scientific reports

Check for updates

OPEN

De novo transcriptome assembly of *Dalbergia sissoo* Roxb. (*Fabaceae*) under *Botryodiplodia theobromae*-induced dieback disease

Ummul Buneen Zafar[1,2,7], Muhammad Shahzaib[1,2,7], Rana Muhammad Atif[2,3], Sultan Habibullah Khan[1,2,4], Muhammad Zeeshan Niaz[5], Khalid Shahzad[6], Nighat Chughtai[6], Faisal Saeed Awan[1], Muhammad Tehseen Azhar[3] & Iqrar Ahmad Rana[1,2,4]✉

Dalbergia sissoo **Roxb. (Shisham) is a timber-producing species of economic, cultural, and medicinal importance in the Indian subcontinent. In the past few decades, Shisham's dieback disease caused by the fungus *Botryodiplodia theobromae* has become an evolving issue in the subcontinent endangering its survival. To gain insights into this issue, a standard transcriptome assembly was deployed to assess the response of *D. sissoo* at the transcriptomic level under the stress of *B. theobromae* infection. For RNA isolation, the control and infected leaf tissue samples were taken from 1-year-old greenhouse-grown *D. sissoo* plants after 20 days of stem-base spore inoculation. cDNA synthesis was performed from these freshly isolated RNA samples that were then sent for sequencing. About 18.14 Gb (Giga base) of data was generated using the BGISEQ-500 sequencing platform. In terms of Unigenes, 513,821 were identified after a combined assembly of all samples and then filtering the abundance. The total length of Unigenes, their average length, N50, and GC-content were 310,523,693 bp, 604 bp, 1,101 bp, and 39.95% respectively. The Unigenes were annotated using 7 functional databases i.e., 200,355 (NR: 38.99%), 164,973 (NT: 32.11%), 123,733 (Swissprot: 24.08%), 142,580 (KOG: 27.75%), 139,588 (KEGG: 27.17%), 99,752 (GO: 19.41%), and 137,281 (InterPro: 26.72%). Furthermore, the Transdecoder detected 115,762 CDS. In terms of SSR (Simple Sequence Repeat) markers, 62,863 of them were distributed on 51,508 Unigenes and on the predicted 4673 TF (Transcription Factor) coding Unigenes. A total of 16,018 up- and 19,530 down-regulated Differentially Expressed Genes (DEGs) were also identified. Moreover, the Plant Resistance Genes (PRGs) had a count of 9230. We are hopeful that in the future, these identified Unigenes, SSR markers, DEGs and PRGs will provide the prerequisites for managing Shisham dieback disease, its breeding, and in tree improvement programs.**

Dalbergia sissoo Roxb. (Shisham), also known as the Indian Rosewood, belongs to the family *Fabaceae*. It is indigenous to the Indian subcontinent and is characterized by the leathery textures of its leaves, pink-white flowers, and the crookedness of the tree itself. The timber heartwood of *D. sissoo* is generally used for making high-grade furniture. Its stem bark has been found to contain dalbergichromene, a neoflavonoid of great significance[1]. In its native habitat, *D. sissoo* can be ranged under the threatened circle for being a high-value timber-producing species[2]. In terms of carbon sequestration, it has been observed that the trees of *D. sissoo* have an average sequestration capacity of 7.56 tha^{-1} (metric tons per hectare) with a carbon intake value of 27.735t

[1]Centre of Agricultural Biochemistry and Biotechnology, University of Agriculture, Faisalabad, Faisalabad 38000, Punjab, Pakistan. [2]Centre for Advanced Studies in Agriculture and Food Security, University of Agriculture, Faisalabad, Faisalabad 38000, Punjab, Pakistan. [3]Department of Plant Breeding and Genetics, University of Agriculture, Faisalabad, Faisalabad 38000, Punjab, Pakistan. [4]National Center for Genome Editing (Gene Editing of Biological Agents for Nutritional, Biochemicals and Therapeutic Purposes), University of Agriculture, Faisalabad, Punjab, Pakistan. [5]Plant Pathology Research Institute, Ayub Agriculture Research Institute, Faisalabad 38850, Punjab, Pakistan. [6]Punjab Forestry Research Institute, Faisalabad 37620, Punjab, Pakistan. [7]These authors contributed equally: Ummul Buneen Zafar and Muhammad Shahzaib. ✉email: iqrar_rana@uaf.edu.pk

CO_2 (metric tons CO_2)[3]. As a whole, *D. sissoo* and other *Dalbergia* tree species determine the extent of relative biodiversity in their native habitats, and they also have a significant economic and ecological impact. Due to these reasons, it is of foremost importance to establish the basis for strategies that can help in the conservation of *D. sissoo* under abiotic and biotic stresses[4]. The implementation of such programs can provide in-depth data on the precursors of resistance against the dieback disease. The genetic diversity analysis of *D. sissoo*, on the other hand, will help us ensure its survival under environmental stresses and conserve its precious natural resources, especially in the native regions[5,6].

In Pakistan, the timber wood harvested from *D. sissoo* plays a vital role in the economy of the country through the woodwork industry. Recently, a constant decline in the population of *D. sissoo* has been observed in the eleven districts of Punjab with a mortality rate of 25–30%. The fungus spp. *B. theobromae* is the main culprit associated with infecting both the underground as well as the aerial parts of the tree. In different studies, when inoculated with *B. theobromae*, the healthy *D. sissoo* plants produced typical symptoms of dieback disease such as wilting and yellowing of leaves, cankers of the stem, and dieback on the branches to various degrees[7,8]. Furthermore, as a countermeasure, the fungicides were experimented with in vitro and it was found that 100 ppm was the optimum concentration of Topsin-M in reducing the growth rate of fungal mycelia. Additionally, the same *D. sissoo* decline patterns have also been observed in different areas of Bangladesh[9]. Another study isolated the contaminated samples from different parts of the *D. sissoo* tree, cultured them on Potato Dextrose Agar (PDA) and Czapek dox agar media, and examined all of them simultaneously through microscopic observation. They found that *B. theobromae* was the most frequently isolated fungus from infection suggesting its strong association with the Shisham dieback[10].

Novel marker-assisted selection processes have shown unmatched potential for assessing the underlying genomic changes, identification of genes/germplasm for biotic and abiotic stress tolerance, genetic diversity, and natural variability among the tree species[11–16]. Several studies have reported molecular markers for *D. sissoo* germplasm such as RAPD and ISSR markers for genetic diversity analysis[17–23]. SSR markers have also proven to be very helpful in evaluating the structure and genetic diversity of plants[24–30]. The SSR markers have many significant comparative advantages such as their transferability to closely related species, reproducibility, abundance, co-dominance properties, and a relatively higher degree of subsequent polymorphism. SSR markers are generally divided into two categories based on their origin of derivation. First, there are conventional SSR markers identified from the genomic sequences, and second, there are Expressed Sequence Tag—Simple Sequence Repeats (EST-SSRs) which can be identified through transcribed RNA sequences[31]. These EST-SSR markers were previously identified and developed using traditional approaches. These approaches were costly, time-consuming, and laborious. Everything changed when transcriptome sequencing (RNA-seq) based on Next Generation Sequencing (NGS) was introduced. Various successful and comprehensive studies on SSR marker development have been performed on different tree species (no reference genome) through the utilization of NGS-based RNA-seq[32–34]. The transcriptomic analysis through the RNA-seq data can also be used to explore and mine the data for molecular and genetic breeding opportunities for threatened species through RNA-transcripts-based genome-wide analysis[35,36].

In this study, we present the first transcriptome of *Dalbergia sissoo* under the stress of *B. theobromae* infection assembled using the BGISEQ-500 sequencing platform. The BGISEQ-500 sequencing platform was used due to the superior relative factors such as cost-effectiveness, high throughput, short-read sequencing, rapid turnaround, ease of use, phasing information, versatility for different applications, and customization options. There is no report to our knowledge that has comprehensively identified SSR markers, Differentially Expressed Genes (DEGs), and Plant Resistance Genes (PRGs) in *D. sissoo* under the stress from *B. theobromae* infection, simultaneously. Somewhat related work is reported on the ornamental tree of Japanese apricot (*prunus mume*). In this study, 1,212 total DEGs and Differentially Methylated Regions (DMR) involved in influencing the biosynthesis of anthocyanins in the chimera of flower color were characterized[37]. Similarly, the DEGs identified in Masson pine (*Pinus massoniana*) and recently in pine wood (*Pinus thunbergia*) conferred resistance against the pine wood nematode that causes the Pine wilt disease[38,39]. Moreover, 1,573 DEGs responded to the drought stress in the mahaleb cherry (*Prunus mahaleb*)[40]. In Mexican Lime (*Citrus aurantifolia*), the DEGs that responded to the biotic stress from Huanglongbing-causing *Candidatus Liberibacter asiaticus* were also identified[41]. We aimed to expand the transcriptomic resource library available for *D. sissoo* through the inclusion of high-quality transcriptome sequencing data under the stress of *B. theobromae* infection. Most of the research on this disease to date is either on conventional plant pathogenic interactions or the use of RAPD and Est-SSRs identified in related species, no genomics tools developed endogenously are present. The findings of this research will help in filling this gap. The discovery of putative DEGs can especially help us understanding the significant disease-related processes like biomarker discovery, pathway-mediating biological processes, disease classification, and subtyping using the generated up- and down-regulated gene expression datasets. The numerous identified PRGs, SSR markers, and Unigenes in conjunction with the genomic data, on the whole, will help in ensuring the survival of *D. sissoo* under environmental stresses, the conservation of its germplasms, and facilitate the hybridization and breeding programs of the future.

Results

Sequence read filtering

Before the beginning of the downstream analysis, all the sequenced reads were filtered for low-quality, adaptor-polluted, and high content of unknown base (N) reads. In terms of clean read quality metrics, a total of 124.83 million (M) raw reads were obtained, out of which 120.95 M were clean reads (Table 1) (see Supplementary Figure S1). Moving on sample-wise, Rep1-Control had 30.91 M, Rep1-Disease contained 45.61 M, and Rep2-Disease had 44.43 M clean reads. Their Q20 percentage was 96.52%, 96.69%, and 96.81% respectively.

Sample	Total raw reads (M)	Total clean reads (M)	Total clean bases (Gb)	Clean reads Q20 (%)	Clean reads Q30 (%)	Clean reads ratio (%)
Rep1-Control	31.93	30.91	4.64	96.52	91.64	96.81
Rep1-Disease	47.33	45.61	6.84	96.69	92.23	96.38
Rep2-Disease	45.57	44.43	6.66	96.81	92.20	97.49

Table 1. Clean reads quality metrics.

BGISEQ-500 transcriptome sequencing, and de novo assembly

As this was a project without a reference genome, the reference sequence was obtained after the clean sequenced reads were assembled using Trinity[42] for subsequent analysis. In total, 1,000,549 individual transcripts and 513,821 Unigenes were obtained. They had a mean length of 483 bp (N50 = 818 bp) and 604 bp (N50 = 1,101 bp) respectively (Table 2) (see Supplementary Figure S2a). The number of Unigenes assembled for Rep1-Control was 205,077, 136,502 for Rep1-Disease, and 221,789 for Rep2-Disease (Table 3) (see Supplementary Table S1). The lengths of the Unigenes ranged from 1179 bp to 251,253 bp with 310,253,693 nucleotides in total (see Supplementary Figure **S2b**).

Functional annotation of Unigenes

After the completion of assembly, 7 functional databases were used to functionally annotate the Unigenes. These functional databases include NR, NT, GO, KOG, KEGG, SwissProt, and InterPro. Out of all the 513,821 Unigenes, 223,132 (43.43%) Unigenes were successfully annotated in at least one of the seven databases. Moreover, 52,244 (10.36%) of them showed annotations in all seven of the databases and the Unigene-associated functional pathway maps (Table 4) (see **Supplementary File S1**).

The NT and NR databases are the official nucleic acid and protein databases of the NCBI, respectively. In terms of NT functional annotations using BLAST (Basic Local Alignment Search Tool), NT constitutes 164,973 (32.11%) while NR had 200,355 (38.9%) functional annotations. The Unigene annotation ratio in the NR database was also calculated for different species (Fig. 1a). In terms of functional classification through KOG (euKaryotic Ortholog Groups), the calculations of 142,580 Unigenes were distributed among 25 functional groups. The largest group was the 'general function prediction only' group with 48,301 (33.87%) Unigene annotations. Following this, the signal transduction mechanism group had 22,486 (15.77%), post-translational modification, protein turnover, and chaperones group had 12,965 (9.09%), and the unknown function group had 12,669 (8.88%) Unigene annotations (Fig. 1b). The Unigenes that were successfully aligned to the NR database through KOG were then annotated using the GO (Gene Ontology)[43] database using Blast2GO[44]. The GO terms distribution was calculated using three categories viz. the biological process, cellular component, and molecular functions (Fig. 1c). In total, 99,752 (19.41%) Unigenes were annotated from the GO database. In the biological processes

Sample	Total number	Total length	Mean length	N50	N70	N90	GC (%)
Rep1-Control	354,390	146,146,365	412	467	267	197	37.88
Rep1-Disease	259,336	144,905,195	558	1,151	399	211	40.14
Rep2-Disease	386,823	197,303,972	510	836	344	210	40.77

Table 2. Quality metrics of transcripts.

Sample	Total number	Total length	Mean length	N50	N70	N90	GC (%)
Rep1-Control	205,077	108,527,100	529	817	345	232	38.47
Rep1-Disease	136,702	99,406,573	727	1,465	669	258	40.47
Rep2-Disease	221,789	151,540,658	683	1,285	530	261	40.95
All-Uni gene	513,821	310,523,693	604	1,101	418	242	39.95

Table 3. Quality metrics of Unigenes.

Values	Total	NT	NR	Swiss-Prot	KEGG	KOG	InterPro	GO	Intersection	Overall
Number	513,821	200,355	164,973	123,733	139,588	142,580	137,281	99,752	53,244	223,132
Percentage	100%	38.99%	32.11%	24.08%	27.17%	27.75%	26.72%	19.41%	10.36%	43.33%

Table 4. Annotation summary.

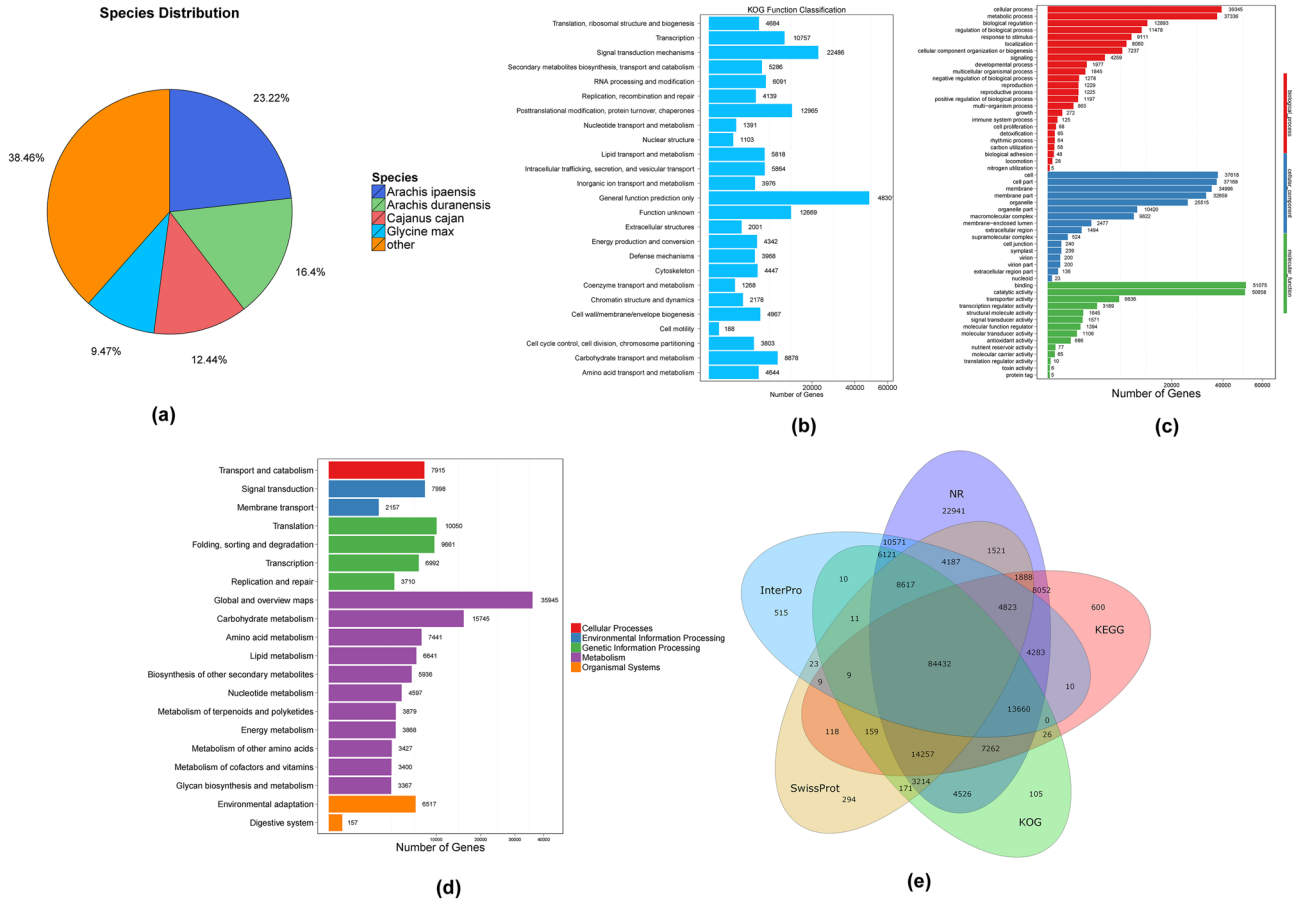

Figure 1. (**a**) Distribution of NR annotated species. (**b**) Functional distribution of KOG annotation. The X-axis represents the number of Unigenes and the Y-axis represents the KOG functional category. (**c**) Functional distribution of GO annotation. The X-axis represents the number of Unigenes and the Y-axis represents the Gene Ontology functional category. (**d**) Functional distribution of KEGG annotation. The X-axis represents the number of Unigenes and the Y-axis represents the KEGG functional category. KEGG metabolic pathway is categorized into 7 branches: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Disease, Metabolism, Organismal Systems, and Drug Development. (**e**) Venn diagram between NR, KOG, KEGG, Swissprot, and Interpro.

category, 39,345 (39.44%) Unigenes were annotated in the cellular process, 37,336 (37.42%) in the metabolic process, and 12,893 (12.92%) in biological regulation. Only 5 Unigenes were annotated in the biological process of nitrogen fixation. In the cellular components category, 37,618 (37.71%) were annotated in the cell, 37,168 (37.26%) were in the cell part, and 34,996 (35.08%) were in the membrane. Lastly, in the molecular functions category, 51,075 (51.20%) were annotated in binding, 50,658 (50.78%) in catalytic activity, and 6,636 (6.65%) in the transporter activity. Only 6 and 5 Unigenes were annotated in the molecular functions of toxin activity and protein tag, respectively.

The KEGG[45] database was also used to annotate and distribute the Unigenes at KEGG Level 1 and KEGG Level 2. In total, 139,588 Unigenes were annotated and all pathways were condensed into 5 clades (cellular processes, environmental information processing, genetic information processing, metabolism, organismal systems) and 20 subgroups (Fig. 1d). Among all the subgroups and clades, the 'global and overview maps' group from the Metabolism clade had 35,945 (25.75%) Unigenes. The 'translation' group from the genetic information processing clade had 10,050 (7.2%) Unigenes. Similarly, the 'signal transduction group' from the environmental information processing clade had 7,998 (5.73%) Unigenes and the 'transport and catabolism singular group' from the cellular processes clade had 7,915 (5.67%) Unigenes. Lastly, the 'environmental adaptation' group from organismal systems clade had 6,517 (4.67%) Unigenes. Additionally, the identified Unigenes were also aligned to the InterPro database using InterProscan software. Furthermore, the Unigene annotations were also performed through Swiss-Prot because the database is based upon the manually reviewed, high-quality annotated, and non-redundant protein sequences from UniProt Knowledgebase (UniProtKB). The annotation results from all the databases have also been illustrated using a Venn[46] diagrammatical representation (Fig. 1e).

The CDS (Coding DNA Sequences) were identified in the candidate coding regions using the TransDecoder software. The longest identified ORF (Open Reading Frame) was curated through BLAST[47] against hmmscan[48] and Swiss-Prot to predict CDS using Pfam protein homology sequences. The total number of identified CDS was 115,762 with a total length of 107,534,790 bp (Fig. 2a) (see Supplementary Table S2a–S2b). The minimum and

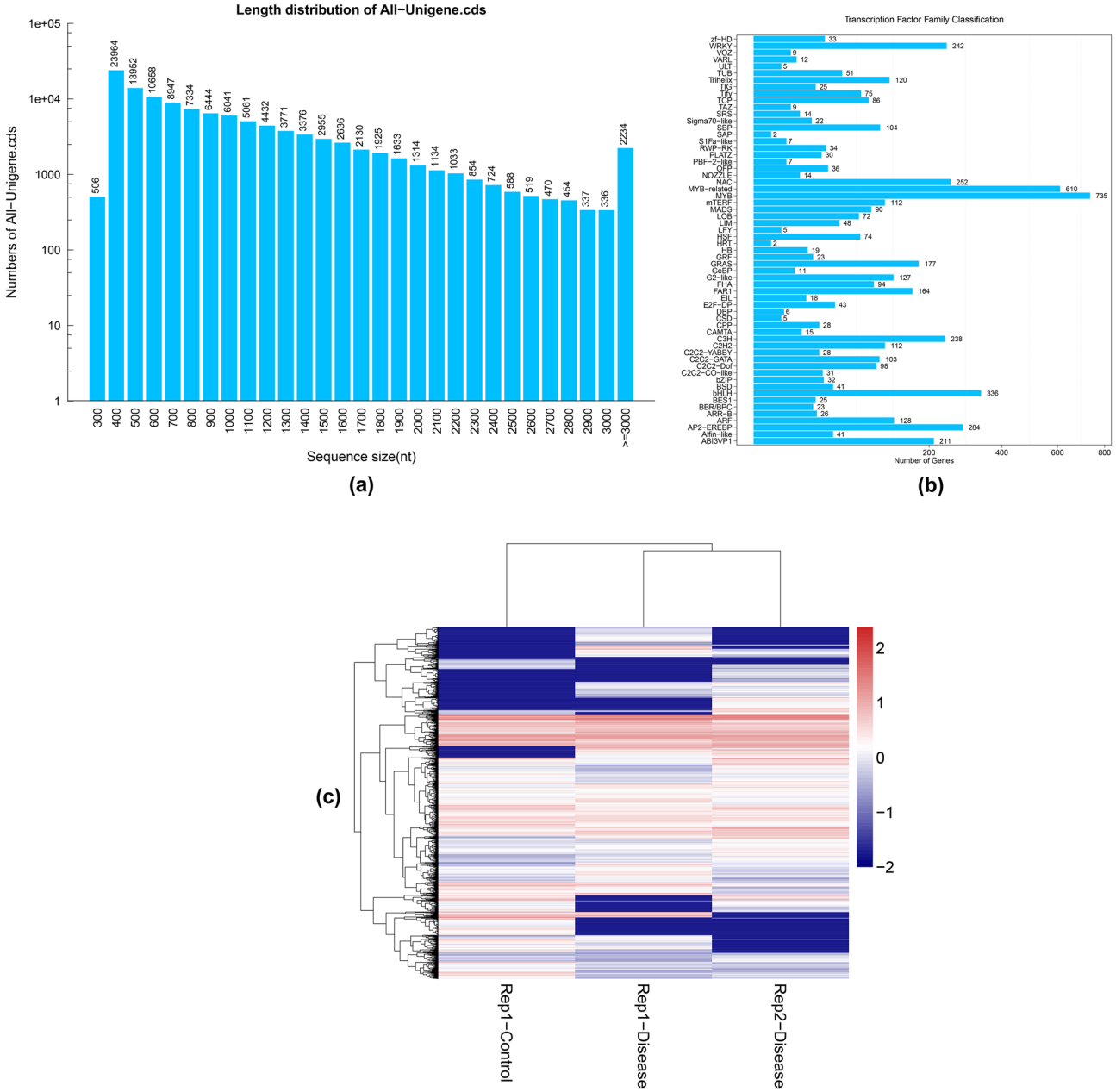

Figure 2. (**a**) CDS length distribution. The X-axis represents the length of CDS and the Y-axis represents the number of CDS. (**b**) Transcription Factor Family Classification of Unigenes. The X-axis represents the number of Unigenes and the Y-axis represents the family of TF. (**c**) Distribution of TF expression level. Each column represents a sample and each row represents a transcription factor.

maximum CDS lengths were 297 and 15,342 respectively. They also had an overall N50 of 1,209 bp with a GC percentage of 44.49% (see Supplementary Table S3). Similarly, a total of 4,673 Transcription Factor (TF) encoding genes were also predicted. The predicted Unigenes were classified into TF families. The most prominent families with their number of predicted Unigenes were MYB (735), MYB-related (610), bHLH (336), AP2-EREBP (284), and NAC (252) (Fig. 2b). The distribution of expression level of TFs among Rep1-Control, Rep1-Disease, and Rep2-Disease was also analyzed (Fig. 2c).

Unigene SSR identification and plant disease resistance gene identification

In terms of SSR markers, 62,863 of them were identified from 513,821 Unigenes (310,523,693 bp) with a distribution density of 202.44 SSRs per Megabase (Mb). In total, 51,508 Unigenes contained SSRs. The maximum number of bases interrupting 2 SSRs in a compound microsatellite was 100. Moreover, the number of sequences containing more than 1 SSR was 8809 and the number of SSRs present in compound formation was 4,175. In terms of repeat motifs, the most abundant of them were di-nucleotide (23,119, 36.78%), followed by tri-nucleotide (18,519, 29.46%) and mono-nucleotide (17,592, 27.98%). These three motif types were observed to be the most dominant constituting about 94.22% of the total. Only 1,410, 1,159, and 1,064 motif repeats were shown by

pentanucleotides, hexanucleotides, and quad-nucleotides respectively (see Supplementary Table S4). A total of 289 different repeat motifs were detected with the mono-nucleotide $(A/T)_n$ accounting for 26.95% of the total. Similarly, 16,944 motif repeats were followed by the di-nucleotides $(AG/CT)_n$ (11,867, 18.88%), $(AT/AT)_n$ (5,827, 9.27%), and $(AC/GT)_n$ (5,342, 8.50%). Moreover, the significant repeats among the tri-nucleotides were shown by $(AAG/CTT)_n$ (4,333, 6.89%) and $(AAT/ATT)_n$ (3,586, 5.70%) (see Supplementary Figure S6). Additionally, a total of 17,364 primers for each Unigene were also designed using Primer3 through the identified Unigene SSR sequences (see Supplementary Table S5). Furthermore, the plant disease resistance gene identification analysis was performed through the Plant Resistance Gene database (PRGdb)[49,50]. It revealed numerous disease-resistance genes in multiple species such as *Populus trichocarpa*, and *Brassica rapa* subsp. *Pekinensis*, and *Vitis vinifera* and more (see Supplementary Table S6).

Unigene expression

To assess the expression level of the genes, Bowtie2[51] was first used to assemble the clean reads into Unigenes as described before, and then RSEM[52] was used to calculate the gene expression level (see Supplementary Table S1). The results of an all-sample alignment showed that for 18,142,156,800 bases, the number of total reads was 120,947,712, with 77,457,276 (64.04%) total mapped reads, and 16,761,366 (13.86%) unique reads. The distribution of Unigene expression level was assessed by illustrating the dispersion of FPKM expression datasets through Box-plots[53] (Fig. 3a). In terms of dispersion and skewness, all the samples are almost in the same Inter Quartile Range (IQR) but vary in terms of their median. The deviation of the median between the control and disease samples showed different expression levels of the genes. Moreover, the genes in all the samples showed almost the same distribution of the FPKM expression outliers. The gene expression density graph with $\log_{10}$FPKM clearly shows the tendency of gene abundance changing with expression quantity. It also mirrors the concentration interval of relative gene expression in all samples (Fig. 3b). Comparatively, the Rep1-Control showed a slightly different second peak of expression density distribution which shows the difference in the level of relative gene expression. For a more intuitive representation of expression values, the intervals were created between different FPKM values (FPKM < = 1, FPKM: 1–10, FPKM > = 10), and the gene amount was calculated for each interval (Fig. 3c). Compared with Rep1-Control, for FPKM < = 1, Rep1-Disease and Rep2-Disease had 1.85- and 1.58-times higher expression levels, respectively. For FPKM: 1 ~ 10, Rep1-Disease and Rep2-Disease had 1.98- and 1.39-times lower expression levels, respectively. Lastly, For the FPKM > = 10, there were no significant differences between the control and disease samples. The higher and lower expression values at different FPKM intervals show the subsequent up and downregulation of genes.

Differentially expressed genes (DEGs)

Detection

The Differentially Expressed Genes or DEGs can be identified based on their expression level between different samples. For this purpose, the Poisson distance correction or distribution algorithm[54,55] was deployed using PoissonDis to detect DEGs. Rep1-Control sample was compared to both the Rep1-Disease and Rep2-Disease samples using $\log_2$FoldChange values of expression (see Supplementary File S2). In summary, the Rep1-Control vs. Rep1-Disease comparison group exhibited 15,532 up- and 19,079 down-regulated DEGs. Similarly, the Rep1-Control vs. Rep2-Disease group showed 16,504 up- and 19,981 down-regulated DEGs (Fig. 4a). This data has also been represented in terms of MA, scatter, volcano, and heatmap plots (Fig. 4b,c).

GO (gene ontology) analysis

The DEGs were classified according to GO functional enrichment terms into three subsequent categories: biological process, cellular component, and molecular function (see Supplementary File S3). In the Rep1-Control vs. Rep1-Disease group, the cellular and metabolic process GO terms were prominent in the biological process category. Similarly, the cell and cell part were prominent in the cellular component category, and binding and catalytic activity were prominent in the category of molecular function (Fig. 5a). The same trend of GO terms prominence was also observed in the Rep1-Control vs. Rep2-Disease group. The up- and down-regulated DEGs were also represented according to the classification of enriched GO terms and an almost equal distribution was observed in all the processes (Fig. 5b). Additionally, the GO functional enrichment DAGs (Directed Acyclic Graphs) in both the aforementioned groups of DEGs showed numerous significantly enriched pathways based on the calculated p-values.

Pathway and protein-interaction network analysis

KEGG[45] database was used to classify the DEGs according to their functional enrichment and pathways. Like the pathway classification and distribution patterns of Unigenes, DEGs in both Rep1-Control vs. Rep1-Disease and Rep1-Control vs. Rep2-Disease group showed a similar trend with 5 individual clades (cellular processes, environmental information processing, genetic information processing, metabolism, and organismal systems) (Fig. 6a). The KEGG pathway enrichment is based on the intermittent rich factor and Q-value[56,57]. The Rep1-Control vs. Rep1-Disease group showed 'Photosynthesis – antenna proteins', Monobactum biosynthesis, and Isoflavonoid biosynthesis pathway as the major ones. Similarly, in Rep1-Control vs. Rep2-Disease group, 'Circadian rhythm—plant', Synthesis and degradation of ketone bodies, 'Photosynthesis – antenna proteins', Vitamin B6 metabolism, and Isoflavonoid biosynthesis pathway were the prominent ones (Fig. 6b). The KEGG pathway maps for each DEG give extensive detail about the underlying pathways and their mechanisms (see Supplementary File S4). Additionally, the enriched KEGG pathways were also represented in terms of both up and downregulated DEGs (Fig. 6c). The three major and significant pathways were Plant hormone signal transduction, Ribosome, and mRNA surveillance pathway. In these pathways, the first one had more down-regulated DEGs while the

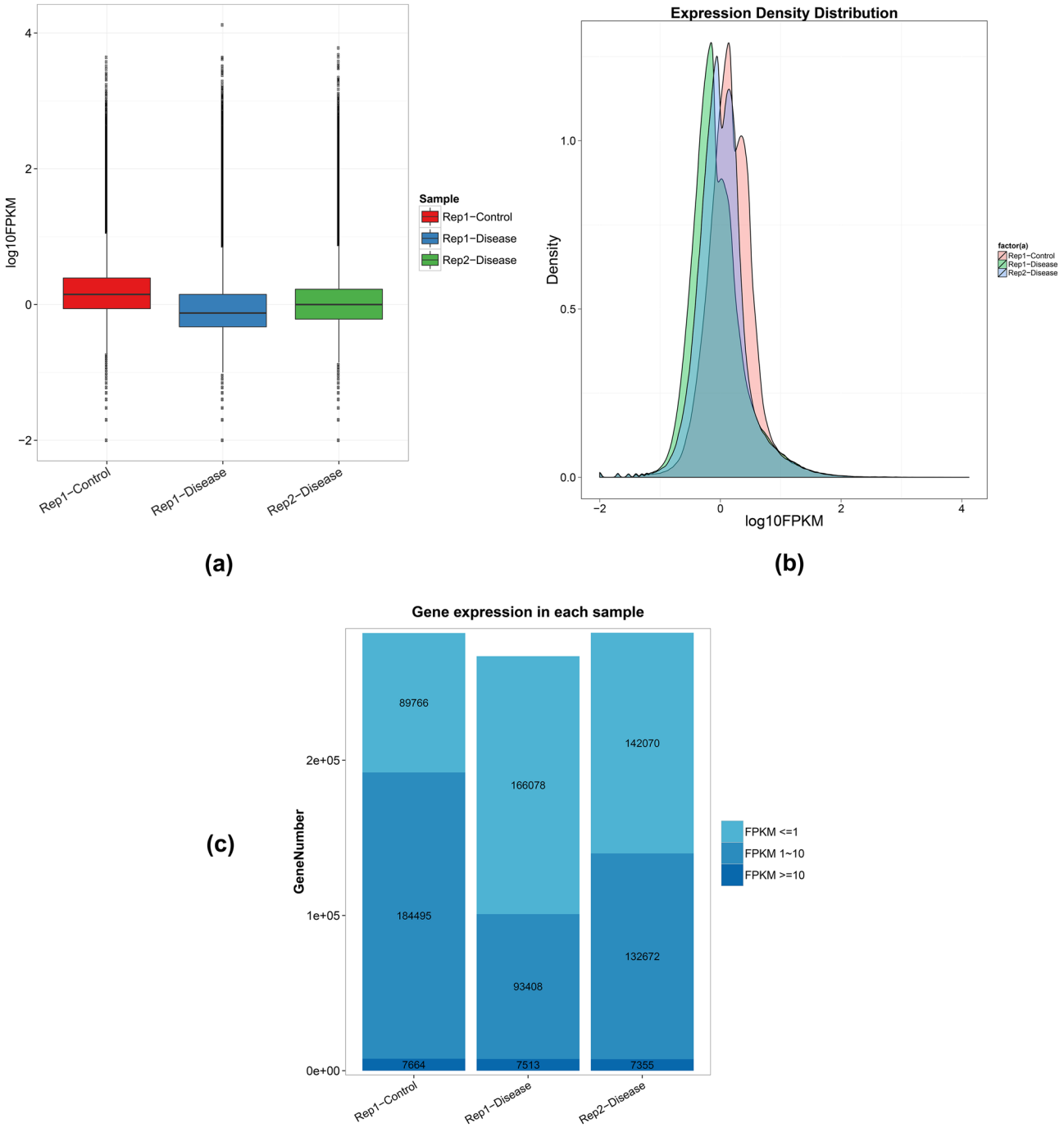

Figure 3. (**a**) Box plot of Gene expression. The X-axis represents the sample name, and the Y-axis is the $\log_{10}$FPKM value. (**b**) Expression Density Distribution. The X-axis represents the $\log_{10}$FPKM value, and the Y-axis represents the expression density, which means the ratio of gene amount under the specific expression level to the total number of expressed genes. (**c**) Gene expression distribution. The X-axis represents the sample name, and the Y-axis represents the gene amount. The depth of the color refers to different gene expression levels: FPKM < = 1 means extremely low expression level, and 110 means high expression level.

latter two had more up-regulated DEGs. The protein-interaction network for Rep1-Control vs. Rep1-Disease and Rep1-Control vs. Rep2-Disease group showed the up and downregulation of Unigenes along with the significance and intensity of the protein-interaction network on an individual gene level (see Supplementary File S5).

Discussion

Shisham (*Dalbergia sissoo*) is a timber-producing deciduous tree species of vast economic importance. Its timber has been used in building furniture and other wood crafts in the Indian sub-continent. It has remained one of the pillars of its economy since old age. The ecological impact of *D. sissoo* comes from its involvement in the

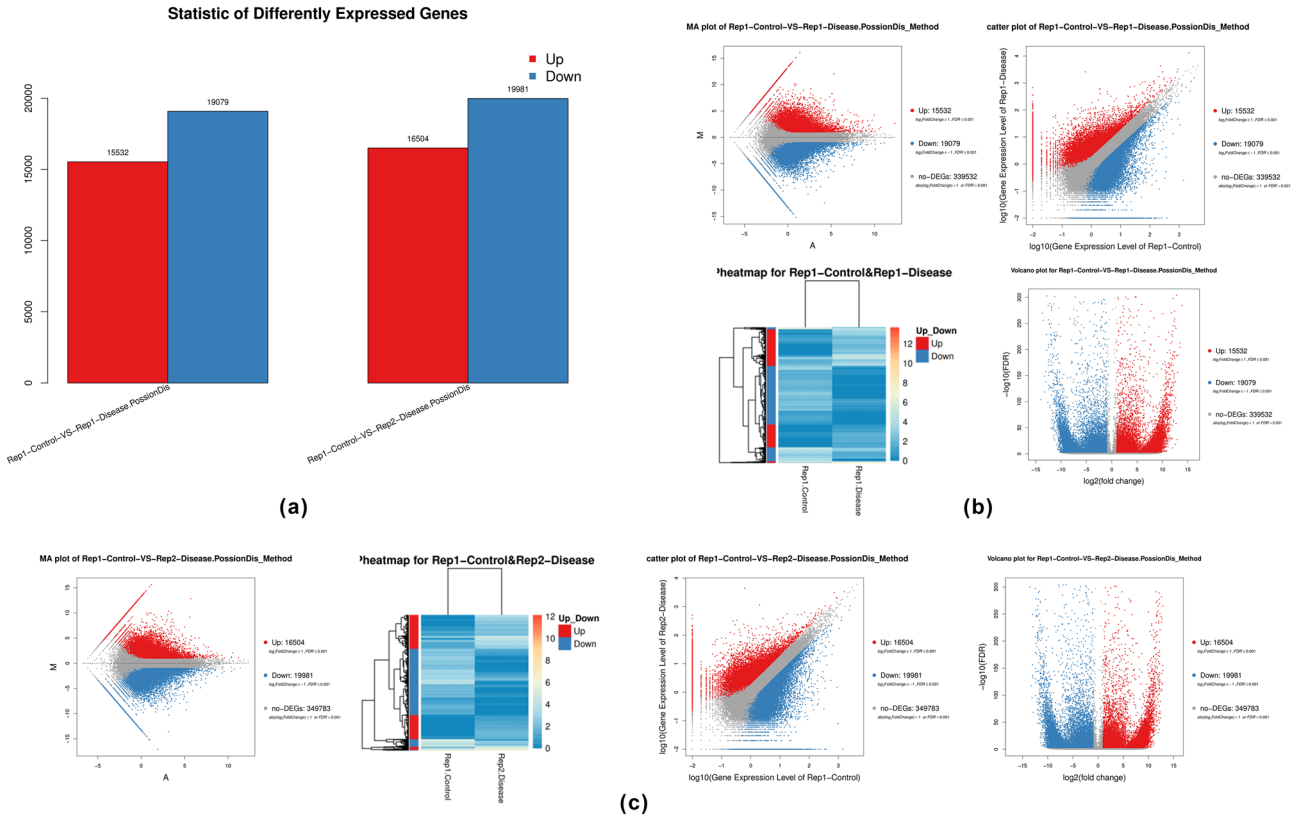

Figure 4. (**a**) Summary of DEGs. The X-axis represents the comparison functionality between each group and the Y-axis represents DEG numbers. The red color represents upregulated DEGs. The blue color represents downregulated DEGs. (**b, c**) MA plot. The X-axis represents value A (log_2 transformed mean expression level) and the Y-axis represents value M (log_2transformed fold change). Red dots represent upregulated DEG. Blue dots represent downregulated DEG. Black points represent non-DEGs. **Scatter plot.** XY-axis represents the log_{10} transformed gene expression level, the blue color represents the up-regulated genes, the red color represents the down-regulated genes, and the grey color represents the non-significant differential genes. Volcano plot. The X-axis represents -log_{10} transformed significance the Y-axis represents log_2 transformed fold change. Red points represent upregulated DEG. Blue points represent downregulated DEG. Black points represent non-DEGs. Heatmap plot. The X-axis represents the sample name and the Y-axis represents DEGs. The dark color means a high expression level while the light color means a low expression level.

associated ways such as soil erosion control, carbon sequestration, water quality maintenance, ecosystem services, and genetic diversity preservation. It also has great medicinal importance in this region and therefore is of great cultural importance as well[58,59]. Even though it is an important tree species with enormous germplasm resources, there is a lack of genomic data under the biotic stress of Shisham dieback. This data is an essential prerequisite for ensuring survival, future breeding, and genetic improvement programs. The diversity of *D. sissoo* is also another endangered factor due to habitat loss and fragmentation, and it requires more attention from the international scientific community for proper preservation and conservation of the germplasm resources.

Fortunately, modern Next Generation Sequencing (NGS) technologies have enabled us to develop large datasets of genomic data for breeding and crop improvement programs both cost-effectively and in an timely-efficient manner[60,61]. In our study, we used the BGISEQ-500 sequencing platform to perform the transcriptome sequencing of a total RNA equimolar mixture isolated from the young leaf tissue samples of the *D. sissoo* tree. The output cDNA library was normalized to maximize the chances of finding less abundant mRNA and robust sampling of complex transcripts[62]. The significant overlapping in paired-end cDNA sequencing for tree species with no reference genome was a critical part of the assembly. In total, 18.14 Gb data was generated. After sequence read filtration treatment, 310,523,693 (96.89%) clean reads were obtained. These clean reads were further assembled using Trinity into 513,821 Unigenes for the reconstruction of reference genomes lacking non-model tree species[63]. The high quality and precision of assembly were confirmed through the N50 value of 1101 bp with an average length of 604 bp. Compared with a recent transcriptome assembly that utilized the Illumina® HiSeq 4000 sequencing platform, our BGISEQ-500-generated paired-end filtered reads were 5.2% more in number (96.89% vs. 91.7%)[64]. Overall, a robust and precise analysis pipeline was implemented to easily assess and report the transcriptome datasets and decipher the underlying molecular mechanisms and disease pathways.

The TransDecoder CDS prediction revealed that most of the Unigenes code for the ORFs of medium to long lengths. The presence of a respective amount of hits (38.99%) in the NR database shows that a significant number of CDS codes for proteins. The CDS that did not exhibit a hit might be very short in length, lack functionally

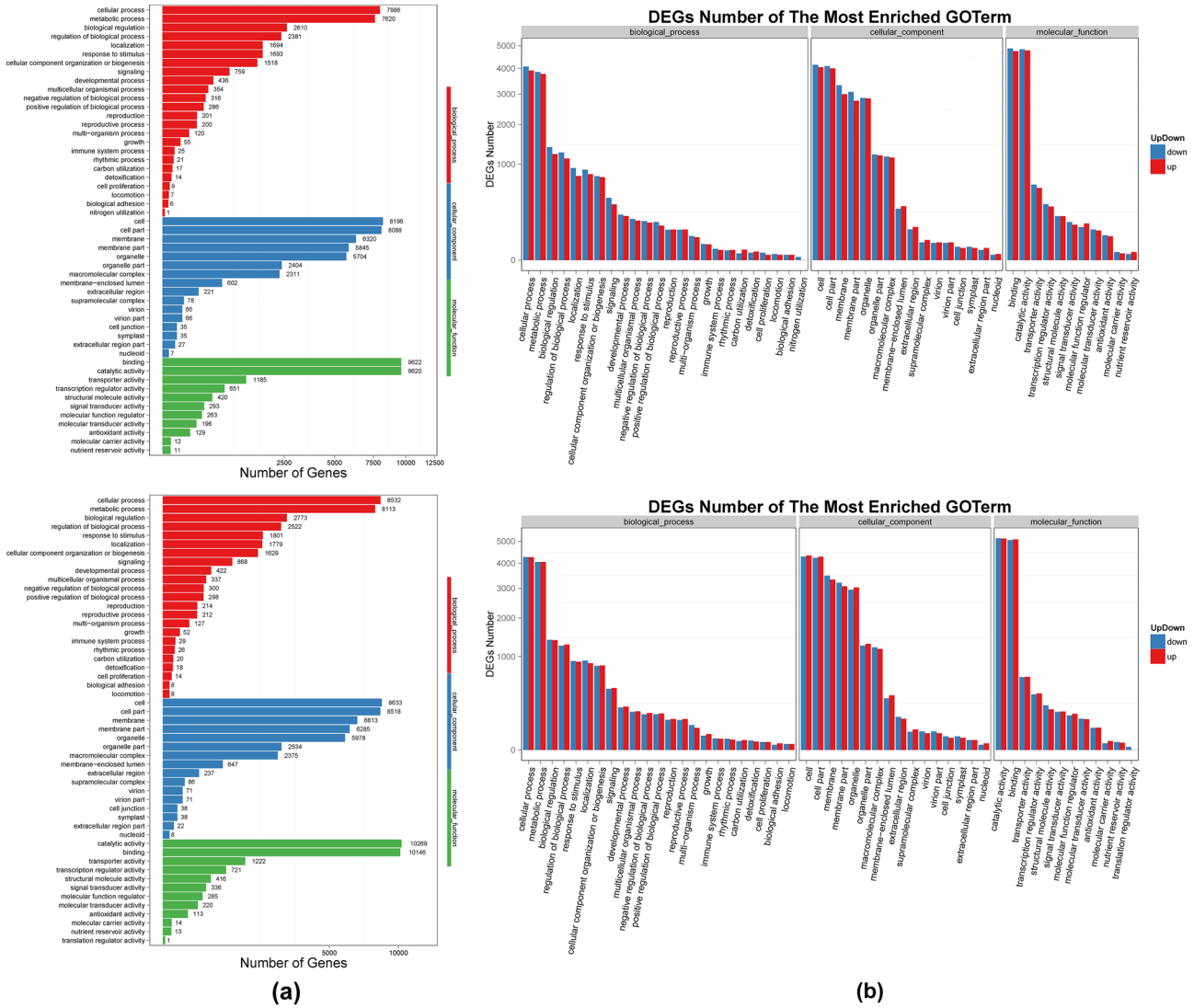

Figure 5. (**a**) GO classification of DEGs. The X-axis represents the number of DEG and the Y-axis represents the GO term. (**b**) GO classification of up-regulate and down-regulate **genes.** The X-axis represents the GO term and the Y-axis represents the amount of up/down-regulated genes.

conserved domain, or were part of non-coding RNAs[65]. In the functional annotation results from the NR database, it was revealed that *D. sissoo* shared maximum similarity with an herb named *Arachis ipaensis*. Both of them originate from the same subfamily of *Faboideae*, suggesting it might serve as a reference genome for *D. sissoo* and other closely related *Fabaceae* species in the future. The GO and KEGG functional term enrichment analysis placed the Unigenes in three categories *i.e.*, biological process, cellular component, and molecular function. The major Unigene-enriched pathways include carbohydrate metabolism, translation, transport and catabolism, signal transduction, and environmental adaptation. The studies on the transcriptomes of other tree species duly support the above-mentioned findings[66–72].

In terms of the generalized FPKM expression level of assembled Unigenes, the top three expressing Unigenes for Rep1-Control were CL9241.Contig10_All, CL20392.Contig2_All, and Unigene15134_All with functional pathway annotations in Ribulose bisphosphate carboxylase small chain 1, Chlorophyll a-b binding protein of LHCII type 1, and serine protease inhibitor-like precursor, respectively. These genes are involved in crucial photosynthetic functions and enzymatic processes[73–75]. There was a partially significant change in the expression of these genes in Rep1-Disease and Rep2-Disease suggesting their moderately associative nature with the biotic stress induced by *B. theobromae*. In Rep1-Disease, the top three expressing Unigenes were CL2816.Contig6_All, CL9241.Contig5_All, and CL33174.Contig2_All with functional pathway annotations in Trypsin inhibitor, Ribulose bisphosphate carboxylase small chain 1, and Photosystem I P700 chlorophyll an apoprotein A2, respectively. When compared to Rep1-Control, all three had a 6.22-, 18.74-, and 1.41-times higher expression level, respectively. These extreme changes in the expression level suggest that the aforementioned protein pathways might have undergone dynamic regulation changes in their functions in response to the biotic stress induced by *B. theobromae*. These functions include plant defense responses and regulation of photosynthetic processes[73,76–78].

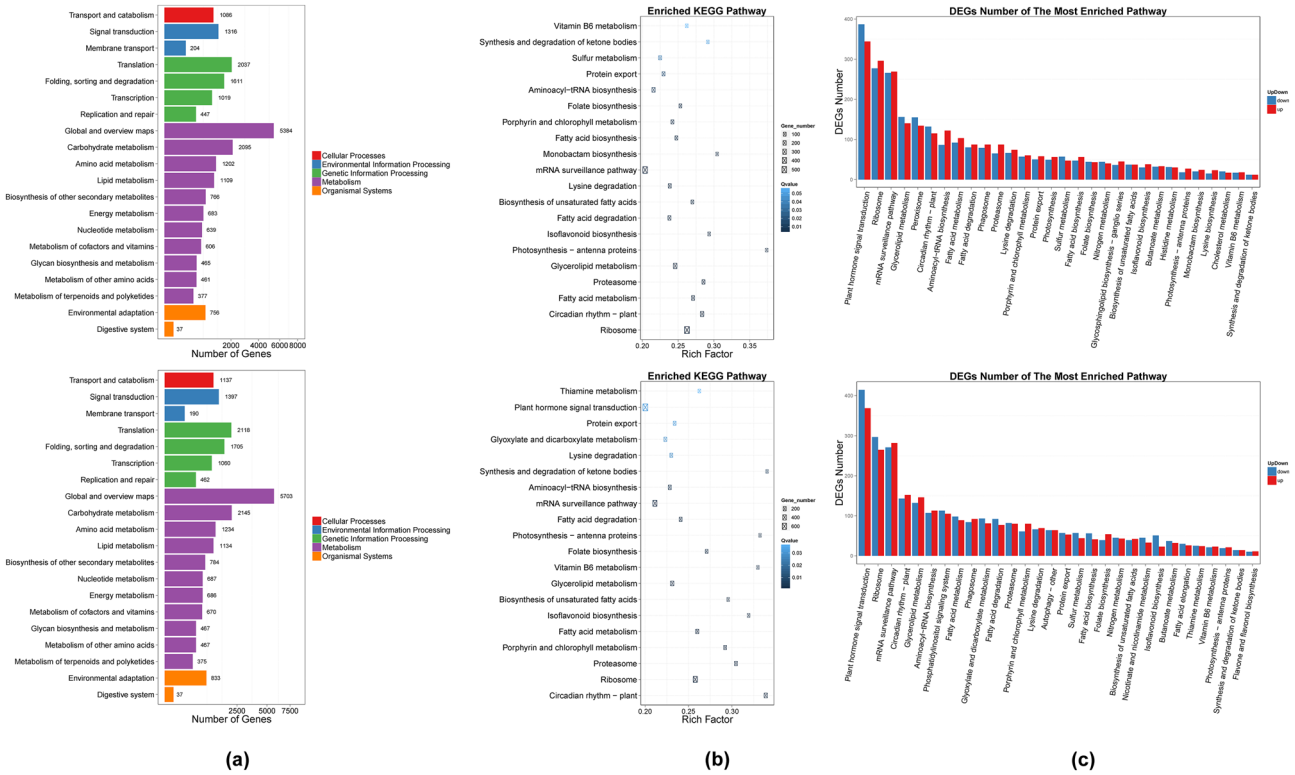

Figure 6. (**a**) Pathway classification of DEGs. The X-axis represents the number of DEG and the Y-axis represents the functional classification of KEGG. There are 7 branches for KEGG pathways: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Disease (For animals only), Metabolism, Organismal Systems, and Drug Development. (**b**) Pathway functional enrichment of DEGs. The X-axis represents the enrichment factor and the Y-axis represents the pathway name. The color indicates the q-value (high: white, low: blue), and the lower q value indicates the more significant enrichment. Point size indicates the DEG number (The bigger dots refer to a larger amount). Rich Factor refers to the value of enrichment factor, which is the quotient of foreground value (the number of DEGs) and background value (total Gene amount). The larger the value, the more significant the enrichment. (**c**) Pathway functional enrichment results for up/down-regulation genes. The X-axis represents the terms of Pathway and the Y-axis represents the number of up/down-regulation genes.

Similarly, in Rep2-Disease, the top three expressing Unigenes were CL9241.Contig10_All, CL33174.Contig2_All, and Unigene176088_All with functional pathway annotations in Ribulose bisphosphate carboxylase small chain 1, photosystem I P700 chlorophyll a apoprotein A2, and Serine protease inhibitor-like precursor. When compared to Rep1-Control, all three had a 1.38-, 1.63-, and 7165.29 times higher expression level, respectively. Aligning with the previous speculation, the first two pathway annotations have similar functions such as plant defense responses and regulation of photosynthetic processes[73,75–78]. However, the third extremely high expression (Unigene176088_All: 7165.29 times the expression of Rep1-Control) pathway annotation of serine protease inhibitor-like precursor has a diverse range of sub-classes and functions. These include extreme defense responses against pathogens by Bowman-Birk Inhibitors (BBI) and Kunitz-type inhibitors, and in abiotic stressors, plant growth, and development processes by Serine Protease Inhibitor-like Proteins (SPIs)[75,79–82]. All these findings suggest that all these genes dynamically change their functions through a significant increase in their expression level to cope with the biotic stress induced by *B. theobromae*.

In terms of the $\log_2$FoldChange expression level of identified DEGs, a significant up and downregulation of genes was observed when both the disease samples were grouped comparatively with the controlled sample. This speculates the dynamic change in the level of expression of different genes under the biotic stress induced by *B. theobromae*. For instance, in the Rep1-Control vs. Rep1-Disease group, the top three up-regulated DEGs were CL18436.Contig5_All, CL11636.Contig1_All, and CL34903.Contig2_All with functional pathway annotations in Aquaporin PIP-type 7a, Histone H3.3, and Kunitz trypsin inhibitor precursor, respectively. These protein pathways are involved in water transport across plasma membranes, defense responses against pathogens, and other abiotic stressors, respectively[79,80,83–86]. In the same group, the top 3 down-regulated DEGs were CL67374.Contig1_All, Unigene263_All, and CL24897.Contig1_All. All these have functional pathway annotations in tRNA-dihydrouridine (20) synthase [NAD(P)+]-like, Pentatricopeptide repeat-containing protein, and HIRA protein, respectively. These protein pathways are involved in response against abiotic stressors and regulation of gene expression[87–90]. Similarly, in the Rep1-Control vs. Rep2-Disease group, the top three up-regulated DEGs were CL8007.Contig2_All, CL30985.Contig9_All, and CL5191.Contig7_All. These have functional pathway annotations in SRC1-like protein, Isoflavone reductase homolog, and NADH dehydrogenase subunit 5. These protein

pathways are involved in response against environmental stressors, pathogen defense, and plant growth and development processes, respectively[91–95]. In the same group, the top three down-regulated DEGs were CL39346.Contig3_All, CL7867.Contig15_All, and CL7253.Contig6_All. These have functional pathway annotations in 29 kDa (kilo-Dalton) ribonucleoprotein A, β-galactosidase 17, and Phosphatase 2C 79 protein, respectively. These protein pathways are involved in stress response, and plant growth and development processes[96–103].

In summary, PIP-type 7a aquaporins have an active involvement in plant defense responses. They are mainly involved in the regulation of water transport across the plasma membrane and thus maintain the water balance of the cell. Due to this active involvement, PIP-type 7a aquaporins can be a potential improvement target to improve the resistance of *D. sissoo* against *B. theobromae* infection. Similarly, Histone H3.3, Kunitz trypsin inhibitor precursor, Pentatricopeptide repeat-containing protein, SRC1-like protein, Isoflavone reductase homolog, 29 kDa ribonucleoprotein A, NADH dehydrogenase subunit 5, β-galactosidase 17, and Phosphatase 2C 79 protein have shown strong association with plant–microbe interaction processes, plant-defense response, cellular expansion, elongation, division and proliferation control, enzyme, secondary metabolite, and hormonal regulations under disease stress, respectively.

DEGs were also identified using the Plant Resistance Gene (PRG) database[49,50] and the assembled Unigenes. In Rep1-Control, the top three genes with the highest FPKM expression values were Unigene176148_All (*Medicago truncatula*), and CL44494.Contig2_All (*Medicago truncatula*), and Unigene6841_All (*Oryza sativa*). These have pathway annotations in Heat shock cognate protein 70–1, Disease resistance protein (TIR-NBS-LRR class) family, and RNI-like superfamily protein, respectively. These pathways are involved in cellular functions like stress responses, pathogen defense, and plant growth and development processes, respectively[104–109]. In Rep1-Disease, the top three genes with the highest FPKM expression values were CL44494.Contig2_All (*Medicago truncatula*), Unigene176148_All (*Medicago truncatula*), and CL58125.Contig2_All (*Glycine max*). These have pathway annotations in the Disease resistance protein (TIR-NBS-LRR class) family, Heat shock cognate protein 70–1, and pleiotropic drug resistance 12, respectively. These pathways are involved in functions like pathogen defense and stress responses[105–111]. Similarly, in Rep2-Disease, the top three genes with the highest FPKM expression values were Unigene123966_All (*Medicago truncatula*), Unigene177302_All (*Oryza sativa*), and CL44494.Contig2_All (*Medicago truncatula*). These have pathway annotations in Heat shock cognate protein 70–1, RNI-like superfamily protein, and Disease resistance protein (TIR-NBS-LRR class) family, respectively. Similar to Rep1-Disease, these protein pathways have functions like stress responses, pathogen defense, and plant growth and development, respectively[104–109]. Compared to Rep1-Control, there was a significant change in the expression level of the aforementioned genes which indicates the dynamic expression adjustments resulting due to the stress induced by *B. theobromae*.

In this study, we have identified Unigenes in the transcriptome of *D. sissoo* under normal and *B. theobromae*-induced stress. We analyzed their GO enrichment terms and KEGG functional pathway annotations, identified their SSR markers, and cross-matched all the Unigenes in 7 functional databases. The CDS and transcription factors were predicted along with expression density. Furthermore, we detected the DEGs, analyzed their GO terms, and KEGG pathways, mapped the protein–protein interaction networks, and curated the Plant Resistance Genes (PRGs) from the assembled Unigenes. Additionally, the FPKM and Fold Change gene expression of Unigenes, DEGs, and PRGs were also comparatively evaluated. On the whole, all this robust, precise, and valuable transcriptomic information will lay the foundations needed to ensure the survival of *D. sissoo* through crop improvement programs, genetic breeding, increased biodiversity, and preserved germplasm resources.

Conclusions

Our study on the transcriptome of *D. sissoo* demonstrates how the stress from dieback disease induced by *B. theobromae* affects the level of gene expression. In terms of de novo transcriptome sequencing, we generated about 18.14 Gb bases in total and 310,523,693 bp clean reads. In total, 513,821 Unigenes were curated with an average length of 604 bp, an N50 of 1,101 bp, and a GC content of 39.95%. All the Unigenes were aligned with 7 functional databases and annotations were given as follows: 200,355 (NR: 38.99%), 164,973 (NT: 32.11%), 123,733 (Swissprot: 24.08%), 142,580 (KOG: 27.75%), 139,588 (KEGG: 27.17%), 99,752 (GO: 19.41%), and 137,281 (InterPro: 26.72%). The GO term enrichment and KEGG functional pathway analysis shed light on the important underlying stress response mechanisms. Moreover, 115,762 CDS and 4,673 TF coding genes were also identified along with 62,863 SSR markers distributed on 51,508 Unigenes. The significant up and down-regulation of 16,018 up- and 19,530 down-regulated DEGs showed the gene expression shifts under the biotic stress conditions. The evaluation of change in the relative expression levels in the case of both the DEGs and 9,230 identified PRGs hints at the dynamic shift in the underlying gene expression levels in response to disease stress. Our primary objective has been the expansion of the transcriptomic resource library accessible for *D. sissoo*. We have achieved this goal by incorporating high-quality transcriptome sequencing data, specifically gathered under the stress of *B. theobromae* infection. The identification of potential DEGs carries significant implications for various disease-related processes, such as biomarker discovery, comprehension of underlying pathway-mediated biological phenomena, disease categorization, and subtyping. All this will be made possible by the generated gene expression datasets highlighting up- and down-regulated genes. The multitude of PRGs, SSR markers, and Unigenes, when combined with genomic data, collectively contributes to the survival of *D. sissoo* under environmental stressors, preservation of germplasms, and the facilitation of future hybridization and breeding initiatives.

Material and methods

Plant materials preparation and RNA extraction

The fungus *B. theobromae* was isolated from the confirmed sources and cultured on BAM Media M127: Potato Dextrose Agar (PDA) according to established protocols[112]. The culture was examined microscopically to assess the viability of the fungus through septation incidence and the shape of spores. After proper assessment, the spore solution was prepared in sterile water and the inoculation in the stem-base portion of greenhouse-grown one-year-old *D. sissoo* vegetative plantations was performed according to standard protocols[113]. The plantations were grown at a temperature of 25 °C, 70% humidity, a light intensity of 350 µmol m^{-2} s^{-1} Photosynthetically Active Radiation (PAR), and with a 12-h light and 12-h dark photoperiod. After successful inoculation, three leaf samples were isolated after 20 days[114,115], one from the control and two from the diseased one[116–118]. The leaf samples were enclosed in desiccated plastic bags and stored at − 30 °C for subsequent extraction. Vegetative plantations of *D. sissoo* were provided by the Punjab Forestry Research Institute (PFRI), Gatwala, Faisalabad. RNA (Ribonucleic acid) was isolated from the three young leaf samples using an RNA Purification Kit according to the manufacturer's instructions (Thermo Fisher Scientific – Waltham, Massachusetts, United States). After extraction, the concentration and quality of RNA for cDNA synthesis were accurately determined using NanoDrop™ 8000 Spectrophotometer (Thermo Fisher Scientific – Waltham, Massachusetts, United States), Agilent 2100 Bioanalyzer system, and Agilent RNA 6000 Nano Kit (Agilent Technologies – Santa Clara, California, United States) and finally, it was stored at − 80 °C.

Sequence read filtering, BGISEQ-500 transcriptome sequencing, and de novo assembly

Extracted RNA from three young leaf samples was used for cDNA synthesis and transcriptome sequencing through the BGISEQ-500 sequencing platform at BGI Genomics Co., Ltd., Shenzhen, Guangdong, China. This sequencing platform has a Paired-End (PE) read length capacity of 100 bp. The raw generated reads were subjected to a filter treatment by implementing strict principles and protocols[119,120]. After filtration, the 'clean reads' are stored in FASTQ format for further analysis[121]. Trinity v2.0.6 was utilized (Parameters: –min_contig_length 150 –CPU 8 –min_kmer_cov 3 –min_glue 3 –bfly_opts '-V 5 –edge-thr = 0.1 –stderr ') to carry out the assembly of clean reads with removed PCR duplications to improve efficiency and Tgicl v2.0.6 was used (Parameters: -l 40 -c 10 -v 25 -O '-repeat_stringency 0.95 -minmatch 35 -minscore 35') to cluster the large transcripts to Unigenes[42,122–124]. Trinity partitions the sequence data using its independent three modules *i.e*., Inchworm, Chrysalis, and Butterfly into many individual de Bruijn graphs. The de Bruijn graphs were used for the representation of the transcriptional complexity of a gene or locus and then Trinity processes each graph simultaneously and independently to extract splicing isoforms of full length[125]. These splicing isoforms are then used to sort out transcripts derived from paralogous genes. The final Unigenes were assembled using Tgicl gene family clustering for each sample for further downstream analysis[126,127]. When several Unigenes have a similarity of more than 70%, they are placed in one unified cluster while the remaining clustered Unigenes are placed as singletons. Furthermore, the RNA-seq transcripts used here have already been submitted to Sequence Read Archive or SRA (https://www.ncbi.nlm.nih.gov/sra) and Gene Expression Omnibus or GEO datasets (https://www.ncbi.nlm.nih.gov/geo/) of the National Center for Biotechnology Information or NCBI (https://www.ncbi.nlm.nih.gov/) under GEO accession ID: GSE220535 and BioProject ID: PRJNA910142. Their sample-wise accession number is GSM6806574 (Leaf Control R1), GSM6806575 (Leaf Treatment R1), and GSM6806576 (Leaf Treatment R2) with SRA IDs SRX18528739 (Rep1—Control), SRX18538740 (Rep1—Treatment), and SRX18538741 (Rep2—Treatment)[128,129].

Functional annotation of Unigenes

The curated Unigenes were analyzed by comparing and searching against public functional databases such as NT (The Nucleotide database) (ftp://ftp.ncbi.nlm.nih.gov/blast/db), NR (The Protein database) (ftp://ftp.ncbi.nlm.nih.gov/blast/db), GO (Gene Ontology) (http://geneontology.org), KOG (EuKaryotic Orthologous Groups) (ftp://ftp.ncbi.nih.gov/pub/COG/KOG), KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg), SwissProt (http://ftp.ebi.ac.uk/pub/databases/swissprot) and InterPro (http://www.ebi.ac.uk/interpro)[43,45,130,131]. Unigenes were aligned using BLAST-N v2.2.23 (http://blast.ncbi.nlm.nih.gov/Blast.cgi), BLAST-X v2.2.23 (http://blast.ncbi.nlm.nih.gov/Blast.cgi), Diamond v0.8.31 (https://github.com/bbuchfink/diamond), Blast2GO v2.5.0 (https://www.blast2go.com), InterProScan5 v5.11–51.0 (https://code.google.com/p/interproscan/wiki/Introduction)[44,47,132,133]. All the software tools were used at their default set perameters. Candidate coding areas were curated for Coding sequences or CDs using TransDecoder v3.0.1 (https://transdecoder.github.io). The longest Open Reading Frames or ORFs were extracted and then the coding regions were predicted using HMMscan and Swiss-Prot using BLAST search for protein homologous sequences in Pfam[48,134]. For the prediction of Unigene Transcription Factors (TFs), each Unigene was exploited using getorf version: EMBOSS: 6.5.7.0 (Parameters: -minsize 150) (http://genome.csdb.cn/cgi-bin/emboss/help/getorf)[135]. TFs were aligned to ORFs using hmmsearch v3.0 (http://hmmer.org) at default parameters and the guidelines provided by PlmTFDB (http://plntfdb.bio.uni-potsdam.de/v3.0/)[136].

Unigene SSR identification, expression, and plant disease resistance gene identification

MISA v1.0 (Parameters: 1–12,2–6,3–5,4–5,5–4,6–4 100 150) (http://pgrc.ipk-gatersleben.de/misa) was used to identify SSRs among Unigenes and then Primer3 v2.2.2 (http://bioinfo.ut.ee/primer3) was used to design primers for each identified SSR at default parameters[137,138]. The clean reads were mapped and aligned with the identified Unigenes using Bowtie2 v2.2.5 (Parameters: -q –phred64 –sensitive –dpad 0 –gbar 99,999,999 –mp 1,1 –np 1 –score-min L,0,-0.1 -I 1 -X 1000 –nomixed –no-discordant -p 1 -k 200) (http://bowtie-bio.sourceforge.net/Bowtie2/index.shtml) and then the level of gene expression was calculated with RSEM v1.2.12 (http://deweylab.biostat.

wisc.edu/RSEM) in FPKM (Fragments Per Kilobase of transcript per Million mapped reads)[51,52,139]. The hierarchal clustering was performed using its hculust function at default parameters[140]. Finally, plant disease resistance genes were detected using Diamond (BLAST) on the identified Unigenes in the PRG database v2.0 (http://prgdb.crg.eu/)[49,50]. Potential disease resistance genes were finalized by evaluating the query coverage and identity values.

Detection, GO, KEGG, and protein-interaction network analysis of differentially expressed genes (DEGs)

PoissonDis (Parameters: Fold Change > = 2.00 and FDR < = 0.001) was used to detect DEGs based on the Poisson distribution[54,55,141,142]. The results of PoissonDis were visualized using a comparative histogram, MA[143–145], scatter[146], volcano[147,148], and heatmap plots[149]. GO (Gene Ontology) annotations and the official classification were used to classify the DEGs and functional enrichment was checked using phyper, a function of R. The p-value was calculated using the hypergeometric test and the False Discovery Rate (FDR) was then calculated for each *p* value[56,141,150]. The GO terms that have an FDR value less than 0.01 were defined as significantly enriched[43,151]. The KEGG annotation results were also used to classify DEGs according to official classification. phyper was used again to check the significant functional enrichment with the abovementioned hypergeometric p-value and FDR calculations[45]. To analyze the protein–protein interaction network, the DEGs were mapped to the STRING database v10.0 (http://string-db.org/)[152].

Data availability

The raw sequencing data and assembled transcriptome generated in this study have been deposited in the Sequence Read Archive or SRA (https://www.ncbi.nlm.nih.gov/sra) and Gene Expression Omnibus or GEO datasets (https://www.ncbi.nlm.nih.gov/geo/) of the National Center for Biotechnology Information or NCBI (https://www.ncbi.nlm.nih.gov/) under GEO accession ID: GSE220535 and BioProject ID: PRJNA910142. Their accession number is GSM6806574 (Leaf Control R1), GSM6806575 (Leaf Treatment R1), and GSM6806576 (Leaf Treatment R2) with SRA IDs SRX18528739 (Rep1—Control), SRX18538740 (Rep1—Treatment), and SRX18538741 (Rep2—Treatment). Researchers can access the data and associated metadata for further analysis and validation. Detailed information on data retrieval and processing can be found in the Materials and Methods section of this paper.

Received: 27 May 2023; Accepted: 26 October 2023
Published online: 22 November 2023

References

1. Mukerjee, S. K., Saroja, T. & Seshadri, T. R. Dalbergichromene: A new neoflavonoid from stem-bark and heartwood of Dalbergia sissoo. *Tetrahedron* **27**, 799–803. https://doi.org/10.1016/S0040-4020(01)92474-3 (1971).
2. So, T., Theilade, I. & Dell, B. Conservation and utilization of threatened hardwood species through reforestation—An example of *Afzelia xylocarpa* (Kruz.) Craib and Dalbergia cochinchinensis Pierre in Cambodia. *Pac. Conserv. Biol.* **16**, 101–116. https://doi.org/10.1071/PC100101 (2010).
3. Mishra, N. N. & Mehera, B. Assessment of biomass and carbon stocks in selected tree species in vindhyas series. *J. Pharm. Phytochem.* **9**, 1010–1013 (2020).
4. Ghazali, H. M. Z. U. *et al.* Fungi species causing dieback and wilt diseases in shisham [*Dalbergia sissoo* (Roxb)] and impact of various fungicides on their management. *J. King Saud. Univ. Sci.* **34**, 101970. https://doi.org/10.1016/j.jksus.2022.101970 (2022).
5. Arif, M., Zaidi, N. W., Singh, Y. P., Rizwanul Haq, Q. M. & Singh, U. S. A comparative analysis of ISSR and RAPD markers for study of genetic diversity in Shisham (*Dalbergia sissoo*). *Plant Mol.r Biol. Rep.* **27**, 488–495 (2009).
6. Sharma, M. K., Singal, R. M. & Pokhriyal, T. C. Dalbergia sissoo in India. 5–16 (2000).
7. Khan, S. H., Idrees, M., Muhammad, F., Mahmood, A. & Zaidi, S. H. Incidence of shisham (*Dalbergia sissoo* Roxb.) decline and in vitro response of isolated fungus spp. to various fungicides. *Int. J. Agric. Biol* **6**, 611–614 (2004).
8. Javaid, A. Research on shisham (*Dalbergia sissoo* Roxb.) decline in Pakistan—A review. *Pak. J. Phytopathol* **20**, 134–142 (2008).
9. Shamsi, S., Sultana, R. & Azad, R. Occurrence of leaf and POD diseases of Dalbergia SISSOO In BANGLADESH. *Banglad. J. Plant Pathol.* **28**, 33 (2012).
10. Ahmad, I., Hanan, A. & Gul, S. Frequency of mycoflora associated with Shisham (*Dalbergia sissoo*) decline in district Faisalabad, Pakistan. *FUUAST J. Biol.* **5**, 225–229 (2015).
11. Schulman, A. H. Molecular markers to assess genetic diversity. *Euphytica* **158**, 313–321 (2007).
12. Mondini, L., Noorani, A. & Pagnotta, M. A. Assessing plant genetic diversity by molecular tools. *Diversity* **1**, 19–35 (2009).
13. Jiang, G.-L. Molecular markers and marker-assisted breeding in plants. *Plant Breed. Lab. Fields* **3**, 45–83 (2013).
14. He, Q. *et al.* Transcriptome profiles of leaves and roots of Goldenrain tree (*Koelreuteria paniculata* Laxm.) in response to cadmium stress. *Int. J. Environ. Res. Public Health* **18**, 12046 (2021).
15. Li, X. *et al.* Transcriptome analysis provides insights into the stress response crosstalk in apple (Malus× domestica) subjected to drought, cold and high salinity. *Sci. Rep.* **9**, 1–10 (2019).
16. Roy, C. B., Liu, H., Rajamani, A. & Saha, T. Transcriptome profiling reveals genetic basis of disease resistance against *Corynespora cassiicola* in rubber tree (*Hevea brasiliensis*). *Curr. Plant Biol.* **17**, 2–16 (2019).
17. Mohit, G., Neelu, G. & Gupta, B. N. Preliminary observations on genetic variability and character association in *Dalbergia sissoo* Roxb. *Indian Forester* **126**, 608–615 (2000).
18. Javaid, A., Akram, W., Shoaib, A., Haider, M. S. & Ahmad, A. ISSR analysis of genetic diversity in *Dalbergia sissoo* in Punjab, Pakistan. *Pak. J. Bot* **46**, 1573–1576 (2014).
19. Tewari, S. K. *et al.* Use of the RAPD marker to determine the genetic diversity of various *Dalbergia sissoo* Roxb. (Shisham) genotypes and their evolutionary relationship in Indian subcontinents. *Vegetos* **35**, 850–857 (2022).
20. Ashraf, M., Mumtaz, A. S., Riasat, R. & Tabassum, S. A molecular study of genetic diversity in Shisham (*Dalbergia sissoo*) plantation of NWFP, Pakistan. *Pak. J. Bot* **42**, 79–88 (2010).
21. Bakshi, M. & Sharma, A. Assessment of genetic diversity in *Dalbergia sissoo* clones through RAPD profiling. *J. For. Res.ch* **22**, 393–397 (2011).
22. Wang, B., Shi, L., Ruan, Z. & Deng, J. Genetic diversity and differentiation in *Dalbergia sissoo* (Fabaceae) as revealed by RAPD. *Genet. Mol. Res.* **10**, 114–120 (2011).

23. Arif, M., Zaidi, N., Singh, Y., Rizwanul Haq, Q. M. & Singh, U. A comparative analysis of ISSR and RAPD markers for study of genetic diversity in Shisham (*Dalbergia sissoo*). *Plant Mol. Biol. Rep.* **27**, 488–495 (2009).
24. Kaur, S. *et al.* Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genom.* **12**, 265. https://doi.org/10.1186/1471-2164-12-265 (2011).
25. Wang, S. *et al.* Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. *Plant Cell Rep.* **31**, 1437–1447. https://doi.org/10.1007/s00299-012-1259-3 (2012).
26. Wang, H. *et al.* Next-Generation Sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome permits large-scale Unigene assembly and SSR marker discovery. *PLOS ONE* **8**, e62293. https://doi.org/10.1371/journal.pone.0062293 (2013).
27. Wu, J., Cai, C., Cheng, F., Cui, H. & Zhou, H. Characterisation and development of EST-SSR markers in tree peony using transcriptome sequences. *Mol. Breed.* **34**, 1853–1866. https://doi.org/10.1007/s11032-014-0144-x (2014).
28. Wei, W. *et al.* Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genom.* **12**, 451. https://doi.org/10.1186/1471-2164-12-451 (2011).
29. Dutta, S. *et al.* Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *BMC Plant Biol.* **11**, 17. https://doi.org/10.1186/1471-2229-11-17 (2011).
30. Zhang, J. *et al.* De novo assembly and characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). *BMC Genom.* **13**, 90. https://doi.org/10.1186/1471-2164-13-90 (2012).
31. Taheri, S. *et al.* Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. *Molecules* **23**, 399 (2018).
32. Dervishi, A., Jakše, J., Ismaili, H., Javornik, B. & Štajner, N. Comparative assessment of genetic diversity in Albanian olive (*Olea europaea* L.) using SSRs from anonymous and transcribed genomic regions. *Tree Genet. Genomes* **14**, 53. https://doi.org/10.1007/s11295-018-1269-6 (2018).
33. Li, N. *et al.* Development and validation of SSR markers based on transcriptome sequencing of *Casuarina equisetifolia*. *Trees* **32**, 41–49. https://doi.org/10.1007/s00468-017-1607-6 (2018).
34. Dong, M. *et al.* Development of EST-SSR markers in Larix principis-rupprechtii Mayr and evaluation of their polymorphism and cross-species amplification. *Trees* **32**, 1559–1571. https://doi.org/10.1007/s00468-018-1733-9 (2018).
35. Zhai, S. H., Yin, G. S. & Yang, X. H. Population genetics of the endangered and wild edible plant *Ottelia acuminata* in Southwestern China using novel SSR markers. *Biochem. Genet.* **56**, 235–254. https://doi.org/10.1007/s10528-018-9840-2 (2018).
36. Vu, D.-D. *et al.* Genetic diversity and conservation of two threatened dipterocarps (Dipterocarpaceae) in southeast Vietnam. *J. For. Res.* **30**, 1823–1831. https://doi.org/10.1007/s11676-018-0735-1 (2019).
37. Jiang, L., Zhang, M. & Ma, K. Whole-genome DNA methylation associated with differentially expressed genes regulated anthocyanin biosynthesis within flower color chimera of ornamental tree Prunus mume. *Forests* **11**, 90 (2020).
38. Liu, Q. *et al.* Transcriptomic profiling reveals differentially expressed genes associated with pine wood nematode resistance in masson pine (*Pinus massoniana* Lamb). *Sci. Rep.* **7**, 1–14 (2017).
39. Wang, X.-Y., Wu, X.-Q., Wen, T.-Y., Feng, Y.-Q. & Zhang, Y. Transcriptomic analysis reveals differentially expressed genes associated with pine wood nematode resistance in resistant Pinus thunbergii. *Tree Physiol.*, tpad018 (2023).
40. Feng, Y. *et al.* Differential expression profiles and pathways of genes in drought resistant tree species *Prunus mahaleb* roots and leaves in response to drought stress. *Scientia Horticulturae* **226**, 75–84 (2017).
41. Arce-Leal, Á. P. *et al.* Gene expression profile of Mexican lime (*Citrus aurantifolia*) trees in response to huanglongbing disease caused by *Candidatus Liberibacter asiaticus*. *Microorganisms,* **8** (2020).
42. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. https://doi.org/10.1038/nbt.1883 (2011).
43. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
44. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676. https://doi.org/10.1093/bioinformatics/bti610 (2005).
45. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
46. Venn, J. On the diagrammatic and mechanical representation of propositions and reasonings. *Philos. Mag.* **9**, 1–18 (1880).
47. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
48. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
49. Calle García, J. *et al.* PRGdb 4.0: An updated database dedicated to genes involved in plant disease resistance process. *Nucleic Acids Res.* **50**, D483-1490. https://doi.org/10.1093/nar/gkab1087 (2021).
50. Sanseverino, W. *et al.* PRGdb 2.0: Towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res.* **41**, D1167–D1171. https://doi.org/10.1093/nar/gks1183 (2013).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. https://doi.org/10.1038/nmeth.1923 (2012).
52. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323. https://doi.org/10.1186/1471-2105-12-323 (2011).
53. Tukey, J. W. *Exploratory data analysis*. Vol. 2 (Reading, 1977).
54. Haight, F. A. *Handbook of the Poisson Distribution*. (Wiley, 1967).
55. Akinkunmi, M. in *Introduction to Statistics Using R* (ed Mustapha Akinkunmi) 175–187 (Springer International Publishing, 2019).
56. Feise, R. J. Do multiple outcome measures require p-value adjustment?. *BMC Med. Res. Methodol.* **2**, 8. https://doi.org/10.1186/1471-2288-2-8 (2002).
57. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445. https://doi.org/10.1073/pnas.1530509100 (2003).
58. Vanshita, A. & Bajpai, M. Phytochemistry and pharmacology of *Dalbergia sissoo* Roxb ex DC: A review. *J. Pharm. Pharmacol.* **75**, 482–501. https://doi.org/10.1093/jpp/rgac106 (2023).
59. Li, M., Liu, M., Wang, B. & Shi, L. Metabonomics analysis of stem extracts from *Dalbergia sissoo*. *Molecules* https://doi.org/10.3390/molecules27061982 (2022).
60. Yadav, P. *et al.* Recent perspective of next generation sequencing: Applications in molecular plant biology and crop improvement. *Proc. Natl. Acad. Sci. India Sect. B: Biol. Sci.* **88**, 435–449 (2018).
61. Ballard, D., Winkler-Galicki, J. & Wesoły, J. Massive parallel sequencing in forensics: Advantages, issues, technicalities, and prospects. *Int. J. Legal Med.* **134**, 1291–1303. https://doi.org/10.1007/s00414-020-02294-0 (2020).
62. Hoang, N. V., Furtado, A., Perlo, V., Botha, F. C. & Henry, R. J. The impact of cDNA normalization on long-read sequencing of a complex transcriptome. *Front. Genet.* **10**, 654 (2019).
63. Amil-Ruiz, F. *et al.* Constructing a de novo transcriptome and a reference proteome for the bivalve *Scrobicularia plana*: Comparative analysis of different assembly strategies and proteomic analysis. *Genomics* **113**, 1543–1553 (2021).
64. Hung, T. H. *et al.* Reference transcriptomes and comparative analyses of six species in the threatened rosewood genus Dalbergia. *Sci. Rep.* **10**, 17749. https://doi.org/10.1038/s41598-020-74814-2 (2020).

65. Wu, G. *et al.* Sequencing, de novo assembly and comparative analysis of *Raphanus sativus* transcriptome. *Front. Plant Sci.* **6**, 198 (2015).
66. Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W. & Buerkle, C. A. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genom.* **11**, 1–16 (2010).
67. Li, Z. *et al.* Transcriptome analysis of *Botrytis cinerea* in response to tea tree oil and its two characteristic components. *Appl. Microbiol. Biotechnol.* **104**, 2163–2178 (2020).
68. Chen, J. *et al.* Deep-sequencing transcriptome analysis of low temperature perception in a desert tree, Populus euphratica. *BMC Genom.* **15**, 1–15 (2014).
69. Pinosio, S. *et al.* First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. *Mol. Ecol. Resour.* **14**, 846–856 (2014).
70. Mantello, C. C. *et al.* De novo assembly and transcriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. *PloS one* **9**, e102665 (2014).
71. Pasha, S. N. *et al.* The transcriptome enables the identification of candidate genes behind medicinal value of Drumstick tree (*Moringa oleifera*). *Genomics* **112**, 621–628 (2020).
72. Liu, F.-M. *et al.* De Novo transcriptome analysis of *Dalbergia odorifera* T. Chen (Fabaceae) and transferability of SSR markers developed from the transcriptome. *Forests* **10**, 98 (2019).
73. Dobberstein, B., Blobel, G. & Chua, N.-H. In vitro synthesis and processing of a putative precursor for the small subunit of ribulose-1, 5-bisphosphate carboxylase of *Chlamydomonas reinhardtii*. *Proc. Natl. Acad. Sci.* **74**, 1082–1085 (1977).
74. Jansson, S. The light-harvesting chlorophyll ab-binding proteins. *Biochimica et Biophysica Acta (BBA)-Bioenergetics* **1184**, 1–19 (1994).
75. Clemente, M. *et al.* Plant serine protease inhibitors: Biotechnology application in agriculture and molecular farming. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms20061345 (2019).
76. Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* **21**, 809–818 (2004).
77. De Leo, F., Bonadé-Bottino, M., Ceci, L. R., Gallerani, R. & Jouanin, L. Effects of a mustard trypsin inhibitor expressed in different plants on three lepidopteran pests. *Insect Biochem. Mol. Biol.* **31**, 593–602 (2001).
78. Jofuku, K. D. & Goldberg, R. B. Kunitz trypsin inhibitor genes are differentially expressed during the soybean life cycle and in transformed tobacco plants. *Plant Cell* **1**, 1079–1093 (1989).
79. Rustgi, S., Boex-Fontvieille, E., Reinbothe, C., von Wettstein, D. & Reinbothe, S. The complex world of plant protease inhibitors: Insights into a Kunitz-type cysteine protease inhibitor of *Arabidopsis thaliana*. *Commun. Integrat. Biol.* **11**, e1368599 (2018).
80. Oliveira, A. S. *et al.* Activity toward bruchid pest of a Kunitz-type inhibitor from seeds of the algaroba tree (*Prosopis juliflora* DC). *Pesticide Biochemi. Physiol.* **72**, 122–132 (2002).
81. Birk, Y. The Bowman-Birk inhibitor. Trypsin-and chymotrypsin-inhibitor from soybeans. *Int. J. Peptide Protein Res.* **25**, 113–131 (1985).
82. Kennedy, A. R. The Bowman-Birk inhibitor from soybeans as an anticarcinogenic agent. *Am. J. Clin. Nutr.* **68**, 1406S-1412S (1998).
83. Stroud, H. *et al.* Genome-wide analysis of histone H3. 1 and H3. 3 variants in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **109**, 5370–5375 (2012).
84. Otero, S., Desvoyes, B. & Gutierrez, C. Histone H3 dynamics in plant cell cycle and development. *Cytogenet. Genome Res.* **143**, 114–124 (2014).
85. Jang, J. Y. *et al.* Ectopic expression of a foreign aquaporin disrupts the natural expression patterns of endogenous aquaporin genes and alters plant responses to different stress conditions. *Plant Cell Physiol.* **48**, 1331–1339 (2007).
86. Mahdieh, M., Mostajeran, A., Horie, T. & Katsuhara, M. Drought stress alters water relations and expression of PIP-type aquaporin genes in *Nicotiana tabacum* plants. *Plant Cell Physiol.* **49**, 801–813 (2008).
87. Ingouff, M. & Berger, F. Histone3 variants in plants. *Chromosoma* **119**, 27–33 (2010).
88. Nie, X., Wang, H., Li, J., Holec, S. & Berger, F. The HIRA complex that deposits the histone H3. 3 is conserved in Arabidopsis and facilitates transcriptional dynamics. *Biol. Open* **3**, 794–802 (2014).
89. Barkan, A. & Small, I. Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* **65**, 415–442 (2014).
90. Bishop, A. C., Xu, J., Johnson, R. C., Schimmel, P. & de Crécy-Lagard, V. Identification of the tRNA-dihydrouridine synthase family. *J. Biol. Chem.* **277**, 25090–25095 (2002).
91. Quiles, M. A. J. & López, N. I. Photoinhibition of photosystems I and II induced by exposure to high light intensity during oat plant growth: Effects on the chloroplast NADH dehydrogenase complex. *Plant Sci.* **166**, 815–823 (2004).
92. Braun, H. P. & Zabaleta, E. Carbonic anhydrase subunits of the mitochondrial NADH dehydrogenase complex (complex I) in plants. *Physiologia Plantarum* **129**, 114–122 (2007).
93. Rípodas, C., Dalla Via, V., Aguilar, O. M., Zanetti, M. E. & Blanco, F. A. Knock-down of a member of the isoflavone reductase gene family impairs plant growth and nodulation in *Phaseolus vulgaris*. *Plant Physiol. Biochem.* **68**, 81–89 (2013).
94. Gang, D. R. *et al.* Evolution of plant defense mechanisms: relationships of phenylcoumaran benzylic ether reductases to pinoresinol-lariciresinol and isoflavone reductases. *J. Biol. Chem.* **274**, 7516–7527 (1999).
95. Poku, S. A., Seçgin, Z. & Kavas, M. Overexpression of Ks-type dehydrins gene OeSRC1 from *Olea europaea* increases salt and drought tolerance in tobacco plants. *Mol. Biol. Rep.* **46**, 5745–5757 (2019).
96. Tähtiharju, S. & Palva, T. Antisense inhibition of protein phosphatase 2C accelerates cold acclimation in *Arabidopsis thaliana*. *Plant J.* **26**, 461–470 (2001).
97. Luan, S. Protein phosphatases in plants. *Annu. Rev. Plant Biol.* **54**, 63–92 (2003).
98. Smith, R. D. & Walker, J. C. Plant protein phosphatases. *Annu. Rev. Plant Biol.* **47**, 101–125 (1996).
99. Moorhead, G. B. G., De Wever, V., Templeton, G. & Kerk, D. Evolution of protein phosphatases in plants and animals. *Biochem. J.* **417**, 401–409 (2009).
100. Sekimata, M., Ogura, K., Tsumuraya, Y., Hashimoto, Y. & Yamamoto, S. A β-galactosidase from radish (*Raphanus sativus* L.) seeds. *Plant Physiol.* **90**, 567–574 (1989).
101. Smith, D. L. & Gross, K. C. A family of at least seven β-galactosidase genes is expressed during tomato fruit development. *Plant Physiol.* **123**, 1173–1184 (2000).
102. Ahsan, N. *et al.* Analysis of arsenic stress-induced differentially expressed proteins in rice leaves by two-dimensional gel electrophoresis coupled with mass spectrometry. *Chemosphere* **78**, 224–231 (2010).
103. Adkins, S. Tomato spotted wilt virus—positive steps towards negative success. *Mol. Plant Pathol.* **1**, 151–157 (2000).
104. Niu, D. *et al.* Bacillus cereus AR156 primes induced systemic resistance by suppressing miR825/825* and activating defense-related genes in Arabidopsis. *J. Integrat. Plant Biol.* **58**, 426–439 (2016).
105. He, X.-F., Fang, Y.-Y., Feng, L. & Guo, H.-S. Characterization of conserved and novel microRNAs and their targets, including a TuMV-induced TIR–NBS–LRR class R gene-derived novel miRNA in Brassica. *FEBS Lett.* **582**, 2445–2452 (2008).
106. Gassmann, W., Hinsch, M. E. & Staskawicz, B. J. The Arabidopsis RPS4 bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J.* **20**, 265–277 (1999).
107. Meyers, B. C., Morgante, M. & Michelmore, R. W. TIR-X and TIR-NBS proteins: Two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. *Plant J.* **32**, 77–92 (2002).

108. Ding, C.-K., Wang, C. Y., Gross, K. C. & Smith, D. L. Reduction of chilling injury and transcript accumulation of heat shock proteins in tomato fruit by methyl jasmonate and methyl salicylate. *Plant Sci.* **161**, 1153–1159 (2001).
109. Jacob, P., Hirt, H. & Bendahmane, A. The heat-shock protein/chaperone network and multiple stress resistance. *Plant Biotechnol. J.* **15**, 405–414 (2017).
110. Stukkens, Y. *et al.* NpPDR1, a pleiotropic drug resistance-type ATP-binding cassette transporter from *Nicotiana plumbaginifolia*, plays a major role in plant pathogen defense. *Plant Physiol.* **139**, 341–352 (2005).
111. Crouzet, J., Trombik, T., Fraysse, Å. S. & Boutry, M. Organization and function of the plant pleiotropic drug resistance ABC transporter family. *FEBS Lett.* **580**, 1123–1130 (2006).
112. Food, É.-U., Administration, D., Safety, C. f. F. & Nutrition, A. *Bacteriological Analytical Manual (BAM)*. (Éditeur non identifié, 2020).
113. Posada, F., Aime, M. C., Peterson, S. W., Rehner, S. A. & Vega, F. E. Inoculation of coffee plants with the fungal entomopathogen *Beauveria bassiana* (Ascomycota: Hypocreales). *Mycol. Res.* **111**, 748–757 (2007).
114. Tao, S.-Q., Auer, L., Morin, E., Liang, Y.-M. & Duplessis, S. Transcriptome analysis of apple leaves infected by the rust fungus *Gymnosporangium yamadae* at two sporulation stages. *Mol. Plant-Microbe Interact.* **33**, 444–461. https://doi.org/10.1094/mpmi-07-19-0208-r (2020).
115. Lorenz, W. W. *et al.* Conifer DBMagic: A database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. *Tree Genet. Genomes* **8**, 1477–1485. https://doi.org/10.1007/s11295-012-0547-y (2012).
116. Wang, F. *et al.* Embryonal control of yellow seed coat locus ECY1 is related to alanine and phenylalanine metabolism in the seed embryo of *Brassica napus*. *G3 (Bethesda)* **6**, 1073–1081. https://doi.org/10.1534/g3.116.027110 (2016).
117. Ullmann, R. *et al.* Genomic adaption and mutational patterns in a HaCaT subline resistant to alkylating agents and ionizing radiation. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms22031146 (2021).
118. Shi, F. *et al.* Whole-transcriptome analysis and construction of an anther development-related ceRNA network in Chinese cabbage (*Brassica campestris* L. ssp. pekinensis). *Sci. Rep.* **12**, 2667. https://doi.org/10.1038/s41598-022-06556-2 (2022).
119. Chen, L.-Y. *et al.* Characterization of transcriptome and development of novel EST-SSR makers based on next-generation sequencing technology in *Neolitsea sericea* (Lauraceae) endemic to East Asian land-bridge islands. *Mol. Breed.* **35**, 187. https://doi.org/10.1007/s11032-015-0379-1 (2015).
120. Yan, L.-P. *et al.* De novo transcriptome analysis of Fraxinus velutina using Illumina platform and development of EST-SSR markers. *Biologia Plantarum* **61**, 210–218. https://doi.org/10.1007/s10535-016-0681-8 (2017).
121. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771. https://doi.org/10.1093/nar/gkp1137 (2010).
122. Pertea, G. *et al.* TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651–652. https://doi.org/10.1093/bioinformatics/btg034 (2003).
123. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512. https://doi.org/10.1038/nprot.2013.084 (2013).
124. Kim, C. S., Winn, M. D., Sachdeva, V. & Jordan, K. E. K-mer clustering algorithm using a MapReduce framework: Application to the parallelization of the Inchworm module of Trinity. *BMC Bioinform.* **18**, 1–15 (2017).
125. Good, I. J. Normal recurring decimals. *J. Lond. Math. Soc.* **S1–21**, 167–169. https://doi.org/10.1112/jlms/s1-21.3.167 (1946).
126. Chen, Y., Ye, W., Zhang, Y. & Xu, Y. High speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Res.* **43**, 7762–7768. https://doi.org/10.1093/nar/gkv784 (2015).
127. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877. https://doi.org/10.1101/gr.9.9.868 (1999).
128. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).
129. Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, o. b. o. t. I. N. S. D. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19-D21 (2010). https://doi.org/10.1093/nar/gkq1019
130. Hunter, S. *et al.* InterPro: The integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
131. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
132. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031 (2014).
133. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60. https://doi.org/10.1038/nmeth.3176 (2015).
134. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121. https://doi.org/10.1093/nar/gkt263 (2013).
135. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Geneti.* **16**, 276–277 (2000).
136. Riaño-Pachón, D. M., Ruzicic, S., Dreyer, I. & Mueller-Roeber, B. PlnTFDB: an integrative plant transcription factor database. *BMC Bioinform.* **8**, 42. https://doi.org/10.1186/1471-2105-8-42 (2007).
137. Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115. https://doi.org/10.1093/nar/gks596 (2012).
138. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585. https://doi.org/10.1093/bioinformatics/btx198 (2017).
139. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628. https://doi.org/10.1038/nmeth.1226 (2008).
140. Murtagh, F. in *International Encyclopedia of Statistical Science* (ed Miodrag Lovric) 633–635 (Springer, 2011).
141. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
142. Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995. https://doi.org/10.1101/gr.7.10.986 (1997).
143. Martin Bland, J. & Altman, D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**, 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8 (1986).
144. Altman, D. G. & Bland, J. M. Measurement in medicine: The analysis of method comparison studies. *J. R. Stat. Soc. Ser. D (Statistician)* **32**, 307–317. https://doi.org/10.2307/2987937 (1983).
145. Cleveland, W. S. *Visualizing data*. (At & T Bell Laboratories ; [Published by Hobart Press], 1993).
146. Bulmer, M. Galton's law of ancestral heredity. *Heredity* **81**, 579–585. https://doi.org/10.1046/j.1365-2540.1998.00418.x (1998).
147. Jin, W. *et al.* The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.* **29**, 389–395. https://doi.org/10.1038/ng766 (2001).
148. Cui, X. & Churchill, G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210. https://doi.org/10.1186/gb-2003-4-4-210 (2003).
149. Bertin, J. *Semiology of Graphics*. (University of Wisconsin Press, 1983).
150. Fisher, R. A.
151. Johnson, N. L., Kemp, A. W. & Kotz, S. *Univariate Discrete Distributions*. Vol. 444 (Wiley, 2005).

152. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646. https://doi.org/10.1093/nar/gkac1000 (2023).

Acknowledgements
We appreciate the helpful discussions and input from Precision Agriculture Lab, CAS-AFS, UAF, and the National Center for Genome Editing for Crop Improvement and Human Health (NCGE). Authors are further oblidged to NCGE for financial support to pay Article processing fee for the publication of this manuscript.

Author contributions
U.B.Z.: Validation, Investigation. M.S.: Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization. R.M.A.: Conceptualization, Methodology. S.H.K.: Writing—Review & Editing, Resources. M.Z.N.: Writing—Review & Editing, Resources. KS: Investigation, Resources. N.C.: Investigation, Resources. F.S.A.: Writing—Review & Editing, Resources, Project Administration. M.T.A.: Writing—Review & Editing. I.A.R: Conceptualization, Methodology, Resources, Supervision, Project Administration, Funding Acquisition, Writing—Review & Editing.

Funding
This research was supported by funding from the Punjab Agriculture Research Board (PARB) under Grant 'PARB Project No. 883'. The PARB had no role in the design of the study, data collection, analysis, interpretation of data, or writing of the manuscript.

Competing interests
The authors declare that they have no conflicts of interest regarding the research, authorship, and publication of this article. The work described in this paper was conducted independently, and the authors have no financial or personal relationships with individuals, organizations, or companies that could potentially bias the findings or interpretations presented herein.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-45982-8.

Correspondence and requests for materials should be addressed to I.A.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.