Journal of Animal & Plant Sciences, 34(2): 2024, Page: 325-335
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2024.2.0720

GENOME-WIDE IDENTIFICATION AND CHARACTERIZATION OF THE GATA TRANSCRIPTION FACTOR FAMILY SUGGESTS FUNCTIONAL EXPRESSION PATTERN AGAINST VARIOUS ENVIRONMENTAL CONDITIONS IN CASSAVA (*Manihot esculenta*)

T. V. Tien[1], V. H. La[2], N. Q. Trung[3], P. C. Thuong[4], B. T. T. Huong[3], L. V. Nguyen[3], D. H. Gioi[3], Q. T. N. Le[5], H. Thi T. Tran[6], H. D. Chu[7,*] and P. B. Cao[8,*]

[1] National Academy of Public Administration, Nguyen Chi Thanh Road, Dong Da District, Hanoi City 122300, Vietnam
[2] Hanoi Pedagogical University 2, Phuc Yen City, Vinh Phuc Province 280000, Vietnam
[3] Vietnam National University of Agriculture, Ngo Xuan Quang Road, Gia Lam District, Hanoi City 122300, Vietnam
[4]Ministry of Science and Technology, Tran Duy Hung Road, Cau Giay District, Hanoi City 122300, Vietnam
[5] Thuyloi University, Dong Da District, Hanoi City 122300, Vietnam
[6]Hanoi National University of Education, Xuan Thuy Road, Cau Giay District, Hanoi City 122300, Vietnam
[7] University of Engineering and Technology, Vietnam National University Hanoi, Xuan Thuy Road, Cau Giay District, Hanoi City 122300, Vietnam
[8] Hung Vuong University, Nguyen Tat Thanh Street, Nong Trang Ward, Viet Tri City, Phu Tho Province 35000, Vietnam
[*]Corresponding author: Ha Duc Chu; Email: cd.ha@vnu.edu.vn and Phi Bang Cao; Email: phibang.cao@hvu.edu.vn

ABSTRACT

GATA transcription factors (TFs) play a significant role in regulating many plant physiological processes. The GATA TF family has been identified and characterized in many important crop species. However, no information is available on the GATA TFs in cassava (*Manihot esculenta*). In this study, 36 *MeGATA* genes have been comprehensively identified, annotated, and characterized in the cassava genome using various bioinformatics tools. The gene structure and duplication of the *MeGATA* genes indicated the redundancy and differences in their gene structural organization. The GATA TFs in cassava could divide into three different groups, as in other plant species. Interestingly, the expression levels of the *MeGATA* genes were significantly changed in various major organs/tissues in the growth and development, especially in response to adverse environmental conditions. Taken together, this study could propose a list of candidate genes for further functional characterization of stress-inducible *MeGATA* genes in cassava.

Keywords: GATA, transcription factor, identification, gene duplication, cassava, expression, characterization.

Published first online January 29, 2024 Published final March 31, 2024

INTRODUCTION

Cassava (*Manihot esculenta*) has been regarded as one of the most important cash crops that are primarily grown in Asia, Africa, and tropical America (Malik *et al.*, 2020). As containing a high starch percentage in storage root, cassava is a primary food source for at least 750 million people (De Souza *et al.*, 2017). This tube crop can also be processed into starch, flour, and alcohol for daily use in food or feed (Malik *et al.*, 2020). However, adverse environmental conditions caused by climate change, such as drought, salt, and heavy metal stress (considered abiotic stress), and cassava brown strike disease (CBSD) (Tomlinson *et al.*, 2018) (considered biotic stresses) are reported as the main factor significantly affecting the growth, development, and productivity of cassava. Thus, understanding the gene regulation related to the defense mechanism in cassava plants plays a crucial role in improving cassava stress tolerance.

It is now well-established that stress tolerance is regulated by some specific genes, such as genes encoding functional and regulatory proteins, particularly transcription factors (TFs), enzymes, chaperones, and metabolites that enable plants to withstand adverse environmental conditions (Agarwal and Jha, 2010; Lindemose *et al.*, 2013; Reddy *et al.*, 2013). Particularly, TFs have been described to regulate gene expression by specifically interacting with *cis*- regulatory elements from the promoter of the targeted genes. Among them, GATA TFs, a group of type IV zinc-finger proteins (Behringer and Schwechheimer, 2015), which specifically bind to the DNA sequence -(A/T)GATA(A/G)- and act as regulators of gene expression (Schwechheimer *et al.*,

2022). This TF has been implicated in the regulation of the development of major organs, including leaves, roots, and flowers (Schwechheimer *et al.*, 2022). Up till now, the GATA TFs have been reported in various higher plant species, including *Arabidopsis thaliana* (Teakle *et al.*, 2002), rice (*Oryza sativa*) (Reyes *et al.*, 2004), soybean (*Glycine max*) (Zhang *et al.*, 2015), apple (*Malus domestica*) (Chen *et al.*, 2017), grape (*Vitis vinifera*) (Zhang *et al.*, 2018), chickpea (*Cicer arietinum*) (Niu *et al.*, 2020) and potato (*Solanum tuberosum*) (Yu *et al.*, 2022). However, the information on the GATA TFs in cassava is still lacking, even though the assembly of this important crop has been released recently (Bredeson *et al.*, 2016).

The purpose of this present study was to comprehensively analyze the GATA TFs in cassava by computational approaches. Firstly, all putative members of the GATA TFs were identified and annotated from the recent assembly of cassava. The major characteristics of the GATA TFs were then analyzed by using various web-based tools. Finally, expression patterns of genes encoding GATA TFs in major organs under various conditions were explored by re-analyzing the previous transcriptome databases.

MATERIALS AND METHODS

Identification and annotation of the GATA TFs: To seek the GATA TFs from the assembly of cassava (Bredeson *et al.*, 2016) and the PlantTFDB platform (Jin *et al.*, 2016) was used to screen all potential members of the GATA TFs. The presence of the conserved domain of GATA TFs (Schwechheimer *et al.*, 2022) was confirmed by screening against the Pfam database (Mistry *et al.*, 2021). All potential members of the GATA TFs were then BlastP-ed against the assembly of cassava (Bredeson *et al.*, 2016) from NCBI and Phytozome (Goodstein *et al.*, 2012) to annotate their identifiers, such as GeneID, ProteinID, and LocusID and seek their sequences, including coding DNA sequence (CDSs), genomic DNA sequence (gDNA) and full-length protein sequence, and chromosomal locations for further analyses.

Prediction of the duplication events of genes encoding the GATA TFs: To analyze the gene duplication, a comprehensive comparison between genes encoding the GATA TFs was carried out as previously described (La *et al.*, 2022; Niu *et al.*, 2020). Particularly, the CDSs of all identified genes encoding GATA TFs were aligned by using ClustalX (Larkin *et al.*, 2007). The identity matrix between these genes was then generated by BioEDIT (Hall, 1999). A duplicated pair was defined with the standard of more than 70% identity (La *et al.*, 2022). The rate of non-synonymous substitutions per non-synonymous site (Ka) and synonymous substitutions per synonymous site (Ks) of each pair were measured by using the DnaSP (Rozas *et al.*, 2017).

Analysis of features of the GATA TFs: To calculate the characteristics of the GATA TFs, the full-length protein sequence of each member was used to apply in the Expasy Protparam (Gasteiger *et al.*, 2005) as following the previous study (Niu *et al.*, 2020). Particularly, several properties of each protein molecule, including protein size (amino acid residues), molecular mass (kilo Dalton, kDa), iso-electric point (pI), instability index (II), aliphatic index (AI), and grand average of hydropathy (GRAVY) were explored (Gasteiger *et al.*, 2005).

Investigation of the subcellular localization of the GATA TFs: To predict the subcellular localization of the GATA TFs, full-length protein sequences of all proteins were used to apply the YLoc tool (Briesemeister *et al.*, 2010) as previously described (La *et al.*, 2022). Particularly, the signal peptide from the full-length protein sequence was mapped into 10 locations in the cell of the plant model, including the nucleus, cytoplasm, mitochondrion, plasma membrane, extracellular space, endoplasmic reticulum, peroxisome, Golgi apparatus, vacuole and chloroplast (Briesemeister *et al.*, 2010).

Construction of the phylogenetic tree of the GATA TFs: To investigate the relationship of members of the GATA TFs, a phylogenetic tree was constructed by using the Molecular Evolutionary Genetics Analysis (MEGA) software (Kumar *et al.*, 2016) as previously reported (Chu *et al.*, 2018; Niu *et al.*, 2020). Particularly, all full-length protein sequences were subjected to the software to generate a Neighbor-Joining tree with 1,000 bootstrap repeats. Other default parameters, such as the model, inter-site ratio, and gap deletion data processing were designed as P-distance, uniform ratio, and partial deletion, respectively (Kumar *et al.*, 2016).

Structural analysis of genes encoding the GATA TFs: To analyze the structure of genes encoding the GATA TFs, the organization of the exon and intron of each gene was explored as previously described (Niu *et al.*, 2020). Briefly, the lengths of gDNA and CDS of each gene were calculated by using the BioEDIT software (Hall, 1999). The gDNA and CDS were then subjected to the Gene Structure Display Server (GSDS) (Hu *et al.*, 2015) to construct the exon/intron structure.

Re-analysis of expression patterns of genes encoding the GATA TFs: To investigate the expression profiles of genes encoding the GATA TFs in cassava, previously reported microarray databases available in the NCBI GEO (Barrett *et al.*, 2013) were used to comprehensively analyze. Briefly, the expression patterns of genes encoding the GATA TFs were explored in 11 major organs/tissues, including leaf blade, leaf mid vein, petiole, stem, lateral bud, shoot apical meristem (SAM),

storage root, fibrous root, root apical meristem (RAM), organized embryogenic structure (OES) and friable embryogenic callus (FEC) by retrieving the GSE82279 dataset as previously provided (Wilson *et al.*, 2017). The fragments per kilobase of transcript per million reads mapped (FPKM) value was used to represent the expression profile of each gene (Wilson *et al.*, 2017). Next, three transcriptome atlas, related to biotic stress, particularly CBSD inoculation (GSE56467) as previously reported (Maruthi *et al.*, 2014), and abiotic stress conditions, including polyethylene glycol 6000 treatment (GSE93098) (Ding *et al.*, 2017) and drought stress (GSE98537) (Zhu *et al.*, 2020) were re-analyzed. Finally, the heatmaps of the *GATA* gene's expression were then clustered by R script.

RESULTS AND DISCUSSION

Genome-wide survey and expansion of the GATA TFs in cassava: As a result, a total of 36 putative members of the GATA TFs was identified in the cassava assembly (Table 1). These members were then annotated as the MeGATAs, and 36 MeGATAs were renamed from MeGATA1 to 36 based on their positions in the chromosomes (Figure 1). The chromosomal distribution of all 36 members of the GATA TFs in cassava was described in Figure 1. Previously, the GATA TFs have been screened in several plant species, such as *A. thaliana* (Teakle *et al.*, 2002), rice (Reyes *et al.*, 2004), soybean (Zhang *et al.*, 2015), chickpea (Niu *et al.*, 2020) and potato (Yu *et al.*, 2022). For example, 29 and 28 members of the GATA TFs have been studied in *A. thaliana* and rice, respectively (Reyes *et al.*, 2004; Teakle *et al.*, 2002). In soybean, the GATA TFs contained 64 members (Zhang *et al.*, 2015), while 19 and 35 GATA TFs were identified and characterized in grape (Zhang *et al.*, 2018) and apple (Chen *et al.*, 2017), respectively. Additionally, 25 members of the GATA TFs have been reported in chickpea (Niu *et al.*, 2020). Recently, the GATA TFs, with 49 members have been identified in potato (Yu *et al.*, 2022). This present study screened 36 members of the GATA TFs in cassava, which were assigned as MeGATA01 to MeGATA36 according to the selected order (Table 1, Figure 1), lower than potato (49 members) (Yu *et al.*, 2022) and soybean (64 members) (Zhang *et al.*, 2015), but higher than grape (19 members) (Z. Zhang *et al.*, 2018), chickpea (25 members) (Niu *et al.*, 2020), rice (28 members) (Reyes *et al.*, 2004), *A. thaliana* (29 members) (Teakle *et al.*, 2002) and apple (35 members) (Chen *et al.*, 2017).

To explain the expansion of genes encoding GATA TFs in cassava, the gene duplication was predicted based on the similarity of their corresponding CDSs. As expected, among 36 *MeGATA* genes, a total of 10 duplication events (with 20 duplicated genes) was found by using various tools as previously described (La *et al.*, 2022; Niu *et al.*, 2020). The similarity of the duplicated *MeGATA* genes was varied from 72.2 (*MeGATA11* and *34*) to 90.1% (*MeGATA03* and *04*) (Table 2). All duplicated *MeGATA* pairs have been produced by segmental duplication events (Table 2, Figure 1). Particularly, two duplicated pairs, *MeGATA02* and *06*, and *MeGATA03* and *04* have occurred from chromosomes 1 and 2, respectively (Table 2, Figure 1), while two (*MeGATA07* and *29*, *MeGATA08* and *31*) and two (*MeGATA10* and *32*, *MeGATA11* and *34*) were found to localize on chromosomes 3 and 15, and chromosomes 3 and 16, respectively (Table 2, Figure 1). Next, two pairs, namely *MeGATA17* and *26*, and *MeGATA20* and *25* were distributed on chromosomes 7 and 10, respectively (Table 2, Figure 1). Two remaining duplicated pairs, namely *MeGATA12* and *28*, and *MeGATA15* and *36* were mapped on chromosomes 4 and 11, and chromosomes 5 and 18, respectively (Table 2, Figure 1).

Next, in order to predict the natural pressure acting on the *MeGATA* genes during evolution, the Ka and Ks values of 10 duplicated pairs were estimated by using DnaSP software (Rozas *et al.*, 2017) according to previous guided (La *et al.*, 2022; Niu *et al.*, 2020). According to Table 2, the Ka/Ks rate of all duplicated *MeGATA* genes from cassava was found to range from 0.32 (*MeGATA03* and *04*) to 1.11 (*MeGATA02* and *06*). It is indicated that the Ka/Ks ratio of a majority of the duplicated pairs (eight out of 10) is less than 1.0, suggesting that the *MeGATA* genes may undergo strong purifying selection pressure during evolution.

Previously, the duplication events were also predicted in genes encoding GATA TFs from higher plant species. For example, eight duplication events (18 duplicated genes) were found as segmental duplications in *CaGATA* genes in chickpea (Niu *et al.*, 2020). In soybean, 23 (out of 25) duplicated pairs of *GmGATA* genes were localized in segmental duplication blocks (Zhang *et al.*, 2015). Recently, 16 duplicated *StGATA* genes were detected to be involved in duplicated genomic blocks in potato (Yu *et al.*, 2022). Taken together, these findings strongly hypothesized that segmental duplication events might play a key role in the expansion of genes encoding GATA TFs in cassava, perhaps in higher plant species.

Analysis of protein features and subcellular localization of the GATA TFs in cassava: Table 1 summarized six features (including size, mass, pI, II, AI, and GRAVY) of 36 members of the GATA TFs in cassava. Among them, MeGATA21 was found as the smallest member of the GATA TFs in cassava (with 106 amino acid residues, while the largest member was MeGATA03 (544 amino acid residues) (Table 1). The molecular mass of the GATA TFs in cassava varied from 12.17 (MeGATA21) to 60.49 kDa (MeGATA03)

Table 1. Summary of the MeGATA TFs in cassava.

STT	Tên gene	GeneID	ProteinID	LocusID	Size (aa)	Mass (kDa)	pI	II	AI	GRAVY	SL
1	*MeGATA01*	Manes.01G224400	XP_021629041	LOC110627103	135	14.99	9.69	61.44	74.52	-0.59	Nucleus
2	*MeGATA02*	Manes.01G135900	XP_021629602	LOC110627559	248	27.74	8.51	68.24	40.60	-0.83	Cytoplasm
3	*MeGATA03*	Manes.01G087500	XP_021632304	LOC110629577	544	60.49	6.58	53.70	70.57	-0.66	Nucleus
4	*MeGATA04*	Manes.02G044400	XP_021604406	LOC110609260	542	60.03	6.55	54.21	69.94	-0.67	Nucleus
5	*MeGATA05*	Manes.02G099500	XP_021602140	LOC110607359	368	39.96	6.04	69.15	61.49	-0.54	Cytoplasm
6	*MeGATA06*	Manes.02G094300	XP_043809693	LOC110608738	217	24.10	9.21	61.30	39.68	-0.83	Cytoplasm
7	*MeGATA07*	Manes.03G201300	XP_021606393	LOC110610666	143	15.60	9.76	61.38	58.04	-1.06	Nucleus
8	*MeGATA08*	Manes.03G154500	XP_021607545	LOC110611478	261	28.96	7.62	56.49	60.15	-0.67	Cytoplasm
9	*MeGATA09*	Manes.03G059100	XP_021608919	LOC110612438	362	39.92	5.88	46.29	57.38	-0.78	Cytoplasm
10	*MeGATA10*	Manes.03G047800	XP_021606819	LOC110611028	297	32.95	8.72	53.11	57.95	-0.78	Cytoplasm
11	*MeGATA11*	Manes.03G033200	XP_021606592	LOC110610821	347	38.98	8.81	55.53	67.00	-0.49	Golgi apparatus
12	*MeGATA12*	Manes.04G132800	XP_021611068	LOC110613946	325	36.35	6.34	55.49	51.94	-0.75	Cytoplasm
13	*MeGATA13*	Manes.04G084400	XP_021611192	LOC110614052	335	37.46	4.95	59.37	65.16	-0.55	Cytoplasm
14	*MeGATA14*	Manes.05G189600	XP_021613255	LOC110615597	285	31.04	6.45	38.43	61.58	-0.70	Nucleus
15	*MeGATA15*	Manes.05G189500	XP_021612757	LOC110615272	355	38.97	4.79	52.33	67.52	-0.67	Nucleus
16	*MeGATA16*	Manes.05G050300	XP_021612232	LOC110614856	263	29.44	7.21	62.72	62.32	-0.67	Nucleus
17	*MeGATA17*	Manes.07G041300	XP_021618775	LOC110619573	297	31.53	5.24	47.08	65.76	-0.61	Nucleus
18	*MeGATA18*	Manes.07G041200	XP_021618597	LOC110619452	364	40.16	4.73	52.66	63.21	-0.77	Nucleus
19	*MeGATA19*	Manes.07G076400	XP_021617953	LOC110618997	271	29.66	8.36	65.25	73.76	-0.50	Nucleus
20	*MeGATA20*	Manes.07G099600	XP_021619011	LOC110619771	351	38.32	6.01	67.74	56.70	-0.63	Cytoplasm
21	*MeGATA21*	Manes.08G113300	XP_043815608	LOC110622414	106	12.17	11.02	65.11	69.81	-0.75	Nucleus
22	*MeGATA22*	Manes.08G149300	XP_021620703	LOC110621019	301	33.59	6.46	59.71	67.41	-0.72	Nucleus
23	*MeGATA23*	Manes.09G174900	XP_021622601	LOC110622414	157	17.65	9.66	68.03	60.32	-0.82	Cytoplasm
24	*MeGATA24*	Manes.09G142600	XP_021622844	LOC110622593	292	32.55	8.37	64.21	64.76	-0.73	Nucleus
25	*MeGATA25*	Manes.10G046800	XP_021626300	LOC110625083	353	38.57	6.66	64.21	63.60	-0.53	Cytoplasm
26	*MeGATA26*	Manes.10G097400	XP_021626705	LOC110625382	304	32.21	5.28	45.96	66.78	-0.57	Nucleus
27	*MeGATA27*	Manes.11G146600	XP_021628794	LOC110626916	315	35.43	8.65	64.06	60.00	-0.66	Cytoplasm
28	*MeGATA28*	Manes.11G034900	XP_021628890	LOC110626985	325	36.00	6.37	53.90	52.89	-0.76	Cytoplasm
29	*MeGATA29*	Manes.15G007100	XP_021593720	LOC110601041	142	15.55	9.79	52.52	59.86	-1.05	Cytoplasm
30	*MeGATA30*	Manes.15G103300	XP_021595107	LOC110601984	225	24.23	7.54	36.11	48.18	-0.88	Cytoplasm
31	*MeGATA31*	Manes.15G049400	XP_021595256	LOC110602133	264	29.58	7.00	64.46	60.19	-0.76	Nucleus
32	*MeGATA32*	Manes.16G080400	XP_021596891	LOC110603465	305	33.80	9.16	67.26	59.28	-0.64	Cytoplasm
33	*MeGATA33*	Manes.16G074900	XP_021596902	LOC110603472	369	40.22	6.00	46.49	58.94	-0.69	Cytoplasm
34	*MeGATA34*	Manes.16G102600	XP_021596550	LOC110603163	299	33.67	8.95	54.16	52.91	-0.81	Cytoplasm
35	*MeGATA35*	Manes.18G056300	XP_021601748	LOC110607013	281	30.49	6.89	37.69	60.75	-0.68	Nucleus
36	*MeGATA36*	Manes.18G056400	XP_021601389	LOC110606735	353	38.58	5.17	46.06	67.65	-0.67	Nucleus

pI - Isoelectric point, II - Instability index, AI - Aliphatic index, GRAVY - Grand average of hydropathy, SL - Subcellular localization

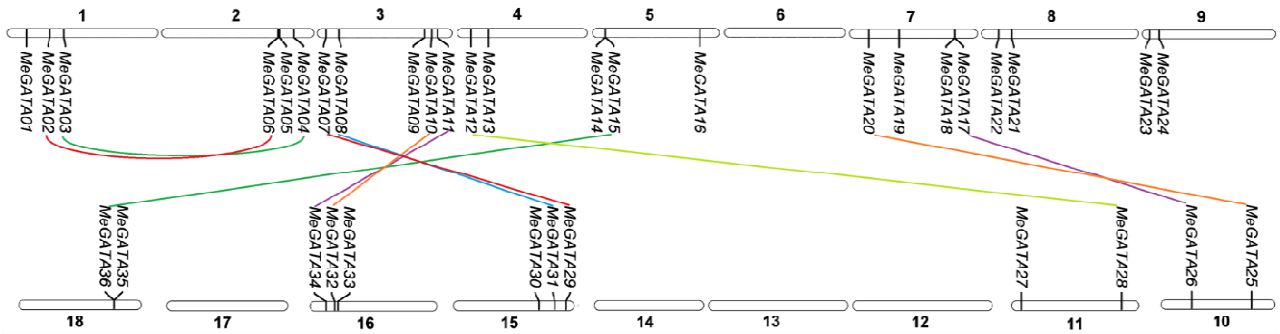

Figure 1. The chromosomal distribution of the *MeGATA* genes in cassava genome.

(Table 1). The pI value of MeGATA21 was 11.02 and MeGATA18 was 4.73, which were the largest and smallest pI values in the GATA TFs in cassava, respectively (Table 1). Additionally, the II and AI scores of the GATA TFs in cassava were found to range from 36.11 (MeGATA30) to 69.15 (MeGATA05) and from 39.68 (MeGATA06) to 74.52 (MeGATA01), respectively (Table 1). The GRAVY value of the GATA TFs in cassava was minus, ranging from -0.49 (MeGATA11) to -1.06 (MeGATA07) (Table 1).

Previously, the general characteristics of the GATA TFs were also comprehensively analyzed in other higher plant species. Zhang *et al.* (2015) revealed that the amino acid residues of the GmGATA proteins in soybean were 80 and the largest was 551, and their masses ranged from 9.1 to 60.8 kDa. The pI values of the GmGATA proteins in soybean varied from 4.63 to 9.66 (Zhang *et al.*, 2015). In grapes, the VvGATA proteins were reported to range from 109 to 386 amino acid residues in size (Zhang *et al.*, 2018). Additionally, the protein sizes of the CaGATA proteins in chickpea were between 133 (14.9 kDa) and 541 amino acid residues (60.2 kDa), and 22 out of 25 CaGATA proteins in chickpea were unstable (II scores were greater than 40) (Niu *et al.*, 2020). The pI scores of the CaGATA proteins in chickpea were varied from 4.27 to 10.27, while the GRAVY scores of all CaGATA proteins in chickpea were less than 0 (Niu *et al.*, 2020). More recently, the length and molecular weight of the StGATA proteins ranged from 118 to 380 amino acid residues and from 13.15 to 60.63 kDa, respectively, while their pI values were found to be great differences (from 4.53 to 10.34) (Yu *et al.*, 2022). Taken together, this study strongly suggested that the GATA TFs in cassava, perhaps in higher plant species exhibited high variation in their general characteristics.

As a part of this study, the subcellular localization of the MeGATA proteins was continued to analyze by using the bioinformatics tool (Briesemeister *et al.*, 2010) as previously reported (La *et al.*, 2022). The prediction of the subcellular localization indicated that most members of the GATA TFs in cassava, specifically 18 and 17 CaGATA proteins were positioned in the cytoplasm and nucleus, respectively, while only MeGATA11 was reported to localize in the Golgi apparatus (Table 1). This finding was also confirmed by a previous study on potato (Yu *et al.*, 2022). Particularly, 28 and 10 (out of 49) members of StGATA proteins were reported to localize in the nucleus and cytoplasm, respectively (Yu *et al.*, 2022).

Classification and structural analysis of the GATA TFs in cassava: The obtained Neighbor-Joining phylogenetic tree showed that the GATA TFs in cassava were clearly divided into three distinct groups, as well-described in Figure 2. Particularly, group 1 comprised three members of the GATA TFs in cassava, namely MeGATA03, 04, and 21, while 10 and 23 members of the GATA TFs in cassava were categorized into groups 2 and 3, respectively (Figure 2).

Previously, a similar phenomenon was also reported for the GATA TFs in other higher plant species. For example, 25 CaGATA proteins in chickpea could be clearly divided into three groups, including group A (17 members), B (five members), and C (three members) (Niu *et al.*, 2020). Moreover, the StGATA proteins in potato were reported to classify into three clades, in which clades 1 and 2 contained three and 13 members, respectively and clade 3 (33 members) contained three subgroups (Yu *et al.*, 2022). The classification into three clades was also observed in other higher plant species, such as *A. thaliana* (Teakle *et al.*, 2002), rice (Reyes *et al.*, 2004), soybean (Zhang *et al.*, 2015), apple (Chen *et al.*, 2017) and grape (Zhang *et al.*, 2018). Next, a structural analysis of genes of the GATA TFs was performed as previously reported (Niu *et al.*, 2020). The CDS lengths of the *MeGATA* genes were varied from 321 (*MeGATA21*) to 6897 bp (*MeGATA03*), while the gDNA lengths of the *MeGATA* genes ranged from 1249 (*MeGATA21*) to 14641 bp (*MeGATA04*) (Figure 2). The amounts of exons of the *MeGATA* genes were reported to be variable, consisting of two to 10 (Figure 2). Particularly, 16 and 10 *MeGATA* genes contained two and three exons, respectively, while five and three *MeGATA* genes consisted of seven and 10 exons (Figure 2). Additionally, only *MeGATA30* and *03* were reported to harbor four and eight exons (Figure 2). Previously, a large number of genes encoding GATA TFs in other

plant species was also indicated to contain two or three exons. Eleven and six (out of 25) *CaGATA* genes in chickpea consisted of two and three exons, respectively (Niu *et al.*, 2020), while 25 (out of 49) *StGATA* genes in potato had two and three exons (Yu *et al.*, 2022).

Expression profiles of the *MeGATA* genes in different tissues during the growth and development: In this study, the FPKM values of the *MeGATA* genes were re-analyzed and constructed a heatmap of the hierarchical clustering to display the expression patterns of the *MeGATA* genes (Figure 3).

Table 2. The information of duplicated events occurred in the *MeGATA* gene family in cassava

Sr. No.	Duplicated genes	Position	Similarity (%)	Ka value	Ks value	Ka/Ks
1	*MeGATA02/06*	Chr1/Chr2	78.0	0.10	0.09	1.11
2	*MeGATA03/04*	Chr1/Chr2	90.1	0.07	0.22	0.32
3	*MeGATA07/29*	Chr3/Chr15	87.7	0.11	0.16	0.68
4	*MeGATA08/31*	Chr3/Chr15	85.3	0.11	0.27	0.41
5	*MeGATA10/32*	Chr3/Chr16	78.7	0.14	0.36	0.39
6	*MeGATA11/34*	Chr3/Chr16	72.2	0.16	0.23	0.70
7	*MeGATA12/28*	Chr4/Chr11	87.3	0.10	0.27	0.37
8	*MeGATA15/36*	Chr5/Chr18	86.8	0.14	0.15	0.93
9	*MeGATA17/26*	Chr7/Chr10	80.7	0.17	0.16	1.06
10	*MeGATA20/25*	Chr7/Chr10	83.7	0.13	0.21	0.62

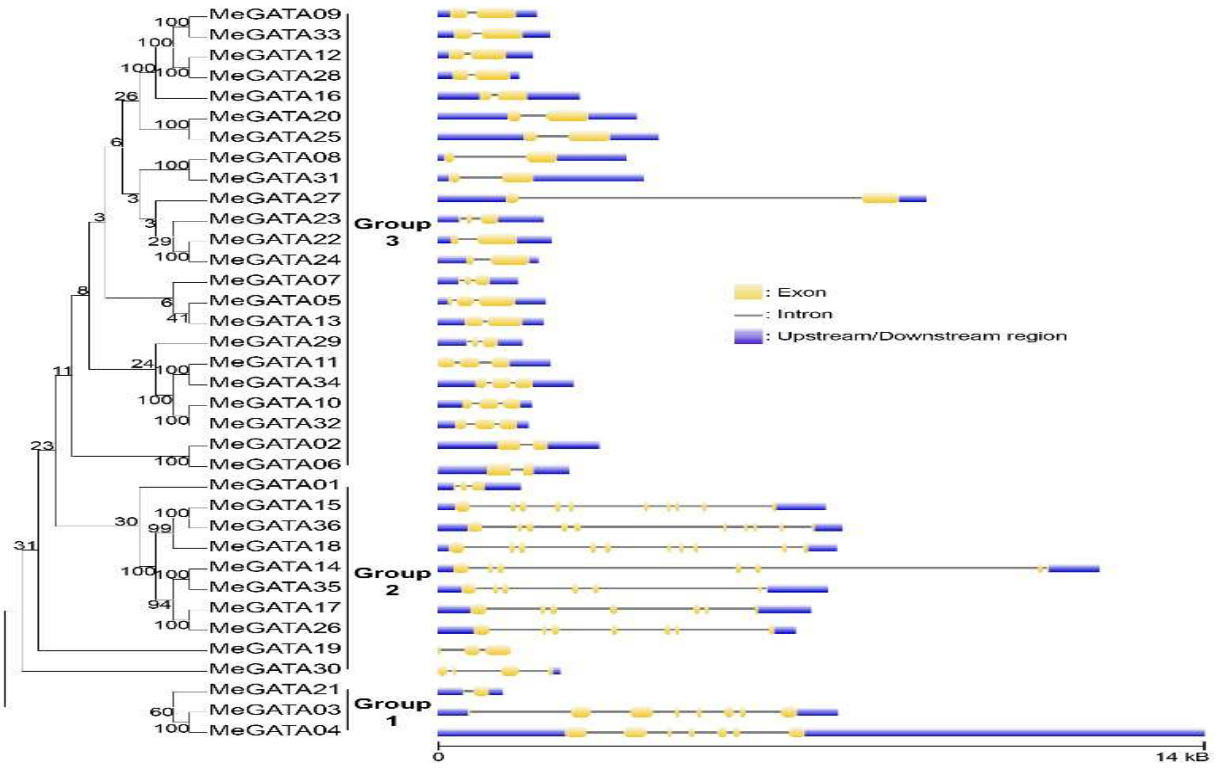

Figure 2. The phylogenetic analysis and structural organization of *MeGATA* TFs in cassava. The exons, introns, and upstream/downstream regions are represented by yellow boxes, black lines, and blue boxes, respectively.

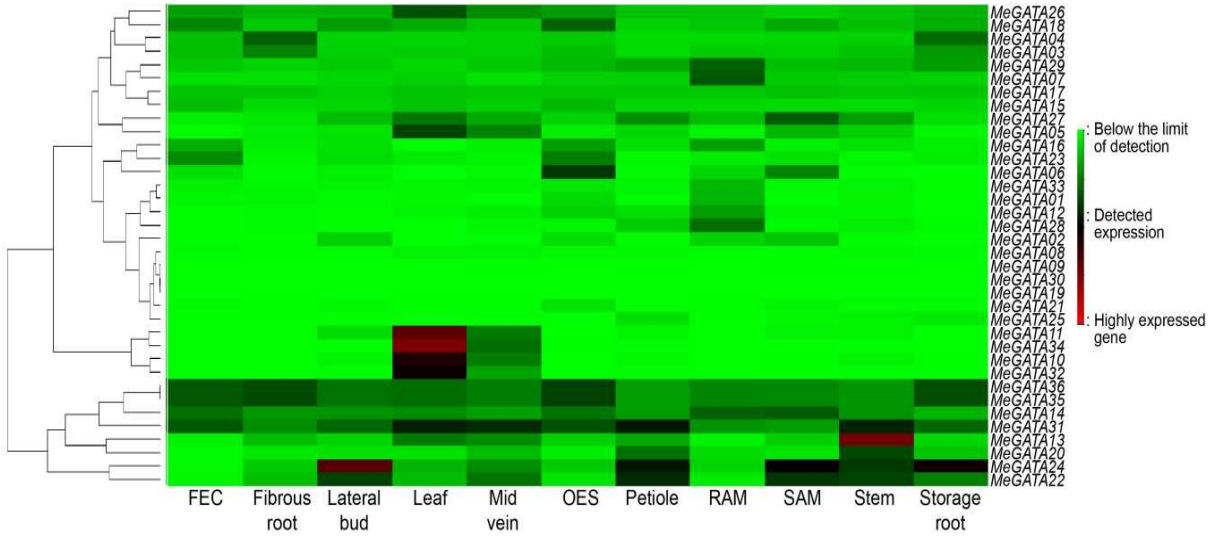

Figure 3. Expression profiles of the *MeGATA* genes in major organs/tissues in the growth and development processes of cassava plants.

This study revealed that the expression of four *MeGATA* genes, including *MeGATA08*, *09*, *19* and *30* was not expressed or low (FPKM values < 10) in any of 11 major organs/tissues, while the remaining *MeGATA* genes (32 out of 36) were expressed (FPKM values ≥ 10) in at least one major organ/tissue (Figure 3). Among them, *MeGATA24* and *13* were noted to be mainly expressed (FPKM values ≥ 100) in lateral bud and stem, respectively, while four *MeGATA* genes, *MeGATA10*, *11*, *32*, and *34* exhibited high transcript abundance in leaf (Figure 3). Interestingly, the expression levels of *MeGATA24* gene tend to be high not only in the petiole and SAM but also in the storage root (Figure 3). This re-analysis suggested that these *MeGATA* genes might play key roles in the tissue development of cassava plants.

Expression analysis of the *MeGATA* genes responding to various stress conditions: To assess the transcript levels of the *MeGATA* genes in major organs/tissues under adverse environmental conditions, the heatmap of the *MeGATA* gene's expression was constructed and provided in Figure 4.

Under drought conditions (Zhu *et al.*, 2020), the expression of 13 and nine *MeGATA* genes was induced and reduced in treated leaf samples (Figure 4). Among them, four genes, namely *MeGATA08*, *10*, *18*, and *23* were highly up-regulated, by 240.75-, 324.13-, 60.31-, and 67.74-fold in drought-treated leaf samples, respectively, whereas *MeGATA33* was noted to be highly down-regulated (-19.45-fold) in treated leaves (Figure 4). Under PEG 6000 treatment (Ding *et al.*, 2017), the expression of 16 *MeGATA* genes was significantly changed in at least one major organ/tissue (Figure 4). For example, one and 11 *MeGATA* genes, including *MeGATA36*, and *MeGATA03*, *13*, *16*, *20*, *23*, *24*, *25*, *27*, *28*, *29*, and *33* were obviously up-regulated and down-regulated in these tested tissues under the PEG 6000 treatment (Figure 4). Interestingly, *MeGATA03* and *20* were reduced in three tissues, including the bottom leaf, root, and folded leaf or and fully expanded leaf under the PEG 6000 treatment, respectively, while three genes, namely *MeGATA24*, *25*, and *28* were down-regulated in two treated tissues, particularly bottom leaf and folded leaf, root and fully expanded leaf, and bottom leaf and root samples, respectively (Figure 4). These findings suggested that these *MeGATA* genes might play an important role in the response to drought or osmotic stresses in cassava.

Furthermore, global cassava production has been critically restricted by CBSD (Tomlinson *et al.*, 2018). One transcriptome atlas of cassava leaf samples after artificial inoculation with CBSD was also explored (Maruthi *et al.*, 2014). The results indicated that five *MeGATA* genes, namely *MeGATA10*, *20*, *25*, *27*, and *33* were down-regulated in treated leaf samples (Figure 5). This study suggested that these five genes might be related to the response to CBSD infection in cassava.

Previously, the GATA TFs have been reported to participate in the response to adverse environmental conditions in plants. For example, *OsGATA16* gene was up-regulated by cold and abscisic treatments but was down-regulated by drought, jasmonic acid, and cytokinin (Zhang *et al.*, 2021). Overexpression of this gene could confer cold tolerance of rice during the seedling period (Zhang *et al.*, 2021). In tomato, overexpression of the *SlGATA17*, a drought-inducible gene could regulate drought resistance by improving the activity of the phenylpropanoid biosynthesis pathway in transgenic

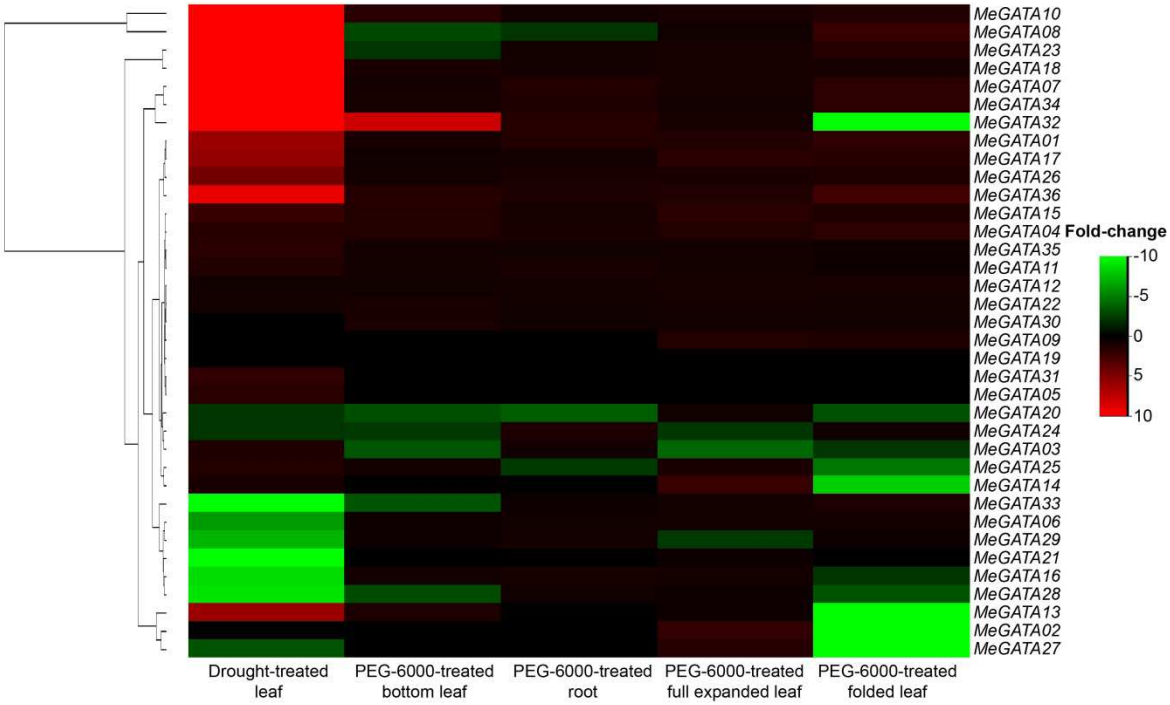

Figure 4. Expression profiles of the *MeGATA* genes in major organs/tissues in the growth and development processes of cassava plants.

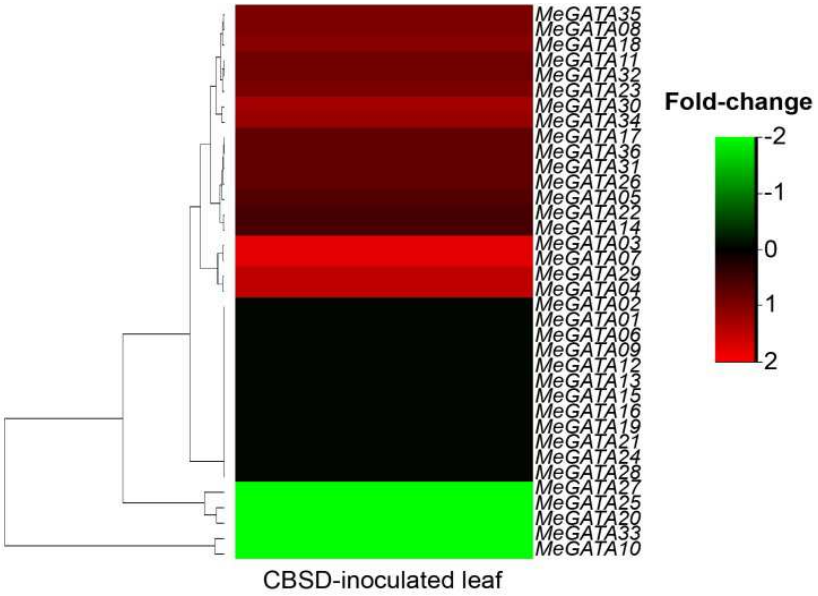

Figure 5. Expression profiles of the *MeGATA* genes in leaf samples under the CBSD inoculation in cassava plants.

plants (Zhao *et al.*, 2021). Overexpression of BdGATA13, a member belonging to the GATA TFs in model grass *Brachpodium distachyon* could enhance drought tolerance by resulting in darker green leaves and later flowering in transgenic *Arabidopsis* plants (Guo *et al.*, 2021). Taken together, the results clearly indicated the function of GATA TFs in plant growth and development processes, especially in the response to adverse environmental stresses.

Conclusions: The present study reported a genome-wide survey and analysis of the MeGATA TF family in cassava, a multi-functional crop in the world. The protein features, gene structures, duplication events, phylogenetic relationship, subcellular localization, and expression profiles of the MeGATA TFs in cassava have been assessed by using bioinformatics tools. These results showed the structural variations in the characteristics of

the MeGATA TFs in cassava. By re-analyzing the previous transcriptome databases, the *MeGATA* genes exhibited differential expression patterns in major organs/tissues in various conditions, more specifically abiotic and biotic stresses. This study could provide a list of potential stress-inducible *MeGATA* genes for further functional characterization.

Authors Contribution: Conceptualization: H.D.C. and P.B.C., Data collection: T.V.T., V.H.L., N.Q.T., P.C.T., B.T.T.H., L.V.N., D.H.G., Q.T.N.L., H.T.T. T., Guidance of data analysis: P.B.C., D.H.C., T.V.T., V.H.L. Manuscript writing: P.B.C, D.H.C, T.V.T, V.H.L. All authors discussed the results and contributed to the final manuscript.

REFERENCES

Agarwal P.K., and B. Jha (2010) Transcription factors in plants and ABA dependent and independent abiotic stress signalling. Biologia Plantarum, 54(2), 201-212. DOI:10.1007/s10535-010-0038-7

Barrett T., S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis and A. Soboleva (2013) NCBI GEO: archive for functional genomics data sets - update. Nucleic Acids Res, 41(Database issue), D991-995. DOI:10.1093/nar/gks1193

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

Bredeson J.V., J.B. Lyons, S.E. Prochnik, GA. Wu, C.M. Ha, E. Edsinger-Gonzales, J. Grimwood, J. Schmutz, I.Y. Rabbi, C. Egesi, P. Nauluvula, V. Lebot, J. Ndunguru, G. Mkamilo, R.S. Bart, T.L. Setter, R.M. Gleadow, P. Kulakow, M.E. Ferguson, S. Rounsley and D.S. Rokhsar (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol, 34*(5), 562-570. DOI:10.1038/nbt.3535

Briesemeister S., J. Rahnenfuhrer and O. Kohlbacher (2010) YLoc--an interpretable web server for predicting subcellular localization. Nucleic Acids Res, 38(Web Server issue), W497-502. DOI:10.1093/nar/gkq477

Chen H., H. Shao, K. Li, D. Zhang, S. Fan, Y. Li and M. Han (2017) Genome-wide identification, evolution, and expression analysis of GATA transcription factors in apple (Malus×domestica Borkh.). Gene, 627, 460-472. DOI:10.1016/j.gene.2017.06.049

Chu H.D., K.H. Nguyen, Y. Watanabe, D.T. Le, T.L.T. Pham, K. Mochida and L.P. Tran (2018) Identification, structural characterization and gene expression analysis of members of the nuclear factor-Y family in chickpea (*Cicer arietinum* L.) under dehydration and abscisic acid treatments. Int J Mol Sci., 19(11), E3290. DOI:10.3390/ijms19113290

De Souza A.P., L.N. Massenburg, D. Jaiswal, S. Cheng, R. Shekar and S.P. Long (2017) Rooting for cassava: insights into photosynthesis and associated physiology as a route to improve yield potential. New Phytologist, 213(1), 50-65. DOI:10.1111/nph.14250

Ding Z., L. Fu, Y. Yan, W. Tie, Z. Xia, W. Wang, M. Peng, W. Hu and J. Zhang (2017) Genome-wide characterization and expression profiling of HD-Zip gene family related to abiotic stress in cassava. PLoS One, 12(3), e0173043. DOI:10.1371/journal.pone.0173043

Gasteiger E., C. Hoogland, A. Gattiker, M.R. Wilkins, R.D. Appel and A. Bairoch (2005) Protein identification and analysis tools on the ExPASy server. In The Proteomics Protocols Handbook (pp. 571-607): Springer.

Goodstein D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam and D.S. Rokhsar (2012) Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res, 40(Database issue), D1178-D1186. DOI:10.1093/nar/gkr944

Guo J., X. Bai, K. Dai, X. Yuan, P. Guo, M. Zhou, W. Shi and C. Hao (2021) Identification of GATA Transcription Factors in Brachypodium distachyon and Functional Characterization of BdGATA13 in Drought Tolerance and Response to Gibberellins. Front Plant Sci, 12. DOI:10.3389/fpls.2021.763665

Hall T.A (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser, 41, 95-98.

Hu B., J. Jin, A.Y. Guo, H. Zhang, J. Luo and G. Gao (2015) GSDS 2.0: An upgraded gene feature visualization server. Bioinformatics, 31(8), 1296-1297. DOI:10.1093/bioinformatics/btu817

Jin J., F. Tian, D.-C. Yang, Y.-Q. Meng, L. Kong, J. Luo and G. Gao (2016) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res, gkw982.

Kumar S., G. Stecher and K. Tamura (2016) MEGA7: Molecular evolutionary genetics analysis version

7.0 for bigger datasets. Mol Biol Evol, 33(7), 1870-1874. DOI:10.1093/molbev/msw054

La H. V., H.D. Chu, Q.T. Ha, T.T.H. Tran, H.V. Tong, T.V. Tran, Q.T.N. Le, H.T.T. Bui, P.B. Cao (2022) SWEET Gene Family in Sugar Beet (*Beta vulgaris*): Genome-Wide Survey, Phylogeny and Expression Analysis. Pakistan J. Biological Sciences: PJBS, 25(5), 387-395. DOI: 10.3923/pjbs.2022.387.395

La H. V., H.D. Chu, C.D. Tran, K.H. Nguyen, Q.T.N. Le, C.M. Hoang, P.B. Cao, A.T.C. Pham, B.D. Nguyen, T.Q. Nguyen, L.V. Nguyen, C.V. Ha, H.T. Le, H.H. Le, T.D. Le and L.-S.P. Tran (2022) Insights into the gene and protein structures of the CaSWEET family members in chickpea (Cicer arietinum), and their gene expression patterns in different organs under various stress and abscisic acid treatments. *Gene, 819*, 146210. DOI:10.1016/ j.gene.2022.146210

Larkin M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson and D.G. Higgins (2007) Clustal W and Clustal X version 2.0. Bioinformatics, *23*(21), 2947-2948. DOI:10.1093/bioinformatics/btm404

Lindemose S., C. O'Shea, M.K. Jensen, and K. Skriver (2013) Structure, function and networks of transcription factors involved in abiotic stress responses. Int J Mol Sci, 14. DOI:10.3390/ijms14035842

Malik A.I., P. Kongsil, V.A. Nguyễn, W. Ou, Sholihin, P. Srean, M.N. Sheela, L.A.B. López-Lavalle, Y. Utsumi, C. Lu, P. Kittipadakul, H.H. Nguyễn, H. Ceballos, T.H. Nguyễn, M.S. Gomez, P. Aiemnaka, R. Labarta, S. Chen, S. Amawan, S. Sok, L. Youabee, M. Seki, H. Tokunaga, W. Wang, K. Li, H.A. Nguyễn, V.Đ. Nguyễn, H.H. Lê and M. Ishitani (2020) Cassava breeding and agronomy in Asia: 50 years of history and future directions. Breed Sci, advpub (2), 145-166. DOI:10.1270/jsbbs.18180

Maruthi M.N., S. Bouvaine, H.A. Tufan, I.U. Mohammed and R.J. Hillocks (2014) Transcriptional response of virus-infected cassava and identification of putative sources of resistance for cassava brown streak disease. PLoS One, 9(5), e96642-e96642. DOI:10.1371/ journal.pone.0096642

Mistry J., S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn and A. Bateman (2021) Pfam: The protein families database in 2021. Nucleic Acids Res, 49(D1), D412-D419. DOI:10.1093/ nar/gkaa913

Niu L., H.D. Chu, C.D. Tran, K.H. Nguyen, H.X. Pham, D.T. Le, W. Li, W. Wang, T.D. Le and L.-S.P. Tran (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 Regulation, 39(4), 1647-1660. DOI:10.1007/s00344-020-10201-5

Reddy D.S., P.B. Mathur, and K.K. Sharma (2013) Regulatory role of transcription factors in abiotic stress responses in plants. In N. Tuteja and S. S. Gill (Eds.), Climate Change and Plant Abiotic Stress Tolerance. Weinheim, Germany: Co. KGaA. DOI:10.1002/9783527675265.ch21

Reyes J.C., M.I. Muro-Pastor and F.J. Florencio (2004) The GATA family of transcription factors in Arabidopsis and rice. Plant Physiol, 134(4), 1718-1732. DOI:10.1104/pp.103.037788

Rozas J., A. Ferrer-Mata, J.C. Sanchez-DelBarrio, S. Guirao-Rico, P. Librado, S.E. Ramos-Onsins and A. Sanchez-Gracia (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol. Biol. Evol., 34(12), 3299-3302. DOI:10.1093/molbev/msx248

Schwechheimer C., P.M. Schröder and C.E. Blaby-Haas (2022) Plant GATA Factors: Their Biology, Phylogeny, and Phylogenomics. Annu Rev Plant Biol. DOI:10.1146/annurev-arplant-072221-092913

Teakle G.R., I.W. Manfield, J.F. Graham and P.M. Gilmartin (2002) Arabidopsis thaliana GATA factors: organisation, expression and DNA-binding characteristics. Plant Mol Biol, *50*(1), 43-57. DOI:10.1023/a:1016062325584

Tomlinson K.R., A.M. Bailey, T. Alicai, S. Seal and G.D. Foster (2018) Cassava brown streak disease: historical timeline, current knowledge and future prospects. Molecular plant pathology, *19*(5), 1282-1294. DOI:10.1111/mpp.12613

Wilson M.C., M.A. Mutka, A.W. Hummel, J. Berry, R.D. Chauhan, A. Vijayaraghavan, N.J. Taylor, D.F. Voytas, D.H. Chitwood and R.S. Bart (2017) Gene expression atlas for the food security crop cassava. New Phytologist, *213*(4), 1632-1641. DOI:10.1111/nph.14443

Yu R., Y. Chang, H. Chen, J. Feng, H. Wang, T. Tian, Y. Song and G. Gao (2022) Genome-wide identification of the GATA gene family in potato (Solanum tuberosum L.) and expression analysis. J. Plant Biochemistry and Biotechnology, 31(1), 37-48. DOI:10.1007/s13562-021-00652-6

Zhang C., Y. Hou, Q. Hao, H. Chen, L. Chen, S. Yuan, Z. Shan, X. Zhang, Z. Yang, D. Qiu, X. Zhou and W. Huang (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., T. Wu, Z. Li, K. Huang, N-E. Kim, Z. Ma, S-W. Kwon, W. Jiang and X. Du (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 Z., C. Ren, L. Zou, Y. Wang, S. Li and Z. Liang (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), 713-723. DOI:10.1139/gen-2018-0042

Zhao T., T. Wu, T. Pei, Z. Wang, H. Yang, J. Jiang, H. Zhang, X. Chen, J. Li and X. Xu (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 Y., X. Luo, M. Wei, A. Khan, F. Munsif, T. Huang, X. Pan and Z. Shan (2020) Antioxidant Enzymatic Activity and Its Related Genes Expression in Cassava Leaves at Different Growth Stages Play Key Roles in Sustaining Yield and Drought Tolerance Under Moisture Stress. J. Plant Growth Regulation, 39(2), 594-607. DOI:10.1007/s00344-019-10003-4.