COMPUTATIONAL IDENTIFICATION OF FUNCTIONAL MODULES AND HUB GENES INVOLVED IN PATHOGENICITY-ASSOCIATED OR DEFENSE RESPONSE ON *FUSARIUM VERTICILLIOIDES*–MAIZE INTERACTIONS

A Dissertation

by

MANSUCK KIM

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Byung-Jun Yoon
Co-Chairs of Committee,	Won-Bo Shim
Committee Members,	Edward R. Dougherty
	Arum Han
Head of Department,	Miroslav M. Begovic

May 2016

Major Subject: Electrical Engineering

ABSTRACT

Fusarium verticillioides is one of the key pathogens for stalk rot and ear rot on maize. While several genes associated with *F. verticillioides* pathogenicity and mycotoxin biosynthesis have been characterized, our knowledge of the cellular and genetic networks for these events is still very limited. Also, underlying molecular and cellular mechanisms associated with the maize defense response against the *F. verticillioides* pathogenicity are complex. Therefore, in order to better understand maize defense as well as *F. verticillioides* pathogenicity, an approach systematically investigating the host-pathogen interactions is needed. In this PhD study, a systematic network-based comparative analysis approach using large-scale *F. verticillioides*-maize RNA-seq data was applied to identify *F. verticillioides* pathogenicity-associated subnetwork modules and also key pathogenicity genes as well as maize subnetwork modules involved in the defense response. For each study, we constructed corresponding co-expression networks through partial correlation based on the given comparable conditions. For the first work, predicting *F. verticillioides* pathogenicity-associated subnetwork modules, we established a pipeline identifying the functional modules by a branch-out technique with probabilistic subnetwork activity inference. For identifying maize defense modules, we first collected candidate maize genes by comparing expression pattern of maize genes and that of the selected four *F. verticillioides* pathogenicity genes through cointegration, correlation, and expression level change. Then, we inferred potential subnetwork modules among the candidate genes by adopting the previously established pipeline. For identifying specific key *F. verticillioides* pathogenicity genes based on the predicted subnetwork modules, we analytically investigated on each gene in its predicted subnetwork module. In this investigation, we considered its influence on others, association to pathogenicity, and distinctive differentiation between the two conditions. Through our systematic investigation of the *F. verticillioides*–maize RNA-seq data, we identified pathogenicity-associated or defensive subnetwork modules, where the member genes were harmoniously coordinated and significantly differentially activated between the two different conditions. Also, we identified specific *F. verticillioides* pathogenicity genes playing a key role in the predicted pathogenicity-associated subnetwork modules.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere appreciation to my advisor Prof. Byung-Jun Yoon for the constant support of my Ph.D study for motivation and enormous knowledge. His guidance assisted me in all my research and writings. Also, I would like to appreciate my co-advisor Prof. Won-Bo Shim for the valuable support. I could not have imagined having a better advisor and co-advisor for my Ph.D study.

Besides my advisor and co-advisor, my sincere gratitude also goes to the other committee members: Prof. Edward R. Dougherty and Prof. Arum Han for their insightful comments and advices, but also for encouragement.

I thank my labmates in Genomic Signal Processing lab. for valuable and stimulating discussions. Also, I thank fellows in Dr. Shim's lab.

Last but not the least, I would like to thank my family: my wife and to my parents and sister for supporting me spiritually. Without their support, it would not be possible to do the Ph.D study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Maize, one of the most significant crops, is susceptible to a variety of pathogens and *Fusarium verticillioides* (teleomorph *Gibberella moniliformis*) is one of the pathogens triggering maize ear and stalk rots. *F. verticillioides* even produces fumonisins that is a group of toxic secondary metabolites potentially harmful to animals and humans [13, 42, 54]. Plant-microbe interactions that have caused economic loss worldwide as well as even harmfulness to animals and humans are difficult to be understood. Plants respond to a number of external stimuli using complicated mechanisms. In particular, unlike animal defense mechanism with adaptive immune system, plant defense response encodes specific group of genes to recognize and respond to certain microbial pathogens [12] [15]. Also, microbes targeting plants have coevolved to prevail over plant immunity. Therefore, characterizing maize defense mechanism against the *F. verticillioides* pathogenicity is essential to comprehend biological functions and processes based on *F. verticillioides*-maize interactions. Recently, high throughput technologies (*i.e.*, microarray technology or next generation sequencing) have been developed to help research on genetic networks involved in biological functions and processes by providing massive datasets. Therefore, innovative computational methods properly analyzing these datasets are needed to characterize underlying biological system. To investigate pathogenicity of microbes or plant–microbe interactions, a number of individual gene-based genetic approaches or few network-based approaches have had some success, however they have also showed their inherent shortcomings mainly due to the lack of fundamental understanding of complicated cellular functions or host-pathogen interaction. In this Ph.D study, we comprehensively investigate mechanism of subnetwork modules or key genes associated with the

F. verticillioides pathogenicity as well as maize defense response based on maize-*F. verticillioides* interactions.

Specifically, we performed a systematic network-based comparative analysis on two different RNA-seq datasets, maize inbred B73 inoculated with two *F. verticillioides* strains (*i.e.*, wild-type vs. *fsr1* mutant) as well as two different *F. verticillioides* phenotypes produced by gene knock-outs of two different genes. First, For a systematic analysis of the infection transcriptome from the dataset that maize inoculated with two strains of *F. verticillioides* (wild type vs. the mutant), we first predicted the co-expression network of the fungus. Subsequently, we identified functional subnetwork modules in the co-expression network consisting of interacting genes that display strongly coordinated behavior in the respective datasets. A probabilistic activity inference method was adopted to identify modules likely to be involved in the pathogenicity of *F. verticillioides*, and a computationally efficient branch-out technique was used to search for potential subnetwork modules. Second, to computationally identify potential maize subnetwork modules associated with defense against *F. verticillioides* using the RNA-seq reads from B73 maize inoculated with wild-type *F. verticillioides* and a loss-of-virulence mutant, we first analyzed the RNA-Seq data based on a cointegration-correlation-expression approach, where maize genes were jointly analyzed with known pathogenicity genes in *F. verticillioides*; (i) cointegration was used to analyze their expression trends over time, (ii) correlation was used to investigate the expression patterns across different replicates, and (iii) expression levels in all replicates were checked to focus on significantly expressed genes. We then searched the maize coexpression network to detect subnetwork modules that are differentially expressed with high significance when inoculated with two different fungal strains. Third, in order to identify potential pathogenicity genes using the both RNA-seq datasets, we first constructed their co-expression networks, and predicted potential subnetwork modules as described in our previous work. Subsequently, we analytically investigated whether each gene in its predicted subnetwork module satisfied several conditions; i) highly impactful in a probabilistic manner, ii) relatively differentially correlated between two strains (wild type vs mutant), iii) relatively more connected in the given module, iv) relatively highly expressed in wild type, v) orthologous to known pathogenic genes in other fungi, and vi) annotated to significant GO terms with other member genes. Through our systematic investigation of the RNA-seq data, we have identified potential subnetwork modules associated with the *F. verticillioides* pathogenicity, potential subnetwork modules involved in maize defense response against the *F. verticillioides* pathogenicity, and also key pathogenicity potential *F. verticillioides* pathogenicity genes. The potential subnetwork modules demonstrated not only harmonious coordination of the member genes, but also strong differentiation between the two phenotypes. Furthermore, the potential functional genes also showed significant influence on other member genes in its subnetwork module.

2. BACKGROUND

2.1 *Fusarium*

Fusarium is a fungal genus that was characterized as *Fusisporium* by Link in 1809. It is known as hyphomycetes, triggering diseases in diverse crops, and also various *Fusarium* species, producing secondary metabolites (*i.e.*, mycotoxins), can be harmful to plants, and even animals and humans [61]. The *Fusarium* mycotoxins can be either detrimental to plants or involved in diseases such as cancer for humans [41]. Since pathogenicity or mycotoxin of *Fusarium* species take place in every developing stages of host plants based on their specific profiles, identification of *Fusarium* activity is essential to prevent the host from potential toxicological risk [46] [11].

After the first identification by Link, Wollenweber and Reinking nearly described 1000 species and devided genera into 16 sections and 65 species in 1935. After that, many researchers characterized taxonomy of *Fusarium* showing different number of species [9]; Snyder and Hansen recognized the number of *Fusarium* species as nine in 1940, and Gerlach and Nirenberg recognized around 90 species through "The genus *Fusarium* – A pictorial Atlas" in 1982 [20], and also Leslie and Summerell recognized approximately 70 species based on morphological, biological and phylogenetic criteria though "The *Fusarium* laboratory manual" in 2006 [36].

In recent decades, typical *Fusarium* species such as *Fusarium graminearum*, *Fusarium oxysporum*, and *Fusarium verticillioides* have been receiving attention based on the following research environment; i) their classical genetic maps have been constructed (*i.e.*, genetic map of *Fusarium verticillioides* [66] [25], genetic map of *Fusarium graminearum* [26] [19]), ii) their reference genome sequences have been available through the Broad Institute (http://www.broadinstitute.org), iii) recently high throughput technologies (*i.e.*, microarray technology and next generation sequencing) producing gene expression data have been developing rapidly. Along with the environment, there is a pressing need for research issues on *Fusarium* species to be addressed due to the following potentials; i) regulation of principle metabolism of *Fusarium* species is different from that of other fungi, ii) atypical biosynthesis of their secondary metabolites has a large economic impact, iii) host-pathogen interactions based on *Fusarium* species are also very different from those of other typical fungi. Therefore, through research on *Fusarium* species that has become very intriguing, new insights into the biological system of *Fusarium* is expected.

2.2 *Fusarium verticillioides*

Fusarium verticillioides (telemorph, *Gibberella moniliformis*), a representative fungal pathogen, is typically responsible for stalk rots and ear rots in maize, thereby leading to economic losses in corn production. Also, the fungal pathogen even produces the fumonisin mycotoxins that are harmful to animal and human while colonializing maize ears and stalks [13] [55].

There are typical infection pathways that *F. verticillioides* can colonize host plants. The principle pathway for kernel infection is assumed to be occurred by airborne conidia landing on corn silks whose kernel is in developing [44]. For ear rot and stalk rot infections, it is also postulated to be taken place by insects such as European corn borer (ECB) or Ostrinia nubilalis damaging corns [43]. Typical symptoms of host plants by the *F. verticillioides* infections are severe such as rotting or wilting. However, it might be also very minor as nearly no signs of infection since *F. verticillioides* is an endophytic fungal species growing with an organized manner in host plants so that the plants can stay asymptomatic throughout the infection cycle [8] [2].

2.3 Pathogenicity of *F. verticillioides*

Some have identified *F. verticillioides* pathogenicity genes directly involved in maize disease by single gene-based genetic investigation. Shim *et al.* [54] discovered *FSR1*, critical for maize stalk rot virulence, and Myung *et al.* [45] discovered *FvVE1*, essential for maize seedling pathogenicity, and also Ridenour *et al.* [50] identified HAP complex genes (*FvHAP2*, *FvHAP3*, and *FvHAP5*), responsible for pathogenicity of maize stalks. Other intriguing pathogenicity-associated genes of *F. verticillioides* have been also identified by single gene-based approach. *FST1* and two *PP2A* subunits (*PPR1* and *PPR2*) were all characterized to be associated with fumonisin biosynthesis by Kim *et al.* [31] and Shin *et al.* [56], respectively, and also Ortiz *et al.* [47] demonstrated that MADA-box transcription factors (*MADS1* and *MADS2*) were involved in fumonisin B1 production. However, although individual gene-based approaches have showed some successful discoveries of the *F. verticillioides* pathogenicity, a limited number of genes have been identified and cellular functions and structures involved in the identified genes are still barely understood. Furthermore, the approaches have been predominantly based on biological knowledge of researchers (i.e., considering conserved functions). Therefore, for other fungi such as *F. graminearum* (FG) or *Magnaporthe grisea* (MGG), some [21, 38, 39] have attempted predicting fungal pathogenicity using protein-protein interaction (PPI) networks by considering gene interrelationship as a group since it is known that genes rely on each other when performing biological functions. However, although they challenged with other knowledge (i.e., orthologs in other fungi or known pathogenicity genes) based on the PPIs, their approaches had inherent shortcomings such as strong dependency on prior knowledge and no careful consideration on coordination of genes as a group.

2.4 Host-pathogen interactions

For maize-pathogen interactions, several methods have been shown using microarray data. For instance, Kelley *et al.* [27] identified maize genes associated with their resistance (or susceptibility) to *Aspergillus flavus*, pathogenic fungus, based on expression alterations. Campos-Bermudez *et al.* [10] identified maize genes and metabolites that displayed expression change after inoculation with *F. verticillioides*. However, their analysis were individual gene-based and the approach just focused on the expression differences. For host-pathogen interaction study other than plant-microbe, analysis based on host-pathogen gene interactions was performed through their correlation. Using Pearson's or Spearman rank correlation coefficients between host-pathogen, Shea *et al.* [53] identified gene pairs in a human-bacterial interactions with Group A *Streptococcus* (GAS). Reid *et al.* [48] identified interactions between mouse and *Plasmodium* as well as mosquito and *Plasmodium*. In addition, Asters *et al.* [1] constructed networks using Euclidean distance calculation based on gene-gene correlation. Although, these analysis identified correlated gene interactions and their networks based on their host-pathogen interactions, further improvements can be performed with systematically analyzing approach for cellular interactions or processes for underlying host-pathogen interactions.

2.5 Network/Pathway based approach

Recently, approaches collectively analyzing genes in the same pathway or subnetwork module have become popular for research on complex diseases since it has been known that genes collaborate each other for performing their biological functions or processes. Therefore, systematic approaches investigating their relationship between genes have received attention recently. [23, 3, 68, 52]. Several pathway-based approaches jointly analyzing expression levels of genes in a given pathway have shown prediction of the pathway activity level [6, 58, 35, 64, 62, 60, 29]. Also, some network-based approaches analyzing gene expression data and network data in an integrative manner to identify subnetwork modules have received increasing attention, as they can predict novel genetic modules associated with specific biological functions of interest [14, 59, 30]. For instance, Chuang *et al.* [14] identified potential subnetworks involved in breast cancer metastasis using gene expression data as well as PPI network. However, these approaches that required several types of information and also strongly relied on the given knowledge are not suitable for investigation on *F. verticillioides* whose discovery is still in early stage.

3. STUDY OF *F. VERTICILLIOIDES* PATHOGENICITY-ASSOCIATED NETWORK MODULES *

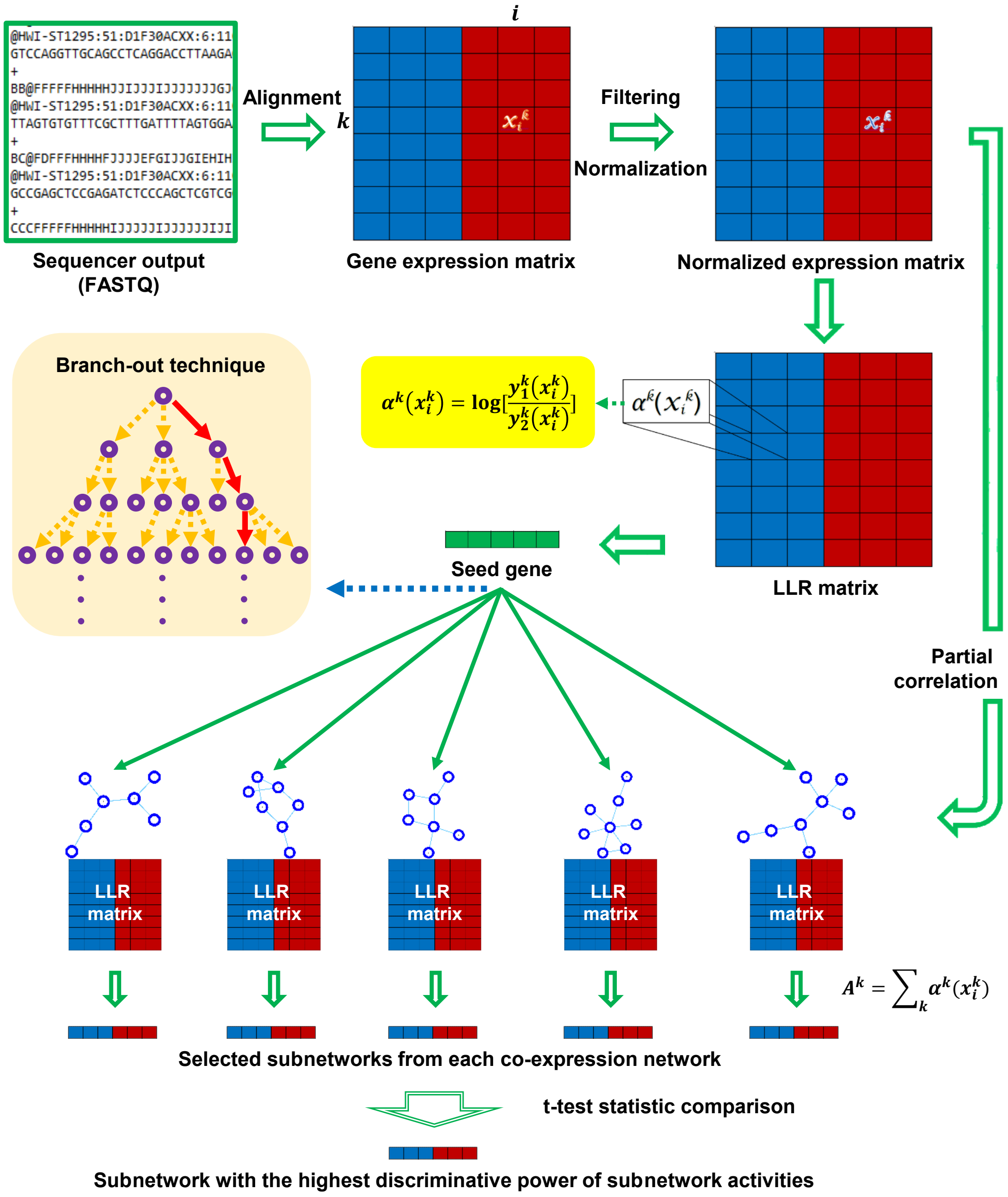
The entire procedure for the network-based RNA-Seq data analysis of *F. verticillioides* to identify potential pathogenic subnetwork modules is illustrated in Figure 3.1. First, for the gene expression data, both filtering for quality control and normalization were performed for the downstream analysis. Second, the co-expression network of *F. verticillioides* through partial correlation at five different levels were predicted. Next, based on the co-expression networks, we searched subnetwork modules from seed genes, significantly differentially expressed genes, by extending them through neighboring genes that strongly enhance probabilistic differential activity between wild type and the mutant. Finally, we assessed the subnetwork modules based on the strength of association and the pathogenicity of the fungi.

3.1 Methods

3.1.1 Preprocessing

Maize inbred B73 was inoculated with *F. verticillioides* wild type and fsr1 mutant as previously described [54]. To capture dynamic changes in gene expression, tissues were harvested from three distinct phases of stalk pathogenesis: establishment of fungal infection (3 days post inoculation [dpi]), colonization and movement in vascular bundle (6 dpi) and host destruction and collapse (9 dpi). As a reference, uninoculated maize stalk tissue as well as WT fungus grown in synthetic medium was used to

* Part of this chapter is reprinted with permission from "Mansuck Kim, Huan Zhang, Charles Woloshuk, Won-Bo Shim, and Byung-Jun Yoon. Computational prediction of pathogenic network modules in *Fusarium verticillioides. IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Web, 2015" [33] © [2015] IEEE.



```
@HWI-ST1295:51:D1F30ACXX:6:11
GTCCAGGTTGCAGCCTCAGGACCTTAAGA
+
BB@FFFFFHHHHHJJIJJJIJJJJJJJGJ
@HWI-ST1295:51:D1F30ACXX:6:11
TTAGTGTGTTTCGCTTTGATTTTAGTGGA
+
BC@FDFFFHHHHFJJJJEFGIJJGIEHIH
@HWI-ST1295:51:D1F30ACXX:6:11
GCCGAGCTCCGAGATCTCCCAGCTCGTCG
+
CCCFFFFFHHHHHIJJJJJIJJJJJJIJI
```

Sequencer output (FASTQ)
Alignment
k
i
x_i^k
Gene expression matrix
Filtering
Normalization
x_i^k
Normalized expression matrix
Branch-out technique
$\alpha^k(x_i^k) = \log[\frac{y_1^k(x_i^k)}{y_2^k(x_i^k)}]$
$\alpha^k(x_i^k)$
Seed gene
LLR matrix
Partial correlation
LLR matrix
LLR matrix
LLR matrix
LLR matrix
LLR matrix
$A^k = \sum_k \alpha^k(x_i^k)$
Selected subnetworks from each co-expression network
t-test statistic comparison
Subnetwork with the highest discriminative power of subnetwork activities

Figure 3.1: Overview of the proposed analysis [33] © [2015] IEEE.

provide a baseline of host and pathogen gene expression. For each sample subjected to sequencing, sectioning was performed on at least three stalk samples from each stage of infection, and isolated tissues were pooled for RNA extraction. Our RNA extraction protocols have been successfully used to extract high quality RNA from maize samples, e.g. kernels, stalks, or leaves, infected with fungal pathogens [54] [31]. RNA samples were processed following standard QA/QC procedures at Texas A&M AgriLife Genomics and Bioinformatics Services prior to sequencing.

After the samples were prepared, RNA sequencing was performed using Illumina HiSeq 2000. Sequence cluster identification, quality prefiltering, base calling, and uncertainty assessment were also performed in real time using Illumina's HCS 1.5.15.1 and RTA 1.13.48.0. In this RNA sequencing, library preparation and RNA isolation were performed by Illumina's simplied sample prep kits and small RNA sample preparation kit, respectively. Through this RNA sequencing process, we acquired FASTQ formatted output files for six independent sample libraries for each time point (*i.e.*, 3 dpi, 6 dpi, and 9 dpi), hence 36 libraries in total. The prepared reads of 36 libraries were aligned to the reference genome of *F. verticillioides* strain 7600 [40] downloaded from Broad Institute (http://www.broadinstitute.org) as well as to the maize B73 reference genome [51]. This alignment was performed to acquire read counts of all *F. verticillioides* genes as well as maize genes by Bowtie2 [34], known to be better for relatively longer reads and gapped alignment, along with a subsequent process using a NGS (next-generation sequencing) analysis tool called Samtools [37].

Table 3.1 shows general statistics of the RNA-seq datasets. The table demonstrates how the loss of *FSR1* influenced the *F. verticillioides* pathogenicity over time. After the alignment, we filtered out genes rarely expressed so that 8072 genes were remained. In this filtering, we filtered genes that were not expressed more than half of the replicates, therefore we further proceeded with significantly differentially expressed genes in the two strains (*i.e.*, wild type vs. the mutant). After the filtering, normalization by each gene length for the read counts was completed to acquire relative expression levels. Subsequent normalization was also performed based on

Table 3.1: General statistics of RNA-seq datasets

	Wild type			Mutant		
	3dpi	6dpi	9dpi	3dpi	6dpi	9dpi
Type of run	single					
Read length	100 (bp)					
Mean # of reads/sample	6,283,647			5,758,106		
Median depth of coverage	5.83	14.18	16.64	1.75	4.48	9.77

the expression level of the β-tubulin genes such as FVEG_05512 and FVEG_04081 since β-tubulin genes, considered as housekeeping genes, are often utilized for normalization of *F. verticillioides*. In this normalization, the mean expression level of FVEG_05512 and FVEG_04081 was applied as criterion.

3.1.2 Constructing co-expression networks

Based on the preprocessed gene expression data, we constructed the co-expression networks of *F. verticillioides* by computing partial correlation coefficients for each gene interaction. Hero et al. [22] demonstrated that partial correlation is more reliable and effective in predicting underlying biological networks since partial correlation only takes into account the strength of association between two genes without considering any impact from the other genes. In the co-expression networks, gene interactions were predicted based on five thresholds of significant partial correlation (*i.e.*, 0.90, 0.91, 0.92, 0.93, and 0.94), thereby letting the interactions to be more reliable and less dependent on a certain threshold level. Consequently, we prepared five co-expression networks of *F. verticillioides*, where the co-expression network at the threshold as 0.90 is the largest and includes the others. The number of genes and interactions of the five co-expression networks are shown in Table 3.2.

Table 3.2: Co-expression networks of wild type and mutant *F. verticillioides*.

Threshold (cut-off partial correlation)	Wild type		Mutant	
	# of genes	# of interactions	# of genes	# of interactions
0.94	5,649	133,214	4,699	124,271
0.93	6,313	203,897	5,446	193,170
0.92	6,856	291,378	6,044	279,193
0.91	7,232	394,807	6,516	381,568
0.90	7,495	514,081	6,891	499,469

3.1.3 Identifying potential pathogenic subnetwork modules

In order to identify potential subnetwork modules, we first searched seed genes from where we started to extend subnetworks. The seed genes were selected to be top 1% differentially expressed genes in the last time point (9 dpi) since the time point at 9 dpi contained the highest RNA abundance level. The degree of differentiation was measured by t-test statistics scores, therefore we prepared the seed genes, significantly differentially expressed between the two phenotypes. Based on the co-expression networks, we greedily expanded the subnetworks from the selected seed genes. For each subnetwork module, the subnetwork started extending from each seed gene through its neighboring genes with certain conditions until the stopping criterion was satisfied. In this extension of subnetwork modules, we applied a computationally efficient branch-out technique. We first evaluated probabilistic activity level of each subnetwork when adding each connected gene into the subnetwork and chose the optimum subnetwork whose discriminative power enhancement (measured by t-test scores) of the activity level exceeded by 10%. Subsequently, we also selected two more suboptimal subnetworks whose discriminative power increase were within 2% difference from the top increment of the optimum subnetwork. During the process, the subnetwork activity level was computed based on the inference method in [58]. Next, we again iteratively tested each connected gene to the subentworks by calculating the activity levels and expanded them by the brach-out method with the two specific conditions. The whole process was continuously repeated for all the seed genes as well as the five co-expression networks. As a result, we identified potential subnetwork modules of *F. verticillioides* possibly corresponding to functional pathways associated with the pathogenicity.

In order to compute the pathogenicity-associated strength of a potential subnetwork, as described in the previous subsection, we needed a method inferring the subnetwork activity level based on expression levels of the member genes. In this probabilistic inference, we adopted the inference method proposed in [58] to probabilistically predict the subnetwork activity level. Suppose $\mathcal{G} = \{g_1, g_2, g_3, \cdots, g_n\}$ is the set of genes that belong to a subnetwork module of interest. Let $\mathbf{x} = \{x^1, x^2, x^3, \cdots, x^n\}$ be expression levels of the member genes in a given subnetwork module. The activity level of the given subnetwork can be measured by

$$\eta(\mathbf{x}) = \sum_{k=1}^{n} \alpha^k(x^k), \tag{3.1}$$

where $\alpha^k(x^k)$ is the log likelihood ratio (LLR) between two phenotypes (*i.e.*, the wild type vs. the mutant) defined as follows

$$\alpha^k(x^k) = log\left[\frac{y_1^k(x^k)}{y_2^k(x^k)}\right]. \tag{3.2}$$

where $y_1^k(x)$ is the conditional probability density function (PDF) of the expression level of gene g_k in one phenotype. $y_2^k(x)$ is the conditional PDF of the expression level of gene g_k in the other phenotype. Through this probabilistic inference, we can predict the activity level $\eta(\mathbf{x})$ of the given subnetwork module based on the expression levels $\mathbf{x}$ of the member genes. Note that the concept of this probabilistic approach comes from naive Bayes model (NBM). We can also evaluate the discriminative power of the probabilistic activities of the given subnetwork module composed of the genes in $\mathcal{G}$ based on t-test statistics as follows

$$t(\mathcal{G}) = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{3.3}$$

where μ_1 and s_1^2 are the mean and the variance of the subnetwork activity level in one phenotype, and μ_2 and s_2^2 are the mean and the variance of the subnetwork activity level in the other phenotype. n_1 is the number of samples for one phenotype, and n_2 is the number of samples for the other phenotype.

3.2 Results

3.2.1 Characteristics of the F. verticillioides *co-expression networks*

Before identifying potential subnetwork modules, we investigated characteristics of the co-expression networks constructed from our RNA-Seq data. Specifically, we tested two important features such as degree exponent and clustering coefficient to understand the structure of predicted biological networks [4]. We first evaluated the degree exponent of each co-expression network to test whether a network is scale-free. Scale-free networks are known as networks composed of a relatively small number of highly connected hubs and a large number of nodes that are barely connected [4]. It is known that the degree distributions in scale-free networks follow the power law $P(k) \sim k^{-\gamma}$, where k is the degree and γ is the degree exponent, and also the degree exponent is normally in the range of $2< \gamma <3$ for real biological networks [4]. In addition, we computed the clustering coefficient whose concept is defined as the

Table 3.3: Degree exponent and clustering coefficient of the networks.

	Thresholds				
	0.94	0.93	0.92	0.91	0.90
Degree exponent	2.98	2.98	2.97	2.96	2.96
Clustering coefficient	0.31	0.33	0.35	0.37	0.38

actual number of interactions between the connected nodes divided by the maximum number of interactions that the node could have. The clustering coefficient of node I can be calculated by

$$\mathcal{C}_I = 2n_I/k_I(k_I - 1), \tag{3.4}$$

where k_I is the number of nodes in the neighborhood of node I so that the maximum number of interactions could be $k_I(k_I - 1)/2$, and n_I is the actual number of interactions between their connected nodes. Higher clustering coefficient indicates stronger cohesiveness among neighbors. The average clustering coefficient of hypothetical hierarchical networks is known to be around 0.6 [4]. The degree exponent and the average clustering coefficient of the *F. verticillioides* co-expression networks are shown in Table 3.3. In Table 3.3, the degree exponent of our co-expression networks was in the range $2.96 < \gamma < 2.98$, which is comparable to that of real biological networks observed in nature. Also, the average clustering coefficient was in the range $0.31 < \overline{\mathcal{C}} < 0.38$. Generally, the higher threshold for a co-expression network, the lower average clustering coefficient that the co-expression network has since the network contains fewer interactions that are more significant.

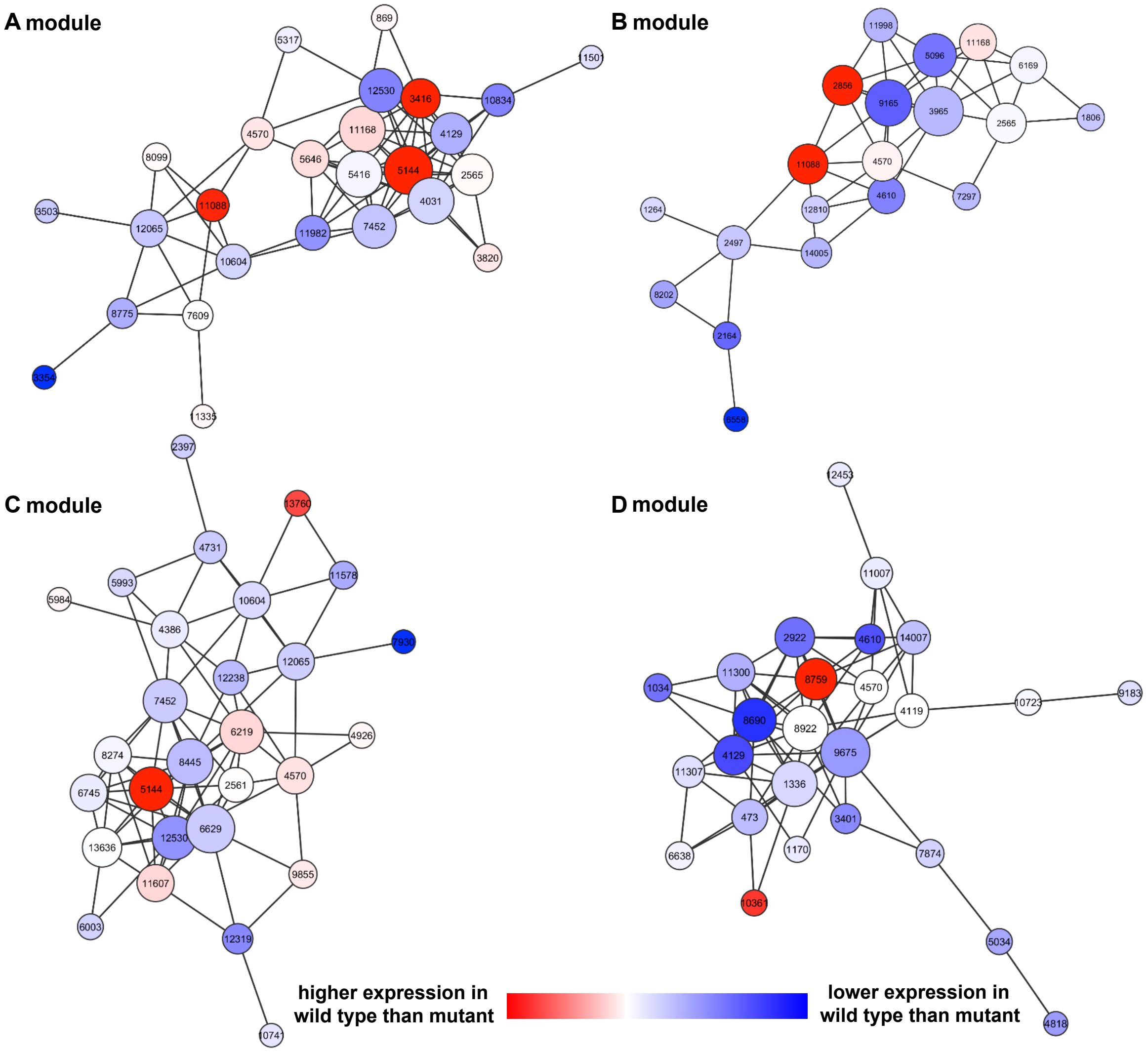


Figure 3.2: Potential subnetwork modules [33] © [2015] IEEE.

3.2.2 *Identification of potential pathogenic subnetwork modules*

Through our proposed analysis approach, we identified four potential *F. verticillioides* pathogenicity-associated subnetwork modules shown in Figure 3.2. We selected seed genes, top 1% differentially expressed genes at the last time point (9 dpi). Based on each co-expression network shown in Table 3.2, we extended the

Table 3.4: Properties of the potential *F. verticillioides* subnetwork modules.

Potential subnetworks	# of genes	# of interactions	*t*-test score	clustering coefficient
module A	26	77	7.3	0.56
module B	20	44	7.1	0.47
module C	28	80	7.1	0.50
module D	26	74	6.8	0.50

subnetwork around each seed gene, by adding a gene from its neighborhood, and allowed extended subnetworks whose discriminative power (measured by *t*-test statistics score) increased at least by 10% at each extension to be candidates. Among these subnetworks, we first chose the subnetwork with the highest discriminative power and subsequently selected up to two additional subnetworks whose discriminative power was very close to the maximum score (*i.e.*, difference less than 2%). We iteratively repeated the process that is a branch-out to extend the subnetworks up to three subnetworks at each step, until they could not be extended further. Next, we selected the subnetwork with the highest discriminative power as a candidate subnetwork for the seed gene. The entire process was repeated for all seed genes in all five co-expression networks. Finally, we chose the top four subnetwork modules with the highest discriminative power shown in Figure 3.2 amongst all candidate subnetworks. Genes that are relatively highly expressed in wild type than the mutant are shown in red, while genes that are relatively highly expressed in the mutant compared to wild type are shown in blue. The discriminative power of the potential subnetworks were 7.3, 7.1, 7.1, and 6.8, respectively, which were significantly higher than other candidate subnetworks. As shown in Table 3.4, the four subnetwork modules had 26, 20, 28 and 26 genes, and 77, 44, 80 and 74 interactions, respectively. We evaluated the average clustering coefficients for the potential modules, which were in the range

0.47< $\overline{\mathcal{C}}$ <0.56 and found that these are much higher than those of the co-expression networks. Therefore, the potential subnetwork modules can be thought to be more cohesive compared to other parts in the networks. Also, the average clustering coefficients of the potential modules are closer to that of an ideal hypothetical hierarchical network [4] so that the potential subnetwork modules can be considered to be a more prominent hierarchical structure compared to other network regions.

3.2.3 Functional coherence of genes in the subnetwork modules

In order to test functional coherence among the genes in the potential subnetwork modules, we analyzed their GO (Gene Ontology) terms. In this GO analysis, we first obtained the 44 functional groups for *F. verticillioides* through g:Profiler (http://biit.cs.ut.ee/gprofiler/) [49]. Next, we evaluated a metric which we call the "classification rate," inspired by a similar concept in [67]. Suppose $\mathcal{G} = \{g_1, g_2, g_3, \cdots, g_m\}$ is the set of genes in a subnetwork module. We define $\mathcal{G}_{i,d} = \{g_{i,1}, g_{i,2}, g_{i,3}, \cdots, g_{i,n(i,d)}\}$ as the set of genes in the subnetwork whose distance from gene g_i is d, where $n(i,d)$ is the number of genes that are at distance d from gene g_i. The classification rate $\Gamma(i,d)$ is defined as follows

$$\Gamma(i,d) = \frac{\sum_k \omega(i,k)}{n(i,d)}, \tag{3.5}$$

where

$$\omega(i,k) = \begin{cases} 1, & \text{functional category of } g_i \\ & \quad = \text{functional category of } g_{i,k} \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

For each of the four subnetwork modules, we computed the average classification rate for all genes in a given module for distance $d = 1, 2, 3$. We also computed the average classification rate for random networks to compare it with that of the potentials. We generated random networks whose number of nodes and interactions were uniformly

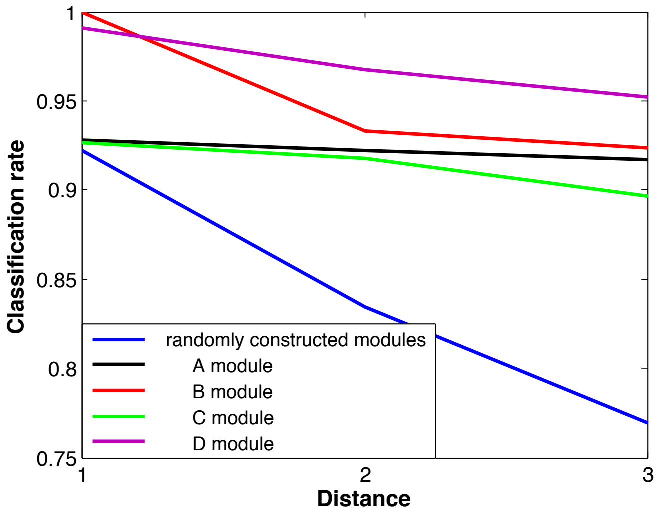


Figure 3.3: Classification rates for randomly constructed modules and the potential subnetwork modules [33] © [2015] IEEE..

distributed over $[20, 30]$ and $[40, 80]$, respectively. In this comparison,, we repeated computing the classification rate of random networks for 5000 times to compute the average classification rate. The results are shown in Figure 3.3. According to Figure 3.3, the average classification rate of the four potential subnetwork modules for up to distance $d = 3$ was mostly above 0.9, which is significantly higher than that of randomly constructed modules. As shown in Figure 3.3, the potential subnetwork modules are functionally coherent since most of the genes in the detected modules belong to the same functional category. It is interesting to note that the average classification rate of random networks is still relatively high ($e.g.$, 0.769 for $d = 3$) and the two main reasons are: (i) many genes are typically assigned with multiple GO terms, and one can belong to 4∼5 functional categories; (ii) many genes in the co-expression networks are mainly associated with a few functional categories ($i.e.$, around 50% of genes in the co-expression networks were associated with at least one of the following functional categories: cellular process, metabolic process, binding and catalytic activity). The main functional categories associated with the four modules are as follows: Module A in Figure 3.2: binding and catalytic activity; Module B in Figure 3.2 cellular process and catalytic activity; Module C in Figure 3.2: metabolic process and binding; Module D in Figure 3.2: metabolic process and binding. Significantly, we identified cellular process of the module B in Figure 3.2 to be a possible pathogenic infection process. The pathogenicity-associated GO terms of the infection process is GO:0006914 [autophagy] → GO:0044248 [cellular catabolic process] → GO:0044237 [cellular metabolic process] → GO:0009987 [cellular process] (http://www.geneontology.org).

3.2.4 Potential pathogenic genes in the subnetwork modules

In addition, we investigated whether orthologous genes of the member genes in the potential subnetwork modules were known pathogenic genes in other fungi. We first collected known pathogenic genes in seven other fungi from the Pathogen Host Interaction Database (PHI-base: http://www.phi-base.org) [65]: *Fusarium graminearum* (FG), *F. oxysporum* (FO), *Aspergillus fumigatus* (AF), *Botrytis cinerea* (BC1G), *Magnaporthe grisea* (MGG), *Ustilago maydis* (UM), and *Cryptococcus neoformans* (CNAG). Next, we searched whether any genes of the potential modules have known pathogenic orthologous genes and identified several orthologous genes as the following: Module A in Figure 3.2 included five potential orthologs FVEG_00869, FVEG_03354, FVEG_07609, FVEG_08099 and FVEG_11501; Module B in Figure 3.2 contained two potential orthologs FVEG_02164 and FVEG_05096; Module C in Figure 3.2 included one potential ortholog, FVEG_06629; Module D in Figure 3.2 contained three potential orthologs FVEG_00473, FVEG_01034 and FVEG_04119. The pathogenic genes whose orthologs are member genes of the potential pathogenic subnetwork modules were previously validated experimentally: (1) FGSG_00408 orthologous to FVEG_00869 (*F. graminearum* mutant deleted of FGSG_00408 was blocked in perithecium formation so that it reduced virulence on kinome of wheat scab [63]); (2) MGG_04556 orthologous to both FVEG_03354 and FVEG_06629 (*M. grisea* mutant of MGG_04556 reduced virulence on rice blast [24]; (3) MGG_00454 orthologous to FVEG_02164 and MGG_03580 orthologous to FVEG_05096 (*M. grisea* mutant of MGG_00454 involved in initiation of autophagy as well as *M. grisea* mutant of MGG_03580 playing a role in phagophore and autophagosome expansion reduced virulence on rice disease [28]; (4) FGSG_00574 orthologous to FVEG_01034 (FGSG_00574 knock-out mutant of *F. graminearum* showed no perithecia development and reduction in virulence and growth on cereal head blight [57]. Consequently, the identified potential subnetwork modules included one or more orthologs of known pathogenic genes in other fungi, and it is possible to expect that these genes might play important roles with the other genes of the modules on the pathogenesis of *F. verticillioides*

4. STUDY OF MAIZE DEFENSE MODULES BASED ON MAIZE–*F. VERTICILLIOIDES* INTERACTIONS *

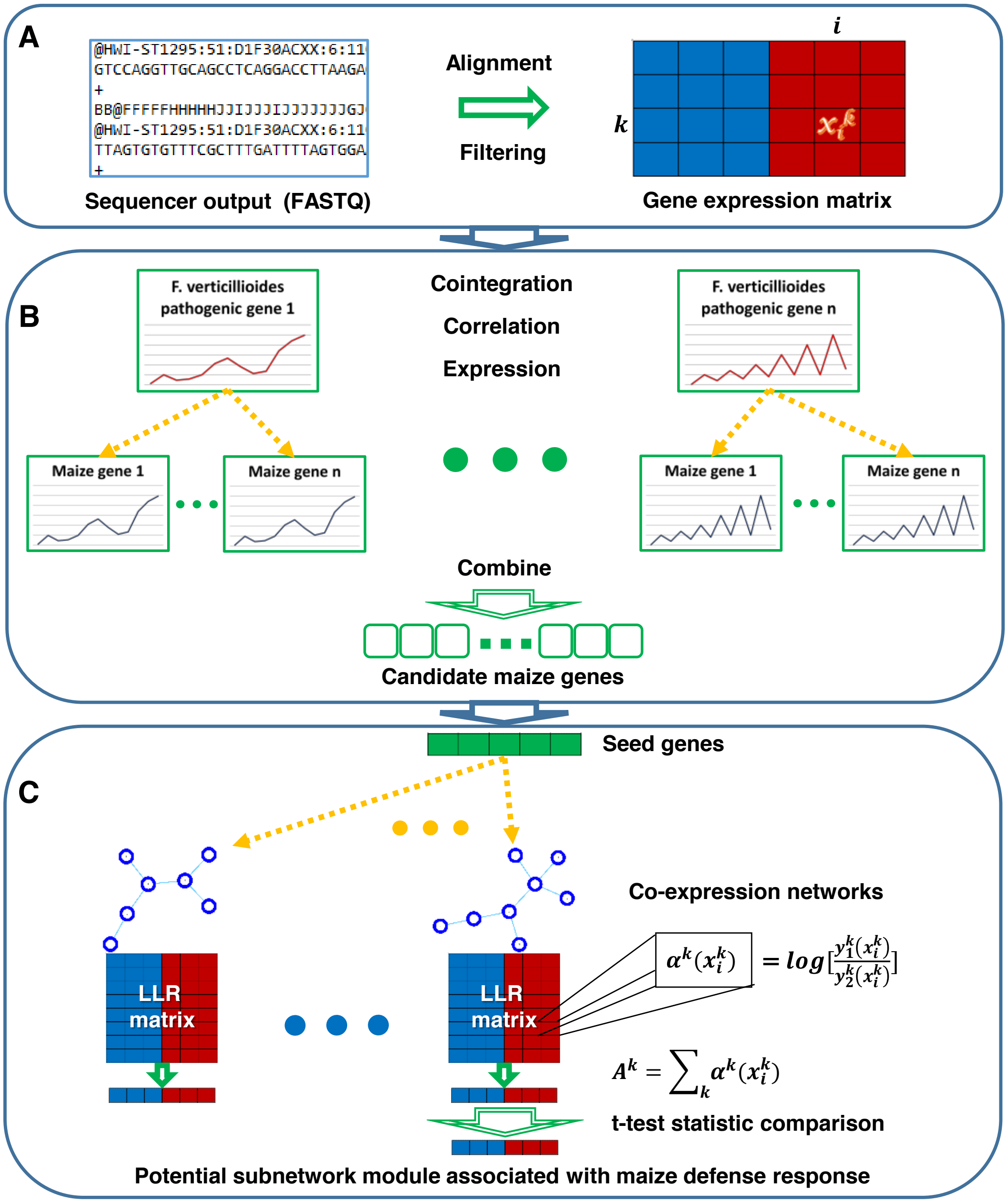
An overview of the network-based comparative analysis pipeline to identify maize defense subnetwork modules based on maize-*F. verticillioides* interactions is shown in Figure 4.1. First, with the preprocessed gene expression data, we applied the cointegration-correlation-expression approach to find candidate maize genes whose expression patterns are corresponding to known *F. verticillioides* pathogenicity genes. Second, we constructed co-expression networks of the candidate maize genes through partial correlation and also converted the gene expression data of the genes into a log-likelihood ratio (LLR) matrix. Next, based on the co-expression networks and the log-likelihood ratio matrix, we extended maize defense subnetwork modules from top 20% significantly differentially expressed genes utilizing a computationally efficient branch-out technique. Finally, we identified potential maize subnetwork modules possibly associated with maize defense response.

4.1 Methods

4.1.1 Preprocessing

Basically, the detailed explanation of preprocessing for *F. verticillioides* (*i.e.*, sample preparation and RNA sequencing) is the same as the previous study since the same dataset was applied in this study. After sample preparation and RNA sequencing, alignment was proceeded as previously illustrated. Table 4.1 demonstrates

* Part of this chapter is reprinted with permission from "Mansuck Kim, Huan Zhang, Charles Woloshuk, Won-Bo Shim, and Byung-Jun Yoon. Computational identification of genetic subnetwork modules associated with maize defense response to *Fusarium verticillioides*. *BMC Bioinformatics*, 16(Suppl 13):S12, 2015" [32] © [2015] BMC.



```
@HWI-ST1295:51:D1F30ACXX:6:11
GTCCAGGTTGCAGCCTCAGGACCTTAAGA
+
BB@FFFFFHHHHHJJIJJJJIJJJJJJJGJ
@HWI-ST1295:51:D1F30ACXX:6:11
TTAGTGTGTTTCGCTTTGATTTTAGTGGA
+
```

A
Sequencer output (FASTQ)
Alignment
Filtering
i
k
x_i^k
Gene expression matrix
B
F. verticillioides pathogenic gene 1
Cointegration
Correlation
Expression
F. verticillioides pathogenic gene n
Maize gene 1
Maize gene n
Maize gene 1
Maize gene n
Combine
Candidate maize genes
Seed genes
C
Co-expression networks
LLR matrix
LLR matrix
$\boxed{\alpha^k(x_i^k)} = log[\frac{y_1^k(x_i^k)}{y_2^k(x_i^k)}]$
$A^k = \sum_k \alpha^k(x_i^k)$
t-test statistic comparison
Potential subnetwork module associated with maize defense response

Figure 4.1: Overview of the proposed network-based comparative analysis [32].

general statistics of the RNA-seq datasets for both maize and *F. verticillioides*. The table not only shows the differences between the two strains (*i.e.*, wild type vs. the mutant), but also indirectly explains how the pathogenicity of *F. verticillioides* have an effect on maize transcription profile over time. After the alignment process, the other preprocessing, filtering process, for maize genes was also completed so that 42.2% of maize genes (57,676 genes) remained for downstream analysis. The basic procedure was the same as the way that *F. verticillioides* was proceeded as previously described.

4.1.2 Selection of pathogenicity genes of F. verticillioides

We chose representative *F. verticillioides* pathogenicity genes to narrow down maize genes to candidates likely involved in maize defense response. Using the selected *F. verticillioides* pathogenicity genes, we compared their expression patterns over time with those of maize genes. In this comparison, four genes were arbitrarily selected among known *F. verticillioides* pathogenicity genes for our analysis; i) *FSR1* (FVEG_09767) - involved in fungal virulence and sexual mating [54]; ii) *FST1* (FVEG_08441) - associated with fungal growth particularly on maize ears [31]; iii) *FvVE1* (FVEG_09521) - involved in strong pathogenesis and toxin production on maize seedlings [45]; iv) *ZFR1* (FVEG_09648) - a significant transcription factor

Table 4.1: General statistics of maize–*F. verticillioides* RNA-seq dataset.

		Wild type			Mutant		
		3dpi	6dpi	9dpi	3dpi	6dpi	9dpi
	Type of run	single					
	Read length	100 (bp)					
F. verticillioides	Mean # of reads aligned	82504	200827	235702	24711	63462	138406.3
	Median depth of coverage	5.8	14.2	16.6	1.8	4.5	9.8
maize	Mean # of reads aligned	4394510	4183565	3377730	3798777	3589860	3577877
	Median depth of coverage	32.1	30.6	24.7	27.8	26.2	26.2

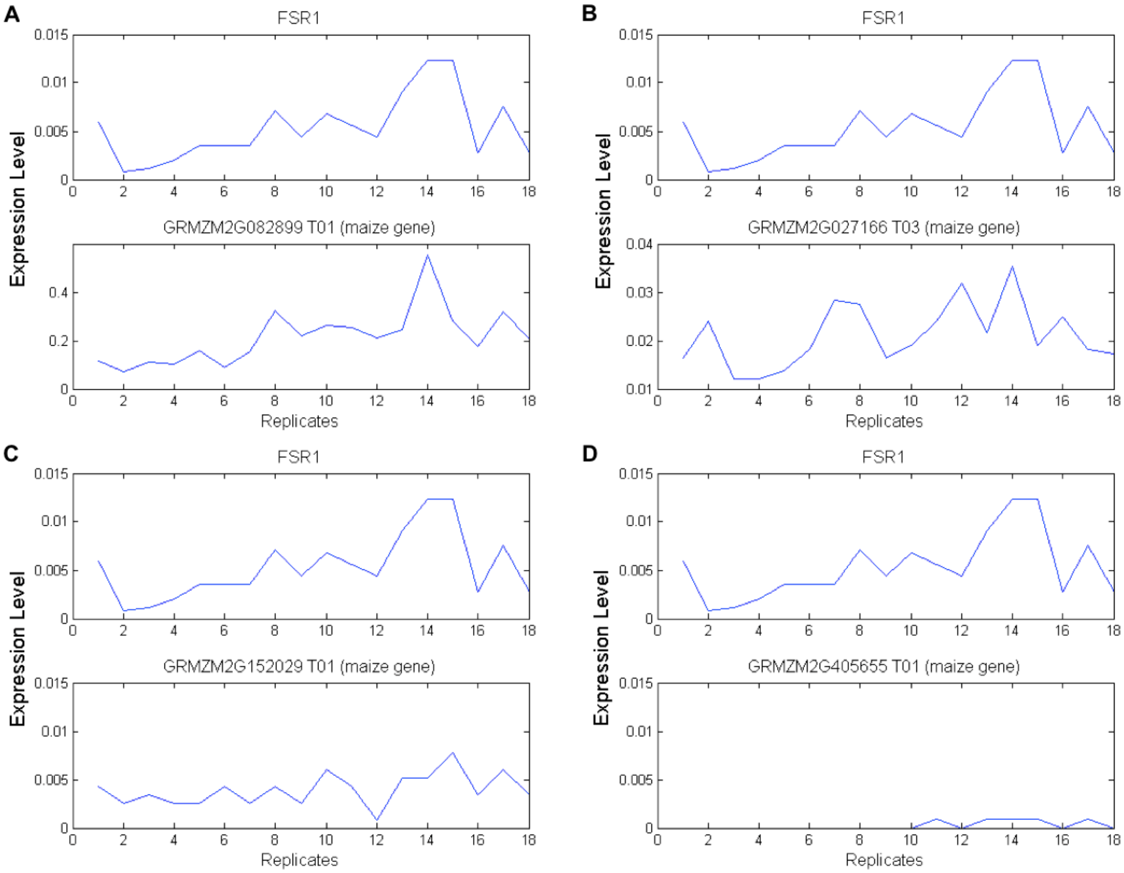


Figure 4.2: Four practical examples in comparison between known *F. verticillioides* pathogenic gene (*FSR1*) and maize genes [32].

controlling fungal growth on maize kernels [7].

4.1.3 Cointegration-correlation-expression approach

Based on the four selected *F. verticillioides* pathogenicity genes, we comprehensively analyzed expression patterns of maize genes over time across all replicates to narrow down maize genes into candidates. For this selection process, three analytical approaches such as cointegration, correlation, and considering expression significance were complementarily applied and cooperated to compare tendency of expression levels between maize genes and the selected *F. verticillioides* pathogenicity genes.

First, we applied cointegration [16] considering nonstationality and time-varying uncertainty in the two species (i.e., maize vs. *F. verticillioides*). In this analysis, we examined a relationship of expression levels across all replicates between maize genes and the selected *F. verticillioides* genes. Specifically, two expression levels between maize and each *F. verticillioides* pathogenicity were considered to be cointegrated when their expression trend was comparable over time across all replicates. The Engle-Granger method was used to investigate single cointegrating relations between the host and the pathogen. For this cointegration, corresponding maize genes whose P-value of the method were lower than 0.05 to each selected *F. verticillioides* pathogenicity genes were considered as candidates. Second, we applied correlation to track expression patterns of the two species over all replicates. This analysis quantified the strength of a linear relationship between maize genes and each selected *F. verticillioides* pathogenicity gene over time across replicates. Corresponding maize genes whose Pearson's correlation coefficients were higher than 0.65 (P-values less than 0.0035) to each selected *F. verticillioides* pathogenicity gene were taken into account as candidates. Third, all maize expression levels over all replicates were observed to remove rarely expressed genes. Maize genes whose mean expression levels were in the top 80% of all genes and expressed in all replicates, were considered as candidates since the selected *F. verticillioides* pathogenicity genes were also expressed in all replicates. Through this combination approach based on cointegration-correlation-expression, each selected *F. verticillioides* pathogenicty gene was used as a criterion in searching candidate maize genes. We finally acquired candidate maize genes by comparison with the four selected pathogenicity genes and combined them for subsequent analysis. Figure 4.2 illustrates the practical examples why this combinational analysis is meaningful in searching candidate maize genes likely associated with maize defense mechanism. Specifically, as shown in Figure 4.2 (B-D), each case demonstrates disagreed relationship between the two species (maize vs. *F. verticillioides*) when one of combination analysis (cointegration-correlation-expression approach) does not satisfy certain levels while the other two meet adequate levels.

4.1.4 Identification of maize subnetwork modules

The basic description to identify subnetwork modules based on the co-expression networks are shown in the previous sections 3.1.2 and 3.1.3. Also, in this identification of genetic subnetwork modules possibly associated with maize defense response, we adopted the probabilistic pathway activity inference scheme, originally proposed in [58] and applied to the previous study as described in details in section 3.1.4. Here, in order to identify maize defense subnetwork modules, we first constructed the co-expression networks of candidate maize genes specifically selected against the *F. verticillioides* pathogenicity genes through the cointegration-correlation-expression approach. The co-expression networks were predicted using certain level of partial correlation based on the preprocessed gene expression matrix of the candidate maize genes. In this network prediction, we constructed four different co-expression networks at four distinct threshold levels as shown in Table 4.2. Therefore, we excluded relatively unsatisfactory interactions between the candidate genes (although we already specifically selected those candidates), and mitigated dependency on a specific threshold level by constructing more than one co-expression network. Subsequently, seed genes, significantly differentially expressed in the two phenotypes among the candidate maize genes, were found to search subnetwork modules possibly involved in maize defense response. For this selection, maize genes whose discriminative power of expression levels between the two conditions was in top 20% were chosen. Next, we extended maize subnetwork modules from the seed genes to search potential genetic modules likely involved in maize defense response based on the co-expression net-

Table 4.2: Co-expression networks of the maize candidates (wild-type vs. mutant).

Threshold (cut-off partial correlation)	Wild type		Mutant	
	# of genes	# of interactions	# of genes	# of interactions
0.9	97	269	93	243
0.8	106	546	101	518
0.7	111	868	107	845
0.6	114	1257	111	1211

works. We first investigated whether expanding the subnetwork module by adding one of neighboring genes to a seed gene would enhance the discriminative power of their probabilistic activities. Subsequently, we decided whether to extend the subnetwork based on two conditions: i) each subnetwork extension by adding a gene improves the discriminative power (measured by t-test statistics score) of the probabilistic activities by at least 5%, ii) the extension is allowed up to three subnetwork modules when the discriminative power difference of the activities between the optimal and suboptimal is within 2% at each expansion. We kept extending the subnetwork modules by appending one of connected genes through this computationally efficient branch-out technique until none of the extended modules could enhance the discriminative power of the activities by at least 5%. We performed the whole process for all seed genes based on the four co-expression networks to identify potential genetic subnetwork modules possibly involved in maize defense response. Consequently, we identified potential maize defense subnetwork modules by considering their differentiation strength of probabilistic activities as well as association to significantly annotated GO terms involved in defense response.

while identifying potential subnetwork modules, we evaluated the goodness of a given subnetwork module by inferring the module activity and assessing its effectiveness in discriminating between the two different conditions. For this purpose, we adopted a probabilistic pathway activity inference scheme, which was originally proposed in [58] and was previously applied to the prediction of pathogenic gene modules in *F. verticillioides* [**?**]. In the following, we present a brief summary of the method. Suppose we have a set of genes $\mathcal{G} = \{g_1, g_2, \cdots, g_n\}$ that belong to a given subnetwork module and the expression levels of these genes are $\mathbf{x} = \{x^1, x^2, \cdots, x^n\}$. The activity level of the given subnetwork module can be measured by

$$\eta(\mathbf{x}) = \sum_{k=1}^{n} \alpha^k(x^k), \tag{4.1}$$

where $\alpha^k(x^k)$ is the log likelihood ratio (LLR) between the two conditions (*i.e.*, maize inoculated with two different strains – wild type vs. the mutant – of *F. verticillioides*) defined as follows

$$\alpha^k(x^k) = log\left[\frac{y_1^k(x^k)}{y_2^k(x^k)}\right]. \tag{4.2}$$

In equation (5.2), $y_1^k(x)$ is the conditional probability density function (PDF) of the expression level of gene g_k in one condition. Similarly, $y_2^k(x)$ is the conditional PDF of the expression level of gene g_k in the other condition. We can estimate the activity level of $\eta(\mathbf{x})$ of the subnetwork module as defined in (5.1) and also assess its discriminative power for differentiating between the two different conditions using the t-test statistics score:

$$t(\mathcal{G}) = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{4.3}$$

where μ_1 and s_1^2 are the mean and the variance of the subnetwork activity level in one condition, and μ_2 and s_2^2 are the mean and the variance of the subnetwork activity level in the other condition. n_1 and n_2 are the number of replicates (or independent measurements) in the respective conditions.

4.2 Results

4.2.1 Characteristics of the candidate maize genes

Before identifying functional subnetwork modules associate with maize defense mechnism, we first tested whether the candidate maize genes predicted by our analysis (*i.e.*, cointegration-correlation-expression approach using the selected *F. verticillioides* pathogenicity genes) were likely involved in maize defense response. To investigate whether the maize candidates were defense-associated to *F. verticillioides* pathogenicity, we compared the candidates associated with *F. veticillioides* pathogenicity with corresponding maize genes to the *F. veticillioides* housekeeping genes, which are constitutively expressed and mainly involved in the maintenance of fundamental cellular functions. Therefore, we selected four common *F. verticillioides* housekeeping genes in this molecular genetic study, such as two beta-tubulin genes (FVEG_04081 & FVEG_05512), pyruvate dehydrogenase E1 component subunit alpha gene (FVEG_07074), and glyceraldehyde 3-phosphate dehydrogenase gene (FVEG_04927), were applied for this comparison. For each *F. verticillioides* housekeeping gene, corresponding maize genes were identified by the cointegration-correlation-expression analysis, and compared with the candidate maize genes against the four selected *F. verticillioides* virulence genes. Since not only the *F. verticillioides* housekeeping genes, but also the selected *F. verticillioides* pathogenicity genes were all relatively significantly expressed in all replicates so, it is possible to presume that their expression patterns might be similar. However, the corresponding maize genes against *F. verticillioides* housekeeping genes and the miaze candidates against the selected *F. verticillioides* virulence genes shared only about 20% common genes in average. As a result, we considered the candidate maize genes against the selected *F. verticillioides* pathogenicity genes obtained through the cointegration-correlation-expression approach were possibly associated with maize defense response.

4.2.2 Identification of potential maize defense subnetwork modules

By performing our proposed network-based comparative analysis pipeline, we identified four potential genetic subnetwork modules possibly associated with maize defense mechanism against *F. verticillioides* pathogenicity. Figure 4.3 & 4.4 show the four identified potential functional subnetwork modules likely involved in maize defense response. In order to search for such potential subnetwork modules, we first selected the four representative *F. verticillioides* pathogenicity genes, and utilize them as criteria in comparison with all maize genes to find the maize candidates. In this comparison, we applied the cointegration-correlation-expression analysis to find the candidates whose expression patterns are comparable with those of the selected *F. verticillioides* pathogenicity genes. Using the identified maize candidates, we constructed co-expression networks, and further selected top 20% significantly differentially expressed genes among the candidates as the seed genes. For each seed gene, we extended the subnetwork module whose discriminative power (measured by *t*-test statistics score) enhanced at least by 5% when adding one of the connected genes. Amongst the extended subnetwork modules, we chose the subnetwork with the optimum discriminative power and up to two more subnetworks with suboptimal discriminative power when the difference from the optimum was within 2%. The extending process was iteratively repeated until non of subnetworks did not show discriminative power improvement by at least 5%. Also, the entire process was reiterated for all the seed genes at all the four co-expression networks. We selected the four potential functional modules possibly associated with maize defense response by considering whether each subnetwork was not only significantly differentiated between the two phenotypes, but also annotated by significant GO terms associated with maize defense system directly or partially. In Figure 4.3 & 4.4, genes relatively highly expressed in the wild-type-infected are indicated in red tone, whereas genes relatively highly expressed in the mutant-infected are illustrated in blue tone. Basic properties of the four identified maize functional modules are shown in Table 4.3. As shown in Table 4.3, the properties such as number of genes and number of interactions ranged from 6 to 8 and from 5 to 8, respectively. Also, the t-test statistics scores ranged from 5.1 to 7.2 that were relatively higher than most other candidate subnetworks.

Table 4.3: Properties of the potential maize defense subnetwork modules.

Potential maize subnetworks	# of genes	# of interactions	t-test score
Fig. 3.3 module A	7	6	5.4
Fig. 3.3 module B	8	8	5.6
Fig. 3.4 module C	8	8	7.2
Fig. 3.4 module D	6	5	5.1

4.2.3 Potential maize subnetwork modules directly associated with maize defense response

Through our proposed network-based comparative analysis pipeline, we identified two potential maize subnetwork modules that includes maize genes whose significantly annotated GO terms were specifically associated with responses to fungi. In this validation, P-values of Benjamini-Hochberg false discovery rate (FDR) method [5] provided by g:Profiler (http://biit.cs.ut.ee/gprofiler/index.cgi) [49] was used to distinguish the most significant GO terms for the identified functional modules. In Figure 4.3, module A contained three known maize genes such as GRMZM2G001696_T01,

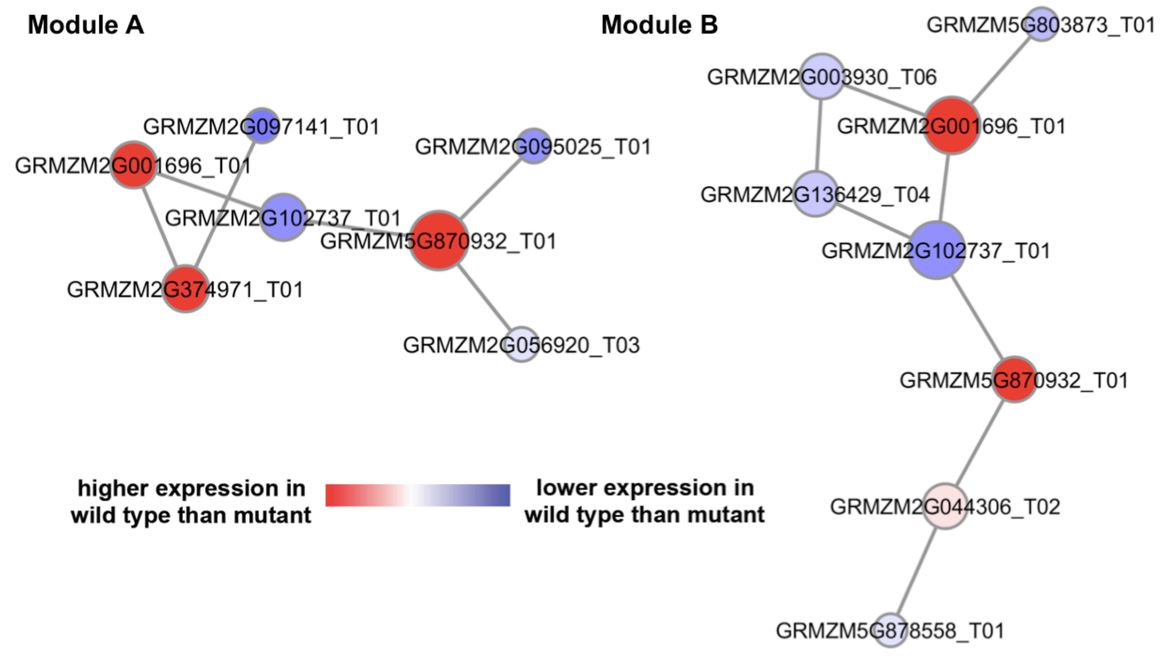


Figure 4.3: Two identified potential maize subnetwork modules [32].

GRMZM2G374971_T01, and GRMZM5G870932_T01 involved in significant GO term, GO:0009817. The term, GO:0009817, is ′defense response to fungus (incompatible interaction)′, whose definition is ′a response of an organism to a fungus that prevents the occurrence or spread of disease′. *P*-value of Benjamini-Hochberg FDR for this term was 1.25e-06. The other module in Figure 4.3, module B, contained three known maize genes such as GRMZM2G001696_T01, GRMZM5G870932_T01, and GRMZM5G878558_T01 associated significant GO term, GO:0009620. This GO term is ′response to fungus′ and the definition is ′any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus from a fungus′; the *P*-value of Benjamini-Hochberg FDR for this term was 5.69e-07. Note that both GO terms are significantly associated with maize defense mechanism against fungal pathogens. It is also possible to view these two functional modules as one whole module since they share key genes such as GRMZM2G001696_T01 and GRMZM5G870932_T01. However it is important to note that they were searched from different seed genes at different co-expression networks. Consequently, the two identified potential maize functional modules, specifically associated with significant GO terms involved in maize defense response, proved that our proposed network-based comparative analysis pipeline performed meaningful and reliable analysis.

4.2.4 Potential maize subnetwork modules indirectly involved in maize defense response

Our proposed analysis pipeline identified two more potential maize functional modules that were not directly associated with specific GO terms involved in the defense response against fungal pathogens, however they also demonstrated potential relevance to defensive mechanism. As shown in Figure 4.4, module A contained four known maize genes such as GRMZM2G003930_T06, GRMZM2G056920_T03, GRMZM2G095025_T01, and GRMZM5G878558_T01 associated with GO term GO:0046914. This GO:0046914 is ′transition metal ion binding′ and the *P*-value of Benjamini-Hochberg FDR for this term was 4.56e-02. The definition of this GO term is 'interacting selectively and non-covalently with a transition metal ions that is an element whose atom has an incompleted-subshell of extranuclear electrons, or which gives rise to a cation or cations with an incompleted-subshell′. Module B in Figure 4.4 contained two known maize genes such as GRMZM2G001696_T01 and GRMZM2G085019_T01, which are associated with GO term, GO:0046686. This GO term is ′response to cadmium ion′ and defined by 'any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a cadmium (Cd) ion

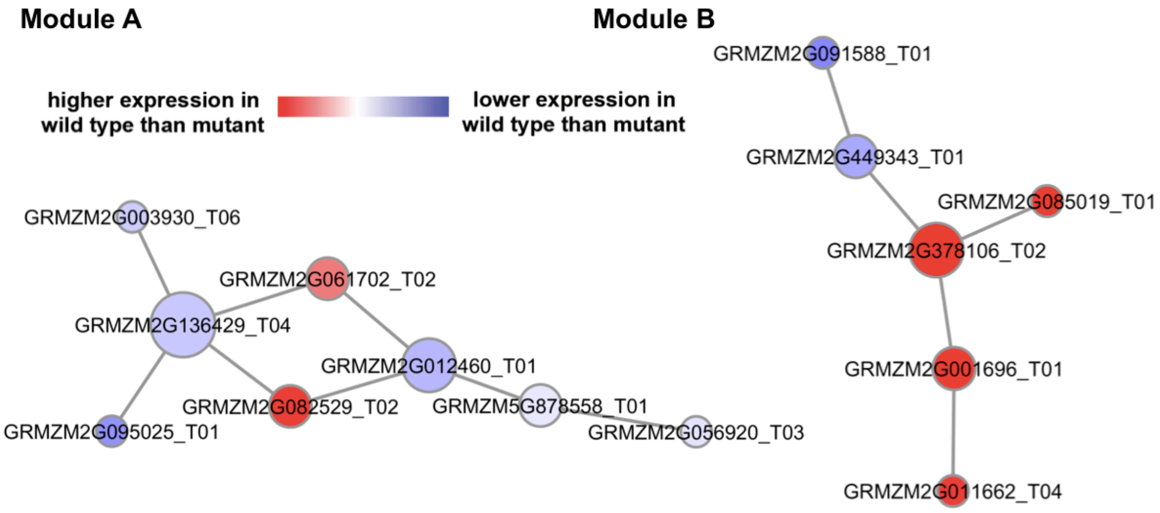


Figure 4.4: Two predicted potential maize subnetwork modules [32].

stimulus′. The P-value of Benjamini-Hochberg FDR for this GO term, GO:0046686, was 3.00e-02. According to these two GO terms (GO:0046914 & GO:0046686), it is known that hyperaccumulation of transition metals tends to reduce the growth of pathogens [17], hence we could expect that transition metals with cadmium (Cd) have a positive effect on plant defense system against fungal pathogenicity. Since both GO terms were all significantly relevant to transition metals, we can take into account the two potential maize subnetwork modules in Figure 4.4 to be likely involved in maize defense mechanism.

In addition, we investigated whether orthologous genes of the two identified maize functional modules in Figure 4.4 are defense-associated. Orthologous genes of *Sorghum bicolor* as well as *Arabidopsis thaliana* provided by RGAP (Rice Genome Annotation Project) website (http://rice.plantbiology.msu.edu) were used. In this cross-check, we found that orthologous genes in Figure 4.4 module A such as SB04G024300, SB04G034470, and SB07G022750 were also associated with GO:0009610, ′response to symbiotic fungus′, other than transition metal ion binding such as AT2G01275, AT2G20030, AT4G28890, AT1G37130, and AT1G77760 for *Arabidopsis thaliana* as well as SB03G007810, SB09G030900, SB04G024300, SB04G034470, and SB07G022750 for *Sorghum bicolor*. Also, orthologous genes of module B in Figure 4.4 such as AT1G59500, AT2G23170, and AT4G37390 for *Arabidopsis thaliana* as well as SB01G032020, SB02G038170, and SB03G035500 for *Sorghum bicolor* were annotated by GO:0010279. This term, GO:0010279, is ′indole-3-acetic (IAA) acid amido synthetase activity′, which is known to be an important controller for plant defense system [18]. This additional analysis using the orthologous genes of the identified subnetwork modules likely involved in maize defense response in Figure 4.4 demonstrated that it is possible to expect the two maize functional modules have a potential to play a significant role in maize defense mechanism.

5. STUDY OF KEY PATHOGENICITY GENES OF *F. VERTICILLIOIDES*

In this study, we performed network-based comparative analysis to identify key pathogenicity genes for two different dataset. For one dataset, maize stalks were inoculated with *F. verticillioides* wild type and *fsr1* mutant as described in the previous studies. For the other dataset, two different genes (*MADS1* and *MADS2*) were knocked out from *F. verticillioides* wild type so that two different *F. verticillioides* mutants (Fmt1 and Fmt2 strains, respectively) were prepared [47]. With the *fsr1* mutant, we identified key pathogenicity genes of *F. verticillioides* by comparing wild type and the *fsr1* mutant. With the two different *F. verticillioides* mutants, we identified key pathogenicity genes of *F. verticillioides* in common using the both phenotypes by comparing the both pairs (*i.e.*, wild–Fmt1 vs. wild–Fmt2).

Figure 5.1 illustrates an overview of the computation procedure for the proposed network-based comparative analysis predicting specific pathogenicity genes of *F. verticillioides*. First, preprocessing such as alignment, filtering, and normalization is performed, and the preprocessed RNA-seq data is not only applied for inferring *F. verticillioides* co-expression networks through partial correlation, but also converted into a log-likelihood ratio (LLR) matrix for downstream analysis. Second, subnetwork modules are extended from seed genes, significantly differentially expressed between the two different conditions (wild vs. mutant), as long as they keep sufficient strength of differential activity between the two conditions. Third, each potential pathogenicity gene is identified in its detected subnetwork module through analytical investigation whether the gene satisfies certain conditions such as influence on other member genes, relevance to the pathogenicity, and so on. We provide the detailed explanation of computational analysis as well as experimental validation steps in the following subsections.

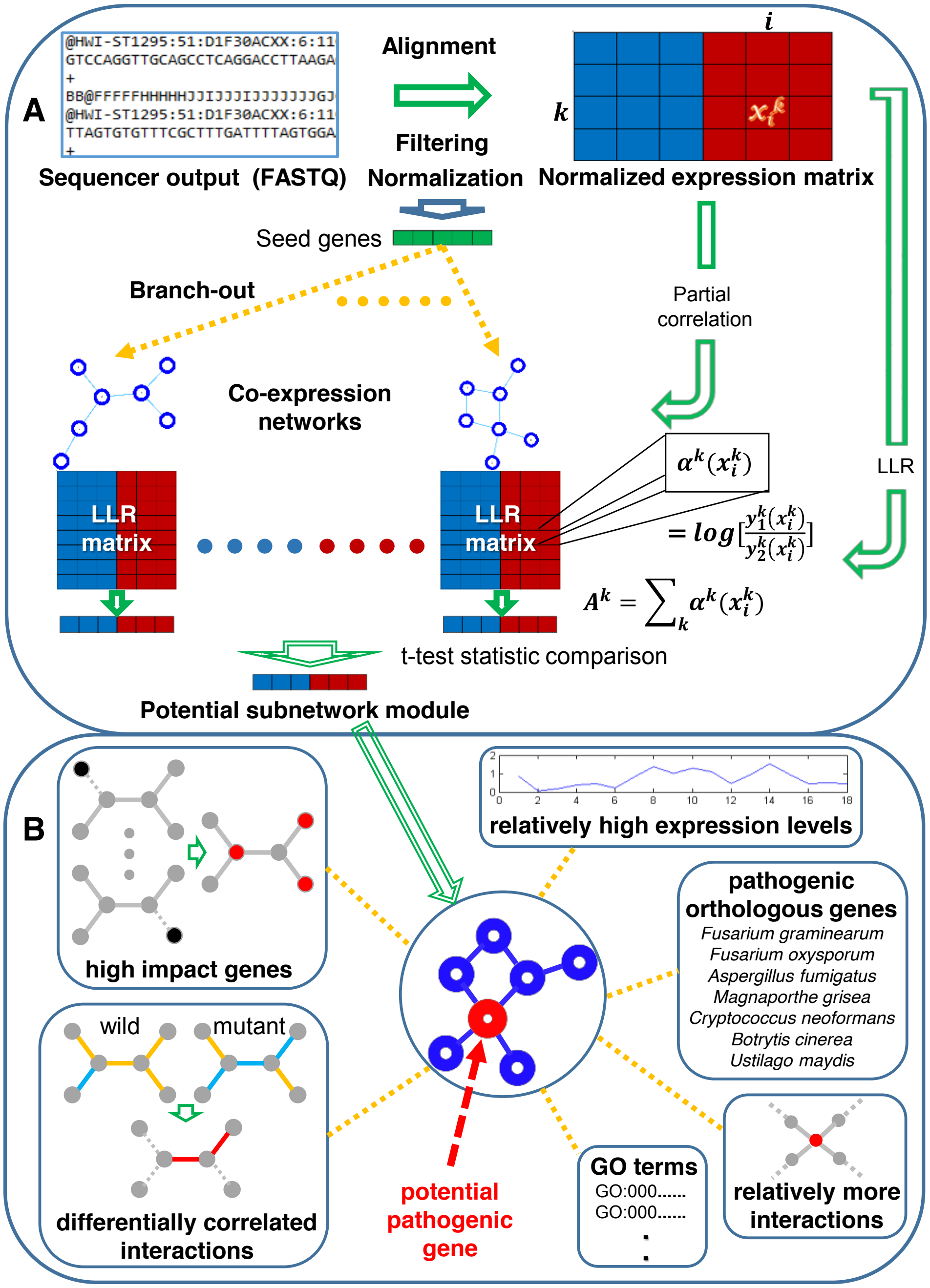

A
@HWI-ST1295:51:D1F30ACXX:6:11
GTCCAGGTTGCAGCCTCAGGACCTTAAGA
+
BB@FFFFFHHHHHJJIJJJIJJJJJJJGJ
@HWI-ST1295:51:D1F30ACXX:6:11
TTAGTGTGTTTCGCTTTGATTTTAGTGGA
+
Sequencer output (FASTQ)
Alignment
Filtering
Normalization
i
k
x_i^k
Normalized expression matrix
Seed genes
Branch-out
Partial correlation
Co-expression networks
LLR
LLR matrix
LLR matrix
$\alpha^k(x_i^k)$
$= log[\frac{y_1^k(x_i^k)}{y_2^k(x_i^k)}]$
$A^k = \sum_k \alpha^k(x_i^k)$
t-test statistic comparison
Potential subnetwork module
B
high impact genes
relatively high expression levels
pathogenic orthologous genes
Fusarium graminearum
Fusarium oxysporum
Aspergillus fumigatus
Magnaporthe grisea
Cryptococcus neoformans
Botrytis cinerea
Ustilago maydis
wild
mutant
differentially correlated interactions
potential pathogenic gene
GO terms
GO:000......
GO:000......
relatively more interactions

Figure 5.1: Overview of the computational prediction analysis.

5.1 Methods

5.1.1 Preprocessing

Since preprocessing of the *fsr1* mutant dataset was explained in the previous studies, the other preprocessing for the two different *F. verticillioides* mutants are provided in this study. After the two different phenotype samples were prepared, RNA sequencing was carried out using Illumina HiSeq 2000 generating 100bp reads in relatively greater quality and coverage. In the sequencing, Illumina's optimized kits performed library preparation and RNA isolation, and also Illumina HCS 1.515.1 and RTA 1.1348.0 simultaneously performed other processes (*i.e.*, quality prefiltering, base calling, uncertainty assessment, and sequence cluster identification). Therefore, five independent sequenced RNA libraries at each time point such as 5 dpi and 7 dpi were processed for wild type and the two mutant *F. verticillioides* (Fmt1 and Fmt2), thereby preparing them in 30 libraries. Using the RNA-seq reads, we performed preprocessing that went through three steps such as alignment, filtering, and normalization for the two datasets (*i.e.*, wild–Fmt1 & wild–Fmt2) as previously explained in the previous studies: i) Sequence aligning of the reads to the reference genome of *F. verticillioides* strain 7600 [40] obtained from the BROAD Institute (http://www.broadinstitute.org) was dealt with in Bowtie 2 default mode [34]. The aligned reads were subsequently processed with another NGS (next-generation sequencing) analysis tool called Samtools [37] to prepare read counts of all the *F. verticillioides* genes for three types such as wild, Fmt1, and Fmt2. ii) Gene filtering was performed for quality control by removing insignificantly expressed genes. During this filtering process, genes expressed more than half of the total replicates and also genes expressed at least 70% of one type (either wild or the mutant) only were retained. Hence, not only significantly expressed genes, but also highly differentially expressed genes only expressed in one of the strains (i.e., wild type vs. Fmt1/Fmt2) were remained for the downstream analysis. We kept 82.3% *F. verticillioides* genes (11,648 genes) for wild–Fmt1 dataset and 82.8% *F. verticillioides* genes (11,722 genes) for wild–Fmt2 dataset for the following analysis iii) Two steps of normalization process were completed for reliable network-based analysis. We first normalized all the read counts by the corresponding gene length, then subsequently normalized the relative expression levels again using the representative *F. verticillioides* housekeeping genes maintaining relatively constant expression levels as criterion. The mean expression level of beta-tubulin genes, typical housekeeping genes of *F. verticillioides*, such as FVEG_04081 and FVEG_05512 were applied to normalize all the *F. verticillioides* gene expression levels for both datasets. Table 5.1 provides general statistics of the RNA-seq datasets. This table shows the influence of *F. verticillioides* pathogenicity as well as the difference between the Fmt1 and the Fmt2 in the time-course data (*i.e.*, 5 dpi and 7 dpi)

Table 5.1: General statistics of RNA-seq datasets

	Wild type		Mutant 1		Mutant 2	
	5dpi	7dpi	5dpi	7dpi	5dpi	7dpi
Type of run	single					
Read length	100 (bp)					
Mean # of reads aligned	2,907,836	4,287,201	3,556,574	4,106,461	3,008,801	3,211,987
Median depth of coverage	205.3	302.7	251.1	290	212.5	226.8

5.1.2 Identification of potential subnetwork modules & key pathogenicity gene

Basic procedures for identifying potential subnetwork modules associated with the *F. verticillioides* pathogenicity based on the predicted co-expression networks are described in the previous studies. Here, we predicted not only potential pathogenicity-associated subnetwork modules, but also key potential pathogenicity genes in the detected modules simultaneously. Subsequently, we could also identify more robust potential subnetwork modules than the previous works using the key pathogenicity genes as criterion while we predicted the key genes.

5.1.2.1 Co-expression networks & subnetwork modules

We first constructed the co-expression networks for wild–Fmt1 as well as wild–Fmt2 based on their preprocessed gene expression data through partial correlation evaluating association strength of gene-pairs. Certain thresholds of partial correlation were used to predict gene interactions forming the co-expression networks, hence significant gene interactions with relatively large partial correlation comprised the co-expression networks. We applied four distinct thresholds, thereby generating four different co-expression networks for both wild–Fmt1 and wild–Fmt2 as shown in Table 5.2 & 5.3. The four thresholds have led to the co-expression networks containing significant gene-gene interactions as well as being properly assigned to different sizes for searching subnetwork modules. Also the various thresholds let the dependency of the proposed analysis on a certain threshold level less. Based on the co-expression networks, we started searching subnetwork modules for each datasets (*i.e.*, wild vs. Fmt1 & wild vs. Fmt2). The detailed steps for searching pathogenicity-associated subnetwork modules are illustrated in the previous studies. We first searched seed genes, a first gene member of a subnetwork, by finding the most significantly differentially expressed genes (*i.e.*, top 1%) in the two strains (*i.e.*, wild vs. mutant) for

Table 5.2: Co-expression networks of the maize candidates (wild-type vs. Fmt1).

Threshold (cut-off partial correlation)	Wild type		Fmt1	
	# of genes	# of interactions	# of genes	# of interactions
0.989	9,755	245,656	7967	77,935
0.985	10,365	444,855	9,371	147,042
0.981	10,715	685,290	10,181	235,580
0.977	10,953	957,324	10,734	340,538

Table 5.3: Co-expression networks of the maize candidates (wild-type vs. Fmt2).

Threshold (cut-off partial correlation)	Wild type		Fmt2	
	# of genes	# of interactions	# of genes	# of interactions
0.989	10,418	297,099	8,153	89,471
0.985	10,828	447,917	9,526	150,473
0.981	11,001	739,748	10,321	239,428
0.977	11,472	964,024	11,089	367,138

both datasets. From each seed gene, we extended it to every neighboring gene and evaluated probabilistic activity levels of them. Then, we allowed a subnetwork with the optimal activity level, and also two more suboptimal subnetworks with specific conditions: i) the discriminative power increase (measured by t-test statistics) of the probabilistic activities of an extended subnetwork by adding a connected gene should exceed at least 10%, ii) the discriminative power difference of subnetwork activities between the optimum and the sub-optimums should be less than 2% at each extension. For each subnetwork, we kept appending each neighboring gene into the subnetwork and computed their subnetwork activities to determine whether new member genes can join to the subnetworks. We continuously searched the candidate subnetwork modules by iteratively adding a neighboring gene using this computationally efficient branch-out technique with the certain conditions until the stopping criterion satisfied.

While branching out the subnetwork modules, we probabilistically inferred differentiated activity levels of the extended modules between the two strains (wild vs. mutant) to identify potential subnetwork modules. In this activity level computation, we adopted a probabilistic inference strategy proposed in [58]. As explained in the previous studies, we predicted the activity level of a subnetwork module by accumulating probabilistically quantified values of the member genes. Suppose we have $\mathcal{G} = \{g_1, g_2, \cdots, g_n\}$, a set of genes in a subnetwork module of interest, and $\mathbf{x} = \{x^1, x^2, \cdots, x^n\}$, expression levels of the given genes. The subnetwork module activity level can be given by

$$\eta(\mathbf{x}) = \sum_{k=1}^{n} \alpha^k(x^k), \tag{5.1}$$

where $\alpha^k(x^k)$ is the log likelihood ratio (LLR) between the two conditions (*i.e.*, wild type vs. the mutant) defined as follows

$$\alpha^k(x^k) = log \left[\frac{y_1^k(x^k)}{y_2^k(x^k)} \right]. \tag{5.2}$$

where $y_1^k(x)$ and $y_2^k(x)$ are the conditional probability density function (PDF) of the expression level of gene g_k in wild type and the mutant, respectively. With this probabilistic quantification, we can compute the activity level of the given subnetwork module and further evaluate the discriminative power of the module between the two conditions based on t-test statistics as follows

$$t(\mathcal{G}) = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{5.3}$$

where μ_1 and s_1^2 are the mean and the variance of the subnetwork activity level in one condition, and μ_2 and s_2^2 are the mean and the variance of the subnetwork activity level in the other condition. n_1 and n_2 are the number of replicates in the respective conditions.

5.1.2.2 Key pathogenicity genes & robust subnetwork modules

While extending the subnetworks, we simultaneously searched specific *F. verticillioides* pathogenicity genes in the candidate subnetwork modules. In order to identify pathogenicity genes playing an important role in the predicted subnetwork modules, we analytically investigated the detected modules with the six analysis approaches as shown in Figure 5.1 (B): i) we analyzed whether each gene in a predicted subnetwork module is impactful based on the probabilistically inferred activity. For each predicted subnetwork module, we computed the discriminative power of the probabilistic activity level by removing each gene from the subnetwork and compared the differences from each other. We subsequently selected genes whose deletion lead to relatively higher difference in the comparison as candidate pathogenicity genes; ii) we investigated whether each gene is relatively differentially correlated with the neighboring genes in the detected subnetwork module between the two conditions. Pearson's correlation and Spearman rank correlation were applied for this analysis. Using the two correlation methods, we compared both correlation coefficients of each interaction in the co-expression networks based on wild type and the mutant. We then chose genes whose interactions in the subnetwork were more significantly differentially correlated between the two conditions as candidate pathogenicity genes; iii) we counted number of interactions of genes in the detected subnetwork module and picked genes relatively more connected to other member genes as candidate pathogenicity genes since it is possible that genes with relatively more interactions to other genes could have a stronger effect on the subnetwork; iv) we examined the expression pattern alteration of the genes in the subnetwork module between the two conditions (wild vs. mutant) and selected genes whose expression levels were significantly higher in wild than in the mutant as candidate pathogenicity genes. It is reasonable to assume that expression patterns of potential pathogenicity genes from the RNA-seq data might follow that of *FSR1* since they are downstream processes under the *FSR1*; v) Using the Pathogen Host Interaction Database (PHI-base: http://www.phi-base.org), we considered *F. verticillioides* genes whose orthologous are known pathogenicity genes in other fungi as candidate pathogenicity genes. In this step, we utilized seven fungal species: *F. graminearum* (FG), *F. oxysporum* (FO), *Aspergillus fumigatus* (AF), *Botrytis cinerea* (BC1G), *Magnaporthe grisea* (MGG), *Ustilago maydis* (UM), and *Cryptococcus neoformans* (CNAG); vi) We investigated whether each gene in the detected subnetwork module was annotated by significant GO term with other member genes. We took into account the genes that can be involved in significant GO terms to be candidate pathogenicity genes. Once we identified a potential *F. verticillioides* pathogenicity gene from its predicted subnetwork module through those systematic analyses, to finetune the detected module was performed for robustness of the approach as well as more reliable subnetwork modules. As described in the previous section, we first applied 10% discriminative power increase as the criterion to overcome at each extension while permitting a subnetwork to be extended up to three branches. However, after identifying a potential pathogenicity gene meeting the six requirements from its predicted subnetwork module, we started to increase the criterion percentage when branching out as long as the potential pathogenicity gene satisfies the abovementioned six qualifications. The criterion percentage escalation was continued until either the pathogenicity gene was no longer meets all the conditions or the predicted subnetwork module became smaller down to the minimum size. The minimum number of genes in a module was assigned to be seven.

5.2 Results

5.2.1 Identification of subnetwork modules

As shown in Figure 5.2, 5.3 and 5.4, we identified potential subnetwork modules through our proposed network-based comparative analysis using two different RNA-seq datasets (one is maize inoculated with two different strains of *F. verticillioides* and the other is two different phenotypes of *F. verticillioides*). In order to identify these potential functional modules in Figure 5.2-5.4, we first constructed co-expression networks, and further selected the seed genes, top 1% differentially expressed genes at the last time point. From each seed gene, we iteratively extended the subnetwork module by adding one neighboring gene, and allowed the subnetwork modules whose enhanced discriminative power (computed by t-test statistics score) was higher than 10% to be extended. Amongst the subnetwork modules permitted to be extended, we first selected the subnetwork module with the top discriminative power, and also selected two additional suboptimal modules whose discriminative power difference from the optimal was less than 2%. We continuously repeated this extending process by branching out up to three extension at each step until the subnetworks were not able to increase the discriminative power by 10%. For both datasets, the entire process of extending the subnetwork modules by our analysis approach was reiterated for all the seed genes at all the co-expression networks. After collecting all extended subnetwork modules, we analytically investigated whether each gene in its predicted subnetwork module satisfied several conditions; i) highly impactful in the probabilistic manner, ii) relatively differentially correlated between the two strains (wild type vs mutant), iii) relatively more connected in the given module, iv) relatively highly expressed in wild-type, v) orthologous to known pathogenic genes in other fungi, and vi) annotated to significant GO terms with other member genes. Through these analyses, we identified key potential genes associated with the *F. verticillioides* pathogenicity.

5.2.2 Results for maize inoculated with two strains - wild type vs. the mutant - of F. verticillioides

Four potential *F. verticillioides* pathogenic subnetwork modules identified by the proposed network-based comparative analysis contained key *F. verticillioides* pathogenicity genes whose annotated GO terms were representative terms with other member genes. In order to validate, we utilized *p*-values of Benjamini-Hochberg false discovery rate (FDR) method [5] provided by g:Profiler (http://biit.cs.ut.ee/gprofiler /index.cgi) [49] to distinguish the most significant GO terms to the identified subnetwork modules. The four potential subnetwork modules are shown in Figure 5.2. Five genes such as FVEG_07930, FVEG_000890, FVEG_08174, FVEG_011886, and FVEG_00594 of Figure 5.2 module-A were annotated by a significant GO term GO:0051234. For this GO:0051234, whose GO term is "establishment of localization", is defined as "the directed movement of a cell, substance or cellular entity, such as a protein complex or organelle, to a specific location". Benjamini-Hochberg FDR *p*-value of the most significant GO term, GO:0051234, for the Figure 5.2 module-A was 0.05. In this subnetwork module, the identified key pathogenicity gene was FVEG_00594. The potential pathogenicity gene, FVEG_00594, whose orthologous gene in *F. graminearum* (FG) has phenotype as "reduced virulence" also showed high impact on the module in the probabilistic manner, significantly differentially correlated pattern between the two conditions, relatively higher connection with other member genes, and relatively significant expression in wild-type. For Fig-

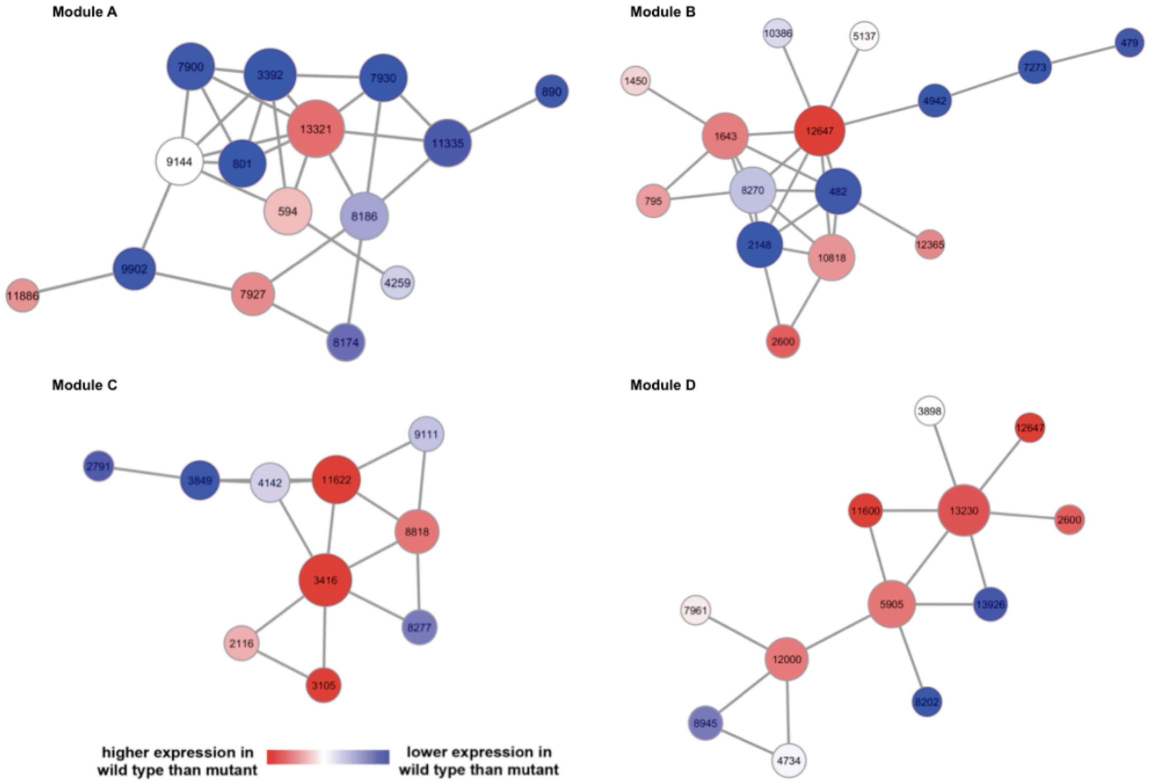


Figure 5.2: Four potential *F. verticillioides* subnetwork modules.

ure 5.2 module-B, four genes such as FVEG_05137, FVEG_00482, FVEG_10818, and FVEG_12365 were annotated by a significant GO term GO:0043168. This GO term is "anion binding" and the definition is "interacting selectively and non-covalently with anions, charged atoms or groups of atoms with a net negative charge"; the *p*-value of Benjamini-Hochberg FDR for this term was 0.05. From this subnetwork module, FVEG_10818 whose orthologous gene in *Magnaporthe grisea* (MGG) has phenotype as "loss of pathogenicity" was identified to be the potential key pathogenicity gene by satisfying the qualifications. For Figure 5.2 module-C, eight genes of the module such as FVEG_02791, FVEG_03105, FVEG_03849, FVEG_08277, FVEG_09111, FVEG_11622, FVEG_02116 and FVEG_04142 were annotated by a significant GO term GO:0044444. This term GO:0044444 is for “cytoplasmic part” and had a *p*-value (for Benjamini-Hochberg FDR) of 0.0077. The GO term is described as “any constituent part of the cytoplasm, all of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures”. In this subnetwork module, the identified key pathogenicity gene was FVEG_11622 whose orthologous gene in *Botrytis cinerea* (BC1G) has phonotype as “reduced virulence” For Figure 5.2 module-D, five genes such as FVEG_13230, FVEG_07961, FVEG_08945, FVEG_03898, and FVEG_05905 were annotated by a significant GO term GO:0016787. This GO term is for “hydrolase activity” and is defined as “catalysis of the hydrolysis of various bonds, e.g. C-O, C-N, C-C, phosphoric anhydride bonds, etc. hydrolase is the systematic name for any enzyme of EC class”. The *p*-value of Benjamini-Hochberg FDR for this GO term, GO:0016787, was 0.05. From this subnetwork module, FVEG_13230 whose orthologous gene in *Magnaporthe grisea* (MGG) has phenotype as “reduced virulence” was identified to be the potential key pathogenicity gene. These four potential subnetwork modules demonstrated that the most significant GO term associated with the member genes including each potential key pathogenicity gene. The potential pathogenicity genes were identified by satisfying the requirements (*i.e.*, being distinguished, pathogenicity-associated, and impactful) among the member genes.

5.2.3 Results for two different F. verticillioides *mutants*

For investigation on dataset with two *F. verticillioides* mutants by comparing the two phenotypes with the network-based comparative analysis, two potential genes such as FVEG_00035 and FVEG_07056 were identified to be pathogenicity-associated in both phenotypes. Phenotypes of orthologs of the two genes (FVEG_00035 and FVEG_07056) were “reduced virulence” and “loss of pathogenicity”, respectively.

Figure 5.3 shows three potential subnetwork modules containing either FVEG_00035 or FVEG_07056 for one dataset (wild vs. Fmt1), and several member genes of the three modules were annotated by significant GO terms. For Figure 5.3 module-A, four genes such as FVEG_07804, FVEG_10460, FVEG_00035, and FVEG_00631 were annotated by a significant GO term GO:1901564. For this GO:1901564, whose GO term is "organonitrogen compound metabolic process", is defined as "the chemical reactions and pathways involving organonitrogen compound". Benjamini-Hochberg FDR *p*-value of the most significant GO term, GO:1901564, for the Figure 5.3 module-A was 1.25e-04. For Figure 5.3 module-B, four genes such as FVEG_06177, FVEG_00991, FVEG_01972, and FVEG_06824 were annotated by a significant GO term GO:0032553. This GO term is "ribonucleotide binding" and the definition is "interacting selectively and non-covalently with a ribonucleotide, any compound

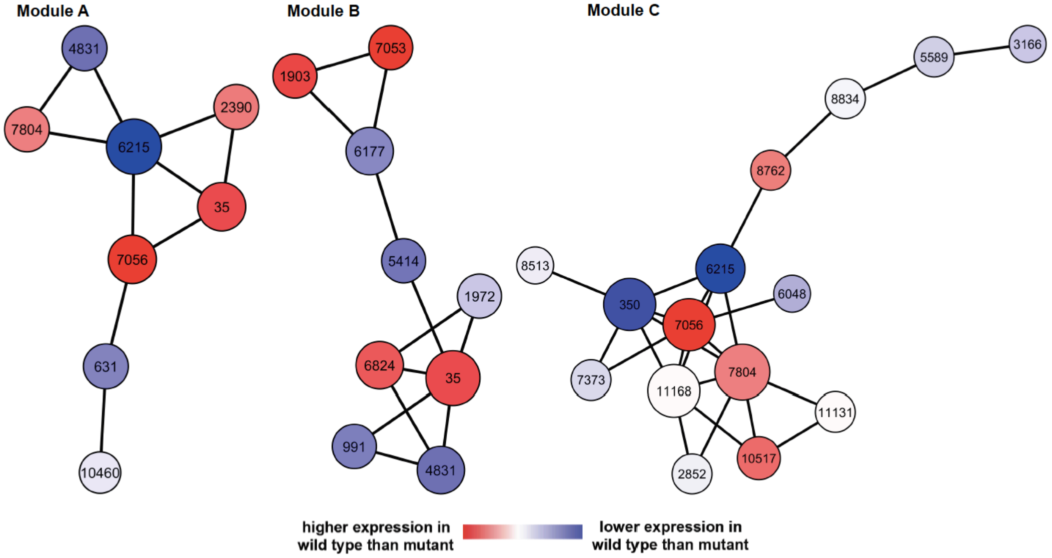


Figure 5.3: Three potential *F. verticillioides* subnetwork modules (wild–Fmt1)

consisting of a ribonucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on the ribose moiety"; the *p*-value of Benjamini-Hochberg FDR for this term was 4.49e-04. For Figure 5.3 module-C, four genes of the module such as FVEG_03166, FVEG_10517, FVEG_05589 and FVEG_08762 were annotated by a significant GO term GO:0050662. This term GO:0050662 is for "coenzyme binding" and had a *p*-value (for Benjamini-Hochberg FDR) of 1.84e-05. The GO term is described as "interacting selectively and non-covalently with a coenzyme, any of various nonprotein organic cofactors that are required, in addition to an enzyme and a substrate, for an enzymatic reaction to proceed".

Three identified potential subnetwork modules in Figure 5.4 including either FVEG_00035 or FVEG_07056 for the other dataset (wild vs. Fmt2) also contained several member genes annotated by significant GO terms. Figure 5.4 module-A, five genes such as FVEG_04540, FVEG_07076, FVEG_07532, FVEG_00035, and FVEG_04805 were annotated by a significant GO term GO:0044763. This GO term is for "single-organism cellular process" and is defined as "Any process that is carried out at the cellular level, occurring within a single organism". The *p*-value of Benjamini-Hochberg FDR for this GO term, GO:0044763, was 0.05. For Figure 5.4 module-B, three genes such as FVEG_05281, FVEG_04805, and FVEG_07056 were annotated by a significant GO term GO:0003676. For this GO:0003676, whose GO term is "nucleic acid binding", is defined as "interacting selectively and non-covalently with any nucleic acid". Benjamini-Hochberg FDR *p*-value of the most significant GO term, GO:0003676, for the Figure 5.4 module-B was 9.16e-03. For Figure 5.4 module-C, six genes such as FVEG_08056, FVEG_13129, FVEG_05927, FVEG_04805, FVEG_02944, and FVEG_03567 were annotated by a significant GO term GO:0000166. This GO term is "nucleotide binding" and the definition is "interacting selectively and non-covalently with a nucleotide, any compound consisting

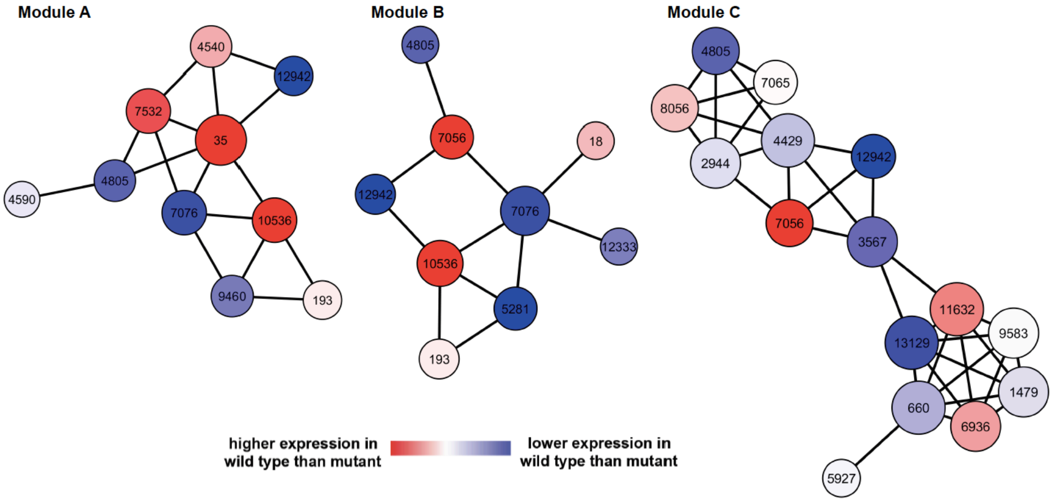


Figure 5.4: Three potential *F. verticillioides* subnetwork modules (wild–Fmt2)

of a nucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on the ribose or deoxyribose"; the *p*-value of Benjamini-Hochberg FDR for this term was 5.12e-05. It is possible to consider the three detected subnetwork modules for each dataset can be viewed as one whole module since they share a few genes each other.

6. CONCLUSION

In this Ph.D study, we performed network-based comparative analysis for *F. verticillioides* pathogenicity as well as host-pathogen (*F. verticillioides* vs. maize) interactions using RNA-seq data (one comes from maize inbred B73 inoculated with *F. verticillioides* wild type and *fsr1* mutant, and the other comes from two different *F. verticillioides* mutants). For the 1st study identifying potential functional subnetwork modules, we searched the inferred co-expression networks to identify the functional modules by the computationally efficient branch-out technique. For the 2nd study identifying potential genetic subnetwork modules associated with maize defense response, we first searched candidate maize genes possibly involved in maize defense mechanism and identified the potential genetic modules by adopting the previously proposed analysis approach. For the 3rd study identifying potential key pathogenicity genes of *F. verticillioides*, we further analyzed inside the potential subnetwork modules detected by the 1st study approach. In this subnetwork analysis, we applied six additional analytical methods for the predicted subnetwork modules. Based on our analysis approaches, we identified four potential pathogenic subnetwork modules that were structurally cohesive with functionally coherent member genes. We also identified four maize potential subnetwork modules directly (or indirectly) associated with the defense response against the *F. verticillioides* pathogenicity. In addition, we identified potential key *F. verticillioides* pathogenicity genes that were coordinated with other member genes and had strong impact on other genes and were also pathogenicity-associated for both datasets. Consequently, our proposed analysis approaches could lead to an improved understanding of the *F. verticillioides* pathogenicity and maize-*F. verticillioides* interaction, particularly at the transcriptome level. Our approaches can also be used to identify potential functional modules or key genes associated with the pathogenicity in other fungal species as well as host genetic modules involved in the defense response in other host plants based on their plant-pathogen interactions.

REFERENCES

[1] Matthew C Asters, W Paul Williams, Andy D Perkins, J Erik Mylroie, Gary L Windham, and Xueyan Shan. Relating significance and relations of differentially expressed genes in response to *Aspergillus flavus* infection in maize. *Scientific Reports*, 4(4815), 2014.

[2] CW Bacon and DM Hinton. Symptomless endophytic colonization of maize by *Fusarium moniliforme*. *Canadian Journal of Botany*, 74(8):1195–1202, 1996.

[3] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16(3):428–435, 2006.

[4] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[6] Andrea H Bild, Guang Yao, Jeffrey T Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M Lancaster, Andrew Berchuck, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–357, 2006.

[7] BH Bluhm, H Kim, RAE Butchko, and CP Woloshuk. Involvement of *ZFR1* of *Fusarium verticillioides* in kernel colonization and the regulation of *FST1*, a putative sugar transporter gene required for fumonisin biosynthesis on maize kernels. *Molecular plant pathology*, 9(2):203–211, 2008.

[8] Daren W Brown, Robert AE Butchko, Mark Busman, and Robert H Proctor. The *Fusarium verticillioides* FUM gene cluster encodes a Zn (II) 2Cys6 protein that affects FUM gene expression and fumonisin production. *Eukaryotic Cell*, 6(7):1210–1218, 2007.

[9] Daren W. Brown and Robert H. Proctor. *Fusarium: genomics, molecular and cellular biology*. Horizon Scientific Press, 2013.

[10] Valeria A Campos-Bermudez, Carolina M Fauguel, Marcos A Tronconi, Paula Casati, Daniel A Presello, and Carlos S Andreo. Transcriptional and metabolic changes associated to the infection by *Fusarium verticillioides* in maize inbreds with contrasting ear rot resistance. *PLoS One*, 8(4):e61580, 2013.

[11] J Chelkowski. *Fusarium: Mycotoxins, Taxonomy, Pathogenicity*. Elsevier, 2014.

[12] Stephen T Chisholm, Gitta Coaker, Brad Day, and Brian J Staskawicz. Host-microbe interactions: shaping the evolution of the plant immune response. *Cell*, 124(4):803–814, 2006.

[13] Yoon-E Choi and Won-Bo Shim. Identification of genes associated with fumonisin biosynthesis in *Fusarium verticillioides* via proteomics and quantitative real-time pcr. *Journal of microbiology and biotechnology*, 18(4):648–657, 2008.

[14] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(140), 2007.

[15] Peter N Dodds and John P Rathjen. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics*, 11(8):539–548, 2010.

[16] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

[17] Helen Fones, Calum AR Davis, Arantza Rico, Fang Fang, J Andrew C Smith, and Gail M Preston. Metal hyperaccumulation armors plants against disease. *PLoS Pathogens*, 6(9):e1001093, 2010.

[18] Jing Fu, Hongbo Liu, Yu Li, Huihui Yu, Xianghua Li, Jinghua Xiao, and Shiping Wang. Manipulating broad-spectrum disease resistance by suppressing pathogen-induced auxin accumulation in rice. *Plant Physiology*, 155(1):589–602, 2011.

[19] Liane R Gale, JD Bryant, Sarah Calvo, Henriette Giese, Talma Katan, Kerry O'Donnell, Haruhisa Suga, Masatoki Taga, Thomas R Usgaard, Todd J Ward, et al. Chromosome complement of the fungal plant pathogen *Fusarium graminearum* based on genetic and physical mapping and cytological observations. *Genetics*, 171(3):985–1001, 2005.

[20] Wolfgang Gerlach, Helgard Nirenberg, et al. *The genus Fusarium–a pictorial atlas.* Mitteilungen aus der Biologischen Bundesanstalt fur Land-und Forstwirtschaft Berlin-Dahlem, 1982.

[21] Fei He, Yan Zhang, Hao Chen, Ziding Zhang, and You-Liang Peng. The prediction of protein-protein interaction networks in rice blast fungus. *BMC Genomics*, 9(1):519, 2008.

[22] Alfred Hero and Bala Rajaratnam. Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory*, 58(9):6064–6078, 2012.

[23] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinfor-*

matics, 18(suppl 1):S233–S240, 2002.

[24] Junhyun Jeon, Sook-Young Park, Myoung-Hwan Chi, Jaehyuk Choi, Jongsun Park, Hee-Sool Rho, Soonok Kim, Jaeduk Goh, Sungyong Yoo, Jinhee Choi, et al. Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nature Genetics*, 39(4):561–565, 2007.

[25] James E Jurgenson, Kurt A Zeller, and John F Leslie. Expanded genetic map of *Gibberella moniliformis* (*Fusarium verticillioides*). *Applied and Environmental Microbiology*, 68(4):1972–1979, 2002.

[26] JE Jurgenson, RL Bowden, KA Zeller, JF Leslie, NJ Alexander, and RD Plattner. A genetic map of *Gibberella zeae* (*Fusarium graminearum*). *Genetics*, 160(4):1451–1460, 2002.

[27] Rowena Y Kelley, W Paul Williams, J Erik Mylroie, Deborah L Boykin, Jonathan W Harper, Gary L Windham, Arunkanth Ankala, and Xueyan Shan. Identification of maize genes associated with host plant resistance or susceptibility to *Aspergillus flavus* infection and aflatoxin accumulation. *PLos One*, 7(5):e36892, 2012.

[28] Michael J Kershaw and Nicholas J Talbot. Genome-wide functional analysis reveals that infection-associated fungal autophagy is necessary for rice blast disease. *Proceedings of the National Academy of Sciences*, 106(37):15967–15972, 2009.

[29] Navadon Khunlertgit and Byung-Jun Yoon. Identification of robust pathway markers for cancer through rank-based pathway activity inference. *Advances in Bioinformatics*, 2013, 2013.

[30] Navadon Khunlertgit and Byung-Jun Yoon. Simultaneous identification of robust synergistic subnetwork markers for effective cancer prognosis. *EURASIP Journal on Bioinformatics and Systems Biology*, 2014(1):1–10, 2014.

[31] Hun Kim and Charles P Woloshuk. Functional characterization of *FST1* in *Fusarium verticillioides* during colonization of maize kernels. *Molecular Plant-Microbe Interactions*, 24(1):18–24, 2011.

[32] Mansuck Kim, Huan Zhang, Charles Woloshuk, Won-Bo Shim, and Byung-Jun Yoon. Computational identification of genetic subnetwork modules associated with maize defense response to *Fusarium verticillioides*. *BMC Bioinformatics*, 16(Suppl 13):S12, 2015.

[33] Mansuck Kim, Huan Zhang, Charles Woloshuk, Won-Bo Shim, and Byung-Jun Yoon. Computational prediction of pathogenic network modules in *Fusarium verticillioides*. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Web, 2015.

[34] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[35] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11):e1000217, 2008.

[36] John F Leslie, Brett A Summerell, and Suzanne Bullock. *The Fusarium laboratory manual*, volume 2. Wiley Online Library, 2006.

[37] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[38] Xiaoping Liu, Wei-Hua Tang, Xing-Ming Zhao, and Luonan Chen. A network approach to predict pathogenic genes for *Fusarium graminearum*. *PLoS One*, 5(10):e13021, 2010.

[39] Artem Lysenko, Martin Urban, Laura Bennett, Sophia Tsoka, Elzbieta Janowska-Sejda, Chris J Rawlings, Kim E Hammond-Kosack, and Mansoor Saqi. Network-based data integration for selecting candidate virulence associated proteins in the cereal infecting fungus *Fusarium graminearum*. *PLoS One*, 8(7):e67926, 2013.

[40] Li-Jun Ma, H Charlotte Van Der Does, Katherine A Borkovich, Jeffrey J Coleman, Marie-Josée Daboussi, Antonio Di Pietro, Marie Dufresne, Michael Freitag, Manfred Grabherr, Bernard Henrissat, et al. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, 464(7287):367–373, 2010.

[41] Antonio N Moretti. Taxonomy of fusarium genus: a continuous fight between lumpers and splitters. *Zbornik Matice srpske za prirodne nauke*, (117):7–13, 2009.

[42] Gary P Munkvold. Cultural and genetic approaches to managing mycotoxins in maize. *Annual Review of Phytopathology*, 41(1):99–116, 2003.

[43] Gary P Munkvold, Richard L Hellmich, and WB Showers. Reduced fusarium ear rot and symptomless infection in kernels of maize genetically engineered for european corn borer resistance. *Phytopathology*, 87(10):1071–1077, 1997.

[44] GP Munkvold, DC McGee, and WM Carlton. Importance of different pathways for maize kernel infection by *Fusarium moniliforme*. *Phytopathology*, 87(2):209–217, 1997.

[45] K Myung, NC Zitomer, M Duvall, AE Glenn, RT Riley, and AM Calvo. The conserved global regulator *VeA* is necessary for symptom production and mycotoxin synthesis in maize seedlings by *Fusarium verticillioides. Plant Pathology*, 61(1):152–160, 2012.

[46] Paul E Nelson, M Cecilia Dignani, and Elias J Anaissie. Taxonomy, biology, and clinical aspects of fusarium species. *Clinical Microbiology Reviews*, 7(4):479, 1994.

[47] Carlos S Ortiz and Won-Bo Shim. The role of mads-box transcription factors in secondary metabolism and sexual development in the maize pathogen *Fusarium verticillioides. Microbiology*, 159(11):2259–2268, 2013.

[48] Adam James Reid and Matthew Berriman. Genes involved in host–parasite interactions can be revealed by their correlated expression. *Nucleic Acids Research*, 41(3):1508–1518, 2012.

[49] Jüri Reimand, Tambet Arak, and Jaak Vilo. g: Profiler - a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, 39(suppl 2):307–315, 2011.

[50] John B Ridenour and Burton H Bluhm. The hap complex in *Fusarium verticillioides* is a key regulator of growth, morphogenesis, secondary metabolism, and pathogenesis. *Fungal Genetics and Biology*, 69:52–64, 2014.

[51] Patrick S Schnable, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A Graves, et al. The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, 2009.

[52] R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24:427–433, 2006.

[53] Patrick R Shea, Kimmo Virtaneva, John J Kupko, Stephen F Porcella, William T Barry, Fred A Wright, Scott D Kobayashi, Aaron Carmody, Robin M Ireland, Daniel E Sturdevant, et al. Interactome analysis of longitudinal pharyngeal infection of *cynomolgus macaques* by group a *Streptococcus. Proceedings of the National Academy of Sciences*, 107(10):4693–4698, 2010.

[54] Won-Bo Shim, Uma Shankar Sagaram, Yoon-E Choi, Jinny So, Heather H Wilkinson, and Yin-Won Lee. *FSR1* is essential for virulence and female fertility in *Fusarium verticillioides* and *F. graminearum. Molecular plant-microbe interactions*, 19(7):725–733, 2006.

[55] Won-Bo Shim and Charles P Woloshuk. Regulation of fumonisin b1 biosynthesis and conidiation in *Fusarium verticillioides* by a cyclin-like (c-type) gene, *FCC1. Applied and Environmental Microbiology*, 67(4):1607–1612, 2001.

[56] Joon-Hee Shin, Jung-Eun Kim, Martha Malapi-Wight, Yoon-E Choi, Brian D Shaw, and Won-Bo Shim. Protein phosphatase 2a regulatory subunits perform distinct functional roles in the maize pathogen *Fusarium verticillioides. Molecular Plant Pathology*, 14(5):518–529, 2013.

[57] Hokyoung Son, Young-Su Seo, Kyunghun Min, Ae Ran Park, Jungkwan Lee, Jian-Ming Jin, Yang Lin, Peijian Cao, Sae-Yeon Hong, Eun-Kyung Kim, et al. A phenome-based functional analysis of transcription factors in the cereal head blight fungus, *Fusarium graminearum. PLoS Pathogens*, 7(10):e1002310, 2011.

[58] Junjie Su, Byung-Jun Yoon, and Edward R Dougherty. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLos One*, 4(12):e8161, 2009.

[59] Junjie Su, Byung-Jun Yoon, and Edward R Dougherty. Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC bioinformatics*, 11(Suppl 6):S8, 2010.

[60] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[61] Brett A Summerell, Matthew H Laurence, Edward CY Liew, and John F Leslie. Biogeography and phylogeography of Fusarium: a review. *Fungal Diversity*, 44(1):3–13, 2010.

[62] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

[63] Chenfang Wang, Shijie Zhang, Rui Hou, Zhongtao Zhao, Qian Zheng, Qijun Xu, Dawei Zheng, Guanghui Wang, Huiquan Liu, Xuli Gao, et al. Functional analysis of the kinome of the wheat scab fungus *Fusarium graminearum*. *PLoS Pathogens*, 7(12):e1002460, 2011.

[64] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[65] Rainer Winnenburg, Thomas K Baldwin, Martin Urban, Chris Rawlings, Jacob Köhler, and Kim E Hammond-Kosack. Phi-base: a new database for pathogen

host interactions. *Nucleic Acids Research*, 34(suppl 1):D459–D464, 2006.

[66] Jin-rong Xu and John F Leslie. A genetic map of *Gibberella fujikuroi* mating population a (*Fusarium moniliforme*). *Genetics*, 143(1):175–189, 1996.

[67] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[68] Byung-Jun Yoon, Xiaoning Qian, and Sayed Mohammad Ebrahim Sahraeian. Comparative analysis of biological networks: Hidden markov model and markov chain-based approach. *Signal Processing Magazine, IEEE*, 29(1):22–34, 2012.