Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative transcriptome analysis reveals evolutionary divergence and shared network of cold and salt stress response in diploid D-genome cotton

Yanchao Xu
Chinese Academy of Agricultural Sciences Cotton Research Institute
Richard Magwanga
Researchers
Dingsha Jin
Chinese Academy of Agricultural Sciences Cotton Research Institute
Xiaoyan Cai
Chinese Academy of Agricultural Sciences Cotton Research Institute
Yuqing Hou
Chinese Academy of Agricultural Sciences Cotton Research Institute
Zhongli Zhou
Chinese Academy of Agricultural Sciences Cotton Research Institute
Kunbo Wang
Chinese Academy of Agricultural Sciences Cotton Research Institute
Fang Liu (liufcri@163.com)
Chinese Academy of Agricultural Sciences Cotton Research Institute https://orcid.org/0000-0002-1900-5798

Research article

Keywords: diploid D-genome cotton, co-expression, Comparative transcriptome, evolutionary divergence, shared network

Posted Date: November 4th, 2019

DOI: https://doi.org/10.21203/rs.2.16759/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on November 12th, 2020. See the published version at https://doi.org/10.1186/s12870-020-02726-4.

Abstract

Background Wild species of cotton are excellent resistance to abiotic stress. Diploid D-genome cotton shown abundant phenotypic diversity and was the putative donor species of allotetraploid cotton which produce largest textile natural fiber.

Results A total of 41,053 genes were expressed in all samples by mapping RNA-seq Illumina reads of G. thurberi (D 1), G. klotzschianum (D 3-k), G. raimondii (D 5) and G. trilobum (D 8) to reference genome. The number of differently expressed genes (DEGs) were significantly higher under cold stress than salt stress. However, 34.1% DEGs under salt stress were overlapped with cold stress in four species. Notably, a potential shared network (cold and salt response, including 16 genes) was digged out by gene co-expression analysis. Totally, 47,180- 55,548 unique genes were identified in four diploid species by De novo assembly. Furtherly, 163, 344, 330 and 161 positive selected genes (PSGs) were detected in thurberi , G. klotzschianum , G. raimondii and G. trilobum by evolutionary analysis, respectively, and 9.5% - 17% PSGs of four species were DEGs in corresponding species under cold or salt stress. What's more, most of PSGs were enriched GO term related to response to stimulation. G. klotzschianum shown best tolerance under both cold and salt stress. Interestingly, we found that a RALF-like protein coding gene not only is PSGs of G. klotzschianum, but also belongs to the potential shared network.

Conclusion Our study provided a new evidence that gene expression variations of evolution by natural selection were essential drivers of the morphological variations related to environmental adaption during evolution. Additionally, there exist shared regulated networks under cold and salt stress, such as Ca 2+ signal transduction and oxidation reduction processes. Our work establishes a transcriptomic selection mechanism for altering gene expression of four diploid D-genome cotton, and provides available gene resource underlying multi-abiotic resistant cotton breeding strategy.

Background

Cotton (*Gossypium)* provides the most natural fiber for the manufacture of textiles [1, 2], and is an economically important crop around the world. At present, *G. hirsutum* and *G. barbadense (*allotetraploid cotton) are widely planted, which were domesticated through a long-time history. Genetic evidence suggests that allotetraploid cotton was formed by diploid A- and D-genome species hybridization event at about 1–2 million years ago [3–5]. Diploid D-genome cotton contains at least 13 species. Among those species, *G. thurberi, G. klotzschianum, G. raimondii* and *G. trilobum,* which distribute four different latitude areas of the Americas [6, 7], were observed distinct difference of morphological character. Molecular evolutionary processes and phylogeny of those four species were revealed through phylogenomic methods [7]. Such as, *G. thurberi* and *G. trilobum* show the close relationship of phylogeny, even though there is distinctly different latitude of natural distribution.

Currently, scientist generally acknowledges that *Gossypium arboreum* and *Gossypium raimondii* are putative A and D genome donor species of allotetraploid cotton, respectively [1, 8, 9]. A-genome diploid

cotton contain two species, *G. arboreum* and *G. herbaceum,* and distributed Southern Africa and Asia. D-genome diploid cotton contain about 13 species, and primarily distributed Mexico, with range extensions into Peru, the Galapagos Islands, and southern Arizona [10]. The main reason of divergent morphological and genomic characteristics of A- and D-genome cotton is geography insulation and division. About 1–2 million years ago, A-genome diploid are hybridization with D-genome diploid cotton, and allotetraploid cotton appeared through subsequent polyploidization events [11]. This requires that A and D genome must have established physical proximity [12], but it is inconceivable to contact across the Pacific. Therefore, the origin and evolution of allotetraploid cotton is also a mystery, although there are many hypotheses or theories.

Cold and salt stresses are important environmental factors that greatly limit cotton production in the world [13, 14]. Plant adaptation to environmental stresses is dependent upon the activation of cascades of molecular networks involved in stress perception, signal transduction, and the expression of specific stress-related genes and metabolites [15]. Stress-inducing factors can occur simultaneously or sequentially and cause osmotic stress, water deficits, ionic imbalances, peroxidation damage, ultimately, growth inhibition [16]. Calcium plays a major role in abiotic stress response as the second messenger [17, 18]. Cold and salt could enhance cytosolic free calcium concentration in plants [19]. Research has showed that OSCA1, reduced hyperosmolality-induced calcium increase 1, is a putative sensor for osmotic stress, involving in cold and salt stress response [20]. MAPK (Mitogen-activated protein kinase) cascades, stimulated by the second messenger, for example calcium, participate in abiotic stress signal transduction [21]. Moreover, SnRK2 (Sucrose non-fermenting–1-related protein kinase 2) family of protein kinases is also involved in signal transduction under salt, drought, and osmotic stress treatments. MAPK and SnRK2 could be rapidly activated by cold and salt in plants [19, 22, 23].

Increasing number of genome sequencing and resequencing, mRNA sequencing and phenotypical assesses of cotton [1, 8, 24–31] provides important resources for studying potential biological mechanism in cotton. Comparative transcriptome analysis usually used to construct regulated model by gene expression changes [32–35]. Phylotranscriptomic analysis provide a new strategy to investigate the gene evolution and expression change during domestication [36–38]. For example, hundreds of candidate genes that have evolved new protein sequences or have changed expression levels in response to natural selection were identified in wild tomato relatives by the phylotranscriptomic analysis, indicating artificial and natural selection have had on the transcriptomes of tomato and its wild relatives and expression change play an important role in the evolution and domestication [37]. The weighted correlation network analysis (WGCNA) is an R package for gene co-expression network (GCN) analysis and can be used as a data exploratory tool or a gene screening (ranking) method to find clusters (modules) of highly correlated genes [39–41]. It was used widely to find hub genes in biomedical science [42–45].

In our research, to reveal the genetic and expression diversity under cold and salt stress, we perform transcriptomic sequencing of four diploid D-genome species, including *G. thurberi, G. klotzschianum, G. raimondii* and *G. trilobum.* The phylogenetic relationship of four species were in line with previous studies

[7]. Six species-specific profiles and four species-specific modules were identified by comparative transcriptomics. Gene expression analysis found more genes was differently expressed under cold stress in contrast to salt stress. The gene evolutionary analysis identified hundreds of PSGs in different genomes or subgenomes (wild species: *G. thurberi, G. klotzschianum, G. raimondii* and *G. trilobum*,cultivated species: *G. arboreum, G. hirsutum* A_t and D_t, *G. barbadense* A_t and D_t). We also found a module was negatively correlated with salt and cold stress by WGCNA. *G. klotzschianum* shown better resistant under cold and salt stress, and 171 common DEGs under cold and salt stress were identified in this species. In summary, gene expression variations were essential drivers of the morphological variations related to environmental adaptation during evolution and there are shared networks that involved in cold and salt stress response, such as signal transduction and oxidation reduction processes. Our work provides an insightful understanding of expression divergence, conservation and response to environmental adaptation during evolution by combining protein-coding sequence and gene expression diversity.

Result

Phenotyping diversity and RNA-seq Data of diploid D genome species

To observe an evolutionary divergence and conservation at the transcriptomic level of wild diploid D-genome cotton, four cotton species including *G. thurberi* (D_1), *G. klotzschianum* ($D_{3\text{-}k}$), *G. raimondii* (D_5) and *G. trilobum* (D_8) with a rich diversity of morphological characteristics were selected for further analysis. These species show tremendous phenotypic variations in flower color and leaf shape, although these are phylogenetically most closely related to each other (Figure 1A). Additionally, four species presented variant cold and salt resistance: *G. klotzschianum* and *G. thurberi* shown excellent resistance to cold and salt stress, followed by *G. raimondii* (Figure 1B). Strangely, *G. trilobum* indicated the lowest resistance, in spite of it is most closely related to *G. thurberi*. For comprehensive evaluation, we carried out RNA-sequencing of D-genome diploid cotton. Four species, *G. thurberi*, *G. klotzschianum, G. raimondii* and *G. trilobum*, were abbreviated as GD1, GD3, GD5 and GD8 for convenience. Ten leaf samples were collected from each species in different time intervals (0h, 6h and 12h after three-leaf stage of seedlings: C0, C6 and C12) after two stress treatment samples (Cold T12 and salt stress S12), two repeats for each sample. A total of 40 data sets with 273.01Gb raw data were obtained. The resulting GC content rates of 44.41%-46.64% are similar in different dataset. 273.01 Gb clean data without an adapter, ploy-N and lower quality reads were obtained after quality control. Every dataset is at least 17.1 million clean reads and more than 89.03% of base Q30. After that, RNA-seq clean reads were mapped to reference genome of *G. raimondii* and ranged from 81.37% to 91.37% clean reads were uniquely mapped to reference genome (Table S1). Based on mapping result more than, 7,374 SNPs (Single Nucleotide Polymorphisms) were identified in each library (Table S2). expression level of 41,053 genes, including 3,548 new transcriptional genes, are quantitated. Except samples of GD1S12, the R1 and R2 libraries in the same sample are with high value of correlation ($R^2 > 0.7$), suggested datasets of other samples are reliable (Figure 1C). The

reason of the lower value of correlation between GD1S12R1 and GD1S12R2 may be effects of developmental and environmental variation on gene expression, although we try to minimize it.

Genetic divergence of four diploid D genome species

The transcript sequence polymorphism analysis contributed to understanding the evolutionary diversity and conservation at the transcriptomic level in four D-genome diploid cotton. The RNA-Seq data were used for sequence polymorphism discovery and 7,374-404,737 SNPs were identified in different libraries (Table S2). More than 59% SNPs have transition type. On account of the datasets was obtained by transcriptomic sequencing, the chromosomal distribution of more than SNPs in introgenic region. Of note is the observation that about 9% SNPs were identified in the intergenic regions of GD1, GD3 and GD8, representing new genic region in other D-genome species. Libraries of GD5 (*G. raimondii*) have less number of SNPs as compared to other three species. In general, transcriptomes of GD5 samples should have the same genome with reference genome. Although transcriptome of GD5 samples and reference genome was very similar to each other (<1 SNP per Kb of most genes), some SNPs were observed between GD5 and reference genome, suggesting that there existed in differences in the leave's transcriptome among plants of the same donor material. A total 3,2651 genes were expected to have high functional effects by SNPs in four diploid D-genome cottons. Go ontology (GO) enrichment analysis revealed that a large number of genes were enriched in oxidation-reduction process, protein modification process, stress response and protein phosphorylation, it is suggested that abiotic stresses have played an important role in driving transcriptional diversity among four species (Table S3). On the contrary, it is also noticeable that there are 8,402 genes have no SNPs function-effect, including many housekeeping genes such as (Tubulin, ribosomal protein and glycolytic enzyme-coding genes). A total of 8,402 genes were involved in several biological pathways including photosynthetic electron transport chain. We used Neighbor-joining methods to construct a phylogeny tree of four species base on transcriptomic data (Figure 2A). And, the principal component cluster was performed to observe the relationship among four species (Figure 2B). A modest number of SNPs separate them among four species (< 5 SNPs per kb of most of genes). The datasets of same species almost overlap with each other. Consistent with previous study, *G. thurberi* showed closer relationships with *G. trilobum*.

Transcriptome De novo assembly

Transcriptome De novo assemblies were performed based on RNA-seq data using Trinity software. Ranged from 47,180 to 55,548 unique genes were identified in four diploid species. Most of unigenes (> *40%*) length was ranged from 200 - 500 bp. To initiate our evolutionary analysis, we identified the strictly orthologous unigenes among 7 species, including three cultivated (A_2-genome: *G. arboreum*; AD_1-genome: *G. hirsutum*; AD_2-genome: *G. barbadense*) and four wild species. For rigorous analysis, the genomes of allotetraploid species were separate A and D subgenome to identify orthologous transcripts. A total of 47,119 orthologous gene pairs were characterized, and 5,312 single-copy transcript pairs of those were used to construct phylogenic tree of the nine genomes by maximum likelihood methods. Phylogenetic analysis using these unigenes revealed a quite similar topology (Figure 2C) of the tree

based on transcriptomic SNP data (Figure 2A), indicating a solid phylogeny for four D-genome diploid species. Same with previous study, A subgenome of *G. hirsutum* and *G. barbadense* originate from *G. arboreum*, which formed a monophyly in our phylogenetic analysis. *G. hirsutum*-D subgenome, *G. barbadense*-D subgenome and GD5 formed a monophyly. These results proved D subgenome of allotetraploid AD_1 and AD_2 genomes have same donors, originated from *G. raimondii*.

Estimate evolutionary rates and identify positively selected genes (PSG)

Non-synonymous (dN) and synonymous (dS) are used to estimate evolutionary rate and positively selected genes (PSGs) in each species. If dN/dS value is *>1*, this is indicative of positive selection, whereas equal to or significantly greater than 1 represent either purifying or neutral selection. Therefore, we evaluated the dN, dS and dN/dS of the identified orthologous single-copy genes among nine genomes. Based on the above-constructed phylogenetic tree, the dN/dS for each orthologous unigene pair was evaluated in the different branches using a free ratio model (model=1), which allows for a separate dN/dS ratio for each branch. We found that wild diploid cotton (GD1, GD3, GD5, GD5_ref, and GD8) had higher dN/dS ratios than the branch of cultivated cotton (GAD1, GAD2, and GA2). Under natural selected pressure, wild species are shown a fast evolutionary ratio based on adaptive choices (Figure 2D). A total of 163, 344, 330, and 161 PSGs in GD1, GD3, GD5 and GD8 were identified by PAML software, respectively (Table S4). PSGs of diploid cottons were enriched in GO terms related to protein modification, protein ubiquitination, RNA processing and ncRNA processing, involved in environmental adaptation (Table S5). What is more, we also found 180, 77, 51, 103 and 70 PSGs in GA2, GAD1_A, GAD1_D, GAD2_A and GAD2_D, and these PSGs were enriched mRNA metabolic process (Table S6).

Gene Expression Divergence and conservation

We detected expression levels of a total of 41,053 transcripts in at least a single sample. Global expression level distributions of 40 data sets were similar to each other. Twenty-four datasets of leaves without stress treatment were used to detect gene expression divergence and conservation of four species. Samples of GD5C0R1 and GD5C0R2 were as a control group to identify different expression transcripts. Patterns of differential gene expression were characterized of four species (Figure 3A). A total of 29,512 DEGs was clustered eight profiles based on gene expression pattern use k-means clustering method. Profiles with a large number of DEGs, display a similar expression pattern (Profile 1, 2 and 3) of four species. The profiles 1 and 2 shown considerable difference in expression over time in leaves of four species. In the Profile 1, the DEGs were down-regulated expression level over time, and enriched in carbon metabolism and carbon fixation in photosynthetic organism. On the contrary, DEGs of the profile 2 were up-regulated and enriched in plant circadian rhythm and ribosome biogenesis in eukaryotes. In the profile 3, DEGs were significantly enriched biosynthesis of unsaturated fatty acids, DNA replication and photosynthesis-antenna proteins process (Figure 3B). Those pathways are indispensable during leaves growth and development, while transcripts expression of those pathways is conservation during divergence of D-genome species of cotton. Profile 4-9 were observed a considerable species-specific expression pattern. In the profile 4 and 5, 3662 and 1048 DEGs displayed GD5-specific expression pattern.

What's more, in the profile 6, 1516 DEGs displayed GD1-specific expression pattern. And, in profile 7 and 8, shown GD3-GD5 and GD1-GD8 special expression pattern. In view of GD1 and GD8 with close relationship, it made perfect sense that GD1 and GD8 contain more conservative biological pathways.

Evolutionary conservation and divergence of the Gene Co-expression Networks

Notwithstanding the analysis of gene expression pattern provided insights of gene expression divergence and conservation, co-expression gene networks, constructed by highly connected genes, is more relevant to important biological processes of growth and development and complex regulation of abiotic stress response. To grasp gene co-expression network conservation and divergence, all of 40 datasets of leaves were used to construct co-expression network by weighted gene co-expression network analysis (WGCNA). In order to improve the accuracy of WGCNA, we removed genes with *FPKM< 1*. In conjunction with connection strengths (soft-threshold power: 5, $R^2 > 0.90$) among 12110 genes (Figure S1), a global view of co-expression network topology among four species was constructed. Genes in the same module shown a higher topological overlap (Figure S2). Finally, a total of 33 modules were used for investigating evolutionary conservation and divergence of gene co-expression networks, which are defined as clusters of highly interconnected genes (Figure 4A). In these modules, the genes number ranged from 31 (darkorange) to 3128 (turquoise) with high correlation coefficients with one another in corresponding modules. Every gene connectivity was evaluated on each module. Highly connected genes (*kME> 0.95*) were identified as hub genes. Ultimately, a total of 425 hub genes (ranging from 1 to 293 within the modules) were detected.

In each module, expression levels of all genes were displayed by a heatmap and were summarized by the eigengene values (the first principle component of module expression profiles). Seven sample conditions were defined for identified significant modules (Figure 4B). Share.T.S group was used for identifying conservative shared networks of cold and salt stress response. Additionally, the module, correlated with cold and salt stress response, was identified by Cold and Salt Groups, respectively. GD1, GD3, GD5 and GD8 Groups were used for identifying genome-specific modules. By association analysis between eigengenes and sample conditions (7 groups: Share.T.S, Cold, Salt, and 4 genomes) via Pearson correlation coefficient analysis, 29 modules were identified with significant genome-specific and/or abiotic stress regulated co-expression patterns (ANOVA, *P<0.05*). Four major modules of highly co-expression genes were most strongly correlated (Pearson's correlation *r>0.9*) with four genomes, respectively. The largest module (turquoise), containing 3128 highly connected genes in GD5, enriched in response to stimulus and immune system process (Table S7). A second module (Blue, 1,193 genes), shown GD3 specific. Other two modules (yellow with 420 genes and purple with 442 genes), which included few highly connected genes, respectively related to GD1 and GD8. Interestingly, all four major modules displayed similar result of GO enrichment analysis. Nearly half genes enriched for GO terms related to response to stimulus and immune system process in each module, revealing that abiotic and biotic stresses have played a major role driving transcriptional variation among these four species. We noticed that characteristics of the seven modules were unique related to corresponding group, and among these modules, five modules were genome-specific. What's more, most of the modules most

strongly correlated with four genome group. These results suggested transcriptional variation is mostly correlated with genome divergence. We observed some modules overlapped among different groups. Overlap modules are helpful to understand the similar biological characteristics among two or more genomes. For example, the darkgreen module was significantly shown a positive correlation with GD1 and GD3, but a negative correlation with GD5 and GD8, and enriched sesquiterpenoid and triterpenoid biosynthesis and flavonoid biosynthesis which were known related with abiotic and biotic stress resistance.

Of particular concern is that co-expression networks related to cold and salt stress, due to four diploid species presented different cold and salt stress resistance. Association analyses between co-expression modules and abiotic stress (Cold and salt stress) revealed that 17 modules correlated with abiotic stress (Shar.T.S: 7 modules; Cold: 8 modules; Salt: 8 modules), thus representing suites of interconnected genes underlying the biological process of abiotic stress response (Table S7). Noticeably, among four species, GD3 and Salt/Cold group had more overlapped modules. Considering GD3 shown better cold and salt tolerance than other three species, suggested GD3 evolved more complete mechanisms in abiotic stress adaptation. On the contrary, almost no one module shown overlap between GD8 and abiotic stress groups, and it is foreseeable that GD8 manifest as significant sensitivity under cold and salt.

Interestingly, we found skyblue3 module displayed significant negative correlation with Share.T.S, Cold and Salt group. One hub gene, RALF (rapid alkalinization factor)-like, was identified in this module. The homologs of this gene in Arabidopsis may regulate plant stress, growth, and development. Fifteen genes in this module that are interconnected with RALF-like protein coding gene. Except five uncharacterized genes, ten out of fifteen interconnected genes are related to GO term of response to stimulus, including *Gorai.005G234900*, *Gorai.007G094200*, *Gorai.N023400*, *Gorai.011G238800*. Homologs of them in Arabidopsis are annotated four important transcription factors: MYB44, TCP9, TCP12 and GATA8, and involved in abiotic stress response. We also observed two xyloglucan endotransglucosylase/hydrolase protein 22 (XTH22)-encoding genes, *Gorai.003G052400* and *Gorai.009G006400*, which involved in carbohydrate transport and metabolism. Homologs of *Gorai.005G094300* in Arabidopsis, EXORDIUM protein-encoding gene, required for cell expansion in leaves, and may be involved in signaling processes that coordinate brassinosteroid (BR) responses to environmental or developmental signals (Figure 5). Most of fifteen genes were decreased the expression level after cold and salt stress treatment.

Characterization of DEGs under cold and salt stress

The false discovery rate (FDR) ≤ 0.001 and $\log_2$ rates ≥ 2 (8 groups: treatment/control) were used to identify DEGs of in four species. The number of DEGs varied from 459 to 3372 among treatments of 8 groups (Figure S3). Totally, 7515 DEGs were found, occupying 20.04% of total detected genes. Interestingly, the number of DEGs among salt stress groups was always smaller than those among cold stress group in four species examined (Figure 6A), indicated that the cold stress response was divergent than salt stress response, and more unique DEGs (6405, 85.2% of all DEGs) in one type of abiotic stress again proved that. But, 1,109 genes (16.1% of cold DEGs; 63.3% of Salt DEGs) were found in two different

abiotic stresses, indicating potential share regulated pathways in cold and salt stress response. As expected, 1109 genes were enriched in the oxidation reduction pathway, which were involved in multiple abiotic stress response.

We thus focused on DEGs in GD3, on account of the it shown higher tolerance under cold and salt stress, and ultimately identified 2759 DEGs in salt and cold groups of GD3, including 2,300 genes cold DEGs, 630 salt DEGs. Among those genes, we found171 share DEGs under salt and cold stress, containing 80 down-regulated and 75 up-regulated DEGs (Figure 6C). Interestingly, 102 genes of share DEGs involved response to stimulate, and 32 genes of those genes were related to response to cold or salt stress, including 13 transcriptional factors encoding genes (NAC, ERF, MYB, G2, HD-ZIP) are putatively related to response to abiotic stress, since some homologs of those genes in Arabidopsis related to abiotic stress response. For example, *Gorai.002G073700* encodes the homolog of Arabidopsis NAC72, which bind to a drought-responsive *cis*-element in the early responsive to dehydration stress 1 promoter [46]. Also, both *Gorai.001G239000* and *Gorai.006G017400* encode homologs of MYB-like protein in Arabidopsis that involve in plant defense response [47]. Notable a total of thirty cold DEGs and three salt DEGs were PSGs. Moreover, three cold DEGs and 12 salt DEGs were in skyblue3 module. Although fewer DEGs (Cold DEGs: 6.1%; Salt DEGs: 1.7%) were identified above all salt DEGs, some DEGs were overlapping with the PSGs or skyblue3 module (Figure 6B). Those results indicate again adaptive evolution drives transcriptional diversity, and a share regulated network involved in cold and salt stress response. We also found *Gorai.006G147500* of skyblue3 module genes is positively selected, and is interconnected with the *hub* gene, RALF-like protein coding gene. It further confirmed the potential regulated network that was identified in previous analysis.

Discussion

In previous study, some researches focus on the relationship of D-genome diploid cotton, for instance the relationship of *G. trilobum* are close with *G. thurberi,* rather far with *G. klotzschianum* and *G. raimondii [7*]. Range from 7,374 to 404,737 million SNPs were detected by aligning our transcriptomic data with the reference genome (*G. raimondii,* JGI). And then, those SNPs were used to construct phylogenetic tree to investigate the relationship of four species. The relationship of the four species is in line with previous studies, and the result further confirmed feasible genomic sequences analysis based on RNA-Seq *de novo* assembly. Allotetraploid cotton contains seven species. Based on the similar distribution of 45S and 5S rDNA between allotetraploid and diploid species in chromosomes, some researcher thought different allotetraploid species of cotton genus have different donor species, such as *G. thurberi, G. klotzschianum, G. raimondii* and *G. trilobum [*48]. But, *G. raimondii* shown most synteny blocks with D-subgenome of *G. hirsutum* and *G. barbadense* by comparative genomic analysis, suggested *G. raimondii* is the possible donor species of *G. hirsutum* and *G. barbadense.* So, some researchers thought allotetraploid cotton may be monophyletic. In our study, 5,312 single copy genes were used to construct phylogenic tree and estimate evolutionary rate. A-genome and A-subgenomes formed a monophyly, and D-genomes and D-subgenomes formed a monophyly. Our result support that *G. raimondii* is donor species ofD

subgenomeof *G. hirsutum* and *G. barbadense.* It is consistent with monophyletic evolutionary theory of allotetraploid cotton [11].

Positive selection plays an important role in plant evolution and adaptation to biotic and abiotic stresses, as gene expression and regulation changes by positive selection have been postulated to be key determinants of the rates of adaptive evolution [49–53]. Our result manifested wild cotton possessed higher dN/dS ratio (genome-wide accelerated evolution) rather than cultivated cotton, suggesting that wild cotton may have undergone adaptive evolution that allows them to cope with their extremely wide range of terrible conditions and environments. Our study found a hundred genes were positive selected during evolutionary, and those gene also were enriched GO terms related to abiotic or biotic stress response. And analyses confirmed that positive selection drive environmental adaptive evolution of wild species. K-means cluster analysis of for wild species found that species-special profiles were enrich GO terms related to abiotic stress response, indicated Expression analysis of four wild species Expression analysis of four wild species also proved that abiotic stresses drive transcriptional diversity among four species [54].

We observed morphological difference between four species. The different ecosystem of four species leads to the divergence of their morphology. Previous genetic evidences indicated gene expression alteration is essential to drive phenotypic diversity during evolution [37, 55]. Our results agreed this conclusion, five species-specific gene expression clusters were identified (profile 4–8), and most of genes are enrichment to the GO terms which related to environmental adaptation in those five profiles. Similarly, most PSGs of wild species were related to environmental adaption. More than 9.6% PSGs were overlapped with genes of corresponding DEG sets. Therefore, we speculate evolutionary selection also could drive gene expression alteration to adapt environment.

Morphological analysis found that *G. klotzschianum*is tolerant to cold and salt stress.*G. klotzschianum,*natural range is the Galapagos Islands, has the ability to adjust itself to the environment. Therefore, we investigated genes in *G. klotzschianum* under positive selection. The number of PSGs are basically low in *G. klotzschianum* (344, ~0.9% of all genes in genome). Nonetheless, 33 PSGs were found to be differentially expressed under salt and cold stress, indicating that gene expression alteration caused by natural selection might have played an essential role improve the environmental adaptation.

Combining the analysis of the preceding context that *G. klotzschianum* shown astonishing tolerance under cold and salt stress, we speculate a shared network that involved in cold and salt stress response were formed during adaptive evolution of *G. klotzschianum.* Thirty-three (~10.6%) PSGs were differently expressed under cold or salt stress. In particular, in WGCNA, we found a module that was negative correlated with salt, cold and share.T.S conditions. Totally, 74% genes in this module were enriched GO terms related to response to stimuli and 36.3% genes were enrichment GO term related to signaling process. This result further confirmed our speculation. Simultaneously, we found stress signaling process is a tentative share regulated network of salt and cold stress response. One hub gene in skyblue3 was found in WGCNA. And, fifteen genes which were high connective with the hub gene were identified.

Interesting, *Gorai.006G147500* among 15 high connective genes, were found. These 16 genes of skyblue3 module were potential regulated network under cold and salt stress. The researcher presume that plant cells must be capable of sensing various environmental signals [19], and some putative sensors were identified in previous studies, such like OSCA (reduced hyperosmolality-induced calcium increase 1) [56], G protein [57], and COLD1 [20]. Salt and cold stress could cause increases in the cytosolic free calcium concentration in plants [58, 59]. homolog of *Gorai.010G168400* in Arabidopsis, encoding *AtOSCA3.1* protein, acts as a hyperosmolarity-gated non-selective cation channel that permeates Ca^{2+} ions [60]. And *Gorai.010G168400,* shown a higher expression level under salt and cold stress compared control groups, was a potential sensor which mediate cold and salt stress in cotton, and calcium ion plays an important role in abiotic stress response as the second messenger [17, 18]. *Gorai.009G294400,* encoding CBL-interacting protein kinase 18, which involved in Ca^{2+} signal transduction, was also found that different express under cold and salt stress. Core stress-signaling pathways involve protein kinases related to the yeast SNF1 and mammalian AMPK [19]. SNF1/AMPK-related kinases mediate the signaling of various abiotic stresses. Among 171 DEGs of salt and cold group, *Gorai.005G081200* and *Gorai.002G103300* which share homology to SNF1/AMPK in their kinase domains.

Oxidation reduction process is also involved in salt and cold stress response by keeping the homeostasis of reactive oxygen species [61]. We found that oxidoreductase activity-related genes such as *Gorai.004G093200, Gorai.013G025400, Gorai.013G059500* and *Gorai.013G176900* were up-regulated under cold and salt stress, correlating with the oxidation reduction process. PSGs and skyblue-module genes also generated many important candidate genes, such as *Gorai.004G227900* and *Gorai.002G226600,* which are related to keeping the homeostasis of reactive oxygen species. Extensive investigations of these genes under cold and salt stress would help us understand how DEGs regulate homeostasis of reactive oxygen species and signal transduction and their function during cold and salt response, thus helping us to increase yields in cotton.

Conclusion

Comparative transcriptome analysis of four diploid D-genome cottons reveal sequence and gene expression variations. Gene evolution analysis of wild and cultivated cottons identifies positively selected genes which involve in the domestication and evolution of cotton and estimate the evolutionary rates. In this work, we found that evolutionary selection could drive gene expression alteration to adapt environment and gene expression variation is the primary evolutionary event during the divergence of four D-genome species. The expression pattern analysis found that six profiles shown distinct species-specific characteristic and genes of those profiles were involved response to stimulate. Thus, gene expression variations were essential drivers of the morphological variations related to environmental adaptation during evolution. More DEGs were identified under cold stress in contrast to DEGs under salt stress, indicated cold stress lead to expression change of more genes. *G. klotzschianum* shown better resistant under cold and salt stress. Compared with other three species, more PSGs were detected in *G.*

klotzschianum, and 9.6% PSGs were differently expressed under cold or salt stress. In *G. klotzschianum,* 27.1% DEGs under salt stress were overlap with that under cold stress and we found skyblue3 module were significantly negatively correlated with cold, salt and share.T.S condition groups. Thus, there are share networks that involved in cold and salt stress response, such as signal transduction and oxidation reduction processes. Based on our multiple analyses, a set of candidate genes involved in cold and salt stress response is putatively proposed, providing genetic resources for multi-abiotic resistant cotton breeding.

Methods

Plant growth and sample collection

Four diploid wild species of D-genome cotton were used in this study, including *G. thurberi* (D_1), *G. klotzschianum* (D_{3-k})*, G. raimondii* (D_5) and *G. trilobum* (D_8). There are 13-14 species in diploid D-genome cottons and they are originally distributed from Southwest Mexico to Arizona, with additional disjunct species distributions in Peru and the Galapagos Islands. Here, *G. thurberi*, *G. klotzschianum, G. raimondii* and *G. trilobum* were obtained from USDA-ARS Southern Agricultural Research Center in College Station, Texas, USA and currently perennially preserved in the National Wild Cotton Nursery, which is located in Sanya, Hainan, China and is supervised by Institute of Cotton Research, Chinese Academy of Agricultural Sciences (ICR-CAAS). The seeds of four wild species were obtained from the wild cotton nursery which is managed by the institute of cotton research, Chinese Academy of Agricultural Sciences, China. The seeds first germinated at 28 ℃, 16h light/8 h dark cycle and light intensity of 150μmol $m^{-2}s^{-1}$ in 15% water content sands. Three days after germination, the properly plants were potted in soil and placed in a growth room in the same condition. Seedlings containing two simple leaves and one heart-shaped leaves (time point: 0h) were separated three groups. First group living in the normal condition (28℃, 16h light/8 h dark cycle and light intensity of 150μmol $m^{-2}s^{-1}$ and 15% water content), a second group was watered with 300 mM NaCl solution (28℃, 16h light/8 h dark cycle and light intensity of 150μmol m-2s-1 and 15% water content) and the third group was growth at low temperature (4℃, 16h light/8 h dark cycle and light intensity of 150μmol $m^{-2}s^{-1}$ and 15% water content). The first group was used to reveal the expression divergence and conservation and leaves of seedling were collected at 0h, 6h and 12h. Second and third groups were used to reveal regulation under cold and salt stress, and leaves were collected at 12h. This experiment had two repeats.

RNA extraction, library construction, and RNA-seq

Total RNA was extracted from each cotton sample using TRIzol Reagent (Life technologies, California, USA) according to the instruction manual. RNA integrity and concentration were checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA). mRNAs were isolated by NEBNext Poly (A) mRNA Magnetic Isolation Module (NEB, E7490). The cDNA libraries were constructed by following the manufacturer's instructions of NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, E7530) and NEBNext Multiplex Oligos for Illumina (NEB, E7500). Briefly, the enriched mRNA was fragmented into

RNAs with approximately 200nt, which were used to synthesize the first-strand cDNA and then the second cDNA. The double-stranded cDNAs were performed end-repair/dA-tail and adaptor ligation. The suitable fragments were isolated by Agencourt AMPure XP beads (Beckman Coulter, Inc.), and enriched by PCR amplification. Finally, the constructed cDNA libraries were sequenced on a flow cell using an Illumina HiSeq™ 2500 sequencing platform. Beijing Biomarker Technologies (http://www.biomarker.com.cn) provides experimental procedures and commercially performed it.

Processing of RNA-seq data

Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. In this step, clean data (clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30, GC-content and sequence duplication level of the clean data were calculated. All the downstream analyses were based on clean data with high quality. These clean reads were then mapped to the reference genome sequence. Only reads with a perfect match or one mismatch were further analyzed and annotated based on the reference genome. Hisat2 tools soft were used to map with reference genome. Beijing Biomarker Technologies (http://www.biomarker.com.cn) provides experimental procedures and commercially performed it.

Gene expression analysis

Quantification of gene expression levels were estimated by fragments per kilobase of transcript per million fragments mapped [62]. Differential expression analysis of two conditions/groups was performed using the DESeq2 [63]. DESeq2 provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted P-value *< 0.01* found by DESeq2 were assigned as differentially expressed.

SNP Calling

Picard - tools v1.41 and samtools v0.1.18 were used to sort, remove duplicated reads and merge the bam alignment results of each sample. GATK2 or Samtools software was used to perform SNP calling. Raw vcffiles were filtered with GATK standard filter method and other parameters (clusterWindowSize: 10; *MQ0 >= 4* and *(MQ0/(1.0*DP)) > 0.1*; *QUAL < 10*; *QUAL < 30.0* or *QD < 5.0* or *HRun > 5*), and only SNPs with distance > 5 were retained.

De Novo Transcriptome Assembly

The left files (read1 files) from all libraries/samples were pooled into one big left.fq file, and right files (read2 files) into one big right.fq file. Transcriptome assembly was accomplished based on the left.fq and right.fq using Trinity [64] with min_kmer_cov set to 2 by default and all other parameters set default. Expression analysis, SNP calling and transcriptome assembly was performed using BMKCloud (www.biocloud.net) in this research.

Ortholog Identification, Phylogenetic Analysis, Evolutionary Rate Estimation and Positive selected genes (PSG) Identification

High-quality draft genomes of two allotetraploid cotton (*G. hirsutum* and *G. barbadense*) and two diploid cotton (*G. arboreum* and *G. raimondii*) were obtained from Cottongen database (https://www.cottongen.org). Genome of allotetraploid species contains two subgenome (A- and D-subgenome). In order to ensure accuracy and reliability of ortholog identification, two subgenome of allotetraploid species were separated. Together with our results of transcriptome assembly (*G. thurberi*, *G. klotzschianum* and *G. trilobum*), totally nine genomes were used for phylogenetic analysis and A-genome clades was as out-group (*G. arboreum*, G. barbadense A subgenome and *G. hirsutum* A subgenome). Orthofinder [65] used to cluster genes into orthologous gene families with S set diamond and all other parameters set default. Single-copy orthologous gene pairs with one copy from each genome and subgenome were used for phylogenetic analysis. Protein sequences of single-copy gene families were aligned by MUSCLE (v3.8.1551) [66]. Well-aligned protein sequences were obtained using Gblocks [67]. Based on 4D sites from the coding sequence (CDS) alignments was used to construct phylogenetic tree by RA x ML [68]. Moreover, a Neighbor–Joining (NJ) tree was built using transcriptome data from four species (*G. thurberi*, *G. klotzschianum*, *G. raimondii* and *G. trilobum*) SNPs using BMK Cloud (www.biocloud.net). Evolutionary rate of each lineage for the nine gnomes was estimated using the Codeml program in the PAML [69] package with a free ration model (*model= 1*). We grabbed the dN, dS, and dN/dS from the result of Codeml. We filter the Genes with dS=0. If the PSGs *dN/dS> 1* were reported as PSGs.

Gene clustering and visualization

To assess the gene expression patterns over time within each species, K-means clustering was used to visualize genes expression pattern (log_2-transformed FPKM values) using BMKCloud (www.biocloud.net).

Weighted gene co-expression network analysis (WGCNA)

The WGCNA [40, 41] package was used to identify gene co-expression network and investigate trait-related modules. WGCNA analysis (*FPKM:>= 1*; variation of FPKM: *cv>= 0.5*;minimum module size: 30; minimum height for merging modules: 0.25) were performed using BMK Cloud (www.biocloud.net). Base on expression data of forty libraries, a matrix of Pearson's correlation between all pair-genes were generated. And then, transformed into an adjacency matrix (a matrix of connection strengths) using the formula: connection strength (adjacency value) = |Pearson's correlation|. Here, parameter represents soft threshold for the correlation matrix, which emphasize strong correlations between genes and penalize weak correlations [39]. A value of 5 was selected accurately selected by determination of soft-thresholding power and evaluation of scale free topology analysis. The dynamic tree-cutting algorithm was used to cut the hierarchal clustering dendrogram and modules were defined after decomposing/combining branches to reach a stable number of clusters [44]. We determined the correlation between each ME with the traits (conditions) as described in previous study [44]. Then, the

association of module with traits were determined, and the correlation matrix was drawn by the R package (ggplot).

Statistical analysis

Statistical analysis of the experimental data was statistically analyzed using the R (v3.5.0).

Data Deposition

The RNA-seq data reported in the article have been deposited in the database of the National Center for Biotechnology Information (NCBI) under accession number PRJNA554555 (https://www.ncbi.nlm.nih.gov/sra/ PRJNA554555).

Abbreviations

WGCNA: Weighted gene co-expression network analysis; NCBI: National Center for Biotechnology Information; PSG: Positive selected genes; DEG: differentiated expressed genes; SNPs: Single Nucleotide Polymorphisms;

Declarations

Ethics approval and consent to participate

No ethical nor consent to participate in this research was sought. The research work was conducted as per the broad mandate of the cotton research institute (CRI), which is the state owed research institute charged with the responsibility to develop, carry out research and approve all the cotton breeding work in China.

Consent to publish

Not applicable

Availability of data and materials

All the relevant data and supplementary data are all availed including the primer sequences used in carrying out the RT-qPCR validation of the significant genes mined for the vital QTLs detected in this research work. All supplementary data supporting this research work are all made available in a public data repository and can be accessed

Competing interests

The authors declare no any form of competing interest

Funding

This research was funded by the National Natural Science Foundation of China, grant number 31530053. The funding numbers provided the financial support to the research programs, but didn't involve in work design, data collection, analysis and preparation of the manuscript.

Author Contributions:

YX, FL, KW, ZZ and ROM designed and conceived the study. YX, YCX, DSJ, and ROM performed the experiments. YX, XC, ZZ, YH, ROM and FL contributed the materials/analysis tools. YX, and ROM carried the bioinformatics analysis; FL, KW, ZZ, and YC supervised the research work; YX, ROM, FL, KW and ZZ interpreted the data and revised the manuscript; FL, KW, ROM and DJ revised the manuscript. All authors approved the manuscript for submission

Acknowledgements

We are deeply indebted to the entire research team for their support during this research work.

References

1.Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM *et al: Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM–1) provides a resource for fiber improvement. Nat Biotechnol* 2015, *33 (*5):531–537.

2.Huang C, Nie XH, Shen C, You CY, Li W, Zhao WX, Zhang XL, Lin ZX: *Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. Plant Biotechnology Journal* 2017, *15(*11):1374–1386.

3.Shan XH, Liu ZL, Dong ZY, Wang YM, Chen Y, Lin XY, Long LK, Han FP, Dong YS, Liu B: *Mobilization of the active MITE transposons mPing and Pong in rice by introgression from wild rice (Zizania latifolia Griseb.). Mol Biol Evol* 2005, *22(*4):976–990.

4.Senchina DS, Alvarez I, Cronn RC, Liu B, Rong JK, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF: *Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol Biol Evol* 2003, *20(*4):633–643.

5.Flagel LE, Wendel JF, Udall JA: *Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. Bmc Genomics* 2012, *13.*

6.Endrizzi JE, Turcotte EL, Kohel RJ: *Genetics, Cytology, and Evolution of Gossypium. Advances in Genetics* 1985, *23:*271–375.

7.Grover CE, Arick MA, 2nd, Thrash A, Conover JL, Sanders WS, Peterson DG, Frelichowski JE, Scheffler JA, Scheffler BE, Wendel JF: *Insights into the Evolution of the New World Diploid Cottons (Gossypium, Subgenus Houzingenia) Based on Genome Sequencing. Genome Biol Evol* 2019, *11(*1):53–71.

8.Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S *et al: The draft genome of a diploid cotton Gossypium raimondii. Nat Genet* 2012, *44(*10):1098–1103.

9.Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J *et al: Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature* 2012, *492(*7429):423–427.

10.Kunbo W, Jonathan W: *Designations for individual genomes and chromosomes in Gossypium. Journal of Cotton Research* 2018, *1:*3.

11.Wendel JF, Flagel LE, Adams KL: *Jeans, Genes, and Genomes: Cotton as a Model for Studying Polyploidy.* In: *Polyploidy and Genome Evolution.* 2012: 181–207.

12.Wendel JF, Cronn RC: *Polyploidy and the evolutionary history of cotton.* In: *Advances in Agronomy.* vol. 78: Academic Press; 2003: 139–186.

13.Maiti RK, Satya P: *Research advances in major cereal crops for adaptation to abiotic stresses. GM Crops Food* 2014, *5(*4):259–279.

14.Fedoroff NV, Battisti DS, Beachy RN, Cooper PJ, Fischhoff DA, Hodges CN, Knauf VC, Lobell D, Mazur BJ, Molden D *et al: Radically rethinking agriculture for the 21st century. Science* 2010, *327(*5967):833–834.

15.Huang GT, Ma SL, Bai LP, Zhang L, Ma H, Jia P, Liu J, Zhong M, Guo ZF: *Signal transduction during cold, salt, and drought stresses in plants. Mol Biol Rep* 2012, *39(*2):969–987.

16.Li X, Li M, Zhou B, Yang Y, Wei Q, Zhang J: *Transcriptome analysis provides insights into the stress response crosstalk in apple (Malus x domestica) subjected to drought, cold and high salinity. Sci Rep* 2019, *9(*1):9071.

17.de Silva K, Laska B, Brown C, Sederoff HW, Khodakovskaya M: *Arabidopsis thaliana calcium-dependent lipid-binding protein (AtCLB): a novel repressor of abiotic stress response. J Exp Bot* 2011, *62(*8):2679–2689.

18.Franz S, Ehlert B, Liese A, Kurth J, Cazale AC, Romeis T: *Calcium-dependent protein kinase CPK21 functions in abiotic stress response in Arabidopsis thaliana. Mol Plant* 2011, *4(*1):83–96.

19.Zhu JK: *Abiotic Stress Signaling and Responses in Plants. Cell* 2016, *167(*2):313–324.

20.Ma Y, Dai X, Xu Y, Luo W, Zheng X, Zeng D, Pan Y, Lin X, Liu H, Zhang D *et al: COLD1 confers chilling tolerance in rice. Cell* 2015, *160(*6):1209–1221.

21.de Zelicourt A, Colcombet J, Hirt H: *The Role of MAPK Modules and ABA during Abiotic Stress Signaling. Trends Plant Sci* 2016, *21(*8):677–685.

22.Yan J, Wang P, Wang B, Hsu CC, Tang K, Zhang H, Hou YJ, Zhao Y, Wang Q, Zhao C *et al: The SnRK2 kinases modulate miRNA accumulation in Arabidopsis. PLoS Genet* 2017, *13(*4):e1006753.

23.Boudsocq M, Barbier-Brygoo H, Lauriere C: *Identification of nine sucrose nonfermenting 1-related protein kinases 2 activated by hyperosmotic and saline stresses in Arabidopsis thaliana. J Biol Chem* 2004, *279(*40):41758–41766.

24.Wang MJ, Tu LL, Lin M, Lin ZX, Wang PC, Yang QY, Ye ZX, Shen C, Li JY, Zhang L *et al: Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. Nature Genetics* 2017, *49(*4):579-+.

25.Fang L, Wang Q, Hu Y, Jia Y, Chen J, Liu B, Zhang Z, Guan X, Chen S, Zhou B *et al: Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet* 2017, *49(7*):1089–1098.

26.Fang L, Gong H, Hu Y, Liu C, Zhou B, Huang T, Wang Y, Chen S, Fang DD, Du X *et al: Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. Genome Biol* 2017, *18(*1):33.

27.Yoo MJ, Wendel JF: *Comparative evolutionary and developmental dynamics of the cotton (Gossypium hirsutum) fiber transcriptome. PLoS Genet* 2014, *10(*1):e1004073.

28.Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C *et al: Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet* 2014, *46(*6):567–572.

29.Wang M, Tu L, Yuan D, Zhu, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G *et al: Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. Nat Genet* 2019, *51(*2):224–229.

30.Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y, Ju L, Deng J, Zhao T, Lian J *et al: Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet* 2019, *51(*4):739–748.

31.Wang M, Wang P, Lin M, Ye Z, Li G, Tu L, Shen C, Li J, Yang Q, Zhang X: *Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat Plants* 2018, *4(*2):90–97.

32.Guo J, Shi G, Guo X, Zhang L, Xu W, Wang Y, Su Z, Hua J: *Transcriptome analysis reveals that distinct metabolic pathways operate in salt-tolerant and salt-sensitive upland cotton varieties subjected to salinity stress. Plant Sci* 2015, *238:*33–45.

33.Savoi S, Wong DC, Arapitsas P, Miculan M, Bucchetti B, Peterlunger E, Fait A, Mattivi F, Castellarin SD: *Transcriptome and metabolite profiling reveals that prolonged drought modulates the phenylpropanoid and terpenoid pathway in white grapes (Vitis vinifera L.). BMC Plant Biol* 2016, *16:*67.

34.Zhou Q, Guo JJ, He CT, Shen C, Huang YY, Chen JX, Guo JH, Yuan JG, Yang ZY: *Comparative Transcriptome Analysis between Low- and High-Cadmium-Accumulating Genotypes of Pakchoi (Brassica chinensis L.) in Response to Cadmium Stress. Environ Sci Technol* 2016, *50(*12):6485–6494.

35.Wei Y, Xu Y, Lu P, Wang X, Li Z, Cai X, Zhou Z, Wang Y, Zhang Z, Lin Z *et al: Salt stress responsiveness of a wild cotton species (Gossypium klotzschianum) based on transcriptomic analysis. PLoS One* 2017, *12(*5):e0178313.

36.Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA *et al: Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A* 2014, *111(*45):E4859–4868.

37.Koenig D, Jimenez-Gomez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, Kumar R, Covington MF, Devisetty UK, Tat AV *et al: Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci U S A* 2013, *110(*28):E2655–2662.

38.Hao Y, Xiong Y, Cheng Y, Song G, Jia C, Qu Y, Lei F: *Comparative transcriptomics of 3 high-altitude passerine birds and their low-altitude relatives. Proc Natl Acad Sci U S A* 2019, *116(*24):11851–11856.

39.Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, Zhao Q: *Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci* 2018, *14(*2):124–136.

40.Pei G, Chen L, Zhang W: *WGCNA Application to Proteomic and Metabolomic Data Analysis. Method Enzymol* 2017, *585:*135–158.

41.Langfelder P, Horvath S: *WGCNA: an R package for weighted correlation network analysis. Bmc Bioinformatics* 2008, *9:*559.

42.Yin L, Cai Z, Zhu B, Xu C: *Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. Genes (Basel)* 2018, *9(*2).

43.Yin L, Cai ZH, Zhu BA, Xu CS: *Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. Genes-Basel* 2018, *9(*2).

44.Garg R, Singh VK, Rajkumar MS, Kumar V, Jain M: *Global transcriptome and coexpression network analyses reveal cultivar-specific molecular signatures associated with seed development and seed size/weight determination in chickpea. Plant Journal* 2017, *91(*6):1088–1107.

45.Oldham MC, Horvath S, Geschwind DH: *Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A* 2006, *103(*47):17973–17978.

46.Tran LS, Nakashima K, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K: *Isolation and functional analysis of Arabidopsis stress-inducible NAC*

transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. Plant Cell 2004, *16(*9):2481–2498.

47.Froidure S, Canonne J, Daniel X, Jauneau A, Brière C, Roby D, Rivas S: *AtsPLA2-α nuclear relocalization by the Arabidopsis transcription factor AtMYB30 leads to repression of the plant defense response. Proceedings of the National Academy of Sciences* 2010, *107(*34):15281–15286.

48.Gan Y, Chen D, Liu F, Wang C, Li S, Zhang X, Wang Y, Peng R, Wang K: *Individual chromosome assignment and chromosomal collinearity in Gossypium thurberi, G. trilobum and D subgenome of G. barbadense revealed by BAC-FISH. Genes Genet Syst* 2011, *86(*3):165–174.

49.Seeholzer S, Tsuchimatsu T, Jordan T, Bieri S, Pajonk S, Yang WX, Jahoor A, Shimizu KK, Keller B, Schulzelefert P: *Diversity at the Mla powdery mildew resistance locus from cultivated barley reveals sites of positive selection. Mol Plant Microbe Interact* 2010, *23(*4):497–509.

50.Khan AM, Khan AA, Azhar MT, Amrao L, Cheema HM: *Comparative analysis of resistance gene analogues encoding NBS-LRR domains in cotton. Journal of the Science of Food & Agriculture* 2016, *96(*2):530–538.

51.Zambounis A, Ganopoulos I, Kalivas A, Tsaftaris A, Madesis P: *Identification and evidence of positive selection upon resistance gene analogs in cotton (Gossypium hirsutum L.). Physiology & Molecular Biology of Plants* 2016, *22(*3):1–7.

52.Sun S, Wang T, Wang L, Li X, Jia Y, Liu C, Huang X, Xie W, Wang X: *Natural selection of a GSK3 determines rice mesocotyl domestication by coordinating strigolactone and brassinosteroid signaling. Nature Communications* 2018, *9(*1):2523.

53.Paape T, Briskine RV, Lischer HEL, Halsteadnussloch G, Shimizuinatsugi R, Hatekayama M, Tanaka K, Nishiyama T, Sabirov R, Sese J: *Patterns of polymorphism, selection and linkage disequilibrium in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nature Communications* 2018, *9(*1):3909.

54.Gao H, Wang Y, Li W, Gu Y, Lai Y, Bi Y, He C: *Transcriptomic comparison reveals genetic variation potentially underlying seed developmental evolution of soybeans. J Exp Bot* 2018, *69(*21):5089–5104.

55.Cong B, Barrero LS, Tanksley SD: *Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. Nat Genet* 2008, *40(*6):800–804.

56.Yuan F, Yang H, Xue Y, Kong D, Ye R, Li C, Zhang J, Theprungsirikul L, Shrift T, Krichilsky B *et al: OSCA1 mediates osmotic-stress-evoked Ca2+ increases vital for osmosensing in Arabidopsis. Nature* 2014, *514(*7522):367–371.

57.Chen L, Ren F, Zhou L, Wang QQ, Zhong H, Li XB: *The Brassica napus calcineurin B-Like 1/CBL-interacting protein kinase 6 (CBL1/CIPK6) component is involved in the plant response to abiotic stress and ABA signalling. J Exp Bot* 2012, *63(*17):6211–6222.

58.Stephan AB, Schroeder JI: *Plant salt stress status is transmitted systemically via propagating calcium waves. Proc Natl Acad Sci U S A* 2014, *111(*17):6126–6127.

59.Yuan P, Yang T, Poovaiah BW: *Calcium Signaling-Mediated Plant Response to Cold Stress. Int J Mol Sci* 2018, *19(*12).

60.Zhang M, Wang D, Kang Y, Wu JX, Yao F, Pan C, Yan Z, Song C, Chen L: *Structure of the mechanosensitive OSCA channels. Nature structural & molecular biology* 2018, *25(*9):850–858.

61.Xu Y, Magwanga RO, Cai X, Zhou Z, Wang X, Wang Y, Zhang Z, Jin D, Guo X, Wei Y *et al: Deep Transcriptome Analysis Reveals Reactive Oxygen Species (ROS) Network Evolution, Response to Abiotic Stress, and Regulation of Fiber Development in Cotton. Int J Mol Sci* 2019, *20(*8).

62.Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol* 2010, *28(*5):511–515.

63.Love MI, Huber W, Anders S: *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol* 2014, *15(*12):550.

64.Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology* 2011, *29:*644.

65.Emms DM, Kelly S: *OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol* 2015, *16:*157.

66.Edgar RC: *Quality measures for protein alignment benchmarks. Nucleic acids research* 2010, *38(7*):2145–2153.

67.Castresana J: *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol* 2000, *17(*4):540–552.

68.Stamatakis A: *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics* 2014, *30(*9):1312–1313.

69.Yang Z: *PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol* 2007, *24(*8):1586–1591.

Figures

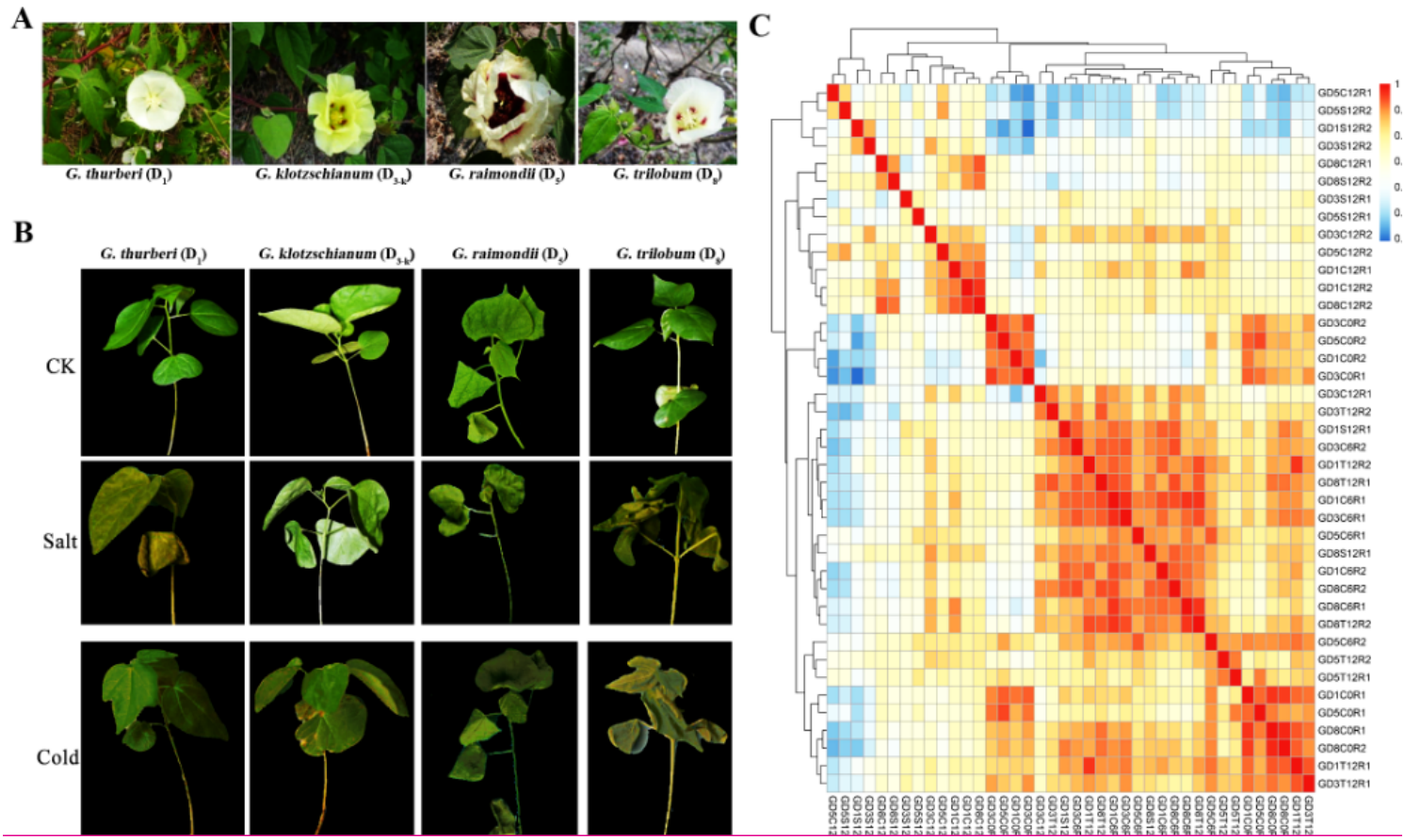

Figure 1

Phenotypic variations and correlation of whole-genome expression of four species. (A) Phenotypic variations in flower color and leaf shape. (B) tolerance divergence of four species in seedlings. (C) heatmap of correlation value (R square) of 40 libraries.

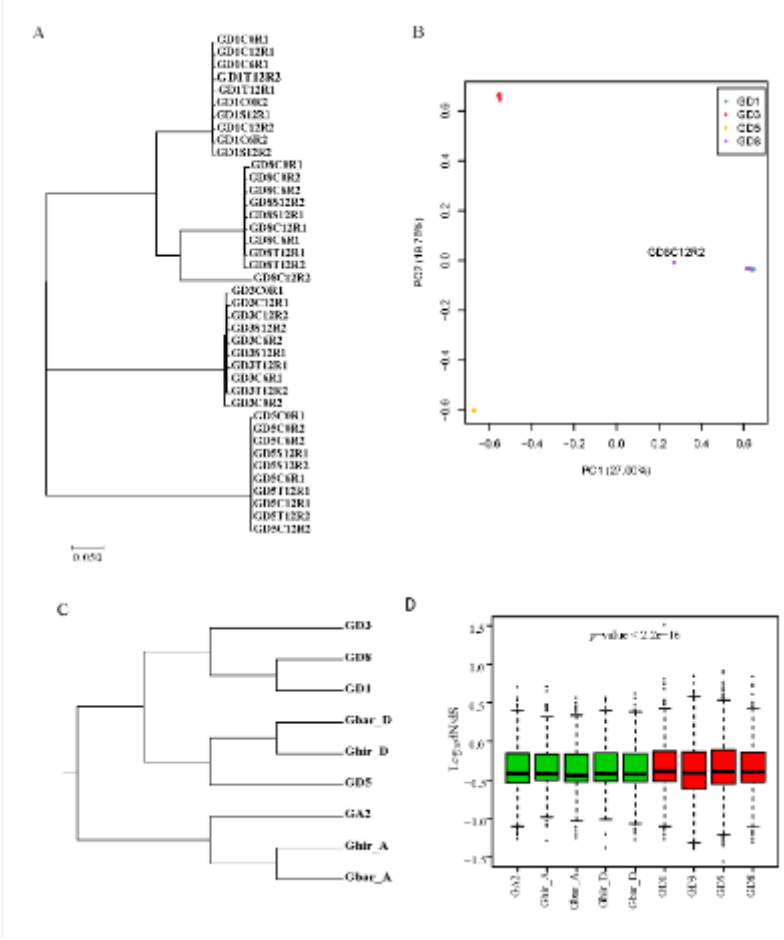

Figure 2

Sequence diversity of four species of cotton. (A) Unrooted phylogenetic tree of four species using SNPs, which obtained from transcriptomic data. The scale bar represents the expected number of substitutions per site. (B) relationship shown by principal component cluster among four species. (C) phylogenetic tree of nine genomes using the identified orthologous genes. The scale bar represents the expected number of substitutions per site. (D) Boxplot of the dN/dS ratio of nine genomes. Wild species, red boxes. Cultivated species, green boxes.

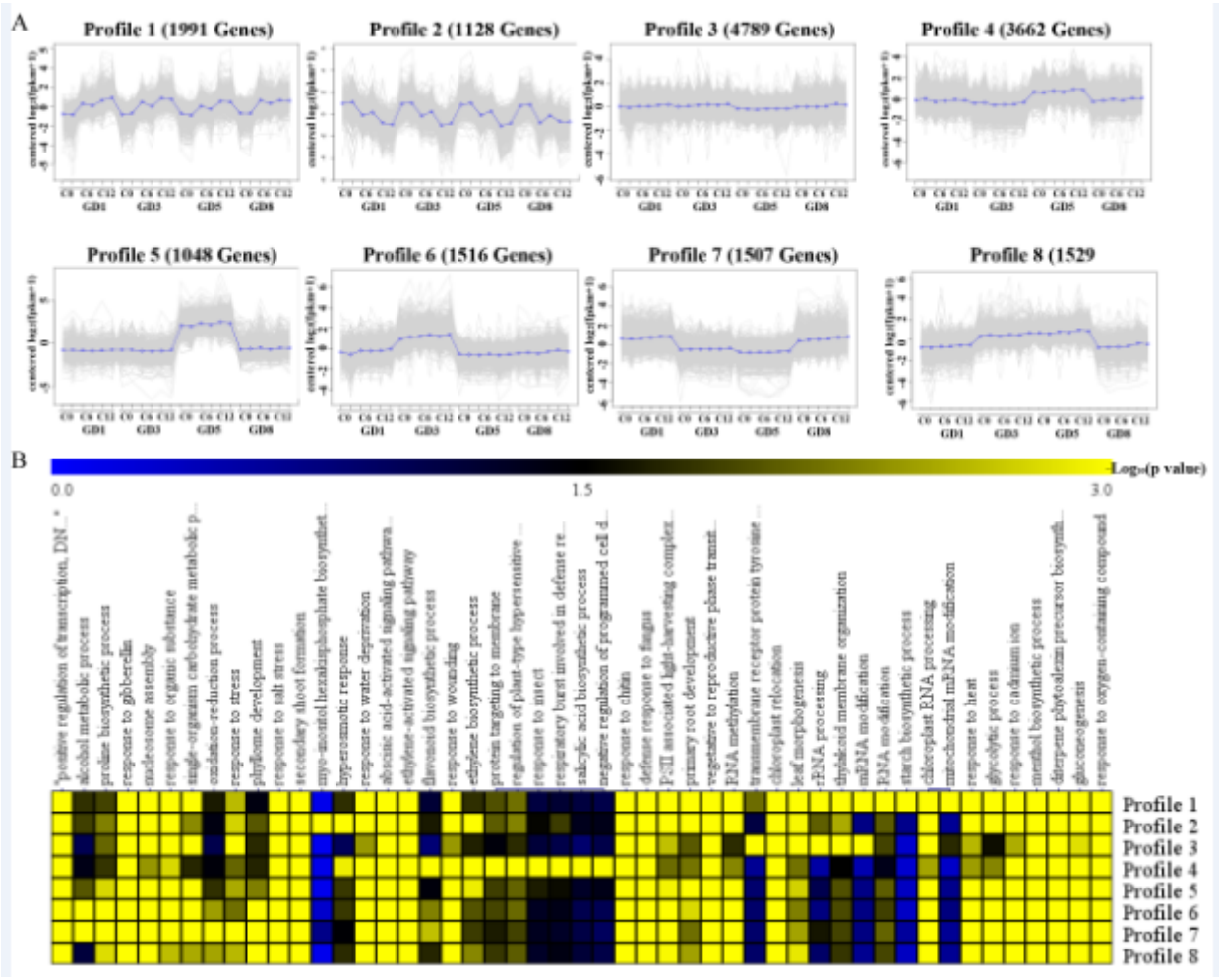

Figure 3

Gene expression pattern across three time points (C0, C6, C12) of four species under normal condition and corresponding top fifteen enrichment GO terms. (A) Gene expression pattern across three time points. Eight gene clusters (profile 1–8) were identified using k-means clustering. In each cluster, the y-axis represents log2(FPKM+1) derived from RNA-seq data for each biological sample, while the x-axis represents the biological samples that are ordered as C0 (R1 and R2), C6(R1 and R2), and C12(R1 and R2) for each species. (B) Heatmap of −log10(p value) of biological process category enrichment among the eight profiles.

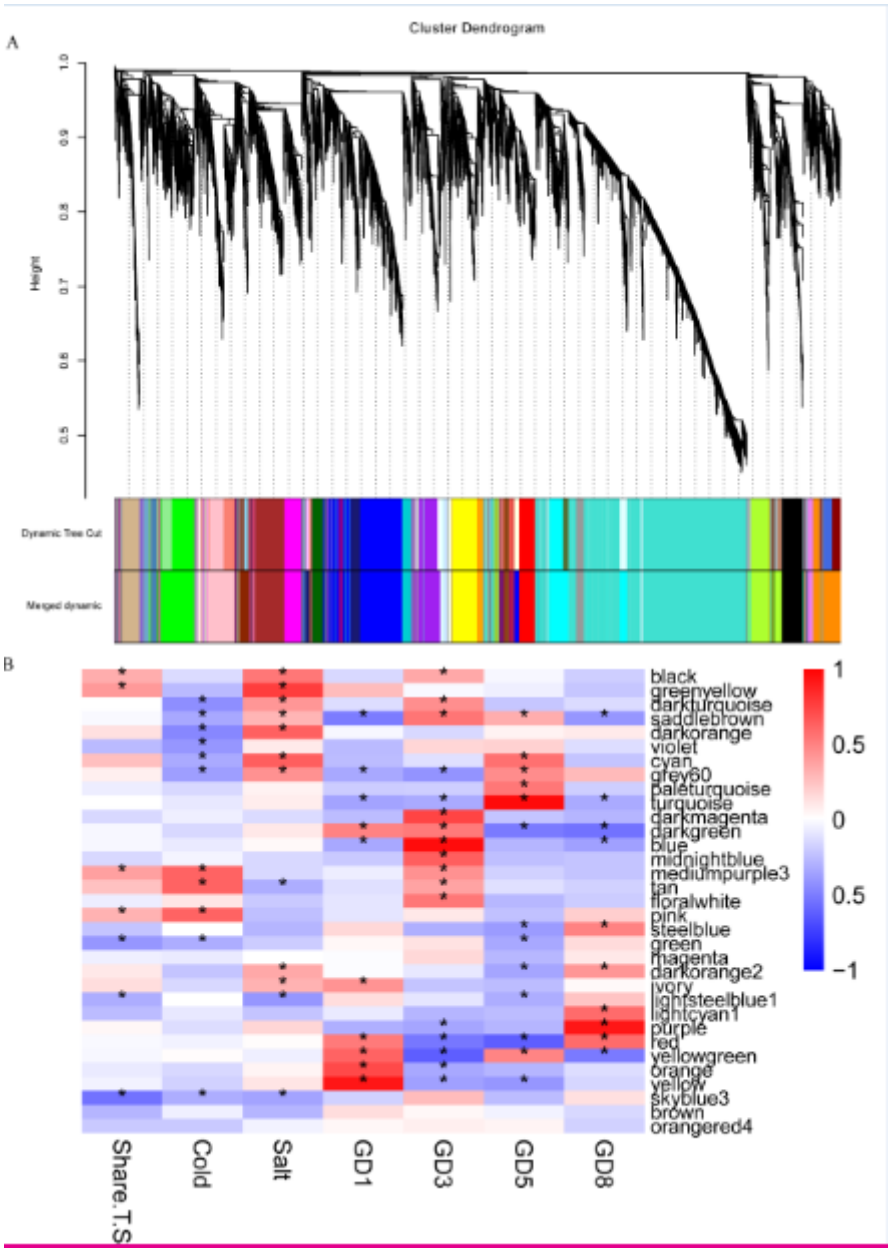

Figure 4

Co-expression network analysis by WGCNA. (A) Hierarchical cluster tree showing co-expression modules identified by WGCNA. Each leaf in the tree represents one gene. The major tree branches constitute 33 modules labeled with different colors. (B) Module–sample association. Each row corresponds to a module labeled with a color as in (A) Modules are distinguished by different colors which were arbitrarily assigned by the WGCNA package. Each column corresponds to a tissue type as indicated. The color of each cell at the row–column intersection indicates the correlation coefficient (R) between the module and the tissue type. *Significance at P<0.05; **Significance at P<0.01

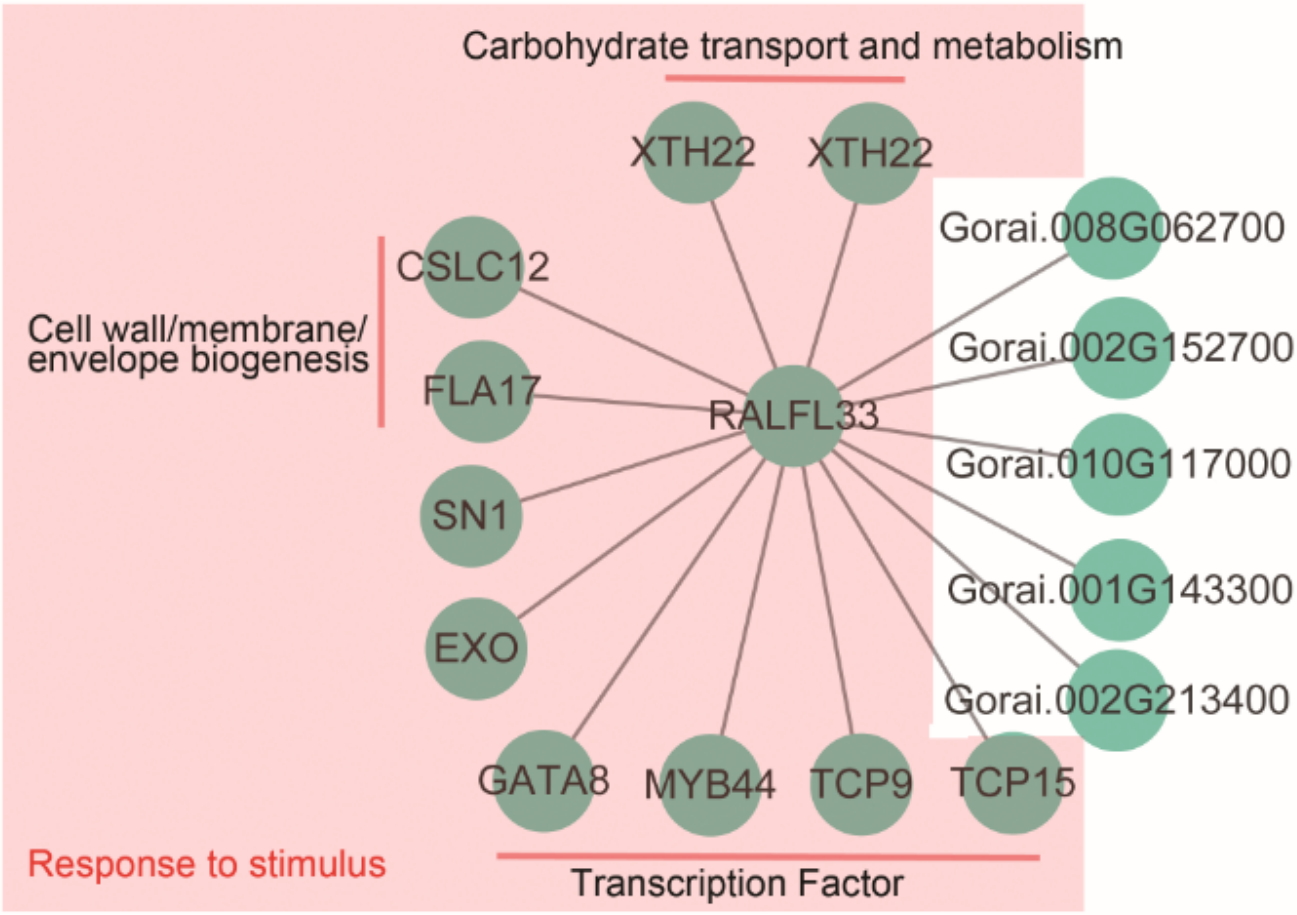

Figure 5

co-expression network of hub gene in skyblue3 modules. Genes was related to response to stimulus in red background. Five function unknown genes were displayed in white background.

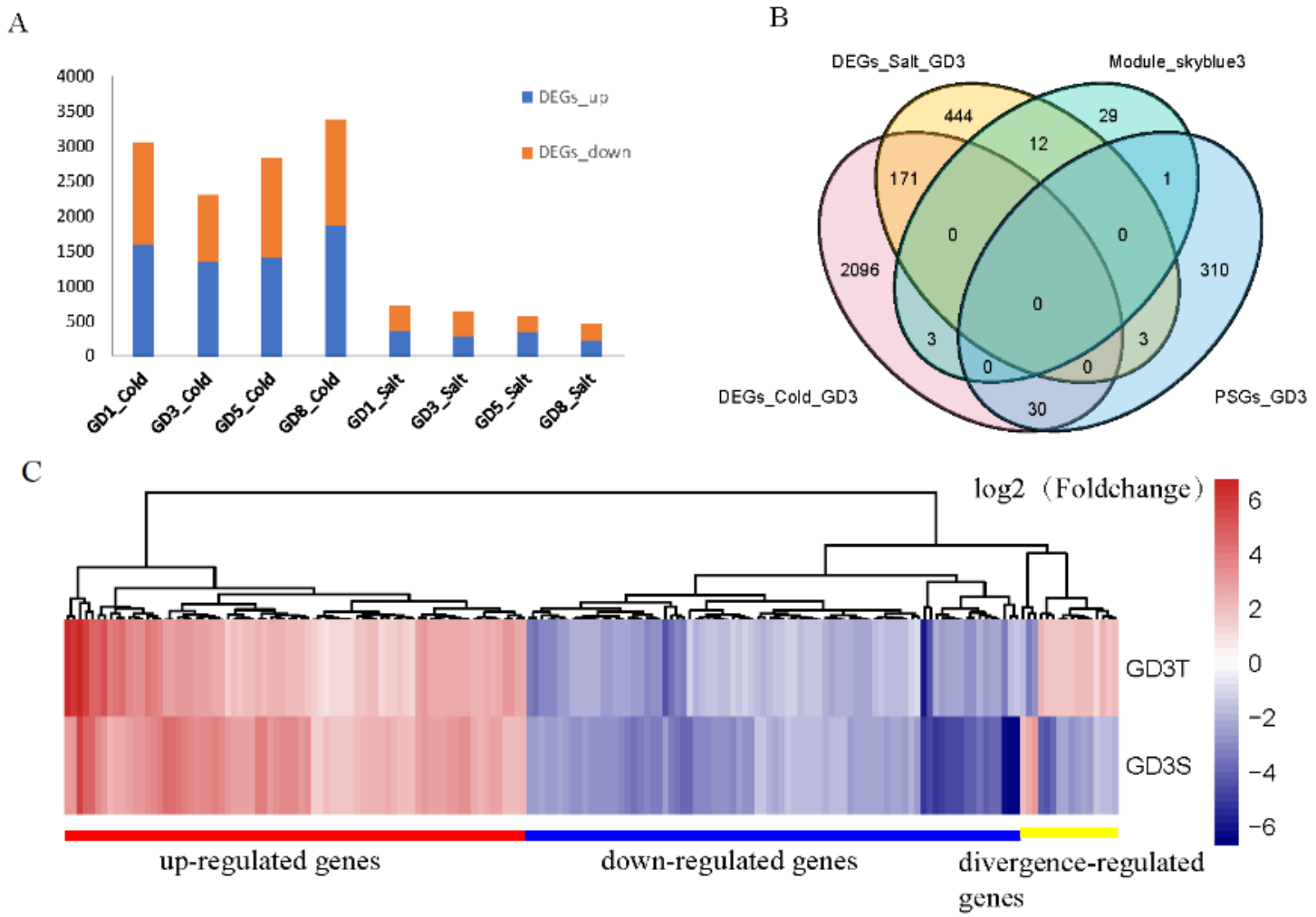

Figure 6

Characterization of DEGs under cold and salt stress. (A) numbers of DEGs of four species under cold and salt. Red block means numbers of down-regulated genes. Blue block means numbers of up-regulated genes. (B) DEGs of GD3 under cold and salt stress; PSGs of GD3; and genes of skyblue3 module shown by Venn. (C) Common DEGs of GD3 under cold and salt stress.

Supplementary Files

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

- TableS1SummaryofRNA.docx
- TableS5GOenrichmentanalysisofPSGsinwildspecies.docx
- TableS7.docx
- TableS6.docx
- FigureS3.tiff
- TableS2SNPsinformationofallsamples.docx

- FigureS2.tiff
- TableS4PSGsidentifiedofninegenome.xlsx
- FigureS1.tiff
- TableS3.docx