Genome-Wide Identification and Characterization of GATA Family Genes in Brassica Napus

Weizhuo Zhu
Zhejiang University
Yiyi Guo
Zhejiang University
Yeke Chen
Zhejiang University
Dezhi Wu (wudezhi@zju.edu.cn)
Zhejiang University https://orcid.org/0000-0002-7900-0542
Lixi Jiang
Zhejiang University

Research article

Keywords: Brassica napus, GATA, Genome-wide, expression patterns, SNP distribution

Posted Date: June 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-30607/v1

Version of Record: A version of this preprint was published on December 4th, 2020. See the published version at https://doi.org/10.1186/s12870-020-02752-2.

Abstract

Background: *GATA* transcription factors are involved in plant developmental processes, which also respond to environmental stresses through binding DNA regulatory regions to regulate the downstream genes. However, little information of *GATA* genes is available in *Brassica napus*, one of important oil crops. The release of *B. napus* genome sequences provides a good opportunity to perform a genome-wide characterization of *GATA* family genes in the rapeseed genome.

Results: In this study, 96 *GATA* genes randomly distributing on 19 chromosomes were identified in *B. napus*. They were classified into four subfamilies on the basis of phylogenetic analysis and the domain structures. The sequences of BnGATAs were obvious divergence among four subfamilies based on their GATA domains, structures and motif compositions. Gene duplication and the synteny between the genomes of *B. napus* and *A. thaliana* were also analyzed to provide insights into the evolutionary characteristics. Moreover, *BnGATAs* showed different expression patterns in various tissues and under diverse abiotic stresses. Single nucleotide polymorphisms (SNPs) distribution analysis suggests functional disparity of *GATA* family genes in different genotypes of a core collection germplasm.

Conclusion: This study examined genomic structures, evolution features, expression patterns and SNP distribution of 96 *BnGATAs.* The results enriched our knowledge of the *GATA* genes in rapeseed.

Introduction

Transcription factors (TFs) regulate gene expression by recognizing and combining *cis*-acting elements on the promoter regions of target genes (Franco et al., 2014). TFs play key roles in the regulation of developmental processes, hormones signaling pathways and disease resistance in plants. There are several well-known transcription factor families in plants including *WRKY*, *MYB* (*V-myb avian myeloblastosis viral oncogene homolog*), *DREB* (*Dehydration-responsive element-binding protein*), *bZIP* (*Basic region-leucine zipper*), *MADS-box* and *GATA* (GATA-binding factor). Among them, the *GATA* IFs are characterized as important regulators for many biological processes, such as flower development, carbon and nitrogen metabolisms (Reyes et al., 2004). *GATA* genes could recognize and bind to (T/A)GATA(A/G) sequences to regulate the transcription of their downstream genes (Lowry et al., 1999; Scazzocchio, 2000). The DNA binding domains of GATA proteins contain a type IV zinc finger structure C-X_2-C-X_{17-20}-C-X_2-C and a conserved basic follow region, and most of them featured with C-X_2-C-X_{18}-C-X_2-C or C-X_2-C-X_{20}-C-X_2-C zinc finger domains (Lowry et al., 1999; Reyes et al., 2004; Yuan et al., 2017; Wang et al., 2019). Generally, the *GATA* family genes could be divided into four subfamilies as subfamily I, II, III and IV in plant species such as *Arabidopsis thaliana* based on phylogenetic relationships, DNA binding domains and intron-exon structures (Reyes et al., 2004; Yuan et al., 2017; Zhang et al., 2018; Zhang et al., 2019).

Many studies have been proved that the *GATA* TFs are responsible for plant growth development, flowering, chlorophyll synthesis, greening and senescence. For instance, the loss-of-function and over-expression plants of *GATA* genes such as *GNC* (*GATA*, Nitrate-inducible, Carbon-metabolism) and *GNL*

(*GNC*-like) can change flowering time and chlorophyll synthesis in *A. thaliana* (Richter et al., 2010, 2013a, 2013b; Yan et al., 2018). The *GNC* TF can regulate genes such as the light-labile factors *PIFs* (*phytochrome interacting factors*) to control chloroplast biogenesis and stomatal index (Richter et al., 2010; Yan et al., 2018). The cross-repressive interaction pathway between *GNC*/ *GNL* and *MADS-box* transcription factor *SOC1* (*Suppressor of Overexpression of Constans1*) affect flowering time (Richter et al., 2013b; Yan et al., 2018). Besides, *GNC* and *GNL* are considerable repressors of gibberellin signaling through being regulated by DELLA and *PIF* regulators (Naito et al., 2007; Richter et al., 2010). Moreover, auxin response factors *ARF2* and *ARF7* can repress the expression of *GNC* and *GNL* genes (Naito et al., 2007; Richter et al., 2010, 2013a, 2013b). The *GATA* member *BnA5.ZML1* in *B. napus* was reported to be a stigma compatibility factor (Duan et al., 2020). The *GATA* member *PdGNC* in *Populus* plays a crucial role in photosynthesis and plant growth (An et al., 2019). In wheat, overexpression of *TaZIM-A1,* a member of the *GATA* family, caused the delay of flowering and the decrease of thousand-kernel weight (Liu et al., 2018).

The *GATA* TFs in plants also respond to diverse abiotic stresses. Under cold stress, the expression levels of *GNC* and *GNL* were significantly increased, while the seedling survival ratio was elevated when *GNC* and *GNL* genes were overexpressed in the transgenic lines of *A. thaliana* (Lee and Lee, 2010). Moreover, under low temperature, *GATA9* gene showed remarkable change of its expression level, resulting in activating its downstream genes in *Vigna subterranea* (Bonthala et al., 2016). Under salinity stress, rice *GATA8* overexpressed lines showed higher biomass accumulation and photosynthetic efficiency than the wild-type and the knockdown seedlings (Kamlesh et al., 2019). In soybean seedlings, the expression of *GATA44* and *GATA58* genes were extremely down-regulated under low nitrogen settlement (Zhang et al., 2015). In *Brassica juncea,* 29 *GATA* genes responded to high temperature and drought treatments based on RNA-seq experiments (Bhardwaj et al., 2015).

Rapessed is an important oil crop of the world. To date, the genomes of Darmor-*bzh* (winter ecotype), Tapitor (winter ecotype), Zhongshuang 11 were successfully sequenced and assembled (Chalhoub et al., 2014; Bayer et al., 2017; Sun et al., 2017). In the present study, 96 *GATA* family members were identified and characterized in the genome of *B. napus*. We methodically analyzed their phylogenetic relationships, GATA domains, gene structures, protein motifs, chromosome distributions, genome syntenys and gene duplications. Moreover, the profiles of *BnGATAs* expression pattern in various tissues and resonding to abiotic stresses were determined. In addition, SNPs of each *BnGATA* genes were systematically identified, in a worldwide collection of rapeseed germplasm (Wu et al., 2019; Xuan et al., 2020). These results enrich our knowledge about *BnGATA* genes, providing a basis for manupulation of the gene and facilitating breeding marker assisted breeding in rapeseed.

Materials And Methods

Identification of *GATAs* in *B. napus*

The amino acid sequences of GATA family members in *A. thaliana* were obtained according to a previous study (Table S1, Reyes et al., 2004), and the homologs of GATAs in *B. napus* were blasted against the reference genome of the rapeseed cultivar "Darmor-*bzh*" (v4.1 genome, http://www.genoscope.cns.fr/brassicanapus/data/). Hidden Markov Model (HMM) and BLASTP programs were applied for the identification of BnGATA proteins. The HMMER profile of GATA zinc finger domain (PF00320) from the Pfam database (http://pfam.janelia.org/) was used to perform the local BLASTP (E-value-20) search. The candidate sequences of GATAs were confirmed in the SMART database (http://smart.embl-heidelberg.de/) (Letunic et al., 2012), the NCBI Conserved Domain database (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-Bauer et al., 2011) and the Pfam database (Finn et al., 2016). Subfamily members were named based on their arrangement order on chromosomes of the *B. napus* genome (Table S2). Moreover, the length of amino acids, molecular weights (MW) and isoelectric point (pI) of GATA proteins were calculated using tools from ExPASy (http://www.expasy.ch/tools/pi_tool.html).

Phylogenetic analysis and classification of *GATAs*

The multiple alignments of GATA amino acids were done using the ClustalW with default parameters (Kumar et al., 2016). A phylogenetic tree was constructed using the MEGA 7.0 by the Neighbor-Joining (NJ) method (Kumar et al., 2016), with the following parameters: poisson model, pairwise deletion and 1000 bootstrap replications. Unrooted NJ tree of GATA proteins from *A. thaliana* and *B. napus* was also constructed using the MEGA 7.0. The *GATA* family members from *A. thaliana* were referred to classify the *GATA* family members in *B. napus*. In addition, the conserved GATA zinc finger domains in proteins were identified using the MEGA 7.0 and the GeneDoc software.

Motifs and gene structures

The Gene Structure Display Server online program (GSDS: http://gsds.cbi.pku.edu.ch) was used to analyzed exon-intron structures of all *GATA* genes (Hu et al., 2015). To identify conserved motifs in GATA proteins, the Multiple Expectation Maximization for Motif Elicitation (MEME) online program (http://meme.sdsc.edu/meme/itro.html) was performed with the following parameters: number of repetition = any, maximum number of motifs = 10; and optimum motif length = 6 to 100 residues (Bailey et al., 2009).

Chromosomal localization and gene duplication analyses

The distribution of 96 *GATA* genes identified in *B. napus* was mapped to 19 chromosomes according to their physical location information using the Circos software (Krzywinski et al., 2009). To identify gene duplication, the *GATA* genes were aligned using BLASTP with the e-value of 1e-10 and MCScanX to classify the duplication patterns including segmental and tandem duplication (Wang et al., 2012). The tandem duplication was definited that a chromosomal region within 300 kb contains two or more genes (Holub et al., 2001). Furthermore, the synteny relationships of *GATA* genes between the genomes of *B. napus* and *A. thaliana* were constructed using python.

Expression patterns of *BnGATA* genes in *B. napus*

In order to understand expression patterns of the *BnGATA* genes in *B. napus,* transcriptome data from 12 tissues of the *B. napus* cultivar “Zhongshuang 11” which was released in 2017 (Sun et al., 2017) were obtained from the NCBI (ID: PRJNA394926). Moreover, transcriptome data of *B. napus* under dehydration, salt, ABA and cold stress conditions were obtained by referring to Zhang et al. (2019). These transcriptome data were available under the project ID: CRA001775 (https://bigd.big.ac.cn/). Expression analysis of *GATAs* was performed using the DSEeq2 R package and the heatmaps of *GATAs* were constructed using the TBtools software (Chen et al., 2018).

SNP distribution of *GATAs* in a core collection of *B. napus*

To reveal natural variation of genomic sequences of *GATA* genes in *B. napus,* SNPs in the coding regions of *GATA* genes were determined in a collection of *B. napus* worldwide germplasm consisting of 300 core accessions in light of the genome resequencing data of our previous studies (Wu et al., 2019; Xuan et al., 2020). High-quality SNPs with MAF larger than 5% and missing rate less than 50% were used for the further analysis.

Results

Identification and phylogenetic analysis of GATA proteins in *B. napus*

A total of 96 proteins with GATA zinc finger domain were identified as the GATA family members in *B. napus* (Table S2). The longest sequence of each protein was remained, and the identification of these proteins was listed in Table S2 and Table S3. The length of 96 GATA proteins was ranged from 101 to 576 amino acids (aa), and the molecular weight was ranged from 11.17 to 64.59 kDa.

To analyze the relationships of GATA proteins between *B. napus* and *A. thaliana*, an unrooted tree was constructed using the full-length amino acids of these GATAs. Totally, 30 proteins from *A. thaliana* and 96 proteins from *B .napus* were identified (Figure 1). In *A. thaliana,* the GATAs were clustered into four subfamilies (Reyes et al., 2004). Here, 96 GATAs in *B. napus* were correspondingly classified into four subfamilies (Figure 1). Among these GATA proteins, 36 members belong to the subfamily I, 43 to the subfamily II, 10 to subfamily III and 7 to the subfamily IV. Each BnGATA protein features with only one GATA domain. Notably, the GATA domain locates mainly in the position 160-230 aa for the subfamily I; 30-150 or 200-260 aa for the subfamily II; 190-330 aa for the subfamily III, and 7-40 aa for the subfamily IV, respectively (Table S2)

Gene structures and protein motifs of *BnGATAs*

As shown in Figure 2B, one to nine exons were determined in *BnGATA* genes. Similar to *GATA* genes in *A. thaliana*, *BnGATA* genes in the subfamilies I and II have 2 to 3 exons except *BnGATA1.6* (4 exons), 3 to 9 exons for the subfamily III, and 6 to 8 exons for the subfamily IV (Figure 2B).

The motif analysis was conducted to display schematic structures of GATA proteins (Figure 2C). The details of 10 kinds of conserved motifs were listed in Supplementary Table S4. The motif 1 and motif 2 were detected in all GATA proteins, the motif 3, 4 and 9 were mainly identified in the members of subfamily I, the motif 6, 8 and 10 were identified in the members of subfamily II, while the motif 5 and 7 were identified in the members of subfamily III. Except for the motif 1 and 2, no other motifs were found in the subfamily IV (Figure 2C). In short, similar gene structures and conserved motifs within a subfamily strongly support the results of subfamily classifications by the phylogenetic analysis.

Moreover, with similar result of GATA domain analysis found in *A. thaliana* (Reyes et al., 2004), BnGATAs in the subfamilies I, II and IV contained 18 residues in the zinc finger loop (C-X_2-C-X_{18}-C-X_2-C), with the exception of BnGATA2.8 and BnGATA2.26, where N-X_2-C-X_{18}-C-X_2-C appears instead of C-X_2-C-X_{18}-C-X_2-C) (Figure 3). All 10 BnGATAs in the subfamily III contained 20 residues between the second and the third Cys residues in the zinc finger (C-X_2-C-X_{20}-C-X_2-C). In addition, several amino acid sites showed high conservation in the GATA domains such as LCNACG residues (Figure 3).

The distribution, genomic synteny and gene duplication of *BnGATA* genes

Totally, 84 out of 96 *BnGATA* genes were distributed over 19 chromosomes, while other 12 genes were assigned into random fragments (6 on the AAnn subgenome and 6 on the CCnn subgenome) (Figure 4 and Table S2). Among 84 *BnGATAs*, 46 genes located on the AA subgenome, including 16 subfamily I genes, 22 subfamily II genes, 5 subfamily III genes and 3 subfamily IV genes; while 50 genes located on the CC subgenome, including 20 subfamily I genes, 21 subfamily II genes, 5 subfamily III genes and 4 subfamily IV genes (Figure 4). Some *BnGATA* genes were formed as clusters in the same chromosomes, such as *BnGATA1.32* and *BnGATA2.36* (Figure 4). However, most *BnGATA* genes were randomly distributed on the AA or CC subgenome. In addition, Chr A1 showed the highest density of *BnGATA*s with 7 genes from the subfamilies II and III (Figure 4).

Using BLAST and MCScanX methods, 82 segmental duplication events of the *GATA*s were identified (Figure 4 and Table S5). Among these events, 80 duplication events occured across chromosomes, while 2 events were detected within a chromosome (*BnGATA1.28/BnGATA1.31*, *BnGATA1.19/BnGATA1.21*). Furthermore, 14 duplication events took place on the AA subgenome, 14 events on the CC subgenome, and 50 events across AA/CC subgenomes. The results suggest that some *BnGATA* genes possibly came into being during gene duplication, and the segmental duplication events could play key roles in the expansion of *BnGATA* genes in *B. napus*.

To better understand the evolution of *BnGATA* genes, the synteny of the *GATA* gene pairs between the genomes of *B. napus* and *A. thaliana* was constructed (Figure 5 and Table S6). Here, 55 *BnGATAs* exhibited syntenic relationship with *AtGATAs*. Some *AtGATAs* were associated with more than one orthologous copies in *B. napus.* For example, *AT2G45050* showed syntenic relationship with *BnGATA1.7, BnGATA1.8, BnGATA1.19* and *BnGATA1.21* (Table S6). Moreover, collinear gene pairs of *GATA* genes fixed on highly conserved syntenic blocks were also detected (Figure 5 and Table S6).

Expression profiles of *BnGATAs* in different tissues

The expression profiles of 96 *BnGATA* genes in 12 tissues of the rapeseed cultivar ZS11 were compared (Figure 6 and Table S7). According to the difference of their expression pattern, these genes were divided into three groups. In details, a total of 39 genes were classified into the group 1 with low expression levels or not detected in the tissues examined. There were 12 *BnGATAs* belonging to the group 2 with high expression levels in these tissues. Meanwhile, 43 *BnGATAs* were included in the group 3 with preferential expression profiles across tissues. For instance, *BnGATA1.11* was not expressed in wilting pistil, expressed with low levels in blossomy pistil and root, but expressed highly in other tissues (Figure 6 and Table S7).

On the other hand, the group 1 contained 9, 28, 1 and 1 genes from the four subfamilies; the group 2 had 6 and 6 genes from the subfamilies I and III, while the group 3 contained 21, 13, 3 and 6 genes from the four subfamilies, respectively (Table S7). Interestingly, it was found that *BnGATAs* from the subfamily II showed low expression levels in all tissues, but the subfamily III members had high expression levels in all tissues (Figure 6 and Table S7). The expression patterns of *GATA* genes in different tissues suggested functional divergences between different subfamilies.

Expression profiles of *BnGATAs* in response to abiotic stresses

Further, we studied the expression pattern of *BnGATA* genes under various abiotic stresses including drought, salinity, ABA induction and cold stresses (Figure 7 and Table S8). Here, 43 out of 96 *BnGATA* genes showed expression values higher than 20 in control group (Table S8). Among these 43 genes, 11 genes were significantly down-regulated in responding to dehydration treatment, with the exception of *BnGATA1.9* which was up-regulated in responding to dehydration treatment. Under salt stress, *BnGATA1.25* showed obviously decreased expression and *BnGATA1.9* showed increased expression, respectively. *BnGATA2.20* was almost not expressed after 4h salt treatment, but highly expressed after 24h salt treatment (Table S8). Moreover, *BnGATA2.5* and *BnGATA2.18* were down-regulated and 7 genes were significantly up-regulated in response to ABA induction. *BnGATA1.11*, *BnGATA1.24* and *BnGATA2.43* were repressed and 8 genes were up-regulated by cold treatment. Interestingly, *BnGATA1.27* was significantly induced by all abiotic stresses (Figure 7). Meanwhile, some *BnGATA* genes such as *BnGATA1.9* and *BnGATA1.29* could respond to diverse abiotic treatments (Figure 7, Table S8).

Sequence variation of *BnGATAs* in a core collection of *B. napus*

Based on our previous resequencing data of the rapeseed accessions (Wu et al., 2019; Xuan et al., 2020), SNPs from 300 core accessions with MAF more than 5% were used for the analysis. In average, 6 SNPs were detected for a *GATA* gene (Table S9). It was found that the SNP density of *BnGATA*s on the AA subgenome was higher than that on the CC subgenome (Table S9). Meanwhile, the SNP density of each subfamily was different, with averagely 6.7, 3.58, 14.2 and 7.14 SNPs for the four subfamilies, respectively.

The SNP density of each *BnGATA* gene within a subfamily was also different. For instance, no SNP was identified for *BnGATA1.27*, while 8 and 10 SNPs were identified for *BnGATA1.29* and *BnGATA2.5*. Moreover, a detailed SNP distribution of *BnGATA1.29* and *BnGATA2.5* were showed in Figure 8. For *BnGATA1.29,* it was found that there were 6 SNP loci in the promoter region, 2 SNPs in the exon/intron region and no SNP in the 3'UTR region (Figure 8A). For *BnGATA2.5,* there were no SNP in the promoter region, 10 SNPs in the exon/intron region and no SNP in the 3'UTR region (Figure 8B) We speculate that sequence variation of these *GATA*s may be related to their expression difference under abiotic stresses.

Discussion

In this study, we identified 96 genes of *GATA* family transcription factors in *B. napus*, designating as *BnGATA1.1* to *BnGATA4.7* based on their subfamily classification. Bioinformatics analyses such as phylogenetic relationships, domains, gene structures, protein motifs, chromosomal locations, homologous and orthologous genes of *GATA* were performed. The expression profiles were analyzed, and the SNPs of *BnGATA*s were identified. The results provide a valuable resource for functional identification of *BnGATA* TFs and molecular breeding in *B. napus*.

The *GATA* family genes were systematically investigated in *A. thaliana* and *O. sativa* (Reyes et al., 2004), *Solanum lycopersicum* (Yuan et al., 2018), *Vitis vinifera* (Zhang et al., 2018), *Phyllostachys edulis* (Wang et al., 2019) and *Gossypium genues* (Zhang et al., 2019). According to these studies, the *GATA* genes from dicotyledons, but not from monocots, could be strictly divided into four subfamilies. Here, we find that the subfamilies I, II and III of the *GATA* genes simultaneously occur in both dicotyledons and monocots, but the subfamily IV genes did not exist in monocots (Reyes et al., 2004; Wang et al., 2019). It demonstrated that the subfamily IV of *GATA* genes appeared after the divergence between dicotyledon and monocot. In *B. napus*, 96 *BnGATA*s were also classified into four subfamilies, including 36, 43, 10 and 7 genes in the subfamilies I, II, III and IV, respectively. Similar GATA protein structures and high homology of these proteins between *B. napus* and *A. thaliana* were found. Among the four subfamilies, many differences were detected in terms of protein structures, motif compositions, SNP densities and expression patterns. For example, in the subfamily III, the GATA domain featured with 20 residues in the zinc finger (C-X_2-C-X_{20}-C-X_2-C) instead of 18 residues as in other three subfamilies. The CCT and TIFY domains which involve in flowering, hypocotyl and root development in *A. thaliana* were only found in the subfamily III (Nishii et al., 2000; Shikata et al., 2004; Vanholme et al., 2007). Almost all genes of the subfamily III were highly expressed in various tissues, and under dehydration condition (Table S7 and Table S8). The SNP density of subfamily III genes (14.2) was averagely higher than those in the other subfamily genes (5.19) (Table S9).

GATA TFs are related to plant growth and development. The subfamily I genes are involved in seed germination and respond to abiotic stresses. For example, *BME3* (*GATA8*) was reported as a positive regulator of seed germination in *A. thaliana* (Liu et al., 2005). The plants, in which *BME3* were knocked out, showed deeper dormancy and more sensitive to cold stratification. Moreover, the decreased expression of *GA20-oxidase* and *GA3-oxidase* in the knockout plants suggested that *BME3* was involved

in GA biosynthesis (Liu et al., 2005). A recent study proposed that *RGL2-DOF6* complex regulates *GATA12* expression to enforce primary dormancy in *A. thaliana* (Ravindran et al., 2017). In this study, *BnGATA1.29* (*BnaC08g25560D*), the ortholog of *BME3* in *A. thaliana*, showed high expression levels in various tissues and significantly responded to ABA and cold stresses, suggesting its functional conservation between *B. napus* and *A. thaliana* (Table S7 and Table S8).

In the subfamily II, *GNC* and *GNL* (*GATA21*) involved in germination, greening, flowering, floral development, senescence and floral organ abscission (Bi et al., 2005; Hudson et al., 2011; Chiang et al., 2012; Richter et al., 2010, 2013a, 2013b; Behringer et al., 2014). *BnGATA2.5* (*BnaA02g08490D*), the ortholog of *GNL* in *A. thaliana*, expressed across all tissues and organs in *B. napus* (Figure 6, Table S7). Meanwhile, the expression of *BnGATA2.5* was down-regulated under ABA inducement, drought and cold treatments, indicating its strong response to abiotic stresses (Figure 7, Table S8). Recently, the association between *BnGATA2.5* expression and plant height, branch initiation height and flowering time in *B. napus* was proved (Shen et al., 2018).

The subfamily III of *GATA* TFs is a novel plant-specific subfamily, which play important roles in flowering, hypocotyl and root development (Nishii et al., 2000; Shikata et al., 2004; Vanholme et al., 2007). For instance, overexpression of *ZIM* (*GATA25*) could upregulate the expression of *XTH33* (*xyloglucosyl transferase 33*), resulting in elongate hypocotyls and prtioles in *A. thaliana* (Shikata et al., 2004; Vanholme et al., 2007). Besides, *ZML1* (*GATA24*) and *ZML2* (*GATA28*) were identified as the two essential components of the *cry1* (Cryptochrome1)-mediated photoprotective response in *A. thaliana* (Shaikhali et al., 2012). In this study, *BnGATA3.1* (*BnaA01g25320D*) as the ortholog of *AtZML1*, were highly expressed in most tissues in *B. napus* (Figure 6, Table S7). The expression of *BnGATA3.1* was slightly changed in response to a variety of abiotic stresses (Figure 7, Table S8). However, so far, little was known about the subfamily IV of the *GATA* TFs in plants. Rapeseed originated from the natural crossing between *B. rapa* (AA) and *B. oleracea* (CC) (Chalhoub et al., 2014). In this study, we identified 46 and 50 *BnGATA* genes located on the AA or CC subgenomes, indicating that *GATA* TFs play similar important roles in both ancestral species. However, the SNP density of *BnGATAs* (7.33) on the AA subgenome was much higher than that (5.02) on the CC subgenome, indicating a higher genetic diversity of *BnGATA* genes on the AA subgenome (Table S9), which could arise from more frequent out crossing between *B. napus* and *B. rapa* than between *B. napus* and *B. oleracea* (Wu et al., 2019).

SNPs in the coding regions are crucial for the generation of new alleles, and allele divergence may lead to gene function alterations, which is vital facilitation for crop species adaptation to environmental stresses (Kumar et al., 2010). For example, 7 functional alleles of powdery mildew resistance gene *Pm3* were isolated from a set of 1,320 bread wheat landraces through allele mining, while other 9 alleles of *Pm3* showed non-function to powdery mildew resistance (Bhullar et al., 2010). In this study, the SNP determination of *BnGATA* genes showed plentiful genetic variation of *BnGATAs* in a core collection of *B. napus*. Haplotypes and allele-specific markers of these genes could be identified for rapeseed molecular-breeding programs.

Taken together, we performed a comprehensive characterization of *GATA* family TFs in *B. napus*. The results enrich our knowledge about *BnGATA* genes, providing a basis for manupulation of the genes and facilitating breeding marker-assisted breeding in rapeseed.

Conclusion

GATA transcription factors play key roles in plant growth, floral development and abiotic stresses response. In present research, genome-wide identification and characterization of *GATA* genes were conducted in *B. napus.* A total of 96 identified *GATA* factors were divided into four subfamilies, which showed high similarity in protein structure and domain composition within the same subfamily. Phylogenetic comparison and synteny analysis of *GATA* genes between *A. thaliana* and *B. napus* provide valuable clues for the evolutionary characteristics of the *BnGATA* genes. Moreover, gene expression and SNP distrubution analysis of *BnGATA* family were also determined. These results provide useful insights into the functional differences, evolutionary relationships and expression profiles of *B.napus GATA* transcription factors.

Abbreviations

ARF: Auxin Response Factors; *AtGATA*: Arabidopsis thaliana *GATA*; BLASTP: Basic local alignment search tool-protein; *BnGATA*: *Brassica napus GATA*; *bZIP*: basic helix loop helix; *cry1*: Cryptochrome1; DREB: Dehydration-responsive element-binding protein; GA: Gibberellin; *GNC*: GATA, Nitrate-inducible, Carbon-metabolism involved; *GNL*: GNC-like; GSDS: Gene structure display server; HMM: Hidden markov mode; MEME: Motif Elicitation; MW: molecular weights; *MYB*: V-myb avian myeloblastosis viral oncogene homolog; pI: isoelectric point; *PIFs*: Phytochrome Interacting Factors; SNP: Single Nucleotide Polymorphisms; *SOC1*: Suppressor of Constans 1; TFs: Transcription factors; *XTH33*: xyloglucosyl transferase 33.

Declaration

Acknowledgements

This work was funded by the National Natural Science Foundation of China (31961143008, 31701411), the Science and Technology Program of Zhejiang Province of China (LGN20C130007), Jiangsu Collaborative Innovation Center for Modern Crop Production, and the 111 project for introduction of foreign experts (B17039).

Availability of data and materials

All data analyzed during this study are included in this article and its Additional files.

Authors’ contributions

WZ Zhu and DZ Wu conceived and designed the research. WZ Zhu and YY Guo performed the experiments and data analyses. WZ Zhu, YY Guo, YK Chen, LX Jiang and DZ Wu wrote the article; all authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

[1] Bailey TL, Elkan C. **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** Proceedings International Conference on Intelligent Systems Molecular Biology. 2 (1994) 28-36.

[2] Bayer PE, Hurgobin B, Golicz AA, Chan CK, Yuan Y, Lee H, Renton M, Meng J, Li R, Long Y, Zou J, Bancroft I, Chalhoub B, King GJ, Batley J, Edwards D. **Assembly and comparison of two closely related Brassica napus genomes.** Plant Biotechnology Journal. 15 (2017) 1602-1610.

[3] Behringer C, Bastakis E, Ranftl QL, Mayer KFX, Schwechheimer C. **Functional Diversification within the Family of B-*GATA* Transcription Factors through the Leucine-Leucine-Methionine Domain.** Plant Physiology. 166 (2014) 293-305.

[4] Bhardwaj AR, Joshi G, Kukreja B, Malik V, Arora P, Pandey R, Shukla RN, Bankar KG, Katiyar-Agarwal S, Goel S, Jagannath A, Kumar A, Agarwal M. **Global insights into high temperature and drought stress regulated genes by RNA-Seq in economically important oilseed crop *Brassica juncea*.** BMC Plant Biology. 15 (2015) 9.

[5] Bhullar NK, Street K, Mackay M, Yahiaoui N, Keller B. **Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the *Pm3* resistance locus.** Proceedings of the National Academy of Sciences of the United States of America. 106 (2009) 9519-9524.

[6] Bi YM, Zhang Y, Signorelli T, Zhao R, Zhu T, Rothstein S. **Genetic analysis of *Arabidopsis GATA* transcription factor gene family reveals anitrate-inducible member important for chlorophyll synthesis and glucose sensitivity.** Plant Journal. 44 (2005) 680-692.

[7] Bonthala VS, Mayes K, Moreton J, Blythe M, Wright V, May ST, Massawe F, Mayes S, Twycross J. **Identification of gene modules associated with low temperatures response in bambara groundnut by network-based analysis.** PLoS ONE. 11 (2016) e0148771.

[8] Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. **Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome.** Science. 345 (2014) 950-953.

[9] Chiang YH, Zubo YO, Tapken W, Kim HJ, Lavanway AM, Howard L, Pilon M, Kieber JJ, Schaller GE. **Functional characterization of the *GATA* transcription factors *GNC* and *CGA1* reveals their key role in chloroplast development, growth, and division in *Arabidopsis*.** Plant Physiology. 160 (2012) 332-348

[10] Chen CJ, Xia R, Chen H, He YH. **TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface.** BioRXiv. (2018).

[11] Duan ZQ, Zhang YT, Tu JX, Shen JX, Yi B, Fu TD, Dai C, Ma CZ. **The Brassica napus GATA transcription factor BnA5.ZML1 is a stigma compatibility factor.** Journal of Integrative Plant Biology. 00 (2020) 1-20.

[12] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangradorvegas A, et al. **The Pfam protein families database: towards a more sustainable future.** Nucleic Acids Research. 44 (2016) 279-285.

[13] Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. **DNA-binding specifi cities of plant transcription factors and their potential to define target genes.** Proceedings of the National Academy of Sciences of the United States of America. 111 (2014) 2367-2372.

[14] Holub EB. **The arms race is ancient history in *Arabidopsis*, the wildflower.** Natural Reviews Genetics. 2 (2001) 516-527.

[15] Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. **GSDS 2.0: An upgraded gene feature visualization server.** Bioinformatics. 31 (2015) 1296-1297.

[16] Hudson D, Guevara D, Yaish MW, Hannam C, Long N, Clarke JD, Bi YM, Rothstein SJ. **GNC and CGA1 modulate chlorophyll biosynthesis and glutamate synthase (GLU1/Fd-GOGAT) expression in *Arabidopsis*.** PLoS ONE. 6 (2011) e26765.

[17] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. **Circos: an information aesthetic for comparative genomics.** Genome Research. 19 (2009) 1639-1645.

[18] Kumar GR, Sakthivel K, Sundaram RM, Neeraja CN. Balachandran SM, Rani NS, Viraktamath BC, Madhav MS. **Allele mining in crops: Prospects and potentials.** Biotechnology Advances. 28 (2010) 451-461.

[19] Kumar S, Stecher G, Tamura K. **MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets.** Molecular Biology and Evolution. 33 (2016) 1870-1874.

[20] Lee J, Lee I. **Regulation and function of *SOC1*, a flowering pathway integrator.** Journal of Experiment Botany. 61 (2010) 2247-2254.

[21] Letunic I, Doerks T, Bork P. **SMART 7: Recent updates to the protein domain annotation resource.** Nucleic Acids Reserch. 40 (2012) 302-305.

[22] Liu H, Li T, Wang YM, Zheng J, Li HF, Hao CY, Zhang XY. ***TaZIM-A1* negatively regulates flowering time in common wheat (Triticum aestivum L.).** Journal of Integrative Plant Biology. 61 (2019) 359-376.

[23] Liu PP, Koizuka N, Martin RC, Nonogaki H. **The *BME3 (Blue Micropylar End 3)* *GATA* zinc finger transcription factor is a positive regulator of *Arabidopsis* seed germination.** Plant Journal. 44 (2005) 960-971.

[24] Lowry JA, Atchley WR. **Molecular evolution of the *GATA* family of transcription factors: conservation within the DNA-binding Domain.** Journal of Molecular Evolution. 50 (1999) 103-115.

[25] Manfield IW, Devlin PF, Jen CH, Westhead DR, Gilmartin PM. **Conservation, convergence, and divergence of light-responsive, circadian-regulated, and tissue-specific expression patterns during evolution of the *Arabidopsis GATA* gene family.** Plant Physiology. 143 (2007) 941-958.

[26] Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL,Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. **CDD: a Conserved Domain Database for the functional annotation of proteins.** Nucleic Acids Research. 39 (2011) 225-229.

[27] Naito T, Kiba T, Koizumi N, Yamashino T, Mizuno T. **Characterization of a unique *GATA* family gene that responds to both light and cytokinin in *Arabidopsis thaliana*.** Bioscience Biotechnology and Biochemistry. 71 (2007) 1557-1560.

[28]Nishii A, Takemura M, Fujita H, Shikata M, Yokota A, Kohchi T. **Characterization of a novel gene encoding a putative single zinc-finger protein, ZIM, expressed during the reproductive phase in *Arabidopsis thaliana*.** Bioscience Biotechnology and biochemistry. 64 (2000) 1402-1409.

[29] Nutan KK, Singla-Pareek SL, Pareek A. **The *Saltol* QTL-localized transcription factor *OsGATA8* plays an important role in stress tolerance and seed development in *Arabidopsis* and rice.** Journal of Experimental Botany. 71 (2020) 684-698.

[30] Ravindran P, Verma V, Stamm P, Kumar PP. **A Novel RGL2-DOF6 Complex Contributes to Primary Seed Dormancy in *Arabidopsis thaliana* by Regulating a *GATA* Transcription Factor.** Molecular Plant. 10 (2017) 1307-1320.

[31] Reyes JC, Muro-Pastor MI, Florencio FJ. **The *GATA* family of transcription factors in *Arabidopsis* and rice.** Plant Physiology. 134 (2004) 1718-1732.

[32] Richter R, Behringer C, Müller IK, Schwechheimer C. **The GATA-type transcription factors *GNC* and *GNL/CGA1* repress gibberellin signaling downstream from DELLA proteins and PHYTOCHROME-INTERACTING FACTORS.** Genes & Development. 24 (2010) 2093-2104.

[33] Richter R, Behringer C, Zourelidou M, Schwechheimer C. **Convergence of auxin and gibberellin signaling on the regulation of the *GATA* transcription factors *GNC* and *GNL* in *Arabidopsis thaliana*.** Proceedings of the National Academy of Sciences of the United States of America. 110 (2013a) 13192-13197.

[34] Richter R, Bastakis E, Schwechheimer C. **Cross-repressive interactions between *SOC1* and the *GATAs GNC* and *GNL/CGA1* in the control of greening, cold tolerance, and flowering time in *Arabidopsis*.** Plant Physiology. 162 (2013b) 1992-2004.

[35] Saitou N, Nei M. **The neighbor-joining method: A new method for reconstructing phylogenetic trees.** Molecular Biology and Evolution. 4 (1987) 406-425.

[36] Scazzocchio C. **The fungal *GATA* factors in Current Opinion.** Microbiology. 3 (2000) 126-131.

[37] Shaikhali J, Barajas-Lopez JD, Otvos K, Kremnev D, Garcia AS, Srivastava V, Wingsle G, Bako L, Strand A. **The CRYPTOCHROME1-Dependent Response to Excess Light Is Mediated through the Transcriptional Activators ZINC FINGER PROTEIN EXPRESSED IN INFLORESCENCE MERISTEM LIKE1 and *ZML2* in *Arabidopsis*.** Plant Cell. 24 (2012) 3009-3025.

[38] Shen YS, Xiang Y, Xu ES, Ge XH, Li ZY. **Major Co-localized QTL for Plant Height, Branch Initiation Height, Stem Diameter, and Flowering Time in an Alien Introgression Derived *Brassica napus* DH Population.** Frontiers in Plant Science. 9 (2018) 390.

[39] Shikata M, Matsuda Y, Ando K, Nishii A, Takemura M, Yokota A, Kohchi T. **Characterization of *Arabidopsis ZIM*, a member of a novel plant-specific *GATA* factor gene family.** Journal of Experimental Botany. 55 (2004) 631-639.

[40] Sun FM, Fan GY, Hu Q, Zhou YM, Guan M, Tong CB, Li JN, et al. **The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype.** The Plant Journal. 92 (2017) 452-468.

[41] Vanholme B, Grunewald W, Bateman A, Kohchi T, Gheysen G. **The tify family previously known as *ZIM*.** Trends Plant Science. 12 (2007) 239-244.

[42] Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H. **MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.** Nucleic Acids Research. 40 (2012) e49.

[43] Wang TT, Yang Y, Lou ST, Wei W, Zhao ZX, Ren YJ, Lin CT, Ma LY. **Genome-Wide Characterization and Gene Expression Analyses of *GATA* Transcription Factors in Moso Bamboo (*Phyllostachys edulis*).** International Journal of Molecular Sciences. 21 (2019) 14.

[44] Wu DZ, Liang Z, Yan T, Xu Y, Xuan LJ, Tang J, Zhou G, Lohwasser U, Hua SJ, Wang HY, Chen XY, Wang Q, Zhu L, Maodzeka A, Hussain N, Li ZL, Li XM , Shamsi IH, Jilani G, Wu LD, Zheng HK, Zhang GP, Chalhoub B, Shen LS, Yu H, Jiang LX. **Whole-Genome Resequencing of a Worldwide Collection of Rapeseed Accessions Reveals the Genetic Basis of Ecotype Divergence.** Molecular Plant. 12 (2019) 30-43.

[45] Xuan LJ, Yan T, Lu LZ, Zhao XZ, Wu DZ, Hua SJ, Jiang LX. **Genome-wide association study reveals new genes involved in leaf trichome formation in polyploid oilseed rape (*Brassica napus* L.).** Plant Cell and Environment. 43 (2019) 675–691.

[46] Yuan Q, Zhang C, Zhao T, Yao M, Xu X. **A genome-wide analysis of *GATA* transcription factor family in tomato and analysis of expression patterns.** Internationl Journal of Agriculture and biology. 20 (2017) 1274-1282.

[47] Yi A, Zhou YY, Han X, Shen C, Wang S, Liu C, Yin WL, Xia XL. **The GATA transcription factor GNC plays an important role in photosynthesis and growth in poplar.** Journal of Experimental Botany. 71 (2020) 1969–1984.

[48] Zhang YT, Ali U, Zhang GF, Yu LQ, Fang S, Lqbal S, Li HH, Lu SP, Guo L **Transcriptome analysis reveals genes commonly responding to multiple abiotic stresses in rapeseed.** Molecular Breeding. 39 (2019) 158.

[49] Zhang C, Hou Y, Hao Q, Chen H, Chen L, Yuan S, Shan Z, Zhang X, Yang Z, Qiu D, et al. **Genome-wide survey of the soybean *GATA* transcription factor gene family and expression analysis under low nitrogen stress.** PLoS ONE. 10 (2015) e0125174.

[50] Zhang Z, Ren C, Zou LM, Wang Y, Li SH, Liang ZC. **Characterization of the *GATA* gene family in *Vitis vinifera*: genome-wide analysis, expression profiles, and involvement in light and phytohormone response.** Genome. 61 (2018) 713-723.

[51] Zhang Z, Zou XY, Huang Z, Fan SM, Qun G, Liu A, Gong JW, Li JW, Gong WK, Shi YZ, Fan LQ, Zhang ZB, Liu RX, Jiang X, Lei K, Shang HH, Xu AX, Yuan YL. **Genome-wide identification and analysis of the evolution and expression patterns of the *GATA* transcription factors in three species of *Gossypium Genus*.** Gene. 680 (2018) 72-83.

[52] Zubo YO, Blakley IC, Franco-Zorrilla JM, Yamburenko MV, Solano R, Kieber JJ, Loraine AE, Schaller GE. **Coordination of Chloroplast Development through the Action of the *GNC* and *GLK* Transcription Factor Families.** Plant Physiology. 178 (2018) 130-147.

Figures

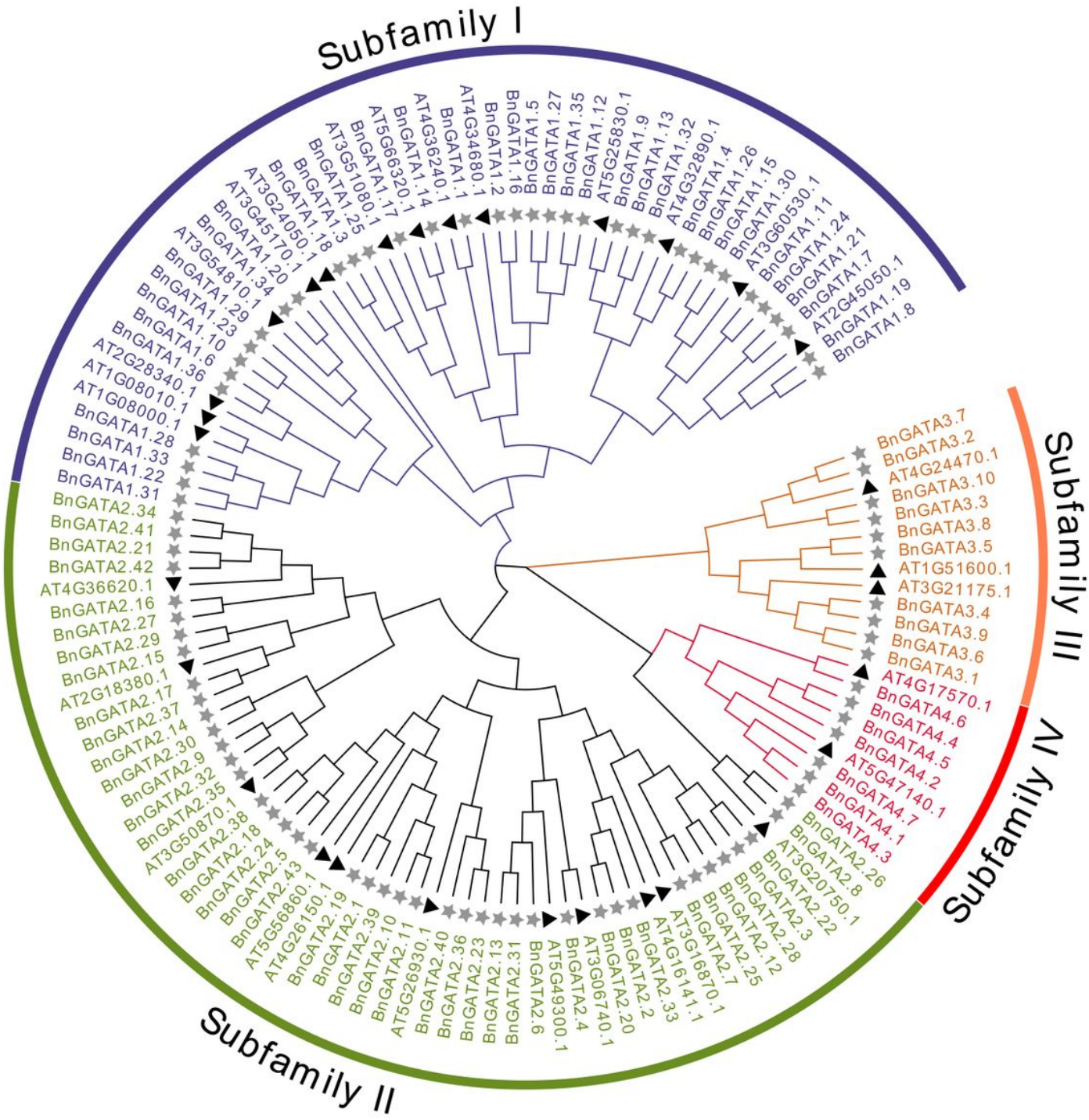

Figure 1

Phylogenetic analysis of GATAs in B. napus and A. thaliana. The different-colored arcs indicate subfamilies of the GATA proteins. The unrooted Neighbour-Joining phylogenetic tree was constructed using MEGA7 with full-length amino acid sequences of 126 GATA proteins, and the bootstrap test

replicate was set as 1000 times. The asterisks and triangles represent the GATA proteins from B. napus and A. thaliana, respectively.

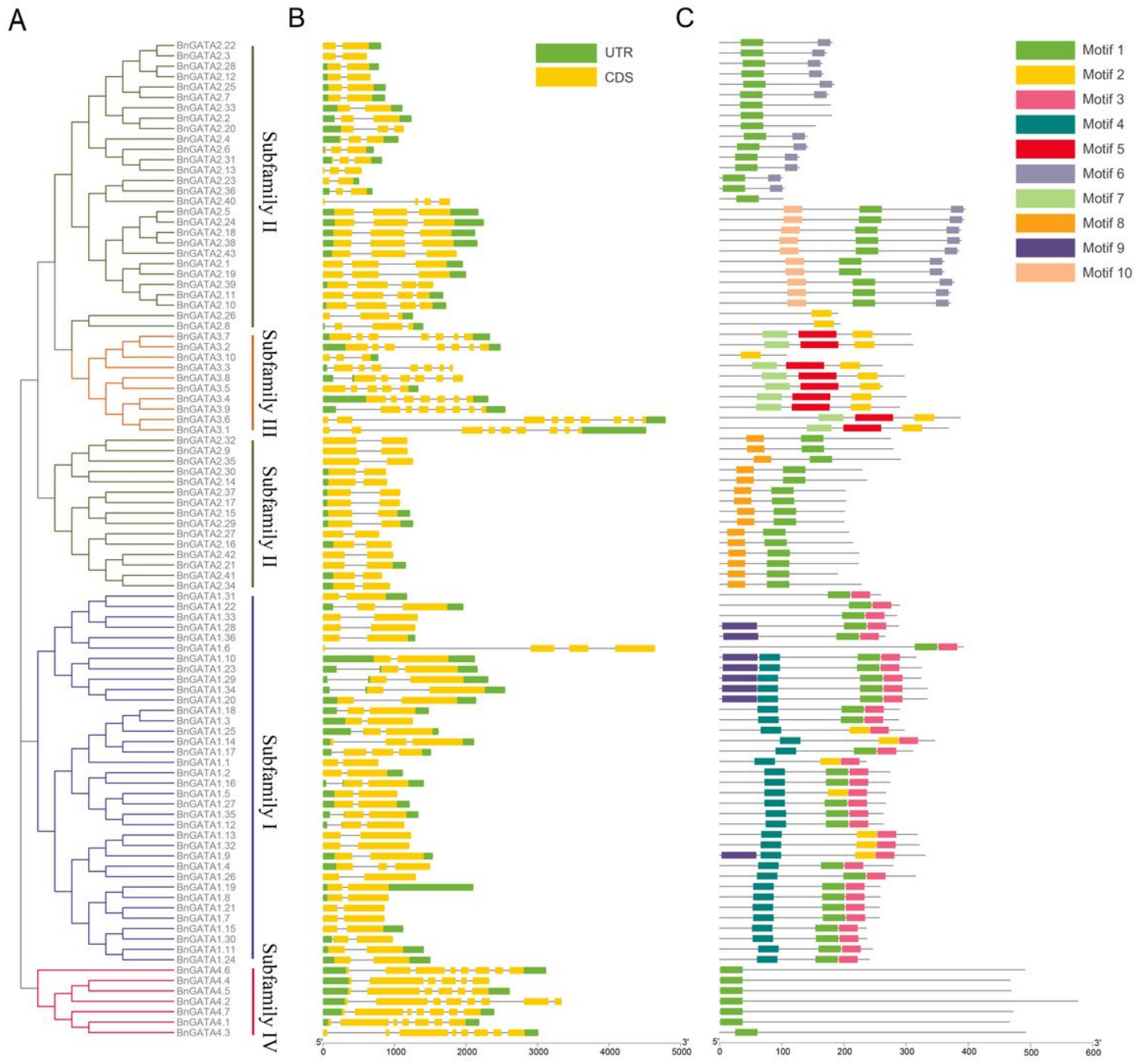

Figure 2

Schematic representation of phylogenetic relationships, gene structures and conserved motifs of the GATA genes from B. napus. A Phylogenetic tree of 96 BnGATA proteins. The unrooted neighbor-joining phylogenetic tree was constructed with MEGA7 using full-length amino acid sequences of 96 BnGATA proteins, and the bootstrap test replicate was set as 1000 times. B Exon/intron structures of BnGATA genes. Yellow boxes represent exons and black lines represent introns. The UTR region of BnGATA genes are indicated in green boxes. The length of exons can be inferred by the scale at the bottom. C The motif

composition of BnGATA proteins. The motifs, numbers 1-10, are displayed in different colored boxes. The sequence information for each motif is provided in Table S4. The length of protein can be estimated using the scale at the bottom.

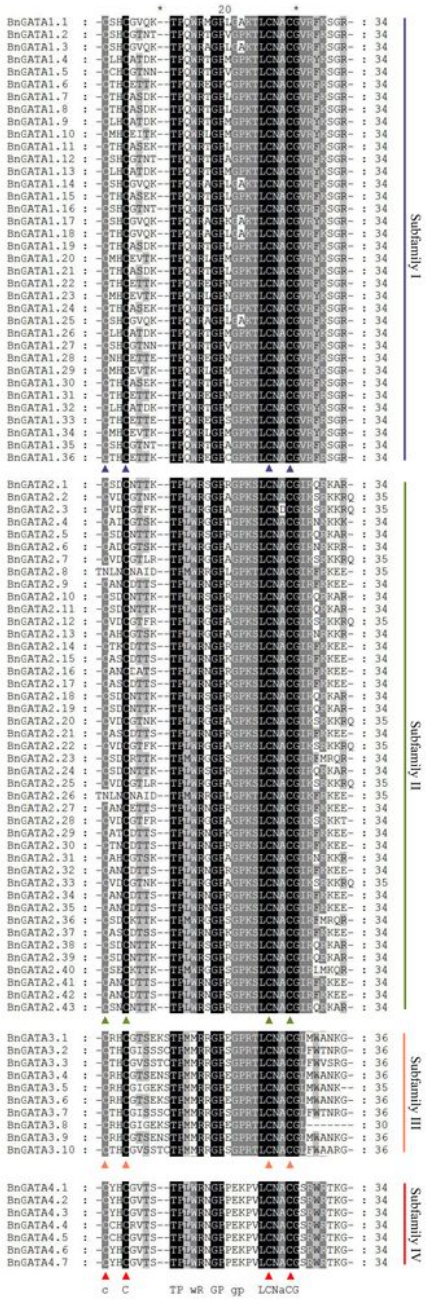

Figure 3

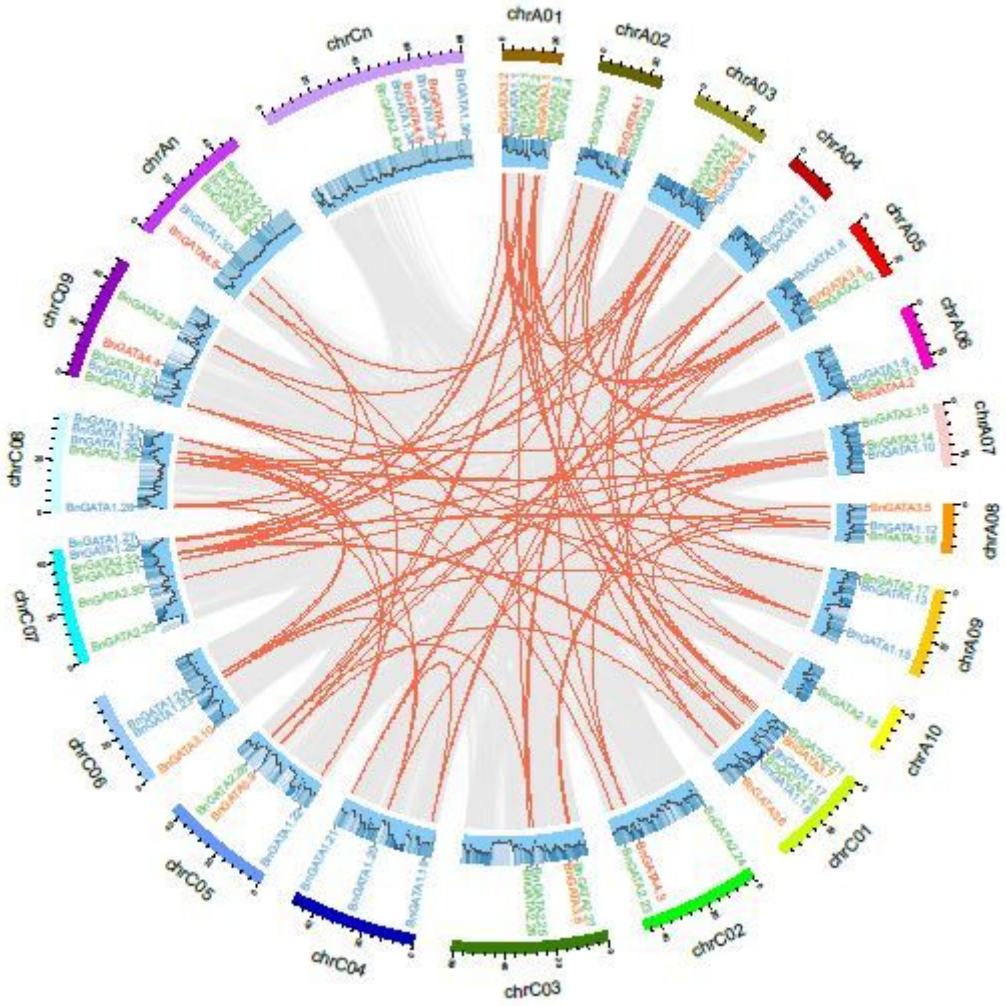

Figure 4

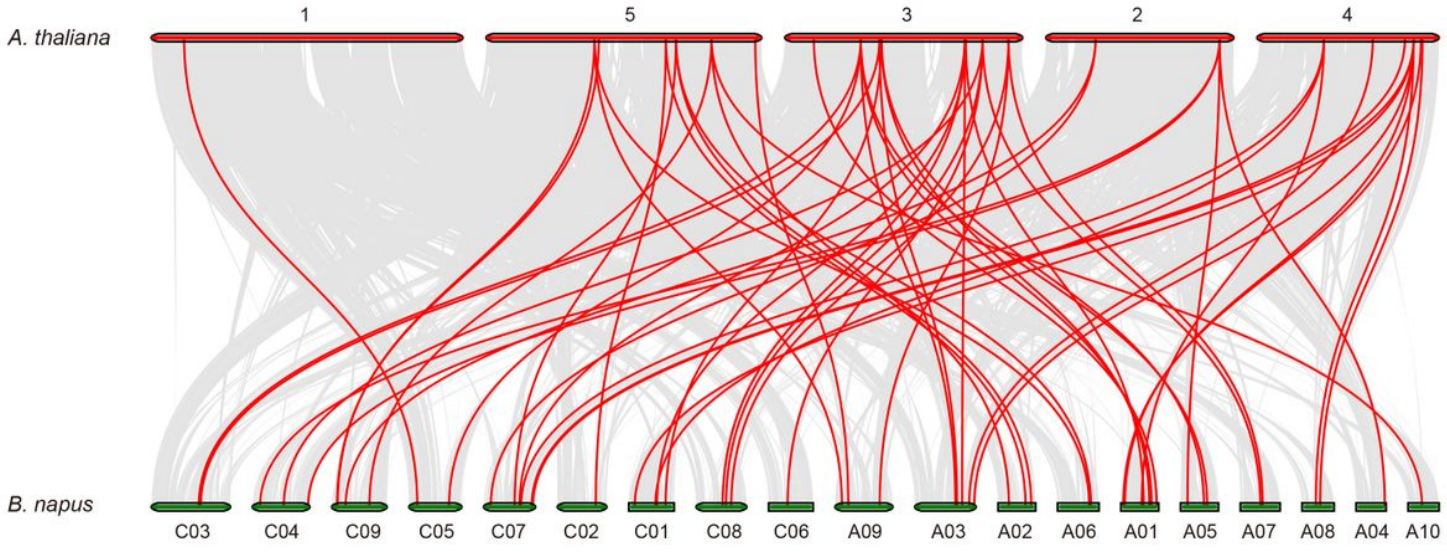

Figure 5

Synteny analysis of GATA genes between B .napus and A. thaliana. Gray lines indicate all collinear blocks within B. napus and A. thaliana, while the red lines depict the orthologous relationships of GATA genes between B. napus and A. thaliana.

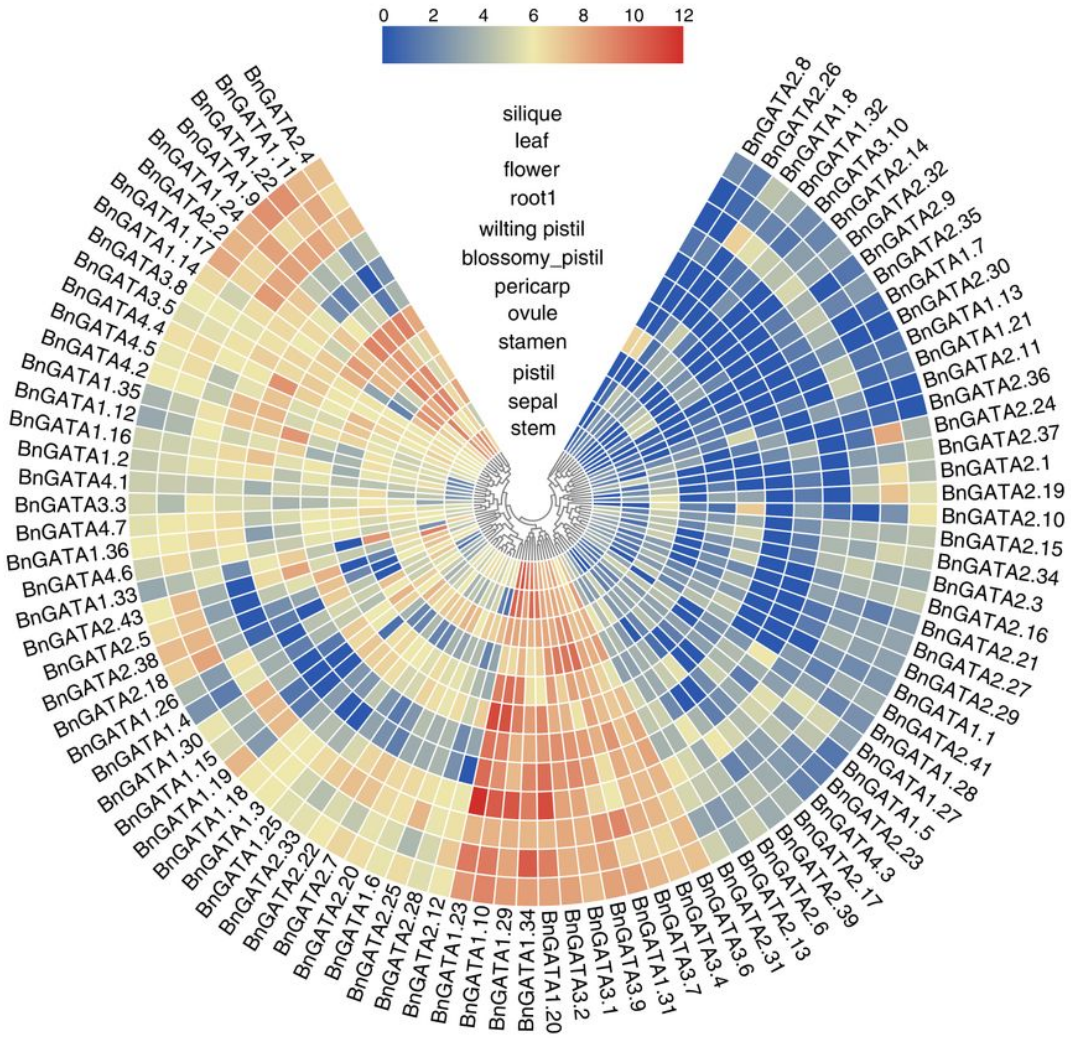

Figure 6

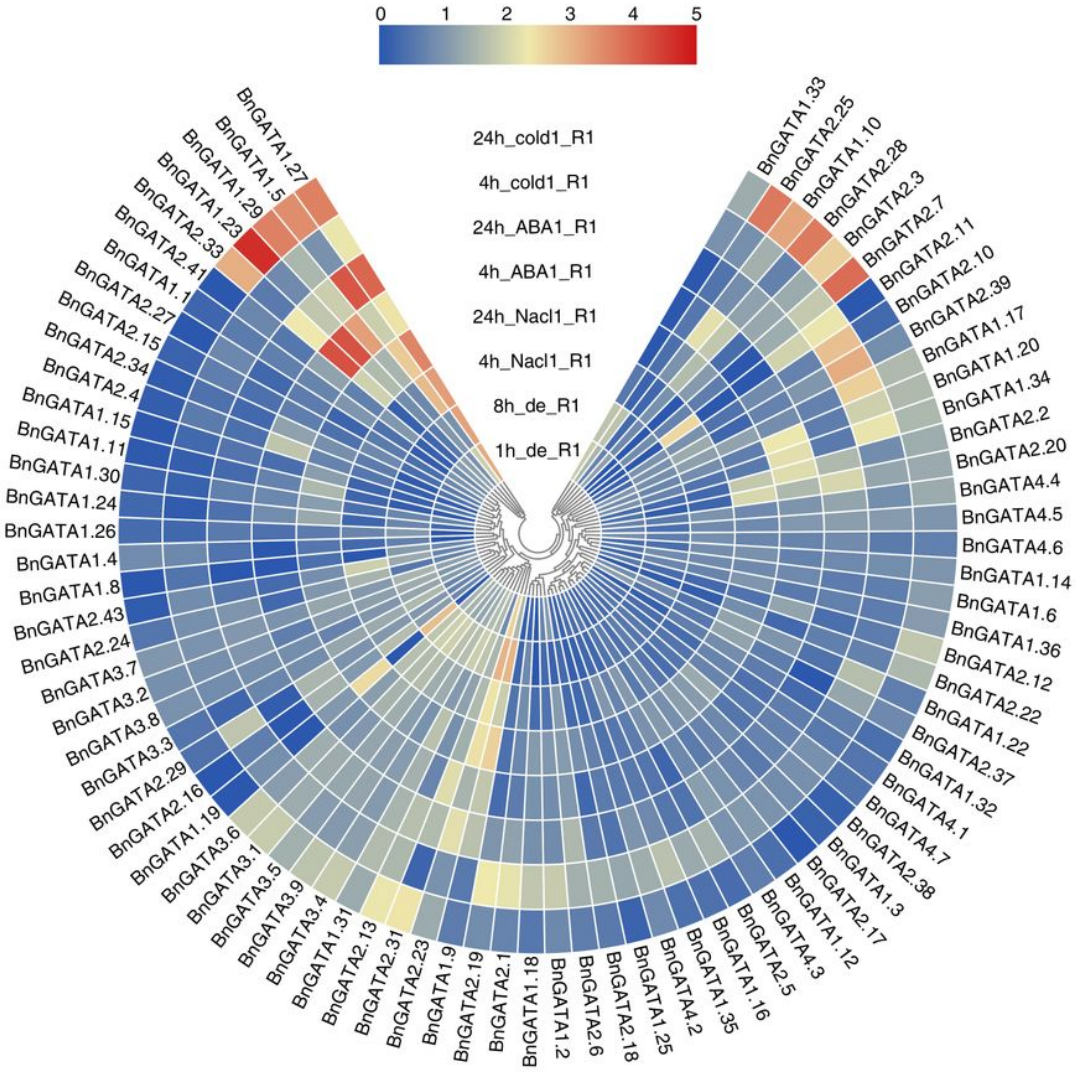
0
1
2
3
4
5
24h_cold1_R1
4h_cold1_R1
24h_ABA1_R1
4h_ABA1_R1
24h_Nacl1_R1
4h_Nacl1_R1
8h_de_R1
1h_de_R1
BnGATA1.27
BnGATA1.5
BnGATA1.29
BnGATA1.23
BnGATA2.33
BnGATA2.41
BnGATA1.1
BnGATA2.27
BnGATA2.15
BnGATA2.34
BnGATA2.4
BnGATA1.15
BnGATA1.11
BnGATA1.30
BnGATA1.24
BnGATA1.26
BnGATA1.4
BnGATA1.8
BnGATA2.43
BnGATA2.24
BnGATA3.7
BnGATA3.2
BnGATA3.8
BnGATA3.3
BnGATA2.29
BnGATA2.16
BnGATA1.19
BnGATA3.6
BnGATA3.1
BnGATA3.5
BnGATA3.9
BnGATA3.4
BnGATA1.31
BnGATA2.13
BnGATA2.31
BnGATA2.23
BnGATA1.9
BnGATA2.19
BnGATA2.1
BnGATA1.18
BnGATA1.2
BnGATA2.6
BnGATA2.18
BnGATA1.25
BnGATA4.2
BnGATA1.35
BnGATA1.16
BnGATA2.5
BnGATA4.3
BnGATA1.12
BnGATA2.17
BnGATA1.3
BnGATA4.7
BnGATA4.1
BnGATA2.38
BnGATA1.32
BnGATA2.37
BnGATA1.22
BnGATA2.22
BnGATA2.12
BnGATA1.36
BnGATA1.6
BnGATA1.14
BnGATA4.6
BnGATA4.5
BnGATA4.4
BnGATA2.20
BnGATA2.2
BnGATA1.34
BnGATA1.20
BnGATA1.17
BnGATA2.39
BnGATA2.10
BnGATA2.11
BnGATA2.7
BnGATA2.3
BnGATA2.28
BnGATA1.10
BnGATA2.25
BnGATA1.33

Figure 7

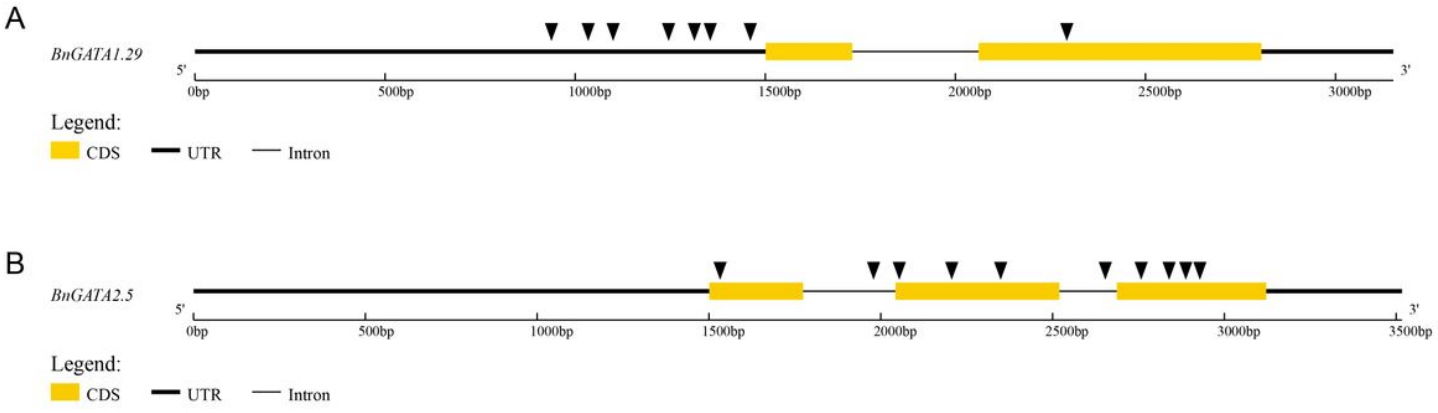
A
BnGATA1.29
5'
3'
0bp
500bp
1000bp
1500bp
2000bp
2500bp
3000bp
Legend:
CDS
UTR
Intron
B
BnGATA2.5
5'
3'
0bp
500bp
1000bp
1500bp
2000bp
2500bp
3000bp
3500bp
Legend:
CDS
UTR
Intron

Figure 8