Proteome Science

METHODOLOGY Open Access

Genome-wide computational analysis of the dirigent gene family in *Solanum lycopersicum*

Muhammad Abu Bakar Saddique[1,7], Ge Guan[1,2], Beibei Hu[1,2], Mudassir khan[3], Muhammad Dawood Amjad[4], Sana Abbas[5], Zahid Hussain[1], Muhammad Faizan Khurram Maqsood[6], Xiumei Luo[1,2*] and Maozhi Ren[1,2*]

Abstract

Background Dirigent (DIR) genes play a key role in the development of organic products in plants. They confer conformational influence on processes that lack stereoselectivity and regioselectivity through processes that are mostly understood. They are required to produce lignans, which are a unique and widely distributed family of plant secondary metabolites with intriguing pharmacological characteristics and potential role in plant development. DIR genes are implicated in the process of lignification and protect plants from environmental stresses, including biotic and abiotic stresses. Nevertheless, no research has been performed on the DIR gene family in *Solanum lycopersicum*. This study provides detailed information on the DIR gene family in *S. lycopersicum*.

Methods and results The conserved domain analysis, phylogenetic analysis, evolutionary adaptation, cis-acting elements, proteomic analysis, signal peptide detection, transmembrane potential analysis, sequence identity and similarity analysis, gene assembly, genomic localization, duplication of gene analysis, and evolutionary linkage of 31 potential DIR genes were studied. All these analyses provide a deep understanding of DIR genes in the *S. lycopersicum* genome that will provide a useful reference for further functional analysis of the DIR genes in *S. lycopersicum*.

Conclusion This research provides an in-depth and comprehensive explanation of the detailed process and structural characterization of DIR genes in the genome of *S. lycopersicum*, laying the groundwork for future plant genetic engineering and crop development exploration. This work will provide valuable information for identifying DIR genes in higher plants and support future research on the DIR gene family.

Keywords Dirigent (DIR), *Solanum lycopersicum*, Genome-wide analysis, Bioinformatics

*Correspondence:
Xiumei Luo
luoxiumei@caas.cn
Maozhi Ren
renmaozhi01@caas.cn
[1]Institute of Urban Agriculture, Chinese Academy of Agricultural Sciences, Chengdu National Agricultural2Science and Technology Center, Chengdu 610213, China
[2]School of Agricultural Science, Zhengzhou University, Zhengzhou 450001, China
[3]Department of Biomedical, Surgical and Dental Sciences, Università degli Studi di Milano, 20122 Milan, Italy
[4]Department of Medical, Oral and Biotechnological Sciences, Università degli Studi 'G. d'Annunzio' Chieti - Pescara, Chieti, Italy
[5]Department of Mathematics & Statistics, Pir Mehr Ali Shah Arid Agriculture University - PMAS AAUR, Rawalpindi, Pakistan
[6]Centre of Agricultural Biochemistry and Biotechnology (CABB), University of Agriculture, Faisalabad, Pakistan
[7]Department of Plant Biotechnology, National University of Sciences and Technology, Islamabad, Pakistan

Background

The name dirigent (DIR) originates from the Latin word dirigere, which means guide or align, and DIR protein was discovered for the very first time in *Forsythia intermedia* [1]. In the pharmacological direction of E-coniferyl alcohol stereospecific interactions, DIR proteins from *Forsythia suspensa* [2], mayapple [3], and western red cedar [4] produce the enantiomer pinoresinol, which is effectively used in the plant defense system. DIR along with the disease resistance response (DRR) gene family possess a DIR-conserved domain [5], and they are thought to regulate the oxygen radical bonding of monolignol plant phenolic compounds in plants to produce lignans and lignins [6]; hence, DIR is implicated in disease resistance adaptations [7].

Lignan has antifungal effects, both constitutively and inducibly, and is considered to be mainly involved in plant defense reactions [8]. Lignin accumulation is thought to act as a protective factor in the defensive reaction against microbial infection [9]. According to the study [10], a thicker morphology of leaf tissue includes lignin, which functions as a protective barrier against microbial attack. Lignin protects the recipient by acting as a nondegradable protective lining for pathogens. Lignin is indeed an essential chemical found largely in terminally specialized cells of supporting and water-conducting components, and it is principally involved as an exoskeleton, a waterway in the xylem, and insect and microbe defense [11].

Determining the roles of the DIR gene family in physiological and biological processes could be a useful approach for evaluating and enhancing crop defense responses against environmental stresses. Nevertheless, recently, no research has been performed on the DIR gene family of *S. lycopersicum*. As a result, the contribution of this work is to identify and analyze the DIR gene family of *S. lycopersicum* using a genome-wide strategy. Following computational analysis, a total of 31 DIR genes from the *S. lycopersicum* genome were retrieved. After that, promoter analysis, gene structure analysis, unique motif identification, tertiary configuration identification, chromosomal dispersal, gene duplication, synteny analysis, phylogenetic analysis, and potential membrane analysis were performed to study the DIR gene family in *S. lycopersicum*.

Methodology

Retrieval of dirigent sequences

The *S. lycopersicum* DIR genes were retrieved from the seed file (PF03018) (https://www.ebi.ac.uk/interpro/entry/pfam), and in the Phytozome13 (*S. lycopersicum* ITAG4.0) database (https://phytozome.jgi.doe.gov/) [12] against the *S. lycopersicum* genome with the default parameters. The *Arabidopsis thaliana* DIR genes were retrieved from The Arabidopsis Information Resource (TAIR) via the BLAST approach. The protein's conserved domain was validated using (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) the conserved domain database in an automated way after redundant and repetitive segments were eliminated once at a time [13, 14].

Structural characterization and phylogenetic evaluation

The ClustalW tool was used to conduct multiple sequence alignments of sequences. The protein sequences were analyzed with the ExPASy ProtParam tool to determine the protein coding sequence (CDS) length, total number of units, aliphatic index, instability index, grand average of hydropathicity (GRAVY), protein molecular weight, and hypothetical isoelectric point (http://web.expasy.org/protparam/). Using the online tool Gene Structure Display Server (GSDS), the exon/intron regions of specific DIR genes in *S. lycopersicum* were studied [15–17]. The tool MEME (http://meme-suite.org/tools/meme) was used with the parameters Zero or one occurrence per sequence (Zoops), a minimum width of motifs of six and a maximum of 50, and a total number of motifs per sequence of 15 to analyze the motifs of *S. lycopersicum* DIR genes, and the results were visualized using TBtools (TBtools v1.09854). Subcellular localization analysis was assessed using WoLF PSORT (https://wolfpsort.hgc.jp/) [18]. The WoLF PSORT findings were evaluated using a heatmap plot generated with the TBtools (v1.09854) [19] program. N-glycosylation sites (ASNs) of the DIR sequence were found online via the NetNGlyc 1.0 server (http://www.cbs.dtu.dk/services/NetNGlyc/).

To perform a phylogenetic analysis of the *S. lycopersicum* DIR gene family, MEGA X (http://www.megasoftware.net/)(20) was used. After aggregation and separation, the differentially expressed genes were divided into different subclasses, and a full dendrogram encompassing *Arabidopsis thaliana* and *S. lycopersicum* was constructed using MEGA X. ClustalW (http://www.ebi.ac.uk/clustalw/) [21] with general parameters was used to realign all of the segments first. Since not all *S. lycopersicum* DIR gene family members are exactly equivalent, gaps were removed, and a more cautious phylogenetic tree was built to improve the study's reliability. MEGA X was used to generate both phylogenetic trees, which were created using the neighbor-joining [22] approach, and bootstrapping tests, which involved 1000 repetitions [23].

Evolutionary adaptation and chromosomal localization

The *S. lycopersicum* DIR genes were mapped to the respective *S. lycopersicum* chromosomes using the Phenogram online program [24] (http://visualization.ritchielab.org/phenograms/plot).

The TBtools program (v1.09854; http://cj-chen.github.io/tbtools/) was used to retrieve chromosome sequence data, as it was used to locate all of the *S. lycopersicum* DIR genes according to their spatial relationship and chromosomal positions, as well as duplicated regions [25]. The ratios of Ks to Ka were calculated using the TBtools program with default settings (v1.09854). The divergence period of the gene pairs was calculated using the rate of substitutions per synonymous site per year. (Yuan et al., 2015) T = Ks/2x (x = 6.56 109).

Analysis of cis-acting elements in the promoter regions

The presence of cis-acting elements in the *S. lycopersicum* DIR gene family was studied using the TBtools program (http://cj-chen.github.io/tbtools/) [19]. The upstream region (up to 2 kb and 200 base pairs) of the *S. lycopersicum* DIR genes were retrieved from the *S. lycopersicum* genome and saved in FASTA file format. PlantCARE was used to submit and evaluate the data (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) [26].

Proteomic analysis and signal peptide prediction

The linkage associations of all the DIR family genes in *S. lycopersicum* were constructed by integrating all the DIR genes of *S. lycopersicum* into the freely accessible tool string database (https://string-db.org/) [27], with the following criteria: (i) minimum confidence: high (score: 0.07), and (ii) the maximum number of potential interconnections: 5, with all the unconnected sequences being eliminated. To illustrate consensus, all *S. lycopersicum* DIR genes in *S. lycopersicum* were mapped using the string database to evaluate their cooccurrence with other closely related taxa. The web hosting server Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) [28] was used to predict the tertiary conformations and homologs of *S. lycopersicum* DIR genes, as reported previously [29]. The presence of signal peptides and the locations of their cleavage sites within the given sequences were predicted using the SignalP 6.0 server online tool (https://services.healthtech.dtu.dk/services/SignalP-6.0/) [30].

Transmembrane potential analysis and sequence identity and similarity analysis

All-amino-acid alignments were evaluated in the TMHMM Server, v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) online program for analyzing the possible transmembrane mechanisms implicated in all *S. lycopersicum* DIR genes. To assess sequence similarity, all of the amino acid sequences were submitted to the online program SIAS (http://imed.med.ucm.es/Tools/sias.html) with the baseline model of the BLOSUM62 method with gap consequences.

(1) The cost of establishing the gap, Po (0-100), is 10.

(2) The cost of expanding the gap, Pe (0-100), is 50.

Synteny analysis

The DIR genes in *Arabidopsis thaliana* and *S. lycopersicum* were assessed using the Circoletto (http://bat.ina.certh.gr/tools/circoletto/) online program, which employs Circos to identify two sequence datasets. DIR genes of *S. lycopersicum* were aligned to *Arabidopsis thaliana* DIR genes. An E value of 10 to -40 (Strict) was used to perform a genome preservation assessment from total local pairings. The *S. lycopersicum* DIR protein sequences and *Arabidopsis thaliana* DIR protein sequences were utilized to search their respective protein repositories, with the best hits obtained based on the E value.

Results

Confirmation of DIR genes in *S. Lycopersicum*

The DIR sequences from the model organism *Arabidopsis thaliana* and the seed file (PF03018) were used as keywords in a BLASTp homology search against the *S. lycopersicum* genome to determine all possible closely related sequences in Phytozome v13 (https://phytozome-next.jgi.doe.gov). By eliminating other repetitive sequences, a maximum of 31 DIR genes, labeled *SlDIR1* to *SlDIR31*, were retrieved from the *S. lycopersicum* genome. All the *SlDIRs* (*S. lycopersicum* DIR genes) genes belonged to the DIR group according to the annotation of Phytozome v13 and also matched with their *Arabidopsis thaliana* orthologs. Subsequently, the existence of a preserved DIR motif was confirmed utilizing SMART and Pfam screening of all *SlDIRs*. TBtool was used to further confirm the presence of a DIR domain in all *SlDIRs* genes, as shown in Fig. 1.

Gene structural characterization, conserved motif analysis, and phylogenetic tree construction

The amino acid contents and stoichiometries of molecules in the *S. lycopersicum* DIR genes family are diverse, and the amount of compound proteins varies greatly among subclasses. The amino acid lengths of the *S. lycopersicum* DIR genes ranged from 60 (*SlDIR23*) to 399 (*SlDIR11*), with an average of 189 base pairs. The lowest molecular weight is 6302.19 kDa, and the maximum is 41413.18 kDa, with an average weight of 20715.22 kDa. The mean isoelectric point (pI) is 7.75, with scores ranging from 4.47 (*SlDIR11*) to 9.88 (*SlDIR11*) (*SlDIR3*). The pI is greater than 7 in 58% of the *S. lycopersicum* DIR gene family members, while it is less than 7 in the remaining genes. As a direct consequence of these findings, there are more basic DIR proteins than acid proteins. The presence of N-glycosylation (Asn) in each DIR sequence and other physiochemical properties can be seen in Table 1.

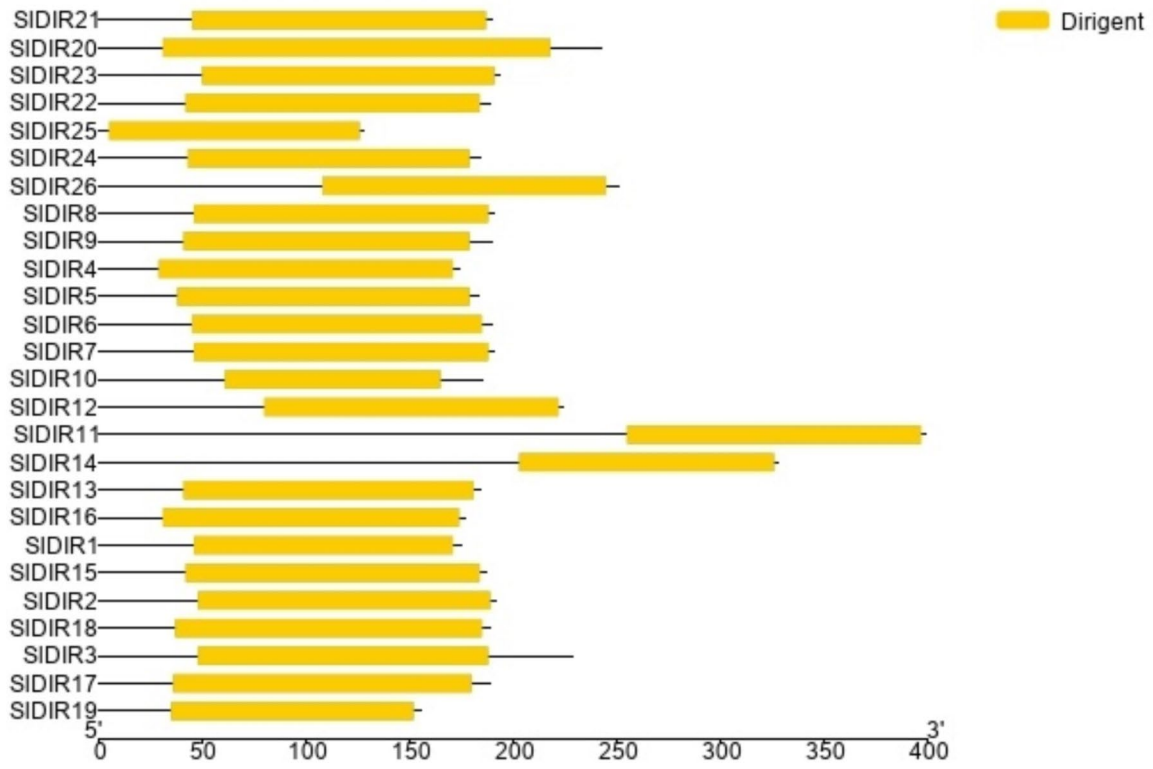

Fig. 1 Confirmation of DIR domains in all the *S. lycopersicum* retrieved sequences

The genomic and protein coding sequence (CDS) of the *S. lycopersicum* were studied, and the genetic makeup of their intron and exon structures were examined to determine how they work. According to the findings of the GSDS 2.0 software program (http://gsds.gao-lab.org/), only six of the 31 DIR genes (16%) had only one intron. Notably, 25 of the 31 proteins did not have introns. The genetic structure of all the genes is shown in Fig. 2. The WoLF PSORT results of subcellular localization were predicted using a heatmap from TBtool software. The highest probability values are highlighted in red, and the lowest probability values are highlighted in light blue color as shown in Fig. 3.

Members of the *S. lycopersicum* DIR genes family from the same subfamily have comparable motif types and quantities; however, there are variations in motif configurations across subfamily members. The precision of the phylogenetic analysis was improved by discovering comparable gene architectures and preserved domains within the same subfamily. Structural variations across subfamilies, on the other hand, imply that the DIR gene family in *S. lycopersicum* has functional variability. The genes and their respective motifs can be visualized in Fig. 4 which shows that each gene has its functionality depending upon the number of motifs present.

ClustalW was employed to evaluate the amino acid patterns of 31 *S. lycopersicum* DIR genes against 26 *Arabidopsis thaliana* DIR genes in addition to assessing the DIR gene family in model plants and *S. lycopersicum* from an evolutionary perspective and to investigate the unique features of the *S. lycopersicum* DIR genes. MEGA X, minimal evolution, and neighbor-joining (NJ) methods were used to study the phylogenetic relation. The *S. lycopersicum* DIR and *Arabidopsis thaliana* DIR genes were clustered together, suggesting that the *S. lycopersicum* DIR genes that can be categorized from *Arabidopsis thaliana* sequences are part of the same grouping. The DIR group of all these sequences may be categorized into seven subclasses (indicated by different colors) based on their similarity to DIR sequences in *Arabidopsis thaliana*, as shown in Fig. 5.

DIR genes promoter analysis

To advance the study of the putative biological responses of *S. lycopersicum* DIR genes during signaling, development, and endurance to abiotic and biotic stress feedback, PlantCARE [26] was utilized to evaluate cis-acting regions within the 2 kb upstream sequence and 200 base pairs upstream from each transcription start site of the *S. lycopersicum* DIR genes. Upon further investigation of the responsive parts of each gene, it was found that, as

Table 1 Physiochemical properties of all the DIR genes in *S. lycopersicum*

Gene ID	Protein Weight kDa	pI	Instability index	Aliphatic index	GRAVY	*N*-Glyc (Asn) position
SlDIR1	21482.33	9.7	26.84	105.05	0.268	132, 175
SlDIR2	10564.04	7.89	15.57	65.32	0.055	36
SlDIR3	26850.33	9.88	18.96	82.6	0.047	80, 233
SlDIR4	19875.05	8.41	18.64	95.84	0.138	9
SlDIR5	19662.64	8.89	7.76	100.22	0.237	65, 176
SlDIR6	21217.62	9.72	26.84	91.89	-0.094	69, 130, 174
SlDIR7	21300.6	9.27	26.28	91.36	0.038	70,131,175
SlDIR8	21403.73	8.78	27.9	91.41	0.002	70,87,131,175
SlDIR9	20738.68	8.85	21.46	90.21	0.027	51
SlDIR10	20640.84	9.61	32.67	98.16	0.009	31,79,178
SlDIR11	41413.18	4.47	39.97	86.02	0.037	70
SlDIR12	23782.26	4.98	30.11	88.75	-0.136	-
SlDIR13	20446.46	6.43	39.06	87.45	0.172	87,124
SlDIR14	33023.95	4.66	23.02	88.72	0.246	5
SlDIR15	20477.75	9.42	20.18	100.05	0.212	69,127
SlDIR16	19444.18	5.53	20.82	103.05	0.237	32,45,171
SlDIR17	21570.59	6.58	25.08	75.34	-0.058	52,65,122,140
SlDIR18	21138.1	6.5	24.51	89.26	0.034	20,55,68,125,143
SlDIR19	17623.43	6.54	20.8	98.08	0.207	47,50,53,58
SlDIR20	27770.79	9.19	38.95	87.1	0.073	10
SlDIR21	21448.92	9.58	32.77	90.26	0.047	24,36,59,69,130,173
SlDIR22	6437.34	6.51	7.38	74.67	-0.04	26
SlDIR23	6302.19	5.85	-3.46	82.83	0.16	26
SlDIR24	20785.85	8.76	18.16	83.54	0.029	56,66,92,127,170
SlDIR25	20861.97	6.64	14.86	82.89	0.01	70,126,169
SlDIR26	22663.37	9.59	16.37	80.87	0.019	1,83,109,126,144,187
SlDIR27	20924.08	9.3	15.81	100.05	0.128	74,99,100,135
SlDIR28	12082.81	8.79	21.01	70.46	-0.365	51,77
SlDIR29	20271.72	9.38	22.01	91.09	0.191	89,124,167
SlDIR30	13834.51	5.6	31.12	73.05	-0.006	-
SlDIR31	26132.8	5.07	29.65	92.47	0.2	-

also shown in Fig. 6a (2 kb base pairs) and Fig. 6b (200 base pairs), each gene seems to have a diverse range of activities in response to environmental stress as well as plant development, growth, and control.

Chromosomal location and gene duplication analysis

The Phytozome dataset v13 provided the chromosomal locations of all *S. lycopersicum* DIR genes. All DIR genes were physically allocated to their appropriate chromosomes by using the phenogram tool, as shown in Fig. 7. The 31 DIR genes were highly heterogeneous and dispersed throughout the *S. lycopersicum* genome on all the 12 chromosomes, excluding chromosome number 03, indicating that biological variability evolved during evolution. Nine genes were found on chromosome 10 (chr10). On the other hand, chromosomes 5, 8, 9, 11, and 12 contained the fewest DIR genes, with only one. Two DIR genes were found on three chromosomes (chromosome number 04, chromosome number 06, and chromosome number 07). Furthermore, chromosomes 2 and 8 contain 3 DIR genes.

The (synonymous rate) Ks, (non-synonymous rate) Ka, and Ka/Ks for these iterations were calculated, and the values were used to predict duplication divergence time. Throughout the genome, there is a wide variety of duplications. The ratio of Ka/Ks indicated that all of the values were less than one, implying that they were purified.

Ka/Ks = 1 indicates neutrality in the process of selection, while Ka/Ks > 1 indicates positive selection, and Ka/Ks < 1 indicates purifying selection. A Ka/Ks < 1 was found for all duplicated DIR gene pairs, indicating purifying selection throughout evolution, except for the *SlDIR2* and *SlDIR3* gene pairs, which indicate positive selection and are highly conserved throughout evolution. In addition, duplication events of duplicated gene pairs were predicted to have happened somewhere between 3.75 and 74.36 million years ago (Table 2).

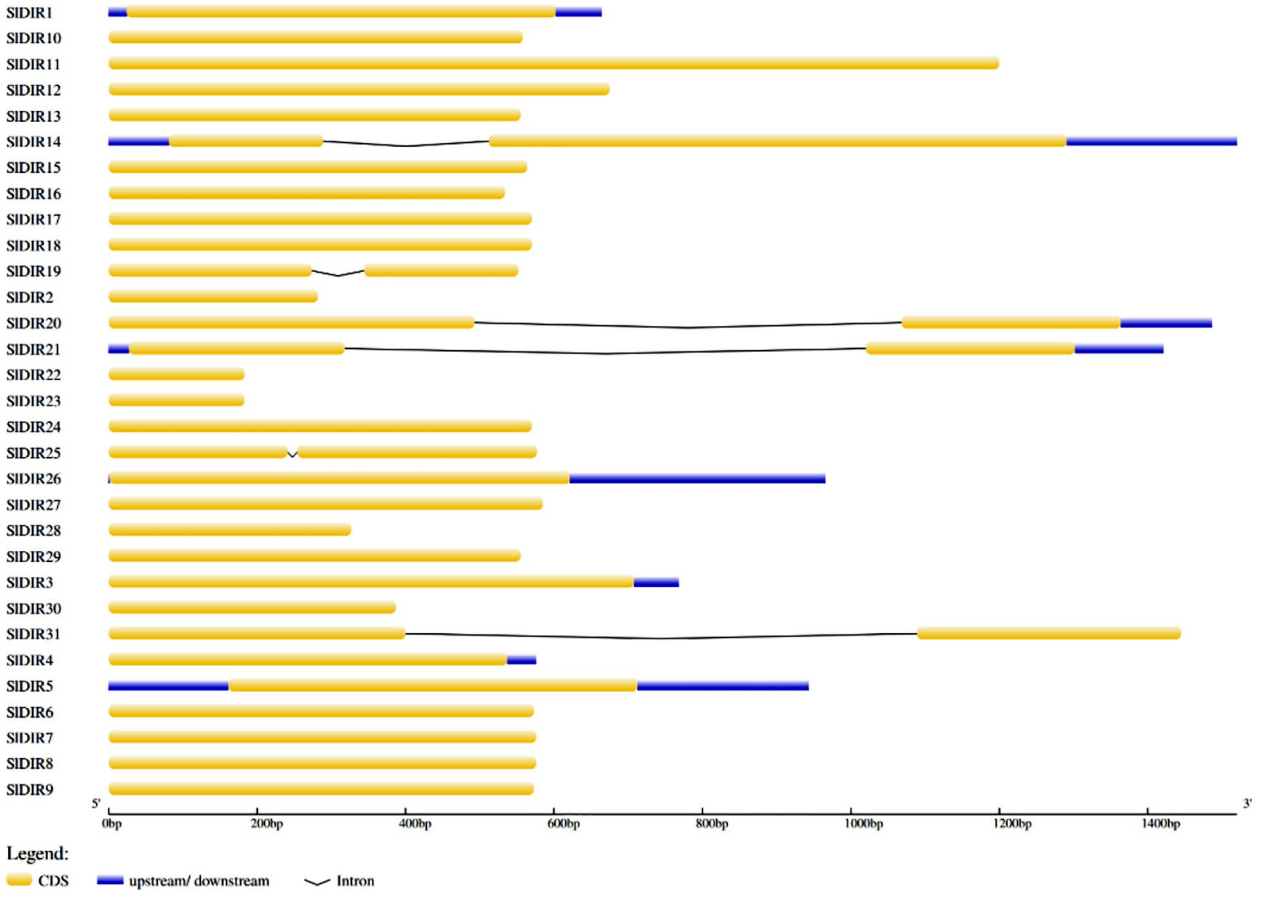

Fig. 2 The intron-exon structure of DIR genes, the exons are shown in light yellow, and the black curve line indicates an intronic region with a blue color indicating the upstream/downstream region

Protein-protein linkage association, signal peptide prediction, and coexpression analysis

To determine the importance of *S. lycopersicum* DIR proteins, data relating to proteins were obtained, and coexpression studies of these proteins with linked taxa were performed. The String Browser revealed significant associations among proteins at different stages. A preliminary shell of contact is observed throughout the intersection, as indicated by all the bright clusters. Figure 8b depicts the evolution, preservation, and coexpression of the differentially expressed proteins in a set of related taxa. It can be visualized from Fig. 8b that DIR sequences are preserved throughout the related taxa, black color indicates high preservation while light color indicates low. Figure 8a depicts the anticipated relationship of 31 *S. lycopersicum* DIR proteins that shows the established linkage among them. SignalP 6.0 was used to predict peptide signals, and the results are shown in Table 3. All the proteins were predicted to have signal peptides, except for nine proteins, *SlDIR1*, *SlDIR2*, *SlDIR4*, *SlDIR10*, *SlDIR12*, *SlDIR22*, *SlDIR23*, *SlDIR28*, *and SlDIR30*. The values of the cleavage site position and marginal probabilities for the signal peptide regions are also given in Table 3.

Synteny analysis and tertiary structure prediction

DIR proteins were analyzed to find orthologous pairs between *S. lycopersicum* and *Arabidopsis thaliana* to further deduce the evolutionary connection. According to synteny analysis, *S. lycopersicum* DIR genes and *Arabidopsis thaliana* DIR genes have collinear gene pairs. A Circos plot [31] was constructed to predict that *S. lycopersicum* DIR genes possess a high degree of evolutionary homology with *Arabidopsis thaliana*, indicating that they could have similar biological activities (Fig. 9). Within the circle, ribbons in four semitransparent colors—blue, green, orange, and red—show the local alignments generated by the BLAST approach. These colors correspond to the four quartiles up to the maximum score; that is, a local alignment scoring 80% of the maximum score is red, while one scoring 20% of the maximum score is blue.For the prediction of protein tertiary structure, the Phyre2 online tool was used, the results of which are shown in

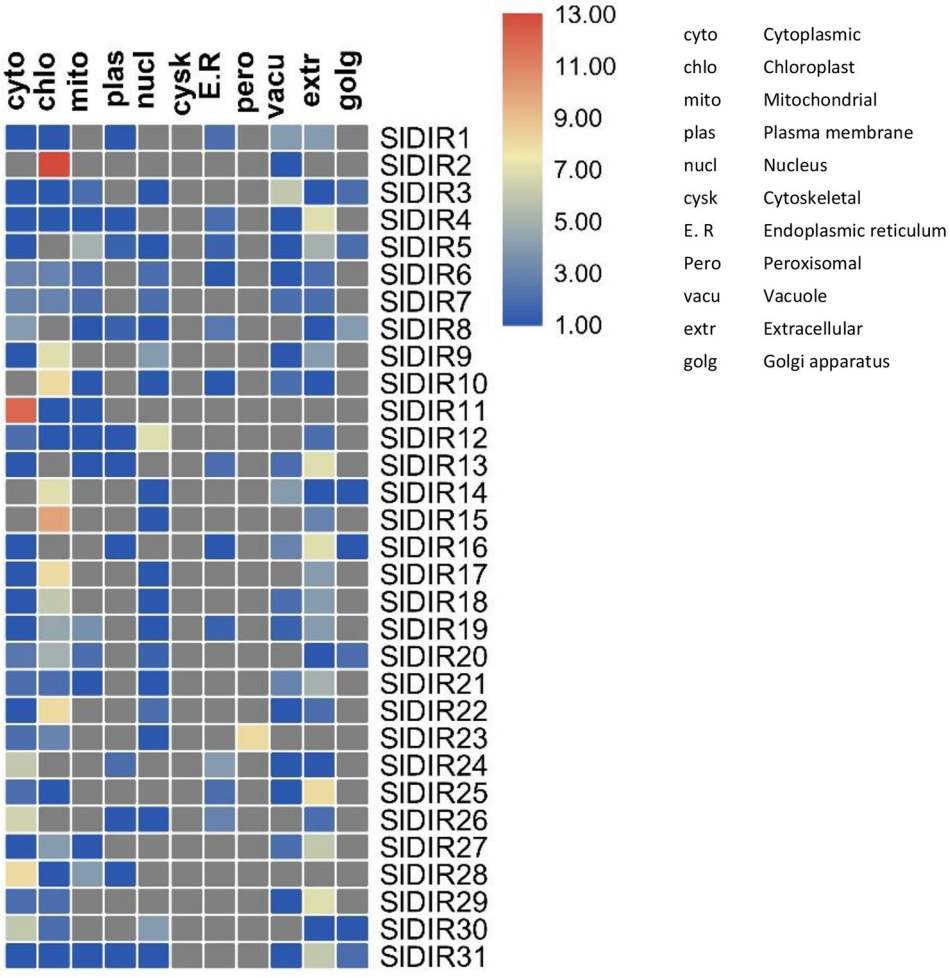

Fig. 3 Heatmap interpretation of the subcellular localization of *S. lycopersicum* DIR genes

Fig. 10. (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)

Transmembrane potential and sequence identity and similarity analysis

To test whether the potential transmembrane outcome is active for all of the studied sequences, TMHMM, an online tool, was used to search for all of the *S. lycopersicum* DIR genes. 15 DIR genes out of the 31 *S. lycopersicum* DIR genes were found to be involved in the essential functions of the cellular membrane. The SIAS assessment was used to determine the sequence distinctiveness, coherence, and global similarity. Table 4 presents common findings in a tabular arrangement. The resemblance and identification ability of *S. lycopersicum* DIR genes were greater than those of the global similarities based on the data (Table 4).

Discussion

Plants are constantly exposed to adverse environmental factors such as salt, drought, and cold, which have significant influences on geographical distribution, proliferation, growth, and plant production. Plants respond to external stresses through a complex set of resistance strategies that are activated and incorporated by the expression of thousands of genes [32]. Such stresses induce the expression of genes that were discovered to be important for plant stress resistance. These gene products act as regulatory proteins, increasing stress tolerance in plants and boosting plant immunity.

The study of transcriptional regulators is a diverse issue in the genetic sciences. Transcription factors influence a range of activities, including physical, biochemical, and evolutionary activities, as well as the activation routes of downstream genes. Numerous transcription factors that

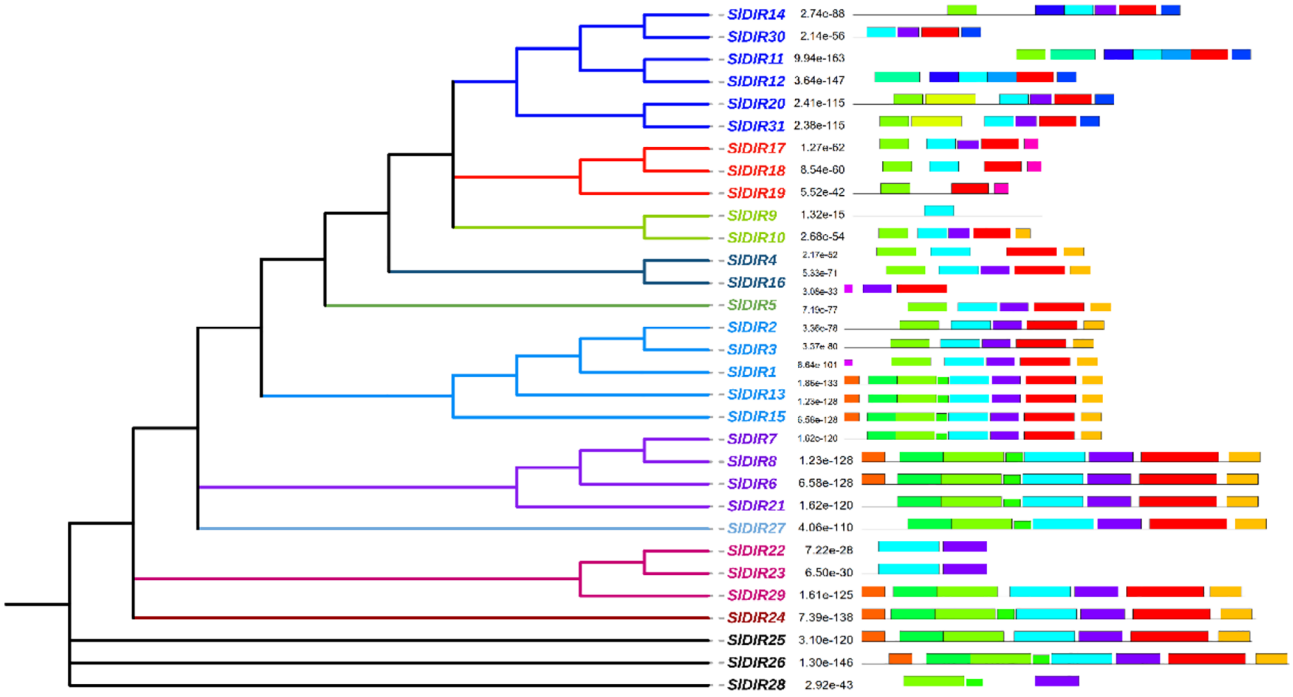

Fig. 4 Phylogenetic analysis and conserved motif analysis of *S. lycopersicum* DIR genes (**A**) Phylogeny of DIR genes via MEGA X with neighbor-joining methodology. (**B**) Different colors represent the various conserved motif domains of DIR genes in all *S. lycopersicum* DIR genes

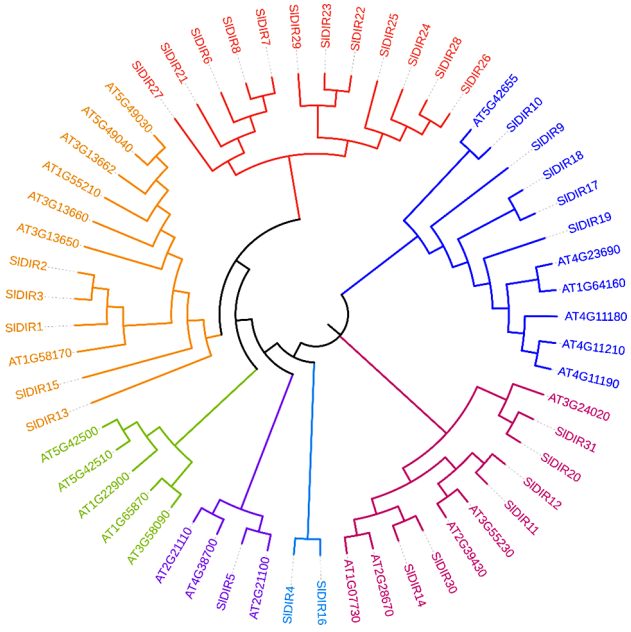

Fig. 5 Phylogenetic relationship of *S. lycopersicum* and *Arabidopsis thaliana* DIR genes

control dehydration, high salt, and other environmental variables have been identified in recent years. In addition, genomic data may be utilized to develop an interpretative layout for gene transformation technology that can be employed to generate highly resistant transgenic organisms.

DIR molecules are a multigene family in plants that respond to pathogen resistance [29, 33, 34]. They serve an essential function in improving stress tolerance in many crops [35]. The DIR genes are vital disease resistance-responsive genes that play a pivotal function in improving stress resistance across numerous plant species. The primary function of DIR, which are often higher oligomers or dimmers, is to protect plant tissues, particularly those involved in seed and heartwood production [34]. DIR and its homologs have been found in all vascular plants [36], and they are thought to play a role in lignin and lignan production [4]. Different plant species have different numbers of DIR genes. There have already been reports of 25 DIR genes in *Arabidopsis thaliana*, 19 DIR genes in *Isatis indigotica*, 54 DIR genes in rice, 35 DIR genes in *Picea glauca*, and 29 DIR genes in *Brassica rapa* [29, 35, 37, 38]. There has never been a genome-wide identification or characterization of the DIR gene family in *S. lycopersicum*.

In this study, we identified and explored via bioinformatics tools, a total of 31 DIR genes in the *S. lycopersicum* genome. Five well-conserved motifs were identified in the amino acid sequence alignments of all 31 *S. lycopersicum* DIR genes that show they have functional variability. Only 16% (6 out of 31) of the DIR had one intron and the rest of the *S. lycopersicum* DIR genes (25 out of 31) contained no intronic regions, according to the gene structural analysis. Like the previously examined DIR genes of *Arabidopsis thaliana* and poplar, the structure of the DIR

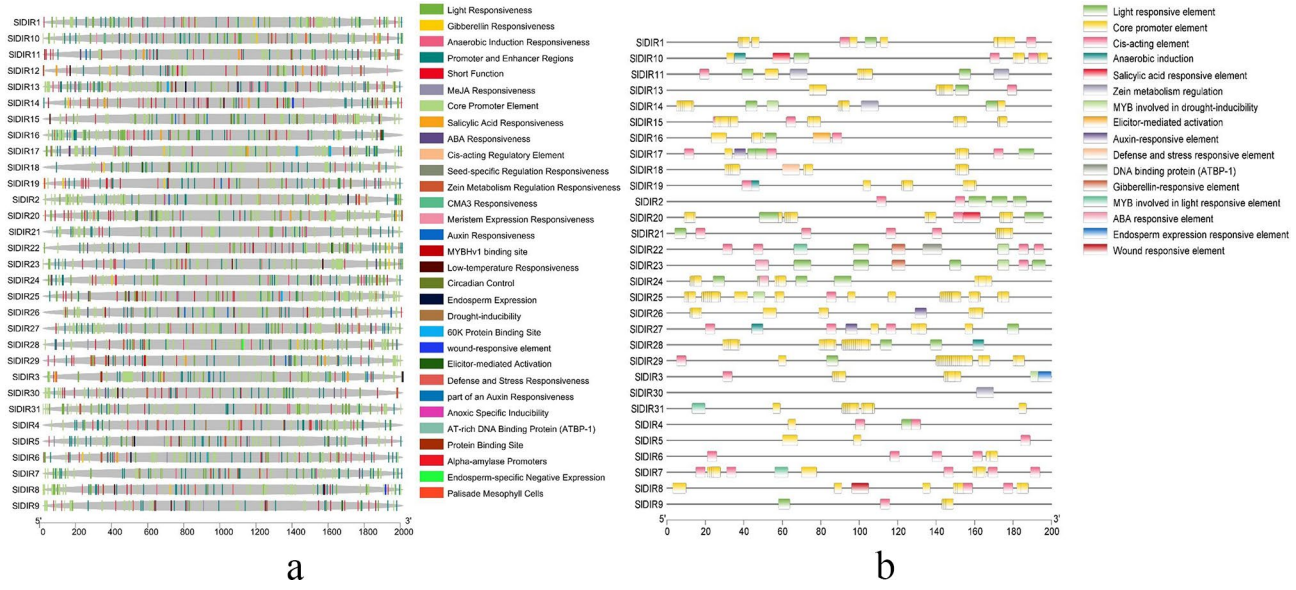

Fig. 6 (**a**) A comprehensive analysis of all *S. lycopersicum* DIR gene promoter analysis (up to 2 kb upstream). (**b**) A comprehensive analysis of all *S. lycopersicum* DIR gene promoter analysis (up to 200 bases upstream region)

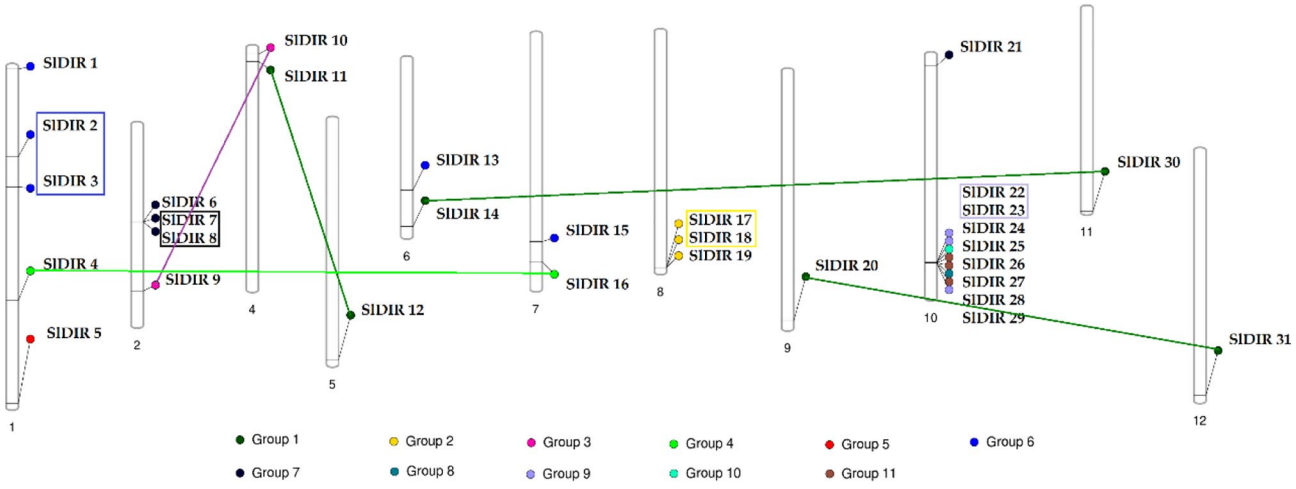

Fig. 7 Allocation of DIR genes across the *S. lycopersicum* genome

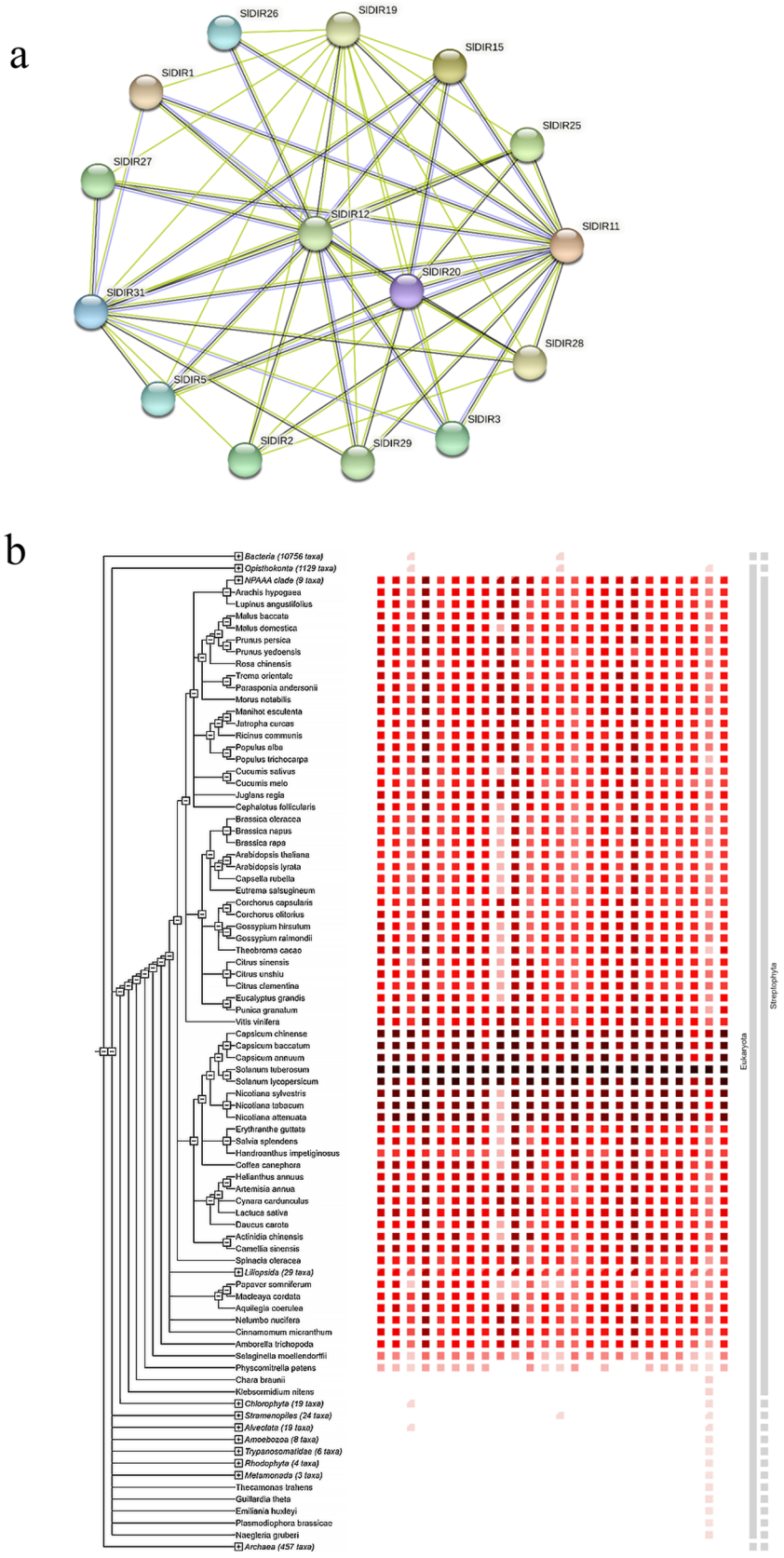

Fig. 8 (**a**) String database prediction of *S. lycopersicum* DIR genes. (**b**) STRING database depicts DIR genes co-expression in related taxa. Black color shows the highest expression in the *S. lycopersicum* genome vs. light color at a different scale

Table 2 Ka, Ks, and Ka/Ks calculations and divergence times of the duplicated *S. Lycopersicum* DIR gene pairs

Seq_1	Seq_2	Ka	Ks	Ka/Ks	T (MYA)
SlDIR14	*SlDIR30*	0.085935924	1.007212224	0.085321	6.549994
SlDIR11	*SlDIR12*	0.083597852	1.861744502	0.044903	6.371787
SlDIR20	*SlDIR31*	0.148327205	0.666594299	0.222515	11.30543
SlDIR17	*SlDIR18*	0.26100604	1.245689901	0.209527	19.89375
SlDIR9	*SlDIR10*	0.975700884	1.790588485	0.544905	74.36745
SlDIR4	*SlDIR16*	0.601743932	1.058155143	0.568673	45.86463
SlDIR2	*SlDIR3*	0.102436322	0.084828357	1.207572	7.807646
SlDIR7	*SlDIR8*	0.048911391	0.09014931	0.54256	3.728002
SlDIR22	*SlDIR23*	0.121149241	0.170426547	0.710859	9.233936

Table 3 Signal peptide prediction, cleavage site position, and marginal probabilities for the signal peptide regions of *S. Lycopersicum* DIR genes

Sr. no	Signal Peptide Prediction	Cleavage Site position	Marginal probabilities for signal peptide regions
SlDIR1	No	-	-
SlDIR2	No	-	-
SlDIR3	Yes	30–31	n-region, h-region, c-region
SlDIR4	No	-	-
SlDIR5	Yes	20–21	n-region, h-region, c-region
SlDIR6	Yes	22–23	n-region, h-region, c-region
SlDIR7	Yes	23–24	n-region, h-region, c-region
SlDIR8	Yes	23–24	n-region, h-region, c-region
SlDIR9	Yes	26–27	n-region, h-region, c-region
SlDIR10	No	-	-
SlDIR11	Yes	30–31	n-region, h-region, c-region
SlDIR12	No	-	-
SlDIR13	Yes	24–25	n-region, h-region, c-region
SlDIR14	Yes	30–31	n-region, h-region, c-region
SlDIR15	Yes	20–21	n-region, h-region, c-region
SlDIR16	Yes	22–23	n-region, h-region, c-region
SlDIR17	Yes	22–23	n-region, h-region, c-region
SlDIR18	Yes	26–27	n-region, h-region, c-region
SlDIR19	Yes	25–26	n-region, h-region, c-region
SlDIR20	Yes	36–37	n-region, h-region, c-region
SlDIR21	Yes	22–23	n-region, h-region, c-region
SlDIR22	No	-	-
SlDIR23	No	-	-
SlDIR24	Yes	19–20	n-region, h-region, c-region
SlDIR25	Yes	23–24	n-region, h-region, c-region
SlDIR26	Yes	36–37	n-region, h-region, c-region
SlDIR27	Yes	26–27	n-region, h-region, c-region
SlDIR28	No	-	-
SlDIR29	Yes	20–21	n-region, h-region, c-region
SlDIR30	No	-	-
SlDIR31	Yes	25–26	n-region, h-region, c-region

genes, which contain minimal introns, was also assessed in this study [13]. Nevertheless, 1–5 introns are found in one-third of the rice genome [34]. This implies that after divergence, rice, *S. lycopersicum*, poplar, and *Arabidopsis thaliana* may have divergent paths. The amino acid (aa) sequences varied from 60 aa (smallest) (*SlDIR23*) to 399 aa (largest) (*SlDIR11*), having an average of 189 base pairs. The lowest molecular weight is 6302.19 kDa, and the maximum is 41413.18 kDa, with an average weight of 20715.22 kDa. The mean isoelectric point (pI) is 7.75, with scores ranging from 4.47 (*SlDIR11*) to 9.88 (*SlDIR11*) (*SlDIR3*). The pI is greater than 7 in 58% of the *S. lycopersicum* DIR gene family members, while it is less than 7 in the rest of the genes, which shows that they are more basic.

Phylogenetic analysis revealed that the *S. lycopersicum* DIR genes and *Arabidopsis thaliana* DIR sequences are likely part of the same group since they were grouped together. Upon examining the proximity of these sequences to DIR sequences in *Arabidopsis thaliana*, it becomes apparent that there are seven distinct subclasses, each characterized by a unique color as shown in Fig. 5. According to findings via promoter analysis, it revealed that every gene seems to have a unique role in plant development, growth, regulation, and response to environmental stresses. This study supports previous research on cis-elements [15], with elements linked to stress and light being discovered in the upstream region, demonstrating that environmental stress and light may have a regulatory function in DIR genes. Furthermore, components sensitive to salicylic acid and methyl jasmonate have been found upstream of many *S. lycopersicum* DIR genes. Gibberellin-responsive domains were also found in the majority of the studied genes and are good signs of plant defense responses. Taken together, these findings suggest that the rhythms of the hormone responses of *S. lycopersicum* DIR genes are extremely complicated. Distinct DIR genes are considered to have multiple functions in a diverse range at various times; nevertheless, they still need to be examined further in the laboratory to demonstrate their functionality.

All *S. lycopersicum* DIR genes were physically allocated on their chromosomes. On all 12 *S. lycopersicum* chromosomes except chromosome 03, the 31 DIR genes were very diverse and scattered, showing biological diversity arose over evolution. Chromosome 10 had nine genes, and chromosomes 5, 8, 9, 11, and 12 have the fewest DIR genes, one. Three chromosomes 04, 06, and 07 had two

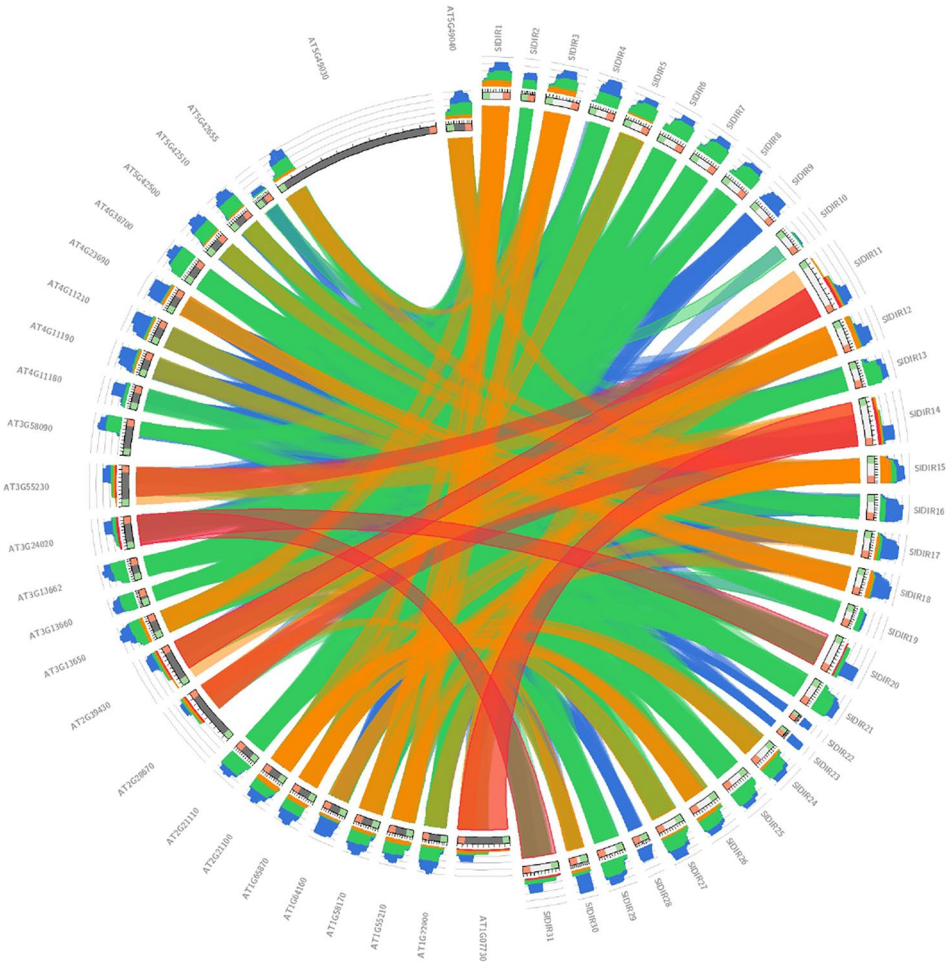

Fig. 9 Synteny map of all the identified *S. lycopersicum* DIR and *Arabidopsis thaliana* DIR genes

DIR genes. Furthermore, chromosomes 2 and 8 have 3 DIR genes. The duplication divergence time is predicted using the ka/ks ratio, which shows duplications are widespread throughout the genome. The Ka/Ks ratio showed that all values were smaller than one, indicating purification. Every duplicated DIR gene pair has a Ka/Ks < 1, showing purifying selection, except for the *SlDIR2* and *SlDIR3* gene pairs, which show positive selection and are highly conserved. Additionally, gene pair duplication events were anticipated to have occurred between 3.75 and 74.36 million years ago.

Based on synteny research, it has been shown that there are collinear gene pairings between *S. lycopersicum* genes and *Arabidopsis thaliana*. A Circos plot predicted that *S. lycopersicum* DIR genes have a significant level of evolutionary similarity with *Arabidopsis thaliana*, suggesting that they may have comparable biological functions. Our finding also shows the cooccurrence of the *S. lycopersicum* DIR genes family in different related taxa which shows their divergence. The Phyre2 tool was used to establish the prediction of protein tertiary structure. The String Browser revealed substantial correlations across proteins at various stages. An initial framework for interaction is seen at the junction. The presence of all these sequences indicates the conservation, preservation, and simultaneous expression of the differentially expressed

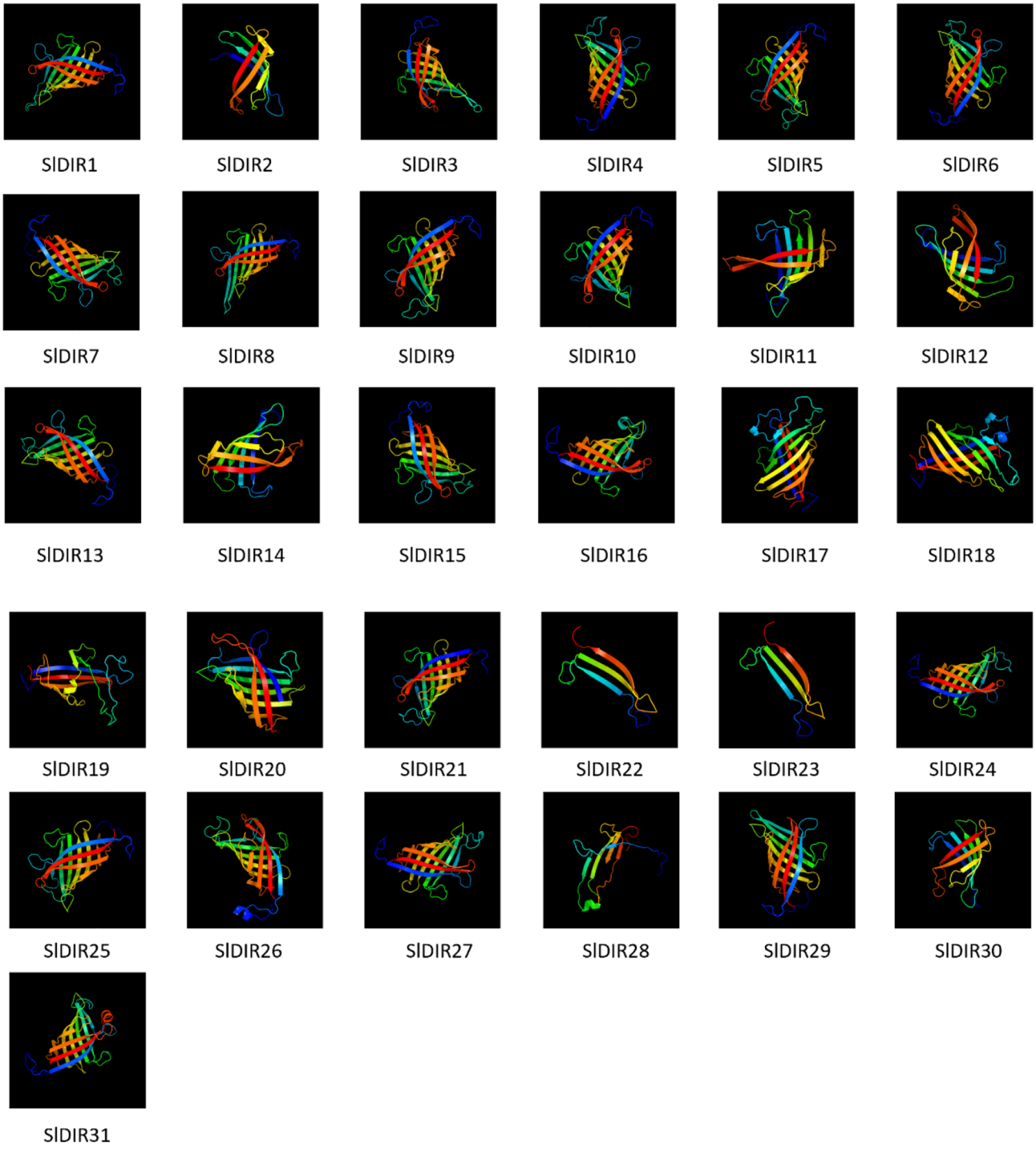

Fig. 10 Tertiary prediction of all *S. lycopersicum* DIR proteins

Table 4 Sequence analysis of *S. lycopersicum* DIR genes

Analysis	Minimum Value	Maximum Value	Mean Value	Standard Deviation
Identity Analysis	1.66	89.52	7.86	6.77
Similarity Analysis	3.33	100	17.25	16.59
Global Similarity (Blosum62)	-0.6	0.91	-0.15	0.12

proteins in a group of closely related organisms. In addition to shedding light on how these *S. lycopersicum* DIR genes carry out their roles, this work paves the way for future investigations into gene functional analysis.

Conclusion

It is possible to study plant species genomes using data analysis and evolutionary approaches. The environmental selection did not displace most of the *S. lycopersicum* DIR genes in the *S. lycopersicum* genome, but they did exhibit remarkable conservation throughout the evolutionary process. Future studies on members of the *S. lycopersicum* DIR gene family involved in the intricate network of plant growth and development will provide a useful reference for further functional analysis of the DIR gene family in *S. lycopersicum*.

Abbreviations

DIR	Dirigent
S. lycopersicum	*Solanum lycopersicum*
CDS protein	coding sequence

Author contributions

Conception and design: M.A.B.S, X.L, R.M. Development of methodology: M.A.B.S, M.K, M.D.A, Z.H, B.H, M.F.K.M. Analysis and interpretation of data: M.A.B.S, M.F.K.M, S.A, M.D.A, M.K. Writing of the manuscript: M.A.B.S, M.K, S.A, M.D.A, B.H, G.G. Figure preparation: M.A.B.S, M.F.K.M, S.A, Z.H. Study supervision: X.L, R.M.

Funding

This work was supported by the National Key R&D Program of China (2023YFE0199400), Central Public-interest Scientific Institution Basal Research Fund (No. Y2024QC33), Sichuan Science and Technology Program (2023YFQ0100), the Science and Technology Innovation Project of the Chinese Academy of Agricultural Sciences (No. 34-IUA-02).

Data availability

Various databases are used that are cited separately.

Declarations

Ethical approval

Not applicable

Competing interests

The authors declare no competing interests.

Received: 14 September 2022 / Accepted: 30 September 2024
Published online: 25 October 2024

References

1. Gang DR, Costa MA, Fujita M, Dinkova-Kostova AT, Wang H-B, Burlat V, et al. Regiochemical control of monolignol radical coupling: a new paradigm for lignin and lignan biosynthesis. Chem Biol. 1999;6(3):143–51.
2. Davin LB, Wang H-B, Crowell AL, Bedgar DL, Martin DM, Sarkanen S, Lewis NG. Stereoselective bimolecular phenoxy radical coupling by an auxiliary (dirigent) protein without an active center. Science. 1997;275(5298):362–7.
3. Xia Z-Q, Costa MA, Proctor J, Davin LB, Lewis NG. Dirigent-mediated podophyllotoxin biosynthesis in Linum flavum and Podophyllum peltatum. Phytochemistry. 2000;55(6):537–49.
4. Kim MK, Jeon J-H, Fujita M, Davin LB, Lewis NG. The western red cedar (Thuja plicata) 8–8′ DIRIGENT family displays diverse expression patterns and conserved monolignol coupling specificity. Plant Mol Biol. 2002;49:199–214.
5. Culley DE, Horovitz D, Hadwiger LA. Molecular characterization of disease-resistance response gene DRR206-d from Pisum sativum (L). Plant Physiol. 1995;107(1):301.
6. Burlat V, Kwon M, Davin LB, Lewis NG. Dirigent proteins and dirigent sites in lignifying tissues. Phytochemistry. 2001;57(6):883–97.
7. Wang Y, Fristensky B. Transgenic canola lines expressing pea defense gene DRR206 have resistance to aggressive blackleg isolates and to Rhizoctonia solani. Mol Breeding. 2001;8:263–71.
8. Lewis NG, Davin LB. Evolution of lignan and neolignan biochemical pathways. ACS; 1994.
9. Moerschbacher BM, Noll U, Gorrichon L, Reisener H-J. Specific inhibition of lignification breaks hypersensitive resistance of wheat to stem rust. Plant Physiol. 1990;93(2):465–70.
10. Fang Y, Mei H, Zhou B, Xiao X, Yang M, Huang Y, et al. De novo transcriptome analysis reveals distinct defense mechanisms by young and mature leaves of Hevea brasiliensis (para Rubber Tree). Sci Rep. 2016;6(1):33151.
11. Zhou J, Lee C, Zhong R, Ye Z-H. MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. Plant Cell. 2009;21(1):248–66.
12. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40(D1):D1202–10.
13. Khan A, Li R-J, Sun J-T, Ma F, Zhang H-X, Jin J-H, et al. Genome-wide analysis of dirigent gene family in pepper (Capsicum annuum L.) and characterization of CaDIR7 in biotic and abiotic stresses. Sci Rep. 2018;8(1):5500.
14. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, et al. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007;35(suppl1):D237–40.
15. Song M, Peng X. Genome-wide identification and characterization of DIR genes in Medicago truncatula. Biochem Genet. 2019;57:487–506.
16. Guo A-Y, Zhu Q-H, Chen X, Luo J-C. GSDS: a gene structure display server. Yi Chuan = Hereditas. 2007;29(8):1023–6.
17. Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics. 2015;31(8):1296–7.
18. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier C, Nakai K. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35(suppl2):W585–7.
19. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13(8):1194–202.
20. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547.
21. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics. 2003(1):2.3. 1-2.3. 22.
22. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985;39(4):783–91.
23. Gu Z, Cavalcanti A, Chen F-C, Bouman P, Li W-H. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol. 2002;19(3):256–62.
24. Wolfe D, Dudek S, Ritchie MD, Pendergrass SA. Visualizing genomic information across chromosomes with PhenoGram. BioData Min. 2013;6:1–12.
25. Poptsova MS, Gogarten JP. BranchClust: a phylogenetic algorithm for selecting gene families. BMC Bioinformatics. 2007;8:1–16.
26. Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, et al. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 2002;30(1):325–7.

27. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2016:gkw937.
28. Kelley LA, Sternberg MJ. Protein structure prediction on the web: a case study using the Phyre server. Nat Protoc. 2009;4(3):363–71.
29. Li Q, Chen J, Xiao Y, Di P, Zhang L, Chen W. The dirigent multigene family in Isatis Indigotica: gene discovery and differential transcript abundance. BMC Genomics. 2014;15:1–13.
30. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. 2022;40(7):1023–5.
31. Darzentas N. Circoletto: visualizing sequence similarity with Circos. Bioinformatics. 2010;26(20).
32. Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, et al. Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J. 2002;31(3):279–92.
33. Jin-Long G, Li-Ping X, Jing-Ping F, Ya-Chun S, Hua-Ying F, You-Xiong Q, Jing-Sheng X. A novel dirigent protein gene with highly stem-specific expression from sugarcane, response to drought, salt and oxidative stresses. Plant Cell Rep. 2012;31:1801–12.
34. Liao Y, Liu S, Jiang Y, Hu C, Zhang X, Cao X, et al. Genome-wide analysis and environmental response profiling of dirigent family genes in rice (Oryza sativa). Genes Genomics. 2017;39:47–62.
35. Ralph S, Park J-Y, Bohlmann J, Mansfield SD. Dirigent proteins in conifer defense: gene discovery, phylogeny, and differential wound-and insect-induced expression of a family of DIR and DIR-like genes in spruce (Picea spp). Plant Mol Biol. 2006;60:21–40.
36. Davin LB, Lewis NG. Dirigent proteins and dirigent sites explain the mystery of specificity of radical precursor coupling in lignan and lignin biosynthesis. Plant Physiol. 2000;123(2):453–62.
37. Arasan SKT, Park J-I, Ahmed NU, Jung H-J, Hur Y, Kang K-K, et al. Characterization and expression analysis of dirigent family genes related to stresses in Brassica. Plant Physiol Biochem. 2013;67:144–53.
38. Ralph SG, Jancsik S, Bohlmann J. Dirigent proteins in conifer defense II: extended gene discovery, phylogeny, and constitutive and stress-induced gene expression in spruce (Picea spp). Phytochemistry. 2007;68(14):1975–91.

Publisher's note