Digitization of trait representation in microarray data analysis of wheat infected by fusarium graminearum

Pan, Youlian; Ouellet, Thérèse; Phan, Sieu; Tchagang, Alain; Fauteux, François; Tulpan, Dan

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/CIBCB.2015.7300304
Conference on Computational Intelligence in Bioinformatics and Computational Biology, pp. 1-9, 2015-08-15

Digitization of Trait Representation in Microarray Data Analysis of Wheat Infected by *Fusarium graminearum*

Youlian Pan[1]*, Thérèse Ouellet[2]*, Sieu Phan[1], Alain Tchagang[1], François Fauteux[1], Dan Tulpan[3]

[1]Information and Communications Technologies, NRC, 1200 Montreal Road, Ottawa, ON, K1A0R6 Canada
[2]Eastern Cereal and Oilseed Research Centre, AAFC, 960 Carling Ave, Ottawa, ON, K1A0C6 Canada
[3]Information and Communications Technologies, NRC, 100 des Aboiteaux Street, Moncton, NB, E1A7R1 Canada

Fusarium head blight (FHB) limits wheat yield and compromises grain quality. We investigated differentially expressed genes after FHB challenge. FHB-susceptible and -resistant common wheat (*Triticum aestivum*) cultivars were challenged with the toxigenic fungus *Fusarium graminearum* and gene expression was analyzed using 61K Affymetrix wheat microarrays. We digitized trait specificity in the susceptible and resistant lines with and without the infection in order to facilitate subsequent data mining. We discovered various patterns of differential gene expression between susceptible and resistant lines in response to the infection. We performed association network analysis among genes in clusters significantly correlated with one or more quantitative trait loci known to contribute to *Fusarium* resistance. We found 11 interconnected hub genes responsive to FHB infection and significantly correlated with wheat resistance to FHB, among which two are predicted to encode a polygalacturonase-inhibiting protein (PGIP1).

Keywords— microarray, digitization, Fusarium head blight, wheat, disease resistance

I. Introduction

Wheat is the crop with the largest production area, and the second in importance by production volume, supplying 19% of the total human calories [2]. The yield of wheat is severely limited by diseases caused by microbial pathogens. The total global potential and actual estimates on wheat crop losses to pathogens are 16% and 10%, respectively [19]. Besides, wheat quality is compromised by pathogen-derived toxins that are hazardous to animal and humans. One of the prevalent wheat diseases, Fusarium head blight (FHB), is caused by ascomycetous fungi of the genus Fusarium. The most common *Fusarium* species in North America is *F. graminearum* [3].

Fusarium graminearum is a filamentous fungus widely distributed on plants and soil. It produces toxic trichothecenes, which are potent inhibitors of peptidyl transferase and inhibit protein synthesis in many eukaryotes, including plants, humans and farm animals. Accordingly, most countries have legislation to protect consumers by setting a limit to the most prevalent *Fusarium* mycotoxins in wheat products. The most prevalent trichothecenes in Canadian wheat is deoxynivalenol (DON), also called vomitoxin. Mycotoxin-contaminated grain is sold at lower prices or is completely rejected. Fusarium head blight is responsible for estimated annual losses of $75M in Manitoba alone [31] and has caused losses of over $3B since 1990 in the USA [28].

Phytopathogenic fungi produce extracellular hydrolytic enzymes that degrade plant cell walls components to facilitate host invasion and pathogen dissemination [14]. Variations in the production of these hydrolytic enzymes may contribute to differences in virulence within and between *Fusarium* species. The action of pectinases result in modifications of the cell wall structure, an increased accessibility of cell wall components for degradation by other enzymes, cell lysis and plant tissue maceration. Polygalacturonase, a pectinase produced by *F. graminearum* and many other fungal pathogens, degrades polygalacturonan, a major component of the plant pectin network, by the hydrolysis of glycosidic bonds between galacturonic acid residues. Mechanisms of wheat resistance to fungal invasion include the inhibition of polygalacturonase to protect its cell wall integrity.

Wheat resistance to FHB is categorized in five types [18]: resistance to initial infection (type I), resistance to spread (type II), resistance to DON accumulation (type III), resistance to kernel infection (type IV), and tolerance (type V). Waldron et al. [33] proposed a similar classification with five types of resistance, in a different order. Types I and II are consistent between these two classifications. These resistance types are inter-related; for example, type II resistance is influenced by DON production, while type I resistance is independent of the mycotoxins and more difficult to ascertain. In this study, we adopted the system in [18].

Several FHB resistant wheat cultivars have been identified and a large number of quantitative trait loci (QTLs) conferring resistance to FHB in wheat have been discovered. Twenty-two chromosomal regions have been identified as contributing consistently to FHB resistance in multiple studies (reviewed in [5]). One of the most effective and best characterized sources of resistance against FHB is the Chinese variety Sumai 3 and its derivatives. These harbor a major QTL for type II resistance (up to 20-25% reduction of disease severity), named Fhb1, that has been mapped to chromosome 3BS as well as a minor QTL (Qfhs.ifa-5A) associated with type I resistance on chromosome 5AS. QTLs in the same chromosomal regions of 3BS and 5AS have been detected in a range of FHB-resistant cultivars, including NuyBay. A minor QTL associated with type II resistance was identified in the Chinese variety Wuhan 1 on chromosome 2DL [29].

Microarray technology has been successfully used in gene expression profiling studies to identify genes and pathways involved in mediating susceptibility or resistance to *Fusarium*

* For correspondence: youlian.pan@nrc.gc.ca, therese.ouellet@agr.gc.ca.
This work was supported by Canadian Wheat Alliance. 978-1-4799-6926-5/15/$31.00

[7],[10],[34]. Numerical representation of experimental designs is commonly used in the statistical analysis of microarray data, generally as a matrix in which rows correspond to arrays and columns to treatments. The popular LIMMA software [26], for example, uses as input a design matrix containing linear model coefficients. In this study, we used a similar representation in which combinations of experimental factors were transformed into numerical vectors and used for subsequent differential expression and correlation analyses. This simple representation facilitated subsequent data analysis, in particular, the construction of mathematical and logical operations on combinations of experimental factors for the selection of differentially expressed genes.

Canadian wheat production is impacted by diseases and pests and changing climatic conditions, and there is a growing need for wheat varieties that are more productive and resilient. To tackle this problem, Agriculture Agri-Food Canada (AAFC), National Research Council Canada (NRC), University of Saskatchewan, and Province of Saskatchewan joined forces and formed the Canadian Wheat Alliance with an 11 year mandate aiming at the production of next generations of wheat varieties resilient under biotic and abiotic stresses. As part of this collective effort, we analyzed two microarray gene expression datasets. In the subsequent sections we describe the datasets, our data mining approach, results, and finally we discuss our findings.

II. The Datasets

In this study, two datasets were analyzed. The first dataset contains 3 spring wheat varieties inoculated with *Fusarium graminearum*: Roblin (very susceptible to FHB, from Canadian source), Wuhan (type II resistant, from Chinese source), and NuyBay (Type III resistant, from Japanese source) [21]. Wheat heads at mid-anthesis were point inoculated in two florets for each spikelet and all fully developed spikelets were inoculated on each head. Inoculated spikelets were sampled 1, 2 and 4 days after inoculation. Two biological replicates were performed. Water injection was performed in the same way in control plants. The second dataset consists of a series of wheat lines, genetically related to cultivars Wuhan and NuyBay, produced by molecular assisted breeding for the presence or absence of specific QTLs. This series include one susceptible line (2-2890) and five resistant lines (2-2614, 2-2618, 2-2598, 2-2543 and 2-2557) [22]. Each of these five resistant lines contains 1, 2, or 3 QTLs among 2DL, 3BS and 5AS, which are known to play a significant role in resistance to *Fusarium*. Inoculation and sampling procedures were similar to that in the first experiment, although in this case samples were taken 4 days after inoculation, and 2 or 3 biological replicates were produced, depending on the line.

The Affymetrix 61K GeneChip Wheat Genome Array (http://www.affymetrix.com/catalog/131517/AFFY/Wheat+Genome+Array) was used in this study. This oligonucleotide microarray contains 61,290 probe sets representing 55,052 transcripts for all 21 pairs of chromosomes in the hexaploid bread wheat genome (*Triticum aestivum*). Contig sequences represented on the chip were obtained from Affymetrix and aligned with wheat genes obtained from Ensemble Plants (version 22, http://plants.ensembl.org/Triticum_aestivum/Info/Index). Contigs were also aligned to 33,326 *Arabidopsis thaliana* coding sequences (CDS) obtained from TAIR (version 10, http://www.arabidopsis.org/). The best BLAST hits were retrieved using OrthoPred (in-house software), using a threshold of 1e-04 on e-values [32]. A total of 44,734 best hits against 27,853 unique wheat genes and 32,836 best hits against 10,409 unique *Arabidopsis* genes were retained after this procedure. The association between Affy probe ID and UniGene from NCBI were updated, which resulted in 45,699 associations mapping to 26,674 unique UniGenes. Ensemble Plants (version 22) gene ontology association with wheat genes was analyzed for gene enrichment using the GOAL software [30].

For the microarray experiments, biological samples were processed in two experimental batches (first and second datasets mentioned above). Arrays were hybridized and washed according to the manufacturer's specifications by the Functional Genomics Platform, McGill University and Genome Quebec Innovation Centre (Montreal, Canada). The raw and normalised data were deposited in NCBI's Gene Expression Omnibus and are part of accession number GSE54556. The probe-level data for the profiling experiments were assembled using the Robust Multichip Analysis (RMA) [12] implemented in the software Acuity 4.0 (Molecular Devices, CA), with quantile normalization and summarization by median polish [19]. RMA normalization produced log2 based output, which was used in subsequent analysis.

III. The Approach

The biological problem was to identify differentially expressed genes between resistant and susceptible lines under *Fusarium* challenge versus control, and to discover marker genes for *Fusarium* resistance traits (types or QTLs) using microarray data. To reduce noise in the data, we selected a subset of genes within the linear range of the sigmoid pattern between mRNA concentration and fluorescence intensity [24], using $Sig_{ratio} \geq 2$ (i.e. max[log2 ratio values] ≥ 2 for differential expression) and $Sig_{signal} \geq 8$ (max[expression values] ≥ 8). Genes satisfying both conditions ($Dif = Sig_{ratio} \bigcap Sig_{signal}$) were considered differentially expressed between wheat lines or between different conditions. After this procedure 2,532 and 2,880 probe sets were retained in the first and second datasets, respectively.

We digitized trait specificity of each wheat line to facilitate subsequent analysis. If a trait existed in a line, we labeled this line "1" for the given trait; otherwise it was labelled "0". For example, in the second dataset, line 2-2614 is *Fusarium* resistant and possesses all three QTLs; the digitization vector {Fg, Q2DL, Q3BS, Q5AS} is {1, 1, 1, 1} with infection by *Fusarium*, and {0, 1, 1, 1} without infection. The time factor (in days) was represented as is. In the first dataset, "FgT" denotes days after *Fusarium* inoculation, and "HT" days after water mock inoculation. The digitization vector for Wuhan 4 days after *Fusarium* treatment {Fg, R2, F3, R, Time, FgT, HT} would thus be {1, 1, 0, 1, 4, 4, 0}.

Trait specificity for the fold change (log2 ratio) data was digitized in a similar way. The ratio introduces additional complexity in trait specificity: trait specificity of both

numerator and denominator had to be considered in the representation. For example, to compare a resistant line with the susceptible line after *Fusarium* infection, three factors must be considered (susceptibility, one or more QTLs related with resistance, and *Fusarium* infection). To simplify the representation in such case, selected traits were excluded from the fold change data using the following mathematical analogy:

$$ax/ay = x/y \qquad (1)$$

For example, when considering differential expression between resistant line 2-2614 and the susceptible line 2-2890 after *Fusarium* infection, both numerator and denominator were Fg positive, and their digitization vector {Fg, Q2DL, Q3BS, Q5AS} were {1, 1, 1, 1} and {1, 0, 0, 0}, respectively. Passing the two vectors through an XOR gate, the resulting vector was {0, 1, 1, 1} as both had the same treatment and the purpose of differential expression analysis was to reveal the "difference" rather than the "similitudes". This XOR gate exclusion model was denoted by "Fg_exc". The effect of "Fg" was excluded in the resulted ratio-trait association. For the purpose of comparison and validation, we passed these two vectors through an OR gate and the resulting vector was {1, 1, 1, 1}. This OR gate inclusion model kept "Fg" in the resulted ratio-trait association. For a trait in a ratio, a result of "1" from the XOR gate was analogous to equation 2.

$$-x/y = -(x/y) \qquad (2)$$

We used the weighted correlation network analysis (WGCNA) R package [15] to generate a Topology Overlap Matrix (TOM). Based on value distribution, we used a soft threshold of 12 and generated TOM values from positive correlation (TOM_{pos}) and from negative correlation (TOM_{neg}) separately. The two matrices were merged as:

$$TOM = (TOM_{pos} + TOM_{neg})/2 \qquad (3)$$

The TOM value remains between 0 and 1. The higher TOM value between a pair of genes meant better topology overlap (higher similarity). Thus, we used 1-TOM as a distance measure in hierarchical clustering.

Clustering was based on the distance in pairwise topology overlap of gene expression profile (1-TOM) or on the distance-in-shape [23] in gene expression profiles across lines and treatments between individual genes. When a cluster was identified, an eigenvector was generated based on expression values of all genes in the cluster. This eigenvector was then used as a representation of the cluster and a Pearson correlation analysis was performed between the eigenvector and the digitization vector for trait representation of respective wheat lines or samples. In addition, a Pearson correlation analysis was performed between the expression profile of each individual gene and the digitization. The WGCNA package was also used to generate gene association networks, which were visualized using Cytoscape [6].

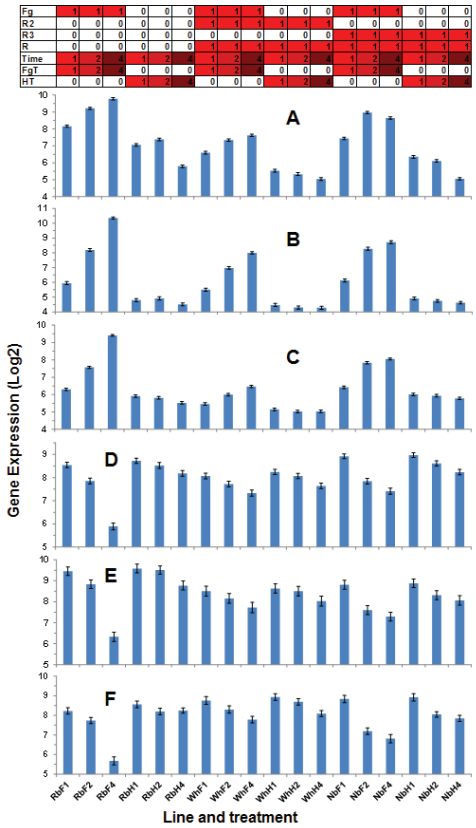

Fig. 1. Six clusters from the first dataset significantly correlated ($p<0.001$) with infection by *Fusarium graminearum*. Error bar = standard error.

IV. Results

From the first dataset, 22 clusters were identified through hierarchical clustering. Fig. 1 shows six clusters significantly correlated with infection by *Fusarium graminearum* over time ($p<0.001$). Clusters A, B, and C contain genes that were generally up-regulated by the infection. Over time, their expression level increased with infection, but generally decreased without infection. After infection, the slope of increase was steeper in the susceptible line Roblin than in the two resistant lines, Wuhan and NuyBay. Gene enrichment analysis revealed that these clusters were enriched in genes involved in plant defense related functions (Table 1). Genes in clusters D, E, and F have an opposite pattern and gene expression decreased over time with or without infection in all

Table 1. Gene Ontology analysis of significant clusters from the first dataset.

Cluster	Size	Enriched Gene Ontology
A B C	358 303 353	cinnamic acid biosynthetic process; L-phenylalanine catabolic process; ammonia-lyase activity; phenylpropanoid metabolic process; response to abscisic acid; response to ethylene; defense responses to fungus and to bacterium; negative regulation of defense response; plant-type hypersensitive response.
D E F	74 37 28	pentose-phosphate shunt; photosynthesis light reaction; starch biosynthetic process; cellular glucan metabolic process; chloroplast thylakoid membrane; photosystem II assembly; photosynthetic electron transport in photosystem I.
G	105	amidase activity; indoleacetamide hydrolase activity; heme binding; protein heterodimerization activity; acyl-CoA oxidase activity; acyl-CoA dehydrogenase activity; hydrolase activity, hydrolyzing O-glycosyl compounds; removal of superoxide radicals, auxin biosynthetic process.
H	134	response to salt stress; aspartic-type endopeptidase activity; glyoxalase III activity; lactate biosynthetic process; sterol binding; proteolysis; response to water deprivation, response to hydrogen peroxide; peroxidase activity; gibberellin biosynthetic process; protein heterodimerization activity.
I	15	glucuronoxylan metabolic process; xylan biosynthetic process; protein heterodimerization activity.
J	135	dolichyl-diphosphooligosaccharide-protein glycotransferase activity; hydrolyzing O-glycosyl compounds; small GTPase mediated signal transduction; xylan catabolic process; response to endoplasmic reticulum stress; defense response signaling pathway, resistance gene-independent; plant-type hypersensitive response; response to hydrogen peroxide; coumarin biosynthetic process; salicylic acid biosynthetic process.
K	43	inorganic diphosphatase activity; lipid transport; magnesium ion binding; defense response to fungus and bacterium.
L	87	protein heterodimerization activity; peroxidase activity; hydroxyethylthiazole kinase activity; terpene synthase activity; heme binding; fatty acid biosynthetic process; magnesium ion binding; metal ion binding.

three lines. When infected, the decrease was more profound throughout all three lines, particularly in the susceptible line Roblin. This pattern suggests that this group of genes may be involved in primary metabolic functions, which was confirmed by GO enrichment analysis (Table 1).

Fig. 2 shows three mirror image pairs of clusters with genes differentially expressed between the wheat lines. Genes in clusters G and H were differentially expressed between the susceptible line and the two resistant lines. Genes in cluster G had higher expression in the susceptible line as compared to the resistant lines and were associated with oxidase, hydrolase, and dehydrogenase activities among others (Table 1). Genes in Cluster H had an opposite pattern and were associated with various stress response functions. Genes in clusters I and J were differentially expressed between Wuhan and the other two lines, the susceptible Roblin and resistant NuyBay. Genes in Cluster I appeared loosely associated with primary metabolism, while those in Cluster J were associated with plant defense against pathogens. Genes in clusters K and L showed differential expression patterns between NuyBay and the other two lines. Unlike the previously described pairs of clusters, there is no clear distinction in metabolic functions of genes in clusters K and L (Table 1).

In the second dataset, 18 hierarchical clusters were identified. One cluster included 44% (1263 genes, Cluster M, Fig. 3) and another cluster 51% (1467 genes, Cluster N, Fig. 3) of the 2880 genes. The former contains 92% genes down-regulated ($p<0.05$), while latter contains 99% genes up-regulated ($p<0.05$) after *Fusarium* infection. The Pearson correlation analysis between expression of individual genes and the digitized trait information revealed similar results (Table 2). Another smaller cluster (19 genes, Cluster O, Fig. 3) contains 53% up-regulated genes ($p<0.05$) after *Fusarium* infection. Clustering results indicated that genes in 9 clusters were differentially expressed as an effect of the three QTLs, with numbers of genes ranging from 3 to 20 per cluster. Interestingly, all 1,466 genes positively correlated ($p<0.05$) with *Fusarium* infection were captured in two clusters (N and O). These two clusters contain 20 additional genes, of which, 19 were moderately correlated ($p<0.2$) with *Fusarium* infection. Similarly, all 1,156 genes negatively correlated with *Fusarium* infection were captured in one cluster; among the additional 107 genes in that cluster, 88 were moderately correlated ($p<0.2$) with *Fusarium* infection (Fig. 3).

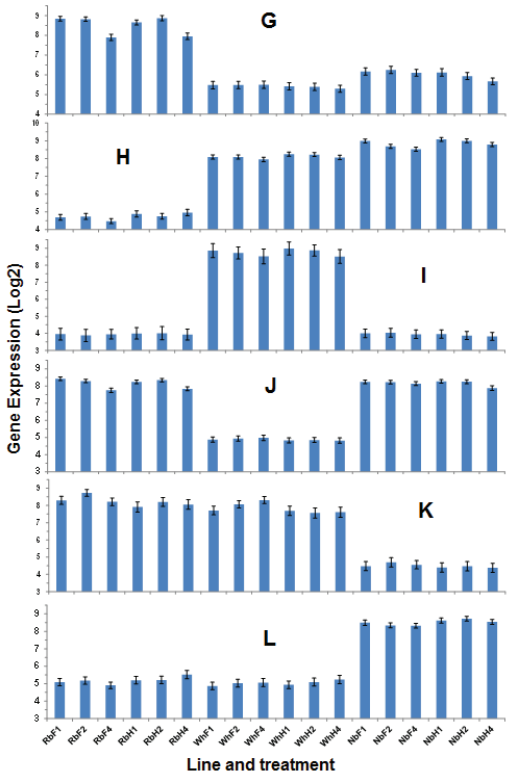

Fig. 2. Six clusters from the first dataset significantly ($p < 10\text{-}10$) and differentially expressed genes between Roblin, Wuhan, and Nuybay.

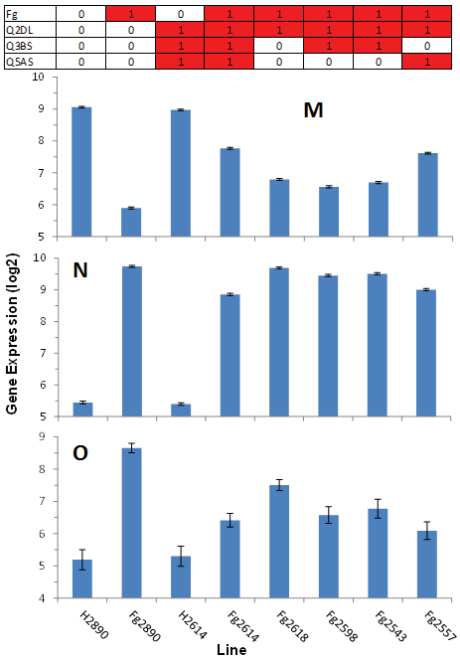

Fig. 3. Three clusters from the second dataset significantly correlated with *Fusarium* infection.

Gene ontology enrichment analysis [30] revealed that genes down-regulated after *Fusarium* infection are predicted to mainly contribute toward growth and other primary metabolic

Table 2. Number of differentially expressed genes under *Fusarium* infection from the second dataset

Analysis	Direction	Fg	Q2DL	Q3BS	Q5AS
Pearson correlation	Positive	1466	44	30	32
	Negative	1156	2	6	7
Clustering	Up	1467	10	16	6
		19	10	3	20
			3	14	
			20		
	Down	1263		9	7
Joint	Positive and Up	1466	34	24	14
	Negative and Down	1156	0	6	5

Affy_ID	SOM clustering						
	3x3	3x4	4x4	4x5	5x5	5x6	6x6
Ta.11139.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.14288.1.A1_x_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.14426.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.14426.1.S1_x_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.26997.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.4574.1.A1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.5248.1.S1_x_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.608.3.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.7740.1.A1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.7740.1.A1_x_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.9169.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.9671.3.S1_a_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
TaAffx.94321.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
TaAffx.98909.1.A1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
TaAffx.98909.1.A1_s_at	C1_0	C0_1	C3_0	C3_2	C0_1	C2_0	C3_1
Ta.14588.1.S1_x_at	C1_0	C0_1	C3_0	C3_2	C0_1	C4_4	C1_3
Ta.14588.2.S1_x_at	C1_0	C0_1	C3_0	C3_2	C0_1	C4_4	C1_3
Ta.18729.1.S1_s_at	C1_0	C0_1	C3_0	C3_2	C0_1	C4_4	C1_3
Ta.19491.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C4_4	C1_3
Ta.28317.1.S1_at	C1_0	C0_1	C3_0	C3_2	C0_1	C4_4	C3_1

Fig. 4. SOM clustering result for cluster P generated from hierarchical clustering process.

functions, such as photosynthesis and chromatin assembly, while those up-regulated are predicted to mainly contribute to stress response and detoxification functions, including defense response to fungus and bacteria, toxin catabolic process, abscisic acid glucosyltransferase activity, responses to salicylic acid, jasmonic acid, abscisic acid, ethylene, gibberellin, cyclopentenone, and chitin. This observation is consistent with the result from the first dataset (Fig. 1).

In order to verify the similarity among genes in the same clusters, we took two additional approaches. First, we used self-organized maps (SOM) in combination with difference-in-shape as a distance measure [23] to independently cluster the second dataset using various parameter settings. Clustering stability [8] was used for the final cluster selection. Secondly, we analyzed the fold change data based on the extent of differential expression between the lines and between the treatments. When a line was infected and had a control (lines 2-2890 and 2-2614), we took the log ratio as a measure of differential expression between the infected plants and the controls. Each of the resistant lines was compared with the susceptible line (2-2890) and this revealed the effect of one or more QTLs as compared with the susceptible line under *Fusarium* infection. In this analysis, we used both the exclusion and the inclusion models in trait representation .

Now, we take a closer look at cluster P produced by hierarchical clustering (Fig 4 & 5) to demonstrate the effect of

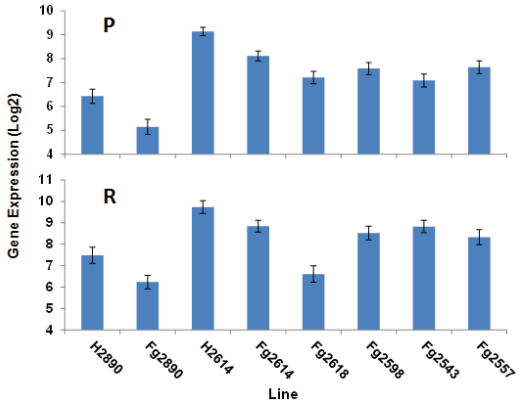

Fig. 5. Gene expression profile of the two clusters significantly correlated with one or more QTLs. Cluster P is the same as Fig. 4.

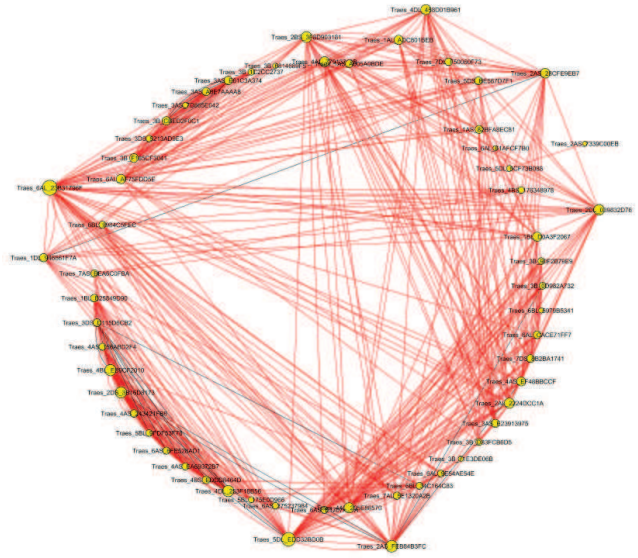

Fig. 6. Association network of genes in 9 clusters that are significantly correlated with one or more QTLs. The size of the node corresponds to the degree of the node. The 11 outstanding hub genes are cited in the text.

SOM. This cluster had 20 probesets representing 16 genes and was highly correlated with QTLs 2DL ($p<0.02$) and 5AS ($p<0.05$), and moderately with 3BS ($p<0.1$). This cluster could be further divided into two or three sub-clusters based on SOM parameter settings of 5x6 and 6x6, respectively. However, with slightly reduced granularity, the other five parameter settings collectively yielded an intact cluster (Fig. 4). Based on the stability principle [8], we considered this set of genes as one cluster. Gene expression in this cluster was up-regulated (FC ≥ 2) in the resistant lines as compared to the susceptible line, but down-regulated by *Fusarium* infection in both the susceptible

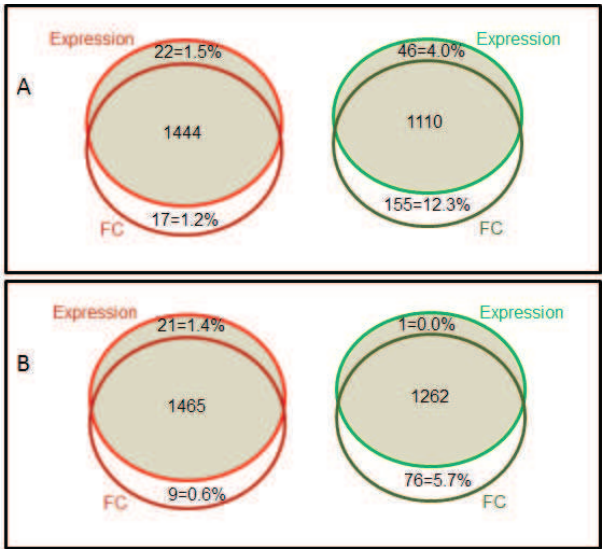

Fig. 7. Vann diagrams demonstrate the consistency between expression data and fold change differential expression data between infected and uninfected wheat. The exclusion model of digitizing trait representation was used for analyzing the fold change data. **A**: Pearson correlation, **B**: Hierarchical clustering.

line and the resistant line. All three QTLs appeared to contribute to resistance against the infection. Gene expression appeared to be less down-regulated in wheat lines having a higher number of QTLs (Fig. 5).

The gene expression profile of cluster P has a similar implication as that of clusters D, E, F from the first dataset in primary metabolism (Fig. 1 and Table 1). Examples of predicted function for genes in this cluster include the early light-induced protein (ELIP, Traes_4DL_458D01B961) involved in photosynthesis and photoprotection (Table 3), and the DNA-directed RNA polymerase II subunit D (NRPB4, Traes_6AS_281AF94BC) involved in RNA splicing, mRNA export from nucleus in response to heat stress, nuclear-transcribed mRNA catabolic process, etc. This cluster is also enriched with genes involved in lipase/hydrolase activities. Genes of this class include a putative adenine nucleotide alpha hydrolases-like protein (Traes_5DS_BE567D7F1), involved in hydrolase activity in response to stress (specifically to molecules of fungal origin), and putative alpha/beta-hydrolases superfamily protein (Traes_1AL_ADC801BEB), and GDSL-like lipase/acylhydrolase (Traes_7DL_F50060F73), which is involved in lipid metabolic processes.

Gene association network analysis indicated that genes in cluster P were best connected among the 9 clusters that were significantly correlated with one or more of the three resistant QTLs; These 9 clusters collectively have 95 probesets representing 61 wheat genes. Genes in cluster P were associated with 24 genes on average in the network. Among the top 10 best connected hub probesets in the network, 6 belonged to this cluster. They include two putative leucine-rich repeat (LRR) family proteins (Traes_2BS_3F8D903181 and Traes_2AS_28CFE9EB7, FORL1, Fig. 6, Table 3). Their common rice ortholog is a polygalacturonase inhibitor 1 precursor (PGIP1, LOC_Os07g38130), which has inhibitory activity against fungal polygalacturonase and is an important factor in regulation of floral organs [13]. Other hub genes include the putative ELIP (Traes_4DL_458D01B961), the two hydrolases (Traes_5DS_BE567D7F1 and Traes_1AL_ADC801BEB) the lipase (Traes_7DL_F50060F73), described above; ELIP is among the top 10 best connected hub probesets.

The remaining 4 of the top 10 best connected hub probesets, representing three genes, were in cluster R (Fig. 5), which was highly correlated with QTL 3BS ($p < 0.02$) and moderately correlated with QTLs 2DL ($p < 0.1$) and 5AS ($p < 0.15$). These genes include a putative SC35-like splicing factor 30A (SCL30A, Traes_2AS_FEB84B3FC) and two potential aquaporin proteins (Traes_6AL_23B31796F, Traes_2DL_039832D76). These two aquaporin-coding genes have a common ortholog in *Arabidopsis* (AT3G53420) that is involved in the response to various stress factors (Table 3). Other important hub genes in the cluster include putative

Table 3. Interesting network hug genes in clusters P and R from the second datasets and their annotations.

Affy ID	Wheat Gene	Ortholog		gene	Annotation
		Rice*	Arabidopsis		
Ta.26997.1.S1_at	Traes_4DL_458D01B96	Os01g0246400	AT3G22840	ELIP	located in chloroplast thylakoid membrane, photosystems I and II and involves in anthocyanin-containing compound biosynthetic process, cellular response to UV-A, UV-B, blue light, red light, far red light, high light intensity, heat, cold, karrikin (plant growth regulators) and sucrose; photosynthesis, photoprotection, regulation of chlorophyll biosynthetic process and seed germination, flavonoid biosynthetic process among others
Ta.19491.1.S1_at	Traes_5DS_BE567D7F1	Os12g0552500	AT2G47710	adenine nucleotide alpha hydrolases-like protein	appears in golgi apparatus, plasma membrane and vacuole, and involves in hydrolase activity in response to stress, specifically to molecule of fungal origin
Ta.7740.1.A1_at Ta.7740.1.A1_x_at	Traes_1AL_ADC801BEB	Os05g0574100	AT2G42690	alpha/beta-Hydrolases superfamily protein	appears in chroloplast and cytoplasm and involved in UV protection, carotenoid biosynthetic process, cellular response to phosphate starvation, cysteine biosynthetic process, galactolipid biosynthetic process, glucosinolate biosynthetic process, lipid catabolic process, response to fructose, response to salt stress, water transport; have molecular function of phosphatidylcholine 1-acylhydrolase activity, triglyceride lipase activity.
Ta.9671.3.S1_a_at	Traes_7DL_F50060F73	N/A	AT3G50400	GDSL-like lipase	located in endomembrane system, involves in the lipid metabolic process, hydrolase activity, acting on ester bonds, carboxylesterase activity
Ta.14588.1.S1_x_at	Traes_2BS_3F8D90318	OS07G0568700	AT1G33670	PGIP1	an inhibitor of fungal polygalacturonase and an important factor for plant resistance to phytopathogenic fungi
Ta.14588.2.S1_x_at	Traes_2AS_28CFE9EB7				
Ta.9544.2.S1_a_at	Traes_2AS_FEB84B3FC	Os07g0633200	AT3G13570	SCL30A	located in spliceosomal complex and involves in RNA splicing, transcription regulation, post-translational protein modification, production of miRNAs involved in gene silencing by miRNA, production of siRNA involved in RNA interference
Ta.28728.1.S1_at	Traes_6AL_23B31796F	Os02g0629200	AT3G53420	aquaporin protein	located in chloroplast and response to abscisic acid, cadmium ion, cold and salt stresses, and temperature stimulus
Ta.28728.1.S1_x_at					
Ta.28728.2.S1_x_at	Traes_2DL_039832D76	Os04g0521100			
Ta.5818.1.S1_a_at	Traes_5DL_EDD32BD0F	Os09g0491756	AT4G24130	U2AF^{35}	Splicing factor U2AF 35 kDa subunit (U2 auxiliary factor 35 kDa subunit) (U2 snRNP auxiliary factor small subunit) located in nucleus, RNA binding, nucleic acid binding, metal ion binding.
Ta.5818.1.S1_at					
Ta.5818.3.S1_x_at					
TaAffx.2456.2.S1_s_at	Traes_1DL_946661F7A	Os05g0364600	AT1G09140	SRp30	alternative splicing regulator, nucleic acid binding

* *Oryza sativa* Japonica

The selected genes in this table are key hub genes in the association network (Fig. 6) and extensively described in the text.

splicing factor U2AF 35 kDa subunit (Traes_5DL_EDD32BD0B) and alternative splicing regulator SRp30 (Traes_1DL_946661F7A), suggesting that this cluster may be important in RNA splicing and alternative splicing.

From the second approach, we found that the exclusion model of trait representation (Equation 1 & 2) outperformed the inclusion model. For example, the effect of *Fusarium* infection was not well revealed by the inclusion model (Fg). Only a small cluster of 12 probesets was significantly correlated with the infection. Of the 12 probesets, only 7 were also grouped by earlier clustering based on expression value

Comparison of	Fg2H2890	Fg2H2614	H2614	Fg2614	Fg2618	Fg2598	Fg2543	Fg2557
			to2890					
Fg	1	1	0	1	1	1	1	1
Fg_exc	1	1	0	0	0	0	0	0
Q2DL	0	1	1	1	1	1	1	1
Q3BS	0	1	1	1	0	1	1	0
Q5AS	0	1	1	1	0	0	0	1
Q2DL_exc	0	0	1	1	1	1	1	1
Q3BS_exc	0	0	1	1	0	1	1	0
Q5AS_exc	0	0	1	1	0	0	0	1

Fig. 8. Digitalization of trait representation of fold change data. **Row headers:** Inclusion model: Fg = Fusarium infected, Q2DL = QTL 2DL, Q3BS = QTL 3BS, Q5AS = QTL 5AS. Exclusion model: Fg_exe = Fusarium infected, Q2DL_exe = QTL 2DL, Q3BS_exe = QTL 3BS, Q5AS_exe = QTL 5AS. **Column headers:** Fg2H2890 = Fg infection over water treated in the 2-2890 susceptible line; the same syntax applies to the resistant line 2-2614. H2614to2890 = under water mock inoculation the resistant line 2-2612 is compared with the susceptible 2-2890 line; the same syntax applies to other resistant lines over the susceptible 2-1890 line with *Fusarium* infection.

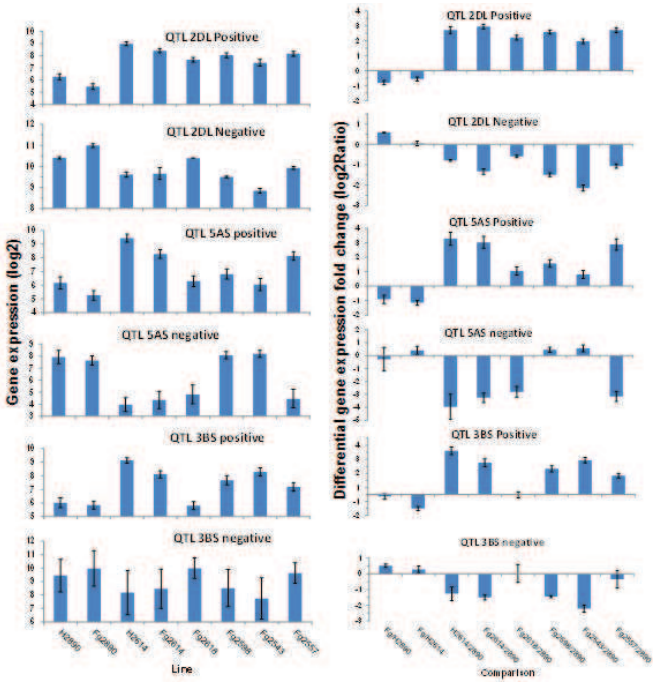

Fig. 9. Expression (left column) and differential expression (fold change, right column) profiles of genes significantly correlated with one or more QTLs. Data shown here are significantly correlated between both the expression data and differential fold change on one side and digitized trait specificity on the other.

and 5 of the 7 were correlated with *Fusarium* infection. The exclusion model (Fg_exe) allowed the detection of 3 clusters with a total of 1,338 probesets negatively correlated with *Fusarium* infection, and 3 clusters with a total of 1,474 probesets positively correlated with the infection. This result is consistent with earlier clustering results (Fig. 7). Nevertheless, neither the exclusion nor the inclusion models worked well for QTL 2DL. The main cause to this problem is related with the fact that QTL 2DL is present in all three resistant lines. Q2DL_exe is exactly the negation of Fg_exe (Fig. 8). In this case, the effect of *Fusarium* infection in the ratio data over-shadowed the effect 2DL and neither model of digitization of trait representation could be directly applied to ratio data involved with this QTL.

Fig. 9 shows the expression profiles of genes that were significantly correlated with the existence of a QTL, both in the expression data and fold change data using the exclusion model. These profiles were similar among the three QTLs. Among the positively correlated genes, *Fusarium* infection appeared to moderately suppress expression of these groups of genes, both in the susceptible and resistant lines. It is interesting to note that QTL 2DL alone (Line 2-2618) had no effect on genes significantly correlated with 3BS.

V. Discussion

Our study shows that the digitization of trait specificity was instrumental for downstream analysis of gene expression in wheat challenged with *Fusarium* infection. This enabled discovery of informative patterns linked with infection by *Fusarium graminearum*, and identification of distinct patterns in resistant lines. The exclusion model, inferred from its analog in mathematics, outperformed the inclusion model. The numerical representation of wheat traits allowed the application of mathematical and logical operations to identify differentially expressed genes linked with resistance to *Fusarium* infection. With this approach, we discovered distinct clusters of genes contributing to the informative patterns, which were predicted to be related, either directly or indirectly, with *Fusarium* infection and resistant types or QTLs. Pattern recognition and cluster discovery are important steps toward discovering genes associated with certain phenotypic traits. Most phenotypic traits are polygenic (contributed by more than one gene). Yet, genes present in a cluster highly correlated with a given phenotypic trait do not necessarily contribute to this trait. For example, each of the clusters M and N from the second dataset has over 1000 genes, and only a small subset of genes from each cluster are probably directly related with *Fusarium* infection. Other differentially expressed genes can be related to plant physiological state because of a decline in overall growth and photosynthesis, or in response to wounds caused by the infection, or changes in the overall plant immune system. Therefore, subsequent gene ontology analysis, network analysis and other systems biology approaches may be necessary to discover functional association between genes and phenotypic traits, and association among the genes.

In this study, we performed a gene association network analysis of genes in the 9 clusters that were significantly correlated with one or more of the three resistant QTLs. Specifically, we focused on 11 inter-connected hub genes from

the two most informative clusters. Polygalacturonase-inhibiting proteins (PGIPs) are plant proteins that weaken the function of fungal polygalacturonases, which are important virulence factors [9]. We found two putative leucine-rich repeat (LRR) family protein-coding genes (Traes_2BS_3F8D903181 and Traes_2AS_28CFE9EB7) in cluster P, believed to encode for PGIP1, which were down-regulated in response to *Fusarium* infection both in the susceptible line 2-2890 and the resistant line 2-2614, but were expressed at a higher level in the resistant lines as compared with the susceptible line with or without the infection. It is interesting to note that the reduction in gene expression caused by the infection was more drastic in the susceptible line than in the resistant line (60% vs. 30%). PGIPs are usually disease induced in response to infection by pathogen, but our results show that without infection the resistant line 2-2614 had over 150% more PGIP1 mRNA (1.3 FC) than the susceptible line 2-2890. With infection, the differential fold change was even higher at 2.1 FC (320% more). This suggests that these putative PGIP1 genes are involved in wheat defense against *Fusarium* infection, most likely through the inhibition of polygalacturonases. The two PGIP1-like genes were directly associated in the network with a putative ELIP (Traes_4DL_458D01B961), another hub gene, which is known to protect from photooxidative stress in *Arabidopsis* [11], and believed to be involved in cold adaptation of *Rhododendron* plants [36].

These three inter-connected genes were all connected to five other hub genes in cluster R, three putative splicing factors (SCL30A, U2AF 35 kDa subunit and SRp30) and two potential aquaporin protein genes. SCL30A is a serine/arginine-rich protein splicing factor and is involved in alternative splicing and the regulation of flowering time. In *Arabidopsis*, SCL30A itself produces seven different transcripts in root, leaf, stem and inflorescence under different hormones and abiotic stresses [25]. The splicing factor U2AF 35 kDa subunit ($U2AF^{35}$) is conserved from fusion yeast to human [35]. It is not well understood how $U2AF^{35}$ functions in plants, but its homologs are found in *Arabidopsis*, rice, maize and other flowering plants [26]. The alternative splicing regulator SRp30 exists in three alternative splicing forms that are differentially expressed in *Arabidopsis* [4] and the alternative splicing pattern of other SR protein genes has been found to be altered in transgenic lines overexpressing *Arabidopsis* SRp30 [16]. The fact that these three splicing factors and regulators were co-expressed in wheat under *Fusarium* infection in the resistant lines as compared with the susceptible line is interesting and deserves further investigation. Aquaporins (AQPs) are integral membrane proteins that form pores called water channels in the membrane of biological cells, and are known to play important roles in abiotic stress tolerance of durum wheat [1]. The expression of aquaporin genes varies depending on the organ, hormone levels and abiotic stress treatments. In our study, expression of two wheat aquaporin genes decreased significantly after *Fusarium* infection, more drastically in the susceptible line 2-2890 (1.7 FC) than in the resistant line 2-2614 (1.1 FC). The resistant line had a higher expression than the susceptible line with (2.9 FC) and without (2.3 FC) infection. Reduction of aquaporin water channels is thus probably a stress response to *Fusarium* infection.

Generally, these hub genes are more or less involved in plant tolerance or stress response to adverse environmental conditions. Some are directly involved in defense again pathogens invasion, such as PGIP1, others are involved in overall tolerance or stress responses, such as AQPs. Yet, others, such as SCL30A, $U2AF^{35}$ and SRp30 are probably involved in downstream response after plants are stressed. There are more hub genes worth studying in this small scope network. A broader network study is beyond the scope of this paper. Subsequent work is in progress to characterize in more detail the wheat response to *Fusarium* infection and mechanisms of resistance to the infection.

VI. Conclusion

We applied digitization of trait associated with *Fusarium* susceptibility and resistance in wheat lines with and without infection. This was necessary for the subsequent analysis that revealed clusters and individual genes that were correlated with certain types of resistance or QTLs, or with the infection. In this study, we focused our attention on genes in 9 clusters that were significantly correlated with one or more resistant QTLs and specifically on the 11 well connected hub genes in the associated network. Our study therefore demonstrates the usefulness of such digitization. Microarray gene expression data are very information rich. Beyond expression itself, differential expression contains additional information. In order to properly explore such information, we devised specific exclusion model for the digitization of trait specificity. Our results demonstrate that the exclusion model outperformed the inclusion model in digitization of differential expression fold change data. The combination of gene expression and differential gene expression data extract in higher confidence groups of genes in certain transcript profiles that are correlated with one or more of the three QTLs in the second dataset. This study is a preliminary step that demonstrates the success of the digitization approach. A broader scale and more detailed systems biology study is in progress and will be reported subsequently.

References

[1] Ayadi M, Cavez D, Miled N, Chaumont F, Masmoudi K (2011). Identification and characterization of two plasma membrane aquaporins in durum wheat (*Triticum turgidum* L. subsp. durum) and their role in abiotic stress tolerance. *Plant Physiol Biochem* 49: 1029-1039.

[2] Atchison J, Head L, Gates A (2010). Wheat as food, wheat as industrial substance; comparative geographies of transformation and mobility. *Geoforum* 41: 236-246.

[3] Bai G, Shaner G (2004). Management and resistance in wheat and barley to Fusarium head blight. *Ann Rev Phytopathol* 42: 135-161.

[4] Barta A, Kalyna M, Reddy ASN (2010). Implementing a rational and consistent nomenclature for serine/arginine-rich protein splicing factors (SR proteins) in plants. *Plant Cell* 22: 2926–2929.

[5] Buerstmayr H, Ban T, Anderson JA (2009). QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: a review. *Plant Breeding* 128: 1–26.

[6] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* 2: 2366 - 2382

[7] Ding L, Xu H, Yi H, Yang L, Kong Z, Zhang L, Xue S, Jia H, Ma Z (2011). Resistance to hemi-biotrophic *F. graminearum* infection is associated with coordinated and ordered expression of diverse defense signaling pathways. *PLoS One* 6:e19008.

[8] Famili AF, Liu G, Liu Z (2004). Evaluation of optimization of clustering in gene expression data analysis. *Bioinformatics* 20:1535–1545.

[9] Ferrari S, Vairo D, Ausubel FM, Cervone F, De Lorenzo G (2003). Tandemly duplicated *Arabidopsis* genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection. *Plant Cell* 15: 93-106.

[10] Foroud NA, Ouellet T, Laroche A, Oosterveen B, Jordan MC, Ellis BE, Eudes F (2012). Differential transcriptome analyses of three wheat genotypes reveal different host response pathways associated with Fusarium head blight and trichothecene resistance. Plant Pathol. 61: 296 - 314.

[11] Hutin C, Nussaume L, Moise N, Moya I, Kloppstech K, Havaux M (2003). Early light-induced proteins protect *Arabidopsis* from photooxidative stress. *PNAS* 100: 4921-4926.

[12] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.

[13] Jang S, Lee B, Kim C, Kim SJ, Yim J, Han JJ, Lee S, Kim SR, An G (2003). The OsFOR1 gene encodes a polygalacturonase-inhibiting protein (PGIP) that regulates floral organ number in rice. *Plant Mol Biol* 53: 357-369.

[14] Kikot GE, Hours RA, Alconada TM (2009). Contribution of cell wall degrading enzymes to pathogenesis of *Fusarium graminearum*: a review. *J Basic Microbiol* 49 :231-241.

[15] Langfelder P, and Norvath S (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559.

[16] Lopato S, Kalyna M, Dorner S, Kobayashi R, Krainer AR, Barta A (1999). atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. *Genes Dev* 13: 987–1001

[17] Makandar R, Nalam V, Chaturvedi R, Jeannotte R, Sparks AA, Shah J (2010). Involvement of salicylate and jasmonate signaling pathways in *Arabidopsis* interaction with *Fusarium graminearum*. *Mol Plant Microbe Interact* 23: 861-70.

[18] Mesterházy Á, Bartók T, Mirocha CG, Komoróczy R (1999). Nature of wheat resistance to Fusarium head blight and the role of deoxynivalenol for breeding. Plant Breeding 118: 97-110

[19] Mosteller F, Tukey J (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley. ISBN 0-201-04854-X.

[20] Oerke EC (2006). Crop losses to pests. Journal of Agricultural Science. 144: 31–43

[21] Ouellet T, Wang L, Schneiderman D, Balcerzak M, Tinker N, Harris L (2005). Comparison of the response of Roblin, Wuhan and NuyBay to *Fusarium* attack using a genomic and proteomic approach. Proceedings of 4th Canadian Workshop on Fusarium Head Blight, Ottawa, Ontario, Canada, November 1-3, 2005. P29.

[22] Ouellet T, Hattori J, Gulden S, Wang L, Soleimani V, Fedak G, Singh J, Pandeya R, Somers D, Tinker N (2008). A database of RNA profiles comparing susceptible and resistant wheat infected with *Fusarium graminearum*. 11th International Wheat Genetics Symposium 2008 Proceedings, Volume 3, Eds. Rudi Appels, Russell Eastwood, Evans Lagudah, Peter Langridge, Michael Mackay, Lynne McIntyre and Peter Sharp. 801-802.

[23] Pan Y, JD Pylatuik, J Ouyang, A Famili and PR Fobert (2004). Discovery of functional genes for systemic acquired resistance in *Arabidopsis thaliana* through integrated data mining. Journal of Bioinformatics and Computational Biology 2: 639-655.

[24] Pan Y, J Zou, Y Huang, Z Liu, S Phan and FA Famili (2009). Goal driven analysis of cDNA microarray data. Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2009), March 30-April 2, Nashville, TN, USA. pp.186-192.

[25] Palusa SG, Ali GS, Reddy ASN (2006). Alternative splicing of pre-mRNAs of *Arabidopsis* serine/arginine-rich proteins: regulation by hormones and stresses. *The Plant Journal* 49: 1091–1107.

[26] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, pp. doi: 10.1093/nar/gkv007.

[27] Ru Y, Wang BB, Brendel V (2008). Spliceosomal proteins in plants. In: Reddy ASN and Golovkin M (eds.), *Nuclear pre-mRNA Processing in Plants, Current Topics in Microbiology and Immunology 326*. Springer-Verlag Berlin Heidelberg 2008. Pp. 1-14.

[28] Schmale III, D.G. and G.C. Bergstrom. 2003 . Fusarium head blight in wheat. *The Plant Health Instructor*. DOI:10.1094/PHI-I-2003-0612-01 (updated 2010).

[29] Suzuki T, Sato M, Takeuchi T (2012). Evaluation of the effects of five QTL regions on Fusarium head blight resistance and agronomic traits in spring wheat (*Triticum aestivum* L.). *Breeding Science* 62: 11–17.

[30] Tchagang AB, Gawronski A, Bérubé H, Phan S, Famili F, Pan Y (2010). GOAL: A software tool for assessing biological significance of genes groups. *BMC Bioinformatics* 11:229.

[31] Tekauz A (1999). Fusarium Head Blight of Barley: A Plant Pathologist's Perspective. In: Proceedings of the Canadian Barley Symposium, February 23-25, 1999, Winnipeg, Manitoba, Canada, pp. 85-91.

[32] Tulpan D, Leger S, Tchagang A, Pan Y (2015). Enrichment of *Triticum aestivum* gene annotations using ortholog cliques and gene ontologies in other plants. BMC Genomics, to appear.

[33] Waldron BL, Moreno-Sevilla B, Anderson JA, Stack RW, Frohberg RC (1999). RFLP Mapping of QTL for Fusarium Head Blight Resistance in Wheat. *Crop Sci* 39: 805–811.

[34] Walters S, Nicholson P, McDoohan F (2009). Action and reaction of host and pathogen during Fusarium head blight disease. *New Phytol* 185: 54-66.

[35] Webb CJ, Wise JA (2004). The Splicing Factor U2AF Small Subunit Is Functionally Conserved between Fission Yeast and Humans. *Mol Cell Biol* 24: 4229–4240.

[36] Wei H, Dhanaraj AL, Rowland LJ, Fu Y, Krebs SL, Arora R (2005). Comparative analysis of expressed sequence tags from cold-acclimated and non-acclimated leaves of *Rhododendron catawbiense* Michx. *Planta* 221:406-416