Highlights in BioScience
ISSN:2682-4043
DOI:10.36462/H.BioSci.202203

Research Article

Open Access

CrossMark

[1] Biotechnology Department, Science Faculty, Helwan university, Helwan, Egypt.
[2] Biotechnology Department, Agriculture Faculty, Al-Azhar university, Cairo, Egypt.

*** To whom correspondence should be addressed: saifeldeenmib99@gmail.com**

Editor: Hatem Zayed, *College of Health and Sciences, Qatar University, Doha, Qatar.*

Reviewer(s):
Jean Legeay, *Université de Lorraine, Inra, UMR IAM - Interactions Arbres-Microorganismes, F-54000 Nancy, France.*
Vivek Chaudhary, *Motilal Nehru National Institute of Technology, Allahabad, Prayagraj, Uttar Pradesh 211004, India.*

Received: September 20, 2022
Accepted: December 3, 2022
Published: December 15, 2022

Citation: Ibrahim MS, Ibrahim SM. Identifying biotic stress-associated molecular markers in wheat using differential gene expression and machine learning techniques. 2022 Dec 15;5:bs202203

Data Availability Statement: All relevant data are within the paper and supplementary materials.

Funding: The authors have no support or funding to report.

Competing interests: The authors declare that they have no competing interests.

Identifying biotic stress-associated molecular markers in wheat using differential gene expression and machine learning techniques

Manar S. Ibrahim[1], Saifeldeen M. Ibrahim*[2]

Abstract

Wheat is an important crop for global food security and a key crop for many developing countries. Thanks to next-generation sequencing (NGS) technologies, researchers can analyze the transcriptome of wheat and reveal differentially expressed genes (DEGs) responsible for essential agronomic traits and biotic stress tolerance. In addition, machine learning (ML) methods have opened new avenues to detect patterns in expression data and make predictions or decisions based on these patterns. We used both techniques to identify potential molecular markers in wheat associated with biotic stress in six gene expression studies conducted to investigate powdery mildew, blast fungus, rust, fly larval infection, greenbug aphid, and Stagonospora nodorum infections. A total of 24,152 threshold genes were collected from different studies, with the highest threshold being 7580 genes and the lowest being 1073 genes. The study identified several genes that were differentially expressed in all comparisons and genes that were present in only one data set. The study also discussed the possible role of certain genes in plant resistance. The *Ta-TLP, HBP-1, WRKY, PPO,* and glucan endo-1,3-beta-glucosidase genes were selected by the interpretable model-agnostic explanation algorithm as the most important genes known to play a significant role in resistance to biotic stress. Our results support the application of ML analysis in plant genomics and can help increase agricultural efficiency and production, leading to higher yields and more sustainable farming practices.

Keywords: Wheat, gene expression, machine learning, biotic stress, disease

Introduction

Agriculture is the primary source of food production on a global scale. In the new era, sustainable agriculture is used to preserve or improve food quality while protecting the environment [1]. wheat (*Triticum aestivum*) is a member of the Poaceae family and is considered the oldest and most commonly cultivated crop [2], with an annual cultivation area of 217 million hectares. It is the most commonly grown crop in the world [3]. Also, it is one of three primary cereal crops with the greatest impact on global food security, second only to maize (*Zea mays*) and third only to rice [4]. Most experts assume that wheat was the first crop to be produced between 10,000 and 8,000 B.C. [2], and this is considered evidence of the importance of wheat from ancient times. Wheat is one of the most essential crops for ensuring global food security [5]. Considering the position of wheat in the world trade, wheat accounts for over 50% of the worldwide grain trade and 30% of the global grain yield [6]. And for its role in nutrition, wheat is a staple crop in more than 40 nations worldwide, feeding 82% to 85% of the world's population [6]. According to the estimations, the annual wheat production is expected to reach 750.4 million metric tonnes in the next few years. Africa would consume 76.5 MMT of wheat, with imported grain accounting for 48.3 MMT, or 63.3 percent of total consumption [7]. Powerful yet user-friendly bioinformatics tools are required by the expanding field of gene expression profiling in order to facilitate systems-level data comprehension [8]. Studies on transcriptomics have a significant potential for unique and exploratory research aimed at new insights into molecular pathways [9].

Next-generation sequencing (NGS) technologies greatly improve studies on genome-wide mRNA expression data. In comparison to microarray technology, NGS provides higher resolution data and more precise transcript level measurements for gene expression studies [10]. By studying the wheat transcriptome, it is possible to identify the differentially expressed genes, genomic annotations, regulatory elements, molecular markers, and expression quantitative trait loci (eQTLs), as well as their sequence variants, which are responsible for significant characteristics [11].

With the advancement of next-generation sequencing technology, the genetic diversity of wheat has recently been studied using a high-throughput, low-cost molecular technique [12].The significant genetic variation seen in the primary, secondary, and tertiary wheat gene pools provides the starting point for the breeding of nutrient-rich wheat genotypes [13]. A variety of molecular markers have been used to study the genetic diversity of wheat [12] to measure biological population divergence's magnitude and each component character's proportionate contribution to the overall divergence [14].

Machine learning (ML) is a rapidly growing field of computer science that has the potential to revolutionize many industries, including plant science. Machine learning algorithms use statistical techniques to identify patterns in data and make predictions or decisions based on those patterns. This makes them well-suited to solving complex problems that are difficult for humans to tackle, such as analyzing vast amounts of data from experiments or monitoring the health of crops. In plant science, machine learning can be used to predict the best time to plant crops, optimize irrigation systems, identify pests and diseases, and even breed new varieties of plants. This could help improve the efficiency and productivity of agriculture, leading to better yields and more sustainable farming practices. Overall, the use of machine learning in plant science has the potential to significantly advance the field and improve our understanding of plants.

Consequently, in this study, the aim of our investigation is to perform expression profiling analysis of wheat to identify unknown molecular markers. We aim to use bioinformatics tools to ;(a) Study the gene expression of wheat under various environmental and developmental conditions. (b) Study the genetic diversity of these genes across the wheat genome. (c) Identify important genes and their interactions. (d) Use machine learning methods to identify genes associated with wheat response to biotic stresses. By doing so, we hope to gain a better understanding of the genetic basis of wheat response to biotic stresses, which could ultimately lead to the development of more resilient and productive varieties of wheat. This is important because biotic stresses, such as pests and diseases, can cause significant yield losses and can threaten the global food supply. By identifying molecular markers that are associated with wheat response to biotic stresses, we can improve our ability to breed new varieties of wheat that are better able to withstand these stresses. This could help to improve the efficiency and productivity of agriculture, leading to better yields and more sustainable farming practices.

Methodology

Data acquisition

The Gene Expression Omnibus (GEO) database, first established in 2005 [15], has continued to be a vital resource for the global scientific community. With over 20 years of experience, GEO remains the most widely used public repository for high-throughput gene expression and other functional genomics data sets. This valuable resource allows researchers to quickly and easily access a wide range of raw and processed data, along with detailed experimental descriptions, at no cost. The convenience and accessibility of GEO make it an invaluable tool for advancing scientific research. In this study, GEO has been used to download the data of six different experiments on wheat for further analysis. These experiments have been conducted under different biotec stresses. such as the response to powdery mildew infection, the response to blast fungus magnaporthe, the resistance pathways of transgenic wheat lines, the response to a Hessian fly larval attack, the response to greenbug aphid feeding, and the response to the necrotrophic effector SnTox3. All GSE accessions are provided with the experiment name of each accession in (table 1).

Differential gene expression analysis

With the help of the R/Bioconductor and Limma package, the online statistical tool GEO2R [16] was used to analyze the raw six gene expression data sets of wheat under different biotic stresses from GEO datasets [17] (table1). The samples have been clustered into two groups (control and infected / treatment). Differential expression genes were analyzed using R packages to create a volcano plot highlighting all significant genes and a heatmap of the top differentially expressed genes based on each group in all samples according to log2FC (fold change) ≥ 1 and adjusted p-value < 0.05 as the threshold for DEGs. Whereby up-regulated DEGs were considered if the logFC (fold change) ≥ 1 and down-regulated DEGs were considered if the logFC $\leq$ -1. A volcano plot shows statistical significance (P value) versus magnitude of change (fold change). It enables quick visual identification of genes with large fold changes that are also statistically significant. These may be the most biologically significant genes [18]. First, a volcano plot was generated using the R package (ggplot2) after reading the data and filtering it according to the adjusted p-value and logFC (blue: down-regulated, red: up-regulated). Second, a heatmap of the top differentially expressed genes in the RNA-seq data set was then generated. To do this, we need to extract the differentially expressed genes from the DE results file. The heatmap for the gene expression data was generated using R package (pheatmap) after filtering the top differential gene expression from the data. Then, a Venn diagram was used to show all the interrelationships of the six wheat experiments. The Venn diagram of the overlapping DEGs was output by the interactiVenn website [19].

Protein-protein interaction and and functional enrichment analyses

The STRING database was utilised for protein-protein interaction analysis [20] to evaluate the link between genes associated with wheat gene expression. The STRING database was chosen because this database seeks to integrate all known and predicted relationships between proteins, including functional and physical interactions [20]. All information about GO, annotated keywords, and protein domains for each GSE dataset has been provided in (table 2)

Machine Learning Model

The results of the differential gene expression study of putative biotic-associated gene biomarkers were employed as attribute values for machine learning (ML) analysis. We retrieved gene expression data from genes that were strongly expressed in gene expression. Their expression data were utilised as training and validation sets for the ML analysis. The gene expression data were adjusted before to the ML analysis. The sklearn and lime0.2.0.1 Python libraries were used to run the "Extra-tree regressor" and local interpretable model-agnostic explanations algorithms.

Results and Discussion

Identification of DEGs

A total of 24,152 threshold genes were collected from different studies. The highest threshold of genes was collected by GSE27320, with 7580 genes, while GSE31760 had the lowest threshold of genes, with only 1073 genes. The volcano plot indicates the upregulated (red color) and downregulated (blue color) genes in *Triticum aestivum* samples. The horizontal axis represents the fold change (log2FC), and the vertical axis represents the adjusted p-values. A total of 3701 upregulated and 3879 downregulated DEGs were identified from the GSE27320 dataset, while 241 upregulated and 832 downregulated DEGs were identified from the GSE31760 dataset.
GSE32151 yielded 3533 upregulated and 5138 downregulated DEGs, GSE34445 yielded 658 upregulated and 1442 downregulated DEGs, GSE45995 yielded 2054 upregulated and 1459 downregulated DEGs, and GSE59723 yielded 74 upregulated and 1141 downregulated DEGs. The volcano plot of each gene expression profile data is shown in (Figure 1).

Users can use heatmaps to examine the expression of a subset of genes. This can provide helpful insight into the expression of various groups and samples without losing sight of the larger study or losing clarity when evaluating patterns averaged over hundreds of genes at once [18]. According to the adjusted p-value ranking, a heatmap is created for the top DE genes.The heatmap clusters the samples into control (blue) and infected / treatment (red) groups. Red indicates samples with relatively high gene expression, while blue indicates samples with relatively low gene expression. Genes with intermediate expression levels are represented by lighter tones and white. A dendrogram hierarchical clustering has been used to reorder the samples and genes (Figure 2).

Venn diagrams of the DEGs between the integrated six GEO data sets are shown in (Figure 2). The number in each circle indicates the quantity of differentially expressed genes across the various comparisons. The overlapping number refers to genes that are differently expressed across all comparisons, while the non-overlapping number refers to genes that are unique to each sample. Genes were identified as being up- or down-regulated.

The GSE32151 and GSE45995 data sets shared 12 genes, while the GSE45995 and GSE59723 data sets shared 6 genes (PDI2, pepc, gstf5, gstf5:1, and ltp9.2). According to the study of Guo-Tian Liu *et al.*[21], gstf5 is a member of the (GSTs) family, which is responsible for catalysing the conjugation of the reduced form of glutathione to a number of electrophilic substrates, they also reported that GSTs contribute to resistance against powdery mildew.
The GSE31760, GSE45995, and GSE59723 data sets shared 3 genes (rgp, PR4, and pox3). The relationship between methyl jasmonate (MeJA) and the expression profiles of nine pathogenesis-related protein genes (PR genes) was examined in the Zongbiao Duan *et al.* [22] study to determine the interactive role in powdery mildew resistance. This investigation revealed that PR4 is one of the pathogenesis-related protein genes (PR genes), and the expressions of PR4 and other PR genes were most significantly activated by MeJA and showed a significant resistance to powdery mildew. The GSE32151 and GSE59723 datasets shared 1 gene (Xip-R1). Xip-R1 is one of the xylanase inhibitors. R.-J. SUN *et al.* [23], xylanase inhibitors (XIs) are plant cell wall proteins found mostly in monocots that limit microbial xylanases' hemicellulose degrading ability. Silvio Tundo *et al.* [24]. The GSE34445 and GSE59723 data sets shared 4 genes (WRKY45, Cht2, WRKY, and ccd1), and the GSE31760 and GSE59723 shared 2 genes (TaAOS and tamdr1). The allene oxide synthase (TaAOS) gene has been identified as being engaged in the JA signalling system, which increases plant resistance to Fusarium head blight (FHB) [25], Tamdr1 has also been identified as a Fusarium head blight resistance gene [26]. The GSE27320 and GSE31760 shared 2 genes (ald and pSBGer4), and the GSE27320 and GSE32151 data sets shared 2 genes (hsp16.9-3LC2 and S276) . The GSE32151 and GSE34445 data sets shared 1 gene (TaAML15), and the GSE27320 and GSE34445 data sets shared 3 genes (Ss1, wpi6, and gstf6b). The GSE34445 and GSE45995 data sets shared 4 genes (TaGlb2b, TaAKT1, TaMRP2, and omet). Pratiksha Singh *et al.* [27] reported that TaGlb2b showed a significant response to Erysiphe graminis and Fusarium graminearum.
The GSE27320, GSE31760, and GSE34445 data sets shared 1 gene (HSP101c) which has an important role in heat tolerance in hexaploid wheat [28]. The GSE27320 and GSE45995 data sets shared 13 genes (Tra2, Tra2:1, TAc23, zip1, TaGI1, PsbP, ctpA, Wcor726, pre-FBPase, lgul, GAPN, 6-FEH, and lgul:1). In a study by Yu Liu *et al.* to examine the response of PsbP to

Table 1. Information for the Six GEO datasets for wheat. Unique identifier for the dataset within GEO (accession), full name of the experiment (experiment name), shortened version of the experiment name (short name), ans the abbreviated version of the experiment name (abbrev).

accession	experiment name	short name	abbrev
GSE27320	Expression data in wheat (T. aestivum L.) near isogenic lines in response to powdery mildew infection	wheat responses to powdery mildew	WRPM
GSE31760	Transcription profiling wheat responses to adapted and non-adapted isolates of the blast fungus, Magnaporthe	wheat responses to blast fungus	WRBF
GSE32151	Lr1-mediated leaf rust resistance pathways of transgenic wheat lines	mediated leaf rust resistance	MLRR
GSE34445	Expression data from wheat following Hessian fly larval attack.	wheat following fly larval	WFFL
GSE45995	Transcriptomics of induced defense responses to Greenbug aphid feeding in near isogenic wheat lines	wheat responses to Greenbug aphid	WRGA
GSE59723	Transcriptional analyses of wheat responses to the necrotrophic effector SnTox3	wheat responses to SnTox3	WRST

Table 2. The SRING database table provides information about GO, annotated keywords, and protein domains for each GSE dataset.

GSE	Category	ID	Description	FDR
GSE27320	Molecular Function	GO:0005200	Structural constituent of cytoskeleton	0.0046
	Cellular Component	GO:0110165	Cellular anatomical entity	2.86E-24
	KEGG Pathways	map04145	Phagosome	0.0343
	Annotated Keywords	KW-0963	Cytoplasm	0.00025
		KW-0206	Cytoskeleton	0.0015
		KW-0493	Microtubule	0.0016
		KW-0342	GTP-binding	0.0065
		KW-0809	Transit peptide	0.0256
	Protein Domains	PF03953	Tubulin C-terminal domain	0.00055
		PF00091	Tubulin/FtsZ family, GTPase domain	0.0012
	Protein Features	IPR000217	Tubulin	2.43E-05
		IPR002453	Beta tubulin	2.43E-05
		IPR003008	Tubulin/FtsZ, GTPase domain	2.43E-05
		IPR008280	Tubulin/FtsZ, C-terminal	2.43E-05
		IPR013838	Beta tubulin, autoregulation binding site	2.43E-05
		IPR017975	Tubulin, conserved site	2.43E-05
		IPR018316	Tubulin/FtsZ, 2-layer sandwich domain	2.43E-05
		IPR023123	Tubulin, C-terminal	2.43E-05
		IPR037103	Tubulin/FtsZ, C-terminal superfamily	2.43E-05
		IPR036525	Tubulin/FtsZ, GTPase domain superfamily	2.71E-05
		IPR000877	Proteinase inhibitor I12, Bowman-Birk	0.0046
	Protein Domains	SM00269	Bowman-Birk type proteinase inhibitor	0.004
GSE31760	Cellular Component	GO:0110165	Cellular anatomical entity	0.001
GSE32151	Cellular Component	GO:0110165	Cellular anatomical entity	6.46E-37
	Annotated Keywords	KW-0325	Glycoprotein	3.38E-05
		KW-1015	Disulfide bond	0.00015
		KW-0732	Signal	0.00024
		KW-0809	Transit peptide	0.0023
		KW-0963	Cytoplasm	0.0023
		KW-0326	Glycosidase	0.0028
		KW-0378	Hydrolase	0.0062
		KW-0676	Redox-active center	0.0206
		KW-0119	Carbohydrate metabolism	0.0217
		KW-0624	Polysaccharide degradation	0.0386
	Protein Features	IPR000877	Proteinase inhibitor I12, Bowman-Birk	0.00055
	Protein Domains	SM00269	Bowman-Birk type proteinase inhibitor	4.17E-05
GSE34445	Cellular Component	GO:0110165	Cellular anatomical entity	1.42E-07
GSE45995	Cellular Component	GO:0110165	Cellular anatomical entity	1.53E-19
	Subcellular localization	GOCC:0005576	Extracellular region	0.0056
	Annotated Keywords	KW-0809	Transit peptide	0.0011
		KW-0150	Chloroplast	0.0022
		KW-0732	Signal	0.0039
		KW-0676	Redox-active center	0.0081
		KW-0597	Phosphoprotein	0.0241
		KW-0963	Cytoplasm	0.0338
	Protein Domains	SM00269	Bowman-Birk type proteinase inhibitor	0.004
GSE59723	Cellular Component	GO:0110165	Cellular anatomical entity	1.75E-08
	Annotated Keywords	KW-1015	Disulfide bond	0.00015
		KW-0732	Signal	0.00097
		KW-0325	Glycoprotein	0.0088
		KW-0326	Glycosidase	0.0214
	Protein Features	IPR000877	Proteinase inhibitor I12, Bowman-Birk	0.0199
	Protein Domains	SM00269	Bowman-Birk type proteinase inhibitor	0.0016

Table 3. The results of the machine learning model represent the most significant genes with their corresponding values.

Accsession	Marker	P/N	Value	Gene.Symbol
GSE27320	Ta.85.1.S1	positive	28944.5	PsbP
	Ta.1842.1.S1_a	positive	309.3	WPEAMT
	Ta.1848.2.S1	positive	4463.8	pip1:2
	Ta.2907.1.S1	positive	4340.4	LOC543101
	Ta.10.1.S1_a	positive	47759.5	LOC543334
	TaAffx.120000.1.S1	positive	465.7	TLK1
	Ta.10.2.S1	positive	63930.8	LOC543334
	Ta.21350.2.S1	positive	216.2	wrsi5-1
	Ta.21348.1.S1	positive	292.1	LOC543233
	Ta.23758.1.S1	positive	265.9	Wcor518
GSE31760	Ta.28.1.S1	negative	16942	LOC543330
	Ta.27762.1.S1	negative	3687.7	Ta-TLP
	Ta.24501.1.S1	negative	2852.3	LOC543292
	Ta.2788.1.A1	negative	768.65	1-SST
	Ta.82.1.S1	negative	8065.5	LOC543287
	TaAffx.115935.1.S1	negative	375.52	LOC543157
	Ta.87.1.S1	negative	654.58	pSBGer4
	Ta.24254.3.S1	negative	333.75	LOC606342
	Ta.25053.1.S1	negative	5220.3	LOC542887
	Ta.22673.1.S1_s	negative	14994	LOC543498
GSE32151	Ta.9320.1.S1	negative	11.7	CCoAMT
	Ta.9320.1.S1	negative	11.81	CCoAMT
	Ta.9402.1.S1	negative	8.97	LOC542918
	Ta.8629.1.A1	negative	5.9	WLTP1
	TaAffx.39351.2.A1	negative	9.64	HrBP1-1
	Ta.9402.1.S1	negative	8.92	LOC542918
	Ta.24806.1.S1	negative	10.79	WLIP19
	Ta.28700.1.S1	negative	4.55	CHS
	Ta.2907.1.S1	negative	11.76	LOC543101
	Ta.28734.1.S1	negative	15.08	TaGRP2
GSE34445	Ta.8614.1.S1	negative	46.58	WRKY45
	Ta.27312.1.S1	negative	63.75	AMT1
	Ta.4050.1.S1	negative	136.86	LOC100037581
	Ta.203.1.S1	negative	86.99	LOC543416
	Ta.1058.1.S1	negative	11290.8	SAMDC1
	Ta.192.1.S1	negative	79	LOC543235
	Ta.4678.2.S1	negative	61.95	WRKY71
	Ta.217.1.S1	negative	1033.27	LOC543244
	Ta.1058.3.S1	negative	15078.5	SAMDC1
	TaAffx.129134.2.S1	negative	281.44	a2b
GSE45995	Ta.28.1.S1	negative	37166.6	LOC543330
	Ta.9226.1.S1	negative	41349.6	PR4
	Ta.2784.1.A1	negative	42724.1	Chi 1
	Ta.14183.1.S1	negative	304.86	LOC543077
	Ta.21342.1.S1	negative	34995.6	Chi 3
	Ta.22871.1.S1_s	negative	13472.6	gamma-TIP
	TaAffx.39351.2.A1	negative	947.06	HrBP1-1
	Ta.27762.1.S1	negative	31690.6	Ta-TLP
	Ta.278.1.S1	negative	41378.7	LOC543422
	Ta.278.1.S1	negative	41179.1	LOC543422
GSE59723	Ta.5428.1.S1	negative	7.41	acT3
	Ta.8228.1.S1	negative	4.69	LOC543097
	TaAffx.128418.102.S1	negative	2.53	ltp9.2
	Ta.706.1.S1_s	negative	7.5	g6pdh
	Ta.81.1.S1	negative	6.86	pepc
	Ta.24501.1.S1	negative	10.2	LOC543292
	Ta.4725.1.S1	negative	5.08	WRKY53-b
	Ta.27762.1.S1	negative	7.64	Ta-TLP
	Ta.234.1.S1	negative	4.48	LCT1
	TaAffx.115935.1.S1	negative	5.11	LOC543157

Figure 1. Visualization of DEGs volcano plots.The representations are as follows: x-axis, log2FC; y-axis, -log10 of an adjusted p-value.

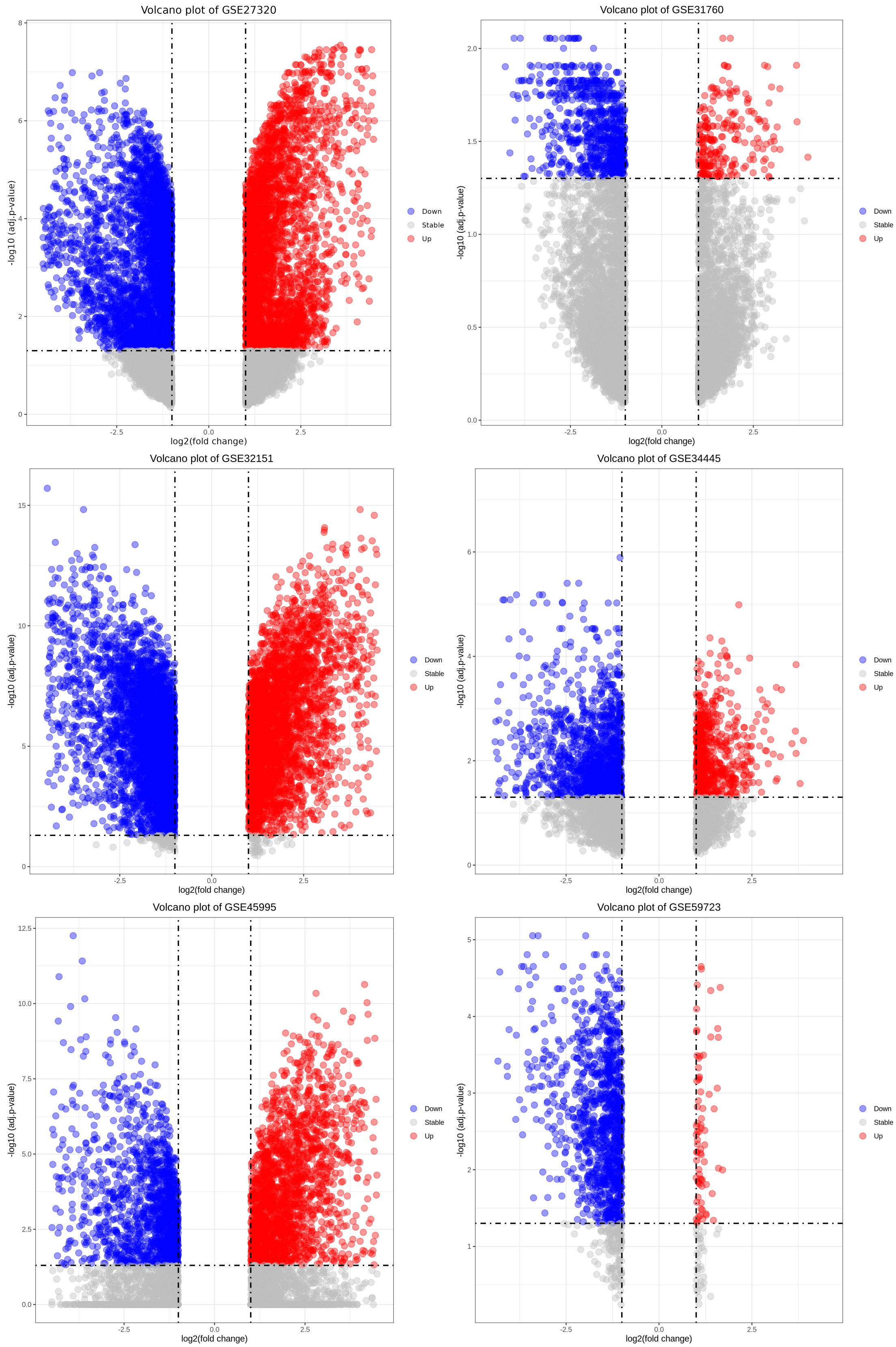


Figure 2. Heatmap and Venn diagram of the top DE genes data

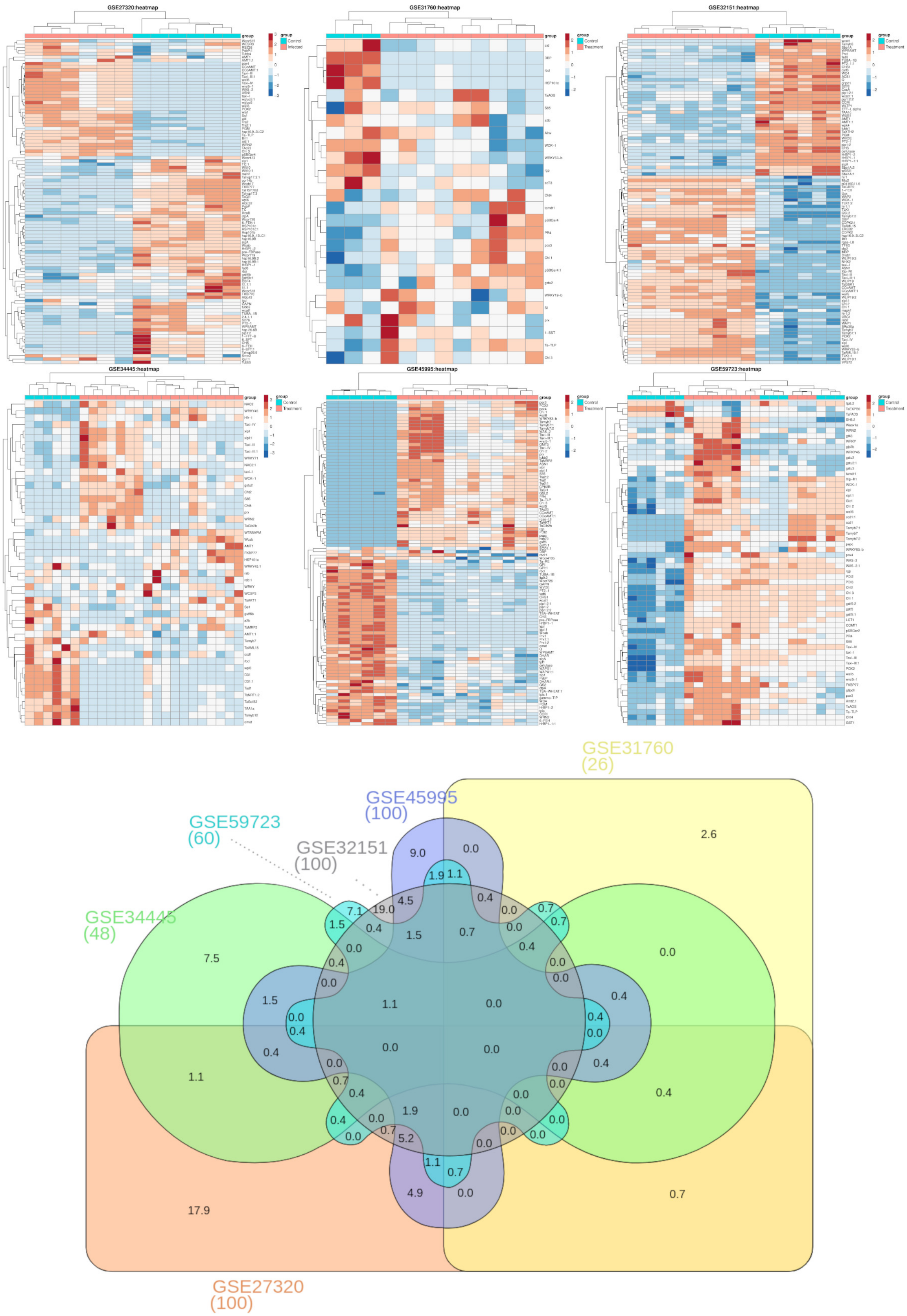


Figure 3. PPI networks show the interaction of DEGs.

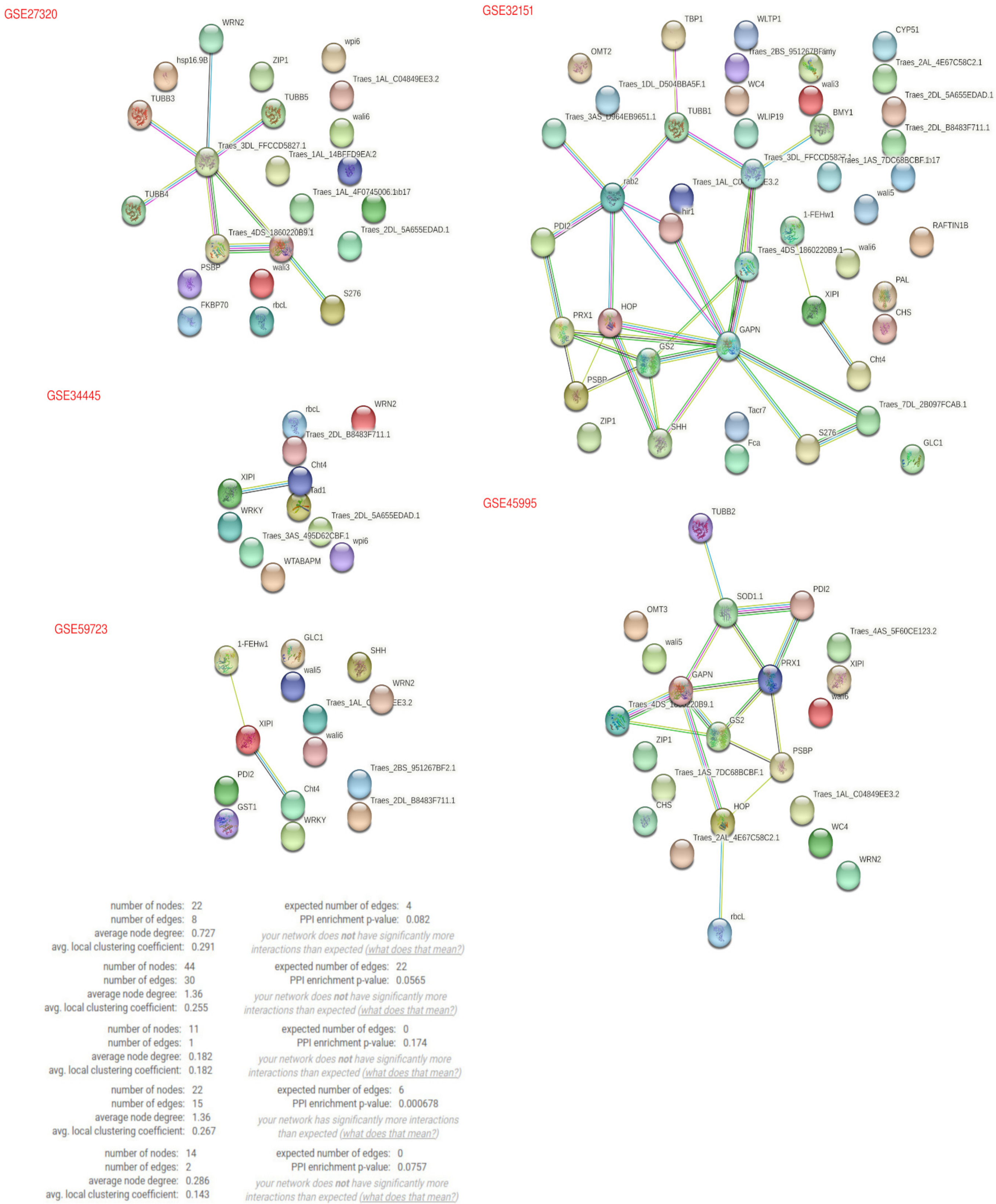
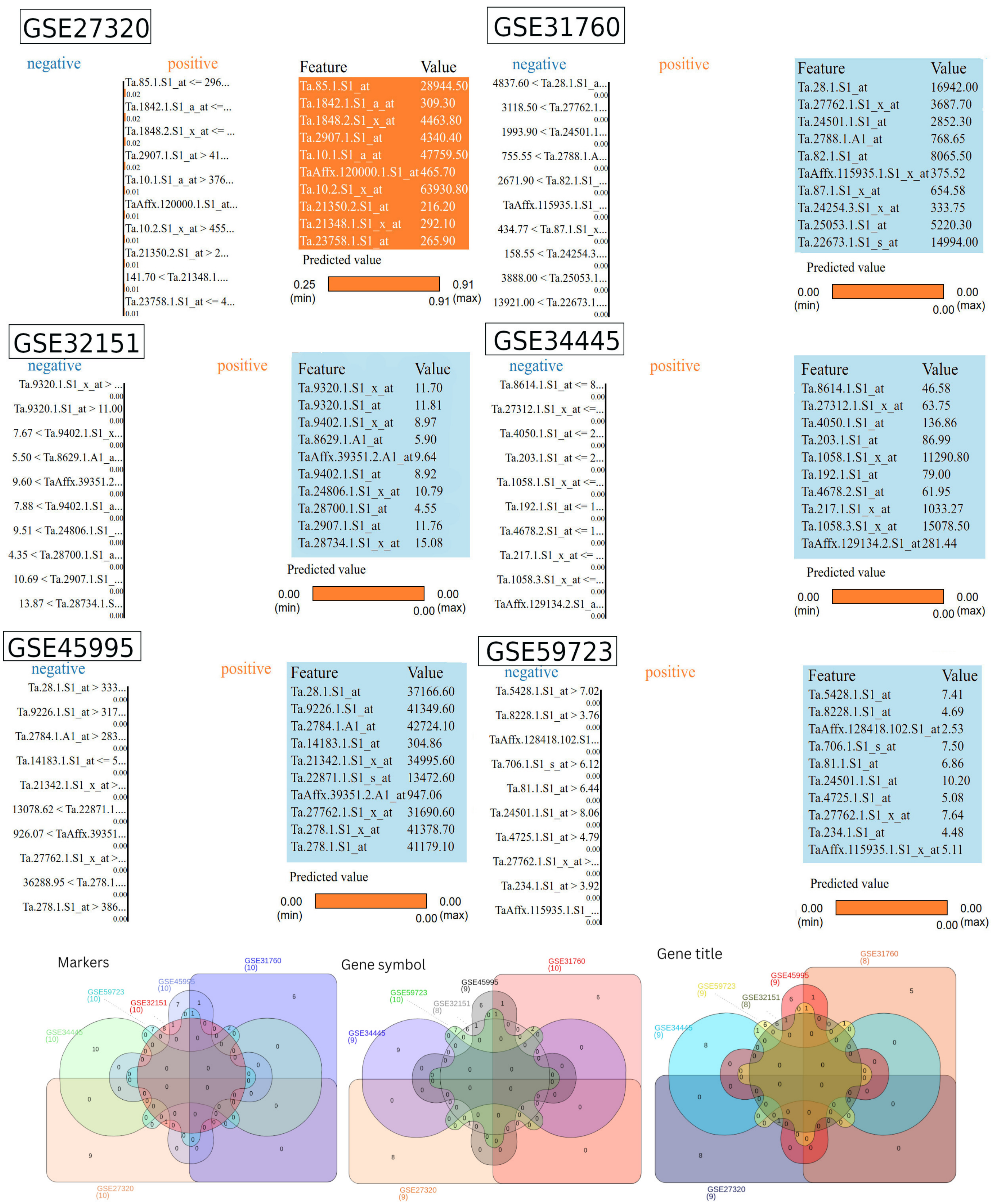


Figure 4. Machine learning results and venn diagrams of the markers, genes, and gene titles that were shared between GSEs.



wheat dwarf virus (WDV), they noticed that the PsbP gene was down-regulated under WDV infection [29].

The GSE27320, GSE45995, and GSE59723 data sets shared 3 genes (pox4, wrsi5-1, and WAS-2). The GSE27320, GSE31760, GSE45995, and GSE59723 data sets shared 2 genes (Ta-TLP and Chi 3).

The GSE27320, GSE32151, and GSE45995 data sets shared 14 genes, while the GSE27320, GSE32151, GSE45995 , and GSE59723 data sets shared 5 genes (Taxi-III, Taxi-III:1, wali6, Taxi-IV, and POX2). The GSE27320, GSE34445, and GSE45995 data sets shared 1 gene (Wcab) which was highly significant in the response to stripe rust in wheat [30]. The GSE27320, GSE31760, GSE34445, and GSE45995 data sets shared 1 gene (rbcl), in an investigation by Ceyda Ozfidan -Konakci *et al.*, the rbcl gene was observed at higher transcription levels in the chloroplasts of wheat as a response to hydrogen sulphide (H2S) and nitric oxide (NO) alleviate cobalt toxicity. The GSE31760, GSE34445, and GSE45995 data sets shared 1 gene (prx), and the GSE27320, GSE34445, and GSE59723 data sets shared 1 gene (FKBP77) which has a high expression under heat stress [31].

The GSE27320, GSE32151, GSE34445, and GSE59723 data sets shared 1 gene (taxi-I).

GSE32151, GSE34445, GSE45995, and GSE59723 data sets shared 3 genes (xipI:1, Tamyb7, and xipI). The Tamyb7 is one of the MYB family, which is considered to have a significant role in resisting stripe rust [32]. GSE27320, GSE34445, GSE45995, and GSE59723 data sets shared 1 gene (WRN2).

GSE31760, GSE34445, GSE45995, and GSE59723 data sets shared 1 gene (S85). The GSE31760, GSE32151, GSE34445, and GSE59723 data sets shared 1 gene (WCK-1), WCK-1 expression is reported to be induced by the fungal elicitor calcium ionophore A23187 as well as drought [33]. GSE31760, GSE34445, and GSE59723 data sets shared 2 genes (Cht4 and gstu2), Cht-4 expression was substantially faster in the resistant cultivar than in the susceptible one, in the response to Fusarium graminearum in wheat [34]. GSE32151, GSE45995, and GSE59723 data sets shared 4 genes (Tamyb7:2, wali5, Chi 2, and Tamyb7:1), Guozhang Kang *et al.* [35] discovered that freezing stress increased the expression of the wali5 protein gene. GSE31760, GSE32151, GSE45995, and GSE59723 data sets shared 2 genes (WRKY53-b and Chi 1).

The GSE31760, GSE32151, and GSE45995 data sets shared 1 gene (DBP), and the GSE27320, GSE32151, and GSE34445 data sets shared 2 genes (AMT1 and AMT1:1).

PPI network construction

Protein-protein interaction (PPI) networks were created using the STRING tool. Six modules were identified in this constructed network, which was made up of 115 nodes and 56 edges. The top significant module was GES32151, which was composed of 44 nodes and 30 edges, and had a PPI enrichment p-value of 0.0565. The lowest module was GSE34445, which was made up of only 11 nodes, 1 edge, and had a PPI enrichment p-value of 0.174 (Figure 3).

Machine learning

The machine learning model identified 53 markers, 48 genes, and 46 gene titles, as shown in (table 3). Also see (Figure 4). Each GSE accession consisted of ten markers. Only GSE27320 had positive values, ranging from 216.2 to 63930.8. The rest of the GSEs had negative values, as the table shows. Some of the GSEs shared the same markers, genes, and gene titles. According to the venn diagram (figure 4) of the markers, GSE32151 and GSE45995 shared TaAffx.39351.2.A1. GSE31760, GSE45995, and GSE59723 shared Ta.27762.1.S1. GSE31760 and GSE59723 shared Ta.24501.1.S1 and TaAffx.115935.1.S1. GSE27320 and GSE32151 shared Ta.2907.1.S1. GSE31760 and GSE45995 shared Ta.28.1.S1.

For genes, GSE32151 and GSE45995 shared HrBP1-1. GSE31760 , GSE45995 , and GSE59723 shared Ta-TLP. GSE31760 and GSE59723 shared LOC543292 and LOC543157. GSE27320 and GSE32151 shared LOC543101. GSE31760 and GSE45995 shared LOC543330. For gene titles, GSE32151 and GSE45995 shared harpin binding protein 1. GSE31760, GSE45995, and GSE59723 shared thaumatin-like protein. GSE34445 and GSE59723 shared WRKY transcription factor. GSE31760 and GSE59723 shared polyphenol oxidase. GSE27320 and GSE32151 shared methylmalonate semialdehyde dehydrogenase. GSE31760 and GSE45995 shared glucan endo-1 and 3-beta-D-glucosidase.

Danielle *et al.* [36] mention in their study that the Ta-TLP gene is a member of pathogenesis-related proteins and has an important role in plant resistance to powdery mildew. Another study by Zhi-Hui He *et al.* [37] reported that harpin-binding protein 1 is thought to be involved in plant disease resistance and drought resilience because it is responsible for plastid glutamine synthase (GS). It was shown to be down-regulated in wheat's single seed descent line 10 (SSDL 10). Furthermore, Thaumatin-like proteins (TLPs), as reported by Weibo Sun *et al.* [38], play roles in plant resistance to pathogen stress by acting as a positive factor in transgenic poplars with enhanced resistance to spots disease.

For WRKY transcription factors, Patel P *et al* [39] mentioned in their study that WRKY transcription factors are thought to have a role in contributing to the level of thermotolerance. It was reported [39] that polyphenol oxidase (PPO) has a role in plant resistance to osmotic stress-tolerant bacteria. They found that bacterized seedlings showed a slight improvement in wheat roots and shoots, and the polyphenol oxidase was greater in the roots and shoots. Hegedus *et al.* [40] reported that the overexpression of glucan endo-1,3-beta-glucosidase has been linked to a variety of physiological and developmental processes, as well as resistance to biotic and abiotic stress. This suggests that these genes may play an important role in plant resistance to various stresses and diseases.

Reference

1. Gimenez E, Salinas M, Manzano-Agugliaro F. World-wide research on plant defense against biotic stresses as improvement for sustainable agriculture. Sustainability. 2018;10(2):391.

2. Igrejas G, Branlard G. The importance of wheat. In: Wheat quality for improving processing and human health. Springer; 2020. p. 1-7.

3. Erenstein O, Jaleta M, Mottaleb KA, Sonder K, Donovan J, Braun HJ. Global Trends in Wheat Production, Consumption and Trade. In: Wheat Improvement. Springer; 2022. p. 47-66.

4. Hayta S, Smedley MA, Clarke M, Forner M, Harwood WA. An efficient Agrobacterium-mediated transformation protocol for hexaploid and tetraploid wheat. Current Protocols. 2021;1(3):e58.

5. Zhou Y, Zhao X, Li Y, Xu J, Bi A, Kang L, et al. Triticum population sequencing provides insights into wheat adaptation. Nature genetics. 2020;52(12):1412-22.

6. Poudel PB, Poudel MR. Heat stress effects and tolerance in wheat: A review. Journal of Biology and Today's World. 2020;9(3):1-6.

7. Jasrotia P, Kashyap PL, Bhardwaj AK, Kumar S, Singh G. Scope and applications of nanotechnology for wheat production: A review of recent advances. Wheat Barley Res. 2018;10(1):1-14.

8. Zhou G, Soufan O, Ewald J, Hancock RE, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic acids research. 2019;47(W1):W234-41.

9. Kumar Kushwaha S, Vetukuri RR, Odilbekov F, Pareek N, Henriksson T, Chawade A. Differential gene expression analysis of wheat breeding lines reveal molecular insights in yellow rust resistance under field conditions. Agronomy. 2020;10(12):1888.

10. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC bioinformatics. 2019;20(1):1-16.

11. Paul S, Duhan JS, Jaiswal S, Angadi UB, Sharma R, Raghav N, et al. RNA-Seq Analysis of Developing Grains of Wheat to Intrigue Into the Complex Molecular Mechanism of the Heat Stress Response. 2022.

12. Yang X, Tan B, Liu H, Zhu W, Xu L, Wang Y, et al. Genetic diversity and population structure of Asian and European common wheat accessions based on genotyping-by-sequencing. Frontiers in Genetics. 2020;11:580782.

13. Velu G, Crespo Herrera L, Guzman C, Huerta J, Payne T, Singh RP. Assessing genetic diversity to breed competitive biofortified wheat with enhanced grain Zn and Fe concentrations. Frontiers in Plant Science. 2019;9:1971.

14. Devesh P, Moitra P, Shukla R, Pandey S. Genetic diversity and principal component analyses for yield, yield components and quality traits of advanced lines of wheat. Journal of Pharmacognosy and Phytochemistry. 2019;8(3):4834-9.

15. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: mining millions of expression profilesdatabase and tools. Nucleic acids research. 2005;33(suppl_1):D562-6.

16. Zhang X, Wang Z, Hu L, Shen X, Liu C. Identification of potential genetic biomarkers and target genes of peri-implantitis using bioinformatics tools. BioMed Research International. 2021;2021.

17. Udhaya Kumar S, Thirumal Kumar D, Bithia R, Sankar S, Magesh R, Sidenna M, et al. Analysis of differentially expressed genes and molecular pathways in familial hypercholesterolemia involved in atherosclerosis: a systematic and bioinformatics approach. Frontiers in Genetics. 2020;11:734.

18. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. JoVE (Journal of Visualized Experiments). 2021;(175):e62528.

19. Sivasakthi P, Sabarathinam S, Vijayakumar TM. Network pharmacology and in silico pharmacokinetic prediction of Ozanimod in the management of ulcerative colitis: A computational study. Health Science Reports. 2022;5(1).

20. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research. 2021;49(D1):D605-12.

21. Liu GT, Wang BB, Lecourieux D, Li MJ, Liu MB, Liu RQ, et al. Proteomic analysis of early-stage incompatible and compatible interactions between grapevine and P. viticola. Horticulture research. 2021;8.

22. Duan Z, Lv G, Shen C, Li Q, Qin Z, Niu J. The role of jasmonic acid signalling in wheat (Triticum aestivum L.) powdery mildew resistance reaction. European journal of plant pathology. 2014;140(1):169-83.

23. Sun RJ, Xu Y, Hou CX, Zhan YH, Liu MQ, Weng XY. Expression and characteristics of rice xylanase inhibitor OsXIP, a member of a new class of antifungal proteins. Biologia plantarum. 2018;62(3):569-78.

24. Tundo S, Mandalà G, Sella L, Favaron F, Bedre R, Kalunke RM. Xylanase Inhibitors: Defense Players in Plant Immunity with Implications in Agro-Industrial Processing. International Journal of Molecular Sciences. 2022;23(23):14994.

25. Hu C, Chen P, Zhou X, Li Y, Ma K, Li S, et al. Arms Race between the Host and Pathogen Associated with Fusarium Head Blight of Wheat. Cells. 2022;11(15):2275.

26. Wang Q, Shao B, Shaikh FI, Friedt W, Gottwald S. Wheat resistances to Fusarium root rot and head blight are both associated with deoxynivalenol-and jasmonate-related gene expression. Phytopathology. 2018;108(5):602-16.

27. Singh P, Song QQ, Singh RK, Li HB, Solanki MK, Yang LT, et al. Physiological and molecular analysis of sugarcane (VarietiesF134 and NCo310) during Sporisorium scitamineum interaction. Sugar Tech. 2019;21(4):631-44.

28. Erdayani E, Nagarajan R, Grant NP, Gill KS. Genome-wide analysis of the HSP101/CLPB gene family for heat tolerance in hexaploid wheat. Scientific reports. 2020;10(1):1-17.

29. Liu Y, Liu Y, Spetz C, Li L, Wang X. Comparative transcriptome analysis in Triticum aestivum infecting wheat dwarf virus reveals the effects of viral infection on phytohormone and photosynthesis metabolism pathways. Phytopathology Research. 2020;2(1):1-13.

30. Lata C, Prasad P, Gangwar O, Adhikari S, Thakur RK, Savadi S, et al. Temporal behavior of wheat–Puccinia striiformis interaction prompted defense-responsive genes. Journal of Plant Interactions. 2022;17(1):674-84.

31. Vishwakarma H, Junaid A, Manjhi J, Singh GP, Gaikwad K, Padaria JC. Heat stress transcripts, differential expression, and profiling of heat stress tolerant gene TaHsp90 in Indian wheat (Triticum aestivum L.) cv C306. PloS one. 2018;13(6):e0198293.

32. Al-Attala MN. The role of Some MYB genes in defense response of wheat against stripe rust pathogen. Egyptian Journal of Desert Research. 2019;69(3):113-29.

33. Liu Y, Yu TF, Li YT, Zheng L, Lu ZW, Zhou YB, et al. Mitogen-activated protein kinase TaMPK3 suppresses ABA response by destabilising TaPYL4 receptor in wheat. New Phytologist. 2022;236(1):114-31.

34. Sorahinobar M, Niknam V, Jahedi A, Ebrahimzadeh H, Moradi B, Behmanesh M, et al. Application of sodium salicylate up-regulates defense response against Fusarium graminearum in wheat spikes. Biologia plantarum. 2019;63:690-8.

35. Kang G, Li G, Yang W, Han Q, Ma H, Wang Y, et al. Transcriptional profile of the spring freeze response in the leaves of bread wheat (Triticum aestivum L.). Acta physiologiae plantarum. 2013;35(2):575-87.

36. Ekom DCT, Benchekroun MN, Udupa SM, Iraqi D, Ennaji MM, et al. Wheat Genetic Transformation as Efficient Tools to Fight against Fungal Diseases. Journal of Agricultural Science and Technology A. 2015;5(3):153-61.

37. He ZH, Li HW, Shen Y, Li ZS, Mi H. Comparative analysis of the chloroplast proteomes of a wheat (Triticum aestivum L.) single seed descent line and its parents. Journal of plant physiology. 2013;170(13):1139-47.

38. Sun W, Zhou Y, Movahedi A, Wei H, Zhuge Q. Thaumatin-like protein(Pe-TLP)acts as a positive factor in transgenic poplars enhanced resistance to spots disease. Physiological and Molecular Plant Pathology. 2020;112:101512.

39. Patel P, Patil T, Maiti S, Paul D, Amaresan N. Screening of osmotic stress-tolerant bacteria for plant growth promotion in wheat (Triticum aestivum L.) and brinjal (Solanum melongena L.) under drought conditions. Letters in Applied Microbiology. 2022;75(5):1286-92.

40. Hegedus G, Kutasy B, Kiniczky M, Decsi K, Juhász Á, Nagy Á, et al.. Liposomal Formulation of Botanical Extracts may Enhance Yield Triggering PR Genes and Phenylpropanoid Pathway in Barley (Hordeum vulgare). Plants 2022, 11, 2969. s Note: MDPI stays neutral with regard to jurisdictional claims in published ; 2022.