Version 4.3-12/08/20 Cotton bZIP Transcription Factors: Characterization of the bZIP Family From Gossypium Hirsutum, Gossypium Arboreum and Gossypium Raimondii

Vaishali Khanale (vaishali.khanale@mahyco.com)
Mahyco Research Centre https://orcid.org/0000-0001-8013-2537
Anjanabha Bhattacharya
Mahyco Research Centre
Rajendra Satpute
Government Institute Of Science Aurangabad
Bharat Char
Mahyco Research Centre

Research article

Keywords: G. hirsutum, G. arboreum, G. raimondii, bZIP proteins, plant development, stress adaptation

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-70685/v1

Abstract

Background

Cotton is an important commodity in the world economy. In this study we have carried out genome-wide identification and bioinformatics characterization of basic leucine zipper domain proteins (bZIPs) from cultivated cotton species *G. hirsutum* along with two subgenome species of allotetraploid cotton, *G. arboreum* and *G. raimondii*. Transcription factors (TFs) are the key regulators in plant development and stress adaptation. Understanding interactions of TFs in cotton crop is important for enhancing stress tolerance and yield enhancement. Among plant TFs, bZIPs plays a major role in seed germination, flower development, biotic and abiotic stress response. Most of the bZIP proteins from cotton remains uncharacterized and can be utilised for crop improvement. In this paper we performed genome-wide identification, phylogenetic analysis, structural characterization and functional role prediction of bZIPs from all three genome species of cotton.

Results

In the present study genome-wide identification, phylogenetic analysis, structural characterization and functional role prediction of bZIP TFs from *G. hirsutum* (AADD) along with two subgenome species *G. arboreum* (A2) and *G. raimondii* (D5) were performed. A total of 228 bZIP genes of *G. hirsutum,* 91 bZIP genes of *G. arboreum* and 86 bZIP genes of *G. raimondii* were identified from CottonGen database. Cotton bZIP genes were annotated in standard pattern according to their match with *Arabidopsis* bZIPs. Multiple genes with similar bZIP designations were observed in cotton. Cotton bZIPs are distributed across all 13 chromosomes with varied density. Phylogenetic characterization of all three cotton species bZIPs with *Arabidopsis* bZIPs classified them into 12 subfamilies, namely A B, C, D, E, F, G, H, I, J, K and S and further into eight subgroups according to functional similarities, viz., A1, A2, A3, C1, C2, S1, S2 and S3.The classification was exclusively based on alignment with *Arabidopsis* bZIPs further supported by structural characteristics like exon number, amino acid length, common functional motifs shared among subfamilies and basic leucine zipper domain (BRLZ) alignment. Subfamily A and S are having maximum number of bZIP genes, subfamily B, H, J and K are single member families. Cotton is carrying only bZIP17 among the group of bZIP17, 28 and 49 which are known to be crucially worked under endoplasmic reticulum (ER) stress. Cotton bZIP protein functions were predicted from identified motifs and orthologs from varied species.

MEME motif analysis identified MYND-Zinc binding domain, tetratricopeptide repeats motif, GluR7, DOG1, (DELAY OF GERMINATION 1) seed dormancy control motif, TGACG sequence specific motif, etc. specifically in some of the subfamily members and presence of bZIP signature domain in all identified bZIPs.

Further we explored the BRLZ domain of *G. raimondii* bZIPs, conserved basic region motif N-X7-R/K is present in almost all subfamily members, variants are GrbZIP62 which is carrying N-X7-I motif and GrbZIP76 with K-X7-R motif. Leucine heptad repeats motif, L-X6-L-X6-L are also present in variant numbers from two to nine with leucine or other hydrophobic amino acid at designated position among 12 subfamily members.

STRING protein interaction network analysis of *G.raimondii* bZIPs observed strong interaction between A-D subfamily members, C-S subfamily members and between GrbZIP17- GrbZIP60. NLS analysis of *G. raimondii* bZIPs observed conserved NLS sequences among subfamilies.

Conclusion

This study analyzed, annotated and phylogenetically classified bZIP proteins from cultivated cotton species *G. hirsutum* along with two subgenome species *G. arboreum* and *G. raimondii*. Cotton bZIPs are classified into twelve subfamilies and eight subgroups. bZIP gene duplications are observed in all three cotton species. We have identified conserved functional motifs among different subfamilies of cotton bZIP proteins and correlated for the prediction of function along with reported function. Explored BRLZ domain structural analysis of *G. raimondii* bZIPs will be useful in further basic characterization of bZIP proteins of cultivated cotton species *G. hirsutum*. STRING protein interaction analysis of *G. raimondii* bZIPs resulted in prediction of interactions among A- D, B-K and C-S subfamily members. Phylogenetic analysis of this study will certainly help in the selection of specific cotton bZIP genes according to the close alignment with *Arabidopsis* orthologs or subgenome homolog.

Introduction

Cotton is an important commodity in the world economy. Approximately 32.93 million hectare was under cotton cultivation around the globe in 2019, with the highest production of 5.77 million metric tons in India followed by US and China (http://ministryoftextiles.gov.in; https://www.statista.com). The textile industry is a significant contributor to the nation's economy and employment, the global textile market share was USD 961.5 billion in 2019 (https://www.grandviewresearch.com).

Functional genomics is a key approach for identifying genes for important traits like yield, biotic-abiotic stress tolerance and fiber quality. Cultivated *G. arboreum,* A subgenome species, and its counterpart, non-spinnable fiber producer *G. raimondii*, D subgenome species are two progenitor species of cultivated allotetraploid cotton, *G.hirsutum* (Li et al 2014). Estimated genome size of *G. hirsutum* is ~2305.2Mb (Chen et al. 2020). Estimated genome size of *G. arboreum* is ~1746Mb which is around two fold larger than *G. raimondii*, ~880Mb (Hendrix et al 2005). Genetic information related to TFs of the two important progenitor species will play major role in the improvement of cultivated allotetraploid cotton. Wang et al. 2012 identified 2,706 transcription factors in *G. raimondii* which includes 208 bHLH and 219 MYB class genes which were preferentially expressed in fiber (Wang et al. 2012) . Kushanov et al. 2016 developed CAPs and DCAPs markers for the *PHYA1*, *PHYB* and *HY5* genes which are associated with the fiber quality and flowering time traits (Kushanov et al. 2016). Over-expression of *G. hirsutum* bZIP transcription factor, ABF2 (bZIP36) resulted in improved drought and salt tolerance in cotton and *Arabidopsis* (Liang et al. 2015).

TFs are the key regulators in plant development and stress adaptation. On the basis of conserved DNA binding domains, TFs are classified into different families among which bZIP-TFs play a major role in seed germination, flower development, and stress response. bZIP proteins specifically bind to DNA with an ACGT core, A-box (TACGTA), C-box (GACGTC) and G-box (CACGTG) motifs and binding specificity is regulated by flanking nucleotides (Jakoby et al. 2002). The bZIP domain carries two structural components located on a contiguous alpha-helix, the first one is a ~16 amino acid residues containing nuclear localization signal, followed by a DNA binding domain invariant N-X7-R/K and the second one is present exactly nine amino acids towards the C-terminus which is a leucines (L) heptad repeat or other bulky hydrophobic amino acids creating an amphipathic helix, L-X6-L-X6-L (Jakoby et al. 2002). In *Arabidopsis* bZIPs, variation observed in leucine repeats, subfamily 'D' bZIPs contains three repeats whereas; more than eight repeats are present in subfamily C and S (Wolfgang et al.2018).

So far, ZIP proteins are classified into 9 to 13 subfamilies in different plant species based on structural and functional characteristics. Marc Jakoby et al. 2002 first classified *Arabidopsis thaliana* bZIP TFs into 10 subfamilies A to I and S. These groups were named with letters considering their family members, for example A for ABF/AREB/ABI5, B for big protein size, C for CPRF2-like, G for GBF, H for HY5, and S for small protein size (Jakoby et al. 2002). The classification of *Arabidopsis* bZIPs is updated by Wolfgang et al. 2018, where the authors divided 78 bZIP proteins of *Arabidopsis* in 13 subfamilies A to I, J, K, M and S (Wolfgang et al. 2018). Wang et al. 2019 classified *Arachis* bZIPs into nine subfamilies A, B, C, D, G, H, I, S and U (Wang et al. 2019). Nijhawan et al. 2008 classified *Oryza* bZIPs into 10 groups A to J (Nijhawan et al. 2008).

If we look at the cotton bZIP protein characterization, Zhang et al. 2018 identified 159 bZIP genes from *G. arboreum* and classified into 13 groups (Zhang et al. 2018). Azeem et al. 2020 identified 87 bZIP genes of *G. arboreum* and 85 bZIP genes of *G. raimondii* from NCBI and classified into 11 subfamilies. Recently Wang et al. 2020 identified 207 bZIP genes of allotetraploid cotton *G.hirsutum* and classified into 13 subfamilies, A to I, S, M, K and J, having major contribution from subfamily 'A' and 'S' bZIPs (Wang et al. 2020).

Emphasising their functional importance, bZIP protein family in plants play crucial roles in developmental and stress responses. Abscisic acid responsive element binding factors (ABFs) from bZIP subfamily 'A' are activated by ABA signalling and are involved in downstream gene regulation in conjunction with stomatal closure mediated adaptation in drought stress (Sirichandra et al 2010, Yoshida et al.2015). Wolfgang and Christoph described bZIP family 'C' genes (bZIP9, bZIP10, bZIP25, bZIP63) and S1 genes (bZIP1, bZIP2, bZIP11, bZIP44, bZIP53) interaction network in response to nutritional starvation and metabolic adaptation under nutritional stress (Wolfgang et al. 2018). TGA or subfamily 'D' , bZIP family proteins binds to the TGACG consensus sequences, to form homo and hetero-dimer. TGA family proteins are predominantly involved in systemic acquired resistance through SA mediated interaction with NPR1 (Fan et al.2002, Fu et al. 2013). TGA family member PERIANTHIA (PAN/bZIP46) plays a role in flower development, interact with ROXY-type GRX (CC-type glutaredoxin, GRX- ROXY1) gene to regulate petal development (Gatz 2013, Gutsche et al. 2017). Recently, Ullah et al 2019 studied soybean TGA family genes and differentiated legumes-specific *TGAs*

structures, involvement in legumes-specific biological processes like legumes-rhizobia symbiotic nodulation and predicted soybean bZIPs role in response to nitrogen.

The bZIP TF family from cotton yet to be fully understand and explored. Thus, this paper attempts phylogenetic analysis, structural characterization and functional role prediction of bZIPs from cultivated cotton species *G. hirsutum* along with two subgenome species *G. arboreum* and *G. raimondii.*

Materials And Methods

A total of 228 bZIP genes of *G. hirsutum,* 91 bZIP genes of *G. arboreum* and 86 bZIP genes of *G. raimondii* were identified from CottonGen database using bZIP domain as a query as well as *Arabidopsis* bZIP protein sequence as a query or through keyword search to recent genome assembly of Gossypium hirsutum (AD1) 'TM-1' genome UTX_v2.1 (published article in may 2020), *Gossypium raimondii* JGI proteins and *Gossypium arboreum* BGI proteins (https://www.cottongen.org/). *Arabidopsis*- bZIP protein sequences were downloaded from TAIR (https://www.arabidopsis.org/). Cotton bZIP genes were annotated on the basis of E-value, bitscore value and percent identity resulted from *Arabidopsis*- bZIP match. All the data regarding gene ID, genomic position, exon number, CDS length, protein length, protein sequences and Arabidopsis match ID with E-value are recorded in supplementary file 1.

Phylogenetic analysis of 227 bZIP genes of *G. hirsutum,* 56 bZIP genes of *G. arboreum* and 57 bZIP genes of *G. raimondii* along with 73 bZIP genes of *Arabidopsis* was performed. Duplicated genes from *G. arboreum* and *G. raimondii* were excluded from phylogenetic analysis and selected first representative gene. For example GabZIP14 is having two entries GabZIP14-1 and GabZIP14-2 so GabZIP14-1 is selected for analysis and designated as GabZIP14. Phylogenetic tree was constructed by maximum likelihood phylogeny method, Dayoff (PAM) protein substitution model and performing 100 bootstrap, using MEGA X (https://www.megasoftware.net/). MEME (http://meme-suite.org) motif search analysis was done for all identified bZIP genes, 228 bZIP genes of *G. hirsutum,* 91 bZIP genes of *G. arboreum* and 86 bZIP genes of *G. raimondii* , selecting 10 motifs identification option and any number of repetition in case of distrubution . Identified motifs were evaluated using SMART (http://smart.embl-heidelberg.de/), Pfam (https://pfam.xfam.org/search/sequence), Interpro (https://www.ebi.ac.uk/interpro/) and TOMTOM. MAST file of all searches were submitted along with supplementary documents.

To understand the interaction network of *G. raimondii* bZIP proteins, STRING (https://string-db.org/) analysis was performed. Network was interpreted for strong protein interactions, sharing functions among themselves with highest edge confidence (0.900), which can be correlated by thickness of the edges.

NLS sequences were listed from plant transcription factor database for *G. raiimondii, Arabidopsis thaliana* bZIP proteins and predicted for *G. raimondii* bZIPs through cNLS Mapper (http://nls-mapper.iab.keio.ac.jp).

Results

G. hirsutum, G. arboreum and *G. raimondii* bZIP protein classification

Phylogenetic analysis of 227 bZIP genes of *G. hirsutum,* 56 bZIP genes of *G. arboreum* and 57 bZIP genes of *G. raimondii* along with 73 bZIP genes of *Arabidopsis* was performed. The phylogenetic analysis indicates that ***G. hirsutum, G. arboreum*** and ***G. raimondii*** bZIPs are closely related to *Arabidopsis* bZIPs and are conserved among subgenomes. Figure 1A shows phylogenetic tree construction. Bootstrap percentage values are mentioned on each branch. Analyzed *G. hirsutum, G. arboreum* and *G. raimondii* bZIP genes are classified into 12 subfamilies A B, C, D, E, F, G, H, I, J, K and S and 8 subgroups A1, A2, A3, C1, C2, S1, S2 and S3. Subgroup classification was done according to clade separation within subfamily and predicated functional similarities. Further this classification is supported by *Arabidopsis*- bZIPs alignment, *G. raimondii* BRLZ domain alignment, exon-intron numbers, protein length and conserved motif identification (Figure 1 and Table 1, 2, 3). For the ease of understanding, in the following elaborated subfamily wise description bZIP proteins from ***G. hirsutum, G. arboreum*** and ***G. raimondii***collectively named as Gossypium bZIPs.

Subfamily A, subgroup A1 contains GossypiumbZIP12, 35, 36, 37, 39, 66 and 67 predicting involvement in ABA stress response and in seed maturation (Liang et al. 2015). Subgroup A2 contains GossypiumbZIP13 and 40 predicting role in drought stress response

(Wang et al. 2019). Subgroup A3 contains FD like protein, involved in positive regulation of flowering GossypiumbZIP14 and 27 (Abe et al. 2005).

Subfamily B contain endoplasmic reticulum (ER) stress response transcription factor bZIP17, involved in salt and osmotic stress resistance in *Arabidopsis* (Liu et al. 2008) and observed interacted with subfamily K member bZIP60 in STRING analysis, which is also involved in ER stress response.

Subfamily C, subgroup C1 contains GossypiumbZIP9, subgroup C2 contains GossypiumbZIP10, 25 and 63, reported to form hetero dimer with S1 subfamily bZIPs , are involved in anther development, positive regulation of seed maturation and response to starvation (Weltmeier et al. 2009; Wolfgang et al. 2018).

Subfamily D of bZIP proteins is involved in flower development, seed development; salicylic acid mediated signalling pathway and pathogen response. MEME motif search identified DOG1- seed dormancy control motif, tetratricopeptide protein-protein interaction motif in these family members. GossypiumbZIP20, 21, 22, 26, 45, 46, 47, 50, 57 and 65 are representing GossypiumbZIPs subfamily D.

Subfamily E of bZIPs functioning in cell wall and pollen development (Gibalova et al. 2009 and 2017) contains GossypiumbZIP34, 61 and 76. In *Arabidopsis thaliana* researchers discovered that a conserved proline residue in the third heptad region of leucine zipper of AtbZIP34 and AtbZIP61 interferes with the formation of homo-dimer whereas change of proline by an alanine in the above mentioned region can form homo dimer and bind to the G-box element (Shen et al. 2007). A conserved proline residue which interferes homodimer formation is also found in the third heptad region of leucine zipper of GossypiumbZIP34 and 61 genes except in **GrbZIP61-1** which is carrying **serine instead of proline**, refer supplementary file 1 for alignment. It is indicated that bZIP34 and bZIP61 is involved in the hetero-dimer formation with I and S subfamily members of bZIPs which are function in vascular development (Shen et al. 2007).

Subfamily F which is well known to be involved in adaptation to zinc deficiency in *Arabidopsis*, wheat and barley contains GossypiumbZIP19, 23 and 24 (Assuncao et al. 2010; Inaba et al. 2015; Nazri et al. 2017; Evens et al. 2017).

Subfamily G plays role in binding to G box motif of genes regulated by hormones and light contains GossypiumbZIP16, 41 and bZIP55.

Subfamily H contains HY like transcription factor GossypiumbZIP56, HY5 which is involved in PhyB signalling pathway and also associated with fiber quality in cotton. HY5 gene is a positive regulator of photo morphogenesis (Wolfgang et al.2018). Kushnov et al 2016 has done comparative sequence analysis of three close relative of fiber quality genes, *PHYA1*, *PHYB*, *HY5* of *G. hirsutum* and *G. barbadense* , developed dCAPS markers which can be utilised in marker –assisted selection breeding for introgression of these genes into the either of the two alloteraploid cotton species. (Kushnov et al. 2016). Abdurakhmonov et al 2017 developed *PHYA1* RNAi *G. hirsutum* with improved fiber quality and yield potential (Abdurakhmonov et al 2017).

Subfamily I which is involved in vascular development contains GossypiumbZIP18, 29, 30, 51, 52, 59 and 69 in 4 clades. Pyo et al. 2006 found expression of bZIP 'I' group members AtbZIP18, 51, 52 and 59 in developing vascular cells and their precursor cells (pyo et al.2006).

Subfamily J and K contains GossypiumbZIP62 and GossypiumbZIP60 respectively. Rolly et al. 2020 studied AtbZIP62 involvement in salt stress tolerance.

Subfamily S, subgroup S1 grouped GossypiumbZIP1, 2, 11, 44 and 53 which are known to be involved in the positive regulation of seed germination, salt stress and in starvation stress response in *Arabidopsis* and peanut. Subgroup S2 contains GossypiumbZIP4, 5, 6 and 7, *Arabidopsis* bZIPs of this group are known as a positive regulator of transcription, express in leaf, pollen, embryo, seed and root. Subgroup S3 contains GossypiumbZIP42, 43, 48 and 58 in one clade, and 70 in separate clade. Nowak et al. 2016 mentioned bHLH109 regulation by AtbZIP4 and AtbZIP43, which is involved in *in-vitro* somatic embryogenesis and stress response. (Hanson et al. 2008; Alonso et al 2009; Weltmeier et al. 2009; Ma et al.2011;Dietrich et al.2011, Wang et al.2019)

GhbZIP44-6D aligned with AtbZIP71 and AtbZIP 72, 74 formed separate clades in which *Gossypium* bZIPs are not grouped.

Conserved motif analysis of *G. hirsutum, G. arboreum* and *G. raimondii*bZIP proteins

Signature bZIP domain is confirmed in all identified bZIP proteins of *G. hirsutum, G. arboreum* and *G. raimondii.* MEME motif analysis deciphering the presence Abscisic acid insensitive 5-Like protein 5-related motif in *G. hirsutum* subfamily 'A' and 'J" members. Sterile alpha motif (SAM)/Pointed domain which is involved in protein- protein interaction is identified in subfamily 'A' bZIPs of all three species. ***G. raimondii*** subfamily 'A' bZIPs are carrying MYND-Zinc binding domain which is involved in protein - protein interaction in the transcriptional regulation context. *G. arboreum* Subfamily 'A' bZIPs, GabZIP18-3 and GabZIP11-3 genes are carrying GluR7, glutamate receptor domain from GluR proteins known to be function in light signal transduction and calcium homeostasis.The presence of two DOG1- seed dormancy control motif and TGA2 like motif is identified in all three gossypium species 'D' subfamily bZIP proteins. Exclusively *G. raimondii* subfamily 'D' bZIPs are carrying the RPA interacting motif, which is conserved in eukaryotic DNA repair and replication proteins, presence of tetratricopeptide repeats motif which are involved in protein-protein interactions. Further presence of tetratricopeptide repeats and other protein protein interaction domains in subfamily D should characterize, as TGA transcription factor family members, are involved in biotic stress tolerance through interaction with NPR1 and similarly STRING analysis also identified D subfamily protein interaction with A subfamily proteins. A typical example of identified motifs in *G. raimondii* bZIP proteins has shown in Figure 2 and some motifs are listed in Table 3.

BRLZ domain structural characterization of *G. raimondii* bZIP proteins

N-X7-R/K basic region

The conserved basic region N-X7-R/K variant is present in almost all analyzed *G. raimondii* bZIP proteins, variation is observed in subfamily E and J. Subfamily E- bZIP76 is having K-X7-R motif, subfamily J- bZIP62 is having N-X7-I motif. (Figure 4 and for BRLZ domain alignment of bZIP69, 76 refer supplementary file1).

Leucine heptad repeats variation in *G. raimondii* bZIPs

Variation in presence of number of leucine heptad repeats motif is observed in *G. raimondii* bZIP proteins. Figure 3 deciphering the detail structure of leucine heptad repeats of *G. raimondii* bZIP proteins, for GrbZIP69 and 76 refer supplementary file 1. First two leucine heptad repeats are conserved among all *G. raimondii* bZIP proteins with exception of **GrbZIP17**, **GrbZIP61, GrbZIP76 andS subfamily proteins GrbZIP 2, 4, 11, and 44** which are carrying either methionine, isoleucine, asparagine or valine at L0, L1 or L2 position (L0 to L9 denotes leucine position in leucine heptad repeats L-X6-L-X6-L motif).

Subfamily-wise leucine heptad repeats are described as follows

Subfamily A bZIP proteins are carrying two to four heptad repeats of leucine or presence of other hydrophobic amino acids at leucine position in last heptad.

Subfamily B, C, E, F and S bZIP proteins are carrying seven to nine heptad sequences with leucine repeats or other different hydrophobic or polar amino acids at leucine position; alanine observed predominantly at leucine position in subfamily C and valine is predominantly present in subfamily S proteins at leucine position.

Subfamily D, G, I and J bZIP proteins are carrying three to six leucine heptad sequences or other hydrophobic amino acids are present at leucine position with predominance of glycine in D subfamily and methionine in I subfamily members.

Pure leucine heptad repeats are present in subfamily H member **GrbZIP56** and subfamily K member **GrbZIP60**, with presence of 4 and 2 heptad repeats of leucine respectively.

STRING protein interaction analysis of *G. raimondii* bZIPs

G. raimondii bZIP proteins interaction analysis is performed using STRING, protein- protein association network software, Figure 4 deciphering the observed network (https://string-db.org/). Very strong interaction observed among subfamily A bZIPs, GrbZIP35 (ABF1/AREB1), GrbZIP39 (ABI5), GrbZIP65, GrbZIP66 (AREB3) and GrbZIP67 (DPBF2) genes, which are known for abscisic acid inducible stress regulation, seed germination and seedling development in plants (Lindemose et al.2013, Yoshida et al. 2015, Skubacz et al.2016). Observed interaction between bZIP subfamily A and D proteins of *G.raimondii* are predicted to be involved in

signal transduction pathways, biotic and abiotic stress tolerance. Another strong interaction observed among GrbZIP17 and GrbZIP60, two endoplasmic reticulum or unfolded protein response modulators known to get activated under environmental stresses (Humbert et al. 2012, Howell, 2013). Interaction of GrbZIP35/ABF1 and GrbZIP55/GBF3 is also observed, these genes are involved in abiotic stress tolerance.

Interacting proteins GrbZIP9, GrbZIP53 and GrbZIP56 (HY5), indicating their common role in developmental processes viz. seed maturation, circadian rhythm (Alonso et al. 2009; Wolfgang et al.2018). Interaction among C group members, *i.e.* GrbZIP9, GrbZIP10 and GrbZIP25 and S group members GrbZIP 2, GrbZIP53 predicting role in starvation adaptation. (Wolfgang et al.2018)

NLS sequence analysis of *G. raimondii* bZIP proteins

NLS sequences from bZIP proteins of *G. raimondii* and *Arabidopsis thaliana* are mentioned in table 4, some of the NLS sequences are sharing similarity among both species as well as in subfamily members. Subfamily C, G and S bZIP proteins are sharing similar NLS sequences.

Discussion

We have identified, structurally characterised and classified in detail the important plant transcription factor family of bZIP proteins, in three genomes of cotton. This information can be used in further crop improvement efforts. Phylogenetic analysis of *G. hirsutum, G. arboreum and G. raimondii* with *Arabidopsis* bZIP proteins shows close relation among functionally similar subfamilies according to their occurrence in same clades, thus suggesting bZIP family proteins are conserved in crucial biological functions in two dicot species. Even plant bZIP proteins are functionally similar in monocot and dicot species, *G. arboreum* and *G. raimondii* bZIPs are phylogenetically analysed with Oryza bZIPs, data not shown. Among *Gossypium* bZIPs, subfamily A and S are representing maximum bZIP genes, subfamily S members are smaller in length, with one or two exons and subfamily B protein is the longest one. Subfamily B, H, J and K of *Gossypium* bZIPs are single member families. *Gossypium* bZIPs Subfamily A, B, C, D, J, K and S members gene orthologs in different crop species are reported to function crucially in biotic-abiotic stress tolerance and starvation adaptation.

During discerning process of Gossypium bZIPs multiple bZIP genes with similar standard designation are observed, possibly due to gene duplication event. Gene duplication may have happen during speciation of dicot species and also during allotetraploidization. Whole-genome duplication in *G.arboreum* and *G.raimondii* before speciation is reported in previous reports. The findings from this study indicate that *Gossypium* bZIP proteins are important in stress adaptation, developmental processes, need further evaluation for trait development like fiber quality, yield and stress tolerance in cotton.

Abbreviations

ABF: ABRE binding factors

ABI5: ABA insensitive 5

ABRE: ABA-responsive element

AREB: ABA responsive element- binding protein

CAPS: Cleaved amplified polymorphisms

dCAPS: Derived-CAPS

CPRF2: Common plant regulatory factor 2

GBF: G-box binding factor

NPR1 (non expressor of pathogenesis-related genes1)

PHY: Phytochrome

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data used in the preparation of this article were obtained from the CottonGen database, Plant transcription factor database. All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

VK initiated, designed and implemented the study, data analysis, drafted the MS

AB conceived the study, data analysis and carefully edited the MS

BC directed the study, carefully edited the MS

RS edited the MS

All authors read and approved the final manuscript.

Acknowledgements

Authors would like to thank Mahyco for allowing to do work and publishing this article.

Authors' information

Affiliations

Mahyco Private Limited.

Vaishali Khanale, Dr. Anjanabha Bhattacharya, Dr. Bharat Char

Government Institute Of Science, Aurangabad

Dr. Rajendra Satpute

References

Abdurakhmonov, Ibrokhim Y.; Buriev, Zabardast T.; Abdukarimov, Abdusattor; Saha, Sukumar; Jenkins, Johnie N.; Pepper, Alan E. (2017). Cotton PHYA1 RNAi improves fiber quality, root elongation, flowering, maturity and yield potential in Gossypium hirsutum L. United States. Patent and Trademark Office; Texas A&M University

Abe M, Kobayashi Y, Yamamoto S, et al. FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. Science. 2005; 309(5737):1052-56.

Alonso R, Oñate-Sánchez L, Weltmeier F, et al. A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of Arabidopsis seed maturation gene expression based on heterodimerization and protein complex formation. Plant Cell. 2009; 21(6):1747-61.

Amir Hossain M, Lee Y, Cho J, et al. The bZIP transcription factor OsABF1 is an ABA responsive element binding factor that enhances abiotic stress signaling in rice. Plant Mol Biol**. 2010; 72:**557–66.

Assunção AG, Herrero E, Lin YF, et al. Arabidopsis thaliana transcription factors bZIP19 and bZIP23 regulate the adaptation to zinc deficiency. Proc Natl Acad Sci U S A. 2010; 107(22):10296-301.

Arabidopsis Information Resource (TAIR). http://arabidopsis.org

Azeem F, Tahir H, Ijaz U. *et al.* A genome-wide comparative analysis of bZIP transcription factors in *G. arboreum* and *G. raimondii* (Diploid ancestors of present-day cotton). Physiol Mol Biol Plants **26,** 433–444 (2020)

Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37: W202-W208.

Chen H, Chen W, Zhou J, et al. Basic leucine zipper transcription factor OsbZIP16 positively regulates drought resistance in rice. Plant Sci. 2012; 193-94:8-17.

Chen, Z.J., Sreedasyam, A., Ando, A. *et al.* Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet **52,** 525–533 (2020).

Choi HI, Hong JH, Ha J, Kang JY, Kim SY. ABFs, a family of ABA-responsive element binding factors. J Biol Chem. 2000; **275**: 1723–30.

cNLS Mapper. http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi.

CLC Genomics Workbench. https://www.qiagenbioinformatics.com/

CottonGen. Cotton Database Resources. http://www.cottongen.org

CottonFGD. Cotton Functional Genomic Database. https://cottonfgd.org

Das P, Lakra N, Nutan KK, Singla-Pareek SL, Pareek A. A unique bZIP transcription factor imparting multiple stress tolerance in Rice. Rice (N Y). 2019; 12(1):58.

Dietrich K, Weltmeier F, Ehlert A, et al. Heterodimers of the Arabidopsis transcription factors bZIP1 and bZIP53 reprogram amino acid metabolism during low energy stress. Plant Cell. 2011; 23(1): 381-95.

Dröge-Laser W, Snoek BL, Snel B, Weiste C. The Arabidopsis bZIP transcription factor family-an update. Curr Opin Plant Biol. 2018; 45:36-49.

Dröge-Laser W, Weiste C. The C/S_1 bZIP Network: A Regulatory Hub Orchestrating Plant Energy Homeostasis. Trends in plant sci. 2018; 23(5): 422-33.

El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427-D432.

Evens NP, Buchner P, Williams LE, Hawkesford MJ. The role of ZIP transporters and group F bZIP transcription factors in the Zn-deficiency response of wheat (Triticum aestivum). Plant J. 2017; 92(2):291-304.

Fan W, Dong X. In vivo interaction between NPR1 and transcription factor TGA2 leads to salicylic acid-mediated gene activation in Arabidopsis. Plant Cell. 2002; 14(6):1377-89.

Fu ZQ, Dong X. Systemic acquired resistance: turning local infection into global defense. Annu Rev Plant Biol. 2013; 64:839-863.

Gatz C. From Pioneers to Team Players: TGA Transcription Factors Provide a Molecular Link Between Different Stress Pathways. Mol Plant-Microbe Inter. 2013; 26(2): 151-59.

Gibalová A, Reňák D, Matczuk K, et al. *AtbZIP34* is required for Arabidopsis pollen wall patterning and the control of several metabolic pathways in developing pollen. Plant Mol Biol. 2009; **70:** 581–601.

Gibalová A, Steinbachová L, Hafidh S, et al. Characterization of pollen-expressed bZIP protein interactions and the role of ATbZIP18 in the male gametophyte. Plant Reprod. 2017; **30:** 1–17.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007; 8(2):R24.

Gutsche N, Holtmannspötter M, Maß L, et al. Conserved redox-dependent DNA binding of ROXY glutaredoxins with TGA transcription factors. Plant Direct. 2017; 1(6):e00030.

Hanson J, Hanssen M, Wiese A, et al. The sucrose regulated transcription factor bZIP11 affects amino acid metabolism by regulating the expression of ASPARAGINE SYNTHETASE 1 and PROLINE DEHYDROGENASE2. The plant Journal. 2008; 53: 935-49.

Howell SH. Endoplasmic Reticulum Stress Responses in Plants. Ann rev of plant bio. 2013; 64:477-99.

Humbert S, Zhong S, Deng Y, Howell SH, Rothstein SJ. Alteration of the bZIP60/IRE1 pathway affects plant response to ER stress in Arabidopsis thaliana. PLoS One. 2012;7(6):e39023.

Inaba S, Kurata R, Kobayashi M, Yamagishi Y, Mori I, Ogata Y, Fukao Y. Identification of putative target genes of bZIP19, a transcription factor essential for Arabidopsis adaptation to Zn deficiency in roots. Plant Journal. 2015; 84(2):323–34.

Jakoby M, Weisshaar B, Dröge-Laser W, et al. bZIP transcription factors in Arabidopsis. Trends Plant Sci. 2002; 7(3):106-111.

Jin JP, Tian F, Yang DC, Meng YQ, Kong L, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017; 45:D1040-5.

Kosugi S, Hasebe M, Entani T, Takayama S, et al. Design of peptide inhibitors for the importin α/β nuclear import pathway by activity-based profiling. Chem. Biol. 2008; 15:940-49.

Kosugi S, Hasebe M, Tomita M, Yanagawa H. Systematic identification of yeast cell cycle-dependent nucleocytoplasmic shuttling proteins by prediction of composite motifs. Proc. Natl. Acad. Sci. USA. 2009; 106: 10171-76.

Kosugi S, Hasebe M, Matsumura N, Takashima H, Miyamoto-Sato E, et al. Six classes of nuclear localization signals specific to different binding grooves of importin α. J. Biol. Chem. 2009; 284: 478-85.

Kumar S, Stecher G, Li M, Knyaz C, and Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution.**2018**; **35**:1547-154

Li F, Fan G, Wang K, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet**. 2014; 46:** 567–72.

Liang C, Meng Z, Meng Z, et al. GhABF2, a bZIP transcription factor, confers drought and salinity tolerance in cotton (Gossypium hirsutum L.). Sci Rep. 2016; 6:35040.

Liao Y, Zou H, Wei W, et al. Soybean *GmbZIP44*, *GmbZIP62* and *GmbZIP78* genes function as negative regulator of ABA signaling and confer salt and freezing tolerance in transgenic *Arabidopsis* . Planta. 2008; **228:** 225–40.

Lindemose S, O'Shea C, Jensen MK, Skriver K. Structure, function and networks of transcription factors involved in abiotic stress responses. Int J Mol Sci. 2013; 14(3):5842-78.

Liu JX, Srivastava R, Howell SH. Stress-induced expression of an activated form of AtbZIP17 provides protection from salt stress in Arabidopsis. Plant Cell Environ. 2008; 31(12):1735-43.

Ma J, Hanssen M, Lundgren K, et al. The sucrose-regulated Arabidopsis transcription factor bZIP11 reprograms metabolism and regulates trehalose metabolism. New Phytol. 2011; 191(3): 733-45.

MEME Suite. http://meme.nbcr.net

Moon SJ, Park HJ, Kim TH, et al. OsTGA2 confers disease resistance to rice against leaf blight by regulating expression levels of disease related genes via interaction with NH1. PLoS One. 2018; 13(11):e0206910.

Nazri AZ, Griffin JHC, Peaston KA, Alexander-Webber DGA, Williams LE. F-group bZIPs in barley-a role in Zn deficiency. Plant Cell Environ. 2017; 40(11):2754-70.

Nijhawan A, Jain M, Tyagi AK, Khurana JP. Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. Plant Physiol. 2008; 146(2):333-50.

Pfam database. https://pfam.xfam.org

Plant transcription factor database. http://planttfdb.gao-lab.org/.

Pyo H, Demura T, Fukuda H. Vascular cell expression patterns of Arabidopsis bZIP group I genes. Plant Biotechnol. 2006; 23:497-501.

Rhee SY, Beavis W, Berardini TZ, et al. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*. 2003; 31(1):224-228.

Rolly NK, Imran QM, Lee IJ, Yun BW. Salinity Stress-Mediated Suppression of Expression of Salt Overly Sensitive Signaling Pathway Genes Suggests Negative Regulation by *AtbZIP62* Transcription Factor in *Arabidopsis thaliana*. Int J Mol Sci. 2020; 21(5):1726.

Sakai H, Lee SS, Tanaka T, et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. 2013; 54(2):e6.

Schwarz, R. and Dayhoff, M. Matrices for Detecting Distant Relationships. In: Dayhoff, M., Ed., Atlas of Protein Sequences. National Biomedical Research Foundation 1979; 353-358.

Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 2000; 28(1):231-234.

Shen H, Cao K, Wang X. A conserved proline residue in the leucine zipper region of AtbZIP34 and AtbZIP61 in *Arabidopsis thaliana* interferes with the formation of homodimer. Biochem Biophys Res Commun. 2007; 362:425–430.

Sirichandra C, Davanture M, Turk BE, et al. The Arabidopsis ABA-activated kinase OST1 phosphorylates the bZIP transcription factor ABF3 and creates a 14-3-3 binding site involved in its turnover. PLoS One. 2010; 5(11):e13935.

Skubacz A, Daszkowska-Golec A, Szarejko I. The Role and Regulation of ABI5 (ABA-Insensitive 5) in Plant Development, Abiotic Stress Responses and Phytohormone Crosstalk. Front Plant Sci. 2016; 7:1884.

SMART database (Simple Modular Architecture Research Tool). http://SMART.embl-heidelberg.de

The String database. https://string-db.org/

Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47(D1):D607-D613.

Tang N, Zhang H, Li X, Xiao J, Xiong L. Constitutive activation of transcription factor OsbZIP46 improves drought tolerance in rice. Plant Physiol. 2012; **158**: 1755–68.

The Rice Annotation Project Database, RAP-DB. http://rapdb.dna.affrc.go.jp/

Ullah I, Magdy M, Wang L, et al. Genome-wide identification and evolutionary analysis of TGA transcription factors in soybean. Sci Rep. 2019; doi: 10.1038/s41598-019-47316-z.

Wang B, Liu C, Zhang D, He C, Zhang J, Li Z. Effects of maize organ-specific drought stress response on yields from transcriptome analysis. BMC Plant Biol. 2019; 19(1):335.

Wang K, Wang Z, Li F. et al. The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet**. 2012; 44:**1098–1103

Wang X, Lu X, Malik W, Chen X, Wang J, et al. Differentially expressed bZIP transcription factors confer multi-tolerances in Gossypium hirsutum L. In J of Bio Macromo. 2020; *146*: 569–78.

Weltmeier F, Rahmani F, Ehlert A, et al. Expression patterns within the Arabidopsis C/S1 bZIP transcription factor network: availability of heterodimerization partners controls gene expression during stress response and development. Plant Mol Biol. 2009; 69(1-2):107-19.

Xiang Y, Tang N, Du H, Ye H, Xiong L. Characterization of OsbZIP23 as a key player of the basic leucine zipper transcription factor family for conferring abscisic acid sensitivity and salinity and drought tolerance in rice. Plant Physiol.2008; **148**:1938–52.

Yoshida T, Fujita Y, Maruyama K, et al. Four Arabidopsis AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signalling in response to osmotic stress. Plant Cell Environ. 2015; 38(1):35-49.

Yoshida T, Mogami J, Yamaguchi-Shinozaki K. Omics Approaches Toward Defining the Comprehensive Abscisic Acid Signaling Network in Plants. Plant Cell Physiol. 2015; 56(6):1043-52.

Yu J, Jung S, Cheng CH, et al. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res. 2014; 42:D1229-D1236.

Zhang B, Liu J, Yang ZE, et al. Genome-wide analysis of GRAS transcription factor gene family in Gossypium hirsutum L. BMC Genomics. 2018; 19(1):348.

Zhang YN, Cai DR, Huang XZ. Identification of BZIP Protein Family in Gossypium arboreum and Tissue Ex-pression Analysis of GaFDs Genes. Acta Agro Sin. 2016; 42:832-43.

Zhu T, Liang C, Meng Z, et al. CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol. 2017; 17(1):101.

Tables

Table 1: *Gossypium* bZIP proteins classification and predicted function

Phylogenetic Group	Gossypium bZIP number	Biological Function
SubfamilyA-Subgroup A1 (A-A1)	Gossypium bZIP12 Gossypium bZIP35 Gossypium bZIP36 Gossypium bZIP37 Gossypium bZIP39 Gossypium bZIP66 Gossypium bZIP67	Express during seed maturation and ABA - stress response, strong interaction with subfamily D
A-A2	Gossypium bZIP13 Gossypium bZIP40	Plant immunity and abiotic stress responses
A-A3	Gossypium bZIP14 Gossypium bZIP27	Reulation of flowering
Subfamily B	Gossypium bZIP17	Response to ER stress , interact with bZIP60
Subfamily C (C-C1) C-C2	Gossypium bZIP9 Gossypium bZIP10 Gossypium bZIP25 Gossypium bZIP63	Anther development, Positive regulation of seed maturation and response to starvation , interacts with S1 family member during starvation adaptation
Subfamily D	Gossypium bZIP20 Gossypium bZIP21 Gossypium bZIP22 Gossypium bZIP26 Gossypium bZIP45 Gossypium bZIP46 Gossypium bZIP47 Gossypium bZIP50 Gossypium bZIP57 Gossypium bZIP65	Involved in anther development, involved in salicylic acid mediated signaling pathway and pathogen response (SAR); MEME motif search found DELAY OF GERMINATION 1 -Seed dormancy control motif
Subfamily E	Gossypium bZIP34 Gossypium bZIP61 Gossypium bZIP76	Cell wall and pollen development
Subfamily F	Gossypium bZIP19 Gossypium bZIP23 Gossypium bZIP24	Involved in the adaptation to zinc deficiency
Subfamily G	Gossypium bZIP16 Gossypium bZIP41 Gossypium bZIP55	Binds to G box motif of genes regulated by hormones and light
Subfamily H	Gossypium bZIP56	Involved in phyB signaling pathway/ Circadian Rhytham/ Fiber development
Subfamily I	Gossypium bZIP18 Gossypium bZIP29 Gossypium bZIP30 Gossypiumb ZIP51 Gossypism bZIP52 Gossypium bZIP59 Gossypium bZIP69	Involved in pollen and vascular development
Subfamily J	Gossypium bZIP62	Salt stress tolerance
Subfamily K	Gossypium bZIP60	Response to ER stress
Subfamily S-Subgroup S1 (S-S1)	Gossypium bZIP1 Gossypium bZIP2 Gossypium bZIP11 Gossypium bZIP44 Gossypium bZIP53	Involved in the positive regulation of seed germination, Anther development, Interaction C subfamily members in starvation adaptation ,
S-S2	Gossypium bZIP4 Gossypium bZIP5 Gossypium bZIP6 Gossypium bZIP7	Involved in positive regulation of transcription, Arabidopsis bZIPs of this group are express in leaf, pollen,embryo, seed and root, Invito somatic
S-S3	Gossypium bZIP42 Gossypium bZIP43 Gossypium bZIP48 Gossypium bZIP58	In vitro somatic embryogeneis and stress response , Embryo deveolpment
	Gossypium bZIP70	Protein dimerization activity and regulation of transcription

Table 2: *G. hirsutum*, *G. arboreum* and *G. raimondii* bZIP genes distribution and structural characteristics.

G. hirsutum, *G. arboreum* and *G. raimondii* bZIP genes distribution and structural characteristics

Subfamily		Number of bZIP genes				Structural characteristics			
		G. hirsutum		G.arboreum	G. raimondii	Total bZIP genes	Exon no./ maximum frequency	Protein length range (aa)	
		A	D						
	bZIP12	1	*	1	1				
	bZIP13	2	4	1	2				
	bZIP14	5	5	2	3				
	bZIP27	*	*	1	1				
Subfamily A	bZIP35	*	1	2	1	45	2 to 5 / 3&4	101 to 453	G. hirsutum
	bZIP36	3	4	2	4	18	2 to 8/ 3&4	202 to 615	G.arboreum
	bZIP37	3	1	1	1	20	2 to 4/ 3&4	181 to 427	G. raimondii
	bZIP39	3	2	2	2				
	bZIP40	*	*	1	1				
	bZIP66	3	4	3	3				
	bZIP67	2	2	2	1				
Subfamily B	bZIP17	1	1	1	1	2	2 &3	676 & 690	G. hirsutum
						1	2	690	G.arboreum
						1	2	689	G. raimondii
	bZIP9	1	2	2	1				
Subfamily C	bZIP10	1	1	1	1	13	5 to 7 / 6	225 to 451	G. hirsutum
	bZIP25	2	2	1	1	5	4 to6/ 6	255 to 426	G.arboreum
	bZIP63	2	2	1	1	4	5&6/ 6	329 to 451	G. raimondii
	bZIP20	*	*	1	1				
	bZIP21	3	3	3	1				
	bZIP22	*	*	1	2	37	1 to 13/ 8.11&12	157 to 510	G. hirsutum
	bZIP26	*	*	1	*				
Subfamily D	bZIP45	4	4	2	1	16	6 to11/ 8&11	183 to 484	G.arboreum
	bZIP46	2	5	2	1	10	1 to 12/ 8	193 to 472	G. raimondii
	bZIP47	2	2	2	1				
	bZIP50	3	3	2	1				
	bZIP57	1	1	1	1				
	bZIP65	2	2	1	1				
Subfamily E	bZIP34	2	4	2	1	23	4&5/ 4	166 to 382	G. hirsutum
	bZIP61	5	3	1	2	4	4	171 to 372	G.arboreum
	bZIP76	4	5	1	1	4	4&5	281 to 375	G. raimondii
Subfamily F	bZIP19	2	1	1	1	7	1&2/ 2	170 to 270	G. hirsutum
	bZIP23	NF	3	1	1	2	1	250&265	G.arboreum
	bZIP24	NF	1	*	*	2	2	258&270	G. raimondii
Subfamily G	bZIP16	2	2	2	1	14	9 to 12 / 11&12	321 to 407	G. hirsutum
	bZIP41	2	2	1	2	6	11&12 /12	366 to 407	G.arboreum
	bZIP55	3	3	3	3	6	10 to 12/ 11	277 to 401	G. raimondii
Subfamily H	bZIP56	1	1	1	1	2	4	170&175	G. hirsutum
						1	4	170	G.arboreum
						1	4	170	G. raimondii
	bZIP18	4	4	3	3				
	bZIP29	3	3	3	2				
Subfamily I	bZIP30	*	*	*	1	26	4 & 5/ 4	206 to 571	G. hirsutum
	bZIP51	2	2	1	2	11	4	213 to 568	G.arboreum
	bZIP52	*	*	1	1	12	2 to5/ 4	213 to 570	G. raimondii
	bZIP59	*	*	1	1				
	bZIP69	3	5	2	2				
Subfamily J	bZIP62	1	NF	1	1	1	8	591	G. hirsutum
						1	8	590	G.arboreum
						1	9	605	G. raimondii
Subfamily K	bZIP60	1	3	1	1	4	2&3	105 to 292	G. hirsutum
						1	3	268	G.arboreum
						1	3	292	G. raimondii
	bZIP1	1	1	1	1				
	bZIP2	1	1	1	1				
	bZIP4	1	1	1	1				
	bZIP5	4	6	4	3				
	bZIP6	*	*	1	1	54	1&2/ 1	117 to 203	G. hirsutum
Subfamily S	bZIP7	1	1	*	1	25	1&2/ 1	117 to 198	G.arboreum
	bZIP11	4	3	4	2	24	1&2/ 1	137 to 209	G. raimondii
	bZIP42	4	3	3	2				
	bZIP43	*	*	1	1				
	bZIP44	3	6	3	5				
	bZIP53	4	4	4	4				
	bZIP58	1	2	1	1				
	bZIP70	1	1	1	1				

Table 2: *G. hirsutum*, *G. arboreum* and *G. raimondii* bZIP genes distribution and structural characteristics, * indicates similar gene may distributed among family.

Table 3: *G. hirsutum*, *G. arboreum* and *G. raimondii* bZIPproteins MEME motif analysis, motif function and occurrence

Sr. No	Motif sequence	pfam/ Interpro/ GenomeNet/Tomtom	Occurrence
1	EKRQRRMISNRESARRSRLRKQAYLQELE / EKRQRRMJSNRESARRSRMRKQAYLEELE/ RRLASNRESARRSRLRKQ/ KIDEKRQRRMJSNRESARRSRMRKQAYLEELESKVQKLREE	bZIP1 domain	228 GhbZIP proteins/ 91 GabZIPs and 86 GrbZIPs
2	LRAENSELKARLTEL	None predicted/ Like Leucine heptad L-X6-L repeat motif	Present in almost all GhbZIPs
3	QLRKENHQLLNKJNFLTZHYHKVEAENSVL	bZIP1 domain (L-X6-L- Leucine repeats)	Subfamily C, G,J and S (GabZIPs)
4	NHQLLDKLNHVSHKYDEVEAENAVLKAZASELRQKLKDLN	Homeobox associated leucine zipper domain	Subfamily C and S of GrbZIPs
5	LRIJVDSVMAHYDELFRLKSTAAKADVFHLJSGMWKTPAERCFLWIGGFR **And** SELLKVLVNQLEPLTEQQLMGICNLQQSSQQAEDALSQGMEALQQSLSDT	DOG1	Subfamily D bZIPs (GhbZIPs) And GhbZIP76-1A &9D, GhbZIP44-6D, 11-2A
6	ETLEGFVRQADNLRQQTLQQMHRILTTRQAARALLAJGEYFSRLRALSSL **And** FDMEYARWLEEHNRQINELRTALNSHLSDAELRIJVESVLAHYDEIFRLK **And** KADVFHLLSGMWKTPAERCFLWJGGFRPSELLKVLVNQLEPLTEQQLMGI	DOG1	Subfamily D bZIPs (GabZIPs)
7	LAHYDEJFRLKSTAAKADVFYLJSGMWKTPAERFFLWIGGFRPSELLKVL **And** SGTVNSGIAAFEMEYARWVEEQNRQICELRTALNAHISDIELRILVESG/ ALEGFVRQADHLRQQTLQQMHRILTTRQAARGLLAJGEYFHRLRALSSL	DOG1	Subfamily D bZIPs (GrbZIPs)
8	DNLRQQTLQQMHRILTTRQAARALLAJGEYFSRLRALSSLW/ CNLQQSSQQAEDALSQGMEALQQSLAETVA/ PQJEPLTEQQLLEVCNLQQSSQQAEDALSQGLEKLQQSLAETVASG	TF TGA2 related	Subfamily D bZIPs (Gh , Gr and Ga) GabZIP25 and GabZIP62
Sr. No	**Motif sequence**	**pfam/ Interpro/ GenomeNet/Tomtom**	**Occurrence**
9	HSNIGSGAAAFDMEYGRWLEEHNRQICELRTALNAHLSDIE	TGACG sequence specific DNA binding protein	Subfamily D bZIPs (GhbZIPs) And GhbZIP9-1A&2D
10	RGKTLGEMTLEEFLVKAGVVE/ TLGQRGKTLGEMTLEEFLVNAGVVEENQQ	Sterile alpha motif (SAM)/Pointed domain Protein- Protein Interaction	Subfamily A bZIPs (GhbZIPs, GrbZIPs, GabZIPs)
11	GGGLVSQGSLQRQGSLTLPRTLSQKTVDEVWKEIQKEQDGG	Abscisic acid Insensitive 5-Like Protein 5-related	Subfamily A and J bZIPs (Gh) and GhbZIP-3A&7D
12	QTLQTEATTLSAQLTLLQRDTTGLTTENSELKLRLQAMEQQAQLRDALNE	coiled-coil region of GIT (G protein-coupled receptor kinase-interacting) proteins, must be involved in Protein interactions	Subfamily I and E members (GabZIPs)
13	GKGLVSQGSLQRQGSLTLPRTLSKKTVDEVWKEIQKE	GluR7, glutamate receptor domain GluRProteins with	Subfamily A of GabZIPs,

		GluR7 domain function in light signal transduction and calcium homeostasis	GabZIP18-3, GabZIP11-3
14	VPQJEPLTEQQLLEICNLQQSSQQAEDALSQGLEKLQQSLAETVASG	Tetratricopeptide repeats, involved in protein-protein interactions	Subfamily D bZIPs (GrbZIPs)
15	FVRQADHLRQQTLQQMHRILTTRQAARGLLAJGEYFHRLRALSSLWAARP	The RPA interacting motif is conserved in eukaryotes DNA repair and replication	Subfamily D bZIPs (GrbZIPs)-
16	SQQYTLVZRDNSVLRAZNSELKQRLQSLE	TOMTOM Protein kinase domain found in regulators of intracellular signal-transduction pathways in eukaryotes	Subfamily C, E,I and S members of GrbZIPs
17	RQPTLGEMTLEDFLVKAGVVREDS	MYND-Zinc binding domain, involved in protein -protein interaction	Subfamily A of GrbZIPs (except GrbZIp14 &27)
Sr. No.	**Motif sequence**	**pfam/ Interpro/ GenomeNet/Tomtom**	**Occurrence**
18	QNTLGGLGKDFGSMNMDELLKNIWTAEENQT	SCOP Domain d1g2na-Ligand binding domain	GrbZIP12,13, 35,36,37,39,66,67 and 4
19	ATTLSAQLTLLQRDSVGLTTZNNELKLRLQAMEQQAQLRDALNEALKAEV	SecD export protein N-terminal TM region	Subfamily C (GrbZIP9), E,G and I of GrbZIPs
20	GASGNVABYMGQMAMAMGKLETLEGFVRQ	None predicted	Subfamily D bZIPs (GhbZIPs)
21	SRLKLTQLEQELQRARQQGIF	None predicted	Subfamily D bZIPs (GhbZIPs)
22	RLKLTQLEQELZRARQQGIFI	None predicted	Subfamily A and D of GabZIPs and GabZIP19
23	ZVEQLREENAELRKK	None predicted	Allmost in all subfamily members of GrbZIPs
24	MSPVVSEILRSGFMINSSLRRRTHLVQSFSVVFLYCGNS	None predicted	GrBZIP2, 11 and 44

Table 4: NLS sequences of *Gossypium raimondii and Arabidopsis thaliana* bZIP proteins

bZIP No.	*Gossypium raimondii*	*Arabidopsis thaliana*
bZIP1	[illegible]	EKKRKRKKL,KKRKRK
bZIP2	RRSRMRKQ	RKRKRMLSNRESARRSRMRKQK, RRSRMRKQ
bZIP3	---	RRSRMRKQ
bZIP4	KKRRMK	RRRTISNRESAKRSRMKKKKR
bZIP5	RRKISNRESAKRSRWRKKR, RKKRYLENLTNQVNKMNIENRRLK	ERKKKRKL, RKKKRKLSNRESAKRSREKKQK,KKKRKL
bZIP6	RRSRMRKQ, VIDERKRRRMI, DERKRRRMIS	KRSRMR
bZIP7	RKRRR, RKRRRMISNRESARRSRMRKQK, RRSRMRKQ	RKKMIQPEMIDERKRKRKRSRMR
bZIP8	---	RKRRR,RRSRMRKQ
bZIP9	LKRLRRKL, RRKLSNRESARRSRKRK, RRSRKRK,RRSRKRKQ, SRKRKQ	SRRRKQ
bZIP10	SRRRKQ,RRSRRRKQ	RRSRRRKQ,SRRRKQ
bZIP11	RKRKRMESNRESARRSRMRKQK, RRSRMRKQ	RKRKRMLSNRESARRSRMKKQK
bZIP12	RKRGAHEDIVEKTVERRGKRMI	---
bZIP13	ARGGRRAVE	---
bZIP14	RSRARKQAYTNELELEVARLLEENVKLRR, DPRHKRKIKNR	---
bZIP16	RRSRLRK,RRSRLRKQ	RRSRLRK,RRSRLRKQ
bZIP17	RRVSKRK,RQNATQLNKGRNRR,RRVSKRK KGNE	---
bZIP18	GGGKGRHRYSNSIDGCSLMESIEAKKAMSPD	---
bZIP19	YREKKKAHTAY	---
bZIP20	RFGRGQMNAIVASDSSDRSKDKTDQKTLRR	---
bZIP21	RLRKKAYVQQLESSRIKLTQLEQELQRAR	---
bZIP22	RKSRLRKKAYVQQLESSRLKLAQLEQELERAR	---
bZIP23	KIVPASTEDEAAADDTEGSRLKKPKKRS	---
bZIP25	SRRRKQ,RRSRRRKQ	RRSRRRKQ,SRRRKQ
bZIP27	RSRARKQAYTNELELEVAHLMEENARLKRQQE	---
bZIP29	DPKRAKRILANRQSAARSKERKMRYIS	---
bZIP30	ARSKQRKMQ	---
bZIP31	---	ARSKQKKLQ
bZIP34	DPKKLKRVISNRLSAQRSRIRKIQRL	---
bZIP36	RRQRRMIKNRESAARSRARKQAYTTELEAKLAKLK	---
bZIP37	RPWGRKRRCLR, PWGRKRRCLRR	---
bZIP39	RSRARKQAVTVELEAELNQLKQENTHLKQALVE, ELERKRKQQV	---
bZIP41	RRSRLRK,RRSRLRKQ	RRSRLRK,RRSRLRKQ
bZIP42	RRSRMRKQ	RRSRMRKQ
bZIP43	RRSRMRKQ, RSRMRKQRHLDELW	RRSRMRKQ
bZIP44	RKRKRMESNRESARRSRMRKQK, RRSRMRKQ	RKRKRK, RKRKRKQSNRESARRSRMRKQK,RRSRMRK
bZIP45	RKSRLRKKAYVQQLESSRLKLTQLEQELQRAR	---
bZIP46	DTNHKNQLHGVQHGAVMVKTIDQSKSKSNDQK	---
bZIP47	RKSRLRKKAYVQQLENSRLKLIQLEQELQRAT	---
bZIP48	RRSRMRKQ, RSRMRKQKHLDELW	RRSRMRKQ
bZIP50	RKSRMRKKAYVQQLESSRLKLAQLEQELERAR	---
bZIP51	DGVKKTMAPDRLAELALIDPKRAKRILA	---
bZIP52	EGGKGRHRYSNSVDGCSLMESIEAKKAMAPD	---
bZIP53	RKRKRMLSNRESARRSRIRKQK, RRSRIRKQ	RKRKRMISNRESARRSRMRKQK,RRSRMRKQ
bZIP54	---	RRSRLRK,RRSRLRKQ
bZIP55	RRSRLRK,RRSRLRKQ	RRSRLRK,RRSRLRKQ
bZIP58	RRSRMRKQ	RRSRMRKQ
bZIP59	EAKKALSATKLAELAIIDPKRAKRIWA	
bZIP60	RRKLRNR, RRKLRNRDAAVRSRERKK	KKRRRR,KRRRRV, KRRRRVRNRDAAVRSRERKK, RRRVRNRDAAVRSRERKK
bZIP62	RKRRKE	RKRRKE
bZIP63	RRSRRRKQ,SRRRKQ	RRSRRRKQ,SRRRKQ
bZIP65	RKSRLRKKAYVQQLESSRIRLTQLEQELQRARA	---
bZIP66	RKRGAPEDIVEKNVERRGKRMI	---
bZIP67	RSRARKQAYTVELELELNQLKQENAKLKQL	---
bZIP68	---	RRSRLRK,RRSRLRKQ
bZIP69	DSKKAMSAAKLAELALIDPKRAKRIWA	
bZIP70	RRSRMRKKQQIEELQSQVNQLQTINRQLSQKLIN	---
bZIP76	EEVRYKNMISAPSWASLEFDYHSKTIDAQLAAFY	---

Table 4: NLS sequences of *Gossypium raimondii and Arabidopsis thaliana* bZIP proteins, black font from plant transcription factor database, colour fonts from cNLS Mapper predictions, red font >8 score (predicting nuclear localization), blue font 7 or <8 score (predicting partial nuclear localization), orange font >3and<7 score (nuclear and cytoplasm localization).

Figures

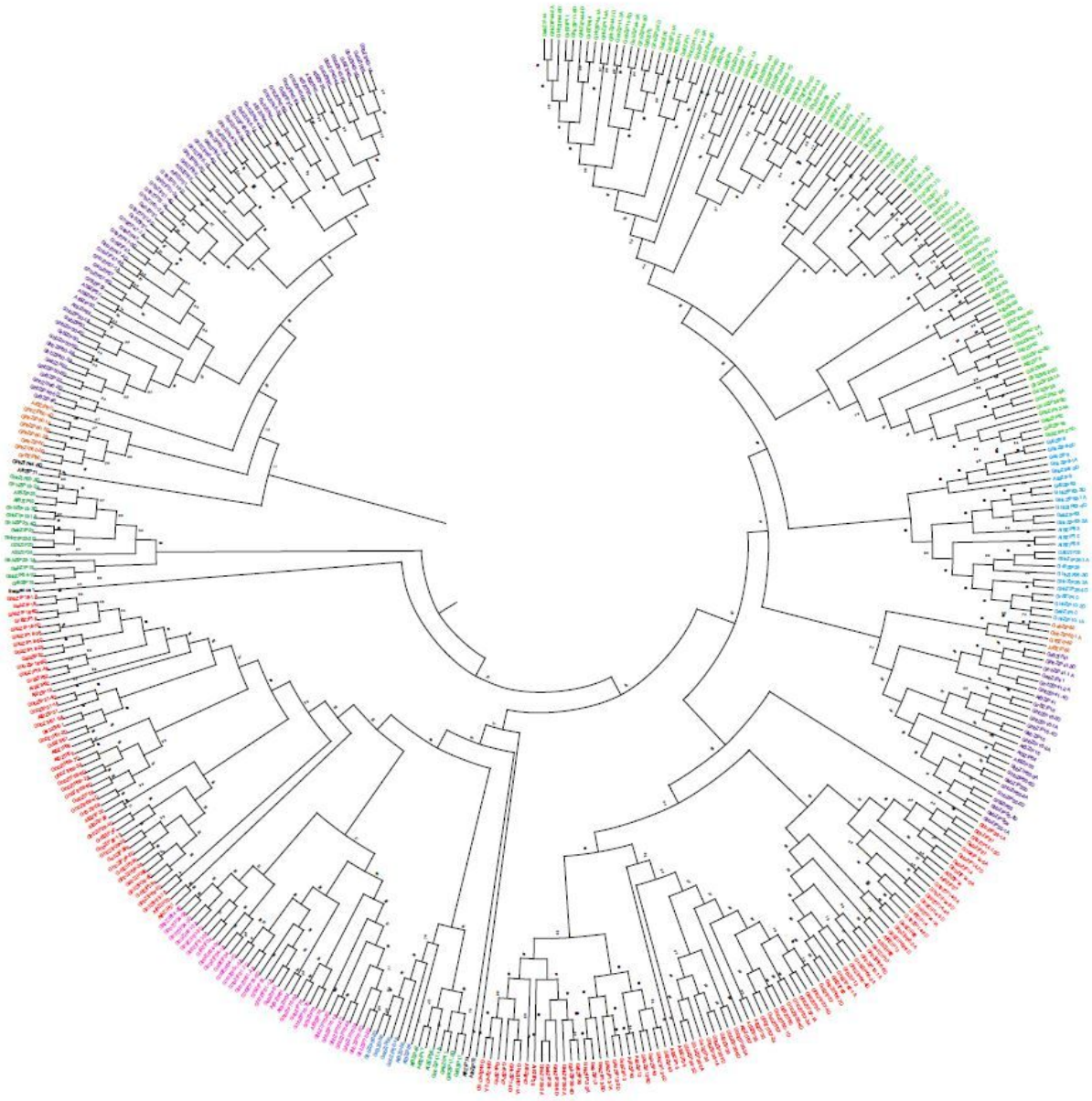

Figure 1

A Phylogenetic tree construction of G. hirsutum, G. arboreum, G.raimondii and Arabidopsis thaliana bZIPs.

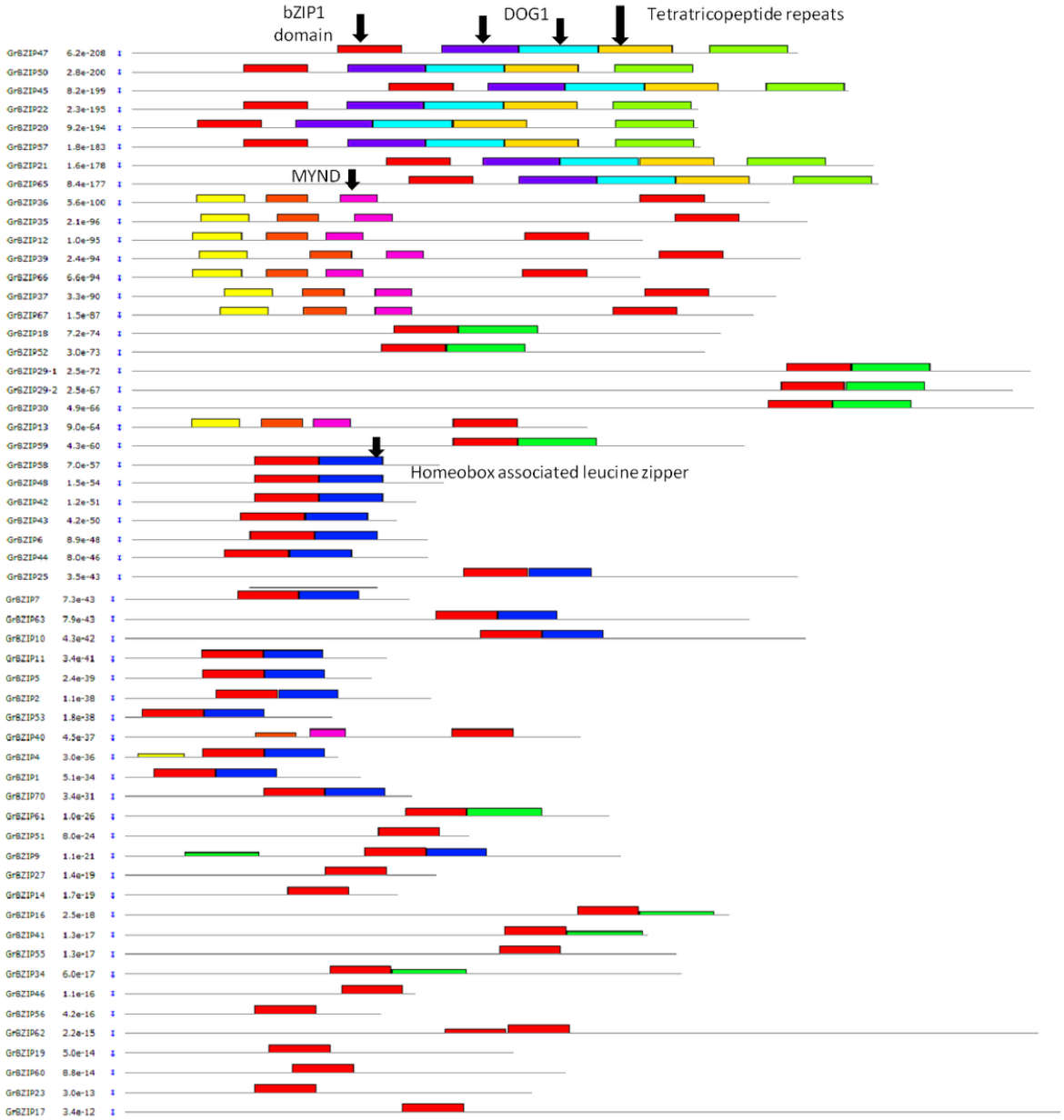

Figure 2

A typical example of MEME motif analysis of G. raimondii bZIPs. MEME motif analysis of Gossypium raimondii bZIPs, bZIP1 domain, DOG1 domain, tetraticopeptide repeat motif, MYND domain, Homeobox associated leucine zipper domain are marked by arrow.

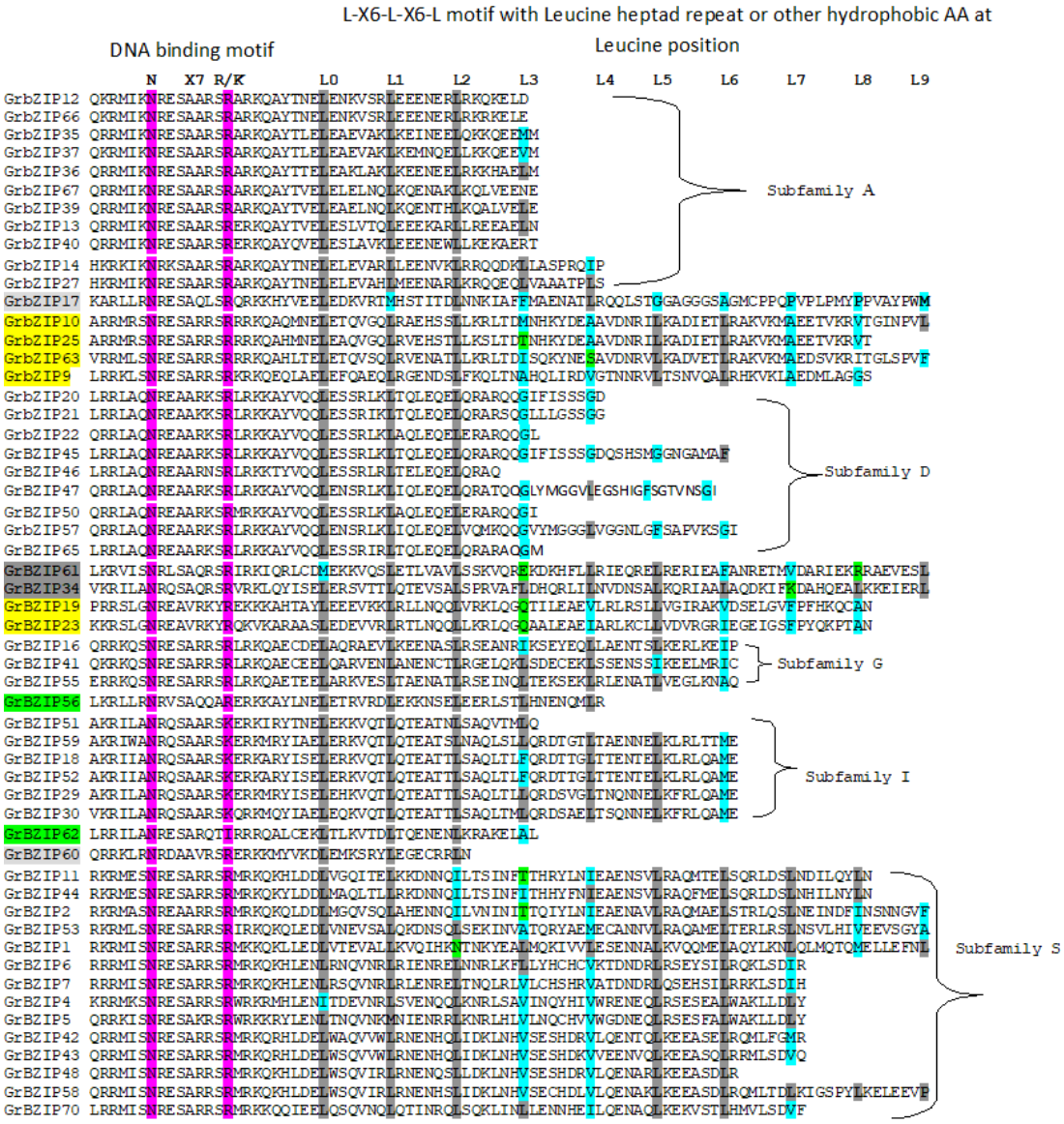

Figure 3

Leucine heptad repeats variation in G. raimondii bZIPs. Leucine heptad repeats variation in G. raimondii bZIPs: basic region motif N-X7-R/K is highlighted in magenta colour, leucine position in luecine heptad motif is highlighted in gray, hydrophobic amino acids other than leucine are highlighted in blue and polar amino acids are highlighted in green

Figure 4

STRING protein interaction network of G. raimondii bZIPs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MASTfor4Ghand1GrremainingbZIPs.html
- MAST86GrbZIPs.html
- MASTGrbZIPsselectedonly1stGeneofeachbZIPs.html
- MAST91GabZIPs.html
- GabZIPMAST1stgeneofeachbZIP.html
- MASTforallGhbZIPs.html
- Supplementaryfile1.xlsx