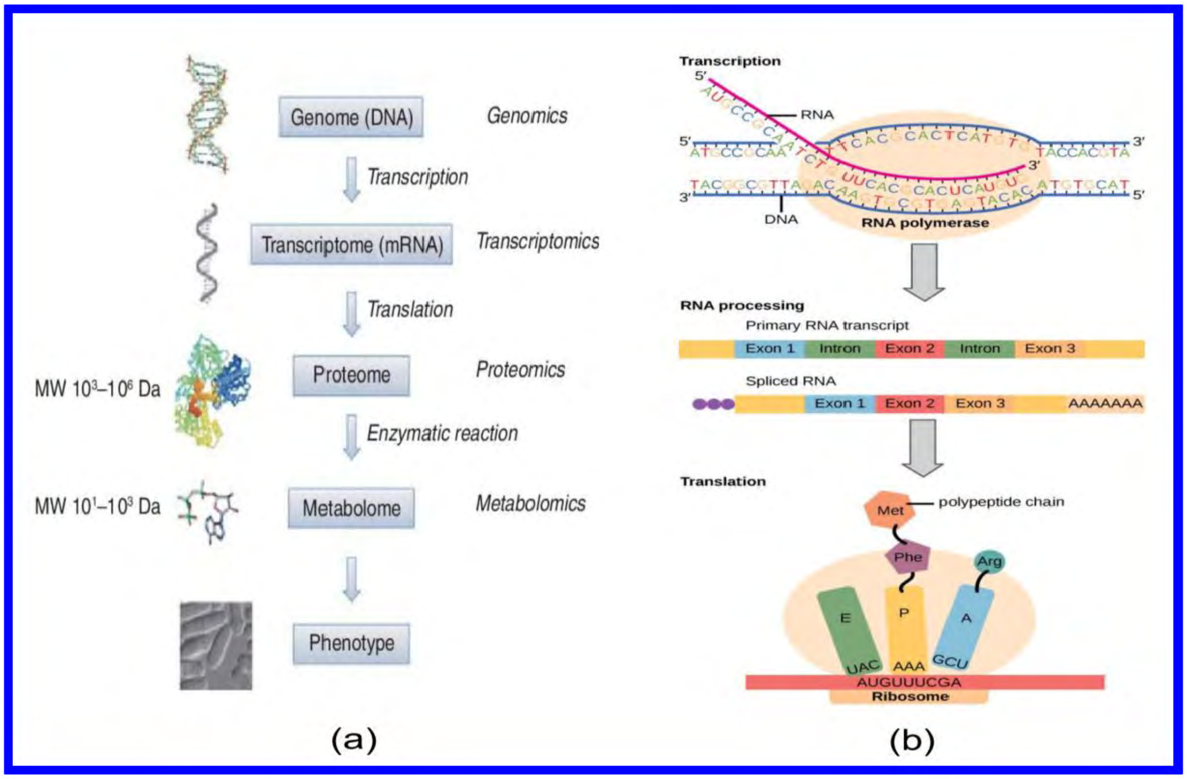


**Fig.1.4.** (a) The Central dogma of molecule biology and (b) Process of gene expression

## 1.3 Field of OMICS Technologies in Genome Research

OMIICS technology is a newly evolved domain of molecular biological research focuses to deal with the generating various omics data and eventually identify the significant DE genes against different experimental conditions. It however refers to the collective technologies used to explore the roles, relationships, and actions of the various types of molecules that make up the cells of an organism. OMICS technologies use high-throughput (HT) computing methods to analyze very large set genes, gene expression, or proteins either in a single procedure or in a combination of procedures. Bioinformatics tools have been used to analyze this huge amount of data generated by OMICS technologies. Bioinformatics is therefore act as the interface between molecular biology, computer, and statistical science. Classical and modern statistical methods as well as up-to-date computing technologies are nowadays widely used by bioinformaticians to cope with the all kinds of molecular data produced from OMICS technology. There are well established a pool of OMICS databases (NCBI, EMBL-EBI, DDBJ, Gene Bank, SwissProt, Uni ProtKB, Reactome, Int Act, PRIDE, PIR etc.) those are extensively used by bioinformaticians to extract and compare genes, gene sequences, gene pathways etc. under different biological conditions. The most commonly research areas of OMICS technologies include: (A) genomics (B) transcriptomics (C) proteomics (D) metagenomics (E) toxicogenomics (F) pharmacogenomics etc.

**(A) Genomics:**

Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, 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 alleles within the genome and other interactions such as epistasis, pleiotropy and heterosis. Genomics harnesses the availability of complete DNA sequences for entire organisms and was made possible by both the pioneering work of Fred Sanger and the more recent next-generation sequencing technology.

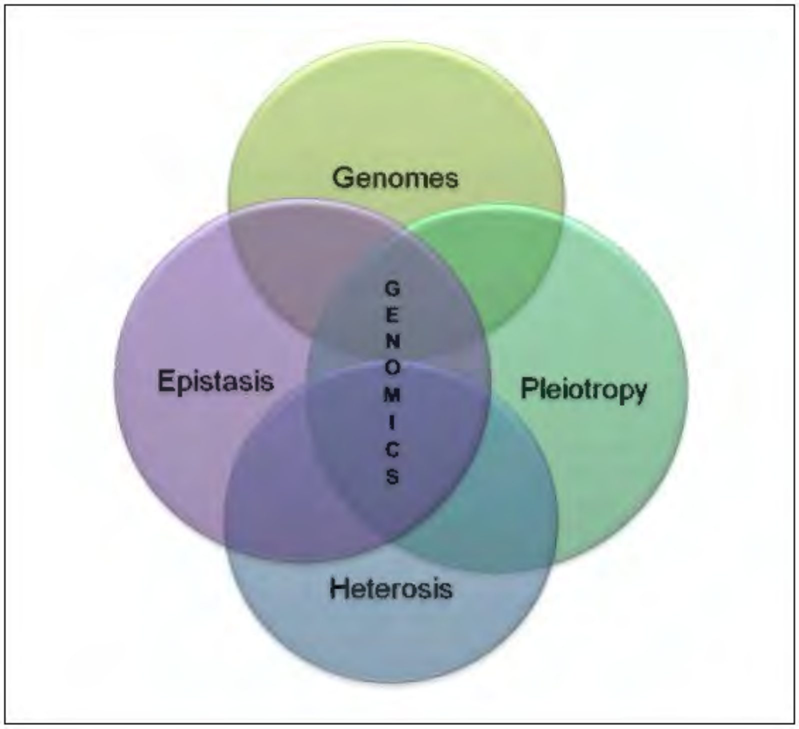


**Fig.1.5.** Genomics studies the genomes of whole organisms and other intragenomic interactions.

### (B) Transcriptomics:

The transcriptome is the complete set of transcripts in a specific type of cell or tissue. It refers to the set of all RNA molecules from protein coding (mRNA) to noncoding RNA, including rRNA, tRNA, lncRNA, pri-miRNA, and others. Transcriptome may apply to an entire organism or a specific cell type. Microarray (or “chip”) technology, and more recently high throughput next generation (NextGen) DNA sequencing, has made assessing the transcriptome a routine laboratory practice. Generally, the goal of transcriptome analysis is to identify genes differentially expressed among different conditions, leading to a new understanding of the genes or pathways associated with the conditions. Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cell populations, or in response to different treatments. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and for understanding development and disease. The key aims of trancriptomics are to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs, to determine the transcriptional structure of genes, in terms of their start sites, 5´ and 3´ ends, splicing patterns and other post-trancriptional modifications and to quantify the changing expression levels of each trancript during development and under different conditions. Various technologies have been developed to deduce and quantify the transcriptome, including hybridization-or sequence-based approaches. Transcriptome analysis requires an appropriate statistical method with a multiple comparison test to interpret global changes in the expression of thousands of genes.

### (C) Proteomics:

Proteomics is the large-scale study of proteomes. A proteome is a set of proteins produced in an organism, system, or biological context(Anderson and Anderson, 1998). It may refer to, for instance, the proteome of a species (for example, *Oryza sativa*) or an organ (for example, flower). The proteome is not constant; it differs from cell to cell and changes over time. To some degree, the proteome reflects the underlying transcriptome. It also differs from Transcriptomics because mRNA is not always translated into protein or single mRNA may produce several proteins due to alternative splicing(Rogers et al., 2008). However, protein activity (often assessed by the reaction rate of the processes in which the protein is involved) is also modulated by many factors in addition to the expression level of the relevant gene .Proteomics is used to investigate (**Fig.1.3**)

- When and where proteins are expressed;
- Rates of protein production, degradation, and steady-state abundance;
- How proteins are modified (for example, post-translational modifications (PTMs) such as phosphorylation);
- The movement of proteins between subcellular compartments;
- The involvement of proteins in metabolic pathways;
- How proteins interact with one another.

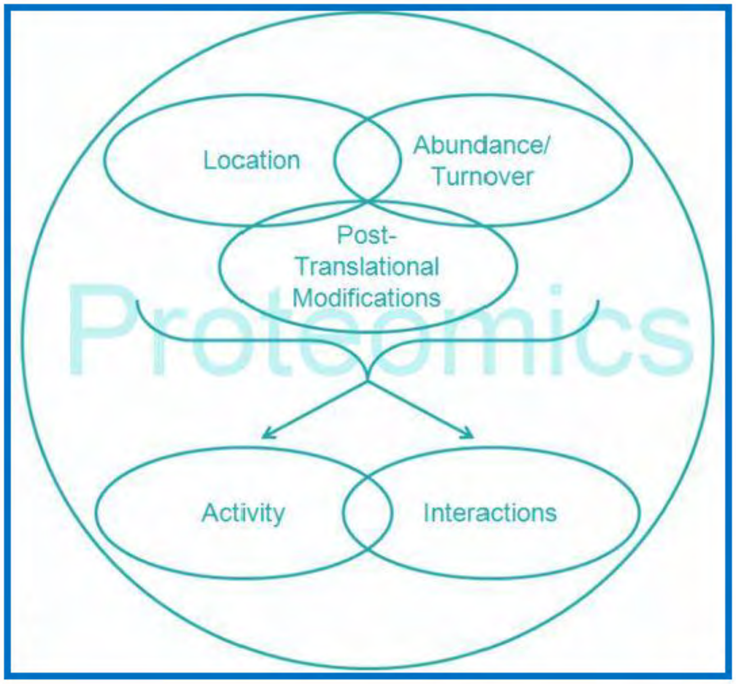


**Fig. 1.6.** Areas of proteomics.

Proteomic experiments generally collect data on three properties of proteins in a sample: location, abundance/turnover, and post-translational modifications. It may be possible to infer a protein's interaction partners among others that are colocalised with it, or to assess whether a protein is active from its post-translational modifications.

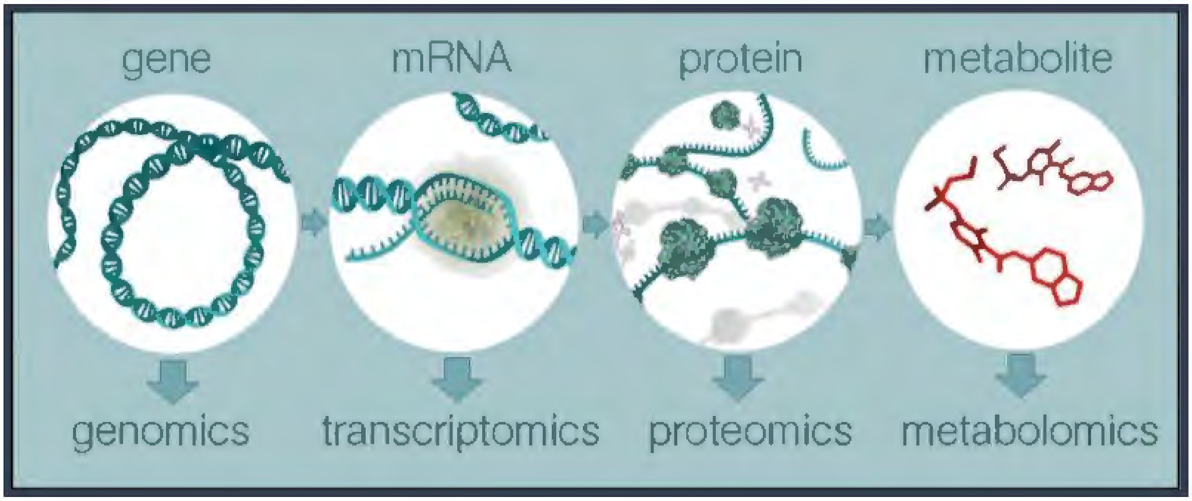


**Fig.1.7.** An overview of the four major "omics" fields, from genomics to metabolomics.

**(D) Metagenomics:**

Metagenomics is an umbrella term, covering a range of study types that use high-throughput DNA sequencing to characterize microbial systems. These include whole-genome shotgun (WGS)-sequenced metagenomic and metatranscriptomic studies, as well as amplicon-based approaches, which target specific marker genes (often rRNA genes) (**Fig.1.5**).

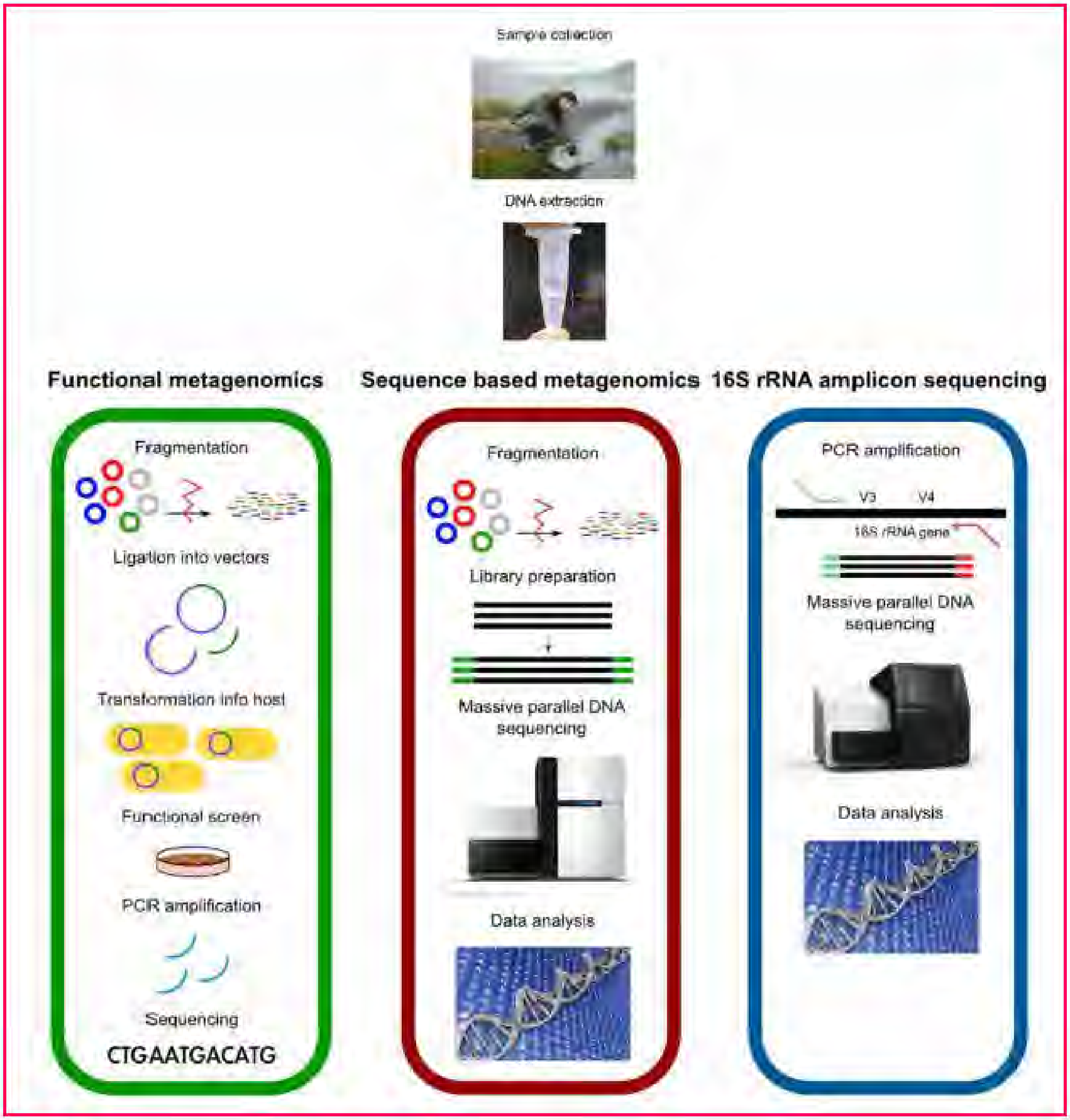


**Fig.1.8.** Different steps in metagenomic data analysis process

## 1.4 Literature Review for Genome Data Analysis to Detect Agricultural Biomarkers

The power of information technology (IT) and robust statistical methods as well as frequent development of modern bioinformatic tools have enabled molecular biologists, statisticians, and computer scientist to uncover the reasons of hidden genomic variability and potentiality in plants and animals in terms of genome mapping, sequencing, data storage, and bioinformatic analyses. Today, next-generation sequence technologies have also led to spectacular 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. In the following subsequent sections discussions have been made for different genome data analysis to detect agricultural biomarkers with the application of robust statistical approach and *in silico* solutions using different bioinformatic tools in the aspect of genetic linkage or mapping analysis e.g. qtl, gene expression clustering, genome-wide study of RNAi genes in wheat (*Triticum aestivum),* robust GWAS and finally classification of metagenomics data.

### 1.4.1 Statistical Approaches for Quantitative Trait Loci (QTL) Detection

Trait variations in animals and plants are observable largely due to the variation of molecular genetic factor that is called DNA or gene or biomarker and sometimes for existing ambient environment. Most of the phenotypes (traits) in organisms are in quantitative in nature(Haley and Knott, 1992). Examples include number of seeds produced in per plant to study the evolutionary fitness, blood pressure to study the hypertension, milk output in dairy breeding etc.(Broman et al., 2003). Variation in such quantitative traits is often due to the effects of multiple genetic loci and for environmental factors. qtl analyses are however specialized techniques that construct the genetic linkage maps to locate loci (qtls) that affect a quantitative trait and estimate the effect of qtls on the trait (Guiderdoni et al., 1992). qtl analysis allows researchers in fields as diverse as agriculture, evolution, and medicine to link certain complex phenotypes to specific regions of chromosomes. The goal of this process is to identify the action, interaction, number, and precise location of these regions (Broman et al., 2003). Due to modern innovation in molecular biology, it has been easier to make fine-scale genetic maps for a large number of organisms by defining the genomic positions of a number of genetic markers (RFPL, isozymes, RAPDs, AFLP, VNTRs, etc.) and to find a comprehensive classification of marker genotypes by means of co-dominant markers(Mollah and Eguchi, 2010; Zeng, 2006). These rapid expansions of associated techniques in molecular biology have enabled the plant breeders, physiologists, pathologists and other plant scientists to gear up and expedite the detailed genetic mapping and analysis of qtls. Thoday first introduced the idea of using two markers to bracket a region for testing qtls (Martínez and Curnow, 1992). Lander and Botstein carried out a similar, but much upgraded, method to use two adjacent markers to test the presence of a qtl in the interval by performing a Likelihood Ratio Test (LRT) at every position in the interval, which is called Standard Interval Mapping (SIM) or simply Interval Mapping (IM) method(Lander and Botstein, 1989). However, SIM can bias identification and estimation of qtls when multiple qtls are located in the same linkage group (Haley and Knott, 1992; Jansen et al., 1995; Lander and Botstein, 1989). Besides, it is also not effective to use only two markers at a time for mapping analysis. To deal with these difficulties, qtl mapping combines SIM with the multiple marker regression analysis is studied by Jansen 1993 and Zeng 2006 termed this combination as Composite Interval Mapping (CIM). It avoids the use of multiple marker intervals to deal with the problems of mapping multiple qtl by conditioning a test for a qtl on some linked or unlinked markers that diffuse the effects of other potential qtls. A comparative evaluation of SIM and CIM approaches has been demonstrated in **Chapter Two** in the scenario of simulation and real data analysis.

### 1.4.2 Statistical Approaches for Gene Expression Data Analysis to Identify Biomarker Genes

In OMICS field, the transcriptomics technology has aimed to study the expression levels of thousands of genes to be investigated simultaneously. Both identification of differentially expressed (DE) genes and clustering or classifications of genes/samples are equally essential in the transcriptomics (microarray/RNA-Seq) data analysis. However, this OMICS dataset poses different computational challenges because of large number of transcripts surveyed with small sample sizes. The detection of genes that are DE between two or more conditions is essential, since it reduces the dimensionality of the transcripts by a set of DE genes that are most informatics under study. Based on this smaller set of DE genes, clustering, or classification become much easier by any supervised/unsupervised learning algorithm. Each of these tasks is conducted by separate statistical algorithm. There are several types of statistical procedures that are used to identify DE genes including classical parametric approach (t-test, F-test/ANOVA and likelihood ratio test), non-parametric approach (e.g. KW-test) and empirical Bayes approaches(Do et al., 2005; Efron et al., 2001; Gottardo et al., 2006; Kendziorski et al., 2003; Kruskal and Wallis, 1952; Newton and Kendziorski, 2003; Robinson et al., 2009; Smyth, 2004; Tusher et al., 2001; Wang et al., 2011; Wilcoxon, 1945). However, most of the aforementioned algorithms are not robust against outliers(Gottardo et al., 2006; Mollah et al., 2007, 2012, 2015). Thus, they might be produced misleading results in the presence of outlying observations. Transcriptomics observations (gene expression) are often corrupted by outliers that arise due to several steps involved in the experimental process, from hybridization to image analysis(Gottardo et al., 2006; Mollah et al., 2012). Non-parametric approaches are somewhat robust against outliers in the case of large sample; however, these approaches are sensitive to outliers in the case of small sample sizes. To overcome this problem, the $\beta$-divergence-based empirical Bayes (BetaEB) approach(Mollah et al., 2012) and $\beta$-ANOVA approach (Mollah et al., 2015) was developed for the robust identification of DE genes. These approaches perform well in the presence of outlying observations with up to 50% genes for both small and large sample cases. However, $\beta$-divergence-based approaches require the appropriate selection of the tuning parameter $\beta$ by cross-validation separately for each gene that is much time consuming. Similarly, there are several methods for genes/samples clustering/classification. For the purpose of RNA-Seq/microarray gene expression data classification, various classification methods have been proposed such as Fisher, Bayes, Naive Bayes (NB), Ada-boost, logistic, neural network, k-nearest neighbors (KNN), support vector machines(SVM) (Sain and Vapnik, 2006), bagging support vector machine(bagSVM), random forests(RF) (Breiman, 2001), classification and regression trees (CART) (Breiman et al., 2017) etc. However, most of these algorithms discussed before are very much sensitive to outlying observations and might be produced misleading results. Generally, there are three ways to obtain robust estimation against outlying observations. however, application of robust methods is complicated than using the traditional methods and deletion of outlying observations loses the information of the dataset. Hence, applying the transformation technique is the suitable option for reducing the outlier effects. Several authors (Atkinson, 2018; Box and Cox,

1964) have been proved that transformation based robust methods perform better than the traditional methods in reducing outlier effects. Thus, in **Chapter Three** we consider logistic transformation for reducing outlier effects from the dataset instead of the other transformations.

### 1.4.3 Statistical Approaches for Genome Wide Analysis to Identify RNA Silencing Machinery Genes in Wheat (*Triticum aestivum*)

RNA silencing is an important molecular occurrence that takes place in eukaryotic groups. This mechanism happens with a small RNA molecule that interferes using a particular nucleotide sequence (Cao et al., 2016). Previous studies implied that there are two types of small RNA molecules nearly size of 21-24 nucleotides are produced in multi-cellular eukaryotes termed as microRNA (miRNA) and short interfering RNA (siRNA) (Qian et al., 2011). These RNA molecules play roles in performing different molecular, biological and cellular processes in plants during development and growth, metabolism, anti-viral and anti-bacterial defense (Carrington and Ambros, 2003; Chen, 2012; Finnegan and Matzke, 2003; Lai, 2003; Van Ex et al., 2011). RNA silencing in plants is however initiated by double-stranded RNAs (dsRNA) that produce small RNAs are known as microRNAs (miRNAs) or small-interfering RNAs (siRNAs)(Qian et al., 2011). Creation and role of these small RNAs largely rely on three key gene families, Dicer-like (DCLs), Argonauts (AGOs) and RNA-dependent RNA polymerases (RDRs) (Baulcombe, 2004; Chapman and Carrington, 2007; Vaucheret, 2006). A complete cycle of RNA silencing process takes three common steps: initiation, maintenance, and signal amplification(Cao et al., 2016). DCL, AGO and RDR are known as the RNA silencing machinery genes work through the generation of small RNA molecules in plants. Among these genes, plant DCL proteins mainly process long double-stranded RNAs into mature small RNAs (Carrington and Ambros, 2003; Chapman and Carrington, 2007; Qian et al., 2011). A DCL protein has been characterized as the presence of six domains such as DEAD, Helicase-C, DUF283, PAZ, RNaseIII and double-stranded RNA-binding motif (DSRM) (Carmell and Hannon, 2004; Margis et al., 2006). DCL gene families are also available in higher-class insects, protozoa, and some fungi (Cao et al., 2016). AGO proteins are very particular small RNA-binding components well known as the core elements of RNAi pathways (Vaucheret, 2008). Small RNAs produced by DCLs are conveyed for gene expression into the AGO-containing RNA-induced silencing complexes (RISCs). Next these small RNAs direct AGOs to the target mRNA, accomplishing sequence-specific regulation of gene expression (Vaucheret, 2008). AGO proteins consists of several functional domains such as DUF283, PAZ, MID, and PIWI(Hutvagner and Simard, 2008). RDR is the third key protein also plays significant roles in RNA interference (RNAi) pathway for eukaryotic gene expression. These enzymes however have a common conserved catalytic domain called RNA-dependent RNA polymerase (RdRp) which is essential for initiation and amplification of the silencing signal (Schiebel, 1998). Successive investigation so far reveals that there are multiple copies of DCL, AGO and RDR genes are present in plants and animals. All members of these gene families take part in diverse roles in RNA silencing pathway. For instance, the genome of Arabidopsis thaliana possesses four DCL proteins (DCL1-DCL4) that definitely generate different kinds and sizes of small RNAs (Bologna and Voinnet, 2014). To activate various important biological functions in eukaryotic cell, different small RNAs produced in cell with diverse functions but they all are the associate members of DCL, AGO and RDR gene families. In Arabidopsis thaliana, four AtDCLs, 10 AtAGOs and six AtRDRs genes were identified(Vaucheret, 2008). Rice (*Oryza sativa*), a crop of monocot group, possesses eight OsDCLs, 19 OsAGOs and five OsRDRs genes, in which OsAGO2 gene exhibited specific up regulation in response to salt and drought(Kapoor et al., 2008; Qin et al., 2018). Also, in tomato (*Solanum lycopersicum*), seven SlDCLs, 15 SlAGOs and six SlRDRs genes were identified(Bai et al., 2012). Five ZmDCLs, 18 ZmAGOs and five ZmRDRs genes were recognized in maize genome(Qian et al., 2011). There are four VvDCLs, 13 VvAGOs and five VvRDRs genes were detected in grapevine (Vitis vinifera) (Zhao et al., 2015). Similarly, a total of five, seven, and eight CsDCLs, CsAGOs, and CsRDRs, respectively, have been identified in cucumber(Gan et al., 2016). On the other hand, the genome of allopolyploid species of Brassica napus possessed eight BnDCLs, 27 BnAGOs, and 16 BnRDRs(Cao et al., 2016; Zhao et al., 2016). Recently, in total four CaDCLs, 12 CaAGOs and six CaRDRs genes have been identified in pepper (Capsicum Annuum L.) (Qin et al., 2018). However, these potential genes in various important plants show considerable divergence and have indispensable role in different genomic functions.

Recent studies have revealed the DCL, AGO and RDR gene families in *Brassica napus*, maize, arabidopsis, rice, tomato, grapevine, cucumber, and pepper (Bai et al., 2012; Cao et al., 2016; Gan et al., 2016; Kapoor et al., 2008; Qian et al., 2011; Qin et al., 2018; Vaucheret, 2008; Zhao et al., 2015, 2016) but not in wheat. Wheat, a cereal grain is a global common staple food. It is the second most produced cereal crops after maize in the world (http://www.fao.org/worldfoodsituation/csdb/en/). It is also the vital source of carbohydrates and is the leading source of vegetal protein in human food. In wheat, rarely was there any investigation about genome-wide identification and characterization of the RNAi machinery components carried out until today. An attempt was made in this study with the help of availability of complete genome sequencing of wheat collected from TIAR to perform a comprehensive bioinformatics analyses to identify the DCL, AGO and RDR gene families that are known as the key components of RNA silencing machinery in wheat (*Triticum* aestivum). Furthermore, these three well known RNA silencing machinery gene sets were validated previously in wet lab experiment to investigate their expression in different organs and tissues as well as at reproductive stages(Gan et al., 2016; Kapoor et al., 2008; Qin et al., 2018). Also the resistance ability of these RNAi proteins were explored against various abiotic( drought, salt, heat, cold etc.) and biotic(diseases viz., *Sclerotinia scletotiorum*, yellow leaf curl virus, tomato mosaic virus, cucumber mosaic virus, potato virus Y etc. ) factors in different crops(Bai et al., 2012; Kapoor et al., 2008; Qin et al., 2018; Zhao et al., 2016) as well. In this study, an attempt was made to explore expression profile of seven TaDCL candidate genes in two organs such as leaves and roots as well as at the same time expression analyses were also carried out against drought in *T. aestivum*. The *in silico* analysis of all identified 62 genes and the expression study of these seven TaDCL genes have been discussed in **Chapter Four** in leaves and roots and in response of drought stress.

### 1.4.4 Statistical Approaches for Genome-Wide Association Studies to Identify Biomarker SNPs

Genome-wide association studies (GWAS) technique is widely used in human genetics research to identify genes associated with complex diseases and in agricultural research to identify genes associated with quantitative traits such as yield and productivity (Huang et al., 2010; Speliotes et al., 2010). Single nucleotide polymorphism (SNP) markers can now cover the genome with high density and are inexpensive to obtain. Evaluations based on SNP genotypes can be computed as soon as DNA can be obtained. In genome-wide association (GWA) studies, hundreds of thousands of single-nucleotide polymorphisms (SNPs) are assayed using high-throughput genotyping technologies and are tested for their associations with experimental outcomes of interest(Liu et al., 2013a). The new genetic associations identified by these studies can be used to improve the detection, treatment and prevention of certain diseases, particularly when used in conjunction with other clinical biomarkers(Liu et al., 2013a). To date, the most frequently used GWA study design has been the case-control design, in which allele frequencies in patients with the disease (cases) are compared to those without the disease (controls) among unrelated individuals, or allele frequencies in patients who responded to the treatment are compared to those who did not respond to the treatment(Liu et al., 2013a). The goal of the case-control studies is to identify SNPs associated with the outcome of interest, such as disease status or responder or non-responder status.

GWA studies involve large amounts of data. Proper statistical methods are needed to analyze such large datasets in order to draw meaningful conclusions. Hidden population structure or stratification and polygenic effects (genetic relatedness) are two most common factors present in such large-scale data needed to address properly. Population stratification (PS) refers to allele frequency differences between cases and controls unrelated to the outcome of interest, but due to sampling from populations with different ancestries. Correcting for population stratification is very important in GWA studies(Campbell et al., 2005; Liu et al., 2013a) since it can cause false positive findings. Large-scale GWA studies with many subjects are particularly vulnerable to population stratification artifacts(Li and Yu, 2008; Xu et al., 2009). Because of the large number of subjects, it is likely that there are some unrecognized hidden population structures may be responsible for systematic differences being detected in SNPs between cases and controls.

Inconsistency of the results across various GWASs might be attributed to the heterogeneity of populations. False discovery might be a possible cause, although many studies have attempted to employ a conservative multiple testing method, the Bonferroni adjustment, to avoid it. This false discovery might come from population stratification; that is, the different allele frequencies between cases and controls are attributed to spurious genetic associations caused by systematic differences in ancestry(Cardon and Palmer, 2003; Marchini et al., 2004; Shin and Lee, 2015).

There is however, a number of statistical approaches proposed earlier for genome-wide association mapping to address the effects of population structure. The most commonly used statistical methods to avoid the bias of population stratification or genetic relatedness are genomic control (GC) (Devlin and Roeder, 1999), structured association (SA)(Pritchard et al., 2002), and principal component analysis (PCA)(Patterson et al., 2006; Price et al., 2006). GC approach modifies the association test statistic by a common factor for all SNPs to correct for PS (Liu et al., 2013a). 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., 2002).The SA method is useful for small datasets(Liu et al., 2013a). Nevertheless, the software package STRUCTURE is computationally intensive and cumbersome for large-scale genome-wide association studies (Price et al., 2006).

Price et al. 2006 suggested another method based on principal component analysis. In this technique, EIGENSTRAT program uses several top principal components (PCs) and applies them as covariates in GWA analysis (Liu et al., 2013a). These top PCs are selected using EIGENSTRAT 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., 2013a). 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., 2013a). The results obtained from PCA adjustment may be misleading too if there are outliers (Liu et al., 2013a). Another improved method was proposed by Li and Yu 2008 to deal with the fact of population stratification for the presence of hidden population structure population-based GWAS. 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).

GWA results based on earlier methods could be misdealing in terms of false discovery rate (FDR) and statistical power if the SNP data contains outliers in the phenotypic trait. Li et al. 2013 made however an improvement in PCA technique to overcome the limitation of analysis in presence of outliers in GWA mapping. In recent times in bioinformatics research, the applications of mixed linear model (MLM) 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. To address the issues of adjustment of population stratification and account for population structure and genetic relatedness (polygenic effects), Kang et al.2010 Zhang et al.2010 and Jeffrey 2011 have proposed such effective approaches to apply linear mixed model for large scale GWAS. Their approaches have been executed in software programs TASSEL(Yu et al., 2006)(Yu et al., 2006), EMMA(Hyun et al., 2008), EMMAX(Kang et al., 2010) ,rrBLUP(Endelman, 2011), GAPIT(Lipka et al., 2012). In GWAS, results produced from these approaches are effected by outliers.

MLM performs significantly in detecting causal genetic variants for phenotypic trait of interest in terms of computational efficiency and consistency of the results such as higher statistical power and lower false discovery rate (FDR). However, there is no investigation has been done yet about the performance evaluation of the mixed linear model for GWAS in presence of outliers in the phenotypic trait. A robust statistical approach has been proposed for correcting population stratification and polygenic effects in presence of outliers in **Chapter Five**. The performance of our approach has been investigated using simulated and real dataset in terms of power and FDR in presence of outlying objects in the phenotypic traits.

### 1.4.5 Statistical Methods for Metagenomic Data Analysis

Metagenomics is the application of modern genomic techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species. The classification of functional metagenomes is the big statistical challenge from the different sources of metagenome dataset. The classification of potential metabolic function from microbial community using metagenomic information is an important task of metagenomics research. The different microbial process has different metagenomic function for several environments(Dinsdale et al., 2008, 2013). Metagenomics is the complete scheme of microbial activity and gives easier interpretation of thousands of proteins using BLAST matches algorithm(Parks and Beiko, 2010). There are many web based tools available for statistical analysis of metagenomic dataset but not all the analysis tools provide accurate and valid results(Arndt et al., 2012). Some traditional multivariate statistical methods such as principal component analysis (PCA), multidimensional scaling (MDS), and canonical discriminate analysis (CDA) are often used for analysis of genomic data and microbial community (Ramette, 2007). The multivariate statistical techniques are plays vital role for classification and visualization of metagenomics data from several microbial community. The metagenomic data profiling from the different environments and its classification is important for separation of functional metagenomes. The MetaGUN is the three-stage gene selection method for gene prediction for metagenomic fragments using support vector machine (SVM)(Jiang et al., 2017; Liu et al., 2011). To explore the universe of metagenome, *k*-nearest neighbor method is significant for the several microbial communities(Aßhauer et al., 2014). AdaBoost is the efficient method for analyzing the gigantic metagenomic data and it is challenging task for bioinformaticians/computer scientists(Sharma et al., 2015). The prediction of ribosomal protein in plants, the machine learning method Random Forest is very much useful (Breiman, 2001; Breiman et al., 2017; Carvalho et al., 2017; De'ath, 2002). The statistical test is very important for the identification of potential metabolic function within and between environments based on the different microbial community. Random forest method is efficient for the robust classification of high dimensional complexity data like as the microbial community data. It is the ensemble learning method for classification and regression multiple patterns datasets.

High dimensional dataset with large number of metabolic features or metabolic functions or metabolic variables is a very basic problem. Therefore, it is essential to select the proper feature selection method for classification of large dimensional metagenomics dataset. In **Chapter Six**, beta *t*-statistic for feature selection of metagenomic data from the several microbial community then applied random forest algorithm for efficient classification of functional metagenomes.

## 1.5 Objectives of the study

Objectives of this study in the thesis have been summarized below based on the previous discussions for suitable understanding of the readers:

1. To evaluate the performance of standard interval mapping (SIM) and composite interval mapping (CIM) methods in qtl detection for backcross population.
2. To apply logistic transformation on gene expression data for both robust DE genes identification using SAM and robust classification of genes using support vector machine (SVM) approach.
3. To explore the genome-wide identification, characterization, phylogenic analysis of RNA silencing machinery genes(DCL,AGO and RDR) and to explore the expression profile of identified seven Dicer-like (DCL) genes in wheat (*Triticum aestivum*).
4. To robustify the Linear Mixed Model using outlier modification rule and its application to identify important SNPs influencing flowering time of rice.
5. To analyze the statistical performance of different classification approaches to group the functional metagenomes.

## 1.6 Outline of the study

This thesis has been organized with seven chapters mentioned below:

**Chapter One** is an introductory chapter, which describes importance and scope of genome research in biology particularly in agriculture, fundamental molecular biological principles, role of different OMICS fields that produce various molecular sequencing data. In section 1.4, a broad literature reviews have been presented successively and at the end of this chapter several objectives of this study have been stated.

In **Chapter Two**, statistical analysis between the existing interval mapping (IM) methods in qtl detection for backcross population in the context of simulated and real rice phenotypic and marker genotypic data has been presented. Standard Interval Mapping (SIM) techniques based on maximum likelihood method (Haley and Knott, Extended Haley and Knott), Multiple Imputation(IM) based on regression technique and Composite Interval Mapping (CIM) approaches were evaluated for multiple qtl mapping for estimation of precise location of the potential markers in different chromosomal location.

In **Chapter Three,** application of logistic transformation on gene expression data for both robust DE genes identification using SAM and robust classification of genes using support vector machine (SVM) approach has been discussed along with its application to identify biomarker genes influencing rice blast disease and finally identified some hub genes which are predicted to possess some resistance power against some important biotic and abiotic stresses.

In **Chapter Four,** a genome-wide investigation in terms of identification and characterization and phylogenetic analysis of RNA silencing machinery genes in wheat (*Triticum aestivum*) with *Arabidopsis thaliana*, rice and expression analysis in wet lab experiment following qRT-PCR of seven wheat Dicer-like (DCL) genes in leaves and roots as well as in response of drought stress have been demonstrated.

In **Chapter Five,** a robust linear mixed model approach has been developed for performance evaluation in terms of false discovery rate (FDR) and statistical power in presence of outlier in the study phenotypic trait in genome-wide association studies (GWAS) in the context of two varying heritability proportions 0.2 and 0.3. Additionally, different bioinformatic analyses were also carried for genome-wide characterization of the identified SNP makers using the proposed approach from SNP data related to rice flowering time

In **Chapter Six,** performance evaluation for classification of functional metagenomes recovered from 10 different environmental samples has been demonstrated using Beta-*t* Random Forest approach compare to other five KNN, classification methods: Naïve Bayes, SVM, KNN, and AdaBoost.

In **Chapter Seven**, a sum up has been presented for general conclusion and future research objectives.

# CHAPTER TWO

## EXPLORING BETTER QTL MAPPING APPROACH BY COMPARATIVE STUDY

# EXPLORING BETTER QTL MAPPING APPROACH BY COMPARATIVE STUDY AND ITS APPLICATION TO IDENTIFY IMPORTANT MARKER GENES INFLUENCING MAIZE PLANT HEIGHT

## 2.1 Introduction

Phenotypic variations in living creature are observed due to the variation of molecular genetic factors that is called DNA or gene or biomarker. Most of the phenotypes (traits) in organisms are in quantitative in nature (Haley and Knott, 1992). Examples include to explore the development  suitability of a plant by taking into consideration the seed count for each plant, insulin level to study the diabetic stage of a patient, number of eggs laying each variety in poultry breeding,  etc. Distinctions within these quantifiable characters are generally for the reason of the influence of various chromosomal loci and for other external issues. In genetics, a quantitative trait locus(qtl) is described by a portion within a genome which is connected for the consequence on a quantitative characteristic (Haley and Knott, 1992). A qtl also explains a particular gene or might be the bunch of concomitant genes, which make influence on the organism's attributes. qtl analyses are however specialized techniques that construct the genetic linkage maps to locate loci (qtls) that impact a quantitative characteristic and assess the outcome of qtls on the characteristic (Guiderdoni et al., 1992). This study technique lets investigators in arenas as many as crop science, developmental trend in organisms, and drug discovery to link some unusual phenotypes to some particular parts of chromosomes. Mating

The basic step for mapping qtl includes organizing a cross for two inborn species opposing largely in a measurable characteristic: separating offspring are recorded simultaneously about characters along with several number of biomarkers (Mollah and Eguchi, 2010) . A mating of two parent ingrained genotypes $M_1$ and $M_2$ is accomplished to generate an $F_1$ population. $F_1$ descendants are completely heterozygotes with the identical genotypes. Usually, the separating descendants are formed by a backcross ($B_1$=$F_1$×parent) or an intercross ($F_2$=$F_1$×$F_1$).

At the advent of up-to-date improvement in different technology, it is now easier and a common work to make fine-tuning large size gene/maker atlases designed for a big number of species in genomics by describing the gene/maker locus for many gene markers (RFPL, isozymes, RAPDs, AFLP, VNTRs, etc.) for determination a detailed cataloging of marker genotype(s) by dint of co-leading markers (Mollah and Eguchi, 2010; Zeng, 2006). These quick developments of allied methods in molecular biology have supported the plant breeders, physiologists, pathologists and other plant scientists to gear up and expedite the detailed genetic mapping and analysis of qtls. A pioneering researcher in the area of genomics Thoday (1960) originally initiated the concept of utilization of two markers to set an area to test qtl. Lander and Botstein carried out a corresponding, but better technique for using two neighboring gene markers for testing the occurrence of a qtl within range by applying a Likelihood Ratio Test (LRT) at each locus within that interval, which is called Standard Interval Mapping (SIM) or Simple Interval Mapping (SIM) or simply Interval Mapping (IM) approach (Lander and Botstein, 1989). However, IM may affect the favor for the detection and measuring of qtls at the time of getting several qtls which are found within similar linkage cluster (Haley and Knott, 1992; Jansen et al., 1995; Lander and Botstein, 1989). Besides, effective results will not produce when two biomarkers are used simultaneously during gene mapping study(Mollah and Eguchi, 2010). To get solved these limitations, qtl mapping technique joins SIM approach along with multiple marker regression examination is studied by Jansen (1993) and Zeng (2006) and termed this combined idea as Composite Interval Mapping (CIM). It omits the usage of several marker intervals for managing related complexities of mapping multiple qtl by acclimatizing an experiment aimed at identifying qtl on nearly concomitant or unrelated markers which drawn-out the influence of further latent qtls.

## 2.2 Materials and Methods

### 2.2.1 Simple Interval Mapping (SIM)

Analysis of variance (ANOVA) is the basic tool for qtl mapping which is called Marker Regression Method (MR). However, the power of this technique decreases at removal of characters of their genotypes are omitted from the set of markers along with markers are broadly spread out (Broman et al., 2003). There are also a number of statistical methods to overcome this weakness of ANOVA for qtl mapping analysis such as Standard Interval Mapping (SIM) established using maximum likelihood(ML) (Lander and Botstein, 1989), regression based(Haley and Knott, 1992) methods namely Haley and Knott (HK) , Extended Haley and Knott (eHK), Multiple Imputation methods(IMP). The steps of this study have been briefly demonstrated in **Fig. 2.1**.

SIM Methods using ML and regression, known as the common and generally implemented as interval-mapping methodologies. The methods make usage of a genomic plot of the input markers and similar to ANOVA, assume the attendance a particular qtl. When using SIM technique, every position is taken into account altogether as well as the logarithm of the odds ratios are determined using the model, which results that there is an actual qtl.

Odd ratio is associated to the Pearson correlation coefficient among phenotypes and marker genotypes for each subject in the experimental cross.

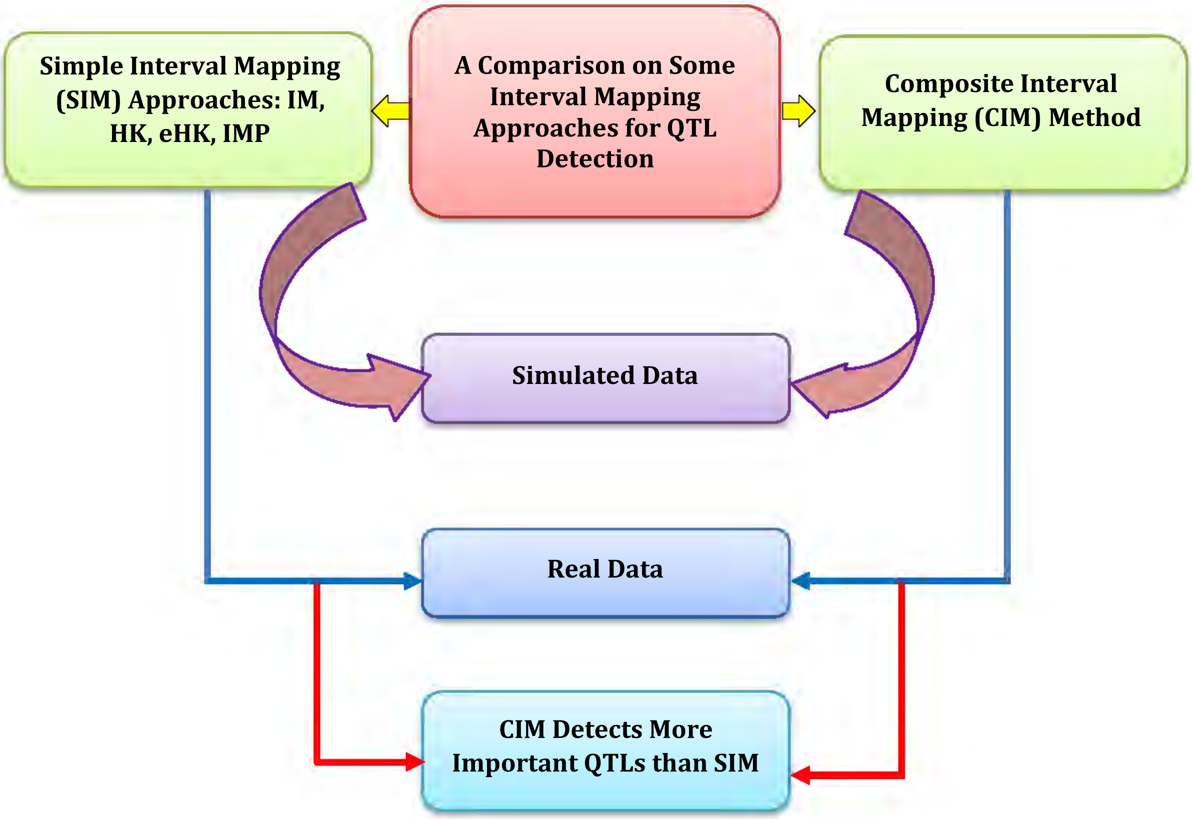

Simple Interval Mapping (SIM) Approaches: IM, HK, eHK, IMP
A Comparison on Some Interval Mapping Approaches for QTL Detection
Composite Interval Mapping (CIM) Method
Simulated Data
Real Data
CIM Detects More Important QTLs than SIM

**Fig.2.1.** Schematic diagram of this study

SIM uses two adjacent markers to check possible occurrence of qtls within intervals using a likelihood ratio test (LRT) at every single point in the interval (Lander and Botstein, 1989). Actually, qtl influences are termed with fixed or random (Xu, 1998). When fixed effects qtl model are used, allelic switch influences commonly predicted and proved along with the qtl discrepancies are measured following projected allelic effect(Xu, 1998). When random effects qtl model, the qtl effects and qtl inconsistency are straight projected and examined(Lander and Botstein, 1989; Xu, 1998). As the conditional expectations of the qtl genotype provide that neighboring marker genotypes are unidentified in ML centered IM approach and termed as the qtl effect model as a random effects model (REM)(Lander and Botstein, 1989). Whereas for HK regression based IM model, the conditional expectation of the qtl genotype provided that neighboring marker genotype is called as fixed and this model can be termed as a fixed effect model (FEM) (Kao, 2000).

### 2.2.1.1 Regression Based SIM

If we assume that no epistasis (qtl×qtl interactions) between qtls, no intervention (qtl× environmental interactions) in crossing ended, along with a qtl within testing positions range. If qtl plotting data includes two segments $x_r(r = 1,.......,n)$ takes the value of quantitative characteristic along with $z_r(r = 1,........,n)$ measures genetic markers and additional descriptive variables, for example, gender and food practice.

The regression based IM model for backcross design can be expressed by

$$x_r = \mu + bz_{r|s} + e_r \;; s = 1, 2; r = 1, 2, ........, n \tag{2.1}$$

$x_r$ measures phenotypic observation taking r$^{th}$ genotype, $\mu$ is the general mean, $z_r$ is the indicative variable which states the qtl genotype of the genotype r that can presented by following ways:

$$z_r = \begin{cases} 1 & when \;\; qtl \;\; genotype \;\; is\, Gg \\ 0 & when \;\; qtl \;\; genotype \;\; is \; gg \end{cases} \tag{2.2}$$

*b* measures the additive influence given by a qtl, $e_r$ is the residual terms follow $N\,(0,\sigma^2)$

As conditional expectation implies same measurement as conditional probabilities of qtl genotype (s) (Kao, 2000), $z_{r|s}(r = 1,........,n)$ is fixed for qtl genotype(s) provided that the neighboring marker genotypes. When $z_{r|s}(r = 1,........,n)$ is fixed, then the model is termed as fixed effect model.

If we consider loci H, with alleles *H* and *h*, and T with alleles *T* and *t*, represent two neighboring markers within a range in the place an assumed qtl is examined. Assume that unnoticed qtl locus G with alleles *G* and *g* are placed within range lined using markers H and T. The conditional probabilities for qtl genotypes *GG* is denoted by $w_{r|1}$ *and* $w_{r|2}$ for Gg genotype provided that the neighboring marker genotypes be in action. Respective conditional probabilities $w_{r|1}$ and $w_{r|2}$ are presented in **Table 2.1** for backcross population. Likelihood of a paired recombination incident within certain range is disregarded.

**Table 2.1.** Conditional probabilities of assumed qtl genotype with respect to neighboring marker genotypes for a backcross population

| | | *QTL genotypes* | |
|---|---|---|---|
| *Marker genotypes* | *Expected frequencies* | *GG($p_{r1}$)* | *Gg ($p_{r2}$)* |
| *HT/HT* | *(1-$R_{HT}$)/4* | *1* | *0* |
| *HT/Ht* | *$R_{HT}$/2* | *1-w* | *w* |
| *HT/hT* | *$R_{HT}$/4* | *w* | *1-w* |
| *HT/ht* | *(1- $R_{HT}$)/2* | *0* | *1* |

Where $p=R_{HG}/R_{HT}$; $R_{HG}$ is the recombination division between the left marker H and the putative qtl and $R_{HT}$ is the recombination section between two neighboring markers H and T.

In order to examine the presence of a qtl for a particular location for a marker, we intend to verify the statement. Null Hypothesis, $H_A : b = 0$ that is no existence of qtl for certain site. Against, $H_0 : b \neq 0$ that is existence of qtl for specific locus.

Importance of qtl genetic consequence (b) is examined by computing

$$F = \frac{(SST - SSE)/(2-1)}{SSE/(n-2)} \tag{2.3}$$

Where SST=Total sum of squares $=\sum_r (x_r - \tilde{\mu})^2$ ; SSE=Residual sum of squares =

$$\sum_r (x_r - \hat{\mu} - z_{s|r}\hat{b})^2$$

If the corresponding p-value of this F-statistic is significant then it can be concluded that qtl employs a noticeable influence for that particular trait for backcross population.

### 2.2.1.2 Maximum Likelihood Based SIM

If we want to use SIM approach using ML estimator, then, for investigating about presence for qtls for locations between a marker interval, we are interested to examine the hypothesis $H_0 : b = 0$ vs $H_1 : b \neq 0$ by considering the fact that the residuals follow normal distribution as well as the attributable term(*x*) between every qtl genotype has the probability density function (pdf) is as $N(\mu + az_{r|s}, \sigma^2)$

The likelihood function about parameters φ = *(μ, b, σ*$^2$*)* can be written as follows

$$L(\varphi \mid X) = \prod_{s-1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(x_s - \mu - bz_{r|s}\right)^2\right] \qquad (2.4)$$

In order to examine $H_a$ vs $H_1$, the likelihood ratio test (LRT) formula can be stated as follows:

$$LRT = -2log\left[\frac{\sup\limits_{\Phi o} L(\varphi \mid X)}{\sup\limits_{\Phi} L(\varphi \mid X)}\right] = 2\left[log \sup_{\Phi} L(\varphi \mid X) - log \sup_{\Phi} L(\varphi \mid X)\right] \qquad (2.5)$$
$$= 4.608295 * LOD$$

where, $\Phi_0$ and $\Phi$ are the controlled (H$_0$) and uncontrolled (H$_1$) factor spaces. Critical boundary point for rejecting null hypothesis may not be only preferred using chi-square distribution for the reason of avoiding the consistency settings of asymptotic principle in consideration of $H_0$. In order to assess the significance boundary point, the number and length of ranges are taken into accounted because the multiple comparisons testing procedure is carried out in plotting qtls. Hypothesis is studied for each location for each single interval along with for the complete intervals for the genome to generate a continuous LRT value sets. For each single locus, the location constant *w* is preidentified as well as simply μ, b, and $\sigma^2$ are considered for estimating and test procedures. When testing results are significant for chromosomal sites, the locations for biggest LRT statistic are concluded for the estimation of the qtl location as well as the ML estimate(s) for that locus provides the estimates of μ, b and $\sigma^2$ determined following the iterative scheme.

The ML estimation for μ, a, and $\sigma^2$ are determined by:

$$\hat{\mu} = x - bz \;, \hat{b} = \frac{\sum_{r=1}^{n}(z_{r|s} - \overline{z})(x_r - \overline{x})}{\sum_{j=1}^{n}(z_{r|s} - \overline{z})^2} \;\; and \;\; \sigma^2 = \frac{1}{n}\sum_{s=1}^{n}(x_r - \hat{\mu} - \hat{b}z_{r|s}) \qquad (2.6)$$

Where $\overline{x} = \frac{1}{n}\sum_{r=1}^{n} x_r$ and $\overline{z} = \frac{1}{n}\sum_{r-1}^{n} z_{r|s} \;, s=1,2$

## 2.2.2 Composite Interval Mapping (CIM)

Traditional techniques to identify qtls are evaluating with respect to single qtl models against assuming there is absence of qtl such as for SIM approaches the likelihood of a single assumed QTK is measured for every site on the genome. Nevertheless, there may be some interacting factors among the testing qtls places somewhere on that genome. As a result of that, the ability of identifying can be cooperated along with for estimating of those loci and influences of the studied qtls could be favored(Lander and Botstein, 1989). Sometimes the unusual termed as ‘ghost’ qtls  can act(Haley and Knott, 1992; Martínez and Curnow, 1992). In that circumtances, the several different qtls should be plotted precise and exact ways following multiple qtl model(s). An important way to control qtl plotting where multiple qtls take part in determination of an attribute implies that an iteration scheme scan the genome along with the fact the addition of known qtls to the regression equation so that qtls are detected(Lander and Botstein, 1989) which is named as Composite Interval Mapping (CIM). This method is able to ascertain the position and effects magnitude simultaneously for qtl perfectly than single the qtl methods mainly in small mapping populations that is weight of association amongst genotypes for plotting population might be tough (Lander and Botstein, 1989; Li et al., 2007). CIM executes interval mapping by taking a subcategory of marker loci consider as cofactors (Mollah and Eguchi, 2008). That subgroup markers work like replacements for the remaining qtls to uplift the resolution of interval mapping by taking into consideration the situation of associated qtls thus decreasing the error variability(Li et al., 2007).

The following model is suggested to define the approach of calculating CIM for examining a qtl between maker intervals on a genome which is given by:

$$x_r = bz_r * + Z_r \xi + \varepsilon_r \tag{2.7}$$

$$z_r * = \begin{cases} 1/2 & when \;\; qtl \;\; genotype \;\; is\, GG \\ -1/2 & when \;\; qtl \;\; genotype \;\; is\, Gg \end{cases}$$

$x_r$ is the phenotypic value of the $rth$ individual; $Z_r$ , is a subset of $Z_r$ which may contain some preferred markers and rest of the explanatory terms; $\xi$ is the partial regression coefficient vector including the mean $\mu$ and $\varepsilon_r$ is residual term that is $\varepsilon_r \sim N(0,\sigma^2)$ .Merits and idea for considering $Z_r$ in quantitative trait loci analysis are explained in Kao and Zeng (Kao and Zeng, 2006), Zeng (Zeng, 1994, 2006). Generally, it could control for the confounding effect of linked qtls and decrease the residual variance in the analysis.

### 2.2.2.1 QTL Analysis by CIM using ML estimation

Methods for analysis of qtl are: a) probability is determined for a group of parameters (mainly qtl effect and qtl site) with respect to experimental observations on phenotypes and marker genotypes. b) Estimation of parameters is considered when the probability is at peak. c) Critical value could be settled following permutation test procedures.

If we consider a set of values of *n* observations then probability function of $\varphi = (w, b, \xi, \sigma^2)$ given by

$$L(\varphi \mid X, Z) = \coprod_{r=1}^{n} \left[ \sum_{s=1}^{2} w_{rs} \theta\left(\frac{x_r - \mu_{rs}}{\sigma}\right) \right] \tag{2.8}$$

In this equation $\theta(.)$ is a standard normal pdf, $\mu_{r1} = b/2 + Z_r\xi$ and $\mu_{r1} = -b/2 + Z_r\xi$. Density for every observation is termed as the mix of three normal densities with three distinct averages along with mixing rates (Mollah and Eguchi, 2008). Mixing rates $w_{rs}$'$s$ are functions of the qtl position parameter, $w$ are conditional probabilities of qtl genotypes given marker genotypes. Expectation Maximization technique are followed to get ML estimations for probability considering the normal mixture model for inaccurate datasets (Kao and Zeng, 2006). Again in case of CIM to examine the hypothesis $H_N : b = 0$ against $H_A : b \neq 0$, the likelihood ratio test (LRT) statistic is defined by:

$$LRT = -2log\left[\frac{\sup\limits_{\Phi o} L(\varphi \mid X)}{\sup\limits_{\Phi} L(\varphi \mid X)}\right] = 2\left[log\sup\limits_{\Phi} L(\varphi \mid X,Z) - log\sup\limits_{\Phi} L(\varphi \mid X,Z)\right] \tag{2.9}$$
$$= 4.608295 * LOD$$

where, $\Phi_0$ and $\Phi$ are the controlled ($H_0$) and uncontrolled ($H_1$) factor spaces. Critical value to reject the null hypothesis cannot be only selected follwoing $\chi^2$ distribution for the cause of omitting the consistency assumptions of asymptotic statements under $H_N : b = 0$.

In order to assess the significance boundary point, the number and length of ranges are taken into accounted because the multiple comparisons testing procedure is carried out in mappin qtls. Hypothesis is studied for each location for each single interval along with for the complete intervals for the genome to generate a continuous LRT results. For each single locus, the location constant *w* is preidentified as well as simply μ, b, $\xi$ and $\sigma^2$ are considered for estimating and test procedures. When testing results are momentous for chromosomal sites, the locations for biggest LRT statistic are concluded as estimate for qtl locus *w*, and the ML estimators for these sites are the estimates of μ, b, $\xi$ and $\sigma^2$ calculated following Expectation Maximization approach.

For general parametric linkage analysis, typically called “logarithm (10 base)-of-odds” (LOD score), is performed following likelihood (odds) ratio. This ratio measure comparative probability between the probabilities of two alternatives $\frac{L_{H_N}}{L_{H_A}}$, where $L_{H_N}$ implies that there is no linkage against $H_A$ (recombination rate,R=0.5) and $L_{H_A}$ is the probability of alternative hypothesis of linkage (R<0.5) developed is a popular statistical tool now widely used by plant breeders in genetics for qtl mapping (Nyholt, 2000). LOD score however make comparisons the probability to obtain experimental data when two positions seem actually linked, to the possibility of detecting similar dataset by complete chance. Positive scores indicate the existence of linkage and the negative scores imply the less likelihood of presence linkage. Generation of LOD results using computer considers as easy procedure to evaluate or study large and composite family lineages to define association amongst the traits and makers of interest. (https://en.wikipedia.org/wiki/ Genetic_linkage #Parametric_linkage_analysis).

LOD output larger than 3.0 measures the presence of related associations or links since it specifies 1000 to 1 odds implying the linkages are detected were not identified accidentally. A LOD result smaller than -2.0 estimated as excluding that linkage (Nyholt, 2000). LOD output that takes the value 3 convert to a *p*-value nearly 0.05, implies no multiple testing adjustments (e.g., Bonferroni adjustment) are needed(Nyholt, 2000; Risch, 1991).

## 2.3 Results and Discussion

### 2.3.1 Simulation Results

For calculation of the performance of the SIM/IM, HK, eHK and IMP in comparison of the CIM approach for qtl search, a backcross population is taken into account for simulation investigation. In this comparison, it is assumed that only single qtl on a chromosome for 6 alike spread out markers, by taking any two successive marker intermission size is 1 cM. Marker points as well genotypes are created by use of R/qtl open source software (Broman et al., 2003) (http://www.qrtl.org/). Here successive marker interval size 1 is considered. To generate the simulated data for backcross population we consider the number of individuals (nind=30), number of chromosomes (nchr=4) and number of markers (nmar=6).The real measurements of the parameters for SIM model are taken as $b$=0.8, $\mu$=0.2.

Determination of the output results of the CIM approach in comparison of the four methods SIM, HK, eHK is calculated based on LOD score. It is observed from the Fig. 2.2 that for four chromosomes with six markers in each chromosome, the four methods IM, HK, eHK and IMP cannot detect any qtl position by any maker in any position of each chromosome whereas the CIM method identified three qtl positions. One is by the 4th marker in chromosome 2 as well as two positions are detected by the 3rd and 5th markers corresponding to chromosome number 4 whereas the other methods fail to detect any qtls in each chromosome.

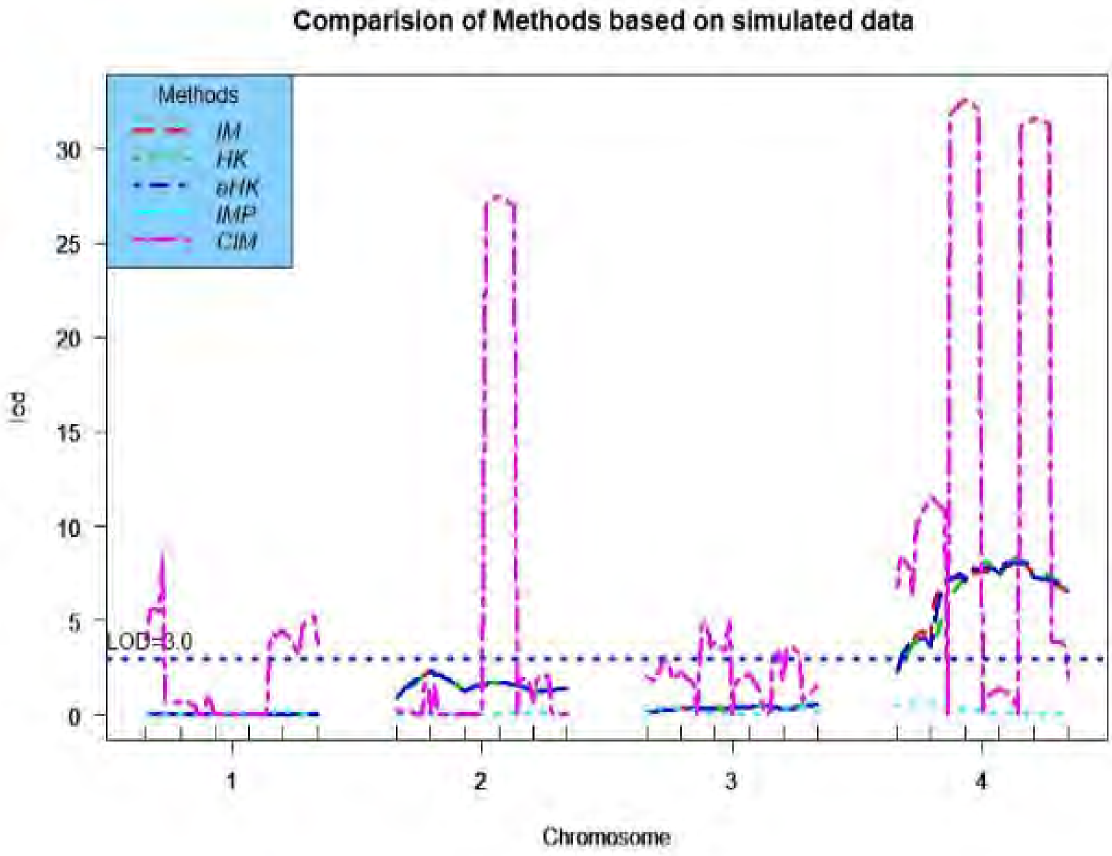


**Fig.2.2.** LOD scores curves for comparison of Interval Mapping (IM), Haley-Knott (HK), Extended Haley-Knott (eHK), Multiple Imputation (IMP), and Composite Interval Mapping (CIM) evaluated based on backcross simulated data.

### 2.3.2 Comparison Analysis Based on Real Data

In order to study the output results CIM in comparison of other four simple interval procedures for qtl analysis in the scenario of real data, we considered a rice mapping population derived from the parent variety of IR64, an irrigated *indica* variety and Azucena, a traditional upland *japonica* variety (Guiderdoni et al., 1992).

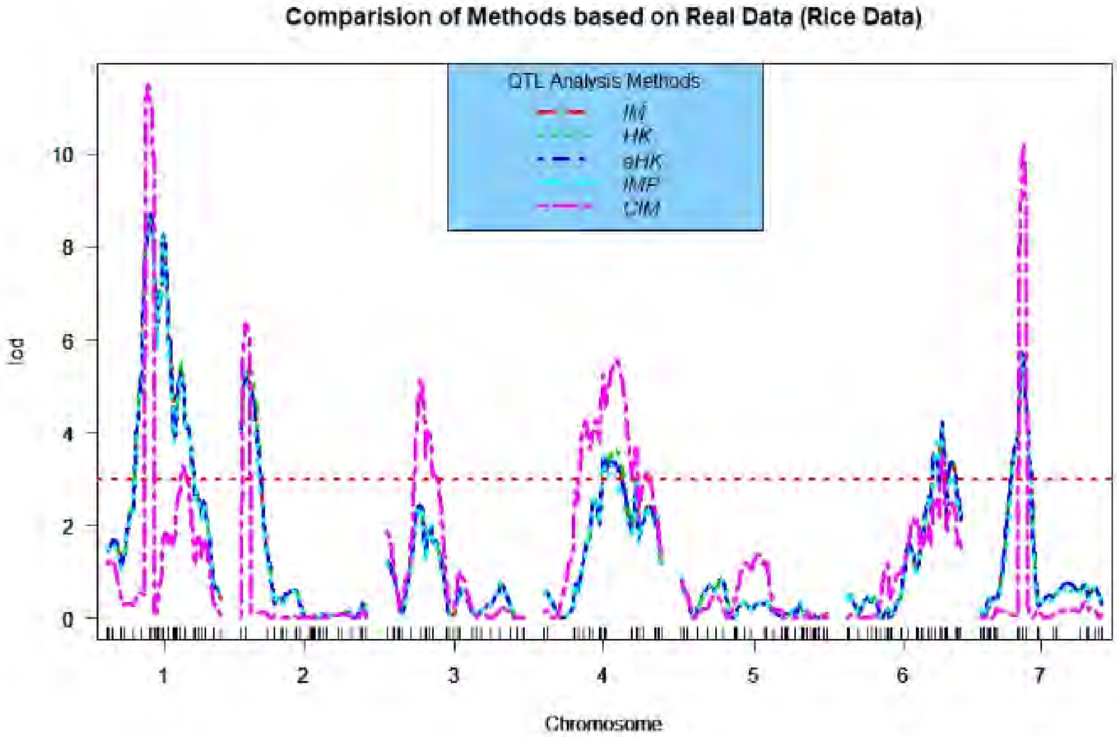


**Fig.2.3.** LOD score curves for comparison of Interval Mapping (IM), Haley-Knott (HK), Extended Haley-Knott (eHK), Multiple Imputation (IMP), and Composite Interval Mapping (CIM) evaluated based on real rice mapping population derived from IR64/Azucena

The dataset used for qtl analysis consisted of molecular marker data of 200 SSR makers from 7 chromosomes. One phenotypic data such as plant height is taken into consideration of backcross population of 200 recombinant inbred lines (RIL) derived from IR64/Azucena (Guiderdoni et al., 1992). It was however observed from the Fig. 2.3 that the qtl mapping tool CIM performs better than the other four methods in detecting qtl positions in real dataset. For each chromosome except the chromosome 5 and 6, CIM method detected qtl positions significantly than the other four interval mapping methods.

## 2.4 Conclusion

The investigation of this comparative study suggests that the Composite Interval Mapping (CIM) performs significantly better than the other four Simple Interval Mapping (SIM) methods in detecting qtl positions in backcross technique both on simulated data and on real rice dataset. CIM detected three makers in chromosome 2 and 4, as well as other four SIM methods were unable in detecting qtls for each of the 4 chromosomes for simulated data. In addition, for a real rice data set from backcross population, the CIM performs mostly in similar fashion for detecting qtls in different positions in each of the 7 chromosomes. CIM were finally able to detect twelve qtls above the LOD threshold 3.0 whereas other SIM methods identified only six marker positions. Although the qtl is a popular technique for biomarker identification in plant breeding but this method cannot be used for human genetics. Due to technological advancement, different omics data on plants and humans are being available for thorough gene expression analysis. In the next chapter different transcriptomic analysis have been demonstrated for identifying and proper classifying of differentially expressed gene biomarkers.

# CHAPTER THREE

## ROBUST STATISTICAL APPROACH FOR GENE EXPRESSION ANALYSIS

# A ROBUST STATISTICAL APPROACH FOR GENE EXPRESSION ANALYSIS AND ITS APPLICATION TO IDENTIFY BIOMARKER GENES INFLUENCING RICE BLAST DISEASE

## 3.1 Introduction

Transcriptomics technology has enabled the expression levels of thousands of genes to be investigated simultaneously. Both identification of differentially expressed (DE) genes and clustering or classifications of genes/samples are equally essential in the transcriptomics (microarray/RNA-Seq) data analysis. However, this OMICS dataset poses different computational challenges because of large number of transcripts surveyed with small sample sizes. The detection of genes that are DE between two or more conditions is essential, since it reduces the dimensionality of the transcripts by a set of DE genes that are most informatics under study. Based on this smaller set of DE genes, clustering, or classification become much easier by any supervised/unsupervised learning algorithm. Each of these tasks is conducted by separate statistical algorithm. There are several types of statistical procedures that are used to identify DE genes including classical parametric approach ($t$-test, $F$-test/ANOVA and likelihood ratio test), non-parametric approach (e.g. KW-test) and empirical Bayes approaches(Do et al., 2005; Efron et al., 2001; Gottardo et al., 2006; Kendziorski et al., 2003; Kruskal and Wallis, 1952; Newton and Kendziorski, 2003; Robinson et al., 2009; Smyth, 2004; Tusher et al., 2001; Wang et al., 2011; Wilcoxon, 1945). However, most of the aforementioned algorithms are not robust against outliers(Gottardo et al., 2006; Mollah et al., 2007, 2012, 2015). Thus, they might be produced misleading results in the presence of outlying observations. Transcriptomics observations (gene expression) are often corrupted by outliers that arise due to several steps involved in the experimental process, from hybridization to image analysis(Gottardo et al., 2006; Mollah et al., 2012). Non-parametric approaches are somewhat robust against outliers in the case of large sample; however, these approaches are sensitive to outliers in the case of small sample sizes. To overcome this problem, the $\beta$-divergence-based empirical Bayes (Beta EB) approach (Mollah et al., 2012) and $\beta$-ANOVA approach (Mollah et al., 2015) was developed for the robust identification of DE genes. These approaches perform well in the presence of outlying observations with up to 50% genes for both small and large sample cases. However, $\beta$-divergence-based approaches require the appropriate selection of the tuning parameter $\beta$ by cross-validation separately for each gene that is much time consuming. Similarly, there are several methods for genes/samples clustering/classification. For the purpose of RNA-Seq/microarray gene expression data classification, various classification methods have been proposed such as Fisher, Bayes, Naive Bayes (NB), Ada-boost, logistic, neural network, k-nearest neighbors (KNN), support vector machines(SVM) (Sain and Vapnik, 2006), bagging support vector machine(bag SVM), random forests(RF) (Breiman, 2001), classification and regression trees (CART) (Breiman et al., 2017) etc.

However, most of these algorithms discussed before are very much sensitive to outlying observations and might be produced misleading results. Generally, there are three ways to obtain robust estimation against outlying observations as follows:

1. Applying the weighted estimators of the model parameters.
2. Applying existing methods on the modified dataset which can be done by modifying/deleting the outlying observations from the dataset
3. Applying the suitable transformation on the dataset and then apply the existing methods.

However, application of robust methods is complicated than using the traditional methods and deletion of outlying observations loses the information of the dataset. Hence, applying the transformation technique is the suitable option for reducing the outlier effects. Several authors (Atkinson, 2018; Box and Cox, 1964) have been proved that transformation based robust methods perform better than the traditional methods in reducing outlier effects. Thus, in this work we consider logistic transformation for reducing outlier effects from the dataset instead of the other transformations.

## 3.2 Materials and Methods

### 3.2.1 Logistic Transformation of Transcriptomics Data for Robust Identification of Biomarker Genes (Proposed)

The transcriptomics dataset is represented by a matrix of dimension $n\times p$ which consists of samples $S=\left\{S_1,S_2,...,S_n\right\}$ and genes $G=\left\{G_1,G_2,...,G_p\right\}$ and we have denoted it by

$D=\left(d_{ij}\right)_{N\times P}$. The logistic transformation on $d_{ij}$ is defined by

$$y_{ij} = f(d_{ij}) = \left(\frac{1}{1+e^{-|d_{ij}|}}\right)\times 10 \qquad (3.1)$$

Obviously, we have $0\leq y_{ij}\leq 10.$ To avoid some other extraneous variations, we then apply the following transformation technique on $y_{ij}$ using $x_{ij}=y_{ij}-\bar{y}_{i.}-y_{.j}+\bar{y}_{..}$

### 3.2.2 Identification DE Genes by SAM using the Proposed

There are several methods for identification of DE genes, among those we considered a popular method named "SAM" (Tusher et al., 2001) to investigate the performance of our proposed logistic transformation. To introduce SAM, let us consider $x_{ij}$ be the *i*th gene expression for the jth samples( i=1,2,…..,$G$ : $j$=1,2,………$n_{k\,;}$ $k$=1,2). We also assume that $\mu_{ij}$ measures the mean of the *i*th gene for *k*th condition. We are interested to test the hypothesis $H_0$: $\mu_{i1}=\mu_{i2}$ versus $H_1$: $\mu_{i1}\neq\mu_{i2}$ that is $H_0$:$\mu_{i1}$-$\mu_{i2}$=0 versus $H_1$ is false. A gene is said to be equally expressed (EE) if $H_0$ is accepted otherwise it is DE. If $\mu_{ik}$ measures the sample mean of *i*th gene for *k*th condition and $s_i^2$ measures the pooled within-class sample variance then the formula of the two sample *t*-test to test the aforesaid null-hypothesis is given as follows:

$$t_i=\frac{d_i}{s_i^2} \qquad (3.2)$$

Where $d_i=m_{i1}-m_{i2}$ (3.3)

and $s_i^2=b\left\{(n_1-1)s_{i1}^2+(n_2-1)s_{i2}^2\right\}$ (3.4)

Here, $b = \frac{\sum_k (1/n_k)}{\sum_k (n_k - 1)}$ $\quad m_{ik} = \hat{\mu}_{ik} = \frac{\sum_j x_{ij}}{n_k}$ 3.5)

$$s_{ik}^{2} = \frac{\sum_j \sum_k \left( x_{ij} - m_{ik} \right)^2}{n_k - 1} \tag{3.6}$$

The test *t*-statistic mentioned in (3.1) follows the *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. This classical *t*-statistic in (3.1) increases the false discovery rate for the case of small-sample cases. Te get sovled this problem, a modified version of this *t*-test statistic proposed by addition a constant $s_0$ to the denominator formula in (3.1) which is called the test statistic of the Significance Analysis of Microarrarys(SAM)(Tusher et al., 2001). The SAM statistic is defined as follows:

$$t_i^{SAM} = \frac{d_i}{s_i^{2} + s_0} \tag{3.7}$$

Where, $s_0$ is the percentile of the distribution of $s_i$. When k>2 conditions, the modified t-test statistic in (3.5) is computed in terms of the Fisher's linear discriminant by considering n samples consists of m non-overlapping subsets such that the response parameter $y_i \in \left\{1, 2, ......., p\right\}$ , $D_k = \left\{ j : y_i = k \right\}$ and $n_k$ is the number of expressions in $D_k$. Then the scores $d_i$ and $s_i^2$ equation (3.6) is substituted by the following two formulas:

$$d_i = \frac{\left[ \left\{ \frac{\sum_{j \in D_k} n_k}{\Pi n_k} \right\} \sum_{k=1}^{p} n_k \left( m_{ik} - m_i \right)^2 \right]^{1/2}}{s_i^{2}} \tag{3.8}$$

$$s_i^{2} = b \left\{ (n_1 - 1) s_{i1}^{2} + (n_2 - 1) s_{i2}^{2} + ........... + (n_k - 1) s_{ik}^{2} \right\} \tag{3.9}$$

where $m_i = \frac{\sum_k n_k m_{ik}}{\sum_k n_k}$ .

To know more description about SAM approach go to https://statweb.stanford.edu/~tibs/SAM/

### 3.2.3 Sample/Gene Group Prediction by SVM using the Proposed Transformed Data

There are several classifiers, among those we considered a popular classifier named "SVM" (Sain and Vapnik, 2006; Zararsiz et al., 2017) to investigate the performance of our proposed logistic transformation. It has attracted great attention because of its strong mathematical background, learning capability, good generalization ability, and wide range of application area such as computational biology, text classification, image segmentation, and cancer classification(Sain and Vapnik, 2006; Zararsiz et al., 2017). SVM is capable of linear/nonlinear classification and deals with high-dimensional data.

Let $x_i$ denotes the training data points, $w$ denotes the weight vector, and $b$ denotes the bias term. The decision function that correctly classifies the data points by their true class labels in a linearly separable space is represented as follows:

$$f(x) = sign(w.x_i + b); i = 1, 2, ..., n \tag{3.9}$$

In a binary classification, the SVM aims to find an optimal separating hyperplane in the feature space, which maximizes the margin and minimizes the misclassification rate by choosing the optimum value of $w$ and $b$. When the cases are not linearly separable, "slack variables" $\{\xi_1, . . ., \xi_n\}$, a penalty term (Sain and Vapnik, 2006; Zararsiz et al., 2017) can be used to allow misclassified data points where $\xi_i > 0$. In most of the classification problems, the separation surface is not linear. In this case, the SVM uses an implicit mapping F of the input vectors to a high-dimensional space defined by a kernel function ($K(x, y)$ = F ($x_i$) F ($x_j$)) and the linear classification then applied in this high-dimensional space. Some of the most widely used kernel functions are linear: $K(x, y) = x_i x_j$, polynomial: $K(x, y) = (x_i x_j + 1)^d$, radial basis function:

$$K (x,y) = exp (-\gamma \| x_i - x_j \|^2)\ etc.$$

The radial basis kernel function is used in the analysis.

### 3.2.4 Performance Evaluation

To evaluate the performance of the proposed logistic transformation in identification of DE genes by SAM approach and classifying samples/genes by SVM classifier, we consider some statistical indices/measures including the receiving operative characteristic (ROC) curve, area under the ROC curve (AUC), partial AUC (pAUC), and misclassification error rate (MER). Outcomes are always divided into four categories for binary classification. The categories are termed as (i) true positive (TP), (ii) false negative (FN), iii) true negative (TN) and (iv) false positive (FP). We then calculate the performance indices/measures using the following confusion matrix as follows:

**Confusion matrix or error matrix**

| | | **Predicted** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Observed** | **Positive** | True positives(nTP) | False negatives(nFN) |
| | **Negative** | False positive (nFP) | True negatives (nTN) |

True positive rate (TPR) or Sensitivity = $\frac{nTP}{nTP+nFN}$

True negative rate (TNR) or Specificity = $\frac{nTN}{nTN+nFP}$

False positive rate (FPR) = $\frac{nFP}{nFP+nTN}$

False negative rate (FNR) = $\frac{nFN}{nFN+nTP}$

False discovery rate (FDR) = $\frac{nFP}{nTP+nFP}$ and

Misclassification error rate (MER) = $\frac{nFP+nFN}{nTP+nTN+nFN+nFP}$

Area under the receiving operating characteristics (ROC) curve, AUC, Power =1-FNR, where, nTP denotes the number of true positive and so on. Each of these performance measures produces values between 0 and 1. A method is said to be best performer if it provides largest values of TPR, TNR, AUC, pAUC, and the smallest values of FPR, FNR, FDR, and MER.

### 3.2.5 Data Source

We compare the performance of the proposed logistic in a comparison of the other transformation methods using both simulated/synthetic and real data analysis

#### 3.2.5.1 Simulated Datasets

To investigate the performance of the proposed logistic transformation by SAM for DE gene identification and SVM for sample/gene classification in presence of outlying observations in the dataset, we simulate transcriptomics dataset using the following data-generating model:

| **Gene Group** | **Sample Group** | | |
|---|---|---|---|
| | **Case** ($P_1$) | **Control** ($P_2$) | |
| **A** | d | d-α | + N(0, $\sigma^2$) |
| **B** | d+α | d | |
| **C** | d | d | |

**Fig.3.1** Schematic diagram for generating synthetic transcriptomic dataset. For RNA-Seq count data, fraction should be omitted from each data point.

#### 3.2.5.2 Data Contamination by Outliers:

Outlier is an observation that deviates from the actual value of the observation and arise in the dataset for some unexpected circumstances. It is a common problem during the analysis of gene expression data. To compare the performance of the proposed approach with SVM in presence of outlying observations in the dataset, we contaminate simulated dataset. Here we contaminated datasets by replacing the some original values(x) of data point by the corresponding outliers (x*), which is defined as $x^* = x \times \kappa \times \text{abs}\,[N(0, 1)]$, where $\kappa \in (50, 100)$. These outliers may arise in the dataset case-wise following the Tukey-Huber contamination model (THCM) (Agostinelli et al., 2015) or independent cell wise following the independent contamination model (ICM) (Alqallaf et al., 2009).

#### 3.2.5.3 Real datasets

##### 3.2.5.3.1 Peanut Species RNA-Seq Dataset (Drought/Control):

We considered two gene expression RNA-seq count dataset (downloaded from https://peanutbase.org/gene_expression/atlas_drought#) (mRNA) selected using fold change(FC) approach ($\log_2$ FC >2 or < -2, FDR <0.05) of two wild peanut species, viz., *Arachis aduranensis* and *Arachis stenosperm* (Brasileiro et al., 2015). To generate the RNA-seq count data, the plants were experimented in two conditions drought-stressed and well-watered (control) and RNA was extracted from leaves and roots. Libraries were sequenced at Fasteris using HI-Seq2000 and the transcripts were mapped into the *A. aduranensis* reference genome. There are 697 up-regulated and 539 down-regulated genes from *A. aduranensis* and 665 up-regulated and 135 down-regulated genes from *A. stenosperm* were considered in comparison.

##### 3.2.5.3.2 Rice Dataset (Control vs Blast Fungus Disease)

The gene expression profile GSE7256 was downloaded from the Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/). GSE7256 dataset contained total 8 samples, including 4 normal and 4 disease (Rice blast fungus) sample with 57381 genes. The summary of the dataset is that, there is two-week old rice plants (cultivar Nipponbare) were treated with either Magnaporthe grisea (virulent isolate FR13) spore suspension in gelatine or gelatine alone. Two time points were taken (3 and 4 days post inoculation- dpi). Disease symptoms were not visible at 3 dpi whereas they were at 4 dpi. Two biological repeats were done. The raw data files we used for the analysis included TXT files.

## 3.3 Results and Discussion

First we will discuss the simulated transcriptomics data analysis results in the sub-section **3.3.1** and then we will discuss the real life transcriptomics data analysis results in the subsection **3.3.2.**

### 3.3.1 Simulation Results

First, we will investigate the performance of our proposed logistic transformation by SAM to detect DE genes in a comparison of other robust transformation (Log, power, rank, & Box-Cox) in the sub-section **3.3.1.1**. Then we will we will investigate the performance of our proposed logistic transformation by SVM classifier for gene prediction in a comparison of the conventional log transformation (Classical) in the sub-section **3.3.1.2.**

#### 3.3.1.1 Performance Investigation by Identification of DE Genes using SAM under the Proposed Logistic Transformation:

We generated simulated transcriptomics datasets using the artificial data generating model as given in **Fig. 3.1** with $d$ = 15, α=5 σ = 1. Each dataset contains $p$ =1000 genes, where 200 DE genes of pattern-A, 200 DE genes of pattern-B and 600 EE genes of pattern-C. Each gene is generated with N =20 sample expressions of which $N_1$=10 expressions are generated from normal condition and $N_2$=10 expressions are generated from the control condition. We replicated this dataset 100 times. Then we contaminated each dataset by 0%, 10%, 15%, and 20% outlier observation.

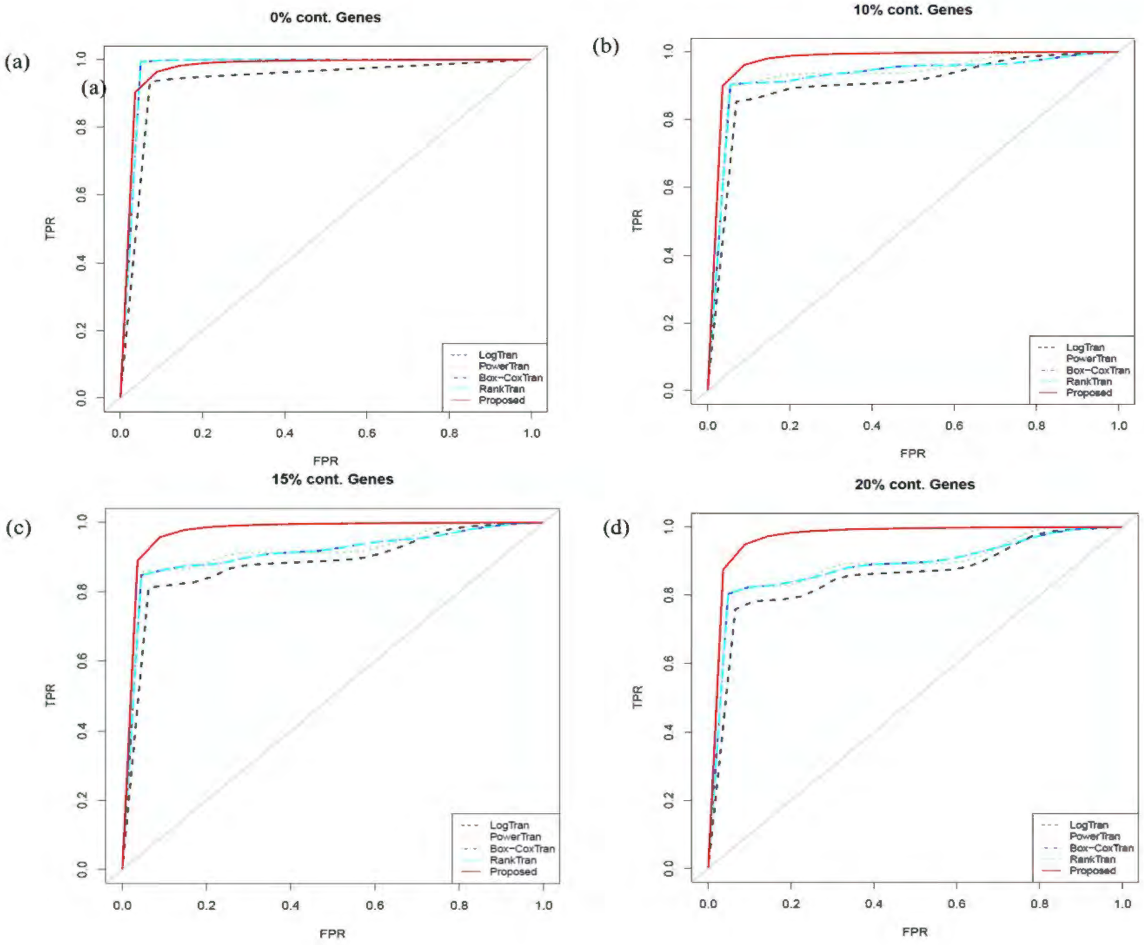


**Fig. 3.2**. Performance evaluation using ROC curve (average TPR vs average FPR) generated by SAM under different transformation. The average TPR and average FPR was calculated based on 100 replications. **(a)** ROC curve without (0%) data contamination **(b)** ROC curve with 10% data contamination, **(c)** ROC curve with 15% data contamination and **(d)** ROC curve with 20% data contamination.

Then we transformed each dataset by Log, Power, Box-Cox, Rank and the proposed logistic transformation approaches to observe the performance of SAM for detection of DE genes in presence of 0%, 10%, 15% and 20% outlier observation, respectively. We computed performance measure AUC, the areas under the ROC curves (**Fig.3.2**) and partial AUC (pAUC) at FPR=0.2 to compare the performance of 5 transformation methods to detect DE genes by SAM in presence of outliers. **Fig 3.2** and **Table 3.1** show that the SAM approach (DE gene identifier) achieves almost equal AUC and pAUC (at FPR=0.2) in absence outliers under all transformation. However, in presence of different levels of outliers, the SAM approach achieves much larger AUC and pAUC (at FPR=0.2) under the proposed logistic transformation only. Thus, the proposed logistically transformed transcriptomics data might be better for identification of DE genes.

**Table 3.1.** Performance evaluation of SAM for Identification of DE genes under 5 transformation including the proposed logistic transformation at four contamination levels (0%, 10%, 15%, and 20%).

| Cont. | Methods | AUC | SE | pAUC | SE |
|---|---|---|---|---|---|
| **0%** | Log_trans | 0.9986 | 0.0005 | 0.1900 | 0.0015 |
| | Power Trans | 0.9886 | 0.0015 | 0.1866 | 0.0019 |
| | Box-Cox Trans | 0.9966 | 0.0030 | 0.1950 | 0.0035 |
| | Rank Trans | 0.9956 | 0.0045 | 0.1958 | 0.0008 |
| | **Proposed** | **0.9897** | **0.00017** | **0.1897** | **0.00012** |
| **10%** | Log-trans | 0.8884 | 0.0240 | 0.1012 | 0.0140 |
| | Power-Trans | 0.9076 | 0.0215 | 0.1366 | 0.0033 |
| | Box-Cox Trans | 0.9206 | 0.0204 | 0.1533 | 0.0022 |
| | Rank-Trans | 0.9066 | 0.0308 | 0.1444 | 0.0215 |
| | **Proposed** | **0.9803** | **0.0020** | **0.1834** | **0.0120** |
| **15%** | Log-trans | 0.8097 | 0.1243 | 0.0803 | 0.0243 |
| | Power-Trans | 0.8586 | 0.0015 | 0.0856 | 0.0015 |
| | Box-Cox Trans | 0.8786 | 0.0012 | 0.1256 | 0.0025 |
| | Rank-Trans | 0.8386 | 0.0013 | 0.1156 | 0.0011 |
| | **Proposed** | **0.9800** | **0.0044** | **0.1880** | **0.0034** |
| **20%** | Log-trans | 0.7759 | 0.1849 | 0.0925 | 0.1849 |
| | Power-Trans | 0.8286 | 0.0005 | 0.0856 | 0.0015 |
| | Box-Cox Trans | 0.8286 | 0.0005 | 0.1156 | 0.0025 |
| | Rank-Trans | 0.8186 | 0.0005 | 0.1056 | 0.0002 |
| | **Proposed** | **0.9780** | **0.0034** | **0.1814** | **0.0036** |

To investigate the performance the proposed transformation in the case of supervised learning, we would like to compare the classification results with the log transformation only in the next sub-section **3.3.1.2,** since log transformation is a common transformation for transcriptomics data analysis.

#### 3.3.1.2 Performance Investigation by Classification of Sample/Genes using SVM under the Proposed Logistic Transformation

We computed the average values of different performance measures such TPR, TNR, FPR, FNR, FDR, AUC, pAUC and MER along with its standard error (SE) based on numerous gene expression count dataset by SVM and proposed method. The average values were calculated using 100 replicated simulated datasets. We computed the average values of the performance indices of AUC and pAUC for the both training and test datasets using SVM and proposed method by introducing 0%, 10%, 15%, and 20% data contamination in each dataset (**Table 3.2**). The AUC and pAUC for the proposed method in the training dataset was 0.9997 and 0.1937 at 10% data contamination whereas the corresponding performance measures for the test dataset were 0.9443 and 0.1524 respectively, which are larger than that of the SVM method (**Table 3.2**). The performance measures AUC and pAUC of the training and test dataset at maximum data contamination rate (20%) were 0.9980, 0.1964 and 0.9093, 0.1366 respectively. The SE measure was getting higher with the increase of data contamination rate for both training and test datasets (**Table 3.2**).

The receiver operating characteristic (ROC) curve was produced for both training and test dataset for the four data contamination levels by taking the average values of TPR and FPR (**Fig. 3.3** and **Fig. 3.4**). The results provided in **Table 3.2** indicate that the proposed method produced higher AUC and pAUC and lower SE for all cases compared to SVM method. The outcomes of the performance indices increased with the progress of percentage of data contamination from 10% to 20% in the training and test dataset. The ROC curves in **Fig.3.3** and **Fig.3.4** however illustrate that the performance of the proposed method was superior in all fashions compared to SVM. **Figure 3.3(a**) only shows that both SVM and proposed method performed nearly in similar manner in presence of 0% data contamination in the training dataset. It is also observed that the proposed method showed very nearly consistent performance in the scenario of training data set with the increase of data contamination although showed higher performance compared to SVM. On the other hand, it is also revealed that the SVM and proposed method showed slightly lower performance in the test data scenario but the proposed method presented higher performance. This performance difference however between SVM and proposed method in test dataset was more distinctive with the increase of data contamination rate and it was highest at 20% data contamination in test data (**Fig.3.4 (d)**). This dataset belongs to data structure $D_1$ and generated by considering # training samples $N = n_1 + n_2;\ n_1 = 40;\ n_2 = 35,$ # test samples $T = t_1 + t_2;\ t_1 = t_2 = 30,$ #Genes $P$=$P_1$+$P_2$ =2×35, #groups α=2, $d$=5, $\sigma$=1, and contamination rate (%) $O_i, i = 0{,}10{,}15{,}20$

**Table 3.2.** Performance evaluation of SVM and proposed method based on simulated gene expression data at four contamination levels (0%, 10%, 15%, and 20%).

| | | **Training set** | | | | **Test set** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Cont.** | **Methods** | **AUC** | **SE** | **pAUC** | **SE** | **AUC** | **SE** | **pAUC** | **SE** |
| **0%** | Classical SVM | 0.9986 | 0.0005 | 0.1956 | 0.0005 | 0.9411 | 0.03097 | 0.1573 | 0.0258 |
| | **Proposed** SVM | **0.9997** | **0.0007** | **0.1937** | **0.0007** | **0.9443** | **0.0309** | **0.1524** | **0.0250** |
| **10%** | Classical SVM | 0.9784 | 0.0140 | 0.1812 | 0.0140 | 0.8463 | 0.0532 | 0.1066 | 0.0250 |
| | **Proposed** SVM | **0.9993** | **0.0020** | **0.1934** | **0.0020** | **0.9379** | **0.0324** | **0.1497** | **0.0243** |
| **15%** | Classical SVM | 0.9197 | 0.1243 | 0.1503 | 0.1243 | 0.7261 | 0.0877 | 0.0654 | 0.0290 |
| | **Proposed** SVM | **0.9989** | **0.0034** | **0.1932** | **0.0034** | **0.9294** | **0.0386** | **0.1441** | **0.0319** |
| **20%** | Classical SVM | 0.9259 | 0.1849 | 0.1625 | 0.1849 | 0.76905 | 0.14281 | 0.0694 | 0.0515 |
| | **Proposed** SVM | **0.9980** | **0.0036** | **0.1964** | **0.0036** | **0.9093** | **0.0387** | **0.1366** | **0.0242** |

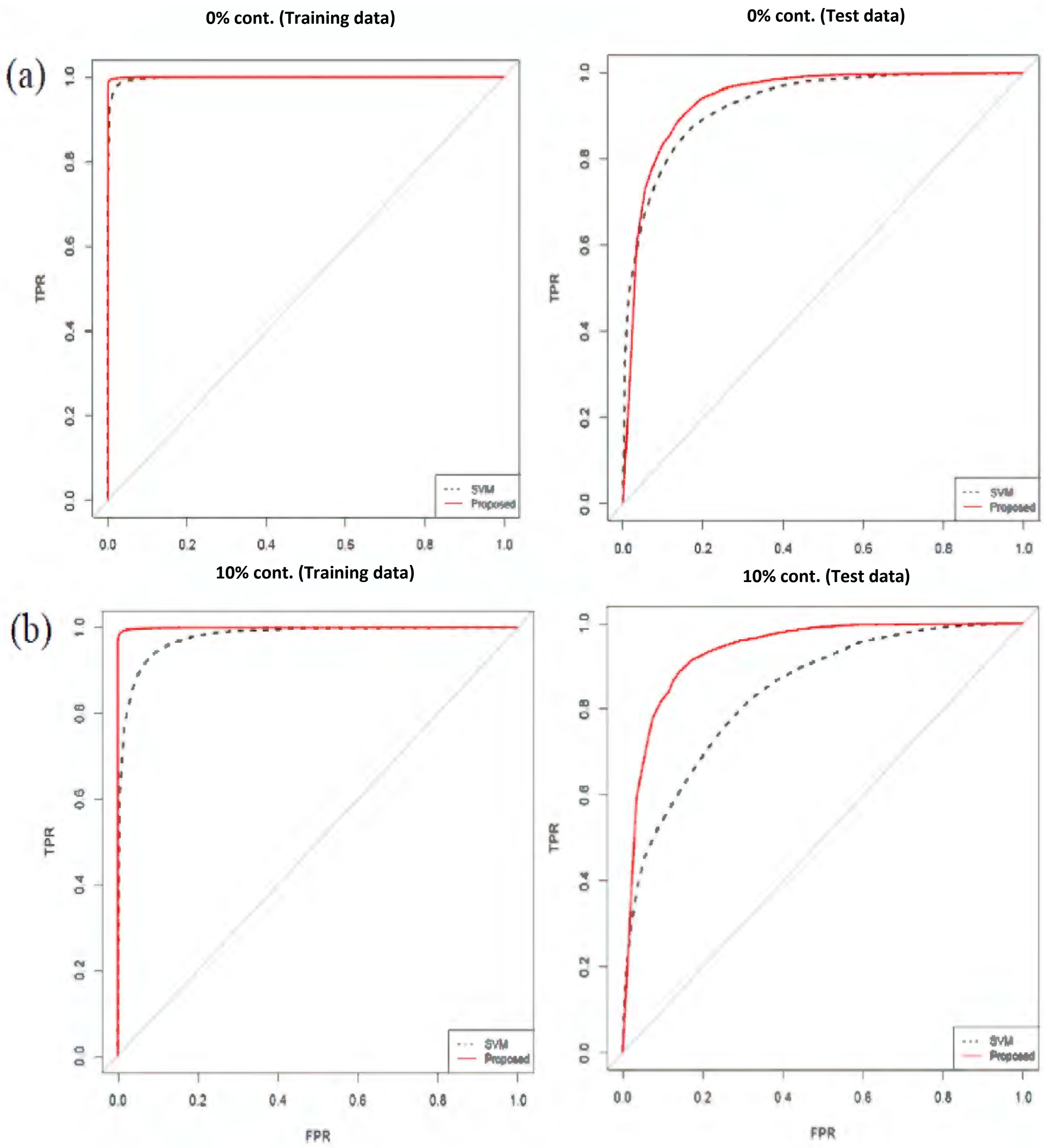
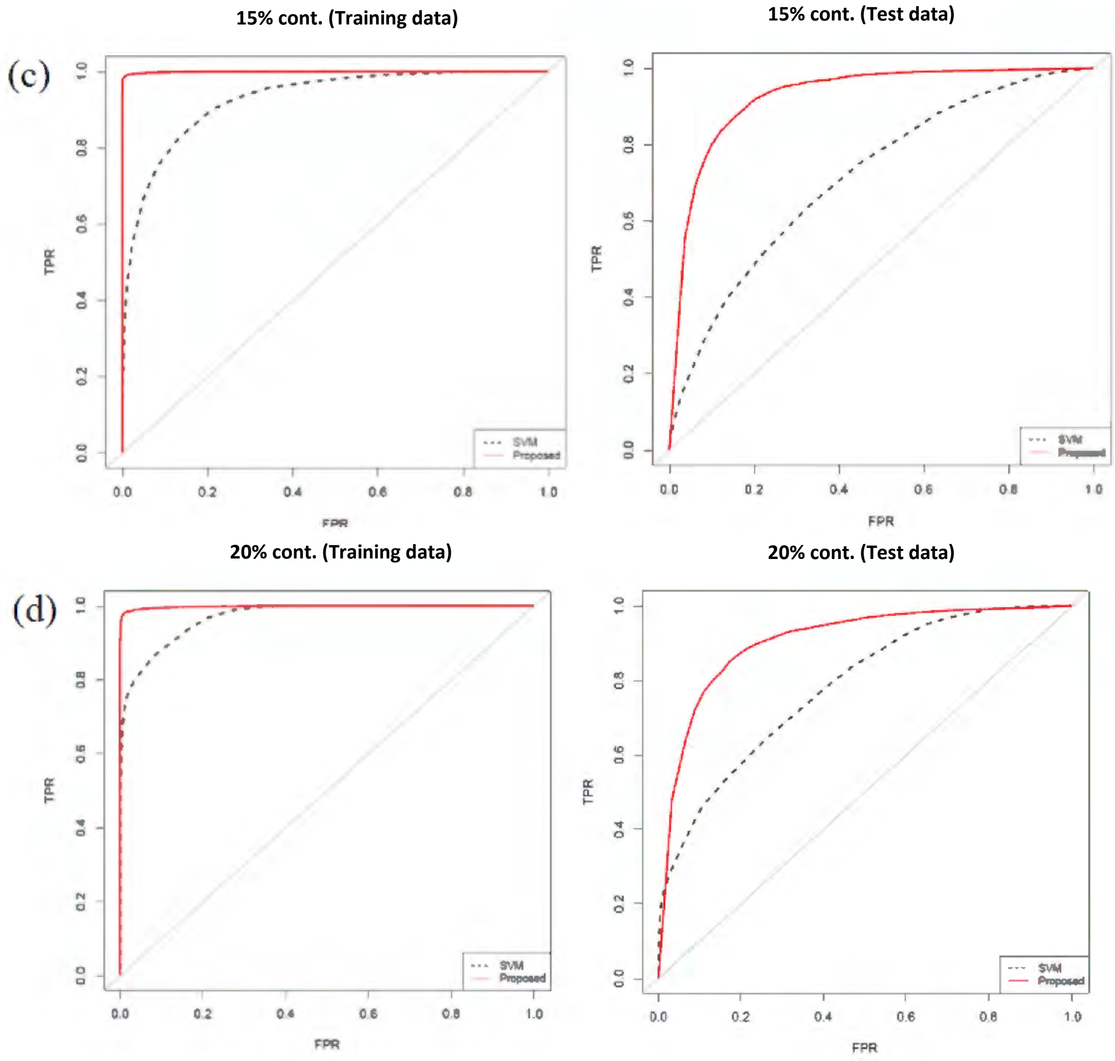


**Fig.3.3**. Performance evaluation using ROC curve (average TPR vs average FPR) generated by SVM and proposed method. **(a)** Represents ROC curve between training and test dataset with 0% data contamination level **(b)** Represents ROC curve between training and test dataset with 10% data contamination level. To obtain the respective results the simulations were carried out 100 times.

We generated $D_2$ dataset under considering different changing parameters of the gene expression model ($d$, $N$, contamination rates (%) and $\sigma$), and computed the performance indices of TPR FPR, TNR, FNR, FDR, for both SVM and proposed method. We considered 100 replications in simulation and estimated their average values to compute each performance indices.


15% cont. (Training data)

15% cont. (Test data)

20% cont. (Training data)

20% cont. (Test data)

**Fig.3.4**. Performance evaluation using ROC curve (average TPR vs average FPR) generated by SVM and proposed method. **(a)** Represents ROC curve between training and test dataset with 15% data contamination level **(b)** Represents ROC curve between training and test dataset with 20% data contamination level. To obtain the respective results the simulations were carried out 1000 times.

**Table 3.3.** Performance measure using average results of 100 times simulated data when $d = d_i; (i = 1, 2, 3, 4, 5)$ in presence of 20% outliers

| **Measures** | **Methods** | ***d*****$_1$= 1** | ***d*****$_2$= 2** | ***d*****$_3$= 3** | ***d*****$_4$= 4** | ***d*****$_5$= 5** |
|---|---|---|---|---|---|---|
| **TPR** | Classical | 0.5230 | 0.5782 | 0.6205 | 0.6279 | 0.6225 |
| | SVM | (0.0626) | (0.0743) | (0.0785) | (0.0802) | (0.0769) |
| | **Proposed** | **0.5623** | **0.7216** | **0.8438** | **0.8822** | **0.8857** |
| | **SVM** | **(0.0549)** | **(0.0745)** | **(0.0710)** | **(0.0650)** | **(0.0630)** |
| **FPR** | Classical | 0.4554 | 0.3804 | 0.3276 | 0.3256 | 0.3303 |
| | SVM | (0.1375) | (0.1081) | (0.1078) | (0.1050) | (0.1050) |
| | **Proposed** | **0.3486** | **0.1988** | **0.1111** | **0.0820** | **0.0742** |
| | **SVM** | **(0.1337)** | **(0.0849)** | **(0.0665)** | **(0.0568)** | **(0.0568)** |
| **TNR** | Classical | 0.5446 | 0.6196 | 0.6724 | 0.6744 | 0.6697 |
| | SVM | (0.1375) | (0.1081) | (0.1078) | (0.105) | (0.1050) |
| | **Proposed** | **0.6514** | **0.8012** | **0.8889** | **0.9180** | **0.9258** |
| | **SVM** | **(0.1337)** | **(0.0849)** | **(0.0665)** | **(0.0568)** | **(0.0568)** |
| **FNR** | Classical | 0.4770 | 0.4218 | 0.3795 | 0.3721 | 0.3775 |
| | SVM | (0.0626) | (0.0743) | (0.0785) | (0.0802) | (0.0769) |
| | **Proposed** | **0.4377** | **0.2784** | **0.1562** | **0.1178** | **0.1143** |
| | **SVM** | **(0.0549)** | **(0.0745)** | **(0.0710)** | **(0.0651)** | **(0.0632)** |
| **FDR** | Classical | 0.3138 | 0.3006 | 0.2793 | 0.2836 | 0.2856 |
| | SVM | (0.1405) | (0.1252) | (0.1301) | (0.1280) | (0.1321) |
| | **Proposed** | **0.2075** | **0.1713** | **0.1077** | **0.0814** | **0.0736** |
| | **SVM** | **(0.1059)** | **(0.0851)** | **(0.0700)** | **(0.0612)** | **(0.0599)** |

A method is said to be perform superior, when it provides higher TPR, TNR and lower FPR, FNR, FDR and MER.

**Case-I:** In this data structure we consider # training samples $N = n_1 + n_2$; $n_1 = 40$; $n_2 = 35$, #test samples $T = t_1 + t_2$; $t_1 = t_2 = 20$, #Genes $P$=2×*35*, #groups α=*2*, *σ=1*, and contamination rate $O = 20\%$. It is observed from **Table 3.3** that the proposed method showed higher performance in terms of generating larger values of TPR, TNR and lower FPR,FNR and FDR. The corresponding standard errors (SE) are arranged in opening parenthesis.

**Case-II:** In this data structure we consider #training samples $N \in n_i + n_j; i = j = 40, 45, 50, 55, 60$; #test samples $T \in t_i + t_j; i = j = 40, 45, 50, 55, 60$; #Genes $P$=2×*35*, #groups α=*2*, $d = 5$; *σ =1*, and contamination rate $O = 20\%$. **Table 3.4** presents the performance indices of SVM and proposed method in terms of TPR, FPR, TNR, FNR, and FDR. The values in the parenthesis measure the corresponding SE. It is however comprehensible from the simulation results of the performance measures mentioned in **Table 3.4** that the proposed method showed higher value of TPR and TNR whereas lower values of FPR, FNR and FDR at different levels of $N = n_{im}$. TPR rate of the proposed method got higher with the improvement of $N = n_{im}$values and this rate was larger compared to SVM method. SVM and proposed method showed TPR 0.6229, 0.8398 when $N = n_{im} = 40$ and TPR values were 0.7611, 0.8535 at the highest level of $N = n_{im} = 60$ respectively. FDR rate also decreased with the increase of $N = n_{im}$values even though the FDRs were much lower for the proposed method than that of the SVM. The FDR was 0.3841 and 0.1091 for SVM and proposed method at the lowest level $N = n_{im}$=40 respectively. For the highest level of $N = n_{im}$=60, the FDR of the SVM and proposed method were 0.2480 and 0.1144.

**Case-III:** Performance indices of SVM and proposed method were also calculated in terms of varying degree of data contamination rates $O_i$ (**Table 3.5**). In this data structure we considered #training samples $N = n_1 + n_2$; $n_1 = 40$; $n_2 = 35$, #test samples $T = t_1 + t_2$; $t_1 = t_2 = 20$, #Genes $P$=2×*35*, #groups *α=2*, $d = 2$; *σ=1*, and contamination rate (%) $O_i, i = 0, 10, 15, 20$ TPR of the proposed method was higher in comparison of SVM. Simulation investigation provided TPR 0.8912 and 0.6252 at 0% data contamination level for the proposed and SVM method respectively whereas these rates were 0.8860 and 0.6177 when data contamination rate was imposed maximum 20% in the simulation study (**Table 3.5**).

FDR rates were also much lower of the proposed method in comparison of SVM at varying level of data contamination. The proposed method however produced the FDR index 0.0749 whereas the corresponding value of the SVM was 0.285 at lowest level of data contamination (0%) in this investigation (**Table 3.5**).

**Table 3.4.** Performance measure using average results of 100 times simulated data when $N = n_{im}; (i = 1, 2, 3, 4, 5; m = 1, 2)$

| Measures | Methods | $n_{1m}$= 40 | $n_{2m}$= 45 | $n_{3m}$= 50 | $n_{4m}$= 55 | $n_{5m}$= 60 |
|---|---|---|---|---|---|---|
| **TPR** | Classical | 0.6229 | 0.6398 | 0.6807 | 0.7158 | 0.7611 |
| | SVM | (0.0848) | (0.1012) | (0.1132) | (0.1212) | (0.1172) |
| | **Proposed** | **0.8398** | **0.8417** | **0.8470** | **0.8501** | **0.8535** |
| | **SVM** | **(0.0510)** | **(0.0477)** | **(0.0448)** | **(0.0419)** | **(0.0403)** |
| **FPR** | Classical | 0.3775 | 0.3623 | 0.3197 | 0.2827 | 0.2363 |
| | SVM | (0.0865) | (0.0982) | (0.1142) | (0.1222) | (0.1181) |
| | **Proposed** | **0.1142** | **0.1122** | **0.1173** | **0.116** | **0.1173** |
| | **SVM** | **(0.0481)** | **(0.0475)** | **(0.0449)** | **(0.0411)** | **(0.0397)** |
| **TNR** | Classical | 0.6225 | 0.6377 | 0.6803 | 0.7173 | 0.7637 |
| | SVM | (0.0865) | (0.0982) | (0.1142) | (0.1222) | (0.1181) |
| | **Proposed** | **0.8858** | **0.8878** | **0.8827** | **0.884** | **0.8827** |
| | **SVM** | **(0.0481)** | **(0.0475)** | **(0.0449)** | **(0.0411)** | **(0.0397)** |
| **FNR** | Classical | 0.3771 | 0.3602 | 0.3193 | 0.2842 | 0.2389 |
| | SVM | (0.0848) | (0.1012) | (0.1132) | (0.1212) | (0.1172) |
| | **Proposed** | **0.1602** | **0.1583** | **0.1530** | **0.1499** | **0.1465** |
| | **SVM** | **(0.051)** | **(0.0477)** | **(0.0448)** | **(0.0419)** | **(0.0403)** |
| **FDR** | SVM | 0.3841 | 0.3705 | 0.3298 | 0.2926 | 0.2480 |
| | | (0.1420) | (0.1491) | (0.1535) | (0.1569) | (0.1493) |
| | **Proposed** | **0.1091** | **0.1072** | **0.1137** | **0.1125** | **0.1144** |
| | **SVM** | **(0.052)** | **(0.0508)** | **(0.0495)** | **(0.0454)** | **(0.0444)** |

**Table 3.5.** Performance measure using average results of 100 times simulated data when contaminations rate (%) are $o_i$; $(i = 1, 2, 3, 4, 5)$

| **Measures** | **Methods** | **$o_1$= 0** | **$o_2$= 10** | **$o_3$= 15** | **$o_5$= 20** |
|---|---|---|---|---|---|
| **TPR** | **Classical** | 0.6252 | 0.6196 | 0.6256 | 0.6177 |
| | **SVM** | (0.0805) | (0.0802) | (0.0777) | (0.0796) |
| | **Proposed** | **0.8912** | **0.889** | **0.8891** | **0.8860** |
| | **SVM** | **(0.0636)** | **(0.0651)** | **(0.0647)** | **(0.0651)** |
| **FPR** | **Classical** | 0.3247 | 0.3283 | 0.3252 | 0.3321 |
| | **SVM** | (0.1122) | (0.1144) | (0.1092) | (0.1142) |
| | **Proposed** | **0.0748** | **0.0718** | **0.0724** | **0.0733** |
| | **SVM** | **(0.0573)** | **(0.0544)** | **(0.0558)** | **(0.0528)** |
| **TNR** | **Classical** | 0.6753 | 0.6717 | 0.6748 | 0.6679 |
| | **SVM** | (0.1122) | (0.1144) | (0.1092) | (0.1142) |
| | **Proposed** | **0.9252** | **0.9282** | **0.9276** | **0.9267** |
| | **SVM** | **(0.0573)** | **(0.0544)** | **(0.0558)** | **(0.0528)** |
| **FNR** | **Classical** | 0.3748 | 0.3804 | 0.3744 | 0.3823 |
| | **SVM** | (0.0805) | (0.0802) | (0.0777) | (0.0796) |
| | **Proposed** | **0.1088** | **0.1110** | **0.1109** | **0.1140** |
| | **SVM** | **(0.0636)** | **(0.0651)** | **(0.0647)** | **(0.0651)** |
| **FDR** | **Classical** | 0.285 | 0.2795 | 0.2826 | 0.2864 |
| | **SVM** | (0.1409) | (0.1345) | (0.1336) | (0.1388) |
| | **Proposed** | **0.0749** | **0.071** | **0.0719** | **0.072** |
| | **SVM** | **(0.0615)** | **(0.0574)** | **(0.0587)** | **(0.0553)** |

The FDR for the proposed and classical SVM method were 0.072 and 0.2864 when the maximum data contamination rate (20%) was introduced in this simulation study. We also obtained higher TNR for the proposed method compared to SVM at different stages of data contamination (**Table 3.5**). The values in the parenthesis measure the corresponding standard error (SE) of each indices.

**Case-IV:** We also investigated the performance of the proposed and SVM method using similar performance measures against different level of σ values (**Table 3.6**). In this data structure we considered #training samples $N = n_1 + n_2$; $n_1 = 40$; $n_2 = 35$, #test samples$T = t_1 + t_2$; $t_1 = t_2 = 20$, $T = t_1 + t_2$; $t_1 = t_2 = 20$, #Genes $P$=2×35, #groups α=$2d = 2$; $\sigma \in \sigma_i$; $i = 1, 1.25, 1.50, 1.75, 2.00$; and contamination rate (%) $O = 20$.

This evaluation criterion also revealed the higher performance of the proposed method compared to SVM in terms of higher TPR and TNR as well as lower values of FPR, FNR, and FDR. The proposed method showed TPR (0.8901) and TNR (0.9272) as well as TPR (0.6103) and TNR (0.702) respectively when σ=1 and σ=2 were considered which were higher than that of the corresponding values of the indices of SVM. In contrast, the FDR values of the proposed method at different levels of σ were also lower in comparison of FDRs of SVM (**Table 3.6**). The values in the parenthesis provide the corresponding values of SE.

**Fig.3.5** and **Fig.3.6 shows** the graphical presentation of the misclassification error rate (MER) of training and test dataset in terms of different levels of *d*, *N*, data contamination rate (%), and σ for the proposed and SVM method. We generated these datasets by following data structure $D_{3.}$ It is observed from the **Fig.3.5(a)** that the MER was comparatively lower for the proposed method than SVM for both training and test dataset at varying level of d. The MER declined significantly with the increase of d values for the test data set but the MER of the proposed method was still lower than that of the SVM. The MERs were consistent for both training and test dataset regardless of the level of *d* values (**Fig. 3.5(b)**).

For the **Fig.3.5(a)** and **Fig.3.5(b)** we consider the datasets 100 times with the following parameters: # training samples $N = n_1 + n_2$; $n_1 = 40$; $n_2 = 35,$#test samples $T = t_1 + t_2$; $t_1 = t_2 = 20,$#Genes $P$=2×*35*, #groups α=*2*, $d \in d_i; i$ =0.10, 0.31, 0.52, 0.73, 0.94, 1.15, 1.36, 1.57, 1.78, 2.00; $\sigma$ =*1*, contamination rate $O = 20\%$ and #training samples $N \in n_i + n_j; i = j = 40, 43, 46, 49, 52, 55, 58, 61, 64, 67, 70$; #test samples $T \in t_i + t_j; i = j = 40, 43, 46, 49, 52, 55, 58, 61, 64, 67, 70$; $T \in t_i + t_j; i = jT \in t_i + t_j; i = j = 40, 43, 46,$ $49, 52, 55, 58, 61, 64, 67, 70$; #Genes $P$=2×*35*, #groups *m=2*, $d = 5$; $\sigma$ =*1*, and contamination rate $O = 20\%$, respectively. The x-axis index of the plot indicates the corresponding values of *d* and *N*.

**Table 3.6.** Performance measure using average results of 100 times simulated data when $\sigma = \sigma_i; (i = 1, 2, 3, 4, 5)$

| Measures | Methods | $\sigma_1$= 1 | $\sigma_2$= 1.25 | $\sigma_3$= 1.50 | $\sigma_4$= 1.75 | $\sigma_5$= 2 |
|---|---|---|---|---|---|---|
| **TPR** | **Classical** | 0.6238 | 0.5898 | 0.5710 | 0.5475 | 0.5355 |
| | **SVM** | (0.0803) | (0.0737) | (0.0684) | (0.0686) | (0.0648) |
| | **Proposed SVM** | **0.8901** | **0.8101** | **0.7320** | **0.6647** | **0.6103** |
| | | **(0.063)** | **(0.0703)** | **(0.0701)** | **(0.0716)** | **(0.0656)** |
| **FPR** | **Classical** | 0.3287 | 0.3646 | 0.3878 | 0.4078 | 0.4350 |
| | **SVM** | (0.1107) | (0.1116) | (0.114) | (0.1134) | (0.1281) |
| | **Proposed SVM** | **0.0728** | **0.1275** | **0.1840** | **0.2448** | **0.2980** |
| | | **(0.0547)** | **(0.0683)** | **(0.0857)** | **(0.0983)** | **(0.1120)** |
| **TNR** | **Classical** | 0.6713 | 0.6354 | 0.6122 | 0.58122 | 0.5650 |
| | **SVM** | (0.1107) | (0.1116) | (0.114) | (0.1174) | (0.1281) |
| | **Proposed SVM** | **0.9272** | **0.8725** | **0.816** | **0.7552** | **0.702** |
| | | **(0.0547)** | **(0.0683)** | **(0.0857)** | **(0.0983)** | **(0.112)** |
| **FNR** | **Classical** | 0.3762 | 0.4102 | 0.4290 | 0.4525 | 0.4645 |
| | **SVM** | (0.0803) | (0.0737) | (0.0684) | (0.0686) | (0.0648) |
| | **Proposed SVM** | **0.1099** | **0.1899** | **0.2680** | **0.3353** | **0.3897** |
| | | **(0.0630)** | **(0.0703)** | **(0.0701)** | **(0.0716)** | **(0.0656)** |
| **FDR** | **Classical** | 0.2888 | 0.3000 | 0.3012 | 0.3054 | 0.3163 |
| | **SVM** | (0.1383) | (0.1344) | (0.1303) | (0.1384) | (0.1394) |
| | **Proposed SVM** | **0.0722** | **0.1198** | **0.1624** | **0.1964** | **0.2150** |
| | | **(0.0573)** | **(0.0724)** | **(0.0905)** | **(0.0969)** | **(0.1043)** |

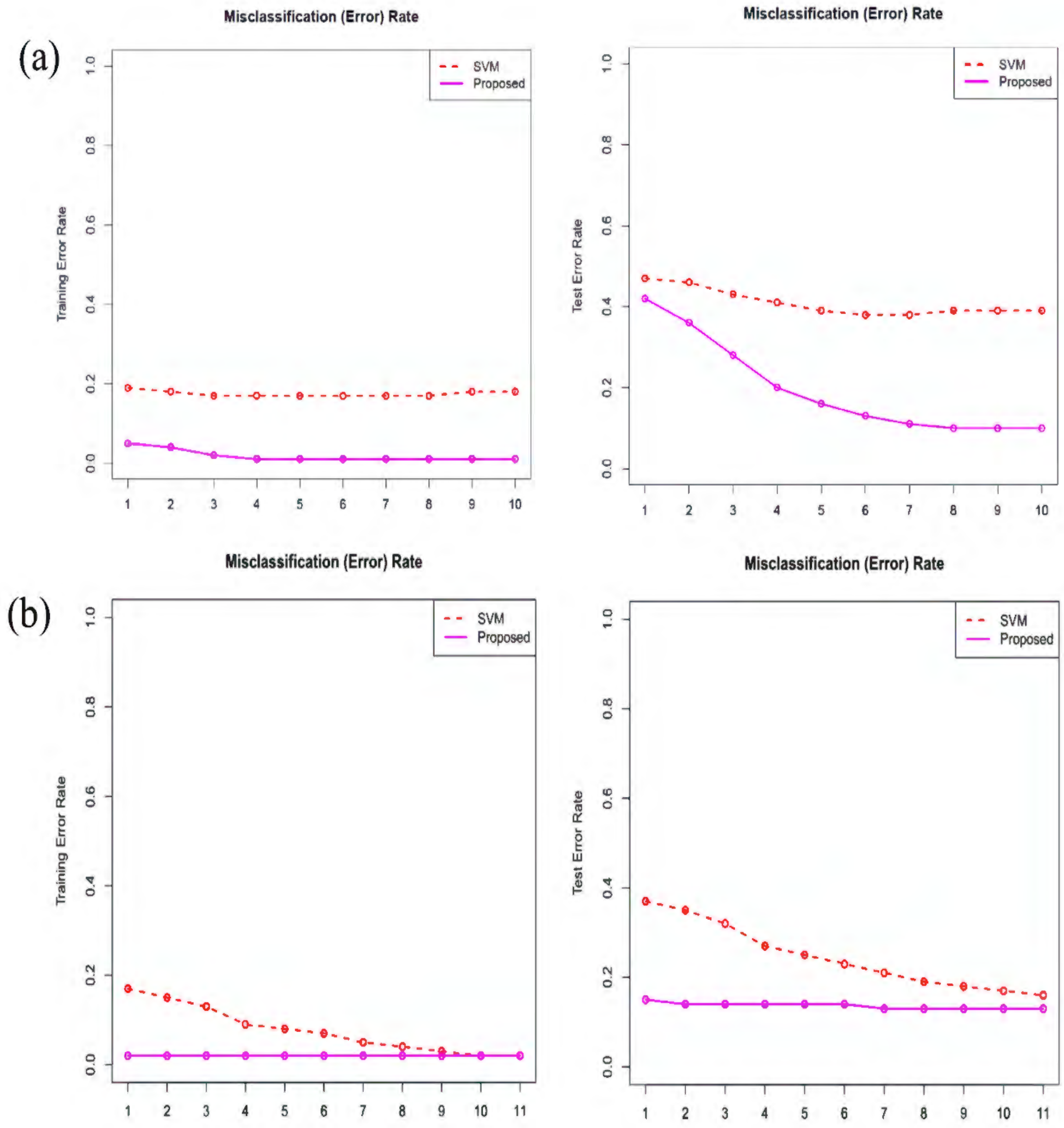


**Fig.3.5.** Misclassification error rate (MER) for training and test dataset. (a) *d* changing and (b) *N* changing for SVM and proposed method

The **Fig.3.6(a)** and **Fig.3.6(b)** also display the MER of SVM and proposed method for the training and test dataset at different data contamination and σ level respectively. It is observed from the **Fig. 3.6(a)** that the MER was lower for the proposed method compared to that of classical SVM while the rate slightly remained in upward trend for test dataset. **Fig. 3.6(b)** also demonstrates that the MER was lower for the proposed method for training and test dataset at varying values of σ though the rate improved with the increase of σ values and the trend was higher in test dataset than that of training dataset. However, in course of classification of test datasets, the MER of our proposed Proposed SVM.was much smaller than that of the classical SVM.

For the **Fig.3.6(a)** and **Fig.3.6(b)** we consider the datasets 1000 times with the following parameters: #training samples $N = n_1 + n_2;\ n_1 = 40;\ n_2 = 35,$ #test samples $T = t_1 + t_2;\ t_1 = t_2 = 20,$ #Genes $P$=*35*, #groups α=*2*, $d = 2$; $\sigma$ =*1*, contamination rate $O \in O_i; i = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50.$ and #training samples $N = n_1 + n_2;\ n_1 = 40;\ n_2 = 35,$ $N = n_1 + n_2;\ n_1 = 40;\ n_2 = 35,$ #test samples $T = t_1 + t_2;\ t_1 = t_2 = 20,$ #Genes $P$=2×*35*, #groups α=*2* ; $\sigma \in \sigma_i; i$ =1.0, 1.07, 1.14, 1.21, 1.28, 1.35, 1.42, 1.50,1.57, 1.64, 1.71, 1.78, 1.85, 1.92, 2.00; $d = 2$ and contamination rate (%) $O = 20.$ The x-axis index of the plot indicates the corresponding values of *d* and *N.*

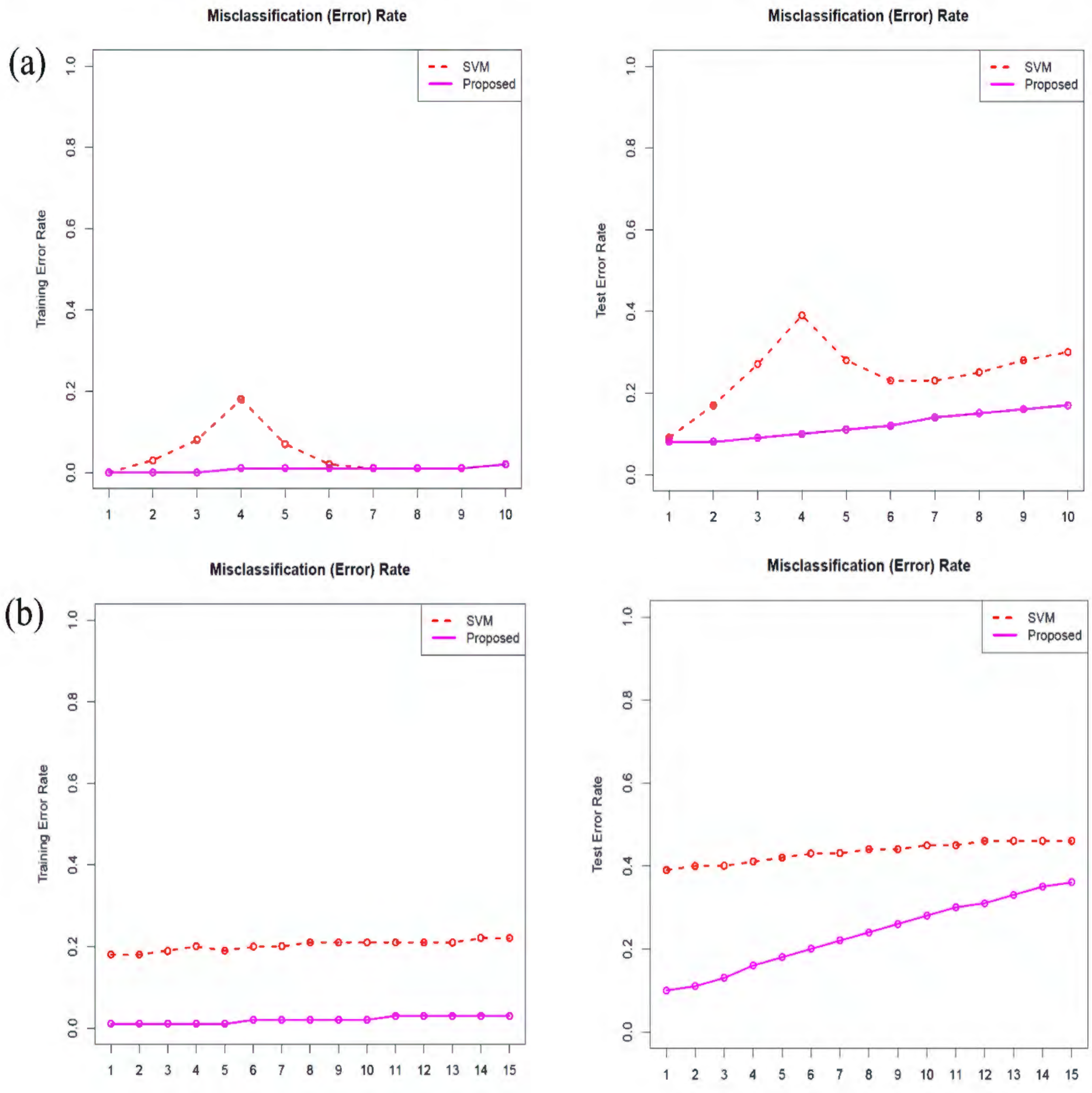


**Fig.3.6.** Misclassification error rate (MER) for training and test dataset. **(a)** Changing contamination rate and **(b)** $\sigma$ changing for classical SVM and proposed SVM method.

In this study, we also checked the performance of SVM and proposed proposed SVM when datasets were not contaminated. In this case, both methods showed equal performances (**Fig. 3.7**).

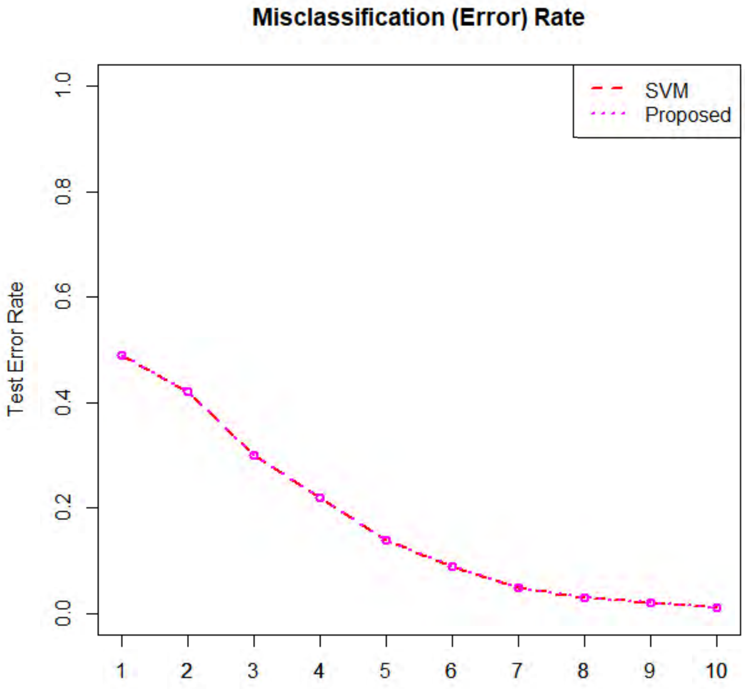


**Fig.3.7.** Performance of classical SVM and proposed SVM when datasets are free of outliers.

The data structure $D_4$ was generated by considering #training samples $N = n_1 + n_2$; $n_1 = n_2 = 20,$ #test samples $T = t_1 + t_2$; $t_1 = t_2 = 20,$ #Genes $P$=*15*, #groups α=*2*, $d$=*5*, $\sigma$ =*0.3*, and contamination rate (%) $O = 20$. The rows of the data matrix indicate the # samples (classes/types) termed as I∈ $\{IA_i, IB_i\}; i = 1,2, \dots ,20$ and columns indicate the # Genes named as G∈ $\{a_i, b_i\}; i = 1,2, \dots ,15$. We visualized and analyzed the datasets in **Fig.3.8** and **Fig.3.9.** Heat maps showed in **Fig.8 (a)** and **Fig.3.8 (b)** demonstrates that the original simulated gene-expression count data structure for training and test data, respectively. Since in real world datasets are not structured accordingly very often, we allocated test dataset and added 20% outliers in both test and training datasets (**Fig. 3.8(c-d)** and **Fig. 3.9(a)**). After applying our proposed approach proposed SVM and classical SVM method (for samples classification) (Sain and Vapnik, 2006), we obtained more than 99% correct classification and the data structure has recovered by our proposed approach (**Fig. 3.9(c)** ) whereas classical SVM failed to recover the correct structure of the datasets when it was contaminated by the outliers (**Fig. 3.9(b)**). It indicates that our proposed SVM approach perform well when the dataset is contaminated by the outliers.

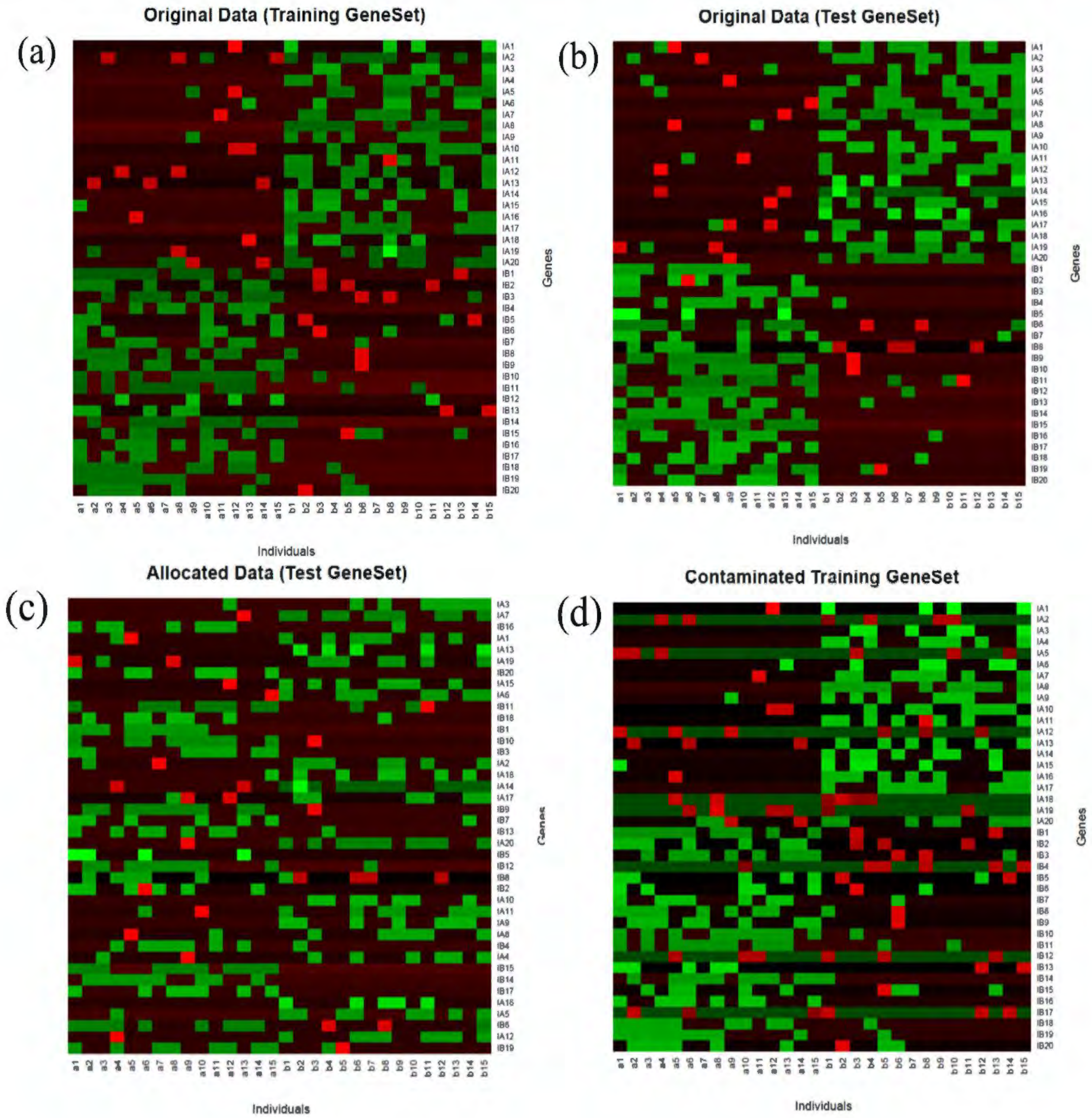


**Fig.3.8**. Classification of the gene-expression data for training and test dataset. **(a)** Illustrates that the classification of the original training gene set. **(b)** Represents the classification of the original test gene set. **(c)** Represents the gene-expression of the allocated test gene set. **(d)** Displays the classification of the training gene-expression RNA-seq count data with 20% data contamination/outliers.

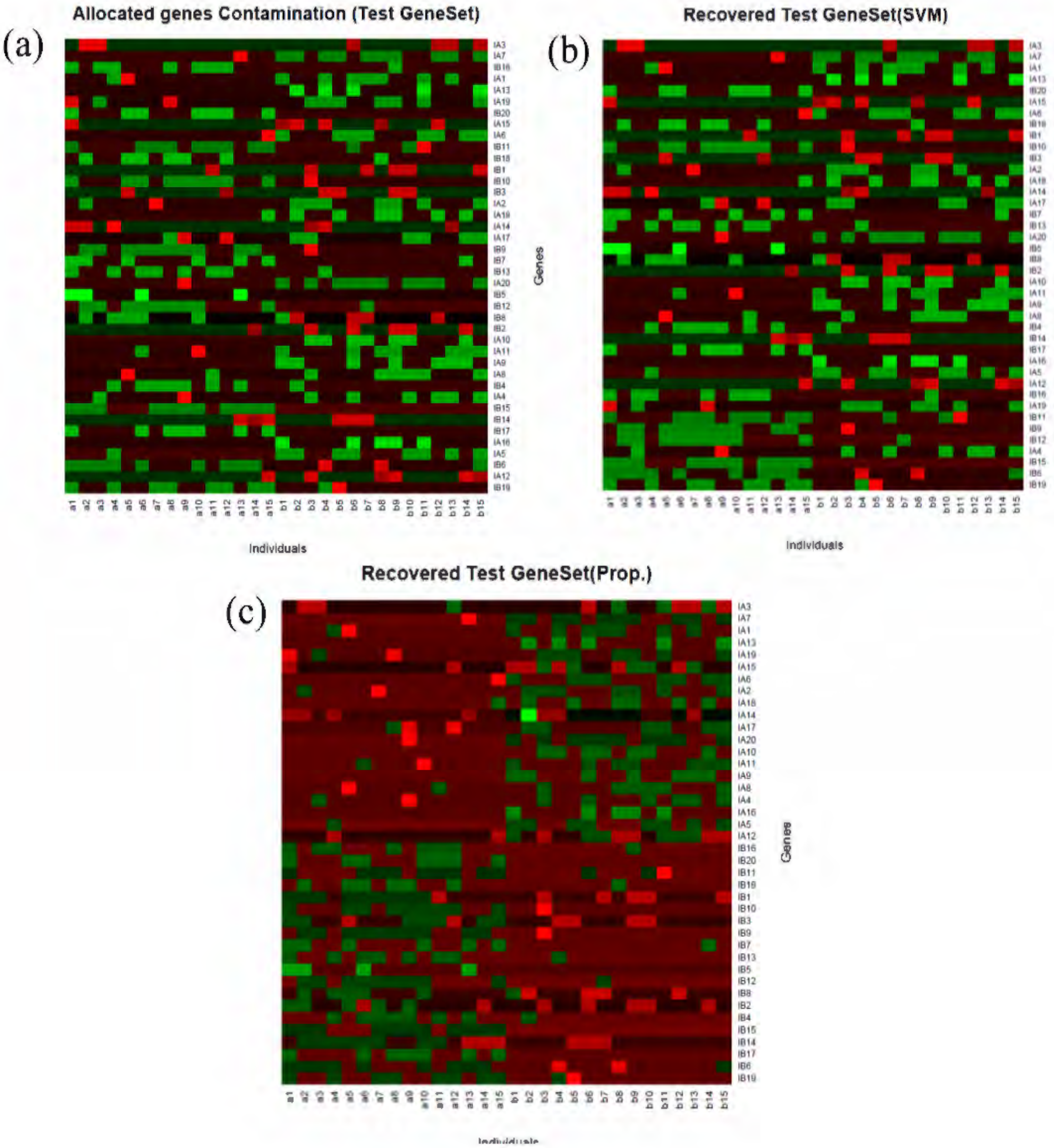


**Fig.3.9.** Classification of the gene-expression data for training and test dataset. **(a)** Represents the classification of the allocated test gene-expression RNA-seq count data with data 20% contamination/outliers. **(b)** Represents the recovered expression by SVM **(c)** Represents the recovered expression structure of the test gene set by proposed SVM.

Now from the simulation investigation it is observed from **Fig 3.3** to **Fig.3.6** and Table **3.2 to Table 3.6** that the classical SVM classifier achieves almost equal values for TPR, TNR, AUC and pAUC (at FPR=0.2) in absence outliers under both traditional log and the proposed logistic transformation. However, in presence of different levels of outliers, the proposed SVM approach achieves much larger values for TPR, TNR, AUC and pAUC (at FPR=0.2) and the smaller values for FNR, FDR and MER under the proposed logistic transformation only. Thus, the proposed logistically transformed transcriptomics data might be better for classifications of genes/samples.

### 3.3.2 Real Transcriptomics Data Analysis

#### 3.3.2.1 Performance Investigation by Identification of DE Genes from the Peanut RNA-Seq Dataset (Drought vs Control) using SAM under the Proposed Logistic Transformation

Differential expression was performed on the training data using the edgeR method. The differentially expressed genes selected in the training data are also selected in the test datasets using 5-fold cross-validation techniques. We selected top 200 up-regulated genes for both dataset and 200 downregulated for *Arachis duranensis* and 135 for *Arachis stenosperm*. We processed dataset by removing zero values of count and made training and test dataset for the classification problem based on the processed up-regulated and downregulated genes label. The performance measurements were also investigated for real RNA-seq count peanut dataset using the statistical indices AUC, pAUC, and SE for both training and test data sets. First, we apply our methodology on the original dataset and then we introduce 10% outliers in the dataset for the measurement of proposed method performance.

**Table 3.7.** Performance measure in absence of outliers for *A. aduranensis*. The values in the parenthesis are the corresponding standard error (SE).

| | Training | | Test | |
|---|---|---|---|---|
| | **AUC** | **pAUC** | **AUC** | **pAUC** |
| **Classical** | 1.0000 | 0.0800 | 0.9916 | 0.1600 |
| **SVM** | (0.0000) | (0.0000) | (0.0186) | (0.0894) |
| **Proposed** | 1.0000 | 0.0000 | 0.9971 | 0.1200 |
| **SVM** | (0.000) | (0.0000) | (0.0063) | (0.1095) |

**Table 3.8.** Performance measure in absence of outliers for *A. stenosperm*. The values in the parenthesis are the corresponding standard error (SE).

| | Training | | Test | |
|---|---|---|---|---|
| | **AUC** | **pAUC** | **AUC** | **pAUC** |
| **Classical** | 0.9122 | 0.1266 | 0.8527 | 0.0133 |
| **SVM** | (0.1510) | (0.1510) | (0.2189) | (0.0421) |
| **Proposed** | 1.0000 | 0.0600 | 0.9452 | 0.0658 |
| **SVM** | (0.000) | (0.0000) | (0.0846) | (0.0866) |

**Table 3.9.** Performance measure in presence of outliers for *A. aduranensis* using. The values in the parenthesis are the corresponding standard error (SE).

| | Training | | Test | |
|---|---|---|---|---|
| | **AUC** | **pAUC** | **AUC** | **pAUC** |
| **Classical** | 0.8252 | 0.0943 | 0.7108 | 0.0000 |
| **SVM** | (0.2071) | (0.2071) | (0.2114) | (0.0000) |
| **Proposed** | 1.0000 | 0.1400 | 0.9742 | 0.0200 |
| **SVM** | (0.0000) | (0.0000) | (0.0660) | (0.0632) |

**Table 3.10** Performance measure in presence of outliers for *A. stenosperma* using. The values in the parenthesis are the corresponding standard error (SE).

| | Training | | Test | |
|---|---|---|---|---|
| | **AUC** | **pAUC** | **AUC** | **pAUC** |
| **Classical SVM** | 0.9122<br>(0.1510) | 0.1266<br>(0.1510) | 0.8527<br>(0.2189) | 0.0133<br>(0.0421) |
| **Proposed SVM** | 1.0000<br>(0.0000) | 0.0600<br>(0.0000) | 0.9452<br>(0.0846) | 0.0658<br>(0.0866) |

We computed the average values of the performance indices of AUC and pAUC for the both training and test datasets using classical SVM and proposed SVM method with and without outliers for training and test data for *A. aduranensis* (**Table 3.7** and **Table 3.8** ) and *A stenosperm* (**Table 3.9** and **Table 3.10**) respectively.

The AUC and pAUC for the proposed method in the training dataset was the same whereas the performance measure AUC and pAUC for the test dataset for *A. aduranensis* were 0.9916 and 0.1600 for SVM and 0.9971 and 0.1200 for the proposed method in absence of outliers The corresponding SE for the proposed method was also lower for the proposed method (**Table 3.7)**. It was also observed from **Table 3.8** that the AUC and pAUC possessed larger and lower SE values for both training and test data sets for *A stenosperm* without outliers. In presence of outliers, our proposed method also showed higher values of AUC and pAUC along with lower SE values (**Table 3.9** and **Table 3.10)** for both training and test dataset.

The receiver operating characteristic (ROC) curves were also produced to check the performance for both training and test dataset for the *A. aduranensis* and *A. stenosperma* with and without outliers (**Fig. 3.10(a)** and **Fig. 3.10(b)**) by taking the average values of TPR and FPR. In absence of data contamination, the SVM and proposed method showed similar performance for both training and test dataset for both *A. aduranensis* and *A. stenosperma*. On the other hand, when we introduce outliers at different levels in the real peanut gene expression data, our proposed method exhibited higher performance for both training and test dataset for *A. aduranensis* and *A. stenosperma* (**Fig.3.11 (a)** and **Fig3.11 (b)**). It is however observed from the ROC **Fig.3.11(a)** and **Fig.3.11(b)** that the performance was more significant for proposed SVM for test data set for the both species compared to training dataset.

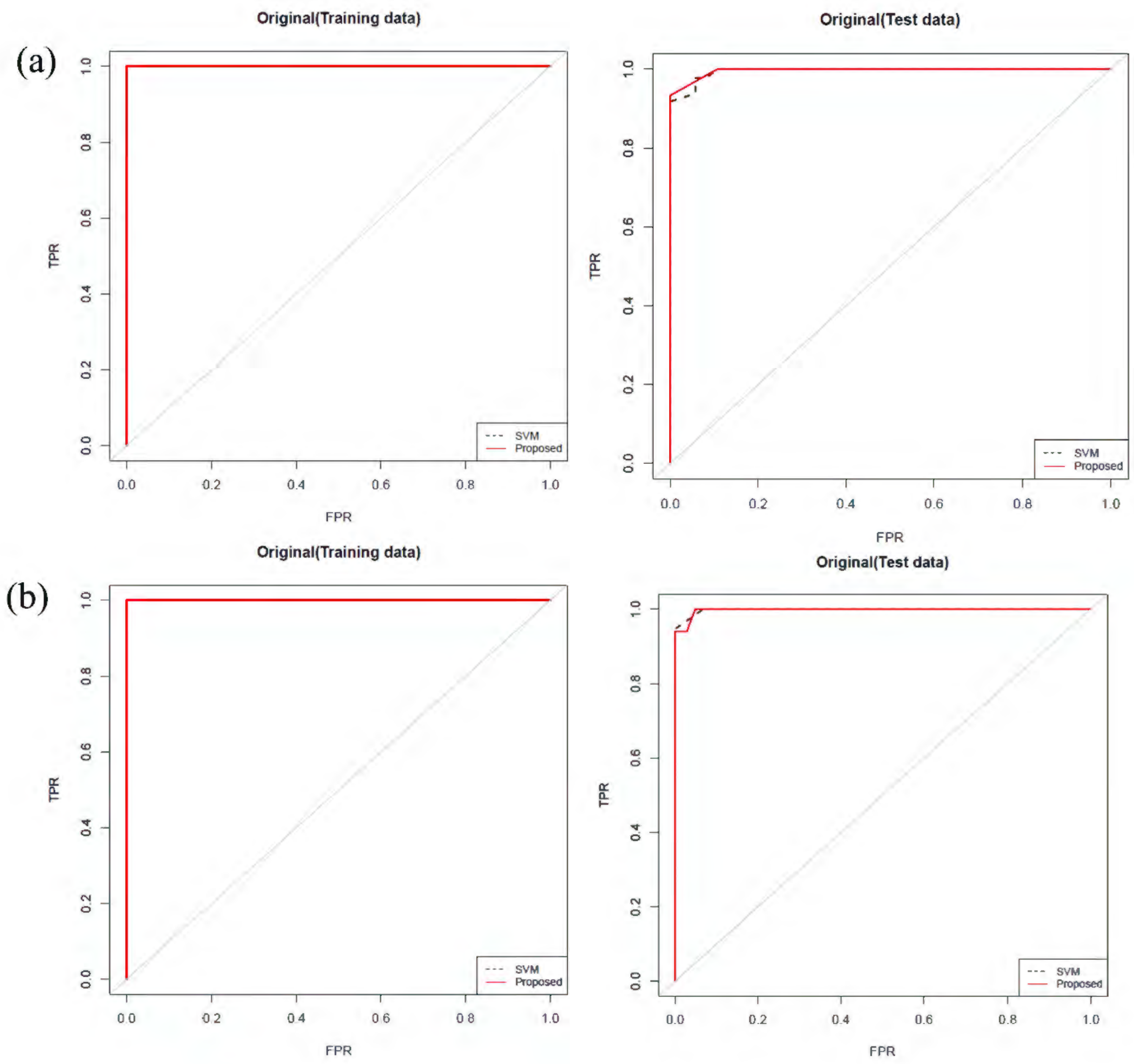


**Fig. 3.10.** Performance evaluation using ROC curve (average TPR vs average FPR) generated by SVM and proposed SVM method. **(a)** Represents ROC curve between training and test dataset without outliers for *A. aduranensis* dataset. **(b)** Represents ROC curve between training and test dataset without data contamination for *A. stenosperma* dataset.

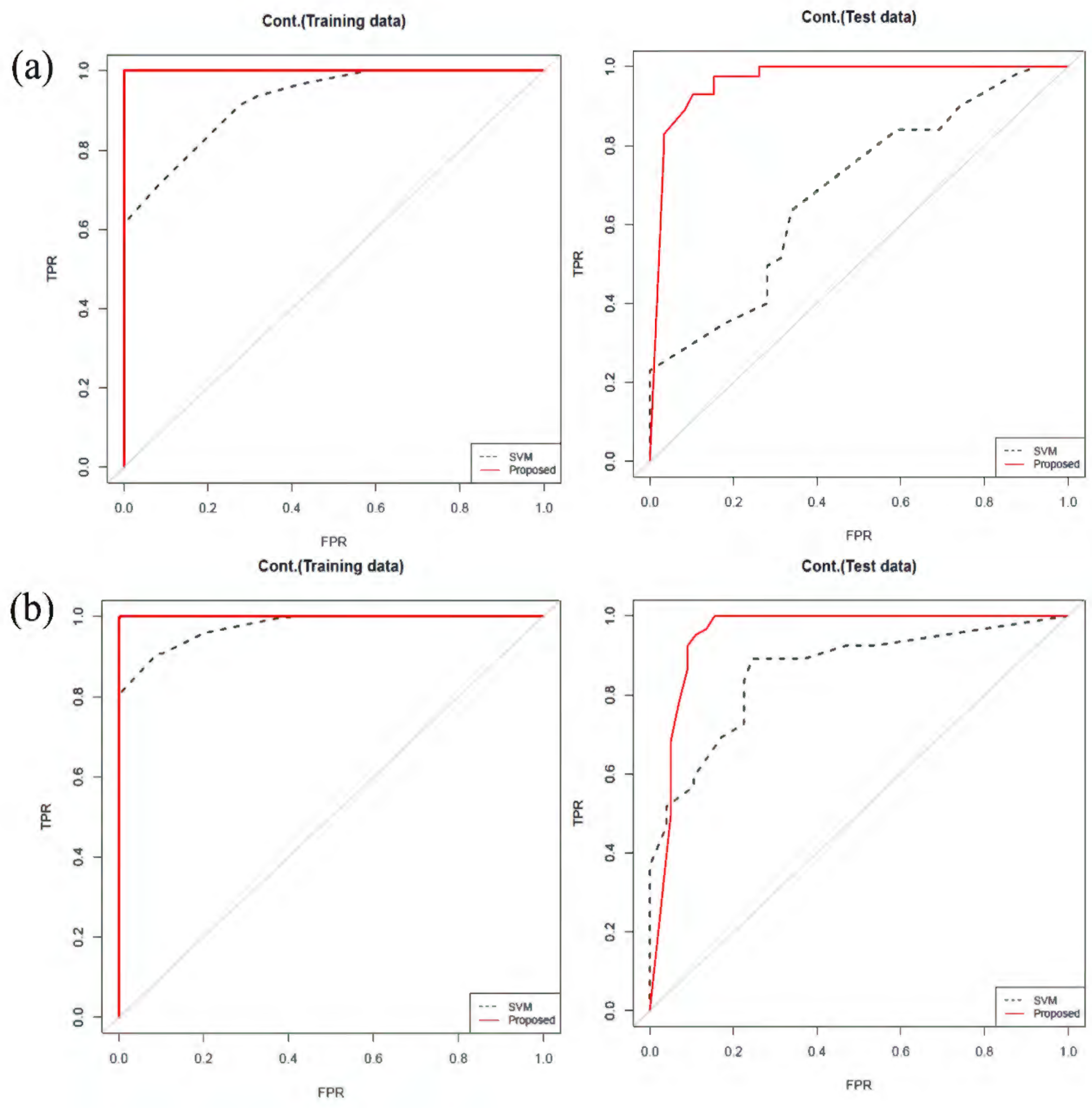


**Fig. 3.11.** Performance evaluation using ROC curve (average TPR vs average FPR) generated by SVM and proposed method. **(a)** Represents ROC curve between training and test dataset with outliers for *A. aduranensis* dataset **(b)** Represents ROC curve between training and test dataset with data contamination for *A. stenosperma* dataset.

From real data analysis and from the **Fig.3.10** and **Fig.3.11** as well as from **Table 3.7** to **Table 3.10** we observe that the classical SVM classifier achieves almost similar results for TPR, TNR, AUC and pAUC (at FPR=0.2) in absence outliers under both traditional log and the proposed logistic transformation for gene classification. However, in presence of different levels of outliers, the proposed SVM approach achieves much larger values for TPR, TNR, AUC and pAUC (at FPR=0.2) and the smaller values for FNR, FDR and MER under the proposed logistic transformation only. Thus, the proposed logistically transformed transcriptomics data might be better for classifications of genes.

### 3.3.2.2 Performance Investigation by Classification of Sample/Genes from the "Rice RNA-Seq Dataset (control vs rice blast fungus)"using SAM under the proposed logistic transformation

Differential expression was performed on the training data using the ***samr*** R package and we applied classical SAM and our proposed method. The P value < 0.05 and log fold change (FC) > 1.5 or log FC < - 1.5 were used as the cut-off criteria. The DE genes with statistical significance between the treatment and control samples were selected and identified. We identified total 6957 DE genes using the classical SAM among those top up-regulated and top down-regulated genes are 200 and 522 respectively. On the other hand, we identified total 4818 DE genes using our proposed method of which 109 and 603 are the top up- and down-regulated genes. The log2FC>0 and log2FC<0 with p-value <0.01 were used for selecting the up- and down-regulated genes for both classical SAM and proposed SAM.

#### 3.3.2.2.1 GO enrichment Analysis

The Gene Ontology (GO) enrichment analysis of the DE genes identified using our proposed SVM method shows that a number of biological molecular and cellular functional pathways are involved in response to rice blast fungus (*Magnaporthe grisea).* Among those the regulation of jasmonic acid mediated signaling pathway (GO:1903507; p-value: 0.00419 ), secondary metabolite biosynthetic process pathway (GO:0044550; p-value: 0.00835), negative regulation of nucleic acid-templated transcription (GO:1903507;

p-value: 0.00187), response to stress(GO:0006950; p-value: 0.00078), defense response (GO:0006952; p-value: 0.00032), protein kinase activity (GO:0004672; p-value: 0.000565), response to wounding(GO:0009611,p value=0.011276994) plasma membrane (GO:0005886, p-value:0.004993245) are statistically significant pathways related to the fungal disease in rice(**Fig.3.12** and **Table A3.3**). The GO analysis also revealed that most of the gene products are allocated to the cell membrane. It clears the relation to the rice blast fungus that mainly affects the cell membrane. The KEGG pathway (**Fig.3.12**) also shows the hub genes that are related to the plant-pathogen interaction pathway.

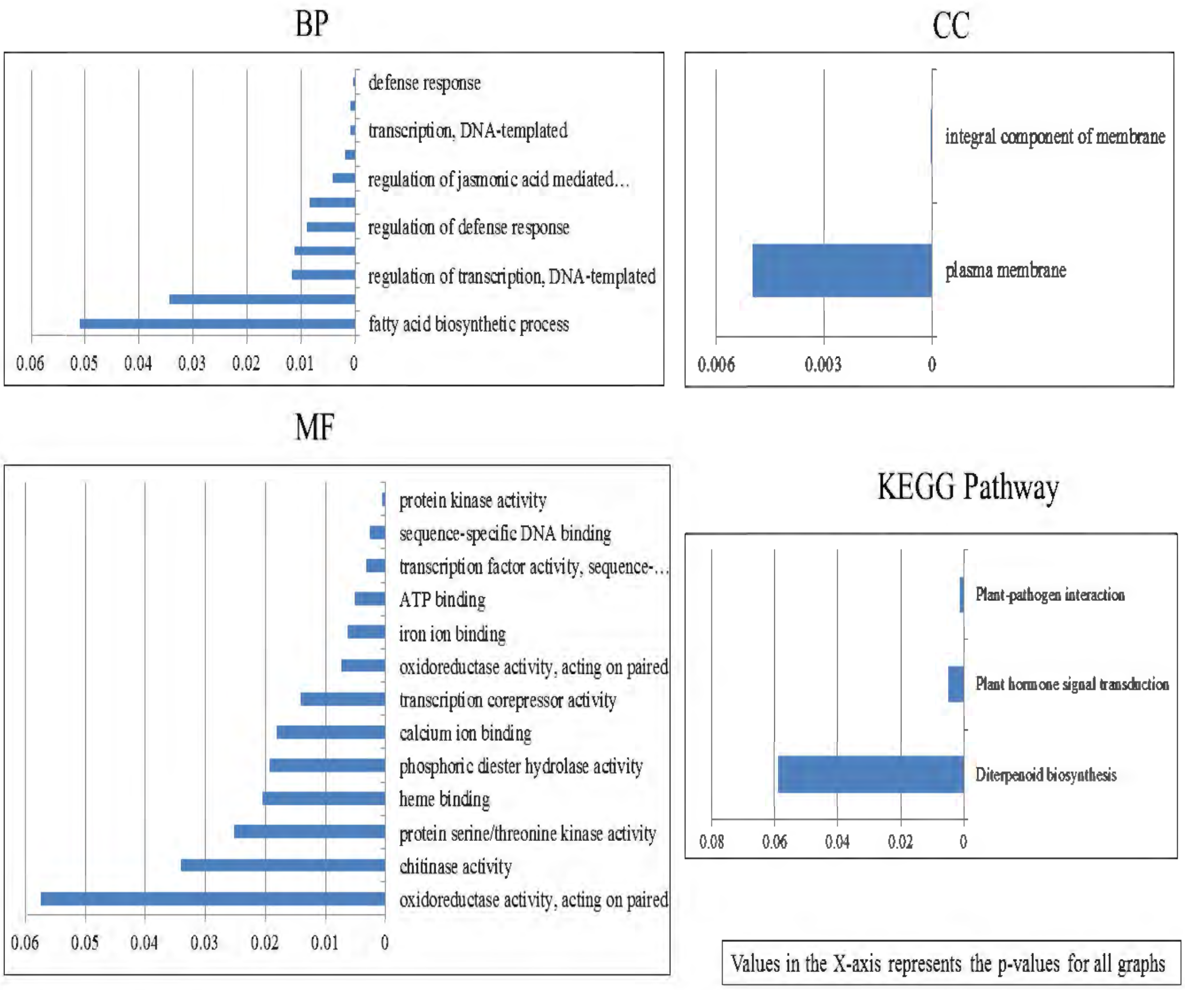


**Fig. 3.12.** GO analysis and KEGG pathways for proposed method genes

On the other hand, the GO enrichment analysis of the DE genes from the classical method also exposed transcription, DNA-templated (GO:0006351; p-value: 1.17E-07), negative regulation of nucleic acid-templated transcription (GO:1903507; p-value: 3.10E-07), regulation of defense response (GO:0031347; p-value: 6.23E-07), response to fungus (GO:0009620; p-value: 0.0035), transcription factor activity, sequence-specific DNA binding (GO:0003700; p-value: 6.86E-06) as significant pathways in rice **(Fig. 3.13 and Table A3.1).** These also provide evidence for the reported genes in presence of *Magnaporthe grisea* disease.

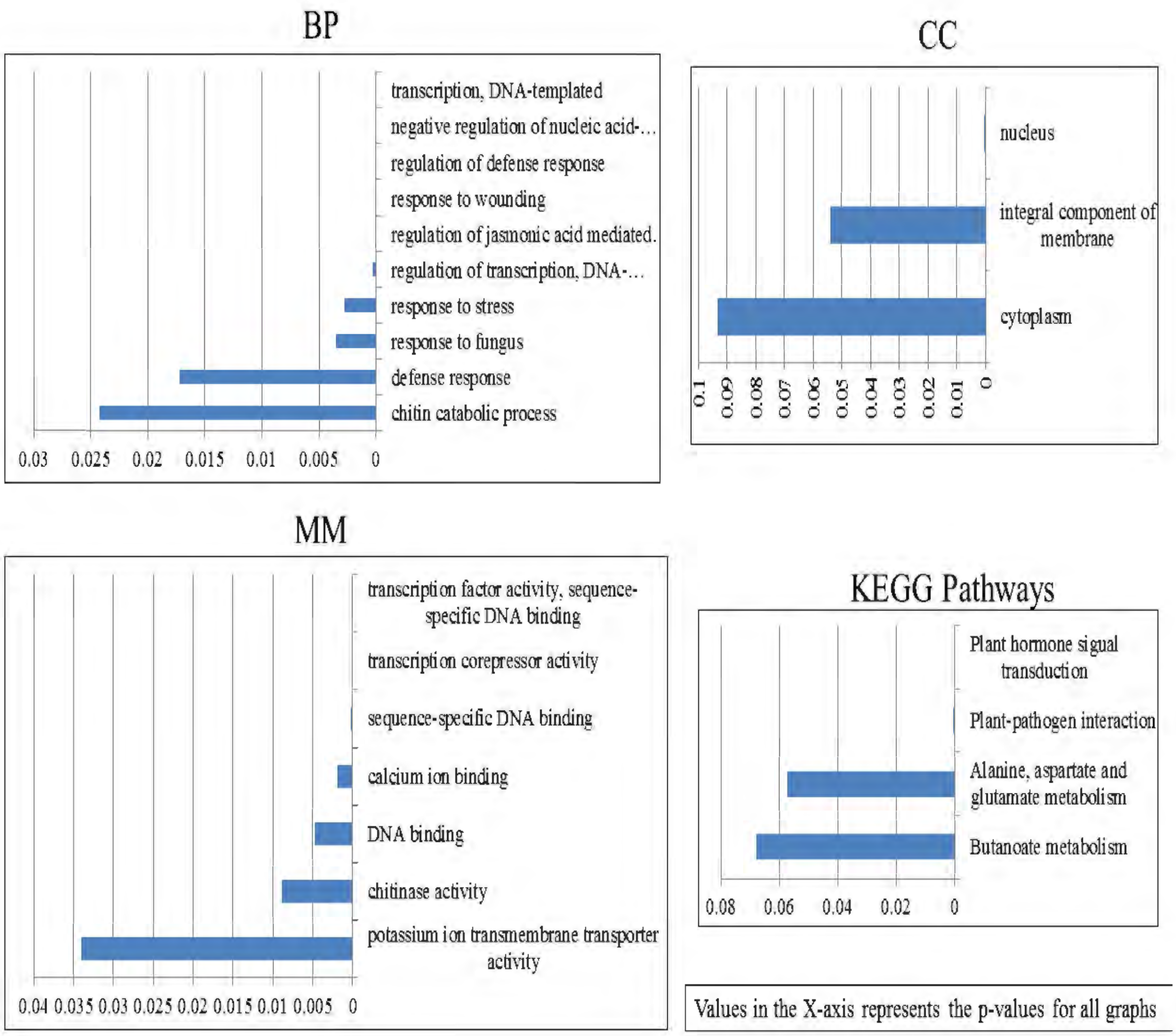


**Fig. 3.13.** GO analysis and KEGG pathways for classical method genes

### 3.3.2.2.2 PPI Network Analysis

The PPI association network was constructed by using STRING database. The network was analyzed in Cytoscape 3.7 and the cluster analysis was conducted by cytoHubba, a plugin software in Cytoscape. The hub proteins were selected based on Degree and Betweeness properties simultaneously. Networking profile of the all identified DE genes using the classical SVM and proposed SVM methods has been presented in **Fig.3.14** and in **Fig.3.15** respectively**.** By the proposed method and classical method, the top 10 hub genes were reported through the PPI network analysis in Cytoscape (**Fig.3.16** and **Fig.3.17**). Among them 4 hub genes (JAZ11, OS02T0205500-01, NAC4, JAZ13) are commonly shared by both methods. The other distinct hub genes JAZ6, JAZ9, JAZ12, OS06T0203600-01, OS12T0555200-01, EL5.1 as well as WRKY24, 4CL5, CML31, Os03g47280, P0505D12.1, DREB1G are reported by the classical SVM and proposed method SVM analysis respectively (**FigA3.16** and **Fig.3.17**). JAsmonate ZIM-domain (JAZ) protein families inhibit the action of transcription factors that perform reactions to the plant hormone jasmonoyl-L-isoleucine (JA-Ile) as well as proteins of these gene families are linked to wound-induced expression and resistance to insect herbivory in rice like other plant species(Thireault et al., 2015). Compared to classical SAM method, the proposed method also retrieved some new genes (WRKY24, 4CL5, CML31, Os03g47280, P0505D12.1, and DREB1G) which are significantly correlated with the membrane disease development or defense like *Magnaporthe grisea* fungal attack. It is observed that the proposed SVM method identifies two important genes such as DREB1G and WRKY24 of which DREB1G plays important roles in cold stress tolerance of in rice like other plants and this gene(transcription factor) could be valuable for developing transgenic rice with higher cold-stress tolerance(Moon et al., 2019). In addition, the WRKY24, a potential gene that is transcription factor of wheat (*Triticum aestivum* L.) has been identified by our proposed SVM technique that is predicted as a key regulatory function in defense against the rice blast fungus (*Magnaporthe grisea*) (Dangl and Jones, 2001; Kaloshian, 2004). In our investigation, the classical SVM however fails to identify these two very important genes highly related to various biotic and abiotic stress tolerances.

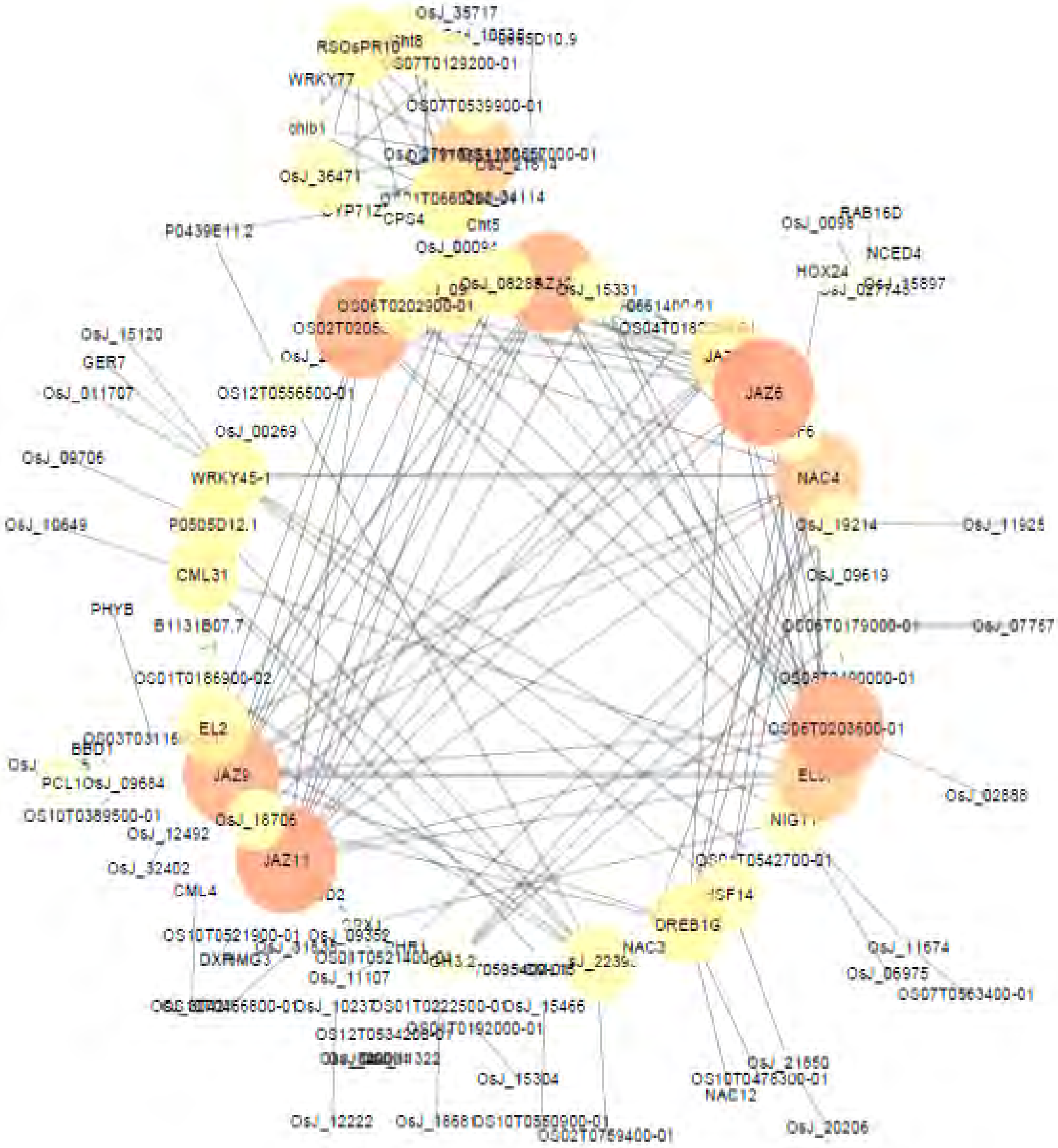


**Fig.3.14.** PPI network of the genes from the classical method

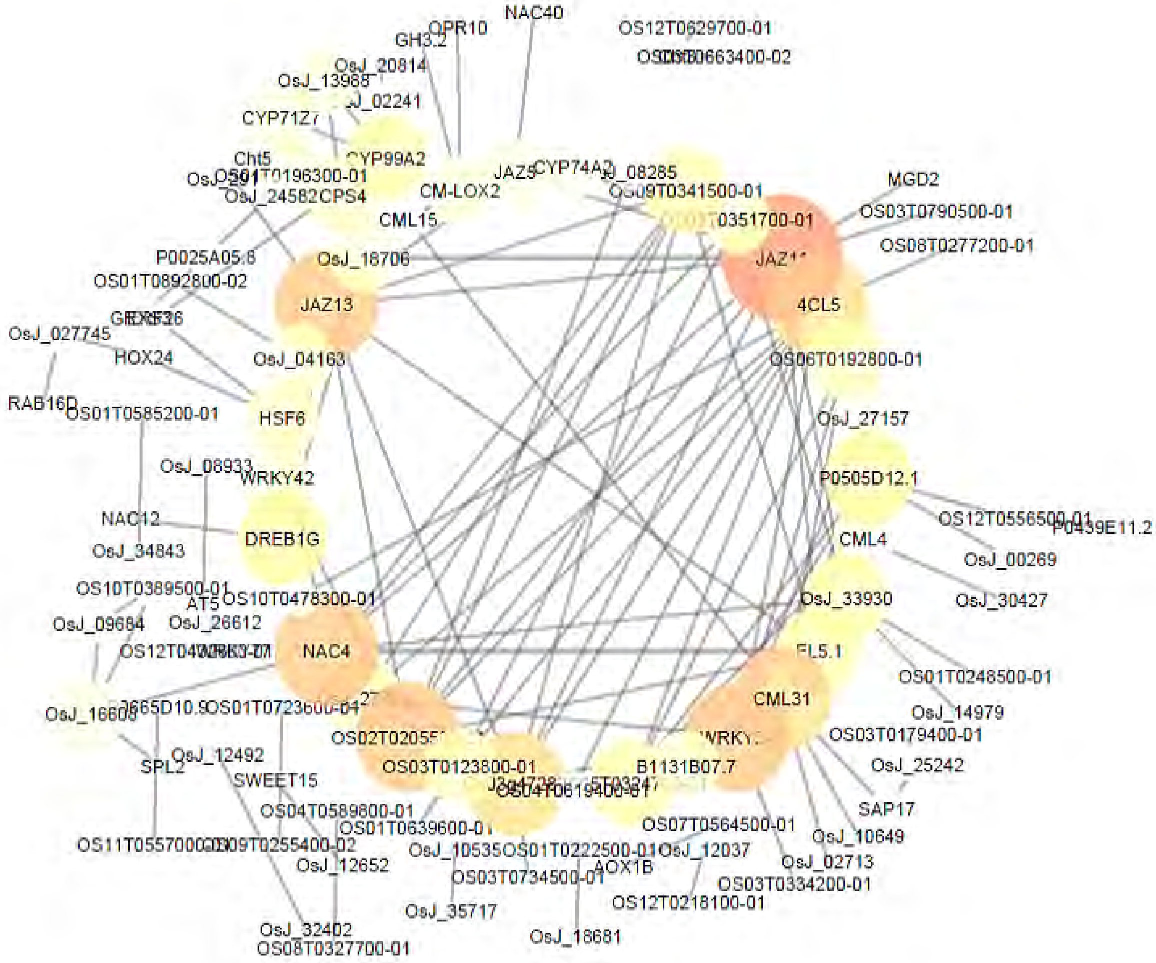


**Fig.3.15.** PPI network of the genes from the proposed method

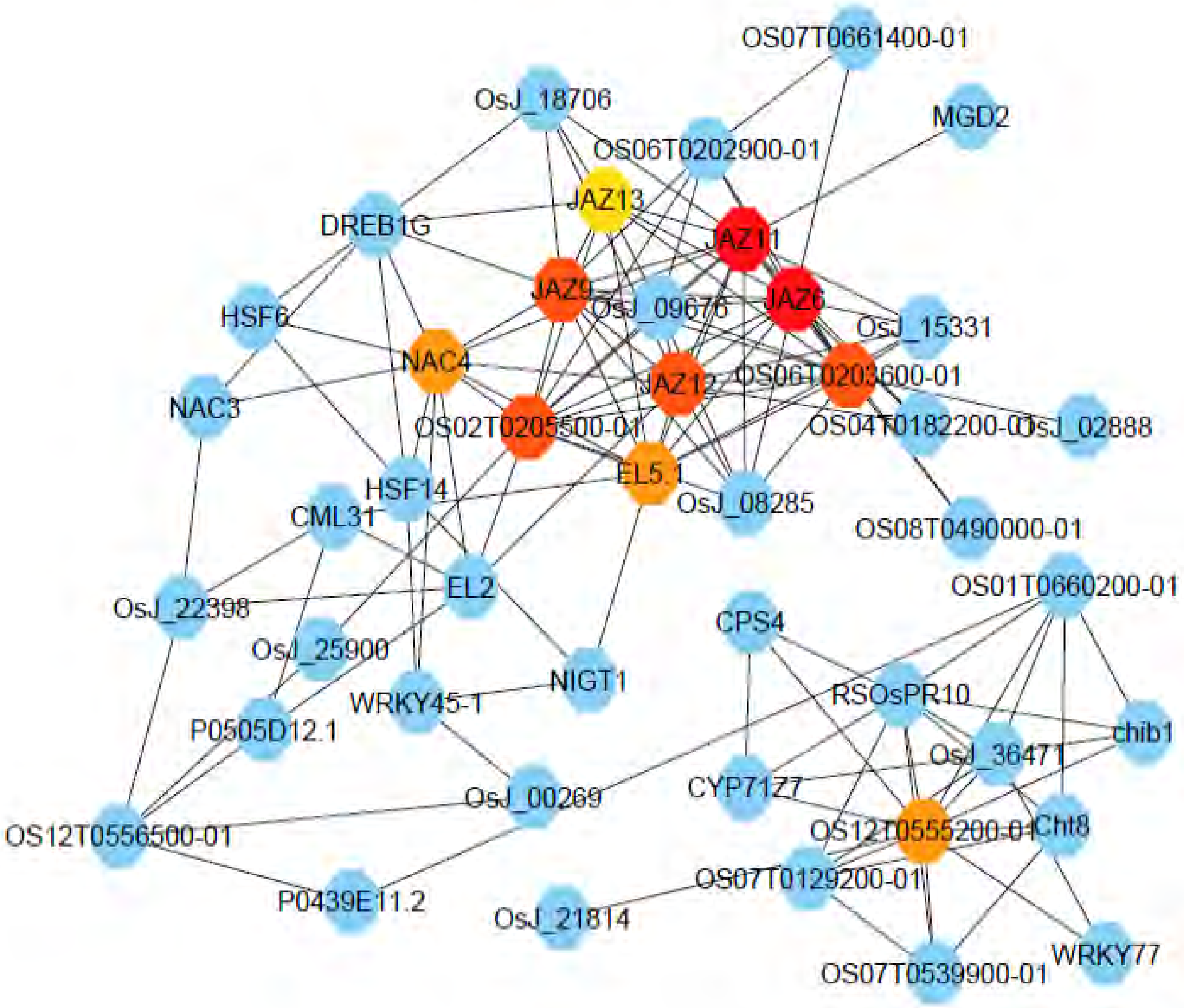


**Fig.3.16.** The top hub genes cluster based on Degree method obtained from classical method.

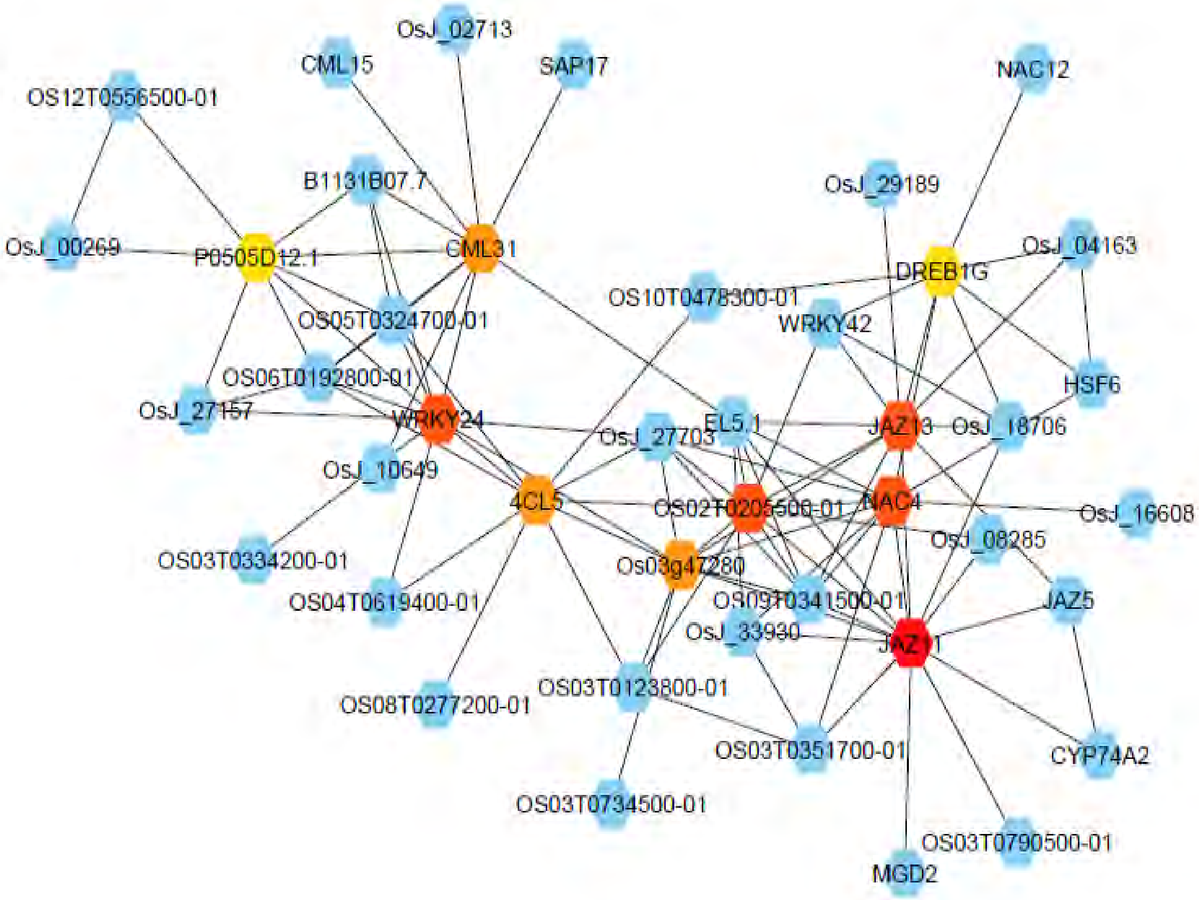


**Fig.3.17.** The top hub genes cluster based on Degree method from the proposed method.

## 3.4 Summary of the Chapter

Identification and suitable classification of DEGs obtained from the gene expression dataset are essential work for predicting genes or transcription factors associated to particular phenotypic and genotypic variation in different plant and animal species. In this work, we applied classical SAM and proposed SAM under logistic transformation in identifying DE genes in absence and in presence of 0%, 10%, 15% and 20% outlier. Our results shows that classical SAM approach produces nearly same AUC and pAUC values whereas proposed SAM approach produces much larger AUC and pAUC (at FPR=0.2) under the proposed logistic transformation only. In case of genes classification identified by out proposed SAM, we observed that the our proposed SVM performed better than classical SVM in terms of higher TPR, TNR, AUC and pAUC as well as lower FNR, FPR and MER. Real data analysis also provides the same performance. In addition, we observe that our proposed SVM approach has been able to identify two important genes such as WRKY24 and DREB1G that potential those are predicted to pay signification role in resistance to rice blast fungus and cold stress resistance in rice like other plants. Therefore, it could be concluded that the proposed logistically transformed transcriptomics data might be better for identification of DE genes and their precise classification.

# CHAPTER FOUR

## GENOME-WIDE ANALYSIS OF RNA SILENCING MACHINERY GENES IN WHEAT

# GENOME-WIDE IDENTIFICATION, CHARACTERIZATION AND DIVERSITY ANALYSIS OF RNA SILENCING MACHINERY GENES IN WHEAT (*Triticum aestivum* L.)

## 4.1 Introduction

Plants as well as different agricultural crops are tend to fight against different biotic and abiotic stresses in their life span such as fungi, bacteria, viruses, drought, salinity etc. This surely requires the development of appropriate genetic processes that involve suitable improvement and expression of resistance related genes to combat against such environmental stressors. In molecular biological research, RNA silencing or RNA interference (RNAi) is an important molecular phenomenon occurs in eukaryotic groups. In this technique a minute RNA fragment interferes using a particular nucleotide sequence (Cao et al., 2016). There are two categories of minute RNA fragments nearly size of 21-24 nucleotides that are formed in multi-cellular species termed as microRNA (miRNA) and short interfering RNA (siRNA) (Qian et al., 2011). These RNA molecules contribute in performing different biological processes such as development and growth, metabolism, anti-viral and anti-bacterial defense (Carrington and Ambros, 2003; Chen, 2012; Finnegan and Matzke, 2003; Lai, 2003; Van Ex et al., 2011). RNA silencing in plants is however initiated by double-stranded RNAs (dsRNA) that produce small RNAs are called microRNAs (miRNAs) or small-interfering RNAs (siRNAs)(Qian et al., 2011). Creation and role of these small RNAs largely rely on three main gene families, Dicer-like (DCLs), Argonauts (AGOs) and RNA-dependent RNA polymerases (RDRs) (Baulcombe, 2004; Chapman and Carrington, 2007; Vaucheret, 2006). A complete cycle of RNA silencing procedure takes three common steps: commencing, organization, and signal strengthening (Cao et al., 2016). DCLs undertake the RNase III-type actions that specifically change complementary double-stranded RNAs (dsRNAs) into small RNAs measuring 21-24 nucleotides in size (Carmell and Hannon, 2004). Next smaller RNAs are included into AGO-containing RNA-induced silencing complexes (RISCs) to function as the sequence specificity in RNA degradation, translational inhibition, or heterochromatin development(Bologna and Voinnet, 2014). When the signal extension starts, RDR enzymes are likely for creation of dsRNAs from single-stranded RNAs (ssRNAs) originals to switch a new round of RNA silencing(Sijen et al., 2001).DCL, AGO and RDR are known as the RNA silencing machinery genes work through the generation of small RNA molecules in plants. Among these genes, plant DCL proteins mostly process large size double-stranded RNAs into matured small RNAs (Carrington and Ambros, 2003; Chapman and Carrington, 2007; Qian et al., 2011). A DCL protein has been characterized as the existence of six domains such as DEAD, Helicase-C, DUF283, PAZ, RNaseIII and double-stranded RNA-binding motif (DSRM) (Carmell and Hannon, 2004; Margis et al., 2006). DCL gene families are also available in upper-class plants, insects, protozoa, and some fungi (Cao et al., 2016). AGO proteins are very particular small RNA-binding components well-known as the essential elements of RNA-silencing conduits (Vaucheret, 2008). Small RNAs produced by DCLs are conveyed for gene expression into the RISC together with AGO genes. Next these small RNAs direct AGOs to the target mRNA, accomplishing sequence-specific regulation of gene expression (Vaucheret, 2008). AGO proteins consists of several functional domains such as DUF283, PAZ, MID, and PIWI(Hutvagner and Simard, 2008). RDR is the third key protein also plays significant roles in RNA interference (RNAi) pathway for eukaryotic gene expression. These enzymes however have a common conserved catalytic domain called RNA-dependent RNA polymerase (RdRp) which is essential for beginning and increase of the silencing signal (Schiebel, 1998). Successive investigation so far reveals that there are numerous copies of DCL, AGO and RDR genes are present in plants and animals. All participants of these gene families take part in diverse roles in RNA silencing pathway. For instance, the genome of *Arabidopsis thaliana* possesses four DCL proteins (DCL1-DCL4) that definitely generate diverse kinds and sizes of small RNAs (Bologna and Voinnet, 2014). To activate various important biological functions in eukaryotic cell, different small RNAs produced in cell with diverse functions but they all are the associate members of DCL, AGO and RDR gene families. In *Arabidopsis thaliana*, 4 AtDCLs, 10 AtAGOs and 6 AtRDRs genes were identified (Vaucheret, 2008). Rice (Oryza sativa), a crop of monocot group, possesses 8 OsDCLs, 19 OsAGOs and 5 OsRDRs genes, in which OsAGO2 gene exhibited particular up-regulation in response to salt and drought (Kapoor et al., 2008; Qin et al., 2018). Also, in tomato (Solanum lycopersicum), 7 SlDCLs, 15 SlAGOs and 6 SlRDRs genes were identified(Bai et al., 2012). Five ZmDCLs, 18 ZmAGOs and 5 ZmRDRs genes were recognized in maize genome (Qian et al., 2011). There are 4 VvDCLs, 13 VvAGOs and 5 VvRDRs genes were detected in grapevine (Vitis vinifera)

(Zhao et al., 2015). Likewise, 5, 7, and 8 CsDCLs, CsAGOs, and CsRDRs, respectively have been identified in cucumber(Gan et al., 2016). On the other hand, the genome of allopolyploid species of Brassica napus contained 8 BnDCLs, 27 BnAGOs, and 16 BnRDRs (Cao et al., 2016; Zhao et al., 2016). Recently, overall 4 CaDCLs, 12 CaAGOs and 6 CaRDRs genes have been identified in pepper (Capsicum Annuum L.) (Qin et al., 2018). However, these potential genes in various important plants show considerable divergence and have indispensable role in different genomic functions.

Recent studies have revealed the DCL, AGO and RDR RNAi gene families in Brassica napus, maize, arabidopsis, rice, tomato, grapevine, cucumber, and pepper (Bai et al., 2012; Cao et al., 2016; Gan et al., 2016; Kapoor et al., 2008; Qian et al., 2011; Qin et al., 2018; Vaucheret, 2008; Zhao et al., 2015, 2016). Furthermore, these three well-known RNA silencing machinery gene-sets were validated previously in wet lab experiment to investigate their expression in various organs and tissues as well as at reproductive stages(Gan et al., 2016; Kapoor et al., 2008; Qin et al., 2018). Also the resistance ability of these RNAi proteins were explored against various abiotic( drought, salt, heat, cold etc.) and biotic(diseases viz., Sclerotinia scletotiorum, yellow leaf curl virus, tomato mosaic virus, cucumber mosaic virus, potato virus Y etc. ) factors in different crops(Bai et al., 2012; Kapoor et al., 2008; Qin et al., 2018; Zhao et al., 2016) as well.

However, there is no study has been made yet to reveal the DCL, AGO and RDR RNAi gene families in wheat, though it is a global common staple food and the second most-produced cereal crops after maize in the world (http://www.fao.org/worldfoodsituation/csdb/en/). Wheat is also the vital source of carbohydrates and is the leading source of vegetal protein in human food. Therefore, an effort has been made in this study to explore the important RNA interference (RNAi) genes from the wheat (Triticum aestivum) genome using various bioinformatic tools with respect to RNAi genes of *Arabidopsis thaliana*. To validate the in-silico predicted RNAi genes, an attempt has been made to investigate the expression profile of the predicted TaDCL candidate genes in two organs (leaves and roots) as well as expression analyses carried out against drought in T. aestivum by wet lab experiment. The information from this study will offer indications to detect biological important genes that contribute in different development and reproductive stages of wheat plant. The systemic schemes of this whole study in this article however have been demonstrated briefly in **Fig. 4.1**.

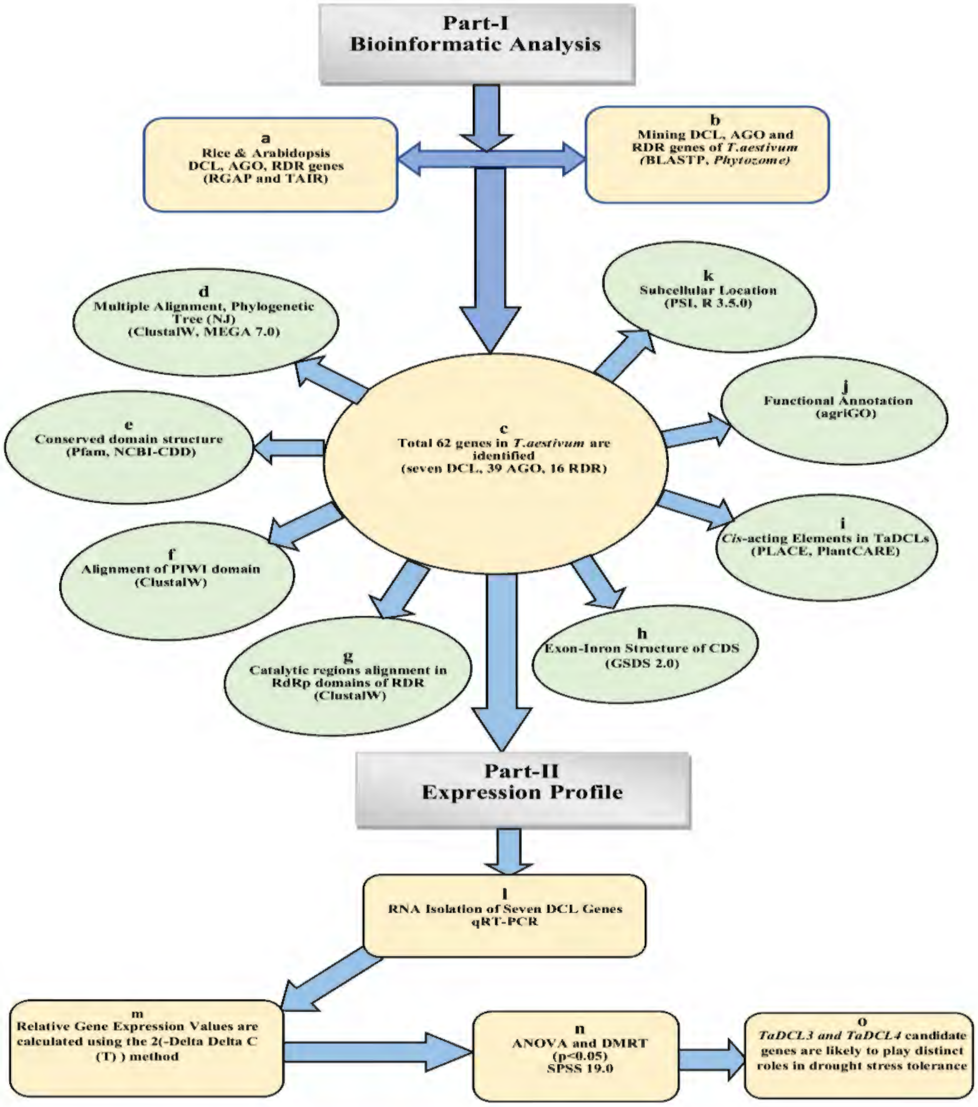


**Fig. 4.1** Schematic diagram of the whole work**.** **(a)** Rice and Arabidopsis protein sequences were obtained from RGAP and TAIR databases. **(b)** Arabidopsis protein sequences were used query to find corresponding protein sequences of *T.aestivum* from Phytozome using BLASTP technique. **(c)** Total 62 DCL, AGO and RDR genes were identified. **(d)** Multiple alignment and phylogenetic analysis were carried out using ClustalW and MEGA 7.0 software. **(e)** Conserved domain structure of each 62 genes were obtained using Pfam and NCBI-CDD database. **(f)** Obtained alignment profile of PIWI domain of amino acids wheat AGO proteins using ClustalW. **(g)** Alignment of catalytic regions in RdRp domains of RDR proteins of wheat, rice, and A. thaliana using ClustalW. **(h)** Exon-intron structure of wheat and A. thaliana DCL, AGO and RDR of full-length coding sequence (CDS) using GSDS 2.0. **(i)** Estimate of cis-Acting Elements of the promoter regios of the TaDCL Genes were calculated using PLACE and PlantCARE database. **(j)** Functional annotation/enrichment analysis of wheat DCL, AGO and RDR genes were performed to identify significantly enriched gene ontology and pathway terms. **(k)** Subcellular locations of each 62 genes were obtained to identify their involvement in metabolism process using PSI. **(l)** Gene expression analysis of TaDCL genes in leaves, roots and under drought stress following quantitative real time PCR (qRT-PCR) technique. **(m)** Gene expressions were obtained using $2^{-\Delta\Delta Ct}$ method. **(n)** Significant genes were calculated following ANOVA and DMRT (P < 0.05). (o) TaDCL3 and TaDCL4 and were significantly induced upon drought treatment and showed higher expression level.

## 4.2 Materials and Methods

### 4.2.1 Identification of DCL, AGO and RDR Genes in Wheat (*Triticum aestivum* L.)

Protein sequences of Arabidopsis and rice DCLs, AGOs and RDRs were obtained from TAIR (http://www.arabidopsis.org) and RGAP (http://rice.plantbiology.msu.edu/) respectively and then these sequences were used as query to search by BLAST-P program against *T. aestivum* genome in the sequenced plant genome database Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html). All retrieved sequences were searched for conserved domains using the Pfam (http://pfam.sanger.ac.uk/) and the National Center for Biotechnology Information Conserved Domain Database (NCBI-CDD: http://www.ncbi.nlm.nih.gov/ Structure/cdd/wrpsb.cgi). Some basic information of these genes such as accession number, genomic location, gene length, and encoded protein length were downloaded from Phytozome-wheat genome database.

### 4.2.2 Sequence Alignment and Phylogenetic Analysis

Protein sequences of DCL, AGO and RDR genes belong to respective *Arabidopsis thaliana*, rice and wheat were used for multiple alignment using ClustalW program(Thompson et al., 1994). After that, phylogenetic tree was constructed using MEGA7: Molecular Evolutionary Genetics Analysis version 7.0(Kumar et al., 2015) for larger datasets by Neighbor-Joining (NJ) method. Bootstrap analysis was done with 1000 replicates to measure statistical support for nodes. All identified genes in this study were named based on their phylogenetic relationship and sequence homologies with corresponding *Arabidopsis thaliana* homologs.

### 4.2.3 Exon-Intron Analysis of DCL, AGO and RDR Genes in *T. aestivum*

The arrangement of respective exon-intron structure of seven DCLs, 39 AGOs and 16 RDRs genes were assessed using the online Gene Structure Display Server (GSDS 2.0 : http://gsds.cbi.pku.edu.cn/index.php) by considering their complete coding sequences (CDS) with their matching genomic sequences downloaded from Phytozome database.

### 4.2.4 Prediction of *Cis*-acing Elements in the TaDCL Genes in *T. aestivum*

To explore *cis*-elements in the promoter sequences of the *TaDCL* genes,1.5kb sequences upstream of the start codon (ATG) were gathered and exposed to stress response-related *cis*-acting element online prediction analysis with Signal Scan search platform in the PLACE database (http://www.dna. affrc.go.jp/PLACE/signalscan.html) and the PlantCARE databank (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/). The sequences data of TaDCLs promoters were taken from phytozome 12. Nine *cis*-elements were used in this investigation: dehydration and cold response (DRE/CRT: RCCGAC), Ethylene Response Factors (ERFs) binding site (GCC box: AGCCGCC, ABA responsive element (ABRE: YACGTGK), ARF1 binding site (AuxRE: TGTCTC), SA-responsive promoter element (SARE: TGACG), environmental signal response (G-box: CACGTG), WRKY binding site (W-box: TTGACY), CAMTA binding site (CG-box: VCGCGB) and sulfur-responsive element (SURE: GAGAC).

### 4.2.5 Gene Ontology (GO) Enrichment Analysis of RNAi Genes in *T. aestivum*

GO enrichment analysis of the identified RNA silencing machinery genes of seven DCLs, 39AGOs and 16 RDRs were obtained using the web-based tool PlantTFDB v4.0(Jin et al., 2017). Fisher's exact test was followed to determine respective *P*-values and Benjamin-Hochberg's adjustment procedure was used as the multiple testing correction performance. Enrichment outcomes with adjusted $P < 0.05$ were considered as statistically important.

### 4.2.6 Subcellular Location of DCL, AGO and RDR in *T. aestivum*

A web-based tool called PSI (Plant Subcellular localization Investigative Predictor) [27] for predicting the subcellular location of the 62 RNA silencing genes and open source software R-3.5.2 [62] were used. A protein was considered located in the certain cellular location if $p < 0.00001$ in the resulting output of the tool.

### 4.2.7 RNA Isolation and Gene Expression Analysis

Complete RNA was taken out using Trizol reagent (TAKARA, Japan) by following the manufacturer's guidelines. RNA was treated with DNaseI (TAKARA, Japan) and reverse-transcribed into cDNA using the PrimeScript RT reagent kit (TAKARA, Japan). The collected cDNAs were used for gene expression studies with quantitative real time PCR (qRT-PCR). The qRT-PCR was accomplished in StepOne Real-Time PCR System (Applied Biosystems, USA) using SYBER Premix Ex Taq reagents (TAKARA, Japan) following the program: 95◦C for 30s, 95◦C for 5s and 60◦C for 45s for 40 cycles. To normalize the sample variance, *T.aestivum* 18s gene was used as internal control. Relative gene expression values were calculasetted using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). The primers used for gene expression analyses are listed at (**Table A4.1).** For the statistical analysis of the gene expression data, ANOVA was performed with SPSS software (Version19.0, IBM, USA). An important difference between mean values was assessed following DMRT ($P < 0.05$) approach.

## 4.3 Results

### 4.3.1 Genome-Wide Identification of DCL, AGO and RDR Genes in Wheat (*T. aestivum*)

To recognize RNA silencing machinery genes in *T. aestivum*, the well-characterized *Arabidopsis* four DCLs, 10 AGOs and six RDRs protein sequences were used as query to BLAST-P search in the whole genome of *Triticum aestivum* in *Phytozome* database. On the basis of the domain composition analysis for the recovered candidate sequences, a full set of seven DCLs, 39 AGOs and 16 RDRs genes were recognized in wheat genome, representative of the maximum number of RNA silencing genes in given plant species identified to date (**Table 4.1**).

**Table 4.1.** List of wheat DCL, AGO and RDR genes

| Gene no. | Gene name | Accession no. | Genomic location | Gene length (bp) | No. of introns | Protein length (aa) | Protein domains |
|---|---|---|---|---|---|---|---|
| | | | **Dicer-Like(DCL)** | | | | |
| 1 | TaDCL1a | Traes_4BL_B3A1B8342.2 | ta_iwgsc_4bl_v1_ 6994694:2688..13103 | 10416 | 18 | 1779 | DEAD, Helicase-C, Dicer-dimer/DUF283, PAZ, Ribonuclease-3, DSRM |
| 2 | TaDCL1b | Traes_5AL_72A7552B9.2 | ta_iwgsc_5al_v1_ 2752914:2855..12171 | 9317 | 17 | 1674 | DEAD, Helicase-C, DUF283, PAZ, Ribonuclease-3, DSRM |
| 3 | TaDCL3a | Traes_1AL_E7144546E.1 | ta_iwgsc_1al_v2_ 3918617:7862..20042 | 12181 | 25 | 1627 | DEAD, Helicase-C, DUF283, PAZ, Ribonuclease-3 |
| 4 | TaDCL3b | Traes_1DL_C646B6990.1 | ta_iwgsc_1dl_v1_ 2224660:33..11929 | 11897 | 22 | 1494 | DEAD, Helicase-C, DUF283, PAZ, Ribonuclease-3 |
| 5 | TaDCL3c | Traes_3AL_562D6614F.1 | ta_iwgsc_3al_v1_ 4401047:3091..12058 | 8968 | 23 | 1586 | DEAD, Helicase-C, DUF283 PAZ,Ribonuclease-3, DSRM |

**Table 4.1**. List of wheat DCL, AGO and RDR genes (Continued)

| Gene no. | Gene name | Accession no. | Genomic location | Gene length (bp) | No. of introns | Protein length (aa) | Protein domains |
|---|---|---|---|---|---|---|---|
| 6 | TaDCL3d | Traes_3DL_2DC78B18A.1 | ta_iwgsc_3dl_v1_ 6937552:192..8538 | 8347 | 23 | 1586 | DEAD, Helicase-C, DUF283 PAZ, Ribonuclease-3,DSRM |
| 7 | TaDCL4 | Traes_2DL_E96DCDCB4.2 | ta_iwgsc_2dl_v1_ 9883052:5868..20610 | 14743 | 18 | 1392 | Helicase-C, DUF283, PAZ, Ribonuclease-3,Ribonuclease-3, DND1-DSRM |
| **Argonaute (AGO)** | | | | | | | |
| 1 | TaAGO1a | Traes_2AL_2512A7F91.1 | ta_iwgsc_2al_v1_ 6428016:16577..22920 | 6344 | 21 | 1117 | Gly-rich Ago1, DUF1785, PAZ, PIWI |
| 2. | TaAGO1b | Traes_2BL_93099ACF4.1 | ta_iwgsc_6al_v1_ 5742333:938..12517 | 11580 | 16 | 645 | PAZ, PIWI |
| 3. | TaAGO1c | Traes_6AL_616161AAB.1 | ta_iwgsc_6al_v1_ 5742333:938..12517 | 11580 | 21 | 978 | DUF1785, PAZ, PIWI |
| 4. | TaAGO1d | Traes_6DL_58620B158.2 | ta_iwgsc_6dl_v1_ 3210843:3..6281 | 6279 | 21 | 910 | DUF1785, PAZ, PIWI |
| 5. | TaAGO1e | Traes_7DL_C255A109C.1 | ta_iwgsc_7dl_v1_ 3396214:4962..9930 | 4969 | 14 | 721 | DUF1785, PAZ, PIWI |
| 6. | TaAGO1f | Traes_6BL_9CFA54D4A.1 | ta_iwgsc_6bl_v1_ 4398549:8497..16669 | 8173 | 21 | 1068 | Gly-rich Ago1, DUF1785, PAZ, PIWI |
| 7. | TaAGO1g | Traes_6AL_317133B3F.2 | ta_iwgsc_6al_v1_ 5770608:7589..20031 | 12443 | 22 | 1085 | Gly-rich Ago1, DUF1785, PAZ, PIWI |
| 8. | TaAGO1h | Traes_6DL_804FB7F75.1 | ta_iwgsc_6dl_v1_ 3318900:4215..12555 | 8341 | 21 | 1067 | Gly-rich Ago1, DUF1785, PAZ, PIWI |
| 9. | TaAGO1i | Traes_7AS_56569A5AC.2 | ta_iwgsc_7as_v1_ 4062808:1122..8394 | 7273 | 21 | 1210 | Gly-rich Ago1, DUF1785, PAZ, PIWI |
| 10. | TaAGO1j | Traes_7DS_4D01B6175.1 | ta_iwgsc_7ds_v1_ 3925271:1281..7139 | 5859 | 21 | 868 | DUF1785, PAZ, PIWI |
| 11. | TaAGO1k | Traes_4AL_A118C6C84.2 | ta_iwgsc_4al_v2_ 7127490:1302..7644 | 6343 | 21 | 1189 | Gly-rich Ago1, DUF1785, PAZ, PIWI |
| 12. | TaAGO2a | Traes_2AL_DFE4C65F6.2 | ta_iwgsc_2al_v1_ 6435863:9..4673 | 4665 | 2 | 952 | DUF1785, PAZ, PIWI |
| 13. | TaAGO2b | Traes_2BL_7713B3533.2 | ta_iwgsc_2bl_v1_ 8046272:2..4842 | 4841 | 1 | 845 | DUF1785, PAZ, PIWI |
| 14. | TaAGO3 | Traes_2DL_A77212060.2 | ta_iwgsc_2dl_v1_ 9902460:2552..5785 | 3234 | 2 | 849 | DUF1785, PAZ, PIWI |
| 15. | TaAGO4a | Traes_3AS_8EE711E2C.2 | ta_iwgsc_3as_v1_ 3376626:7020..14138 | 7119 | 21 | 999 | DUF1785, PAZ, PIWI |

**Table 4.1**. List of wheat DCL, AGO and RDR genes (Continued)

| Gene no. | Gene name | Accession no. | Genomic location | Gene length (bp) | No. of introns | Protein length (aa) | Protein domains |
|---|---|---|---|---|---|---|---|
| **Argonaute (AGO)** | | | | | | | |
| 16. | TaAGO4b | Traes_3DS_57EA31670.1 | ta_iwgsc_3ds_v1_2601767:2..5053 | 5052 | 17 | 799 | DUF1785, PAZ, PIWI |
| 17. | TaAGO4c | Traes_3B_F4E4667F8.1 | ta_iwgsc_3b_v1_10588096:6226..11781 | 5556 | 20 | 918 | DUF1785, PAZ, PIWI |
| 18. | TaAGO5a | Traes_2BS_8368F6B5D.1 | ta_iwgsc_2bs_v1_5155056:10..6355 | 6346 | 21 | 815 | DUF1785, PAZ, PIWI |
| 19. | TaAGO5b | Traes_2DS_4CC8FD7E3.1 | ta_iwgsc_2ds_v1_5380297:2..6639 | 6638 | 21 | 838 | DUF1785, PAZ, PIWI |
| 20. | TaAGO5c | Traes_5BL_F505BF164.1 | ta_iwgsc_5bl_v1_10819703:15714..21283 | 5570 | 19 | 732 | DUF1785, PAZ, PIWI |
| 21. | TaAGO5d | Traes_4AL_7CC35DF1D.2 | ta_iwgsc_4al_v2_7146652:9555..15647 | 6093 | 21 | 875 | DUF1785, PAZ, PIWI |
| 22. | TaAGO5e | Traes_4DS_88D2821C6.2 | ta_iwgsc_4ds_v1_2318247:117..6407 | 6291 | 21 | 916 | DUF1785, PAZ, PIWI |
| 23. | TaAGO5f | Traes_3AS_3F8424E4E.1 | ta_iwgsc_3as_v1_1779497:2..5110 | 5109 | 20 | 832 | DUF1785, PAZ, PIWI |
| 24. | TaAGO5g | Traes_3B_CA99AB66C.1 | ta_iwgsc_3b_v1_9317732:662..5865 | 5204 | 20 | 841 | DUF1785, PAZ, PIWI |
| 25. | TaAGO6a | Traes_1BL_05F7B7DFA.1 | ta_iwgsc_1bl_v1_2609948:748..7252 | 6505 | 19 | 893 | DUF1785, PAZ, PIWI |
| 26. | TaAGO6b | Traes_5AL_07EFD5712.1 | ta_iwgsc_5al_v1_2801172:3925..11424 | 7500 | 21 | 883 | DUF1785, PAZ, PIWI |
| 27. | TaAGO6c | Traes_5DL_672EE3605.1 | ta_iwgsc_5dl_v1_4504514:5254..15341 | 10088 | 21 | 882 | DUF1785, PAZ, PIWI |
| 28. | TaAGO6d | Traes_5BL_F611D65E0.1 | ta_iwgsc_5bl_v1_10861763:7394..17699 | 10306 | 21 | 883 | DUF1785, PAZ, PIWI |
| 29. | TaAGO6e | Traes_7AL_D88450A3C.2 | ta_iwgsc_7al_v1_944917:567..5236 | 4670 | 17 | 768 | DUF1785, PAZ, PIWI |
| 30. | TaAGO7a | Traes_2AL_3F3117458.1 | ta_iwgsc_2al_v1_6334464:10651..14850 | 4200 | 2 | 934 | DUF1785, PAZ, PIWI |
| 31. | TaAGO7b | Traes_2BL_24111235C.1 | ta_iwgsc_2bl_v1_7955795:2637..7172 | 4536 | 2 | 1023 | DUF1785, PAZ, PIWI |
| 32. | TaAGO8 | Traes_7AL_1BAB53DCE.1 | ta_iwgsc_7al_v1_4478819:3359..8183 | 4825 | 21 | 901 | DUF1785, PAZ, PIWI |
| 33. | TaAGO9a | Traes_1AL_095416BC0.1 | ta_iwgsc_1al_v2_3876661:1149..9699 | 8551 | 20 | 925 | DUF1785, PAZ, PIWI |
| 34. | TaAGO9b | Traes_1BL_7C037D478.2 | ta_iwgsc_1bl_v1_3899789:1061..10207 | 9147 | 21 | 927 | DUF1785, PAZ, PIWI |
| 35. | TaAGO9c | Traes_1DL_64B330BBB.2 | ta_iwgsc_1dl_v1_2275968:1227..11215 | 9989 | 21 | 922 | DUF1785, PAZ, PIWI |
| 36. | TaAGO10a | Traes_6AS_FBB2AFAAB.1 | ta_iwgsc_6as_v1_4364847:5932..12979 | 7048 | 20 | 948 | DUF1785, PAZ, PIWI |
| 37. | TaAGO10b | Traes_6DS_9DD64BD48.1 | ta_iwgsc_6ds_v1_2082993:2072..10005 | 7934 | 20 | 883 | DUF1785, PAZ, PIWI |
| 38. | TaAGO10c | Traes_7AL_96766587F.2 | ta_iwgsc_7al_v1_4543530:3452..7823 | 4372 | 15 | 583 | PAZ, PIWI |
| 39. | TaAGO10d | Traes_7DL_C538856D4.1 | ta_iwgsc_7dl_v1_3364674:1..5438 | 5438 | 18 | 664 | PAZ, PIWI |

**Table 4.1**. List of wheat DCL, AGO and RDR genes (Continued)

| Gene no. | Gene name | Accession no. | Genomic location | Gene length (bp) | No. of introns | Protein length (aa) | Protein domains |
|---|---|---|---|---|---|---|---|
| **RNA-dependent RNA Polymerase (RDR)** | | | | | | | |
| 1. | TaRDR1a | Traes_6DL_4B89E8742.2 | ta_iwgsc_6dl_v1_3272251:691..6878 | 6188 | 3 | 1120 | RBD, RdRP |
| 2. | TaRDR1b | Traes_6BL_78BEF51DD.1 | ta_iwgsc_6bl_v1_4353480:4946..8011 | 3066 | 2 | 727 | RdRP |
| 3. | TaRDR1c | Traes_6AL_393C6B853.1 | ta_iwgsc_6al_v1_5823227:4526..9827 | 5302 | 3 | 1119 | RBD, RdRP |
| 4. | TaRDR1d | Traes_6BL_0A9D15EDC.2 | ta_iwgsc_6bl_v1_4254864:3..1709 | 1707 | 1 | 484 | RdRP |
| 5. | TaRDR1e | Traes_6BL_0BB5C493D.1 | ta_iwgsc_6bl_v1_4369818:3771..6377 | 2607 | 2 | 333 | RdRP |
| 6. | TaRDR1f | Traes_6AL_13BC97E04.1 | ta_iwgsc_6al_v1_5769298:833..7192 | 6360 | 3 | 1116 | RBD, RdRP |
| 7. | TaRDR1g | Traes_6BL_DF680C2AF.1 | ta_iwgsc_6bl_v1_4369819:1195..5786 | 4592 | 4 | 947 | RdRP |
| 8. | TaRDR2a | Traes_2DL_6DB81005E.1 | ta_iwgsc_2dl_v1_9891666:1..7150 | 7150 | 2 | 732 | RdRP |
| 9. | TaRDR2b | Traes_4AS_8D6311711.1 | ta_iwgsc_4as_v2_5940771:10843..26735 | 15893 | 3 | 1127 | RdRP |
| 10. | TaRDR2c | Traes_4DL_2E9CE89D9.2 | ta_iwgsc_4dl_v3_14391157:7408..18499 | 11092 | 4 | 1156 | RdRP |
| 11. | TaRDR2d | Traes_4DL_A54C80661.1 | ta_iwgsc_4dl_v3_14383226:1730..7149 | 5420 | 2 | 869 | RdRP |
| 12. | TaRDR3 | Traes_7BL_8CEC8F99B.2 | ta_iwgsc_7bl_v1_6687438:7..10355 | 10349 | 13 | 616 | RdRP |
| 13. | TaRDR4 | Traes_3AS_F27BB108C.2 | ta_iwgsc_3as_v1_3345716:2930..10167 | 7238 | 11 | 436 | RdRP |
| 14. | TaRDR5 | Traes_3B_2C6DB84FB.2 | ta_iwgsc_3b_v1_10500163:2098..12737 | 10640 | 6 | 560 | RdRP |
| 15. | TaRDR6a | Traes_3B_DC77B5E89.1 | ta_iwgsc_3b_v1_10414255:3..4337 | 4335 | 1 | 923 | RdRP |
| 16. | TaRDR6b | Traes_3DL_F32B49981.1 | ta_iwgsc_3d_v1_6978:1..3009 | 3009 | 1 | 543 | RdRP |

### 4.3.2 Clustering DCL, AGO and RDR Genes Based on Phylogenetic Analysis in *T. aestivum*

With the aim of determining the phylogenetic link and evolutionary difference of TaDCLs, TaAGOs and TaRDRs with those of Arabidopsis and rice homologs, three independent phylogenetic trees were created from the alignments of their encoded protein sequences using neighbor-joining methods (**Fig. 4.2**). The TaDCL proteins showed high sequence conservation compared with their *Arabidopsis* and rice counterpart. **Fig. 4.2(a)** shows that the seven TaDCLs were obviously classified into four distinct clades. All the DCL members of the corresponding group are clustered with the similar orthologs AtDCLs and OsDCLs at a high similarity. In relation to the phylogenetic association and sequence homology with AtDCLs, the TaDCLs were named as TaDCL1a, TaDCL1b, TaDCL3a, TaDCL3b, TaDCL3c, TaDCL3d, and TaDCL4. However, there were no TaDCL2 gene(s) corresponding to AtDCL2 and OsDCL2.

Based on the NJ phylogenetic analysis and the protein sequence homology with AtAGOs, the 39 candidate of TaAGOs family consisted of 11 AGO1s (TaAGO1a-TaAGO1k), two AGO2s (TaAGO2a, TaAGO2b), one AGO3 (TaAGO3), three AGO4s (TaAGO4a, TaAGO4b, TaAGO4c), seven AGO5s (TaAGO5a-TaAGO5g), five AGO6s (TaAGOa-TaAGO6e), two AGO7s (TaAGO7a and TaAGO7b), one AGO8 (TaAGO8), three AGO9s (TaAGO9a-TaAGO9c) and four AGO10s (TaAGO10a-TaAGO10d) (**Fig. 4.2(b)**). It is however observed from the phylogenetic tree that TaAGO8 and three TaAGO9s (TaAGO9a-TaAGO9c) form a group with only OsAGO4b not with the corresponding AtAGO8 and AtAGO9s. On the other hand, it is unusually obvious that TaAGO5a and TaAGO5b groups with OsAGO14 gene. TaAGO5c, TaAGO5d, TaAGO5e form a group with OsAGO11 and OsAGO12 as well as TaAGO5f and TaAGO5g specifically with OsAGO18 but not with any AtAGOs. The characteristic of these genes indicate that TaAGO8, three TaAGO9s, and seven TaAGO5s are the monocot specific that originated during evolution.

Finally, like DCLs and AGOs, RDR genes in *T. aestivum* were named after the Arabidopsis homologs, which were the evolutionary neighboring in the phylogenetic tree, and exhibit the top protein sequence homologies. **Fig.4.2(c)** illustrates that the phylogenetic tree however produced from the RDR protein sequences is separated into four major clusters. RDR1 clade has seven members (TaRDR1a, TaRDR1b, TaRDR1c, TaRDR1d, TaRDR1e, TaRDR1f, TaRDR1g,) form a cluster with their AtRDR1 and OsRDR1 counterparts. Clade RDR2 has four members (TaRDR2a, TaRDR2b, TaRDR2c, TaRDR2d,) and clearly generates a distinct cluster with their AtRDR2 and OsRDR2 homologs. TaRDR3, TaRDR4, and TaRDR5 together form a cluster with their similar genes to AtRDR3, AtRDR4, AtRDR5, OsRDR3, and OsRDR4. Clade RDR6 has two members TaRDR6a and TaRDR6b produced a distinct group with AtRDR6 and OsSHL2 genes in the phylogenetic tree imply that these TaRDR protein families covey the similar genetic information of their AtRDR and OsRDR counterparts.

### 4.3.3 Conserve Domain Composition of DCL, AGO and RDR Genes in *T. aestivum*

Conserved domain organization of *T. aestivum* DCL, AGO and RDR proteins were almost similar to the model plant *Arabidopsis* except some differences in AGOs and RDRs. By NCBI and Pfam conserved domain databases analysis showed that all TaDCLs retain the same domains DEAD, Helicase-C, DUF283, PAZ, and RNaseIII as reported for Arabidopsis DCLs. The remaining TaDCL4 gene is short of the DEAD domain, which is different from the AtDCL4 (**Fig.4.3(a**)). The gene length and the protein length varied slightly between and within groups of this gene family. The gene length of TaDCL4 was the largest (14743bp) but protein length was lowest (1392aa) followed by TaDCL3a (12181bp), TaDCL3b (11897bp). However, the TaDCL3c and TaDCL3d had two distinct gene lengths of 8968bp and 8347bp respectively but both these gene had the same size of protein length 1586aa. The TaDCL1a and TaDCL1b had the varied number of gene and protein length (10416bp, 1779aa) and (9317bp, 1674aa) respectively.

The domain arrangement of TaAGOs was the similar to that of AtAGOs. All TaAGOs retained DUF1785, PAZ, and PIWI domains. Six TaAGO1s (TaAGO1a, TaAGO1f, TaAGO1g, TaAGO1h, TaAGO1i and TaAGO1k) proteins however possessed an additional domain called Gly-rich Ago1 (**Fig. 4.3(b**)).The gene length of TaAGOs varied from 12443bp for TaAGO1g to 3234bp for TaAGO3.Moreover, earlier investigation showed that the PIWI domain presenting wide-ranging homology to RNase H binds the siRNA 5′ end to the target RNA(Höck and Meister, 2008) and cleaves target RNAs that show sequence complementary to small RNAs (Baumberger and Baulcombe, 2005; Rivas et al., 2005). Three conserved metal chelating residues in the PIWI domain such as aspartate, aspartate, and histidine (DDH) are associated in this regard (Kapoor et al., 2008). This metal chelating residues function as catalytic triad were first identified in Arabidopsis AGO1, and a conserved histidine at position 798 (H798) was also found to be critical for AGO1 for *in vitro* endonuclease activity (Baumberger and Baulcombe, 2005). To find out the existence of conserved catalytic residues (DDH/H), which are responsible for the endonuclease activity of AGO proteins involved in RNAi, the PIWI domains of all TaAGOs were aligned using CLUSTALX (**Fig.4.4**). The PIWI domain sequence alignment showed that total 16 members namely 8 TaAGO1s (TaAGO1a, TaAGO1b, TaAGO1c, TaAGO1d, TaAGO1e, TaAGO1i, TaAGO1j, TaAGO1k), four TaAGO5s (TaAGO5c, TaAGO5d, TaAGO5f, TaAGO5g), two TaAGO7s (TaAGO7a, TaAGO7b), two TaAGO10s (TaAGO10c, TaAGO10d) possessed the four important active residues (DDH/H), implying that they might have endonuclease activity(Rivas et al., 2005) (**Fig.4.4**). The rest of the 23 TaAGOs showed at least one dissimilarity among these catalytic traid residues. The D760 residue in TaAGO8 was changed by Tyrosine (Y) whether for other's was remain same as AtAGO1. In the TaAGO2a, TaAGO2b, TaAGO3 and TaAGO6a, TaAGO6e the H986 residue was replaced by aspartate (D) and "-" respectively (**Fig.4.4** and **Table 4.2**). The changes in the H798 catalytic residue was happened by three amino acid residue called arginine (R) (TaAGO1f, TaAGO1g and TaAGO1h); Proline (P) (TaAGO4a, TaAGO4b, TaAGO4c, TaAGO5a and TaAGO5b) and Tyrosine (Y) (TaAGO10c, TaAGO10d) (**Fig.4.4** and **Table 4.2**). The replacement/ mutational changes of the D760, D845, H986 and H798 (DDH/H) catalytic residues in the reported wheat AGO genes are clearly indicating the functional distinction compared to the AGO proteins in Rice and Arabidopsis. The further investigation of these genes is demanded for deeper understanding about the PIWI domain nuclease activities as well as the biological functionality in Wheat.

Total 16 TaRDR proteins possessed an RdRp domain (**Fig.4.3(c)**). In addition, the proteins TaRDR1a, TaRDR1c, and TaRDR1f carried an additional domain named RNA-binding domain (RBD) of their *Arabidopsis* counterpart. The gene length of the TaRDR varied from 1707 bp for TaRDR1d and 15893bp for TaRDR2b with potentially encoding 434 and 1127 amino acids respectively (see Table 1). Additionally, TaRDRs proteins were studied to identify the presence of the catalytic motif in the RdRp domain. The sequence alignment revealed that all members of TaRDR1, TaRDR2, and TaRDR6 groups shared a common DLDGD motif in the catalytic domain, whereas the proteins TaRDR3, TaRDR4, and TaRDR5 retained DFDGD motif (**Fig. 4.5**).

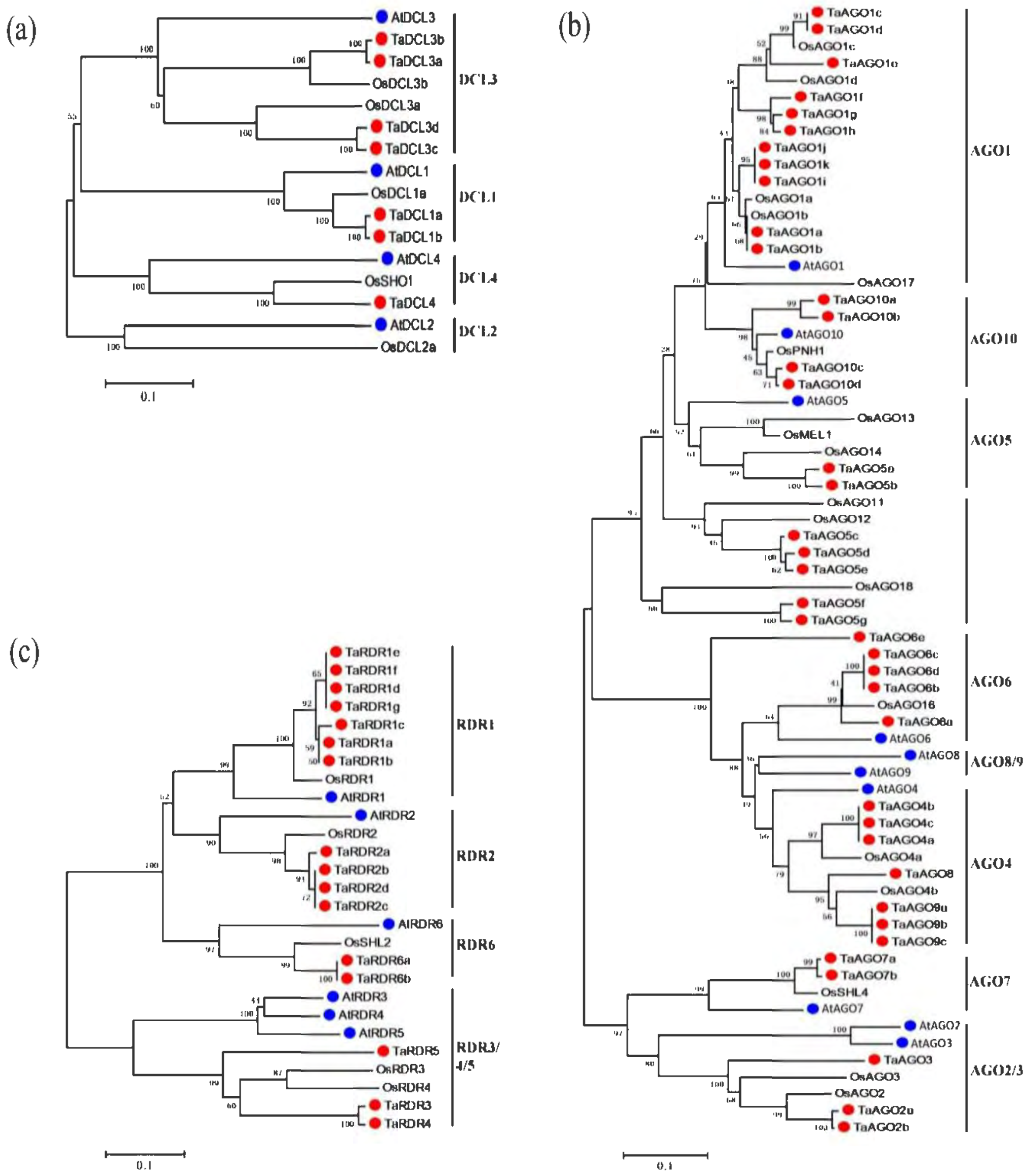


**Fig.4.2** Phylogenetic trees of *A. thaliana*, rice (*Oryza sativa*) wheat (*T. aestivum)* **(a)** *DCL* **(b)** *AGO* and **(c)** *RDR* proteins sequences. The trees were created by MEGA 7.0 software using Neighbor-Joining (NJ) method with bootstrap of 1000.*T. aestivum* proteins are indicated with a red filled circle before the protein names.

To find out the existence of conserved catalytic residues (DDH/H), which are responsible for the endonuclease property of AGO proteins involved in RNAi, the PIWI domains of all TaAGOs were aligned using CLUSTALX (**Fig.4.4**). The PIWI domain sequence alignment showed that total 15 members namely seven TaAGO1s (TaAGO1a, TaAGO1b, TaAGO1c, TaAGO1d, TaAGO1i, TaAGO1j, TaAGO1k), four TaAGO5s (TaAGO5c, TaAGO5d, TaAGO5f, TaAGO5g), two TaAGO7s (TaAGO7a, TaAGO7b), two TaAGO10s (TaAGO10c, TaAGO10d) possessed the four key active residues (DDH/H), implying that they might have endonuclease activity(Rivas et al., 2005) (**Fig.4.4**). From the remaining 24 TaAGOs, the 14 members comprise of three TaAGO1s (TaAGO1f, TaAGO1g, TaAGO1h), two AGO4s (TaAGO4a, TaAGO4c), three TaAGO5s (TaAGO5a, TaAGO5b, TaAGO5e), three TaAGO6s (TaAGO6b, TaAGO6c, TaAGO6d), three AGO9s (TaAGO9a, TaAGO9b, TaAGO9c) possessed the four residues DPD/H where the second aspartate was replace by proline and the histidine in the third position was replaced by aspartate (**Table** 4.**2**). The three members namely TaAGO4b, TaAGO6a, TaAGO6e had the same replacement of residues DPD/H in the second and third position by proline and aspartate respectively and one residue was missing of each TaAGO1e, TaAGO4b, TaAGO6a, TaAGO6e in the fourth position (**Fig. 4.4** and **Table 4.2**). The three genes TaAGO2a, TaAGO2b, TaAGO3 possessed same four residues DHD/D (**Fig. 4.4** and **Table 4.2**). The member TaAGO8 had four residues YPD/H where the first aspartate was replaced by tyrosine (**Fig. 4.4** and **Table 4.2**). TaAGO10a and TaAGO10b possessed the same residues DYD/H where the aspartate in the second position was replaced by tyrosine (**Table 4.2**).

All the 16 TaRDR proteins possessed an RdRp domain (**Fig.4.3(c)**). In addition, the proteins TaRDR1a, TaRDR1c, and TaRDR1f carried an additional domain named RNA-binding domain (RBD) of their *Arabidopsis* counterpart. The gene length of the TaRDR varied from 1707 bp for TaRDR1d and 15893bp for TaRDR2b with potentially encoding 434 and 1127 amino acids respectively (**Table 4.1**). Additionally, TaRDRs proteins were studied to identify the presence of the catalytic motif in the RdRp domain. The sequence alignment revealed that all members of TaRDR1, TaRDR2, and TaRDR6 groups shared a common DLDGD motif in the catalytic domain, whereas the proteins TaRDR3, TaRDR4, and TaRDR5 retained DFDGD motif (**Fig. 4.5**).

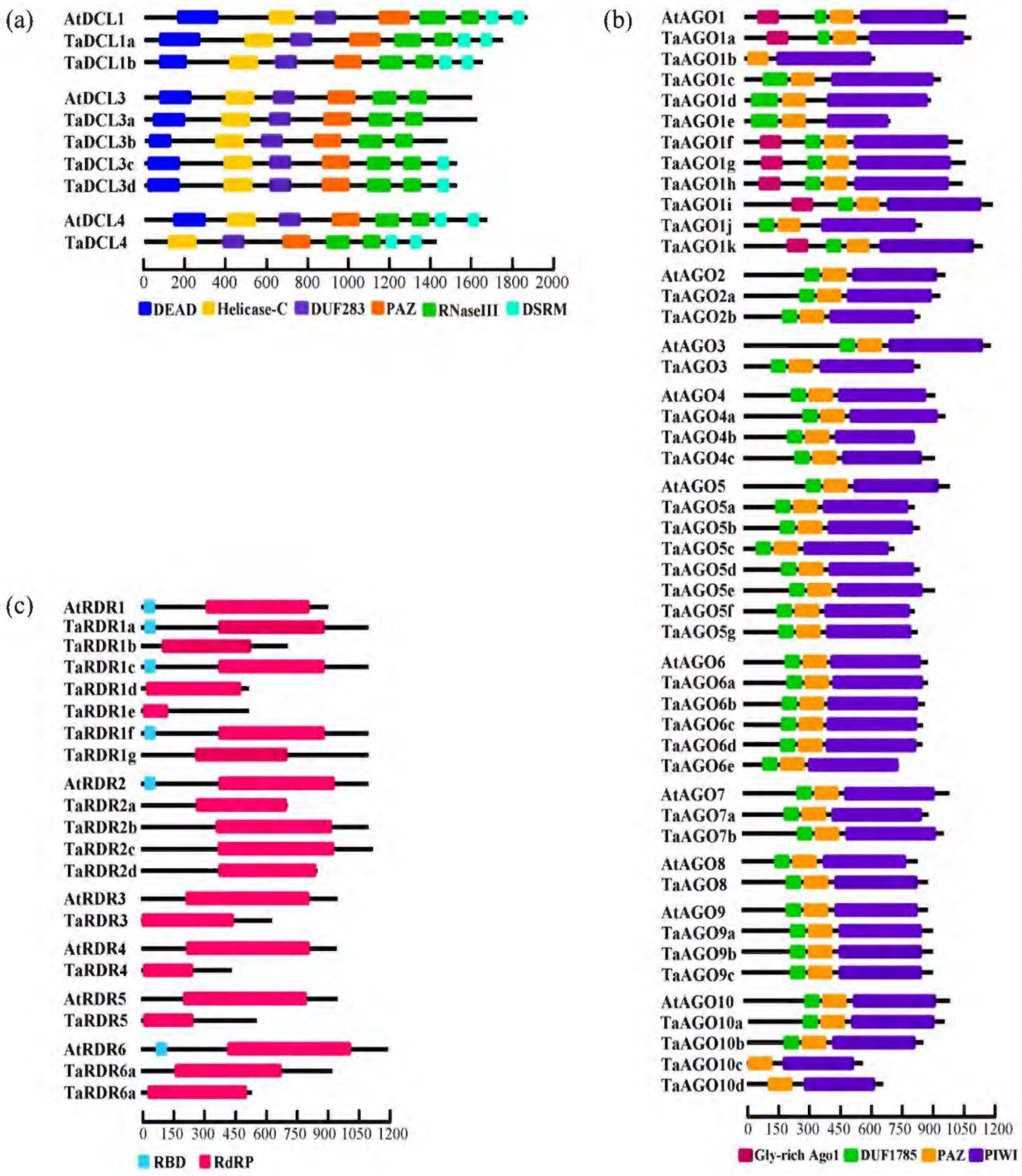
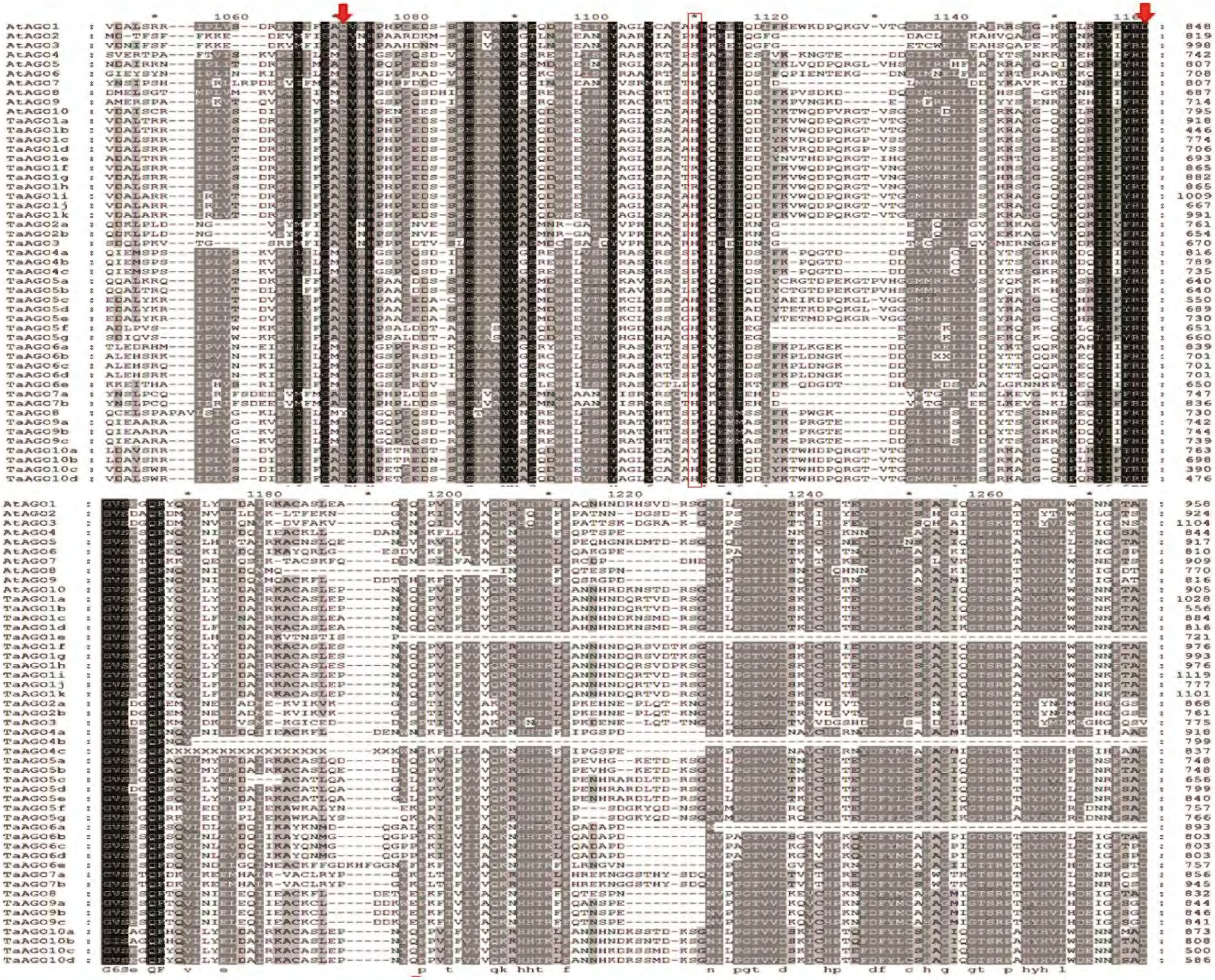


**Fig.4.3.** Protein structure and functional domain/motifs analysis of **(a)** DCL **(b)** AGO **(c)** RDR in wheat. Domains are indicated as boxes in different colors. The diagrams were drawn to scale.



Continued as Fig.4.4 in Page 127

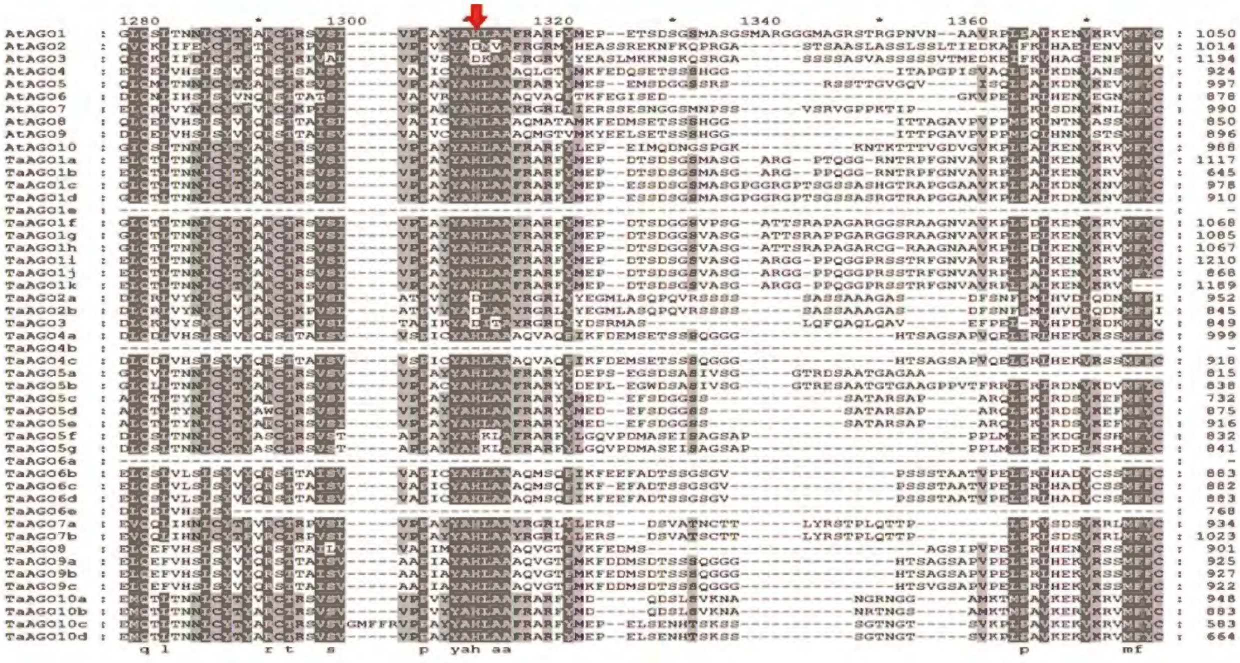


**Fig.4.4.** Alignment profile of PIWI domain amino acids of wheat AGO proteins. The protein sequences were aligned using MEGA 7.0. The conserved Asp, Asp and His (DDH) traid residues, as well as His (H) are shaded in colors. Amino acid positions corresponding to each protein are indicated at the end of each line.

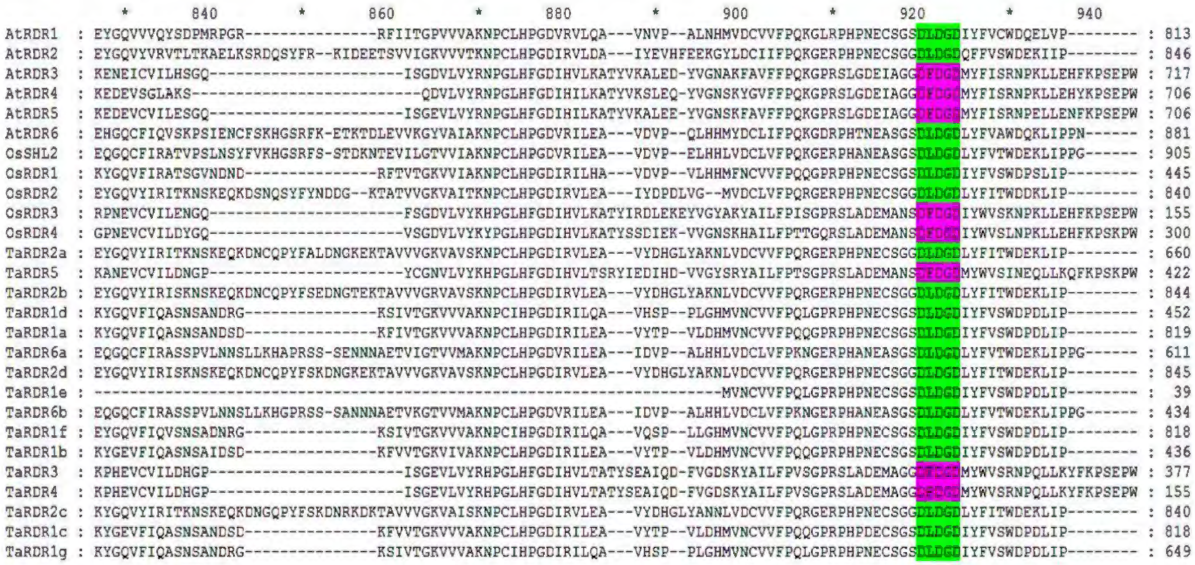


**Fig.4.5** Alignment of catalytic regions in RdRp domains of RDR proteins of *T. aestivum*, rice, and *A. thaliana*. The conserved motif DLDGD and DFDGD were colored. The alignments were performed using MEGA 7.0. A ‘’-‘’ indicates of an amino acid missing of the corresponding protein sequence. The positions corresponding to each protein are mentioned at the end of line.

**Table 4.2**. Comparison of argonaute proteins with missing catalytic residue(s) present in PIWI domain of wheat, rice[b] and Arabidopsis[b]

| Serial no. | Wheat | | Rice | | *Arabidopsis thaliana* | |
|---|---|---|---|---|---|---|
| | Argonaute (AGO) | Motif[a] | Argonaute (AGO) | Motif[b] | Argonaute (AGO) | Motif[b] |
| 1 | TaAGO1a | DDH/H | OsAGO1 | DDH/P | AGO2 | DDD/H |
| 2 | TaAGO1b | DDH/H | OsAGO2 | DDD/H | AGO3 | DDD/H |
| 3 | TaAGO1c | DDH/H | OsAGO3 | DDD/H | AGO4 | DDH/S |
| 4 | TaAGO1d | DDH/H | OsAGO4a | DDH/P | AGO6 | DDH/P |
| 5 | TaAGO1e | DDH/H | OsAGO4b | DDH/P | AGO9 | DDH/R |
| 6 | TaAGO1f | DDH/R | OsAGO11 | GDH/H | | |
| 7 | TaAGO1g | DDH/R | OsAGO13 | -DH/H | | |
| 8 | TaAGO1h | DDH/R | OsAGO15 | DDH/P | | |
| 9 | TaAGO1i | DDH/H | OsAGO16 | DDH/P | | |
| 10 | TaAGO1j | DDH/H | OsAGO17 | HDR/C | | |
| 11 | TaAGO1k | DDH/H | OsAGO18 | DDH/S | | |
| 12 | TaAGO2a | DDD/H | | | | |
| 13 | TaAGO2b | DDD/H | | | | |
| 14 | TaAGO3 | DDD/H | | | | |
| 15 | TaAGO4a | DDH/P | | | | |
| 16 | TaAGO4b | DDH/P | | | | |
| 17 | TaAGO4c | DDH/P | | | | |
| 18 | TaAGO5a | DDH/P | | | | |

**Table 4.2**.Comparision of argonaute proteins with missing catalytic residue(s) present in PIWI domain of wheat, rice[b] and Arabidopsis[b](continue)

| **Serial no.** | **Wheat** | | **Rice** | | ***Arabidopsis thaliana*** | |
|---|---|---|---|---|---|---|
| | Argonaute (AGO) | Motif[a] | Argonaute (AGO) | Motif[b] | Argonaute (AGO) | Motif[b] |
| 19 | TaAGO5b | DDH/P | | | | |
| 20 | TaAGO5c | DDH/H | | | | |
| 21 | TaAGO5d | DDH/H | | | | |
| 22 | TaAGO5e | DDH/P | | | | |
| 23 | TaAGO5f | DDH/H | | | | |
| 24 | TaAGO5g | DDH/H | | | | |
| 25 | TaAGO6a | DDH/P | | | | |
| 26 | TaAGO6b | DDH/P | | | | |
| 27 | TaAGO6c | DDH/P | | | | |
| 28 | TaAGO6d | DDH/P | | | | |
| 29 | TaAGO6e | DDH/P | | | | |
| 30 | TaAGO7a | DDH/H | | | | |
| 31 | TaAGO7b | DDH/H | | | | |
| 32 | TaAGO8 | YDH/P | | | | |
| 33 | TaAGO9a | DDH/P | | | | |
| 34 | TaAGO9b | DDH/P | | | | |
| 35 | TaAGO9c | DDH/P | | | | |
| 36 | TaAGO10a | DDH/Y | | | | |
| 37 | TaAGO10b | DDH/Y | | | | |
| 38 | TaAGO10c | DDH/H | | | | |
| 39 | TaAGO10d | DDH/H | | | | |

[a]Motifs are correspond to conserved D760, D845, H986/H798 of Arabidopsis AGO1; D(aspartate), H(Histidine), K(Lysine), A(Alanine), P(Proline), Q(Glutamine), G(Glycine), R(Arginine), C(Cysteine), S(Serine); '–' represents the missing catalytic residue [b] Reviewed in Kapoor *et al* (2008)

### 4.3.4 Assembly of Exon-Intron of DCL, AGO and RDR Genes in *T. aestivum*

The exon-intron organization of DCL, AGO and RDR genes was studied for more broad understanding regarding their probable structural progression. Our analyses for all three-gene families exhibited that intron number was usually conserved in members of the same groups though varied considerably in different groups of the same family. The intron number varied from 17 to 25 among TaDCLs genes and the members of the group TaDCL3 possessed 7~8 additional introns than the members of group of TaDCL1 and TaDCL4 (**Table 4.1 and Fig. 4.6(a)).**The intron numbers varied between 14 and 22 among all members of TaAGO genes except the members of the groups TaAGO2, TaAGO3, TaAGO7. The members of these groups contained as few as 1~2 introns which were significantly different from the other TaAGO groups (**Table 4.1 and Fig.4.6(b))**. On the other hand, there was no considerable variation in intron numbers among the members of the TaRDR groups except TaRDR3 and TaRDR4. All most for all members of TaRDR groups, the intron numbers varied ranging from 1~6 while TaRDR3 had 13 introns highest in number and TaRDR contained 11 introns second highest in number (**Table 4.1 and Fig.4.6(c)).** It was observed that the exon-intron structure of DCL, AGO and RDR genes was very similar within members of the same groups though substantially different through various groups of the same family indicate that these gene families particularly AGO and RDR families have experienced recurrent gene replication and recombination during evolution.

### 4.3.5 Cis-acting Regulatory Elements in the Promoters of TaDCL Genes in *T. aestivum*

Gene transcription in plants is activated by an appropriate transcription factor that binds to a specific motif called *cis*-element (Liu et al., 2013c). Some *cis*-elements that are involved in stress responses have been well identified in plants, including dehydration and cold response (DRE/CRT) (Sakuma et al., 2002), Ethylene Response Factors (ERFs) binding site (GCC box) (Cheng et al., 2013), ABA responsive element (ABRE) (Osakabe et al., 2014), ARFs binding site (AuxRE) (Ulmasov et al., 1997), SA-responsive promoter element (SARE) (Pieterse and Van Loon, 2004), environmental signal response (G-box)(Williams, 1992), WRKY binding site (W-box) (Chen et al., 2012), CAMTA binding site (CG-box) (Yang and Poovaiah, 2002), and sulfur-responsive element (SURE) (Maruyama-Nakashita et al., 2005). We scanned the ~1,500 bp upstream promoter regions of TaDCL genes with nine stress-related cis-elements to obtain preliminary clues on how the TaDCL genes respond to stress stimuli. The results showed that there were various stress/stimulus response-related *cis*-acting elements in the promoter of TaDCL genes (**Table 4.3**).

**Table 4.3**. Predicted stress response-related *cis*-elements in the 1.5 kb sequence upstream of ATG of the *TaDCL* genes

| Gene Name | DRE /CRT | GCC box | ABRE | AuxRE | SARE | G-box | W-box | CG-box | SURE |
|---|---|---|---|---|---|---|---|---|---|
| *TaDCL1a* | 1 | 4 | 1 | 1 | 0 | 1 | 1 | 3 | 9 |
| *TaDCL1b* | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 3 |
| *TaDCL3a* | 2 | 1 | 2 | 1 | 1 | 1 | 0 | 3 | 4 |
| *TaDCL3b* | 3 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 1 |
| *TaDCL3c* | 0 | 0 | 2 | 0 | 1 | 1 | 4 | 3 | 2 |
| *TaDCL3d* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 10 |
| *TaDCL4* | 1 | 0 | 3 | 1 | 1 | 3 | 2 | 0 | 2 |

Interestingly, the promoters of TaDCL1a, TaDCL3a, and TaDCL4 contained a large number of stress-related cis-elements. Totally, one DRE/CRT, four GCC box, one ABRE, one AuxRE, one G-box, one W-box, three CG-box and nine SURE were contained in the promoter of TaDCL1a; two DRE/CRT, one GCC box, two ABRE, one AuxRE, one SARE, one G-box, three CG-boxes and four SURE were located on the promoter of TaDCL3a; one DRE/CRT, three ABRE, one AuxRE, one SARE, three G-box, two W-box and two SURE included in the promoter of TaDCL4. Every TaDCL gene contained at least one type of stress response-related *cis*-element, but the type of *cis*-element(s) in TaDCL genes was distinguishable (**Table 4.3**). Collectively, the stress-responsive *cis*-element analyses indicated that the TaDCLs are likely to be involved in plant response to various stresses and hormone signals.

### 4.3.6 GO Analysis of DCL, AGO and RDR Genes in TaDCL Genes in *T. aestivum*

GO enrichment examination of DCL, AGO and RDR genes was inferred for biological processes, molecular functions, and cellular components. RNAi genes have close involvement with post-transcriptional gene silencing (PTGS) in plants (Fire et al., 1998). The GO results showed that a pool of biological, molecular and cellular functions (pathways) are associated with the RNA silencing machinery genes in wheat. AGO proteins are assumed to be related to cleavage named endonucleolytic activities(Lingel and Izaurralde, 2004). Biological process results predicted from GO analysis implied that 24-26 genes are related to metabolic functions or pathways in wheat (**Table A4.3)**. The predicted metabolic functions are macromolecule metabolic process (GO: 0043170, *P* =2.30e-08), cellular metabolic process (GO: 0044237, *P* = 4.20e-06), organic substance metabolic process (GO: 0071704, *P* = 3.20e-05), primary metabolic process (GO: 0044238,*P* = 9.10e-05), cellular macromolecule metabolic process (GO: 0044260, *P* = 2.00e-08),nitrogen compound metabolic process ( GO:0006807, *P* =1.50e-10),cellular nitrogen compound metabolic process (GO:0034641, *P* = 2.60e-11),organic cyclic compound metabolic process (GO:1901360, *P* = 2.00e-13), cellular aromatic compound metabolic process (GO:0006725, *P*= 1.30e-13), heterocycle metabolic process (GO:0046483, *P*=1.10e-13), nucleobase-containing compound metabolic process (GO:0006139,*P* = 1.90e-14), nucleic acid metabolic process (GO:0090304, *P* = 1.60e-16), RNA metabolic process (GO:0016070, *P* = 1.30e-18). GO analysis of biological process also indicated that each nine different biosynthetic pathways are regulated by 16 RNAi genes such as RNA biosynthetic process (GO 0032774, *P* = 1.30e-10), nucleobase-containing compound biosynthetic process (GO 0034654, *P* = 1.60e-09), aromatic compound biosynthetic process(GO:0019438, *P* = 6.60e-09), heterocycle biosynthetic process(GO:0018130, *P* = 6.90e-09), organic cyclic compound biosynthetic process(GO:1901362, *P* = 1.20e-08), macromolecule biosynthetic process (GO:0009059, *P* = 1.70e-06), cellular nitrogen compound biosynthetic process (GO:0044271, *P* = 2.40e-06), cellular biosynthetic process (GO:0044249,*P*=7.00e-05), biosynthetic process (GO:0009058, *P* = 0.00015). Few genes (GO: 0035194, *P* = 8.10e-25 and GO: 0016441, *P* = 5.00e-24) are associated with post-transcriptional gene silencing mechanism (**Table A4.3**). The dsRNA fragmentation (GO: 0031050, *P* = 2.80e-23), response to dsRNA (GO:

0043331, *P* = 2.80e-23) and cellular response to dsRNA (GO: 0071359, *P* = 2.80e-23) pathway each is maintained by some 10 RNA silencing genes. It is however observed that 11 and 9 genes are activated in response to chemical (GO 0042221, *P* = 6.70e-08) and virus (GO 0009615, *P* = 3.40E-20) response mechanism respectively in wheat. On the other hand GO analysis of molecular pathways(**Table A4.3**) suggested that 42-47 genes are responsible for different molecule binding functions such as heterocyclic compound (GO: 1901363, *P* = 1.80e-10), organic cyclic (GO: 0097159, *P* = 1.80e-10) and nucleic acid (GO:0003676, *P* = 5.00e-26) binding while 13 genes are involved in protein binding (GO:0005515, *P* = 2.80e-17). RNA polymerase activity (GO: 0034062, *P* = 6.60e-16) and transferase activity (transferring phosphorus-containing groups) (GO: 0016772, *P* = 0.00437) are associated each with 13 RNAi genes. The identified TaAGOs contained the PIWI and PAZ domain, which play significant role in making complex with target RNA to specific nucleic acid position. The PAZ domain has a nucleic acid-binding fold that promotes the domain to bind to the specific position of the nucleic acids [43, 44]. Additionally, the molecular functions such as endoribonuclease activity (GO: 0004521, *P* = 3.90e-13), ribonuclease activity (GO: 0004540, *P* = 2.20e-12), endonuclease activity (GO: 0004519, *P* = 1.20e-10), nuclease activity (GO: 0004518, *P* = 6.70e-09) and hydrolase activity (acting on ester bonds) (GO: 0016788, *P* = 3.00e-04) each maintained by eight RNA silencing genes (**Table A4.3**). However only a few (9-5) cellular component pathways are involved in gene silencing mechanisms in wheat (**Table A4.3**). Red color box however implies that the predicted pathways (GO terms) have higher statistical significance in RNAi mechanism in wheat and inside the box: GO terms and GO description are mentioned (**Fig.A4.1, A4.2, and A4.3**). The Venn diagram illustrates the number of predicted biological, molecular, and cellular pathways shared by the TaDCL, TaAGO and TaRDR genes (**Fig.4.7(d)**). It is observed that there are 91 biological functions are common among DCL, AGO and RDR genes to contribute wheat RNAi regulation (**Fig.4.7(d)**). However, only four molecular functions are shared by RNAi genes but seven molecular pathways are shared between TaDCL and TaAGO genes (**Fig.4.7(e)**). On the other hand, predicted all six cellular functions are shared by the three RNA silencing machinery genes (**Fig.4.7(f)**).The heat maps (**Fig.4.7(a), 4.7(b)** and **4.7(c)**) have been produced using *P*-values (-log10(*P*)) of the GO terms for biological, molecular and cellular component pathways involved in RNAi regulation using open source software R-3.5.2 (R Core Team 2018). Heat map shaded with two colors across the three maps implies the present (statistically significant) or absent (statistically insignificant) of predicted pathways for corresponding RNAi gene regulation in wheat.

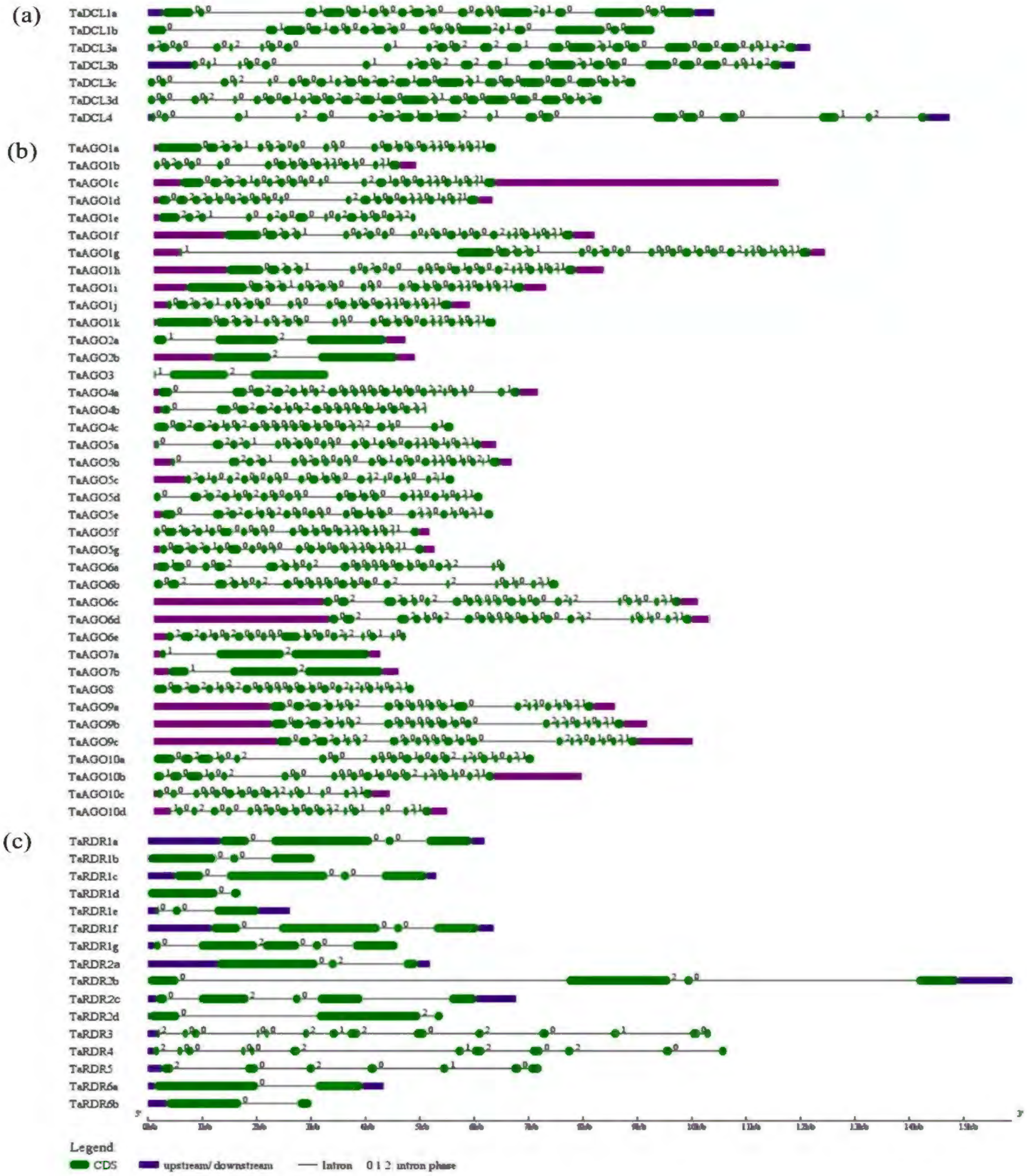


**Fig.4.6** Exon-intron structure of *T. aestivum* and *A. thaliana* **(a)** DCL, **(b)** AGO and **(c)** RDR. Exons (green box), intron (black lines), and intron phases (0, 1, and 2) are mentioned. The gene structure was estimated using the online GSDS1.0 (http:// gsds.cbi.pku.edu.cn/) by comparing their full-length coding sequences (CDS) with their corresponding genomic sequence.

### 4.3.7 Subcellular Location of DCL, AGO and RDR Genes in *T. aestivum*

A number of RNA silencing genes were found locating in more than one cellular locations (**Fig. 4.8(b), Fig. 4.8(c)**). It is however observed that most of the RNA silencing genes were found to be located in cytosol (DCL 71.4 %, AGO 87.2 % and RDR 87.5 %) followed by plastid (DCL 14.3 %, AGO 33.3 % and RDR 31.2 %). A few of the AGO (20.5 %) and RDR (12.5 %) genes were located in mitochondria but no DCL protein was found in mitochondria. On the other hand, ER, golgi and peroxisome contain no RNA silencing genes (**Fig. 4.8(d) and TableA4.2**).

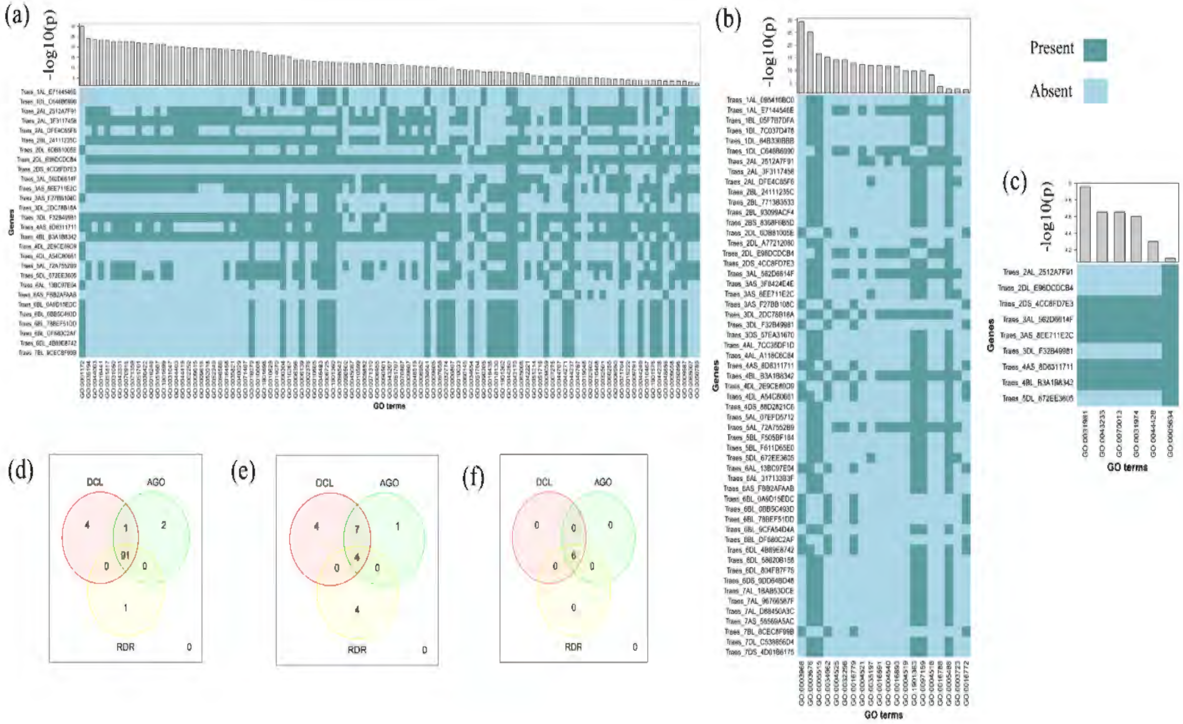


**Fig.4.7** The heat map for the predicted GO terms corresponding to the RNAi genes are illustrated for **(a)** biological process **(b)** molecular function **(c)** cellular components whether the genes are related (Present) or not (Absent). Histograms corresponding to each GO terms represent the p-value (-log (p-value)) showed in the opposite site of GO terms in the heat map. The Venn diagrams display the common GO terms shared by three gene families in terms of **(d)** biological process **(e)** molecular functions **(f)** cell components.

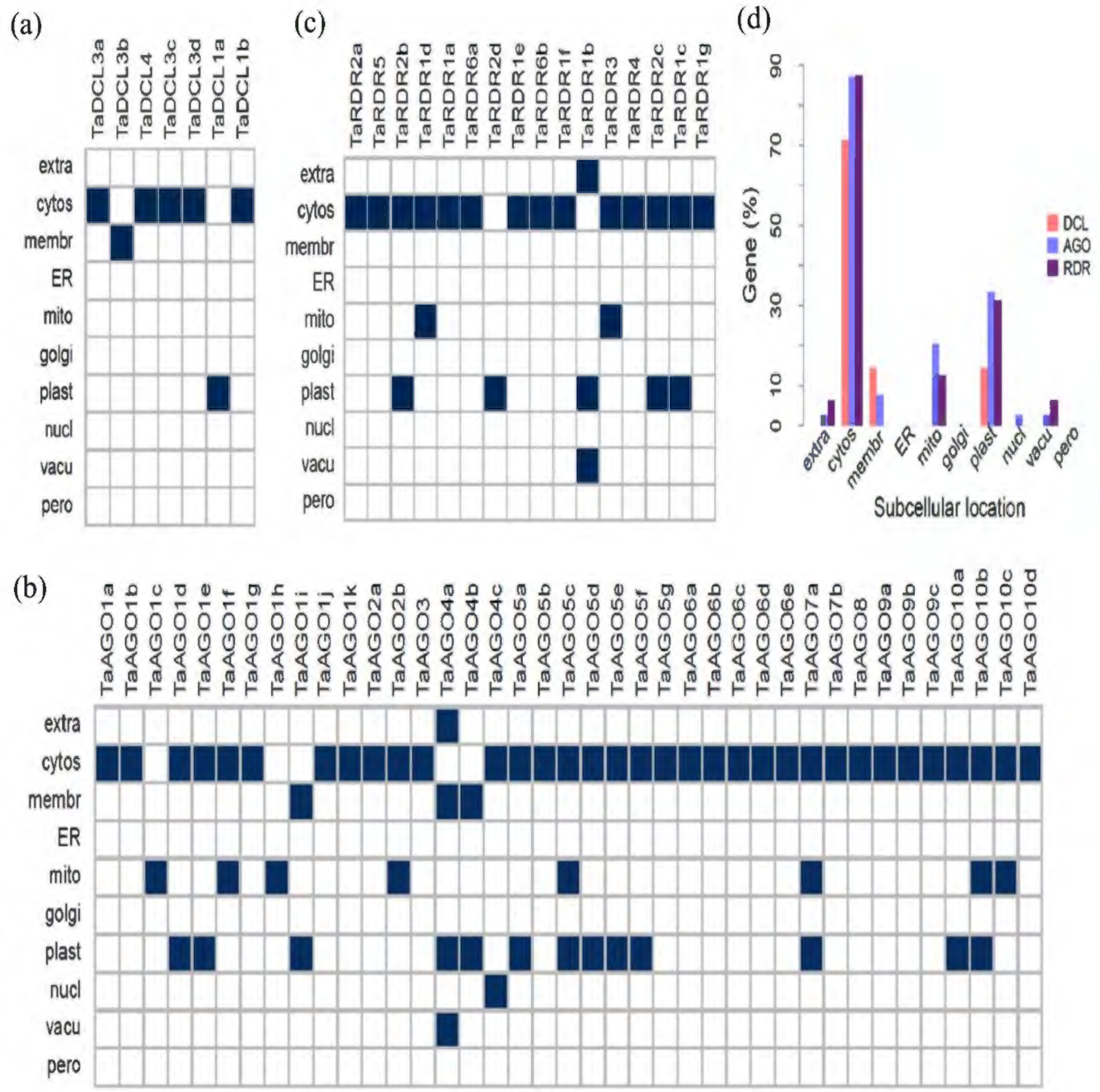


**Fig.4.8** Prediction of subcellular locations: cytosol (cytos), endoplasmic reticulum (ER), extracellular (extra), golgi apparatus (golgi), membrane (membr), mitochondria (mito), nuclear (nucl), peroxisome (pero), plastid (plast) and vacuole (vacu) for each of the **(a)** DCL, **(b)** AGO and **(c)** RDR gene in *T. aestivum* with the help of PSI and R-3.5.2.

### 4.3.8 Expression Pattern of TaDCL Genes in Leaves and Roots as well as in Response to Drought

To obtain evidence for the likely functions of the TaDCL genes, their expression profile in leaves and roots of two-week aged plants were analyzed by qRT-PCR. The results showed that two out of seven TaDCL genes (TaDCL3a and TaDCL3b) were expressed highly in roots than that of leaves indicating root specific function of TaDC3a and TaDCL3b (**Fig.**

**4.9(a)**). To explore the role of TaDCL genes in response to drought stress tolerance, the expression patterns of TaDCLs were measured at 3-days post treatment against drought. The results showed that drought treatment positively induced the expression of all the seven TaDCL candidate genes. Among them, TaDCL3b and TaDCL4 showed the significant change in expression in response to drought signifying these genes might play a vital role against drought stress in wheat (**Fig. 4.9(b)**).

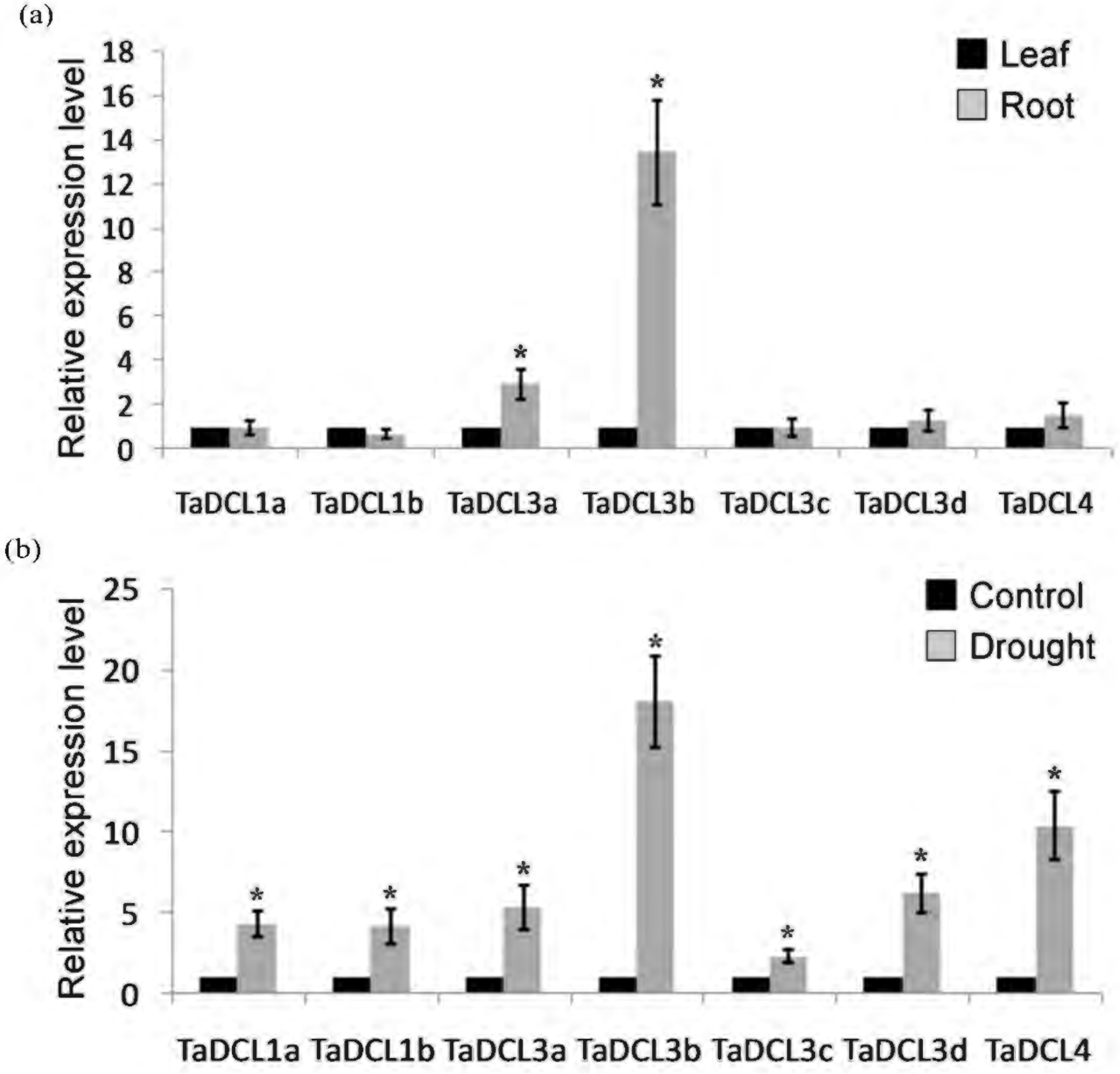


**Fig.4.9** qRT-PCR expression of TaDCL genes in **(a)** leaves and roots and **(b)** under drought stress. Ta18s was considered as the reference gene, and three biological replicates were accomplished for the experiments. Error bars indicate the standard error. Asterisks indicate the significance difference between control and treatment at P < 0.05.

## 4.4 Discussion

RNA silencing in organisms is a very useful molecular mechanism. It performs a wide-ranging biological activity in plants and animals by means of inhibiting transcript accumulation of genes necessary to accomplish these activities (Kapoor et al., 2008; Qian et al., 2011). DCLs, AGOs, and RDRs genes however contribute central roles in these processes. In this study, three gene families encoding the key elements of RNA silencing in wheat (*T. aestivum*) were identified. Phylogenetic analysis showed the evolutionary relationship among the members of these three gene families and gene multiplicity in wheat. Other structural characterization provided the basic functional genomic information for the gene families that would help as substantial source for further biological experimentation and broad investigation against different biotic and abiotic stressors as well as plant growth and development as a whole.

### 4.4.1 Wheat (*T. aestivum*) Dicer-Like (DCL) Genes

It is well studied that DCL, AGO and RDR genes play major role in RNA silencing pathway. Our analysis provides that wheat (*T.aestivum*) contain seven TaDCL, 39 TaAGO and 16 TaRDR protein-coding genes. This replicated gene numbers are considerably higher than those in *A. thaliana*. Dicer-like (DCL) proteins are vital machineries in miRNA and siRNA biogenesis mechanism and function to transform double-stranded RNAs into matured small RNAs(Qian et al., 2011). Comparative to animals and fungi, the remarkable extension of DCL protein elements in monocots and dicots may reveal the placement of RNA silencing in antiviral resistance(Deleris et al., 2006; Finnegan and Matzke, 2003). For instance, four DCL proteins are encoded in *A. thaliana* and seven putative DCL proteins were identified in wheat. Genomic investigation in Arabidopsis presented both specified and corresponding functions of DCL proteins(Deleris et al., 2006; Fahlgren et al., 2006). AtDCL1 and AtDCL3 activities overlap to stimulate Arabidopsis flowering (Henderson et al., 2006). It was identified that AtDCL2 and AtDCL4 take part in functional overlap in antiviral resistance (Deleris et al., 2006). Similarly, AtDCL2, AtDCL3, and AtDCL4 reveal overlying roles in siRNA generation (Henderson et al., 2006). Hardly is it known about the functions of DCL proteins in wheat. Phylogenetic demonstration however provided that the functional divergence of DCLs happened earlier the expansion of monocots and dicots nearly 200 million of years back (Henderson et al., 2006; Margis et al., 2006).

In this study, four subfamilies of DCL genes were identified in wheat. Phylogenetic study of DCL proteins from wheat, rice, and Arabidopsis showed that four sub-families of DCL1, DCL2, DCL3, DCL4 homologs were different. The members of the group TaDCL1s (TaDCL1a, TaDCL1b) were analogous to their Arabidopsis AtDCL1 and rice OsDCL1a counterpart. Surprisingly, there was no TaDCL2 member identified similar to AtDCL2 and OsDCL2. Four members of the TaDCL3s (TaDCL3a-TaDCL3d) were identical to Arabidopsis AtDCL3 and rice OsDCL3s. Conserved domain analyses using Pfam and NCBI further implied that there was all most similar structural identification in DCL families from wheat, rice, and Arabidopsis. In consideration of these resemblances, it is recommended that a similar evolutionary association in the functional divergence of the wheat DCL gene families and rice as well as Arabidopsis. Barely are there any reports regarding the biochemical or genetic analyses of DCL genes in wheat exist.

It was revealed from the cis-acting analyses of promoters of the TaDCL genes that though there was a variation among the TaDCL transcripts in terms of possessing cis-acting elements with the exception that the promoters of TaDCL1a, TaDCL3a, and TaDCL4 contained a vast quantity of stress-related *cis*-elements. It is however observed that every TaDCL gene possessed at least one stress response-related *cis*-element (**Table 4.3**). Eventually the stress-responsive *cis*-element studies pointed out that the TaDCLs in wheat are probable to be associated in plant reaction to many stresses and hormone signals. Gene ontology classification or functional annotation showed that a significant number of biological and molecular functions or pathways are linked to maintaining DCL genes in wheat. Among those, metabolic, biosynthetic, post-transcriptional gene silencing process, molecular binding, chemical response, immune system and virus responses are common pathways ($P < 0.01$) in RNAi mechanism (**Table A4.3**). Cytosol is the place of maximum metabolism in plants and most of the proteins in the cell are located in cytosol(Kholodenko, 2003; Ohlrogge et al., 1979). Five out of seven DCL proteins are localized in cytosol of *T. aestivum*. Previous study demonstrated that RNA silencing genes plays a central role in plant tolerance to abiotic stress (Bai et al., 2012; Qin et al., 2018). Dicer enzymes work to process double-stranded RNAs into small RNA of diverse size that initiate the RNA silencing pathway(Carrington and Ambros, 2003; Chapman and Carrington, 2007; Qian et al., 2011). Therefore, the transcript levels of TaDCLs were measured. TaDCL3a and TaDCL3b showed a higher expression level in roots than leaves **(Fig. 4.9a)**. It implies that TaDCL3a and CaDCL3b might also be involved in wheat root development. Moreover, the transcripts of TaDCL3b and TaDCL4 were considerably induced upon drought treatment **(Fig. 4.9b)**. This finding suggests that the genesTaDCL3 and TaDCL4 are involved in wheat root development and drought stress tolerance.

### 4.4.2 Wheat (*T. aestivum*) Argonaute (AGO) Genes

Argonautes (AGOs) are another key RNA silencing machinery genes first recognized in plants and elements are commonly well-defined by the existence of PAZ and PIWI domains. AGO gene family contains multiple copies of genes in species and organisms and are much conserved. Eukaryotic organisms most possess AGO multi-gene families, the components of which have particular biological function, as discovered by a range of mutant phenotypes(Carmell et al., 2002). *A. thaliana* possesses 10 AGO genes two of which are identified to act explicitly in different forms of RNA silencing. Hence, it is possible that functional divergence of RNA silencing is related to dissimilarity between AGO family members, resembling that of animals. AtAGO1 is related to the miRNA and transgene-silencing pathways (Fagard et al., 2000; Vaucheret et al., 2004), and AtAGO4 with endogenous siRNAs, which contribute in epigenetic silencing (Fagard et al., 2000; Vaucheret et al., 2004). Furthermore, AtAGO7 functions in the evolution from young to mature plant growing phases (Hunter et al., 2016) and meristem maintenance(Lynn et al., 1999; Moussian et al., 1998), respectively. Limited investigation has been completed about the functional divergence of monocot AGO genes to date. Nearly double and three times higher in numbers of AGO genes were identified in wheat compare to rice and *A. thaliana* respectively. It is suggested that the functional expansion of AGOs happened afterward the divergence of monocots and dicots almost 200 million years before(Qian et al., 2011). Besides, the phylogenetic analysis and Pfam as well as NCBI conserved domain comparisons of AGO genes from wheat, rice, and *A. thaliana* implied that these gene members hold maximum relationships between the species studied. These findings are abundant source for further investigation into the functional change of the wheat (*T. aestivum*) AGO gene group.

AGO genes are the members of particularly fundamental RNA-binding proteins which possess PAZ and PIWI domains. AGO proteins go through endonuclease action which is initially linked to the PIWI domain, that holds four well-preserved metal-chelating amino acid residues (DDH/H). It is worth mentioning that various AGO proteins are endonucleolytically inactive, while the catalytic residues are well-preserved (Qian et al., 2011). For example, in Arabidopsis and rice 5 and 11 AGO genes, which do not code for the conserved catalytic residues, respectively (Qian et al., 2011). Similarly, in wheat PIWI domains there are as many as 24 AGO genes lacked the conserved catalytic residues. The nonexistence of conserved catalytic residues could result in loss of function of target RNA treating by endonucleolytic cleavage in these proteins (Kapoor et al., 2008).

AGO genes also conserve a large number of biological and molecular pathways ($P < 0.05$) in RNAi technique similar to DCL in wheat. AGO proteins in cytosol are also highly linked to perform multiple cell process. Theses TaAGO proteins may however have essential functions in processes as different as embryonic development, cell differentiation and transposon silencing (Höck and Meister, 2008). PSI investigation showed that almost all TaAGO proteins have substantial molecular activity in cytosol whereas only one-third TaAGOs proteins are located in plastid.

### 4.4.3 Wheat (*T. aestivum*) RNA-dependent RNA Polymerase (RDR) Genes

RNA-dependent RNA polymerases generally increase RNA-interference (RNAi) silencing signals through the generation of additional aberrant RNA population (Sijen et al., 2001). Astier Manifacier and Cornuet (1971) originally described the activities of RDR in Chinese cabbage(Qian et al., 2011).To date, a number of RDR gene paralogs have been recognized in different plants such as *Arabidopsis thaliana* , rice (*Oryza sativa* ) ,tomato (*Solanum lycopersicum*) , maize (*Zea mays*), grapevine (*Vitis vinifera*), cucumber (*Cucumis sativus*), *Brassica napus*, pepper(*Capsicum Annuum* L.), *Nicotiana benthamiana* (Cao et al., 2016; Qin et al., 2018). However, three RDR genes in Arabidopsis make functions in distinctive and overlapping biological manners for example viral defense, chromatin silencing (Kapoor et al., 2008). Among these RDR genes, AtRDR1 is stimulated by salicylic acid (SA) or viral contamination in several other plants (Diaz-Pendon et al., 2007; Qi et al., 2009; Yu et al., 2003). Moreover, AtRDR2 plays role in repressive of chromatin modifications on some particular transgenes, endogenous genes with the production of 24 nucleotide interfering RNA molecules (Matzke et al., 2007; Zaratiegui et al., 2007). AtRDR6 magnifies some anomalous RNAs produced from transgenes or reversed replications to start humiliation of opposite of RNA species(Luo and Chen, 2007). In this work, 16 RDR transcripts were identified in wheat. Like rice and Arabidopsis, these genes were grouped into four different subclasses. Phylogenetic relationships and Pfam as well as NCBI comparisons showed that the presence of wheat RDR genes analogous RDR orthologs in each sub-class from rice and Arabidopsis. Finally, these outcomes implied that the RDR genes of wheat, rice, and Arabidopsis diverged from the similar common predecessor and hence accomplished same activities of the three taxa. GO analysis showed that wheat RDR transcripts are linked to few biological and four molecular functional activities at gene silencing regulation. Similar results were found for cytosol in *T. aestivum* that the maximum number of RDR proteins is located in that molecular organ. 14 TaRDR proteins out of 16 TaRDR are found in cytosol in *T. aestivum*.

## 4.5 Summary of the Chapter

DCL, AGO and RDR are called RNA silencing machinery genes play significant role in the regulation of gene expression through the generation of small RNA (sRNA) molecules in plants. These genes act in the way of RNA interference (RNAi) pathway by restricting the transcript buildup in cells. The wheat genome possessed seven DCL, 39 AGO and 16 RDR genes. The aim of this study was genome-wide identification and characterization of these genes and analyses of expression pattern of seven TaDCLs. The phylogenetic analysis provided that all subfamilies of these three gene sets maintain their evolutionary relationships similar to their rice and Arabidopsis counterpart. First subfamily of DCL, AGO and RDR possessed the multiple copies of genes higher than that of corresponding rice and Arabidopsis DCL, AGO and RDR genes. Conserved domain structure analysis suggested that these genes were also contained consistent domain structure similar to rice and Arabidopsis. Although there was a difference in possessing *cis*-acting elements of the promoter regions of TaDCL genes but the promoters of TaDCL1a, TaDCL3a, and TaDCL4 hold several number of stress-related *cis*-elements. GO investigation identified different metabolic process, biosynthetic process, molecular binding such as RNA, nucleic acid, protein binding, posttranscriptional, transferase and nuclease activities are significantly connected to RNAi gene regulation in wheat. Cytosol is however found to be the key molecular component that possesses the maximum number of wheat RNA silencing proteins. Expression study implied that TaDCL3 and TaDCL4 genes are expected to play distinct roles in development and drought stress tolerance. This work provides the important indication of DCL, AGO and RDR genes evolutionary resemblances of their rice and Arabidopsis counterpart. These results may however provide valuable sources for further biological implementation and justification to draw more specific conclusion regarding any particular gene(s) and its domain of activity against different biotic and abiotic stresses as well as growth and improvement in plants and animals.

# CHAPTER FIVE

## ROBUST STATISTICAL APPROACH FOR GWAS TO IDENTIFY BIOMARKER SNPs

# ROBUSTIFICATION OF LINEAR MIXED MODEL USING OUTLIER MODIFICATION RULE AND ITS APPLICATION TO IDENTIFY IMPORTANT SNPs INFLUENCING RICE FLOWERING TIME

## 5.1 Introduction

One of the major challenges in recent research in genetics is to explore the genetic biomarkers or factors, which are associated with complex traits of living organisms. Trait differences in living organisms are importantly related to genetic molecular variations in genes. These variations are observed largely at physiological, developmental, and morphological stages. Identification of genetic basis such as causal genetic variants for such distinction in phenotypic traits is identifiable at single nucleotide polymorphism (SNP) levels. The techniques to explore the SNP contribution in phenotypic variation are termed as Genome-Wide Association Studies (GWAS). SNPs however are commonly examined for association study across the whole genome with the trait of interest. The important SNPs identified by GWAS can be used for treatment and prevention of certain diseases or complex traits. 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 is the main concerning issue when extensive genome-wide association analysis with numerous subjects is in consideration (Li and Yu, 2008; Liu et al., 2013a; 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., 2013a). Due to higher false discovery rate (FDR), it is imperative to correct the observed population stratification in GWAS (Campbell et al., 2005; Liu et al., 2013a).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., 2002), 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., 2013a). 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(SA) analysis technique suggests locating the samples to discrete subpopulation clusters and then collecting evidence of association within each cluster (Pritchard et al., 2002). The SA method is useful for small datasets(http://pritch.bsd.uchicago.edu/ software/structure2_1.html) (Liu et al., 2013a). 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 Analysis(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 genome wide analysis (GWA) (Liu et al., 2013a). These top PCs are selected using Eigenstrat (Forster et al., 2019) 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., 2013a; 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., 2013c).The results obtained from PCA adjustment may be misleading too if there are outliers (Liu et al., 2013a). Outlying data were introduced at genotypic level to check the performance of the robust PCA approach (Liu et al., 2013a).

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).

GWAS results based on earlier methods could be misdealing in terms of false discovery rate (FDR) and statistical power if few phenotypic observations are contaminated by outliers. Recently, an improvement in PCA technique was made to overcome the limitation of analysis in presence of outliers in GWA mapping(Liu et al., 2013a). In recent times in 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. Their approaches have been executed in software programs TASSEL (Yu et al., 2006) (Yu et al., 2006), EMMA(Hyun et al., 2008), EMMAX (Kang et al., 2010), rrBLUP (Endelman, 2011), GAPIT(Lipka et al., 2012).

While LMM performs significantly in detecting a causal genetic variants for phenotypic trait of interest in terms of computational efficiency and consistency of the results such as higher statistical power and lower false discovery rate (FDR) but there is hardly any investigation regarding the performance evaluation of the linear mixed model on GWAS in presence of outliers in the phenotypic trait. In this work, an attempt has been made to robustify the LMM by modifying the outlying observations using minimum $\beta$-divergence method (Mollah et al., 2007; Nurul Haque Mollah et al., 2010) from the dataset. The performance of our approach has been investigated using simulated and real rice crop dataset related to flowering time in terms of power and FDR in presence of phenotypic outliers.

## 5.2 Materials and Methods

### 5.2.1 Proposed Method

If we consider that, there are *m* genotypes with *n* measurements of a phenotype. For genome-wide association studies the mixed linear model approach (LMM) is widely used. Efficient mixed-model association (EMMA) (Hyun et al., 2008) is such a model generally expressed as-

$$\boldsymbol{y} = \boldsymbol{X}\gamma + \boldsymbol{Z}\boldsymbol{u} + \varepsilon \qquad (5.1)$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_n)^{/}$ is the $n \times 1$ vector of phenotypic observations, and $\boldsymbol{X} = (x_{ij})$ is an $n \times q$ matrix of fixed effects including mean, SNPs and other confounding variables. $\gamma$ is a $q \times 1$ vector representing coefficients of the fixed effects. $\boldsymbol{Z}$ is an $n \times m$ incidence (design) matrix mapping each phenotype to one of the $m$ genotypes. $\boldsymbol{u}$ is the vector of random polygenic effects $\boldsymbol{u} \mid N(0, \boldsymbol{K}\sigma_g^{\;2})$ ,where $\sigma_g^{\;2}$ is the polygenic variance component, and $\boldsymbol{K} = (k_{ij})$ is the $m \times m$ genomic relationship matrix with elements of pairwise relationship coefficients estimated using genotypes of 36,901 SNPs. The genomic pairwise relationship coefficient between two individuals, *j* and *t*, is defined as follows

$$k_{jt} = \frac{1}{T_\varphi} \sum_{i=1}^{T_\varphi} \frac{(x_{ij} - 2f_i)(x_{it} - 2f_i)}{2f_i(1 - f_i)} \tag{5.2}$$

where $T_\varphi$ is the total number of SNPs, $x_{ij}$ and $x_{it}$ measure the numbers (0,1,2) of the minor allele(s) for the $i^{th}$ SNP of the $j^{th}$ and $t^{th}$ individuals, respectively, and $f_i$ is the frequency of the minor allele. The overall phenotypic variance-covariance matrix can be represented as $V = \sigma_g^{\;2}\boldsymbol{ZKZ}' + \sigma_\varepsilon^{\;2} I$ where $I$ is the $n \times n$ identity matrix. $\varepsilon$ is the vector of random error effects($\varepsilon \sim N(0, I\sigma_\varepsilon^{\,2})$), where $\sigma_\varepsilon^{\,2}$ is the error variance component. The variance components for polygenic effects and errors were estimated by restricted maximum likelihood (REML) using spectral decomposition instead of iterative EM algorithm (Hyun et al., 2008). The full-likelihood function is maximized when $\beta$ is $\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}H^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}H^{-1}\boldsymbol{y}$, and the optimal variance component is $\hat{\sigma}_F^{\;2} = R/n$ for full likelihood and $\hat{\sigma}_R^{\;2} = R/(n-q)$ for restricted likelihood, where $R = (\boldsymbol{y} - \boldsymbol{X}\hat{\gamma})' H^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\gamma})$ is a function of $\delta$ as well. Where $H = \sigma^{-1}V = \boldsymbol{ZKZ} + \delta I$ is a function of $\delta$, defined as $\delta = \sigma_\varepsilon^{\;2} / \sigma_g^{\;2}$ , $\sigma = \sigma_g$ .

When the maximum likelihood (ML) or restricted maximum likelihood (REML) variance component $\hat{V} = \hat{\sigma}_g^{\;2} K + \hat{\sigma}_\varepsilon^{\;2} I$ is estimated, the classical *F*-statistic testing the null hypothesis $M\boldsymbol{\gamma} = 0$ for an arbitrary full- rank $p \times q$ matrix *M* (Kennedy et al., 1992; Yu et al., 2006).

$$F = \frac{(M\hat{\boldsymbol{\gamma}})'(M(\boldsymbol{X}'\hat{V}^{-1}\boldsymbol{X})^{-1}M')^{-1}(M\hat{\boldsymbol{\gamma}})}{p} \tag{5.3}$$

with *p* numerator degrees of freedom and *n*-*q* denominator degrees of freedom. The Satterthwaite degrees of freedom may also be computed, avoiding computationally intensive matrix operations.

The identification of causal SNPs using the LMM approach may produce misleading results due to the presence of outliers in the phenotypic trait. In this work, we therefore consider the minimum $\beta$-divergence method (Mollah et al., 2007; Nurul Haque Mollah et al., 2010) to improve the robustness and efficiency in terms of higher statistical power and lower FDR.

We can obtain robust estimation of model parameters by following three ways:

1. Applying the weighted estimators of the model parameters.
2. Applying existing methods on the modified dataset, which can be done by modifying/deleting the outlying observations from the dataset
3. Applying the suitable transformation on the dataset and then apply the existing methods.

The following steps were considered and implemented for outlier modification in the phenotypic observation:

- Select the most significantly associated SNP with the phenotypic variations using robust ANOVA (Mollah et al., 2015).

  Let $\boldsymbol{y} = (y_1, y_2, ..., y_n)' = (y_{11}, y_{12}, ..., y_{1n_1}, ..., y_{m1}, y_{m2}, ..., y_{mn_m})'$,

  where, $(n = n_1 + n_2 + ... + n_m)$

- Divide the phenotypic data into *m* groups corresponding to the *m* genotypic labels of the selected SNP.
- Detect the outlying observations from the $l^{\text{th}}$ ($l$=1,2,….,$m$) group using the $\beta$-weight function defined by

$$W_\beta(y_{li} \mid \hat{\theta}_l) = \exp\left\{-\frac{\beta}{2\sigma_l^2}(y_{li}-\hat{\mu}_l)^2\right\} \quad ; i=1, 2, \ldots\ldots, n \text{ and } l=0,1,2 \qquad (5.4)$$

The minimum $\beta$-divergence method estimators $\hat{\theta}_{l,\beta} = (\hat{\mu}_{l,\beta}, \hat{\sigma}^2_{l,\beta})$ of the parameters $\theta_{l,\beta} = (\mu_{l,\beta}, \sigma^2{}_{l,\beta})$ are computed iteratively as follows:

$$\mu_{l,t+1} = \frac{\sum_{i=1}^{n_l} W_\beta(y_{li} \mid \theta_{l,t})\, y_{li}}{\sum_{i=1}^{n_l} W_\beta(y_{ji} \mid \theta_{l,t})} \qquad (5.5)$$

and

$$\sigma^2{}_{l,t+1} = \frac{\sum_{i=1}^{n_l} W_\beta(y_{lk} \mid \theta_{l,t})(y_{li}-\mu_{l,t})^2}{(\beta+1)^{-1}\sum_{i=1}^{n_l} W_\beta(y_{li} \mid \theta_{l,t})} \qquad (5.6)$$

The notation $\theta_{t+1}$ represents the update to $\theta_t$ in the $(t+1)^{th}$ iteration. The robustness of these estimators is discussed in the background of influence function (Mollah et al., 2007) and their reliability (Nurul Haque Mollah et al., 2010). It is noteworthy that the minimum $\beta$-divergence estimators $\hat{\theta}_{l,\beta} = (\hat{\mu}_{l,\beta}, \hat{\sigma}^2_{l,\beta})$ reduce to the classical maximum likelihood estimators (MLEs) $\hat{\theta}_l = (\hat{\mu}_l, \hat{\sigma}^2_l)$ for $\beta = 0$.

It is considerable that the MLEs of a Gaussian distribution are consistent and asymptotically efficient in the absence of outlying objects (Zhang et al., 2005). Therefore, in this article, an attempt has been made to develop a robust mixed linear model approach in which the classical MLEs $\hat{\theta}_l$ are used in the absence of outlying objects and minimum $\beta$

-divergence estimators $\hat{\theta}_{l,\beta}$ stated in equation (5) and (6) are used in the presence of outliers for estimation of $\theta_l$ in the mixed model. The minimum $\beta$-divergence method suggests two approaches for combining the robustness and efficiency of estimation in LMM. The tuning parameter $\beta$ is selected through cross-validation (CV) technique (Mollah et al., 2007). CV process produces $\beta$=0 for the minimum $\beta$-divergence method estimators and is then equivalent to the classical estimators. When there are outlying subjects in the phenotypic traits, the technique generates $\beta$>0 for the minimum $\beta$-divergence estimators. To overcome the challenges of outlying observations in GWA, an alternative approach that is the $\beta$-weight function mentioned in (1) has been proposed with $\beta$=0.2 for outlier detection. This weight function imposes smaller weights (≥0) to outlying observations and larger weights (≤1) to uncontaminated/usual objects.

An outlying phenotypic observation $x_{jk}$ in the $j^{th}$ group is defined based on the $\beta$-weight function mentioned below:

$$W_\beta(x_{jk} \mid \hat{\theta}_{j,\beta}) = \begin{cases} > \tau_j, if \quad x_{jk} \; is\ not\ an\ outlier, \\ \leq \tau_j, if \quad x_{jk} \; is\ an\ outlier, \end{cases} \tag{5.7}$$

Where the threshold value $\tau_j$ is the $p^{th}$ quantile value of empirical distribution of $W_\beta(x_{jk} \mid \hat{\theta}_{j,\beta})$.

- Then replace the outlying phenotypic observation of *jth* group by its robust mean $\mu_{j,\beta}$ (*j=1, 2,….,m*) where m is the number of genotype in the selected SNP.

- After that apply efficient mixed model association (EMMA) to the modified dataset discussed in the previous step.

### 5.2.2 Simulated Data

#### 5.2.2.1 Genotype Simulation

To investigate the performance of our proposed method, a set of synthetic genotype and phenotype data were generated. Synthetic genotype dataset were simulated reflecting population structure. For this purpose, *m*=1000 SNPs were generated for *n*=1000 individuals and these individuals were taken from k=3 distinct populations by considering different minor allele frequencies (MAFs). To do this, first, a set of *m* latent vectors viz., $v = v_1, v_2, ......v_m$ was generated from multivariate normal distribution with mean zero and variance-covariance matrix $cov(v_j, v_i) = \rho^{|j-i|}$ (Li et al., 2014; Wang and Abbott, 2008). In our simulation we considered $\rho$ =0.5 to avoid the linkage disequilibrium (LD) between the SNPs. Finally, two cutoff values *s₁* and *s₂* were used to convert latent vectors $v_{lj}$ to genotype score $z_{ji}$ (*i= 1,2,…n*; *j=1,2,…,m*) as follows

$$z_{j,i} = \begin{cases} 0, & v_{ij} < s_1 \\ 1, & s_1 \le v_{ij} \le s_2 \\ 2, & v_{ij} > s_2 \end{cases}$$

Where $s_1$ and $s_2$ determine the minor allele frequency.

#### 5.2.2.2 Phenotype Simulation

Phenotypic datasets were produced by considering several factors comprising of genetic variation, error variation and population stratification. To generate phenotype data, two distinct scenarios were considered with the heritability rate 0.2 and 0.3. In every scenario, 4 SNPs were considered as causal variants and the remaining SNPs were assigned as polygenic variants. The genetic effects of the SNPs were simulated from normal distribution such that it satisfies certain proportion of genetic variance for different genetic effects (main effect and polygenic effect). The continuous trait values were simulated using the linear model $y_j = \mu + \sum_{k=1}^{m_1} a_k x_{ij} + \sum_{k=1}^{m_2} b_k x_{ij} + \varepsilon_j$. To check the performance of the proposed method in presence of outlier, we contaminated 0%, 1%, 2%, 3%, 4%, and 5% of the phenotypic data by outlying observation. We compared the performance of the methods in terms of statistical power and FDR. The statistical power and FDR of the methods were calculated using the formula, $Power = \frac{p_T}{p_k}$ and $FDR = \frac{\tau}{p_T + \tau}$ respectively, where $p_T$ is the number of truly detected SNPs and $p_k$ is the total number of causal variants and $\tau$ is the number of falsely detected SNPs. For each scenario, 1000 replications were performed to account the average value of the power and FDR for comparison.

### 5.2.3 The Real Datasets

To validate the performance of the proposed method, we analyzed a real dataset obtained from rice crop. The genotypic and phenotypic data used to carry out the analysis in this investigation were collected from the rice diversity database (www.ricediversity.org). The data set contain 413 accessions along with 36,901 SNPs of *Oryza Sativa*(Zhao et al., 2011). All selected SNPs were taken into consideration in the analysis with call rate >70% and minor allele frequency (MAF)>0.05(Zhao et al., 2011). The individual with missing observation in the phenotype were not considered in this study. Experimental data of the flowering time were collected as the number of days until the inflorescence was 50% emerged from the flag leaf calculated from the day of planting. The phenotype data used in this analysis for flowering time were recorded in Faridpur dristrict in Bangladesh.

### 5.2.4 Gene set Enrichment Analysis

To identify classes of genes/SNPs that may have a significant association with the phenotypic variations, we performed the following gene set enrichment analysis (GSEA).

#### 5.2.4.1 Gene Ontology and Pathway Analysis of the Identified SNP Markers

Gene ontology (GO) classification as well as the functional pathway enrichment analysis of detected SNPs/Genes to predict the associated molecular functional routes or variables in terms of biological process (BP), molecular functions (MF) and cellular components (CC), were performed using the online-based database MSU RGAP7 (http://rice.plantbiology.msu.edu).

#### 5.2.4.2 Prediction of the Subcellular Location of the Identified SNP Makers

To predict the subcellular location (SCL) of the detected SNPs/Genes, an online-based tool called Plant Subcellular localization Integrative (PSI) predictor (Liu et al., 2013b) was used to predict the corresponding molecular organs or locations in the plant cell.

## 5.3. Results

### 5.3.1 Simulation Study

To investigate the performance of the proposed method in comparison of the linear regression model (LRM) and linear mixed model (LMM), we computed statistical power and false discovery rate (FDR) by each of the LRM, LMM, and proposed methods in both absence and presence of outliers in the simulated datasets described in sub-sections 2.2.1-2.2.2. A method is said to be better with higher statistical power and smaller FDR. **Fig. 5.1(a)** and **Fig.5.1(b)** show the results of statistical power and FDR respectively for detection of four causal SNP variants out of 1000 SNPs from the simulated dataset having genetic heritability rate 0.2 computed by each of LRM, LMM and the proposed methods with the cutoff-value $10^{-5}$ in presence of 0% to 5% phenotypic outliers. It is observed that LMM and the proposed method show almost similar performance in terms of higher statistical power (> 60%) and lower FDR in absence of outliers, whereas LRM shows higher power with much larger FDR. It is also observed that the statistical power of both LRM and proposed method reduces very steadily with the increasing rate of outliers, but the power of LMM decreases significantly. On the other hand, FDR of both LMM and the proposed method remains much lower and stable, but LRM produces much higher FDR with increasing rate of outliers. Similar results are also found when the cutoff value is $10^{-7}$ (**Fig. 5.1(c)** and **5.1(d)).** Thus the proposed method shows better performance than the LRM and LMM for the dataset with the genetic heritability=0.2 at different levels of phenotypic outliers.

**Fig. 5.2(a)** and **Fig.5.2(b)** also show the results of statistical power and FDR respectively for detection of four causal SNPs variants out of 1000 SNPs from the simulated dataset having genetic heritability=0.3 computed by each of LRM, LMM and the proposed methods with the cutoff-value $10^{-5}$ in presence of 0% to 5% phenotypic outliers same as before. It is observed that LMM and the proposed method show almost similar performance in terms of higher statistical power and lower FDR in absence of outliers, whereas LRM shows lower statistical power with much larger FDR. It is also observed that the statistical power of both LRM and proposed method remain almost stable with the increasing rate of outliers, but the power of LMM decreases significantly. On the other hand, FDR of both LMM and the proposed method remains much lower and stable, but LRM produces much higher FDR though it gets insignificantly lower at h=0.3 with increasing rate of outliers as before. Similar results are also found when the cutoff value is $10^{-7}$ (**Figure 5.2(c)** and **5.2(d)).** Hence the proposed method also shows superior performance than the LRM and LMM for the dataset with the genetic heritability=0.3 at different levels of phenotypic outliers. It should be noted here that the statistical power of all methods becomes larger and FDR smaller when the genetic heritability was larger at 0.3.

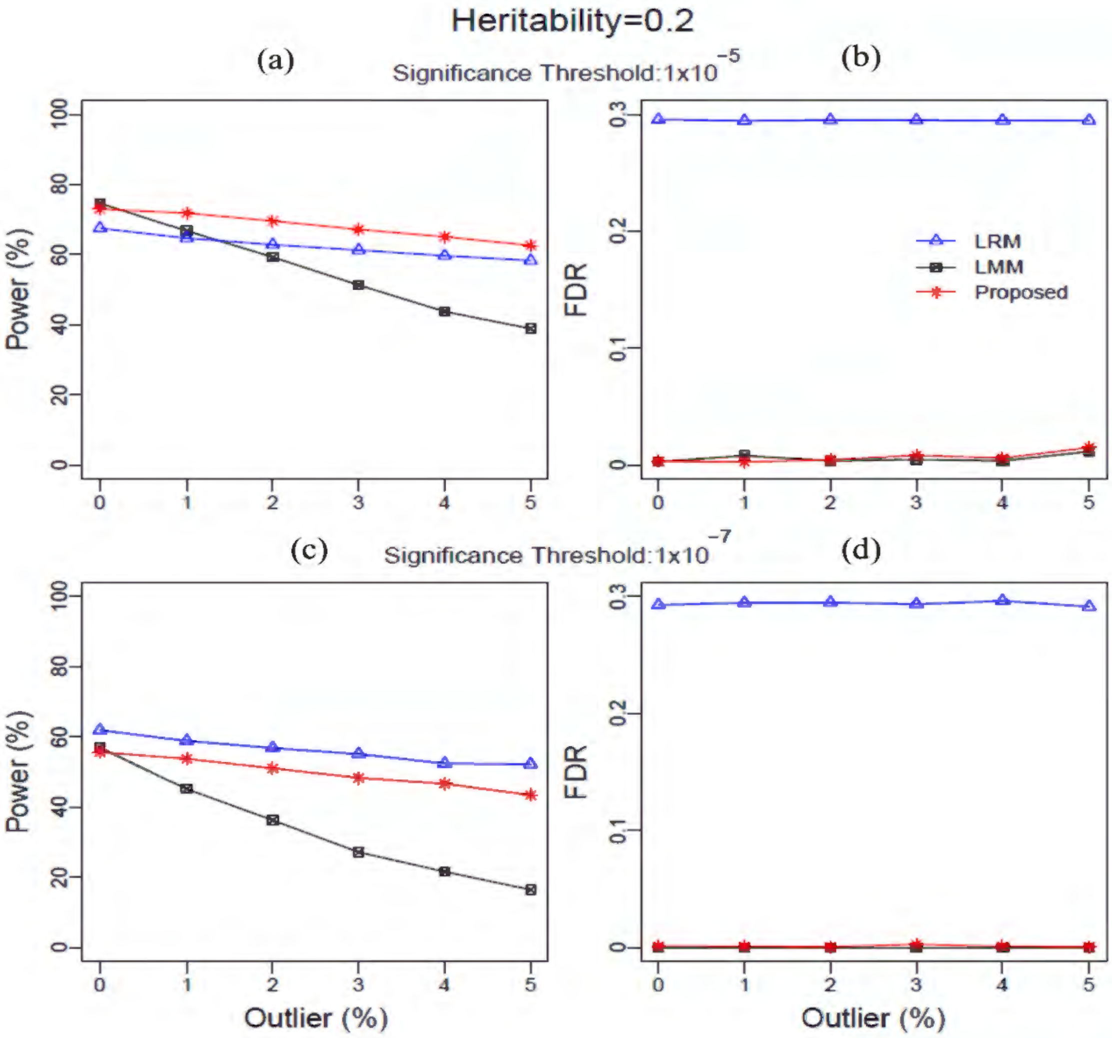


**Fig.5.1.** Statistical power and false discovery rate (FDR) estimates in the genome-wide association study (GWAS) using simulated quantitative trait obtained from 1000 replicates. (a) Represents statistical power at h=0.2 and at cutoff $10^{-5}$. (b) Represents FDR at h=0.2 and at cutoff=$10^{-5}$. (c) Represents statistical power at h=0.2a and cutoff $10^{-7}$. (d) Represents FDR at h=0.2 and cutoff $10^{-7}$

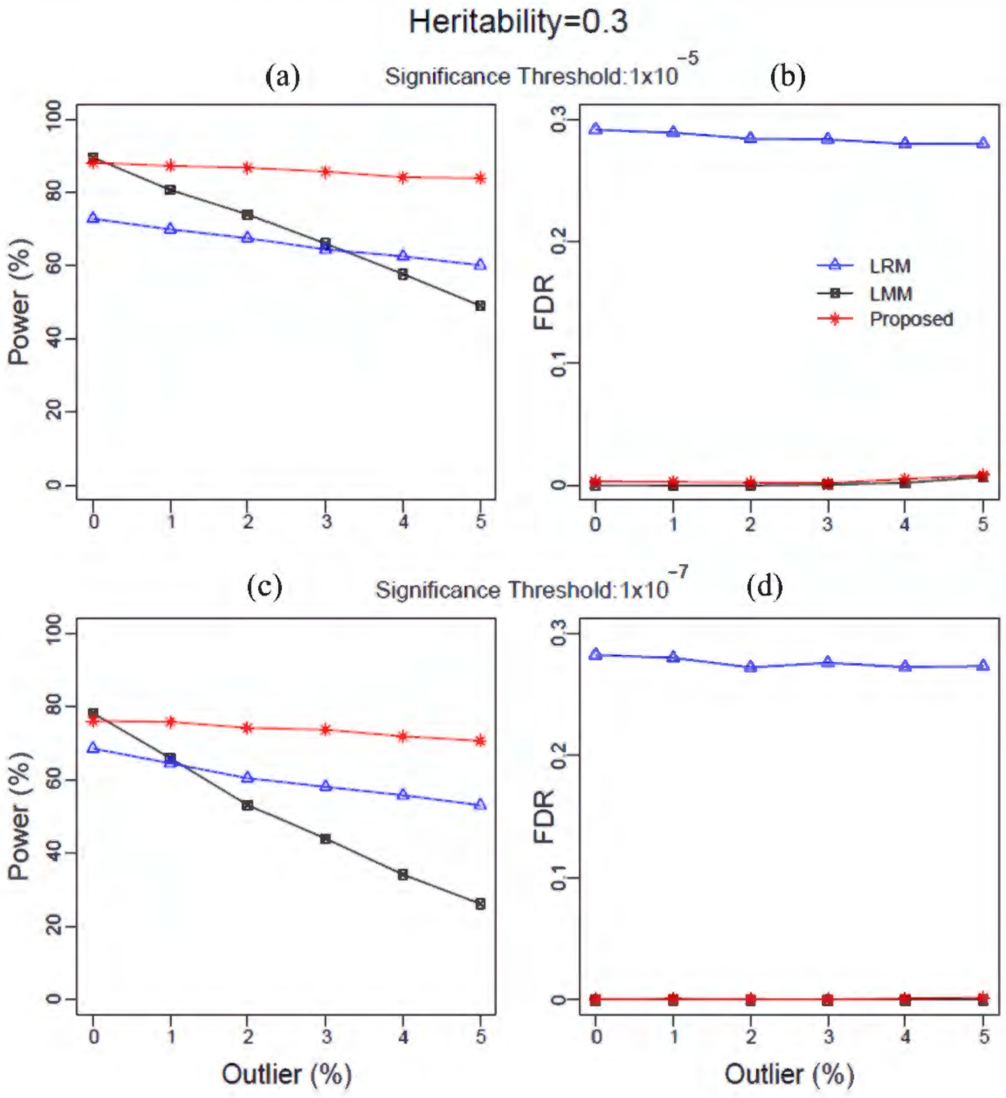


**Fig.5.2** Statistical power and false discovery rate (FDR) estimates in the genome-wide association study (GWAS) using simulated quantitative trait obtained from 1000 replicates.(a) Represents statistical power at h=0.3 and at cutoff $10^{-5}$.(b) Represents FDR at h=0.3 and at cutoff=$10^{-5}$.(c) Represents statistical power at h=0.3 and cutoff $10^{-7}$.(d) Represents FDR at h=0.3 and cutoff $10^{-7}$.

### 5.3.2 Genome-Wide Characterization of the Identified Rice SNPs using Proposed Method

To demonstrate the performance of the proposed in the case of real data analysis, we considered real rice flowering time SNP dataset described in sub-section 5.2.3. We then calculated the *p*-values of 36,901 SNPs using the $F_\beta$-statistic in equation (3) to test the hypothesis that there is no association between SNP and the phenotypic trait, that is $M\gamma = 0$.

Using the resulting *p*-values a Manhattan plot has been drawn by taking the values of –$\log_{10}$ (*p*-values) (**Fig.5.3**). SNPs corresponding to the color points lying above the threshold 0.05 are termed as significant SNPs. From the **Fig.5.3,** it is clearly observable that 11 SNPs fall above the threshold, where six SNPs lie in chromosome 2, one SNP marker belong to each of chromosomes 6 and 7 as well as three SNP markers are belong to chromosome 8. Then we used corresponding locus ID of these 11 identified SNPs from the rice genomics annotation database "Rice Genome Annotation Project (RGAP) release 7 (http://rice.plantbiology.msu.edu/)" for collecting necessary genomic information given in **Table 5.1**. This table describes the genomic information of 11 identified SNPs including chromosome, SNP position, chromosomal location, distance, protein-coding genes, CDS coordinates (3'-5') and each SNP containing gene product names.

There are few special domains containing transposable proteins or transcription factors that are found in these 11 rice SNP markers, probably associated with some important biological and cellular functions in plant cell physiology, stress, and development in rice. Pentatricopeptide repeats (PPR) domain containing protein found in LOC_Os02g21070 is assumed to participate in the biological molecule modification (Sharma and Pandey, 2016). Previous study also showed that this protein is also actively involved in the cellular biosynthesis, regulate plant secondary metabolism comprising abiotic and biotic stress stimuli (Schaper and Anisimova, 2015; Sharma and Pandey, 2016). The coiled-coil domain containing (CCDC) protein identified in LOC_Os02g21880 is predicted to involve in performing some cellular processes in nucleus. This protein acts as the activity of regulation of protein positioning in the cell during cell division by separating and organizing signaling pathways in a temporal and spatial manner (Rose, 2004). The sec 1 family protein is likely to have important link between cell cycle progression and the membrane fusion apparatus found in gene LOC_Os02g24134 (Assaad et al., 2001). The mobile retrotransposon element of Ty 1 copia class type is one of the most fluids found in LOC_Os02g24770 is putative to play key roles in the structural evolution of the rice genome (Todorovska, 2007). The protein kinase domain containing proteins are commonly assumed to expand in the flowering plant lineage (Lehti-Shiu and Shiu, 2012). This type of protein found in different flowering plants is often likely to play conserved regulatory roles in metabolism and cell division in the cellular process (Lehti-Shiu and Shiu, 2012).This protein kinase domain however found in rice gene LOC_Os06g18000 is also have the same putative regulatory roles in metabolism and cell division like other flowering plants. The novel BSD domain-containing protein or transcription factor found in LOC_Os08g25060 is assumed to play crucial role in somatic embryogenesis (SE) for various developmental process in rice. BSD is also likely to involve in the cell proliferation during SE. These similar BSD type transcription factor was also found in banana for performing the same functions(Shivani et al., 2017).

Then we performed GSEA for the detected 11 SNP makers using GO and SCL analysis to identify more valid SNPs out of 11 that have significant association with the flowering time and other trait variations in rice. We also studied their expressions in various organs in rice to find the link with the flowering time.

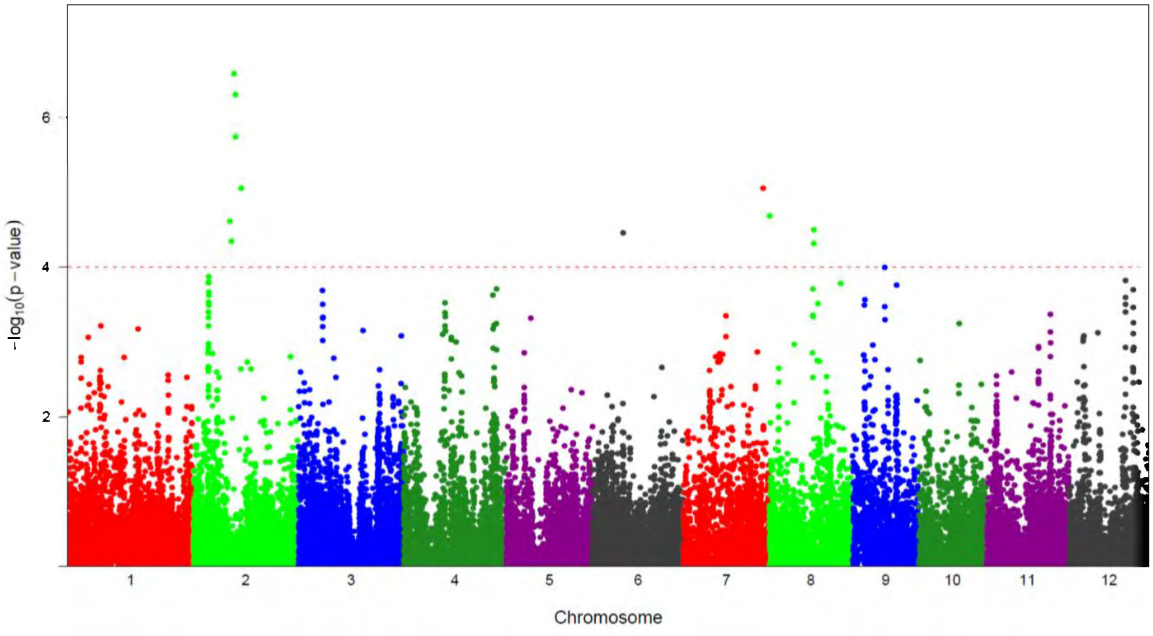


**Fig.5.3** Manhattan plot of the 36,901 SNPs plotted across 12 chromosomes. Each dot denotes a single SNP. The X-axis represents the corresponding genomic location and the Y-axis measures the association level. Threshold $10^{-5}$ was taken into consideration to differentiate the potential SNPs/markers in rice.

**Table 5.1.** Identified 11 SNPs of Rice Genome

| ID | P-value | Chr | SNP Pos | Loc | Distance | Start | End | Description |
|---|---|---|---|---|---|---|---|---|
| id2005644 | 2.44E-05 | 2 | 12488337 | LOC_Os02g21070 | 4273 | 12492610 | 12493561 | PPR repeat domain containing protein, putative, expressed |
| id2005743 | 4.62E-05 | 2 | 13011782 | LOC_Os02g21880 | including | 13006951 | 13013160 | coiled-coil domain-containing protein, putative, expressed |
| id2005919 | 2.63E-07 | 2 | 13975952 | LOC_Os02g24134 | 1657 | 13977609 | 13987430 | Sec1 family transport protein, putative, expressed |
| ud2000772 | 1.82E-06 | 2 | 14370758 | LOC_Os02g24770 | Including | 14368741 | 14371953 | retrotransposon protein, putative, Ty1-copia subclass, expressed |
| id2005983 | 4.88E-07 | 2 | 14376159 | LOC_Os02g24780 | including | 14374834 | 14379670 | retrotransposon protein, putative, unclassified, expressed |
| id2006587 | 8.89E-06 | 2 | 16434820 | LOC_Os02g27750 | including | 16432240 | 16438080 | transposon protein, putative, unclassified, expressed |
| wd6000761 | 3.44E-05 | 6 | 10471943 | LOC_Os06g18000 | 715 | 10469299 | 10471228 | protein kinase domain containing protein, expressed |
| ud7002027 | 8.82E-06 | 7 | 27420180 | LOC_Os07g45950 | 918 | 27421098 | 27423642 | expressed protein |
| id8000022 | 2.11E-05 | 8 | 51045 | LOC_Os08g01070 | 2995 | 54040 | 58330 | retrotransposon protein, putative, unclassified, expressed |
| id8004076 | 4.92E-05 | 8 | 15199041 | LOC_Os08g25040 | including | 15198870 | 15199151 | expressed protein |
| id8004083 | 3.14E-05 | 8 | 15206184 | LOC_Os08g25060 | including | 15203190 | 15211952 | BSD domain-containing protein, putative, expressed |

* including indicates the protein coding genes

### 5.3.2.1 Molecular Functional Enrichment Analysis

There are 11 SNP candidate makers identified in chromosomes 2, 6, 7, 8 using the proposed method. Only six of those gene/SNP markers (LOC_Os02g21070, LOC_Os02g21880, LOC_Os02g24134, LOC_Os06g18000, LOC_Os07g45950, and LOC_Os08g25060) are detected to participate in different functional pathways in rice genome. **Fig.5.4(a)** shows that only the gene LOC_Os06g18000 (*p-value*=0.0000344) performs the maximum BP functions (abscission, flower development, response to stress, protein modification, signal transduction, cell death). Molecular transport mechanism and cellular process are assumed to conserve by LOC_Os02g24134 (*p-value*=0.000000263). Some other biological processes are anticipated to be controlled by two SNP makers LOC_Os06g1880 (*p-value*=0.0000462) and LOC_Os02g21070 (*p-value* 0.0000244) in rice. On the other hand, six MF are conserved by the SNP makers in rice (**Fig.5.4(a)**). The protein binding functional activity is maintained by LOC_Os08g25060 (*p-value*=0.0000314) and LOC_Os02g24134 (*p-value*=0.000000263). Among the predicted six molecular functional pathways, nucleotide binding and kinase activity were found to relate the gene LOC_Os06g18000. While the marker LOC_Os02g21880 was identified to initiate some molecular functions in rice but the gene, LOC_Os02g21070 was found not to involve in any molecular functional pathways in rice. A few cellular component functional pathways are predicted to maintain by the identified six SNP makers in rice such as LOC_Os02g21070 in plastid, LOC_Os02g24134 in membrane, vacuole, and golgi apparatus. Only LOC_Os06g18000 is involved in plasma membrane level (**Fig.5.4(a)**).

### 5.3.2.2 Prediction of Subcellular Location of the Identified SNP Markers in Rice

Subcellular localization of the identified makers in rice implies that cytosol is the highest number of genes container molecular location in rice (**Fig.5.4(b)**). LOC_Os02g27750, LOC_Os06g18000, and LOC_Os07g45950 and LOC_Os08g25040 were found located in plastid (plast) and the latter two genes were predicted to be located in extracellular (extra) and membrane (membr) of which LOC_Os07g45950 were also available in vacuole (vacu) (**Fig. 5.4(b)**). It is however observed from the **Fig.5.4 (b)** that three genes named LOC_Os02g21070, LOC_Os02g21880, and LOC_Os08g25040 are predicted to active in nuclear activity. However, any identified SNP makers out of 11 markers were predicted not to belong from the cellular locations viz., endoplasmic reticulum (ER), peroxisome (pero) and mitochondria (mito) in rice (**Fig.5.4 (b)**).

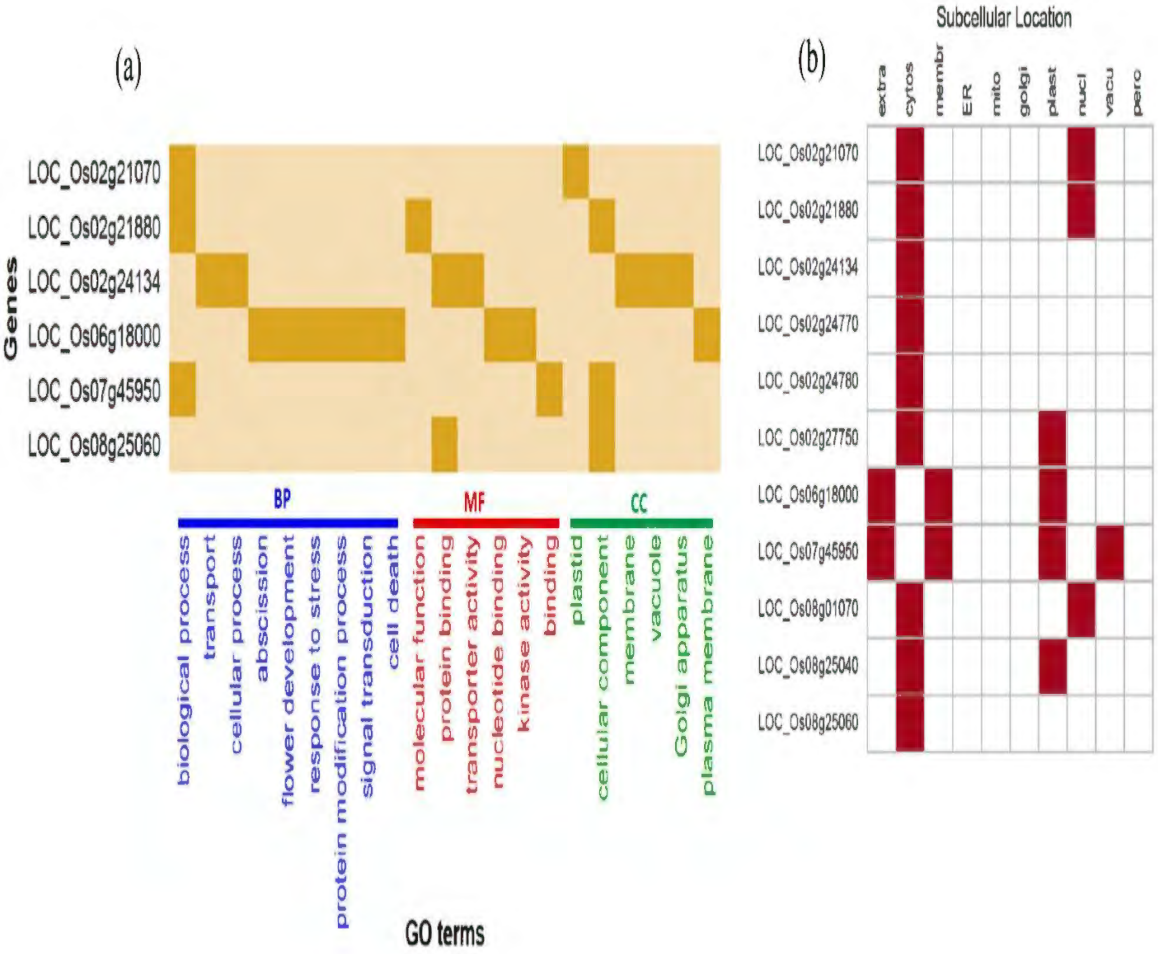


**Fig.5.4** (a) Represents the expression map of the functional pathways viz., biological process (BP), molecular function (MF) and cellular component (CC) of the six SNP markers in rice. (b) Represents the subcellular location(SCL) of the 12 SNP markers in 10 molecular organs viz., cytosol (cytos), endoplasmic reticulum (ER), extracellular (extra), golgi apparatus (golgi), membrane (membr), mitochondria (mito), nuclear (nucl), peroxisome (pero), plastid (plast) and vacuole (vacu). These two plots were produced using the open source software R.

### 5.3.2.3 Expression Analysis of the SNP Candidate Genes in Rice

Expression level or power of the 11 identified SNP maker genes from the proposed technique at different rice plant organs or tissues such as seedling, vascular cell, root, leaves, post-emergence, pre-emergence, seed, endosperm, embryo, shoots, anther, pistil and panicle were estimated using RGAP 7. The heatmap for studying expression level of the identified 11 SNPs was created via R open source software(R CoreTeam, 2017). Heatmap showed in **Fig.5.5** exhibits various expression levels in the 13 distinct rice plant organs or tissues viz., seedling, vascular cell, root, leaves, post and pre-emergence inflor, seed, endosperm, embryo, shoots, anther, pistil and panicle. It was observed from the expression plot that the two SNP makers LOC_Os02g21880 and LOC_Os02g24134 showed high-level expression in seedling, root, shoot, and panicle in rice while these two genes exhibited only high level expression in vascular cell at 14DAP (**Fig.5.5**).

The maker LOC_Os02g21070 compared to the remaining 10 makers only presented optimum expression in leaves. Moreover, seedling, vascular cell at 14DAP and shoots specific expressions were maximum for the maker LOC_Os06g18000. Though the maker LOC_Os08g01070 had highest peak in expression only in panicle but in other organs the expression intensity were lower along with the no expression in vascular cell. In addition, the SNP marker LOC_Os08g25060 provided topmost expression in panicle, root tip, and in vascular cell at 7DAP whereas in the other organs viz. seedling, endosperm, shoots, anther, pistil, this marker showed reduced expression. Surprisingly the three makers LOC_Os08g25040, LOC_Os07g45950 LOC_Os02g27750, and LOC_Os02g24780 had hardly expression in different organs or tissues in rice (**Fig.5.5**).

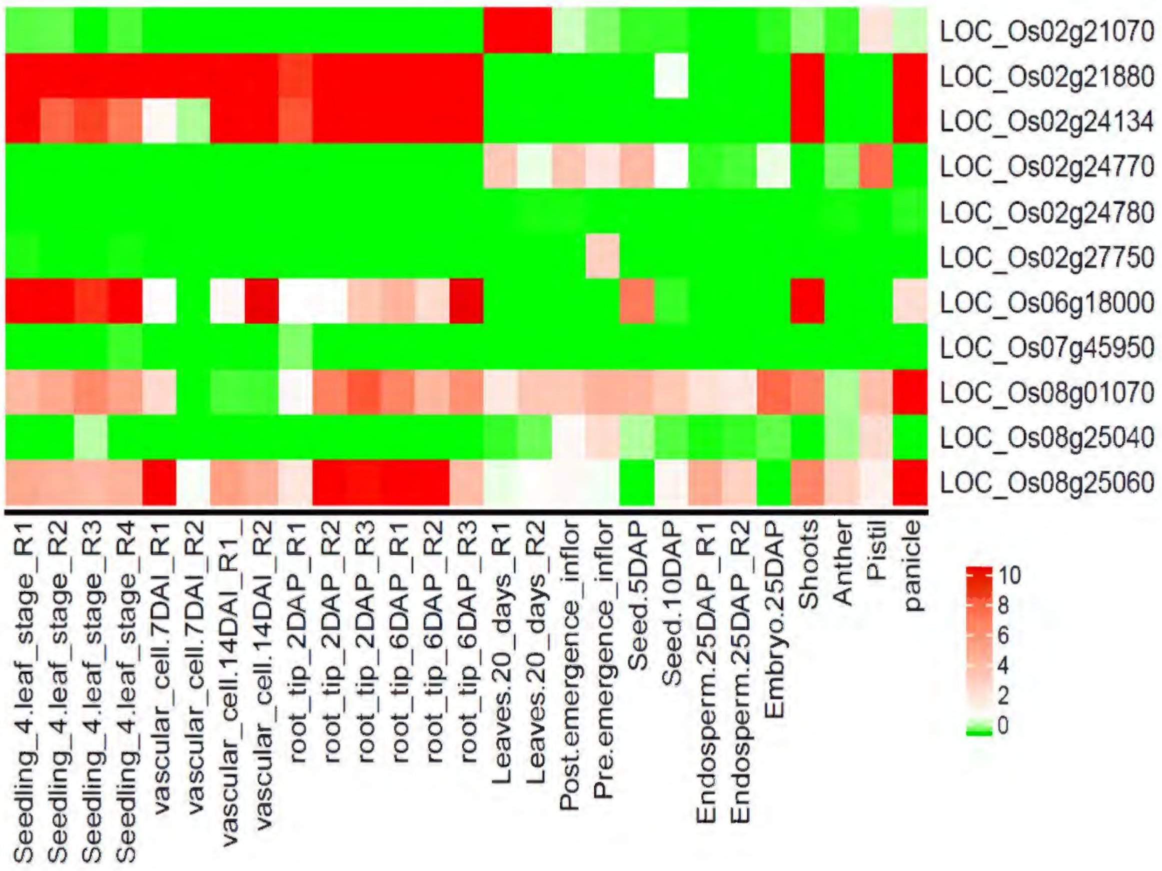


**Fig.5.5** Heatmap showing the expression pattern of the 11 identified SNP makers identified by proposed approach in various organs (seedling, vascular cell stage, root, leaves, post and pre-emergence inflor, seed, endosperm, embryo, shoots, anther, pistil and panicle) of rice. The color scale bar of the figure represents log2 transformed FPKM values. Subsequently the heatmap was generated by representing individual value with different colors. Map was produced using the open source R software.

## 5.4. Discussion and Conclusion

GWAS technique has recently made widespread scopes to explore the novel biomarker genes in species at SNPs level. This analysis procedure can simultaneously screen a very large set of accessions for evaluating genetic variation underlying diverse complex traits whereas the traditional quantitative trait locus (qtl) approach fails to investigate such big data to explore the genomic potentials(Zhao et al., 2011). Two key challenges such as existing embedded population structures or stratification along with polygenic effects or genetic relatedness among individuals are associated when GWAS is made(Hyun et al., 2008; Shin and Lee, 2015). To overcome these complexities recently LMM has been widely used. It is however observed that this method is sensitive to outliers and this may lead to produce inflated FDR and lower statistical power if the outliers are not properly handled. Previous study showed that a robust PCA approach combined with k-medoids clustering was proposed to address the issue of subject outliers in GWAS but it had some limitations when the dataset is large with varying population structures as well as if there are some other sample structures such as family structures or cryptic relatedness(Liu et al., 2013a). These significant issues are easily overcome by the recently introduced LMM while sensitive to outliers(Liu et al., 2013a). We therefore proposed a robust LMM approach for handling outliers and minimizing the confounding effects of the population stratification and genetic relatedness in GWA analysis. We introduced a $\beta$-weight function termed as minimum $\beta$- divergence method accompanied by a tuning parameter $\beta$ to overcome the problem of data contamination. From the performance evaluation, it is observed that the proposed approach performs superior in terms of lower FDR and higher statistical power compared to LRM and LMM at different level of outliers against two heritability proportions 0.2 and 0.3.

We further identified 11 rice SNP maker genes using the proposed method. We performed GSEA for the identified 11 SNP makers using GO and SCL analysis to detect more valid SNPs out of 11 that have significant association with the flowering time and other trait variations in rice. We also studied their expressions in various organs in rice to find the link with the flowering time. Flowering time is one of the most important agronomic traits that determines rice yield(Weng et al., 2014). Previous investigation suggested that traditional flowering time genes might have roles in plant development and stress response(Weng et al., 2014). From GO analysis, it is observed that the gene LOC_Os06g18000 might play functional roles in flower development and for response to stress in rice. Amongst the 11 genes, LOC_Os02g21880, LOC_Os06g18000, LOC_Os02g24134 exhibited larger expression in seedling, vascular cell, root, shoot, and panicle. Earlier study also suggested that leaves, shoot and panicles have significant roles in regulating flowering time(Lee and An, 2015; Weng et al., 2014). In addition, the gene

LOC_Os08g25060 is predicted to provide maximum expression in vascular cell, root, and panicle. Our findings resulted from real data analysis also supported by the other outcomes(Ahsan et al., 2018; Assaad et al., 2001; Pasam et al., 2012; Rose, 2004; Shivani et al., 2017; Zhao et al., 2011). Cytosol is the place where the occurrence of the maximum different metabolisms in plants and most of the proteins in the cell are localized in cytosol(Kholodenko, 2003; Ohlrogge et al., 1979). Our results also support that the cytosol contains the maximum number of genes. Plastid is an important molecular organ found in plant cell mostly involve in photosynthesis and other gene expression(Jansen et al., 2005). Photosynthesis is the key physiological parameters in rice relates ultimately in many aspects to increase the rice productivity(Hidayati et al., 2016). Increase photosynthesis rate can utilize the solar radiation properly that lead to create early flowering time because flowering signals are produced in leaves (Karki et al., 2013; Lee and An, 2015). In our study, SCL analysis shows that the expression of the gene LOC_Os06g18000 in plastid may act as flowering promoter. This gene expression in plastid likely to enhance the photosynthesis process which regulate the leaf anatomy for earlier flowering in rice. In GO analysis, it is also observed that this gene expression is associated to flowering in rice. Finally, it is concluded that phenotypic outliers may significantly influence the analysis results in GWAS. Our proposed robust method outperforms the existing LRM and LMM methods in presence of outliers and the genomic information presented may however provide basic platform for further biological investigations.

# CHAPTER SIX

## ROBUST CLASSIFICATION OF FUNCTIONAL METAGENOMES

# ROBUST CLASSIFICATION OF THE FUNCTIONAL METAGENOMES RECOVERED FROM DIFFERENT ENVIRONMENTAL SAMPLES

## 6.1 Introduction

Metagenomics refers to one of the influential branch of omics technology to study the numerous microbial derived from diverse environments. It is very necessary to analyze in terms of clustering or classifying the functional metagenomes collected from various sources on this earth. Microbiologists, statisticians and computational researches or individuals often face limitations for analyzing such big metagenomic dataset obtained from newly introduced NGS method. One of the major challenges is the proper and precise classification or grouping of the derived metagenomic dataset for in-depth phylogenic or evolutionary association analysis. Various microbial route possess several metagenomic functional magnitudes for distinct environment from where they are collected(Dinsdale et al., 2008, 2013). The study of the metagenomic provides the structure of the concerned microbes and their related activities. This also help to get easy explanation and analyzing results for thousands of protein sequence by applying BLAST matching techniques (Parks and Beiko, 2010). Numerous online software tools, platform, or databases are accessible to carry out suitable and thorough analysis of the functional metagenomes or different metagenomic datasets. But these analysis tools or software somtimes do not deliver sufficient up-to-date and effective results (Arndt et al., 2012). A set of multivariate statistical techniques, for example, principal component analysis(PCA), multi-dimensional scaling (MDS) , canonical discriminant analysis(CDA), linear discriminant analysis(LDA), tree mapping etc. are often commonly employed for some genomic and metagenomic dataset(Ramette, 2007). Though these multivariate statistical techniques provide some comprehensive results in classifying, clustering and visualizing large-set of metagenomic data obtained diverse environmental locations or conditions but sometimes researchers experience poor results in terms of higher misclassification rates. It is very crucial to make proper profile of the functional metagenomes separated from same or diverse environmental places or conditions. MetaGUN is the three-stage gene identification approach for predicting of metagenomic transcripts or proteins using classification support vector machine(SVM) technique (Jiang et al., 2017; Liu et al., 2011). K-Nearest Neighborhood (KNN) ,AdaBoost, LogitBoot are some classification approach can be applied for metagenomic and other genomic dataset(Sharma et al., 2015). Besides Random Forest (RF) (Breiman, 2001) is one of the popular efficient classification technique suited for large-scale metagenomic DNA or transcripts (Dinsdale et al., 2008, 2013). This technique uses the ensemble learning procedures for classification along with regression multiple patters datasets (Dinsdale et al., 2008, 2013). Large voluminous dataset along with big set of metagenomic functions or variables is the of the initial difficulties. It is therefore very essential to identify potential metagenomic functional variables influencing active metagenomes to form different class of the derived metagenomes using robust statistical approach. In this work, a beta *t*-statistic (Akond et al. 2018) has been introduced for robust identification of metagenomic functional variables from large metagenomic datasets then employ RF classification technique for making efficient classes or groups as well to make lower error rate. The steps of this study have been presented in **Fig.6.1**.

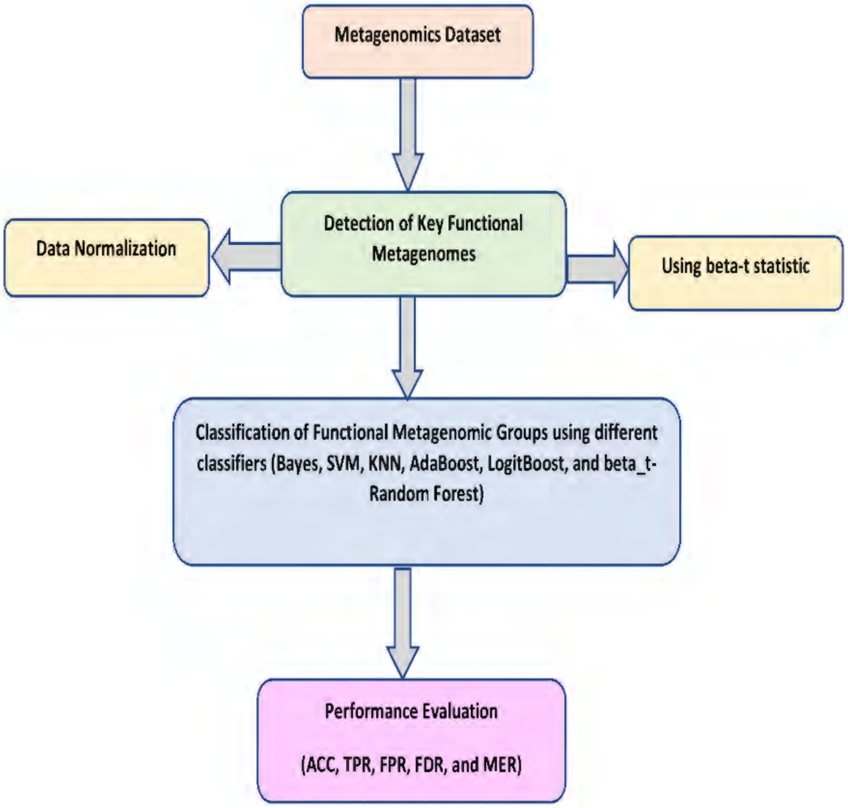

Metagenomics Dataset
Data Normalization
Detection of Key Functional Metagenomes
Using beta-t statistic
Classification of Functional Metagenomic Groups using different classifiers (Bayes, SVM, KNN, AdaBoost, LogitBoost, and beta_t-Random Forest)
Performance Evaluation
(ACC, TPR, FPR, FDR, and MER)

**Fig.6.1** Schematic diagram of this study

## 6.2 Materials and Methods

Bayesian classifier is commonly applied for simple probabilistic classifier. Transcipts or DNA sequence features are used for the input $X = (x_1 x_2 \ldots x_p)$ to the Bayesian classifier. For each metagenome, Bayesian classifier created a mult-iclass and the Bayesian classifier was trained by set of categorized as training dataset ($X$, $C$). Support Vector Machine (SVM) another machine learning classification technique, KNN, AdaBoost, LogitBoost (De'Ath and Fabricius, 2000), Random Forest(Venables and Ripley, 2002) Beta-*t* statistic (Akond et al. 2018) are employed for classification and comparison of functional metagenomes derived from distinct microbial communal. The analyses results of this work are carried out using open source statistical R programming language (R CoreTeam, 2017).

### 6.2.1 Dataset

Dataset used in this study are gathered from an earlier published article(Dinsdale et al., 2013). That data sample consists of 212 metagenomes derived from 10 different environmental conditions along with 26 metagenomics functional variables.

## 6.3 Results and Discussion

To identify the key functional metagenomes, we used the beta *t*-test statistic. This method is described in details in the previously published paper(Akond et al. 2018) . Using the method we select the top nine key functional metagenomes (AAD, CDCC, CVPGP, DNAM, MT, MC, NN, Plasmids and SM) based the on the P-values at 5% level of significance (**Table 6.1**). The key functional metagenomes are abbreviated in the alphabetical letter case those are selected from the 10 different microbial community and 212 metagenomes.

The Pearson correlation network plot (**Fig.6.2**) showed the correlation among the key functional metagenom. The ADD (amino acids and derivatives) is strongly correlated ($r > 0.81$) with the other metagenomes CDCC (Cell Division and Cell Cycle), CVPGP (Cofactors Vitamins Prosthetic Groups Pigments), DNAM (DNA Metabolism), and MT (Membrane Transport).

**Table 6.1.** Key Metabolite functions selected by the beta-*t* test statistic

| **Key Metabolite Functions** | **Metabolite Function Abbreviation** | ***P*-value*** |
|---|---|---|
| *Amino Acids and Derivatives* | AAD | 0.041 |
| *Cell Division and Cell Cycle* | CDCC | 0.034 |
| *Cofactors Vitamins Prosthetic Groups Pigments* | CVPGP | 0.043 |
| *DNA Metabolism* | DNAM | 0.005 |
| *Membrane Transport* | MT | 0.015 |
| *Motility and Chemotaxis* | MC | 0.025 |
| *Nucleosides and Nucleotides* | NN | 0.007 |
| *Plasmids* | Plasmids | 0.014 |
| *Sulfur Metabolism* | SM | 0.026 |

*$P < 0.05$, statistical significant at 5% level of significance

The highly positive correlations among the metagenomic variables imply that they are similar in direction from AAD functional metagenomes. On the other hand, ADD is negatively correlated with the MC (Motility and Chemotaxis), NN (Nucleosides and Nucleotides), Plasmids, and SM (Sulfur Metabolism). The opposite relationship existed among the functional metagenomes in the different microbial community.

To investigate the performance of the different classifiers we divided full dataset into three different parts using the cross validation (CV) method such as *10*-fold, *5*-fold, and *3*-fold cross validation dataset and checking the performance. In case of full dataset, performance of different classifiers (Bayes, SVM, KNN, AdaBoost, LogitBoost and Beta-*t* Random Forest) is shown in the **Table 6.2**. The performance measure of all the methods using accuracy (AAC), true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR), and misclassification error rate (MER). Bayes classifier showed the lowest performance in terms of ACC (57%), TPR (57%), FPR (42%), FDR (48%), and MER (43%) whereas the highest performance is observed for beta-*t* Random Forest in terms of ACC, TPR, FPR, FDR and MER with the results of 94%, 94%, 5%, 6% and 6% respectively.

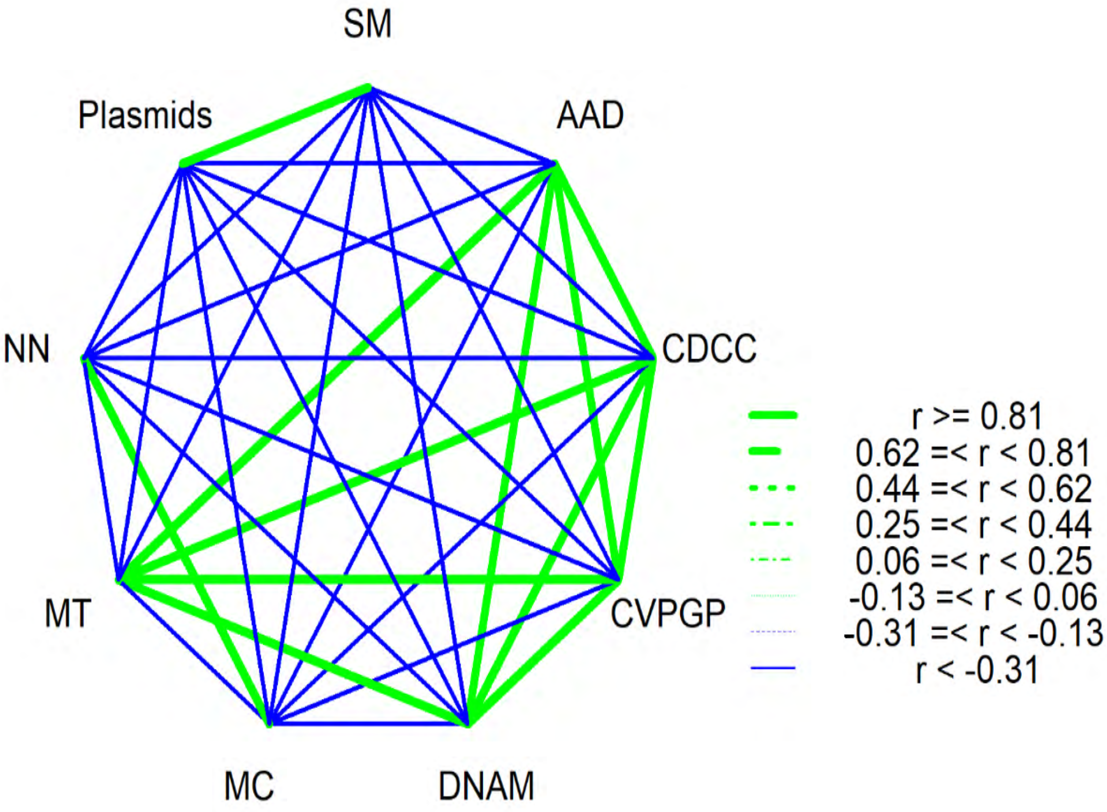


**Fig.6.2.** Pearson correlation network for nine key functional metagenomes

Finally, we showed that beta-*t* Random Forest provided the better performance for full dataset. For the 10-fold cross validation dataset, the Bayes classifier showed the lowest performance and LogitBoost and beta-*t* Random Forest showed approximately equal performance but eventually beta-*t* Random Forest was considered as better classifier than the other methods. In case of *5*-fold and *3*-fold cross validation dataset, it is found that the beta-*t* Random Forest method showed better ACC, TPR, FPR, FDR, and MER respectively.

**Table 6.2.** Classification performance of different classifiers

| Methods | ACC | TPR | FPR | FDR | MER |
|---|---|---|---|---|---|
| | | | ***Full Dataset*** | | |
| *Bayes* | 0.566 | 0.569 | 0.422 | 0.477 | 0.434 |
| *SVM* | 0.514 | 0.509 | 0.475 | 0.577 | 0.486 |
| *KNN* | 0.844 | 0.849 | 0.149 | 0.154 | 0.156 |
| *AdaBoost* | 0.894 | 0.878 | 0.083 | 0.081 | 0.106 |
| *LogitBoost* | 0.933 | 0.901 | 0.025 | 0.024 | 0.067 |
| *Beta_t+Random Forest* | **0.937** | **0.935** | **0.054** | **0.056** | **0.063** |
| | | | ***10-Fold Cross Validation*** | | |
| *Bayes* | 0.557 | 0.549 | 0.421 | 0.417 | 0.443 |
| *SVM* | 0.501 | 0.506 | 0.509 | 0.498 | 0.499 |
| *KNN* | 0.855 | 0.852 | 0.138 | 0.139 | 0.145 |
| *AdaBoost* | 0.907 | 0.912 | 0.094 | 0.097 | 0.093 |
| *LogitBoost* | 0.955 | 0.944 | 0.032 | 0.032 | 0.045 |
| *Beta_t+Random Forest* | **0.955** | **0.962** | **0.050** | **0.052** | **0.045** |
| | | | ***5-Fold Cross Validation*** | | |
| *Bayes* | 0.596 | 0.592 | 0.402 | 0.429 | 0.404 |
| *SVM* | 0.503 | 0.503 | 0.485 | 0.530 | 0.497 |
| *KNN* | 0.793 | 0.798 | 0.202 | 0.205 | 0.207 |
| *AdaBoost* | 0.887 | 0.892 | 0.107 | 0.111 | 0.113 |
| *LogitBoost* | 0.946 | 0.922 | 0.023 | 0.022 | 0.054 |
| *Beta_t+Random Forest* | **0.972** | **0.968** | **0.021** | **0.022** | **0.028** |
| | | | ***3-Fold Cross Validation*** | | |
| *Bayes* | 0.664 | 0.664 | 0.329 | 0.300 | 0.336 |
| *SVM* | 0.502 | 0.496 | 0.486 | 0.458 | 0.498 |
| *KNN* | 0.824 | 0.823 | 0.168 | 0.169 | 0.176 |
| *AdaBoost* | 0.901 | 0.886 | 0.078 | 0.078 | 0.099 |
| *LogitBoost* | 0.962 | 0.939 | 0.011 | 0.011 | 0.038 |
| *Beta_t+Random Forest* | **0.988** | **0.980** | **0.003** | **0.003** | **0.012** |

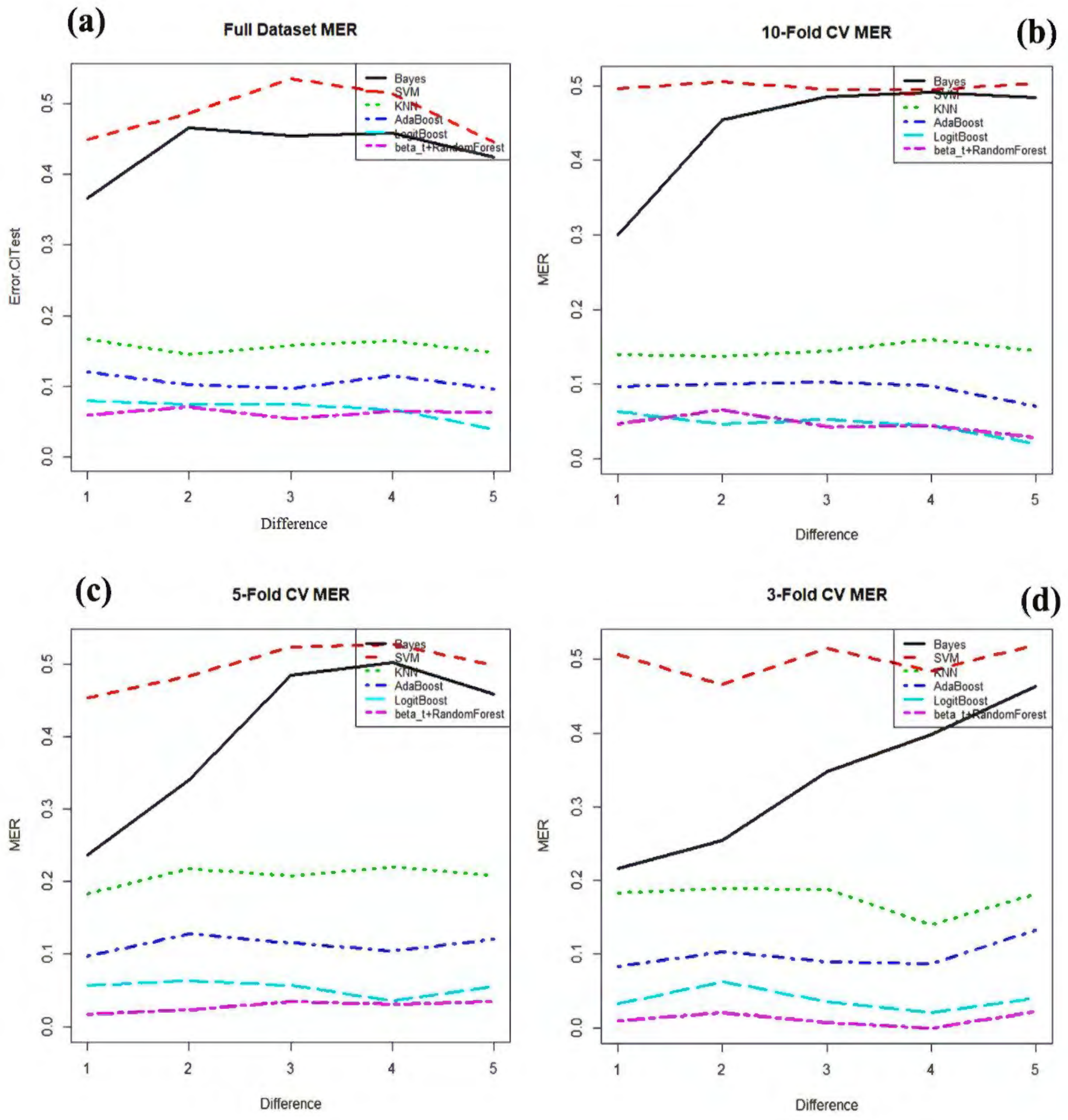


**Fig.6.3.** Misclassification error rate (a) Full dataset, (b) 10-fold cross validation, (c) 5-fold cross validation and (d) 3-fold cross validation for different classifiers.

From the **Fig.6.3** it is revealed that among the misclassification error rate (MER) of the six different classifiers, the SVM classifier provided the highest MER and beta-*t* Random Forest showed the lowest MER for full dataset. Similarly, for other datasets (10-fold, 5-fold, and 3-fold CV) SVM also showed the highest MER and beta-*t* Random Forest provided the lowest MER. It is however demonstrated that the beta-*t* Random Forest showed the lowest MER for all datasets.

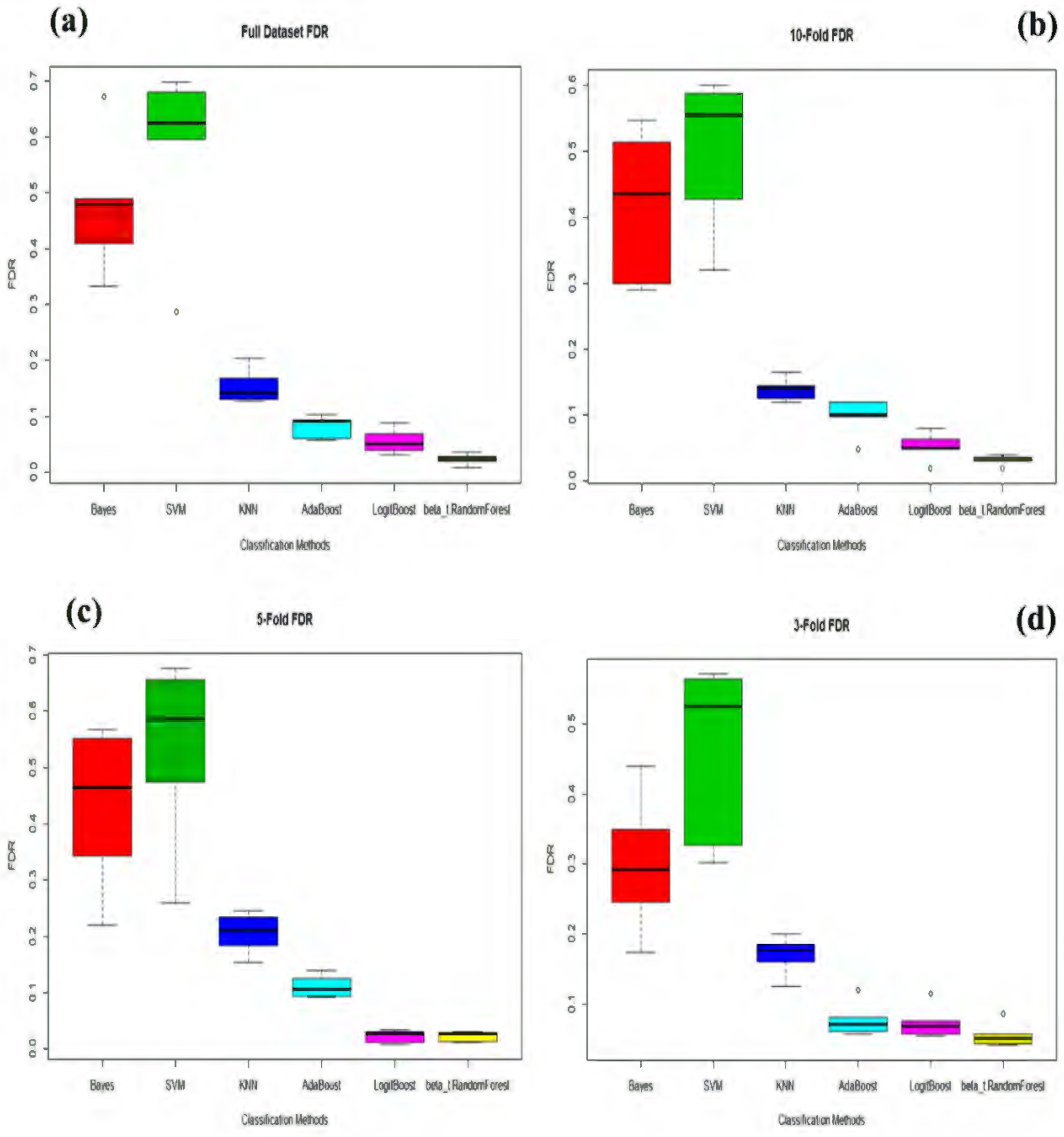


**Fig.6.4.** False discovery rate (a) Full dataset, (b) 10-fold, (c) 5-fold and (d) 3-fold cross validation for different classification methods of metagenome dataset.

The false discovery rate (FDR) was calculated for each of the dataset. **Fig.6.4** illustrates that SVM produced largest FDR for all datasets followed by Bayes classifier and KNN. On the other hand, among these six classifiers, the beta-*t* Random Forest produced lowest FDR to classify the functional metagenomes from several microbial communities.

The radar plot (**Fig.6.5**) shows the different performance measurement methods for popular classifiers in the literature to classify the functional metagenomes from the different microbial community. The beta-t Random Forest classifier showed the highest TPR and lowest FDR and MER for classification of the metagenomes.

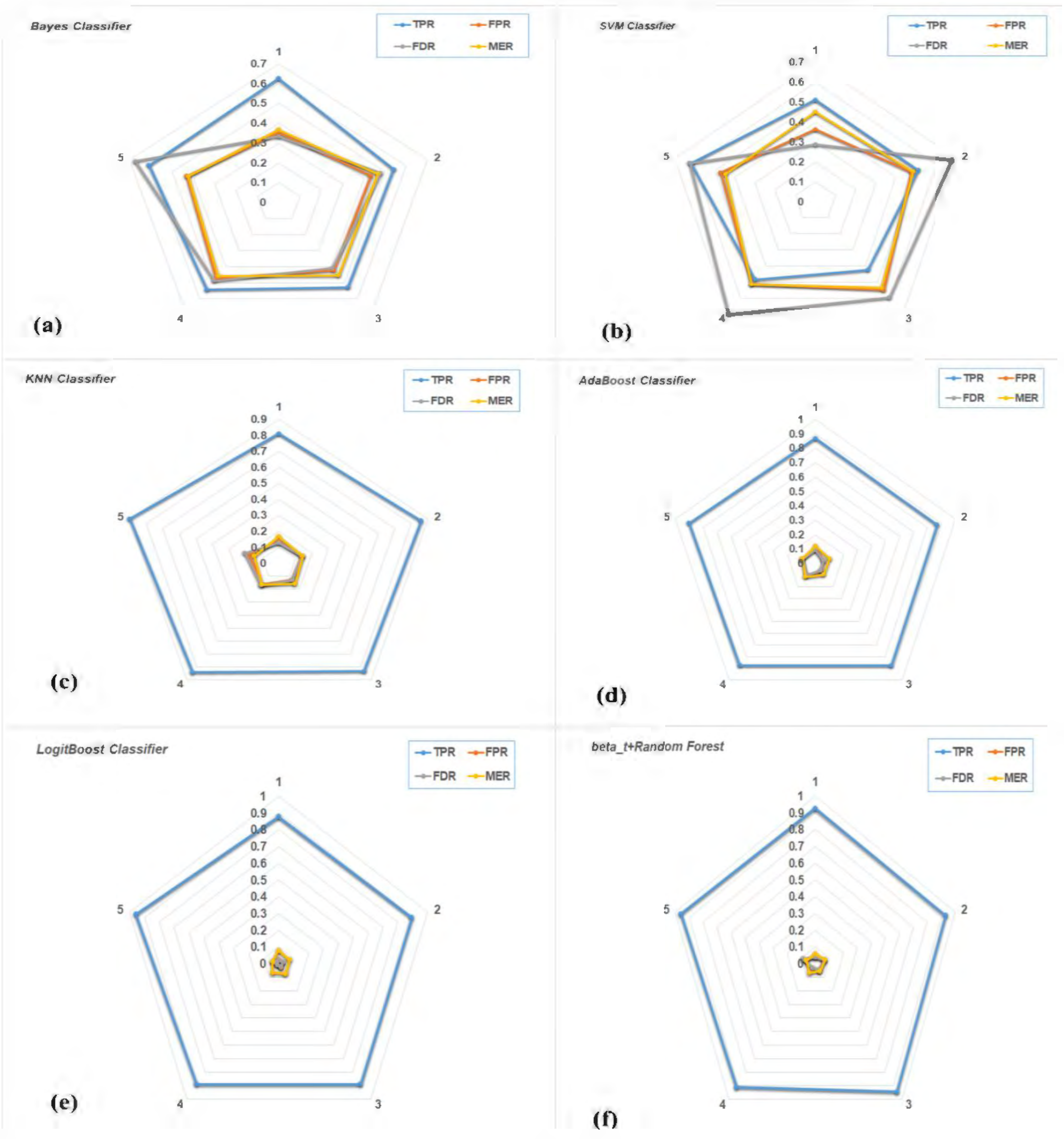


**Fig.6.5.** The average classification performances of (a) Bayes, (b) SVM, (c) KNN, (d) AdaBoost, (e) LogitBoost and (f) beta-t+Random Forest classification methods for metagenome dataset.

## 6.4 Summary of the Chapter

Classification of the metagenomic data obtained from different microbial community is an important task in the context of their associated functional metagenomic variables. In this study the beta-t random forest classifier showed the lowest FDR and MER along with highest TPR in all cases of data compared to Bayes, SVM, KNN, AdaBoost and LogitBoost classifiers. Therefore, the beta-t random forest classifier is considered the best classifier in grouping the metagenomes derived from different environmental samples.

# CHAPTER SEVEN

## DISCUSSION, CONCLUSION AND FUTURE RESEARCH DIRECTION

# DISCUSSION, CONCLUSION AND FUTURE RESEARCH DIRECTION

## 7.1 Discussion and Conclusion

There are several statistical methods for the analysis of various types of high-dimensional genomic datasets including (i) phenotypic-genotypic (ii) microarray (iii) DNA-Seq (iv) RNA-Seq (v) SNPs and (vi) metagenomic datasets. There are several statistical algorithms for each types of data analysis. But in some situations, performances of some algorithms are not so satisfactory label. So appropriate method selection is very important to achieve the better results from the data analysis as early mentioned. In this thesis, we proposed some methods suitable for the data analysis mentioned above. Comprehensive discussions have been presented in **Chapter One** in the perspective of different types of genomic data analysis. In addition, a brief overview has been stated afterward in the aspect of statistical and bioinformatic analyses for different genomic dataset.

QTL mapping technique for genetic linkage analysis is a part of genomics helps largely conventional plant breeders in genetics to identify the precise location of the makers in the chromosome linked quantitatively to certain phenotypic traits. A qtl may be single gene/maker or a cluster of genes. There are some challenges in qtl mapping such as suitable selection of statistical approach for optimum detection of the makers of interest in presence of single locus or multiple locus in consideration of linked and unlinked situation. There are a number of statistical approaches to test the presence of a qtl in the interval of two adjacent makers using LOD scores, which is called Standard Interval Mapping (SIM). However, SIM can bias identification and estimation of qtls when multiple qtls are located in the same linkage group. To deal with these complexities, qtl mapping combines SIM with the multiple marker regression analysis termed as Composite Interval Mapping (CIM). In **Chapter Two,** a comparison between four SIM approaches and CIM has been showed to investigate the performance of optimum detection of qtl location in different chromosome using LOD score. The investigation of this comparative study suggests that the Composite Interval Mapping (CIM) performs significantly better than the other four Simple Interval Mapping (SIM) methods in detecting qtl positions in backcross technique both on simulated data and on real rice dataset. CIM detected three makers in chromosome two and four, as well as other four SIM methods were unable in detecting qtls for each of the four chromosomes for simulated data. In addition, for a real rice data set from backcross population, the CIM performs mostly in similar fashion for detecting qtls in different positions in each of the 7 chromosomes. CIM were finally able to detect twelve qtls above the LOD threshold 3.0 whereas other SIM methods identified only six marker positions.

Usually transcriptomics data (microarray and RNA-Seq) analysis requires identification of differentially expressed (DE) genes between two or more conditions and classification/clustering of samples (DE genes) based on the DE genes (samples). There are several statistical methods for these types of data analysis. However, most of them are suffering from the small sample size and outlying observations. So transcriptomics data analysis by most of the conventional algorithms might be produced misleading results, since transcriptomics datasets are also often contaminated by outlying observations due to several steps involve in the data generating processes. To overcome these problems, we proposed logistic transformation of transcriptomics data for (i) robust identification of DE genes by SAM approach and (ii) robust classification of samples (DE genes) based on the reduced DE gene-set (samples) **in chapter-3**. Simulation and real transcriptomics data analysis results showed that the proposed procedure outperform over the conventional procedure in presence of outliers, otherwise it keeps almost equal performance. It should be mention here that in the case of real rice genome data (control vs blast fungus disease) analysis, our proposed method detected two additional genes that were significantly associated with the rice blast fungus disease. This report is also supported by the literature review.

In molecular biological research that is in functional genomics research, RNA silencing is an important molecular phenomenon occurs in eukaryotic groups. It is a way of mechanism so that a small RNA molecule interferes with a particular nucleotide sequence nearly size of 21-24 nucleotides that are produced in multi-cellular eukaryotes termed as microRNA (miRNA) and short interfering RNA (siRNA). These RNA molecules contribute in performing different biological and molecular processes at development and growth level as well as work in metabolism, anti-viral and anti-bacterial defense activities.

DCL, AGO and RDR are called RNA silencing machinery genes or RNA interference (RNAi) genes play significant role in the regulation of gene expression through the generation of small RNA (sRNA) molecules in plants. These genes are involved in the activities of RNA interference (RNAi) pathway by restricting the transcript accumulation in cells. In **Chapter Four,** it has been showed that the wheat genome possessed 7 DCL, 39 AGO and 16 RDR RNAi genes. The objective of this study was genome-wide identification and characterization of these genes and analyses of expression pattern of seven TaDCLs using qRT-PCR. The phylogenetic investigation implied that all subfamilies of these three gene sets maintain their evolutionary relationships similar to their rice and Arabidopsis counterpart. First subfamily of DCL, AGO and RDR possessed the multiple copies of genes higher than that of corresponding rice and Arabidopsis DCL, AGO and RDR RNAi genes. Conserved domain structure analysis suggested that these genes also contained consistent domain structure similar to rice and Arabidopsis. Although there was a difference in possessing *cis*-acting elements of the promoter regions of 7 TaDCL genes but the promoters of TaDCL1a, TaDCL3a, and TaDCL4 contained a large number of stress related *cis*-elements. GO analysis identified different metabolic process, biosynthetic process, molecular binding such as RNA, nucleic acid, protein binding, post transcriptional, transferase and nuclease activities that are projected to significantly connected to RNAi gene regulation in wheat(*Triticum aestivum*). Cytosol is however found to be the key molecular component that possesses the maximum number of wheat RNA silencing proteins. Expression analyses indicated that TaDCL3 and TaDCL4 genes are likely to play distinct roles in development and drought stress tolerance. This work provides the important indication of DCL, AGO and RDR genes evolutionary resemblances of their rice and Arabidopsis counterpart. These results may however provide valuable sources for further biological implementation and justification to draw more specific conclusion regarding any particular gene(s) and its domain of activity against different biotic and abiotic stresses as well as growth and development in plants and animals.

Genome-wide association studies (GWAS) have been popular for the identification of nucleotide sequence variants (SNP biomarkers) underlying certain trait(s) disease(s) of interest in different plant and animals. The inconsistency of the results across various GWAS might be attributed to hidden population structure and possible genetic relatedness. False discovery might be a possible case that might arise due to population stratification; that is, the different allele frequencies between cases and controls are attributed to false genetic associations caused by heterogeneity of populations. Several statistical methods and computational software tools have been developed so far to solve these problems. Recent improvement in the correction for population stratification and genetic relatedness was achieved by employing a linear mixed model (LMM) approach to a large-scale GWA. When associations between nucleotide variants and traits of interest are tested, the LMM methodology reflects the polygenic effects explained by the genetic relationships among individuals using genomic information. Use of LMM methodology decreases the inflation of FDR, increases the statistical power. It is now commonly used by the genetics researchers or bioinformaticians for large-scale GWA analysis in terms of existing different population structure and polygenic effects. However, the GWAS might be misleading using this LMM technique if there are subjects outliers in the data. In **Chapter Five**, we hereafter proposed a robust LMM methodology for handling outliers and minimizing the differential effects of the population structures and genetic relatedness in GWA. We introduced a $\beta$-weight function termed as minimum $\beta$- divergence method accompanied by a tuning parameter $\beta$ to solve the problem of data contamination. The proposed approach performed better in terms of lower FDR and higher statistical power compared to LRM and LMM at varying subject outliers against two heritability proportions 0.2 and 0.3.

We also identified 11 rice SNP maker genes using the proposed method. Genome-wide identification and characterization of these genes provided biological insights such as their possible molecular functional pathways, availability in certain molecular locations and lastly their expression patterns in various rice plant organs. It was observed that the gene LOC_Os06g18000 play functional roles in flower development and for response to stress in rice. Amongst the 11 genes, LOC_Os02g21880, LOC_Os06g18000, LOC_Os02g24134 revealed higher expression in seedling, vascular cell, root, shoot, and panicle. Besides, the gene LOC_Os08g25060 is predicted to provide maximum expression in vascular cell, root, and panicle. Our results also found that the cytoplasm (cytos) contains the maximum number of gene markers. Plastid is an important molecular organ found in plant cell mostly involve in photosynthesis and other gene expression. Photosynthesis is the chief physiological factor in rice links eventually in many parts to increase the rice yield. Our analysis results of SCL shows that the expression of the gene LOC_Os06g18000 in plastid may work as flowering promoter. This gene expression in plastid likely to boost the photosynthesis procedure, which is predicted to regulate the leaf anatomy for prior flowering in rice. From GSEA, it is also observable that this gene expression is connected to flowering in rice. It is however concluded that data contamination or presence of outliers may significantly influence the analysis results in GWAS. Our proposed robust approach outperform the existing LRM and LMM methods in presence of subject outliers and the resulting genomic information presented might provide substantial source for more in-depth biological research for the development of more rice important traits.

Metagenomics is one of the promising branch in genome research for studying the microbial community available in different environments. Proper classification of the functional metagenomes with respect to their functional variables using suitable statistical tools is however a major issue for metagenomic researchers. The different microbial community possesses different metagenomic function or pathways in terms of their associated functional metagenomic variables obtained from several environments. In **Chapter Six**, the beta-t random forest classifier showed the lowest FDR and MER along with highest TPR in all cases of data compared to Bayes, SVM, KNN, AdaBoost and LogitBoost classifiers. Therefore, the beta-t random forest classifier is considered to the best classifier in grouping the metagenomes derived from different environmental samples.

## 7.2 Areas of Future Research

Extensive various genome data are continuously generated by the bioinformatics and biotechnological researchers. Suitable computational techniques or methods are needed to select properly for valid genomic analysis and to draw precise conclusions in identifying potential gene biomarkers associated to particular phenotypic variation(traits of interests) or causing specific harmful disease(s). The following are the future research areas related to genome data analysis:

- Development of robust statistical method(s) and software tools for multiple qtl mapping for large-scale gene markers.
- Application of the proposed logistic transformation based SAM and SVM procedure to identify agricultural biomarkers from RNA-Seq or microarray gene expression profile.
- In-depth expression profile analysis for the validation of the predicted RNAi genes of AGO, DCL and RDR against different biotic and abiotic stresses in potential organs in wheat (Triticum aestivum).
- Application of the proposed robust linear mixed model (LMM) to identify Agricultural biomarker SNPs.

# CHAPTER EIGHT
# BIBLIOGRAPHY

# BIBLIOGRAPHY

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*. doi:10.1007/s11749-015-0450-6.

Ahsan, A., Monir, M., Meng, X., Rahaman, M., Chen, H., and Chen, M. (2018). Identification epistasis loci underlying rice flowering time by controlling population stratification and polygenic effect. *DNA Research*. doi:10.1093/dnares/dsy043.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., et al. (2017). *Molecular Biology of the Cell*. doi:10.1201/9781315735368.

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Annals of Statistics*. doi:10.1214/07-AOS588.

Anderson, N. L., and Anderson, N. G. (1998). Proteome and proteomics: New technologies, new concepts, and new words. in *Electrophoresis* doi:10.1002/elps.1150191103.

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*. doi:10.1371/journal.pgen.0010060.

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., et al. (2012). METAGENassist: A comprehensive web server for comparative metagenomics. *Nucleic Acids Research*. doi:10.1093/nar/gks497.

Assaad, F. F., Huet, Y., Mayer, U., and Jürgens, G. (2001). The cytokinesis gene KEULE encodes a Sec1 protein that binds the syntaxin KNOLLE. *Journal of Cell Biology*.

Aßhauer, K. P., Klingenberg, H., Lingner, T., and Meinicke, P. (2014). Exploring neighborhoods in the metagenome universe. *International Journal of Molecular Sciences*. doi:10.3390/ijms150712364.

Atkinson, A. C. (2018). Regression Diagnostics, Transformations and Constructed Variables. *Journal of the Royal Statistical Society: Series B (Methodological)*. doi:10.1111/j.2517-6161.1982.tb01181.x.

Bai, M., Yang, G. S., Chen, W. T., Mao, Z. C., Kang, H. X., Chen, G. H., et al. (2012). Genome-wide identification of Dicer-like, Argonaute and RNA-dependent RNA polymerase gene families and their expression analyses in response to viral infection and abiotic stresses in Solanum lycopersicum. *Gene* 501, 52–62. doi:10.1016/j.gene.2012.02.009.

Baulcombe, D. (2004). RNA silencing in plants.Baulcombe, D. (2004). RNA silencing in plants. Nature, 431(7006), 356–363. https://doi.org/10.1038/nature02874. *Nature*. doi:10.1038/nature02874.

Baumberger, N., and Baulcombe, D. C. (2005). Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0505461102.

Berg, J., Tymoczko, J., and Stryer, L. (2002). *Biochemistry. 5th Edition, New York: W H Freeman*.

Bologna, N. G., and Voinnet, O. (2014). The Diversity, Biogenesis, and Activities of Endogenous Silencing Small RNAs in *Arabidopsis*. *Annual Review of Plant Biology*. doi:10.1146/annurev-arplant-050213-035728.

Box, G. E. P., and Cox, D. R. (1964). An Analysis of Transformations An Analysis of Transformations. *Analysis*.

Boyle, J. (2008). Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Biochemistry and Molecular Biology Education*. doi:10.1002/bmb.20192.

Brasileiro, A. C. M., Morgante, C. V., Araujo, A. C. G., Leal-Bertioli, S. C. M., Silva, A. K., Martins, A. C. Q., et al. (2015). Transcriptome Profiling of Wild Arachis from Water-Limited Environments Uncovers Drought Tolerance Candidate Genes. *Plant Molecular Biology Reporter*. doi:10.1007/s11105-015-0882-x.

Breiman, L. (2001). Random Forrest. *Machine Learning*. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification and regression trees*. doi:10.1201/9781315139470.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: qtl mapping in experimental crosses. *Bioinformatics*. doi:10.1093/bioinformatics/btg112.

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*. doi:10.1038/ng1607.

Cao, J.-Y., Xu, Y.-P., Li, W., Li, S.-S., Rahman, H., and Cai, X.-Z. (2016). Genome-Wide Identification of Dicer-Like, Argonaute, and RNA-Dependent RNA Polymerase Gene Families in Brassica Species and Functional Analyses of Their Arabidopsis Homologs in Resistance to Sclerotinia sclerotiorum. *Frontiers in Plant Science* 7, 1–17. doi:10.3389/fpls.2016.01614.

Cardon, L. R., and Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet*. doi:10.1016/S0140-6736(03)12520-2.

Carmell, M. A., and Hannon, G. J. (2004). RNase III enzymes and the initiation of gene silencing. *Nature Structural and Molecular Biology*. doi:10.1038/nsmb729.

Carmell, M. A., Xuan, Z., Zhang, M. Q., and Hannon, G. J. (2002). The Argonaute family: Tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes and Development*. doi:10.1101/gad.1026102.

Carrington, J. C., and Ambros, V. (2003). Role of MicroRNAs in Plant and Animal Development. *Science* 301, 336 LP – 338. doi:10.1126/science.1085242.

Carvalho, T. F. M., Silva, J. C. F., Calil, I. P., Fontes, E. P. B., and Cerqueira, F. R. (2017). Rama: A machine learning approach for ribosomal protein prediction in plants. *Scientific Reports*. doi:10.1038/s41598-017-16322-4.

Chapman, E. J., and Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nature Reviews Genetics*. doi:10.1038/nrg2179.

Chen, L., Song, Y., Li, S., Zhang, L., Zou, C., and Yu, D. (2012). The role of WRKY transcription factors in plant abiotic stresses. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. doi:10.1016/j.bbagrm.2011.09.002.

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., et al. (2016). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science*. doi:10.1126/science.aad7977.

Chen, X. (2012). Small RNAs in development - insights from plants. *Current Opinion in Genetics and Development*. doi:10.1016/j.gde.2012.04.004.

Cheng, M.-C., Liao, P.-M., Kuo, W.-W., and Lin, T.-P. (2013). The Arabidopsis ETHYLENE RESPONSE FACTOR1 Regulates Abiotic Stress-Responsive Gene Expression by Binding to Different cis-Acting Elements in Response to Different Stress Signals. *PLANT PHYSIOLOGY*. doi:10.1104/pp.113.221911.

Dangl, J. L., and Jones, J. D. G. (2001). Plant pathogens and integrated defence responses to infection. *Nature*. doi:10.1038/35081161.

De'ath, G. (2002). Multivariate regression trees: A new technique for modeling species-environment relationships. *Ecology*. doi:10.1890/0012-9658 (2002) 083 [1105:MRTANT]2.0.CO;2.

De'Ath, G., and Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Deleris, A., Gallago-Bartolome, J., Bao, J., Kasschau, K. D., Carrington, J. C., and Voinnet, O. (2006). Hierarchical action and inhibition of plant dicer-like proteins in antiviral defense. *Science*. doi:10.1126/science.1128214.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics*. 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*. doi:10.1006/tpbi.2001.1542.

Diaz-Pendon, J. A., Li, F., Li, W.-X., and Ding, S.-W. (2007). Suppression of Antiviral Silencing by Cucumber Mosaic Virus 2b Protein in *Arabidopsis* Is Associated with Drastically Reduced Accumulation of Three Classes of Viral Small Interfering RNAs. *The Plant Cell*. doi:10.1105/tpc.106.047449.

Dinsdale, E. A., Edwards, R. A., Bailey, B. A., Tuba, I., Akhter, S., McNair, K., et al. (2013). Multivariate analysis of functional metagenomes. *Frontiers in Genetics*. doi:10.3389/fgene.2013.00041.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., et al. (2008). Functional metagenomic profiling of nine biomes (Nature (2008) 452, (629-632 )). *Nature*. doi:10.1038/nature07346.

Do, K. A., Müller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society. Series C: Applied Statistics*. doi:10.1111/j.1467-9876.2005.05593.x.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*. doi:10.1038/35103511.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*. doi:10.1198/016214501753382129.

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal*. doi:10.3835/ plantgenome 2011.08.0024.

Fagard, M., Boutet, S., Morel, J.-B., Bellini, C., and Vaucheret, H. (2000). AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.200217597.

Fahlgren, N., Montgomery, T. A., Howell, M. D., Allen, E., Dvorak, S. K., Alexander, A. L., et al. (2006). Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA Affects Developmental Timing and Patterning in Arabidopsis. *Current Biology*. doi:10.1016/j.cub.2006.03.065.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: New players in cell differentiation and development. *Nature Reviews Genetics*. doi:10.1038/nrg3606.

Finnegan, E. J., and Matzke, M. A. (2003). The small RNA world. *Journal of Cell Science* 116, 4689 LP – 4693. doi:10.1242/jcs.00838.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature*. doi:10.1038/35888.

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*. doi:10.1038/s41587-018-0009-7.

Gan, D., Liang, D., Wu, J., Zhan, M., Yang, F., Xu, W., et al. (2016). Genome-Wide Identification of the Dicer-Like, Argonaute, and RNA-Dependent RNA Polymerase Gene Families in Cucumber (Cucumis sativus L.). *Journal of Plant Growth Regulation*. doi:10.1007/s00344-015-9514-9.

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*. doi:10.1111/j.1541-0420.2005.00397.x.

Group, N. P. (2006). WHAT IS A GENE ? ‘ G. *Nature*.

Guiderdoni, E., Galinato, E., Luistro, J., and Vergara, G. (1992). Anther culture of tropical japonica ?? indica hybrids of rice (Oryza sativa L.). *Euphytica*. doi:10.1007/BF00041756.

Haley, C. S., and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*. doi:10.1038/hdy.1992.131.

Hannon, G. J., Rivas, F. V., Murchison, E. P., and Steitz, J. A. (2006). The expanding universe of noncoding RNAs. in *Cold Spring Harbor Symposia on Quantitative Biology* doi:10.1101/sqb.2006.71.064.

Henderson, I. R., Zhang, X., Lu, C., Johnson, L., Meyers, B. C., Green, P. J., et al. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genetics*. doi:10.1038/ng1804.

Hidayati, N., Triadiati, and Anas, I. (2016). Photosynthesis and Transpiration Rates of Rice Cultivated Under the System of Rice Intensification and the Effects on Growth and Yield. *HAYATI Journal of Biosciences*. doi:10.1016/j.hjb.2016.06.002.

Höck, J., and Meister, G. (2008). The Argonaute protein family. *Genome Biology*. doi:10.1186/gb-2008-9-2-210.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide asociation studies of 14 agronomic traits in rice landraces. *Nature Genetics*. doi:10.1038/ng.695.

Hunter, L. J. R., Brockington, S. F., Murphy, A. M., Pate, A. E., Gruden, K., Macfarlane, S. A., et al. (2016). RNA-dependent RNA polymerase 1 in potato (Solanum tuberosum) and its relationship to other plant RNA-dependent RNA polymerases. *Scientific Reports* 6, 1–11. doi:10.1038/srep23082.

Hutvagner, G., and Simard, M. J. (2008). Argonaute proteins: Key players in RNA silencing. *Nature Reviews Molecular Cell Biology*. doi:10.1038/nrm2321.

Hyun, M. K., 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*. doi:10.1534/genetics.107.080101.

Jansen, R. C., Van Ooijen, J. W., Stam, P., Lister, C., and Dean, C. (1995). Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. *Theoretical and Applied Genetics*. doi:10.1007/BF00220855.

Jansen, R. K., Raubeson, L. A., Boore, J. L., DePamphilis, C. W., Chumley, T. W., Haberle, R. C., et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology*. doi:10.1016/S0076-6879(05)95020-9.

Jiang, Y., Wang, J., Xia, D., and Yu, G. (2017). EnSVMB: Metagenomics Fragments Classification using Ensemble SVM and BLAST. *Scientific Reports*. doi:10.1038/s41598-017-09947-y.

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*. doi:10.1093/nar/gkw982.

Kaloshian, I. (2004). Gene-for-gene disease resistance: Bridging insect pest and pathogen defense. *Journal of Chemical Ecology*. doi:10.1007/s10886-004-7943-1.

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*. doi:10.1038/ng.548.

Kao, C.-H., and Zeng, Z.-B. (2006). General Formulas for Obtaining the MLEs and the Asymptotic Variance- Covariance Matrix in Mapping Quantitative Trait Loci When Using the EM Algorithm. *Biometrics*. doi:10.2307/2533965.

Kao, C. H. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*.

Kapoor, M., Arora, R., Lama, T., Nijhawan, A., Khurana, J. P., Tyagi, A. K., et al. (2008). Genome-wide identification, organization and phylogenetic analysis of Dicer-like, Argonaute and RNA-dependent RNA Polymerase gene families and their expression analysis during reproductive development and stress in rice. *BMC Genomics* 9, 1–17. doi:10.1186/1471-2164-9-451.

Karki, S., Rizal, G., and Quick, W. P. (2013). Improvement of photosynthesis in rice (Oryza sativa L.) by inserting the C4 pathway. *Rice*. doi:10.1186/1939-8433-6-28.

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*. doi:10.1002/sim.1548.

Kennedy, B. W., Quinton, M., and van Arendonk, J. A. (1992). Estimation of effects of single genes on quantitative traits. *Journal of animal science*. doi:10.2527/1992.7072000x.

Kholodenko, B. N. (2003). Four-dimensional organization of protein kinase signaling cascades: the roles of diffusion, endocytosis and molecular motors. *Journal of Experimental Biology*. doi:10.1242/jeb.00298.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*. doi:10.1038/nrg.2015.17.

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

Kumar, S., Stecher, G., Evolution, K. T.-M. biology and, and 2016, U. (2015). {MEGA}7: {Molecular} {Evolutionary} {Genetics} {Analysis} {Version} 7.0 for {Bigger} {Datasets}. *Molecular Biology and Evolution*. doi:10.1093/molbev/msw054.

Lai, E. C. (2003). microRNAs: Runts of the Genome Assert Themselves. *Current Biology*. doi:10.1016/j.cub.2003.11.017.

Lander, E. S., and Botstein, S. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*.

Lee, Y. S., and An, G. (2015). Regulation of flowering time in rice. *Journal of Plant Biology*. doi:10.1007/s12374-015-0425-x.

Lehti-Shiu, M. D., and Shiu, S. H. (2012). Diversity, classification and function of the plant protein kinase superfamily. *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi:10.1098/rstb.2012.0003.

Li, H., Ye, G., and Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics*. doi:10.1534/genetics.106.066811.

Li, J., Zhong, W., Li, R., and Wu, R. (2014). A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *Annals of Applied Statistics*. doi:10.1214/14-AOAS771.

Li, Q., and Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*. doi:10.1002/gepi.20296.

Lingel, A., and Izaurralde, E. (2004). RNAi: Finding the elusive endonuclease. *RNA*. doi:10.1261/rna.7175704.

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*. doi:10.1093/bioinformatics/bts444.

Liu, L., Zhang, D., Liu, H., and Arendt, C. (2013a). Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics*. doi:10.1186/1471-2105-14-132.

Liu, L., Zhang, Z., Mei, Q., and Chen, M. (2013b). PSI: A Comprehensive and Integrative Approach for Accurate Plant Subcellular Localization Prediction. *PLoS ONE*. doi:10.1371/journal.pone.0075826.

Liu, Y., Guo, J., and Zhu, H. (2011). Gene prediction in metagenomic fragments based on the SVM algorithm. in *Proceedings - 2011 4th International Conference on Biomedical Engineering and Informatics, BMEI 2011* doi:10.1109/ BMEI.2011.6098588.

Liu, Y., Wang, L., Xing, X., Sun, L., Pan, J., Kong, X., et al. (2013c). ZmLEA3, a multifunctional group 3 LEA protein from maize (Zea mays L.), is involved in biotic and abiotic stresses. *Plant and Cell Physiology*. doi:10.1093/pcp/pct047.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods*. doi:10.1006/meth.2001.1262.

Luo, Z., and Chen, Z. (2007). Improperly Terminated, Unpolyadenylated mRNA of Sense Transgenes Is Targeted by RDR6-Mediated RNA Silencing in Arabidopsis. *THE PLANT CELL ONLINE*. doi:10.1105/tpc.106.045724.

Lynn, K., Fernandez,  a, Aida, M., Sedbrook, J., Tasaka, M., Masson, P., et al. (1999). The PINHEAD/ ZWILLE gene acts pleiotropically in Arabidopsis development and has overlapping functions with the ARGONAUTE1 gene. *Development (Cambridge, England).*

Manolio, T. A. (2010). Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine*. doi:10.1056/nejmra0905980.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*. doi:10.1038/ng1337.

Margis, R., Fusaro, A. F., Smith, N. A., Curtin, S. J., Watson, J. M., Finnegan, E. J., et al. (2006). The evolution and diversification of Dicers in plants. *FEBS Letters*. doi:10.1016/j.febslet.2006.03.072.

Martínez, O., and Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics*. doi:10.1007/BF00222330.

Maruyama-Nakashita, A., Nakamura, Y., Watanabe-Takahashi, A., Inoue, E., Yamaya, T., and Takahashi, H. (2005). Identification of a novel cis-acting element conferring sulfur deficiency response in Arabidopsis roots. *Plant Journal*. doi:10.1111/j.1365-313X.2005.02363.x.

Mattick, J. S. (2001). Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Reports*. doi:10.1093/embo-reports/kve230.

Mattick, J. S. (2003). Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*. doi:10.1002/bies.10332.

Mattick, J. S., and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Molecular Biology and Evolution*. doi:10.1093/ oxfordjournals.molbev.a003951.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A. J. M. (2007). Targets of RNA-directed DNA methylation. *Current Opinion in Plant Biology*. doi:10.1016/j.pbi.2007.06.007.

Molecular Biology of the Cell (4th Ed) (2002). *Journal of Biological Education*. doi:10.1080/00219266.2002.9655847.

Mollah, M. M. H., Jamal, R., Mokhtar, N. M., Harun, R., and Mollah, M. N. H. (2015). A hybrid one-way ANOVA approach for the robust and efficient estimation of differential gene expression with multiple patterns. *PLoS ONE*. doi:10.1371/journal.pone.0138810.

Mollah, M. M. H., Mollah, M. N. H., and Kishino, H. (2012). β-empirical Bayes inference and model diagnosis of microarray data. *BMC bioinformatics*.

Mollah, M. N. H., Akond, Z., and Alam, M. (2018). Biomarker Identification from RNA-Seq Data using a Robust Statistical Approach. *Bioinformation*. doi:10.6026 97320630014153.

Mollah, M. N. H., and Eguchi, S. (2008). Robust composite interval mapping for qtl analysis by minimum ??-divergence method. in *Proceedings - IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2008* doi:10.1109/ BIBM.2008.43.

Mollah, M. N. H., and Eguchi, S. (2010). Robust qtl analysis by minimum &beta;-divergence method. *International Journal of Data Mining and Bioinformatics*. doi:10.1504/ijdmb.2010.034199.

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*. doi:10.1007/s11063-006-9023-8.

Moon, S. J., Min, M. K., Kim, J. A., Kim, D. Y., Yoon, I. S., Kwon, T. R., et al. (2019). Ectopic expression of OsDREB1G, a member of the OsDREB1 subfamily, confers cold stress tolerance in rice. *Frontiers in Plant Science*. doi:10.3389/ fpls.2019.00297.

Moussian, B., Schoof, H., Haecker, A., Jürgens, G., and Laux, T. (1998). Role of the ZWILLE gene in the regulation of central shoot meristem cell fate during Arabidopsis embryogenesis. *EMBO Journal*. doi:10.1093/emboj/17.6.1799.

Newton, M. A., and Kendziorski, C. (2003). "Parametric Empirical Bayes Methods for Microarrays," in doi:10.1007/0-387-21679-0_11.

Nissen, P., Hansen, J., Ban, N., Moore, P. B., and Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science*. doi:10.1126/science.289.5481.920.

Nurul Haque Mollah, M., Sultana, N., Minami, M., and Eguchi, S. (2010). Robust extraction of local structures by the minimum β-divergence method. *Neural Networks*. doi:10.1016/j.neunet.2009.11.011.

Nyholt, D. R. (2000). All LODs are not created equal. *American journal of human genetics*. doi:10.1086/303029.

Ohlrogge, J. B., Kuhn, D. N., and Stumpf, P. K. (1979). Subcellular localization of acyl carrier protein in leaf protoplasts of Spinacia oleracea. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.76.3.1194.

Osakabe, Y., Yamaguchi-Shinozaki, K., Shinozaki, K., and Tran, L. S. P. (2014). ABA control of plant macroelement membrane transport systems in response to water deficit and high salinity. *New Phytologist*. doi:10.1111/nph.12613.

Parks, D. H., and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*. doi:10.1093/bioinformatics/btq041.

Pasam, R. K., Sharma, R., Malosetti, M., van Eeuwijk, F. A., Haseneyer, G., Kilian, B., et al. (2012). Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biology*. doi:10.1186/1471-2229-12-16.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*. doi:10.1371/journal.pgen.0020190.

Pearson, T. A., and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA - Journal of the American Medical Association*. doi:10.1001/jama.299.11.1335.

Pennisi, E. (2007). DNA study forces rethink of what it means to be a gene. *Science*. doi:10.1126/science.316.5831.1556.

Pieterse, C. M. J., and Van Loon, L. C. (2004). NPR1: The spider in the web of induced resistance signaling pathways. *Current Opinion in Plant Biology*. doi:10.1016/ j.pbi.2004.05.006.

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*. doi:10.1038/ng1847.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2002). Association Mapping in Structured Populations. *The American Journal of Human Genetics*. doi:10.1086/302959.

Qi, X., Bao, F. S., and Xie, Z. (2009). Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. *PLoS ONE*. doi:10.1371/journal.pone.0004971.

Qian, Y., Cheng, Y., Cheng, X., Jiang, H., Zhu, S., and Cheng, B. (2011). Identification and characterization of Dicer-like, Argonaute and RNA-dependent RNA polymerase gene families in maize. *Plant Cell Reports* 30, 1347–1363. doi:10.1007/s00299-011-1046-6.

Qin, L., Mo, N., Muhammad, T., and Liang, Y. (2018). Genome-wide analysis of DCL, AGO, and RDR gene families in pepper (Capsicum Annuum L.). *International Journal of Molecular Sciences*. doi:10.3390/ijms19041038.

R CoreTeam, D. C. (2017). A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. doi:10.1007/978-3-540-74686-7.

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*. doi:10.1111/j.1574-6941.2007.00375.x.

Risch, N. (1991). A note on multiple testing procedures in linkage analysis. *American journal of human genetics*.

Rivas, F. V., Tolia, N. H., Song, J. J., Aragon, J. P., Liu, J., Hannon, G. J., et al. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. *Nature Structural and Molecular Biology*. doi:10.1038/nsmb918.

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*. doi:10.1093/bioinformatics/btp616.

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., et al. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*. doi:10.1093/ bioinformatics/btn553.

Rose, A. (2004). Genome-Wide Identification of Arabidopsis Coiled-Coil Proteins and Establishment of the ARABI-COIL Database. *PLANT PHYSIOLOGY*. doi:10.1104/pp.103.035626.

Rossi, J. J. (2004). Ribozyme diagnostics comes of age. *Chemistry and Biology*. doi:10.1016/j.chembiol.2004.07.002.

Sain, S. R., and Vapnik, V. N. (2006). The Nature of Statistical Learning Theory. *Technometrics*. doi:10.2307/1271324.

Sakuma, Y., Liu, Q., Dubouzet, J. G., Abe, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2002). DNA-Binding Specificity of the ERF/AP2 Domain of Arabidopsis DREBs, Transcription Factors Involved in Dehydration- and Cold-Inducible Gene Expression. *Biochemical and Biophysical Research Communications* 290, 998–1009. doi:10.1006/BBRC.2001.6299.

Schaper, E., and Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. *New Phytologist*. doi:10.1111/nph.13184.

Schiebel, W. (1998). Isolation of an RNA-Directed RNA Polymerase Specific cDNA Clone from Tomato. *the Plant Cell Online*. doi:10.1105/tpc.10.12.2087.

Sharma, A. K., Gupta, A., Kumar, S., Dhakan, D. B., and Sharma, V. K. (2015). Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics*. doi:10.1016/j.ygeno.2015.04.001.

Sharma, M., and Pandey, G. K. (2016). Expansion and Function of Repeat Domain Proteins During Stress and Development in Plants. *Frontiers in Plant Science*. doi:10.3389/fpls.2015.01218.

Shin, J., and Lee, C. (2015). A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. *Genomics*. doi:10.1016/j.ygeno.2015.01.006.

Shivani, Awasthi, P., Sharma, V., Kaur, N., Kaur, N., Pandey, P., et al. (2017). Genome-wide analysis of transcription factors during somatic embryogenesis in banana (Musa spp.) cv. Grand Naine. *PLoS ONE*. doi:10.1371/journal.pone.0182242.

Siderowf, A., Aarsland, D., Mollenhauer, B., Goldman, J. G., and Ravina, B. (2018). Biomarkers for cognitive impairment in Lewy body disorders: Status and relevance for clinical trials. *Movement Disorders*. doi:10.1002/mds.27355.

Sijen, T., Fleenor, J., Simmer, F., Thijssen, K. L., Parrish, S., Timmons, L., et al. (2001). On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell*. doi:10.1016/S0092-8674(01)00576-1.

Simon, C., and Daniel, R. (2011). Metagenomic analyses: Past and future trends. *Applied and Environmental Microbiology*. doi:10.1128/AEM.02345-10.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*. doi:10.1038/ng.686.

Team, R. C. (2018). R: A Language and Environment for Statistical Computing. *Vienna, Austria*.

Thireault, C., Shyu, C., Yoshida, Y., St. Aubin, B., Campos, M. L., and Howe, G. A. (2015). Repression of jasmonate signaling by a non-TIFY JAZ protein in Arabidopsis. *Plant Journal*. doi:10.1111/tpj.12841.

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*. doi:10.1093/nar/22.22.4673.

Todorovska, E. (2007). Retrotransposons and their role in plant—genome evolution. *Biotechnology and Biotechnological Equipment*. doi:10.1080/ 13102818.2007. 10817464.

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 of the United States of America*. doi:10.1073/pnas.091062498.

Ulmasov, T., Hagen, G., and Guilfoyle, T. J. (1997). ARF1, a transcription factor that binds to auxin response elements. *Science*. doi:10.1126/science.276.5320.1865.

Van Ex, F., Jacob, Y., and Martienssen, R. A. (2011). Multiple roles for small RNAs during plant reproduction. *Current Opinion in Plant Biology*. doi:10.1016/ j.pbi.2011.07.003.

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: Mechanisms and regulations. *Genes and Development*. doi:10.1101/gad.1410506.

Vaucheret, H. (2008). Plant ARGONAUTES. *Trends in Plant Science*. doi:10.1016/ j.tplants.2008.04.007.

Vaucheret, H., Vazquez, F., Crété, P., and Bartel, D. P. (2004). The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. *Genes and Development*. doi:10.1101/ gad.1201404.

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with \proglang{S}*. doi:10.1016/ j.electacta.2013.08.022.

Wang, K., and Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology*. doi:10.1002/ gepi.20266.

Wang, Y., Wu, C., Ji, Z., Wang, B., and Liang, Y. (2011). Non-parametric change-point method for differential gene expression detection. *PLoS ONE*. doi:10.1371 /journal.pone.0020060.

Wei, H., Zhou, B., Zhang, F., Tu, Y., Hu, Y., Zhang, B., et al. (2013). Profiling and Identification of Small rDNA-Derived RNAs and Their Potential Biological Functions. *PLoS ONE*. doi:10.1371/journal.pone.0056842.

Weng, X., Wang, L., Wang, J., Hu, Y., Du, H., Xu, C., et al. (2014). Grain Number, Plant Height, and Heading Date7 Is a Central Regulator of Growth, Development, and Stress Response. *PLANT PHYSIOLOGY*. doi:10.1104/pp.113.231308.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. doi:10.2307/3001968.

Williams, M. E. (1992). Sequences Flanking the Hexameric G-Box Core CACGTG Affect the Specificity of Protein Binding. *THE PLANT CELL ONLINE*. doi:10.1105/ tpc.4.4.485.

Wirta, V. (2006). *Mining the transcriptome methods and applications Valtteri Wirta*.

Wold, B., and Myers, R. M. (2008). Sequence census methods for functional genomics. *Nature Methods*. doi:10.1038/nmeth1157.

Xu, H., Sarkar, B., and George, V. (2009). A new measure of population structure using multiple single nucleotide polymorphisms and its relationship with FST. *BMC Research Notes*. doi:10.1186/1756-0500-2-21.

Xu, S. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics*.

Yang, T., and Poovaiah, B. W. (2002). A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. *The Journal of biological chemistry* 277, 45049–58. doi:10.1074/jbc.M207941200.

Yu, D., Fan, B., MacFarlane, S. A., and Chen, Z. (2003). Analysis of the Involvement of an Inducible *Arabidopsis* RNA-Dependent RNA Polymerase in Antiviral Defense. *Molecular Plant-Microbe Interactions*. doi:10.1094/MPMI.2003.16.3.206.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., 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*. doi:10.1038/ng1702.

Zararsiz, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G. E., Duru, I. P., et al. (2017). A comprehensive simulation study on classification of RNA-Seq data. *PLoS ONE*. doi:10.1371/journal.pone.0182507.

Zaratiegui, M., Irvine, D. V., and Martienssen, R. A. (2007). Noncoding RNAs and Gene Silencing. *Cell*. doi:10.1016/j.cell.2007.02.016.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*.

Zeng, Z. B. (2006). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.90.23.10972.

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*. doi:10.1093/nar/gki475.

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*. doi:10.1038/ng.546.

Zhao, H., Zhao, K., Wang, J., Chen, X., Chen, Z., Cai, R., et al. (2015). Comprehensive Analysis of Dicer-Like, Argonaute, and RNA-dependent RNA Polymerase Gene Families in Grapevine (Vitis Vinifera). *Journal of Plant Growth Regulation*. doi:10.1007/s00344-014-9448-7.

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*. 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*. doi:10.1038/ncomms1467.

Zhao, X., Zheng, W., Zhong, Z., Chen, X., Wang, A., and Wang, Z. (2016). Genome-wide analysis of RNA-interference pathway in Brassica napus, and the expression profile of BnAGOs in response to Sclerotinia sclerotiorum infection. *European Journal of Plant Pathology* 146, 565–579. doi:10.1007/s10658-016-0942-6.

# APPENDIX

**Table A3.1** GO Analysis of the Rice Blast Disease Genes obtained from Classical SAM

| GO | Go Terms | Count | P-Value | List Total | FDR |
|---|---|---|---|---|---|
| Biological Process(BP) | | | | | |
| GO:0006351 | transcription, DNA-templated | 47 | 1.17E-07 | 243 | 1.51E-04 |
| GO:1903507 | negative regulation of nucleic acid-templated transcription | 7 | 3.10E-07 | 243 | 4.01E-04 |
| GO:0031347 | regulation of defense response | 8 | 6.23E-07 | 243 | 8.07E-04 |
| GO:0009611 | response to wounding | 8 | 1.21E-06 | 243 | 0.001571 |
| GO:2000022 | regulation of jasmonic acid mediated signaling pathway | 7 | 2.06E-06 | 243 | 0.002665 |
| GO:0006355 | regulation of transcription, DNA-templated | 35 | 3.29E-04 | 243 | 0.425384 |
| GO:0006950 | response to stress | 8 | 0.002786 | 243 | 3.548868 |
| GO:0009620 | response to fungus | 3 | 0.003502 | 243 | 4.441429 |
| GO:0006952 | defense response | 12 | 0.017226 | 243 | 20.1491 |
| GO:0006032 | chitin catabolic process | 5 | 0.024223 | 243 | 27.20668 |
| GO:0002238 | response to molecule of fungal origin | 2 | 0.048614 | 243 | 47.5521 |
| GO:0016998 | cell wall macromolecule catabolic process | 3 | 0.051298 | 243 | 49.4366 |
| GO:0030001 | metal ion transport | 6 | 0.053503 | 243 | 50.93721 |
| GO:0080163 | regulation of protein serine/threonine phosphatase activity | 3 | 0.064357 | 243 | 57.74436 |
| GO:0009738 | abscisic acid-activated signaling pathway | 5 | 0.065937 | 243 | 58.65923 |
| GO:0006629 | lipid metabolic process | 6 | 0.068131 | 243 | 59.89871 |
| GO:0006040 | amino sugar metabolic process | 3 | 0.085732 | 243 | 68.67311 |
| GO:0010200 | response to chitin | 2 | 0.094874 | 243 | 72.49583 |
| GO:0009617 | response to bacterium | 2 | 0.094874 | 243 | 72.49583 |

| GO | Go Terms | Count | P-Value | List Total | FDR |
|---|---|---|---|---|---|
| Molecular Function(MF) | | | | | |
| GO:0003700 | transcription factor activity, sequence-specific DNA binding | 33 | 6.86E-06 | 328 | 0.008628 |
| GO:0003714 | transcription corepressor activity | 7 | 1.05E-05 | 328 | 0.013256 |
| Molecular Function(MF) | | | | | |
| GO:0043565 | sequence-specific DNA binding | 21 | 2.48E-04 | 328 | 0.311556 |
| GO:0005509 | calcium ion binding | 17 | 0.001883097 | 328 | 2.344131 |
| GO:0003677 | DNA binding | 43 | 0.004683124 | 328 | 5.736283 |
| GO:0004568 | chitinase activity | 6 | 0.00879852 | 328 | 10.52547 |
| GO:0015079 | potassium ion transmembrane transporter activity | 4 | 0.033937006 | 328 | 35.24116 |
| GO:0010427 | abscisic acid binding | 3 | 0.061899886 | 328 | 55.25276 |
| GO:0004791 | thioredoxin-disulfide reductase activity | 3 | 0.061899886 | 328 | 55.25276 |
| GO:0004872 | receptor activity | 3 | 0.068999002 | 328 | 59.33235 |
| GO:0004864 | protein phosphatase inhibitor activity | 3 | 0.068999002 | 328 | 59.33235 |
| GO:0004252 | serine-type endopeptidase activity | 6 | 0.072222651 | 328 | 61.06934 |
| GO:0008061 | chitin binding | 4 | 0.086919343 | 328 | 68.1564 |
| GO:0008889 | glycerophosphodiester phosphodiesterase activity | 2 | 0.098534057 | 328 | 72.89472 |
| Cellular Component(CC) | | | | | |
| GO:0005634 | nucleus | 75 | 3.29E-05 | 321 | 0.033442 |
| GO:0016021 | integral component of membrane | 117 | 0.053722 | 321 | 42.93479 |
| GO:0005737 | cytoplasm | 38 | 0.093217 | 321 | 62.99388 |

**Table A3.2** KEGG Pathway Analysis of the Rice Blast Disease Genes obtained from Classical SAM

| | Term | Pathways | Count | P-Value | List Total | FDR |
|---|---|---|---|---|---|---|
| KEGG PATHWAY | osa04075 | Plant hormone signal transduction | 14 | 7.01E-05 | 82 | 0.070764 |
| | osa04626 | Plant-pathogen interaction | 10 | 3.82E-04 | 82 | 0.384398 |
| | osa00250 | Alanine, aspartate and glutamate metabolism | 4 | 0.057163128 | 82 | 44.78889 |
| | osa00650 | Butanoate metabolism | 3 | 0.067952152 | 82 | 50.84297 |

**Table A3.3** Gene Ontology (GO) Analysis of the Rice Blast Disease Genes obtained from Proposed SAM

| GO | Term | Count | p-value | List Total | FDR |
|---|---|---|---|---|---|
| **Biological Process(BP)** | | | | | |
| GO:0006952 | defense response | 14 | 3.25E-04 | 195 | 0.404781 |
| GO:0006950 | response to stress | 8 | 7.82E-04 | 195 | 0.97141 |
| GO:0006351 | transcription, DNA-templated | 31 | 8.43E-04 | 195 | 1.047387 |
| GO:1903507 | negative regulation of nucleic acid-templated transcription | 4 | 0.001868464 | 195 | 2.30736 |
| GO:2000022 | regulation of jasmonic acid mediated signaling pathway | 4 | 0.004191808 | 195 | 5.108122 |
| GO:0044550 | secondary metabolite biosynthetic process | 8 | 0.008348571 | 195 | 9.935508 |
| GO:0031347 | regulation of defense response | 4 | 0.008831168 | 195 | 10.48108 |
| GO:0009611 | response to wounding | 4 | 0.011276994 | 195 | 13.1996 |
| GO:0006355 | regulation of transcription, DNA-templated | 25 | 0.011731855 | 195 | 13.69673 |
| GO:0046856 | phosphatidylinositol dephosphorylation | 3 | 0.034344621 | 195 | 35.3527 |
| GO:0006633 | fatty acid biosynthetic process | 6 | 0.051015506 | 195 | 47.98261 |
| GO:0006040 | amino sugar metabolic process | 3 | 0.058298562 | 195 | 52.7518 |
| GO:0006032 | chitin catabolic process | 4 | 0.058629026 | 195 | 52.95834 |
| GO:0000272 | polysaccharide catabolic process | 3 | 0.074727253 | 195 | 62.07028 |
| GO:0009620 | response to fungus | 2 | 0.076616763 | 195 | 63.02584 |
| GO:0031408 | oxylipin biosynthetic process | 3 | 0.080495459 | 195 | 64.91836 |

| GO | Term | Count | p-value | List Total | FDR |
|---|---|---|---|---|---|
| GO:0030001 | metal ion transport | 5 | 0.081280842 | 195 | 65.29055 |
| GO:0009790 | embryo development | 2 | 0.094840196 | 195 | 71.16997 |
| **Molecular Function(MF)** | | | | | |
| GO:0008061 | chitin binding | 4 | 0.079120338 | 315 | 63.34326 |
| GO:0004497 | monooxygenase activity | 8 | 0.070743803 | 315 | 59.07065 |
| **Molecular Function(MF)** | | | | | |
| GO:0016705 | oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen | 8 | 0.057465371 | 315 | 51.35277 |
| GO | Term | Count | p-value | List Total | FDR |
| **Molecular Function(MF)** | | | | | |
| GO:0004568 | chitinase activity | 5 | 0.034037019 | 315 | 34.40254 |
| GO:0004674 | protein serine/threonine kinase activity | 22 | 0.025238508 | 315 | 26.7458 |
| GO:0020037 | heme binding | 19 | 0.020564761 | 315 | 22.35288 |
| GO:0008081 | phosphoric diester hydrolase activity | 3 | 0.019341099 | 315 | 21.16347 |
| GO:0005509 | calcium ion binding | 14 | 0.018106189 | 315 | 19.9462 |
| GO:0003714 | transcription corepressor activity | 4 | 0.01408988 | 315 | 15.86695 |
| GO:0016709 | oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen | 9 | 0.007424451 | 315 | 8.673877 |
| GO:0005506 | iron ion binding | 17 | 0.006268247 | 315 | 7.37017 |
| GO:0005524 | ATP binding | 63 | 0.005197102 | 315 | 6.147152 |
| GO:0003700 | transcription factor activity, sequence-specific DNA binding | 25 | 0.003251062 | 315 | 3.887196 |
| GO:0043565 | sequence-specific DNA binding | 18 | 0.00266318 | 315 | 3.194724 |
| GO:0004672 | protein kinase activity | 23 | 5.65E-04 | 315 | 0.685482 |
| **Cellular Component(CC)** | | | | | |
| GO:0005886 | plasma membrane | 28 | 0.004993 | 285 | |
| GO:0016021 | integral component of membrane | 129 | 1.94E-06 | 285 | |

**Table A3.4** KEGG Pathway Analysis of the Rice Blast Disease Genes obtained from Proposed SAM

| Category | Term | Pathways | Count | PValue | List Total | FDR |
|---|---|---|---|---|---|---|
| KEGG PATHWAY | osa04626 | Plant-pathogen interaction | 8 | 0.00106 | 59 | 1.007396 |
| | osa04075 | Plant hormone signal transduction | 9 | 0.0048 | 59 | 4.488238 |
| | osa00904 | Diterpenoid biosynthesis | 3 | 0.058942 | 59 | 43.99766 |

**Table A4.1.** Primer sequences of seven TaDCL genes for qRT-PCR analysis

| Gene Name | Forward Primers (5'-3') | Reverse Primers (3'-5') | Product (bp) |
|---|---|---|---|
| *TaDCL1a* | CCCTGAAAAGCCTGACGGT | CCAGAGCTGAAGAGCACTGAA | 128 |
| *TaDCL1b* | TTGCTGGTGCAGTATTCCTGG | TGGAAGGGTCTCTGGCGTTA | 101 |
| *TaDCL3a* | CGCTGACAATGCTCCACAAG | CACTGTGGGATGGAGGATCAG | 133 |
| *TaDCL3b* | TTCCCAGGATTCAGCATCGC | ATCCCCAGGTTAACGAGCCT | 129 |
| *TaDCL3c* | AGGTAGAACAAAGCACGCCT | TATCCGCAAAGCAATCCCCT | 117 |
| *TaDCL3d* | CCGTGCAGGACTGTATGAGTT | TGAACCCTTGACCTTGAGGC | 107 |
| *TaDCL4* | ATCTCTGGATTCCATGGCCC | GTCCTATCACCCATACGCCA | 128 |
| *18S* | ATCGGCGGATGTTGCTTA | TTAGCAGGCTGAGGTCTCGT | 152 |

**Table A4.2.** Percentage of three groups of RNA silencing genes involved in different cellular location

| | Names of Sub Cellular Locations in Wheat | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **extra** | **cytos** | **membr** | **ER** | **mito** | **golgi** | **plast** | **nucl** | **vacu** | **pero** |
| **DCL** | 0% | 71.4% | 14.3% | 0% | 0% | 0% | 14.3% | 0% | 0% | 0% |
| **AGO** | 2.56% | 87.2% | 7.69% | 0% | 20.5% | 0% | 33.3% | 2.56% | 2.56% | 0% |
| **RDR** | 6.25% | 87.5% | 0% | 0% | 12.5% | 0% | 31.2% | 0% | 6.25% | 0% |

extra(extracellular), cytos (cytpsol),membr (membrane);ER (endoplasmic reticulum); mito(mitochondria);golgi (golgi apparatus);plast (plastid);nucl (nuclear);vacu (vacuole);pero (peroxisome)

**Table A4.3** Gene Ontology (GO) analysis of RNA silencing machinery genes

| NO. | GO ID | Description | Annotated | Count | Expected | P-value | Genes |
|---|---|---|---|---|---|---|---|
| **Biological Process(BP)** | | | | | | | |
| 1 | GO:0001172 | transcription, RNA-templated | 51 | 13 | 0.04 | 1.00E-30 | Traes_2DL_6DB81005E,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 2 | GO:0035194 | posttranscriptional gene silencing by RNA | 61 | 11 | 0.04 | 8.10E-25 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 3 | GO:0044003 | modification by symbiont of host morphology or physiology | 21 | 9 | 0.01 | 2.70E-24 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 4 | GO:0016441 | posttranscriptional gene silencing | 71 | 11 | 0.05 | 5.00E-24 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 5 | GO:0051817 | modification of morphology or physiology of other organism involved in symbiotic interaction | 23 | 9 | 0.02 | 7.60E-24 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 6 | GO:0031050 | dsRNA fragmentation | 48 | 10 | 0.03 | 2.80E-23 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 7 | GO:0043331 | response to dsRNA | 48 | 10 | 0.03 | 2.80E-23 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 8 | GO:0070918 | production of small RNA involved in gene silencing by RNA | 48 | 10 | 0.03 | 2.80E-23 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 9 | GO:0071359 | cellular response to dsRNA | 48 | 10 | 0.03 | 2.80E-23 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 10 | GO:0051701 | interaction with host | 29 | 9 | 0.02 | 9.30E-23 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 11 | GO:0030422 | production of siRNA involved in RNA interference | 31 | 9 | 0.02 | 1.90E-22 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5DL_672EE3605 |
| 12 | GO:0016246 | RNA interference | 33 | 9 | 0.02 | 3.60E-22 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5DL_672EE3605 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 13 | GO:0051607 | defense response to virus | 34 | 9 | 0.02 | 4.80E-22 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 14 | GO:1901699 | cellular response to nitrogen compound | 64 | 10 | 0.04 | 6.60E-22 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 15 | GO:0031047 | gene silencing by RNA | 129 | 11 | 0.09 | 5.00E-21 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 16 | GO:0044403 | symbiosis, encompassing mutualism through parasitism | 44 | 9 | 0.03 | 6.50E-21 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 17 | GO:0044419 | interspecies interaction between organisms | 47 | 9 | 0.03 | 1.30E-20 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 18 | GO:0002252 | immune effector process | 51 | 9 | 0.04 | 2.80E-20 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 19 | GO:0009615 | response to virus | 52 | 9 | 0.04 | 3.40E-20 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 20 | GO:0009616 | virus induced gene silencing | 13 | 7 | 0.01 | 6.10E-20 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 | GO:0052018 | modulation by symbiont of RNA levels in host | 13 | 7 | 0.01 | 6.10E-20 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 22 | GO:0052249 | modulation of RNA levels in other organism involved in symbiotic interaction | 13 | 7 | 0.01 | 6.10E-20 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 23 | GO:0098586 | cellular response to virus | 14 | 7 | 0.01 | 1.20E-19 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 24 | GO:0016458 | gene silencing | 172 | 11 | 0.12 | 1.30E-19 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 25 | GO:0035821 | modification of morphology or physiology of other organism | 65 | 9 | 0.05 | 2.90E-19 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 26 | GO:0040029 | regulation of gene expression, epigenetic | 188 | 11 | 0.13 | 3.60E-19 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 27 | GO:0071407 | cellular response to organic cyclic compound | 118 | 10 | 0.08 | 4.10E-19 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 28 | GO:0016070 | RNA metabolic process | 4883 | 24 | 3.41 | 1.30E-18 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 29 | GO:0010608 | posttranscriptional regulation of gene expression | 213 | 11 | 0.15 | 1.40E-18 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 30 | GO:1901698 | response to nitrogen compound | 159 | 10 | 0.11 | 8.90E-18 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 31 | GO:0010629 | negative regulation of gene expression | 315 | 11 | 0.22 | 1.10E-16 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 32 | GO:0014070 | response to organic cyclic compound | 208 | 10 | 0.15 | 1.40E-16 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 33 | GO:0090304 | nucleic acid metabolic process | 5981 | 24 | 4.18 | 1.60E-16 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2D |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | C78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 34 | GO:0010267 | production of ta-siRNAs involved in RNA interference | 16 | 6 | 0.01 | 5.20E-16 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 35 | GO:0006396 | RNA processing | 916 | 13 | 0.64 | 1.20E-14 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 36 | GO:0006139 | nucleobase-containing compound metabolic process | 7349 | 24 | 5.13 | 1.90E-14 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 37 | GO:0002376 | immune system process | 231 | 9 | 0.16 | 3.70E-14 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 38 | GO:0010605 | negative regulation of macromolecule metabolic process | 565 | 11 | 0.39 | 6.80E-14 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 39 | GO:0046483 | heterocycle metabolic process | 7926 | 24 | 5.53 | 1.10E-13 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 40 | GO:0006725 | cellular aromatic compound metabolic process | 8000 | 24 | 5.58 | 1.30E-13 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 41 | GO:1901360 | organic cyclic compound metabolic process | 8145 | 24 | 5.69 | 2.00E-13 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 42 | GO:0098542 | defense response to other organism | 460 | 10 | 0.32 | 3.90E-13 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 43 | GO:0090502 | RNA phosphodiester bond hydrolysis, endonucleolytic | 98 | 7 | 0.07 | 4.70E-13 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 44 | GO:0045087 | innate immune response | 200 | 8 | 0.14 | 9.50E-13 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 45 | GO:0010599 | production of lsiRNA involved in RNA interference | 5 | 4 | 0 | 9.50E-13 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_4BL_B3A1B8342 |
| 46 | GO:0009892 | negative regulation of metabolic process | 723 | 11 | 0.5 | 9.80E-13 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 47 | GO:0071310 | cellular response to organic substance | 505 | 10 | 0.35 | 9.80E-13 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 48 | GO:0006955 | immune response | 211 | 8 | 0.15 | 1.50E-12 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 49 | GO:0090501 | RNA phosphodiester bond hydrolysis | 123 | 7 | 0.09 | 2.40E-12 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 50 | GO:0043207 | response to external biotic stimulus | 585 | 10 | 0.41 | 4.20E-12 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 51 | GO:0051707 | response to other organism | 585 | 10 | 0.41 | 4.20E-12 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 52 | GO:0070887 | cellular response to chemical stimulus | 590 | 10 | 0.41 | 4.50E-12 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 53 | GO:0009607 | response to biotic stimulus | 649 | 10 | 0.45 | 1.20E-11 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 54 | GO:0048519 | negative regulation of biological process | 917 | 11 | 0.64 | 1.30E-11 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 55 | GO:0006952 | defense response | 688 | 10 | 0.48 | 2.10E-11 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 56 | GO:0034641 | cellular nitrogen compound metabolic process | 10059 | 24 | 7.02 | 2.60E-11 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 57 | GO:0009605 | response to external stimulus | 784 | 10 | 0.55 | 7.30E-11 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 58 | GO:0097659 | nucleic acid-templated transcription | 3520 | 16 | 2.46 | 1.30E-10 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 59 | GO:0032774 | RNA biosynthetic process | 3526 | 16 | 2.46 | 1.30E-10 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 60 | GO:0006807 | nitrogen compound metabolic process | 10862 | 24 | 7.58 | 1.50E-10 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 61 | GO:0010033 | response to organic substance | 1345 | 11 | 0.94 | 7.40E-10 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 62 | GO:0060145 | viral gene silencing in virus induced gene silencing | 4 | 3 | 0 | 1.20E-09 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C |
| 63 | GO:0034654 | nucleobase-containing compound biosynthetic process | 4174 | 16 | 2.91 | 1.60E-09 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 64 | GO:0051704 | multi-organism process | 1127 | 10 | 0.79 | 2.40E-09 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 65 | GO:0090305 | nucleic acid phosphodiester bond hydrolysis | 376 | 7 | 0.26 | 6.00E-09 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 66 | GO:0019438 | aromatic compound biosynthetic process | 4589 | 16 | 3.2 | 6.60E-09 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 67 | GO:0018130 | heterocycle biosynthetic process | 4604 | 16 | 3.21 | 6.90E-09 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 68 | GO:1901362 | organic cyclic compound biosynthetic process | 4790 | 16 | 3.34 | 1.20E-08 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 69 | GO:0044260 | cellular macromolecule metabolic process | 15085 | 25 | 10.53 | 2.00E-08 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 70 | GO:0043170 | macromolecule metabolic process | 16912 | 26 | 11.81 | 2.30E-08 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 71 | GO:0065008 | regulation of biological quality | 1058 | 9 | 0.74 | 2.60E-08 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 72 | GO:0042221 | response to chemical | 2079 | 11 | 1.45 | 6.70E-08 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 73 | GO:0051214 | RNA virus induced gene silencing | 2 | 2 | 0 | 4.70E-07 | Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F |
| 74 | GO:0051716 | cellular response to stimulus | 2180 | 10 | 1.52 | 1.20E-06 | Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 75 | GO:0009059 | macromolecule biosynthetic process | 6780 | 16 | 4.73 | 1.70E-06 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 76 | GO:0007275 | multicellular organismal development | 1758 | 9 | 1.23 | 1.90E-06 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_6AS_FBB2AFAAB |
| 77 | GO:0044707 | single-multicellular organism process | 1802 | 9 | 1.26 | 2.30E-06 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_6AS_FBB2AFAAB |
| 78 | GO:0044271 | cellular nitrogen compound biosynthetic process | 6955 | 16 | 4.86 | 2.40E-06 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 79 | GO:0044237 | cellular metabolic process | 20927 | 26 | 14.61 | 4.20E-06 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 80 | GO:0044767 | single-organism developmental process | 1955 | 9 | 1.36 | 4.50E-06 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_6AS_FBB2AFAAB |
| 81 | GO:0019048 | modulation by virus of host morphology or physiology | 5 | 2 | 0 | 4.70E-06 | Traes_2AL_DFE4C65F6,Traes_3AS_8EE711E2C |
| 82 | GO:0032502 | developmental process | 1996 | 9 | 1.39 | 5.40E-06 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_6AS_FBB2AFAAB |
| 83 | GO:0010468 | regulation of gene expression | 3247 | 11 | 2.27 | 5.60E-06 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 84 | GO:0032501 | multicellular organismal process | 2096 | 9 | 1.46 | 8.00E-06 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_6AS_FBB2AFAAB |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 85 | GO:0060255 | regulation of macromolecule metabolic process | 3655 | 11 | 2.55 | 1.80E-05 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 86 | GO:0006950 | response to stress | 3116 | 10 | 2.18 | 2.80E-05 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 87 | GO:0071704 | organic substance metabolic process | 22783 | 26 | 15.9 | 3.20E-05 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 88 | GO:0019222 | regulation of metabolic process | 4186 | 11 | 2.92 | 6.20E-05 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 89 | GO:0009791 | post-embryonic development | 1039 | 6 | 0.73 | 6.90E-05 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342 |
| 90 | GO:0044249 | cellular biosynthetic process | 8961 | 16 | 6.26 | 7.00E-05 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 91 | GO:0010467 | gene expression | 6922 | 14 | 4.83 | 7.20E-05 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 92 | GO:1901576 | organic substance biosynthetic process | 9041 | 16 | 6.31 | 7.80E-05 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 93 | GO:0044238 | primary metabolic process | 21860 | 25 | 15.26 | 9.10E-05 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 94 | GO:0048856 | anatomical structure development | 1710 | 7 | 1.19 | 0.00014 | Traes_2AL_2512A7F91,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3DL_F32B49981,Traes_4BL_B3A1B8342,Traes_6AS_FBB2AFAAB |
| 95 | GO:0009058 | biosynthetic process | 9498 | 16 | 6.63 | 0.00015 | Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 96 | GO:0050896 | response to stimulus | 5562 | 12 | 3.88 | 0.00017 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 97 | GO:0009987 | cellular process | 26936 | 27 | 18.8 | 0.00021 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_6DB81005E,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3AS_F27BB108C,Traes_3DL_2DC78B18A,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_5AL_72A7552B9,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 98 | GO:0065007 | biological regulation | 6400 | 12 | 4.47 | 0.00066 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 99 | GO:0050789 | regulation of biological process | 5910 | 11 | 4.13 | 0.00132 | Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2BL_24111235C,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| **Molecular Function(MF)** | | | | | | | |
| 100 | GO:0005488 | binding | 35883 | 47 | 37.15 | 0.00217 | Traes_1AL_095416BC0,Traes_1AL_E7144546E,Traes_1BL_05F7B7DFA,Traes_1BL_7C037D478,Traes_1DL_64B330BBB,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2BL_7713B3533,Traes_2BL_93099ACF4,Traes_2BS_8368F6B5D,Traes_2DL_A77212060,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,T |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | raes_3AL_562D6614F,Traes_3AS_3F8424E4E,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DS_57EA31670,Traes_4AL_7CC35DF1D,Traes_4AL_A118C6C84,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_4DS_88D2821C6,Traes_5AL_07EFD5712,Traes_5AL_72A7552B9,Traes_5BL_F505BF164,Traes_5BL_F611D65E0,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6AL_317133B3F,Traes_6AS_FBB2AFAAB,Traes_6BL_9CFA54D4A,Traes_6DL_4B89E8742,Traes_6DL_58620B158,Traes_6DL_804FB7F75,Traes_6DS_9DD64BD48,Traes_7AL_1BAB53DCE,Traes_7AL_96766587F,Traes_7AL_D88450A3C,Traes_7AS_56569A5AC,Traes_7DL_C538856D4,Traes_7DS_4D01B6175 |
| 101 | GO:1901363 | heterocyclic compound binding | 21060 | 45 | 21.81 | 1.80E-10 | Traes_1AL_095416BC0,Traes_1AL_E7144546E,Traes_1BL_05F7B7DFA,Traes_1BL_7C037D478,Traes_1DL_64B330BBB,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2BL_7713B3533,Traes_2BL_93099ACF4,Traes_2BS_8368F6B5D,Traes_2DL_A77212060,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_3F8424E4E,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DS_57EA31670,Traes_4AL_7CC35DF1D,Traes_4AL_A118C6C84,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_4DS_88D2821C6,Traes_5AL_07EFD5712,Traes_5AL_72A7552B9,Traes_5BL_F505BF164,Traes_5BL_F611D65E0,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6AL_317133B3F,Traes_6AS_FBB2AFAAB,Traes_6BL_9CFA54D4A,Traes_6DL_4B89E8742,Traes_6DL_58620B158,Traes_6DL_804FB7F75,Traes_6DS_9DD64BD48,Traes_7AL_1BAB53DCE,Traes_7AL_96766587F,Traes_7AL_D88450A3C,Traes_7AS_56569A5AC,Traes_7DL_C538856D4,Traes_7DS_4D01B6175 |
| 102 | GO:0097159 | organic cyclic compound binding | 21062 | 45 | 21.81 | 1.80E-10 | Traes_1AL_095416BC0,Traes_1AL_E7144546E,Traes_1BL_05F7B7DFA,Traes_1BL_7C037D478,Traes_1DL_64B330BBB,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2BL_7713B3533,Traes_2BL_93099ACF4,Traes_2BS_8368F6B5D,Traes_2DL_A772120 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | 60,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_3F8424E4E,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DS_57EA31670,Traes_4AL_7CC35DF1D,Traes_4AL_A118C6C84,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_4DS_88D2821C6,Traes_5AL_07EFD5712,Traes_5AL_72A7552B9,Traes_5BL_F505BF164,Traes_5BL_F611D65E0,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6AL_317133B3F,Traes_6AS_FBB2AFAAB,Traes_6BL_9CFA54D4A,Traes_6DL_4B89E8742,Traes_6DL_58620B158,Traes_6DL_804FB7F75,Traes_6DS_9DD64BD48,Traes_7AL_1BAB53DCE,Traes_7AL_96766587F,Traes_7AL_D88450A3C,Traes_7AS_56569A5AC,Traes_7DL_C538856D4,Traes_7DS_4D01B6175 |
| 103 | GO:0003676 | nucleic acid binding | 8266 | 44 | 8.56 | 5.00E-26 | Traes_1AL_095416BC0,Traes_1AL_E7144546E,Traes_1BL_05F7B7DFA,Traes_1BL_7C037D478,Traes_1DL_64B330BBB,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2BL_7713B3533,Traes_2BL_93099ACF4,Traes_2BS_8368F6B5D,Traes_2DL_A77212060,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_3F8424E4E,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DS_57EA31670,Traes_4AL_7CC35DF1D,Traes_4AL_A118C6C84,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_4DL_A54C80661,Traes_4DS_88D2821C6,Traes_5AL_07EFD5712,Traes_5AL_72A7552B9,Traes_5BL_F505BF164,Traes_5BL_F611D65E0,Traes_5DL_672EE3605,Traes_6AL_13BC97E04,Traes_6AL_317133B3F,Traes_6AS_FBB2AFAAB,Traes_6BL_9CFA54D4A,Traes_6DL_4B89E8742,Traes_6DL_58620B158,Traes_6DL_804FB7F75,Traes_6DS_9DD64BD48,Traes_7AL_1BAB53DCE,Traes_7AL_96766587F,Traes_7AL_D88450A3C,Traes_7AS_56569A5AC,Traes_7DL_C538856D4,Traes_7DS_4D01B6175 |
| 104 | GO:0005515 | protein binding | 11888 | 42 | 12.31 | 2.80E-17 | Traes_1AL_095416BC0,Traes_1AL_E7144546E,Traes_1BL_05F7B7DFA,Traes_1BL_7C037D478,Traes_1DL_64B330BBB,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2AL_3F3117458,Traes_2AL_DFE4C65F6,Traes_2BL_24111235C,Traes_2BL_7713B3533,Traes_2BL_93099ACF4,Traes_2BS_8368F6B5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | D,Traes_2DL_A77212060,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_3F8424E4E,Traes_3AS_8EE711E2C,Traes_3DL_2DC78B18A,Traes_3DS_57EA31670,Traes_4AL_7CC35DF1D,Traes_4AL_A118C6C84,Traes_4BL_B3A1B8342,Traes_4DS_88D2821C6,Traes_5AL_07EFD5712,Traes_5AL_72A7552B9,Traes_5BL_F505BF164,Traes_5BL_F611D65E0,Traes_5DL_672EE3605,Traes_6AL_317133B3F,Traes_6AS_FBB2AFAAB,Traes_6BL_9CFA54D4A,Traes_6DL_58620B158,Traes_6DL_804FB7F75,Traes_6DS_9DD64BD48,Traes_7AL_1BAB53DCE,Traes_7AL_96766587F,Traes_7AL_D88450A3C,Traes_7AS_56569A5AC,Traes_7DL_C538856D4,Traes_7DS_4D01B6175 |
| 105 | GO:0003968 | RNA-directed RNA polymerase activity | 40 | 13 | 0.04 | 4.00E-30 | Traes_2DL_6DB81005E,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 106 | GO:0034062 | RNA polymerase activity | 435 | 13 | 0.45 | 6.60E-16 | Traes_2DL_6DB81005E,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 107 | GO:0016779 | nucleotidyltransferase activity | 651 | 13 | 0.67 | 1.10E-13 | Traes_2DL_6DB81005E,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |
| 108 | GO:0016772 | transferase activity, transferring phosphorus-containing groups | 5652 | 13 | 5.85 | 0.00437 | Traes_2DL_6DB81005E,Traes_3AS_F27BB108C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4DL_2E9CE89D9,Traes_4DL_A54C80661,Traes_6AL_13BC97E04,Traes_6BL_0A9D15EDC,Traes_6BL_0BB5C493D,Traes_6BL_78BEF51DD,Traes_6BL_DF680C2AF,Traes_6DL_4B89E8742,Traes_7BL_8CEC8F99B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 109 | GO:0004521 | endoribonuclease activity | 114 | 8 | 0.12 | 3.90E-13 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 110 | GO:0004540 | ribonuclease activity | 141 | 8 | 0.15 | 2.20E-12 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 111 | GO:0004519 | endonuclease activity | 231 | 8 | 0.24 | 1.20E-10 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 112 | GO:0004518 | nuclease activity | 387 | 8 | 0.4 | 6.70E-09 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 113 | GO:0016788 | hydrolase activity, acting on ester bonds | 1663 | 8 | 1.72 | 3.00E-04 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2AL_2512A7F91,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 114 | GO:0004525 | ribonuclease III activity | 36 | 7 | 0.04 | 7.00E-15 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 115 | GO:0032296 | double-stranded RNA-specific ribonuclease activity | 36 | 7 | 0.04 | 7.00E-15 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 116 | GO:0016891 | endoribonuclease activity, producing 5'-phosphomonoesters | 73 | 7 | 0.08 | 1.30E-12 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 117 | GO:0016893 | endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters | 83 | 7 | 0.09 | 3.40E-12 | Traes_1AL_E7144546E,Traes_1DL_C646B6990,Traes_2DL_E96DCDCB4,Traes_3AL_562D6614F,Traes_3DL_2DC78B18A,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9 |
| 118 | GO:0003723 | RNA binding | 1811 | 7 | 1.88 | 0.00257 | Traes_2AL_2512A7F91,Traes_2AL_DFE4C65F6,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_4BL_B3A1B8342,Traes_5AL_72A7552B9,Traes_5DL_672EE3605 |
| 119 | GO:0035197 | siRNA binding | 4 | 4 | 0 | 1.00E-12 | Traes_2AL_2512A7F91,Traes_2AL_DFE4C65F6,Traes_3AS_8EE711E2C,Traes_5DL_672EE3605 |
| 120 | GO:0005634 | nucleus | 4695 | 9 | 2.25 | 8.00E-05 | Traes_2AL_2512A7F91,Traes_2DL_E96DCDCB4,Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_3DL_F32B49981,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342,Traes_5DL_672EE3605 |
| 121 | GO:0031981 | nuclear lumen | 672 | 5 | 0.32 | 1.10E-05 | Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 122 | GO:0043233 | organelle lumen | 775 | 5 | 0.37 | 2.20E-05 | Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 123 | GO:0070013 | intracellular organelle lumen | 775 | 5 | 0.37 | 2.20E-05 | Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 124 | GO:0031974 | membrane-enclosed lumen | 795 | 5 | 0.38 | 2.50E-05 | Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |
| 125 | GO:0044428 | nuclear part | 918 | 5 | 0.44 | 5.00E-05 | Traes_2DS_4CC8FD7E3,Traes_3AL_562D6614F,Traes_3AS_8EE711E2C,Traes_4AS_8D6311711,Traes_4BL_B3A1B8342 |

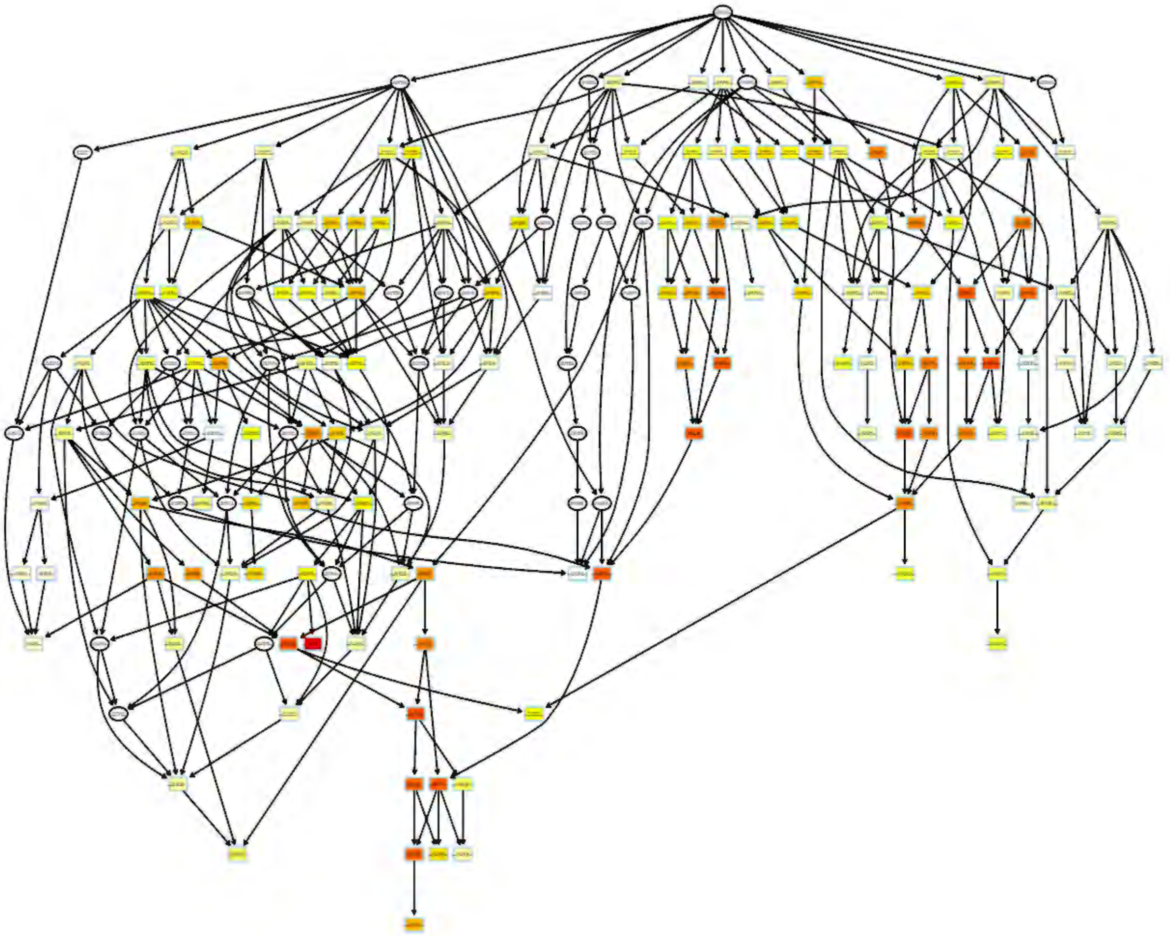


**Fig. A4.1** Biological Process of DCL, AGO and RDR Genes in *T. aestivum* respectively, based on Gene Ontology (GO) categorization. A web-based tool Plant TFDB v4.0 was used to carry out the GO analysis of genes (http://planttfdb.cbi.pku.edu.cn/). Three graphical outputs in the form of hierarchical image containing all statistical significant GO terms. Red color implies that the terms have higher statistical significance. Inside the box: GO terms and GO description are mentioned.

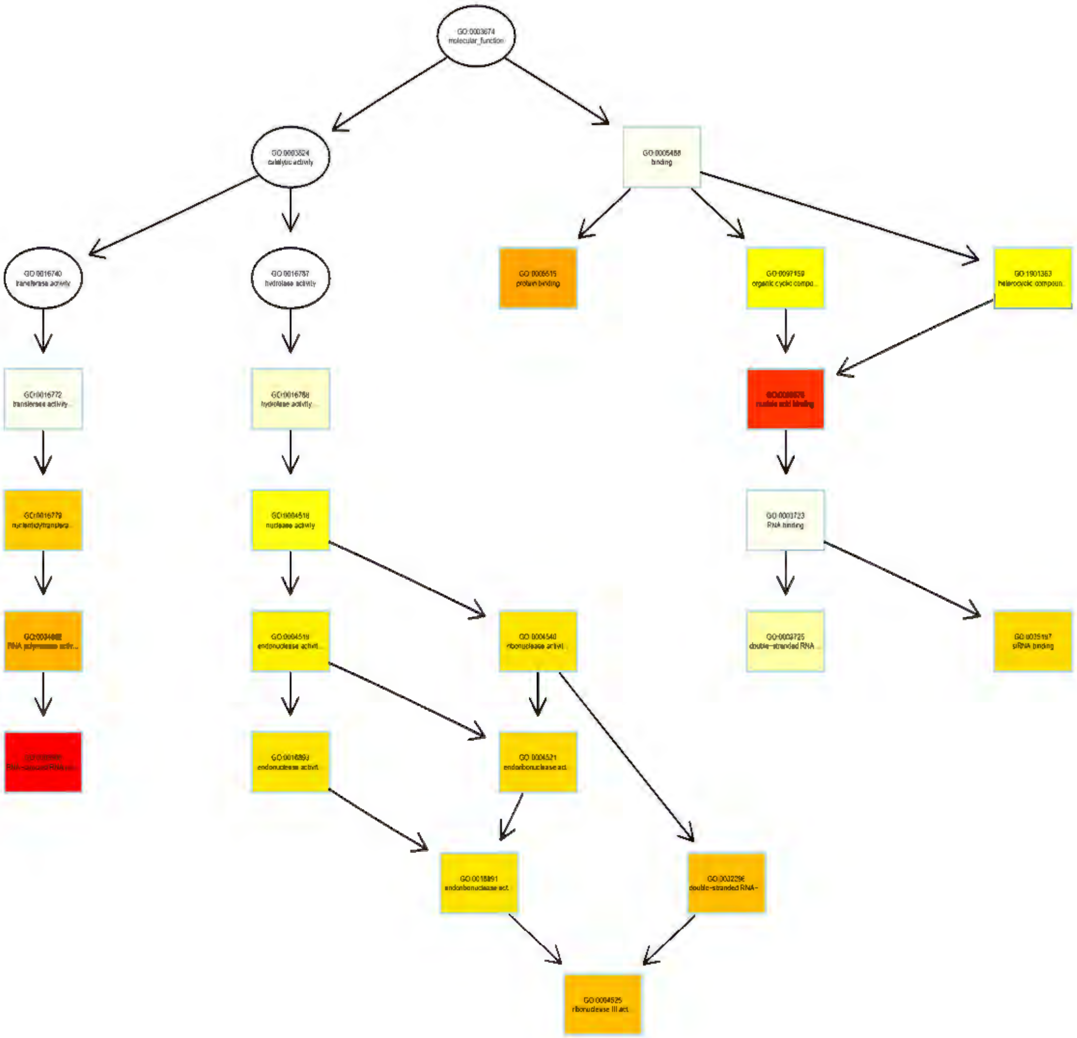


**Fig. A4.2** Molecular Function function of DCL, AGO and RDR Genes in *T. aestivum* respectively, based on Gene Ontology(GO) categorization. A web-based tool PlantTFDB v4.0 was used to carry out the GO analysis of genes (http://planttfdb.cbi.pku.edu.cn/). Three graphical outputs in the form of hierarchical image containing all statistical significant GO terms. Red color implies that the terms have higher statistical significance. Inside the box: GO terms and GO description are mentioned.

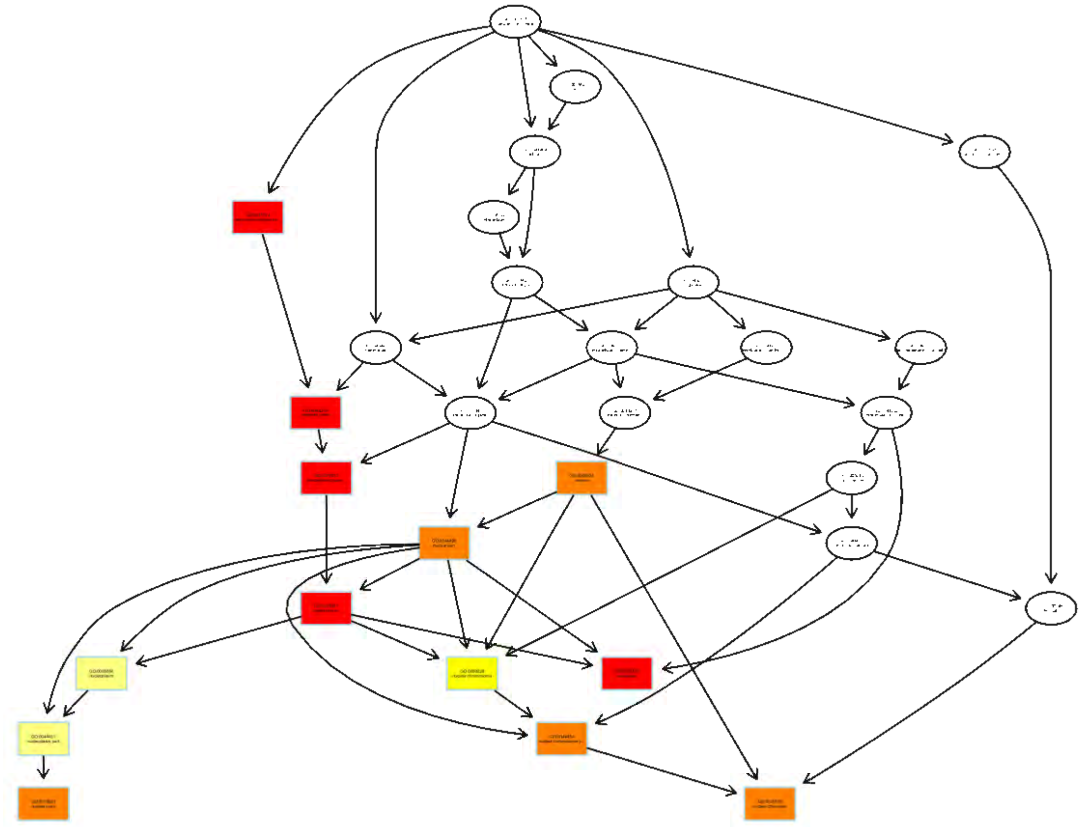


**Fig.A4.3** Cellular Component function of DCL, AGO and RDR Genes in *T. aestivum* respectively, based on Gene Ontology (GO) categorization. A web-based tool PlantTFDB v4.0 was used to carry out the GO analysis of genes (http://planttfdb.cbi.pku.edu.cn/). Three graphical outputs in the form of hierarchical image containing all statistical significant GO terms. Red color implies that the terms have higher statistical significance. Inside the box: GO terms and GO description are mentioned.

# LIST OF PUBLICATIONS

# A. PUBLISHED ARTICLES (JOURNAL)

1. **Zobaer Akond**, Munirul Alam,, Mohammad Sakil Ahmed, Md. Nurul Haque Mollah Multivariate Statistical Techniques for Metagenomic Analysis of Microbial Community Recovered from Environmental Samples. Journal of Bio-Science, 24: 45-53(2016), ISSN 1023-8654.
2. **Zobaer Akond**, Munirul Alam, Md. Nurul Haque Mollah. Biomarker Identification from RNA-Seq Data using a Robust Statistical Approach. Bioinformation 14(4): 153-163 (2018) (**ISI Indexed** and **RG:IF 0.80**)
3. **Zobaer Akond**, Mohammad Nazmol Hasan, Md. Jahangir Alam, Munirul Alam, Md. Nurul Haque Mollah. Classification of Functional Metagenomes Recovered from Different Environmental Samples. Bioinformation 15(1): 26-31 (2019) (**ISI Indexed** and **RG :IF 0.80**).
4. **Zobaer Akond**, Md. Jahangir Alam, Mohammad Nazmol Hasan , Md. Shalim Uddin, Munirul Alam, Md. Nurul Haque Mollah. A Comparison on Some Interval Mapping Approaches for QTL Detection, Bioinformation 15(2): 90-94 (2019) (**ISI Indexed** and **RG :IF 0.80**).
5. **Zobaer Akond**, Hafizur Rahman, Md. Asif Ahsan, Md. Jahangir Alam , Md. Parvez Mosharaf, Munirul Alam, Md. Nurul Haque Mollah.Genome-Wide Identification and Characterization of RNA Silencing Machinery Genes in Wheat (*Triticum aestivum* L.) (Submitted for publication in BMC Plant Biology).
6. **Zobaer Akond,** Md. Matiur Rahaman, Munirul Alam, Md. Nurul Haque Mollah, A Robust Statistical Approach for Gene Expression Analysis And Its Application To Identify Biomarker Genes Influencing Rice Blast Disease (Manuscript is ready to submit for publication in a well-reputed International Journal).
7. **Zobaer Akond**, Md. Asif Ahsan, Munirul Alam, Md. Nurul Haque Mollah, Robustification of Linear Mixed Model Using Outlier Modification Rule and Its Application to Identify Important SNP(s) Influencing Rice Flowering Time (Manuscript is ready to submit for publication in a well-reputed International Journal).
8. MS Ahmed, M Kamruzzaman, MM Rana, **Z Akond**, MNH Mollah.In Silico Analysis of Human Collagen Protein Function. Journal of. Bio-Sci. 24: 55-65, 2016. ISSN 1023-8654
9. Mohammad 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) (**ISI Indexed** and **RG :IF 0.80**).
10. Md. Parvez Mosharaf, Md. Asif Ahsan, Hafizur Rahman, **Zobaer Akond**, Fee Faysal Ahmed, Md. Mazharul Islam, Mohammad Ali Moni, Md. Nurul Haque Mollah1.In Silico Identification, Characterization and Diversity Analysis of RNAi Pathway Gene Families in Sweet Orange (Submitted in Nature Scientific Report).

## B. CONFERENCE PROCEEDINGS

1. Statistical Computation for Metagenomics Data Analysis. **Z. Akond**, M. Alam, M.N. Mollah.The Second International Conference on Theory and Applications of Statistics Dhaka University, Bangladesh. December 27-29, 2015.Dhaka University Statistics Department Alumni Association (DUSDAA).

2. Statistical Methods for Functional Analysis of Metagenomes, **Zobaer Akond** and Md. Nurul Haque Mollah. International Conference on Computer Communication, Chemical, Materials & Electronic Engineering, 24-25 March 2016.

3. Genome-wide identification and phylogenetic analysis of RNA silencing machinery genes in wheat (*Triticum aestivum*). **Zobaer Akond**, Hafizur Rahman, Munirul Alam, M.M. Hossain, Md. Nurul Haque Mollah. International Conference on Bioinformatics and Biostatistics for Agriculture Health and Environment, Department of Statistics, University of Rajshahi. 20-23 January 2017. ISBN: 978-984-34-0996-6

## C. ABSTRACT SUBMISSION

1. Bioinformatic and Phylogenetic Analysis of Dicer-like, Argonaute and RNA-dependent RNA polymerase Gene Families in Carrot (Daucus carota). **Zobaer Akond**, Hafizur Rahman, Md. Nurul Haque Mollah. International Conference on Analysis of Repeated Measures Data.25-26 November 2016. East West University, Aftabnagar,Dhaka.

2. Robust Multivariate Analysis of functional metagenomics. Asma Ul Husna, **Zobaer Akond**, Md Motiur Rahman, Md Nurul Haque Mollah. International Conference on Bioinformatics and Biostatistics for Agriculture Health and Environment, Department of Statistics, University of Rajshahi. 20-23 January 2017. ISBN: 978-984-34-0996-6.