

Transcriptome Analysis Reveals Drought Stress response Mediated by Biological Processes and Key Pathways in *Gossypium Darwinii*

Cuilian Xu

State Key Laboratory of Cotton Biology, Tarium University

Muhammad Kashif Ilyas

Pakistan Agricultural Research Council

Richard Odongo Magwanga

Chinese Academy of Agricultural Sciences Cotton Research Institute

Hejun Lu

Chinese Academy of Agricultural Sciences Cotton Research Institute

M Kashif Riaz Khan

NIAB: Nuclear Institute for Agriculture and Biology

Zhongli Zhou

Chinese Academy of Agricultural Sciences Cotton Research Institute

Yujun Li

Cotton Sciences Research Institute of Hunan

Zhengcheng Kuang

Cotton Sciences Research Institute of Hunan

Abdul Ghafoor (✉ ghafoor59pk@yahoo.com)

Pakistan Agricultural Research Council <https://orcid.org/0000-0002-3079-1021>

Asif Javaid

PARC: Pakistan Agricultural Research Council

Danish Ibrar

PARC: Pakistan Agricultural Research Council

Kunbo Wang

CAAS CRI: Chinese Academy of Agricultural Sciences Cotton Research Institute

Fang Liu

CAAS CRI: Chinese Academy of Agricultural Sciences Cotton Research Institute

Haodong Chen

CAAS CRI: Chinese Academy of Agricultural Sciences Cotton Research Institute

Research Article

Keywords: *Gossypium darwinii*, Abiotic stress, Transcriptome sequencing, Gene expression

Posted Date: April 20th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-534041/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background The allotetraploid wild cotton species, *Gossypium darwinii*, possess important traits of stress tolerance and good genetic stability that can be used for cotton improvement.

Materials and Methods In this study, we analyzed the RNA-seq transcriptomes from leaves of *G. darwinii* seedlings with and without drought stress. A total of 86.7 million valid reads with an average length of 95.79 bp were generated from the two samples and 58,960 transcripts with a length of more than 500 bp were assembled. We searched the known proteins on the strength of sequence similarity; these transcripts were annotated with COG, KEGG and GO functional categories. According to gene expression abundance RPKM value, we carried out RT-qPCR analysis to determine the expression pattern of the obtained transcription factors.

Results A total of 58,961 genes was differentially expressed (DEG), with 32,693 and 25,919 genes found to be upregulated and downregulated, respectively. Through gene ontology and KEGG pathways, the upregulated genes were found to associate with all the GO terms, molecular functions (MF), biological process (BP) and cellular components (CC), which are highly linked to enhancing drought stress tolerance.

Conclusion The results obtained provide valuable information and a large number of transcripts, which can be exploited by cotton breeders in developing a more drought stress tolerant cotton germplasm.

Introduction

Drought is one of the major abiotic stresses that affect plant growth and yield [1]. The drought stress is the dominant form of abiotic stress with huge economic impact globally [2]. Growth and productivity of cotton are severely reduced especially during the period of seed germination and seedlings under drought [3]. Improvement of cultivated cotton through conventional methods of breeding has achieved minimal progress, this due to the genome nature of the cultivated cotton [4]. The genetic diversity of the cultivated cotton has been extensively eroded and narrowed, due to intensive selection and inbreeding [3]. Adoption of recent techniques in molecular breeding, will grateful help in solving the perennial problem of reduction in yield and yield quality due to effects of various abiotic stress factors [5, 6]. The wild progenitors harbors a tremendous amount of untapped genetic diversity that can be exploited to broaden the genetic base of cotton[7]. Moreover, important stress responsive genes, I have been identified from populations developed from *Gossypium hirsutum*, an elite cultivars and *Gossypium tomentosum*, a wild cotton genotype known to be highly adaptive to saline and dry areas [3]. Further, by use of simple sequence repeat (SSR) markers, Kirungu et al [8], identified 27 key genes from an F_{2:3} generation derived from two wild cotton species of the D genome, *Gossypium davidsonii* and *Gossypium klotzschianum*.

Plants have evolved a number of survival strategies, one of which is the induction of various stress responsive genes [9]. The nature of the abiotic stresses are complex and thus for plants to adopt in the ever hostile environments, plants adopt physiological, morphological and biochemical mechanisms to improve their tolerance levels [10, 11]. Drought tolerance mechanism in plants is being controlled by multi-gene effects and a series of complex regulatory systems [12, 13]. Several genes have been identified from model plants, such as *Arabidopsis thaliana*, which are involved in the metabolism of different levels, signal transduction, stress response, gene expression and regulation [14, 15]. Next-generation sequencing (NGS) techniques, such as the Solexa-seq has been employed to aid in a deeper understanding of gene expression pattern in plants when exposed to drought stress condition [16]. Moreover, the effects of drought stress on cotton growth and development is exerted through several pathways.

Gossypium is the largest plant genus among the terrestrial plants with over 50 species [17]. It is an important cash crop, which play an important role in the economies of several countries globally, among them China[18]. Among the cultivated cotton cultivars, 90% of the cotton production is obtained from *G. hirsutum*, a tetraploid cotton [19]. But due to the narrow genetic base of the cultivated species, and sensitivity to various stress factors, breeding for high yielding, superior quality and resistance to various stresses remains a major challenge globally [20]. Therefore, it is imperative to devise various mechanisms to unravel the problem of a narrow genetic base and improve the genetic diversity of the cultivated cotton. One of the most certain ways is through the exploration of various plants transcription factors, mainly from the wild progenitors, being several of the novel genes have been identified among the wild germplasms [20]. In the current study, a wild cotton, *Gossypium darwinii*, was used in

determining the possible alleles responsible for enhancing their drought stress tolerance. RNA sequencing was carried out by using young tender leaf samples of *G. darwinii* exposed to drought stress and normal condition. The sequencing was done through Solexa high-throughput sequencing, where high throughput sequencing data were generated, and then analyzed through bioinformatics tools, with the result of assembling of total expressed genes, and unigenes annotation through blast the public database, gene expression information compared with the RPKM.

Materials And Methods

Experimental material

We used RNA samples prepared from an accession AD₅₋₃ of *Gossypium darwinii*, allotetraploid wild cotton, under drought stress and normal environments. This accession was grown in the National Wild Cotton Nursery, managed by the Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, in Sanya, Hainan Province in China.

Plant material and growth

To ensure maximum germination, a small slit was made on the seed coat, and they were then pre-germinated in the incubator using sterilized moist filter paper for 2–3 days until the radicle was approximately 1 cm long and the condition of the incubator was set at 28°C with >90% ambient humidity. The seedlings were then transplanted into sterilized vermiculite filled in a small conical pot with dimensions of 5 cm bottom region, 7 cm top region, and 8 cm in depth. After seed planting 14 days, at phase three-leaf stage, the treated group (drought stress) was stopped irrigation 6 days, while the contrast normal group was continued watering to keep optimum growth moisture contents. The moisture contents were detected constantly by weighing method. The first sampling was taken when the moisture contents of all pots of drought-stress group (treated group) reduced to 10%-8% naturally after watering was withdrawn after two days, then, further three samples were collected at 2 days intervals and the moisture contents reduced down to 7%-6%, 6%-5%, 5%-4% respectively. Five leaves were sampled every time for both treated and control group. All the sampling stages continued for 6 days.

The pretreatment and assembly of the sequence

Isolated poly (A) RNA from total RNA with using beads with Oligo (dT) from the cotton tissue samples. Then, interrupted the mRNA and into short fragments. Suitable fragments were selected for the PCR amplification as templates to prepare cDNA library, the size of the inserts is about 200 bp. The Illumina/Solexa approach involved the sequencing of cDNA fragments, observed and counting the number of times to a particular fragment. Many raw reads were got by Solexa paired-end sequencing, and then be pretreated to decrease error rate. The sliding window approach involved the filtering process. First, we removed reads that did not pass the built-in Illumina's software Failed-Chastity filter according to the relation "failed-chastity ≤ 1 ", using a chastity threshold of 0.6, on the first 25 cycles. Secondly, we discarded all reads with adaptor contamination. Thirdly, we ruled out low-quality reads with ambiguous sequences "N". Finally, the reads with more than 10% Q < 20 bases were also removed. After the low quality, dirty raw reads were filtered, we combined all the Treated-samples, CK-samples, respectively, the reassembly process was completed with short reads assembling program Trinity. Finally, the unigenes of each sample were used for further processes of sequence splicing and then using sequence clustering software removed the redundancy sequence, which acquires non-redundant unigenes (here called 'transcripts').

Reads quality pretreatment and denovo assembly

A large amount of sample data was obtained by paired-end sequencing of Solexa RNA. In view of the impact of the Solexa data error rate on the results, the raw data was quality preprocessed. Quality pretreatment steps: 1) Sliding window method to remove low quality segments: quality threshold 20 (error rate = 1%), window size 5 bp, length threshold 35 bp 2). The short reads were filtered by removing adaptor contaminations and low-quality reads with base N's or more than 10% of bases with a quality score < 20. The de novo assembly of the WGS short reads was performed by SOAPdenovo (<http://soap.genomics.org.cn/SOAPdenovo.html>), with kmer = 83. The assembled sequences were filtered by removing those aligned to microbial and human genomes (minimum aligned length is 100 bp) downloaded from GenBank with BWA-mem (<http://bio-bwa.sourceforge.net/>) default parameters. Excision of N-part sequence in reads: length threshold 35bp. The valid reads of 2 samples of cotton were combined for de novo splicing, using the software Trinity, version trinityrnaseq_r2011-11-26, using the

paired-end splicing method. The splicing sequence was de-duplicated, and finally 400,708 transcripts larger than 100 bp in length were obtained, and the size was 125 Mb. The longest transcript under each Loci was obtained as Unigene, and 207,241 Unigenes were obtained, and the size was 64 Mb. The sequencing libraries were prepared with the standard protocol from PacBio and sequenced by Wuhan Nexomics (<http://www.nextomics.cn>) on the PacBio long read sequencing instrument, model RSII, with both P5 Polymerase Binding and C3 Chemistry Kits and P6 Polymerase Binding and C4 Chemistry Kits.

Comparison of genes and public databases

The sample genes were compared to the public data gene, and functional annotation was performed by the similarity of the genes. Compared with nr, SWISS-PROT, TrEMBL, Cdd and pfam libraries. Gene similarity alignments are primarily based on the BLAST algorithm. BLAST, the full name of the Basic Local Alignment Search Tool, is a "search tool based on local alignment algorithm", published by Altschul et al. Blast is able to perform a function of comparing the similarity between two nucleic acids or protein sequences. It can quickly find similar sequences between two sequences and compare the regions to determine the similarity. The sample gene sequences were aligned with SWISS-PROT, TREMBL, CDD, PFAM, and NR libraries, respectively, with similarities > 30%, and $e < 1e-5$ annotations, and all annotation details obtained from the combined genes.

Functional annotation of the transcripts

The transcripts with more than 500 bp were used to carry out further analysis of the transcriptions annotation. A BLASTx alignment (Similarity > 30%, E value < 10^{-5}) between transcripts and protein databases like Conserved Domain Database (CDD, <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>), Swiss-Prot protein database (<http://www.expasy.ch/sprot>), protein families (PFAM) database (<http://pfam.xfam.org/>), NCBI non-redundant protein database (<http://www.ncbi.nlm.nih.gov>) and TrEMBL (<http://www.ebi.ac.uk/uniprot>). Then combine annotation genes on these different databases, and the sequence direction of unigenes based on the best aligning results. The Nr, Swiss-Prot, KEGG, and COG were followed to decide the direction is a priority order when the results of different databases conflicted. And software named ESTScan [21] will decide its coding regions and sequence direction of when a unigene happened to be unaligned to none of the above databases. And using the RPKM method to calculate the expression of transcripts [22] below showed the computational formula of RPKM. $RPKM = 10^6C / (NL10^{-3})$, (N is the total number of reads that uniquely aligned to all transcripts in the specific sample, C is the number of reads that uniquely aligned to the transcript, and L denotes the number of bases of the transcript. The FDR method was applied to determine the threshold of P-values in multiple tests. We use " $FDR \leq 0.001$ and the absolute value of $\log_2$ ratio ≥ 1 " as the critical value to evaluate the significance of the expression difference gene).

Results

RNA presequencing and data quality

The data retention rate is high and the data quality is good. The total RNA-seq data is 86,700, and the average length is 95.79 bp. The samples are all in accordance with the 4G requirements, which meets the customer and analysis needs.

Table 1
The statistical results of the data before and after the quality pretreatment are as follows

Sample	Raw Data		Valid Data			Valid Ratio (reads)
	Read	Base	Read	Base	Average length	
Treat	51104338	5110433800	44318328	4240686660	95.69	86.72%
Control	48321894	4832189400	42394518	4065816222	95.90	87.73%
All	99426232	9942623200	86712846	8306502882	95.79	87.21%

3.2 De novo Assembly

The valid reads of 2 samples of cotton were combined for de novo splicing, using the software Trinity, version trinityrnaseq_r2011-11-26, using the paired-end splicing method. The splicing sequence was reduplicated, and finally 400,708 transcripts larger than

100 bp in length were obtained, and the size was 125 Mb. The longest transcript under each locus was obtained as Unigene, and 207,241 Unigenes were obtained, and the size was 64 Mb.

Table 2
The statistical table of results of de novo splicing

	All (>= 100bp)	>=500 bp	>=1000 bp	N50	N90	Total Length	Max Length	Min Length	Average Length
Transcript	400708	58960	15884	397	145	125386375	11381	101	312.91
Unigene	207241	30858	9802	429	133	64511685	11381	101	311.29

The transcriptome sequencing, stitching and the GC contentdetermination of the analysed transcription data

After pretreatment, 9.94 million raw reads were obtained, and after filtering 8.67 million valid reads with average length of 95.79bp were produced by Solexa sequence. Moreover, 400,708 transcripts with lengths greater than or equal to 100bp were achieved by denovo assembly after filtering out the redundant sequence (Fig. 1A). Only one longest transcript of every locus was reserved with 207,241 unigenes. Furthermore, only transcripts with length greater than or equal to 500bp were further analyzed. Moreover, the GC analysis of all the transcripts (≥ 100 bp) revealed the feature of the ratio of canonical bases that shows the average content ratio of GC of all these transcripts that is 39.7%, and AT is 60.3%. The least GC content was 8.49% while the highest was 81.67%; and 90% of all the sequences contained GC of more than 90% (Fig. 1B).

Gene annotation and function classification

We annotated the transcripts (≥ 500 bp) based on 5 different protein database, namely, conserved domain database (cdd), non-redundant (nr), protein families (pfam), sprot and TrEMBL, in which 21,609; 38,175; 20,993; 38,322 and 32,370 transcript were annotated, however, non-redundant and pfam databases provided the highest proportion of annotated transcript with 100% similarity (Table xxx). Further analysis on the annotated transcripts were analysed by blasting in other plant genomes, in which most of the sequences were matched by cotton transcript belong to *Vitisvinifera* L. (Vitis), *Ricinuscommunis*L. (Ricinus) and *Populus*, with the number 10,360, 9,856, 8,192, and the ratio in all transcripts were 27.1%, 25.8%, 21.2%, respectively (Figure xxxx). Moreover, only 1,381 transcripts matched with the known sequences from cotton with the ratio of 3.6% (Fig. 2A, Table 1)

Table 3
Sequencing and identification of the transcription based on various databases

Percentage identity similarities												
Protein databases	100%	90-99.9%	80-89.9%	70-79.9%	60-69.9%	50-59.9%	40-49.9%	30-39.9%	20-29.9%	10-19.9%	0-9.9%	Totals
cdd	6	435	1,351	2029	2896	4245	5447	4755	443	2	0	21609
nr	353	3787	8,812	10388	7756	4601	1946	531	1	0	0	38175
sprot	50	761	2,627	3968	4318	4121	3339	1793	16	0	0	20993
trembl	188	2770	7,562	10268	8468	5635	2845	582	4	0	0	38322
pfam	248	5010	9,183	7351	4876	3204	1751	733	14	0	0	32370
Totals	845	12763	29535	34004	28314	21806	15328	8394	478	2	0	151,469

Clusters of Orthologous Groups of proteins (COGs)of the annotated transcripts.

The rational of classifying the proteins encoded in sequenced genomes is important for making the genome sequences maximally useful for functional and evolutionary studies [24]. Out of the five databases, nr annotated transcripts were further analyzed and out of 38,175 Nr hits, 3,586 sequences were assigned to the COG classifications (Fig. 2B). Among the 23 COG categories, the

cluster for General function prediction only (584, 14%) represented the largest group, followed by Transcription (163, 5%), Replication, metabolism and carbohydrate transport (1,200, 4%), recombination and repair (217, 7%), protein turnover, Post-translational modification and chaperones (369, 9%), and Translation, Signal transduction mechanisms (1,487, 3%), biogenesis and ribosomal structure (11,615, 75%), only 5 unigenes were assigned to Cell motility and secretion (Fig. 2B)

The result of KEGG pathway analysis showed that 13,088 (22.2% of all the annotation genes) transcripts (several transcripts hit multiple pathways) mapped to 284 pathways belong to all the five categories of KEGG, including metabolism, genetic information processing, environmental information processing, cellular processes, and organismal systems. Out of the ten pathways, the plant hormone signal transduction pathway was regulated by the highest number of transcripts, 446 transcripts, which were mainly responsible for the signal transport, a critical process being the very first step initiated by various molecules in the plants when plant are exposed to drought stress. Secondly, the metabolism transport, which included the nucleotide and sucrose metabolism, this metabolism progress, mainly induces more genes, which are involved in the oxidative phosphorylation pathway; moreover, many genes were found to be involved in plant-pathogen interaction.

Gene Ontology (GO) analysis

Gene ontology provides the very fundamental role or functions of the various genes. The genes function are classified into three GO functional annotation, the biological process (BP), cellular component (CC) and molecular function (MF) [25]. We carried out GO functional analysis using Blast2GO program [26]. A higher proportion of the transcripts were found to be associated with various GO functional terms, 164, 660 transcripts harbored GO functions. All the GO functions were detected, in relation to cellular component, 342 functions were obtained and found to be associated with 83,572 transcripts, with the cell part (GO:0044464) and cell (GO:0005623) had the highest number of transcripts 11,332 in each, accounting for 13.6% of all the transcripts. In molecular functions, 818 GO functions were obtained, which associated with 26,490 transcripts, with binding (GO:0005488), and catalytic activity (GO:0003824) found to be associated with the highest number of transcripts, 18,778 and 16,604, respectively. Finally, in relation to the biological process, 998 GO terms were detected, with were linked to 54,598 transcripts. The metabolic process (GO:0008152), cellular process (GO:0009987), primary metabolic process (GO:0044238), and cellular metabolic process (GO:0044237), were the dominant biological processes associated with the highest number of transcripts, 16,448 (30.1%), 14,386 (26.3%), 12,061 (22.1%), and 11,717 (21.5%) transcripts, respectively. The various GO functions detected, have been found to associated with a number of stress responsive genes, such as the late embryogenesis abundant (LEA)[27], the multidrug and toxic compound extrusion (*MATE*) gene family [28], cyclin dependent kinase (*CDK*) genes [10], which showed that the transcripts could be playing a significant role in enhancing drought stress tolerance in *Gossypium darwinii*.

The analysis of differential expression genes (DEGs) and enrichment analysis

We evaluated all the transcripts detected in this, a total of 58, 639 were found to be differentially expressed genes (DEGs), in which 32,693 (55.75%) were upregulated genes, 127 (0.2%) not expressed and 25,919 (44.2%) were downregulated. The ratio of the differential expression genes indicated that the expression change of cotton genes to drought stress mainly is upregulated. We performed GO analysis of the upregulated genes, which was aimed to validate their biological function within the plants under drought stress conditions, 17,028 of the upregulated genes were associated by various GO terms, majority of which were highly associated with various forms of abiotic stress factors. The results showed that *Gossypium darwinii* perhaps has a higher level of drought stress tolerance and could be used in breeding for more drought stress tolerant cotton germplasms. The KEGG was performed for the signal pathway analysis of high abundant differential expression genes. Upregulated genes mainly involved in the progress of glycerophospholipid metabolism, Glycolysis/Gluconeogenesis, amino sugar and nucleotide sugar metabolism, Lysosome, Alanine, aspartate and glutamate Fatty acid metabolism, metabolism, Pyruvate metabolism, Galactose metabolism, Cysteine and methionine metabolism (Fig. 4A). Among the various KEGG pathways, of significance was the ko04120, Ubiquitin mediated proteolysis. When plants are exposed to drought stress, the equilibrium and delicate balance between reactive oxygen species (ROS) release and elimination, becomes altered, leading to excessive accumulation of the ROS within the cell [29]. Excess accumulation of ROS triggers oxidative stress, which eventually lead to plant death [30]. The ubiquitin mediated proteolysis, aids in the elimination of various reactive oxygen species, thus reducing the lethal effects, and maintains the plants survival under stress [31].

Table 4
KEGG pathways

	Pathway	Count	Percentage	Pathway ID
1	Plant hormone signal transduction	446	3.41%	ko04075
2	RNA transport	361	2.76%	ko03013
3	Spliceosome	339	2.59%	ko03040
4	Purine metabolism	283	2.16%	ko00230
5	Oxidative phosphorylation	272	2.08%	ko00190
6	Ribosome	256	1.96%	ko03010
7	Glycolysis/ Gluconeogenesis	234	1.79%	ko00010
8	Starch and sucrose metabolism	220	1.68%	ko00500
9	Plant-pathogen interaction	207	1.58%	ko04626
10	Protein processing in endoplasmic reticulum	198	1.51%	ko04141

Discussion

De novo assembly of G. darwinii mRNA-seq data Sets of G. darwinii seedlings under drought stress

Being member of tetraploid genus, *Gossypium darwinii* has many excellent economical characters in the aspect of fiber quality and resistance to unfavorable environment, which mostly lacked in cultivated cotton. The first step to explore fine gene factors from *G. darwinii* and transgene them to cultivated cotton is to obtain the genetic information of *G. darwinii*. In this study, our samples at multiple times from the material grown in the soil with water content of 10%-4%, so the stress related genes could express as sufficient as possible. Through transcriptome de novo assembly, we assembled obtained 207,241 transcripts successfully in *G. darwinii* (N50 = 397 bp and length ≥ 100 bp). Through BLAST against the known SWISS-PROT, TREMBL, CDD, PFAM, NR database, there were 94,251 transcripts with hits, the others 112,990 (54.5%) may be considered as novel transcripts. A large amount of the cotton seed EST sequence got from *G. darwinii*, not only further enrich the EST sequence information in NCBI database, but also provide the sequence foundation for EST-SSR development.

High abundant differential expression genes

Studies have shown that under normal circumstances, the generation and cleaning of reactive oxygen were in a state of dynamic balance, various kinds of adversity stress would break the balance, which can make the excessive accumulation of active oxygen in cells and produce toxic action to organism. The second biggest kind of high abundant differentially expressed genes was stress reaction gene, the expression of which increased about 9 times under the drought stress from a high level. The third biggest kind of high abundant differentially expressed genes are phosphor-ethanolamine N-methyltransferase which are involved in metabolic reactions, the main function of this enzyme is a catalytic phosphate amine methylation, the final synthesis of choline phosphate, and this substance is a precursor of synthetic lipid Tai choline and betaine in plant. The liquidity of phosphatidyl choline and plant cell membrane; betaine is a kind of small molecular osmotic agent, structure and stability of permeability of cell regulation, large biological molecules have certain effect, also can affect the intracellular distribution of ions. In addition to that in the control expression is extremely low or nearly no expression, but after drought treatment, significantly expressed genes, such as cysteine endopeptidase (cysteine-type endopeptidase activity) compared with the control expression increased nearly 40 times, this enzyme as proteolytic enzymes responsible for the degradation of damaged or denatured proteins under stress conditions [23].

When *G. darwinii* grow under drought stress, the number of the downregulated genes is more than upregulated. There are 34 (23.4%) among all of 145 downregulated, genes are involved in photosynthesis process. The highest downregulated gene is light-harvesting complex II chlorophyll a/b binding protein, this protein take participate in light-harvesting and transition, and it is the most abundant antenna protein in thylakoid of plant chloroplast [24]. The higher downregulated gene is thiamine biosynthetic

enzyme, carbonic anhydrase, F-type H⁺-transporting ATPase subunit, (S)-2-hydroxy-acid oxidase, plastocyanin, ferredoxin, superoxide dismutase activity, especially the expression of xyloglucosyl transferase gene was significantly decreased under drought stress. Through the analysis of high abundant differential expression genes, we found that under drought stress, *G. darwinii* seedlings express many stress response genes, keep clear of intracellular accumulation of oxygen free radicals in time from poison injury. Through the osmotic regulation to protect the integrity of the cell membrane and a series of physiological and biochemical reaction to resist drought damage. There are 66 sequences of the 207 high abundant differential expression genes, did not match with existing database, so its function is unknown. This part of the sequence is likely to be involved drought resistance mechanism in *G. darwinii*; the study on the function of them has important significance to further clarify *G. darwinii* drought tolerance mechanism.

Conclusions

Transcriptome analysis of *G. darwinii* seedlings under drought stress was found that the pathway and gene of photosynthetic system were significantly downregulated. The higher downregulated gene including biosynthetic enzyme, carbonic anhydrase, F-type H⁺-transporting ATPase subunit, (S)-2-hydroxy-acid oxidase, plastocyanin, ferredoxin, superoxide dismutase, these genes involved in carbohydrate synthesis, transport system, electron transfer system and osmoregulatory system. It seems clear that the effects of drought stress on cotton growth and development is exerted through several pathways. The high mRNA levels of cysteine endopeptidase gene and the low mRNA levels of xyloglucosyl transferase gene were present under drought stress implied a severe premature senility in cotton. So we studied means that it will be an important direction to study and reply to drought stress from study earlier senescence caused by drought stress. Comparative transcriptome studies showed the gene expression under drought stress, the Unigene information provides the latest materials for SSR developing and the deeper study of the mechanism of cotton drought tolerance. The study will help to strengthen the tolerance to drought stress by molecular biotechnology to do an in-depth study on the commonality of the mechanism of response to drought between cotton and other plants.

Declarations

Author Contributions: H.C, C.X, M.K.I, R.O.M & F.L designed the experiments. C.X, H.C, M.K.R & H.L conceived the experiments and analyzed the results. C.X., H.C, H.L & M.K.R carried out all computational analyses. X.W, M.K.I, X.C, Y.L, F.L, participated in part of experiments directly or indirectly contributed reagents/materials/analysis tools. H.C, A.G, A.J, R.O.M, H.L, M.K.R, M.K.I, Y.L, & F.L drafted the manuscript and X.W., A.G, X.C, R.O.M & F.L. proofread and R.O.M, M.K.I. & A.G. revised the manuscript. All authors read and approved the final manuscript.

Funding: This program was financially sponsored by National Key R&D Program of China (2021YFE0101200), Hunan Natural Science Foundation Youth Fund (2020JJ5291), “Huxiang Young Talents Plan” Support Project of Hunan Province (2019RS2048), State Key Laboratory of Cotton Biology Open Fund (CB2021A14).

Acknowledgments: We wish to thank Sangon Biotech (Shanghai) Co., Ltd. for help in sequencing, analysis support and assembly. Authors are thankful to Shahbaz Khan (Pakistan) for inputs regarding manuscript formatting according the journal pattern.

Conflicts of Interest: The authors declare no conflict of interest.

Ethics approval

Authors declare that this study was conducted on wild cotton (*Gossypium Darwinii*) crop, and no experimentations were conducted on human or animal subjects using research material during the current studies. The study was approved by the relevant ethics committees.

References

1. Lipiec J, Doussan C, Nosalewicz A, Kondracka K (2013) Effect of drought and heat stresses on plant growth and yield: a review. Int Agrophysics 27:463–477. doi:10.2478/ntag-2013-0017

2. Azhar MT, Rehman A (2018) Overview on Effects of Water Stress on Cotton Plants and Productivity. In Biochemical, Physiological and Molecular Avenues for Combating Abiotic Stress Tolerance in Plants 297–316
3. Magwanga RO, Lu P, Kirungu JN, Diouf L, Dong Q, Hu Y, Cai X, Xu Y, Hou Y, Zhou Z, Wang X, Wang K, Liu F (2018) GBS mapping and analysis of genes conserved between *Gossypium tomentosum* and *Gossypium hirsutum* cotton cultivars that respond to drought stress at the seedling stage of the BC2F2 generation. *Int J Mol Science* 9(6):1614. doi:10.3390/ijms19061614
4. Bourland F, Myers GO (2015) Conventional Cotton Breeding
5. Bhau BS, Sharma DK, Bora M, Gosh S, Puri S, Borah B, Kumar DG, Wann SB (2016) Molecular Markers and Crop Improvement. In Abiotic Stress Response in Plants 381–410
6. Gill MB, Zeng F, Shabala L, Zhang G, Yu M, Demidchik V, Shabala S, Zhou M (2019) Identification of QTL related to ROS formation under hypoxia and their association with waterlogging and salt tolerance in Barley. *Int J Mol Sci* 20(3):699. doi:10.3390/ijms20030699
7. Shim J, Mangat PK, Angeles-shim RB (2018) Natural Variation in Wild *Gossypium* Species as a Tool to Broaden the Genetic Base of Cultivated Cotton. *J Plant Sci Curr Res* 2:005. doi:10.24966/PSCR-3743/100005
8. Magwanga RO, Lu P, Kirungu JN, Cai X, Zhou Z, Wang X, Diouf L, Xu Y, Hou Y, Hu Y, Dong Q, Wang K, Liu F (2008) Whole genome analysis of cyclin dependent kinase (CDK) gene family in cotton and functional evaluation of the role of *CDKF4* gene in drought and salt stress tolerance in plants. *Int J Mol Sci* 19(9):2625. doi:10.3390/ijms19092625
9. Anjum S, Xie X, Wang L (2011) Morphological, physiological and biochemical responses of plants to drought stress. *Afr J Agric Res* 6:2026–2032. doi:10.5897/AJAR10.027
10. Sourour A (2017) A review: Morphological, physiological, biochemical and molecular plant responses to water deficit stress. *Int J Eng Sci* 6:01–04. doi:10.9790/1813-0601010104
11. Molina-Bravo R, Zamora-Meléndez A (2016) QTLs for genetic improvement under global climate changes. In *Advances in Plant Breeding Strategies: Agronomic, Abiotic and Biotic Stress Traits*. 2:471–513. ISBN 9783319225180
12. Zhao C, Piao S, Wang X, Huang Y, Ciais P, Elliott J, Huang M, Janssens IA, Li T, Lian X, Liu Y, Müller C, Peng S, Wang T, Zeng Z, Peñuelas J (2016) Plausible rice yield losses under future climate warming. *Nat Plants* 3:16202. doi:10.1038/nplants.2016.202
13. Ashraf M (2010) Inducing drought tolerance in plants: Recent advances. *Biotechnol Adv* 28:169–183. doi:10.1016/j.biotechadv.2009.11.005
14. Amombo E, Li H, Fu J (2017) Research Advances on Tall Fescue Salt Tolerance: From Root Signaling to Molecular and Metabolic Adjustment. *J Am Soc Hortic Sci* 142:337–345. doi:10.21273/jashs04120-17
15. Kahl G (2015) Next-next-next generation sequencing (next 3 generation sequencing, third-generation sequencing). In *The Dictionary of Genomics. Transcriptomics and Proteomics* 1:1
16. Wendel JF, Cronn RC (2003) Polyploidy and the Evolutionary History of Cotton Polyploidy. and the Evolutionary History of Cotton
17. Magwanga RO, Kirungu JN, Lu P, Yang X, Dong Q, Cai X, Xu Y, Wang X, Zhou Z, Hou Y, Nyunja R, Agong SG, Hua J, Zhang B, Wang K, Liu F (2019) Genome wide identification of the trihelix transcription factors and overexpression of *Gh_A05G2067 (GT-2)*, a novel gene contributing to increased drought and salt stresses tolerance in cotton. *Physiol Plant* 167(3):447–464. doi: 10.1111/ppl.12920
18. Guo W, Cai C, Wang C, Zhao L, Wang L, Zhang TA (2008) A preliminary analysis of genome structure and composition in *Gossypium hirsutum*. *BMC Genomics* 9:314. doi:10.1186/1471-2164-9-314
19. Yuan Y, Wang X, Wang L, Xing H, Wang Q, Saeed M, Tao J, Feng W, Zhang G, Song XL, Sun XZ (2018) Genome-Wide Association Study Identifies Candidate Genes Related to Seed Oil Composition and Protein Content in *Gossypium hirsutum* L. *Front. Plant Sci* 9:1359. doi:10.3389/fpls.2018.01359
20. Volk GM, Richards CM, Henk AD, Reilley A, Miller DD, Forsline PL (2009) Novel diversity identified in a wild apple population from the kyrgyz republic. *Hort Sci* 44:516–518

21. Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* 138–148

22. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628. doi:10.1038/nmeth.1226

23. Dhillon A, Sharma K, Rajulapati V, Goyal A (2016) Proteolytic Enzymes. In *Current Developments in Biotechnology and Bioengineering: Production, Isolation and Purification of Industrial Products* 149–173. ISBN 9780444636621

24. Klimyuk VI, Persello-Cartieaux F, Havaux M, Contard-David P, Schuenemann D, Meierhoff K, Gouet P, Jones JDG, Hoffman NE, Nussaume LA (2007) Chromodomain Protein Encoded by the Arabidopsis *CAO* Gene Is a Plant-Specific Component of the Chloroplast Signal Recognition Particle Pathway That Is Involved in LHCP Targeting. *Plant Cell* 11:87. doi:10.2307/3870840

Figures

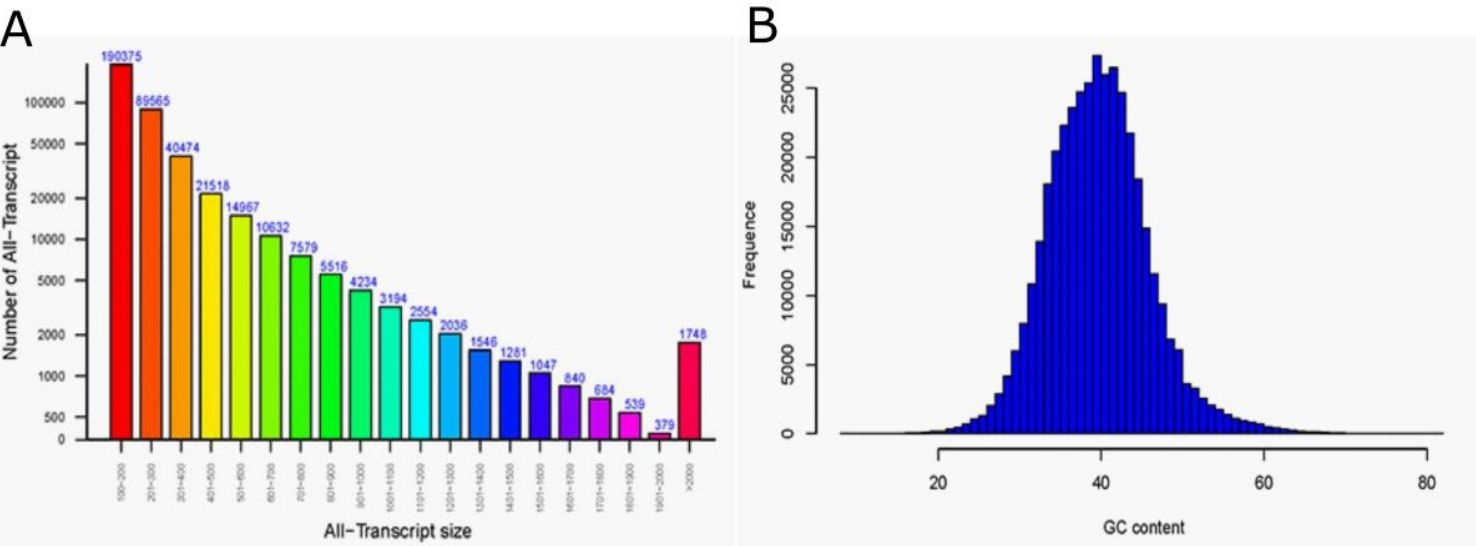
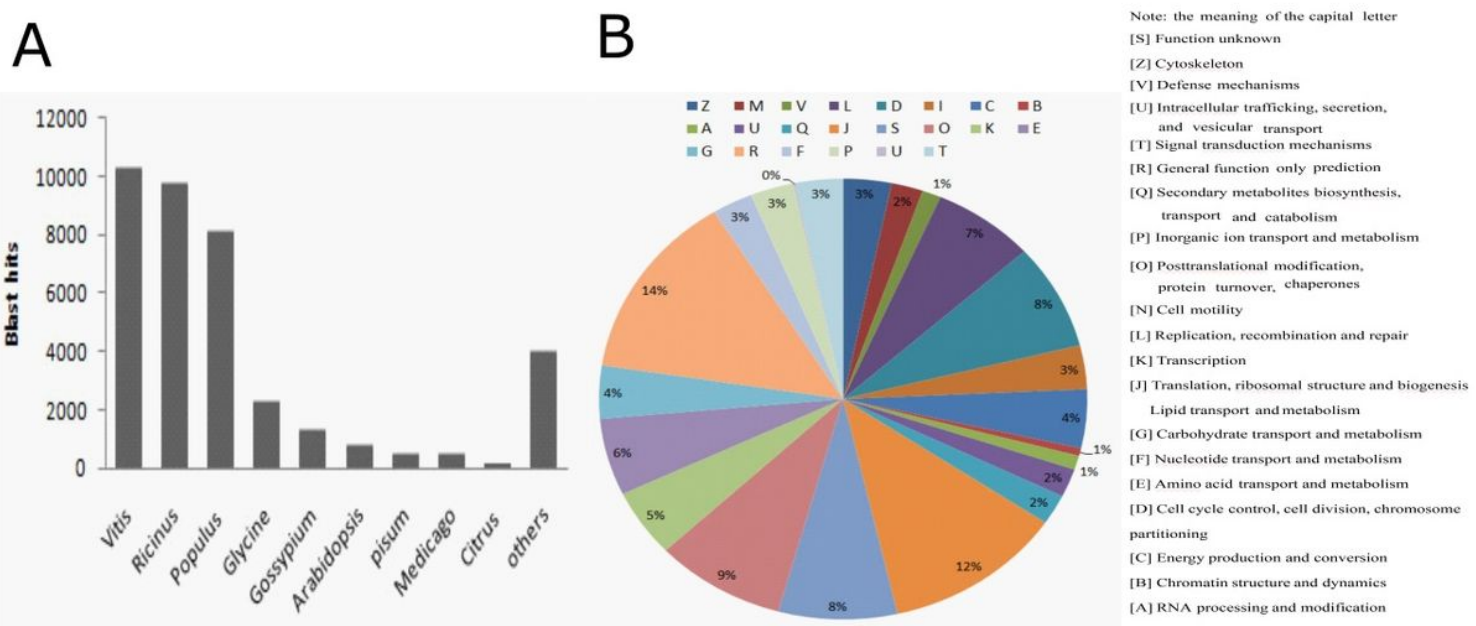


Figure 1

Transcriptome assembly and sequencing. (A) The length Distribution of the entire Transcript. (B). The GC content frequency distribution of transcript



BLAST hits retrieved from the NCBI nr database (A), COG function classification of Genes (B).

The GO classification of Gene



Pathway of glycerophospholipid metabolism (A), Pathway of photosynthesis (B).