


# Transcriptome Sequencing of the Sweet Potato Progenitor (*Ipomoea Trifida* (H.B.K.) G. Don.) and Discovery of Drought Tolerance Genes

Qinghe Cao<sup>1,2</sup> · Ang Li<sup>1</sup> · Jinyang Chen<sup>1,2</sup> · Yu Sun<sup>3</sup> · Jun Tang<sup>1,2</sup> · An Zhang<sup>1,2</sup> · Zhilin Zhou<sup>1,2</sup> · Donglan Zhao<sup>1,2</sup> · Daifu Ma<sup>1</sup> · Shan Gao<sup>3</sup> 

Received: 3 December 2015 / Accepted: 23 February 2016 / Published online: 11 March 2016  
© Springer Science+Business Media New York 2016

**Abstract** *Ipomoea trifida* (H.B.K.) G. Don. is the closest wild relative of cultivated sweet potato (*I. batatas*). The diploid *I. trifida* is important for sweet potato breeding and construction of transgenic systems. It can also be used to discover functional genes, particularly stress tolerance genes which had been lost during the domestication of sweet potato. Compared to the abundant *I. batatas* transcriptome data, the nucleotide sequences of diploid *I. trifida* are rare. Using high-throughput Illumina RNA-seq technology, a total of 66,329,578 paired-end 101 bp reads were sequenced and *de novo* assembled to produce 90,684 *I. trifida* transcripts. Based on sequence similarity searches, 69,540, 39,236, 19,768, 2848 and 2766 transcripts were annotated by their homologous proteins from NCBI NR database, GO terms, KEGG pathways, known transcription factors and protein kinases, respectively. The *I. trifida* transcriptome has a medium

heterozygous rate of 0.04 %. The difference between *I. trifida* and *I. batatas* transcriptome is out of expectation. In this study, we first reported a comprehensive transcriptome for the diploid *I. trifida*, which contained gene expression information in root, leaf, stem and flower tissues. The *I. trifida* transcript sequences enriched the gene resources for sweet potato molecular research and breeding. In addition, we demonstrated that these sequences could be used to design SSR markers and clone functional genes. Particularly, we cloned a potential drought tolerance gene ItWRKY1 from *I. trifida* and validated its function using *Agrobacterium*-mediated tobacco transformants.

**Keywords** *Ipomoea trifida* · Transcriptome · RNA-seq · Drought tolerance · Transgene

Communicated by: Desiree M. Hautea

Qinghe Cao and Ang Li contributed equally to this paper.

**Electronic supplementary material** The online version of this article (doi:10.1007/s12042-016-9162-7) contains supplementary material, which is available to authorized users.

✉ Qinghe Cao  
caoqinghe@jaas.ac.cn

✉ Shan Gao  
gao\_shan@mail.nankai.edu.cn

<sup>1</sup> Key Laboratory of Biology and Genetic Improvement of Sweetpotato, Ministry of Agriculture, Xuzhou, Jiangsu 221131, People's Republic of China

<sup>2</sup> Jiangsu Xuhuai Regional Xuzhou Institute of Agricultural Sciences, Xuzhou 221131, Jiangsu, People's Republic of China

<sup>3</sup> College of Life Sciences, Nankai University, Tianjin 300071, People's Republic of China

## Introduction

The genus *Ipomoea* contains 600–700 species including the cultivated sweet potato (*Ipomoea batatas* (L.) Lam.) and its wild progenitor (*Ipomoea trifida* (H.B.K.) G. Don.) (Austin and Huaman 1996). Sweet potato, ranked thirteenth globally in production value among agricultural commodities, is one of the only seven world food crops with an annual production of more than 100 million metric tons per year. Asia accounts for over 80 % of world production (most of that is in China), with Africa producing about 15 % and the rest of the world producing about 5 %. While *I. batatas* exists in the form of hexaploid ( $2n = 6 \times = 90$ ), *I. trifida* could be diploid ( $2n = 2 \times = 30$ ), tetraploid ( $2n = 4 \times = 60$ ) or hexaploid ( $2n = 6 \times = 90$ ), which are cross-compatible with each other and also with *I. batatas*. Although the botanical origin of *I. batatas* remains unclear, rDNA FISH data indicated that the diploid *I. trifida* appeared to be the closest wild relative of *I. batatas* (Srisuwan et al. 2006). Using data from noncoding chloroplast sequences,

nuclear ITS sequences and nuclear SSRs, (Roullier et al. 2013) proved *I. batatas* could have had either an autopolyploid origin from the diploid *I. trifida*, or an allopolyploid origin involving genomes of *I. trifida* and *I. triloba*.

Although high-throughput sequencing technology facilitates whole genome sequencing, the high heterozygosity and the hexaploid nature make it difficult to obtain *I. batatas* genome for further studies in the evolution or molecular functions of sweet potato. Fortunately, *de novo* assembly of transcriptome using RNA-seq technology provides a rapid and cost-effective approach to obtain the expressed gene sequences for non-model organisms. Up to date, transcriptomes of three sweet potato cultivars and two varieties have been acquired by *de novo* assembly using 454 pyrosequencing technology or Illumina sequencing technology. They are *Ipomoea batatas* (L.) Lam. var. Tanzania (Schafleitner et al. 2010), *Ipomoea batatas* (L.) Lam. cv. Guangshu87 (Wang et al. 2010b), *Ipomoea batatas* (L.) Lam. cv. Jingshu6 (Xie et al. 2012), *Ipomoea batatas* (L.) Lam. cv. Xushu18 (Tao et al. 2012) and *Ipomoea batatas* (L.) Lam. var. Georgia Jet (Firon et al. 2013). In the year 2010, hybrid assembly of 454 sequenced reads with EST sequences from the Genbank database produced the Tanzania transcriptome containing 66,418 transcripts (31,685 contigs plus 34,733 singletons), 59.17 % (39,299/66,418) of which had significant matches in the UniRef100 protein database. Using Illumina paired end sequencing, Wang et al. (2010b) assembled the Guangshu87 transcriptome containing 56,516 transcripts, 48.54 % (27,435/56,516) of which were annotated by the known proteins in the NCBI Non-Redundant Protein (NR) database. Xie et al. (2012) assembled the Jingshu6 (purple sweet potato) transcriptome containing 58,800 transcripts, 68.5 % (40,280/58,800) of which had homologous proteins in the NR database. Tao et al. (2012) reported the Xushu18 transcriptome containing 128,052 transcripts ( $\geq 100$  bp), 40.42 % (51,763/128,052) of which had significant blastx hits in the NR database. Firon et al. (2013) reported the Georgia Jet transcriptome containing 55,296 transcripts, 72.84 % (40,278/55,296) of which could match known proteins in the NR database.

Compared to the abundant *I. batatas* transcriptome data, the nucleotide sequences of *I. trifida*, particularly in the form of diploid, are rare. Until May 2015, only 2517 EST and 139 gene sequences of *I. trifida* had been found in the NCBI Nucleotide database. In this study, we sequenced and analyzed the diploid *I. trifida* transcriptome based on the following reasons. Firstly, the small genome size of the diploid *I. trifida* makes this species more amenable to genetic analysis than the hexaploid *I. batatas*, e.g. for the self-incompatibility study in *Ipomoea* (Kowyama et al. 2000). Secondly, the diploid *I. trifida* can be used to discover functional genes, particularly genes conferring resistance or tolerance to biotic and abiotic stresses, which had been possibly lost in the cultivated sweet

potato during its domestication. For example, Tokui et al. made the discovery that *I. trifida*'s resistance to root knot nematode was controlled by at least two dominant genes (Tokui et al. 1992, 1993). Another study was conducted to investigate *I. trifida*'s resistance to two races of *Meloidogyne incognita* and its resistance mechanism (Komiyama et al. 2006). Thirdly, the diploid *I. trifida* plays an important role in sweet potato breeding and construction of transgenic systems. The known examples include: 1) developing the synthetic hexaploid *I. trifida* that could be subsequently crossed with *I. batatas* (Iwanaga et al. 1991) (Freyre et al. 1991); 2) developing 4 $\times$  interspecific hybrids between *I. batatas* and *I. trifida* as storage-root initiators for wild species (Orjeda et al. 1991). 3) developing *Agrobacterium*-mediated transformation system in *I. trifida* to identify the S genes (Kakeda et al. 2009).

To the best of our knowledge, this study was the first exploration to characterize the diploid *I. trifida* through a large-scale transcriptome sequencing with data analysis. Our research objectives included: 1) to produce adequate and accurate transcript sequences of the diploid *I. trifida* with functional annotation and SSR markers for fundamental research or breeding; 2) to find some drought tolerance genes (particularly in the WRKY super family) that could be utilized in the breeding of drought-tolerance sweet potato cultivars.

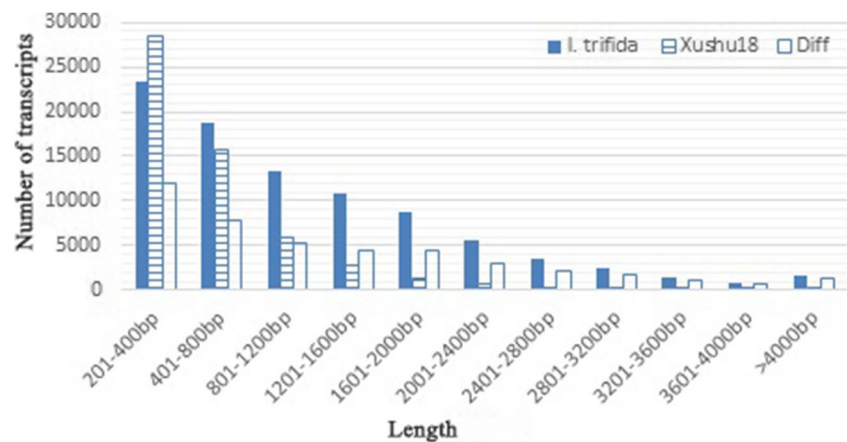
## Results

### Sequencing and *de novo* Assembly of *I. trifida* Transcriptome

Total RNA was extracted from root, leaf, stem and flower tissues and pooled together to construct one RNA-seq library, which was sequenced on the Illumina HiSeq 2000 system. After data cleaning and quality control, a total of 66,329,578 paired-end 101 bp raw reads (6.7 Gbp data) were processed to 55,363,424 cleaned reads, with the Q20 percentage of 99.3 %. These cleaned reads were *de novo* assembled into the *I. trifida* transcriptome containing 40,439 genes and 90,684 transcripts with the mean length 1178 bp and the N50 length 1784 bp, filtering out contigs shorter than 200 bp. The lengths of 90,684 *I. trifida* transcripts are distributed between 201 bp and 19,722 bp (Fig. 1). Although 23,390 transcripts with lengths less than 400 bp, we still obtained 48,480 (53.46 % of 90,684) transcripts with lengths more than 800 bp. Particularly, the *I. trifida* transcriptome has more long transcripts ( $\geq 2000$  bp) compared to the *I. batatas* Lam. cv. Xushu18 transcriptome (Tao et al. 2012), which has been thought of as the best assembled and annotated one out of all transcriptomes in sweet potato cultivars.

Using blastn, 51.66 % (46,843/90,684) of the total *I. trifida* transcripts were mapped to the best hits from the *I. batatas* cv. Xushu18 transcriptome with query coverage  $\geq 50\%$

**Fig. 1** Length distribution of *I. trifida* transcriptome. Xushu18 represents the *Ipomoea batatas* (L.) Lam. cv. Xushu18 transcriptome. Diff represents the 43,841 *I. trifida* transcripts which are covered less than 50 % of their lengths by the Xushu18 transcriptome



**(METHODS).** Among these 46,843 queries, 9475 (20.22 %) have identities above 99 % with the best hits, 55.32 % (25,912/46,843) having identities between 90 % and 99 %, 11.9 % (5572/46,843) between 80 % and 90 %, and 12.56 % (5884/46,843) below 80 %. Among 43,841 unmapped *I. trifida* transcripts, there are 31,652 transcripts with lengths more than 400 bp (Fig. 1). These results suggest these two species have unexpected difference on the transcriptome level, although *I. trifida* is the closest wild relative of *I. batatas*.

#### Identification of Heterozygous Sites and SSR Markers

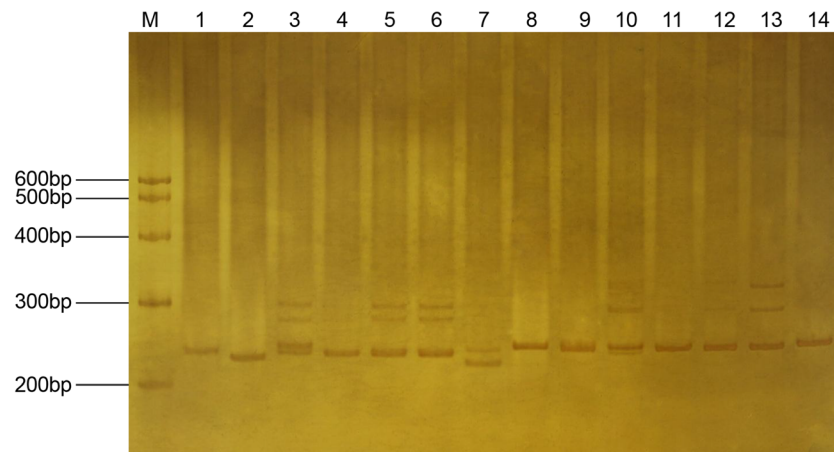
We identified 43,242 heterozygous sites, which count for 0.04 % of a total of 106,832,648 nucleotides in the *I. trifida* transcriptome (**Supplementary file 1**). This percentage is lower than those percentages in some high heterozygous species, e.g. 0.17 % in *Chrysanthemum morifolium* (Xu et al. 2013). Among 90,684 transcripts, 16,840 (18.57 %) carry at least one heterozygous site. Of these, 8135 transcripts have only one heterozygous site and 8374 transcripts have heterozygous sites between two and 10. The remaining 331 transcripts have heterozygous sites between 11 and 29.

SSR (Simple Sequence Repeat) markers are the most useful microsatellite markers to construct high-density genetics maps and identify traits of interest in plants. For further assessment of the assembly quality and development of molecular markers, 17,279 SSRs of six classes including monomer, dimer, trimer, quadmer, pentamer and hexamer were identified in 15.78 % (14,307/90,684) of the total *I. trifida* transcripts (**Supplementary file 2**). Out of 14,307 transcripts containing SSRs, 83.01 % (11,877/14,307) have only one SSR and 13.87 % (1984/14,307) have two SSRs. As for the remaining transcripts, 372, 58, 12, 2 and 2 of them have 3, 4, 5, 6 and 7 SSRs, respectively. The number of mono-, di-, tri-, quad-, penta- and hexa-nucleotide SSRs is 7095 (41.06 % of 17,279), 5546 (32.1 %), 4342 (25.13 %), 228 (1.32 %), 39 (0.23 %) and 29 (0.17 %). The most abundant repeat type for monomer, dimer, trimer, quadmer, pentamer and hexamer

is A<sup>10</sup> (1476, 8.54 % of 17,279), AG<sup>6</sup> (730, 4.22 %), CTT<sup>5</sup> (405, 2.34 %), AAAG<sup>5</sup> (29), GATTT<sup>5</sup> (3) and CTTTTT<sup>5</sup> (5). Using the transcriptome sequence information, 160 SSRs were randomly selected for validation. The polymorphism of these SSRs was detected in seven accessions of different ploidy levels in the genus *Ipomoea*. They were *I. tenussima* (2×), *I. ochracea* (2×), *I. trifida* (6×), *I. triloba* (2×), *I. batatas* cv. Nancy hall (6×), *I. batatas* cv. Xushu18 (6×) and *I. trifida* (2×). Using 160 designed pairs of primers (**Supplementary file 3**), 82.5 % (132/160) of SSRs were successfully PCR-amplified and showed polymorphism in the seven *Ipomoea* accessions. For example, Electrophoresis of PCR products using the primers of p92\_1 and p105\_1 are shown in Fig. 2.

#### Annotation of *I. trifida* Transcriptome

Using blastx (Ye et al. 2006), 69,540 *I. trifida* transcripts were annotated by the functional description of their top 20 similar sequences (hits), if they existed, from the NR database (**Supplementary file 4**). As far as we know, plant transcriptome without the genome information could be annotated with a percentage varying greatly from 13.03 % (Wang et al. 2010a) to 83.8 % (Ness et al. 2011). Thus, this functional annotation of the *I. trifida* transcriptome with a comparatively high percentage 76.68 % (69,540/90,684) provides abundant information for further analysis (Fig. 3a). The distribution of query coverage shows 37.01 % (25,737/69,540) of the mapped transcripts covered by the best hits with percentages above 80 %, while 35.47 % (24,668/69,540), 18.11 % (12,594/69,540) and 9.4 % (6541/69,540) have percentages ranging from 60 % to 80 %, from 40 % to 60 % and smaller than 40 %, respectively (Fig. 3b). The average similarity distribution of the top hits shows that 48.84 % (33,962/69,540) of the mapped transcripts have more than 80 % similarity with the known sequences, 42.6 % (29,621/69,540) of the mapped transcripts having similarity values between 60 % and 80 %, and 8.57 %

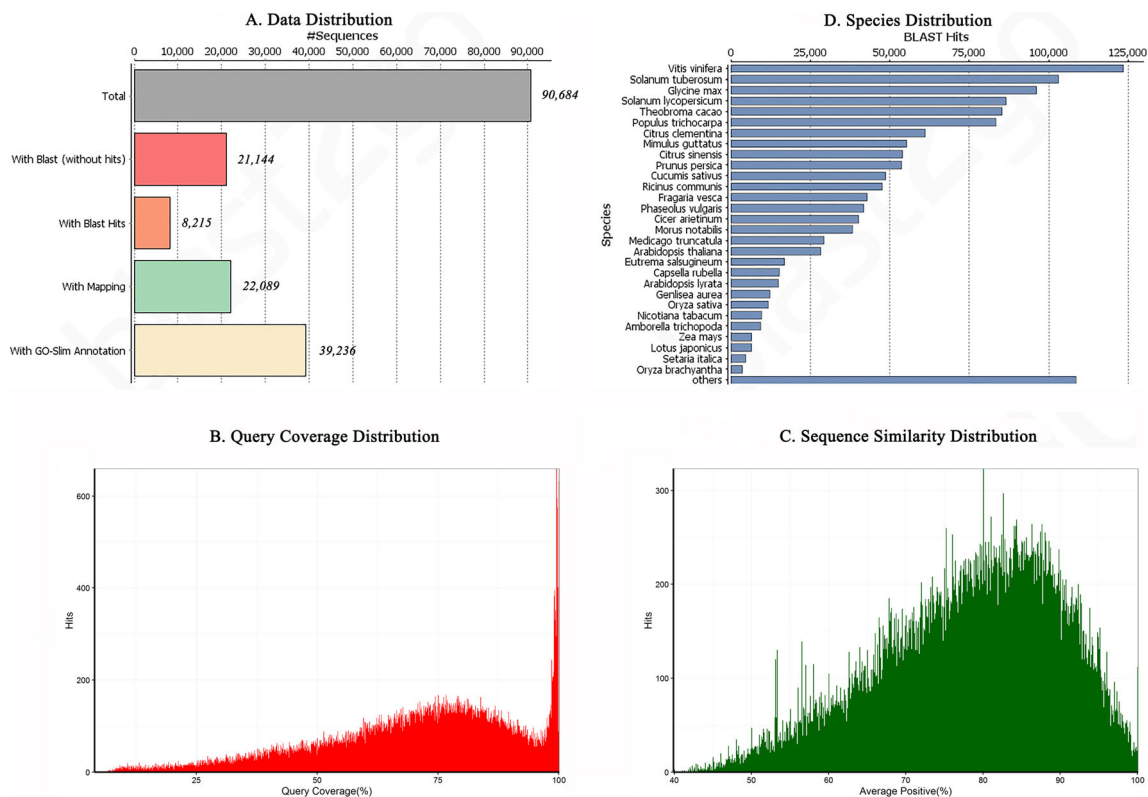


**Fig. 2** Electrophoresis of PCR products amplified by SSR primers. Lane M is markers. Lane 1 and 8 is *I. tenussima* (2×). Lane 2 and 9 is *I. ochracea* (2×). Lane 3 and 10 is *I. trifida* (6×), Lane 4 and 11 is *I. triloba* (2×). Lane 5 and 12 is *I. batatas* cv. Nancy hall (6×). Lane 6

and 13 is *I. batatas* cv. Xushu18 (6×). Lane 7 and 14 is *I. trifida* (2×). The primers of p92\_1 were used in lane 1–7. The primers of p105\_1 were used in lane 8–14. The primer sequences can be seen in the supplementary file 3

(5957/69,540) having similarity values between 40 % and 60 % (Fig. 3c). The top three species which contribute to most of the blastx hits are *Vitis vinifera*, *solanum tuberosum* and *Glycine max* (Fig. 3d). The high similarity between the top species in *I. trifida* and the top species in *I. batatas* confirmed the genus of this transcriptome data.

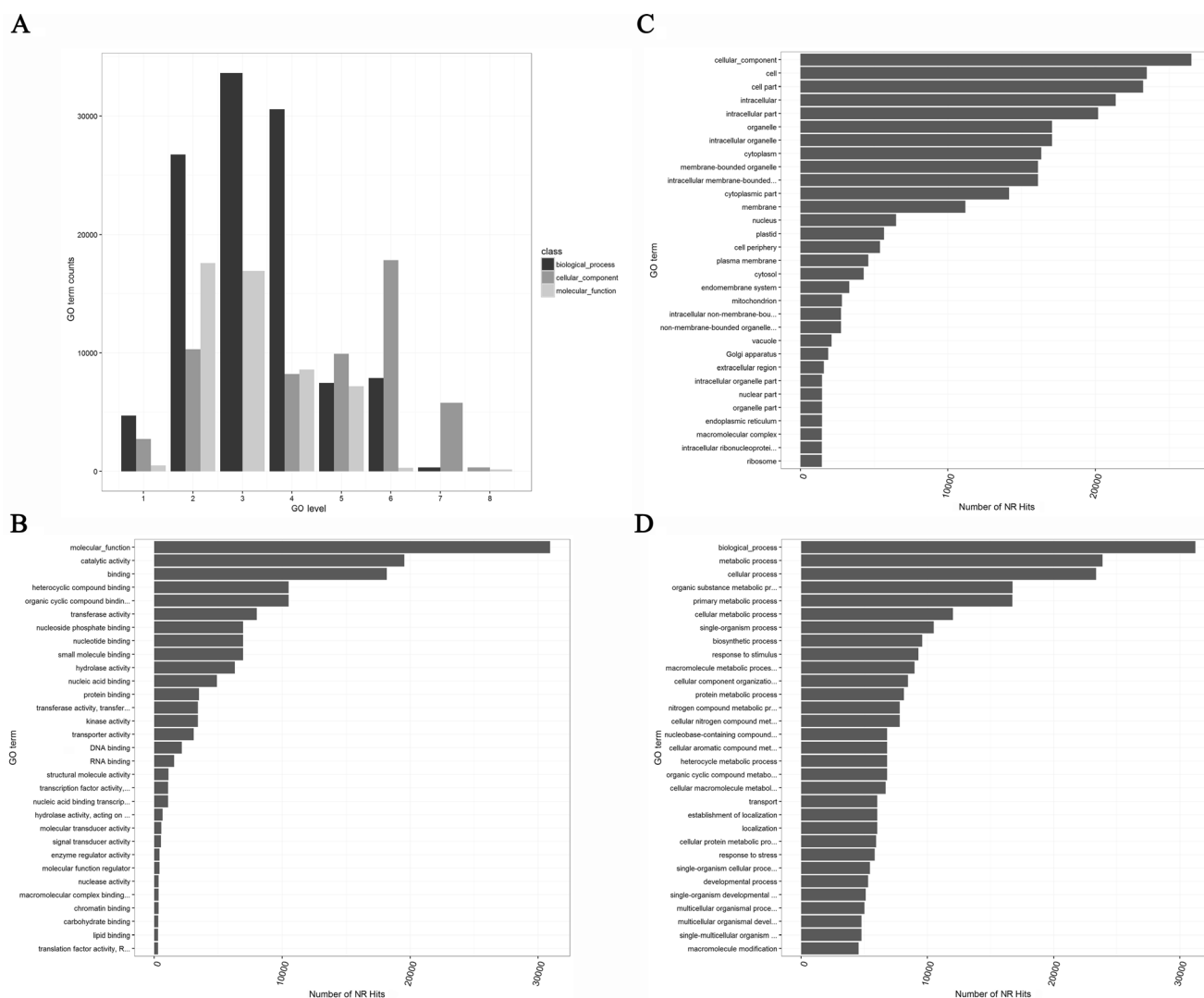
Through homologs from the NR database, 43.27 % (39,236/90,684) of the total *I. trifida* transcripts were annotated by Gene Ontology categories (GO terms) at eight levels in three GO domains which are molecular function (MF), cellular component (CC) and biological process (BP) (Fig. 4a and Supplementary file 5). The total number of transcripts re-



**Fig. 3** NR annotation of *I. trifida* transcriptome. **a** Blast hits are from the NR database. GO Slim terms are for plants. **b** Query coverage represents the percentage of query sequences covered by their best hits. **c** Average positive represents the average percentage of positive aligned amino acid,

which is equal to alignment length divided by the positive aligned amino acids between the query and the top 20 hits. **d** The top species are ranked using the number of blastx hits from the NR database

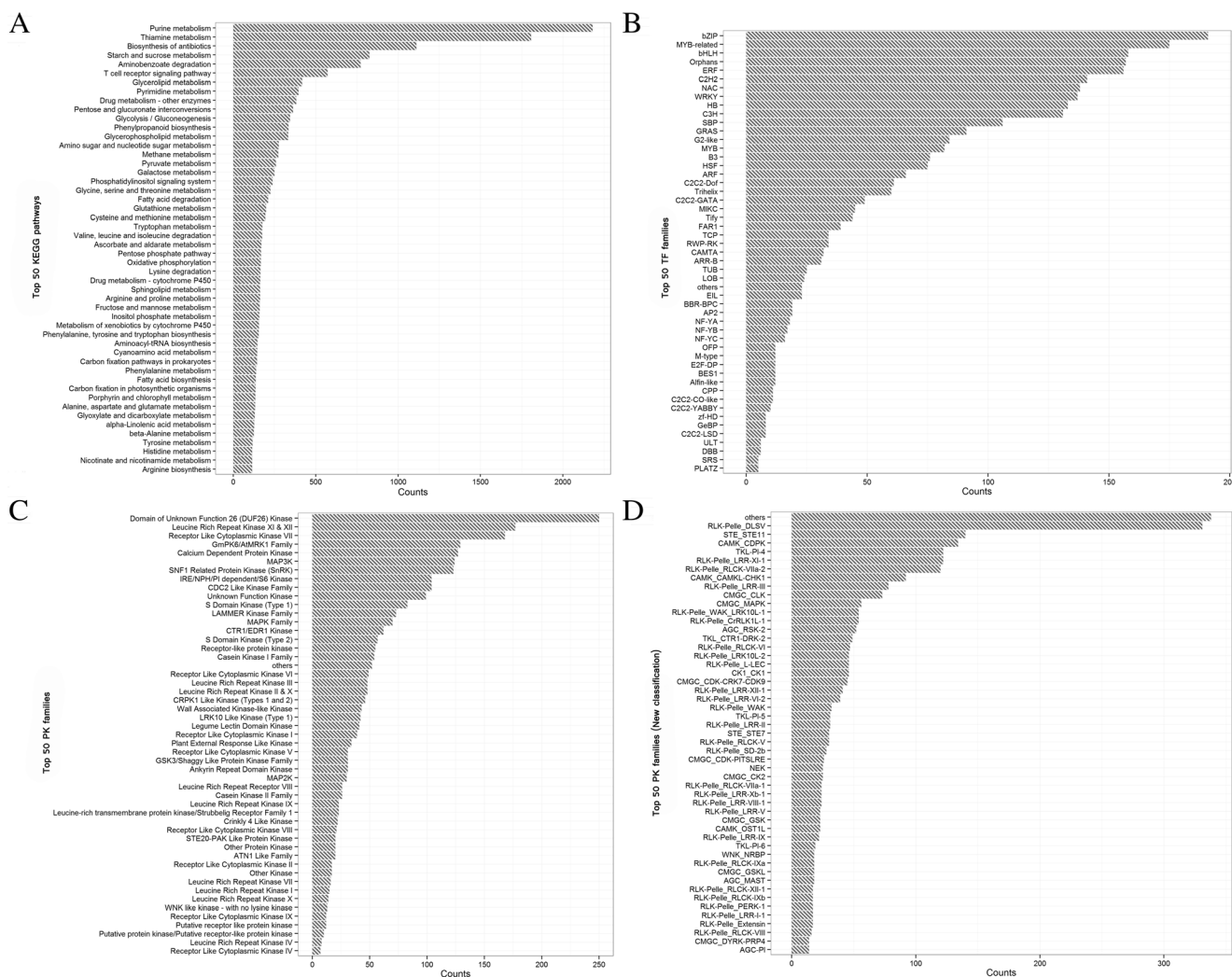




**Fig. 4** GO annotation of *I. trifida* transcriptome. **a** Level distribution of GO Slim terms assigned to 39,236 *I. trifida* transcripts. **b** Top GO Slim terms in the molecular function domain. **c** Top GO Slim terms in the cellular component domain. **d** Top GO Slim terms in the biological process domain

annotated by GO Slim terms for plants in three GO domains is 30,943, 26,510 and 31,224 for MF, CC and BP, respectively. In the MF domain, 19,553 (63.2 % of 30,943) transcripts with GO Slim annotation involves in the ‘catalytic activity’ category (Fig. 4b). The following categories are binding (18,181, 58.76 %), organic cyclic compound binding (10,497, 33.92 %), heterocyclic compound binding (10,497, 33.92 %), and so on. The highly represented categories in the CC domain are cell (23,484, 88.59 % of 26,510), cell part (23,242, 87.67 %) and intracellular (21,386, 80.67 %) (Fig. 4c). In the BP domain, they are metabolic process (23,870, 76.45 % of 31,224), cellular process (23,361, 74.82 %) and organic substance metabolic process (16,751, 53.65 %) (Fig. 4d). It is worth mentioning that “response to stress” represents 5833 transcripts with GO Slim annotation in the BP domain, which could provide useful information for discovery of the drought tolerance genes in *I. trifida*.

By mapping GO Slim terms to enzyme codes and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2004), 19,768 (21.8 % of 90,684) transcripts were assigned to 142 KEGG pathways, which represent the biosynthesis, degradation and metabolism in the cell (Supplementary file 6). The most abundant group containing 2181 (11.03 % of 19,768) transcripts is related to purine metabolism (Fig. 5a). The following is 9.14 % (1807/19,768) related to thiamine metabolism, 5.63 % (1112/19,768) related to biosynthesis of antibiotics, 4.18 % (827/19,768) related to starch and sucrose metabolism, 3.91 % (773/19,768) related to aminobenzoate degradation, 2.89 % (572/19,768) related to T cell receptor signaling pathway, and 2.11 % (418/19,768) related to glycerolipid metabolism. As we expected, the wild *I. trifida* has abundant genes involved in the biosynthesis of antibiotics and starch and sucrose metabolism.



**Fig. 5** Pathways, TFs and PKs in *I. trifida* transcriptome. **a** Top KEGG pathways. **b** Top 50 largest transcription factor families. **c** Top 50 largest protein kinase families. **d** Top 50 largest protein kinase families (new classification system)

Transcription factors (TFs) and protein kinases (PKs) play important roles in plant responses to abiotic and biotic stresses. Among 90,684 *I. trifida* transcripts, 2848 (3.14 % of 90,684) transcripts were identified as TFs and classified into 51 different families (Fig. 5b). The largest family of TFs is bZIP (191, 6.71 % of 2848), followed by MYB-related (175, 6.14 %), bHLH (158, 5.55 %), Orphans (157, 5.51 %), ERF (156, 5.48 %), C2H2 (141, 4.95 %), NAC (138, 4.85 %) and WRKY (137, 4.81 %). To our surprise, the Orphans and ERF family with such a large number of members had never been reported in sweet potato. ERF was initially discovered in tobacco and proved to increase tolerances to salt, drought and diseases in transgenic tobacco (Zhang et al. 2009). WRKY is one of the important plant-specific transcription factor super families, involved in the regulation of various physiological progresses including biotic and abiotic defenses, senescence and trichome development (Eulgem et al. 2000; Fan et al. 2015; Jiang and Yu 2015; Li et al. 2015; Sarris

et al. 2015; Schluttenhofer and Yuan 2015). All of the WRKY proteins contain a 60 aa (amino acid) region named WRKY domain that is highly conserved amongst family members. The WRKY cDNA was first cloned from sweet potato (*Ipomoea batatas*, SPF1), then from wild oat (*Avena fatua*, ABF1,2), parsley (*Petroselinum crispum*, PcWRKY1,2,3) and Arabidopsis (ZAP1), based on the ability of the encoded proteins to bind specifically to the DNA sequence motif (T)(T)TGAC(C/T) that is known as the W box (Eulgem et al. 2000). WRKY proteins are classified into group I, group II and group III, which contains two WRKY domains, one WRKY domain and one WRKY domain, respectively. Generally, the WRKY domains of group I and group II members have the same type of finger motif, while the WRKY domains of group III members have different types of finger motif.

Among 90,684 *I. trifida* transcripts, 2766 (3.05 % of 90,684) were identified as PKs belonging to 51 different families

(Fig. 5c). These 2766 PKs were also classified into 51 families using a new classification system (Lehti-Shiu and Shiu 2012) (Fig. 5d). In Fig. 5c, the largest family of PKs is the Domain of Unknown Function 26 (DUF26) Kinase (250, 9.04 % of 2766), followed by Leucine Rich Repeat Kinase XI & XII (177, 6.4 %), Receptor Like Cytoplasmic Kinase VII (168, 6.07 %) and GmpK6/AtMRK1 Family (129, 4.66 %). Using the new classification system, the top largest PK families are RLK-Pelle\_DLSV (331, 11.97 %), STE\_STE11 (140, 5.06 %), CAMK\_CDPK (134, 4.84 %), RLK-Pelle\_LRR-XI-1 (122, 4.41 %), TKL-PI-4 (122, 4.41 %), RLK-Pelle\_RLCK-VIIa-2 (120, 4.34 %) and so on. Plant RLK/Pelle family members play diverse roles in development, regulating cell-type specificity and organ identity, as well as in defense response and, to a lesser extent, in abiotic stress response (Lehti-Shiu and Shiu 2012). Plant RLK/Pelle family had undergone the most significant degree of expansion among protein kinase families in all land plant lineages. It was thought that their expansion and subsequent functional divergence could have allowed plants to perceive or respond to various environmental signals.

### ItWRKY1 Gene Cloning from *I. trifida*

Based on the *I. trifida* transcriptome annotation, we found two longest WRKY transcripts, which could be mapped to one transcript in the Xushu18 transcriptome with identity above 97 % (Supplementary file 8). These two *I. trifida* transcripts have a 1653 bp CDS region coding a 550 aa (amino acid) protein, which has an identity of 98.36 % (541/550) with the protein SPF1 (Uniprot: Q40090). This 550 aa protein named ItWRKY1 has two identical WRKY domains to those in the protein SPF1. So the ItWRKY1 protein belongs to the WRKY group I.

We performed RT-PCR to amplify a 686 bp region of the ItWRKY1 mRNA using total RNA extracted from fibrous root, tuberous root, stem and leaf tissues, respectively (Fig. 6a). The results showed the ItWRKY1 gene was well expressed in the tuberous root and poorly expressed in the fibrous root (Fig. 6b). Then, we confirmed the existence of the ItWRKY1 gene at the DNA level by PCR-amplification of the 1653 bp ItWRKY1 CDS (Fig. 6c). Finally, the RT-PCR product of the 1653 bp ItWRKY1 CDS using tuberous root RNA was inserted into the pEASY-T1 vector for gene cloning. The Sanger sequence of positive clones was consistent with the assembled ItWRKY1 CDS sequence from the *I. trifida* transcriptome.

### ItWRKY1 Response to Drought Stress in Transgenic Tobacco

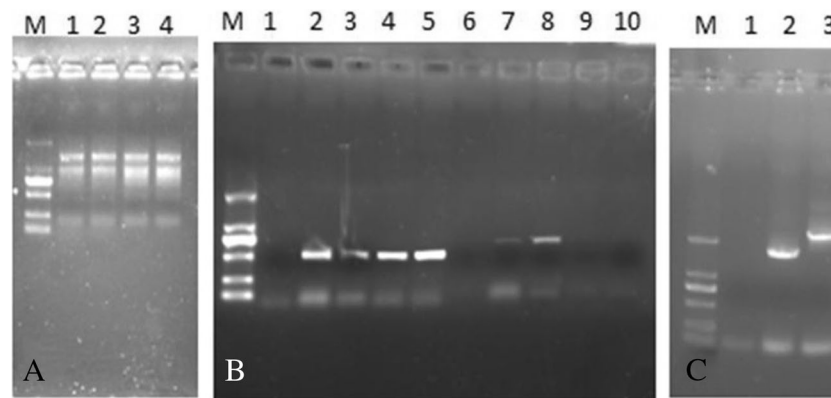
To investigate ItWRKY1's response to drought stress, we developed an experimental platform for *Agrobacterium*-mediated transformation of tobacco. The 1653 bp ItWRKY1 CDS was

transferred from the pEASY-T1 plasmid to the pCAMBIA1301 plasmid. The recombinant pCAMBIA1301 plasmid was transferred into *Agrobacterium tumefaciens*, which infected tobacco to acquire seven transgenic tobacco lines (METHODS). PCR amplification of DNA showed these seven lines had the ItWRKY1 CDS integrated into their genome (Fig. 7a). Then, we performed RT-PCR to amplify the ItWRKY1 CDS using total RNA from the seven transgenic lines (Fig. 7b). Using the Actin gene as control (Fig. 7c), we found seven transgenic lines expressed the ItWRKY1 CDS except No. 7 line (Fig. 7d). To simulate drought stress, we used 30 mL PEG solution (20 %) to irrigate seven transgenic lines and one non-transgenic line every day for 10 days. We also used 30 mL water to irrigate another non-transgenic line as control. The growth of seven transgenic and two non-transgenic tobacco lines were recorded on the 1st day (not processed) and from the 2nd to the 11th day (PEG processed). In general, the seven transgenic lines showed significantly different phenotypes (Supplementary file 9) depending on the ItWRKY1 expression level. For example, No.6 tobacco which had the highest ItWRKY1 expression level grew well during 10 days of exposure to PEG (Fig. 7e), while No.7 tobacco that did not show any expression of ItWRKY1 clearly began to wilt with signs observed from the 4th day (Fig. 7f). These results suggest ItWRKY1 CDS could be involved in the drought tolerance.

### Discussion

In this study, we performed a large-scale transcriptome sequencing of the dipliod *Ipomoea trifida* using high-throughput Illumina RNA-seq technology. A total of 66,329,578 paired-end 101 bp reads were cleaned and *de novo* assembled to produce 90,684 *I. trifida* transcripts. Based on sequence similarity searching, the *I. trifida* transcriptome was annotated by their homologous proteins from the NR database, GO terms, KEGG pathways, known transcription factors and protein kinases, respectively. The *I. trifida* transcriptome has a medium heterozygous rate of 0.04 %. In addition, it has unexpected difference from the Xushu18 (*I. batatas*) transcriptome. Using the transcriptome sequence information, six classes of 17,279 SSRs were identified in the diploid *I. trifida* and 160 SSRs were randomly selected for validation. Among them, 132 SSRs were successfully PCR-amplified and the PCR products demonstrated polymorphism in seven *Ipomoea* accessions.

Using the *I. trifida* transcriptome annotation, we found two transcripts which have a 1653 bp CDS region coding a 550 aa protein with an 98.36 % identity with the protein SPF1 in *I. batatas*. Then, we PCR-amplified this ItWRKY1 CDS in *I. trifida* at both cDNA and DNA level. The Sanger sequence of the ItWRKY1 CDS clones was consistent with the assembled sequence from the *I. trifida* transcriptome. We used



**Fig. 6** ItWRKY1 gene cloning from *I. trifida*. **a** Lane M is DL2000 makers. Lane 1–4 is total RNA from fibrous root, tuberous root, stem and leaf tissues. **b** Lane M is DL2000 markers. Lane 1–5 is RT-PCR product of Actin (516 bp) from water (negative control), fibrous root RNA, tuberous root RNA, stem RNA and leaf RNA. Lane 6–10 is RT-

PCR product of ItWRKY1 (686 bp) from water (negative control), fibrous root RNA, tuberous root RNA, stem RNA and leaf RNA. **c** Lane M is DL2000 markers. Lane 1–3 is RT-PCR product of ItWRKY1 (1669 bp) from water (negative control), tuberous root RNA and PCR product of ItWRKY1 (more than 2000 bp) from tuberous root DNA

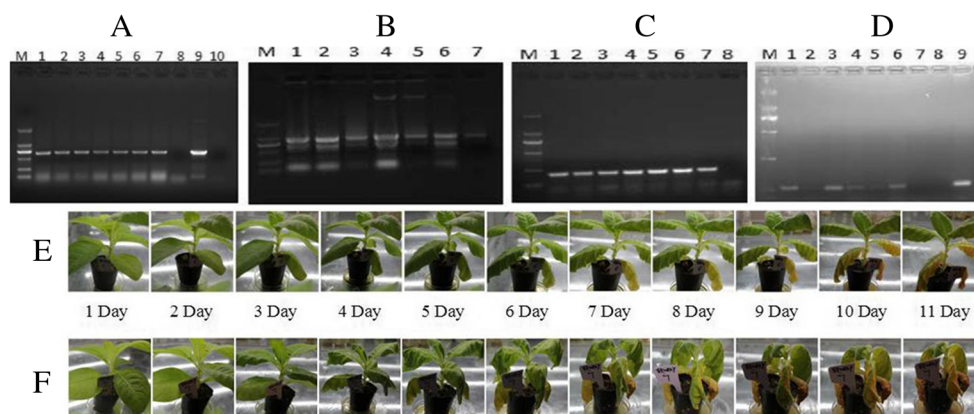
*Agrobacterium*-mediated tobacco transformants to investigate the response of ItWRKY1 to drought stress. The results suggest the 1653 bp ItWRKY1 CDS could be involved in the drought tolerance.

## Methods

### Sample Preparation, RNA-Seq Library Construction and Sequencing

Total RNA was extracted from four tissues (root, leaf, stem and flower) of the diploid *I. trifida* (DLP 4597), which had been planted in the field of the Chinese Sweet Potato Research

Institute in April 2012 and collected in September 2012. The fresh tissues were immediately frozen with liquid nitrogen and preserved at  $-80^{\circ}$  for further processing. The total RNA of each tissue was extracted separately with the Invitrogen™ Trizol Reagent (Thermo Fisher Scientific, Waltham, MA, USA). After removing the residual genomic DNA by RNA-free DNase I (New England Biolabs, Ipswich, MA, USA), the RNA quality was measured with Nanodrop. The qualified RNA ( $OD_{260}/OD_{280} = 1.6$ ) of four tissues was mixed to construct one RNA-seq library by the following main steps. The mRNA was isolated using oligo(dT) beads and sheared into short fragments. Taking these short fragments as templates, random hexamer-primers were used to synthesize the first-strand cDNA. The second-strand cDNA was synthesized



**Fig. 7** ItWRKY1 response to drought stress in transgenic tobacco. **a** Lane M is DL2000 markers. Lane 1–10 is PCR products of the 1653 ItWRKY1 CDS using genomic DNA from seven transgenic tobacco lines (No.1 to 7), one non-transgenic line (negative control), the pCambia1301 recombinant plasmid (positive control) and water (negative control). **b** Lane M is DL2000 markers. Lane 1–7 is total RNA from seven transgenic tobacco lines (No.1 to 7). **c** Lane M is DL2000 markers.

Lane 1–8 is RT-PCR products of Actin (209 bp) using total RNA from seven transgenic tobacco lines and negative control. **d** Lane M is DL2000 markers. Lane 1–9 is RT-PCR products of ItWRKY1 (110 bp) using total RNA from seven transgenic tobacco lines, negative and positive control. **e** The growth status of No.6 transgenic tobacco line. **f** The growth status of No.7 transgenic tobacco line. All the tobacco lines were irrigated by 30 mL 20 % PEG solution from the 2nd day to the 11th day



using buffer, dNTPs, RNase H and DNA polymerase I. Short fragments were purified with QIAquick PCR Purification Kit (QIAGEN, Hilden, Nordrhein-Westfalen, Germany) and resolved with EB buffer for end repair and dA tailing. After the adapter ligation and the agarose gel electrophoresis, the suitable fragments were selected for the PCR amplification. Finally, the RNA-seq library was sequenced using the Illumina HiSeq™ 2000 system in Shanghai Majorbio Bio-pharm Technology Corporation.

### RNA-Seq Data Processing and Transcriptome Analysis

Considering the fact that quality of *de novo* transcriptome assembly is highly dependent on the quality of data (Zhou et al. 2016) (Feder et al. 2015) (Holl et al. 2015) (Chen et al. 2014) (Xu et al. 2013), we used Fastq\_clean pipeline (Zhang et al. 2014) that is optimized to clean the raw reads from Illumina platforms. This pipeline trimmed low quality (<Q20) nucleotides on both ends of the raw reads and removed the trimmed reads which contain ambiguous nucleotides (“N”) more than two. Then, the pipeline trimmed the adapter and PCR primer sequences from the 3′ end of the reads by flexible string matching and produced the cleaned reads with lengths more than 25 bp. All of the cleaned reads were used to assemble the *I. trifida* transcriptome by the program Trinity (Grabherr et al. 2011). The heterozygous site calling was conducted using a Perl pipeline based on the samtools and bcftools (Li et al. 2009). One heterozygous site was required to contain at least two genotype of A, T, C or G. On each heterozygous site, the minority allele was required to have the read depth 5 and a frequency not less than 0.1.

The *I. batatas* cv. Xushu18 transcriptome sequences were acquired from Key Laboratory of Bio-resources and Eco-environment, Ministry of Education, Sichuan Key Laboratory of Molecular Biology and Biotechnology. There are 128,052 transcripts (≥100 bp) in the Xushu18 transcriptome but we only used 55,181 of them (≥200 bp) in this study. The comparison between the *I. trifida* transcriptome and the Xushu18 transcriptome was conducted using blastn. SSR identification and SSR primer design was conducted using the software Msatcommander v0.8.2 for windows (Faircloth 2008) and Primer v5.0 with default parameters.

The *I. trifida* transcripts were blasted against the NR database with an e-value threshold of  $10^{-6}$  and other default parameters. For each transcript, the top 20 hits from the NR database, if existed, were used to annotate this transcript. Functional annotation by Gene Ontology (GO) Slim terms for plants was carried out based on the top 20 hits using the Blast2GO v3.0 (Conesa et al. 2005). KEGG pathways, transcription factors and protein kinases were predicted from the *I. trifida* transcriptome using the Blast2GO v3.0 and iTAK pipeline v1.5 (<http://bioinfo.bti.cornell.edu/tool/itak>).

### ItWRKY1 Gene Cloning from *I. trifida*

Total RNA was extracted from fibrous root, tuberous root, stem and leaf tissues using TRIzol Reagent Kit (CWBIO, Beijing, CHINA), following the manufacturer’s instructions. DNAase I (Promega, Shanghai, CHINA) was used to remove the residual genomic DNA. Total RNA from different tissues was reverse-transcribed to generate first strand cDNA by the EasyScript Reverse Transcriptase Kit (TransGen Biotech, Beijing, CHINA). The RT-PCR product of the 1653 bp ItWRKY1 CDS was purified by Gel Extraction Kit (TIANGEN, Beijing, CHINA) and ligated into the pEASY-T1 vector. The recombinant pEASY-T1 plasmid was transferred into *E. coli* strain Trans1-T1. The plasmids extracted from positive clones were validated by Sanger sequencing after PCR amplification. More detailed information can be seen in the **Supplementary file 9**.

### Agrobacterium-Mediated Transformation of Tobacco

The pCAMBIA1301 vector and the recombinant pEASY-T1 vector were digested using two enzyme KpnI and XbaI. Then, the pCAMBIA1301 vector and the 1653 bp ItWRKY1 CDS were gel purified, ligated together and transferred into *E. coli* strain DH5-α. The recombinant pCAMBIA1301 vector containing the ItWRKY1 CDS was validated by colony PCR and restriction enzyme digestion. The plasmids containing recombinant pCAMBIA1301 vectors were transferred into *Agrobacterium tumefaciens* strain LBA4404.

Full developed tobacco leaves were processed by removing the upper half parts and edges. The remaining parts of leaves were cut into segments (0.5 × 1.0 cm). These leaf segments were infected by dipping into a suspension of the ItWRKY1 transformed *Agrobacterium tumefaciens* LBA4404 for 20 min. Then, the leaf segments were blotted on sterile paper towels and cultured on MS medium at temperature 22–25 °C in the dark. After three days of co-cultivation, the leaf segments were washed with sterile water and transferred to selective medium containing 300 mg/L Timentin (inhibiting *A. tumefaciens*) to be screened for four weeks. Calli from leaf segments were subcultured once every two weeks. Seedlings from calli were transferred to rooting medium. Among 10 rooted transgenic tobacco lines, seven of them grew up in the small flowerpots under standard conditions. More detailed information can be seen in the **Supplementary file 9**.

**Acknowledgments** We appreciate the help from Associate Professor Zhangjun Fei in Boyce Thompson Institute for Plant Research, Cornell University and Professor Wenjun Bu in College of Life Sciences, Nankai University. The data analysis in this study was supported by National Scientific Data Sharing Platform for Population and Health Translational Cancer Medicine Specials. This work was supported by grants from National Natural Science Foundation of China (31461143017 and 31371681), Jiangsu Natural Sciences Foundation (BK20141144), China

Agriculture Research System (CARS 11-B-02), Jiangsu Science & Technology Pillar Program (BE2014311) and Xuzhou International Science & Technology Cooperative Project (XM13B022, KC14H0141).

## Compliance with Ethical Standards

**Competing Interests** Non-financial competing interests

## References

- Austin DF, Huaman Z (1996) A synopsis of ipomoea (convolvulaceae) in the Americas. *Taxon* 45:3–38
- Chen Y-R et al. (2014) Transcriptome responses of the host trichoplusia ni to infection by the baculovirus autographa californica multiple nucleopolyhedrovirus. *J Virol* 88:13781–13797
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676
- Eulgem T, Rushton PJ, Robatzek S, Somssich IE (2000) The WRKY superfamily of plant transcription factors. *Trends Plant Sci* 5:199–206
- Faircloth B (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol Ecol Resour* 8:92–94
- Fan X et al. (2015) Transcriptome-wide identification of Salt-responsive members of the WRKY Gene Family in *Gossypium aridum*. *PLoS One* 10:e0126148
- Feder A, Burger J, Gao S, Lewinsohn E, Katzir N, Schaffer AA, Meir A, Davidovich-Rikanati R, Portnoy V, Gal-On A, Fei Z, Kashi Y, Tadmor Y (2015) A Kelch domain-containing F-box coding gene negatively regulates flavonoid accumulation in *Cucumis melo* L. *Plant Physiology*. doi:10.1104/pp.15.01008
- Firon N et al. (2013) Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC Genomics* 14:460
- Freyre R, Iwanaga M, Orjeda G (1991) Use of *Ipomoea-Trifida* (Hbk) G Don germ plasm for sweet-potato improvement .2. Fertility of synthetic hexaploids and triploids with 2n gametes of *Ipomoea-Trifida*, and their interspecific crossability with sweet-potato. *Genome* 34:209–214
- Grabherr MG et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 29:644–652
- Holl HM, Gao S, Fei Z, Andrews C, Brooks SA (2015) Generation of a de novo transcriptome from equine lamellar tissue. *BMC Genomics* 16:739
- Iwanaga M, Freyre R, Orjeda G (1991) Use of *Ipomoea-Trifida* (Hbk) G Don germ plasm for sweet-potato improvement .1. Development of synthetic hexaploids of *Ipomoea-Trifida* by ploidy-level manipulations. *Genome* 34:201–208
- Jiang Y, Yu D (2015) WRKY transcription factors: links between phytohormones and plant processes. *Science China Life Sciences* 58:501–502
- Kakeda K, Urabayashi T, Ohashi T, Oguro T, Kowayama Y (2009) Agrobacterium-mediated transformation of *ipomoea trifida*, a diploid relative of sweet potato. *Breed Sci* 59:95–98
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32:D277–D280
- Komiyama A, Sano Z, Murata T, Matsuda Y, Yoshida M, Saito A, Okada Y (2006) Resistance to two races of *Meloidogyne incognita* and resistance mechanism in diploid *ipomoea trifida*. *Breed Sci* 56:81–83
- Kowayama Y, Tsuchiya T, Kakeda K (2000) Sporophytic self-incompatibility in *ipomoea trifida*, a close relative of sweet potato. *Ann Bot-London* 85:191–196
- Lehti-Shiu MD, Shiu S-H (2012) Diversity, classification and function of the plant protein kinase superfamily. *Philos Trans R Soc Lond B Biol Sci* 367:2619–2639
- Li H et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079
- Li P et al. (2015) Chrysanthemum WRKY gene CmWRKY17 negatively regulates salt stress tolerance in transgenic chrysanthemum and Arabidopsis plants. *Plant Cell Rep* 34(8):1365–1378
- Ness RW, Siol M, Barrett SC (2011) De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. *BMC Genomics* 12:298
- Orjeda G, Freyre R, Iwanaga M (1991) Use of *ipomoea trifida* germ plasm for sweet potato improvement. 3. development of 4× interspecific hybrids between *ipomoea batatas* (L.) lam.(2n = 6× = 90) and *I. trifida* (HBK) G. Don.(2n = 2× = 30) as storage-root initiators for wild species. *Theor Appl Genet* 83:159–163
- Roullier C et al. (2013) Disentangling the origins of cultivated sweet potato (*Ipomoea batatas* (L.) Lam.). *PLoS One* 8:e62707
- Sarris PF et al. (2015) A Plant immune receptor detects pathogen effectors that Target WRKY transcription factors. *Cell* 161:1089–1100
- Schafleitner R et al. (2010) A sweetpotato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers. *BMC Genomics* 11:604
- Schluttenhofer C, Yuan L (2015) Regulation of specialized metabolism by WRKY transcription factors. *Plant Physiol* 167:295–306
- Srisuwan S, Sihachakr D, Siljak-Yakovlev S (2006) The origin and evolution of sweet potato (*Ipomoea batatas* Lam.) and its wild relatives through the cytogenetic approaches. *Plant Sci* 171(3):424–433
- Tao X, Gu YH, Wang HY, Zheng W, Li X, Zhao CW, Zhang YZ (2012) Digital gene expression analysis based on integrated de novo transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam]. *PLoS One* 0037:e36234
- Tokui M, Noro K, Nakamura M, Shiotani I, Yamamoto T (1992) Inheritance of resistance to root-knot nematode in diploid *ipomoea trifida* (H.B.K.) Don, closely related species to sweet potato. *Jpn J Breed* 42:398–399(in Japanese)
- Tokui M, Nakamura G, Takahashi E, Shiotani I (1993) Dominant genes controlling resistance to root-knot nematode in the *Ipomoea trifida* strains. *Jpn J Breed* 43:247(in Japanese)
- Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS (2010a) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* 11:400
- Wang ZY et al. (2010b) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* 11:726
- Xie F, Burklew CE, Yang Y, Liu M, Xiao P, Zhang B, Qiu D (2012) De novo sequencing and a comprehensive analysis of purple sweet potato (*Ipomoea batatas* L.) transcriptome. *Planta* 236:101–113
- Xu Y et al. (2013) Transcriptome sequencing and whole genome expression profiling of chrysanthemum under dehydration stress. *BMC Genomics* 14:662
- Ye J, McGinnis S, Madden TL (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res* 34:W6–W9
- Zhang G, Chen M, Li L, Xu Z, Chen X, Guo J, Ma Y (2009) Overexpression of the soybean GmERF3 gene, an AP2/ERF type transcription factor for increased tolerances to salt, drought, and diseases in transgenic tobacco. *J Exp Bot* 60(13):3781–3796
- Zhang M, Sun H, Fei Z, Zhan F, Gong X, Gao S (2014) Fastq\_clean: An optimized pipeline to clean the Illumina sequencing data with quality control. In: Bioinformatics and Biomedicine (BIBM), 2014 I.E. International Conference on, IEEE, pp 44–48
- Zhou D et al. (2016) De novo sequencing transcriptome of endemic gentiana straminea (gentianaceae) to identify genes involved in the biosynthesis of active ingredients. *Gene* 575:160–170