THE FULL-LENGTH TRANSCRIPTOME BY THE SINGLE-MOLECULE LONG-READ SEQUENCING REVEALS A HEAT-RESISTANT MECHANISM IN CAPER BUSH (*CAPPARIS SPINOSA* L.)

LIU, Z.[1] – ZHOU, K.[2] – WANG, L.[3,4] – LI, S.[1] – CHEN, G.[1] – SUN, Z.[5] – SUN, R.[2*] – QANMBER, G.[2*]

[1]*Zhengzhou Research Base, State Key Laboratory of Cotton Biology, School of Agricultural Sciences, Zhengzhou University, Zhengzhou, 450001 Henan, China*

[2]*State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, 455000 Henan, China*

[3]*State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, 830011 Urumqi, China*

[4]*University of Chinese Academy of Sciences, 100049 Beijing, China*

[5]*Development Center for Science and Technology, Ministry of Agriculture and Rural Affairs, 100122 Beijing, China*

**Corresponding authors*
e-mail: sunruibin@caas.cn; gqkhan12@gmail.com

(Received 21st May 2021; accepted 20th Sep 2021)

Abstract. Recent global climate change affects the living conditions of many plant and animal species on Earth. *Capparis spinosa* L. is a perennial xerophytic shrub with significant adaptability to arid environments and a candidate species for the prevention of soil erosion. However, its reference genome or large-scale full-length cDNA sequences are still lacking. Here, we applied a combination of second and third-generation sequencing technologies to sequence the full-length transcriptome in *Capparis spinosa* L. In total, 14.96 GB of clean reads were generated, including 828,518 reads of insert (ROI) and 370,258 full-length non-chimeric (FLNC) reads. We reported 216,240 consensus isoforms by transcript clustering analysis. After removing redundant reads, we identified 191,599 non-redundant isoforms of which 171,893 were coding isoforms. Through second-generation sequencing (SGS), we obtained 434 differentially expressed genes (DEGs) from samples collected under different temperatures for further analysis. Among these DEGs, we found 29 transcription factors belonging to seven different TF families in which *CsMYB28* and *CsMYB73* were up-regulated as the temperature rose. Overall, this is the first study to perform SMRT sequencing of full-length transcriptome of *Capparis spinosa* L. Our results will contribute to increasing interest for *Capparis spinosa* L as a candidate crop for future environmental challenges.
Keywords: *global warming, Capparis spinosa, drought tolerance, SMRT, caper*

Introduction

In recent years, climate change has been reported as a factor that has adverse effects on biological survival around the world (Bellard et al., 2012). Climate change has a significant impact on biodiversity, especially on changes in the distribution of species (Peterson et al., 2004). As the environment changes more drastically, the distribution of species will become more uncontrollable (Lagouvardos and Kotroni, 2007). Moreover, climate change has significantly affected agricultural production. High temperature, heat and drought stress make it impossible to ensure global food security (Campbell,

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

2015). Environmental factors such as drought, salinity and extreme temperatures adversely affect plant growth and productivity.

However, the introduction of stress-tolerant crops and varieties into agricultural systems is not a rapid process. Xerophytic (drought-tolerant) plants are those that are able to adapt to drought and can grow, develop and reproduce as usual under such conditions (Henckel, 1964). Through the evolution, drought-tolerant plants acquired a series of morphological and biochemical changes in roots, stems and leaves in order to adapt to this stress. Such crops may be an effective way to promote sustainable agriculture (Thiry et al., 2016). This study pay attention to the crop that can adapt to heat stress namely caper (*Capparis spinosa* L.).

Capparis spinosa L. is a typical heat-tolerant plants, which could flourish under extremely high temperatures and poor soil conditions (Nabavi et al., 2016). It is usually grown on the sandy loam with low alkalinity. Since it has a deep and extensive root system that can grow in harsh environments, it is recommended to prevent land degradation and control soil erosion (Nabavi et al., 2016). *C. spinosa* is a dicotyledonous 20-30 cm tall shrub that is widely distributed in the Mediterranean countries such as Greece, Italy, Turkey, Morocco and Spain (Inocencio et al., 2000). In China, it is distributed in Xinjiang and Tibet especially in Turpan, Korla and Aksu areas of Xinjiang province. *C. spinosa* is also known as Caper, (wild watermelon in China), Cappero (in Italy) and Alcapparo (in Spain) (Tlili et al., 2011). *C. spinosa* is a winter-deciduous perennial and evergreen shrub, growing and flowering from May to October during summer (Rhizopoulou and Psaras, 2003). The flowers are open pollinated at night and die soon after sunrise (Rhizopoulou and Psaras, 2003). It withstands not only over 45 ℃ in summer and over 50 ℃ in the gravel Gobi, but also hurricanes with winds of about 40 days every year (Rhizopoulou and Psaras, 2003). Its main roots are flourishing which extend up to 30-40 m underground with well-developed vascular system, so the groundwater resources can be effectively utilized (Rhizopoulou and Psaras, 2003).

The adaptation of *C. spinosa* to drought is based on its osmotic adjustment, regulation of stomatal opening, modification of cell wall properties, and extensive root systems. The stomata remains open throughout the day, with a high transpiration rate, resulting in leaf temperature of 3.9 ℃ less than air temperature (Rhizopoulou and Psaras, 2003). The stomatal length of *C. spinosa* is 28 um, which is much larger than 134 desert plants (Gadgoli and Mishra, 1999). The stomatal density is also high among all desert plants (Gadgoli and Mishra, 1999). Dense root system and transport tissues enabled *C. spinosa* to adapt environmental stress. Further, the sugar and proline content are high in the *C. spinosa*, as both are important cell osmotic adjustment substance, and regulate the flow of water between the membranes and maintaining the balance of water in tissues (Rhizopoulou and Psaras, 2003). In *C. spinosa*, unsaturated fatty acids are the main components of lipids in petals, which affect the fluidity of the membrane. The solute and water potential of petals are low when flowing in the night, which keep enough water in cells while after sunrise, the water and solute potential begin to rise (Rhizopoulou et al., 2006). Strong drought resistance properties of *C. spinosa* play an important ecological and economic role in desert areas.

In the past years, structural and functional genomics have laid the foundation for exploring plant biology. In order to carry out these studies efficiently, it is necessary to obtain high quality genomic and transcriptome sequences (Margulies et al., 2005). Second-generation sequencing (SGS) technology has changed DNA sequencing and

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

genomic/transcriptome studies (Shendure and Ji, 2008). However, one of the disadvantages of SGS is that they have a shorter read length (i.e., hundreds of base pairs). As a result, shorter read length reduces the accuracy of sequence assembly and increase the difficulty of bioinformatics analysis in the future. Recently, a third-generation sequencing (TGS) platform of single-molecule real-time (SMRT) sequencing carries out in the PacBio RS (Pacific Biosciences of California) that is widely used in genome sequencing because of its long reads (average 4-8 kb) sequencing technology (Eid et al., 2009). TGS has long read length and contributes to the de novo genome and transcriptome assembly in higher organisms and makes TGS the best choice for full-length transcripts (Sharon et al., 2013). However, relatively high error rate of TGS may be problematic in sequence alignment and bioinformatics analysis. But this problem can be minimized and improved by high-accuracy of SGS (Li et al., 2014). Previously it has been shown that combining SGS and TGS technologies could provide high quality and more complete assembly in genomic and transcriptome studies (Sharon et al., 2013). Recently, transcriptome sequencing by SGS and TGS technologies has been widely applied in plant development and stress response research (Minio et al., 2019. Filichkin et al., 2018; Pan et al., 2020).

Though the excellent adaptation ability of *C. spinosa* to the extreme conditions, the molecular mechanisms were still unknown. In this study, we constructed a full-length cDNA library from *C. spinosa* derived from four different tissues (root, stem, leaves and shoots) by using SMRT sequencing. In addition, leaves collected at three different time points (8:00 am, 2:00 pm, 8:00 pm) of a day were used as sample for RNA sequencing. The annotation of full-length transcriptome analysis of *C. spinosa* helps to understand the complexity of the *C. spinosa* genome and provides a reference sequence for gene cloning and mechanism analysis of *C. spinosa*'s adaptation. This will help improve this species and develop more intensive research, especially in responding to climate change.

Methods

Sample collection, RNA extraction, library construction and sequencing

The roots, stems, leaves and shoots of *C. spinosa* (Specimen number 05SYIIM213:2) were taken from State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, China. C. spinosa plants were sampled from Turfan desert in July, where is one of the most dry with 25-100 mm of annual precipitation and the hottest area in southern Xinjiang. At three different time points (8:00 am, 2:00 pm, 8:00 pm) we collected leaves and measured the temperature respectively. Total RNA was extracted by RNAprep Pure Plant Kit (TIANGEN), NanoDrop 2000 microvolume spectrophotometer instrument (Thermo Scientific, USA) was used to measure the purity and concentration of RNA. The quality and integrity of RNA was tested by agarose gel electrophoresis and Agilent Bioanalyzer 2100 system (Agilent Technologies, USA). Qualified RNA samples were used for following experiments. For SMRT sequencing, a portion of RNA samples were taken and mixed with equal molar ratio and oligo-dT magnetic beads were used to enrich mRNA. Reverse transcription PCR was conducted by using SMARTer™ PCR cDNA Synthesis Kit (Clontech, Japan) to synthesize full-length cDNA. BluePippin (Sage Science, USA) was used to select specific size of full-length cDNA for construction of library with specific size of insert cDNA. After re-amplification, and end repairing and

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

adenylation, adaptors with a hairpin loop structure were ligated to cDNA to obtain SMRT bell sequencing library. The constructed SMRT library was quantified by Qubit 2.0 (Thermo Scientific, USA), and quality assessed using Agilent Bioanalyzer 2100 system. The qualified SMRT library was sequenced by using a PacBio RSII sequencer (Pacific Biosciences, USA). For common Illumina transcriptome sequencing, purified mRNA was fragmented using the fragmentation agents and served as templates for cDNA synthesis primed by random hexamers. Then the end of cDNA fragments was repaired and added by single nucleotide A (adenine) to ligate with adaptors. By using agarose gel electrophoresis, about 250 bp were selected for constructing library. The amplified libraries were quantified and quality assessed by Qubit 2.0 and Agilent Bioanalyzer 2100 system.

Processing of SMRT sequencing data

The raw SMRT sequencing polymerase reads were processed into reads of insert (ROIs) by removing adapter sequences, ROIs with length of less than 50 bp and quality of less than 0.75 were discarded. During the SMRT library construction, some cDNA sequences may ligate directly with each other without adapter, this leads to some chimeric ROIs with sequencing primer existed in the inner part of sequence. These chimeric ROIs were discarded, only non-chimeric ROIs were further processed. Filtered non-chimeric ROIs within the 5′sequencing primer, 3′sequencing primer and poly A tail before 3′sequencing primer at terminals were recognized as full-length reads, while ROIs lacking one of these three elements were non-full-length reads. After removing terminal primer sequences, the full-length transcripts were used for iterative clustering and error correction referring non-full-length transcripts by ICE （isoform-level clustering for error correction）algorithm by using SMRT analysis 2.3.0 software. After error correction, the obtained polished full-length transcripts with accuracy > 99% were recognized as high quality. Further, the high-quality full-length transcripts were processed to eliminate redundancy using CD-HIT software (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi). The final non-redundancies high quality full-length transcripts were considered as full-length transcript isoforms.

Functional annotation

Transdecoder 3.0.0 software (https://portal.rc.fas.harvard.edu/p3/build-reports/TransDecoder%2F3.0.0-fasrc01) was used to identify candidate coding sequence regions and corresponding proteins of final non-redundancy full-length transcripts with default parameters and integrating the blast against Swissprot (https://web.expasy.org/docs/swiss-prot) and Pfam (https://pfam.xfam.org/) database search results. All obtained full-length transcripts were functionally annotated by mapping transcripts sequences to six sequence databases including GO, KEGG, KOG, NR, NT, and Swissprot.

Transcription factors (TFs) prediction and analysis

The corresponding protein amino acid sequences of full-length transcripts were submitted to iTAK database (http://itak.feilab.net/cgi-bin/itak/index.cgi) to predict genome-wide transcription factors (TFs) using iTAK online v1.6 program with default parameters. The resulted candidate TFs were classified into corresponding TF families automatically.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

NGS sequencing reads mapping, transcripts quantification and differentially expressed genes (DEGs) analysis

Using non-redundancy high quality full-length transcripts as reference, NGS sequencing reads were mapped by Bowtie 2.1.0 software (Langmead et al., 2012), transcripts quantification was conducted using RSEM 1.2.15 software (Li et al., 2011) based on mapping results, expression level of transcripts were computed and represented with fragments per kilo base (kb) of transcript per million fragments mapped (FPKM) value. edgeR 3.14.0 software (Chen et al., 2014) was used for differentially expressed genes (DEGs) identification, and genes with absolute value of expression fold change (FC) > 2 and false discovery rate (FDR) < 0.05 were recognized as differentially expressed genes. The total DEGs were identified in different sample comparison pairs (8:00 am vs 2:00 pm, 8:00 am vs 8:00 pm and 2:00 pm vs 8:00 pm) that were used for subsequent analyses. Expression heatmap within hierarchy clustering of DEGs was drawn using R software package pheatmap based on log10-transformed FPKM values. GO and KEGG pathway enrichment analyses were conducted using topGO 2.24.0 (Alexa et al., 2016) and KOBAS 2.0 (Bu et al., 2021) software respectively.

Quantitative real-time PCR experiment

Remove genomic DNA from RNA samples and reverse cDNA using PrimeScript^{TM}RT reagent Kit with gDNA Eraser (Perfect Real Time) kit (Takara, Dalian, China). By a qRT-PCR experiment, Premix Ex Taq™ II (Takara) was used along with the LightCycler 480 system (Roche). The *CsUBQ6* (C. spinosa_1-6k_c193395/f1p19/1544) was used as the reference gene. The relative expression values were calculated by the $2^{-\Delta CT}$ method.

Results

General quality and data output of SMRT sequencing

C. spinosa is a shrub plant (*Fig. 1A, B*), and is found to be adapted well under drought and poor soil conditions. To obtain a full-length of *C. spinosa* transcriptome, we mixed different tissues for library sequencing. Full-length cDNA library with an insert size of 1-6 kb was sequenced with 2 SMRT cells. In total, we obtain 893869 polymerase reads with 14.96 Gb of clean reads after preprocessing (*Table 1*). A total of 828518 reads of insert (ROIs) were produced with full passes $\geq$ 50 and the predicted consensus accuracy > 0.75, including 44.69% (370528) of full-length non-chimeric reads and 4.98% of full-length chimeric reads. While, the remaining were non-full-length reads (50.33%) (*Table 2*). The mean length of ROI is 1820 bp, and the quality of 0.90 and 8 passes (*Table 2*). Similarly, we obtained 216240 consensus isoforms sequences and ICE (Isoform-level Clustering Algorithm) was used to cluster and polish the non-chimeric transcripts. After clustering and polishing, redundant transcripts were removed. Finally, we obtained 191599 non-redundant isoforms, of which 171893 were coding non-redundant isoforms. The average lengths of non-redundant andcoding non-redundant isoforms are 1674 bp and 985 bp, respectively (*Table 3*). Among non-redundant isoforms, most of them were mainly distributed in length from 1 k to 3 k and among these, 53.05% of the isoforms were 1-2 k in length. The percentage of isoforms of length in 2-3 k was 20.41%. Only very few isoforms were longer than 5 k (*Fig. 1C*).

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

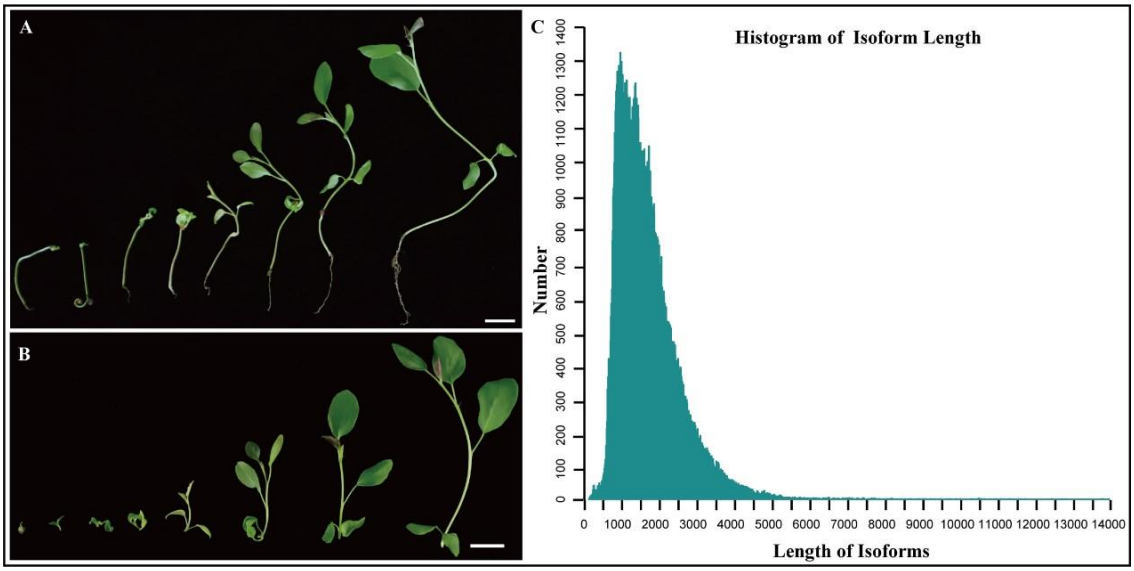

Figure 1. *(A, B) Phenotype of C. spinosa whole plant and shoots. (C) Length distribution of C. spinosa isoforms*

Table 1. *Statistics of polymerase reads*

Sample name	Cell ID	cDNA size	Polymerase read bases	Polymerase reads	Polymerase read N50	Polymerase read length
C. spinosa	A01	1-6 K	7.69 GB	387853	42750	19828
C. spinosa	B01	1-6 K	7.27 GB	506016	34750	14360

Table 2. *Statistics of reads of insert (ROI)*

Sample	*C. spinosa*
cDNA size	1-6 k
Reads of insert	828518
Read bases of insert	1508295266
Number of five prime reads	457118
Number of three prime reads	558636
Number of poly-A reads	503802
Number of filtered short reads	0
Number of non-full-length reads	416980
Number of full-length reads	411538
Number of full-length non-chimeric reads	370258
Full-length non-chimeric reads percentage (%)	44.69%
Mean read length of insert	1820
Mean read quality of insert	0.904459165655
Mean number of passes	8

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

Table 3. *Statistics of non-redundant isoforms*

Sample	*C. spinosa*
cDNA size	1-6 k
Number of consensus isoforms	216240
Number of non-redundant isoforms	191599
Average length of non-redundant isoforms	1674
Number of coding isoforms	171893
Average length of coding isoforms	985

Functional annotation and transcription factors analysis

To predict and analyze the function of 191,599 non-redundant isoforms, we used BLAST to perform functional annotation in GO, KEGG, KOG, NR, Swissprot databases. In total, 186840 isoforms were successfully annotated in at least one out of five subjected databases (*Table 4*), accounting 97.51% of the total isoforms. For functional classification of *C. spinosa* transcripts, GO annotation was performed using BLAST2GO, and transcripts divided into three categories (molecular function, cellular component, and biological process) (*Fig. 2*). In the classification of biological process, major categories were "cellular process" (104104, 18.77% of the total) and "metabolic process" (97177, 17.52% of the total). While in the category of cellular component, isoforms involved in the "cell" (131964, 22.17% of the total) and "cell part" (131635, 22.11% of the total) were main component in our analysis. However, the major subgroups of molecular function were "binding" (89893, 43.44% of the total) and "catalytic activity" (80841, 39.07% of the total).

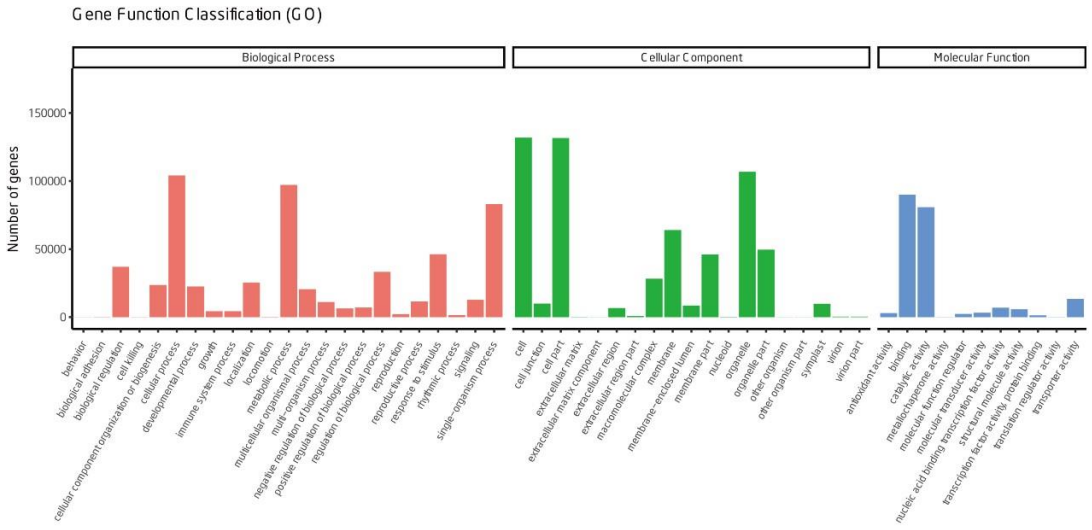

Figure 2. *Distribution of Gene Ontology (GO) terms of C. spinosa full-length transcriptome*

Further, to study the functions and interactions of isoforms in *C. spinosa*, we performed KEGG analysis. During KEGG analysis, a total of 92992 isoforms were annotated and divided into 19 functional categories (*Fig. 3*). Among these categories, the "carbohydrate metabolism" pathway contained a maximum number of isoforms accounting for a total of 12728 isoforms.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

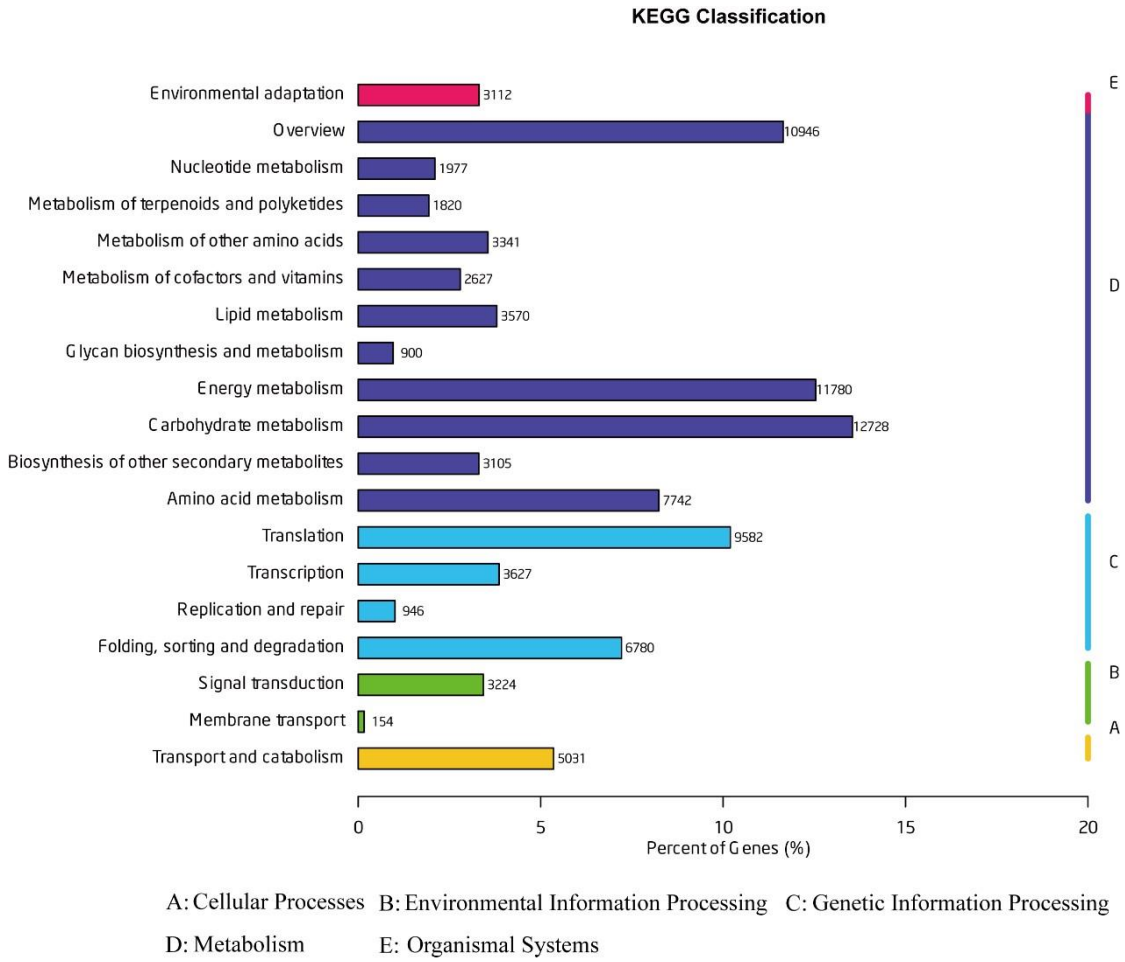

Figure 3. *KEGG pathway classification of C. spinosa full-length transcriptome*

Table 4. *Statistics of functional annotation of isoforms*

Anno_Database	Annotated_Number
GO	161075
KEGG	93966
KOG	82834
NR	182680
Swissprot	143370
All_Annotated	186840

For the prediction of transcription factor, we used full length transcripts and found that a total of 7188 transcripts were annotated as transcription factors belonging to 64 TF (transcription factor) families. AP2/ERF and bHLH were the two largest transcription factor families accounting for 10.2% and 6.6% of the genes, respectively. Similarly, other transcription factor such as C3H, bZIP, GRAS, MYB-related, C2H2, WRKY, NAC, and MYB accounted for 58.2% (each accounted for 3.7% to 8.4%) of the transcription factor genes (*Fig. 4*).

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

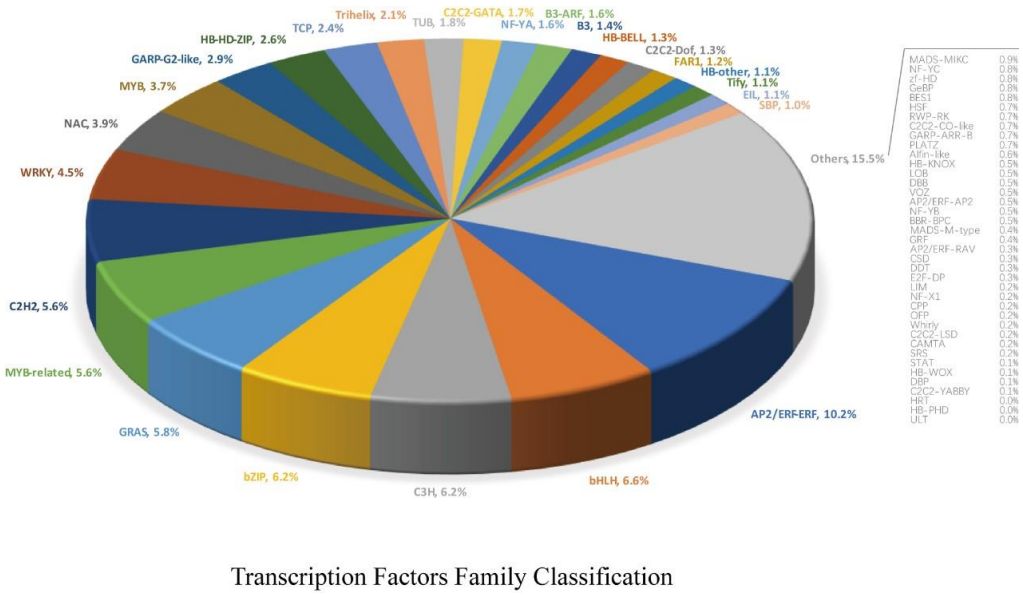

Figure 4. *Distribution of different transcription factor in C. spinosa full-length transcriptome. The number represents the proportion of transcription factors*

Quality improvement and error correction of SMRT Isoforms by RNA-Seq reads

Since SMRT sequencing produces high error rates, it is necessary to perform error correction by using high quality SGS short read corrections. We filtered raw data and obtain a total of 737 million clean reads which were further used for gene expression analysis (*Table 5*). The RNA-Seq reads were aligned to the reference isoforms with the alignment ratio greater than 89% for whole alignment. Further, we identified differentially expressed genes (DEGs) between different comparative pairs (8:00 am vs 2:00 pm, 8:00 am vs 8:00 pm, 2:00 pm vs 8:00 pm) and 434 DEGs were obtained. The numbers of DEGs in three comparative pairs were shown in *Figure 5A*. In 8:00 am vs 8:00 pm and 2:00 pm vs 8:00 pm comparative pairs, more DEGs were up-regulated (10 and 100 genes respectively) than down-regulated (8 and 24 genes respectively), whereas, in 8:00 am vs 2:00 pm more genes were down-regulated (244 genes) than up-regulated genes (100 genes) (*Fig. 5A*). Moreover, *Figure 5A* showed the degree of overlap of DEGs between different comparative pairs. In 8:00 am vs 2:00 pm, 8:00 am vs 8:00 pm and 2:00 pm vs 8:00 pm comparative pairs, the number of DEGs were 344, 18, and 124 respectively (*Fig. 5A*).

Table 5. *Mapping rate of NGS RNA-sequencing reads to reference isoforms*

Sample	Total reads	Mapped reads (ratio)
8pm-1	101873366	92175720 (90.48%)
8pm-2	80236132	71433332 (89.03%)
8pm-3	70479896	63334948 (89.86%)
8am-1	84737406	75985496 (89.67%)
8am-2	80664968	72437594 (89.80%)
8am-3	79395082	71806346 (90.44%)
2pm-1	80060560	72265336 (90.26%)
2pm-2	72857102	65459946 (89.85%)
2pm-3	87460746	79245078 (90.61%)

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

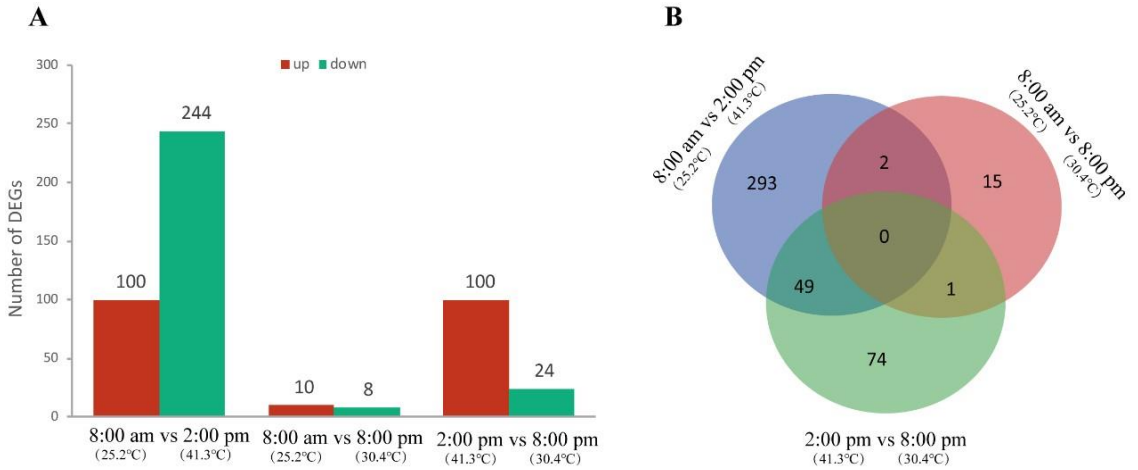

Figure 5. *DEGs between different comparative pairs (8:00 am vs 2:00 pm, 8:00 am vs 8:00 pm, 2:00 pm vs 8:00 pm). The numbers in brackets represent temperature of C. spinosa leaves. (A) The number of DEGs between different group. (B) Venn diagram showing overlapping areas in different group*

Additionally, we clustered the expression level of 434 DEGs. The results showed that almost half of the genes (190) hardly expressed in the "8:00 am" samples while expressed in the "2:00 pm" and "8:00 pm" samples. Similarly, 94 genes were expressed in the "8:00 am" samples, but hardly expressed in the "2:00 pm" and "8:00 pm" samples (*Fig. 6*).

GO and KEGG analysis of differentially expressed genes

We performed GO classification and enrichment analysis of 434 DEGs. GO analysis divided DEGs into three major functional categories: molecular function, cellular component, and biological process. In terms of molecular function, chlorophyll binding was most highly enriched category. Besides, DEGs were also enriched significantly in pigment binding, aldehyde dehydrogenase activity and transcription factor binding. Under cellular component category, DEGs were significantly enriched in the regulation of transcription. Most importantly, Abscisic acid and cold response process pathways were highly enriched (*Fig. 7*). Next, to explore the biological pathways, all DEGs were annotated from KEGG databases, and results indicated that most of DEGs showed enrichment in the microbial metabolism in diverse environments pathways, carbon metabolism, and circadian rhythm pathways (*Fig. 8*).

Expression pattern of transcription factors in DEGs

In all 434 DEGs, a total of 29 transcription factors were identified. These transcription factors were assigned to seven different TF families. Among these TFs, most of them around 14 members (48.27%) belonged to MYB family (*Fig. 9*; *Table 6*). However, TFs belonging to bHLH, bZIP, C2C2-CO-like, C2H2, DBB, GARP-G2-like families were also identified. Expression patterns analysis of these transcription factors shown two MYB genes C.spinosa_1-6k_c837860/f1p6/1454 and C.spinosa_1-6k_c571484/f1p41/1176 were up-regulated in "2:00 pm" and "8:00 pm" samples indicating these two MYB genes may respond to heat and drought stresses (*Fig. 9*).

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

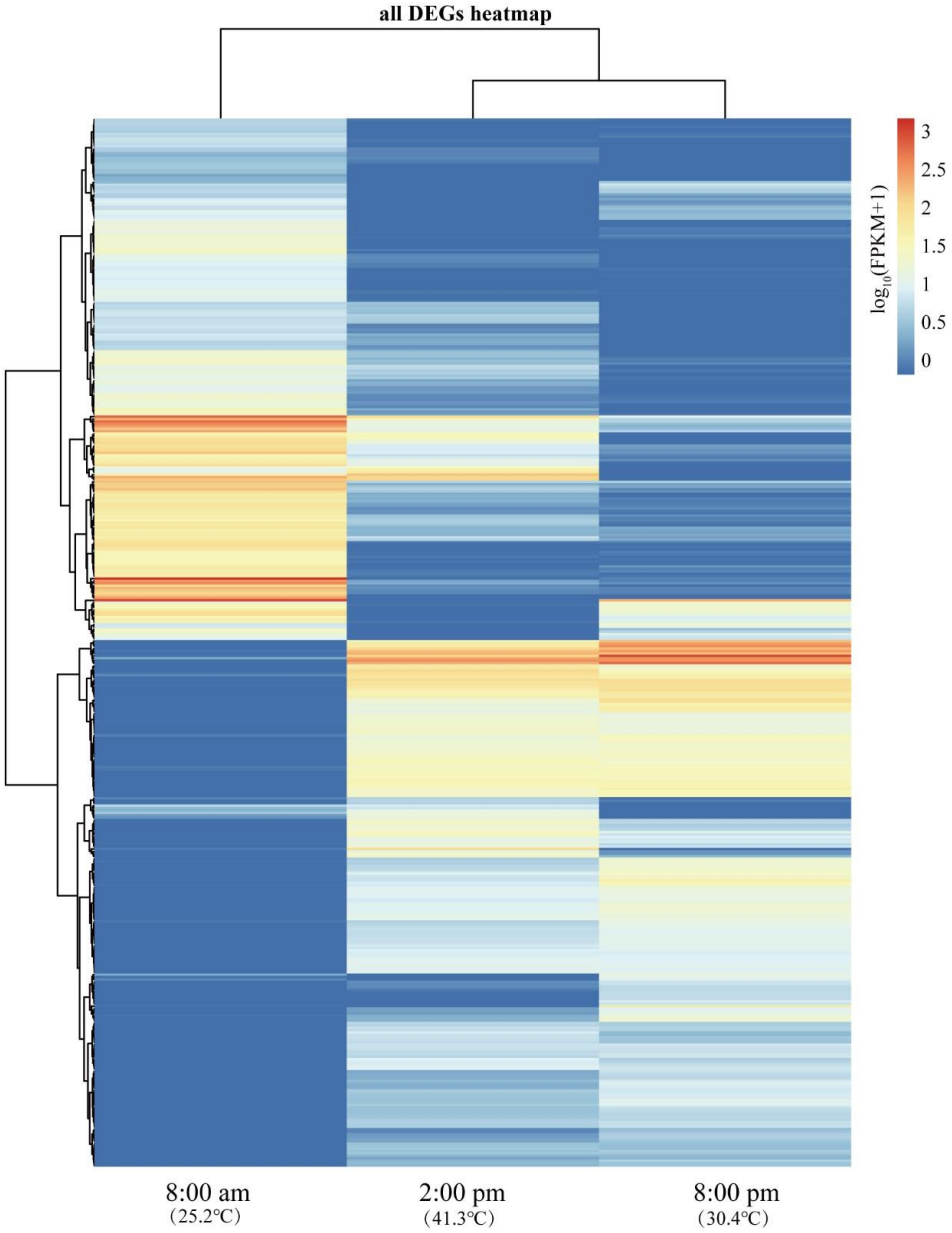

Figure 6. *The heatmap of 434 DEGs in "8:00 am", "2:00 pm" and "8:00 pm" samples. The $log_{10}FPKM + 1$ value was used to construct the heat map*

Table 6. *Statistics and classification of differentially expressed transcription factors*

Transcription factors name	Numbers
MYB->MYB-related	14
C2C2->C2C2-CO-like	5
DBB	4
C2H2	2
GARP-G2-like	2
bHLH	1
bZIP	1
Total	29

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

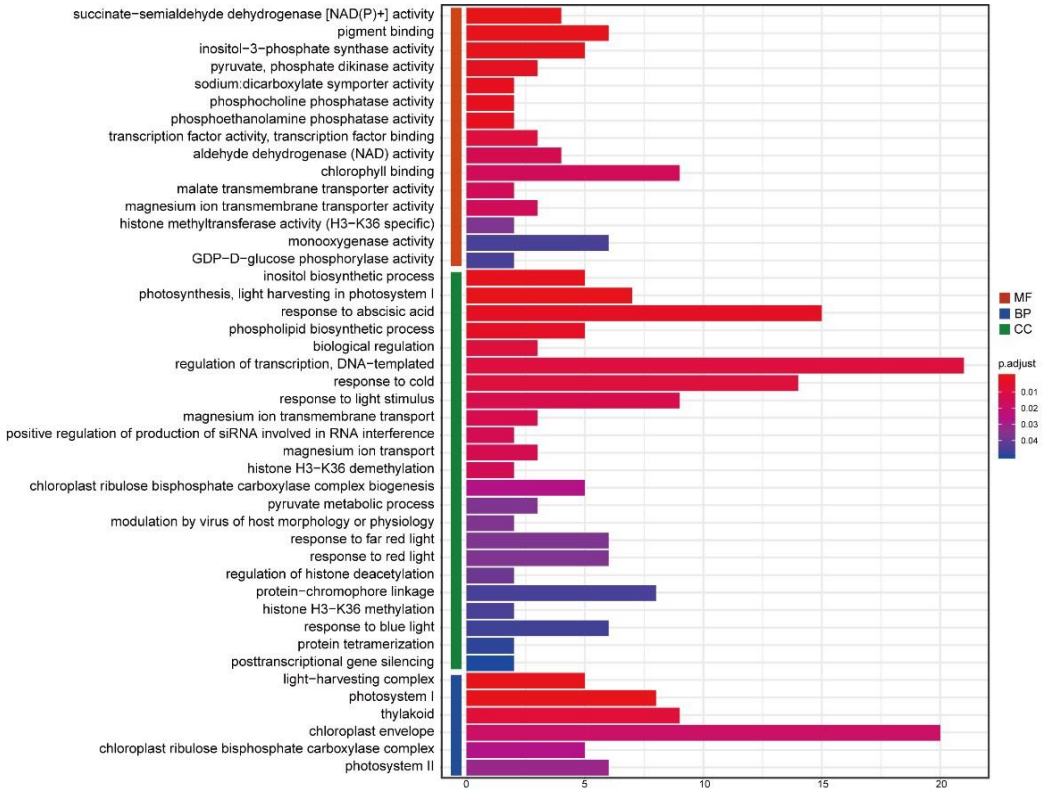

Figure 7. *GO Enrichment of DEGs. Molecular function (red color), biological process (blue color), and cellular component (green color)*

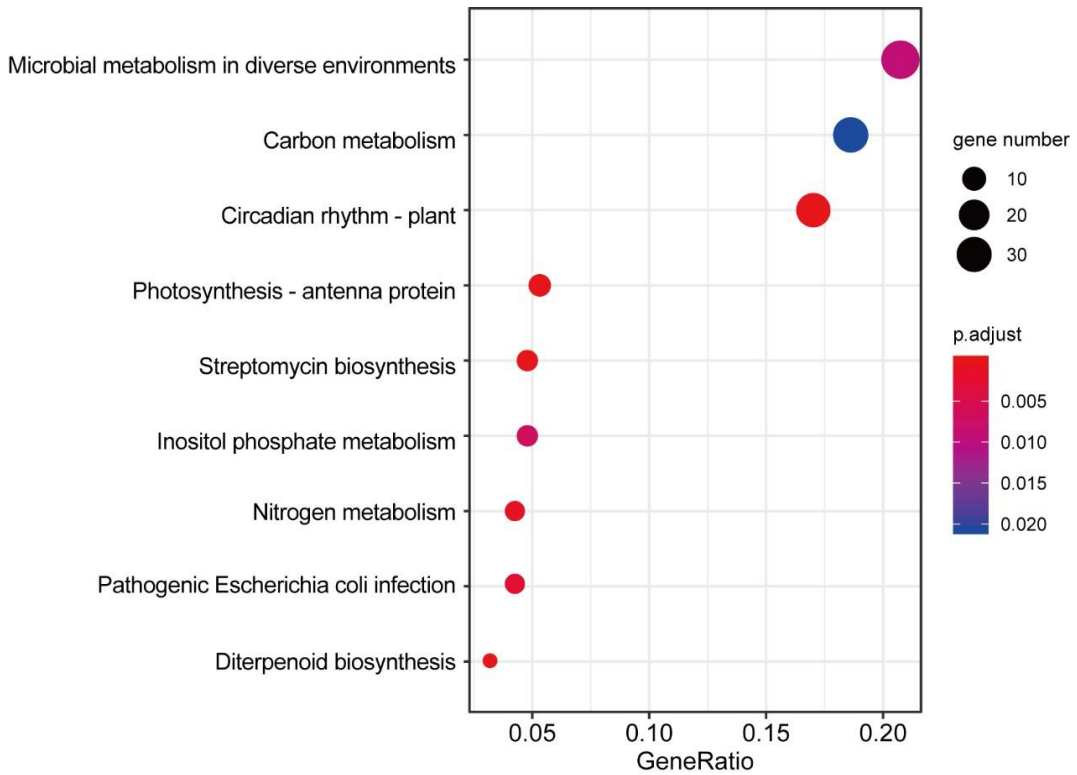

Figure 8. *KEGG pathway classification of C. spinosa DEGs*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

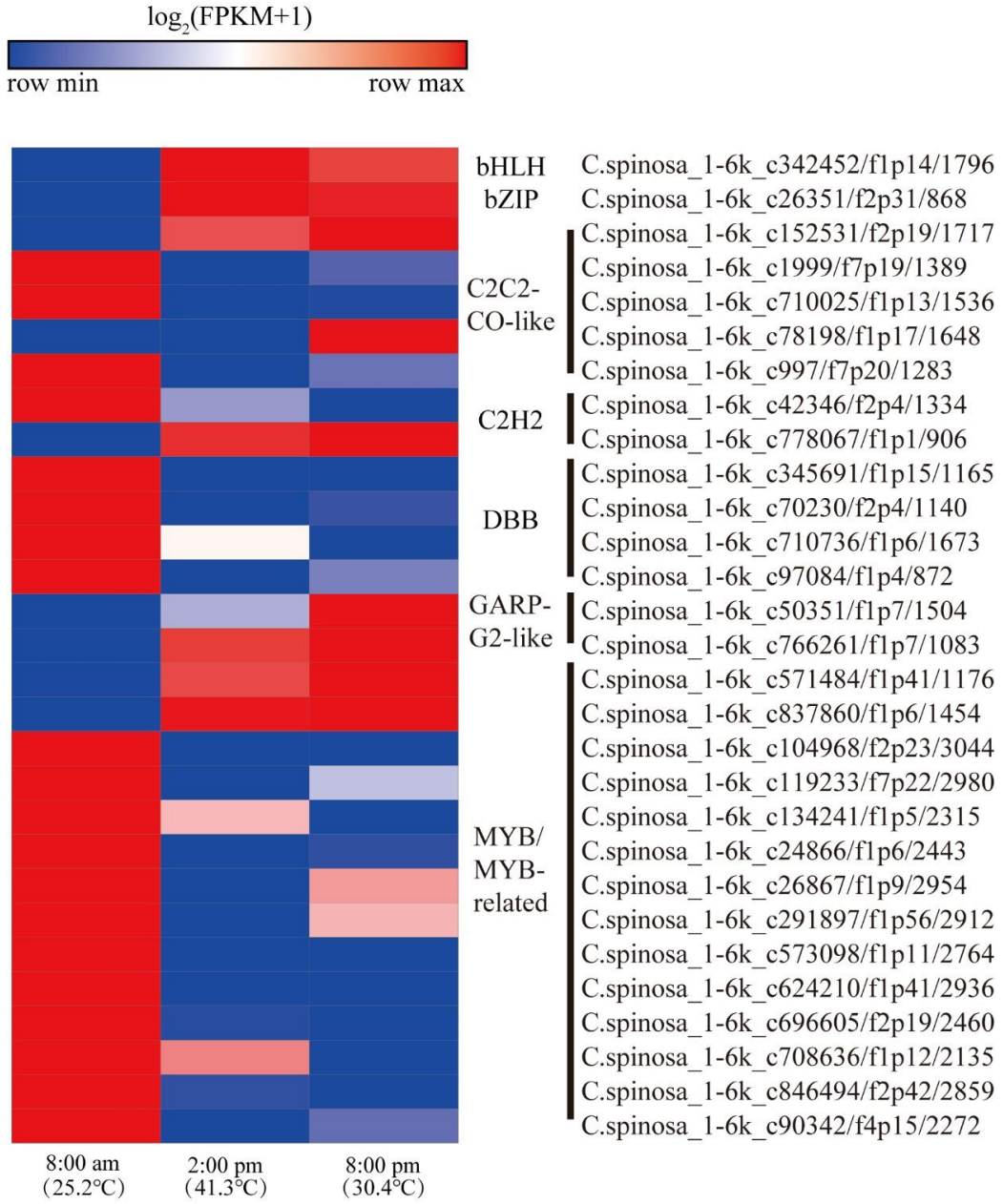

Figure 9. *Expression pattern of 29 transcription factors from DEGs in "8:00 am", "2:00 pm" and "8:00 pm" samples. The log_2FPKM + 1 value was used to construct the heat map*

Discussion

Climate change has adverse effects on biological survival and biodiversity especially in the distribution of species. Drastic environmental changes have affected the agricultural production. *C. spinosa* is an important source of secondary metabolites required by humans. Many biologically active chemical constituents were also identified from different parts (e.g. roots and seeds) of the *C. spinosa*. Therefore, some studies focused on the phytochemical and pharmacological properties of *C. spinosa* (Nabavi et al., 2016). *C. spinosa* is a shrub that has significant adaptability to harsh environments and an important source of candidate genes related to drought and heat resistance (Chedraoui et al., 2017). Nevertheless, since there was no reference genome

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

for *C. spinosa*, the progress of gene excavating and their functional study is slow. However, SMRT sequencing has been applied in many species with the advantage of obtaining full-length transcripts (Hoang et al., 2017). In this study, we provided a transcript-level reference genome for *C. spinosa* that laid the foundation for subsequent research on gene cloning and functional analysis. SMRT sequencing produced 14.96 Gb of clean data, including 828518 ROI and 370528 full-length non-chimeric reads. A total of 216240 consensus isoforms sequences were identified which included 191599 non-redundant isoforms. Previous studies demonstrated long reads sequencing ability of SMRT technology (Abdel-Ghany et al., 2016). In this study, the average size of non-redundant isoforms is 1678 bp. Besides this, Illumina-based RNA-seq is widely used in transcriptome analysis since RNA-seq has the advantages of highly accurate reads coupled with low costs (Li et al., 2017). However, the disadvantage of SGS is short reads. The SMRT sequencing can generate long reads (average 4-8 kb) without PCR amplification which usually represents full-length transcript. In this paper, we combined SMRT sequencing with SGS to obtain full-length transcripts of *C. spinosa* with high accuracy. Until now only one study was reported to clone a *CsHSP70* gene by using RACE (Luan and Dong, 2009). The *CsHSP70* gene contained a 1950 bp coding sequence (including the stop codon) encoding a 70-kDa protein with 650 amino acids. The study found that *CsHSP70* heterologous expressed in E. coli can increase tolerance to temperature stress both at 50 ℃ and 4 ℃ (Luan and Dong, 2009). We searched *CsHSP70L* (*CsHSP70*-Like) gene using *CsHSP70* protein sequence by Blastp in our protein database for *C. spinosa*. Finally, we found a sequence with highest similarity of 97.53% with *CsHSP70* (*Fig. 10B*) with the gene number C. spinosa_1-6k_c514583/f1p102/4431:160-2109(-) so we designated this gene as *CsHSP70L*. *CsHSP70L* contained a 1950 bp coding sequence (including the stop codon) encoding 650 amino acids. However, we were not found FPKM about this gene in the transcriptome data. We detected the expression of this gene by qRT-PCR. For qRT-PCR, we first used homologous alignment to identify reference genes in *caper CsUBQ6* which had high similarity with *AtUBQ6* (*Fig. 10A*). Our results show that this gene expressed highly in the "2:00 pm" samples (*Fig. 10E*). As the temperature in Xinjiang is very high at noon, high expression of *CsHSP70L* at noon indicated that it may be involved in response to heat.

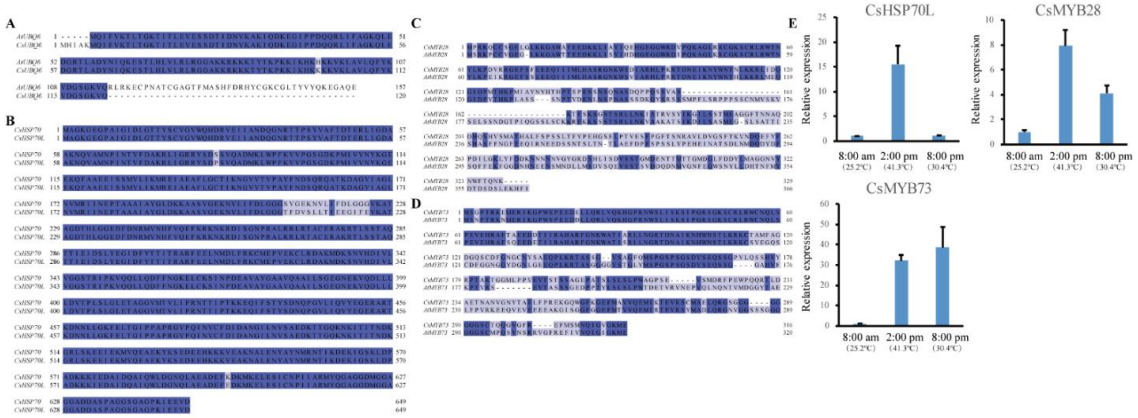

Figure 10. *Analysis of drought related genes in C. spinosa. (A) Alignment of AtUBQ6 with CsUBQ6 to confirm it as an internal control. (B) Alignment of CsHSP70 with CsHSP70L. (C) Alignment of CsMYB28 with AtMYB28. (D) Alignment of CsMYB73 with AtMYB73. (E) Relative expression pattern of CsHSP70L, CsMYB73 and CsMYB28 using qRT-PCR analysis*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

The MYB family is one of the largest transcription factors family in plants. Some studies reported that MYB participated in heat and drought stress (Yu et al., 2016; Zhao et al., 2019). The RNA-seq data show two MYB genes (C. spinosa_1-6k_c837860/f1p6/1454 and C. spinosa_1-6k_c571484/f1p41/1176) were up-regulated in "2:00 pm" and "8:00 pm" samples (*Fig. 9*). We predicted their homologous genes of in *Arabidopsis* and found that *AtMYB28* and *AtMYB73* were their putative homologous gene pairs respectively (*Fig. 10C, D*). So, we designated these genes as *CsMYB28* and *CsMYB73* respectively. We found *CsMYB28* and *CsMYB73* increased in "2:00 pm" and "8:00 pm" samples by qRT-PCR (*Fig. 10E*). This result is consistent with transcriptome data. *AtMYB28* was previously reported to regulate the biosynthesis of aliphatic glucosinolates (Sønderby et al., 2007, 2010). Recently over-expression of *AtMYB28* showed hypersensitivity to exogenous ABA during seed germination, cotyledon greening and early seedling growth, suggesting that *AtMYB28* positively involved in response to ABA (Yu et al., 2016). *AtMYB73* is a negative regulator in response to salt stress in *Arabidopsis*. While, *GhMYB73* showed increased tolerance to salt stress in transgenic *Arabidopsis* (Zhao et al., 2019). In the present research, *CsMYB28* and *CsMYB73* were up-regulated in "2:00 pm" and "8:00 pm" samples which may indicate that these two genes participated in heat and drought stresses for *C. spinosa*. However, this speculation may need further research to investigate the function of *CsMYB28* and *CsMYB73*. Our results provide the possibility to isolate and clone the gene in *C. spinosa* which will help the functional and molecular studies of *C. spinosa* in future.

Conclusions

C. spinosa is known for its medicinal properties. Meanwhile it is a candidate species to maintain and promote agricultural development in extreme climate change and extreme drought areas. *C. spinosa* is a good candidate species for addressing climate risks. This is the significant ability of *C. spinosa* to adapt different climates. In this study, complete SMRT sequencing of the full-length transcriptome of *C. spinosa* was performed first time. Further, full-length isoforms improved *C. spinosa* transcriptome annotation a lot. Additionally, the full-length transcriptome provides a more accurate depiction of gene transcription. It provides a strong factual basis for the development and utilization of *C. spinosa* genes in the future. Our results help discover related functional genes in *C. spinosa* and to lay the foundation for screening heat-tolerant genotypes. It is of great practical value to obtain more stress-resistant species by genetic engineering.

Funding. This work was sponsored by State Key Laboratory of Cotton Biology Open Fund (No. CB2020A02).

Authors' contributions. RS and GQ conceived and designed the experiments; LW collected the samples; ZL, RS, SL and GC performed the experiments and analyzed the data; ZL, RS and GQ drafted the manuscript; KZ, LW, ZS, RS and GQ revised the manuscript. All authors have read and approved the final manuscript.

Conflict of interests. The authors declare that they have no competing interests.

Availability of data and materials. The datasets generated during the current study are available from the corresponding author on reasonable request.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

REFERENCES

[1] Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016): A survey of the sorghum transcriptome using single-molecule long reads. – Nat. Commun. 7: 1-11.
[2] Alexa, A., Rahnenfuhrer, J. (2016): topGO: enrichment analysis for Gene Ontology. – R package version 2.28. 0. Cranio.
[3] Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., Courchamp, F. (2012): Impacts of climate change on the future of biodiversity: biodiversity and climate change. – Ecol. Lett. 15: 365-377.
[4] Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021): KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. – Nucleic Acids Res.
[5] Campbell, J. R. (2015): Development, global change and traditional food security in Pacific Island countries. – Reg Environ Change 15(7): 1313-1324.
[6] Chedraoui, S., Abi-Rizk, A., El-Beyrouthy, M., Chalak, L., Ouaini, N., Rajjou, L. (2017): Capparis spinosa L. in a systematic review: a xerophilous species of multi values and promising potentialities for agrosystems under the threat of global warming. – Front. Plant Sci. 8: 1845.
[7] Chen, Y., Lun, A. T., Smyth, G. K. (2014): Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. – In: Datta, S., Nettleton, D. (eds.) Statistical Analysis of Next Generation Sequencing Data. Springer, Cham, pp. 51-74.
[8] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009): Real-time DNA sequencing from single polymerase molecules. – Science 323: 133-138.
[9] Filichkin, S. A., Hamilton, M., Dharmawardhana, P. D., Singh, S. K., Sullivan, C., Ben-Hur, A., et al. (2018): Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching. – Front. Plant Sci. 9: 5.
[10] Gadgoli, C., Mishra, S. H. (1999): Antihepatotoxic activity of p-methoxy benzoic acid from Capparis spinosa. – J. Ethnopharmacol. 66: 187-192.
[11] Henckel, P. A. (1964): Physiology of plants under drought. – Annu. Rev. Plant Physiol. 15: 363-386.
[12] Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017): A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. – BMC Genomics 18: 1-22.
[13] Inocencio, C., Rivera, D., Alcaraz, F., Tomás-Barberán, F. A. (2000): Flavonoid content of commercial capers (Capparis spinosa, C. sicula and C. orientalis) produced in Mediterranean countries. – Eur. Food Res. Technol. 212: 70-74.
[14] Lagouvardos, K., Kotroni, V. (2007): TRMM and lightning observations of a low-pressure system over the Eastern Mediterranean. – Bull. Am. Meteorol. Soc. 88: 1363-1368.
[15] Langmead, B., Salzberg, S. L. (2012): Fast gapped-read alignment with Bowtie 2. – Nat. Methods 9(4): 357-359.
[16] Li, B., Dewey, C. N. (2011): RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. – BMC Bioinformatics 12(1): 1-16.
[17] Li, Q., Li, Y., Song, J., Xu, H., Xu, J., Zhu, Y., et al. (2014): High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. – New Phytol. 204: 1041-1049.
[18] Li, Y., Dai, C., Hu, C., Liu, Z., Kang, C. (2017): Global identification of alternative splicing via comparative analysis of SMRT-and Illumina-based RNA-seq in strawberry. – Plant J. 90: 164-176.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617

[19] Luan, D., Dong, Y. (2009): Cloning and its prokaryotic expression of a 70 kD heat shock protein gene from Capparis spinosa. – Acta Bot. Boreali-Occident. Sin. 29: 1291-1297.
[20] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005): Genome sequencing in microfabricated high-density picolitre reactors. – Nature 437: 376-380.
[21] Minio, A., Massonnet, M., Figueroa-Balderas, R., Vondras, A. M., Blanco-Ulate, B., Cantu, D. (2019): Iso-Seq allows genome-independent transcriptome profiling of grape berry development. – G3-Gene Genom Genet 9(3): 755-767.
[22] Nabavi, S. F., Maggi, F., Daglia, M., Habtemariam, S., Rastrelli, L., Nabavi, S. M. (2016): Pharmacological effects of Capparis spinosa L. – Phytother. Res. 30: 1733-1744.
[23] Pan, C., Wang, Y., Tao, L., Zhang, H., Deng, Q., Yang, Z., et al. (2020): Single-molecule real-time sequencing of the full-length transcriptome of loquat under low-temperature stress. – PLoS One 15(9): e0238942.
[24] Peterson, A. T., Martínez-Meyer, E., González-Salazar, C., Hall, P. W. (2004): Modeled climate change effects on distributions of Canadian butterfly species. – Can. J. Zool. 82: 851-858.
[25] Rhizopoulou, S., Psaras, G. K. (2003): Development and structure of drought-tolerant leaves of the Mediterranean shrub Capparis spinosa L. – Ann. Bot. 92: 377-383.
[26] Rhizopoulou, S., Ioannidi, E., Alexandredes, N., Argiropoulos, A. (2006): A study on functional and structural traits of the nocturnal flowers of Capparis spinosa L. – J. Arid Environ. 66: 635-647.
[27] Sharon, D., Tilgner, H., Grubert, F., Snyder, M. (2013): A single-molecule long-read survey of the human transcriptome. – Nat. Biotechnol. 31: 1009-1014.
[28] Shendure, J., Ji, H. (2008): Next-generation DNA sequencing. – Nat. Biotechnol. 26: 1135-1145.
[29] Sønderby, I. E., Hansen, B. G., Bjarnholt, N., Ticconi, C., Halkier, B. A., Kliebenstein, D. J. (2007): A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. – PLoS One 2: e1322.
[30] Sønderby, I. E., Burow, M., Rowe, H. C., Kliebenstein, D. J., Halkier, B. A. (2010): A complex interplay of three R2R3 MYB transcription factors determines the profile of aliphatic glucosinolates in Arabidopsis. – Plant Physiol. 153: 348-363.
[31] Thiry, A. A., Chavez Dulanto, P. N., Reynolds, M. P., Davies, W. J. (2016): How can we improve crop genotypes to increase stress resilience and productivity in a future climate? A new crop screening method based on productivity and resistance to abiotic stress. – J. Exp. Bot. 67: 5593-5603.
[32] Tlili, N., Elfalleh, W., Saadaoui, E., Khaldi, A., Triki, S., Nasri, N. (2011): The caper (Capparis L.): Ethnopharmacology, phytochemical and pharmacological properties. – Fitoterapia 82: 93-101.
[33] Yu, Y.-T., Wu, Z., Lu, K., Bi, C., Liang, S., Wang, X.-F., et al. (2016): Overexpression of the MYB transcription factor MYB28 or MYB99 confers hypersensitivity to abscisic acid in Arabidopsis. – J. Plant Biol. 59: 152-161.
[34] Zhao, Y., Yang, Z., Ding, Y., Liu, L., Han, X., Zhan, J., et al. (2019): Over-expression of an R2R3 MYB Gene, GhMYB73, increases tolerance to salt stress in transgenic Arabidopsis. – Plant Sci. 286: 28-36.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2001_601617