Journal of Tropical Biodiversity and Biotechnology
Volume 09, Issue 03 (2024): jtbb88015
DOI: 10.22146/jtbb.88015

Research Article

# Mining GATA Transcription Factor Encoding Genes in The Cocoa Tree (*Theobroma cacao* L.) Suggests Their Potential Roles in Embryo Development and Biotic Stress Response

**Ngoc Thi Bich Chu[1], Thi Man Le[1], Ha Duc Chu[2], Huyen Thi Thanh Tran[3], Lan Thi Mai Tran[1], Hong Viet La[4], Quyen Thi Xuan Vu[1], Huynh Huy Phung[1,5], Phi Bang Cao[1]***

1)Faculty of Natural Sciences, Hung Vuong University, Phu Tho Province 35000, Vietnam
2)Faculty of Agricultural Technology, University of Engineering and Technology, Vietnam National University Hanoi, Xuan Thuy Road, Cau Giay District, Hanoi City 122300, Vietnam
3)Faculty of Biology, Hanoi National University of Education, Xuan Thuy Road, Cau Giay District, Hanoi City 122300, Vietnam
4)Institute of Research and Application, Hanoi Pedagogical University 2, Phuc Yen City, Vinh Phuc Province 280000, Vietnam
5)Thanh Thuy Junior High School, Phu Tho Province 35000, Vietnam
* Corresponding author, email: phibang.cao@hvu.edu.vn

**Keywords:**
Cocoa plant
Characterization
Expression profiles
GATA
Transcription factor
Identification
**Submitted:**
14 August 2023
**Accepted:**
12 March 2024
**Published:**
12 July 2024
**Editor:**
Furzani Binti Pa'ee

## ABSTRACT

GATA transcription factors (TFs) are widely recognized as significant regulators, characterized by a DNA-binding domain that consists of a type IV zinc finger motif. This TF family has been widely investigated in numerous higher plant species. The purpose of the present work was to comprehensively analyze the GATA TF in cocoa plant (*Theobroma cacao* L.) by using various bioinformatics tools. As a result, a total of 24 members of the GATA TFs have been identified and annotated in the assembly of the cocoa plant. According to phylogenetic analysis, these TcGATA proteins were classified into four distinct groups, including groups I (10 members), II (seven members), III (five members), and IV (two members). Next, our investigation indicated that the TcGATA proteins in different groups exhibited a high variation in their physicchemical features due to their different protein lengths, gene structures, and conserved motif distributions, whereas the TcGATA proteins in the same clade might share the common conserved motifs. Additionally, the gene duplication of the *TcGATA* genes in the cocoa plant was also investigated. Of our interest, the relative expression levels of the *TcGATA* genes were investigated according to available transcriptome databases. The results exhibited differential expression patterns of all *TcGATA* genes in various developmental stages of zygotic and somatic embryogenesis, indicating that these *TcGATA* genes divergently function during various developmental stages of the zygotic and somatic embryos. Moreover, *TcGATA* genes were differently expressed under *Phytophthora megakarya* treatment across different points of treatment and cocoa varieties. To sum up, our findings could provide a basis for a further deep understanding of the GATAs in the cocoa plant.



## INTRODUCTION

*Theobroma cacao* L. (2n = 2x = 20), origin in Amazonian lowland rainforests, was domesticated over 1,500 years ago (Motamayor et al. 2002) and

has been grown in more than 50 countries in the world (Diaz-Valderrama et al. 2020; Jaimez et al. 2022). The chocolate tree is considered an economically important species due to its use in the cosmetics, and confectionery industries (Tan et al. 2021), chocolate production (Figueira & Scotton 2020), and medicinal benefits (Pucciarelli 2013). Chocolate tree plantations are often grown in agroforestry ecosystems alongside other fruit and commercial crops, thus providing lasting economic and environmental benefits (Guiltinan et al. 2008). Furthermore, the cocoa tree supplies essential livelihoods for 40 - 50 million people worldwide (Wickramasuriya & Dunwell 2018). It would be significant to investigate the growth and developmental processes of this important chocolate-producing tree at the molecular level. Although the cocoa tree has been viewed as an experimental organism with some limitations in the research field (Figueira & Scotton 2020), its genome was an excellent resource permitting accelerated progress in plantation, breeding, and allowing the understanding of the biochemistry of this tree crop (Motamayor et al. 2013).

Plants are exposed to a variety of environmental stresses that reduce and limit crop productivity. To adapt and survive under these adverse conditions, plants are equipped with specific genes that confer tolerance to such stresses and also regulate their developmental processes. The link between these stress-response mechanisms and gene regulation is predominantly facilitated by transcription factors (TFs). Briefly, TFs are proteins that bind to specific DNA sequences, thereby controlling the transfer of genetic information from DNA to mRNA, playing a crucial role in turning genes on or off in response to environmental stimuli. Among them, GATA has been regarded as one of the ubiquitous TF families in plants that plays an essential role in many processes of plant development, metabolism and signal transduction (Reyes et al. 2004; Behringer & Schwechheimer 2015; Schwechheimer et al. 2022), and abiotic stress signalling (Gupta et al. 2017; Zhang et al. 2021; Zhao et al. 2021). Specifically, the GATA proteins shared a highly conserved type IV zinc finger motif (Schwechheimer et al. 2022) followed by a basic region facilitating DNA binding (Merika & Orkin 1993; Teakle et al. 2002; Reyes et al. 2004; Behringer & Schwechheimer 2015; Schwechheimer et al. 2022). Due to the large number of plant genomes available, the GATAs have been investigated in numerous dicotyledonous and monocotyledonous species, such as *Arabidopsis thaliana* (Teakle et al. 2002; Kim et al. 2021a), soybean (*Glycine max*) (Zhang et al. 2015), rice (*Oryza sativa*) (Gupta et al. 2017), apple (*Malus* × *domestica*) (Chen et al. 2017), grape (*Vitis vinifera*) (Zhang et al. 2018), cotton (*Gossypium* spp.) (Zhang et al. 2019), chickpea (*Cicer arietinum*) (Niu et al. 2020), pepper (*Capsicum annuum*) (C Yu et al. 2021), sweet cherry (*Prunus avium*) (Manzoor et al. 2021), cucumber (*Cucumis sativus*) (Zhang et al. 2021), potato (*Solanum tuberosum*) (R Yu et al. 2021), four Rosaceae species (Manzoor et al. 2021), purple false brome (*Brachypodium distachyon*) (Peng et al. 2021), *Populus* (Kim et al. 2021b), wheat (*Triticum aestivum*) (Du et al. 2022; Feng et al. 2022), foxtail millet (*Setaria italica*) (Lai et al. 2022), and peanut (*Arachis hypogaea*) (Li et al. 2023). Among them, the function of GATA TFs have been well-investigated in many plants (Gupta et al. 2017; Zhu et al. 2020; Zhang et al. 2021; C Yu et al. 2021; Feng et al. 2022; Li et al. 2023; Le et al. 2023). Unfortunately, even though the genome of the cocoa tree has been published recently (Motamayor et al. 2013), the characterization of the GATAs of this important economical species has not been described.

In this current study, we aimed to conduct a systematic investigation of the GATA TFs in the cocoa genome by using bioinformatics ap-

proaches. Here, we performed genome-wide identification and characterization of the cocoa GATAs. To establish the evolutionary relationship of GATA TFs in cocoa with other plants, we also performed a comparative genomic analysis. Finally, the expression levels of the *GATA* genes were investigated by using previous RNA-Seq datasets. The obtained results will provide a cornerstone to understand various plant TF characteristics including evolutionary insights.

## MATERIALS AND METHODS

### Identification and annotation of GATA family in cocoa tree

To identify all putative members of the TcGATA family in the cocoa genome, a TBLASTN search (Gertz et al. 2006) in the NCBI, Phytozome (Goodstein et al. 2012) and PlantTFDB databases (Jin et al. 2017) has been conducted against the recent genome of this species (BioProject accession: PRJEB14326) (Motamayor et al. 2013) using well-characterized AtGATA proteins from *A. thaliana* (Teakle et al. 2002) as queries (cut-off value < 10e-4). The Pfam database (Mistry et al. 2021) was then used to confirm all potential candidates that included the conserved GATA zinc finger domain (Pfam accession: PF00320). Afterward, the full-length protein, coding DNA (CDS) and genomic DNA (gDNA) sequences and the corresponding identifier of each TcGATA member were collected for further analyses.

### Characterization of GATAs in cocoa tree

The exon/intron organization of *TcGATA* genes was constructed from the CDS and gDNA of each TcGATA member by using Gene Structure Display Server v2 (GSDS) (Hu et al. 2015) and the physicochemical features of the TcGATA proteins were calculated by the Expasy Protparam online tool (Gasteiger et al. 2003; Gasteiger et al. 2005) as previously described (Niu et al. 2020; Wang et al. 2021). Sub-cellular localization prediction of TcGATA proteins was performed by using the SherLoc2 program (Briesemeister et al. 2009). The gene ontology of *TcGATA* genes, including biological functions, cellular content, and molecular functions, was estimated by NETGO 2.0 (Yao et al. 2021) with scores higher than 0.8. The conserved motifs in the GATA TFs were screened by using the MEME web-based tool (Bailey et al. 2006).

### Phylogeny and gene duplication analysis of the GATAs in cocoa tree

To generate the phylogenetic tree, the MAFFT program (Katoh & Standley 2013) was used to align the full-length protein sequences of TcGATA members from cocoa and well-characterized GATA proteins from other higher plant species, including *A. thaliana* (Teakle et al. 2002; Kim et al. 2021a), apple (Chen et al. 2017), *Populus* (Kim et al. 2021b), grape (Zhang et al. 2018) and rice (Gupta et al. 2017). The Maximum likelihood (ML) phylogenetic tree was generated using MEGA version 11 software (Tamura et al. 2021) with the bootstrap test replicated 1000 times.

Gene duplications were determined as previously described (Guo et al. 2015). The ratio between Ka (the number of nonsynonymous substitutions per non-synonymous site) and Ks (the number of synonymous substitutions per synonymous site) values were calculated by using MEGA version 11 (Tamura et al. 2021) and DNASp version 6 tools (Rozas et al. 2017).

### Analysis of the expression profiles of the *GATA*s in cocoa

The expression features of *TcGATA* genes were detected at different de-

velopmental stages of zygotic and somatic embryos by investigating data in a public database (GEO accession: GSE55476) (Maximova et al. 2014) available from the NCBI GEO (Barrett et al. 2013). Additionally, the expression profiles of the *TcGATA* genes under pathogen infection were investigated by analyzing the previous microarray atlas (GEO accession: GSE116041) (Pokou et al. 2019). Relative expression values of *TcGATA* genes were estimated using *Actin 11*, the most stable expressed gene in various cocoa tissues (Pinheiro et al. 2011), as a reference gene, following the previous description (Cao 2022). Up- and down-regulated genes were defined by a fold-change cut-off (|fold-change| ≥ 1.5-fold) between 6, 24, and 72 hours after inoculation (hai) and 0 hai.

## RESULTS AND DISCUSSION

### Identification and annotation of the TcGATA proteins in cocoa tree

A total of 24 putative *TcGATA* genes were identified in the cocoa genome (Table 1), along with their annotations, like Phytozome locus and their corresponding sequences. Finally, we assigned these 24 GATA full-length protein sequences to TcGATA01 to TcGATA24 based on their physical location on the genome (Table 1).

Recently, there has been a significant effort to identify and characterize GATA TFs in various higher plant species, including both dicotyledonous and monocotyledonous plants (Table S1). Compared to other plants, the TcGATA family found in the cocoa genome is larger than in sweet cherry (18 genes) (Manzoor et al. 2021), *Ophiorrhiza pumila* (18

**Table 1.** Summary of the TcGATAs in cocoa.

| Gene | Phytozome locus | Gene size | Protein size | MW | pI | AI | GRAVY | SCL |
|---|---|---|---|---|---|---|---|---|
| TcGATA01 | Thecc.01G024200 | 1414 | 389 | 42.57 | 5.84 | 54.94 | -0.70 | Nucl |
| TcGATA02 | Thecc.01G034600 | 875 | 249 | 27.78 | 8.47 | 41.16 | -0.86 | Nucl |
| TcGATA03 | Thecc.01G136600 | 747 | 248 | 28.43 | 5.54 | 61.29 | -0.55 | Golgi, Extra, Vacu, Nucl |
| TcGATA04 | Thecc.01G308700 | 489 | 119 | 13.25 | 9.98 | 64.79 | -0.76 | Nucl |
| TcGATA05 | Thecc.01G385800 | 939 | 273 | 30.63 | 7.71 | 46.92 | -0.89 | Nucl |
| TcGATA06 | Thecc.02G040700 | 2410 | 238 | 26.42 | 7.66 | 80.71 | -0.52 | Nucl |
| TcGATA07 | Thecc.02G076600 | 1071 | 322 | 35.51 | 5.52 | 55.12 | -0.62 | Nucl |
| TcGATA08 | Thecc.02G128000 | 5007 | 353 | 38.95 | 5.00 | 62.41 | -0.73 | Nucl |
| TcGATA09 | Thecc.02G128100 | 3563 | 313 | 33.18 | 5.01 | 62.27 | -0.58 | Nucl |
| TcGATA10 | Thecc.03G242400 | 4664 | 538 | 59.88 | 6.52 | 70.67 | -0.66 | Nucl, Cyto |
| TcGATA11 | Thecc.04G215200 | 1572 | 243 | 26.93 | 8.43 | 57.82 | -0.75 | Nucl |
| TcGATA12 | Thecc.04G294500 | 642 | 147 | 16.08 | 9.76 | 63.06 | -0.94 | Nucl, Cyto |
| TcGATA13 | Thecc.05G293000 | 1121 | 302 | 33.38 | 8.79 | 66.19 | -0.67 | Nucl, Cyto |
| TcGATA14 | Thecc.05G319600 | 1047 | 171 | 18.64 | 10.14 | 61.58 | -0.61 | Nucl, Cyto |
| TcGATA15 | Thecc.06G060500 | 9887 | 308 | 33.85 | 6.59 | 60.71 | -0.85 | Nucl |
| TcGATA16 | Thecc.06G097700 | 1193 | 320 | 34.94 | 6.14 | 55.22 | -0.67 | Nucl |
| TcGATA17 | Thecc.08G072900 | 4282 | 299 | 32.11 | 5.69 | 63.95 | -0.57 | Nucl |
| TcGATA18 | Thecc.08G073000 | 3928 | 355 | 38.47 | 4.69 | 67.52 | -0.68 | Nucl |
| TcGATA19 | Thecc.09G046000 | 1493 | 363 | 39.92 | 5.88 | 57.47 | -0.71 | Nucl |
| TcGATA20 | Thecc.09G053800 | 1562 | 311 | 34.00 | 9.21 | 57.81 | -0.67 | Nucl, Cyto |
| TcGATA21 | Thecc.09G089800 | 1189 | 302 | 33.64 | 9.15 | 59.83 | -0.84 | Nucl, Cyto |
| TcGATA22 | Thecc.09G218100 | 2714 | 255 | 28.15 | 9.04 | 45.57 | -1.11 | Nucl, Cyto |
| TcGATA23 | Thecc.09G342100 | 2326 | 342 | 37.04 | 8.75 | 60.38 | -0.63 | Nucl |
| TcGATA24 | Thecc.10G075500 | 1460 | 341 | 37.09 | 6.67 | 65.51 | -0.57 | Nucl |

Note: -: No information, protein size (amino acid residues), MW: Molecular weight (kDa), AI: Aliphatic index, pI: Iso-electric point, GRAVY: Grand average of hydropathicity, SCL: Sub-cellular localization, Nucl: Nuclear, Cyto: Cytoplasm, Golgi: Golgi apparatus, Vacu: Vacuolar, Extra: Extracellular

genes) (Shi et al. 2022), castor bean (*Ricinus communis*) (19 genes), grape (19 genes) (Zhang et al. 2018), Japanese apricot (*Prunus mume*) (20 genes), and peach (*Prunus persica*) (22 genes) (Manzoor et al. 2021), sugar beet (*Beta vulgaris*) (16 genes) (Le et al. 2023). However, it is significantly smaller than in other plants, such as *A. thaliana* (30 genes) (Reyes et al. 2004), *P. bretschneideri* (Manzoor et al. 2021), apple (35 genes) (Chen et al. 2017), seven *Populus* spp. (33 to 40 members), *G. hirsutum* (87 genes) (Zhang et al. 2019), and wheat (79 members) (Feng et al. 2022). These comparisons reveal that the number of GATA members varies greatly across different plant species.

## Analysis of the physical and chemical features of the TCGATA proteins in cocoa

The predicted full-length amino acid sequences of 24 TcGATA members in cocoa were analyzed using the ExPaSy Protparam tool (Gasteiger et al. 2003, 2005). The investigation provided information on the physical and chemical properties of the TcGATAs in cocoa, which are summarized in Table 1. The full-length of the predicted protein sequences encoded by the 24 TcGATAs varied from 119 (TcGATA04) to 538 amino acid residues (TcGATA10). The weights of the TcGATAs ranged from 13.25 kDa (TcGATA04) to 59.88 kDa (TcGATA07). The theoretical isoelectric point (pI) values of the TcGATAs were distributed between 4.69 (TcGATA18) and 10.14 (TcGATA14), with 12 TcGATAs being acidic (pI values ranging from 4.69 to 6.67) and the remaining sequences being basic (pI values varying from 7.71 to 10.14). The aliphatic index (AI) values of the TcGATAs ranged from 41.16 (TcGATA02) to 77.86 (TcGATA06). Additionally, the grand average of hydropathicity (GRAVY) values for all members of the TcGATAs in cocoa were less than 0, indicating that the TcGATAs were hydrophilic proteins (Table 1).

The obtained results were in agreement with the previously comprehensive analysis conducted on the general characteristics of GATA TFs in various higher plant species. For instance, GATAs in Rosaceae woody species, like *Pyrus bretschneideri*, *Prunus avium*, *P. mume*, and *P. persica*, were reported to range from 119 to 548 amino acid residues in full-length sequences and from 12.99 to 60.23 kDa in molecular weight, respectively (Manzoor et al. 2021). In grapes, the protein full-lengths of GATA TFs ranged from 109 to 386 amino acid residues (Zhang et al. 2018), while apples had GATAs with amino acid residues varying from 90 to 1161 (9.9 to 129.74 kDa) (Chen et al. 2017). Additionally, seven *Populus* species were found to possess a total of 389 predicted GATA proteins, with sequence lengths ranging from 82 to 791 amino acid residues, except for PtsGATA29, which had only 46 amino acid residues (Kim et al. 2021b). Additionally, the pI values of GATA TFs in higher plant species were found to a wide range from acidic to base, with pI scores in four Rosaceae woody species ranging from 4.71 to 10.07 (Manzoor et al. 2021), and peanut (*Arachis hypogaea)* ranging from 4.75 to 10.21 (Li et al. 2023), respectively. These varying pI scores are due to their different protein lengths. Interestingly, the GRAVY values of all members of GATAs in apples (Chen et al. 2017), four Rosaceae species (Manzoor et al. 2021), and peanuts (Li et al. 2023) were evidently negative, indicating that these GATAs may be hydrophilic (Schwechheimer et al. 2022). Overall, the physical and chemical properties of TcGATAs in cocoa, and possibly other plant species, were highly variable based on the study results. The dissimilarity of the physical and chemical properties proposed that GATAs might play different functions in plants.

## Phylogenetic analysis, Gene structure and Conserved Motif of the GATAs in cocoa

A phylogenetic tree comprising all of the 24 TcGATAs and well-characterized GATAs from *A. thaliana* (Teakle et al. 2002; Kim et al. 2021a), grape (Zhang et al. 2018), and *P. trichocarpa* (Kim et al. 2021b) has been constructed in order to clarify the phylogenetic relationships of the TcGATAs in cocoa (Figure 1).

According to the ML estimation, 24 TcGATA proteins were divided into four different groups, namely groups I, II, III, and IV, respectively, as shown in Figure 1. Specifically, group I had the largest number of cocoa TcGATA proteins (10 TcGATA members), followed by group II (seven TcGATA members), and group III (five TcGATA members). Group IV had the least number of cocoa TcGATA proteins, with only two members, including TcGATA06 and TcGATA10, respectively (Figure 1).

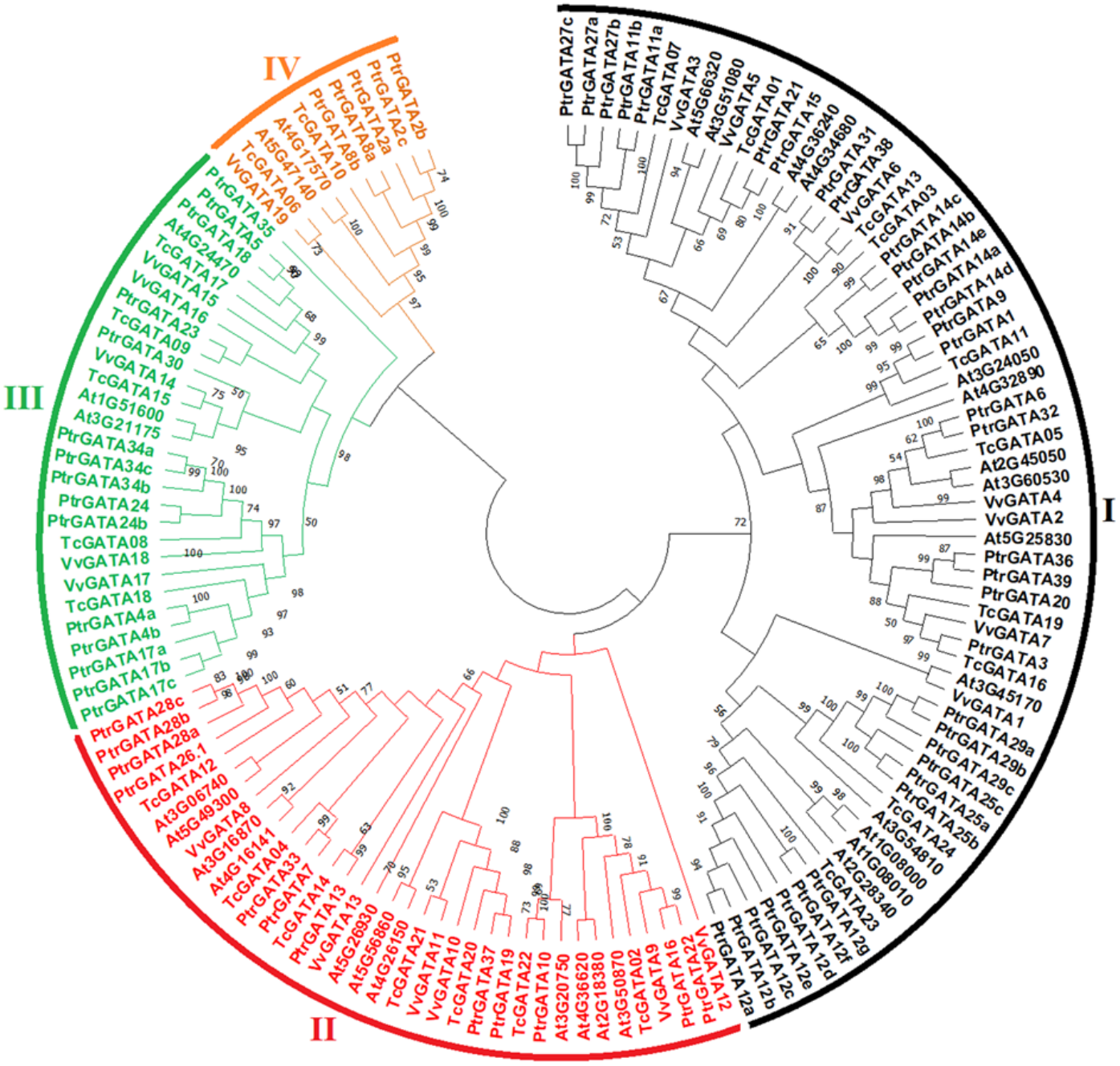


**Figure 1.** Phylogenetic tree of GATA family from *Arabidopsis thaliana* (At), *Populus trichocarpa* (Ptr), *Vitis vinifera* (Vv), and *Theobroma cacao* (Tc).

Previously, the classification of GATA TFs had been established for different higher plant species (Table S1). GATA TFs from many a large number of dicotyledonous plants were also classified into four main clades, like *A. thaliana* (Teakle et al. 2002), four Rosaceae species (Manzoor et al. 2021), peanuts (Li et al. 2023), and grape (Zhang et al. 2018). Out of four clades, clades I and IV contained the largest and smallest number of GATAs, respectively. Particularly, 49 GATA TFs found in potato (*Solanum tuberosum*) were divided into five groups, group II had the largest number of GATA proteins (15 GATA members), followed by groups IV (13 GATA members), V (10 GATA members), III (eight GATA members), and I (only three GATA members) (R Yu et al. 2021). But the phylogenetic tree was constructed from only the 49 GATAs of potato. Therefore, the classification of potato GATAs might need further comparison with other species. The clade V, VI, and VII GATAs were reported only in rice with two, four, and two members, respectively (Gupta et al. 2017).

The exon/intron arrangement of the cocoa *TcGATA* genes was then examined. The results showed that newly discovered cocoa *TcGATA* family genes have exon counts ranging from 1 to 11 (Figure 2). Interestingly, the *TcGATA* genes in the same clade may share the seminar gene organization (Figure 1, 2). For example, eight (out of ten) members in group I included two exons, except for *TcGATA03* had only one exon, and *TcGATA01* contained three exons (Figure 2). Seven (out of nine) *TcGATA* members of group II contained three exons, while two others had two exons (*TvGATA02* and *TcGATA04*). In group III, a majority member had seven exons, whereas two remaining members had 10 (*TcGATA18*) and 11 exons (*TcGATA08*) (Figure 2). Two group IV-belonging *TcGATA* genes included four (*TcGATA06*) and eight (*TcGATA10*) exons, respectively (Figure 2). Our findings confirmed the wide range of variability in exon numbers of *GATA* genes identified in four Rosaceae species (from 1 to 10 exons), peanut (Chen et al. 2017), grape (from 1 to 11 exons) (Zhang et al. 2018), and *Populus* species (from 1 to 12 exons) (Kim et al. 2021b). The unique gene architecture of TcGATA15 and TcGATA17, characterized by their extremely short exons and

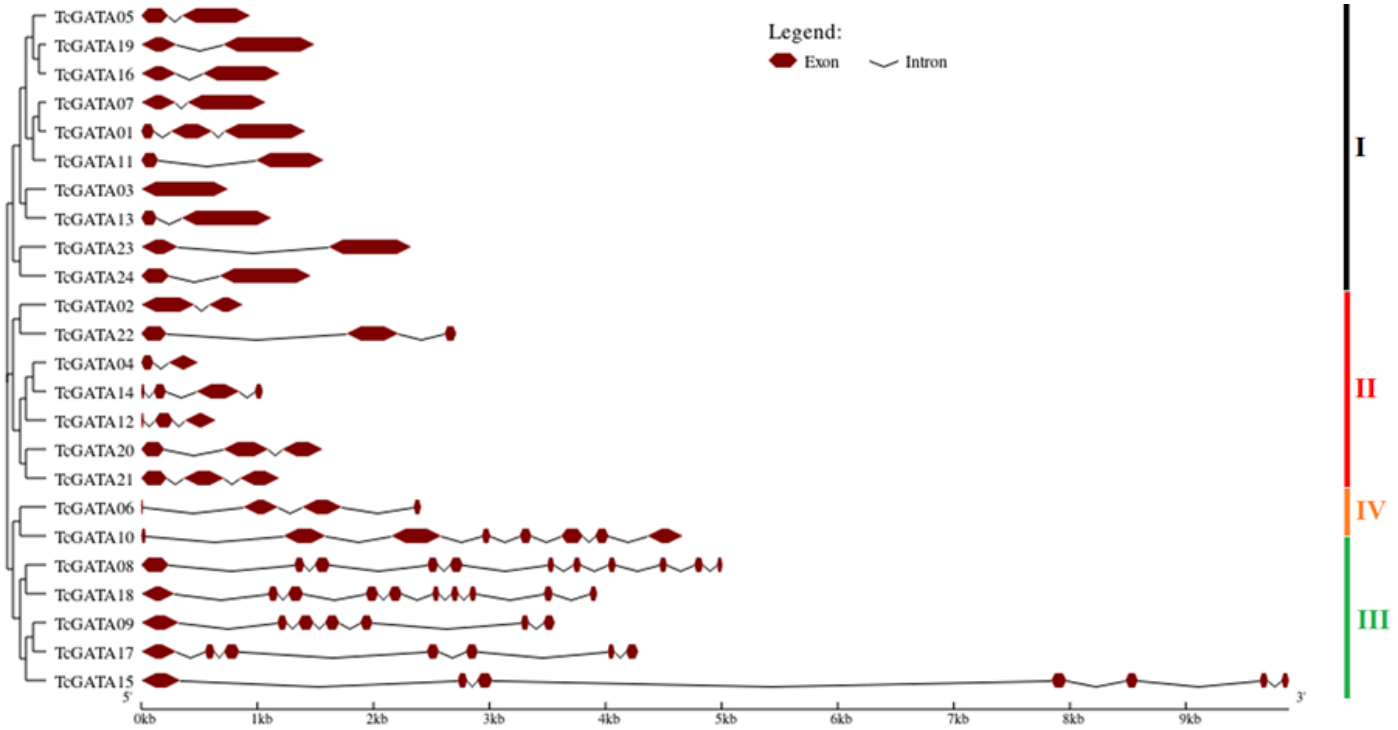


**Figure 2.** Gene exon/intron organizations of the *GATA* family in cocoa.

long introns, offers fascinating insights into evolutionary processes. This atypical structure, contrasting with more common gene architectures, aligns with theories suggesting that the evolution of introns is linked to alternative splicing and the resulting functional diversity in proteins. The presence of long introns in these two genes could potentially delay transcriptional output, providing a mechanism for the suppression of gene expression under adverse conditions. This hypothesis aligns with the broader concept that gene architecture can be an adaptive trait in evolution, where specific structural features, like long introns, may confer selective advantages in response to environmental challenges. The divergence in the gene structure of the *GATA* family of cocoa suggests that the *TcGATA* genes underwent an evolutionary change, which might have generated the functional separation of the *GATA* family and might enable genes to have new functions that can help plants better adapt to environmental changes (Fan et al. 2014).

The evolutionary relationship and classification of the TcGATA family were validated by analyzing their conserved motifs predicted and confirmed by the MEME program (Bailey et al. 2006) (Figure 3). The conserved motif 1 was present in all TcGATA proteins, and the majority of members of the same TcGATA group exhibited similar patterns. Group IV had the lowest number of conserved motifs (1), while group I had the highest (6). However, some proteins had distinct conserved mo-

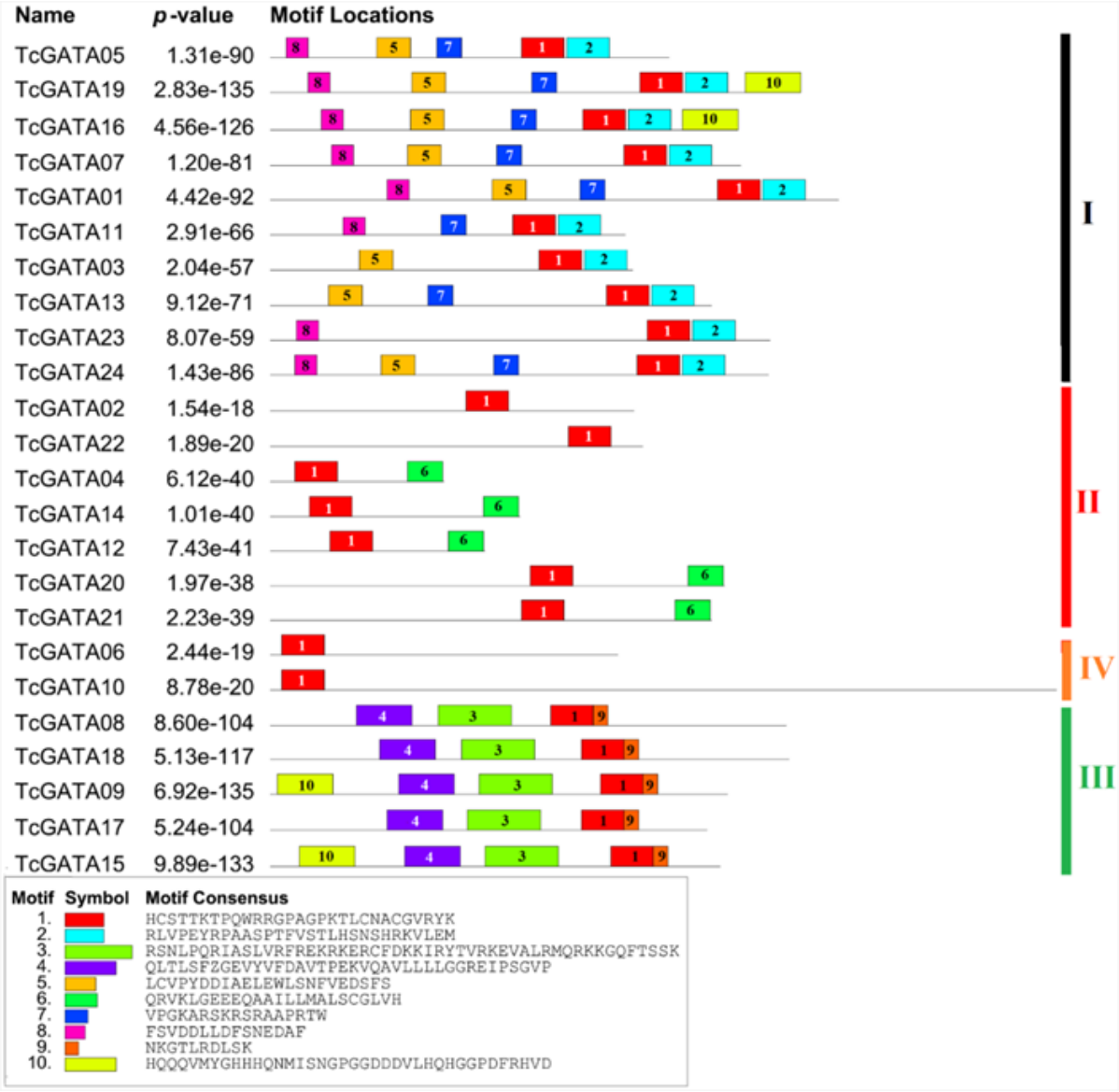


**Figure 3.** Conserved motifs of the GATA family members in cocoa generated by MEME.

tifs across different groups. For instance, motif 1 was unique to all groups. All members of group I contained motifs 2, 5 except for two genes (TcGATA11 and TcGATA23), motif 7 except for two genes (TcGATA03 and TcGATA23), and motif 8 except for two genes (TcGATA03 and TcGATA13). In addition, motif 10 was found in two members, TcGATA16 and TcGATA19, respectively. For group II, motif 6 was detected in five (out of seven) members (TcGATA04, TcGATA12, TcGATA14, TcGATA20 and TcGATA21). All members of group III contained motifs 3, 4, and 9. Moreover, motif 10 was recorded in two genes, TcGATA09 and TcGATA15, respectively. The common motif detected in all TcGATA was the zinc finger loop (C-X2-C-X18–20-CNAC) domain. GATA members in groups I, II, and IV had the C-X2-C-X18-CNAC conserved domain, while the group III members harboured the C-X2-C-X20-CNAC domain (Figure 6). Additionally, the conserved amino acid motif TPQWRXGPXGXKTL was identified between the second and third cysteine residues in the C-X2-C-X18-CNAC zinc finger loop of group I while the conserved amino acid motif TX2T-PLWRXGPXGPKXL was detected between the second and third cysteine residues in the C-X2-CX18-CNAC zinc finger loop of group II. Moreover, the conserved amino acid motif GXSX3TPXMRRGPXGPRXL was detected between the second and third cysteine residues in the C-X2-CX20-CNAC zinc finger loop of group III and the conserved amino acid motif GX2STPLWRNGPPEK-PVL was identified between the second and third cysteine residues in the C-X2-CX18-CNAC zinc finger loop (Figure 4).

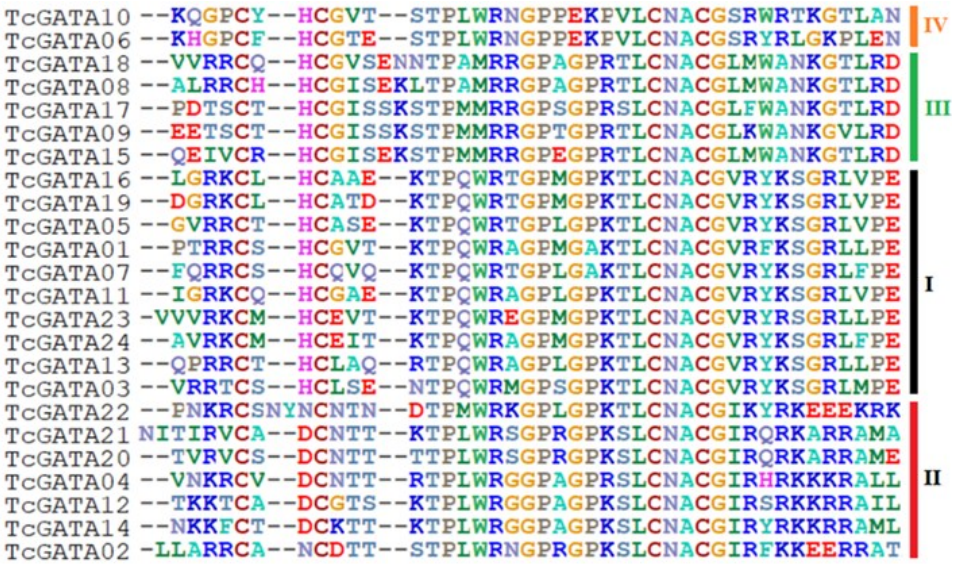


**Figure 4.** Alignments of GATA domains of all identified TcGATA family members in cocoa tree.

Our findings showed that the motif distributions of TcGATA proteins were comparable within each subfamily. The presence of the same motif in all groups or in each group suggested that they might have fundamental functions. The conserved GATA domains and motifs found between the second and third cysteine residues in the C-X2-CX18-20-CNAC zinc finger loop found in different groups of TcGATAs were consistent with conserved structures previously identified uin peanut (Li et al. 2023), chickpea (Niu et al. 2020), and *Populus* species (Kim et al. 2021b). The examination of gene architectures and conserved motifs reveals that GATA members within a group exhibit relatively high conservation properties in various species and that members among groups exhibit reasonably high conservation properties.

## Physical distribution and gene duplication of the GATAs in cocoa

The distribution of the 24 *TcGATA* genes across the cocoa genome was investigated in this study. Results showed that the *TcGATA* gene family was distributed randomly across the genome (Figure 5). The quantity of *TcGATA* genes differs across various chromosomes, with chromosomes 9 and 16 containing the largest number of *TcGATA* gene distributions with five members, followed by chromosome 2 with four members, and chromosomes 4, 5, 6, and 8 with two members each (Figure 5). It is noteworthy that chromosomes 3 and 10 each only had one *GATA* gene, while chromosome 7 had no *TcGATA* gene (Figure 5).

As an intriguing aspect of this research, the duplication events that occurred in the *TcGATA* gene family in cocoa were predicted as previously described (Niu et al. 2020), with details provided in Figure 5 and Table 2. Three duplicate genes were found in the *TcGATA* family, with nucleotide similarities ranging from 53.3 (*TcGATA08* and *TcGATA18*) to 57.3% (between *TcGATA09* and *TcGATA17*). These findings indicate that whole genome duplication (WGD) and segmental duplication (SD) events played a significant role in the expansion of the *TcGATA* gene family. Additionally, the Ka/Ks ratios for the three duplicated genes were all less than 1, ranging from 0.26 (*TcGATA08* and *TcGATA18*) to 0.30 (*TcGATA20* and *TcGATA21*), indicating that the *TcGATA* genes were under strong purifying selection.

These findings showed a similar trend in chromosome distribution and evolution of *GATA* genes in many plant species. The random distribution of *GATA* genes across the genome was reported in seven *Populus* species (Kim et al. 2021b), four Rosaceae species (Manzoor et al. 2021) and chickpea (Niu et al. 2020). Interestingly, there was no tandem duplication observed in the *TcGATA* family, while two WGD and one SD events accounted for the duplication events in the *TcGATA* family (Table 2). This result confirmed that the WGD and SD events played a significantly important role in the evolution of the *GATA* genes compared to tandem duplication events, as also observed in chickpeas (Niu et al. 2020), *Populus* species (Kim et al. 2021b), grape (Zhang et al. 2018), and perhaps many other plants (Zhang et al. 2019; Li et al. 2023).

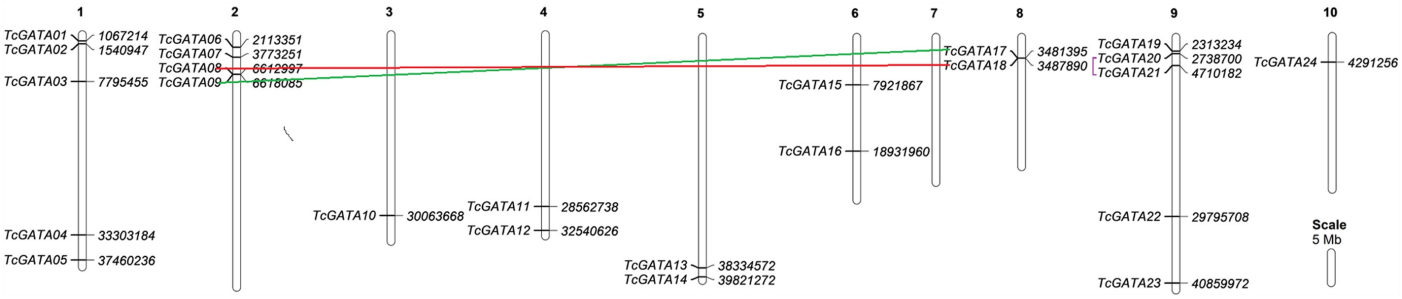


**Figure 5.** The chromosomal distribution of *TcGATA* genes in the cocoa genome. The red lines indicated the duplication events. The chromosome number is indicated above for each chromosome.

**Table 2.** Prediction of the duplication events in the Tc*GATA* gene family in cocoa.

| Duplicated gene | Duplicated gene | Duplication event | Similar level (%) | Ka | Ks | Ka/Ks |
|---|---|---|---|---|---|---|
| *TcGATA08* | *TcGATA18* | WGD | 53.3 | 0.342 | 1.31 | 0.26 |
| *TcGATA09* | *TcGATA17* | WGD | 57.3 | 0.363 | 1.35 | 0.27 |
| *TcGATA20* | *TcGATA21* | SD | 55.2 | 0.441 | 1.47 | 0.30 |

Note: WGD: Whole genome duplication, SD: Segmental duplication, Ka: the number of nonsynonymous substitutions per non-synonymous site, Ks: the number of synonymous substitutions per synonymous site.

## Gene ontology analysis of the GATAs in cocoa

In this study, gene ontology (GO) analysis was used to annotate the probable roles of the TcGATA TFs. Appropriately, 24 TcGATAs were then categorized into 55 functional groups and divided into three main ontologies, including cellular component, biological process, and molecular function (Figure 6). As a result, in the cellular component category, all 24 TcGATAs anticipated their function in the nuclear, intracellular organelle, while only one member awaited the role in the intracellular, non-membrane-bounded organelle (Figure 6). The GO analysis also indicated that all TcGATAs were distributed in the nucleus (Figure 6), which was also confirmed by the sub-cellular localization prediction by the SherLoc2 tool (Table 1). All TcGATAs were localized in the nuclear compartment followed by the cytoplasm (seven out of 24) (Figure 6). While only one member of the TcGATA was predicted to localize on the Golgi apparatus, vacuolar, plasma membrane, or extracellular (Table 1).


**Figure 6.** GO analysis involving in molecular function, biological processes, and cellular components of TcGATAs investigated by NETGO 2.0.

It has been thought that the determination of the sub-cellular localization of proteins can provide insight into their potential roles (Goodin 2018). In the molecular function category, all TcGATA proteins were predicted to act as TFs (DNA binding, nucleic acid binding, and sequence-specific DNA binding) (Figure 6). Under the biological process annotation, all TcGATAs were associated with biological processes, and 18 out of 24 TcGATAs anticipated their function in the regulation of biological processes. In addition, 11 out of 24 TcGATAs were predicted to function in response to stimuli. These obtained results were also in agreement with the previous reports that the 32 PbGATAs of Chinese white pear anticipated their functionality in DNA binding, and nucleic acid binding TF (Manzoor et al. 2021).

## Expression patterns of the cocoa TcGATAs

In this study, of our interest, the expression pattern of the *TcGATA* genes in different stages of cocoa embryo development was investigated (Figure 7). In general, most *TcGATA* genes were expressed in all stages of embryo development, except for *TcGATA14* (Figure 7). The expressed *TcGATA* genes exhibited different expression levels during different developmental stages of the embryo (Figure 7). Sixteen *TcGATA* genes were differentially expressed during zygotic embryo maturation, with 11 *TcGATA* genes displaying higher expression levels in the mature zygotic embryo tissues than in other developmental stages, including *TcGATA03*, *TcGATA05*, *TcGATA07*, *TcGATA11*, *TcGATA22* (group I), *TcGATA04*, *TcGATA20*, *TcGATA21* (group II), *TcGATA08*, *TcGATA18* (group III), and *TcGATA10* (group IV), respectively (Figure 7). However, four *TcGATA* genes belonging to group I showed lower expression levels in mature zygotic embryo samples than in early developmental stages, including *TcGATA02*, *TcGATA03*, *TcGATA23*, and *TcGATA24* (Figure 7). Similarly, five *TcGATA* genes exhibited higher expression at mature (M-SE) than late torpedo (LT-SE) developmental stages of somatic embryogenesis, including *TcGATA05*, *TcGATA07*, *TcGATA11*, *TcGATA04*, and *TcGATA18*, respectively. However, *TcGATA16* showed a lower expression level at M-SE than LT-SE developmental stages of somatic embryogenesis (Figure 7). At the same developmental stages, differential gene expression between zygotic and somatic embryogenesis was recorded. At the torpedo stage, four *TcGATA* genes (*TcGATA02*, *TcGATA03*, *TcGATA16*, and *TcGATA23*, respectively) were more expressed in zygotic embryos compared to somatic embryo while three other genes (*TcGATA04*, *TcGATA07*, and *TcGATA10*, respectively) were less expressed. At the mature stage, 11 *TcGATA* genes (*TcGATA01*, *TcGATA03*, *TcGATA04*, *TcGATA08*, *TcGATA09*, *TcGATA17*, *TcGATA18*, *TcGATA20*, *TcGATA21*, *TcGATA22*, and *TcGATA23*, respectively) had higher expression levels in zygotic embryos compared to somatic embryo while three other genes (*TcGATA07*, *TcGATA16*, and *TcGATA19*, respectively) had lower expression levels (Figure 7). Overall, the expression of *TcGATA* genes in embryogenesis suggested that this transcription family played an important role in the seed development of cocoa. The differential expression patterns of different genes in various developmental stages of zygotic and somatic embryogenesis indicated that different *TcGATA* genes divergently function during various developmental stages of the zygotic and somatic embryos. Despite the large number of reports of the genome-wide analysis of GATAs in plants, the function of this family in embryogenesis has been poorly communicated. Earlier, the GATA factor HANABA TARANU was reported to be required to position the proembryo boundary in the early embryo of *A. tha-*

*liana* (Nawy et al. 2010). On the other hand, the expression of two *GATAs* (*GATA NITRATE-INDUCIBLE CARBON-METABOLISM-INVOLVED* and *CYTOKININ-RESPONSIVE GATA1*) in *Arabidopsis* seedlings has been described (Chiang et al. 2012). In addition, the BME3 (Blue Micropylar End 3) GATA TF has been previously described as a positive regulator of *Arabidopsis* seed germination (Liu et al. 2005). So, our findings provided evidence that indicated the function of *GATAs* in embryo development in plants. Moreover, further deep investigation might be required to explore the role of GATAs in the seed development of the seed crop species.

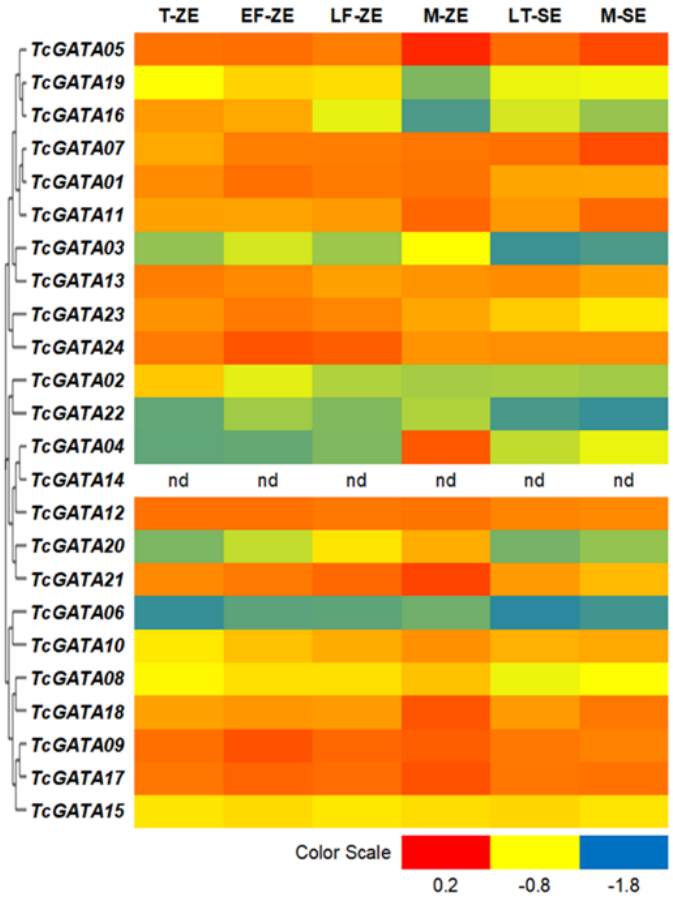


**Figure 7.** Expression patterns of the *T. cacao GATA* genes during Zygotic (ZE) and Somatic Embryo (SE) maturation. Values represented log2 of the relative expression level of *TcGATA* genes per expression level of *Actin 11* gene which was the most stable expressed gene in various tissues (Pinheiro et al. 2011). T-ZE: Torpedo zygotic embryo, EF-ZE: Early-full zygotic embryo, LF-ZE: Late-full zygotic embryo, M-ZE: Mature zygotic embryo, LT-SE: Late Torpedo somatic embryo, M-SE: Mature somatic embryo, nd: non-determined.

A significant number of *TcGATA* genes showed differential expression under *Phytophthora megakarya* treatment across different points of treatment and cocoa varieties (Figure 8). Particularly, the expression of *TcGATA22* was not detected in any treatments of both Nanay and Scavina genotypes. At 6 hours after inoculation (hai), only two genes, *TcGATA06* and *TcGATA23*, showed an increase in relative expression level in the Nanay genotype. However, in the Scavina genotype, four genes, including *TcGATA05*, *TcGATA08*, *TcGATA12* and *TcGATA18*, were up-regulated by *P. megakarya* treatment, whereas *TcGATA03* and *TcGATA04* were down-regulated. At 24 and 72 hai, in Nanay genotype, five genes, including *TcGATA04*, *TcGATA05*, *TcGATA13*, *TcGATA17*, and *TcGATA19*, were down-regulated by *P. megakarya* treatment, and only

two genes, *TcGATA06* and *TcGATA07*, were up-regulated. Differently, in the Scavina genotype, eight genes, including *TcGATA01*, *TcGATA04*, *TcGATA05*, *TcGATA06*, *TcGATA12*, *TcGATA16*, and *TcGATA20*, were up-regulated. Moreover, at 72 hai, most of the expressed *TcGATA* genes were up-regulated by *P. megakarya* treatment, except for *TcGATA03*, which was down-regulated and four genes, *TcGATA07*, *TcGATA10*, *TcGATA15*, and *TcGATA19*, which were not regulated by *P. megakarya* treatment (Figure 8). In summary, *TcGATA* genes showed different expression patterns in the susceptible (Nanay) and resistant (Scavina) cocoa genotypes under *P. megakarya* treatment at different time points (6, 24, and 72 hours) after inoculation. These discovered results indicated that *TcGATA* genes function differently under *P. megakarya* treatment in various genotypes of cocoa tree. The increase in relative expression level from 6 hai to 72 hai in the tolerance genotype contributed to explaining the function of *TcGATA*s in the biotic stress response in cocoa. In the literature, expression pattern analysis exhibited that *GATA* genes responded to diverse abiotic stresses, such as high temperature, salinity, cold, and drought treatments, in many plants, such as rice (Gupta et al. 2017), wheat (Feng et al. 2022), *oilseed rape* (Zhu et al. 2020), cucumber (Zhang et al. 2021), and pepper (R Yu et al. 2021). However, knowledge about the function of *GATA*s in the biotic response was limited until recently. For example, overexpression of *TaGATA1* showed high resistance to *Rhizoctonia cerealis* in wheat (Liu et al. 2020). A further detailed investigation into the role of *TcGATA* in *Phytophthora* might be necessary, as cocoa undergoes significant annual losses to the water mold *Phytophthora* spp. (Oomycetes) (ranging between 20 and 25% of global losses) (Adeniyi 2019).

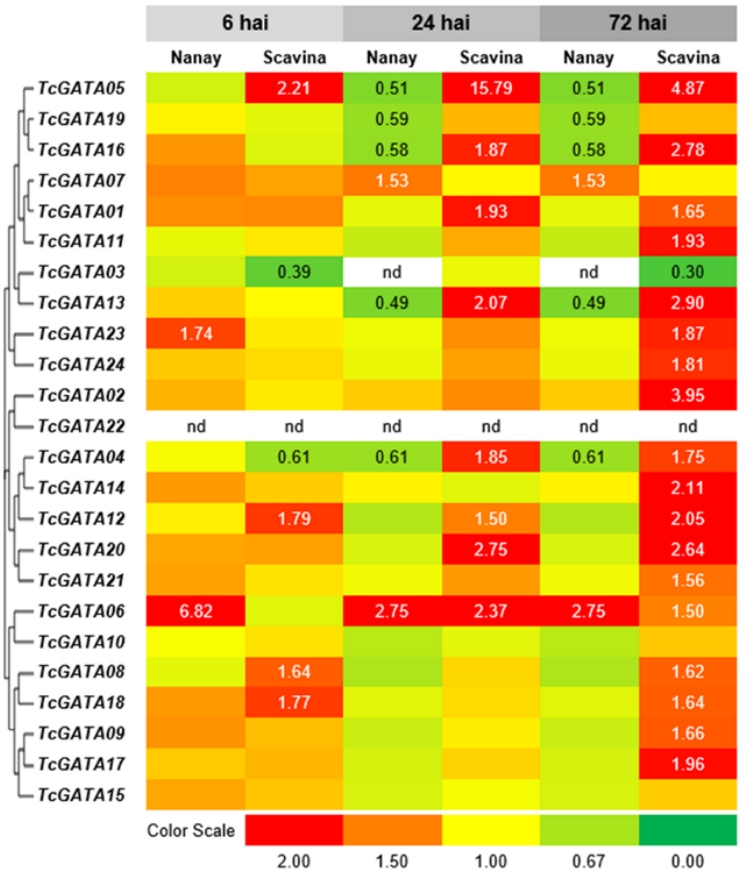


**Figure 8.** Expression patterns of the *TcGATA* gene family under inoculation of *Phytophthora megakarya*. hai: hour after inoculation, Nanay: Nanay variety (susceptible genotype), Scavina: Scavina (tolerance genotype), nd: non-determined.

## CONCLUSIONS

This present study focused on the identification and characterization of the GATA TF family in cocoa tree. A total of 24 *TcGATA* genes were identified in the assembly of cocoa. By using various tools, the physicobiochemical features, gene structure, and conserved motifs of the TcGATA proteins were analyzed. The gene expression patterns of the *TcGATA*s were investigated during the development of zygotic and somatic embryos. Moreover, their expression patterns under inoculation with *P. megakarya* were also analyzed. The results provide valuable information for further understanding the different functions of *TcGATA*s during seed development and in response to *P. megakarya* in cocoa plants. Additionally, these findings offer insightful information for comparative genomics studies in plants based on the characterization, evolution and expression of GATA gene family.

## AUTHOR CONTRIBUTION

N.T.B.C. contributed to the research design, data collection and analysis, and preparation of the first draft of the manuscript, T.M.L. contributed to data collection, H.D.C. contributed to data collection and analysis, and preparation of the first draft of the manuscript. T.T.T.H. contributed to data collection and analysis. L.T.M.T. contributed to data collection and analysis, H.V.L. contributed to the research design, data collection and analysis, Q.T.X.V. contributed to data collection and analysis, H.H.P. contributed to data collection and analysis, V.T.T. contributed to data collection, P.B.C. contributed to the research design, data collection and analysis, and preparation and editing of the manuscript and to supervise all the process.

## ACKNOWLEDGMENTS

This work was funded by the fundamental research program of Hung Vuong University under the project grant No. 01/2023/KHCN (HV01.2023).

## CONFLICT OF INTEREST

The authors declare no conflict of interest regarding the research or the research funding.

## REFERENCES

Ao, T. et al., 2015. Identification and characterization of GATA gene family in Castor Bean (*Ricinus communis*). *Plant Diver*, 37, pp. 453-462. doi: 10.7677/ynzwyj201514151

Apuli, R.P. et al., 2020. Inferring the genomic landscape of recombination rate variation in European aspen (*Populus tremula*). *G3: Genes, Genomes, Genetics*, 10(1), pp. 299-309. doi: 10.1534/g3.119.400504

Adeniyi, D., 2019. Diversity of cacao pathogens and impact on yield and global production. In *Theobroma Cacao-Deploying Science for Sustainability of Global Cocoa Economy* (pp. 1-20). doi: 10.5772/intechopen.81993

Bailey, T.L. et al., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(suppl 2), pp.W369-W373.

Barrett, T. et al., 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*, 41(Database issue), pp.D991-D995. doi: 10.1093/nar/gks1193

Behringer, C. & Schwechheimer, C., 2015. B-GATA transcription factors – insights into their structure, regulation, and role in plant development. *Front Plant Sci*, 6, 90. doi: 10.3389/fpls.2015.00090

Briesemeister, S. et al., 2009. SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J Proteome Res*, 8 (11), pp.5363-5366. doi: 10.1021/pr900665y

Cao, P.B., 2022. *In silico* structural, evolutionary, and expression analysis of small heat shock protein (shsp) encoding genes in cocoa (*Theobroma cacao* L.). *J Anim Plant Sci*, 32(5), pp.1394-1402. doi: 10.36899/JAPS.2022.5.0546

Chen, H. et al., 2017. Genome-wide identification, evolution, and expression analysis of GATA transcription factors in apple (*Malus x domestica* Borkh.). *Gene*, 627, pp.460-472. doi: 10.1016/j.gene.2017.06.049

Chiang, Y.H. et al., 2012. Functional Characterization of the GATA Transcription Factors GNC and CGA1 Reveals Their Key Role in Chloroplast Development, Growth, and Division in Arabidopsis. *Plant Physiol*, 160(1), pp.332-348. doi: 10.1104/pp.112.198705

Diaz-Valderrama, J.R. et al., 2020. The History of Cacao and Its Diseases in the Americas. *Phytopathology*, 110(10), pp.1604-1619. doi: 10.1094/PHYTO-05-20-0178-RVW

dos Santos, T.B. et al., 2022. Physiological responses to drought, salinity, and heat stress in plants: A review. *Stresses*, 2(1), pp.113-135. doi: 10.3390/stresses2010009

Du, X. et al., 2022. Genome-wide analysis of wheat GATA transcription factor genes reveals their molecular evolutionary characteristics and involvement in salt and drought tolerance. *Int J Mol Sci*, 24(1), 27. doi: 10.3390/ijms24010027

Fan, K. et al., 2014. Molecular evolution and expansion analysis of the NAC transcription factor in Zea mays. *PLoS One*, 9(11), e111837. doi: 10.1371/journal.pone.0111837

Feng, X. et al., 2022. Genome-wide identification and characterization of GATA family genes in wheat. *BMC Plant Biol*, 22(1), 372. doi: 10.1186/s12870-022-03733-3

Figueira, A. & Scotton, D. C., 2020. 13.1 Theobroma cacao Cacao. In *Biotechnology of Fruit and Nut Crops* (2 ed.). CABI, pp.282-313.

Gasteiger, E. et al., 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*, 31(13), pp.3784-3788. doi: 10.1093/nar/gkg563

Gasteiger, E. et al., 2005. Protein identification and analysis tools on the ExPASy server. In *The proteomics protocols handbook*. Springer, pp. 571-607.

Gertz, E.M. et al., 2006. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology*, 4(1), 41. doi: 10.1186/1741-7007-4-41

Goodin, M.M., 2018. Chapter six - Protein localization and interaction studies in plants: Toward defining complete proteomes by visualization. *Adv Virus Res*, 100, pp.117-144. doi: 10.1016/bs.aivir.2017.10.004.

Goodstein, D.M. et al., 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res*, 40(Database issue), pp.D1178-D1186. doi: 10.1093/nar/gkr944

Guiltinan, M.J. et al., 2008. Genomics of *Theobroma cacao*, "the Food of the Gods". In *Genomics of Tropical Crop Plants*. New York, USA: Springer New York, pp.145-170

Guo, M. et al., 2015. Genome-wide analysis of the CaHsp20 gene family in pepper: comprehensive sequence and expression profile analysis under heat stress. *Front Plant Sci*, 6, 806. doi: 10.3389/fpls.2015.00806

Gupta, P. et al., 2017. Abiotic stresses cause differential regulation of alternative splice forms of GATA transcription factor in rice. *Front Plant Sci*, 8, 1944. doi: 10.3389/fpls.2017.01944

Hu, B. et al., 2015. GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*, 31(8), pp.1296-1297. doi:10.1093/bioinformatics/btu817

Jaimez, R.E. et al., 2022. Theobroma cacao L. cultivar CCN 51: a comprehensive review on origin, genetics, sensory properties, production dynamics, and physiological aspects. *PeerJ*, 10, e12676. doi:10.7717/peerj.12676

Jiang, L. et al., 2020. Identification, phylogenetic evolution and expression analysis of GATA transcription factor family in maize (*Zea mays*). *Int J Agric Biol*, 23(3), pp. 637-643. doi: 10.17957/IJAB/15.1334

Jin, J. et al., 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res*, 45 (D1), pp.D1040-D1045. doi: 10.1093/nar/gkw982

Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4), pp.772-780. doi: 10.1093/molbev/mst010

Kim, M. et al., 2021a. Genome-wide comparative analyses of GATA transcription factors among 19 Arabidopsis ecotype genomes: Intraspecific characteristics of GATA transcription factors. *PLoS One*, 16 (5), e0252181. doi: 10.1371/journal.pone.0252181

Kim, M. et al., 2021b. Genome-wide comparative analyses of GATA transcription factors among seven Populus genomes. *Sci Rep*, 11(1), 16578. doi: 10.1038/s41598-021-95940-5

Lai, D. et al., 2022. Genome-wide identification, phylogenetic and expression pattern analysis of GATA family genes in foxtail millet (*Setaria italica*). *BMC Genomics*, 23(1), 549. doi: 10.1186/s12864-022-08786-0

Le, T.M. et al., 2023. Comprehensive characterization and expression profiling of the GATA transcription factor in sugar beet (*Beta vulgaris* L.) suggests their potential roles in taproot development and biotic stress response. *Jordan J Biol Sci*, 16(4), pp.611-619. doi: 10.54319/jjbs/160406

Li, X. et al., 2023. Genome-wide identification and expression analysis of GATA gene family under different nitrogen levels in *Arachis hypogaea* L. *Agron*, 13(1), 215. doi: 10.3390/agronomy13010215

Liu, P.P. et al., 2005. The BME3 (Blue Micropylar End 3) GATA zinc finger transcription factor is a positive regulator of Arabidopsis seed germination. *Plant J*, 44(6), pp.960-971. doi: 10.1111/j.1365-313X.2005.02588.x

Liu, X. et al., 2020. The wheat LLM-domain-containing transcription factor TaGATA1 positively modulates host immune response to Rhizoctonia cerealis. *J Exp Bot*, 71(1), pp.344-355. doi: 10.1093/jxb/erz409

Manzoor, M.A. et al., 2021. Comprehensive comparative analysis of the GATA transcription factors in four Rosaceae species and phytohormonal response in Chinese pear (*Pyrus bretschneideri*) fruit. *Int J Mol Sci*, 22(22), 12492. doi: 10.3390/ijms222212492

Maximova, S.N. et al., 2014. Genome-wide analysis reveals divergent patterns of gene expression during zygotic and somatic embryo maturation of *Theobroma cacao* L., the chocolate tree. *BMC Plant Biol*, 14, 185. doi: 10.1186/1471-2229-14-185

Merika, M., & Orkin, S.H., 1993. DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol*, 13(7), pp.3999-4010. doi: 10.1128/mcb.13.7.3999-4010.1993

Mistry, J. et al., 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res*, 49(D1), pp.D412-D419. doi: 10.1093/nar/gkaa913

Motamayor, J.C. et al., 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol*, 14(6), r53. doi: 10.1186/gb-2013-14-6-r53

Motamayor, J.C. et al., 2002. Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity (Edinb)*, 89(5), pp.380-386. doi: 10.1038/sj.hdy.6800156

Nawy, T. et al., 2010. The GATA factor HANABA TARANU is required to position the proembryo boundary in the early Arabidopsis embryo. *Dev Cell*, 19(1), pp.103-113. doi: 10.1016/j.devcel.2010.06.004

Niu, L. et al., 2020. The GATA gene family in chickpea: Structure analysis and transcriptional responses to abscisic acid and dehydration treatments revealed potential genes involved in drought adaptation. *J Plant Growth Regul*, 39(4), pp.1647-1660. doi: 10.1007/s00344-020-10201-5

Peng, W. et al., 2021. Genome-wide characterization, evolution, and expression profile analysis of GATA transcription factors in *Brachypodium distachyon*. *Int J Mol Sci*, 22(4), 2026. doi: 10.3390/ijms22042026

Pinheiro, T.T. et al., 2011. Establishing references for gene expression analyses by RT-qPCR in Theobroma cacao tissues. *Genet Mol Res*, 10(4), pp.3291-3305. doi: 10.4238/2011.November.17.4

Pokou, D.N. et al., 2019. Resistant and susceptible cacao genotypes exhibit defense gene polymorphism and unique early responses to Phytophthora megakarya inoculation. *Plant Mol Biol*, 99(4-5), pp.499-516. doi: 10.1007/s11103-019-00832-y

Pucciarelli, D.L., 2013. Cocoa and heart health: a historical review of the science. *Nutrients*, 5(10), pp.3854-3870. doi: 10.3390/nu5103854

Reyes, J.C. et al., 2004. The GATA family of transcription factors in Arabidopsis and rice. *Plant Physiol*, 134(4), pp.1718-1732. doi: 10.1104/pp.103.037788

Rozas, J. et al., 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol*, 34(12), pp.3299-3302. doi: 10.1093/molbev/msx248

Schwechheimer, C. et al., 2022. Plant GATA factors: Their biology, phylogeny, and phylogenomics. *Annu Rev Plant Biol*, 73, pp.123-148. doi: 10.1146/annurev-arplant-072221-092913

Shabbir, R. et al., 2022. Combined abiotic stresses: Challenges and potential for crop improvement. *Agron*, 12(11), 2795. doi: 10.3390/agronomy12112795

Shi, M. et al., 2022. Genome-wide survey of the GATA gene family in camptothecin-producing plant *Ophiorrhiza pumila*. *BMC Genomics*, 23, 256. doi: 10.1186/s12864-022-08484-x

Tamura, K. et al., 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol*, 38(7), pp.3022-3027. doi: 10.1093/molbev/msab120

Tan, T.Y.C et al., 2021. The health effects of chocolate and cocoa: A systematic review. *Nutrients*, 13(9), 2909. doi: 10.3390/nu13092909

Teakle, G.R. et al., 2002. *Arabidopsis thaliana* GATA factors: organisation, expression and DNA-binding characteristics. *Plant Mol Biol*, 50(1), pp.43-57. doi: 10.1023/a:1016062325584

Wang, Q. et al., 2021. Genome-wide survey and characterization of AC-D6-like genes in leguminous plants. *Biologia*, 76(10), pp.3137-3147. doi: 10.1007/s11756-021-00829-3

Wickramasuriya, A.M., & Dunwell, J.M., 2018. Cacao biotechnology: current status and future prospects. *Plant Biotechnol J*, 16(1), pp.4-17. doi:10.1111/pbi.12848

Yao, S. et al., 2021. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res*, 49(W1), pp.W469-W475. doi: 10.1093/nar/gkab398

Yu, C. et al., 2021. Genome-wide identification and function characterization of GATA transcription factors during development and in response to abiotic stresses and hormone treatments in pepper. *J Appl Genet*, 62(2), pp.265-280. doi: 10.1007/s13353-021-00618-3

Yu, R. et al., 2021. Genome-wide identification of the GATA gene family in potato (*Solanum tuberosum* L.) and expression analysis. *J Plant Biotech Biochem*, 31(1), pp.37-48. doi: 10.1007/s13562-021-00652-6

Yuan, Q. et al., 2018. A Genome-wide analysis of GATA transcription factor family in tomato and analysis of expression patterns. Int *J Agricul Biol*, 20(6), pp. 1274-1282. doi: 10.17957/IJAB/15.0626

Zhang, C. et al., 2015. Genome-wide survey of the soybean GATA transcription factor gene family and expression analysis under low nitrogen stress. *PLoS One*, 10(4), e0125174. doi: 10.1371/journal.pone.0125174

Zhang, H. et al., 2021. OsGATA16, a GATA transcription factor, confers cold tolerance by repressing OsWRKY45–1 at the seedling stage in rice. *Rice*, 14(1), 42. doi: 10.1186/s12284-021-00485-w

Zhang, K. et al., 2021. Genome- wide identification, phylogenetic and expression pattern analysis of GATA family genes in cucumber (*Cucumis sativus* L.). *Plants (Basel)*, 10(8). doi: 10.3390/plants10081626

Zhang, Z. et al., 2018. Characterization of the GATA gene family in *Vitis vinifera*: genome-wide analysis, expression profiles, and involvement in light and phytohormone response. *Genome*, 61(10), pp.713-723. doi: 10.1139/gen-2018-0042

Zhang, Z. et al., 2019. Genome-wide identification and analysis of the evolution and expression patterns of the GATA transcription factors in three species of *Gossypium* genus. *Gene*, 680, pp.72-83. doi: 10.1016/j.gene.2018.09.039

Zhao, T. et al., 2021. Overexpression of SlGATA17 promotes drought tolerance in transgenic tomato plants by enhancing activation of the phenylpropanoid biosynthetic pathway. *Front Plant Sci*, 12, 634888. doi: 10.3389/fpls.2021.634888

Zhu, W. et al., 2020. Genome-wide identification, phylogenetic and expression pattern analysis of GATA family genes in Brassica napus. *BMC Plant Biol*, 20(1), 543. doi: 10.1186/s12870-020-02752-2

## APPENDICES

**Table S1.** Number of GATA genes in each group of some plant species used in genome-wide identification of the GATA gene family.

| Plant genome names | Number of each group of GATA genes | | | | | | | Total | Ref. |
|---|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI | VII | | |
| *Arabidopsis thaliana* | 14 | 11* | 3 | 2 | 0 | 0 | 0 | 30 | (Reyes et al. 2004) |
| *Arachis hypogaea* | 26 | 13 | 6 | 0 | 0 | 0 | 0 | 45 | (Li et al. 2023) |
| *Brassica napus* | 36 | 43 | 10 | 7 | 0 | 0 | 0 | 96 | (Zhu et al. 2020) |
| *Glycine max* | 30 | 17 | 9 | 8 | 0 | 0 | 0 | 64 | (Zhang et al. 2015) |
| *Gossypium arboreum* | 20 | 13 | 8 | 5 | 0 | 0 | 0 | 46 | (Zhang et al. 2019) |
| *Gossypium hirsutum* | 36 | 25 | 16 | 10 | 0 | 0 | 0 | 87 | (Zhang et al. 2019) |
| *Gossypium raimondii* | 19 | 14 | 8 | 5 | 0 | 0 | 0 | 46 | (Zhang et al. 2019) |
| *Malus domestica* | 20 | 8 | 4 | 3 | 0 | 0 | 0 | 35 | (Chen et al. 2017) |
| *Populus deltoides* | 18 | 9 | 9 | 2 | 0 | 0 | 0 | 38 | (Kim et al. 2021b) |
| *Populus euphratica* | 18 | 12 | 8 | 2 | 0 | 0 | 0 | 40 | (Kim et al. 2021b) |
| *Populus pruinosa* | 17 | 11 | 7 | 2 | 0 | 0 | 0 | 37 | (Kim et al. 2021b) |
| *Populus tremula* | 17 | 7 | 8 | 1 | 0 | 0 | 0 | 33 | (Kim et al. 2021b) |
| *Populus tremula x alba* | 18 | 10 | 8 | 2 | 0 | 0 | 0 | 38 | (Kim et al. 2021b) |
| *Populus tremuloides* | 17 | 7 | 9 | 2 | 0 | 0 | 0 | 35 | (Kim et al. 2021b) |
| *Populus trichocarpa* | 18 | 10 | 9 | 2 | 0 | 0 | 0 | 39 | (Kim et al. 2021b) |
| *Populus trichocarpa* | 18 | 10 | 9 | 2 | 0 | 0 | 0 | 39 | (Apuli et al. 2020) |
| *Prunus avium* | 6 | 7 | 4 | 1 | 0 | 0 | 0 | 18 | (Manzoor et al. 2021) |
| *Prunus mume* | 7 | 7 | 5 | 1 | 0 | 0 | 0 | 20 | (Manzoor et al. 2021) |
| *Prunus persica* | 8 | 6 | 6 | 2 | 0 | 0 | 0 | 22 | (Manzoor et al. 2021) |
| *Pyrus bretschneideri* | 13 | 11 | 6 | 2 | 0 | 0 | 0 | 32 | (Manzoor et al. 2021) |
| *Ricinus communis* | 7 | 7 | 4 | 1 | 0 | 0 | 0 | 19 | (Ao et al. 2015) |
| *Solanum lycopersicum* | 14 | 9 | 4 | 3 | 0 | 0 | 0 | 30 | (Yuan et al. 2018) |
| *Solanum tuberosum* | 3 | 15 | 8 | 13 | 10 | 0 | 0 | 49 | (R Yu et al. 2021) |
| *Ophiorrhiza pumila* | 7 | 5 | 5 | 1 | 0 | 0 | 0 | 18 | (Shi et al. 2022) |
| *Vitis vinifera* | 7 | 6 | 5 | 1 | 0 | 0 | 0 | 19 | (Zhang et al. 2018) |
| *Oryza sativa* | 8 | 9 | 3 | 1 | 2 | 4 | 2 | 29 | (Gupta et al. 2017) |
| *Phyllostachys edulis* | 12 | 13 | 6 | 0 | 0 | 0 | 0 | 31 | (Wang et al. 2020) |
| *Setaria italica* | 14 | 8 | 4 | 2 | 0 | 0 | 0 | 28 | (Lai et al. 2022) |
| *Triticum aestivum* | 21 | 23 | 31 | 4 | 0 | 0 | 0 | 79 | (Du et al. 2022) |
| *Triticum aestivum* | 35 | 21 | 12 | 4 | 0 | 0 | 0 | 79 | (Feng et al. 2022) |
| **Total** | **483** | **342** | **217** | **85** | **2** | **4** | **2** | **1131** | |

*This number is from the recent analysis (Kim et al. 2021b).

***In the case of four species, different classification, group A, B, C, and/or D, was used so that it is also omitted (group A: 15 GATA genes, group B: 5 GATA genes, group C: 7 GATA genes, and group D: 1 GATA genes in *Brachypodium distachyon* (Peng et al. 2021), group A: 12 genes, group B: 9 genes, group C: 4 genes, and group D: 3 genes in *Capsicum annuum* (C Yu et al. 2021), group A: 17 GATA genes, group B: 5 GATA genes, and group C: 3 GATA genes in *Cicer arietinum* (Niu et al. 2020), group A: 11 genes, group B: 9 genes, group C: 4 genes, and group D: 2 genes in *Cucumis sativus* (Zhang et al. 2021)).

In case of *Zea mays*, different classification, goup I, II, III, IV, V, and VI, was used so that it is also omitted (group I: 5 genes, group II: 0 gene, goup III: 7 genes, goup IV: 3 genes, group V: 5 genes, group VI: 3 genes (Jiang et al. 2020)