Physiol Mol Biol Plants (January–March 2016) 22(1):33–49
DOI 10.1007/s12298-016-0349-z

RESEARCH ARTICLE

Genome- wide characterization of Nuclear Factor Y (NF-Y) gene family of sorghum [*Sorghum bicolor* (L.) Moench]: a bioinformatics approach

Neha Malviya[1] · **Parul Jaiswal**[1] · **Dinesh Yadav**[1]

Received: 12 September 2015 / Revised: 11 March 2016 / Accepted: 28 March 2016 / Published online: 1 April 2016

Abstract Nuclear factor Y (NF-Y) is a heterotrimeric transcription factor (TF) complex with preferential binding to CCAAT elements of promoters, regulating gene expression in most of the higher eukaryotes. The availability of plant genome sequences have revealed multiple number of genes coding for the three subunits, namely NF-YA, NF-YB and NF-YC in contrast to single NF-Y gene for each subunit reported in yeast and animals. A total of 33 NF-YTF comprising of 8 NF-YA, 11 NF-YB and 14 NF-YC subunits were accessed from the sorghum genome. The bioinformatic characterization of NF-Y gene family of sorghum for gene structure, chromosome location, protein motif, phylogeny, gene duplication and *in-silico* expression under abiotic stresses have been attempted in the present study. The identified *SbNF-Y* genes are distributed on all the 10 chromosomes of sorghum with variability in the frequency and 18 out of 33 *SbNF-Y*s were found to be intronless. Segmental duplication event was found to be predominant feature based on gene duplication pattern study. Several orthologs and paralogs groups were disclosed through the comprehensive phylogenetic analysis of SbNF-Y proteins along with 36 *Arabidopsis* and 28 rice NF-Y proteins. *In-silico* expression analysis under abiotic stresses using rice transcriptome data revealed several of the sorghum *NF-Y* genes to be associated with salt, drought, cold and heat stresses.

Keywords NF-Y · CCAAT box · Orthologs · Paralogs · Phylogenetic tree · Transcription factor

Electronic supplementary material The online version of this article (doi:10.1007/s12298-016-0349-z) contains supplementary material, which is available to authorized users.

✉ Dinesh Yadav
dinesh_yad@rediffmail.com

[1] Department of Biotechnology, D.D.U Gorakhpur University, Gorakhpur, Uttar Pradesh 273 009, India

Introduction

Nuclear Factor Y (NF-Y) is a transcription factor complex with three subunits namely NF-YA, NF-YB and NF-YC that binds to the CCAAT box of eukaryotic promoters and regulate gene expression (Dolfini and Mantovani 2013; Laloum et al. 2013; Petroni et al. 2012). NF-YA also called HAP3 (Heme Activator Protein 3) or CBF-A (CCAAT-Binding Factor A), are characterized having two domains which are highly conserved in all higher eukaryotes (Maity and de Crombrugghe 1998). The N-terminal domain is involved in NF-YB: NF-YC heterodimer interaction and the C-terminal domain is concerned with DNA binding site recognition (Xing et al. 1994). At the stage of transcription activation NF-YB forms a heterotrimer, with NF-YA and NF-YC (Kim et al. 1996). Like NF-YC proteins consist of HFM domain, but these are more closely related to the core histone H2A (Dolfini et al. 2012).

These subunits are required for their associationship and transcriptional regulation in both vertebrates and plants (Sinha et al. 1995). A single gene encodes each NF-Y subunit in animals and yeast, but in plants the three subunits are encoded by multiple genes (Edwards et al. 1998; Gong et al. 2004; Riechmann et al. 2000). Based on the availability of plant genome sequences NF-Y gene families of *Arabidopsis*, rice, wheat, *Brachypodium distachyon*, canola, soybean, common bean have been reported (Cao et al. 2011; Liang et al. 2014; Ripodas et al. 2014; Siefers et al. 2009; Stephenson et al. 2007; Thirumurugan et al. 2008).

Individual NF-Y subunits are identified to be involved in number of important plant processes, yet no complete NF-Y

complex for gene regulation has been studied in the plants. The functional diversity of NF-Y genes corresponding to the three subunits have been elucidated in different plants (Laloum et al. 2013; Petroni et al. 2012). The role of NF-Y in embryogenesis, photoperiod-dependent flowering time regulation, seed oil production, root nodule development, chloroplast biogenesis, seed germination (Petroni et al. 2012), heat stress (Sato et al. 2014), drought tolerance (Zhang et al. 2015), enhanced yield (Yadav et al. 2015) have been reported.

Sorghum bicolor is a grass species belonging to the subtribe Saccharinae, within the grass family Poaceae. Sorghum is a diploid with 10 chromosomes and has a genome of about 730 Mb (Paterson et al. 2009). There are few studies related with transcription factors like MYB, AUX_ARF, bZIP, AP2, WRKY, basic helix loop helix, NAC and Dof in sorghum (Kushwaha et al. 2011; Lu et al. 2013; Sekhwal et al. 2015; Yan et al. 2013) though genome wide analysis of NF-Y gene family of sorghum is still lacking. This manuscript reports genome wide *in-silico* characterization of *SbNF-Y* gene family highlighting putative gene structures, chromosomal localization, motif analysis, gene duplication, ancestral protein sequence analysis, comprehensive phylogenetic analysis with rice and *Arabidopsis* NF-Y gene family. The *in-silico* expression profiling of NF-Y genes of sorghum under stress conditions have also been assessed based on the available rice transcriptome data after deciphering the corresponding orthologs identified in the phylogenetic tree.

Materials and methods

Databases searches for the identification of NF-Y gene family of sorghum and its annotation

The NF-Y A, B and C sequences of *A.thaliana* were retrieved from plant transcription factor database Plant TFDB (http://plantfdb.cbi.pku.edu.cn/) version 3.0 (Jin et al. 2014). The nucleotide sequences of NF-Y domain were used to search the potential NF-Y domain homologs hit in the whole genome shotgun (wgs) sequence and nucleotide collection (nr/nt) of sorghum through tblastn at the NCBI database. The 2 kb upstream and 2 kb downstream sequences of NF Y domain homologs were retrieved from whole genome shotgun sequence of sorghum for fishing out the putative NF-Y genes using BioEdit software version 7.2.5 (Hall 1999). The NF-Y gene sequences identified were tentatively designated as *SbNF-YA*s, *SbNF-YB*s and SbNF-*YCs* with corresponding numbers. The annotated sequences were further subjected to bioinformatics server namely FGENESH (Solovyev et al. 2006) for prediction of full length genes with putative CDS and protein sequences. The putative NF-Y protein sequences of sorghum were subjected to protein functional analysis using PFAM version 27 (Punta et al. 2012), PROSITE version 20.93 (Sigrist et al. 2013), INTERPROSCAN version 4.0 (Quevillon et al. 2005) and MOTIFSCAN databases (Falquet et al. 2002). The physio-chemical feature of NF-Y proteins like isoelectric point (PI) and molecular weight were calculated using Expasy server (Gasteiger et al. 2003).

Chromosomal distribution and intron/exon gene structure prediction

The annotated *SbNF-Y* genes were positioned on chromosomes through NCBI-BLAST search and manually marked. FGENESH server was used to analyze the gene structure of predicted *SbNF-Y* genes.

NF-Y protein alignment and phylogenetic analysis

ClustalX 2.0.10 (Thompson et al. 1997) was utilized for Multiple sequence alignment of predicted SbNF-Y proteins. Neighbor joining approach (with 500 reiterations) was used to construct phylogenetic tree using MEGA 6.0 software (Tamura et al. 2013).

Inference of duplication time

The Ks and Ka value of paralogous sequences were calculated by DnaSP software version 5.0 (Librado and Rozas 2009). The Ks value was calculated which was further used to calculate the approximate date of the duplication event ($T = Ks/2\lambda$), using the mean value of clock-like rates (λ) of synonymous substitution (6.5×10^{-9}) (Gaut et al. 1996).

Motif prediction and subcellular localization

Motif analysis of SbNF-YA, SbNF-YB and SbNF-YC protein sequences were analyzed through MEME (Multiple EM for Motif Elicitation) program software version 4.4.0 (Bailey and Elkan 1994). For the identification of conserved motifs the maximum number of motifs was set for 10, with a width range between 5 to 55, while other factors were at default selections. The NF-Y protein sequences were passed through the available prediction algorithms and the raw data were used to make preliminary location predictions. ARAMEMNON (http://aramemnon.botanik.uni-koeln.de) server (Schwacke et al. 2003) was used for subcellular localization and identification of transmembrane domain.

Prediction of ancestral protein sequences

For the prediction of ancestral sequences, FAST-ML 2.02 (www.tau.ac.il/~talp/supplementary/fastml/fastml.2.02) (Pupko et al. 2002) was used. The unrooted input tree was used, which was automatically rooted by midpoint rooting in FAST-ML Ancestral Sequence Prediction analysis.

In-silico expression analysis

Due to non-availability of transcriptome data for the identified *SbNF-Y* genes of sorghum, attempts was made to analyze the transcriptome data of rice orthologs to generate heatmap as both the crops belong to same family and show high degree of synteny (Wang et al. 2009). The coding sequences of the orthologous pair of sorghum genes were used as queries for BLAST search and the corresponding rice gene sequences were identified based on more than 90 % sequence identities. The Rice eFP browser (http://bar.utoronto.ca/efprice/cgi-bin/efpWeb.cgi) (Toufighi et al. 2005) was used to retrieve RMA normalized expression data for abiotic stresses namely drought stress, salt stress, cold stress and heat shock recorded for seedling stages. All transcript data was analyzed with HCE version 2.0 beta web tools (Seo et al. 2004).

Result and discussion

Genome wide characterization of NF-Y transcription factor genes of sorghum

The nucleotide and amino-acid sequences of the conserved NF-Y domains corresponding to A, B and C subunits subjected to BLAST search resulted into a total of 33 *NF-Y* genes comprising of eight *NF-YA*, 11 *NF-YB* and14 *NF-YC* from the whole genome of sorghum. The predicted sorghum eight *NF-YA* genes, 11 *NF-YB* and 14 *NF-YC* genes were named as *SbNF-YA1* to *SbNF-YA8*, *SbNF-YB1* to *SbNF-YB11* and *SbNF-YC1* to *SbNF-YC14* respectively. There exists great variability in terms of number of *NF-YA*, *NF-YB* and *NF-YC* subunits in different crops. A total of 10 *NF-YA*, 13 *NF-YB* and 13 *NF-YC* genes were predicted in *Arabidopsis* (Siefers et al. 2009), 14 *NF-YA*, 14 NF-YB, five NF-YC in *Brassica napus* (Liang et al. 2014) while in *Brachypodium* seven *NF-YA*, 17 *NF-YB* and 12 *NF-YC* genes were annotated (Cao et al. 2011). In case of wheat 10 *NF-YA*, 11 *NF-YB* and 14 *NF-YC* genes have been reported (Stephenson et al. 2007) while in rice, there exists five *NF-YA*, 10 *NF-YB* and 10 *NF-YC* genes (Yang et al. 2005). Recently genome wide analysis of NF-Y genes of two legumes namely soybean (21 *GmNF-YA*, 32 *GmNF-YB* and 15 *GmNF-YC*) and common bean (nine *NF-YA*, 14 *NF-YB* and seven *NF-YC*) has been reported (Quach et al. 2015; Ripodas et al. 2014). The presence of the conserved NF-YA, NF-YB and NF-YC domains in the predicted SbNF-Y protein was a typical feature for considering it as a family member of NF-Y transcription factor.

The annotated sequences were further subjected to FGENESH server for prediction of ORFs, CDS and protein sequences. The result of FGENESH for *SbNF-YA*, *SbNF-YB* and *SbNF-YC* gene sequences of sorghum is summarized in Table 1.The identified *SbNF-Y* genes encodes peptides ranging from 90 to 790 aa with the pI value varying from 4.26 to 10.83, and the molecular weight ranging from 10.21 kD to 83.52 kDa as estimated from expasy server (http://web.expasy.org/compute_pi/).

Protein functional analysis and subcellular localization

The protein functional analyses of these 33 putative SbNF-Y proteins have been performed using different servers namely PFAM version 27, PROSITE version 20.93, INTERPROSCAN version 4.0, MOTIFSCAN databases and NUCPRED software. The members of SbNF-YA proteins subjected to INTERPROSCAN revealed identity with InterPro accession number IPR001289. These proteins were also similar to signature accession namely PS51152 from PROSITE database and Pfam database confirming their identity to NF-YA like proteins. All the SbNF-YA, SbNF-YB and SbNF-YC proteins showed nuclear localization signal (NLS) with RRR amino-acid residuewhen subjected to NUCPRED software. Further Motifscan integrated with PeroxiBase profiles, PROSITE patterns, PROSITE profiles, HAMAP profiles, Pfam HMMs (local models), Pfam HMMs (global models) databases revealed the presence of alanine, arginine, glutamine and serine rich regions in different SbNF-YA proteins.

Similarly, for SbNF-YB proteins showed identity with InterPro accession number IPR003957 along with a signature accession namely PF00808 for Pfam database and PS00685 for PROSITE database confirming their identity to NF-YB like proteins. In case of SbNF-YB proteins only six out of 11 sequences showed NLS signal with variable amino-acid residues. The NLS signal for SbNF-YB1, SbNF-YB3, SbNF-YB5 and SbNF-YB6 was found to be KRK while for SbNF-YB7 and SbNF-YB8, the NLS signal of KRRK and RRK was observed. Motifscan integrated with PeroxiBase profiles, PROSITE patterns, PROSITE profiles, HAMAP profiles, Pfam HMMs (local models), Pfam HMMs (global models) databases revealed the presence of asparagine, glutamine, glycine, methionine, histidine, alanine and proline rich regions in different SbNF-YB proteins.

The SbNF-YC sequences showed identity with InterPro accession number IPR027170 along with PF00808 accession of Pfam database while PROSITE database could not identify any putative hits. Only five SbNF-YC sequences (SbNF-YC1, SbNF-YC3, SbNF-YC5, SbNF-YC6, SbNF-YC8) out of 14 have NLS with KRR residue when subjected to NUCPRED software. Motifscan integrated with PeroxiBase profiles, PROSITE patterns, PROSITE profiles, HAMAP profiles, Pfam HMMs (local models), Pfam HMMs (global models) databases revealed the presence of glutamine, proline, alanine, aspartic acid, glycine and leucine in different SbNF-YC proteins.

Table 1 List of *NF-Y* genes identified in sorghum with their corresponding proteins, CDS and chromosome positions

Sb NF-Y Gene	Source accession number	Chr. no.	Chromosome location (bp)	No. of introns	mRNA length(bp)	Amino acid length (a.a)	Protein Mol. Wt.(kDa)	Protein PI value
SbNF-YA1	XM 002465725.1	1	68598278–68598532	0	690	229	24.87	8.68
SbNF-YA2	ABXC01000113.1	1	10088592–10088732	1	273	90	10.21	10.83
SbNF-YA3	ABXC01000830.1	2	3779232–3779891	2	636	211	22.83	9.96
SbNF-YA4	ABXC01005669.1	8	52974564–52974779	1	360	119	13.05	10.05
SbNF-YA5	ABXC01001587.1	2	73106344–73106742	5	918	305	33.37	8.49
SbNF-YA6	ABXC01000526.1	1	55640044–55640457	4	783	260	28.35	9.76
SbNF-YA7	ABXC01000138.1	1	12431437–12431853	1	549	182	19.61	9.94
SbNF-YA8	ABXC01005686.1	8	53656544–55656918	3	873	290	30.78	9.88
SbNF-YB1	ABXC01000523.1	1	55505670–55506635	4	2373	790	83.52	5.38
SbNF-YB2	ABXC01002897.1	4	5945582–59456306	0	831	276	29.26	6.45
SbNF-YB3	ABXC01004591.1	7	6276612–6277436	0	828	275	27.67	6.00
SbNF-YB4	XM002459009.1	3	72403131–72403568	0	441	146	16.46	7.14
SbNF-YB5	XM002459011.1	3	72412515–72413060	0	544	182	19.09	6.15
SbNF-YB6	XM002463118.1	2	73025022–73025675	0	657	218	22.73	6.30
SbNF-YB7	002G135100.1	2	20148061–20146210	1	520	174	19.17	4.74
SbNF-YB8	007G117100.1	7	49678386–49682600	0	4214	297	33.44	5.14
SbNF-YB9	010G119200.1	10	2352136–2353131	0	822	273	28.66	7.36
SbNF-YB10	003G346500.1	3	66773363–66774961	5	388	167	18.03	6.11
SbNF-YB11	009G239600.1	9	57918864–57921191	3	419	137	15.04	4.87
SbNF-YC1	ABXC01004605.1	7	6754771–6755478	0	762	253	28.38	5.04
SbNF-YC2	ABXC01000653.1	1	64266744–64267415	0	744	247	26.18	5.31
SbNF-YC3	ABXC01001448.1	2	63138811–63139227	0	609	202	21.42	5.37
SbNF-YC4	ABXC01003563.1	5	51186980–51187096	5	558	185	20.34	8.26
SbNF-YC5	ABXC01004477.1	6	61341870–61342172	0	387	128	13.58	5.74
SbNF-YC6	ABXC01005047.1	7	63565543–63565956	2	726	241	25.81	5.87
SbNF-YC7	ABXC01006233.1	9	53768950–53769063	3	357	118	12.85	5.80
SbNF-YC8	ABXC01006864.1	10	56155673–56156437	0	798	255	28.39	5.11
SbNF-YC9	003G040500.1	3	3812916–3813569	0	654	217	22.95	4.29
SbNF-YC10	005G089100.1	5	12927331–12928204	1	793	263	29.35	4.26
SbNF-YC11	007G054700.1	7	5576465–5577850	0	1386	416	51.82	4.73
SbNF-YC12	008G43000.1	8	4255910–4257304	0	1395	464	51.95	4.72
SbNF-YC13	008G071900.1	8	9705760–9706977	0	1218	405	45.58	5.03
SbNF-YC14	007G070100.1	7	7704760–7707221	0	675	224	25.14	4.45

The sub-cellular localization of SbNF-Y proteins based on consensus sequences showed chloroplast as important target site. Few of the SbNF-Y proteins were also destined for mitochondria while few followed secretary pathway (Table 2). Several of the SbNF-Y proteins are attached to a membrane by hydrophobic anchors. Alpha helix transmembrane spans with average hydrophobicity were predicted in twelve of the NF-Y proteins. Anchoring of protein to GPI (glysoylphosphatidyl inositol) *via* C-terminal attachment was predicted in three of the NF-YB proteins namely SbNFYB1, SbNFYB3 and SbNFYB6. Through gene fusion experiments it was demonstrated that proteins destined to receive a GPI anchor carry a C-terminal signal sequence for GPI-anchorage. (Caras et al. 1987).

Chromosomal locations and gene structure analysis of SbNF-Y genes

The identified 33 *SbNF-Y* genes of sorghum are distributed on all the ten chromosomes of *S. bicolor* as given in GenBank chromosome data. In case of *SbNF-YA* genes, a maximum of four genes *viz. SbNF-YA1*, *SbNF-YA6*, *SbNF-YA7* and *SbNF-YA2* were identified on chromosome one while chromosome two and eight possess two genes each. The chromosome location of *SbNF-YB* sequences revealed a maximum of three genes on the chromosome three namely *SbNF-YB5*, *SbNF-YB4* and *SbNF-YB10* while chromosome two and seven each have two genes. For *SbNF-YC*, a maximum of four genes i.e. *SbNF-YC6*,

Table 2 Identification of subcellular localization and transmembrane domain of SbNF-Y proteins using ARAMEMNON server

Gene	Transmembrane spans	Subcellular location			Lipid modification
		Chloroplast	Mitochondria	Secretory pathway	
SbNF-YA1	α helix TMspans	✓			
SbNF-YA2		✓			
SbNF-YA3	α helix TMspans				
SbNF-YA4					
SbNF-YA5		✓			
SbNF-YA6	α helix TMspans	✓			
SbNF-YA7		✓			
SbNF-YA8	α helix TMspans	✓	✓		
SbNF-YB1	α helix TMspans				GPI-SOM
SbNF-YB2			✓	✓	
SbNF-YB3	α helix TMspans				GPI-SOM
SbNF-YB4			✓	✓	
SbNF-YB5		✓	✓		
SbNF-YB6	α helix TMspans			✓	GPI-SOM
SbNF-YB7					
SbNF-YB8		✓		✓	
SbNF-YB9		✓		✓	
SbNF-YB10		✓			
SbNF-YB11		✓			
SbNF-YC1					
SbNF-YC2	α helix TMspans	✓			
SbNF-YC3		✓			
SbNF-YC4	α helix TMspans	✓	✓		
SbNF-YC5					
SbNF-YC6					
SbNF-YC7	α helix TMspans	✓	✓		
SbNF-YC8					
SbNFYC9					
SbNF-YC10	α helix TMspans	✓	✓		
SbNF-YC11					
SbNF-YC12					
SbNF-YC13	α helix TMspans	✓			
SbNF-YC14					

SbNF-YC14, *SbNF-YC1* and *SbNF-YC11* were observed on chromosome seven while chromosomes five and eight have two genes each. The distribution of *SbNF-Y* genes on 10 chromosomes of sorghum is shown in Fig. 1. The variability of *NF-Y* gene density on chromosomes of different crops have been reported like *PvNF-Y* genes were distributed on 10 out of the 11 common bean chromosomes (Ripodas et al. 2014), *BdNF-YA* and C subunits were absent on chromosome five, and no *BdNF-YB* on chromosome four and five (Cao et al. 2011), *OsHAP* gene were dispersed on 11 chromosome out of the 12 rice chromosomes (Thirumurugan et al. 2008).

The exon-intron organization provides an insight into evolutionary relationships among genes or organisms and needs substantial investigation (Koralewski and Krutovsky 2011). The gene structure analysis could provide possible mechanisms of structural evolution of *NF-Y*genes in sorghum and thus an attempt was made to compare the exon-intron structures of all the 33 annotated *NF-Y* genes. The putative gene structure of these *SbNF-Y* gene families is shown in Fig. 2. Predominance of intronless *SbNF-Y* genes was evident based on the fact that a total of 18 out of 33 *SbNF-Y* genes lacked intron. In case of *SbNF-YA* genes, there exists great variability in terms of introns varying from intronless *SbNF-YA1* to a maximum of five introns for *SbNF-YA5*. For *SbNF-YB* gene family seven out of 11 were found to be intronless while a maximum of five introns was observed for *SbNF-YB10*.

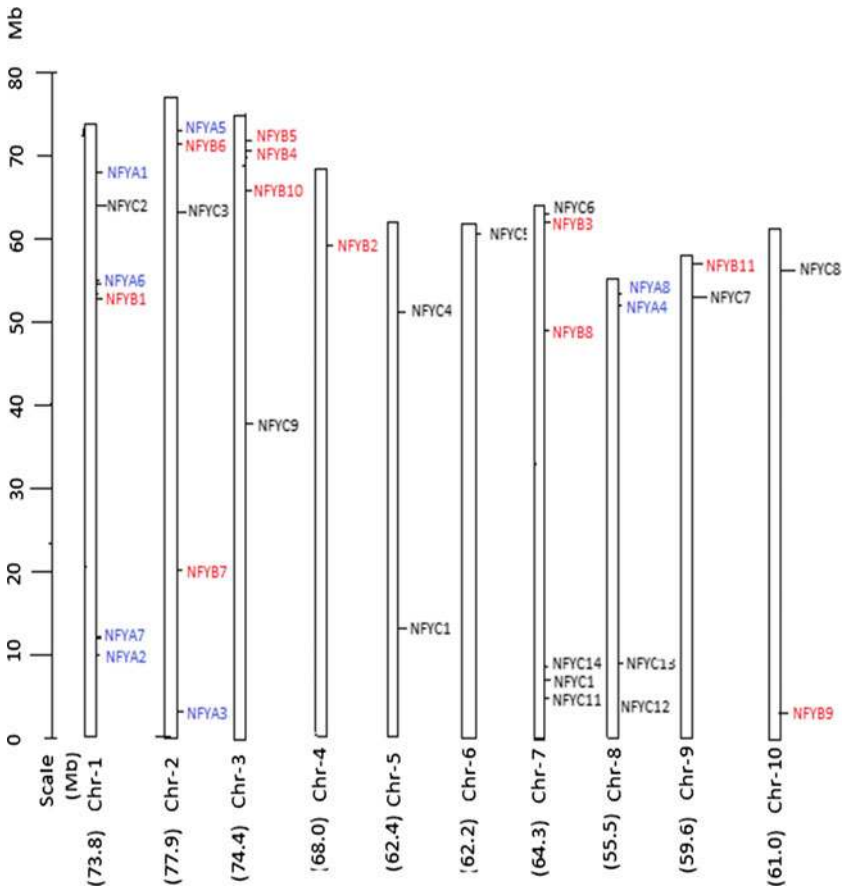


Fig. 1 Chromosomal localization of 33 *SbNF-Y* genes on ten chromosomes of *Sorghum bicolor*. The chromosome numbers and its size are indicated at the bottom of each bar. Sizes of chromosomes are represented on using *vertical scale*

Similarly for *SbNF-YC* gene family, 10 out of 14 were intronless and *SbNF-YC4* possessed a maximum of 5 introns. Thus *SbNF-Y* gene family shows great variability in terms of intron/exon distribution pattern. The relevance of the gene structure analysis is based on the fact that introns are believed to be the essential entities of eukaryotic genes as their loss or gain causes structural diversity and complexity, which might contribute to the evolution of multiple gene families like the NF-Y gene family. Evidence now exists that introns have many functions, including regulation and exon shuffling and alternative splicing (Fedorova and Fedorov 2003; Le et al. 2003). Further introns might be associated with enhancement of intragenic recombination and moderating the evolutionary rate of genes (Roy and Gilbert 2006).

Multiple sequence alignment and phylogenetic assessment of SbNF-Yproteins

The 33 SbNF-Y proteins were aligned using ClustalX 2.0.10 revealing conserved DNA binding domains for SbNF-YA, SbNF-YB and SbNF-YC proteins as shown in Fig. 3. The SbNF-Y proteins have a highly conserved region flanked by largely non-conserved sequences as reported in *Arabidopsis* NF-Y members (Siefers et al. 2009). These conserved regions, in mammals and yeast, are found to be necessary and sufficient for heterodimerization, heterotrimerization and DNA interactions at CCAAT sites (Kim et al. 1996; McNabb et al. 1997; Romier et al. 2003; Testa et al. 2005; Xing et al. 1993, 1994). The core region sequences of the SbNF-YA peptides are highly conserved in comparison to SbNF-YB and SbNF-YC (Fig. 3). Structural and functional studies suggests that NF-YA proteins are more evolutionarily constrained as compared to NF-YB or NF-YC proteins and further mutations in this pentamer region gives complete or near complete loss of NF-Y binding (Romier et al. 2003).

The NF-YA proteins, consists of two domains that are highly conserved in all higher eukaryotes studied till date (Laloum et al. 2013). The first conserved domain of 20 amino-acid as revealed by studies on yeast and mammalian system forms a alpha helix and found to be essential for the interaction with NF-YB and NF-YC (Mantovani 1999; Romier et al. 2003; Xing et al. 1993). The next neighboring domain of 21 amino-acids, separated from first domain by a conserved linker sequence, is required for specific binding to CCAAT boxes. Other than the conserved domains, plant NF-YA proteins are variable in size and structure as evident for sorghum NF-YA proteins. These conserved domains are placed in the C-terminus of the protein in mammals whereas they are present more centrally in plant NF-YAs. The three histidine (H) and three arginine (R) residues known to be essentially required for

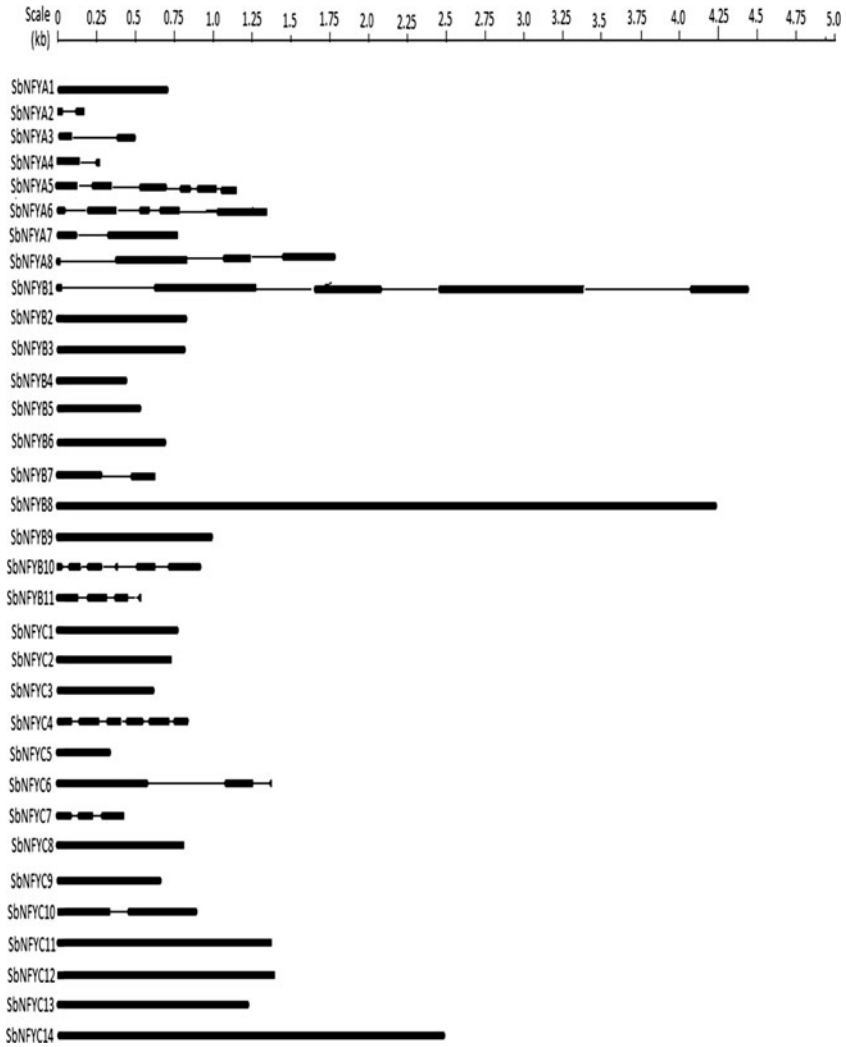


Fig. 2 Intron/exon structure of *SbNF-Y* genes, *black bars* represents the exon while the *lines* represent the intron. The size of exons and introns can be estimated using the horizonal scale

DNA binding as in the case of mammalian NF-YA (Xing et al. 1993) was also observed to be highly conserved in all the eight SbNF-YAproteins (Fig. 3a). Similarly in *Arabidopsis*, rice, *Brachypodium* and wheat these conserved residues were observed (Cao et al. 2011; Siefers et al. 2009; Stephenson et al. 2007; Thirumurugan et al. 2008). The NF-YB and NF-YC subunits in contrast are considered to make DNA contact with the neighboring region of CCAAT sequences and are loosely conserved in comparison to NF-YA (Romier et al. 2003).

The evolutionary relationship among different SbNF-Y proteins was assessed by subjecting the deduced amino-acid sequences of 33 SbNF-Yproteins for multiple sequence alignment followed by phylogenetic tree construction. The phylogenetic tree revealed three major clusters: cluster I, II and III for SbNF-YA, SbNF-YB and SbNF-YC respectively (Fig. 4). The closely related members within the clusters showed almost similar amino-acid sequence lengths, isoelectric point and molecular weights as listed in Table 1. The result revealed that SbNF-YB8 was much closer to SbNF-YA proteins, while other SbNF-YB proteins showed close relationship with SbNF-YC proteins. From this phylogenetic tree we could presume that SbNF-YB has been evolved from SbNF-YA gene and are closely related to SbNF-YC genes.

Assessment of gene duplication time for SbNF-Y proteins

The phylogenetic tree constructed for SbNF-YA, SbNF-YB and SbNF-YC genes identified a total of eight pairs of paralogous genes in the terminal nodes as supported by strong bootstrap values. The presence of these paralogous genes on different chromosomes (Fig. 1) clearly suggests the possibility of segmental duplication event, which might be associated with the expansion of the NF-Y gene family in sorghum. Further, study has been done to predict the immediate ancestral protein sequences of these eight pairs of paralogous genes and the identified protein motif structure of the duplicated genes (Fig. 5 and Supplementary Table 1). The ratio of non-synonymous (Ka) to synonymous (Ks) nucleotide substitution rates denotes the selective pressures on genes. This can be used to identify combinations of genes in the phylogenetic

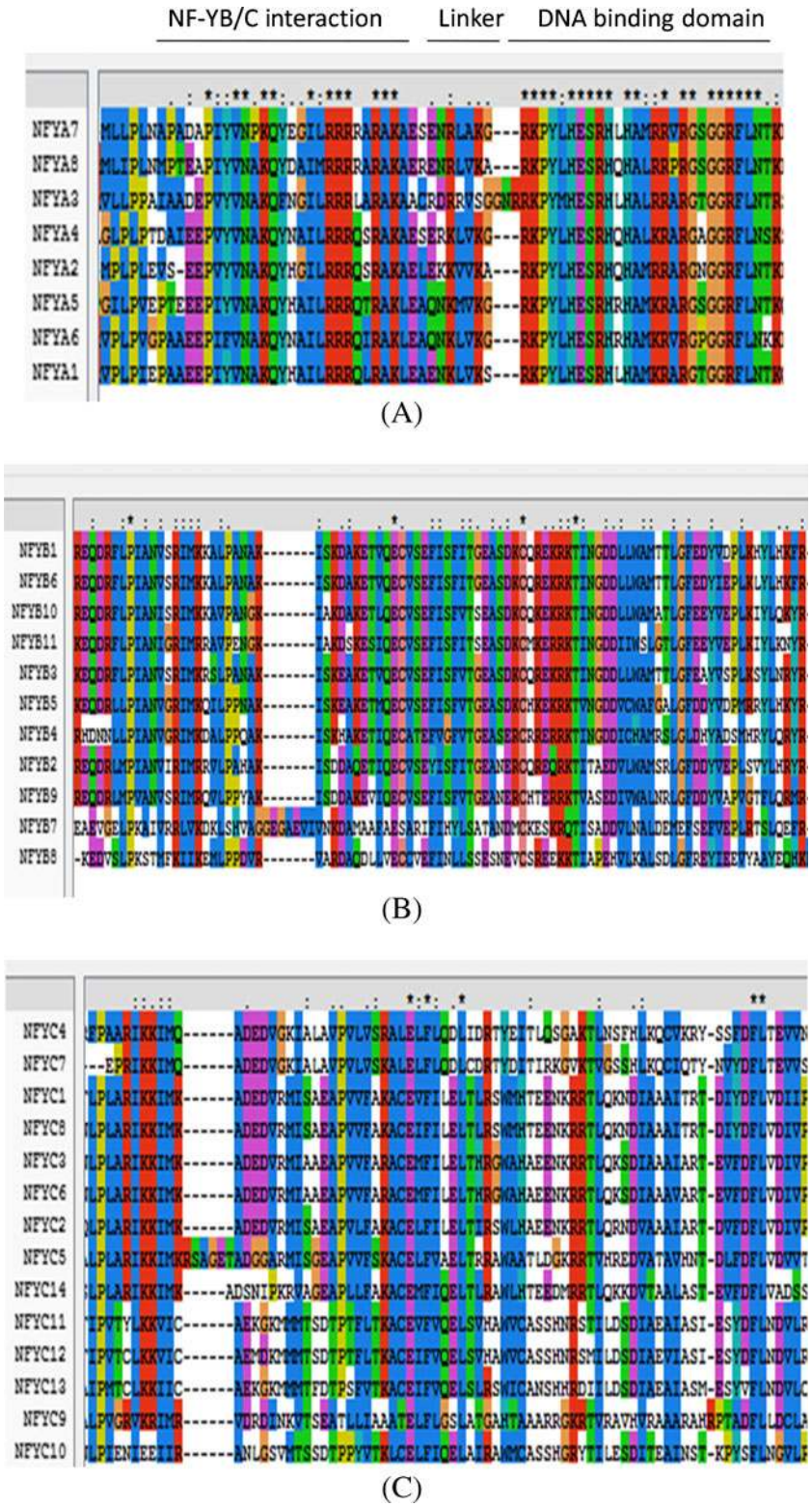


Fig. 3 *S. bicolor* NF-Y protein sequence alignments. Sequence alignments were obtained using Clustal X 2.0.10 program. Sequences were identified and in Genomic databases and renamed as SbNF-YA 1–8 (**a**); SbNF-YB 1–11 (**b**) and SbNF-YC 1–14 (**c**), respectively. *Dashes* indicate gap in the sequence. Strongly conserved residues are indicated above the alignment with *asterisks*, *colon* indicate the residues variation occurring within the strongly conserved groups and *dots* indicates the residues variation occurring within weaker conserved residue groups

trees, where encoded proteins might be subjected to functional alteration. This ratio when greater than 1 indicates positive selective pressure, whereas when it is around 1, it indicates either neutral evolution at the protein level or an averaging of sites under positive and negative selective pressures. A ratio less than 1 indicates the possibility of selective pressures associated with conserved protein sequences (Hurst 2002; Yang and Bielawski 2000).

The Ks and Ka value of paralogous sequences calculated using DnaSP software is shown in the Table 3.The

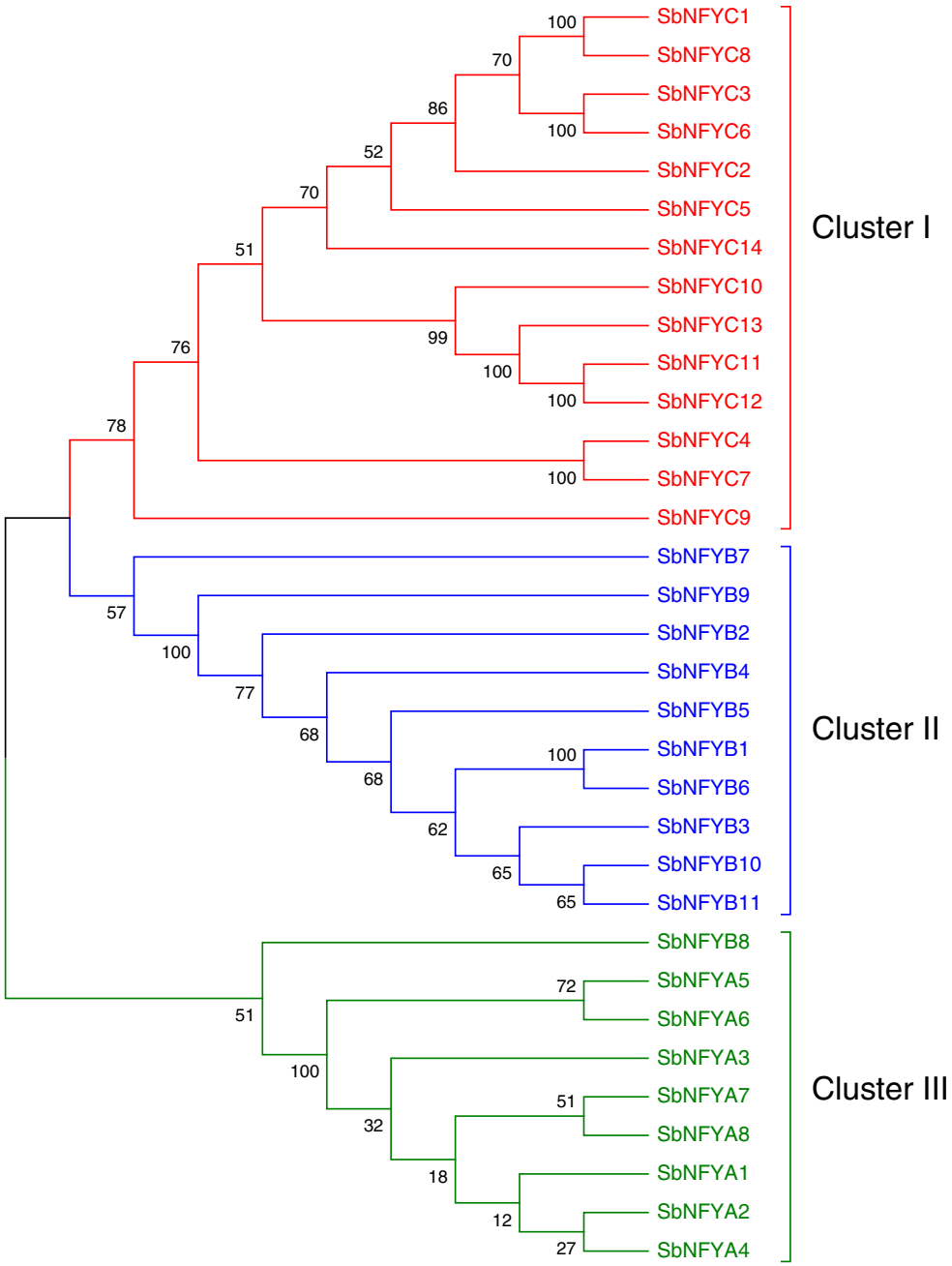


Fig. 4 The combined phylogenetic tree of 8 Sb NF-YA, 11 Sb NF-YB and 14 Sb NF-YC proteins constructed by neighbor-joining methods using MEGA 6.0 with a bootstrap of 500 reiterations

paralogous genes namely SbNF-YA5/SbNF-YA6 and SbNF-YB11/SbNF-YB12 reveals the possibility of neutral mutation owing to almost equal Ka and Ks values while paralogous genes *SbNF-YC7*/*SbNF-YC8*; *SbNF-YB10*/*SbNF-YB11*; *SbNF-YC1*/*SbNF-YC8*; *SbNF-YC3*/*SbNF-YC6*; *SbNF-YB10*/*SbNF-YB11*; *SbNF-YC3*/*SbNF-YC6* and Sb*NF-YC*4/Sb*NF-YC*7 showed positive mutation. One of the paralogous genes *SbNF-YB1*/*SbNF- YB6* showed negative (purifying) selection as Ka value is greater than Ks. Therefore, statistics of the two variables evaluated for different paralogous genes from different evolutionary lineages could provide a powerful tool for quantifying molecular evolution. The date of duplication events were also estimated approximately using the Ks value. The segmental duplications of the *SbNF-Y* genes is found to be originated from 10.1 Mya (million years ago, Ks = 0.1324) to 155.3 Mya (Ks = 2.0192), with an average mean of 59.7 Mya. This data suggests that the sorghum *NF-Y* family gene expansion is due to segmental duplication events, and further these *SbNF-Y* genes retained their function after duplication.

Motif analysis

The distributions of conserved motifs were assessed for SbNF-Y proteins by means of MEME software (Fig. 6). The

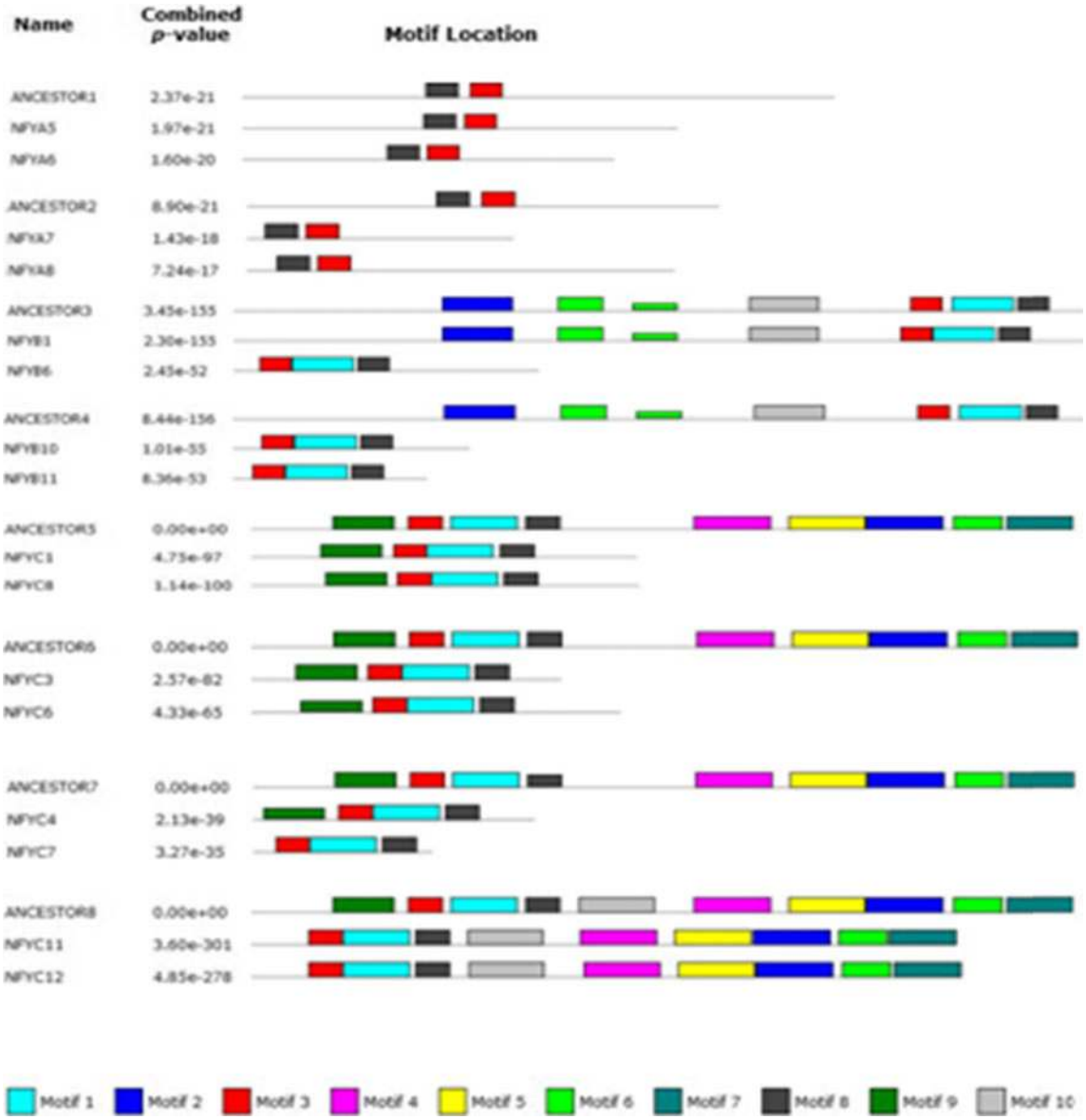


Fig. 5 Distribution of motifs highlighting gene duplicates and their ancestors

Table 3 The gene duplication time identified in different paralogous pairs of *NF-Y* gene families. The Ka represents the number of non-synonymous substitution per non-synonymous site while Ks is the number of synonymous substitution per synonymous site and Ka/Ks represents the ratios of non-synonymous (Ka) versus synonymous (Ks) mutations

Paralogous genes		Location on chromosome	Duplication event	Ks value	Ka value	Ka/Ks	Date (Millions years ago)
SbNF-YA5	*SbNF-YA6*	Chr2/Chr1	Segmental	0.5934	0.4467	0.7527	45.6
SbNF-YA7	*SbNF-YA8*	Chr1/Chr8	Segmental	0.9666	0.3352	0.3467	74.3
SbNF-YB1	*SbNF-YB6*	Chr1/Chr2	Segmental	0.1731	0.3741	2.1611	13.3
SbNF-YB10	*SbNF-YB11*	Chr3/Chr9	Segmental	1.0555	0.2092	0.1981	81.1
SbNF-YC1	*SbNF-YC8*	Chr5/Chr10	Segmental	0.9015	0.1409	0.1562	69.3
SbNF-YC3	*SbNF-YC6*	Chr2/Chr7	Segmental	0.3720	0.1103	0.2965	28.6
SbNF-YC4	*SbNF-YC7*	Chr5/Chr9	Segmental	2.0192	0.3094	0.1532	155.3
SbNF-YC11	*SbNF-YC12*	Chr7/Chr8	Segmental	0.1324	0.1169	0.8829	10.1

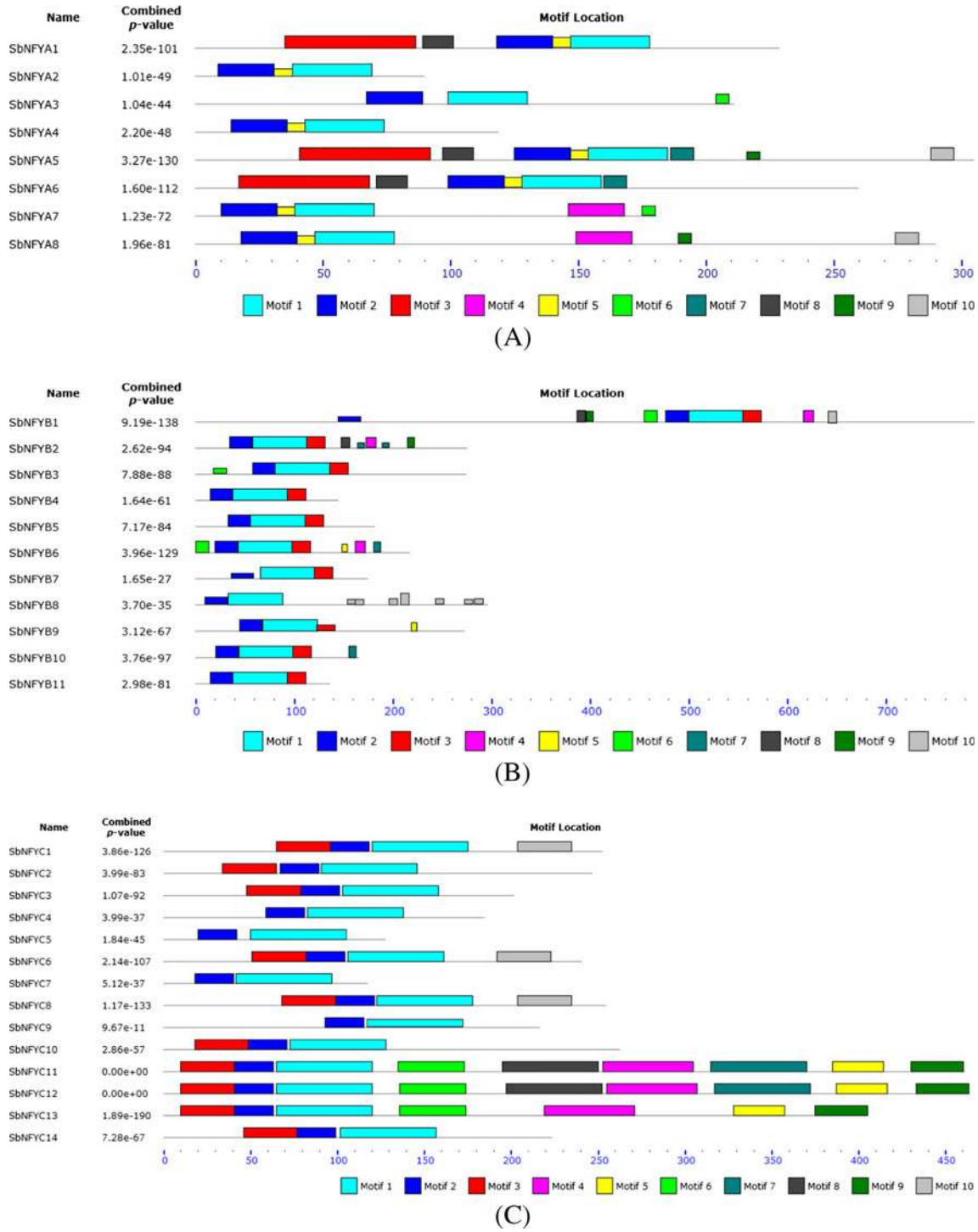


Fig. 6 Motif distribution among SbNF-YA (**a**); SbNF-YB (**b**) and SbNF-YC (**c**) proteins respectively using MEME ver. 4.4.0

variation in motif distributions among NF-YA, NF-YB and NF-YC proteins of sorghum is listed in the Supplementary Table 2. In case of SbNF-YA sequences, motif 1 and 2 were ubiquitously present while rest of the 8 motifs showed variability in terms of distribution among the SbNF-YA proteins. Similarly Sb NF-YB proteins revealed uniform presence of highly conserved motif 1 among all the 11 SbNF-YA proteins. A motif of 22 amino-acid and 18 amino-acid sequence was observed outside the conserved core region of SbNF-YB. The motif analysis of NF-YC proteins of sorghum also revealed a highly conserved motif 2 observed in 13 out of 14 SbNF-YC proteins. These conserved motifs reflects typical diagnostic features for different subunits of NF-Y proteins in general and hence provides confirmatory

identification of SbNF-Y proteins from the sorghum genome.

Comprehensive phylogenetic analysis of NF-Y proteins of sorghum, rice and *Arabidopsis*

A comprehensive phylogenetic tree comprising of 33 NF-Y proteins of sorghum along with 36 *Arabidopsis* and 28 Rice NF-Y proteins was constructed using software MEGA 6.0 by NJ method with bootstrap (500 reiterations) analysis (Fig. 7). The phylogenetic tree gives an insight into the functional attributes of identified NF-Y members of sorghum based on the availability of functionally characterized NF-Y members of *Arabidopsis* and rice. The phylogenetic tree of NF-YA members of sorghum, rice and *Arabidopsis* revealed few orthologous groups like SbNF-YA8 was placed closely with two NF-YA proteins of *Arabidopsis* namely At5g12840 (AtNF-YA1) (Wenkel et al. 2006) and At3g20910 (AtNF-YA9) (Levesque-Lemay et al. 2003). Regulation of At5g12840 has been implicated in late flowering and thus SbNF-YA8 might represent the most likely candidates for similar function in sorghum. Similarly SbNF-YB2, SbNF-

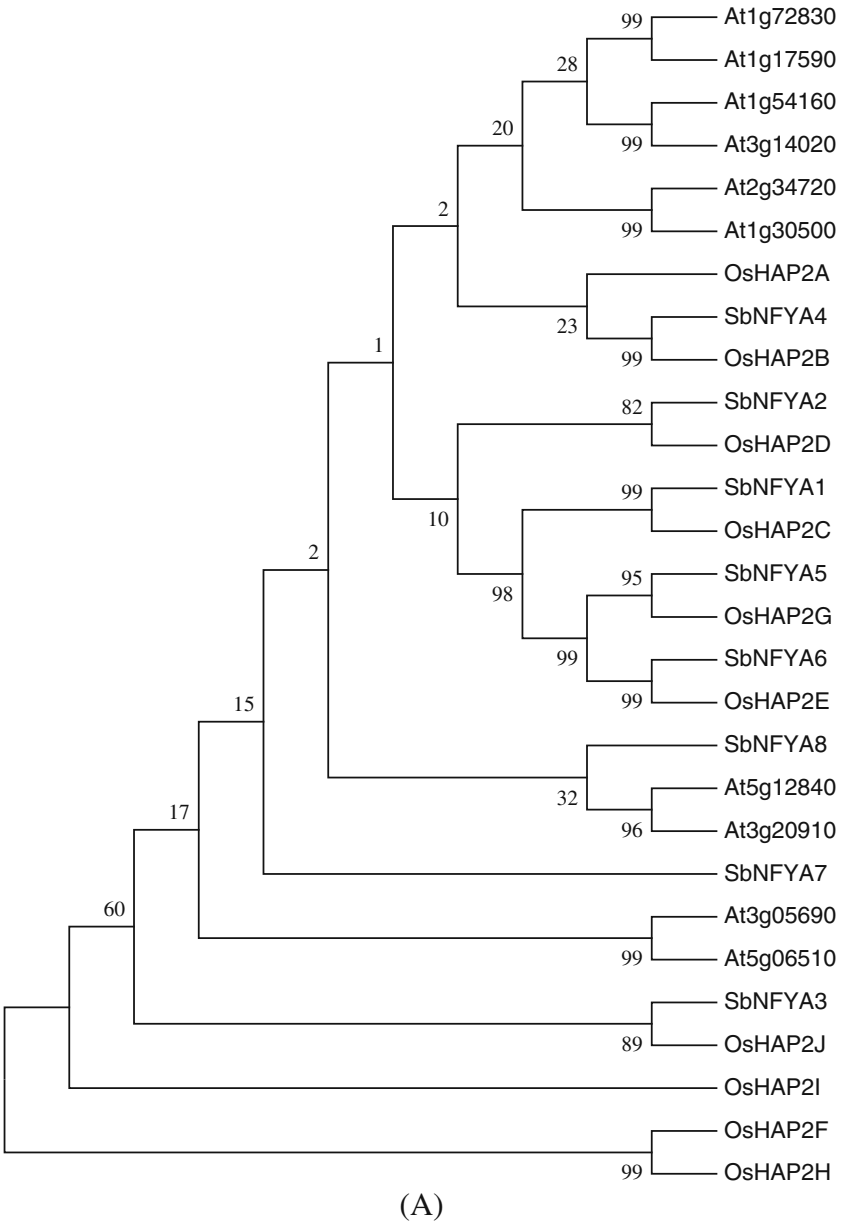
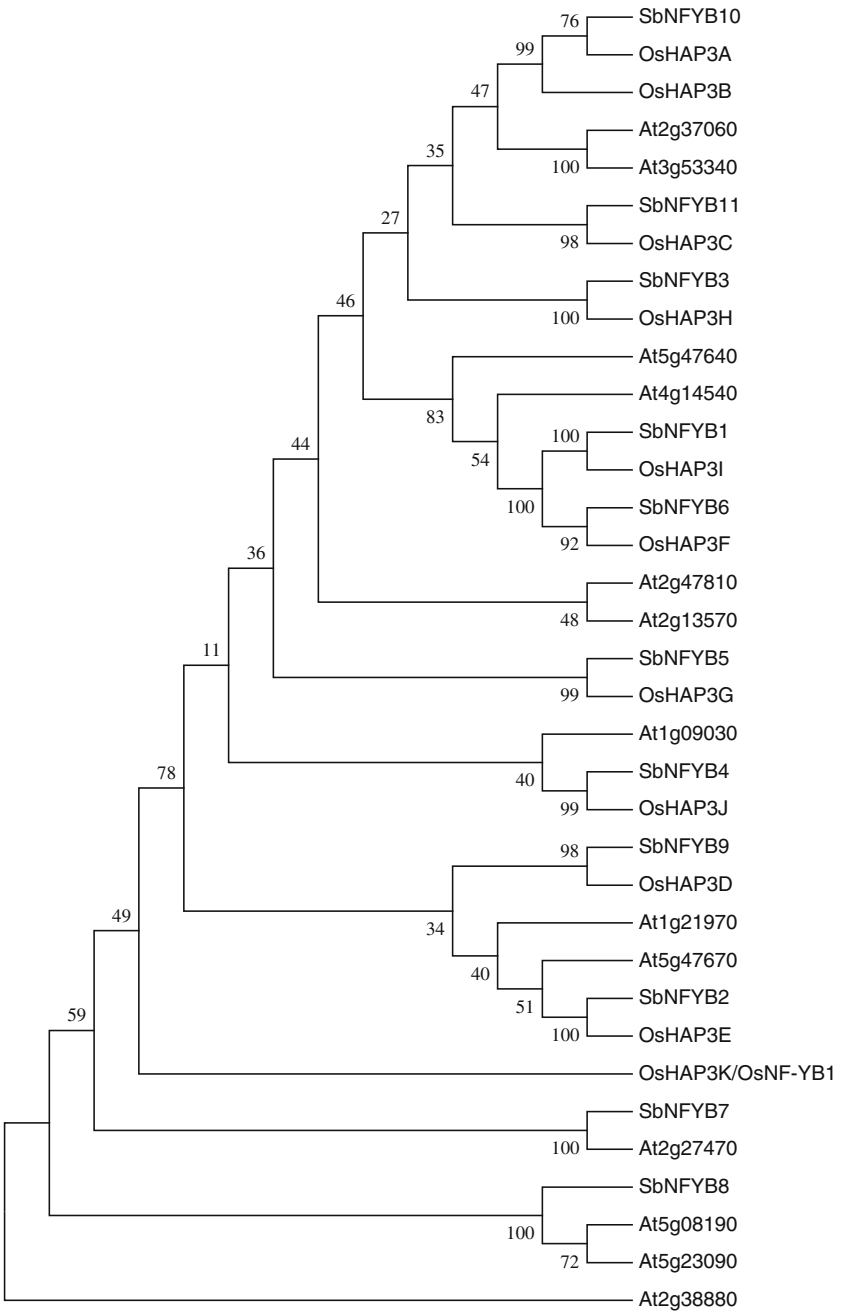


Fig. 7 Phylogenetic tree constructed for (**a**) a total of 28 NF-Y sequences comprising 10 *Arabidopsis thaliana*, 10 *Oryza sativa* and 8 *Sorghum bicolor* NF-YA proteins; (**b**) a total of 35 NF-Y sequences comprising 13 *Arabidopsis thaliana*, 11 *Oryza sativa* and 11 *Sorghum bicolor* NF-YB proteins and (**c**) a total of 34 NF-Y sequences comprising 13 *Arabidopsis thaliana*, 14 *Oryza sativa* and 14 *Sorghum bicolor* NF-YC proteins using MEGA 6.0 by NJ method with 500 bootstrap reiterations



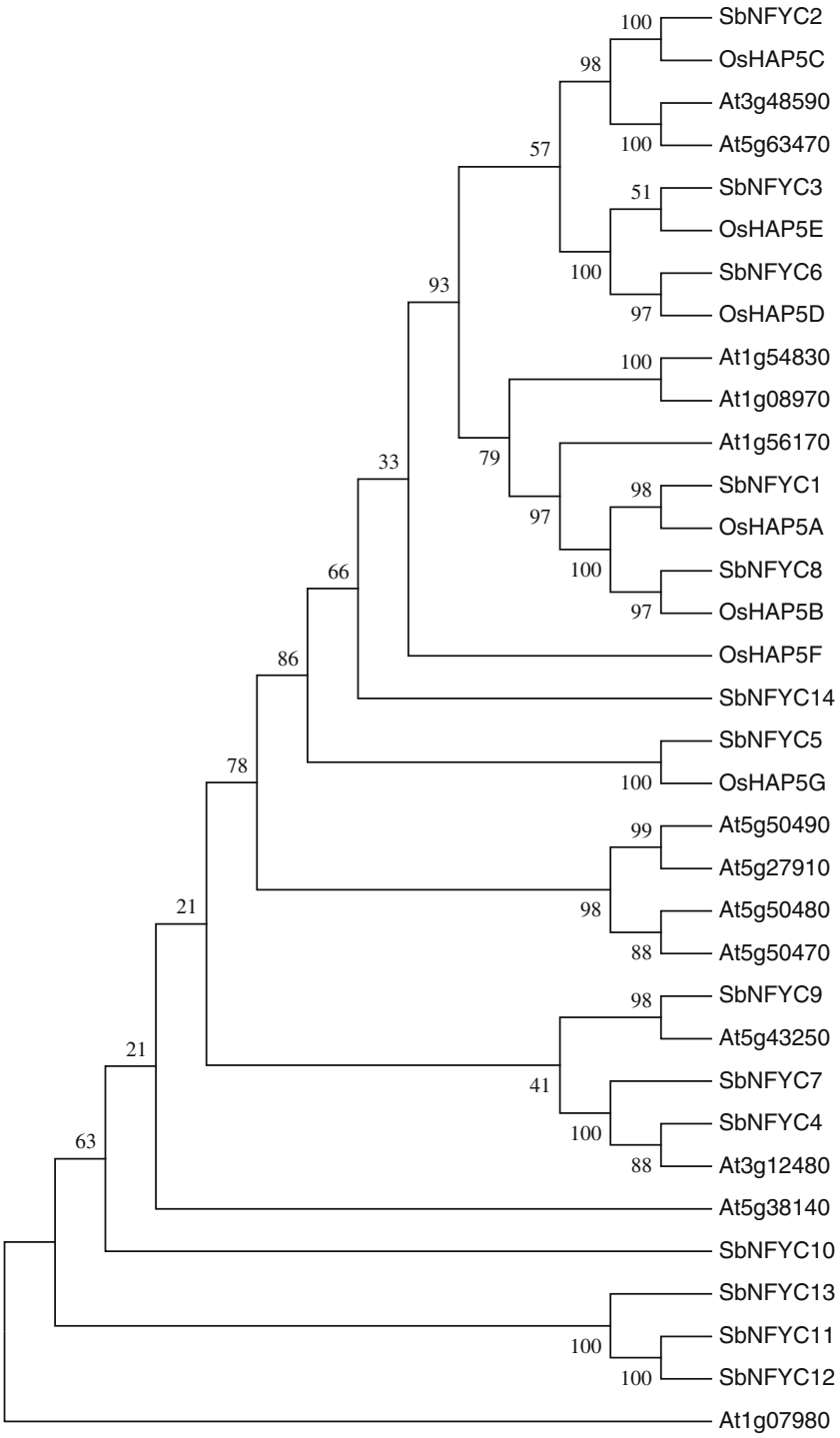
(B)

Fig. 7 (continued)

YB3, SbNF-YB10 and SbNF-YB11 showed similarity with OsHAP3E, OsHAP3H, OsHAP3A and OsHAP3C respectively. OsHAP3A and OsHAP3C are known to be important in regulation of chloroplast biogenesis (Miyoshi et al. 2003), OsHAP3H are important regulator of photoperiodic flowering (Wei et al. 2010) and OsHAP3E are associated with floral meristem development (Ito et al. 2011). Several HAP5 genes of rice showed similarity with SbNF-YC sequences. The SbNF-YC2 are closely placed with At3g48590 (AtNF-YC1) and At5g63470 (AtNF-YC4) known to associated with regulation of flowering (Hackenberg et al. 2012) and germination (Liu and Howell 2010; Warpeha et al. 2007). Thus based on phylogenetic tree, attempts can be made to predict the putative functions of SbNF-Y members prior to subjecting it for validation by wet lab experimentations.

In-silico expression analysis under abiotic stresses

To get an insight into the functional attributes of *NF-Y* genes of sorghum, attempt was made to analyze *in-silico* expression



(C)

Fig. 7 (continued)

using transcriptome data of rice for abiotic stresses due to non-availability of sorghum transcriptome data. Publicly available gene expression values for abiotic stresses of rice seedling for 21 *NF-Y* genes orthologs to sorghum *NF-Y* genes were retrieved (Fig. 7) and a hierarchical clustering based heat map was constructed (Supplementary Figure). The data reveals differential expression of *SbNF-Y* genes based on its identity with the corresponding *OsNF-Y* genes identified in the phylogenetic tree constructed. In case of *SbNF-YA* group, *SbNF-YA2*, *SbNF-YA3* and *SbNF-YA5* seems to be expressed at higher level under drought and salt stress condition based on the enhanced transcript level of the rice orthologs *OsHAP2D*,

OsHAP2J and *OsHAP2G* respectively. The *OsHAP2J* has been associated with drought stress (Jiao et al. 2009). Similarly *SbNF-YA1* and *SbNF-YA6* showed relatively higher expression value under cold stress condition as evident from the higher transcript level of the corresponding rice ortholgos *OsHAP2C* and *OsHAP2E* respectively. In case of *SbNF-YB* group, *SbNF-YB3*, *SbNF-YB4* and *SbNF-YB10* seems to be potential candidate owing to comparatively higher transcript level under salt, drought and cold stress condition respectively as extrapolated from the evident higher transcript level of rice orthologs namely *OsHAP3H*, *OsHAP3J* and *OsHAP3A* respectively. The functional role of *OsHAP3A* in chloroplast biogenesis (Miyoshi et al. 2003) and *OsHAP3H* in response to photoperiodic flowering, grain yield and plant height (Wei et al. 2010) has been reported. The NF-YB genes of sorghum namely *SbNF-YB1*, *SbNF-YB2*, *SbNF-YB5*, *SbNF-YB6*, *SbNF-YB9* and *SbNF-YB11* corresponding to rice orthologs *viz. OsHAP3I*, *OsHAP3E*, *OsHAP3G*, *OsHAP3F*, *OsHAP3D* and *OsHAP3C* respectively revealed comparatively higher transcript level under heat shock condition. The functional attributes of some of the *OsHAP3* genes have been reported. The *OsHAP3C* is associated with flowering and grain yield (Wei et al. 2010), *OsHAP3E* to floral meristem development, *OsHAP3G* and *OsHAP3I* are related to embryogenesis (Ito et al. 2011). In case of *SbNF-YC* group, *SbNF-YC1* corresponding to rice ortholog *OsHAP5A* seems to be promising based on higher transcript level under drought stresses. Similarly two of the SbNF-YC genes namely *SbNF-YC5* and *SbNF-YC8* showed higher transcript level under cold stress condition and could be important for further validation by transgenic technology by making relevant constructs. The higher expression level of *SbNF-YC1* as evident from its corresponding rice ortholog OsHAP5A under drought stress could be another potential stress related *NF-Y* gene of sorghum. The *in-silico* expression profiling attempted in the present study provides an opportunity to target some of the identified potential SbNF-Y genes showing comparatively higher expression level under abiotic stresses for its confirmation by appropriate wet lab experimentations.

Conclusion

The availability of genome sequences provides an opportunity for genome wide identification and characterization of transcription factor gene family as they account for more than 5 % of genome and are known to play a significant role in gene regulation. The genome wide identification and characterization of NF-Y gene family of rice, wheat, *Arabidopsis*, *Brachypodium*, *Brassica napus*, soybean and common bean have led to functional validation of several NF-Y subunits associated with important agronomic traits. The multiple members of NF-Y subunits in plants reflect the redundancy and differentiated functions of these proteins which need to be explored by expression profiling. Using bioinformatics tools attempts have been made in the present study to identify the putative members of NF-Y subunits of sorghum and subject it to extensive *in-silico* characterization for gene structures, motif analysis, chromosomal distribution, conserved motifs, duplication status, ancestral protein sequences and phylogenetic tree construction. The *in-silico* expression profiling based on the available rice transcriptome data for different abiotic stress conditions revealed several potential SbNF-Ys showing higher expression level under drought, salt, cold and heat stresses.

Acknowledgments The financial support by Department of Science and Technology, Government of India, New Delhi in the form of Women Scientist-A fellowship (SR/WOS-A/LS-110/2012(G) to N. Malviya is thankfully acknowledged. DY would like to acknowledge DST Govt. of India for BOYSCAST Fellowship (No.SR/BY/L-02/10) availed at Australian Centre for Plant Functional Genomics, University of Adelaide, South Australia for working on Nuclear Factor Y transcription factors. The author wishes to acknowledge the Head, Department of Biotechnology, D.D.U Gorakhpur University, Gorakhpur, INDIA for infrastructural support.

References

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol; ISMB International Conference on Intelligent Systems for Molecular Biology 2:28–36

Cao S, Kumimoto RW, Siriwardana CL, Risinger JR, Holt BF 3rd (2011) Identification and characterization of NF-Y transcription factor families in the monocot model plant *Brachypodium distachyon*. PLoS One 6:e21805. doi:10.1371/journal.pone.0021805

Caras IW, Weddell GN, Davitz MA, Nussenzweig V, Martin DW Jr. (1987) Signal for attachment of a phospholipid membrane anchor in decay accelerating factor. Science 238:1280–1283

Dolfini D, Mantovani R (2013) Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? Cell Death Differ 20:676–685. doi:10.1038/cdd.2013.13

Dolfini D, Gatta R, Mantovani R (2012) NF-Y and the transcriptional activation of CCAAT promoters. Crit Rev Biochem Mol Biol 47:29–49. doi:10.3109/10409238.2011.628970

Edwards D, Murray JA, Smith AG (1998) Multiple genes encoding the conserved CCAAT-box transcription factor complex are expressed in Arabidopsis. Plant Physiol 117:1015–1022

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A (2002) The PROSITE database, its status in 2002. Nucleic Acids Res 30:235–238

Fedorova L, Fedorov A (2003) Introns in gene evolution. Genetica 118:123–131

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784–3788

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 U S A 93:10274–10279

Gong W et al. (2004) Genome-wide ORFeome cloning and analysis of Arabidopsis transcription factor genes. Plant Physiol 135:773–782. doi:10.1104/pp.104.042176

Hackenberg D, Keetman U, Grimm B (2012) Homologous NF-YC2 subunit from Arabidopsis and Tobacco is activated by Photooxidative stress and Induces flowering. Int J Mol Sci 13: 3458–3477. doi:10.3390/ijms13033458

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98 Key: citeulike: 691774

Hurst LD (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet: TIG 18:486

Ito Y, Thirumurugan T, Serizawa A, Hiratsu K, Ohme-Takagi M, Kurata N (2011) Aberrant vegetative and reproductive development by overexpression and lethality by silencing of OsHAP3E in rice. Plant Sci: an International Journal of Experimental Plant Biology 181:105–110. doi:10.1016/j.plantsci.2011.04.009

Jiao Y et al. (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. Nat Genet 41: 258–263. doi:10.1038/ng.282

Jin JP, Zhang H, Kong L, Gao G, Luo JC (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res 42:D1182–D1187

Kim IS, Sinha S, de Crombrugghe B, Maity SN (1996) Determination of functional domains in the C subunit of the CCAAT-binding factor (CBF) necessary for formation of a CBF-DNA complex: CBF-B interacts simultaneously with both the CBF-A and CBF-C subunits to form a heterotrimeric CBF molecule. Mol Cell Biol 16:4003–4013

Koralewski TE, Krutovsky KV (2011) Evolution of exon-intron structure and alternative splicing. PLoS One 6:e18055. doi:10.1371/journal.pone.0018055

Kushwaha H, Gupta S, Singh VK, Rastogi S, Yadav D (2011) Genome wide identification of Dof transcription factor gene family in sorghum and its comparative phylogenetic analysis with rice and Arabidopsis. Mol Biol Rep 38:5037–5053. doi:10.1007/s11033-010-0650-9

Laloum T, De Mita S, Gamas P, Baudin M, Niebel A (2013) CCAAT-box binding transcription factors in plants: Y so many? Trends Plant Sci 18:157–166. doi:10.1016/j.tplants.2012.07.004

Le HH, Nott A, Moore MJ (2003) How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci 28:215–220

Levesque-Lemay M, Albani D, Aldcorn D, Hammerlindl J, Keller W, Robert LS (2003) Expression of CCAAT-binding factor antisense transcripts in reproductive tissues affects plant fertility. Plant Cell Rep 21:804–808. doi:10.1007/s00299-003-0588-7

Liang M, Yin X, Lin Z, Zheng Q, Liu G, Zhao G (2014) Identification and characterization of NF-Y transcription factor families in Canola (*Brassica napus* L.). Planta 239:107–126. doi:10.1007/s00425-013-1964-3

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452. doi:10.1093/bioinformatics/btp187

Liu JX, Howell SH (2010) bZIP28 and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in Arabidopsis. Plant Cell 22:782–796. doi:10.1105/tpc.109.072173

Lu M, Zhang DF, Shi YS, Song YC, Wang TY, et al. (2013) Expression of SbSNAC1, a NAC transcription factor from sorghum, confers drought tolerance to transgenic *Arabidopsis*. PCTOC 115:443–455

Maity SN, de Crombrugghe B (1998) Role of the CCAAT-binding protein CBF/NF-Y in transcription. Trends Biochem Sci 23:174–178

Mantovani R (1999) The molecular biology of the CCAAT-binding factor NF-Y. Gene 239:15–27

McNabb DS, Tseng KA, Guarente L (1997) The *Saccharomyces cerevisiae* Hap5p homolog from fission yeast reveals two conserved domains that are essential for assembly of heterotetrameric CCAAT-binding factor. Mol Cell Biol 17:7008–7018

Miyoshi K, Ito Y, Serizawa A, Kurata N (2003) OsHAP3 genes regulate chloroplast biogenesis in rice. Plant J: for Cell and Molecular Biology 36:532–540

Paterson AH et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556. doi:10.1038/nature07723

Petroni K et al. (2012) The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. Plant Cell 24:4777–4792. doi:10.1105/tpc.112.105734

Punta M et al. (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301. doi:10.1093/nar/gkr1065

Pupko T, Pe'er I, Hasegawa M, Graur D, Friedman N (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. Bioinformatics 18:1116–1123

Quach TN, Nguyen HT, Valliyodan B, Joshi T, Xu D, Nguyen HT (2015) Genome-wide expression analysis of soybean NF-Y genes reveals potential function in development and drought response. Mol Genet Genomics MGG 290:1095–1115. doi:10.1007/s00438-014-0978-2

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120. doi:10.1093/nar/gki442

Riechmann JL et al. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105–2110

Ripodas C, Castaingts M, Clua J, Blanco F, Zanetti ME (2014) Annotation, phylogeny and expression analysis of the nuclear factor Y gene families in common bean (*Phaseolus vulgaris*). Front Plant Sci 5:761. doi:10.3389/fpls.2014.00761

Romier C, Cocchiarella F, Mantovani R, Moras D (2003) The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. J Biol Chem 278:1336–1345. doi:10.1074/jbc.M209635200

Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7:211–221. doi:10.1038/nrg1807

Sato H et al. (2014) Arabidopsis DPB3-1, a DREB2A interactor, specifically enhances heat stress-induced gene expression by forming a heat stress-specific transcriptional complex with NF-Y subunits. Plant Cell 26:4954–4973. doi:10.1105/tpc.114.132928

Schwacke R et al. (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol 131:16–26. doi:10.1104/pp.011577

Sekhwal MK, Swami AK, Sharma V, Sarin R (2015) Identification of drought-induced transcription factors in *Sorghum bicolor* using GO term semantic similarity. Cell Mol Biol Lett 20:1–23. doi:10.2478/s11658-014-0223-3

Seo J, Bakay M, Chen YW, Hilmer S, Shneiderman B, Hoffman EP (2004) Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays. Bioinformatics 20:2534–2544. doi:10.1093/bioinformatics/bth280

Siefers N, Dang KK, Kumimoto RW, Bynum WE, Tayrose G, Holt BF 3rd (2009) Tissue-specific expression patterns of Arabidopsis NF-Y transcription factors suggest potential for extensive combinatorial complexity. Plant Physiol 149:625–641. doi:10.1104/pp.108.130591

Sigrist CJA et al. (2013) New and continuing develop-ments at PROSITE. Nucleic Acids Res 41(D1):D344–D347

Sinha S, Maity SN, Lu J, de Crombrugghe B (1995) Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A and CBF-B and with yeast HAP2 and HAP3. Proc Natl Acad Sci U S A 92:1624–1628

Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7(Suppl 1:S10):11–12. doi:10.1186/gb-2006-7-s1-s10

Stephenson TJ, McIntyre CL, Collet C, Xue GP (2007) Genome-wide identification and expression analysis of the NF-Y family of transcription factors in *Triticum aestivum*. Plant Mol Biol 65:77–92. doi:10.1007/s11103-007-9200-9

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729. doi:10.1093/molbev/mst197

Testa A, Donati G, Yan P, Romani F, Huang TH, Vigano MA, Mantovani R (2005) Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. J Biol Chem 280:13606–13615. doi:10.1074/jbc.M414039200

Thirumurugan T, Ito Y, Kubo T, Serizawa A, Kurata N (2008) Identification, characterization and interaction of HAP family genes in rice. Mol Genet Genomics MGG 279:279–289. doi:10.1007/s00438-007-0312-3

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The botany array resource: e-Northerns, expression angling, and promoter analyses. Plant J: for Cell and Molecular Biology 43:153–163. doi:10.1111/j.1365-313X.2005.02437.x

Wang X, Tang H, Bowers JE, Paterson AH (2009) Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. Genome Res 19:1026–1032. doi:10.1101/gr.087288.108

Warpeha KM et al. (2007) The GCR1, GPA1, PRN1, NF-Y signal chain mediates both blue light and abscisic acid responses in Arabidopsis. Plant Physiol 143:1590–1600. doi:10.1104/pp.106.089904

Wei X et al. (2010) DTH8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. Plant Physiol 153:1747–1758. doi:10.1104/pp.110.156943

Wenkel S, Turck F, Singer K, Gissot L, Le Gourrierec J, Samach A, Coupland G (2006) CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of Arabidopsis. Plant Cell 18:2971–2984. doi:10.1105/tpc.106.043299

Xing Y, Fikes JD, Guarente L (1993) Mutations in yeast HAP2/HAP3 define a hybrid CCAAT box binding domain. EMBO J 12:4647–4655

Xing Y, Zhang S, Olesen JT, Rich A, Guarente L (1994) Subunit interaction in the CCAAT-binding heteromeric complex is mediated by a very short alpha-helix in HAP2. Proc Natl Acad Sci U S A 91:3009–3013

Yadav D et al. (2015) Constitutive overexpression of the TaNF-YB4 gene in transgenic wheat significantly improves grain yield. J Exp Bot. doi:10.1093/jxb/erv370

Yan L, Xu C, Kang Y, Gu T, Wang D, Zhao S, Xia G (2013) The heterologous expression in *Arabidopsis thaliana* of sorghum transcription factor SbbHLH1 downregulates lignin synthesis. J Exp Bot 64:3021–3032. doi:10.1093/jxb/ert150

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yang J, Xie Z, Glover BJ (2005) Asymmetric evolution of duplicate genes encoding the CCAAT-binding factor NF-Y in plant genomes. New Phytol 165:623–631. doi:10.1111/j.1469-8137.2004.01260.x

Zhang T, Zhang D, Liu Y, Luo C, Zhou Y, Zhang L (2015) Overexpression of a NF-YB3 transcription factor from Picea wilsonii confers tolerance to salinity and drought stress in transformed *Arabidopsis thaliana*. Plant Physiol Biochem: PPB/Societe francaise de physiologie vegetale 94:153–164. doi:10.1016/j.plaphy.2015.05.001