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

Cuilian Xu
Chinese Academy of Agricultural Sciences Cotton Research Institute
Richard Magwanga (magwangarichard@yahoo.com)
Researchers
M Kashif Riaz Khan
Chinese Academy of Agricultural Sciences Cotton Research Institute
Zhongli Zhou
Chinese Academy of Agricultural Sciences Cotton Research Institute
Xiaoyan Cai
chinese academy of agricultural sciences
Yujun Li
Chinese Academy of Agricultural Sciences
Stephene Gaya Agong
Jaramogi Oginga Odinga University of Sciences and Technology
Kunbo Wang
Chinese Academy of Agricultural Sciences
Fang Liu
Chinese Academy of Agricultural Sciences
Haodong Chen
Chinese Academy of Agricultural Sciences

Research

Keywords: G. darwinii, Abiotic stress, Transcriptome sequencing, Gene expression

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-534041/v1

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. In this study, we analyzed the RNA-seq transcriptomes from leaves of *G. darwinii* seedlings with and without drought stress.

Results: 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. 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. 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.

Background

Drought is one of the major abiotic stresses that affect plant growth and yield (Lipiec et al., 2013). Growth and productivity of cotton are severely reduced especially during the period of seed germination and seedlings under drought (Magwanga et al., 2018b). Drought tolerance mechanism in plants is being controlled by multi-gene effects and a series of complex regulatory systems (Molina-Bravo & Zamora-Meléndez, 2016; Zhao et al., 2016). 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 (Ashraf, 2010; Amombo et al., 2017). Next-generation sequencing (NGS) techniques, such as the Solexa-seq has been employed to aid in a deeper understanding of gene expression patterns in plants when exposed to drought stress conditions (Kahl, 2015). Moreover, the effects of drought stress on cotton growth and development are exerted through several pathways. *Gossypium* is the largest plant genus among the terrestrial plants with over 50 species (Wendel & Cronn, 2003). It is an important cash crop, which plays an important role in the economies of several countries globally, among them China (Wang et al., 2018). Among the cultivated cotton cultivars, 90% of the cotton production is obtained from *G. hirsutum*, tetraploid cotton (Guo et al., 2008). 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 (Yuan et al., 2018). Therefore, it is imperative to devise various mechanisms to unravel the problem of a narrow genetic base and improve the genetic diversity of 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 (Volk et al., 2009, 2019).

In the current study, 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. 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.

Materials And Methods

Experimental material

We used RNA samples prepared from an accession $AD_{5\text{-}3}$ 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 conditions

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 28^0C 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 true leaves, 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 the weighing method. The first sampling was taken when the moisture contents of all pots of the 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 in both treated and control groups. All the sampling stages continued for 6 days.

Sequencing and analysis

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 the 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 the 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, then 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’).

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 TrEMBLI (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 (Iseli et al., 1999) 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 (Mortazavi et al., 2008) below showed the computational formula of RPKM.

$RPKM=10^6C/ (NL\ 10^{-3})$

(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 ≤0.001 and the absolute value of log2Ratio ≥1” as the critical value to evaluate the significance of the expression difference gene).

Results

The transcriptome sequencing, stitching and the GC content determination of the analyzed transcription data

After pre-treatment, 9.94 million raw reads were obtained, and after filtering 8.67 million valid reads with an average length of 95.79 bp were produced by Solexa sequence. Moreover, 400,708 transcripts with lengths greater than or equal to 100 bp were achieved by denovo assembly after filtering out the redundant sequence (Figure 1A). Only one longest transcript of every locus was reserved with 207,241 unigenes. Furthermore, only transcripts with length greater than or equal to 500 bp 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% (Figure 1B).

Gene annotation and function classification

Gene annotation

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 transcripts were annotated, however, non-redundant and Pfam databases provided the highest proportion of annotated transcript with 100% similarity (Table 1). Further analysis of the annotated transcripts was analyzed by blasting in other plant genomes, in which most of the sequences were matched by cotton transcript belong to *Vitis vinifera* L. (Vitis), *Ricinus communis* 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 2A-B). Moreover, only 1,381 transcripts matched with the known sequences from cotton with a ratio of 3.6% (Fig.2 A, Table 1).

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 (Karmirantzou & Hamodrakas, 2002). 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 (Figure 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 (Figure 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 plants are exposed to drought stress. Secondly, the metabolism and transport processes, which includes the nucleotide and sucrose metabolism, this metabolism progress, mainly induced 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) (Tiirikka et al., 2014). We carried out a GO functional analysis using the Blast2GO program (Conesa & Götz, 2008). 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, concerning cellular components, 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, concerning the biological process, 998 GO terms were detected, which 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, 16448 (30.1%), 14386 (26.3%), 12061 (22.1%), and 11717 (21.5%) transcripts, respectively (Figure 3 and Supplementary Table 1). The various GO functions detected, have been found to associated with many stress responsive genes, such as the late embryogenesis abundant (LEA) (Magwanga et al., 2018c), the multidrug and toxic compound extrusion (MATE) gene family (Lu et al., 2019), cyclin dependent kinase (CDK) genes (Magwanga et al., 2018a), 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 a 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 with various GO terms, the 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 sugars and nucleotide sugar metabolism, Lysosome, Alanine, aspartate, and glutamate Fatty acid metabolism, metabolism, Pyruvate metabolism), Galactose metabolism), Cysteine and methionine metabolism (Figure 4A-B, Table 2). 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 (Cruz de Carvalho, 2008). Excess accumulation of ROS triggers oxidative stress, which eventually leads to plant death (Saed-Moucheshi et al., 2014). The ubiquitin mediated proteolysis, aids in the elimination of various reactive oxygen species, thus reducing the lethal effects, and maintains the plant's survival under stress (Tör et al., 2003).

Discussion

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

Being a member of the tetraploid genus, *G. darwinii* has many excellent economical characters in the aspect of fiber quality and resistance to unfavorable environments, 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 a 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 cottonseed EST sequence got from *G. darwinii*, not only further enrich the EST sequence information in the 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 an organism. The second biggest kind of high abundant differentially expressed genes were 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 the plant. The liquidity of phosphatidylcholine 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 a 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 (Dhillon et al., 2016).

When *G. darwinii* grows under drought stress, the number of downregulated genes is more than upregulated. There are 34 (23.4%) among all of 145 downregulated, genes are involved in the 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 (Klimyuk et al., 2007). 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 reactions to resist drought damage. There are 66 sequences of the 207 high abundant differential expression genes, did not match with the 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.

Conclusion

Transcriptome analysis of *G. darwinii* seedlings under drought stress was found that the pathway and gene of the 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 are 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 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.

Declarations

Ethics approval and consent to participate

Approval was granted by the institute of cotton research to carryout the research and use the available facilities within the institute for perform the research work.

Consent for publication

No consent was sought, it is a key mandate for researchers to desiminate their fundings through publication

Availability of data and materials

All data generated in this research are made available including supplementary files

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict, thus the authors declare that they have no competing interests.

Funding

This program was financially sponsored by The National Key Research and Development Program of China (2016YFD0100203, 2016YFD0101409), State Key Laboratory of Cotton Biology Open Fund, Hunan Natural Science Foundation Youth Fund and Hunan Natural Science Foundation of Changde Mutual Funds (2016JJ5014).

Authors' contributions

FL, CX & HC designed the experiments. CX, HC, MKR & HL conceived the experiments and analyzed the results. CX, HC, HL & MKR carried out all computational analyses. MKI, XC, YL, FL participated in part of experiments directly or indirectly like contributed reagents/materials/analysis tools, etc. HC, HL, MKR, MKI, YL& FL drafted the manuscript, and XW, XC &FL proofread and revised the manuscript.

Acknowledgements

We wish to thank Sangon Biotech (Shanghai) Co., Ltd. for help in sequencing, analysis support, and assembly.

References

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.

Ashraf M. 2010. Inducing drought tolerance in plants : Recent advances. Biotechnol. Adv. **28**: 169–183.

Conesa A. & Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics. **2008**: .

Cruz de Carvalho M.H. 2008. Drought stress and reactive oxygen species. Plant Signal. Behav. **3**: 156–165.

Dhillon A., Sharma K., Rajulapati V. & Goyal A. 2016. Proteolytic Enzymes. pp. 149–173. Current Developments in Biotechnology and Bioengineering: Production, Isolation and Purification of Industrial Products,

Guo W., Cai C., Wang C., Zhao L., Wang L. & Zhang T. 2008. A preliminary analysis of genome structure and composition in Gossypium hirsutum. BMC Genomics. **9**: .

Iseli C., Jongeneel C. V & Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proceedings. Int. Conf. Intell. Syst. Mol. Biol. 138–48.

Kahl G. 2015. Next-next-next generation sequencing (next 3 generation sequencing, third-generation sequencing) . pp. 1–1. The Dictionary of Genomics, Transcriptomics and Proteomics,

Karmirantzou M. & Hamodrakas S.J. 2002. A Web-based classification system of DNA-binding protein families. Protein Eng. Des. Sel. **14**: 465–472.

Klimyuk V.I., Persello-Cartieaux F., Havaux M., Contard-David P., Schuenemann D., Meiherhoff K., Gouet P., Jones J.D.G., Hoffman N.E. & Nussaume L. 2007. A 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.

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.

Lu P., Magwanga R.O., Kirungu J.N., Hu Y., Dong Q., Cai X., Zhou Z., Wang X., Zhang Z., Hou Y., Wang K. & Liu F. 2019. Overexpression of Cotton a DTX/MATE Gene Enhances Drought, Salt, and Cold Stress Tolerance in Transgenic Arabidopsis. Front. Plant Sci. **10**: .

Magwanga R.O., Lu P., Kirungu J.N., Cai X., Zhou Z., Wang X., Diouf L., Xu Y., Hou Y., Hu Y., Dong Q., Wang K. & Liu F. 2018a. 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**: .

Magwanga R.O., Lu P., Kirungu J.N., Diouf L., Dong Q., Hu Y., Cai X., Xu Y., Hou Y., Zhou Z., Wang X., Wang K. & Liu F. 2018b. 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 BC2F2generation. Int. J. Mol. Sci. **19**: .

Magwanga R.O., Lu P., Kirungu J.N., Lu H., Wang X., Cai X., Zhou Z., Zhang Z., Salih H., Wang K. & Liu F. 2018c. Characterization of the late embryogenesis abundant (LEA) proteins family and their role in drought stress tolerance in upland cotton. BMC Genet. **19**: .

Molina-Bravo R. & Zamora-Meléndez A. 2016. QTLs for genetic improvement under global climate changes. pp. 471–513. Advances in Plant Breeding Strategies: Agronomic, Abiotic and Biotic Stress Traits,

Mortazavi A., Williams B.A., McCue K., Schaeffer L. & Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. **5**: 621–8.

Saed-Moucheshi A., Pakniyat H., Pirasteh-Anosheh H. & Azooz M.M. 2014. Role of ROS as Signaling Molecules in Plants. pp. 585–620. Oxidative Damage to Plants: Antioxidant Networks and Signaling,

Tiirikka T., Siermala M. & Vihinen M. 2014. Clustering of gene ontology terms in genomes. Gene. **550**: 155–164.

Tör M., Yemm A. & Holub E. 2003. The role of proteolysis in R gene mediated defence in plants. Mol. Plant Pathol. **4**: 287–296.

Volk G.M., Henk A.D., Richards C.M., Miller D.D., Forsline P.L. & Reilley A. 2019. Novel Diversity Identified in a Wild Apple Population from the Kyrgyz Republic. HortScience. **44**: 516–518.

Volk G.M., Richards C.M., Henk A.D., Reilley A., Miller D.D. & Forsline P.L. 2009. Novel diversity identified in a wild apple population from the kyrgyz republic. HortScience. **44**: 516–518.

Wendel J.F. & Cronn R.C. 2003. Polyploidy and the Evolutionary History of Cotton Polyploidy and the Evolutionary History of Cotton. .

Yuan Y., Wang X., Wang L., Xing H., Wang Q., Saeed M., Tao J., Feng W., Zhang G., Song X.-L. & Sun X.-Z. 2018. Genome-Wide Association Study Identifies Candidate Genes Related to Seed Oil Composition and Protein Content in Gossypium hirsutum L. Front. Plant Sci. **9**: .

Zhao C., Piao S., Wang X., Huang Y., Ciais P., Elliott J., Huang M., Janssens I.A., 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**: .

Tables

Table 1. Percentage identity similarities as per various databases

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

Table 2. KEGG Pathway annotation for transcript sequence

No.	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 the endoplasmic reticulum	198	1.51%	ko04141

Figures

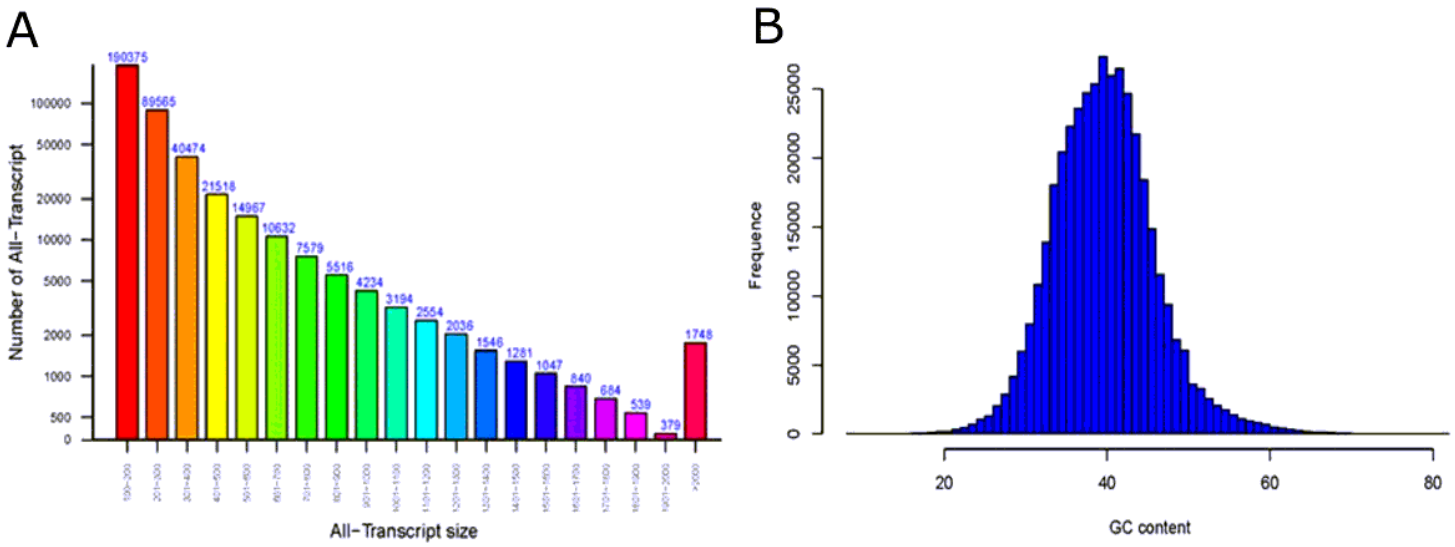

Figure 1

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

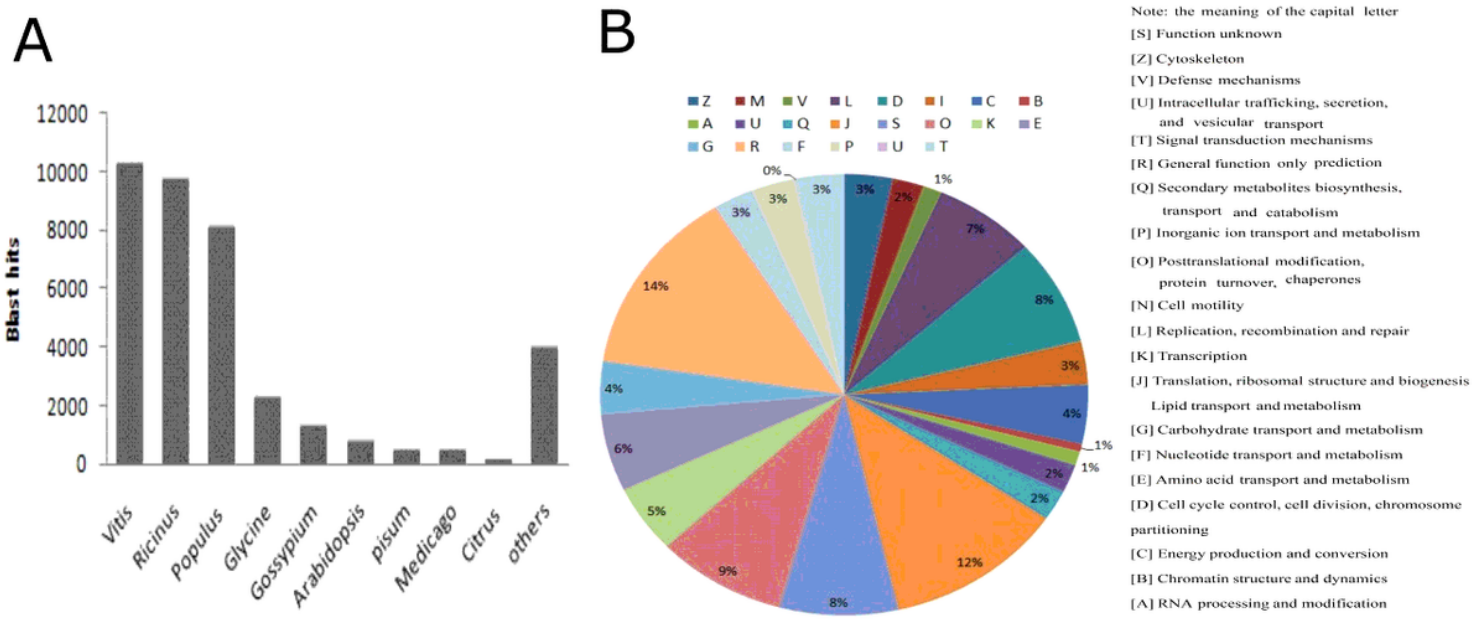

Figure 2

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

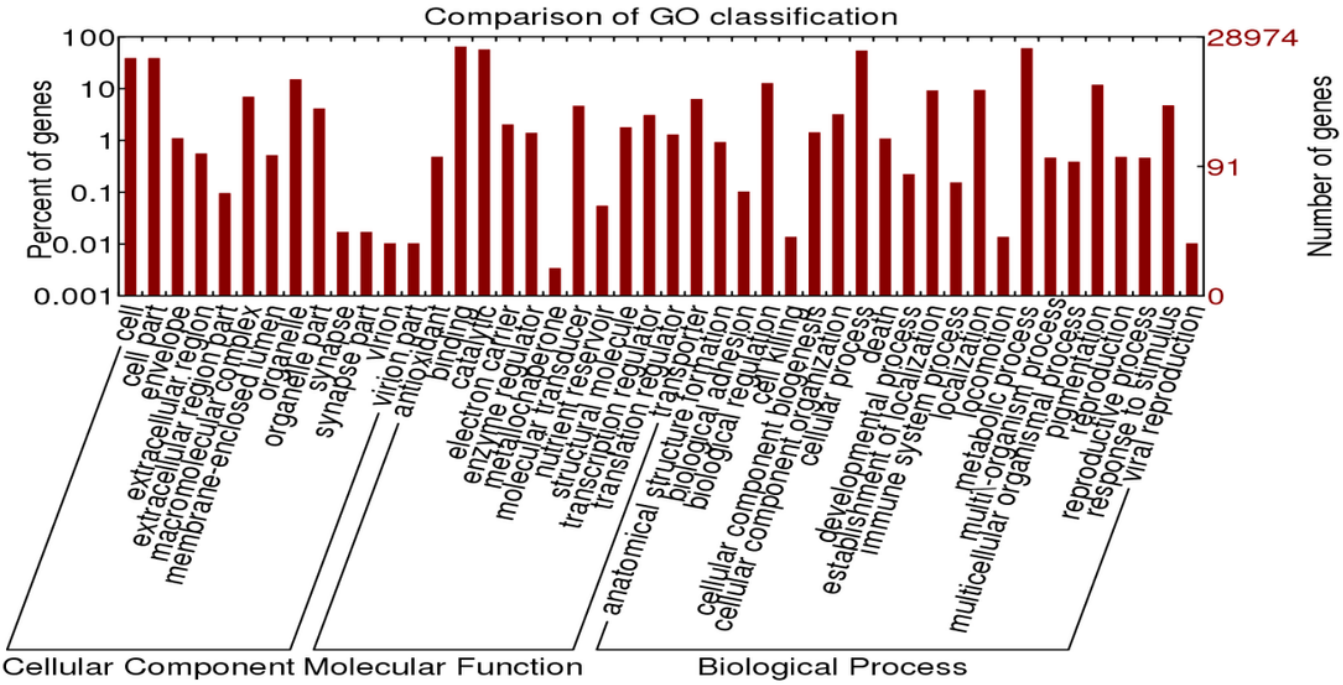

Figure 3

The GO classification of Gene

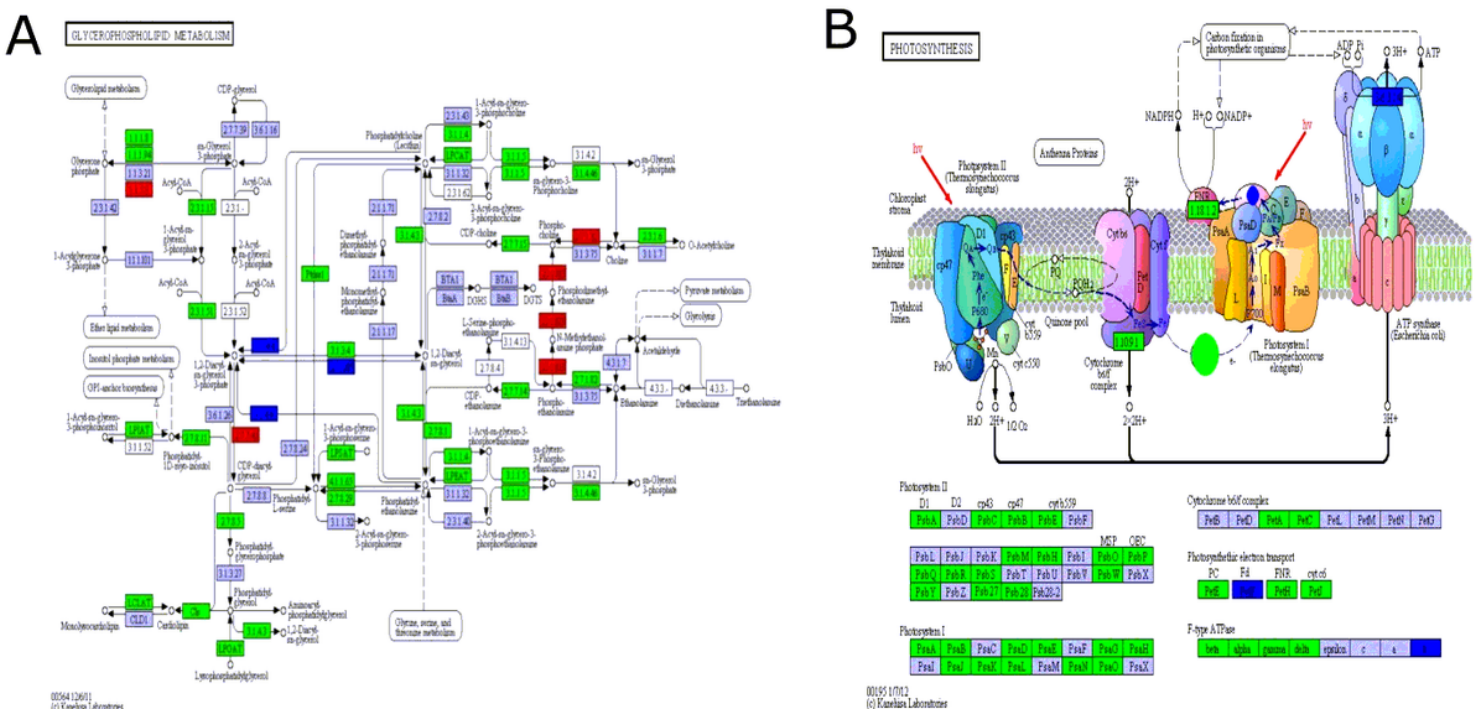

Figure 4

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1GOfunctionalclassfication.xlsx