Research Paper / 33-45

Bioinformatics genome-wide characterization of the *WRKY* gene family in *Sorghum bicolor*

Jafar Ahmadi[1*], Kayvan Agahi[1], Sedigheh Fabriki Ourang[1]

[1]Department of Genetics and Plant Breeding, Imam Khomeini International University, P. O. Box: 34148-96818, Qazvin, Iran.
[*]Corresponding author, Email: j.ahmadi@eng.ikiu.ac.ir. Tel: +98-28-33901227. Fax: +98-28-33780074.

Received: 09 Mar 2020; Accepted: 22 Apr 2020.
DOI: 10.30479/ijgpb.2020.12875.1266

Abstract

The *WRKY* gene family encodes a large group of transcription factors that regulate genes involved in plant response to biotic and abiotic stresses. Sorghum is a notable grain and forage crop in semi-arid regions because of its unusual tolerance against hot and dry environments. In this study, we performed a genome-wide analysis of *WRKY*s using a genome assembly of *sorghum bicolor.* First, all possible *WRKY* gene sequences as well as all possible WRKY protein sequences in the *Sorghum bicolor* genome database were identified using the NCBI website. A set of 85 *WRKY* genes was identified in the *S. bicolor* genome and classified into three groups (I–III). Among the members of the group I, the SbWRKY13 had a different and novel zinc finger motif as compared to other members of this group. It was also found that mutations occurred at R, K, Y and Q in the conserved WRKYGQK sequence. No complete gene duplication was found in gene copy number investigation, suggesting that the expansion of *SbWRKY* genes was not necessarily based on the gene duplication events or duplication of *SbWRKY* genes had probably happened in the past times. Gene cluster analysis showed that the number of genes on chromosomes were positively correlated with the number of clusters. The study of amino acid composition revealed that totally in all groups, Alanine and Proline were the most frequent residues while Cysteine had the lowest frequency. The study of introns in the SbWRKY domain showed that the majority of *SbWRKY* genes had two types of introns in their WRKY domains (phase 0 and phase 2). Also, investigation of conserved known motifs revealed that there were six, two and one known motifs outside of the region of the SbWRKY domain for groups IIa, IIb and IIc, respectively. The results describe evolution and functional differentiations of WRKY transcription factors in *Sorghum bicolor*.

***Key words*:** DNA-binding protein, *Sorghum bicolor*, Zing finger motif, *WRKY*.

INTRODUCTION

Environmental stresses are of the most important limiting factors that affect growth and yield of crops. The *WRKY* gene family encodes a large group of transcription factors that regulate genes involved in various physiological pathways including, development processes as well as responses to biotic and abiotic stresses such as pathogens (Gao *et al.,* 2020), high temperatures (Wang *et al.,* 2017), low temperatures (Romero *et al.,* 2019), salt and drought (Chanwala *et al.,* 2020), H_2O_2 (Vandenabeele *et al.,* 2003), UV radiation (Hu *et al.,* 2020), nematode damage (Chinnapandi *et al.,* 2019), wounding (An *et al.,* 2019), dormancy and germination (Chen *et al.,* 2016) and plant senescence (Gu *et al.,* 2019). Therefore, the study of this gene family would be helpful in order to adopt strategies for increasing tolerance of crop plants against stresses. Also, the phylogenetic analysis of *WRKY* genes would be useful for studying and understanding their roles in plants.

The name of the WRKY family itself is derived from the most prominent feature of these proteins,

the WRKY domain which constitutes by about 60 amino acid residues. In WRKY domain, a conserved WRKYGQK sequence at the N-terminal is followed by a metal chelating zinc finger motif (C–X4-5 – C–X22-23 –H–X–H, (C2 H2) or C–X5–8 –C–X25–28 –H–X1-2 –C, (C2 HXC)) at the C-terminal end (Xu *et al.,* 2020). However, in some *WRKY* genes, the WRKY domain can be characterized as WRRY, WSKY, WKRY, WVKY, or WKKY (Xie *et al.,* 2005). Studies have shown that WRKY transcription factors interact with the W-box (TTGAC[T/C]) sequence in promoter regions to modulate gene expression (Eulgem *et al.,* 1999; Ciolkowski *et al.,* 2008).

Classification of genes is important for the functional analysis of a gene family (Sun *et al.,* 2020). All known WRKY proteins contain either one or two WRKY domains. They can be classified on the basis of both the number of WRKY domains and the features of their zinc-finger-like motif. WRKY proteins with two WRKY domains belong to group I, whereas most proteins with one WRKY domain belong to group II. Generally, the WRKY domains of group I and group II members have the same type of finger motif, whose pattern of potential zinc ligands (CX4–5CX22–23HXH) is unique among all described zinc-finger-like motifs. The single finger motif of a small subset of WRKY proteins is distinct from that of group I and II members. Instead of a C2–H2 pattern, their WRKY domains contain a C2–HC motif (CX7CX23HXC). Owing to this distinction, they were assigned to group III. Based on a phylogenetic analysis of the WRKY family, the members of group II can be divided into five subgroups: IIa, IIb, IIc, IId, and IIe (Eulgem *et al.,* 2000). Sorghum, a C4 grass that diverged from maize just 15 million years ago, is the fifth most important cereal grown worldwide. This grain and forage crop is especially notable in the semiarid tropics because of its unusual tolerance of hot and dry environments. Sorghum has been recognized as a key plant species in the comparative analysis of grass genomes and as a source of beneficial genes for agriculture. (Mullet *et al.,* 2002).

Little research has been conducted on genome-wide WRKY transcription factor in *S. Bicolor*. Such information is useful for in-depth research in evolutionary biology, with the goal of elucidating the origin and evolution of species and enhancing their economic value. Therefore, in this study, we analyzed 86 putative *WRKY* from the *S. Bicolor* genome. Also, we conducted a phylogenetic analysis to include information on the *WRKY* gene family evolution in *S. Bicolor*.

MATERIALS AND METHODS

Gathering the *Sorghum bicolor* WRKY sequence database

First, a Hidden Markov Model (HMM) profile of the WRKY domain (PF03106) was downloaded from the Pfam database (http://pfam.sanger.ac.uk/) (Finn *et al.,* 2014). Then, all possible *WRKY* gene sequences in the *Sorghum bicolor* genome database were identified using the TBLASTN program against the non-redundant sequences (nr) database at the NCBI website (http://blast.ncbi.nlm.nih.gov). The gene overlapping was carefully scanned with regard to the start and endpoints of the nucleotide sequences and only non-overlapping *WRKY* sequences were used for further analysis. Similarly, all possible WRKY protein sequences were also identified using BLASTP programs (Expect threshold$\leq 10^{-5}$). Subsequently, to confirm the presence of the WRKY domain, all derived protein sequences were evaluated through a search using a DELTA-BLAST algorithm (Domain Enhanced Lookup) at the NCBI web site. The MapChart software version 2.3 (Voorrips, 2002) downloaded from https://www.wageningenur.nl/en/show/Mapchart-2.30.htm was used to show chromosomal locations of *Sorghum bicolor WRKY* genes. In order to study the conserved intron splicing sites in the SbWRKY domains, the unspliced DNA sequences were downloaded from http://pgsb.helmholtz-muenchen.de/plant/sorghum/index.jsp. The *Arabidopsis* WRKY protein sequences were downloaded from www.arabidopsis.org. Amino acid composition (AAC) of the SbWRKY protein sequences was calculated using the ProtParam tool at ExPASy website http://web.expasy.org/protparam. Subsequently, the AAC was subjected to a one-way analysis of variance. Then, the averages of the AACs were compared according to the least significant difference (LSD) method using SPSS software release 19.0.0 (α=0.05).

Sequence alignments and phylogenetic analysis

The alignment of the amino acid sequences of the WRKY domain was performed using the CLUSTALW program in MEGA software version 6 (Tamura *et al.,* 2013). The parameters used in the alignment were: gap open penalty: 10.00, gap extension penalty: 0.3, protein weight matrix: gonnet series, residue-specific penalties: hydrophilic penalties: on, gap separation distance: 0, end gap separation: on, use negative matrix: off, delay divergent cut off (%): 30. The alignment results were demonstrated and highlighted using Gendoc software version 2.7.

Phylogenetic trees were built using the Neighbor-

joining method and the Jones-Taylor-Thornton (JTT) model using MEGA software version 6. The test of the phylogeny was done using the bootstrap method with 1000 replicates.

The presence of possible known motifs on the outside of the main SbWRKY domain was evaluated using motif search program (http://www.genome.jp/tools/motif/) against Pfam, NCBI CDD (Conserved Domain Database) PROSITE Profile library and only those groups that had a known motif on the outside of the main SbWRKY domain were displayed.

RESULTS AND DISCUSSION

Identification, classification and sequence alignments of *S. bicolor WRKY* genes

In this study, a set of 92 *WRKY* genes was recognized in the *S. bicolor* genome. Among these sequences, 7 *WRKY* genes were excluded because of gene overlapping or lack of specific domains or motifs. Analysis of the amino acid sequences encoded by these 7 genes showed that both deletion and insertion occurred in CX4C and WRKYGQK motifs. Amongst the 85 remaining *WRKY* genes, 10 *WRKY* genes were placed in group I, 50 *WRKY* genes in group II, and 25 *WRKY* genes in group III, based on the number of WRKY domains and the type of zinc finger motifs (Table 1).

In the present study, we found that amongst 10 members of the group I, the SbWRKY13 had a different and novel zinc finger motif as compared to other members of this group. It contained two WRKY domains similar to group I while its zinc finger motif was similar to group III and not group I (C2HC instead of C2H2). Therefore, it was placed in a separate and new category as Group Ia and the rest members of the group was classified in Group Ib (Figure 1). This kind of WRKY domain structure could be important with regard to the point that the third SbWRKY group might have arisen as a result of loss in one of its WRKY domains.

WRKY genes are commonly found in land plants and many *WRKY* genes have been identified and classified in *Arabidopsis thaliana* (Eulgem *et al.,* 2000), *Oryza sativa* (Wu *et al.,* 2005; Xie *et al.,* 2005) and *Hordeum vulgare* (Mangelsen *et al.,* 2008). In the present work, a set of 85 *WRKY* genes was evaluated in the *S. bicolor* genome. Previous studies have demonstrated that the number of *WRKY* genes and genome size was independent of each other. For example, although *Populus trichocarpa* and *Crocus sativus*, have an approximately equal genome size (458 Mb and 487 Mb). However, the number of identified *WRKY* genes in *P. trichocarpa* and *C. sativus* were 104 and 55 genes, respectively which is about 2 times greater in *P. trichocarpa* than in *C. sativus*.

Multiple alignment analysis of SbWRKY domains was demonstrated in Figure 1. It has been proposed that the amino acid residues of WRKYGQK are the distinguishing regions of the WRKY transcription factor (Eulgem *et al.,* 2000; Rushton *et al.,* 2010; Song *et al.,* 2014). In this study, it was found that mutations occurred at R (Arginine), K (Lysine), Y (Tyrosine) and Q (Glutamine), in the conserved WRKYGQK sequence. In total, the amino acid substitutions included 6 Q to E (i.e. in 6 cases, the amino acid Q was replaced by E), 5 Q to K, 1 Q to S, and only 1 R to T in the conserved WRKYGQK residues (Figure 1). Thus, the highest amino acid replacement occurred at the Q position.

Further study showed that subgroup Ian had the CX4CX22-23HXH zinc finger motif. Also, a CX4CX23HXH zinc finger motif was observed for subgroup Iac, CX5CX23HXC for subgroup Ibn, CX7CX24HXC for subgroup Ibc, CX5CX23HXH for subgroups IIa, IIb and IIe, and CX4CX22-24HXH for subgroup IIc. Furthermore, group III had the CX5-7CX23-35HXC zinc finger motif (Figure 1). On the other hand, the alignment analysis revealed that in subgroup Ibc, the WRKY domain was replaced by WTKY. It seems that the subgroup IIc genes have been formed after group Ia, when it lost one of its WRKY domains. Likewise, group III is originated from the subgroup Ib. Previous research studies proposed that group II and III *WRKY* genes evolved from the group I through the elimination of the N-terminal in WRKY domain (Eulgem *et al.,* 2000; Zhang & Wang, 2005).

For better classification of *SbWRKY* genes of group II, a phylogenetic tree was constructed with 50 SbWRKY group II protein sequences together with the *Arabidopsis* WRKY protein sequences as a template (Figure 2). The results showed that the group II *SbWRKY* genes were divided into 5 subgroups, including 5 members in subgroup IIa, 6 in subgroup IIb, 21 in subgroup IIc, 7 in subgroup IId, and 11 in subgroup IIe.

Chromosomal location and the number of gene clusters of *SbWRKY* genes

A total of 85 *WRKY* genes were mapped to chromosomes 1–10. Eleven *WRKY* genes including 2 group I, 7 group II, and 2 group III genes were situated on chromosome 1. Also, eight genes were assigned to chromosome 2. The highest number of *WRKY* genes belonged to

chromosome 3 (23 genes including 1 group I, 16 group II, and 6 group III genes), whereas only four *WRKY* genes were mapped to chromosome 5. Moreover, five genes were mapped to each of chromosomes 6, 7 and 10. Furthermore, six (wholly from group II genes), seven (1 group I, 1 group II, and 5 group III genes) and eleven (2 group I, 7 group II, and 2 group III genes) *WRKY* genes were found on chromosomes 4, 8 and 9, respectively (Figure 3).

In *Arabidopsis thaliana* and *Oryza sativa* genomes, gene duplication events seem to have a more important role rather than gene expansion (Eulgem *et al.,* 2000; Wu *et al.,* 2005). In this study, however, we found no complete gene duplication. One nearly tandem duplication was observed on chromosome 10 though (*SbWRKY84* and *SbWRKY85*, Figure 3). This result suggested that in *S. bicolor*, the expansion of *SbWRKY* genes has not been necessarily based on gene

Table 1. The WRKY gene families in *Sorghum bicolor* genome.

Gene name	GeneBank ID	Chr.	Group	Location	Exon Count	Gene name	GeneBank ID	Chr.	Group	Location	Exon Count
SbWRKY1	Sb01g000696	1	IIe	676390	3	SbWRKY44	Sb04g009800	4	IIe	12451791	3
SbWRKY2	Sb01g005070	1	IId	4145185	4	SbWRKY45	Sb04g016540	4	IId	38362358	3
SbWRKY3	Sb01g007480	1	Ia	6420226	4	SbWRKY46	Sb04g030930	4	IIe	60932443	3
SbWRKY4	Sb01g007570	1	IIc	6524859	3	SbWRKY47	Sb04g033240	4	IIc	63147733	4
SbWRKY5	Sb01g008550	1	IId	7382814	3	SbWRKY48	Sb04g034440	4	IIb	64267731	5
SbWRKY6	Sb01g012870	1	IIc	11944441	2	SbWRKY49	Sb05g001170	5	III	1225736	2
SbWRKY7	Sb01g014180	1	IId	13370277	4	SbWRKY50	Sb05g001220	5	III	1305458	3
SbWRKY8	Sb01g027770	1	IIe	48360316	2	SbWRKY51	Sb05g017130	5	IIc	42145584	2
SbWRKY9	Sb01g032120	1	Ia	54964329	4	SbWRKY52	WGI	5	III	1212386	1
SbWRKY10	Sb01g036180	1	III	59806395	3	SbWRKY53	Sb06g013835	6	III	38137577	3
SbWRKY11	Sb01g036870	1	III	60434864	3	SbWRKY54	Sb06g019710	6	Ia	49279763	5
SbWRKY12	Sb02g011050	2	III	17809234	2	SbWRKY55	Sb06g024220	6	IIc	53339360	4
SbWRKY13	Sb02g021226	2	Ib	52728488	8	SbWRKY56	Sb06g027290	6	IIe	56196796	4
SbWRKY14	Sb02g022280	2	III	55170648	3	SbWRKY57	Sb06g027710	6	IId	56529226	3
SbWRKY15	Sb02g022290	2	III	55213197	3	SbWRKY58	Sb07g006230	7	IIc	8885371	4
SbWRKY16	Sb02g024760	2	IIa	59264547	3	SbWRKY59	Sb07g006980	7	IId	10712171	1
SbWRKY17	Sb02g024765	2	IIa	59281563	2	SbWRKY60	Sb07g016330	7	Ia	40441219	4
SbWRKY18	Sb02g027950	2	IIc	63184303	5	SbWRKY61	Sb07g019400	7	III	50025392	1
SbWRKY19	Sb02g043030	2	III	76813673	3	SbWRKY62	Sb07g028430	7	Ia	63403512	3
SbWRKY20	Sb03g000240	3	IIb	65767	5	SbWRKY63	Sb08g002520	8	III	2586470	3
SbWRKY21	Sb03g003360	3	IIc	3574879	3	SbWRKY64	Sb08g002560	8	III	2649149	3
SbWRKY22	Sb03g003370	3	IIb	3577819	3	SbWRKY65	Sb08g002570	8	III	2657310	3
SbWRKY23	Sb03g003640	3	IIc	3846956	3	SbWRKY66	Sb08g002590	8	III	2673615	3
SbWRKY24	Sb03g011800	3	IIb	13453957	6	SbWRKY67	Sb08g005080	8	III	6492893	3
SbWRKY25	Sb03g026170	3	IIc	52684056	3	SbWRKY68	Sb08g016240	8	Ia	43243578	4
SbWRKY26	Sb03g026280	3	IIb	52879966	2	SbWRKY69	Sb08g020270	8	IId	51276577	3
SbWRKY27	Sb03g028440	3	IIe	56296964	3	SbWRKY70	Sb09g005700	9	IIc	7601936	3
SbWRKY28	Sb03g028530	3	IIc	56464472	3	SbWRKY71	Sb09g023270	9	Ia	52888054	5
SbWRKY29	Sb03g029920	3	III	58204806	3	SbWRKY72	Sb09g023500	9	III	53146881	3
SbWRKY30	Sb03g030480	3	IIc	58723799	3	SbWRKY73	Sb09g026350	9	IIc	55640076	2
SbWRKY31	Sb03g032800	3	IIc	61257721	3	SbWRKY74	Sb09g026830	9	IIc	56013261	3
SbWRKY32	Sb03g033640	3	IIe	61888491	2	SbWRKY75	Sb09g028660	9	IIc	57513946	3
SbWRKY33	Sb03g033780	3	IIc	62036742	2	SbWRKY76	Sb09g028750	9	IIb	57590098	6
SbWRKY34	Sb03g034670	3	IIe	62841197	3	SbWRKY77	Sb09g029050	9	III	57799582	3
SbWRKY35	Sb03g038170	3	III	66100608	3	SbWRKY78	Sb09g029810	9	IIc	58435466	3
SbWRKY36	Sb03g038180	3	III	66110932	3	SbWRKY79	Sb09g029850	9	IIe	58485871	3
SbWRKY37	Sb03g038190	3	III	66121144	3	SbWRKY80	WGI	9	Ia	39917265	3
SbWRKY38	Sb03g038200	3	III	66129723	3	SbWRKY81	Sb10g004000	10	III	3530829	3
SbWRKY39	Sb03g038210	3	III	66144796	3	SbWRKY82	Sb10g019923	10	IIe	42454390	5
SbWRKY40	Sb03g038510	3	Ia	66433156	4	SbWRKY83	Sb10g020010	10	IIe	42717205	2
SbWRKY41	Sb03g039550	3	IIc	67223920	3	SbWRKY84	Sb10g025590	10	IIa	54957968	3
SbWRKY42	Sb03g047350	3	IIc	74261998	3	SbWRKY85	Sb10g025600	10	IIa	54965360	3
SbWRKY43	Sb04g005520	4	IIa	5399378	2						

Chr.: Chromosome, WGI: without GeneBank ID, *SbWRKY*: *WRKY* genes of *Sorghum bicolor*.

duplication events or maybe duplication of *SbWRKY* genes has happened in the past times.

As suggested by Holub (2001) a gene cluster is defined as a chromosome region with two or more genes located within a 200 kb sequence. Accordingly, based on the Holub's criterion, we found 12 *WRKY* gene clusters containing a total of 29 genes. Only one gene cluster was found on each of chromosomes 1, 5, 8 and 10, whereas chromosomes 2, 3 and 9 contained two, four and two gene clusters, respectively (Figure 3).

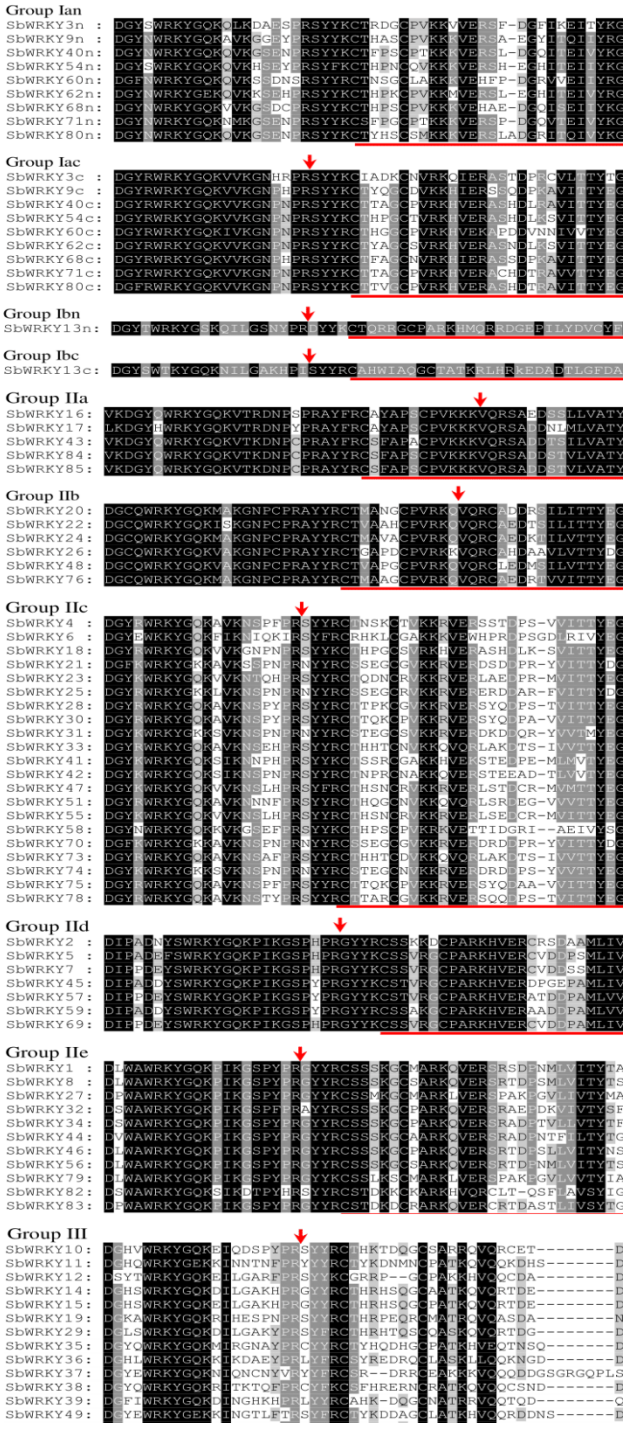

Figure 1. Alignment of *Sorghum bicolor* WRKY domains. Alignment was done using the Clustal W program and was highlighted using Gendoc software. The conserved residues within each group have been shown in black color. The position of the conserved intron was indicated by a small vertical red arrowhead line. The horizontal red lines indicate the conserved zinc finger motif positions.

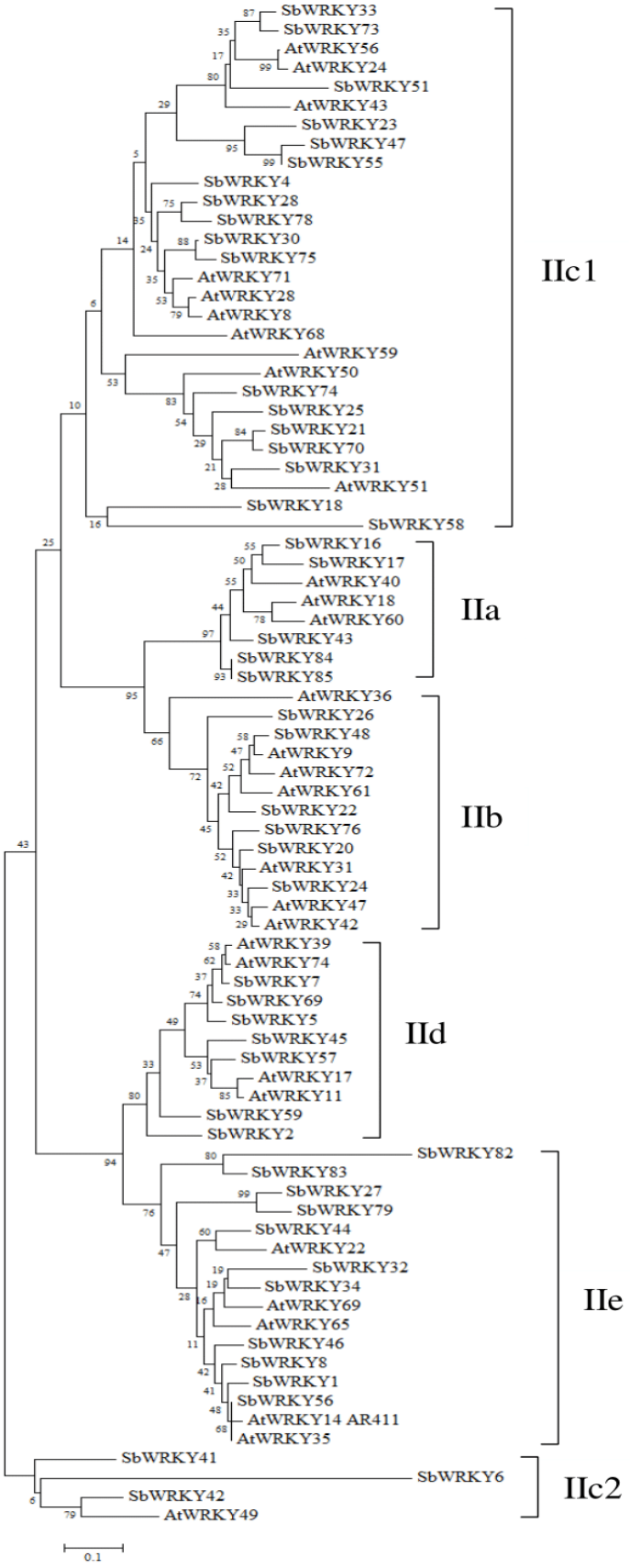

Figure 2. Phylogenetic tree of *Arabidopsis* WRKY and *Sorghum bicolor* WRKY proteins. The phylogenetic tree was drawn using the Neighbor-Joining (NJ) method with 1000 bootstrap replicates in MEGA 6.0 software. The AtWRKY protein sequences were downloaded from www.arabidopsis.org.

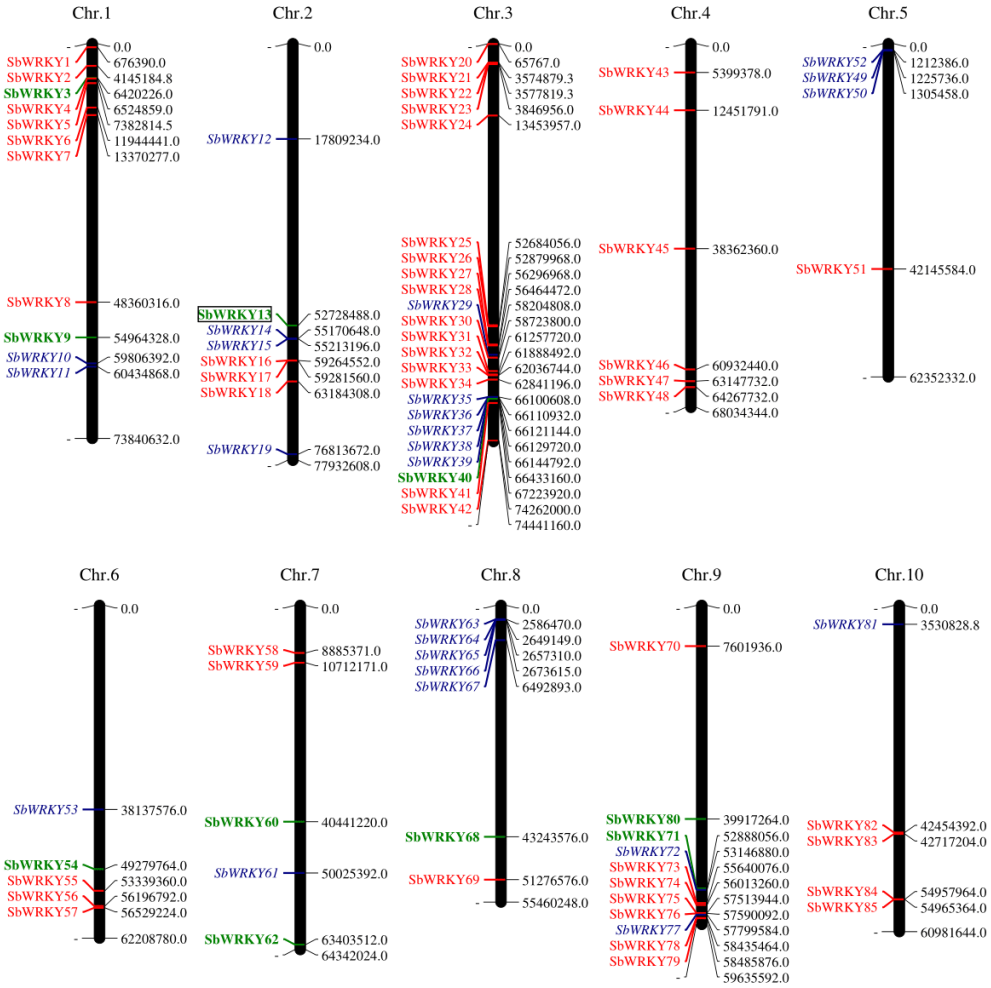

Figure 3. Chromosomal locations of *Sorghum bicolor WRKY* genes. The WRKY genes are shown on the left side of each chromosome. The chromosome numbers have been shown on the top of each chromosome. The color assigned to each gene shows the groups to which each *WRKY* gene belongs (group I in green (Bold); group II in red and group III in blue (Italic)). The numbers on the right side correspond to the start position of each *WRKY* gene. The *SbWRKY*13 is displayed inside a rectangle because of its zinc finger motif type.

No clusters were found on chromosomes 4, 6 and 7. Gene cluster analysis showed that the number of genes on chromosomes was positively correlated with the number of gene clusters (Pearson correlation=0.882, *P-Value*=0.001).

Amino acid composition (AAC) of the SbWRKY proteins

The study of the AAC for the SbWRKY proteins (presented in Table 2 and Figures 4 and 5) revealed that the frequencies of nine amino acids were statistically different among SbWRKY proteins belonging to different groups (Figures 4 and 5). Totally in all groups, Alanine and Proline were the most frequent residues while Cysteine had the lowest frequency (Figure 4). Comparisons between mean percentages of the amino acids showed that threonine, glutamate and aspartate residues were significantly more frequent in SbWRKY III while a major part of the SbWRKY II proteins had been constituted by alanine and proline. Since glutamate and aspartate belong to the negatively charged amino acid group therefore, SbWRKY III proteins should have a negative net charge. Nakashima and Nishikawa (1992) reported that the AAC was different between cytoplasmic and extracellular peptides of membrane proteins. Alanine and Arginine residues were preferentially sited on the Cytoplasmic side, while the threonine and Cysteine/Cystine were preferentially sited on the extracellular side.

Research studies have shown that plants significantly accumulate proline when subjected to environmental stresses especially drought stress (Routley, 1966), Also, Thomas and Shanmugasundaram (1991) believe

that as osmoregulation, alanine is able to reduce the damage caused by salt stress. Therefore, high amounts of these two amino acids in the studied SbWRKY proteins might have an association with the response of the plant against drought or salinity stresses.

Phylogenetic Analysis

The WRKY domain phylogenetic tree can be divided into nine clades: Ian, Iac, IIa, IIb, IIc1, IIc2, IId, IIe, and III. Proteins of the group I contain two different WRKY domains placed at the N-terminal domain (In) or the C-terminal (Ic). Clade Iac contained 10 members including nine members with WRKY domains at the C-terminal region and one group IIc member (SbWRKY18). This group IIc member was clustered with SbWRKY62Iac, demonstrating a common origin of their domains. Also, Clade IaN had 11 members including nine members with WRKY domains at the N-terminal region and two group IIc members (SbWRKY6 and SbWRKY58) (Figure 6). This result confirms the above conclusion about the origin of the subgroup IIc.

Group II can be divided into six clades. Figure 6 shows that clusters IIa and IIb can be combined into one clade suggesting the domains had a recent gene ancestor. On the other hand, each of the clusters IId and IIe were placed in separate clades. The 21 members of the group IIc were clustered differently. Two group IIc members (SbWRKY41 and SbWRKY42) formed one clade (IIc2), while the rest members of the group formed the other clade (IIc1). Although all members of IIc1 and IIc2 belong to group IIc, however, they clustered into two clades, indicating the existence of divergence in group IIc. Also, Clade III contained 27 members comprising of 25 group III members together with IbN and IbC domains, representing a high affinity between domains of group III and group Ib. Moreover, the phylogenetic tree showed that clade III was the nearest neighbor to the clades IId and IIe, suggesting that they have had a common ancestor before divergence.

Intron splicing sites in the SbWRKY domains

For intron analysis, the unspliced *SbWRKY* gene sequences were compared with the corresponding CDS sequences. The results showed that the majority of *SbWRKY* genes had two types of the intron in their WRKY domains (phase 0 and phase 2). A phase 0 intron was located before the V position (the V-type intron), five amino acids after the second C residue in the C2H2 zinc finger motif (Figure 1). Another conserved intron which was a phase 2 intron was found after the R residue (the R-type intron) exactly

Table 2. Analysis of variance (ANOVA) for amino acid composition pattern in different groups of the *Sorghum bicolor* WRKY proteins.

Source	df	Mean of square																		
		Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Between group	2	32.21*	3.39**	14.74**	1.1	0.89	19.63	0.45	5.65*	7.54*	5.55	0.02	7.08**	27.75*	15.76**	3.45	0.88	7.94*	1.26	0.61	2.71
Within group	82	9.93	0.68	2.35	2.5	1.28	7.55	1.82	1.52	1.84	3.48	0.94	1.34	6.54	1.97	2.04	4.7	2.49	2.3	0.24	1.39
Total	84																				

* and **: Significant differences at 0.05 and 0.01 level of probability.

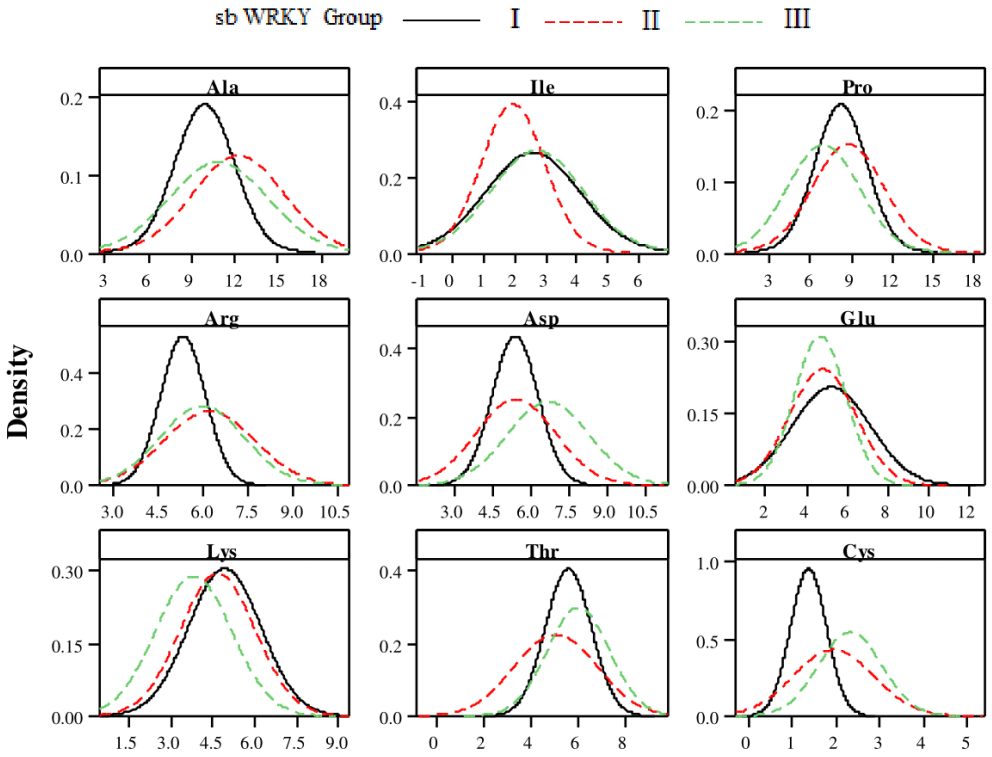

Figure 4. Histogram showing distributions of the frequency of nine amino acids among different groups of the *Sorghum bicolor* WRKY proteins.

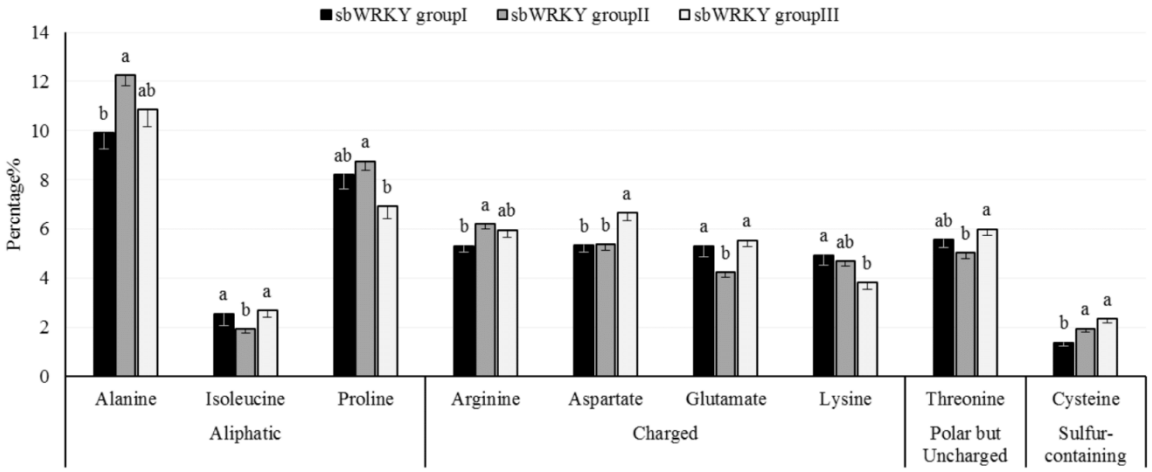

Figure 5. Amino acid composition pattern in different groups of the *Sorghum bicolor* WRKY proteins. Comparisons were performed according to LSD method (α=0.05). Columns that do not share any letter are statistically different.

at the 9th residue position after the WRKYGQK motif (Figure 1). We found no intron in the WRKY domain of subgroup Ian. In group III as well as subgroups Iac, Ibn, Ibc, IIc, IId, IIe, an R-type intron was found eight residues after the WRKYGQK motif region in the WRKY domain of genes. In subgroups IIa and IIb, a V-type intron was detected five amino acids after the second C, and 17 residues before the first H within the zinc finger motif region in the WRKY domain (Figures 1 and 6).

So far, no V type intron has been reported in *Arabidopsis WRKY* genes. Also in *Arabidopsis* subgroup IIa and IIb *WRKY* genes, R-type introns are placed at the fourth amino acid (K residue) after the

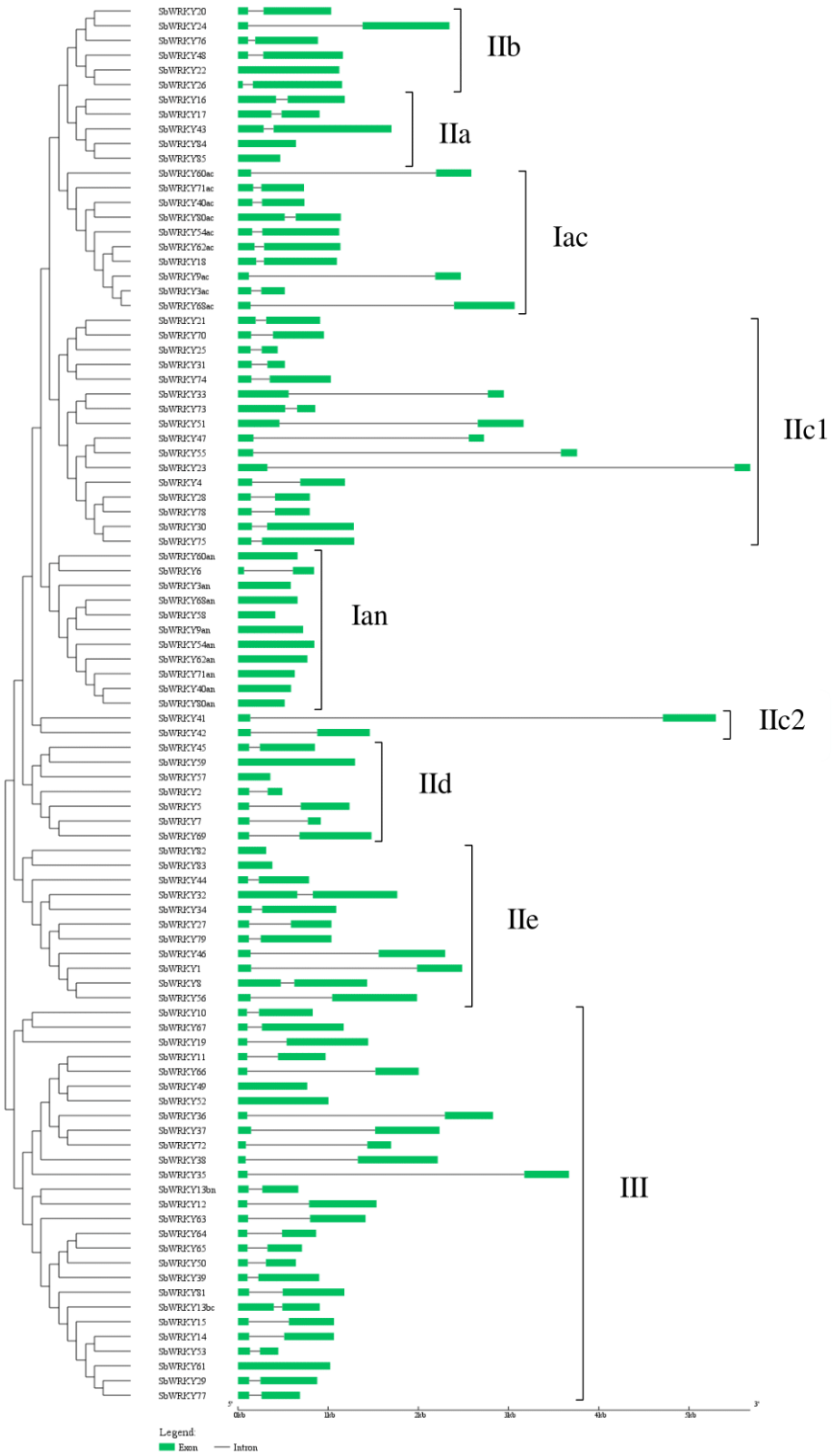

Figure 6. Phylogenetic tree plus Exon-intron structures of WRKY domain genes of *Sorghum bicolor*. Exons and introns were represented by green boxes and black lines, respectively. First, the phylogenetic tree was constructed using the Neighbor-Joining (NJ) method with 1000 bootstrap replicates in MEGA 6.0 software. Then the tree file in Newick format together with the *Sorghum bicolor* WRKY domain genes was inserted in Gene Structure Display Server 2.0 program (http://gsds.cbi.pku.edu.cn/).

second C residue in the zinc finger motif (Eulgem *et al.,* 2000). Therefore, introns in subgroup IIa and IIb *WRKY* genes may have different origins. Also, eight studied *SbWRKY* genes have lost their introns. Intron loss can be considered as the consequence of intron turnover, the result of homologous recombination between an intron-containing allele and a mature mRNA (Wu *et al.,* 2005).

Conserved Motifs in *S. bicolor* proteins outside of the region of the WRKY domain

In this study, we investigated the presence of conserved known motifs outside of the region of the main SbWRKY domain. Except for the conserved 60 amino acid residues of the main WRKY domain, no important known motif was previously reported from the remainder of the WRKY protein sequences. However, the results of the present study showed that there were some known motifs that should be taken into consideration. Based on the results, no conserved motif was detected for groups Ia, IIe and III. However other groups had some known conserved motifs. For example, for groups IIa, IIb and IIc we found six, two and one motifs outside of the main WRKY domain region (Figure 7 and Table 3). Interestingly, we observed a DNA polymerase III motif (subunits gamma and tau) in group IIb. Analysis of the function of this motif in *SbWRKY* genes would be interesting. Also, a GCM motif was found for group IIc before the conserved WRKY domain. The GCM motif is a DNA binding domain that recognizes preferentially the non-palindromic octamer 5'-ATGCGGGT-3' therefore, the role of this motif in the mentioned group may be interesting.

In conclusion, we identified genome-wide WRKY transcription factors from the *Sorghum bicolor* genome and analyzed phylogenetic relationships, gene structures, intron splicing sites and gene duplication events. The amino acid composition as well as the conserved motifs in proteins were also analyzed. Our results can be used for further studies on evolutionary relationships and systematic taxonomy of WRKY transcription factors in other sorghum species.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers whose constructive comments and suggestions have improved the present article.

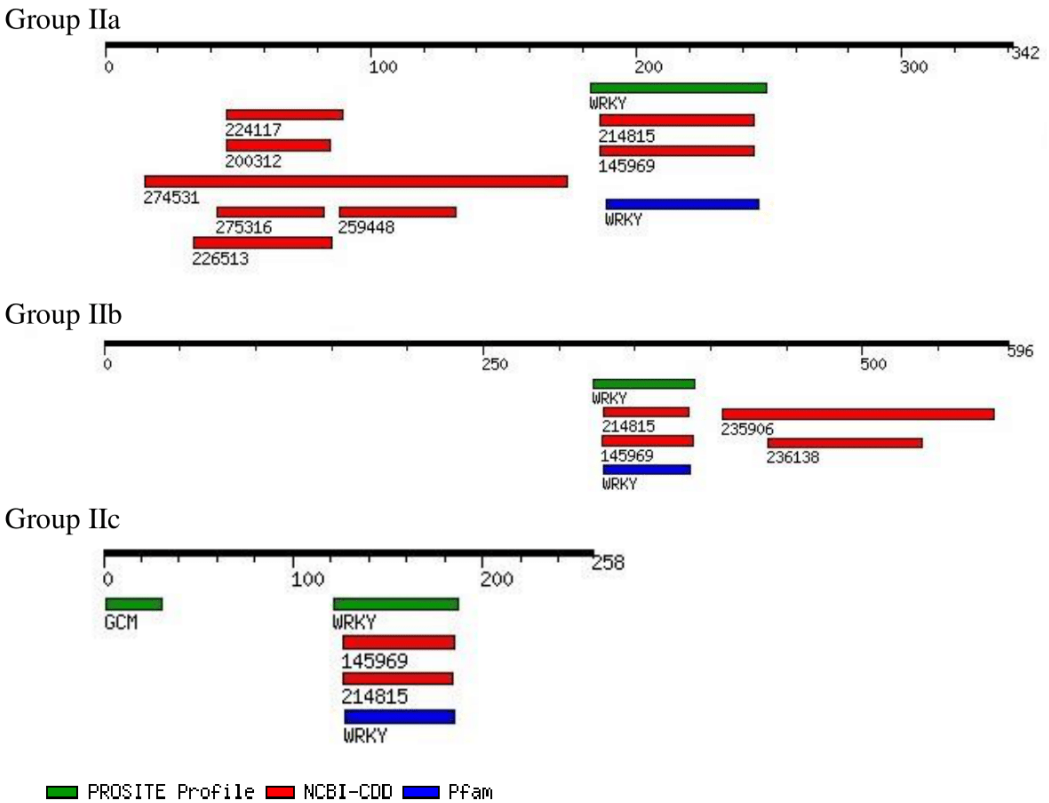

Figure 7. Distribution of conserved known motifs in *Sorghum bicolor* WRKY proteins. Motifs were identified by using MOTIF search program. Among all SbWRKY groups, only those groups (IIa IIb and IIc) that had a known motif on the outside of the main SbWRKY domain were displayed. The characteristics of each motif are shown in Table 3.

Table 3. Characteristics of SbWRKY motif in *Sorghum bicolor* proteins.

Group	Motif library	Motif	Position (Score, E-value)	Description
IIa	PROSITE PROFILE	WRKY	183..249(2324)	PS50811, WRKY domain profile.
	NCBI-CDD	200312	48..87(33.1, 0.22)	TIGR03689, pup_AAA, proteasome ATPase.
		224117	48..92(31.2, 0.92)	COG1196, Smc, chromosome segregation ATPases [Cell division and chromosome partitioning].
		226513	35..87(31.0, 0.81)	COG4026, COG4026, uncharacterized protein containing TOPRIM domain, potential nuclease [General function prediction only].
		259448	90..134(31.5, 0.45)	pfam15315, FRG2, facioscapulohumeral muscular dystrophy candidate 2.
		274531	17..176(32.7, 0.30)	TIGR03348, VI_IcmF, type VI secretion protein IcmF.
		275316	44..84(31.9, 0.54)	TIGR04523, conserved hypothetical protein, helix-rich Mycoplasma protein.
	Pfam	WRKY	189..246(1.7e-25)	PF03106, WRKY DNA -binding domain
IIb	PROSITE PROFILE	WRKY	323..389(2603)	PS50811, WRKY domain profile.
	NCBI-CDD	214815	329..386(129, 9e-36)	smart00774, WRKY, DNA binding domain.
		145969	328..388(123, 1e-33)	pfam03106, WRKY, WRKY DNA -binding domain.
		235906	408..587(42.9, 5e-04)	PRK07003, PRK07003, DNA polymerase III subunits gamma and tau; validated.
		236138	438..539(42.5, 6e-04)	PRK07994, PRK07994, DNA polymerase III subunits gamma and tau; validated.
	Pfam	WRKY	329..387(5.6e-26)	PF03106, WRKY DNA -binding domain
IIc	PROSITE PROFILE	WRKY	122..187(3299)	PS50811, WRKY domain profile.
		GCM	1..31(575)	PS50807, GCM domain profile.
	NCBI-CDD	145969	127..185(131, 1e-38)	pfam03106, WRKY, WRKY DNA -binding domain.
		214815	127..184(124, 9e-36)	smart00774, WRKY, DNA binding domain.
	Pfam	WRKY	128..185(3.8e-26)	PF03106, WRKY DNA -binding domain

REFERENCES

An J.-P., Zhang X.-W., You C.-X., Bi S.-Q., Wang X.-F., and Hao Y.-J. (2019). MdWRKY40 promotes wounding-induced anthocyanin biosynthesis in association with MdMYB1 and undergoes MdBT2-mediated degradation. *New Phytologist,* 224: 380-395.

Chanwala J., Satpati S., Dixit A., Parida A., Giri M. K., and Dey N. (2020). Genome-wide identification and expression analysis of WRKY transcription factors in pearl millet (*Pennisetum glaucum*) under dehydration and salinity stress. *BMC Genomics,* 21: 231.

Chen M., Tan Q., Sun M., Li D., Fu X., Chen X., Xiao W., Li L., and Gao D. (2016). Genome-wide identification of WRKY family genes in peach and analysis of WRKY expression during bud dormancy. *Molecular Genetics and Genomics,* 291: 1319-1332.

Chinnapandi B., Bucki P., Fitoussi N., Kolomiets M., Borrego E., and Braun Miyara S. (2019). Tomato SlWRKY3 acts as a positive regulator for resistance against the root-knot nematode *Meloidogyne javanica* by activating lipids and hormone-mediated defense-signaling pathways. *Plant Signaling & Behavior,* 14: 1601951.

Ciolkowski I., Wanke D., Birkenbihl R. P., and Somssich I. E. (2008). Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. *Plant Molecular Biology,* 68: 81-92.

Eulgem T., Rushton P. J., Robatzek S., and Somssich I. E. (2000). The WRKY superfamily of plant transcription

factors. *Trends in Plant Science,* 5: 199-206.

Eulgem T., Rushton P. J., Schmelzer E., Hahlbrock K., and Somssich I. E. (1999). Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. *The EMBO Journal,* 18: 4689-4699.

Finn R. D., Bateman A., Clements J., Coggill P., Eberhardt R. Y., Eddy S. R., Heger A., Hetherington K., Holm L., Mistry J., Sonnhammer E. L. L., Tate J., and Punta M. (2014). Pfam: the protein families database. *Nucleic Acids Research,* 42: D222-D230.

Gao Y.-F., Liu J.-K., Yang F.-M., Zhang G.-Y., Wang D., Zhang L., Ou Y.-B., and Yao Y.-A. (2020). The WRKY transcription factor WRKY8 promotes resistance to pathogen infection and mediates drought and salt stress tolerance in *Solanum lycopersicum*. *Physiologia Plantarum,* 168: 98-117.

Gu L., Dou L., Guo Y., Wang H., Li L., Wang C., Ma L., Wei H., and Yu S. (2019). The WRKY transcription factor *GhWRKY27* coordinates the senescence regulatory pathway in upland cotton (*Gossypium hirsutum* L.). *BMC Plant Biology,* 19: 116.

Holub E. B. (2001). The arms race is ancient history in *Arabidopsis*, the wildflower. *Nature Reviews Genetics,* 2: 516-527.

Hu J., Fang H., Wang J., Yue X., Su M., Mao Z., Zou Q., Jiang H., Guo Z., Yu L., Feng T., Lu L., Peng Z., Zhang Z., Wang N., and Chen X. (2020). Ultraviolet B-induced MdWRKY72 expression promotes anthocyanin synthesis in apple. *Plant Science,* 292: 110377.

Mangelsen E., Kilian J., Berendzen K. W., Kolukisaoglu Ü. H., Harter K., Jansson C., and Wanke D. (2008). Phylogenetic and comparative gene expression analysis of barley (*Hordeum vulgare*) WRKY transcription factor family reveals putatively retained functions between monocots and dicots. *BMC Genomics,* 9: 194.

Mullet J. E., Klein R. R., and Klein P. E. (2002). Sorghum bicolor—an important species for comparative grass genomics and a source of beneficial genes for agriculture. *Current Opinion in Plant Biology,* 5: 118-121.

Nakashima H., and Nishikawa K. (1992) The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Letters,* 303: 141-146.

Romero I., Alegria-Carrasco E., Gonzalez de Pradena A., Vazquez-Hernandez M., Escribano M. I., Merodio C., and Sanchez-Ballesta M. T. (2019). WRKY transcription factors in the response of table grapes (cv. *Autumn Royal*) to high CO_2 levels and low temperature. *Postharvest Biology and Technology,* 150: 42-51.

Routley D. (1966). Proline accumulation in wilted ladino clover leaves. *Crop Science,* 6: 358-61.

Rushton P. J., Somssich I. E., Ringler P., and Shen Q. J. (2010). WRKY transcription factors. *Trends in Plant Science,* 15: 247-258.

Song H., Wang P., Nan Z., and Wang X. (2014). The WRKY transcription factor genes in lotus japonicus. *International Journal of Genomics,* 2014: 420128.

Sun W., Ma Z., Chen H., and Liu M. (2020). Genome-wide investigation of WRKY transcription factors in Tartary buckwheat (*Fagopyrum tataricum*) and their potential roles in regulating growth and development. *PeerJ,* 8: e8727.

Tamura K., Stecher G., Peterson D., Filipski A., and Kumar S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution,* 30: 2725-2729.

Thomas S. P., and Shanmugasundaram S. (1991). Osmoregulatory role of alanine during salt stress in the nitrogen fixing cyanobacterium *Anabaena* sp. 287. *Biochemistry International,* 23: 93-102.

Vandenabeele S., Van Der Kelen K., Dat J., Gadjev I., Boonefaes T., Morsa S., Rottiers P., Slooten L., Van Montagu M., and Zabeau M. (2003). A comprehensive analysis of hydrogen peroxide-induced gene expression in tobacco. *Proceedings of the National Academy of Sciences,* 100: 16113-16118.

Voorrips R. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *Journal of Heredity,* 93: 77-78.

Wang J., Tao F., Tian W., Guo Z., Chen X., Xu X., Shang H., and Hu X. (2017). The wheat WRKY transcription factors TaWRKY49 and TaWRKY62 confer differential high-temperature seedling-plant resistance to *Puccinia striiformis* f. sp. tritici. *Public Library of Science (PloS) one,* 12: e0181963-e.

Wu K.-L., Guo Z.-J., Wang H.-H., and Li J. (2005). The WRKY family of transcription factors in rice and *Arabidopsis* and their origins. *DNA Research,* 12: 9-26.

Xie Z., Zhang Z.-L., Zou X., Huang J., Ruas P., Thompson D., and Shen Q. J. (2005). Annotations and functional analyses of the rice *WRKY* gene superfamily reveal positive and negative regulators of abscisic acid signaling in aleurone cells. *Plant Physiology,* 137: 176-89.

Xu Y.-H., Sun P.-W., Tang X.-L., Gao Z.-H., Zhang Z., and Wei J.-H. (2020). Genome-wide analysis of WRKY transcription factors in *Aquilaria sinensis* (Lour.) Gilg. *Scientific Reports,* 10: 3018.

Zhang Y., and Wang L. (2005). The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evolutionary Biology,* 5: 1.