Louisiana State University
LSU Digital Commons

LSU Doctoral Dissertations | Graduate School

2016

Genome-wide Transcriptome Analysis of Cotton (Gossypium hirsutum L.) to Identify Genes in Response to Aspergillus flavus Infection, and Development of RNA-Seq Data Analysis Pipeline

Renesh Bedre
Louisiana State University and Agricultural and Mechanical College, rhbedre@gmail.com

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Recommended Citation

Bedre, Renesh, "Genome-wide Transcriptome Analysis of Cotton (Gossypium hirsutum L.) to Identify Genes in Response to Aspergillus flavus Infection, and Development of RNA-Seq Data Analysis Pipeline" (2016). *LSU Doctoral Dissertations*. 2251.
https://digitalcommons.lsu.edu/gradschool_dissertations/2251

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contactgradetd@lsu.edu.

GENOME-WIDE TRANSCRIPTOME ANALYSIS OF COTTON (*GOSSYPIUM HIRSUTUM* L.) TO IDENTIFY GENES IN RESPONSE TO *ASPERGILLUS FLAVUS* INFECTION, AND DEVELOPMENT OF RNA-SEQ DATA ANALYSIS PIPELINE

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Plant Environment Management and Soil Sciences

by
Renesh Hanumanrao Bedre
M.S. Indian Institute of Information Technology, India, 2011
August 2016

This dissertation is dedicated to my parents Mr. Hanumanrao Bedre and Mrs. Sharda Bedre for their love, affection and sacrifice, and making me what I am today.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank and express my deepest appreciation to my major professor, Dr. Niranjan Baisakh, who has given me an opportunity to pursue a Ph.D. and provided me with continuous support to perform the research. In the last five years, he has spent his valuable time and provided expertise in guiding and motivating me throughout my research work. His expertise in plant molecular biology and bioinformatics immensely helped me at all the times of the research. I greatly appreciate his skills, knowledge, understanding, and patience which contributed vastly to my graduate experience.

I would like to express my sincere gratitude to my Ph.D. committee members, Dr. Stephen Harrison, Dr. Gerald Myers, Dr. Steven Brandt, Dr. Kenneth Damann and the Dean's Representative Dr. Zhi-Yuan Chen for their valuable time, knowledge, insightful comments, and guidance towards the success of the project.

I would also like to express my gratitude to my master's supervisor Dr. Pritish Varadwaj for his guidance, motivation and help in the beginning of my career, and for recommending me for the Ph. D. study with my major professor.

I also must thank my fellow mates of Baisakh Lab and friends at the Louisiana State University for their help, camaraderie and sharing unforgettable moments in life.

I would also thank my family who supported and encouraged me throughout my life to date.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

NGS	Next generation sequencing
RNA-Seq	RNA sequencing
PCR	Polymerase chain reaction
PPB	Parts per billion
ACP	Annealing control primer
SRA	Sequence nucleotide archive
SAGE	Serial analysis of gene expression
CAGE	Cap analysis of gene expression
JA	Jasmonic acid
ET	Ethylene
FPKM	Fragments per kilobase of transcripts per million mapped fragments
Log2FC	Log2 fold change
SRAP	Standalone RNA-Seq analysis pipeline

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Aflatoxins are toxic and potent carcinogenic metabolites produced by *Aspergillus flavus* and *A. parasiticus*. Aflatoxins can contaminate cottonseed under conducive environmental conditions. Much success has been achieved by the application of atoxigenic strains of *A. flavus* for controlling aflatoxin contamination in cotton, peanut and maize. Development of aflatoxin-resistant cultivars overexpressing resistance-associated genes and/or knocking down aflatoxin biosynthesis of *A. flavus* could be an effective strategy for controlling aflatoxin contamination in cotton. In this study, differentially expressed genes (DEGs) were identified in response to infection with both toxigenic and atoxigenic strains of *A. flavus* pericarp and seed of cotton through genome-wide transcriptome profiling. The genes involved in antifungal response, oxidative burst, transcription factors, defense signaling pathways and stress response were highly differentially expressed in pericarp and seed tissues in response to *A. flavus* infection. The cell-wall modifying genes and genes involved in the production of antimicrobial substances were more active in pericarp than seed. Genes involved in defense response in cotton were highly induced in pericarp. The DEGs will serve as the source for identifying biomarkers for breeding, potential candidate genes for transgenic manipulation, and will help in understanding complex plant-fungal interaction for future downstream research.

The increasing volume of sequence data generated by the rapidly decreasing cost of RNA sequencing (RNA-Seq) necessitates the development of software pipeline(s) that can analyze the massive amounts of RNA-Seq data in an efficient manner. Through the present study, a comprehensive and flexible Standalone RNA-Seq Analysis Pipeline (SRAP) implemented with the parallel programming approach was developed, which can analyze transcriptome for any genome. SRAP consists of high-level modules, including sequence reads filtering, mapping to reference genome (or transcriptome), sequence assembly, gene expression analysis and variant discovery along with low-level modules for other common NGS utilities. The high-level modules, unlike low-level modules, require intense computation in terms of memory and processor. SRAP is developed with in-house developed scripts (Python), parallel computing and open source bioinformatics tools. It can be executed as a batch and/or individual mode for single or multiple sample files. SRAP generates RNA-Seq data analysis output files with statistical summary and graphic visualization.

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

Cotton (*Gossypium* sp.) is the most important fiber crop of high economic value. It is a polyploid model species for evolutionary studies in plants. Among different cotton species, *Gossypium hirsutum* (upland cotton) is the main source of cotton fiber and cottonseed which are used in the textile industry, as medical supplies, in the oil industry, as food for dairy cows, in currency manufacturing and for production of other diverse major consumer products [1] . However, the use of cottonseed is limited by its contamination with the highly potent carcinogenic aflatoxin (potent carcinogenic toxin produced from fungus *Aspergillus flavus*), which affects the economy of the crop worldwide. Contamination with aflatoxin is common during environmental stress in cotton crops. Co-inoculation of atoxigenic strains of *Aspergillus* was shown to greatly reduce the aflatoxin contamination in plants such as cotton, maize and peanuts [2-6]. However, inoculation causes wound injury which predisposes cotton to other diseases and insect pests. Host plant resistance has been the mainstay of integrated pest management in several agronomic crops. Unfortunately, there is no germplasm in cotton gene pool that is resistant to *A. flavus* [7]. Understanding the regulation of the genes involved in response to *A. flavus* infection in cotton will not only help unravel the complex genetic network of the host-pathogen interaction, but also will lead to identification of key candidate genes for genetic manipulation to develop *A. flavus* resistant cotton. The genetic complexity of cotton has hindered development of a complete reference genome of *G. hirsutum*. The proposed project will focus on the transcriptome analysis through next generation sequencing (NGS) such as RNA-Seq of cotton tissues (pericarp and seed) to analyze the global response of cotton to toxigenic strain of *A. flavus* infection. Bioinformatics analysis integrated with gene expression validation through quantitative real-time PCR will enable identification of differentially expressed genes and novel transcripts, regulated by aflatoxigenic and non-aflatoxigenic strains of *Aspergillus flavus* infection. The novel genes identified in response to fungal infection and the reference transcriptome developed will serve as a significant resource for genetic transformation of cotton to develop *A. flavus* resistant varieties.

RNA-Seq serves as a powerful tool to identify new genes and splice variants, low abundance transcripts and functional molecular markers. RNA-Seq has become the method of choice over traditional Sanger sequencing and microarray technologies in a couple of years due to its rapidly falling cost and high depth of sequencing. In addition, RNA-Seq allows the researcher to solve the research problems, which were out of scope by microarray and traditional Sanger sequencing technologies [8]. The analysis of RNA-Seq data involves multiple steps from filtering the raw reads to identifying differentially regulated genes and novel transcripts in a given sample. Many methods and tools are available to analyze the huge data points generated by RNA-Seq method [9-11]. In general, multiple steps require many methods that run in different software programs and computer languages, which make RNA-Seq data handling a herculean task. Each of these methods and software has its merits and demerits. Various tools that are available for RNA-Seq data analysis [12-14] focus on a single task and lack flexibility in the analysis. A few proprietary packages are available for menu-driven analysis of RNA-Seq data in a single platform, but their use is cost-prohibitive. Lack of a single and efficient publicly available pipeline which combines the entire data analysis steps in one program prompted us to devise a robust and flexible pipeline with the in-built computational tools that could be available to the scientific community for better biological understanding through analysis of massive NGS data.

1.2 ORIGIN, EVOLUTION AND DIVERSITY OF COTTON

Cotton belongs to the genus *Gossypium* under the family Malvaceae, which includes four domesticated species *G. hirsutum*, *G. barbadense* , *G. arboreum* and *G. herbaceum*. Among these, *G. hirsutum* and *G. barbadense* belong to New World allopolyploids (2n=52), and *G. arboreum* and *G. herbaceum* belong to Old World diploids (2n=26) [15]. The DNA sequence data of the polyploid cotton suggested that cotton originated during the mid-Pleistocene era, about 1-2 million years ago following rapid colonization [15, 16]. The exact place of origin of genus *Gossypium* is unknown, but it has three primary centers of diversity including Australia (particularly Kimberley region), the Horn of Africa and the southern part of the Arabian Peninsula, and the western part of central and southern Mexico [15]. Genetic survey of more than 500 accessions of *G. hirsutum* established that southern Mexico-Guatemala and the Caribbean were the centers of genetic diversity [17].

Phylogenetic analysis based on chloroplast DNA restriction sites [17], 5S ribosomal sequences [18], the chloroplast gene ndhF and nuclear 5.8S genes [19] revealed that four major lineages of *Gossypium* diploid species spread over three continents i.e. Australia (C, G and K genomes, Old World), America (D genome, New World) and Africa/Arabia. Africa/Arabia contains two lineages (one containing A, B and F genomes and the other contains E-genome species, Old World). During evolutionary events, trans-oceanic dispersal of A-genome cotton (female) to the America caused hybridization with native D-genome diploid (male) resulting in allopolyploid cotton [15, 16]. The trans-oceanic dispersal and hybridization mechanisms played a vital role in diversification and speciation within *Gossypium*.

Although some of the old world Gossypium are cultivated as important crops in some part of the word, the new world tetraploid cultivars presently dominate the cotton production worldwide [15]. *G. hirsutum* or "Upland" cotton contributes to more than 90% of the cotton crop production in textile industry, oil industry and animal food purposes internationally, ranging from native Meso-America to more than 50 countries around the world [15, 16]. *G. hirsutum* has mostly tropical and subtropical distribution and grows well in hot and dry weather with sufficient irrigation. *G. barbadense* has long, strong and fine fibers. However, it contributes to less than 10% of the total world cotton production due to its low yield [15].

1.3 CONTAMINATION OF AFLATOXINS IN COTTON

Aflatoxins include four mycotoxins (B1, B2, G1 and G2) that are highly toxic and carcinogenic chemicals produced as secondary metabolites from the asexual fungi *Aspergillus flavus* and *A. parasiticus* [7, 20]. Aflatoxin B1 is the most potent carcinogen to humans and animal, and widely occurring aflatoxin in nature. Aflatoxins are known to cause suppression of the immune system, cancer, retardation in growth, and in extreme cases death of both humans and animals. Aflatoxins have the ability to contaminate a variety of crops such as corn, cotton, peanut and tree nuts during their growth and development, amounting to an estimated economic loss of ~$270 M annually worldwide [7, 20]. The occurrence of aflatoxins in agricultural products is highly regulated. U.S. Food and Drug Administration (FDA) has imposed strict limits on the levels of aflatoxin contamination in foods and feeds; the permitted aflatoxin levels in human food and milk is 20 parts per billion (ppb) and 0.5 ppb, respectively [21], but for the cereals, nuts, dried fruits more stringent aflatoxin standards are 4 ppb for total aflatoxin content and 2 ppb for aflatoxin B1 [21, 22].

The infection by *A. flavus*, the soil borne saprophytic fungus, in cotton is well known. *A. flavus* grows well in high temperatures ranging from 28 °C to 37 °C, which is also favorable for cotton crop growth [21]. *A. flavus* contaminates cottonseed with aflatoxin by a two-phase process. The first phase involves the damage of cotton bolls and partial suture opening that predisposes the bolls to *A. flavus*, which is followed by a second phase involving exposure of bolls to high humidity and warm temperature, during either pre- or post-harvest. The contamination of cottonseed with aflatoxins is of great concern to the cotton industry because cottonseed is used as a preferred protein source for dairy cows. It is also used for vegetable oil production. Cottonseed contributes ~15% of income of the farmers from cotton. Further, the aflatoxins can readily transfer from foods to milk of cows as aflatoxin M1 (hydroxylated derivative of metabolized aflatoxin B1) that ultimately will affect humans [21]. The price of cottonseed is largely determined by the content of aflatoxin. Aflatoxin contamination accounts to high annual economic losses in the USA. The highly affected states in USA are desert regions of Arizona, the Imperial Valley of California, South Texas, and to some extent, the Gulf Coast.

1.4 CONTROL MEASURES OF AFLATOXINS IN COTTON

Considering the declining economy of the cottonseed industry due to infection by *A. flavus*, it is highly important to manage aflatoxin contamination. Both pre- and post-harvest strategies have been used to lessen aflatoxin contamination in crops such as cotton, peanuts and maize. Pre-harvest strategies include application of insect pests and proper irrigation to control the aflatoxin contamination. The postharvest strategies include use of controlled storage conditions that are less favorable to fungal growth, and detoxification of aflatoxins from contaminated seeds and grains [23]. Some plant metabolites such as linoleic acid derivative 13(S)-hydroperoxide are known to inhibit aflatoxin synthesis [24, 25]. Further, the bio-competition by application of atoxigenic strains *A.flavus* and/or *A. parasiticus* to outcompete toxigenic strains in the fields has been shown to be an effective strategy to reduce the aflatoxin contamination in the crops [3, 20, 21]. Atoxigenic strains of *Aspergillus* were reported to reduce the contamination of aflatoxins by ~70-90% in peanut and cotton [4-6]. Bio-competition strategy is of utmost importance in cotton due to the fact that cotton has limited genetic diversity and to date, no aflatoxin-resistant genotype is available in cotton [7, 21]. For long term control of *A. flavus* infections in cotton, it is essential to develop germplasm, which can resist the fungal invasion and/or shut down toxin production [21]. This necessitates detailed investigation into the host-pathogen interaction to identify genes that are induced in cotton in response to *A. flavus* invasion or by toxin production.

Strategies such as expressed sequence tag (EST) and oligonucleotide microarray technologies have been used for the identification of genes induced or regulated in response to *A. flavus* infection in crops such as maize, peanuts and cotton [7, 26, 27]. Genetic engineering of genes induced or upregulated in response to *A. flavus* infection in cotton provides a promising approach to develop cotton varieties resistant to *A. flavus*. To this end, Lee et al. [7] reported, for the first time, a set of up- or down-regulated genes, such as (a)biotic stress responsive genes, storage protein genes and transcription factors, in response to artificial *A. flavus* infection in cotton. But this study employed an annealing control primer (ACP) system that covered a small number of genes, considering the size of *G. hirsutum* genome (~2.83 Gbp) [28, 29]. To identify key regulators in the interaction of *A. flavus* infection with cotton, it is necessary to discover a gene expression atlas through a high-coverage transcriptome analysis approach.

In the past few years, high-throughput next-generation sequencing (NGS) technologies have been considered superior over the traditional methods of sequencing, in terms of time and cost, in addition to the amount of data generated. Various commercial sequencing platforms such as Roche/454, Illumina/Solexa, Ion Torrent and Applied Biosystems SOLiD are able to produce thousands to millions of sequencing reads in a single run with very low cost [30]. The NGS technologies have revolutionized the area of genetics and genomics research, and are currently the methods of choice for understanding complex biological phenomenon through systems biology approaches. Despite high economic loss due to aflatoxin contamination in cotton (cottonseed), a high-throughput transcriptome study for discovering the underlying response mechanism of cotton to *A. flavus* invasion followed by aflatoxin contamination is lacking. In the present study, a high throughput RNA-Seq technology will be utilized to identify differentially regulated genes in response to aflatoxin contamination in cotton because of its high depth of sequence capture, the capability to generate longer and more accurate contigs despite short reads, lower cost and lower error rates compared to other technologies [31, 32].

1.5 DEVELOPMENT OF RNA-SEQ DATA ANALYSIS PIPELINE (COMPUTER PROGRAM/SOFTWARE)

Since its introduction in 2005, RNA-Seq technology has become increasingly popular within the biological research community, which is clearly evident by the large volume of sequence data submitted to the Sequence Nucleotide Archive (SRA) database each year. The conventional Sanger sequencing method is capable of producing sequence reads up to only 1kb long from a single sample at one time, and handling a maximum of 96 samples at one time with an advanced capillary-based sequencer [33]. Serial Analysis of Gene Expression (SAGE) and Cap Analysis of Gene Expression (CAGE), which provides count-based quantification of expressed transcripts involves high sequencing costs and not high-throughput [34]. With the advancement of next and third generation sequencing technologies, the sequencing platforms (e.g. Illumina/Solexa) are capable of generating several thousands to millions of sequence reads in parallel [10, 33, 35]. Due to rapidly decreasing cost and multifold increase in the output from NGS platforms, the NGS technologies have been applied extensively in transcriptome analysis projects for discovery of novel transcripts and gene quantification in addition to genome sequencing [34]. The RNA-Seq, which involves sequencing of mRNA to study transcript structure, allelic information, and high resolution gene expression under a particular condition in individuals by sequencing complete transcriptome, has become a method of choice in most laboratories due to rapidly declining cost of NGS and the ability to explore non-model organisms.

The Illumina sequencers, such as HiSeq2000, HiSeq2500, HiSeq1000, Genome Analyzer IIx and MiSeq, can produce hundreds of GB of sequence output. The HiSeq2500 is the advanced version of HiSeq2000, which was introduced by Illumina in 2012, and is capable of producing an output of 120GB sequence data in 27 hours [33]. This massive amount of data generated from NGS platforms requires intense computational processing, which makes data analysis a daunting task. Further, the rate of increase in computational speed as compared to sequencing data output from NGS platforms is far behind [36]. Many different methodologies have been published to analyze diverse steps in RNA-Seq data analysis, but integrating them into a single pipeline has been a challenging task [37-39]. The steps and configuration parameters used in RNA-Seq data analysis are dependent on each other, and therefore influence the downstream analysis. Non-availability of a one-go pipeline/package to handle and analyze such huge data imposes a limitation on the usefulness of RNA-Seq technology. Therefore, more robust and efficient tools are necessary to analyze such big data generated by NGS technologies.

Several pipelines developed for RNA-Seq analysis [37-40] require computational knowledge and bioinformatics background for their use. For example, RNA-Seq analysis pipeline developed by Goncalves et al. [38] is built in R package, which requires knowledge of R language. A RNA-Seq pipeline “Grape” lacks the statistical analysis for differential expression of genes [37]. Besides, “Grape” uses a lot of computational configurations for installing the pipeline, for example configuration of MySQL database [37], which is not an easy interface for an inexperienced user. The pipeline introduced by Wang et al. [39] lacks the flexibility of using an aligner for mapping the sequence reads to a reference genome/transcriptome, and uses only BWA [13] aligner for mapping. Therefore, the present project is undertaken to develop an RNA-Seq pipeline that would overcome most of the limitations described above. The RNA-Seq pipeline was designed by considering life science researchers from non-computational backgrounds and limited Bioinformatics skills. The proposed software pipeline also keeps up with the exponential increase in data output from sequencing technologies by providing parallel execution of multiple tasks.

1.6 GOALS AND OBJECTIVES

With a long term goal of developing cotton germplasm with resistance to *A. flavus*, this project is formulated to identify underlying existing or novel genes that are regulated by toxigenic strains of *A. flavus*. Another goal is to provide the scientific community with an easy-to-handle, efficient, flexible and robust RNA-Seq pipeline. The proposal is envisaged with the following objectives to accomplish the goals.

A. Identification of differentially expressed genes in cotton (*Gossypium hirsutum* L.) in response to infection by *Aspergillus flavus*;

B. Development and design of an efficient and flexible pipeline for analyzing the RNA-Seq data.

1.7 REFERENCES CITED

1. Rapp RA, Haigler CH, Flagel L, Hovav RH, Udall JA, Wendel JF: **Gene expression in developing fibres of Upland cotton (Gossypium hirsutum L.) was massively altered by domestication**. *Bmc Biol* 2010, **8**.

2. Brown RL, Cotty PJ, Cleveland TE: **Reduction in Aflatoxin Content of Maize by Atoxigenic Strains of Aspergillus-Flavus**. *J Food Protect* 1991, **54**(8):623-626.

3. Cotty PJ: **Influence of Field Application of an Atoxigenic Strain of Aspergillus-Flavus on the Populations of Aspergillus-Flavus Infecting Cotton Bolls and on the Aflatoxin Content of Cottonseed**. *Phytopathology* 1994, **84**(11):1270-1277.

4. Dorner JW: **Biological control of aflatoxin contamination of crops**. *J Toxicol-Toxin Rev* 2004, **23**(2-3):425-450.

5. Pitt JI, Hocking AD: **Mycotoxins in Australia: biocontrol of aflatoxin in peanuts**. *Mycopathologia* 2006, **162**(3):233-243.

6. Dorner JW: **Management and prevention of mycotoxins in peanuts**. *Food Addit Contam* 2008, **25**(2):203-208.

7. Lee S, Rajasekaran K, Ramanarao MV, Bedre R, Bhatnagar D, Baisakh N: **Identifying cotton (Gossypium hirsutum L.) genes induced in response to Aspergillus flavus infection**. *Physiol Mol Plant P* 2012, **80**:35-40.

8. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population**. *Nature* 2010, **464**(7289):773-U151.

9. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.

10. Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, Maloof JN, Sinha NR: **A high-throughput method for Illumina RNA-Seq library preparation**. *Front Plant Sci* 2012, **3**.

11. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B: **A beginners guide to SNP calling from high-throughput DNA-sequencing data**. *Hum Genet* 2012, **131**(10):1541-1554.

12. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3).

13. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.

14. Li RQ, Yu C, Li YR, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.

15. Wendel JF, Brubaker CL, Seelanan T: **The origin and evolution of Gossypium**. In: *Physiology of cotton.* Springer; 2010: 1-18.

16. Wendel JF, Flagel LE, Adams KL: **Jeans, genes, and genomes: cotton as a model for studying polyploidy**. In: *Polyploidy and genome evolution.* Springer; 2012: 181-207.

17. Wendel JF, Albert VA: **Phylogenetics of the Cotton Genus (Gossypium) - Character-State Weighted Parsimony Analysis of Chloroplast-DNA Restriction Site Data and Its Systematic and Biogeographic Implications**. *Syst Bot* 1992, **17**(1):115-143.

18. Cronn RC, Zhao XP, Paterson AH, Wendel JF: **Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons**. *J Mol Evol* 1996, **42**(6):685-705.

19. Seelanan T, Schnabel A, Wendel JF: **Congruence and consensus in the cotton tribe (Malvaceae)**. *Syst Bot* 1997, **22**(2):259-290.

20. Yin YN, Yan LY, Jiang JH, Ma ZH: **Biological control of aflatoxin contamination of crops**. *J Zhejiang Univ-Sc B* 2008, **9**(10):787-792.

21. Yu JJ: **Current Understanding on Aflatoxin Biosynthesis and Future Perspective in Reducing Aflatoxin Contamination**. *Toxins* 2012, **4**(11):1024-1057.

22. Van Egmond H, Jonker M, Abbas H: **Worldwide regulations on aflatoxins**. *Aflatoxin and food safety* 2005:77-93.

23. Lillehoj E, Wall J: **Decontamination of aflatoxin-contaminated maize grain**. In: *US Universities-CIMMYT Maize Aflatoxin Workshop, El Batan, Mexico (Mexico), 7-11 Apr 1986: 1987*: CIMMYT; 1987.

24. Wright MS, Greene-McDowelle DM, Zeringue HJ, Bhatnagar D, Cleveland TE: **Effects of volatile aldehydes from Aspergillus-resistant varieties of corn on Aspergillus parasiticus growth and aflatoxin biosynthesis**. *Toxicon* 2000, **38**(9):1215-1223.

25. Criseo G, Bagnara A, Bisignano G: **Differentiation of aflatoxin-producing and non-producing strains of Aspergillus flavus group**. *Lett Appl Microbiol* 2001, **33**(4):291-295.

26. Guo BZ, Fedorova ND, Chen XP, Wan CH, Wang W, Nierman WC, Bhatnagar D, Yu JJ: **Gene Expression Profiling and Identification of Resistance Genes to Aspergillus flavus Infection in Peanut through EST and Microarray Strategies**. *Toxins* 2011, **3**(7):737-753.

27. Luo M, Brown RL, Chen ZY, Menkir A, Yu JJ, Bhatnagar D: **Transcriptional Profiles Uncover Aspergillus flavus-Induced Resistance in Maize Kernels**. *Toxins* 2011, **3**(7):766-786.

28. Grover CE, Kim HR, Wing RA, Paterson AH, Wendel JF: **Incongruent patterns of local and global genome size evolution in cotton**. *Genome Res* 2004, **14**(8):1474-1482.

29. Guo WZ, Cai CP, Wang CB, Zhao L, Wang L, Zhang TZ: **A preliminary analysis of genome structure and composition in Gossypium hirsutum**. *Bmc Genomics* 2008, **9**.

30. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genom Hum G* 2008, **9**:387-402.

31. Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, Sagnon N, Lazzaro BP: **De Novo Transcriptome Sequencing in Anopheles funestus Using Illumina RNA-Seq Technology**. *Plos One* 2010, **5**(12).

32. Luo CW, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT: **Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample**. *Plos One* 2012, **7**(2).

33. Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research**. *Mol Ecol* 2012, **21**(8):1794-1805.

34. Lindner R, Friedel CC: **A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq**. *Plos One* 2012, **7**(12).

35. D'Antonio M, De Meo PD, Pallocca M, Picardi E, D'Erchia AM, Calogero RA, Castrignano T, Pesole G: **RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application**. *Bmc Genomics* 2015, **16**.

36. Schatz MC, Langmead B, Salzberg SL: **Cloud computing and the DNA data race**. *Nat Biotechnol* 2010, **28**(7):691-693.

37. Knowles DG, Roder M, Merkel A, Guigo R: **Grape RNA-Seq analysis pipeline environment**. *Bioinformatics* 2013, **29**(5):614-621.

38. Goncalves A, Tikhonov A, Brazma A, Kapushesky M: **A pipeline for RNA-seq data processing and quality assessment**. *Bioinformatics* 2011, **27**(6):867-869.

39. Wang L, Si YQ, Dedow LK, Shao Y, Liu P, Brutnell TP: **A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq**. *Plos One* 2011, **6**(10).

40. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, Nie J, Tang X, Baheti S, Doughty JB *et al*: **MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing**. *BMC Bioinformatics* 2014, **15**:224.

CHAPTER 2: IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES IN COTTON (*GOSSYPIUM HIRSUTUM* L.) IN RESPONSE TO INFECTION BY *ASPERGILLUS FLAVUS*

2.1 INTRODUCTION

Aflatoxins comprise a group of four polyketide-derived mycotoxins (B1, B2, G1 and G2) that are highly toxic and carcinogenic chemicals produced as secondary metabolites from toxigenic isolates of the saprophytic fungi *Aspergillus flavus* and *A. parasiticus* [1-6]. Aflatoxin B1 is the most widely occurring structure that is carcinogenic to humans and animals [2-4]. Aflatoxins cause suppression of the immune system, cancer, retardation in growth, and in extreme cases death of both humans and animals. Aflatoxins have the ability to contaminate a variety of crops including corn, cotton, peanut and tree nuts during their growth and development, accounting to an estimated economic loss of ~$270M annually worldwide [4], [5], [7]. The occurrence of aflatoxin in agricultural products is highly regulated. U.S. Food and Drug Administration (FDA) has imposed strict limits on the levels of aflatoxin contamination in foods and feeds; the permitted aflatoxin levels in human food and milk is 20 parts per billion (ppb) and 0.5 ppb, respectively [8], but for the cereals, nuts and dried fruits, aflatoxin standards are more stringent, which is 4 ppb for total aflatoxin content and 2 ppb for aflatoxin B1 [8], [9]. The cottonseeds alone contribute ~15% of the income of the farmers from cotton. The contamination of cottonseed with aflatoxin is of high concern to the cotton industry because cottonseed are used as a preferred protein meal for dairy cows and cottonseed is also used for oil production. Further, cows fed with contaminated cottonseed meal can metabolize the aflatoxin B1 to M1 (hydroxylated derivative of metabolized aflatoxin B1), which in their milk will ultimately affect humans [8]. The prices of cottonseeds are largely determined by the content of aflatoxin present.

Aflatoxin contamination is a major problem in the arid cotton growing regions of Arizona, the Imperial Valley of California, South Texas, and to some extent in Louisiana in the U.S., and accounts to high annual economic losses.

Considering the declining economy of the cottonseed industry due to the infection of cotton by *A. flavus*, it is highly important to take necessary steps to manage aflatoxin contamination in cotton. Both pre- and post-harvest strategies have been used to lessen the aflatoxin contamination in cotton and other crops. Pre-harvest strategies include control of insect pests and proper irrigation to manage aflatoxin contamination. The post-harvest strategies include control of storage conditions that are less favorable to fungal growth, and detoxification of aflatoxin from contaminated seeds and grains [10]. Some plant metabolites, such as linoleic acid derivative 13(S)-hydroperoxide, are known to inhibit aflatoxin synthesis [8, 11]. Further, the bio-competition by application of atoxigenic strains *A. flavus* and/or *A. parasiticus* to outcompete toxigenic strains in the fields has been shown to be an effective strategy to reduce the aflatoxin contamination [5, 8, 12]. Atoxigenic strains of *Aspergillus* were reported to reduce the contamination of aflatoxin by ~70–90% in cotton and peanut [13-15]. This bio-competition strategy is of utmost importance in cotton because cotton has limited genetic diversity, and to date, no aflatoxin-resistant genotype is available in cotton [4, 8].

The defense responses in plants depend on the type of pathogen [6, 16]. Among different mechanisms, defense responses in plants are known to be regulated by the phytohormones, such as salicylic acid (SA), jasmonic acid (JA), ethylene (ET), cytokinin (CK) and auxins [6, 16, 17]. As a general rule, plant resistance to biotrophic pathogens is controlled by SA. In contrast, the resistance to necrotrophic pathogens is controlled by JA- and ET-dependent signaling pathways [6, 16, 17]. Moreover, resistance to necrotrophic fungal pathogens is known to be quantitative in nature and regulated by multiple genes [6, 18]. Toxigenic strain of *A. flavus* is characterized with the features of a necrotrophic fungal pathogen [6]. It is essential to develop germplasm that can resist the fungal invasion and/or shut down toxin production for long-term control of *A. flavus* infections [4, 19]. However, conventional breeding for resistance to *A. flavus* in cotton has been handicapped due to the unavailability of the genetic resistance in the available cotton gene pool. Genetic engineering of cotton with genes induced or upregulated in response to *A. flavus* infection will provide a promising approach to develop cotton varieties resistant to *A. flavus*. This necessitates detailed investigation into the host-pathogen interaction to identify genes that are induced in cotton in response to *A. flavus* invasion or by toxin production. Further, understanding the largely unknown molecular basis of bio-competition strategy in controlling toxigenic *A. flavus* infection using atoxigenic strain of *A. flavus* could lead to identification of candidate genes for their use in manipulation of *A. flavus* resistance in cotton. Strategies such as small-scale expressed sequence tag (EST) library sequencing and oligonucleotide microarray have been used for the identification of genes induced or regulated in response to *A. flavus* infection in crops, such as maize, peanuts and cotton [4, 20, 21]. These small-scale targeted strategies based on the identification one or a few genes are not sufficient to understand the complex host-pathogen interaction responses [22]. Therefore, to identify key regulators in the interaction of *A. flavus* infection with cotton, it is necessary to discover genes on a global scale using high-coverage transcriptome analysis approach. We report here the identification of differentially expressed/regulated genes in the pericarp and seed tissues of cotton in response to *A. flavus* infection with an objective to understand the complex genetics involved in response of cotton to both toxigenic and atoxigenic strains of *A. flavus* infection.

2.2 MATERIALS AND METHODS

2.2.1 Fungal Culture Preparation and Cotton Boll Inoculation

Fungal cultures of toxigenic (AF13) and atoxigenic (AF36) strains of *A. flavus* were prepared as described earlier [4]. Briefly, the strains were grown on maltose extract agar medium at 30°C for a week. Conidia were harvested by scrapping the mycelium in 9 µl of potato dextrose broth (PDB), and the suspension was adjusted to a concentration of 10^4 conidia/ml. Cotton variety 'Coker 312' was grown in the greenhouse for the present study as described earlier [4]. A hole to a depth of 5–10 mm was made in the center of one of the locules (L1) of cotton bolls (28–30 dpa) using a 3 mm dia cork borer. Ten µL of the conidia suspension was applied into the hole using a Pasteur pipet. Bolls inoculated with only PDB without conidia served as the control. Pericarp and fiber-free seeds from non-inoculated and inoculated locule (L1) and adjacent/distal (Adj) locules of cotton bolls were harvested and placed in liquid nitrogen at 6, 24, 48, and 72 h after inoculation, and stored at -80°C for RNA isolation. Three bolls each from two different plants (biological replicates) were used for each treatment.

2.2.2 RNA Extraction, Library Preparation and Sequencing

The total RNA was separately extracted from seed and pericarp tissues from L1 and Adj locules collected at each time point by using Spectrum total RNA isolation kit (Sigma-Aldrich, St. Louis, MO). RNA quantity and integrity were assessed as described earlier [4]. For library preparation, 2 µg of RNA from each different time points and replications were mixed for each tissue and experimental condition in order to minimize the cost of library preparation. Altogether, six libraries–non-inoculated pericarp (NIP) and seed (NIS), Pericarp (NTP) and seed (NTS) inoculated with atoxigenic strain, and pericarp (TP) and seed (TS) inoculated with toxigenic strain were prepared as per the manual of Illumina RNA-library construction kit. The libraries were single-end sequenced using the Illumina HiSeq-2000 platform at the sequencing facilities of the Iowa State University, Johnston, IA.

2.2.3 Read Filtering and Sequence Assembly

The single-end raw short Illumina sequencing reads (100 bp) were subjected to filtering to obtain high quality reads for downstream analysis. The raw reads were filtered and trimmed for adapter contamination and low quality, ambiguous and uncalled nucleotide bases. The reads containing more than 5% of uncalled bases and of average quality <= 20 over a window size of 5 bp in 5' to 3' direction were discarded. The filtering and trimming of the raw reads were performed by an in-house pipeline developed with Python programming. Subsequently, the high quality reads were assembled de novo using Trinity (release 2013-02-25) [23] with parameters of k-mer size 25, minimum contig length 200 bp and min_kmer_cov 2. The assembly was performed individually for reads of each of the six libraries. The overlapping k-mers were assembled into linear transcripts followed by clusters of overlapping transcripts. The transcripts for alternative spliced form and paralogous genes from these overlapping transcripts were obtained. All bioinformatics data analysis was performed using the Louisiana State University High Performance Computing (HPC) resource SuperMike-II configured with 16 CPUs. After the assembly was performed for each library, exactly duplicate (100% similar) transcripts were removed to determine the total unigenes. We have used the term "transcript" here to describe individual sequence assembly and "unigene" to denote the longest transcript from a particular alternatively spliced isoforms cluster.

2.2.4 Functional Annotation

The assembled transcripts were subjected to functional annotation using homology search against publicly available databases, such as *G. raimondii* protein database (http://phytozome.jgi.doe.gov/pz), *G. arboreum* protein database (http://cgp.genomics.org.cn/page/species/index.jsp), NCBI's non-redundant (nr) plant nucleotide sequence database (http://www.ncbi.nlm.nih.gov), and UniProtKB (SwissProt and Tr-EMBL plant sequences) database (http://www.uniprot.org). The homology search was performed using BLASTx algorithm [24] against the databases at an E-value cut-off of 1e-05. If the annotation from different databases conflicted with each other, priority was given to the match with *G. raimondii* protein database, NR and UniprotKB, in that order. The Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg) database was utilized for assigning biological pathways to the transcripts. Further, based on the sequence similarity with *G. raimondii* protein database, GO annotations for biological process, molecular function and cellular component were assigned to the assembled transcripts. The GO enrichment analysis was performed with agriGO analysis toolkit [25] with default P-value and false discovery rate (FDR). For identification of enrichment of metabolic pathways, the PathExpress analysis tool with criteria of $P<0.05$ was used [26].

2.2.5 Mapping Reads to Reference Sequence

The cleaned reads from each library were mapped individually to the D genome *G. raimondii* (http://www.phytozome.net/) and the A genome *G. arboreum* (http://cgp.genomics.org.cn/ page/species/index.jsp) [27, 28] using Tophat (version 2.0.9) spliced aligner [29] and Bowtie2 aligner [30] with number of threads set to 10. The unaligned reads were mapped by Bowtie2 that split these into smaller segments for realigning and finding potential spliced sites. Mapped and unmapped reads were reported as BAM files, which were used for downstream analysis.

2.2.6 Differential Gene Expression Analysis

Differentially expressed genes from the six RNA-Seq libraries were identified by using respective BAM files for alignment with Tophat-Cufflink pipeline (version 2.1.1) that produced the transcript assembly [31, 32]. The read counts were normalized as Fragments per Kilobase of Transcripts per Million mapped fragments (FPKM). The assembly files created across infected and uninfected control conditions were pooled in a single file using Cuffmerge for differential expression analysis. Finally, the pooled file from Cuffmerge was fetched to Cuffdiff for calculating expression level and statistical significance of genes across control and inoculated conditions. Cuffdiff employed a blind dispersion model, which conservatively treated all conditions as a replicate of each other in the absence of transcripts from biological replicates as is the case in the present study [31, 32]. The codes used in Cuffdiff for identifying DEGs under NIPvsTP were as follows: cuffdiff-o cuffdiff_out_NIP_TP-b GraimondiiGenome.fa-p 10-L NIP, TP -max-bundle-frags 10000000 -FDR -u merged.gtf accepted_hits_NIP.bam accepted_hits_TP. bam (Generalized code: cuffdiff-o cuffdiff_out_NIP_TP -b genome.fa -p 10 -L NIP,TP -u merged.gtf accepted_hits_NIP.bam accepted_hits_TP.bam). The same code was used for other experimental conditions. The heatmaps for gene expression analysis (log2 fold change) were plotted using heatmap.2 function within R package (version 3.1.2). Three-way comparisons– NIvsNT, NIvsT, and NTvsT were performed to understand the modulation of gene expression between different experimental conditions for both pericarp and seed. The (digital) expression of 10 selected genes showing maximum log2 fold change under *A. flavus* infection in comparison to non-inoculated control was validated by reverse-transcription PCR using gene specific primers (APPENDIX I; sheet 1) following the method described earlier [4].

2.3 RESULTS AND DISCUSSION

2.3.1 Illumina Sequencing, Quality Control and Alignment to the Reference Genome

Sequencing of the six libraries generated 911,040,814 reads of 100 bp long resulting in a total of 91.1 Gbp sequence data (Table 2.1). The average quality of the reads in the libraries after

Table 2.1 Sequencing and assembly statistics of cotton transcriptome

	Pericarp			Seed			
Parameter	**NIP**	**NTP**	**TP**	**NIS**	**NTS**	**TS**	**Total**
# Raw single-end reads	163,090,907	138,340,968	196,958,025	193,242,949	137,348,154	82,059,811	911,040,814
Bases sequenced (Gbp)	16.30	13.83	19.69	19.32	13.73	8.205	91.1
Sequence coverage	65.24X	55.34X	78.78X	77.3X	54.94X	32.82X	364.42X
# Cleaned reads used in assembly	162,591,684	137,657,801	196,080,070	192,530,847	136,780,468	81,716,978	907,357,848
Assembled unique transcripts	99,772	107,294	122,657	100,510	82,896	73,440	586,569
# Unigene	31,425	35,831	40,010	31,635	28,019	23,638	190,558
Transcript N50 (bp)	1887	1842	1976	1927	1703	1787	1864
Transcript N90 (bp)	701	647	699	705	634	658	675
Max. transcript size (bp)	15441	14102	15414	15383	10154	14933	15441
Min. transcript size (bp)	201	201	201	201	201	201	201
Average transcript size (bp)	1339.58	1275.85	1351.4	1353.22	1217.53	1278.19	1307.8
Transcriptome size (Mbp)	133.65	136.89	165.75	136.01	100.92	93.87	767.11

NIP = non-inoculated pericarp NTP = atoxigenic pericarp; TP = toxigenic pericarp; NIS = non-inoculated seed; NTS = atoxigenic seed; TS = toxigenic seed.

filtering was > 33. There was negligible amount of adapter/primer contamination in the reads. Altogether, 907,357,848 high quality reads were obtained from all six libraries, which totaled to 90.73 Gbp sequencing data (Table 2.1). The raw sequences have been deposited to the NCBI. SRA database (http://www.ncbi.nlm.nih.gov/sra/?term=PRJNA275482). Out of high quality filtered reads, 59.6% mapped to the D genome of cotton (*Gossypium raimondii*), whereas 75.6% mapped to the A genome (*G. arboreum*) across all six libraries. The higher percentage of alignment of the reads to the A genome could be due to its higher genome size as compared to the D genome.

2.3.2 De Novo Sequence Assembly

The high quality reads from each of the six libraries from different experimental conditions were assembled independently into transcripts with a length more than 200 bp. The lengthwise distribution of transcripts of six independent libraries of cotton under different experimental conditions is shown in (Figure 2.1). Trinity has a better resolving power than others in identifying alternative spliced transcript, and thus produces less duplicates and chimeric transcripts [33]. However, redundancy was encountered in the assembled transcriptome due to high sequencing depth, duplication and assembly process. Therefore, exactly duplicate transcripts were removed from all six libraries to obtain 586,569 unique transcripts (Table 2.1). Further, total transcripts from six libraries clustered into 190,558 unigenes (Table 2.1). The size of the transcripts ranged from 201 to 15,441 bp with a mean of 1307.80 bp. Similarly, unigenes size ranged from 201 to 15,441 bp with a mean of 841.88 bp (Table 2.1). Of the total reads, 87% aligned to the assembled transcripts indicating good coverage of the transcriptome. The high N50 and N90 values for unique transcriptome were 1864 bp and 675 bp, respectively, further suggesting a good quality assembly. Furthermore, complete alignment of the longest transcript comp54729_c0_seq1_NIP (15,441 bp) from NIP library to a gene coding for auxin transport protein of *Theobroma cacao* (TCM_019010) in the NCBI database demonstrated that the transcript was not chimeric which could have occurred due to repetitive regions in the genes. These results strongly supported a high quality transcriptome assembly of cotton.

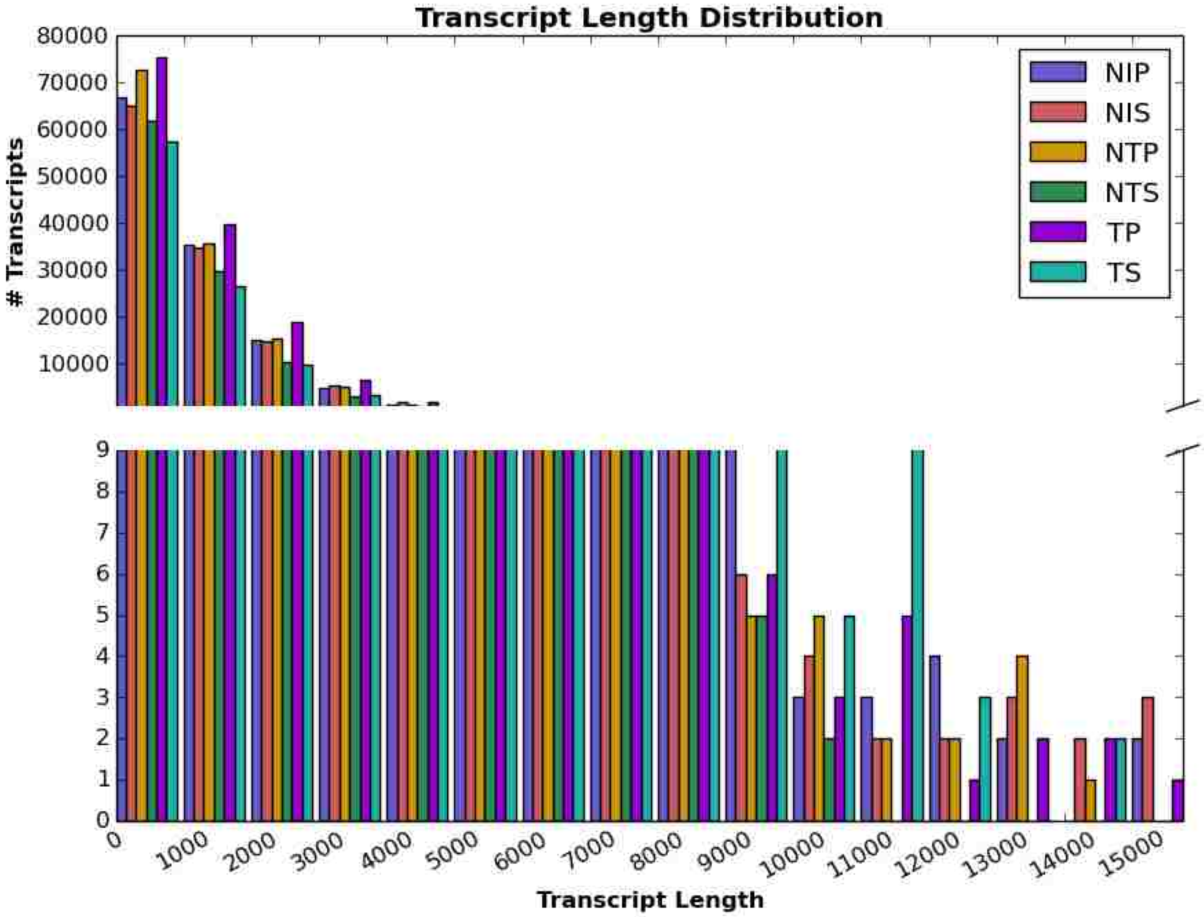


Figure 2.1 Lengthwise (in bp) distributions of transcripts in RNA-Seq libraries from pericarp and seed tissues of cotton with and without *Aspergillus flavus* infection. NIP = non-inoculated pericarp, NTP = non-toxigenic pericarp, TP = toxigenic pericarp, NIS = non-inoculated seed, NTS = non-toxigenic seed, TS = toxigenic seed.

2.3.3 Functional Annotation

Out of the 586,569 total unique transcripts, 466,054 (79.45%) were assigned functions based on their similarity to cotton protein database. The remaining un-annotated sequences were searched against NCBI nr and UniProtKB plant databases. Longer sequences produced significant blast hits as compared to the shorter sequences. Out of the total annotated transcripts, 405,652 (87.04%) transcripts with more than 500 bp length showed similarity to proteins in the cotton database. Of the remaining unmapped transcripts, 8,755 transcripts mapped to plant NCBI nr/nt and UniProtKB database. Further, the cotton unique transcriptome was mapped to the protein sequences of *G. arboreum*. In total, 19,750 un-annotated unique transcripts matched with *G. arboreum* protein sequences. Thus, we annotated total 494,559 transcripts (84.31%) of the cotton transcriptome. The transcripts which did not match to known genes may represent novel genes or genes that may have diverged from their homologs or noncoding RNAs [34]. The homology search showed 79.45% and 82.89% unique transcripts matching to *G. raimondii* and *G. arboreum* proteins, respectively.

2.3.4 Identification of Differentially Expressed Genes (DEGs) in Response to *Aspergillus flavus* Infection

Statistically significant differentially expressed genes (DEGs) in terms of FPKM (fragments per kb per million mapped reads) were calculated using combination of log2FC and P-value criteria based on mapping of the cotton reads against the *G. raimondii* genome as reference. In pericarp tissue, 1265, 832 and 396 genes were up-regulated (log2FC≥2, $P<0.05$) under NIPvsNTP, NIPvsTP and NTPvsTP conditions, respectively. On the other hand, 247, 123 and 869 genes were down-regulated (log2FC≤-2, $P<0.05$) under same experimental conditions, respectively. Similarly, in the seed tissue, 680, 492 and 369 genes were up-regulated under NISvsNTS, NISvsTS and NTSvsTS conditions, respectively, whereas, 321, 80 and 302 genes were down-regulated under same experimental conditions, respectively (Figure 2.2, A, B and C).

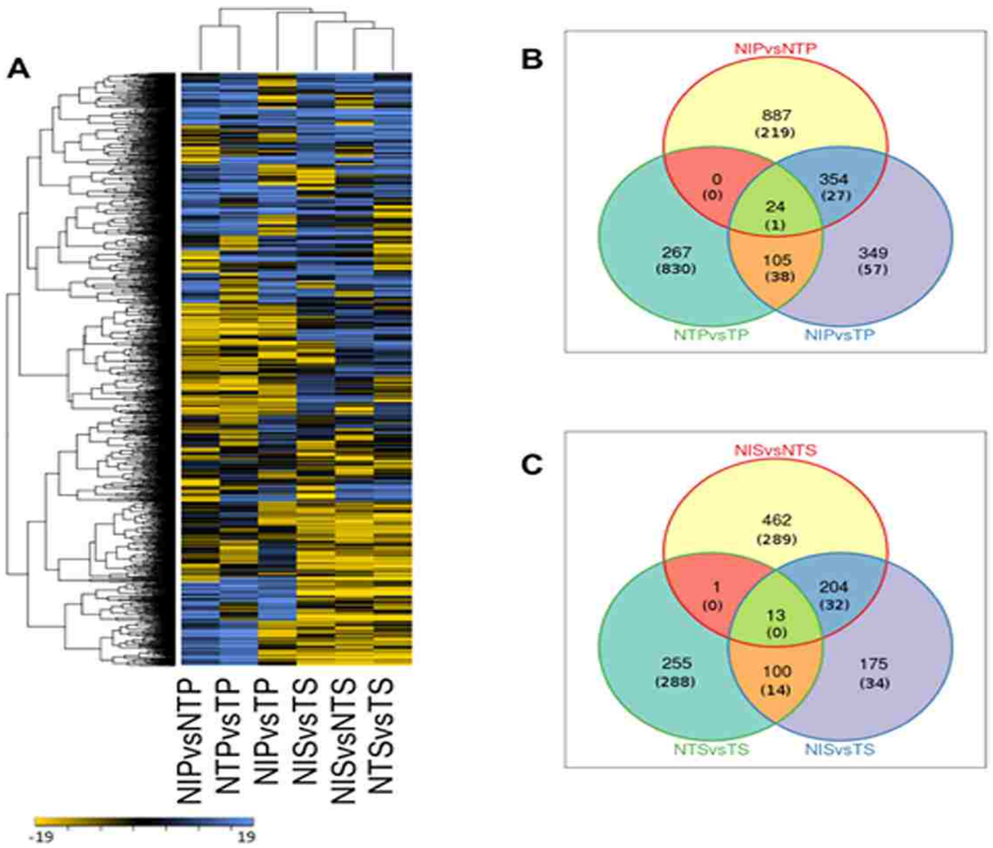


Figure 2.2 Gene expression profile of cotton in pericarp and seed tissue in response to *A. flavus* infection**. A)** Heatmap showing differentially expressed genes (DEGs) of cotton in response to infection by atoxigenic and toxigenic strains of *Aspergillus flavus*. The up-regulated genes (log2FC≥2 and P<0.05) and down-regulated genes (log2FC≤-2 and P<0.05) are represented by blue and yellow color, respectively. Genes with similar expression profiles were clustered together by hierarchical clustering. For description of the gene names represented in the heatmaps, please refer to the APPENDIX I, sheet 2. Venn diagram shows the unique and common DEGs in pericarp **(B)** and seed **(C)** tissues under different experimental conditions.

Principal component analysis (PCA) showed distinct response of pericarp and seed tissues to toxigenic and atoxigenic strains of *A. flavus* (Figure 2.3). The total variance contributed by three principal components was 72% (Figure 2.3). The results further showed significant differences in the expression profile of genes in response to atoxigenic and toxigenic infection in pericarp, whereas, in seed, the difference in response was not significant with the infection by both strains of *A. flavus*. This suggested that the pericarp tissue, being the primary tissue for inoculation, exhibited higher level of differential response of genes as compared to the seed tissue. Thus, identification of specific category of highly up-regulated genes from different clusters, and characterization of their biochemical response would provide potential candidates for functional characterization through genetic manipulation toward improvement of resistance to *A. flavus* infection. The distributions of DEGs under different conditions are shown in three-way Venn

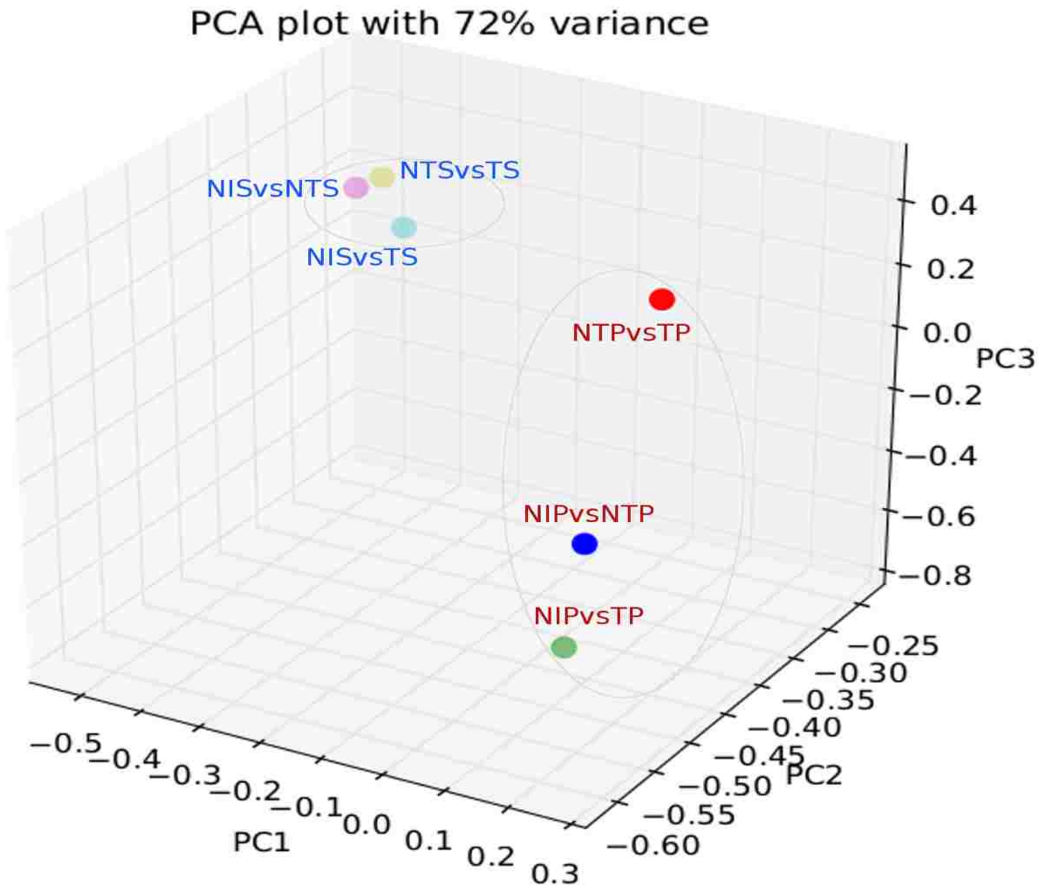


Figure 2.3 Principal component analysis (PCA) showing the variability (72% variance) of DEGs of cotton in pericarp and seed tissue in response to infection by toxigenic and atoxigenic strains of *Aspergillus flavus*. Expression of genes under different experimental conditions in seed (small oval) and pericarp (large oval) were distinct with the variability of expression higher in pericarp compared to seed. PC1, PC2 and PC3 explained 32%, 24% and 16% of the total variance.

diagrams for pericarp (Figure 2.2, B) and seed (Figure 2.2, C). The DEGs were further characterized into different groups based on their putative functional significance as described below. Because *A. flavus* is a necrotrophic fungus, JA and ET dependent signaling pathways presumably function in defense response and regulate the expression of defense related genes, genes involved in oxidative burst, synthesis of antimicrobial compounds, regulation of transcription factors and localized programmed cell death in cotton in response to the fungal infection.

2.3.4.1 Genes interfering with fungal virulence and growth

Eleven transcripts encoding chitinases were differentially expressed under infection by both atoxigenic and toxigenic strains in pericarp and seed. The transcripts encoding B-CHI and CHIV were induced by infection with both atoxigenic and toxigenic strains in pericarp and seed. But, *CTL2* was up-regulated specifically in pericarp by both strains, whereas, *CHIA* was specifically induced in seed by both strains. Among the transcripts encoding CTL2, *Gorai.006G078900* and *Gorai.011G198500* were induced under atoxigenic and toxigenic strain infection, respectively. The three genes encoding β-1,3-glucanases (BG) were up-regulated specifically in seed. *BG3* (*Gorai.006G134600*) and *BG1* (*Gorai.006G134700*) were highly induced by the atoxigenic strain, whereas, another *BG3* transcript (*Gorai.010G003600*) was up-regulated specifically by the toxigenic strain. Plant pathogenic fungi infect the plants through wounds or release of hydrolytic enzymes, such as pectinases, proteases and amylases for successful colonization [22]. Therefore, identification and characterization of plant genes, which interfere with invasion of fungus in plants, can be useful to reduce fungal pathogenicity. The hydrolytic enzymes chitinase and β-1,3-glucanase genes possess antifungal activity by degrading the fungal cell wall containing chitins [22, 35, 36]. Plant chitinases possess lysozyme activity and are highly active in inhibiting fungal growth [37]. Moreover, over-expression of chitinase genes has conferred resistance to fungal infection in plants, such as tobacco, peanuts and rice [38-40].

Five transcripts encoding trypsin and protease inhibitor proteins (TPI) were induced in pericarp, and only one *TPI* was induced in seed (Figure 2.4, A). The two transcripts, *Gorai.011G254400* and *Gorai.012G027700*, were up-regulated under infections by both the strains in pericarp, but the *TPI* (*Gorai.011G254900*) was specifically induced under the atoxigenic infection in pericarp. Among the *TPI* genes induced in pericarp, *Gorai.011G254500* and *Gorai.011G254600* were highly up-regulated by the toxigenic strain and down-regulated by the atoxigenic strain (Figure 2.4, A). For example, *Gorai.011G254500* and *Gorai.011G254600* were up-regulated by 2.9- and 6.5-fold, and 9.8- and 12.7-fold higher by the toxigenic strain in comparison with the non-inoculated control and the atoxigenic strain, respectively (APPENDIX I, sheet 3). In seed, the *TPI* (*Gorai.012G027600*) was induced under the toxigenic strain infection only. Reduced growth of *A. flavus* has direct impact on aflatoxin production [41]. Trypsin inhibitors are known to possess antifungal activity [22, 41] and inhibit conidial germination and hyphal growth of *A. flavus* [41]. Four genes encoding serine protease inhibitors (*SPI*) were also differentially expressed in pericarp and seed. *Gorai.012G105800* and *Gorai.007G143500* were up-regulated specifically under the atoxigenic strain infection in pericarp and the toxigenic strain infection in seed, respectively. Serine protease inhibitor gene has been shown to be induced in response to infection with *A. flavus* in peanut [20].

Plants accumulate hydroxyproline-rich glycoproteins (HRGPs), phytoalexins and lignin-like substances as a resistance mechanism in response to fungal pathogen infection [42-44]. The HRGPs and lignification are involved in fortifying the cell-wall structure of plants and contribute to the resistance to pathogen invasion [45]. Among 20 DEGs encoding HRGP, 10 and two genes were specifically induced in pericarp and seed, respectively (Figure 2.4, A). In pericarp, most of these genes were induced under the atoxigenic strain infection. Similarly, in seed the two genes were induced by the atoxigenic strain infection. The increase in the level of HRGP transcripts has been observed in several plants in response to wound or pathogen infection, and is associated

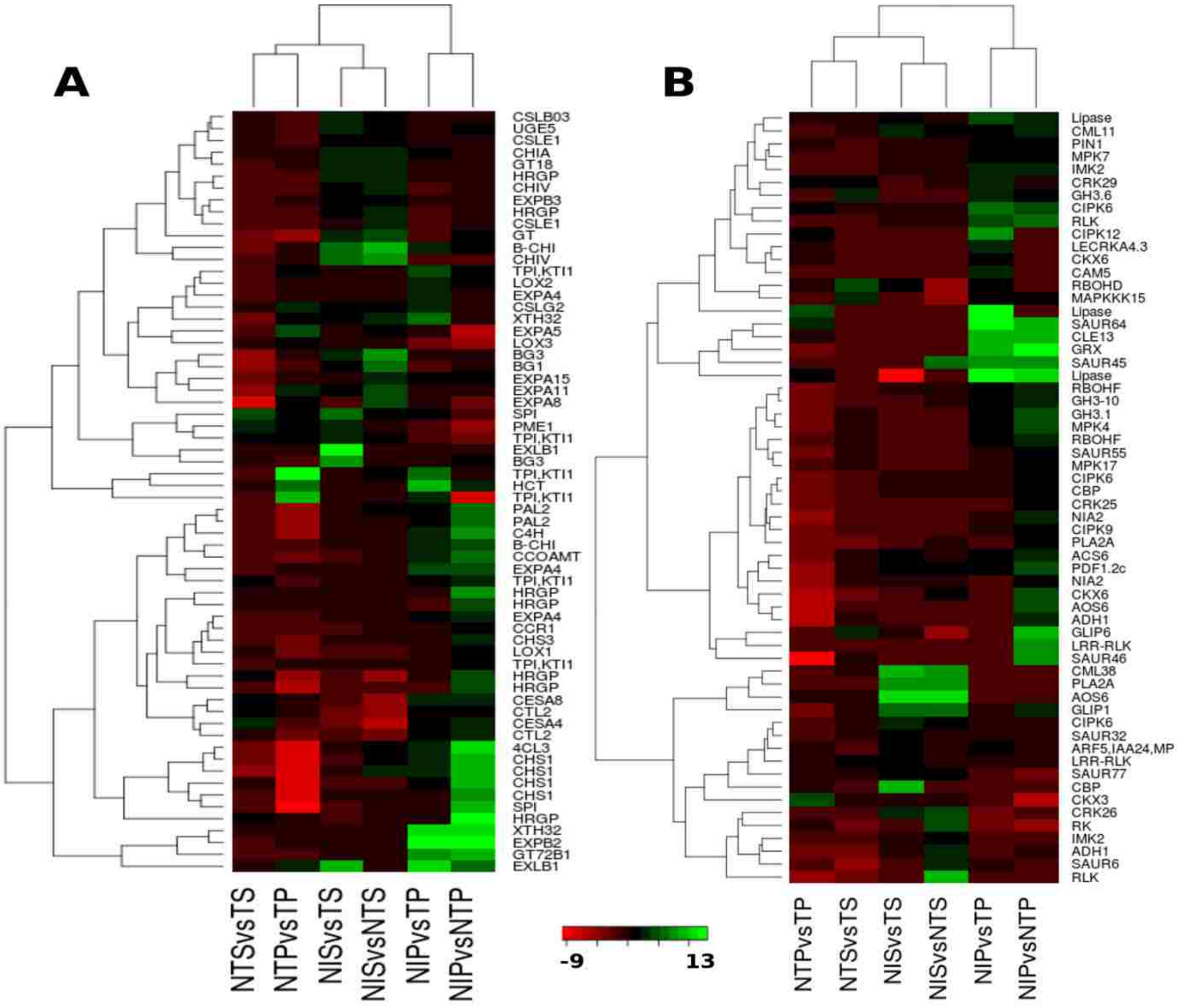


Figure 2.4 Heatmaps showing DEGs involved in interference of fungal virulence and growth **(A)** and DEGs involved in defense signaling **(B)**. The green color represents up-regulated (log2FC≥2) genes and red color represents down-regulated (log2FC≤- 2) genes. For description of the gene names represented in the heatmaps, please refer to the APPENDIX I, sheet 3.

with higher resistance to pathogen [37]. The genes involved in the metabolism of phenylpropanoids, which serve as a source for furanocoumarin and isoflavonoid phytoalexins and lignins, were differentially expressed in the pericarp and seed tissues. The phenylpropanoid pathway precursors are also involved in the synthesis of lignin and phenolic substances, and possess antifungal activities [37, 42, 43]. The increase in cell wall lignification as a structural modification has been observed in plants for defense in response to fungal pathogen [37]. The enzymes phenylalanine ammonia-lyase (*PAL*), 4-coumarate CoA ligase (*4CL*) and chalcone synthase (*CHS*) are involved in the phenylpropanoid pathway [37, 42, 43]. Two genes encoding each *PAL2* and *4CL1* were highly up-regulated in pericarp under the atoxigenic strain infection (Figure 2.4, A). The transcript encoding *4CL3* was up-regulated in both tissues, but had the higher fold change in gene expression in pericarp specifically under the atoxigenic strain infection as compared to the toxigenic strain infection. The enzyme chalcone synthase (CHS) was highly induced in pericarp (Figure 2.4, A). The three genes encoding *CHS* (*Gorai.005G035100*, *Gorai.006G000200* and *Gorai.009G339300*) were specific to pericarp and showed higher expression under the atoxigenic strain infection. The two *CHS* genes (*Gorai.011G161200* and *Gorai.011G161300*) were up-regulated in both tissues under the atoxigenic infection, but only in pericarp under the toxigenic infection, the fold change for these two *CHS* genes were higher in pericarp as compared to seed (Figure 2.4, A). Other genes involved in the lignin biosynthesis, such as cinnamate-4-hydroxylase (*C4H*), cinnamoyl-CoA reductase 1 (*CCR*), hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (*HCT*) and caffeoyl-CoA 3-O-methyltrans- ferase (*CCOAMT*), were specifically up-regulated in the pericarp tissue under fungal infection (Figure 2.4, A). Two transcripts encoding *C4H*, one transcript encoding *CCR1* and *CCOAMT* each were highly up-regulated under the atoxigenic strain infection, while *HCT* was highly induced under the toxigenic strain infection in pericarp. Simultaneous up-regulation of *PAL*, *4CL* and *CHS* genes of the phenylpropanoid pathway suggested that these genes are coordinately regulated in response to fungal infection and that wounding and fungal infection at pericarp induced and accumulated these transcripts at a much higher level in pericarp as a part of an induced plant defense response to the invading fungus. The genes involved in the phenylpropanoid pathway are also induced in response to wounding [46]. The production of phytoalexins and accumulation of *PAL*, *CHS* and *HRGP* upon infection indicates hypersensitivity response, thus establishing immunity of uninfected distant cells [37, 42]. The increase in lignification was also shown to be associated with hypersensitive resistant reaction in wheat in response to *Puccinia graminis* f.sp. *tritici* infection [37, 47].

Lipoxygenase (*LOX*) genes of the LOX biosynthesis pathway were differentially expressed in pericarp tissue under *A. flavus* infection. The *LOX2* gene was up-regulated under infection by both the strains in pericarp, whereas, the *LOX1* (*Gorai.004G241400*) was specifically up-regulated under the atoxigenic strain infection in pericarp. The *LOX3* gene was down-regulated in the pericarp tissue under infection by both strains (Figure 2.4, A). However, the *LOX* genes did not show any change in response to *A. flavus* infection in seed. Most volatile compounds and hydroperoxy fatty acids are the products of the LOX biosynthesis pathway [22]. These results suggested that *LOX* genes expression was more abundant in pericarp as a possible mechanism of resistance against *A. flavus* infection. Plant-derived volatile compounds have the capability to inhibit *A. flavus* growth and aflatoxin biosynthesis [22]. Previous studies have reported anti-fungal role of *LOX* genes in peanut, corn and soybean [20]. In humans, *LOX* genes were shown to be involved in degradation of aflatoxin B1 by oxidative metabolism [20].

Wounding and pathogen infection have also been shown to regulate the genes involved in cell wall biosynthesis and modifications, such as pectins, cellulose and hemicellulose [46]. The transcripts encoding xyloglucan endotransglucosylase (*XTH*), cellulose synthase (*CS*), UDP-D-galactose 4-epimerase (*UGE*), pectin methylesterase (*PME*), expansin (*EXP*) and glycosyltransferase (*GT*), which are involved in the cell wall modification, were regulated by *A. flavus* infection in both pericarp and seed (Figure 2.4, A). The results also suggested that cell-wall modifying genes and genes involved in the production of antimicrobial substances were more active in pericarp as compared to seed. It is thus evident that the atoxigenic strain of *A. flavus* played a major role in activation of the antifungal and cell-wall modifying genes in the pericarp and seed tissues. Further, the characteristic response of these genes under specific fungal strain infection and tissue will help elucidate the mechanism of defense.

2.3.4.2 The cross-talk between JA/ET and SA signaling pathway

Salicylic acid (SA), jasmonic acid (JA), ethylene (ET) and abscisic acid (ABA) are known to be involved in defense signaling pathways in response to pathogen infections and wounding [16, 17, 46, 48, 49]. SA- induced defense response generally involves resistance to biotrophic and hemibiotrophic pathogens, whereas JA and ET initiate the defense reaction in response to necrotrophic pathogens [16, 17]. The activation of SA and JA/ET pathways are pathogen dependent that involves mutually antagonistic activities [16]. JA and ET work synergistically in response to pathogen infection in plants [16]. In the present study, the transcripts encoding phospholipase, GDSL lipase, allene oxide synthase (*AOS*) and alcohol dehydrogenase (*ADH*), which are involved in the JA biosynthesis [46, 49], were induced in both pericarp and seed tissues (Figure 2.3, B). Similarly, the gene 1-aminocyclopropane-1-carboxylic acid synthase (*ACS*), involved in ET biosynthesis [46], was up-regulated in both pericarp and seed tissues (Figure 2.4, B). A higher number of lipase-encoding transcripts were induced in pericarp as compared to seed (Figure 2.4, B). In pericarp, the lipases genes showed similar expression pattern under both atoxigenic and toxigenic strains infection, whereas in seed the genes were more active under the toxigenic strain infection (Figure 2.3, B). *AOS* and *ACS6* genes were up-regulated under the atoxigenic strain infection in pericarp, while in seed AOS was up-regulated by both strains, and ACS6 was specific to toxigenic strain infection (Figure 2.3, B). The jasmonate zim-domain protein (*JAZ*) inhibits the JA signaling in plants by interacting with JIN1/MYC2 gene and represses the expression of JA responsive genes [16]. The transcript encoding *JAZ8* (*Gorai.006G092400*) was down-regulated in both pericarp and seed tissues under *A. flavus* infection (Figure 2.4, B). Further, SA mediated signaling pathway was inhibited by JA/ET signaling pathway in both pericarp and seed tissue in response to the fungal infection. The non-expresser of PR genes (NPR) which are a vital component of SA signaling pathway [16] was down-regulated in both pericarp and seed in response to *A. flavus* infection (APPENDIX I). The mitogen activated protein kinase gene, *MPK4*, acts as a positive regulator of JA and a negative regulator of SA signaling pathway in plants [16]. *MPK4* was up-regulated under the atoxigenic strain infection in pericarp. Interestingly, the glutaredoxin (*GRX*), which is identified as a negative regulator of JA/ET signaling pathway [16], was up-regulated in pericarp, but down-regulated in seed. The enhanced expression of JA-responsive marker gene plant defensin 1.2 (*PDF1.2*) is known to be associated with resistance to necrotrophic pathogens [16]. In the present study, the transcript encoding *PDF1.2c* was specifically up-regulated under the atoxigenic *A. flavus* strain infection in pericarp tissue (Figure 2.4, B). The transcription factor *ERF1* that acts as a positive regulator of JA and ET signaling pathway in *Arabidopsis* [16] was highly induced in seed by the infection with both strains of *A. flavus*. These results indicated that

JA/ET signaling pathway may be a component in the resistance mechanism of cotton to necrotrophic *A. flavus*.

2.3.4.3 Genes involved in defense signaling pathways

Plant receptor protein kinases (*RPK*) represent PRRs that are involved in the perception of pathogen signal and trigger inducible defense [50]. The transcripts similar to receptor-like protein kinase (*RLK*), leucine-rich repeat receptor-like protein kinase (*LRR-RLK*), cysteine-rich RLK (*CRK*), lectin receptor kinase, inflorescence meristem receptor-like kinase (*IMK*) and receptor kinase (*RK*) were differentially expressed in both tissues under the atoxigenic and the toxigenic *A. flavus* strains infection (Figure 2.4, B). In seed, the transcripts encoding RLK were highly induced by the atoxigenic strain. The transcripts for LRR-RLK were specifically induced by the atoxigenic strain in pericarp and the toxigenic strain in seed. The *CRK* genes were induced in both tissues by both atoxigenic and toxigenic strains (Figure 2.4, B). Plants produce elicitors that are perceived by *RPK* to amplify immunity and resistance to fungal infection [51, 52]. The elicitor CLAVATA3 (*CLV3*), secreted by shoot apical meristem, binds to the LRR-RLK and activates it [51]. In pericarp, the transcript encoding CLAVATA3 (*CLV3*) was up-regulated under both atoxigenic and toxigenic strains infection (Figure 2.4, B).

The increase in the level of Ca^{2+} is indicative of the activation of plant's innate immunity. The increase in Ca^{2+} levels is the result of release of pathogen elicitors after the infection in plants. The elevated levels of calcium under stress conditions are recognized by calcium binding proteins such as calcium dependent kinases (CDPKs), calmodulins (CaMs) and calcineurin B-like proteins (CBL), which in turn induce downstream target gene expression [46, 51, 53]. The up-regulation of a large number of transcripts encoding calcium binding proteins including CaMs, CBL and calcium binding EF-hand family proteins (CBP) in the present study suggested that there were elevated levels of Ca^{2+} after fungal infection in both pericarp and seed tissues (Figure 2.4, B). In pericarp, most of the transcripts encoding CBP were induced under the atoxigenic strain infection, whereas in seed these were under the toxigenic strain infection (Figure 2.4, B). Recent studies on plant-pathogen interactions reported that Ca^{2+}-mediated activation of CaMs, CBL interacting protein kinases (CIPKs) and CDPKs are involved in plant's immunity responses [51]. The transcripts similar to *CIPK9*, *CIPK12*, *CIPK6* and *CIPK21* were differentially expressed in both tissues under fungal infection (Figure 2.4, B). *CIPK9* and *CIPK6*, and *CIPK12* and *CIPK6* genes were induced under the atoxigenic and the toxigenic infection in pericarp, respectively (Figure 2.4, B). However, in seed, only *CIPK6* was induced under the toxigenic infection. The *CIPK* genes are involved in late immune responses (3-24 h after infection) and promote accumulation of phytoalexin, and expression of cell death and PR genes in response to fungal infection [51].

The activation of MAPK pathway is also one of the defense responses that contribute to resistance to fungal infection in plants starting as early as 1 min after infection. Wounding and pathogen elicitors induce fast activation of MAPK cascade signaling [51]. The MAPK cascade signaling involves three components: MAPK kinase kinase (MAPKKK), MAPK kinase (MAPKK) and MAPK. In MAPK signaling cascade, the MAPKKK activates MAPKK, which in turn activates MAPK. The four transcripts encoding MAPKKK15, MPK17, MPK7 and MPK4 were highly up-regulated specifically under the atoxigenic strain infection in pericarp (Figure 2.4, B). MAPK cascade signaling was not induced in seed tissue in response to the fungal infection. The activation of MAPK signaling cascade regulates the downstream transcription factors, which further induce the expression of defense related genes leading to enhanced long-term defense response and resistance to fungal infection by regulating the synthesis of antimicrobial peptides and chemicals, programmed cell death, stress hormones (JA and ET), nitric oxide (NO) and reactive oxygen species (ROS) [51, 52]. The NO synthesis in plants is catalyzed by the enzyme nitrate reductase (NIA), which plays an important role in plant defense responses [54, 55]. Under the atoxigenic strain infection, the two transcripts encoding NIA2 were up-regulated in pericarp (Figure 2.4, B). In tobacco cells, the fungal elicitors contributed to prolonged activation of MAPKs, which regulate the expression of NO and ROS [51, 56]. The ROS, which play an important role in plant defense responses together with NO, are synthesized by the phagocyte enzyme complex of NADPH oxidase [54, 56]. The respiratory burst oxidase homolog (RBOH) which is plant NADPH oxidase [56, 57] was differentially expressed in both pericarp and seed tissues of cotton under fungal infection (Figure 2.4, B). The *RBOH* is regulated by MAPK signaling cascade and its increased expression is associated with resistance to pathogens [51, 56]. The three homologs of *RBOH* including *RBOHF* (*Gorai.003G085100*, *Gorai.008G199100*) and *RBOHD* (*Gorai.009G202500*) were highly induced specifically in pericarp (Figure 2.4, B). The *RBOHF* transcripts were induced under both atoxigenic and toxigenic strains infection in pericarp, whereas *RBOHD* was specifically up-regulated under the toxigenic strain infection (Figure 2.4, B). The production of NO and ROS together are necessary for inducing the hypersensitive response (HR) and cell death in plants [54].

The phytohormone auxin (Aux), besides regulating growth and developments of plants, plays an important role in the defense responses to pathogens, [16, 58, 59]. Aux regulates the expression of *Aux/IAA*, Gretchenhagen-3 (*GH3*) family, Auxin response factor (*ARF*) and small auxin-up RNA (*SAUR*) genes [16, 58, 60]. Over-expression of *OsGH3.1* in rice enhanced the resistance to fungal pathogen by reducing the auxin level and suppressing the expression of expansin genes [16, 59]. In this study, the transcripts similar to *GH3.1* and *GH3.10* were specifically induced in pericarp by both strains (Figure 2.4, B). The *GH3.6* (*Gorai.005G208000*) was specifically induced under the toxigenic strain infection in pericarp. Similarly, *SAUR* family genes were differentially expressed in pericarp and seed tissues under *A. flavus* infection (Figure 2.4, B). *SAUR* genes have inhibitory activity on auxin biosynthesis and transport [60]. The over-expression of *SAUR39* transcript in rice showed reduced free IAA level and auxin transport [60].

The role of phytohormone cytokinin has also been elucidated in disease resistance reaction in *Arabidopsis* [16, 61]. *Arabidopsis* lines overexpressing cytokinin oxidase/dehydrogenase (*CKX*) showed enhanced resistance to clubroot disease [61]. The transcripts similar to *CKX6* (*Gorai.012G081000* and *Gorai.011G295400*) were induced under atoxigenic and toxigenic strains infection only in pericarp. Contrastingly, *CKX3* was down-regulated in pericarp under the atoxigenic strain infection (Figure 2.4, B). The detailed characterization of these genes is necessary to understand the role and cross-talk of auxin- and cytokinin-signaling pathways in *A. flavus*-mediated defense response in cotton.

2.3.4.4 Genes encoding transcription factors (TFs)

Transcription factors control the transcriptional regulation by activating or suppressing the expression of downstream genes in response to pathogens infection [62]. The transcription factor *GhWRKY3* is known to be induced under wounding and fungal infection in cotton [62]. Mutation in *WRKY70* and *WRKY33* enhanced the susceptibility of *Arabidopsis* to necrotrophic *B. cinerea* fungal infection [17]. *WRKY40*, *WRKY33*, *WRKY53*, *WRKY22*, *WRKY11*, *WRKY15* and *WRKY60* are known to be induced under wounding in *Arabidopsis* [46]. In the present study, 28 WRKY-related transcripts were differentially expressed under fungal infection in pericarp and seed of cotton (Figure 2.5, A). The *WRKY75* (*Gorai.001G057600* and *Gorai.005G164300*) and *WRKY72* were induced in both pericarp and seed in response to *A. flavus* infection. Most of the

WRKY TFs were up-regulated under the atoxigenic strain infection in pericarp and the toxigenic strain infection in seed (Figure 2.5, A). *WRKY75* (*Gorai.006G043200*) and *WRKY40* (*Gorai.009G124000*) were specifically up-regulated under the atoxigenic strain infection in both pericarp and seed. *WRKY6* (*Gorai.001G214800*), *WRKY53* (*Gorai.008G253300*), *WRKY50* (*Gorai.001G021500*) and *WRKY41* (*Gorai.007G014600*) were down-regulated in pericarp. *WRKY50* is reported to negatively regulate the plant-fungus interaction, and mutation in *WRKY50* has been associated with enhanced resistance to pathogen [17]. It was also reported that the WRKY TFs regulate the expression of RLK genes, which are induced in response to pathogen infection [46, 63].

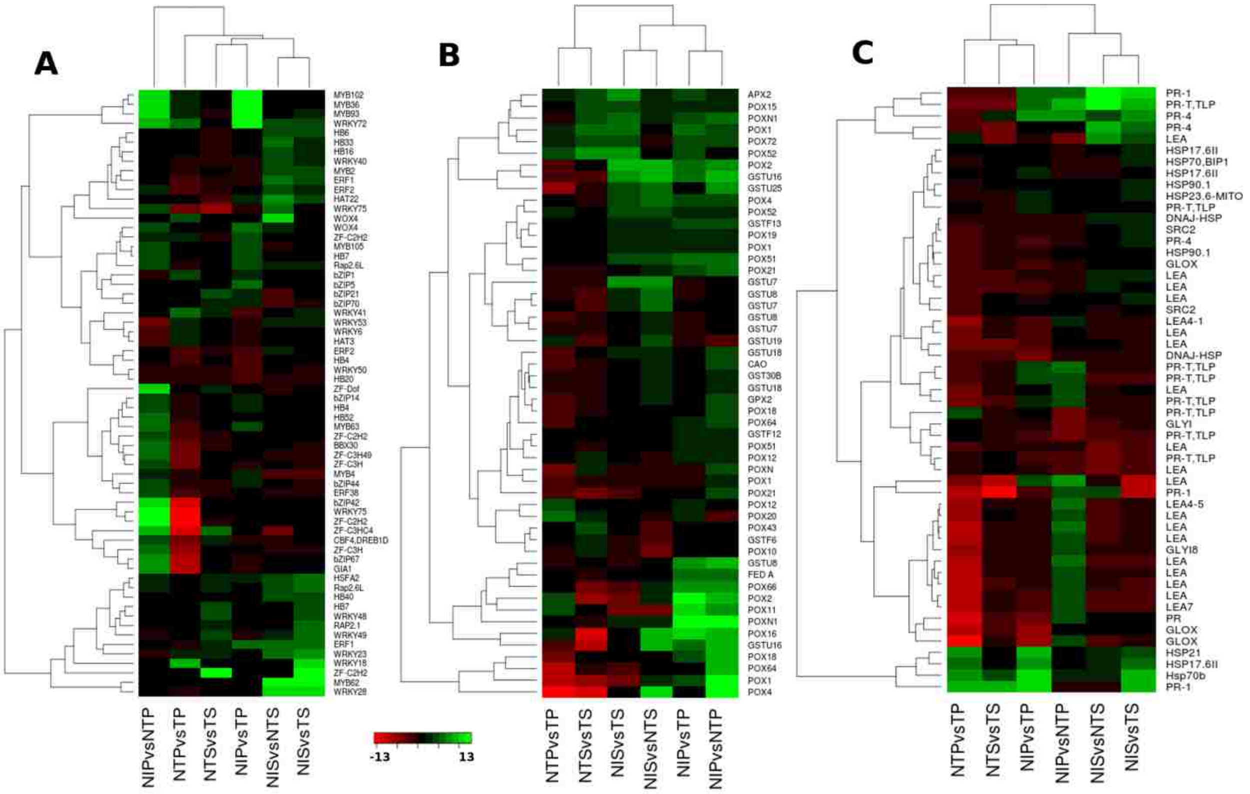


Figure 2.5 Heatmaps showing DEGs involved in transcriptional regulation (A), involved in oxidative burst (B) and stress response (C). The green color represents up-regulated (log2FC>=2) genes and red color represents down-regulated (log2FC≤-2) genes. For gene names represented in the heatmaps, please refer to the APPENDIX I.

The APETALA2 (AP2)/Ethylene responsive factor (ERF) family of TFs are known to be involved in the activation of defense-related genes in response to pathogen infection through ET and JA pathways [17, 64]. The members of AP2/ERF family were differentially expressed in response to the necrotrophic *A. flavus* infection in pericarp and seed (Figure 2.5, A). The transcripts similar to *RAP2.6L* (*Gorai.012G125700*) and *ERF38* were up-regulated specifically in pericarp (Figure 2.5, A). *RAP2.6L* was induced under both atoxigenic and toxigenic strains infection, whereas, *ERF38* expression was specific to the atoxigenic strain infection. The number of AP2/ERF TFs up-regulated in seed was higher as compared to pericarp. *ERF1* (*Gorai.005G049300*) and *RAP2.6L* (*Gorai.005G197100*) were induced specifically in seed under both strain infections. In seed, *ERF2* (*Gorai.009G165500*) and *RAP2.1* were up-regulated under the atoxigenic and the toxigenic strain infection, respectively. However, *ERF2* (*Gorai.010G156600*) was down-regulated under the toxigenic strain infection in pericarp. The AP2/ERF TF family genes are known to be highly induced in response to wounding in *Arabidopsis* [46]. Overexpression of an ERF TF in Arabidopsis conferred enhanced resistance to necrotrophic fungus *B. cinerea* [17]. A transcript encoding *DREB1D/CBF4* gene was induced under the atoxigenic strain infection in pericarp (Figure 2.5, A). *DREB/CBF* TFs belong to *AP2/ERF* TF family and have been reported to be induced in response to wounding, in addition to abiotic stresses, such as cold, heat, salt and drought in *Arabidopsis* [46].

The role of bZIP family TFs in biotic stress responses of plants is well established [65]. The bZIP-family TFs were induced in pericarp of cotton by *A. flavus* infection. The five bZIP TFs, *bZIP14*, *GIA1*, *bZIP67*, *bZIP44* and *bZIP42*, were specifically induced under the atoxigenic strain infection, whereas *bZIP1* and *bZIP5* were induced under the toxigenic strain infection

(Figure 2.5, A). The bZIP TFs were not induced by the fungus in seed; *bZIP70* and *bZIP21* were down-regulated under the atoxigenic strain infection in seed.

Many MYB TFs were also differentially expressed under *A. flavus* infection in cotton. More MYB genes were up-regulated in pericarp as compared to seed (Figure 2.5, A). *MYB93*, *MYB63*, *MYB102*, *MYB105*, *MYB102* and *MYB36* were induced under both the atoxigenic and the toxigenic strains infection in pericarp, whereas *MYB62* and *MYB2* were induced in seed (Figure 2.5, A). The gene expression pattern of MYB TFs has been studied in Arabidopsis in response to wounding and pathogen infection [46]. The MYB TFs regulate the expression of flavonoid genes, PR genes and genes involved in secondary metabolism [46]. In the present study, *MYB4*, which is a repressor of phenylpropanoid pathway [49], was down-regulated under both strains infection in seed (Figure 2.5, A).

The gene expression profiles of zinc finger (ZF), heat shock factors (HSF) and homeobox (HB) type TFs were altered in response to wounding in Arabidopsis [46]. In cotton, a number of transcripts encoding ZF TFs were up-regulated in response to the atoxigenic strain infection in pericarp, whereas in seed, only one transcript similar to *ZF-C2H2* (*Gorai.002G223300*) was induced under the toxigenic strain infection (Figure 2.5, A). The HB type TFs showed high activity in seed than pericarp, which was evident from the up-regulation of a large number of HB genes in seed as compared to pericarp (Figure 2.5, A). The HB TFs, *HAT3*, *HB20* and *HB4*, were down-regulated in pericarp under *A. flavus* infection. The HSFs did not show any activity in pericarp, and were specifically induced in seed (Figure 2.5, A). The transcript showing similarity with *HSFA2* was highly up-regulated in seed under both strains infections. Further characterization of these TFs is necessary to understand their regulatory mechanisms in controlling expression of downstream genes and their roles in defense response of cotton to *A. flavus* infection.

2.3.4.5 Genes involved in oxidative burst

Oxidative burst is one of the earliest defense responses of plants against pathogen infection and wounding, and is considered as the hallmark of pathogen recognition [46, 56, 57, 66]. ROS production is observed in both plant triggered immunity (PTI) and effector triggered immunity ETI [57]. In addition to their involvement in direct defense reaction by killing the pathogens, ROS are also involved in the activation of defense-related genes through signaling mechanism [57]. ROS can regulate the TFs and produce antimicrobial phytoalexins and other secondary metabolites, which have inhibitory activity on pathogen growth [57]. As discussed earlier, the MAPK pathway activated the expression of *RBOH* genes in response to *A. flavus* infection, which could trigger the plant apoplastic oxidative burst. The transcripts encoding glutathione S-transferase (GST), ascorbate peroxidase (APX), copper amine oxidase (CAO), ferredoxin (FED) and peroxidase (POX), which are involved in ROS processing and scavenging, showed increased activity under *A. flavus* infection in cotton (Figure 2.5, B). Most of these genes were induced in pericarp, and their expression patterns were different under the atoxigenic and the toxigenic strains infection in both pericarp and seed tissues. The transcripts similar to copper amine oxidase (CAO) under atoxigenic strain infection and ferredoxin under both strains infections were specifically induced in pericarp tissue (Figure 2.5, B). Most of the transcripts encoding glutathione S-transferase (GST) were induced under the atoxigenic strain infection in both pericarp and seed (Figure 2.5, B). The expression patterns for peroxidase genes (POX) were similar under the atoxigenic and the toxigenic strain infection in pericarp, but in seed POX genes were highly induced under the toxigenic strain infection (Figure 2.5, B). The peroxidase 2

(*POX2*) genes were specifically induced in pericarp tissue under both the atoxigenic and the toxigenic strains infection. The transcript similar to glutathione peroxidase (*GPX2*) was induced under the atoxigenic strain infection in pericarp (Figure 2.5, B). Hydrogen peroxide can also act as a secondary messenger and initiate regulation of defense related genes [46]. These results suggested that wounding and subsequent *A. flavus* infection activated the ROS-regulated defense response in both pericarp and seed tissues of cotton.

2.3.4.6 Genes involved in stress response

Many stress responsive genes have been implicated in plant's response to fungal infection [4, 22]. The late embryogenesis abundant (LEA) storage protein was shown to be induced in response to *A. flavus* infection in cotton [4]. In this study, 12 transcripts similar to LEA genes were specifically up-regulated in response to the atoxigenic *A. flavus* infection in pericarp (Figure 2.5, C). In seed, only 4 *LEA4* transcripts were differentially induced under the atoxigenic and the toxigenic strains infection. The heat shock proteins (HSP) play a major protective role in biotic and abiotic stresses by controlling chaperone activity and other cellular processes [22, 46]. The expression of HSPs is under the regulation of HSFs [46]. There are several HSPs and HSFs that have been reported to be induced by pathogen infection and wounding [22, 46]. Most of the transcripts similar to HSPs, in this study, were induced in seed as compared to pericarp tissue (Figure 2.5, C). *HSP70B* was specifically induced under the atoxigenic strain infection in pericarp. *HSP23.6-MITO*, *HSP90.1*, *HSP21* and *HSP17.6II* were all induced in both pericarp and seed tissues by both the strains. *HSP90.1* and *DNAJ-HSP* were up-regulated by both the strains specifically in seed, whereas *HSP70* was up-regulated under the toxigenic strain infection. The *DNAJ-HSP* (*Gorai.008G099300*) was down-regulated in pericarp tissue. The up-regulation of a large number of HSPs in seed as compared to pericarp could be correlated to the induction of

HSFs specifically in seed tissue, as discussed earlier. Stress-inducible cold regulated gene (*SRC2*) was specifically up-regulated in the seed tissue (Figure 2.5, C). Two transcripts encoding glyoxal oxidase-related protein were specifically induced under the toxigenic strain infection in seed (Figure 2.5, C). The overexpression of glyoxal oxidase gene was shown to enhance the resistance of grape plant to fungal infection [67]. The stress responsive gene glyoxalase I was also known to be induced in response to abiotic and biotic stresses in plants [22, 68]. Glyoxalase I was up-regulated in response to necrotrophic hemibiotroph fungus *T. basicola* in *G. hirsutum* [68]. The expression of a transcript coding for a glyoxalase I family protein, *GLYI8*, was specific to pericarp and up-regulated in response to the atoxigenic *A. flavus* infection (Figure 2.5, C). The pathogenesis related genes (PR) that are associated with the resistance reactions [22, 68, 69] were also differentially expressed in the pericarp and seed tissues (Figure 2.5, C). Among the 15 differentially expressed transcripts that were similar to PR genes, seven and three were specific to pericarp and seed, respectively. In pericarp, *PR* and *PR-1* (*Gorai.006G115900*) were up-regulated specifically under the atoxigenic strain infection, and *PR-T* (*Gorai.007G193600*) was up-regulated under the toxigenic strain infection. Two PR genes (*PR-4* and *PR-1*) were equally induced in seed, while *PR-T* (*Gorai.009G194000*) was down-regulated under both strains infection in seed. Most of these stress responsive genes have also been known to be induced under abiotic stress conditions.

2.4 GO AND KEGG ENRICHMENT ANALYSIS OF DEGS

In this study, GO enrichment analysis of the DEGs identified the functional categories, such as biological process, molecular function and cellular component that were distinctly represented by the atoxigenic and the toxigenic strain infection in pericarp and seed tissues of cotton. The GO categories under biological process, such as response to stimulus (GO:0050896), response to wounding (GO:0009611), response to external stimulus (GO:0009605), response to chemical stimulus (GO:0042221) and flavonoid biosynthesis (GO:0009813) were highly represented under the atoxigenic strain infection in pericarp. On the other hand, response to biotic stimulus (GO:0009607), response to stress (GO:0006950), regulation of defense response (GO:0031347), response to chitin (GO:0010200), defense response to fungus (GO:0050832) and signal transduction (GO:0007165) were enriched under the toxigenic strain infection in seed (Figure 2.6). Under molecular function category, most of the responses were highly enriched in pericarp as compared to seed. In pericarp, transcription regulator activity (GO:0030528), transcription factor activity (GO:0003700), transmembrane transporter activity (GO:0022857)

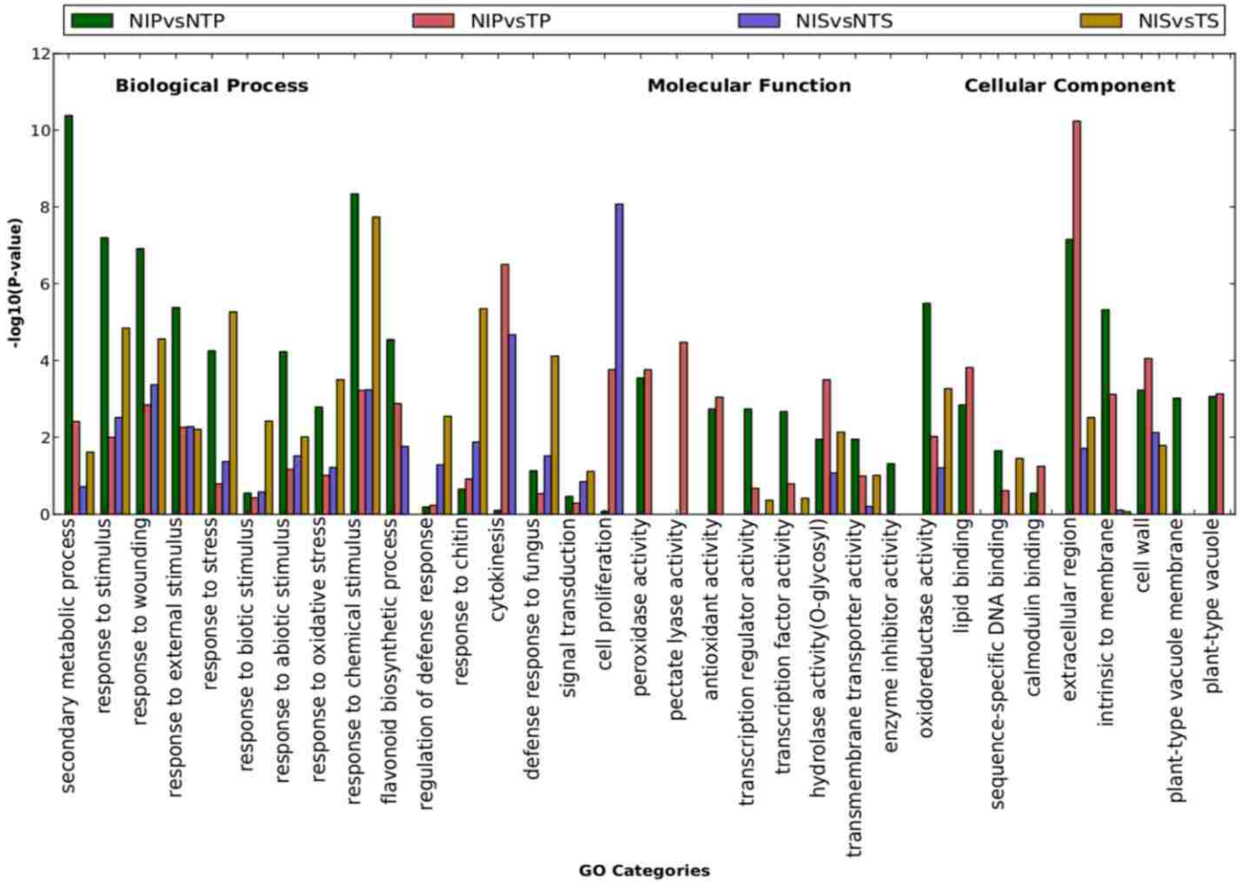


Figure 2.6 Gene ontology enrichment analysis of DEGs in pericarp and seed tissues of cotton in response to infection with atoxigenic and toxigenic strains of *Aspergillus flavus*. The X-axis represents the GO categories and Y-axis represents enrichment in terms of P-value.

and enzyme inhibitor activity (GO:0004857) were enriched under the atoxigenic strain infection in comparison with the toxigenic strain infection (Figure 2.6). Peroxidase activity (GO:0004601), pectate lyase activity (GO:0030570), antioxidant activity (GO: 0016209) and hydrolase activity (GO:0004553) were enriched under the toxigenic strain infection in pericarp as compared to the atoxigenic strain infection (Figure 2.6). Most of the cellular components were enriched in pericarp as compared to seed tissue (Figure 2.6). Component of cell wall, membrane and vacuoles were differentially enriched under the atoxigenic and the toxigenic strain infection in pericarp and seed.

Analysis of the biochemical pathways represented by the DEGs showed that 94, 77, 59, and 63 KEGG pathways were represented under the atoxigenic and the toxigenic strains infection in pericarp and seed, respectively. Highly enriched pathways ($P < 0.05$) in pericarp and seed are shown in Figure 2.7. The phenylpropanoid pathway, which is involved in the production of antimicrobial phytoalexins, lignins and phenolic substances [37, 42], was enriched in the toxigenic strain infection in pericarp, followed by the atoxigenic infection in pericarp and seed. The flavonoid biosynthesis pathway was the most highly enriched under the atoxigenic strain infection in seed followed by pericarp. Genes in the flavonoid pathway are involved in the production of antifungal compounds and are associated with defense reactions [70]. The alkaloid biosynthesis pathway was highly enriched under both atoxigenic and toxigenic strains infection in pericarp as compared to seed (Figure 2.7). In tobacco plants, alkaloid biosynthesis is induced in response to insect damage and application of jasmonate [71]. This suggests that JA-regulated defense response was activated in cotton in response to *A. flavus* infection. Further, enrichment of arachidonic acid (AA) metabolism was observed under the toxigenic strain infection in seed followed by the atoxigenic infection in pericarp (Figure 2.7). AA acts as a signaling molecule,

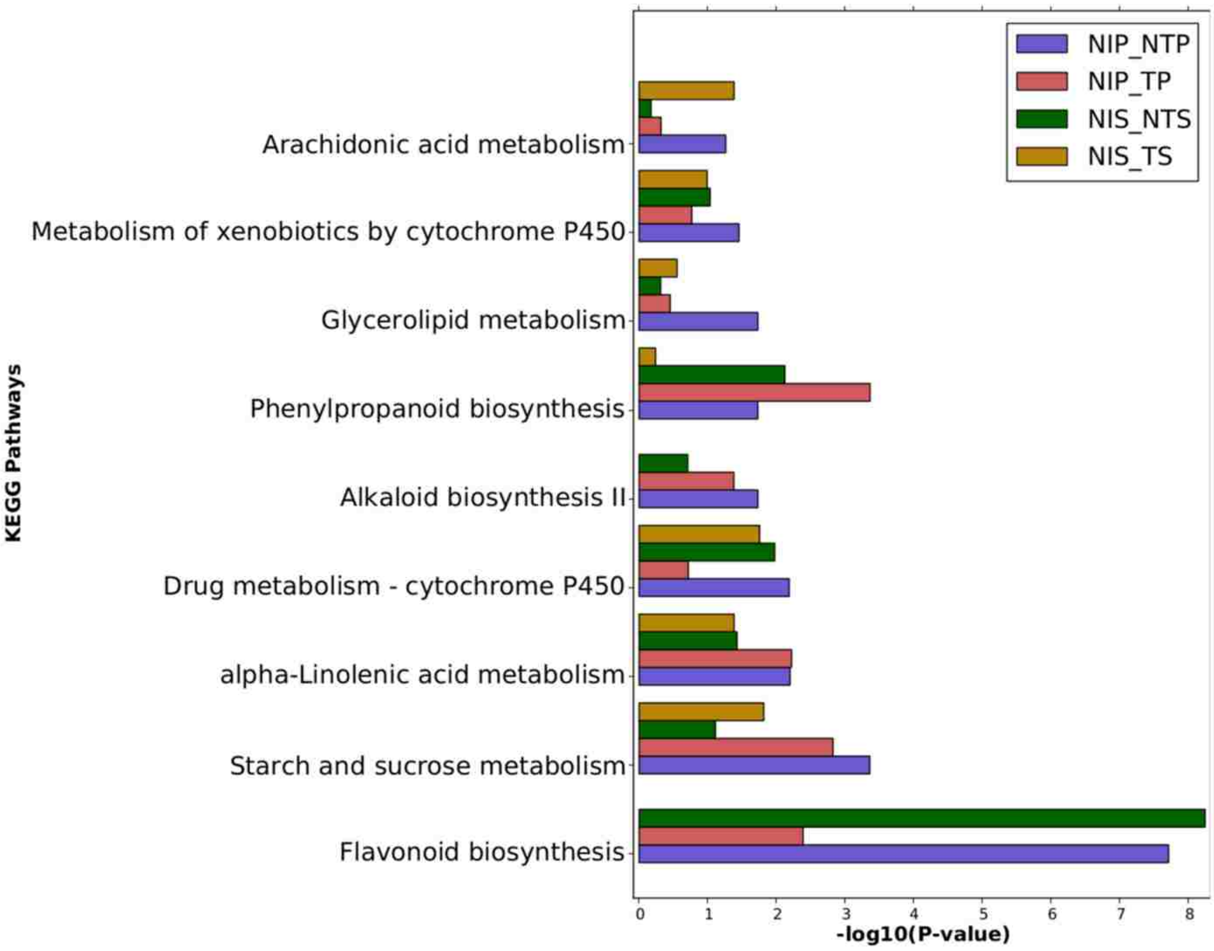


Figure 2.7 Highly represented KEGG metabolic pathways in pericarp and seed tissues of cotton under *Aspergillus flavus* infection. The X-axis represents the enrichment in terms of P-value and Y-axis represents the biochemical pathways.

and activates plant's defense responses through fatty acids. AA is a potent elicitor present in the pathogen, which activates plant innate immunity leading to programmed cell death and defense responses [72]. The alpha-linolenic acid metabolism pathway was enriched in pericarp in comparison to seed under both atoxigenic and toxigenic strains infection (Figure 2.7). JA and its derivatives, which are key regulators of plant defense responses to necrotrophic pathogens, are synthesized from the alpha-linolenic acid pathway [73, 74]. The primary metabolic pathways, such as starch and sucrose metabolism and glycerolipid metabolism, were also highly enriched under the atoxigenic strain infection in pericarp (Figure 2.7). The up-regulation of carbohydrate, amino acids and lipid metabolisms was suggested to regulate the signal transduction cascade during plant defense responses [75]. The biochemical pathways involved in response to the atoxigenic strain of *A. flavus* infection in pericarp and seed tissue of cotton can be manipulated for stress tolerance in cotton.

Validity of the next generation sequence data was confirmed by reverse-transcription PCR of 10 genes belonging to different functional categories with fold change expression of 5-fold or above (from sequence data) under experimental conditions relative to non-inoculated control. The results showed significant up-regulation of their mRNA accumulation under infection by the atoxigenic or the toxigenic strain infection in a tissue-dependent manner

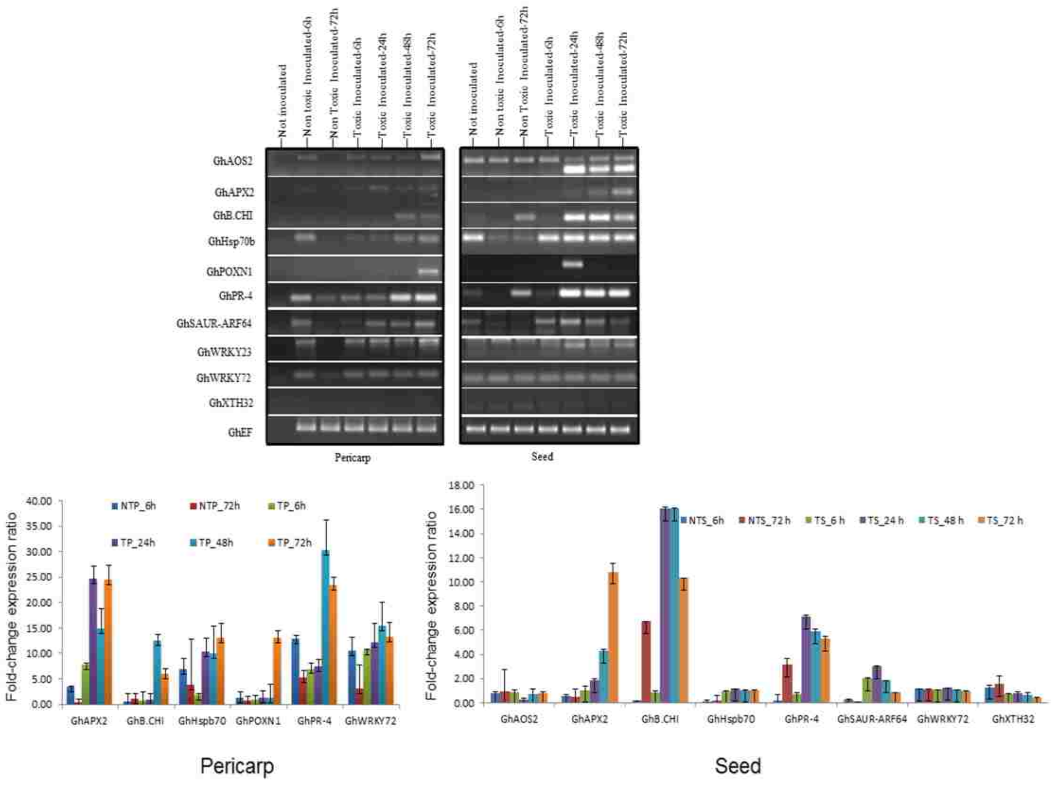


Figure 2.8 Gel image (upper panel) showing semiquantitative RT-PCR and fold-change expression through quantitative RT-PCR (lower panel) of genes under infection by atoxigenic and toxigenic strains of *Aspergillus flavus* in pericarp and seed tissues of cotton

2.5 REFERENCES CITED

1. Cleveland TE, Dowd PF, Desjardins AE, Bhatnagar D, Cotty PJ: **United States Department of Agriculture - Agricultural Research Service research on pre-harvest prevention of mycotoxins and mycotoxigenic fungi in US crops**. *Pest Manag Sci* 2003, **59**(6-7):629-642.

2. Wogan GN: **Impacts of chemicals on liver cancer risk**. *Semin Cancer Biol* 2000, **10**(3):201-210.

3. (IARC) IAfRoC: **Some Naturally Occurring Substances: Food Items and Constituents, Heterocyclic Aromatic Amines and Mycotoxins**. In: *IARC Monographs on the evaluation of the carci- nogenic risks to humans.* vol. 56. Lyon, France: INTERNATIONAL AGENCY FOR RESEARCH ON CANCER; 1993: 245–540.

4. Lee S, Rajasekaran K, Ramanarao MV, Bedre R, Bhatnagar D, Baisakh N: **Identifying cotton (Gossypium hirsutum L.) genes induced in response to *Aspergillus flavus* infection**. *Physiol Mol Plant P* 2012, **80**:35-40.

5. Yin YN, Yan LY, Jiang JH, Ma ZH: **Biological control of aflatoxin contamination of crops**. *J Zhejiang Univ-Sc B* 2008, **9**(10):787-792.

6. Kelley RY, Williams WP, Mylroie JE, Boykin DL, Harper JW, Windham GL, Ankala A, Shan XY: **Identification of Maize Genes Associated with Host Plant Resistance or Susceptibility to *Aspergillus flavus* Infection and Aflatoxin Accumulation**. *Plos One* 2012, **7**(5).

7. Richard JL PG: **Mycotoxins: Risks in Plant, Animal, and Human Systems**. Ames, IA: Council for Agricultural Science and Technology; 2003.

8. Yu JJ: **Current Understanding on Aflatoxin Biosynthesis and Future Perspective in Reducing Aflatoxin Contamination**. *Toxins* 2012, **4**(11):1024-1057.

9. Van Egmond H, Jonker M, Abbas H: **Worldwide regulations on aflatoxins**. *Aflatoxin and food safety* 2005:77-93.

10. Lillehoj E, Wall J: **Decontamination of aflatoxin-contaminated maize grain**. In: *US Universities-CIMMYT Maize Aflatoxin Workshop, El Batan, Mexico (Mexico), 7-11 Apr 1986: 1987*: CIMMYT; 1987.

11. Wright MS, Greene-McDowelle DM, Zeringue HJ, Bhatnagar D, Cleveland TE: **Effects of volatile aldehydes from Aspergillus-resistant varieties of corn on Aspergillus parasiticus growth and aflatoxin biosynthesis**. *Toxicon* 2000, **38**(9):1215-1223.

12. Cotty PJ: **Influence of Field Application of an Atoxigenic Strain of Aspergillus-Flavus on the Populations of *Aspergillus flavus* Infecting Cotton Bolls and on the Aflatoxin Content of Cottonseed**. *Phytopathology* 1994, **84**(11):1270-1277.

13. Dorner JW: **Biological control of aflatoxin contamination of crops**. *J Toxicol-Toxin Rev* 2004, **23**(2-3):425-450.

14. Pitt JI, Hocking AD: **Mycotoxins in Australia: biocontrol of aflatoxin in peanuts**. *Mycopathologia* 2006, **162**(3):233-243.

15. Dorner JW: **Management and prevention of mycotoxins in peanuts**. *Food Addit Contam* 2008, **25**(2):203-208.

16. Bari R, Jones J: **Role of plant hormones in plant defence responses**. *Plant Mol Biol* 2009, **69**(4):473-488.

17. Birkenbihl RP, Somssich IE: **Transcriptional plant responses critical for resistance towards necrotrophic pathogens**. *Front Plant Sci* 2011, **2**.

18. St Clair DA: **Quantitative Disease Resistance and Quantitative Resistance Loci in Breeding**. *Annu Rev Phytopathol* 2010, **48**:247-268.

19. Yu J, Bhatnagar D, Cleveland TE, Payne G, Nierman WC, Bennett JW: **15 *Aspergillus flavus* Genetics and Genomics in Solving Mycotoxin Contamination of Food and Feed**. *OMICs technologies: Tools for Food Science* 2012:367.

20. Guo BZ, Fedorova ND, Chen XP, Wan CH, Wang W, Nierman WC, Bhatnagar D, Yu JJ: **Gene Expression Profiling and Identification of Resistance Genes to *Aspergillus flavus* Infection in Peanut through EST and Microarray Strategies**. *Toxins* 2011, **3**(7):737-753.

21. Luo M, Brown RL, Chen ZY, Menkir A, Yu JJ, Bhatnagar D: **Transcriptional Profiles Uncover *Aspergillus flavus*-Induced Resistance in Maize Kernels**. *Toxins* 2011, **3**(7):766-786.

22. Cleveland TE, Yu JJ, Bhatnagar D, Chen ZY, Brown RL, Chang PK, Cary JW: **Progress in elucidating the molecular basis of the host plant - *Aspergillus flavus* interaction, a basis for devising strategies to reduce aflatoxin contamination in crops**. *J Toxicol-Toxin Rev* 2004, **23**(2-3):345-380.

23. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-U130.

24. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

25. Du Z, Zhou X, Ling Y, Zhang ZH, Su Z: **agriGO: a GO analysis toolkit for the agricultural community**. *Nucleic Acids Res* 2010, **38**:W64-W70.

26. Goffard N, Weiller G: **PathExpress: a web-based tool to identify relevant pathways in gene expression data**. *Nucleic Acids Res* 2007, **35**:W176-W181.

27. Wang KB, Wang ZW, Li FG, Ye WW, Wang JY, Song GL, Yue Z, Cong L, Shang HH, Zhu SL *et al*: **The draft genome of a diploid cotton Gossypium raimondii**. *Nat Genet* 2012, **44**(10):1098-+.

28. Li FG, Fan GY, Wang KB, Sun FM, Yuan YL, Song GL, Li Q, Ma ZY, Lu CR, Zou CS *et al*: **Genome sequence of the cultivated cotton Gossypium arboreum**. *Nat Genet* 2014, **46**(6):567-572.

29. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

30. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-U354.

31. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-U174.

32. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nat Biotechnol* 2013, **31**(1):46-+.

33. Drew DP, Dueholm B, Weitzel C, Zhang Y, Sensen CW, Simonsen HT: **Transcriptome Analysis of Thapsia laciniata Rouy Provides Insights into Terpenoid Biosynthesis and Diversity in Apiaceae**. *Int J Mol Sci* 2013, **14**(5):9080-9098.

34. Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TWA, Canesin LEC, Pinto LR, Carneiro MS, Garcia AAF, de Souza AP, Vicentini R: **De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties**. *Plos One* 2014, **9**(2).

35. Tohidfar M, Mohammadi M, Ghareyazie B: **Agrobacterium-mediated transformation of cotton (Gossypium hirsutum) using a heterologous bean chitinase gene**. *Plant Cell Tiss Org* 2005, **83**(1):83-96.

36. Wang SY, Ye XY, Chen J, Rao PF: **A novel chitinase isolated from Vicia faba and its antifungal activity**. *Food Res Int* 2012, **45**(1):116-122.

37. Collinge DB, Slusarenko AJ: **Plant Gene-Expression in Response to Pathogens**. *Plant Mol Biol* 1987, **9**(4):389-410.

38. Prasad K, Bhatnagar-Mathur P, Waliyar F, Sharma KK: **Overexpression of a chitinase gene in transgenic peanut confers enhanced resistance to major soil borne and foliar fungal pathogens**. *J Plant Biochem Biot* 2013, **22**(2):222-233.

39. Rohini VK, Rao KS: **Transformation of peanut (Arachis hypogaea L.) with tobacco chitinase gene: variable response of transformants to leaf spot disease**. *Plant Sci* 2001, **160**(5):889-898.

40. Baisakh N, Datta K, Oliva N, Ona I, Rao G, Mew T, Datta S: **Rapid development of homozygous transgenic rice using anther culture harboring rice chitinase gene for enhanced sheath blight resistance**. *Plant Biotechnology* 2001, **18**(2):101-108.

41. Chen Z-Y, Brown R, Russin J, Lax A, Cleveland T: **A corn trypsin inhibitor with antifungal activity inhibits *Aspergillus flavus* α-amylase**. *Phytopathology* 1999, **89**(10):902-907.

42. Lawton MA, Lamb CJ: **Transcriptional Activation of Plant Defense Genes by Fungal Elicitor, Wounding, and Infection**. *Mol Cell Biol* 1987, **7**(1):335-341.

43. Chappell J, Hahlbrock K: **Transcription of Plant Defense Genes in Response to Uv-Light or Fungal Elicitor**. *Nature* 1984, **311**(5981):76-78.

44. Ecker JR, Davis RW: **Plant Defense Genes Are Regulated by Ethylene**. *P Natl Acad Sci USA* 1987, **84**(15):5202-5206.

45. HammondKosack KE, Jones JDG: **Resistance gene-dependent plant defense responses**. *Plant Cell* 1996, **8**(10):1773-1791.

46. Cheong YH, Chang H-S, Gupta R, Wang X, Zhu T, Luan S: **Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in Arabidopsis**. *Plant Physiol* 2002, **129**(2):661-677.

47. Beardmore J, Ride JP, Granger JW: **Cellular Lignification as a Factor in the Hypersensitive Resistance of Wheat to Stem Rust**. *Physiol Plant Pathol* 1983, **22**(2):209-&.

48. Sanchez-Vallet A, Lopez G, Ramos B, Delgado-Cerezo M, Riviere MP, Llorente F, Fernandez PV, Miedes E, Estevez JM, Grant M *et al*: **Disruption of Abscisic Acid Signaling Constitutively Activates Arabidopsis Resistance to the Necrotrophic Fungus Plectosphaerella cucumerina**. *Plant Physiol* 2012, **160**(4):2109-2124.

49. Dolezal AL, Shu XM, OBrian GR, Nielsen DM, Woloshuk CP, Boston RS, Payne GA: ***Aspergillus flavus* infection induces transcriptional and physical changes in developing maize kernels**. *Front Microbiol* 2014, **5**.

50. Garcia-Brugger A, Lamotte O, Vandelle E, Bourque S, Lecourieux D, Poinssot B, Wendehenne D, Pugin A: **Early signaling events induced by elicitors of plant Defenses**. *Mol Plant Microbe In* 2006, **19**(7):711-724.

51. Tena G, Boudsocq M, Sheen J: **Protein kinase signaling networks in plant innate immunity**. *Curr Opin Plant Biol* 2011, **14**(5):519-529.

52. Dodds PN, Rathjen JP: **Plant immunity: towards an integrated view of plant-pathogen interactions**. *Nat Rev Genet* 2010, **11**(8):539-548.

53. Asano T, Hayashi N, Kikuchi S, Ohsugi R: **CDPK-mediated abiotic stress signaling**. *Plant Signal Behav* 2012, **7**(7):817-821.

54. Yoshioka H, Mase K, Yoshioka M, Kobayashi M, Asai S: **Regulatory mechanisms of nitric oxide and reactive oxygen species generation and their role in plant immunity**. *Nitric Oxide-Biol Ch* 2011, **25**(2):216-221.

55. Wang P, Du Y, Li Y, Ren D, Song C-P: **Hydrogen peroxide–mediated activation of MAP kinase 6 modulates nitric oxide biosynthesis and signal transduction in Arabidopsis**. *The Plant Cell* 2010, **22**(9):2981-2998.

56. Asai S, Ohta K, Yoshioka H: **MAPK signaling regulates nitric oxide and NADPH oxidase-dependent oxidative bursts in Nicotiana benthamiana**. *Plant Cell* 2008, **20**(5):1390-1406.

57. Torres MA: **ROS in biotic interactions**. *Physiol Plantarum* 2010, **138**(4):414-429.

58. Ghanashyam C, Jain M: **Role of auxin-responsive genes in biotic stress responses**. *Plant Signal Behav* 2009, **4**(9):846-848.

59. Domingo C, Andres F, Tharreau D, Iglesias DJ, Talon M: **Constitutive Expression of OsGH3.1 Reduces Auxin Content and Enhances Defense Response and Resistance to a Fungal Pathogen in Rice**. *Mol Plant Microbe In* 2009, **22**(2):201-210.

60. Kant S, Bi YM, Zhu T, Rothstein SJ: **SAUR39, a Small Auxin-Up RNA Gene, Acts as a Negative Regulator of Auxin Synthesis and Transport in Rice**. *Plant Physiol* 2009, **151**(2):691-701.

61. Siemens J, Keller I, Sarx J, Kunz S, Schuller A, Nagel W, Schmulling T, Parniske M, Ludwig-Muller J: **Transcriptome analysis of Arabidopsis clubroots indicate a key role for cytokinins in disease development**. *Mol Plant Microbe In* 2006, **19**(5):480-494.

62. Guo R, Yu F, Gao Z, An H, Cao X, Guo X: **GhWRKY3, a novel cotton (Gossypium hirsutum L.) WRKY gene, is involved in diverse stress responses**. *Mol Biol Rep* 2011, **38**(1):49-58.

63. Du L, Chen Z: **Identification of genes encoding receptor-like protein kinases as possible targets of pathogen- and salicylic acid-induced WRKY DNA-binding proteins in Arabidopsis**. *Plant J* 2000, **24**(6):837-847.

64. Lorenzo O, Piqueras R, Sanchez-Serrano JJ, Solano R: **ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense**. *Plant Cell* 2003, **15**(1):165-178.

65. Singh KB, Foley RC, Onate-Sanchez L: **Transcription factors in plant defense and stress responses**. *Curr Opin Plant Biol* 2002, **5**(5):430-436.

66. Bolwell GP, Wojtaszek P: **Mechanisms for the generation of reactive oxygen species in plant defence–a broad perspective**. *Physiol Mol Plant P* 1997, **51**(6):347-366.

67. Guan X, Zhao H, Xu Y, Wang Y: **Transient expression of glyoxal oxidase from the Chinese wild grape Vitis pseudoreticulata can suppress powdery mildew in a susceptible genotype**. *Protoplasma* 2011, **248**(2):415-423.

68. Coumans JV, Poljak A, Raftery MJ, Backhouse D, Pereg-Gerk L: **Analysis of cotton (Gossypium hirsutum) root proteomes during a compatible interaction with the black root rot fungus Thielaviopsis basicola**. *Proteomics* 2009, **9**(2):335-349.

69. Pritsch C, Muehlbauer GJ, Bushnell WR, Somers DA, Vance CP: **Fungal development and induction of defense response genes during early infection of wheat spikes by Fusarium graminearum**. *Mol Plant Microbe Interact* 2000, **13**(2):159-169.

70. Treutter D: **Significance of flavonoids in plant resistance and enhancement of their biosynthesis**. *Plant Biol (Stuttg)* 2005, **7**(6):581-591.

71. Todd AT, Liu E, Polvi SL, Pammett RT, Page JE: **A functional genomics screen identifies diverse transcription factors that regulate alkaloid biosynthesis in Nicotiana benthamiana**. *Plant J* 2010, **62**(4):589-600.

72. Savchenko T, Walley JW, Chehab EW, Xiao Y, Kaspi R, Pye MF, Mohamed ME, Lazarus CM, Bostock RM, Dehesh K: **Arachidonic acid: an evolutionarily conserved signaling molecule modulates plant stress signaling networks**. *Plant Cell* 2010, **22**(10):3193-3205.

73. Wasternack C: **Action of jasmonates in plant stress responses and development--applied aspects**. *Biotechnol Adv* 2014, **32**(1):31-39.

74. Robert-Seilaniantz A, Grant M, Jones JD: **Hormone crosstalk in plant disease and defense: more than just jasmonate-salicylate antagonism**. *Annu Rev Phytopathol* 2011, **49**:317-343.

75. Rojas CM, Senthil-Kumar M, Tzin V, Mysore KS: **Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense**. *Front Plant Sci* 2014, **5**:17.

CHAPTER 3: DEVELOPMENT OF AN AUTOMATED RNA-SEQ DATA ANALYSIS PIPELINE

3.1 INTRODUCTION

The high-throughput Next Generation Sequencing (NGS) technologies have greatly revolutionized research in biology and have been increasingly used in life sciences in recent years over traditional technologies such as microarray and EST-based sequencing [1]. The million to billion short sequence reads produced by NGS platforms are widely used to study the genome, transcriptome and epigenome of organisms. Genome-wide transcriptome sequencing (RNA-Seq: sequencing of RNA in the form of cDNA in a biological sample) has been widely used as method of choice to study RNA regulation in a biological sample. The RNA-Seq is an advanced technology that overcomes the limitations imposed by previous technologies such as microarray where prior knowledge of the organism is necessary to study gene regulation [2, 3]. The RNA-Seq is a high resolution technique that provides a digital measure of gene expression and it allows studying allele-specific expression, isoform level gene regulation and transcript structure, which were not possible with previous technologies [1, 2, 4, 5]. In addition, RNA-Seq provides an opportunity to study alternative spliced sites, and identify novel transcripts, non-coding RNAs, fusion transcripts and single nucleotide polymorphisms [6].

The datasets produced from NGS platforms such as Illumina for RNA-Seq are massive and complex with multiple biological samples comprising a million to a billion sequence reads (25 to 300 bp), which corresponds to hundreds of gigabytes of data. The analysis of RNA-Seq data involves various steps (Figure 3.1) and intensive computational processing, which further complicates the tasks of handling, retrieving and scientific and/or biological interpretation of the analyzed data. A typical workflow of RNA-Seq data analysis pipeline is depicted in Figure 3.1. The analysis of RNA-Seq data is not straightforward and requires skilled bioinformaticists.

Typically, in RNA-Seq data analysis, the sequence reads generated from NGS platforms are filtered to get high quality sequence reads and subsequently mapped to reference genome or transcriptome of the organism. The task of mapping involves identification of the locations on the genome where sequence reads are identical with genomic sequences. The high quality filtered reads and mapped sequence data are then used for their assembly into transcripts, and differential gene expression analysis and variant discovery.

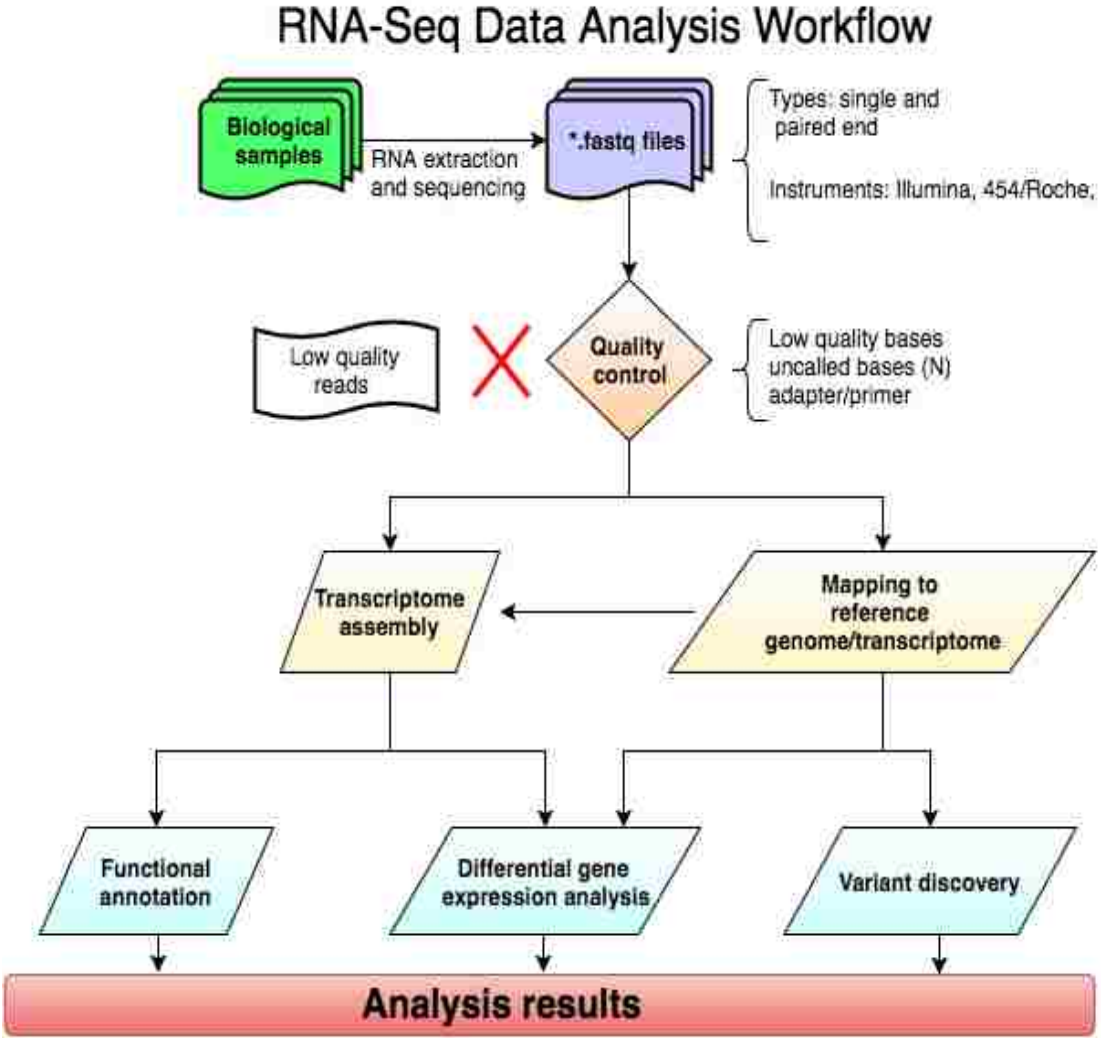

RNA-Seq Data Analysis Workflow
Biological samples
RNA extraction and sequencing
*.fastq files
Types: single and paired end
Instruments: Illumina, 454/Roche,
Low quality reads
Quality control
Low quality bases
uncalled bases (N)
adapter/primer
Transcriptome assembly
Mapping to reference genome/transcriptome
Functional annotation
Differential gene expression analysis
Variant discovery
Analysis results

Figure 3.1 Typical overview of workflow of Standalone RNA-Seq analysis pipeline (SRAP)

As the sequencing output from NGS platforms and biological samples under study are increasing continuously [7, 8], it is highly necessary to develop automated computational tools that can analyze massive amounts of RNA-Seq data with high accuracy, speed, flexibility and minimum manual intervention. Several tools are available to analyze RNA-Seq data [9-14], but those tools have limited capability and focus on a single point of analysis, such as assembly, splice sites, quantification or variant discovery. Further, the implementation of an automated software pipeline, which can process different steps in RNA-Seq workflow, is more difficult than processing single steps each time because of the parameters set up for each step, mathematical and statistical assumptions, and the intermediate files generated, which generally have different formats. Manual processing of each step of RNA-Seq is time-consuming, and requires additional effort and computational skill for processing output data from the previous step to the next step of downstream analysis. The various steps involved in the RNA-Seq data analysis are dependent on each other and therefore thorough knowledge is required for processing and analyzing massive RNA-Seq datasets. At present, no automated RNA-Seq data analysis pipeline is available that covers all the steps in RNA-Seq data analysis and provides the flexibility in the analysis parameters and wide range of tools. To overcome these limitations in RNA-Seq data analysis, an automated RNA-Seq data analysis pipeline was developed that can analyze different modules as a comprehensive automated flow or individual modules at a time with the parallel computing approach. The present pipeline integrates in-house developed algorithms along with open-source tools to provide users with broader option and flexibility to perform comprehensive RNA-Seq data analysis. The automated pipeline was tested on single and paired end sequence reads obtained from Illumina NGS platforms. The analysis pipeline produces the statistical summary and the visualization of the output dataset for each module. The proposed software SRAP (Standalone RNA-Seq Analysis Pipeline) is a comprehensive RNA-Seq data analysis pipeline and it allows the life science researchers with minimal computational expertise to perform daunting RNA-Seq data analysis task in a single platform.

3.2 IMPLEMENTATION

SRAP is a standalone software pipeline developed combining in-house coded scripts with open source bioinformatics tools using Python, matplotlib and Bash to analyze massive data from RNA-Seq and other NGS applications in an efficient manner. The software pipeline is user-friendly without requiring extensive computational expertise. SRAP is implemented with the parallel programming approach to effectively analyze huge datasets generated from RNA-Seq experiments with multiple numbers of samples. SRAP can run on Linux/Unix based personal computers, workstations and high performance computers (HPC). It can also run on Windows based system using Virtual Box software.

The python packages, such as numpy, pysam, multiprocessing, matplotlib, itertools, datetime, math, shutil, subprocess, termcolor, glob, gzip and collections, need to be pre-installed on the given computational environment to run SRAP. If these packages are not installed, the installation module of SRAP will prompt and guide the users for installing all necessary packages. SRAP may still work in the absence of the required packages, but the performance will be slow, limited, and may result in errors. For running SRAP there is no minimum requirement of the physical memory, and it will run efficiently on all modest computers with memory ideally ≥2GB, depending on the size of datasets. The README file associated with SRAP provides complete details about dependencies, other third party tools and different modes to run the software pipeline.

SRAP comprises of different modules required for the analysis of RNA-Seq data and data from other NGS applications, including filtering of raw reads, mapping the reads to reference genome/transcriptome sequences, assembly of cleaned sequences to form cotings, differential gene expression (DGE) analysis, variant discovery (single nucleotide polymorphism – SNP, insertion/deletion – indel) and other common NGS utilities. A complete workflow of SRAP to run as a batch and individual mode is shown in Figure 3.2.

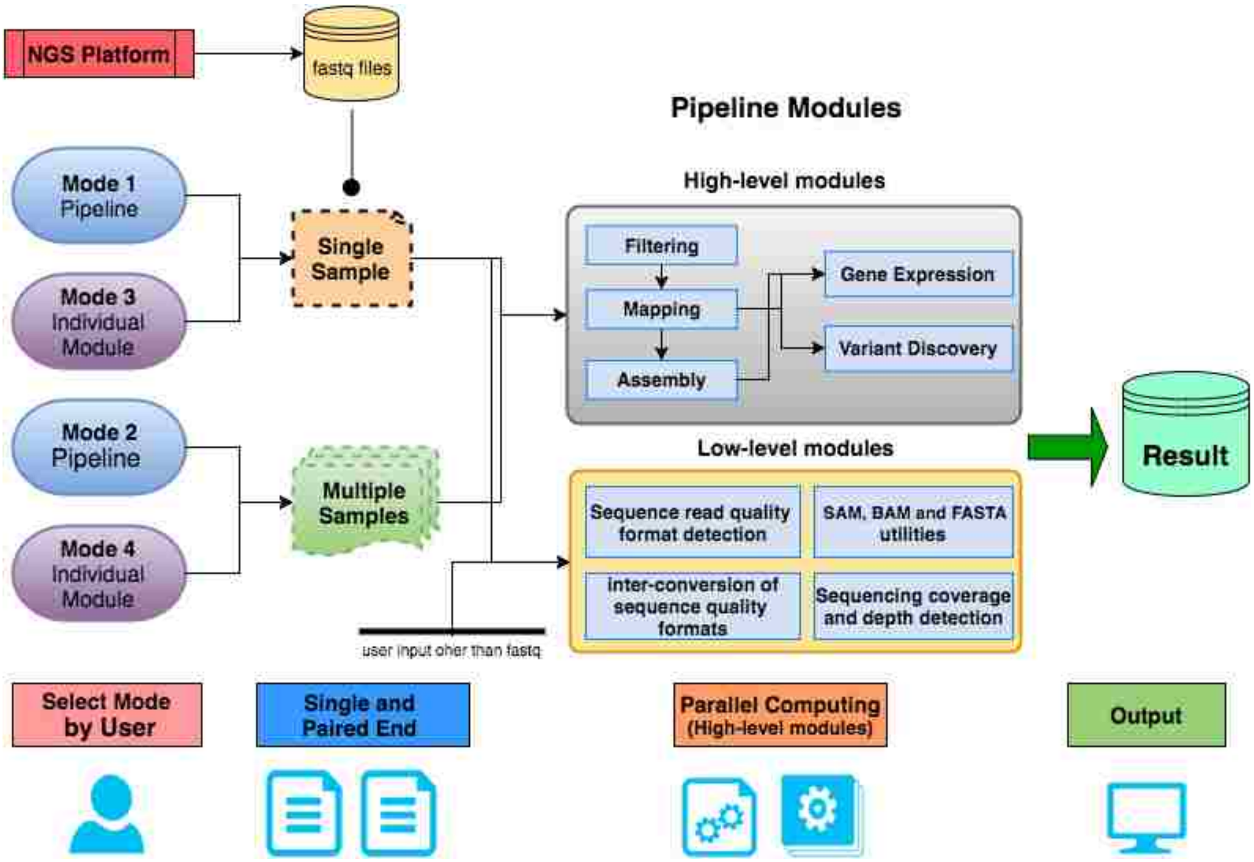


Figure 3.2 A schematic representation of the workflow of standalone RNA-Seq analysis pipeline (SRAP)

At this point, SRAP does not offer web interface, but it is under development. This pipeline implemented with shared parallel computing and distributed computing will be available soon. SRAP supports single (short reads from one end of RNA fragments) as well as paired end (short reads from both ends of RNA fragments) data. SRAP is unique in its ability to analyze the data with different modes (Figure 3.2). Each mode has unique path, which performs specific task in a parallel fashion. SRAP covers a wide range of applications for RNA-Seq data analysis and their parameters can be adjusted according to the user requirement.

3.3 RESULTS AND DISCUSSION

SRAP is a comprehensive automated standalone software pipeline designed for RNA-Seq data analysis, which comprises of four different modes to run the pipeline (Figure 3.1) for batch and individual NGS-produced single and paired end sequence reads files. SRAP is comprised of high level major modules for RNA-Seq data analysis, including sequence quality filtering, reference sequence mapping, sequence transcriptome assembly (Phase 1), differential gene expression analysis (Phase 2) and variant (SNP/Indel) discovery (Phase 3). Along with these five major modules, SRAP also comprises of low-level modules for different NGS applications, such as quality format detection, inter-conversion of quality formats, BAM and FASTA file utilities, and sequence coverage detection. The high-level modules require robust computation resources, which use high memory and employs multiple processors to perform data-intensive tasks, and lacks in low-level modules.

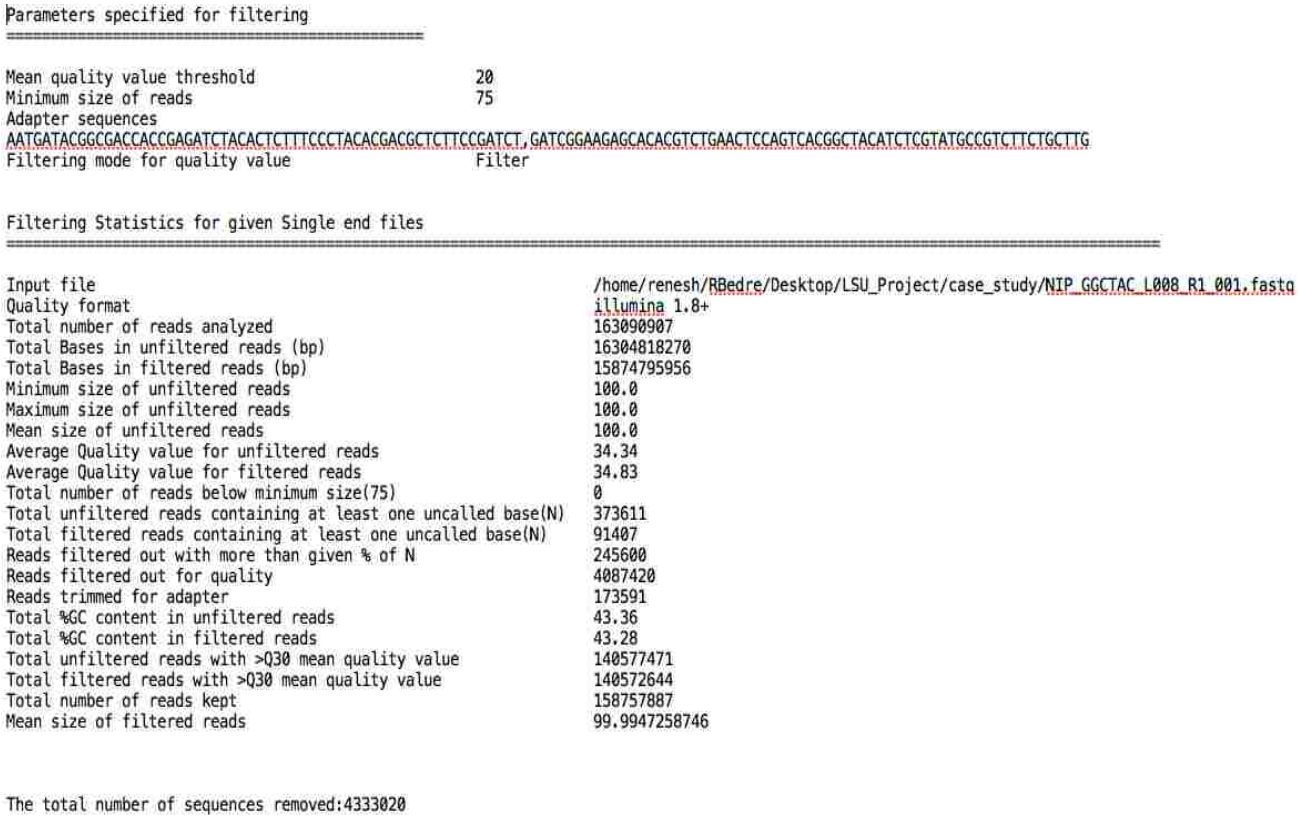
SRAP can be executed as an automated pipeline through different phases for complete analysis or through individual modules to carry out a specific task (Table 3.1). The automated pipeline requires a configuration file where parameters for each module are mentioned and can be customized based on the user’s requirements. The configuration file is optional for an individual module (Phase 4) where parameters can be customized on the command line. The default parameters, which are used in the pipeline, are well optimized and suited for most of the analysis tasks. The configuration file can be constructed and edited by the user to modify, add and/or remove analysis modules since the input and output from each module is compatible with the next module. SRAP analyzes the RNA-Seq data through different phases (Table 3.1) based on the number of input (sample) files and data analysis mode (Figure 3.1). The software pipeline supports data in various formats, such as the FASTQ format generated by most sequencing platforms, compressed FASTQ file format (gz), FASTA format and aligned SAM/BAM format [15, 16]. Each module in the pipeline produces an output report file with the summary of the data and visualization.

Table 3.1 Various phases of SRAP for performing different RNA-Seq data analysis tasks

Phases	Tasks	Configuration file	Description
Phase 1	Filter, map and assemble	fil_map_assembly.conf	Filtering of sequence reads, mapping to reference and assembling of sequence reads to construct contigs
Phase 2	Analyze gene expression	gene_exp.conf	Mapping of the sequence reads to reference transcriptome, differential gene expression
Phase 3	Discover variant	Optional	Identifying SNPs and Indels
Phase 4	Individual tasks	Optional	Performing individual module analysis

In filter, map and assemble phase, the single and/or multiple sample sequence reads files are analyzed in filtering, sequence assembly and mapping modules. In the filtering step, the developed standalone filtering module effectively checks the sequence reads for various quality parameters, including adapter/primer contamination, low quality bases based on Phred score (<20) and content of uncalled bases (N). This module filters out or trims low quality sequence reads and keep the high quality sequence reads, which are utilized by different modules during the entire RNA-Seq data analysis steps. The sequence filtering was performed on the NIP (non-inoculated pericarp) sample from cotton RNA-Seq data [17] and the results are shown in Figure 3.3 and 3.4. The filtered high quality sequence data, which is indicated by green line (Figure

3.4A), has the Phred quality score more than 20. In the mapping module, the sequence reads are mapped to the reference genome or transcriptome to know the origin of sequenced reads on the genome. The in house-developed python script along with open-source tools such as Bowtie2 [9], TopHat2 [10] and BWA [12] was deployed in the pipeline to cover a broader range of sequencing analysis options for users. The reference genome or transcriptome sequence must be provided by the user for mapping and/or assembly modules. The reference species sequences can be downloaded from the respective species sequence database. For example, the Rice genome sequence and annotation can be downloaded from Rice Genome Annotation Project database [18] or phytozome (www.phytozome.net). The sequence reads mapping data obtained by aligning the high quality RNA-Seq reads from a NIP sample to reference cotton *G. raimondii* genome is shown in the Table 3.2. The sequence reads obtained from NGS platforms do not represent full length genes, therefore construction of full length genes by assembly of these sequence reads is important to study transcribed genes and their structure. The high quality mapped sequence reads obtained from filtering and mapping modules respectively are retrieved for transcript construction. The in house-developed python script along with open-source tools such as Trinity [11], Cufflinks [13] and StringTie [14] was deployed in the pipeline to cover a broader range of sequencing analysis options for users. Trinity [11] was integrated into the pipeline for genome-guided and *de-novo* (without reference genome) assembly of transcripts to form full/partial length genes along with their transcript isoforms. With the *de novo* assembly method, novel transcripts can be determined, but it is less accurate as compared to genome-guided assembly. The other genome guided and *de novo* assembly tools such as Cufflinks [13] and StringTie [14] were also included in SRAP to provide flexibility in the analysis to the users. The StringTie assembler combines both genome-guided and *de novo* assembly approaches and identifies 36-60% transcripts more accurately than cufflinks [14]. In contrast to Cufflinks where identification of transcripts and their quantification are performed in different steps, the StringTie assembles and quantifies the expression levels of transcripts simultaneously [14]. This phase also provides the opportunity for quantifying the mRNA levels of the expressed genes using the parameters provided in the configuration file (Table 3.1). Along with high-throughput Cufflinks and StringTie tools for transcript assembly and quantification, SRAP is also integrated with the htseq-count [19] for quantifying the mapped sequence reads in absolute values (raw counts) instead of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) count produced by Cufflinks and StringTie [13, 14].



```
Parameters specified for filtering
===============================================

Mean quality value threshold                    20
Minimum size of reads                           75
Adapter sequences
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT,GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG
Filtering mode for quality value                Filter

Filtering Statistics for given Single end files
=====================================================================================================================

Input file                                                   /home/renesh/RBedre/Desktop/LSU_Project/case_study/NIP_GGCTAC_L008_R1_001.fastq
Quality format                                               illumina 1.8+
Total number of reads analyzed                               163090907
Total Bases in unfiltered reads (bp)                         16304818270
Total Bases in filtered reads (bp)                           15874795956
Minimum size of unfiltered reads                             100.0
Maximum size of unfiltered reads                             100.0
Mean size of unfiltered reads                                100.0
Average Quality value for unfiltered reads                   34.34
Average Quality value for filtered reads                     34.83
Total number of reads below minimum size(75)                 0
Total unfiltered reads containing at least one uncalled base(N)  373611
Total filtered reads containing at least one uncalled base(N)    91407
Reads filtered out with more than given % of N               245600
Reads filtered out for quality                               4087420
Reads trimmed for adapter                                    173591
Total %GC content in unfiltered reads                        43.36
Total %GC content in filtered reads                          43.28
Total unfiltered reads with >Q30 mean quality value          140577471
Total filtered reads with >Q30 mean quality value            140572644
Total number of reads kept                                   158757887
Mean size of filtered reads                                  99.9947258746


The total number of sequences removed:4333020
```

Figure 3.3 A screenshot of the filtering statistics output for the non-inoculated pericarp (NIP) library of cotton RNA-Seq data (Ref. 17)

Table 3.2 Sequence reads alignment statistics. The sequence reads from RNA-Seq dataset of non-inoculated pericarp (NIP) library of cotton (*Gossypium hirsutum*) was aligned to the *G. raimondii* reference genome

Parameters	**NIP library**
Total sequence reads	158,757,887
Sequence reads aligned	102,190,094 (64.36%)
Sequence reads aligned to multiple locations[a]	6,312,694 (3.97%)
Sequence reads aligned to a single location[b]	95,877,400 (60.39%)

a: Same sequence reads from the data mapped to multiple locations on the genome sequence
b: Same sequence reads from the data mapped to a single location on the genome sequence

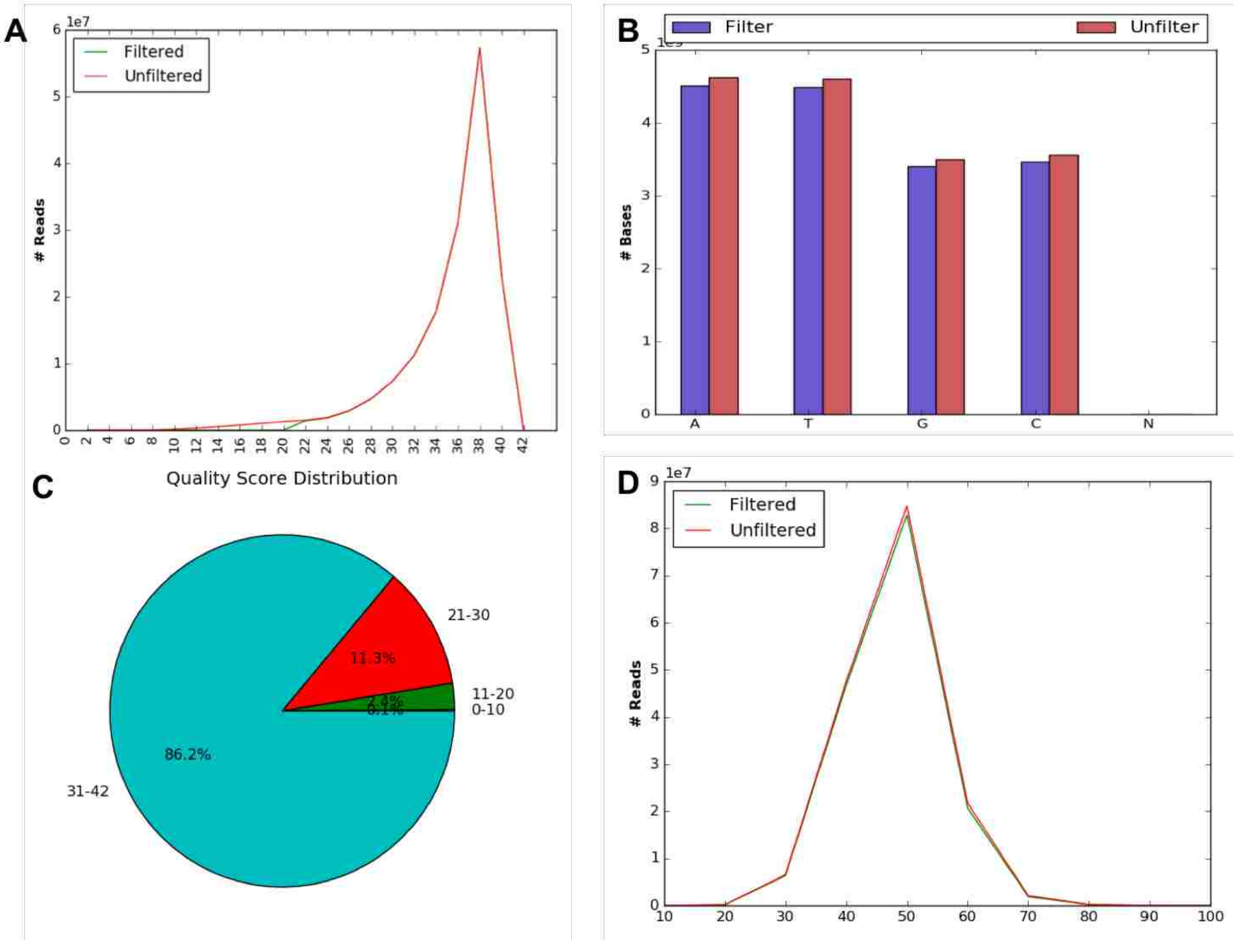


Figure 3.4 Filtering analysis of the RNA-Seq reads from cotton non-inoculated pericarp NIP library (Ref. 17). **A)** The comparison of filtered and unfiltered reads (raw sequence reads). The filtered sequence reads (green line) has the Phred quality score >20, whereas the unfiltered

sequence reads have the Phred quality score <20. The x-axis and y-axis represent Phred quality score and sequence read count, respectively. **B)** The distribution of nucleotide bases (A, T, G, and C) in filtered and unfiltered sequence reads. A large number of low quality bases has been removed from the unfiltered sequence reads. C) The Phred quality score distribution of unfiltered sequence reads. D) The GC content distribution of filtered and unfiltered reads. The x-axis and y-axis represent % GC content and sequence read count, respectively.

In Phase 2, along with filtering, mapping and assembly, differential gene expression analysis is performed to measure the differences in the mRNA abundance of the genes between the control and an experimental condition (untreated vs treated, unstressed vs stressed, etc.) based on the counts obtained from the transcript quantification module. The Figure 3.5 represents the volcano plot obtained from SRAP, which compares the expression of the genes between control (NIP-noninoculated pericarp) and experimental (TP-pericarp inoculated with toxigenic strain of *A. flavus*) tissues of cotton [17]. In the variant discovery phase (Phase 3), after the completion of the filtering and mapping modules, the SNPs and Indels in samples are identified in comparison with the reference sequences, using default parameters (Figure 3.6). The users can customize the parameters as per requirement such as reference sequences, number of processors, and algorithm by editing in the configuration file for each phase of analysis. The configuration file also allows the users to change the tools of their choice from the available options. In the gene expression analysis with RNA-Seq experiments, the accuracy of the differential gene expression depends on the resolution of expression at gene and isoform level from the counts obtained from mapping data and sources of variability across the replicates. To address the issues that complicate the transcript level expression and to reduce false positive rates, Cuffdiff 2 [20] methodology was adopted for performing differential gene and transcript expression. Though replicates are necessary to reduce the rate of false positive detection in the differentially expressed transcripts, Cuffdiff 2 has a high precision in detecting the differentially expressed transcripts regardless of the number of replicates [20]. In the case of absence of replicates in the experiment, Cuffdiff 2 counts biological samples in the control and experimental condition as replicates of each other and measures the variance [20].

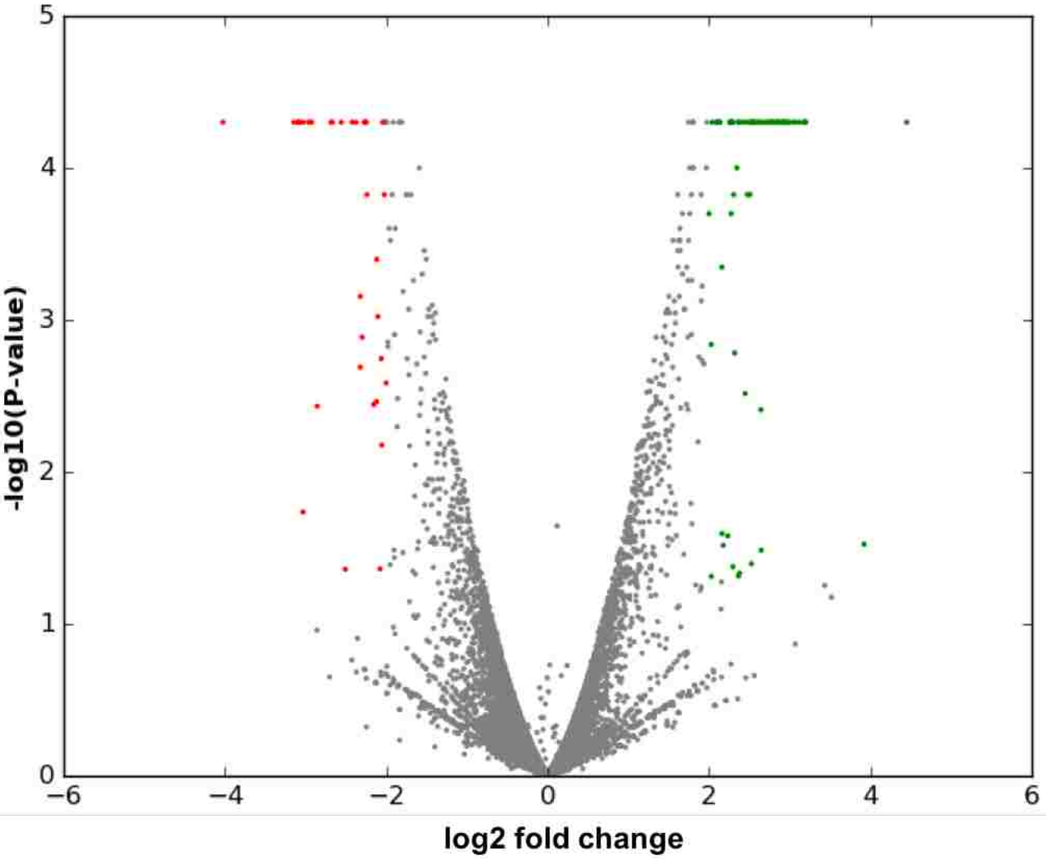
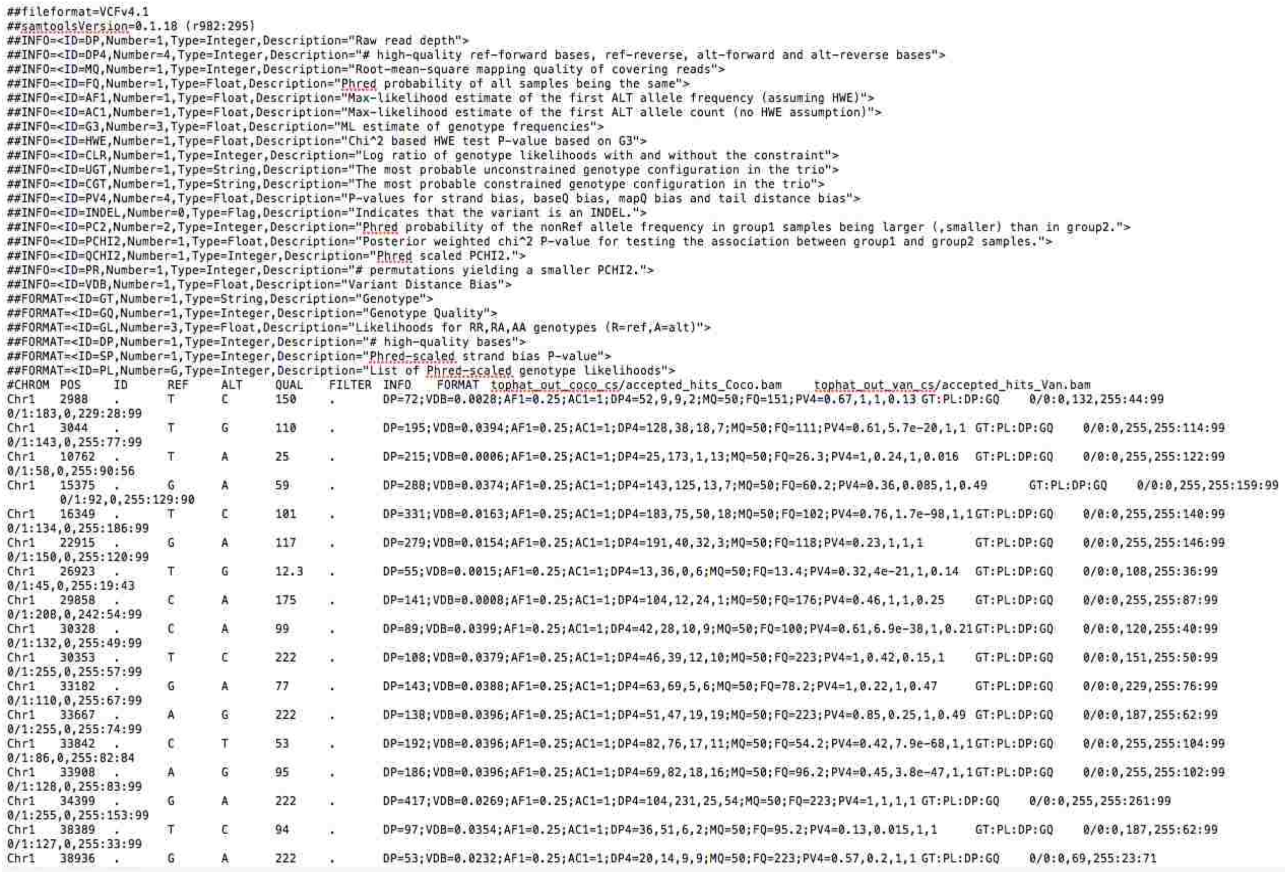


Figure 3.5 Differential gene expression analysis in the cotton experimental RNA-Seq dataset (pericarp inoculated with toxigenic *Aspergillus flavus*, TP) in comparison to the control (non-inoculated pericarp, NIP). Green and red dots represent the up-regulated genes (log2 fold change ≥ 2, P<0.05) and down-regulated genes (log2 fold change ≤ -2, P<0.05), respectively

Along with the different phases of SRAP, which runs high-level modules, the software pipeline also supports common utilities that are essential in the NGS data analysis. The common NGS utilities include format conversion (FASTQ to FASTA, SAM to BAM, BAM to SAM, TAB to FASTA and FASTA to TAB), detection of FASTQ quality variants, sequence reads quality variants interconversion, finding the length of sequence reads, sequence coverage or depth analysis and merging bam files. These common NGS utilities are low-level modules and do not require intense computation unlike the high-level modules. The low-level modules are executed on command lines without a configuration file.



```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.18 (r982:295)
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype likelihoods with and without the constraint">
##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstrained genotype configuration in the trio">
##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrained genotype configuration in the trio">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than in group2.">
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
#CHROM POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  tophat_out_coco_cs/accepted_hits_Coco.bam     tophat_out_van_cs/accepted_hits_Van.bam
Chr1    2988    .       T       C       150     .       DP=72;VDB=0.0028;AF1=0.25;AC1=1;DP4=52,9,9,2;MQ=50;FQ=151;PV4=0.67,1,1,0.13 GT:PL:DP:GQ    0/0:0,132,255:44:99
0/1:183,0,229:28:99
Chr1    3044    .       T       G       110     .       DP=195;VDB=0.0394;AF1=0.25;AC1=1;DP4=128,38,18,7;MQ=50;FQ=111;PV4=0.61,5.7e-20,1,1  GT:PL:DP:GQ    0/0:0,255,255:114:99
0/1:143,0,255:77:99
Chr1    10762   .       T       A       25      .       DP=215;VDB=0.0006;AF1=0.25;AC1=1;DP4=25,173,1,13;MQ=50;FQ=26.3;PV4=1,0.24,1,0.016  GT:PL:DP:GQ    0/0:0,255,255:122:99
0/1:58,0,255:90:56
Chr1    15375   .       G       A       59      .       DP=288;VDB=0.0374;AF1=0.25;AC1=1;DP4=143,125,13,7;MQ=50;FQ=60.2;PV4=0.36,0.085,1,0.49      GT:PL:DP:GQ    0/0:0,255,255:159:99
        0/1:92,0,255:129:90
Chr1    16349   .       T       C       101     .       DP=331;VDB=0.0163;AF1=0.25;AC1=1;DP4=183,75,50,18;MQ=50;FQ=102;PV4=0.76,1.7e-98,1,1GT:PL:DP:GQ    0/0:0,255,255:140:99
0/1:134,0,255:186:99
Chr1    22915   .       G       A       117     .       DP=279;VDB=0.0154;AF1=0.25;AC1=1;DP4=191,40,32,3;MQ=50;FQ=118;PV4=0.23,1,1,1        GT:PL:DP:GQ    0/0:0,255,255:146:99
0/1:150,0,255:120:99
Chr1    26923   .       T       G       12.3    .       DP=55;VDB=0.0015;AF1=0.25;AC1=1;DP4=13,36,0,6;MQ=50;FQ=13.4;PV4=0.32,4e-21,1,0.14  GT:PL:DP:GQ    0/0:0,108,255:36:99
0/1:45,0,255:19:43
Chr1    29858   .       C       A       175     .       DP=141;VDB=0.0008;AF1=0.25;AC1=1;DP4=104,12,24,1;MQ=50;FQ=176;PV4=0.46,1,1,0.25     GT:PL:DP:GQ    0/0:0,255,255:87:99
0/1:208,0,242:54:99
Chr1    30328   .       C       A       99      .       DP=89;VDB=0.0399;AF1=0.25;AC1=1;DP4=42,28,10,9;MQ=50;FQ=100;PV4=0.61,6.9e-38,1,0.21GT:PL:DP:GQ    0/0:0,120,255:40:99
0/1:132,0,255:49:99
Chr1    30353   .       T       C       222     .       DP=108;VDB=0.0379;AF1=0.25;AC1=1;DP4=46,39,12,10;MQ=50;FQ=223;PV4=1,0.42,0.15,1     GT:PL:DP:GQ    0/0:0,151,255:50:99
0/1:255,0,255:57:99
Chr1    33182   .       G       A       77      .       DP=143;VDB=0.0388;AF1=0.25;AC1=1;DP4=63,69,5,6;MQ=50;FQ=78.2;PV4=1,0.22,1,0.47      GT:PL:DP:GQ    0/0:0,229,255:76:99
0/1:110,0,255:67:99
Chr1    33667   .       A       G       222     .       DP=138;VDB=0.0396;AF1=0.25;AC1=1;DP4=51,47,19,19;MQ=50;FQ=223;PV4=0.85,0.25,1,0.49 GT:PL:DP:GQ    0/0:0,187,255:62:99
0/1:255,0,255:74:99
Chr1    33842   .       C       T       53      .       DP=192;VDB=0.0396;AF1=0.25;AC1=1;DP4=82,76,17,11;MQ=50;FQ=54.2;PV4=0.42,7.9e-68,1,1GT:PL:DP:GQ    0/0:0,255,255:104:99
0/1:86,0,255:82:84
Chr1    33908   .       A       G       95      .       DP=186;VDB=0.0396;AF1=0.25;AC1=1;DP4=69,82,18,16;MQ=50;FQ=96.2;PV4=0.45,3.8e-47,1,1GT:PL:DP:GQ    0/0:0,255,255:102:99
0/1:128,0,255:83:99
Chr1    34399   .       G       A       222     .       DP=417;VDB=0.0269;AF1=0.25;AC1=1;DP4=104,231,25,54;MQ=50;FQ=223;PV4=1,1,1,1 GT:PL:DP:GQ    0/0:0,255,255:261:99
0/1:255,0,255:153:99
Chr1    38389   .       T       C       94      .       DP=97;VDB=0.0354;AF1=0.25;AC1=1;DP4=36,51,6,2;MQ=50;FQ=95.2;PV4=0.13,0.015,1,1      GT:PL:DP:GQ    0/0:0,187,255:62:99
0/1:127,0,255:33:99
Chr1    38936   .       G       A       222     .       DP=53;VDB=0.0232;AF1=0.25;AC1=1;DP4=20,14,9,9;MQ=50;FQ=223;PV4=0.57,0.2,1,1 GT:PL:DP:GQ    0/0:0,69,255:23:71
```

Figure 3.6 A screenshot of the variant call format (VCF) file depicting the SNPs from Rice RNA-Seq dataset (Unpublished)

A user can select the automated pipeline per se or its individual modules to execute the relevant analysis. While the execution of the software pipeline and/or individual modules is ongoing, the users can monitor and track the progress of the analysis with verbose output on the screen. The output of the analysis including the graphical and statistical summary report for all the modules will be in the same directory from where SRAP is executed.

3.4 Availability and Requirements

SRAP software pipeline is a standalone application and can be downloaded from https://dl.dropboxusercontent.com/u/57407558/RSP.zip. The requirements, installation and usage of SRAP are described in the README file in the base directory of SRAP. SRAP is in zip compressed format and need to be extracted before installation. The installation module of SRAP will guide the users for pre-requisites and installation.

3.5 REFRENCES CITED

1. Goncalves A, Tikhonov A, Brazma A, Kapushesky M: **A pipeline for RNA-seq data processing and quality assessment**. *Bioinformatics* 2011, **27**(6):867-869.

2. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results**. *Genome Biol* 2010, **11**(12).

3. Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research**. *Mol Ecol* 2012, **21**(8):1794-1805.

4. Li J, Hu J, Newman M, Liu KJ, Ge HY: **RNA-Seq Analysis Pipeline Based on Oshell Environment**. *Ieee Acm T Comput Bi* 2014, **11**(5):973-978.

5. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population**. *Nature* 2010, **464**(7289):773-U151.

6. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, Nie J, Tang X, Baheti S, Doughty JB *et al*: **MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing**. *BMC Bioinformatics* 2014, **15**:224.

7. Poland JA, Rife TW: **Genotyping-by-Sequencing for Plant Breeding and Genetics**. *Plant Genome-Us* 2012, **5**(3):92-102.

8. Luo J, Wu M, Gopukumar D, Zhao Y: **Big Data Application in Biomedical Research and Health Care: A Literature Review**. *Biomed Inform Insights* 2016, **8**:1-10.

9. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-U354.

10. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**(4).

11. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc* 2013, **8**(8):1494-1512.

12. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.

13. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-578.

14. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nat Biotechnol* 2015, **33**(3):290-+.

15. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Res* 2010, **38**(6):1767-1771.

16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

17. Bedre R, Rajasekaran K, Mangu VR, Timm LES, Bhatnagar D, Baisakh N: **Genome-Wide Transcriptome Analysis of Cotton (Gossypium hirsutum L.) Identifies Candidate Gene Signatures in Response to Aflatoxin Producing Fungus Aspergillus flavus**. *Plos One* 2015, **10**(9).

18. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu JZ, Zhou SG *et al*: **Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data**. *Rice* 2013, **6**.

19. Anders S, Pyl PT, Huber W: **HTSeq-a Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2015, **31**(2):166-169.

20. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nat Biotechnol* 2013, **31**(1):46-+.

CHAPTER 4: CONCLUSIONS AND FUTURE PERSPECTIVES

4.1 CONCLUSIONS

- Understanding the expression profile of genes, especially in response to the atoxigenic strain infection, could provide clues to the molecular mechanisms of resistance, in addition to the physical barriers, conferred by the atoxigenic strains against the toxigenic strain.
- Comparative analysis of the genes involved in specific gene ontology categories of the atoxigenic vis-à-vis the toxigenic strain infection will lead to the identification of promising candidates for genetic manipulation of cotton toward development of varieties resistant to *A. flavus*. For example, genes with transcriptional regulation involved in response to stress stimulus, involved in flavonoid biosynthesis and lipid biding in extracellular regions (Fig 6) could be considered promising candidates for further validation through functional characterization.
- The sequencing reads and the assembled transcripts that were developed and utilized in the present study will enrich the cotton genomic resources in public databases. The sequencing reads data is publicly available and can be downloaded from the NCBI SRA database (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA275482).
- The automated SRAP (Standalone RNA-Seq analysis pipeline) developed through this study will provide a powerful resource for the life scientists to analyze massive RNA-Seq data in performing a complete analysis tasks, including filtering, mapping, sequence assembly, gene expression analysis and variant discovery.
- The implementation of SRAP with parallel computing approach, its flexibility in the analysis, ability to handle multiple biological samples, ability to analyze any genome,

comprehensive output report with visualization from the modules and extensive statistical analysis make SRAP as a powerful tool for analysis of RNA-Seq data.

- The output data obtained from the RNA-Seq pipeline can be utilized by other bioinformatics tools for downstream analysis, such as gene ontology and biological pathway enrichment analysis of differentially expressed genes.

4.2 FUTURE PERSPECTIVES

- The comparative analysis of the cotton transcriptome with available corn and peanut transcriptome, induced under *A. flavus* infection, will provide a better understanding of the genetic and biochemical basis of *A. flavus*-cotton interaction and also identify conserved orthologous genes in cotton for their functional translation in conferring resistance to *A. flavus* through genetic manipulation of cotton.
- SRAP offers a unique platform for complete RNA-Seq data analysis. Other downstream applications, such as GO and pathway enrichment analysis for differentially expressed genes, would make the pipeline more attractive and competitive.
- As SRAP is implemented with shared parallel computing approach, the implementation of the pipeline with distributed computing approach will enhance by multifold the speed of analysis.
- The availability of the pipeline on web interface and as menu driven on Windows without command line will make it convenient for users especially those with no working knowledge of Linux OS.

APPENDIX I: DETAILS OF DIFFERENTIALLY EXPRESSED GENES UNDER INFECTION BY ATOXIGENIC AND TOXIGENIC STRAINS OF *ASPERGILLUS FLAVUS* IN SEED AND PERICARP TISSUES OF COTTON

Description:

Sheet 1, Nomenclature of genes and primer sequences used for expression analysis through RT-PCR
Sheet 2, All differentially expressed genes discussed in the manuscript and used for heatmap in Figure 2.2
Sheet 3, Genes from different classes used in the generation of heatmaps for Figure 2.3 and Figure 2.4

File:

http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0138025.s002

APPENDIX II: PERMISSION TO REPRINT PUBLISHED MANUSCRIPT (BEDRE ET AL., 2015) IN CHAPTER 2

Citation:
Bedre R, Rajasekaran K, Mangu VR, Timm LES, Bhatnagar D, Baisakh N. Genome-Wide Transcriptome Analysis of Cotton (*Gossypium hirsutum* L.) Identifies Candidate Gene Signatures in Response to Aflatoxin Producing Fungus *Aspergillus flavus*. Plos One. 2015;10(9).

Dear Renesh,

Thank you for your message. PLOS ONE publishes all of the content in the articles under an open access license called "CC-BY." This license allows you to download, reuse, reprint, modify, distribute, and/or copy articles or images in PLOS journals, so long as the original creators are credited (e.g., including the article's citation and/or the image credit). Additional permissions are not required. You can read about our open access license here: http://www.plos.org/about/open-access/.

There are many ways to access our content, including HTML, XML, and PDF versions of each article. Higher resolution versions of figures can be downloaded directly from the article.

Thank you for your interest in PLOS ONE and for your continued support of the Open Access model. Please do not hesitate to be in touch with any additional questions.

Kind Regards,

Jackie

Jackie Surplice
EO Staff
PLOS ONE

Case Number: 04623535

VITA

Renesh Hanumanrao Bedre was born to Hanumanrao and Sharda Bedre in 1988 at Aurangabad, Maharashtra, India. He finished his high school in 2005. He completed a bachelor degree in Pharmacy (B. Pharm) from Dr. Babasaheb Ambedkar Marathwada University in Aurangabad, Maharashtra, India in 2009. During his B.S., he developed interests toward bioinformatics. In 2009, he successfully qualified the Graduate Aptitude Test in Engineering (GATE) test conducted by the Govt. of India and secured the scholarship for master studies. On the basis of his GATE score, he secured the admission for M.Tech in Bioinformatics at the Indian Institute of Information Technology (IIIT), Allahabad, India, which is one of the top engineering institutes of India. At IIIT, he successfully completed several projects under the supervision of Dr. Pritish Varadwaj and earned the expertise in computational programming languages.

Towards completion of master study, he got an opportunity for Ph.D. at LSU to work under Dr. Niranjan Baisakh and joined the Baisakh Lab in August, 2011. During Ph.D., he has been involved in several bioinformatics projects, including transcriptomics and genomics data analysis of plants under biotic and abiotic stresses, and development of NGS data analysis tools with novel algorithms and automated pipelines for large-scale datasets in a high-throughput manner. During his Ph.D., he learned the use of bioinformatics tools to address biological questions, earned the expertise in NGS data analysis, high performance computing, bioinformatics software development, web development and database management. He has received a Travel Grant Award from the LSU Graduate School to attend the XXII Plant and Animal Genome Conference in 2014. He is also the recipient of Gerald O. Mott Meritorious Student Award from the Crop Science Society of America in 2015. He was a member of the

“Indian Student Association (ISA)” which is one of the big organizations at LSU. He has also served as a member of the Graduate Student Association (GSA) of the Department of School of Plant, Environmental and Soil Sciences at LSU. He is a Ph.D. candidate and will receive his Ph.D. degree in August, 2016.