Planta (2018) 247:745–760
https://doi.org/10.1007/s00425-017-2821-6

ORIGINAL ARTICLE

Evolution of intron-poor clades and expression patterns of the glycosyltransferase family 47

Junfeng Tan[1,2] · Zhenyan Miao[1,2,3] · Chengzhi Ren[1,3] · Ruxia Yuan[1,3] · Yunjia Tang[1,4] · Xiaorong Zhang[2] · Zhaoxue Han[1,3] · Chuang Ma[1,2,3]

Received: 19 October 2017 / Accepted: 24 November 2017 / Published online: 1 December 2017

Abstract

Main conclusion **A large-scale bioinformatics analysis revealed the origin and evolution of GT47 gene family, and identified two clades of intron-poor genes with putative functions in drought stress responses and seed development in maize.**

Glycosyltransferase family 47 (GT47) genes encode β-galactosyltransferases and β-glucuronyltransferases that synthesize pectin, xyloglucans and xylan, which are important components of the plant cell wall. In this study, we performed a systematic and large-scale bioinformatics analysis of GT47 gene family using 352 GT47 proteins from 15 species ranging from cyanobacteria to seed plants. The analysis results showed that GT47 family may originate in cyanobacteria and expand along the evolutionary trajectory to moss. Further analysis of 47 GT47 genes in maize revealed that they can divide into five clades with diverse exon–intron structures. Among these five clades, two were mainly composed with intron-poor genes, which may originate in the moss. Gene duplication analysis revealed that the expansion of GT47 gene family in maize was significantly driven from tandem duplication events and segmental duplication events. Significantly, almost all duplicated genes are intron-poor genes. Expression analysis indicated that several intron-poor GT47 genes may be involved in the drought stress response and seed development in maize. This work provides insight into the origin and evolutionary process, expansion mechanisms and expression patterns of GT47 genes, thus facilitating their functional investigations in the future.

Keywords Evolution · Expansion · Origin · Stress · Seed · Transcriptional regulation

Introduction

Glycosyltransferases (GTs; EC 2.4.x.y) are a ubiquitous group of carbohydrate-active enzymes that catalyze the transfer of sugar moieties to a variety of acceptor molecules (such as sugars, lipids, proteins, and so on), forming glycosidic bonds (Lairson et al. 2008). Glycosylation of these

Junfeng Tan, Zhenyan Miao and Chengzhi Ren contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00425-017-2821-6) contains supplementary material, which is available to authorized users.

✉ Zhaoxue Han
zxhan@nwsuaf.edu.cn

✉ Chuang Ma
cma@nwafu.edu.cn; chuangma2006@gmail.com

[1] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Life Sciences, Northwest A&F University, Yangling 712100, Shaanxi, China

[2] Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Yangling 712100, Shaanxi, China

[3] Key Laboratory of Biology and Genetics Improvement of Maize in Arid Area of Northwest Region, Ministry of Agriculture, Northwest A&F University, Yangling 712100, Shaanxi, China

[4] Biomass Energy Center for Arid and Semi-Arid Lands, Northwest A&F University, Yangling 712100, Shaanxi, China

acceptor molecules often has a significant impact on their solubility, chemical stability, toxicity, or bioactivity (Weis et al. 2008). The importance of GTs has also been recognized due to their ubiquitous presence in many plant species, from the lowest plants (such as *Chlamydomonas reinhardtii* and *Physcomitrella patens*) to higher plants [such as *Arabidopsis* and *Zea mays* (maize)] (Lombard et al. 2014). On the basis of amino acid sequence similarity, GTs have been classified into more than 100 families (Lombard et al. 2014), among which structural and functional diversity are frequently reported (Wu et al. 2017; Yin et al. 2010; Taujale and Yin 2015). For instance, proteins in the GT gene family 1 (GT1) contain a PSPG (Plant Secondary Product Glycosyltransferase) motif (44-amino acid) and some of which play roles in the plant natural products biosynthesis (Jones and Vogt 2001) and the sexuality of maize (Hayward et al. 2016). Proteins in the GT8 gene family contain a Glyco_transf_8 domain (about 256 amino acids) and some of which are involved in plant cell wall biosynthesis (Yin et al. 2010). Considering the high number of subfamilies and genes, assigning biological functions to all GT genes in plant species of interests is a challenging work in plant biology.

Recently, an integrative bioinformatics analysis strategy that combines genome-wide gene identification, evolution and expression analysis has become a popular strategy to investigate the biological function of genes in a specific family. The application of this integrative strategy has enriched our knowledge of evolutionary process and biological functions of genes belonging to the largest GT family (GT1) in several plant species [maize (Li et al. 2014), peach (Wu et al. 2017), soybean (Rehman et al. 2016), *Arabidopsis* (Li et al. 2001), *Bombyx mori* (Huang et al. 2008), *Linum usitatissimum* (Barvkar et al. 2012), chickpea (Sharma et al. 2014), and other species (Caputi et al. 2012; Yu et al. 2017; Yonekura-Sakakibara and Hanada 2011)] as well as GT8 in 15 plant genomes (Yin et al. 2010) and GT43 in charophycean green algae (Taujale and Yin 2015). Different from these reported GT families (e.g., GT8 and GT43), genes in the GT47 family encode proteins containing an exostosin domain (Pfam ID: PF03016), some of which have been characterized as important genes involved in the biosynthesis of different components of the plant cell wall, such as, xyloglucan (Madson et al. 2003), and pectin (Iwai et al. 2002; Zhong and Ye 2003). A representative gene of the GT47 family in the model specie *Arabidopsis thaliana* is *IRX10*, which has been identified as a xylan xylosyltransferase (Brown et al. 2009). Inactivation of *OsIRX10*, the orthologous gene in rice to *Arabidopsis IRX10*, caused a decrease in xylan content in culm cell walls as well as an improvement of enzymatic cell wall saccharification efficiency (Chen et al. 2013). RNAi silencing of the *Arabidopsis IRX10* orthologue in wheat (*TaGT47*_2) resulted in a markedly decrease in arabinoxylan content in transgenic wheat lines (Lovegrove et al. 2013). Another representative gene of the GT47 family in *Arabidopsis* is *MUR3*, which encodes a xyloglucan galactosyltransferase belonging to a large family of type-II membrane proteins. The *mur3* mutant of *Arabidopsis* contains a severely altered structure of xyloglucans (Madson et al. 2003; Tedman-Jones et al. 2008). The galactose-depleted xyloglucan leads to dwarfism in *Arabidopsis* with curled rosette leaves, short petioles, and short inflorescence stems (Kong et al. 2015). Given the fact that the focus has been mainly on orthologous of *Arabidopsis IRX10* and *MUR3*, a systematic analysis of GT47 genes is very important for understanding the molecular mechanisms of plant cell wall synthesis. Recently, several genes in GT1 and GT8 families in *Arabidopsis* have been indicated to play roles in abiotic stress responses and seed development (Le Gall et al. 2015; Li et al. 2017; Yin et al. 2010; Rehman et al. 2016). This also raises the interesting questions of whether some GT47 genes are involved in abiotic stress responses and seed development in grasses like maize.

A recent bioinformatics analysis has identified 39 GT47 genes both in *Arabidopsis* and in *Sorghum bicolor*, and defined five clades according to the phylogenetic classification (Rai et al. 2016). However, to date, the dynamic evolution of GT47 gene family has not been investigated yet. Moreover, the evolution and putative function of the GT47 family are still unknown in maize or other grasses. Maize is one of the most important crops for both human consumption and livestock feed (Batidzirai et al. 2016), and has been selected as a model crop for basic and applied researches (Strable and Scanlon 2009). Very recently, a new version of maize B73 reference genome (version B73 RefGen v4) has been generated by the combination of next- and third-generation sequencing technologies, which has higher coverage of genome sequences (covered ~ 97% maize B73 genome), and more accurate genome sequences than previous genome (Jiao et al. 2017). This improved maize reference genome provides us a novel opportunity to perform genome-wide identification and evolution study of gene families in maize.

In this study, a comprehensive bioinformatics analysis was performed on the GT47 gene family. We identified 352 GT47 proteins from 15 species ranging from cyanobacteria to seed plants, and explored the phylogeny, exon–intron structures, expansion patterns, expression profiles, and regulatory relationships of this gene family. These analyses provided new insights into the evolution patterns of GT47 gene family, and indicated putative functions of intron-poor GT47 genes in drought stress responses and seed development.

Materials and methods

Identification of GT47 family genes in maize and other 14 species

To identify putative GT47 genes in maize, the Hidden Markov Model (HMM) profile of GT47 exostosin domain (PF03016) was downloaded from the Pfam database (http://pfam.xfam.org) (Finn et al. 2016). This HMM profile was input into hmmsearch (a program in the HMMER v3.0 software; http://hmmer.janelia.org) (Eddy 2011) for scanning all annotated protein sequences in the maize B73 reference genome (version B73 RefGen v4). The default *E* value threshold of 1.0E−3 was used to implement the hmmsearch program. The identified proteins were further screened by examining the presence of exostosin domain using NCBI's Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-Bauer et al. 2011). For multiple proteins assigned to a gene, only the one with the lowest *E* value of the hmmsearch was retained for the downstream analysis. Information regarding maize B73 reference genome sequences in FASTA format and the corresponding genome annotation in GFF3 format used in this study was downloaded from the maizeGDB database (http://www.maizegdb.org). A similar process was also performed to identify GT47 genes in *Arabidopsis* and another 13 species (Online Resource 1), the protein sequences of which were obtained from the Phytozome database (version 11.0; https://phytozome.jgi.doe.gov/pz/portal.html) (Goodstein et al. 2012).

Physicochemical features and subcellular location prediction of GT47 proteins

Physiochemical features, including isoelectric points (p*I*) and molecular weight (kDa), of each GT47 protein were calculated using the online tool Compute p*I*/Mw (http://web.expasy.org/compute_pi) (Gasteiger et al. 2005). Subcellular localizations of GT47 proteins were predicted using DeepLoc (http://www.cbs.dtu.dk/services/DeepLoc) (Almagro Armenteros et al. 2017), which uses the deep learning technology to learn sequence-based features for accurately predicting protein types (membrane or soluble) and protein subcellular localizations (nucleus, cytoplasm, extracellular, mitochondrion, cell membrane, endoplasmic reticulum, plastid, Golgi apparatus, lysosome/vacuole or peroxisome).

Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignment of all predicted GT47 proteins were performed using the MAFFT program (http://mafft.cbrc.jp/alignment/software) (Katoh and Standley 2013) with default settings (the 200 PAM log-odds matrix; gap opening penalty 2.4; gap extension penalty 0.06). A phylogenetic tree was then constructed using the FastTree software (http://www.microbesonline.org/fasttree) (Price et al. 2010), and bootstrap analysis was performed using 1000 replicates. A cyanobacterial GT47 protein from *Synechococcus* sp. *WH 7803* was used as outgroup.

Visualization of GT47 genes in maize

The exon–intron structures and corresponding genomic coordinates of maize GT47 genes were extracted from the maize B73 genome annotation file in the standard GFF3 format. Their genomic localizations were displayed using the circular visualization tool Circos (version 0.69; http://www.circos.ca) (Krzywinski et al. 2009), and their exon–intron structures were visualized using the Gene Structure Display Server (GSDS) online server (http://gsds.cbi.pku.edu.cn) (Hu et al. 2015).

Duplication analysis of GT47 genes in maize

Duplicated chromosomal segments in the maize genome were detected by using the SynMap tool available at the CoGe server (https://genomevolution.org/CoGe/SynMap.pl) (Lyons et al. 2008), which identifies the syntenic blocks (tightly linked conserved gene clusters of three or more genes) by sequence similarity from maize B73 genome sequences. Maize GT47 genes within the duplicated chromosomal segments were identified according to their genomic coordinates. Tandem duplication events associated with the GT47 gene family in maize were analyzed using the criteria described in previous work (Ouyang et al. 2007). In brief, genes were considered to be tandem duplicates, if they belonged to the same gene family, were located within a 100-kb distance, and separated by five or less genes

For each pair of segmentally duplicated genes, a phylogenetic tree was constructed using these two genes and their paralogues and orthologues in different taxa with the TreeBeST method from TreeFam software (Schreiber et al. 2013). The phylogenetic tree was automatically generated using the EnsemblPlants database (http://plants.ensembl.org). The corresponding duplication time was estimated using synonymous substitutions per synonymous site (dS)

per year as $T = dS/2\lambda$. ($\lambda = 6.5 \times 10^{-9}$) (Gaut et al. 1996). The number of dS, and the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site (d*N*/d*S*) were calculated using the codeml program implemented in phylogenetic analysis by maximum likelihood in PAMLX software (Xu and Yang 2013).

Spatio-temporal expression patterns of GT47 genes in maize

Two spatio-temporal RNA-seq datasets consisted of 158 samples that were used to characterize the expression profiles of GT47 genes in maize B73 inbred line. The first RNA-seq dataset was used to profile the expression patterns of GT47 genes in three tissues (leaf, ear, and tassel) at four developmental stages (V12, V14, V18, and R1) under two environmental conditions (well-watered and drought) of maize B73. The raw reads of this RNA-seq dataset (50nt single reads, ~ 1481 millions) were generated with Illumina HiSeq 2500 and retrieved from the Gene Expression Omnibus (GEO) database of NCBI (Accession number: GSE71723) (Thatcher et al. 2016). The second RNA-seq dataset was used to profile the expression patterns of GT47 genes during maize seed development from 0 DAP to 38 DAP, the raw reads of which (100 nt paired-end reads, ~ 1862 millions) were generated using Illumina HiSeq 2500 and downloaded from the Sequence Read Archive (SRA) database of NCBI (Accession number: SRP037559) (Chen et al. 2014).

For each RNA-seq dataset, clean reads were firstly separated from low-quality reads (mean quality score < 20; reads length < 20) by using Trimmomatic (version 0.36; http://www.usadellab.org/cms/?page=trimmomatic) (Bolger et al. 2014), and then aligned to the maize B73 reference genome (version B73 RefGen v4) using TopHat (version 2.1.1; https://ccb.jhu.edu/software/tophat/index.shtml) (Trapnell et al. 2009). By the Cufflinks program (version 2.2.1; http://cole-trapnell-lab.github.io/cufflinks) (Trapnell et al. 2012), the expression abundance of GT47 genes was finally characterized using the terms of FPKM (fragments per kilobase of transcript per million map reads) values (Chen et al. 2014; Thatcher et al. 2016; Miao et al. 2017). Based on the FPKM values, heatmaps of the expression of GT47 genes were generated, using 'pheatmap' function in R programming language (https://www.r-project.org).

Differential expression analysis was performed between well-watered and drought conditions. Genes were considered to be differentially expressed if they satisfied the following criteria: |log2 (fold-change)| ≥ 1 and FDR (false discovery rate)-adjusted *p* value ≤ 0.05.

Ribosome profiling of GT47 genes in maize

The ribosome profiling approach records the footprints of fully assembled ribosomes along mRNAs, and thus has the ability of globally measuring translation (e.g., translational efficiency) by deep sequencing of ribosome-protected mRNA fragments (Shirokikh et al. 2017; Brar and Weissman 2015). In order to systematically explore the effect of drought stress on gene expressions at transcriptional and translational levels, both RNA-seq and ribosome profiling were performed on 14-day seedlings of maize B73 inbred line under two environmental conditions (well-watered and drought-stressed), with two biological replicates each sample. Detailed information about this experiment can be found in Lei et al. (2015). The raw reads of RNA-seq and ribosome profiling were obtained from the SRA database (Accession number: SRP052520). All RNA-seq data were processed as the procedure described in previous subsection, while all ribosome profiling data were processed following the procedure described previously (Lei et al. 2015). For each gene, the FPKM was used to measure gene expression at the translational level (i.e., translational abundance), and the translational efficiency (TE) was calculated by the formula (Lei et al. 2015):

$$TE = FPKM_{(translational\ level)}/FPKM_{(transcriptional\ level)}.$$

Prediction of *cis*-acting elements in the proximal promoter of GT47 genes

The 1500 bp upstream sequences from the start codon of maize GT47 genes were regarded as proximal promoters, in which stress- and development-related *cis*-acting elements were predicted using the PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html) (Lescot et al. 2002).

Predicted regulatory network of GT47 genes in maize

The regulatory relationships between transcription factors (TFs) and GT47 genes in maize were first obtained from the Plant Transcription Factor Database v4.0 (PlantTFDB; http://planttfdb.cbi.pku.edu.cn) (Jin et al. 2017), in which TFs' target genes were annotated according to the computational prediction of TF binding sites and the 'wet' experimental results in literature. Then, a gene co-expression analysis was performed on the drought stress- or seed development-related gene expression data with the Gini correlation coefficient (GCC) algorithm, which can accurately infer regulatory relationships from transcriptome data (Ma and Wang 2012; Ma et al. 2014). The GCC algorithm was implemented using the 'rsgcc' R package (https://cran.r-project.org/web/packages/rsgcc). The significance level of GCC values was calculated using 1000 permutations for tests with a level of 0.001. The predicted regulatory relationship between a TF and a GT47 gene was retained in the final regulatory network if the corresponding GCC value satisfied

the following criteria: |GCC| ≥ 0.5; p value < 0.05. The drought stress- and seed development-related regulatory networks were visualized using Cytoscape (version 3.5.0; http://www.cytoscape.org) (Kohl et al. 2011), in which nodes represent GT47 genes or TF families, edges denote regulatory relationships between GT47 genes and TFs.

Real-time quantitative RT-PCR (qRT-PCR) of GT47 genes in maize under drought stress

Maize (*Zea mays* cv. B73) seeds were soaked in deionized water for 12 h and then placed on a sheet of moist filter paper in a Petri dish to allow germination at 28 °C for 3 days. Germinated seeds were transferred to a floating foam sheet in the hydroponic boxes (40 × 20 × 15 cm) containing continuously aerated water in the growth chamber (28 °C day/26 °C night, 16 h light/8 h dark photoperiod, 30–50% relative humidity) for about 1 week, and then the seedlings were cultivated in 1/2 Hoagland solution for approximately 1 week. To induce expression of target genes, seedlings at the three-leaf stage were subjected to simulated dehydration stress treatment. The treatment was carried out by submerging the roots of the plants in 1/2 Hoagland solution of 16% (w/v) polyethylene glycol (PEG; MW 8000) for different periods of time. After the treatment, the leaf and root tissues of the well-watered and drought-stressed seedlings were harvested as described above.

Total RNA was isolated from plants using TRIZOL reagent (Invitrogen, CA, USA). For each sample, 20 µg of total RNA was digested in a volume of 20 µL with RNase-free DNase I (TaKaRa, Dalian, China) according to the manufacturer's instructions for treatment to remove genomic DNA contamination. The first-strand cDNA synthesis was performed using 5 µg of DNase-treated total RNA with the PrimeScript™ II 1st Strand cDNA Synthesis Kit (TaKaRa, Dalian, China) in a reaction volume of 20 µL following the manufacturer's instructions.

The gene-specific primers were designed for seven candidate drought-responsive GT47 genes in maize using the Primer Premier 5 software (Online Resource 2). Each reaction contained 7.5 µL SYBR Green PCR Master Mix (TaKaRa, Dalian, China), 2 µL of 1:40 (v/v) dilution of the first strand cDNA, and 0.2 µM of each primer in a final volume of 15 µL. The qRT-PCR reactions were carried out on the QuantStudio™ 6 Flex Real-Time PCR System (The Applied Biosystems, CA, USA). The reaction procedures were as follows: 95 °C for 10 min, followed by 44 cycles of 95 °C for10 s, 58 °C for 15 s and 72 °C for 20 s. The maize *GAPDH* gene (Accession number X07156.1) was used as an internal control. Amplification specificity was verified with a heat dissociation protocol (melting curves in 65–95 °C range) in the final step of PCR. All primer pairs showed a single peak on the melting curve, and a single band of the expected size was visualized after separation by agarose gel electrophoresis. All reactions were repeated three times. The relative expression levels were calculated using the formula (Pfaffl 2001):

$$\text{Ratio} = (E_{\text{target}})^{\Delta \text{Cp}}\text{target}^{(\text{control}-\text{treatment})} / (E_{\text{ref}})^{\Delta \text{Cp}}\text{ref}^{(\text{control}-\text{treatment})}.$$

Results

Characteristic features of GT47 genes in maize, sorghum and *Arabidopsis*

A total of 47 putative GT47 genes were identified in the maize B73 reference genome (Online Resource 3). The protein sequences encoded by maize GT47 genes varied in length from 278 to 887 amino acids, and the corresponding length distribution was similar to that of GT47 proteins in *Arabidopsis* and sorghum (Fig. 1a, Online Resource 4). Most of maize GT47 proteins (72.34%; 37/47) had relatively high isoelectric points (pI > 7.0). The proportion is comparable to that in sorghum (70.27%; 27/37) but is markedly lower than that in *Arabidopsis* (87.18%; 34/39) (Fig. 1b, Online Resource 5).

For each GT47 protein, we identified whether it was membrane-integral or soluble by DeepLoc software (Almagro Armenteros et al. 2017). We observed that a small proportion of GT47 proteins were predicted to be soluble in maize (12.77%; 6/47) and in sorghum (7.69%; 3/39), while no *Arabidopsis* GT47 protein was predicted to be soluble (Fig. 1c, Online Resource 6). Subcellular localization prediction showed that GT47 proteins in maize are predominantly located in endoplasmic reticulum and Golgi apparatus. The proportion (80.85%; 38/47) is lower than that in sorghum (89.19%; 33/37) and in *Arabidopsis* (97.44%; 38/39) (Fig. 1d, Online Resource 7).

The GT47 gene family in maize can be classified into five clades with diverse exon–intron structures

To investigate the evolution and origin of GT47 genes, we performed the phylogenetic analysis using 352 identified GT47 proteins from 15 species ranging from cyanobacteria to seed plants, including one species (*Synechococcus* sp. *WH 7803*) in cyanobacteria, two species (*Galdieria sulphuraria*, *Chondrus crispus*) in red algae, three species (*Micromonas* sp. *RCC299*, *Micromonas pusilla CCMP1545*, and *Chlamydomonas reinhardtii*) in green algae, one species (*Mpolymorpha*) in moss, one species (*Selaginella moellendorffii*) in fern, one species (*Picea abies*) in gymnosperms, one species (*Amborella trichopoda*) in basal angiosperm, three species (*Brachypodium distachyon*, sorghum, and maize) in monocots, and two species (*Vitis vinifera*, *Arabidopsis*) in dicots (Fig. 2a; Online Resource 1).

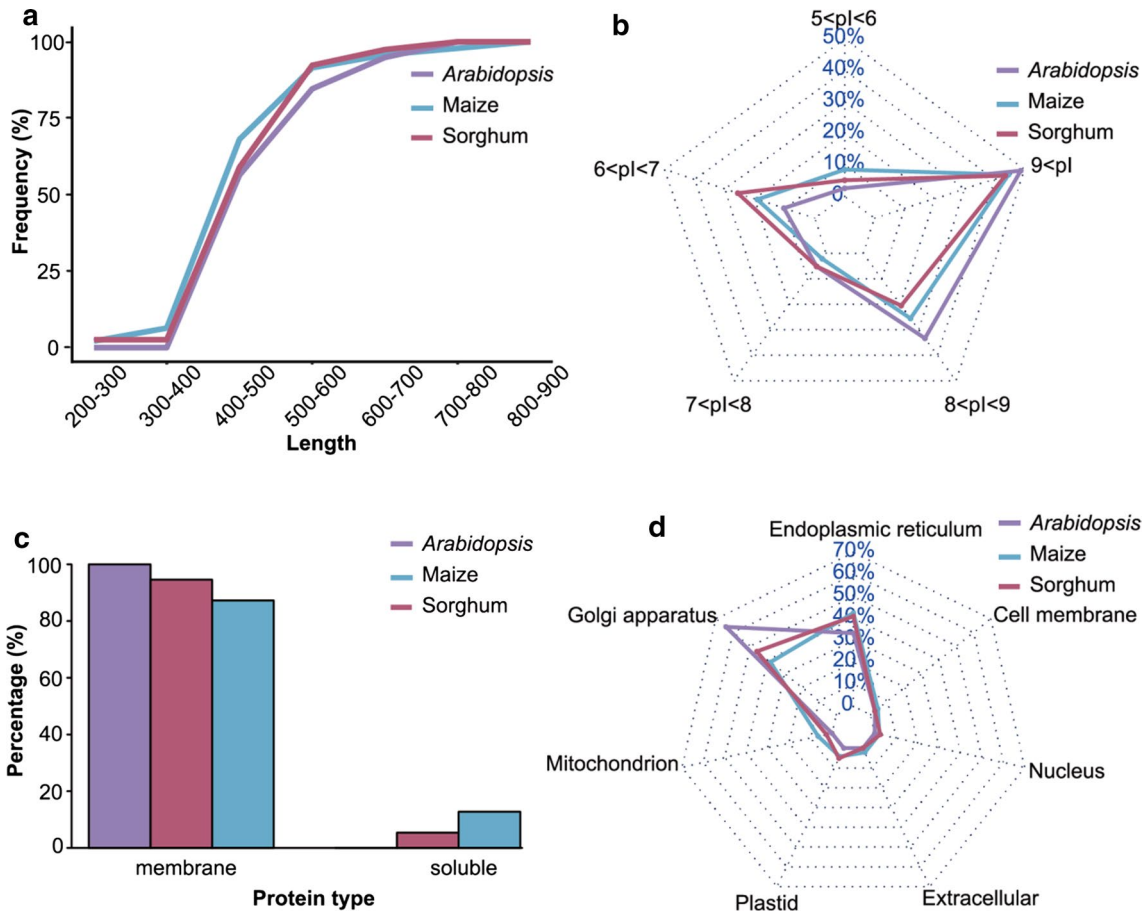

Fig. 1 Sequence characteristics of GT47 proteins in maize, sorghum, and *Arabidopsis*. **a** Cumulative length distribution of GT47 protein sequences. **b** Radar chart of the PI distribution of GT47 protein sequences. **c** Percentage of GT47 protein types. **d** Radar chart of the distribution of predicted subcellular localization of GT47 proteins

These 352 GT47 genes were divided into six clades (I, II, III, IV, V, and VI) (Fig. 2a). The clade III, IV, V, and VI contained 198 GT47 proteins from red algae (4), green algae (44), moss (13), fern (13), and seed plant (124), indicating that these four clades are ancient clades and may originate from cyanobacteria. The other two clades I and II encompasses 153 GT47 proteins from moss (6, 2), fern (10, 6), and seed plant (81, 48), suggesting that these two clades are recent clades and may form in moss plants.

The 47 maize GT47 genes belonged to the five clades I, II, III, IV, and V with diverse exon–intron structures (Fig. 2b, Online Resource 8). The clade I contained 17 maize GT47 genes, all of which were intron-poor genes ($\leq$ 2 introns per gene) including 13 genes with no intron, three genes (*Zm00001d027639*, *Zm00001d041181*, and *Zm00001d026066*) with only one intron, and one gene (*Zm00001d052301*) with two introns. For the clade II, all members were intron-poor, except *Zm00001d042333* with three introns and *Zm00001d043134* with seven introns. In contrast, the majority of members in the clade III and IV were intron-rich ($\geq$ 3 introns per gene). For example, 64.29% (9/14) members in the subgroup III contained at least three introns. Particularly, *Zm00001d013199*, *Zm00001d042820*, and *Zm00001d038905* in the clade IV had six, seven, and nine introns, respectively. The clade V only contained one maize GT47 gene (*Zm00001d048298*), which had 11 introns.

The GT47 gene family in maize has expanded through tandem duplication and segmental duplication

The 47 GT47 genes are unevenly distributed in the maize B73 reference genome, some of which are located closely to each other on the same chromosome (Fig. 3). This promoted us to examine whether the GT47 gene family is involved in tandem duplication events. On the basis of the criteria that tandem duplicated genes are located within a 100-kb distance and separated by five or less genes (Ouyang et al. 2007), we found that there are 15 maize GT47 genes located within five tandem duplication regions, representing 31.9%

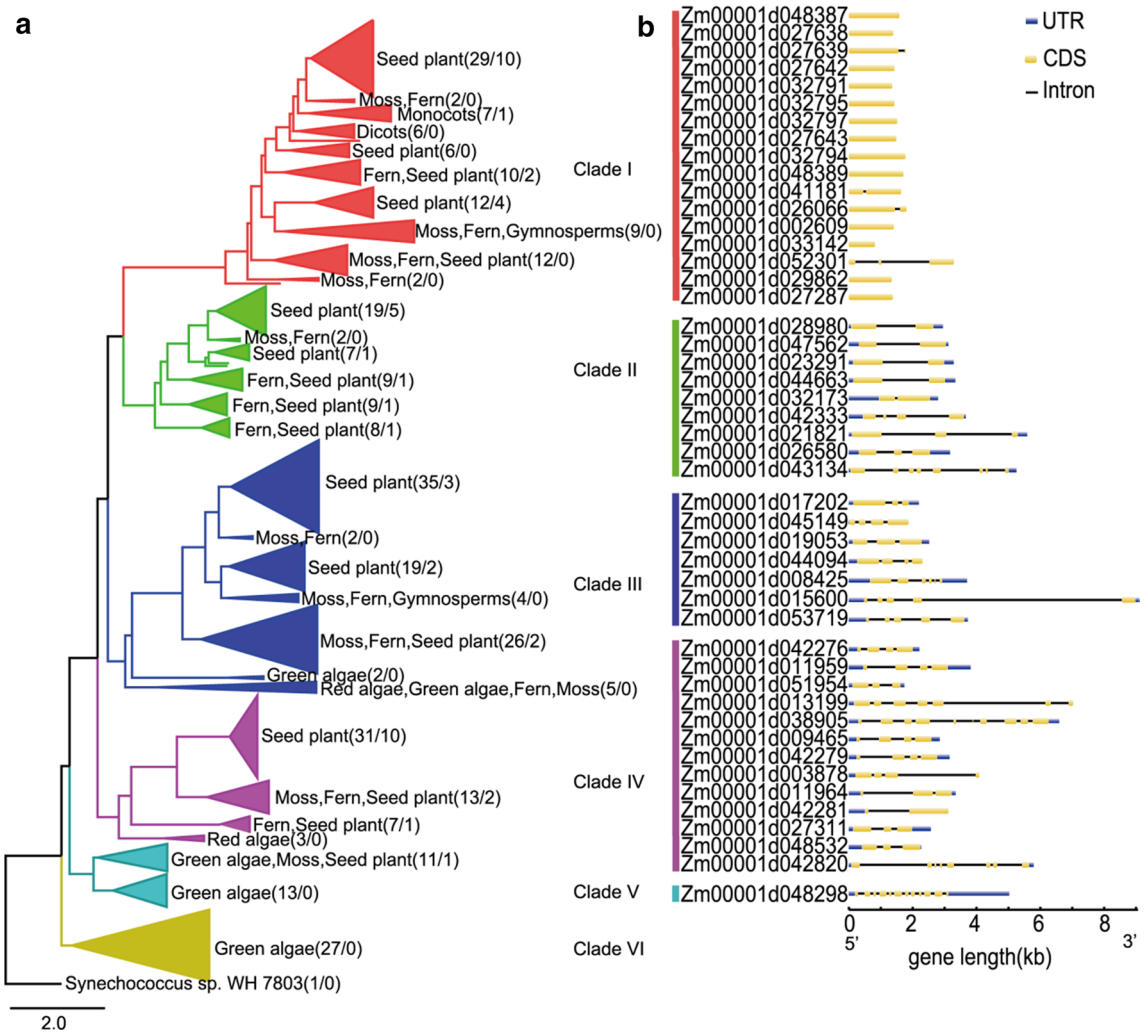

Fig. 2 Phylogenetic tree of 352 GT47 proteins from 15 species and exon–intron structures of 47 GT47 genes from maize. **a** Phylogenetic tree of the GT47 gene family. 352 GT47 protein sequences were divided into six clades (I, II, III, IV, V and VI) shown with different colors. Two numbers in parentheses represent the number of GT47 proteins from all analyzed species and that only from maize, respectively. **b** Exon–intron structures of GT47 genes in maize. Untranslated regions (UTRs), coding regions (CDS), and introns are indicated by blue boxes, yellow boxes, and black lines, respectively

of all GT47 genes in maize (Fig. 3). Among these 15 GT47 genes, 12 are intron-poor genes. Interestingly, three tandem duplication regions contain 10 GT47 genes, all of which are intron-poor. These 10 intron-poor GT47 genes can be grouped in a branch of the clade I consisting of only seed plant species (Fig. 3). The other two tandem duplication regions contain five GT47 genes, two of which are intron-poor genes. These five GT47 genes can be grouped in a seed-plant-specific branch of the clade IV. These results indicated that tandem duplications have made significant contribution to the expansion of intron-poor genes in the GT47 family in maize during the seed plant evolution.

Besides tandem duplication, segmental duplication is another mechanism of gene family expansion frequently reported in plants (Zhu et al. 2014; Leister 2004; Cannon et al. 2004). To examine whether segmental duplication is involved in the GT47 family expansion or not, segmental duplication regions in the whole maize genome were identified by the SynMap tool available at the CoGe server (Lyons et al. 2008). We found that 25.5% (12/47) of GT47 genes in maize are located within six segmental duplication regions, forming six segmentally duplicated genes pairs (Fig. 3). Among these 12 GT47 genes, 10 are intron-poor (Fig. 3). These gene pairs were grouped in the fern/seed-plant branch of the clade I (*Zm00001d026066/Zm00001d002609*), the seed-plant-specific branch of the clade II (*Zm00001d028980/Zm00001d047562*; *Zm00001d023291/Zm00001d044663*), the seed-plant-specific branch (*Zm00001d038905/Zm00001d009465*; *Zm00001d011964/Zm000*

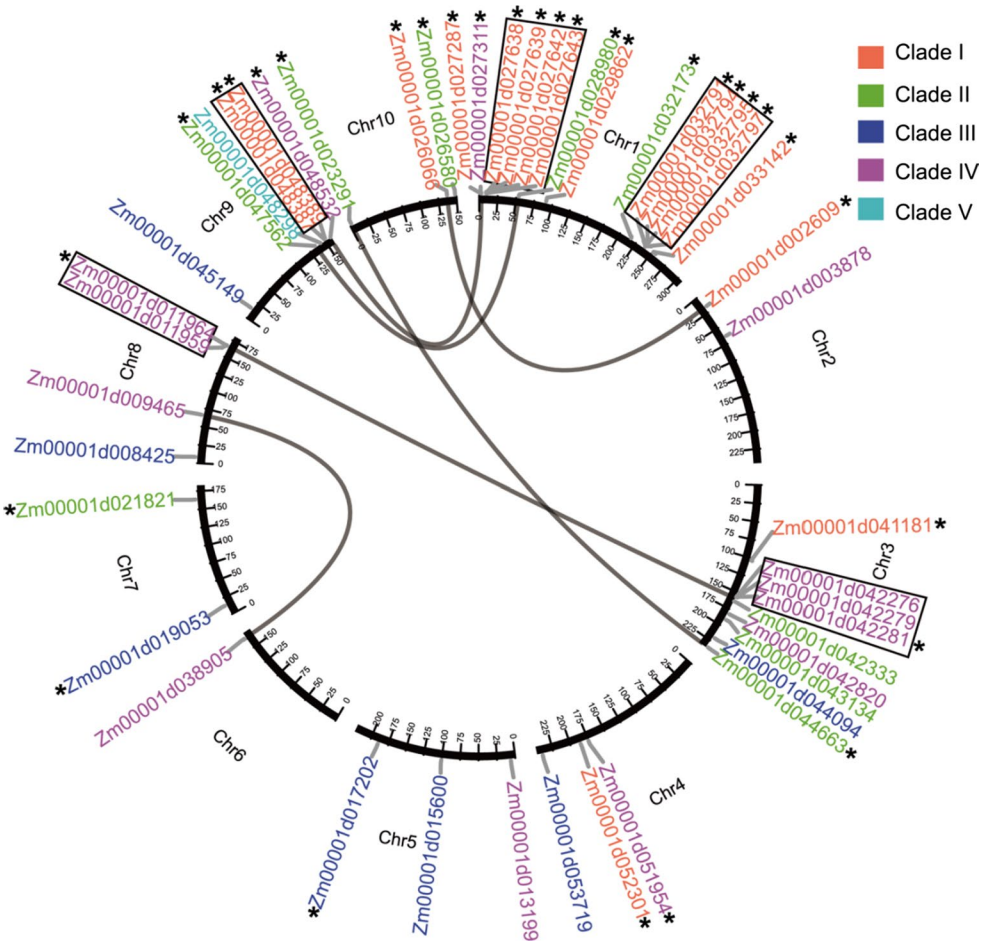

Fig. 3 Genomic location of 47 GT47 genes in maize. Segmentally duplicated genes are connected by gray lines, and tandemly duplicated genes are indicated by a black box. Asterisks indicate intron-poor genes. The colors of gene names represent the clades defined in Fig. 2

01d042281) and the moss/fern/seed-plant branch (*Zm00001d027311/Zm00001d048532*) of the clade IV, respectively.

The duplication time of these six gene pairs dated back to around 37.4, 7.8, 14.6, 15.7, 8.2, 13.3 million years ago (mya), respectively (Online Resource 9). This result showed that segmental duplications expanded the GT47 gene family in maize after the divergence of the grasses, around 60 mya (Salse et al. 2008). Some segmentally duplicated gene pairs (*Zm00001d026066/Zm00001d002609*; *Zm00001d027311/Zm00001d048532*) existed before the divergence of maize from sorghum, while some (*Zm00001d011964/Zm00001d042281*; *Zm00001d028980/Zm00001d047562*) likely occurred after the divergence (Online Resource 10). Although, as expected, GT47 genes in segmentally duplicated regions were under strong purifying selection (one gene pair with d*N*/d*S* of 0.82 and the other five gene pairs with d*N*/d*S* < 0.40) (Online Resource 9), changes in the exon–intron structure of three segmentally duplicated gene pairs were observed (Fig. 4). In addition, expression divergences were also observed though co-expression analysis of two genes in each segmentally duplicated gene pair. The

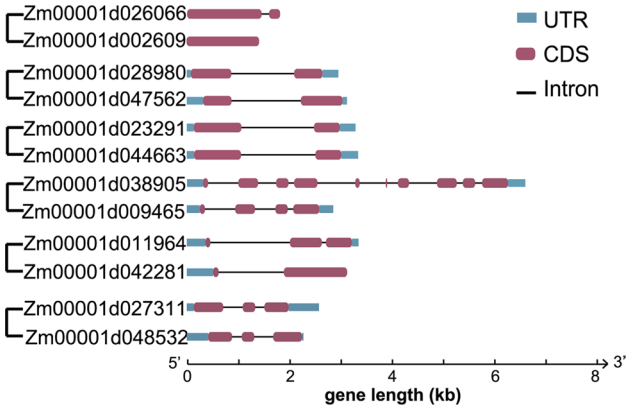

Fig. 4 Exon–intron structures of segmental duplication genes. Exons, introns, and untranslated regions (UTRs) are indicated by dark red boxes, black lines and blue box, respectively

GCC value ranges from 0.33 to 0.97 for the drought stress-related transcriptomic map (Fig. 5), and ranges from − 0.27

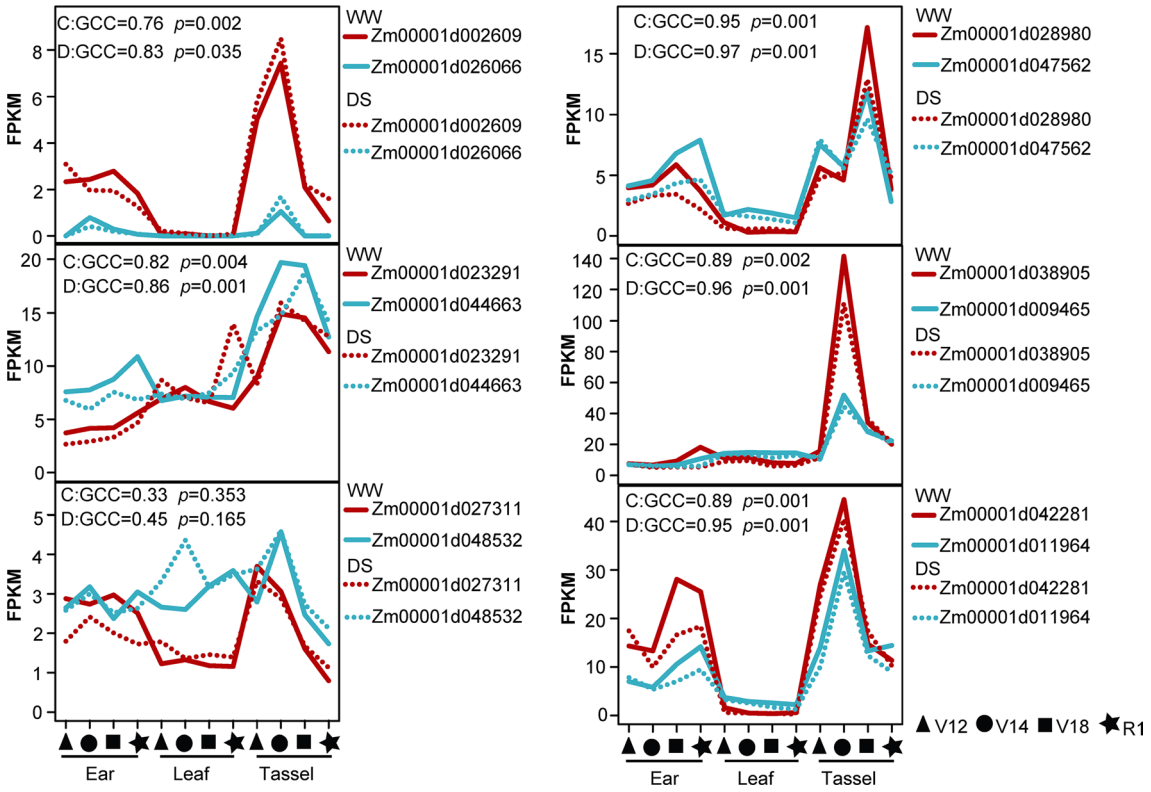

Fig. 5 Expression profiles of segmentally duplicated genes in three tissues of maize at four developmental stages under well-watered (WW) and drought-stressed (DS) conditions. The *x*-axis represents the tissues and stages, and the *y*-axis represents gene expression levels in terms of FPKM. The red and blue lines represent gene expression of segmental duplication gene pairs. The solid and dashed lines represent experiments under well-watered and drought-stressed conditions, respectively. GCC represents the Gini correlation coefficient between two genes

to 0.97 for the seed development-related transcriptomic map (Fig. 6).

The GT47 gene family is involved in drought stress responses

The analysis of *cis*-acting elements identified four environmental stress-related elements in the region 1500-bp upstream of GT47 genes in maize (Online Resource 11), suggesting possible roles of GT47 gene family in responding to environmental stresses. To investigate this possibility, we performed differential expression analysis for 47 maize GT47 genes using the spatio-temporal transcriptomic map obtained from three tissues (leaf, ear, and tassel) of B73 maize at four developmental stages (V12, V14, V18, and R1) under well-watered and drought stress conditions. Among the 47 GT47 genes in maize, 11 genes were non-expressed (FPKM < 1) in all tissues under all experimental conditions (Online Resource 12); 29 genes exhibited slight differences in expression patterns between well-watered and drought-stressed conditions (Online Resource 12); seven showed significant responses to drought stress with different manners (|log2 (fold-change)| ≥ 1; FDR < 0.05), covering two intron-poor genes (*Zm00001d048389* and *Zm00001d032795*) in the clade I (Fig. 7a). In the maize ear tissue, *Zm00001d019053* and *Zm00001d032795* were remarkably up-regulated at the V12 and R1 stage, respectively; *Zm00001d038905* and *Zm00001d048389* were significantly down-regulated at the R1 stage. In the maize tassel tissue, *Zm00001d017202* showed up-regulated expression at the R1 stage; in the maize leaf tissue, *Zm00001d023291* and *Zm00001d042279* showed up-regulated expression at the R1 stage.

Further qRT-PCR experiments also indicated the drought-stress response of three GT47 genes (*Zm00001d048389, Zm00001d038905* and *Zm00001d042279*) in the leaf and root tissues of 12-day-old maize plants subjected to 5 and 10 h of PEG8000 for simulating drought stress treatment (Fig. 7b). *Zm00001d048389* and *Zm00001d038905* were significantly up-regulated in the maize leaf tissue at 5 and 10 h after PEG treatment, respectively. *Zm00001d042279* was significantly up-regulated in the maize root at 10 h after PEG treatment. Of note, *Zm00001d023291* showed slight expression changes; while the other three genes (*Zm00001d017202*,

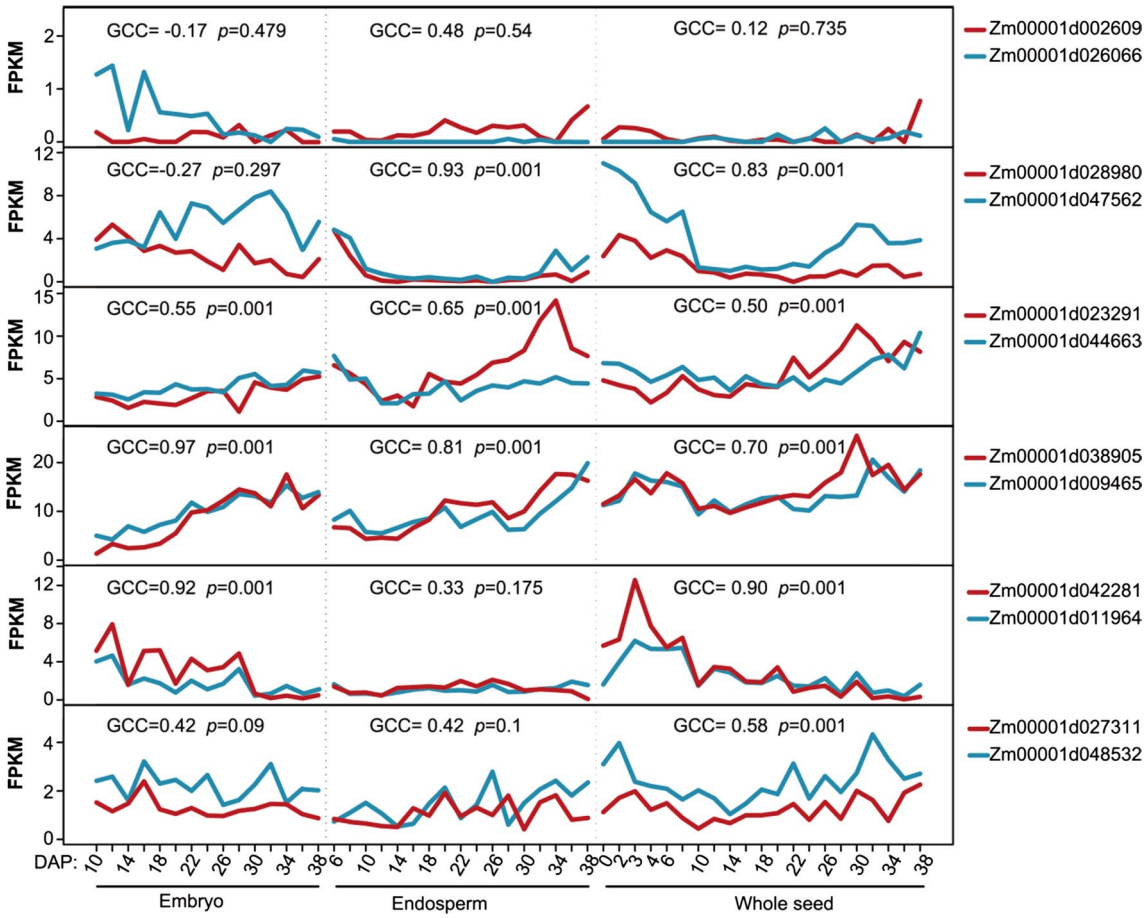

Fig. 6 Expression profiles of segmentally duplicated genes during the seed development of maize. The *x*-axis represents different tissues and seed developmental stages, and the *y*-axis represents the gene expression levels in terms of FPKM. The red and blue lines represent gene expression of segmentally duplicated genes, respectively. DAP: day after pollination

Zm00001d019053 and *Zm00001d032795*) were not expressed in maize leafs and roots at the three-leaf stage.

To further characterize the possible roles of GT47 gene family in drought stress responses, we systematically examined the gene expression at both transcriptional and translational levels in leaf tissue of maize seedlings that were grown in well-watered and drought-stressed conditions (Fig. 8a). Using the commonly used criteria of defining an "expressed" gene (Jeong et al. 2014; Miao et al. 2017), we found that 25 GT47 genes expressed at transcriptional level (FPKM ≥ 1), 72% of which (18/25) were also expressed at translational level. Among these 25 expressed GT47 genes, we detected three up-regulated genes (*Zm00001d042279*; *Zm00001d023291; Zm00001d048532*) and four down-regulated genes (*Zm00001d042276*; *Zm00001d042281*; *Zm00001d011964*; *Zm00001d011959*) at transcriptional level. Meanwhile, three up-regulated genes (*Zm00001d042279*; *Zm00001d023291*; *Zm00001d044663*) and three down-regulated genes (*Zm00001d042276*; *Zm00001d042281*; *Zm00001d011964*) were detected at translation level (Fig. 8c; Online Resource 13). Of note, *Zm00001d048532* and *Zm00001d011959* showed differential expression at only transcriptional level; *Zm00001d044663* exhibited differential expression at only translational level (Fig. 8c; Online Resource 13). These results indicated that drought stress differentially altered gene expression of GT47 genes at both transcriptional and translational levels (Fig. 8b). One of the possible reasons is that drought stress resulted in the change of translational efficiency of several GT47 genes. For example, drought stress enhanced the translational efficiency of *Zm00001d027311,* the value increased from 0.19 to 0.51 (Online Resource 14). Drought stress also inhibited the translational efficiency of some GT47 genes. The TE value of *Zm00001d048298* decreased from 0.52 to 0.24 (Online Resource 14).

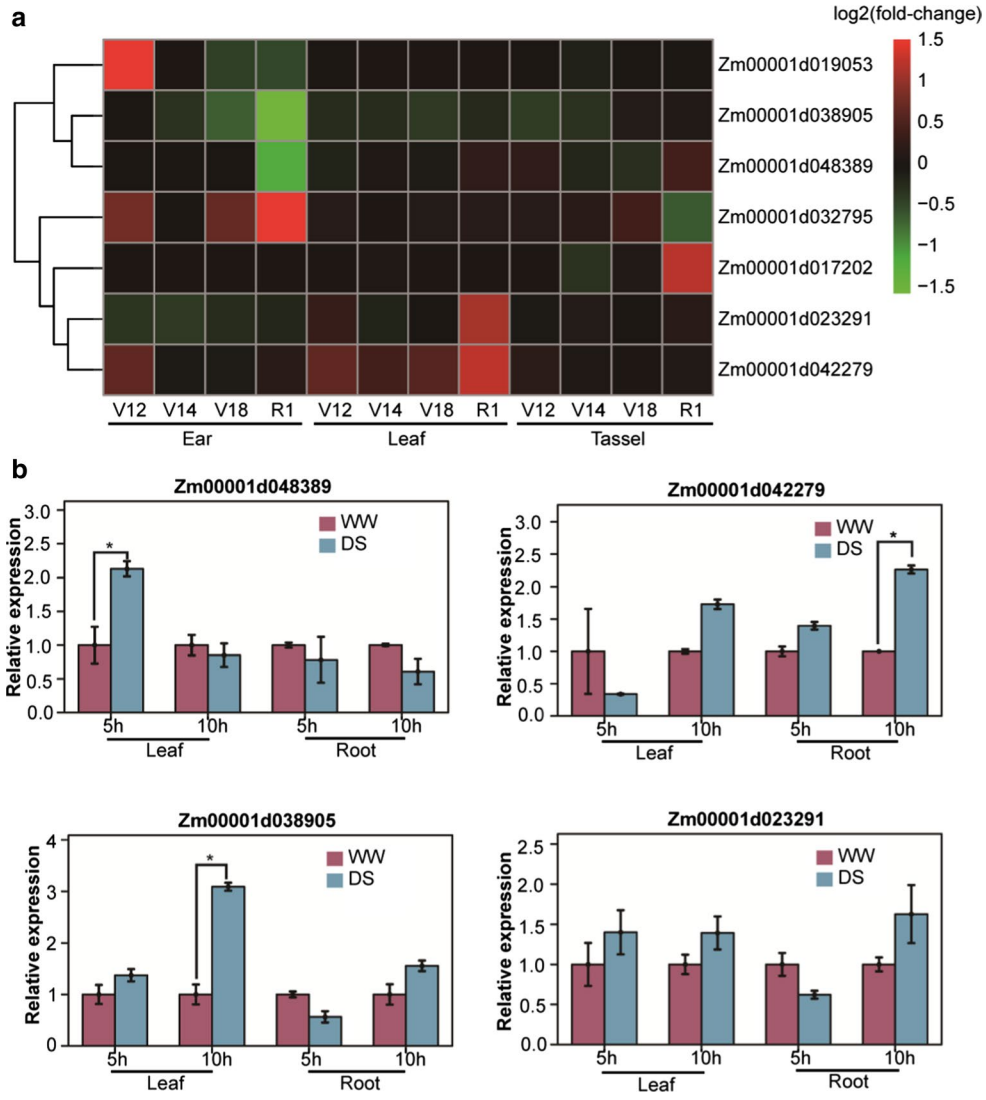

Fig. 7 Expression levels of maize GT47 genes under drought stress. **a** Heat map of the expression profile of seven maize GT47 genes in response to drought stress. The log2 (fold-change) value was used to construct the heat map. **b** qRT-PCR expression levels of four GT47 genes in maize under drought stress. Expression levels of GT47 s under PEG8000 treatment. Relative expression values were the average of three independent biological samples ± SE. Significantly difference between the well-watered (WW) and drought-stressed (DS) experiments (Student's *t* tests; p value < 0.05) were indicated by an asterisk

Expression behavior of maize GT47 genes during seed development

The expression patterns of GT47 genes were also explored in three tissues (embryo, endosperm, and whole seed) during maize seed development from 0 DAP to 38 DAP. We found that 30 GT47 genes were expressed with FPKM higher than one in the development of seeds. Among these 30 GT47 genes, there were 16 intron-poor genes (*Zm00001d009465 Zm00001d011964 Zm00001d013199 Zm00001d015600 Zm00001d017202 Zm00001d019053 Zm00001d021821 Zm00001d023291 Zm00001d026066 Zm00001d026580 Zm00001d027311 Zm00001d027642 Zm00001d032791 Zm00001d032794 Zm00001d032797 Zm00001d033142*). Based on hierarchical clustering of z-score normalized FPKM values, these genes were divided into four groups in tissue-specific manners (Fig. 9). 11 genes in the group 2 exhibit higher expression levels in the later stages of endosperm and whole seed development than those in the early and middle stages. Three genes in the group 3 show higher expression levels in embryo than endosperm and whole seed. While in the group 4, 14 genes exhibit higher expression levels in embryo and whole seed than endosperm. The remaining two genes that belonged to group 1 showed irregular expression patterns. These results showed that the GT47 gene family might function in specific tissue and stage during maize seed development.

Discussion

This study reports, to our knowledge, the first systematic and large-scale bioinformatics analysis of GT47 gene family. Our study identified 352 GT47 proteins from 15 species ranging from cyanobacteria to seed plants. Further phylogenetic analysis expanded our knowledge of the GT47 gene family

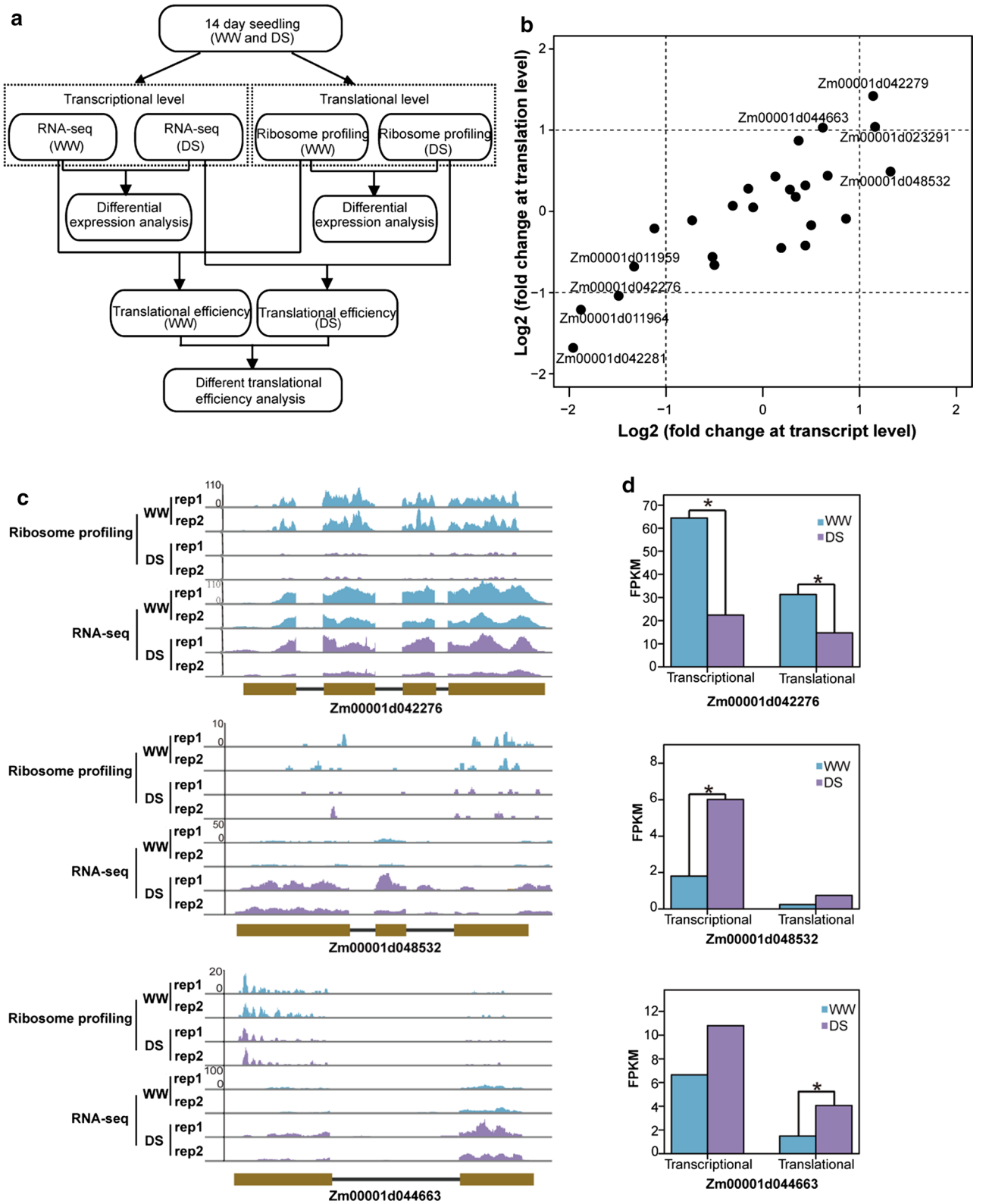

Fig. 8 Expression profiles of maize GT47 genes at transcriptional and translational levels in maize seedlings under well-watered (WW) and drought-stressed (DS) conditions. **a** Flowchart of RNA-seq and ribosome profiling data analysis. **b** Fold-change of gene expression of GT47 genes at transcriptional and translational levels. **c** Examples of three GT47 genes with different coverages of ribosome profiling reads and RNA-seq reads under drought stress. **d** Examples of three GT47 genes with significant changes in gene expression at transcriptional and/or translational level. Genes with significant differences in gene expression were indicated by an asterisk

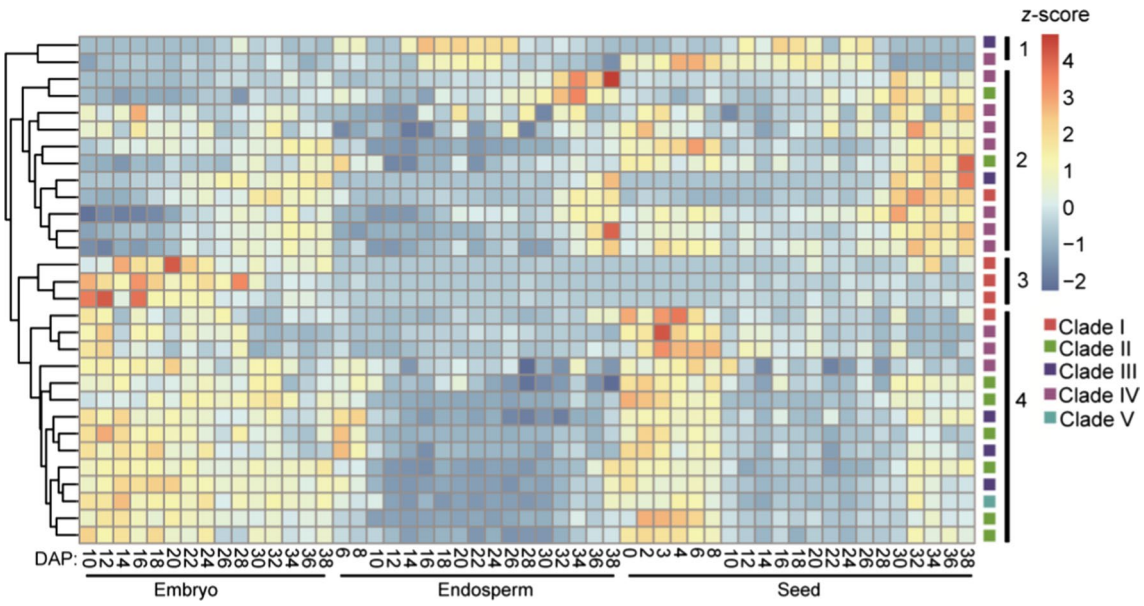

Fig. 9 Heat map of the expression profile of maize GT47 genes during seed development. The *z*-score of GT47 gene expression values were used to construct the heat map. The color of the square represents the gene clade

evolution, which was limited according to the results obtaining from the phylogenetic tree consisting of only *Arabidopsis* and sorghum GT47 proteins (Rai et al. 2016). The comprehensive bioinformatics analysis also identified two intron-poor clades of the GT47 gene family in maize, which may form in moss plants. The results presented here also suggested two of the major mechanisms (tandem duplication and segmental duplication) for the GT47 gene family expansion in maize. Finally, our analyses have provided novel observations about expression patterns of GT47 genes in two spatio-temporal transcriptomic maps of maize, indicating the putative biological function of the GT47 gene family involving in drought stress responses and seed development.

Although belonging to the same GT superfamily, GT47 gene family is different from the other previously reported gene families (GT1, GT7 and GT43) in several aspects (Online Resource 15). First, the domain of GT47 gene family is different from that of GT1, GT7 and GT43 in the protein sequence and length. The GT47 gene family encodes proteins containing an exostosins domain with length of 300 amino acids, while proteins in the GT1, GT8, and GT43 containing a C-terminal consensus sequence (Plant Secondary Product Glycosyltransferase [PSPG] motif; 44 amino acids), Glyco_transf_8 domain (256 amino acids), and Glyco_transf_43 (206 amino acids), respectively. These differences in domain indicate that the function diversity may exist between proteins from different GT gene families. Second, differential evolutionary paths may experience by these four GT gene families. In our study, we found that the GT47 gene family likely originated in cyanobacteria and can be classified into six clades. While previous studies reported that the GT1, GT8, and GT43 gene families might origin in chloroplast (Li et al. 2014), cyanobacteria (Yin et al. 2010), and green algae (Taujale and Yin 2015), respectively. These three gene families (GT1, GT8, and GT43) can be divided into 14, eight, and four clades, respectively. More information regarding the evolutionary histories and relationships between different GT families are required to be further explored. Third, the composition of intron-poor and intron-rich genes is different in these GT gene families. In maize, 66.0% (31/47) of GT47 genes are intron-poor ($\leq$ 2 introns per gene), while the proportions in GT1, GT8, and GT43 are 97.3% (143/147), 100% (13/13), and 47.9% (23/48), respectively. Function of these intron-poor genes in different GT families is also needed to be investigated in the future.

The maize GT47 gene family has significantly expanded by tandem duplication and segmental duplication in its evolutionary process. Fifteen GT47 genes in maize experienced tandem duplication, and were grouped in the seed-plant-specific branch of clade I and clade IV, indicating that the tandem duplication-driven GT47 gene family expansion may appear in seed plants. In addition, 66.7% (10/15) of these tandemly duplicated genes are intron-poor in the clade I. Known that the importance of tandem duplication in plant adaptation to environmental stimuli (Hanada et al. 2008), future researches are suggested to explore the correlation between intron-poor gene clade and plant adaptation to environmental stresses. We found that there are 12 GT47 genes (including 10 intron-poor genes) in maize involved in segmental duplication, forming six segmentally duplicated

gene pairs with different evolutionary histories and varying degrees of differentiation (Figs. 3, 4). The duplication of these six gene pairs occurred after the divergence of the grasses and went through purifying selection (Online Resource 9). However, gene structural and expression divergences have been observed (Figs. 4, 5, 6). For instance, *Zm00001d027311* and *Zm00001d048532* have different length of untranslated region and showed low Gini correlation (GCC = 0.33, p value = 0.353) in three tissues (leaf, ear, and tassel) of B73 maize at four developmental stages (V12, V14, V18, and R1). The expression divergence was also detected for some duplicated gene pairs in different tissues. For instance, *Zm00001d027311* and *Zm00001d048532* exhibited low correlation in embryo (GCC = 0.42, p value = 0.09) and endosperm (GCC = 0.42, p value = 0.1) but high correlation (GCC = 0.58, p value = 0.001) in whole seed. These differences in gene structure and expression indicated that novel function (neofunctionalization) may arise for some duplicated genes in different biological processes like seed development and drought stress responses.

Besides 12 duplicated genes, 24 other GT47 genes (12 intron-poor genes and 12 intron-rich genes) in maize have also been identified to be associated with drought stress, through the analysis of spatio-temporal transcriptomic transcriptome data of B73 maize under normal and drought-stressed conditions. Several interesting clues were observed in Online Resource 12. For example, *Zm00001d017202*, *Zm00001d011959*, *Zm00001d013199*, *Zm00001d027287*, *Zm00001d032173*, *Zm00001d032797*, *Zm00001d038905*, *Zm00001d042276*, *Zm00001d044094*, and *Zm00001d045149* showed tassel-specific expression patterns. Besides, the expression of *Zm00001d002609*, *Zm00001d015600*, *Zm00001d026066*, *Zm00001d026580*, *Zm00001d042281*, *Zm00001d042820*, and *Zm00001d048389* were markedly suppressed in leaf tissue which suggested that these genes might be involved in differentiation between reproductive and vegetative tissues in early development. As shown in Fig. 8, *Zm00001d017202*, *Zm00001d019053* and *Zm00001d032795* were up-regulated at the R1 stage in reproductive tissue which implied these genes might play positive roles in response to drought stress. In contrast, *Zm00001d038905* and *Zm00001d048389* were significantly down-regulated suggesting negative regulation possibly. *Zm00001d023291* and *Zm00001d042279* showed up-regulated expression at the R1 stage in leaf tissue specifically. These indicated the functional differentiation of GT47 members among multiple tissues. Taken together, our study provides information on when and where all of GT47 genes are actively expressed. The spatio-temporal gene expression diversity, combined with changes in translational levels in maize seedlings during drought stress were very useful for accelerating understanding of their roles. This transcriptome data was further used to construct a predicted transcriptional regulatory network, which consists of 52 transcription factors (TFs) and 33 (22 intron-poor genes) GT47 genes (Online Resource 16; Online Resource 17). The MYB and C2H2 TF family are heavily connected with GT47 genes in the network, highlighting their importance in response to drought stresses. The putative regulatory mechanisms of GT47 gene family in maize seed development can also be explored using the transcriptional regulatory network covering 52 TFs (7 development-related TF families) and 32 GT47 genes (22 intron-poor genes) (Online Resource 17). The TF gene family ERF contains 17 TFs, which regulate 13 GT47 genes and may be key regulators of GT47 genes during seed development. These results highlight the functional versatility of the GT47 gene family in maize and motivate the future research in investigating regulatory mechanisms of GT47 genes (especially intron-poor GT47 genes) in drought stress response and seed development.

Author contribution statement Designed the experiments: CM, ZH and ZM. Performed the experiments: JT, CR and RY. Analyzed the data: JT, CR, RY and XZ. Wrote the paper: JT, CM, TY and ZM. All authors read and approved the final manuscript.

Acknowledgements This work was supported by the Special Fund for Basic Scientific Research of Central College (QN2011114 and 2452015412), the Fund of Northwest A & F University (Z111021603 and Z111021403), the Youth Talent Program of State Key Laboratory of Crop Stress Biology for Arid Areas (CSBAAQN2016001), and Projects of Youth Technology New Star of Shaanxi Province (2017KJXX-67).

Compliance with ethical standards

Conflict of interest We declare that we have no competing interests.

References

Almagro Armenteros JJ, Kaae Sønderby C, Kaae Sønderby S, Nielsen H, Winther O (2017) DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33:3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Barvkar VT, Pardeshi VC, Kale SM, Kadoo NY, Gupta VS (2012) Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in *Linum usitatissimum* identified genes with varied expression patterns. BMC Genom 13:175. https://doi.org/10.1186/1471-2164-13-175

Batidzirai B, Valk M, Wicke B, Junginger M, Daioglou V, Euler W, Faaij A (2016) Current and future technical, economic and environmental feasibility of maize and wheat residues supply for biomass energy application: illustrated for South Africa. Biomass Bioenergy 92:106–129. https://doi.org/10.1016/j.biombioe.2016.06.010

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brar GA, Weissman JS (2015) Ribosome profiling reveals the what, when, where, and how of protein synthesis. Nat Rev Mol Cell Biol 16:651–664. https://doi.org/10.1038/nrm4069

Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR (2009) Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in Arabidopsis. Plant J 57:732–746. https://doi.org/10.1111/j.1365-313X.2008.03729.x

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol 4:10. https://doi.org/10.1186/1471-2229-4-10

Caputi L, Malnoy M, Goremykin V, Nikiforova S, Martens S (2012) A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. Plant J 69:1030–1042. https://doi.org/10.1111/j.1365-313X

Chen X, Vega-Sánchez ME, Verhertbruggen Y, Chiniquy D, Canlas PE, Fagerström A, Prak L, Christensen U, Oikawa A, Chern M (2013) Inactivation of *OsIRX10* leads to decreased xylan content in rice culm cell walls and improved biomass saccharification. Mol Plant 6:570–573. https://doi.org/10.1093/mp/sss135

Chen J, Zeng B, Zhang M, Xie S, Wang G, Hauck A, Lai J (2014) Dynamic transcriptome landscape of maize embryo and endosperm development. Plant Physiol 166:252–264. https://doi.org/10.1104/pp.114.240689

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285. https://doi.org/10.1093/nar/gkv1344

Gasteiger E, Hoogland C, Gattiker A, Se Duvaud, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook. Humana Press, Totowa, pp 571–607

Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci 93:10274–10279. https://doi.org/10.1073/pnas.93.19.10274

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186. https://doi.org/10.1093/nar/gkr944

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148:993–1003. https://doi.org/10.1104/pp.108.122457

Hayward AP, Moreno MA, Howard TP, Hague J, Nelson K, Heffelfinger C, Romero S, Kausch AP, Glauser G, Acosta IF (2016) Control of sexuality by the sk1-encoded UDP-glycosyltransferase of maize. Sci Adv 2:e1600991. https://doi.org/10.1126/sciadv.1600991

Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G (2015) GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31:1296–1297. https://doi.org/10.1093/bioinformatics/btu817

Huang F-F, Chai C-L, Zhang Z, Liu Z-H, Dai F-Y, Lu C, Xiang Z-H (2008) The UDP-glucosyltransferase multigene family in *Bombyx mori*. BMC Genom 9:563. https://doi.org/10.1186/1471-2164-9-563

Iwai H, Masaoka N, Ishii T, Satoh S (2002) A pectin glucuronyltransferase gene is essential for intercellular attachment in the plant meristem. Proc Natl Acad Sci 99:16319–16324. https://doi.org/10.1073/pnas.252530499

Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, Wang H, Hannah R (2014) Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat Genet 46:17–23. https://doi.org/10.1038/ng.2836

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell M, Stein JC, Wei X, Chin C-S (2017) Improved maize reference genome with single molecule technologies. Nature 546:524–527. https://doi.org/10.1038/nature22971

Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, Gao G (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45:D1040–D1045. https://doi.org/10.1093/nar/gkw982

Jones P, Vogt T (2001) Glycosyltransferases in secondary plant metabolism: tranquilizers and stimulant controllers. Planta 213:164–174. https://doi.org/10.1007/s004250000492

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

Kohl M, Wiese S, Warscheid B (2011) Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol 696:291–303. https://doi.org/10.1007/978-1-60761-987-1_18

Kong Y, Peña MJ, Renna L, Avci U, Pattathil S, Tuomivaara ST, Li X, Reiter W-D, Brandizzi F, Hahn MG (2015) Galactose-depleted xyloglucan is dysfunctional and leads to dwarfism in *Arabidopsis*. Plant Physiol 167:1296–1306. https://doi.org/10.1104/pp.114.255943

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. https://doi.org/10.1101/gr.092759.109

Lairson LL, Henrissat B, Davies GJ, Withers SG (2008) Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem 77:521–555. https://doi.org/10.1146/annurev.biochem.76.061005.092322

Le Gall H, Philippe F, Domon J-M, Gillet F, Pelloux J, Rayon C (2015) Cell wall metabolism in response to abiotic stress. Plants 4:112–166. https://doi.org/10.3390/plants4010112

Lei L, Shi J, Chen J, Zhang M, Sun S, Xie S, Li X, Zeng B, Peng L, Hauck A (2015) Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. Plant J 84:1206–1218. https://doi.org/10.1111/tpj.13073

Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends Genet 20:116–122. https://doi.org/10.1016/j.tig.2004.01.007

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30:325–327. https://doi.org/10.1093/nar/30.1.325

Li Y, Baldauf S, Lim E-K, Bowles DJ (2001) Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. J Biol Chem 276:4338–4343. https://doi.org/10.1074/jbc.M007447200

Li Y, Li P, Wang Y, Dong R, Yu H, Hou B (2014) Genome-wide identification and phylogenetic analysis of Family-1 UDP glycosyltransferases in maize (*Zea mays*). Planta 239:1265–1279. https://doi.org/10.1007/s00425-014-2050-1

Li P, Li YJ, Zhang FJ, Zhang GZ, Jiang XY, Yu HM, Hou BK (2017) The Arabidopsis UDP-glycosyltransferases UGT79B2 and UGT79B3, contribute to cold, salt and drought stress tolerance via modulating anthocyanin accumulation. Plant J 89:85–103. https://doi.org/10.1111/tpj.13324

Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. https://doi.org/10.1093/nar/gkt1178

Lovegrove A, Wilkinson MD, Freeman J, Pellny TK, Tosi P, Saulnier L, Shewry PR, Mitchell RA (2013) RNA interference suppression of genes in glycosyl transferase families 43 and 47 in wheat starchy endosperm causes large decreases in arabinoxylan content. Plant Physiol 163:95–107. https://doi.org/10.1104/pp.113.222653

Lyons E, Pedersen B, Kane J, Freeling M (2008) The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. Trop Plant Biol 1:181–190. https://doi.org/10.1007/s12042-008-9017-y

Ma C, Wang X (2012) Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis. Plant Physiol 160:192–203. https://doi.org/10.1104/pp.112.201962

Ma C, Xin M, Feldmann KA, Wang X (2014) Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in *Arabidopsis*. Plant Cell 26:520–537. https://doi.org/10.1105/tpc.113.121913

Madson M, Dunand C, Li X, Verma R, Vanzin GF, Caplan J, Shoue DA, Carpita NC, Reiter W-D (2003) The MUR3 gene of *Arabidopsis* encodes a xyloglucan galactosyltransferase that is evolutionarily related to animal exostosins. Plant Cell 15:1662–1670. https://doi.org/10.1105/tpc.009837

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229. https://doi.org/10.1093/nar/gkq1189

Miao Z, Han Z, Zhang T, Chen S, Ma C (2017) A systems approach to a spatio-temporal understanding of the drought stress response in maize. Sci Rep 7:6590. https://doi.org/10.1038/s41598-017-06929-y

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res 35:D883–D887. https://doi.org/10.1093/nar/gkl976

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:e45. https://doi.org/10.1093/nar/29.9.e45

Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490

Rai KM, Thu SW, Balasubramanian VK, Cobos CJ, Disasa T, Mendu V (2016) Identification, characterization, and expression analysis of cell wall related genes in *Sorghum bicolor* (L.) moench, a food, fodder, and biofuel crop. Front Plant Sci 7:1287. https://doi.org/10.3389/fpls.2016.01287

Rehman HM, Nawaz MA, Bao L, Shah ZH, Lee J-M, Ahmad MQ, Chung G, Yang SH (2016) Genome-wide analysis of family-1 UDP-glycosyltransferases in soybean confirms their abundance and varied expression during seed development. J Plant Physiol 206:87–97. https://doi.org/10.1016/j.jplph.2016.08.017

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell 20:11–24. https://doi.org/10.1105/tpc.107.056309

Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A (2013) TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res 42:D922–D925. https://doi.org/10.1093/nar/gkt1055

Sharma R, Rawat V, Suresh C (2014) Genome-wide identification and tissue-specific expression analysis of UDP-glycosyltransferases genes confirm their abundance in *Cicer arietinum* (Chickpea) genome. PLoS One 9:e109715. https://doi.org/10.1371/journal.pone.0109715

Shirokikh NE, Archer SK, Beilharz TH, Powell D, Preiss T (2017) Translation complex profile sequencing to study the in vivo dynamics of mRNA-ribosome interactions during translation initiation, elongation and termination. Nat Protoc 12:697–731. https://doi.org/10.1038/nprot.2016.189

Strable J, Scanlon MJ (2009) Maize (*Zea mays*): a model organism for basic and applied research in plant biology. Cold Spring Harb Protoc. https://doi.org/10.1101/pdb.emo132

Taujale R, Yin Y (2015) Glycosyltransferase family 43 is also found in early eukaryotes and has three subfamilies in charophycean green algae. PLoS One 10:e0128409. https://doi.org/10.1371/journal.pone.0128409

Tedman-Jones JD, Lei R, Jay F, Fabro G, Li X, Reiter WD, Brearley C, Jones JD (2008) Characterization of *Arabidopsis* mur3 mutations that result in constitutive activation of defence in petioles, but not leaves. Plant J 56:691–703. https://doi.org/10.1111/j.1365-313X.2008.03636.x

Thatcher SR, Danilevskaya ON, Meng X, Beatty M, Zastrow-Hayes G, Harris C, Van Allen B, Habben J, Li B (2016) Genome-wide analysis of alternative splicing during development and drought stress in maize. Plant Physiol 170:586–599. https://doi.org/10.1104/pp.15.01267

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562. https://doi.org/10.1038/nprot.2012.016

Weis M, Lim E-K, Bruce NC, Bowles DJ (2008) Engineering and kinetic characterisation of two glucosyltransferases from *Arabidopsis thaliana*. Biochimie 90:830–834. https://doi.org/10.1016/j.biochi.2008.01.013

Wu B, Gao L, Gao J, Xu Y, Liu H, Cao X, Zhang B, Chen K (2017) Genome-wide identification, expression patterns, and functional analysis of UDP glycosyltransferase family in peach (*Prunus persica* L. Batsch). Front plant Sci 8:389. https://doi.org/10.3389/fpls.2017.00389

Xu B, Yang Z (2013) pamlX: a graphical user interface for PAML. Mol Biol Evol 30:2723–2724. https://doi.org/10.1093/molbev/mst179

Yin Y, Chen H, Hahn MG, Mohnen D, Xu Y (2010) Evolution and function of the plant cell wall synthesis-related glycosyltransferase family 8. Plant Physiol 153:1729–1746. https://doi.org/10.1104/pp.110.154229

Yonekura-Sakakibara K, Hanada K (2011) An evolutionary view of functional diversity in family 1 glycosyltransferases. Plant J 66:182–193. https://doi.org/10.1111/j.1365-313X.2011.04493.x

Yu J, Hu F, Dossa K, Wang Z, Ke T (2017) Genome-wide analysis of UDP-glycosyltransferase super family in *Brassica rapa* and *Brassica oleracea* reveals its evolutionary history and functional characterization. BMC Genom 18:474. https://doi.org/10.1186/s12864-017-3844-x

Zhong R, Ye Z-H (2003) Unraveling the functions of glycosyltransferase family 47 in plants. Trends Plant Sci 8:565–568. https://doi.org/10.1016/j.tplants.2003.10.003

Zhu Y, Wu N, Song W, Yin G, Qin Y, Yan Y, Hu Y (2014) Soybean (*Glycine max*) expansin gene superfamily origins: segmental and tandem duplication events followed by divergent selection among subfamilies. BMC Plant Biol 14:93. https://doi.org/10.1186/1471-2229-14-93