The GASA Gene Family in *Theobroma cacao*: Genome wide Identification and Expression Analysis

Abdullah^{1,*}, Sahar Faraji², Furrukh Mehmood¹, Hafiz Muhammad Talha Malik³, Ibrar Ahmed³, Parviz Heidari^{*,4}, Peter Poczai^{*,5,6,7}

¹*Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, 45320 Islamabad, Pakistan*

²*Department of Plant Breeding, Faculty of Crop Sciences, Sari Agricultural Sciences and Natural Resources University (SANRU), Sari 4818166996, Iran*

³*Alpha Genomics Private Limited, 45710 Islamabad, Pakistan*

⁴*Faculty of Agriculture, Shahrood University of Technology, 3619995161 Shahrood, Iran*

⁵*Finnish Museum of Natural History, P.O. Box 7, FI-00014 University of Helsinki, Finland*

⁶*Faculty of Biological and Environmental Sciences, PO Box 65 FI-00065 University of Helsinki, Finland*

⁷*Institute of Advanced Studies Kőszeg (iASK), PO Box 4, H-9731 Kőszeg, Hungary*

* Correspondence: Abdullah (abd.ullah@bs.qau.edu.pk)

Parviz Heidari (heidarip@shahroodut.ac.ir)

Péter Poczai (peter.poczai@helsinki.fi)

Abstract

The gibberellic acid-stimulated *Arabidopsis* (*GASA/GAST*) gene family is widely distributed in plants. The role of the *GASA* gene family has been reported previously in various physiological and biological processes, such as cell division, root and seed development, stem growth, and fruit ripening. These genes also provide resistance to abiotic and biotic stresses including antimicrobial, antiviral, and antifungal. Here, we report 17 *tcGASA* genes in *Theobroma cacao* L. distributed on six chromosomes. The gene structure, promoter-region sequences, protein structure, and biochemical properties, expression, and phylogenetics of all *tcGASAs* were analyzed. Phylogenetic analyses divided *tcGASA* proteins into five groups. The nine segmentally duplicating genes form four pairs and cluster together in phylogenetic tree. Purifying selection pressure was recorded on *tcGASA*, including duplicated genes. Several stress/hormone-responsive cis-regulatory elements were also recognized in the promoter region of *tcGASAs*. Differential expression analyses revealed that most of the *tcGASA* genes showed elevated expression in the seeds (cacao food), implying their role in seed development. The black rot disease of genus *Phytophthora* caused up to 20–25% loss (700,000 metric tons) in world cacao production. The role of *tcGASA* genes in conferring fungal resistance was also explored based on RNAseq data against *Phytophthora megakarya*. The differential expression of *tcGASA* genes was recorded between the tolerant and susceptible cultivars of cacao plants, which were inoculated with the fungus for 24h and 72h. This differential expression indicating possible role of *tcGASA* genes to fungal resistant in cacao. Our findings provide new insight into the function, evolution, and regulatory system of the *GASA* family genes in *T. cacao* and provide new target genes for development of fungi-resistant cacao varieties in breeding programs.

Keywords: Gibberellic acid-stimulated *Arabidopsis* (*GASA*); Gene expression; Phylogenetics; *Phytophthora megakarya*, Abiotic stresses, Biotic stresses; *Theobroma cacao*, Malvaceae

Introduction

The gibberellic acid-stimulated *Arabidopsis* (*GASA/GAST*) gene family is widely distributed in plants and performs various functions (Ahmad et al., 2020b; Fan et al., 2017). *GAST1* was the first gene identified among the *GASA* family's genes in tomato (Shi et al., 1992). *GASA* proteins are comprised of three domains, including a signal peptide of up to 18–29 amino acids at the N-terminal, hydrophilic and high variable regions of up to 7–31 amino acids in the center, and a conserved domain at the C-terminal of up to 60 amino acids which mostly includes 12 cysteine residues (Rezaee et al., 2020; Silverstein et al., 2007; Zhang and Wang, 2008). This third domain of the C-terminal is the characteristics of all identified GASAs (Ahmad et al., 2019; Fan et al., 2017; Rezaee et al., 2020; Zhang and Wang, 2008). These cysteine-rich peptides play a vital role in various plant processes. Their roles have been stated in organ development (Fuente et al., 2006), lateral root development (Zimmermann et al., 2010), stem growth (Zhang et al., 2009), cell division (Nahirñak et al., 2012; Roxrud et al., 2007), fruit ripening and development (Moyano-Cañete et al., 2013), flowering time (Ahmad et al., 2019; Fan et al., 2017; Zhang et al., 2009), seed development (Ahmad et al., 2020b; Roxrud et al., 2007), and bud dormancy (Yang et al., 2019). The detailed expression analyses of various tissues of *Arabidopsis*, tomato, rice, soybean, and apple showed the tissue-specific expression of the *GASA* genes (Ahmad et al., 2019; Fan et al., 2017; Rezaee et al., 2020). For example, *GASA* genes in tomato, including *Solyc11g011210*, *Solyc12g089300*, and *Solyc01g111075*, showed high expression at the fruit-ripening stage, whereas *Solyc03g113910*, *Solyc01g111075*, *Solyc06g069790*, and *Solyc12g042500* showed high expression level at the flowering stage (Rezaee et al., 2020). Moreover, some studies also showed the contrast effect of *GASA* proteins. For instance, *AtGASA4* promotes flowering (Roxrud et al., 2007), while *AtGASA5* induces the opposite effect (Zhang et al., 2009).

The *GASA* genes also play crucial roles in response to various biotic, abiotic, and hormone-related stresses. For instance, a member of the *GASA* gene family, *GmSNI*, provides resistance to viruses in soybean and *Arabidopsis* (He et al., 2017). Similarly, a high expression level of *CcGASA4* was reported in citrus leaves after infection with *Citrus tristeza* virus (Wu et al., 2020). The antimicrobial properties of various proteins of the *GASA* family have also been reported (Almasia et al., 2008; Berrocal-Lobo et al., 2002; Kovalskaya and Hammond, 2009). The antifungal activity of the *GASA* proteins has been reported in almost all tissues of potato,

including root, tubers, leaves, stem, stolon, axillary bud, and flowers (Almasia et al., 2008; Berrocal-Lobo et al., 2002; Kovalskaya and Hammond, 2009). Similarly, the antifungal activity of GASA members has been found in Arabidopsis, tomato, Alfalfa, and Jujuba (Balaji and Smart, 2012; García et al., 2014; Qu et al., 2016; Zhang and Wang, 2008). The GASAs also showed resistance to various abiotic stresses, such as salt and drought (Wang et al., 2018). The induction of *GASA4* and *GASA6* has been reported in Arabidopsis by growth hormones such as auxin, BR, GA, and cytokinin. In contrast, repression has been stated by stress hormones including SA, ABA, and JA (Qu et al., 2016).

Theobroma cacao L. belongs to the family Malvaceae (Purseglove, 1968). This is an economically important tree and grows in up to 50 countries located in the humid tropics (Motamayor et al., 2013). *Theobroma cacao* L. seeds are enclosed in pods and are used for chocolate production, confectionery, and cosmetics (Litz, 2005). This plant is adapted to high humidity areas, and is therefore predisposed to various fungal diseases (Bridgemohan and Mohammed, 2019; McElroy et al., 2018). Pod rot, or black rot, is caused by *Phytophthora* species of fungus (*P. megakarya*, *P. palmivora*, and *P. capsici*), leading to 20–30% loss in yield and 10% death of trees (Bridgemohan and Mohammed, 2019). The elucidation of the whole genome is helping to understand the genetic bases of biotic and abiotic stresses (Motamayor et al., 2013). The availability of the high-quality chromosome-level genome assembly of *Theobroma cacao* (Argout et al., 2017; Motamayor et al., 2013) provides quality resources for the characterization of various gene families to elucidate the role of different gene families in cacao development. However, few gene families such as WRKY (Dayanne et al., 2017) and NAC (Shen et al., 2020) have been elucidated in *Theobroma cacao*. Moreover, genome-wide analyses were also performed to reveal the divergent patterns of gene expression during zygotic and somatic embryo maturation in cacao (Maximova et al., 2014). Here, we are interested in characterizing the GASA gene family and providing data about the comprehensive and expression analyses of this gene family specifically regarding fungus-related diseases.

The expression pattern and evolutionary relationships of GASA genes were studied in Arabidopsis (Zhang and Wang, 2008), apple (Fan et al., 2017), common wheat (Cheng et al., 2019), grapevine (Ahmad et al., 2020b), soybean (Ahmad et al., 2019), and *Solanum tuberosum* subsp. *tuberosum* cv. Kennebec (Nahiriñak et al., 2016). In the current study, we are interested in identifying and characterizing GASAs of *T. cacao* and exploring their roles in various biotic and

abiotic stresses. We provide data of GASAs related to their distribution in genome, chemical properties, subcellular localization, and cis-regulatory elements of promoter regions for the first time in cacao. We also explored the roles of GASA in various abiotic and biotic stresses. This helps us to identify GASAs that show high expressions against infections of fungus *Phytophthora megakarya*.

2. Materials and Methods

2.1 Identification of GASA genes in the genome of *Theobroma cacao* and analyses for conserved GASA domain

We retrieved the GASA family's protein sequences from Arabidopsis Information Resource (TAIR10) database (<ftp://ftp.arabidopsis.org>). We used them as a query in BLAST for the identification of GASA genes in the *Theobroma cacao* genome. The GASA genes were identified in the latest version of the *Theobroma cacao* genome (*Theobroma cacao* Belizian Criollo B97-61/B2) (Argout et al., 2017). We retrieved protein sequences, coding DNA sequences (CDS), genomics, and promoter sequence (1500 bp upstream of gene). The retrieved protein sequences were further analyzed for the presence of the GASA domain using CDD (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) database of the National Center for Biotechnology Information (NCBI). All the sequences that showed the presence of the conserved GASA domain were selected for further analyses, whereas all the proteins with an absent/truncated domain were discarded.

2.2 Chromosome mapping and characterization of physiochemical properties

The position of each gene, including chromosome number and position on chromosome, was noted. All the genes were renamed according to the location of the chromosome and their position, as shown in table 1. The MapChart software (Voorrips, 2002) through Ensemble was used to show the position of each *tcGASA* gene along the position of the chromosome. Various physiochemical properties, including length of the protein, molecular weight (MW), isoelectric point (PI), instability index, and the grand average of hydropathy (GRAVY), were determined using the ExPASy tool (Gasteiger et al., 2005). The subcellular localization of GASA genes was also predicted using the BUSCA webserver (Savojardo et al., 2018).

2.3 Gene structure and promoter region analyses

We analyzed CDS sequences for exons–introns within all *tcGASA* genes using the Gene Structure Display Server (<http://gsds.cbi.pku.edu.cn>). PlantCare (Lescot et al., 2002) was used to study cis-regulatory elements in the 1500 bp promoter region.

2.4 Prediction of post-translational modifications of GASA proteins

The phosphorylation site of the GASA proteins was predicted by the NetPhos 3.1 server (<http://www.cbs.dtu.dk/services/NetPhos/>) (Blom et al., 2004) with a potential value > 0.5. N-glycosylation sites were predicted using the NetNGlyc 1.0 server (<http://www.cbs.dtu.dk/services/NetNGlyc/>) (Gupta and Brunak, 2002) with default parameters.

2.5 Phylogenetic and conserved motif analyses

The phylogenetic relationship of the *GASA* genes of *T. cacao* was inferred with *GASA* genes of six other species, including *Arabidopsis*, *Gossypium raimondii*, *Vitis vinifera*, rice, *Brachypodium distachyon*, and maize. A similar approach, which was given for the identification of *GASA* genes in *Theobroma cacao*, was employed for the identification of the genes of all other species. Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) (Sievers et al., 2011) was used for multiple alignment of the protein sequences of all species. The unrooted neighbor-joining tree was drawn using MEGA X (Kumar et al., 2018) and visualization of the tree was improved by using an interactive tree of life (iTOL) (Letunic and Bork, 2019). The conserved motif distribution into *GASA* proteins was performed using MEME v5.3.0 server (<http://meme-suite.org/tools/meme>) (Bailey et al., 2009). We searched for a maximum number of five motifs with a minimum width of motif 6 and a maximum motif width of 30.

2.6. Gene duplications and estimation of Ka/Ks values

The identity of > 85% in nucleotide sequences of genes is considered a sign of duplication (Zheng et al., 2010). Hence, we aligned DNA coding sequences using Clustal Omega (Sievers et al., 2011), and the extent of the identity of the genes with each other was determined using Geneious R8.1 (Kearse et al., 2012). Gene duplication events, as compared to other species, were determined using the MCScan v0.8 program (Wang et al., 2012) through the Plant Genome Duplication Database.

The selection pressure on *GASA* genes was determined by calculating non-synonymous (K_a) and synonymous (K_s) substitutions and the ratio of non-synonymous to synonymous substitutions (K_a/K_s) using DnaSP v6 software (Rozas et al., 2017). The time divergence and duplication were assessed by a synonymous mutation rate of λ substitutions per synonymous site

per year as $T = (Ks/2\lambda)$ ($\lambda = 6.5 \times 10^{-9}$) $\times 10^{-6}$ (Yang et al., 2008). The synteny relationships of *GASA* genes among the orthologous pairs of cacao–rice, cacao–*Arabidopsis*, cacao–maize, cacao–*Brachypodium distachyon*, cacao–*Gossypium raimondii*, and cacao–*Vitis vinifera* at both gene and chromosome levels were visualized using Circos software (Krzywinski et al., 2009).

2.7 Three-dimensional protein modeling and molecular docking

We used iterative template-based fragment assembly simulations in I-TASSER (Yang et al., 2015) to build three-dimensional protein structures of GASAs after selection of best models by the 3D-refine program (Bhattacharya et al., 2016). We draw a Ramachandran plot using the RAMPAGE program to validate the predicted structures by measuring backbone psi (Ψ) and dihedral phi (Φ) angles (Lovell et al., 2003). We also used P2Rank in PrankWeb software (Jendele et al., 2019) and CASTp tool (Tian et al., 2018) to analyze the refined structure of *GASA* proteins to predict protein pockets and cavities. Finally, PyMOL (DeLano, 2002) was used to visualize results.

2.8. In silico expression analysis of *GASA* Genes through RNA-seq data

The publicly available RNA-seq data related to the cacao genome was employed for expression assays to measure *GASA* family members in multiple tissues and during various biotic and abiotic stimuli exposure. The RNA-seq data of cacao under various time courses of biotic stress (*Phytophthora megakarya*) were downloaded from GEO DataSets under accession number GSE116041 (Pokou et al., 2019). This data was also log2 transformed to generate heatmaps via the TBtools package (Chen et al., 2020). Furthermore, the expression levels of *GASA* genes for tissue specific expression and in under multiple abiotic stresses, including cold, osmotic, salt, drought, UV, wounding and heat, have been detected in the *Arabidopsis* orthologous genes for *tcGASAs* (SAMEA5755003 and PRJEB33339).

3. Results

3.1 Identification of *GASA* genes and their distributions on chromosomes within genomes

We detected 17 *GASA* genes in the genome of *T. cacao* distributed on six chromosomes out of ten. These genes were named from *tcGASA1* to *tcGASA17* based on their distribution on chromosomes starting from chromosome 1. When two or more genes were present on the same chromosome, then the gene present at the start of a chromosome was named first (Table 1 and

Figure 1). Six genes were distributed on chromosome 8 and five genes were distributed on chromosome 4. Chromosome 1 and chromosome 2, each contained two genes, whereas chromosome 5 and chromosome 9 contained one gene each. This data showed the unequal distribution of *tcGASA* genes within the cacao genome. The location of each gene on the chromosome is mentioned in Table 1, as well as the start and end. Moreover, the sequences of genes, proteins, coding regions, and promoter regions are provided in Table S1.

3.2 Protein length, molecular weight, and isoelectric point of *tcGASA* proteins

In the current study, *tcGASAs* were characterized based on their physiochemical properties (Table 1). The identified *tcGASA* proteins were low molecular weight proteins ranging in length from 88 (*tcGASA13*) to 320 (*tcGASA5*) amino acids with molecular weight (MW) ranging from 9.64 to 33.92 kDa. Except for *tcGASA5* and *tcGASA8*, with molecular weights of 13.42 kDa and 33.92 kDa, respectively, the MW of all other *tcGASAs* was found to be less than 13 kDa. The isoelectric point also showed similarities among *tcGASA* and suggested the alkaline nature of the proteins. Except for *tcGASA11*, which has an isoelectric point of 6.67, all other *tcGASA* proteins have isoelectric points of more than 8, ranging from 8.27 to 9.83.

3.3 Analyses of instability index, GRAVY, and subcellular localization of *tcGASA* proteins

The instability index provides information about the stable and non-stable features of proteins in the various biochemical processes. The instability index indicated 4 stable *tcGASAs* including *tcGASA3*, *tcGASA11*, *tcGASA13*, and *tcGASA15*, as well as 13 unstable *tcGASAs* (Table 1). The positive value of GRAVY indicates its hydrophobic nature, whereas the negative value indicates the hydrophilic nature of proteins. The GRAVY value was recorded as negative for sixteen *tcGASA* proteins ranging from -0.75 to -0.023, but as positive (0.004) for *tcGASA12* (Table 1). Hence, the data indicate the hydrophilic nature of most *tcGASA* proteins. Subcellular localization provides information about the function of proteins. Based on BUSCA, we predicted extracellular localization of *tcGASA* proteins, except for *tcGASA8*, which localized in the plasma membrane (Table 1).

3.4 *tcGASA* proteins 3D structure analyses and post-translational modifications

The predicted 3D structure of all *tcGASA* proteins showed that these proteins contain β sheets, α helices, random coils, and extended strands. The random coils were the most abundant and were more extensive than α helices, while the β sheets were the least (Figure 2). In the present study, the post-translational modifications of *tcGASAs* were predicted in terms of phosphorylation and

glycosylation (Figure 3, Table S2). We predicted a total of 224 potential phosphorylation events on amino acids serine, threonine, and tyrosine within tcGASA proteins. Most of the phosphorylation events were predicted related to serine (92) followed by threonine (86) and then by tyrosine (46). Among tcGASA proteins, most of phosphorylation sites (57 sites) were predicted in tcGASA05, whereas in other proteins' phosphorylation events ranged from 9 to 14 sites. Three tcGASA including tcGASA10, tcGASA15, and tcGASA17, were also identified with a potential glycosylation site (Table S2).

3.5 Phylogenetic analyses of tcGASA proteins

The phylogenetic inference of ninety-nine GASA-protein sequences from seven species (mentioned in methodology), including 17 sequences of tcGASA, resolved into five groups (Figure 4). The GASA proteins of cacao showed high variations and were distributed into all five groups. Group I include only one sequence tcGASA16 (Tc08v2_t014690). Group II was confined to five protein sequences including tcGASA01 (Gene ID: Tc01v2_t001590), tcGASA4 (Tc02v2_t021110), tcGASA5 (Tc04v2_t000240), tcGASA11 (Tc08v2_t003650), and tcGASA12 (Tc08v2_t003660). Group III comprised two sequences—tcGASA14 (Tc08v2_t014670) and tcGASA15 (Tc08v2_t014680). Group IV included tcGASA2 (Tc01v2_t005890), tcGASA03 (Tc02v2_t009150), and tcGASA13 (Tc08v2_t007890). Group V consisted of six protein sequences, including tcGASA6 (Tc04v2_t006890), tcGASA7 (Tc04v2_t009440), tcGASA8 (Tc04v2_t016520), tcGASA9 (Tc04v2_t016530), tcGASA10 (Tc05v2_t012230), and tcGASA17 (Tc09v2_t020100). The tcGASA4 (Tc02v2_t021110) shared a node with one of the protein sequences of *Vitis vinifera*, whereas the other 16 proteins showed close relationships with the sequences of another species of family Malvaceae (*Gossypium raimondii*).

3.6 Gain and loss of intron(s), proteins motifs, and their linked to phylogenetic grouping

We also determined numbers of introns-exons within the genes and motifs in protein sequences. We drew a separate phylogeny of the tcGASA genes to compare tcGASA genes for presence and location of introns, and tcGASA proteins for conserved motifs, based on phylogenetic clusters that they formed. Each phylogenetic group was shown with different colors for clarity (Figure 5.a). The analyses of genomic sequences showed the absence of intron in one gene, presence of one intron in five genes, two introns in eleven genes, and three introns in one gene (Figure 5b). Genes clustered together showed differences in their number and in intron-exon distributions

(Figure 5a, b). Five motifs were revealed in tcGASA. Four motifs (1–4) were distributed in all tcGASAs, whereas a fifth motif was limited to Tc04v2_t016520 (tcGASA8), Tc05v2_t012230 (tcGASA10), and Tc09v2_t020100 (tcGASA17) (Figure 5c). The motifs of proteins that clustered together within the phylogenetic tree presented similarities in the distribution of motifs to some extent (Figure 5 a, c).

3.7 Duplications, divergence, and synteny among GASA genes

We analyzed the paralogous relationships among the *GASA* genes within cacao and analyzed the orthologous relationships of *GASA* genes by comparing them with the other six species (mentioned in methodology). Segmental duplications were found in nine *GASA* genes that were paired into four groups as: *tcGASA2-tcGASA3-tcGASA13*; *tcGASA8-tcGASA9*; *tcGASA10-tcGASA17*; and *tcGASA14-tcGASA15* (Table 2) (as shown with similar colors in Figure 1). The analyses of synonymous and non-synonymous substitutions revealed that high purifying selection pressure exists on these genes after duplication. The analyses of divergence time indicated that the event of duplication of these pairs occurred 50 MYA to 204 MYA (Table 2). The synteny analyses with *GASAs* of other species showed high resemblance and identified orthologous genes among *T. cacao* and compared species (Figure 6).

3.8 Promoter regions analysis

The analyses of cis-regulatory elements in promoter regions revealed the presence of binding sites for key transcription factors related to light-responsive elements (48.88%), hormone-responsive elements (25.40%), stress-related elements (19.06%), growth-response elements (5.40%), and DNA- and protein-related binding sites (1.26) (Figure 7a). Regulatory sites were found for various hormones such as auxin, salicylic acid, abscisic acid, gibberellin, and MeJA (Figure 7b). Similarly, regulatory elements were identified for drought, elicitor, anaerobic induction, low temperature, and plant defense/stress (Figure 7c). The complete detail of each element, along with sequence and function is provided in Table S3.

3.9 Tissue-specific expression of *tcGASA* genes

We evaluated the expression of orthologous *tcGASAs* in various tissues to evaluate their role in the functions of *T. cacao* (Figure 8a). Differential expression was noted for *tcGASA* genes in various tissues of cacao. Five genes, including *tcGASA02*, *tcGASA03*, *tcGASA08*, *tcGASA09*, and *tcGASA13*, showed high expression in the beans (food part of cacao). In addition, *tcGASA16* showed a high expression in leaves and entire seedlings. However, *tcGASA02* and *tcGASA03*

were significantly downregulated in leaves compared to beans. *In-silico* expression results showed that *tcGASA12* and *tcGASA17* were less induced in pistil tissues.

3.10 Expression analyses of *tcGASAs* in abiotic and biotic stresses

We explicated the role of *tcGASAs* under various abiotic stress by analyzing orthologous genes of Arabidopsis (Figure 8b). The orthologous genes of *tcGASA01*, *tcGASA05*, *tcGASA12*, and *tcGASA14* showed an upregulation in response to wound healing and the orthologs of *tcGASA05* were more expressed in response to cold stress. The expression of orthologous *GASA* genes in Arabidopsis was less induced in response to drought and UV stresses. The *tcGASA01* and *tcGASA14* were highly upregulated in response to osmotic pressure and salt stress, while *tcGASA016* and *tcGASA17* were downregulated.

3.11 Expression analyses of *tcGASAs* in biotic stress (*P. megakarya*)

The role of *tcGASAs* against *P. megakarya* was accessed using RNA-seq data of cacao inoculated with *P. megakarya* for 0 hour (0h) 6h, 24h, and 72 h in fungal resistant cultivar Scavina (SCA6) and susceptible cultivar Nanay (NA32) (Figure 9a, b). Four genes *tcGASA01*, *tcGASA08*, *tcGASA09*, *tcGASA15* were not expressed. The analyses showed differential expression of *tcGASAs* expression under biotic stress of fungus *P. megakarya* in SCA6 and NA32 (Figure 10). The *tcGASA03*, *tcGASA05*, *tcGASA06*, *tcGASA16*, and *tcGASA17* showed high expression after 24 hours (24h) of treatment and *tcGASA02*, *tcGASA04*, and *tcGASA13* highly expressed after 72h only in SCA6. This showed their possible role to fungal resistant.

4. Discussion

In the current study, we identified 17 *tcGASAs* and analyzed their genomic distributions and chemical properties. RNAseq data analyses explicated their possible roles in bean development (food of cacao), various abiotic stresses, and the biotic stress of *P. megakarya*.

Cacao is an economically important plant and its seeds are used for chocolate, and it is also the main source of income for 40–50 million farmers (Abdullah et al., 2019b; Litz, 2005; Motamayor et al., 2013). Despite this importance, only two gene families, *WRKY* (Dayanne et al., 2017) and *NAC* (Shen et al., 2020), are studied in cacao. The role of the *NAC* family was not explained in abiotic and biotic stresses. The *GASA* family's role was reported in plant development, function regulation, and biotic and abiotic stresses, as mentioned in the introduction (*vide infra*). Cacao production is also affected by various biotic and abiotic stresses.

In the current study, we focused on the GASA gene family in cacao. We identified 17 GASA genes, which were unequally distributed on six chromosomes among ten. The GASA proteins have low molecular weight with conserved GASA and cysteine domains. Our findings are in agreement with previous studies that revealed that GASA genes mostly exist in lower numbers, have low molecular weight, and are unequally distributed on chromosomes within genomes as reported for rice (9; Rezaee et al., 2020), Grapevine (14; Ahmad et al., 2020b), Arabidopsis (15; Fan et al., 2017), and tomato (19; Rezaee et al., 2020). However, a somewhat high number of GASA have also reported, such as in apple (26; Fan et al., 2017) and soybean (37; Ahmad et al., 2019). The number of introns were variable in tcGASA within genes that cluster together in phylogeny. The loss and gain of introns occurs in the course of evolution within protein-coding genes of the plants and is also reported within the GASA of other plant species (Abdullah et al., 2020a; Ahmad et al., 2019, 2020b; Ahmadizadeh et al., 2020; Ding et al., 2020; Heidari et al., 2019).

The tcGASAs were predicted to be alkaline, hydrophilic, and mostly unstable proteins. However, we also detected four stable proteins: tcGASA3, tcGASA11, tcGASA13, and tcGASA15. The stability demonstrates the lifetime of proteins related to cellular enzymatic reactions (Heidari et al., 2019). Hence, these four proteins may play extensive roles in various enzymatic activities. The data of subcellular localization also provides insight into the function of proteins (Zhang and Wang, 2011). Apart from tcGASA8 (which localizes to the plasma membrane), other proteins locate in the extracellular space. Extracellular localization of GASA proteins in several plants has also been reported previously (Ahmad et al., 2019; Rezaee et al., 2020; Zhang et al., 2009). Localization of GASA proteins in the plasma membrane, cytoplasm, and the nucleus has also been described, however (Wang et al., 2009). Variations in subcellular localization may occur due to various factors, such as protein-protein interaction and post-translational modifications (Nahirňak et al., 2016). Post-translational modifications are processes of chemical modifications of proteins and they produce diversity in structure and function, including subcellular localization, protein-protein interaction, and regulating enzyme activity by allosteric phenomena (Duan and Walther, 2015; Nahirňak et al., 2016; Webster and Thomas, 2012). We predicted phosphorylation sites in all tcGASA, ranging in number from 7 to 57. The phosphorylation of proteins is also vital for cell signaling, regulation of various protein mechanisms, and as a substrate for various kinases (Heidari et al., 2020; Nawaz et al., 2019; Silva-Sanchez et al.,

2015). Hence, tcGASA may be a good target for various studies to elucidate their complete role. Similarly, glycosylation is also an abundant and varied modification that plays an essential role in the biological and physiological functions of a living organism (Yang et al., 2017). We detected glycosylation sites on the N-terminal of tcGASA10, tcGASA14, and tcGASA16. These tcGASAs may play significant roles in plant function and regulations.

We performed the phylogenetic analyses of tcGASAs with GASA of 6 other species. We also included *Gossypium raimondii*, a closely related species from the plant family Malvaceae. The phylogenetic analyses distributed GASA proteins into 5 groups and the tcGASAs were also distributed into all five groups. Tc08v2_t01469 (tcGASA16) was the only protein which grouped with Arabidopsis, while other tcGASAs showed sister relationships with GASA of *Gossypium*. The cacao belongs to the basal group, whereas *Gossypium* belongs to the crown group of the family Malvaceae (Abdullah et al., 2019a, 2020c, 2020b), but the close phylogenetic relationships of proteins of these two species support their family-level relationship. The motifs and exon-intron analyses showed variations within proteins that cluster together in phylogeny. This shows that GASA of some groups may have evolved during evolution, which led to variations in motifs and introns in some groups. The same was observed in GASA of other species and has also been reported for other gene families (Ahmad et al., 2019, 2020b; Chattha et al., 2020; Lei et al., 2020). The phylogenetic grouping of GASA was not correlated with the function of the GASA, and we observed contrasting gene expression, including in the same group. These findings show that some other processes are related to the function of proteins instead of their close phylogenetic relationships. A similar observation was reported in *Nicotiana tabacum* L. (Nawaz et al., 2019). However, some studies also proposed that the closely related proteins on a phylogenetic tree have similar function (Nawaz et al., 2014; Nuruzzaman et al., 2010).

The tandem and segmental duplicated genes play an important role in evolution, domestication, functional regulation, and biotic and abiotic stresses (Ahmad et al., 2020a; Li et al., 2020; Liu et al., 2020; Piégu et al., 2020; Song et al., 2020). We determined segmental duplication of nine genes that form four groups as: *tcGASA2-tcGASA3-tcGASA13*, *tcGASA8-tcGASA9*, *tcGASA14-tcGASA15*, and *tcGASA10-tcGASA17*. These gene pairs were present in the same group within the phylogenetic tree. Similar results were reported in genome-wide analyses of GASAs in soybean (Ahmad et al., 2019). A previous study also proposed that segmentally duplicated genes

also showed similar functions and stable expressions. (Ahmad et al., 2019; Faraji et al., 2020). In the current study, each pair of segmentally duplicated genes did not agree with the previous finding. The pairs *tcGASA8-tcGASA9* and *tcGASA14-tcGASA15* showed similarities in gene expression in different tissues. The *tcGASA2-tcGASA3-tcGASA13* showed a similar expression in the entire seedling and bean but showed differential expression in pistils and leaves. In biotic stress of fungus (*P. megakarya*), *tcGASA2-tcGASA3-tcGASA13*, and *tcGASA8-tcGASA9*, *tcGASA10-tcGASA17* showed similar expression, while *tcGASA14-tcGASA15* and showed different expression as *tcGASA14* showed high expression whereas *tcGASA15* was found unexpressed. These findings suggest that segmentally duplicated pairs may also perform different functions. Hence, functional analyses of each segmentally duplicated gene can provide authentic information about their roles in various physiological and biochemical processes. Moreover, our study, along with the previous report (Ahmad et al., 2019), suggests that genes within the same groups have more chances of segmental duplication events among them. The $Ka/Ks < 1$ indicates that purifying selection pressure exists on the *GASA* genes after duplication, as reported previously (Ahmad et al., 2020b). We also observed mostly purifying selection pressure on *tcGASAs*, including duplicating genes (Table S1).

The cis-acting regulatory elements involved in transcription of regulation genes are induced through independent signal transduction pathways under biotic and abiotic stresses (Ahmadizadeh and Heidari, 2014; Nawaz et al., 2019). We observed several key cis-regulating elements in response to light, hormones, stresses, and growth in the promoter site of *tcGASAs*. The cis-regulating elements for drought, anerobic induction, low-temperature, and plant defense were also evident. The existence of diverse cis-regulating elements in promoter regions indicates their roles in the regulation of the *tcGASAs* and different pathways of cacao. Further study can shed light on the roles of these cis-regulating elements.

Different strategies are used to develop cacao cultivars that produce food in high quantities and are resistant to abiotic and biotic stresses (Bridgemohan and Mohammed, 2019). Here, we studied the role of *GASA* genes in cacao development and in protection against biotic and abiotic stresses. Tissue-specific expression was observed for *tcGASAs*, which showed their role in the development and functional regulation of cacao. Up to eight *tcGASAs* were highly expressed in the cacao bean, which is the food part used for making chocolate (Abdullah et al., 2020d). This high expression may reveal the conserved function of these genes in the development of the

bean. The role of *tcGASAs* was also stated in the development of grapevine seeds (Ahmad et al., 2020b). Further data based on cloning can provide new insight into their roles in bean development. Expression analyses of the orthologous genes of *Arabidopsis* also indicated the role of *GASA* genes in various abiotic stresses, including drought, which significantly affects the growth of cacao (Medina and Laliberte, 2017). Hence, these genes may also be important to produce drought-resistant cultivars. Expression analyses also provide insight into the gene's function in response to biotic stresses (Nawaz et al., 2019). The black rod disease of genus *Phytophthora* caused up to 20–25% loss (700,000 metric tons) to the world cacao production annually. In some regions, the *Phytophthora* caused losses of 30–90% of the crops (Adeniyi, 2019). Here, we explored the function of *tcGASAs* based on RNAseq data against black rod causing pathogen *P. megakarya* and observed highly expressed *tcGASA* genes in plants inoculated at 24h and 72h in the tolerant cultivar SCA6 as compared to susceptible cultivar NA32. This data indicates that *tcGASAs* respond to fungus. Hence, the complete characterization of these upregulated genes can provide target genes for the development of resistant cultivars to the disease of genus *Phytophthora* to enhance the production of cacao for the welfare of not only farmers involved in cacao cultivation, but also for the welfare of all humanity to provide high-quality delicious chocolate with quality nutrition.

In conclusion, our study provides new insight into the identification, characterization, and expression of the *GASA* genes in the *Theobroma cacao* plant. Expression analyses revealed the role of the *GASA* genes in seed development. We also determined that *GASA* genes provide resistance against the fungus *Phytophthora megakarya*, which causes significant losses to cacao production each year. Our study provides a base for the generation of cultivars that are resistant to fungus of the genus *Phytophthora*.

Data availability

The public data set is analyzed in the manuscript. All the analyzed data are available in the main manuscript or as supplementary materials.

Conflict of Interest

Ibrar Ahmed and Hafiz Muhammad Talha Malik was employed of Alpha Genomics. All other author declared that no conflict of interest exists.

References

- Abdullah, Henriquez, C. L., Mehmood, F., Shahzadi, I., Ali, Z., Waheed, M. T., et al. (2020a). Comparison of chloroplast genomes among Species of Unisexual and Bisexual clades of the monocot family Araceae. *Plants* 9, 737. doi:10.3390/plants9060737.
- Abdullah, Mehmood, F., Shahzadi, I., Ali, Z., Islam, M., Naeem, M., et al. (2020b). Correlations among oligonucleotide repeats, nucleotide substitutions and insertion-deletion mutations in chloroplast genomes of plant family Malvaceae. *J. Syst. Evol.* doi:10.1111/jse.12585.
- Abdullah, Mehmood, F., Shahzadi, I., Waseem, S., Mirza, B., Ahmed, I., et al. (2020c). Chloroplast genome of Hibiscus rosa-sinensis (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics* 112, 581–591. doi:10.1016/j.ygeno.2019.04.010.
- Abdullah, Shahzadi, I., Mehmood, F., Ali, Z., Malik, M. S., Waseem, S., et al. (2019a). Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. *Plant Gene* 19, 100199. doi:10.1016/J.PLGENE.2019.100199.
- Abdullah, Waseem, S., Mirza, B., Ahmed, I., and Waheed, M. T. (2019b). Comparative analyses of chloroplast genome in Theobroma cacao and Theobroma grandiflorum. *Biologia (Bratisl)*.
- Abdullah, Waseem, S., Mirza, B., Ahmed, I., and Waheed, M. T. (2020d). Comparative analyses of chloroplast genomes of Theobroma cacao and Theobroma grandiflorum. *Biologia (Bratisl)*. 75, 761–771. doi:10.2478/s11756-019-00388-8.
- 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* (IntechOpen). doi:10.5772/intechopen.81993.
- Ahmad, B., Azeem, F., Ali, M. A., Nawaz, M. A., Nadeem, H., Abbas, A., et al. (2020a). Genome-wide identification and expression analysis of two component system genes in Cicer arietinum. *Genomics* 112, 1371–1383. doi:10.1016/j.ygeno.2019.08.006.
- Ahmad, B., Yao, J., Zhang, S., Li, X., Zhang, X., Yadav, V., et al. (2020b). Genome-wide characterization and expression profiling of GASA genes during different stages of seed development in grapevine (Vitis vinifera L.) predict their involvement in seed development. *Int. J. Mol. Sci.* 21, 1–16. doi:10.3390/ijms21031088.

- Ahmad, M. Z., Sana, A., Jamil, A., Nasir, J. A., Ahmed, S., Hameed, M. U., et al. (2019). A genome-wide approach to the comprehensive analysis of GASA gene family in Glycine max. *Plant Mol. Biol.* 100, 607–620. doi:10.1007/s11103-019-00883-1.
- Ahmadizadeh, M., and Heidari, P. (2014). Bioinformatics study of transcription factors involved in cold stress. *Biharean Biol.* 8.
- Ahmadizadeh, M., Rezaee, S., and Heidari, P. (2020). Genome-wide characterization and expression analysis of fatty acid desaturase gene family in Camelina sativa. *Gene Reports* 21, 100894. doi:10.1016/j.genrep.2020.100894.
- Almasia, N. I., Bazzini, A. A., Hopp, H. E., and Vazquez-Rovere, C. (2008). Overexpression of snakin-1 gene enhances resistance to *Rhizoctonia solani* and *Erwinia carotovora* in transgenic potato plants. *Mol. Plant Pathol.* 9, 329–338. doi:10.1111/j.1364-3703.2008.00469.x.
- Argout, X., Martin, G., Droc, G., Fouet, O., Labadie, K., Rivals, E., et al. (2017). The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. *BMC Genomics* 2017 181 18, 1–9. doi:10.1186/S12864-017-4120-9.
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi:10.1093/nar/gkp335.
- Balaji, V., and Smart, C. D. (2012). Over-expression of snakin-2 and extensin-like protein genes restricts pathogen invasiveness and enhances tolerance to *Clavibacter michiganensis* subsp. *michiganensis* in transgenic tomato (*Solanum lycopersicum*). *Transgenic Res.* 21, 23–37. doi:10.1007/s11248-011-9506-x.
- Berrocal-Lobo, M., Segura, A., Moreno, M., López, G., García-Olmedo, F., and Molina, A. (2002). Snakin-2, an antimicrobial peptide from potato whose gene is locally induced by wounding and responds to pathogen infection. *Plant Physiol.* 128, 951–961. doi:10.1104/pp.010685.
- Bhattacharya, D., Nowotny, J., Cao, R., and Cheng, J. (2016). 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Res.* 44, W406–W409.
- Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4, 1633–1649. doi:10.1002/pmic.200300771.

- Bridgemohan, P., and Mohammed, M. (2019). “ The Ecophysiology of Abiotic and Biotic Stress on the Pollination and Fertilization of Cacao (Theobroma cacao L.; formerly Sterculiaceae family) ,” in *Abiotic and Biotic Stress in Plants* (IntechOpen).
doi:10.5772/intechopen.84528.
- Chattha, W. S., Atif, R. M., Iqbal, M., Shafqat, W., Farooq, M. A., and Shakeel, A. (2020). Genome-wide identification and evolution of Dof transcription factor family in cultivated and ancestral cotton species. *Genomics* 112, 4155–4170. doi:10.1016/j.ygeno.2020.07.006.
- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13, 1194–1202. doi:10.1016/j.molp.2020.06.009.
- Dayanne, S. M. D. A., Oliveira Jordão Do Amaral, D., Del-Bem, L. E., Bronze Dos Santos, E., Santana Silva Raner, J., Peres Gramacho, K., et al. (2017). Genome-wide identification and characterization of cacao WRKY transcription factors and analysis of their expression in response to witches’ broom disease. *PLoS One* 12. doi:10.1371/journal.pone.0187346.
- DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsl. protein Crystallogr.* 40, 82–92.
- Ding, Q., Yang, X., Pi, Y., Li, Z., Xue, J., Chen, H., et al. (2020). Genome-wide identification and expression analysis of extensin genes in tomato. *Genomics* 112, 4348–4360.
doi:10.1016/j.ygeno.2020.07.029.
- Duan, G., and Walther, D. (2015). The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLoS Comput. Biol.* 11, 1004049.
doi:10.1371/journal.pcbi.1004049.
- Fan, S., Zhang, D., Zhang, L., Gao, C., Xin, M., Tahir, M. M., et al. (2017). Comprehensive analysis of GASA family members in the Malus domestica genome: Identification, characterization, and their expressions in response to apple flower induction. *BMC Genomics* 18, 1–19. doi:10.1186/s12864-017-4213-5.
- Faraji, S., Filiz, E., Kazemitabar, S. K., Vannozzi, A., Palumbo, F., Barcaccia, G., et al. (2020). The AP2/ERF Gene Family in Triticum durum: Genome-Wide Identification and Expression Analysis under Drought and Salinity Stresses. *Genes (Basel)*. 11, 1464.
- Fuente, D. L. J. I., Amaya, I., Castillejo, C., Sánchez-Sevilla, J. F., Quesada, M. A., Botella, M. A., et al. (2006). The strawberry gene FaGAST affects plant growth through inhibition of

- cell elongation. *J. Exp. Bot.* 57, 2401–2411. doi:10.1093/jxb/erj213.
- García, A. N., Ayub, N. D., Fox, A. R., Gómez, M. C., Diéguez, M. J., Pagano, E. M., et al. (2014). Alfalfa snakin-1 prevents fungal colonization and probably coevolved with rhizobia. *BMC Plant Biol.* 14, 248. doi:10.1186/s12870-014-0248-9.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). “Protein Identification and Analysis Tools on the ExPASy Server,” in *The Proteomics Protocols Handbook* (Humana Press), 571–607. doi:10.1385/1-59259-890-0:571.
- Gupta, R., and Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function - PubMed. *Pacific Symp. Biocomput.*, 310–322. Available at: <https://pubmed.ncbi.nlm.nih.gov/11928486/> [Accessed December 19, 2020].
- He, H., Yang, X., Xun, H., Lou, X., Li, S., Zhang, Z., et al. (2017). Over-expression of GmSN1 enhances virus resistance in Arabidopsis and soybean. *Plant Cell Rep.* 36, 1441–1455. doi:10.1007/s00299-017-2167-3.
- Heidari, P., Ahmadizadeh, M., Izanlo, F., and Nussbaumer, T. (2019). In silico study of the CESA and CSL gene family in Arabidopsis thaliana and Oryza sativa: Focus on post-translation modifications. *Plant Gene* 19. doi:10.1016/j.plgene.2019.100189.
- Heidari, P., Mazloomi, F., Nussbaumer, T., and Barcaccia, G. (2020). Insights into the SAM synthetase gene family and its roles in tomato seedlings under abiotic stresses and hormone treatments. *Plants* 9. doi:10.3390/plants9050586.
- Jendele, L., Krivak, R., Skoda, P., Novotny, M., and Hoksza, D. (2019). PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Res.* 47, W345–W349.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi:10.1093/bioinformatics/bts199.
- Kovalskaya, N., and Hammond, R. W. (2009). Expression and functional characterization of the plant antimicrobial snakin-1 and defensin recombinant proteins. *Protein Expr. Purif.* 63, 12–17. doi:10.1016/j.pep.2008.08.013.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular

- Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096.
- Lei, P., Wei, X., Gao, R., Huo, F., Nie, X., Tong, W., et al. (2020). Genome-wide identification of PYL gene family in wheat: Evolution, expression and 3D structure analysis. *Genomics*. doi:10.1016/j.ygeno.2020.12.017.
- Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi:10.1093/nar/30.1.325.
- Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi:10.1093/nar/gkz239.
- Li, Z., Gao, Z., Li, R., Xu, Y., Kong, Y., Zhou, G., et al. (2020). Genome-wide identification and expression profiling of HD-ZIP gene family in *Medicago truncatula*. *Genomics* 112, 3624–3635. doi:10.1016/j.ygeno.2020.03.008.
- Litz, R. (2005). *Theobroma Cacao*. In *Biotechnology of fruit and nut crops*. Available at: [https://books.google.com.pk/books?hl=en&lr=&id=AxbUJntXepEC&oi=fnd&pg=PR9&dq=Theobroma+cacao+\(Cacao\).+In+Biotechnology+of+fruit+and+nut+crops.+Edited&ots=hxySo_7J_O&sig=bnriIwslrWbltZsdjF8xtwiLK9E](https://books.google.com.pk/books?hl=en&lr=&id=AxbUJntXepEC&oi=fnd&pg=PR9&dq=Theobroma+cacao+(Cacao).+In+Biotechnology+of+fruit+and+nut+crops.+Edited&ots=hxySo_7J_O&sig=bnriIwslrWbltZsdjF8xtwiLK9E) [Accessed September 16, 2018].
- Liu, C., Wu, Y., Liu, Y., Yang, L., Dong, R., Jiang, L., et al. (2020). Genome-wide analysis of tandem duplicated genes and their contribution to stress resistance in pigeonpea (*Cajanus cajan*). *Genomics*. doi:10.1016/j.ygeno.2020.10.003.
- Lovell, S. C., Davis, I. W., Arendall III, W. B., De Bakker, P. I. W., Word, J. M., Prisant, M. G., et al. (2003). Structure validation by α geometry: ϕ , ψ and $C\beta$ deviation. *Proteins Struct. Funct. Bioinforma.* 50, 437–450.
- Maximova, S. N., Florez, S., Shen, X., Niemenak, N., Zhang, Y., Curtis, W., 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.
- McElroy, M. S., Navarro, A. J. R., Mustiga, G., Stack, C., Gezan, S., Peña, G., et al. (2018). Prediction of Cacao (*Theobroma cacao*) Resistance to *Moniliophthora* spp. Diseases via Genome-Wide Association Analysis and Genomic Selection. *Front. Plant Sci.* 9, 343.

doi:10.3389/fpls.2018.00343.

- Medina, V., and Laliberte, B. (2017). *A review of research on the effects of drought and temperature stress and increased CO₂ on Theobroma cacao L., and the role of genetic diversity to address climate change*. Bioversity. Costa Rica.
- Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., Cornejo, O., 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. doi:10.1186/gb-2013-14-6-r53.
- Moyano-Cañete, E., Bellido, M. L., García-Caparrós, N., Medina-Puche, L., Amil-Ruiz, F., González-Reyes, J. A., et al. (2013). FaGAST2, a strawberry ripening-related gene, acts together with fagast1 to determine cell size of the fruit receptacle. *Plant Cell Physiol.* 54, 218–236. doi:10.1093/pcp/pcs167.
- Nahirñak, V., Almasia, N. I., Hopp, H. E., and Vazquez-Rovere, C. (2012). Snakin/GASA proteins. *Plant Signal. Behav.* 7, 1004–1008. doi:10.4161/psb.20813.
- Nahirñak, V., Rivarola, M., De Urreta, M. G., Paniego, N., Hopp, H. E., Almasia, N. I., et al. (2016). Genome-wide analysis of the Snakin/GASA gene family in Solanum Tuberosum cv. Kennebec. *Am. J. potato Res.* 93, 172–188.
- Nawaz, Z., Kakar, K. U., Saand, M. A., and Shu, Q. Y. (2014). Cyclic nucleotide-gated ion channel gene family in rice, identification, characterization and experimental analysis of expression response to plant hormones, biotic and abiotic stresses. *BMC Genomics* 15, 1–18. doi:10.1186/1471-2164-15-853.
- Nawaz, Z., Kakar, K. U., Ullah, R., Yu, S., Zhang, J., Shu, Q. Y., et al. (2019). Genome-wide identification, evolution and expression analysis of cyclic nucleotide-gated channels in tobacco (Nicotiana tabacum L.). *Genomics* 111, 142–158. doi:10.1016/j.ygeno.2018.01.010.
- Nuruzzaman, M., Manimekalai, R., Sharoni, A. M., Satoh, K., Kondoh, H., Ooka, H., et al. (2010). Genome-wide analysis of NAC transcription factor family in rice. *Gene* 465, 30–44. doi:10.1016/j.gene.2010.06.008.
- Piégu, B., Arensburger, P., Beauclair, L., Chabault, M., Raynaud, E., Coustham, V., et al. (2020). Variations in genome size between wild and domesticated lineages of fowls belonging to the Gallus gallus species. *Genomics* 112, 1660–1673. doi:10.1016/j.ygeno.2019.10.004.
- Pokou, D. N., Fister, A. S., Winters, N., Tahi, M., Klotioloma, C., Sebastian, A., et al. (2019).

- Resistant and susceptible cacao genotypes exhibit defense gene polymorphism and unique early responses to *Phytophthora megakarya* inoculation. *Plant Mol. Biol.* 99, 499–516. doi:10.1007/s11103-019-00832-y.
- Purseglove, J. W. (1968). Tropical crops: Dicotyledons 1 and 2. *Trop. Crop. Dicotyledons 1 2*. Available at: <https://www.cabdirect.org/cabdirect/abstract/19691701550> [Accessed September 16, 2018].
- Qu, J., Kang, S. G., Hah, C., and Jang, J. C. (2016). Molecular and cellular characterization of GA-Stimulated Transcripts GAS4 and GAS6 in *Arabidopsis thaliana*. *Plant Sci.* 246, 1–10. doi:10.1016/j.plantsci.2016.01.009.
- Rezaee, S., Ahmadizadeh, M., and Heidari, P. (2020). Genome-wide characterization, expression profiling, and post-transcriptional study of GAS4 gene family. *Gene Reports* 20, 100795. doi:10.1016/j.genrep.2020.100795.
- Roxrud, I., Lid, S. E., Fletcher, J. C., Schmidt, E. D. L., and Opsahl-Sorteberg, H. G. (2007). GAS4, one of the 14-member *Arabidopsis* GAS4 family of small polypeptides, regulates flowering and seed development. *Plant Cell Physiol.* 48, 471–483. doi:10.1093/pcp/pcm016.
- Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302.
- Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018). BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, W459–W466. doi:10.1093/nar/gky320.
- Shen, S., Zhang, Q., Shi, Y., Sun, Z., Zhang, Q., Hou, S., et al. (2020). Genome-wide analysis of the NAC domain transcription factor gene family in *Theobroma cacao*. *Genes (Basel)*. 11. doi:10.3390/genes11010035.
- Shi, L., Gast, R. T., Gopalraj, M., and Olszewski, N. E. (1992). Characterization of a shoot-specific, GA3- and ABA-Regulated gene from tomato. *Plant J.* 2, 153–159. doi:10.1111/j.1365-313X.1992.00153.x.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi:10.1038/msb.2011.75.

- Silva-Sanchez, C., Li, H., and Chen, S. (2015). Recent advances and challenges in plant phosphoproteomics. *Proteomics* 15, 1127–1141. doi:10.1002/pmic.201400410.
- Silverstein, K. A. T., Moskal, W. A., Wu, H. C., Underwood, B. A., Graham, M. A., Town, C. D., et al. (2007). Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J.* 51, 262–280. doi:10.1111/j.1365-313X.2007.03136.x.
- Song, X., Li, E., Song, H., Du, G., Li, S., Zhu, H., et al. (2020). Genome-wide identification and characterization of nonspecific lipid transfer protein (nsLTP) genes in *Arachis duranensis*. *Genomics* 112, 4332–4341. doi:10.1016/j.ygeno.2020.07.034.
- Tian, W., Chen, C., Lei, X., Zhao, J., and Liang, J. (2018). CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* 46, W363–W367.
- Voorrips, R. E. (2002). MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J. Hered.* 93, 77–78. doi:10.1093/jhered/93.1.77.
- Wang, H., Wei, T., Wang, X., Zhang, L., Yang, M., Chen, L., et al. (2018). Transcriptome Analyses from Mutant *Salvia miltiorrhiza* Reveals Important Roles for SmGASA4 during Plant Development. *Int. J. Mol. Sci.* 19, 2088. doi:10.3390/ijms19072088.
- Wang, L., Wang, Z., Xu, Y., Joo, S., Kim, S., Xue, Z., et al. (2009). OsGSR1 is involved in crosstalk between gibberellins and brassinosteroids in rice. *Plant J.* 57, 498–510.
- Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49–e49.
- Webster, D. E., and Thomas, M. C. (2012). Post-translational modification of plant-made foreign proteins; glycosylation and beyond. *Biotechnol. Adv.* 30, 410–418. doi:10.1016/j.biotechadv.2011.07.015.
- Wu, T., Cheng, C., Zhong, Y., Lv, Y., Ma, Y., and Zhong, G. (2020). Molecular characterization of the gibberellin-stimulated transcript of GASA4 in Citrus. *Plant Growth Regul.* 91, 89–99. doi:10.1007/s10725-020-00589-1.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* 12, 7–8.
- Yang, Q., Niu, Q., Tang, Y., Ma, Y., Yan, X., Li, J., et al. (2019). PpyGAST1 is potentially involved in bud dormancy release by integrating the GA biosynthesis and ABA signaling in ‘Suli’ pear (*Pyrus pyrifolia* White Pear Group). *Environ. Exp. Bot.* 162, 302–312.

doi:10.1016/j.envexpbot.2019.03.008.

- Yang, S., Höti, N., Yang, W., Liu, Y., Chen, L., Li, S., et al. (2017). Simultaneous analyses of N-linked and O-linked glycans of ovarian cancer cells using solid-phase chemoenzymatic method. *Clin. Proteomics* 14, 3.
- Yang, S., Zhang, X., Yue, J.-X., Tian, D., and Chen, J.-Q. (2008). Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics* 280, 187–198.
- Zhang, S., and Wang, X. (2008). Expression pattern of GASA, downstream genes of DELLA, in Arabidopsis. *Chinese Sci. Bull.* 53, 3839–3846. doi:10.1007/s11434-008-0525-9.
- Zhang, S., and Wang, X. (2011). Overexpression of GASA5 increases the sensitivity of Arabidopsis to heat stress. *J. Plant Physiol.* 168, 2093–2101.
- Zhang, S., Yang, C., Peng, J., Sun, S., and Wang, X. (2009). GASA5, a regulator of flowering time and stem growth in Arabidopsis thaliana. *Plant Mol. Biol.* 69, 745–759. doi:10.1007/s11103-009-9452-7.
- Zheng, L., Ying, Y., Wang, L., Wang, F., Whelan, J., and Shou, H. (2010). Identification of a novel iron regulated basic helix-loop-helix protein involved in Fe homeostasis in Oryza sativa. *BMC Plant Biol.* 10, 166.
- Zimmermann, R., Sakai, H., and Hochholdinger, F. (2010). The Gibberellic Acid Stimulated-Like gene family in maize and its role in lateral root development. *Plant Physiol.* 152, 356–365. doi:10.1104/pp.109.149054.

Figure legends

Figure 1. Location of *tcGASA* genes on the chromosome in *Theobroma cacao*. Each gene is shown with a gene identification number along with the number given to each gene in the current study. The pairs of segmentally duplicated genes are shown with same color.

Figure 2. 3D structure analyses of *tcGASA* proteins.

Figure 3. Post translational phosphorylation and glycosylation modification of *tcGASA* proteins. The circle is drawn on proteins in which the glycosylation site was predicted.

Figure 4. The phylogenetic inference of *tcGASA* proteins along with protein sequences of six other species. Each sequence is presented with gene number. The start of each sequence contained code for the species as: Tc: *Theobroma cacao*; AT: *Arabidopsis thaliana*; B456: *Gossypium raimondii*; VIT: *Vitis vinifera*; Os: *Oryza sativa* (rice); BRADI: *Brachypodium distachyon*; Zm: *Zea mays* (Maize).

Figure 5. The phylogenetic analyses in comparison to introns-exons in genes and motifs in proteins of *tcGASAs*. (a) Phylogeny of *tcGASA*, (b) introns-exons in *tcGASA* genes, (c) Motifs in *tcGASA* proteins.

Figure 6. Synteny analysis of *GASA* genes. The syntenic blocks of cacao *GASA* genes are compared with six other species. (a) *Arabidopsis*, (b) *Gossypium raimondii*, (c) *Vitis vinifera*, (d) rice, (e) *Brachypodium distachyon*, and (f) maize.

Figure 7. Cis-regulatory elements in promoter regions of *tcGASA* genes. (a) represents the extent of various types of regulatory elements based on function, such as light, hormone, growth, stress, and binding domain excluding the TATA box and CAAT box. (b) Distribution of different types of hormone-related cis-regulatory elements. (c) Distribution of cis-regulatory elements related to various types of environmental stresses.

Figure 8. Expression analyses of orthologous *tcGASA* genes of *Arabidopsis*. (a) Expression in different types of tissues, (b) expression in response to various abiotic stresses.

Figure 9. Expression analyses of *tcGASA* genes in cacao plants after inoculation with *P. megakarya* after 0h, 6h, 24h, and 72h. (a) Nanay (NA-32) susceptible cultivar, (b) Scavina (SCA6) tolerant cultivar.

Figure 10. Venn diagram to represent differential expression in susceptible cultivar Nanay (32) and tolerant Cultivar (SCA6) of cacao. (a) 0h (b) 6h (c)24h (d) 72h

Table 1- List of the identified GASA genes and their characteristics in *Theobroma cacao*

Gene locus ID	Gene ID	Location	Length(aa)	MW (kDa)	pI	Instability index	GRAVY	Subcellular localization
Tc01v2_t001590	<i>TcGASA1</i>	1: 830,757-831,466	105	11.18	8.93	Unstable	-0.023	Extracellular space
Tc01v2_t005890	<i>TcGASA2</i>	1: 3,268,900-3,269,687	107	11.44	8.98	Unstable	-0.057	Extracellular space
Tc02v2_t009150	<i>TcGASA3</i>	2: 5,635,674-5,636,363	90	9.90	9.01	Stable	-0.054	Extracellular space
Tc02v2_t021110	<i>TcGASA4</i>	2: 30,717,916-30,719,013	106	11.91	9.21	Unstable	-0.195	Extracellular space
Tc04v2_t000240	<i>TcGASA5</i>	4: 270,172-272,304	320	33.92	9.83	Unstable	-0.750	Extracellular space
Tc04v2_t006890	<i>TcGASA6</i>	4: 15,324,400-15,325,648	106	11.83	9.32	Unstable	-0.343	Extracellular space
Tc04v2_t009440	<i>TcGASA7</i>	4: 19,321,294-19,322,037	95	10.37	9.26	Unstable	-0.072	Extracellular space
Tc04v2_t016520	<i>TcGASA8</i>	4: 26,297,870-26,298,743	123	13.42	8.27	Unstable	-0.026	Plasma membrane
Tc04v2_t016530	<i>TcGASA9</i>	4: 26,301,782-26,302,647	95	10.62	8.88	Unstable	-0.251	Extracellular space
Tc05v2_t012230	<i>TcGASA10</i>	5: 25,218,365-25,220,141	114	12.76	9.46	Unstable	-0.327	Extracellular space
Tc08v2_t003650	<i>TcGASA11</i>	8: 1,941,175-1,942,036	102	10.92	6.67	Stable	-0.031	Extracellular space
Tc08v2_t003660	<i>TcGASA12</i>	8: 1,943,144-1,944,201	100	10.57	8.66	Unstable	0.004	Extracellular space
Tc08v2_t007890	<i>TcGASA13</i>	8: 4,555,655-4,556,418	88	9.64	9.32	Stable	-0.062	Extracellular space
Tc08v2_t014670	<i>TcGASA14</i>	8: 15,650,553-15,651,438	102	10.98	8.72	Unstable	-0.164	Extracellular space
Tc08v2_t014680	<i>TcGASA15</i>	8: 15,651,664-15,652,892	92	10.01	8.90	Stable	-0.148	Extracellular space
Tc08v2_t014690	<i>TcGASA16</i>	8: 15,653,408-15,654,378	110	12.13	9.64	Unstable	-0.263	Extracellular space
Tc09v2_t020100	<i>TcGASA17</i>	9: 30,243,952-30,245,432	112	12.32	9.46	Unstable	-0.252	Extracellular space

Table 2. Gene duplications, synonymous and non-synonymous substitutions and time of divergence

Gene 1	Gene 2	Duplication type	K_a	K_s	K_a/K_s	p-value	Divergence time (MYA)
<i>TcGASA2</i>	<i>TcGASA3</i>	Segmental	0.2679	2.6607	0.1007	1.04E-06	204.67
<i>TcGASA2</i>	<i>TcGASA13</i>	Segmental	0.1708	2.2495	0.0759	2.52E-08	173.04
<i>TcGASA8</i>	<i>TcGASA9</i>	Segmental	0.1515	0.6514	0.2325	1.30E-06	50.11
<i>TcGASA10</i>	<i>TcGASA17</i>	Segmental	0.283	1.0283	0.2752	3.28E-05	79.1
<i>TcGASA14</i>	<i>TcGASA15</i>	Segmental	0.3234	0.7971	0.4058	1.46E-08	61.32