genes

MDPI

Article

Genome-Wide Identification and Analysis of the MADS-Box Gene Family in *Theobroma cacao*

Qianqian Zhang [1], Sijia Hou [1], Zhenmei Sun [2], Jing Chen [1], Jianqiao Meng [1], Dan Liang [1], Rongling Wu [1] and Yunqian Guo [1,*]

[1] Center for Computational Biology, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, China; awayzqq@163.com (Q.Z.); hsj381552790@163.com (S.H.); shixiaohuaa0201@163.com (J.C.); mjq990521@163.com (J.M.); liangdanyx2014@163.com (D.L.); rwu@bjfu.edu.cn (R.W.)

[2] Institute of Marine Materials Science and Engineering, College of Ocean Science and Engineering, Shanghai Maritime University, Shanghai 201306, China; sun1120817625@163.com

* Correspondence: guoyunqian@bjfu.edu.cn

Abstract: The MADS-box family gene is a class of transcription factors that have been extensively studied and involved in several plant growth and development processes, especially in floral organ specificity, flowering time and initiation and fruit development. In this study, we identified 69 candidate MADS-box genes and clustered these genes into five subgroups (Mα: 11; Mβ: 2; Mγ: 14; Mδ: 9; MIKC: 32) based on their phylogenetical relationships with *Arabidopsis*. Most *TcMADS* genes within the same subgroup showed a similar gene structure and highly conserved motifs. Chromosomal distribution analysis revealed that all the *TcMADS* genes were evenly distributed in 10 chromosomes. Additionally, the cis-acting elements of promoter, physicochemical properties and subcellular localization were also analyzed. This study provides a comprehensive analysis of MADS-box genes in *Theobroma cacao* and lays the foundation for further functional research.

Keywords: MADS-box transcription factors; *Theobroma cacao*; bioinformatics analysis; genome-wide characterization; gene family

Citation: Zhang, Q.; Hou, S.; Sun, Z.; Chen, J.; Meng, J.; Liang, D.; Wu, R.; Guo, Y. Genome-Wide Identification and Analysis of the MADS-Box Gene Family in *Theobroma cacao*. *Genes* **2021**, *12*, 1799. https://doi.org/10.3390/genes12111799

Academic Editor: Serena Aceto

Received: 13 October 2021
Accepted: 12 November 2021
Published: 15 November 2021

1. Introduction

MADS-box genes encode eukaryotic transcription factors that play a prominent role in plant development processes. MADS-box proteins contain a highly conserved DNA-binding MADS-domain of approximately 50–60 amino acids in length in their N-terminal region, and this domain could be involved in recognizing and binding the CArG motif of their target gene [1]. The name itself is given by the initials of the four first-discovered transcription factors in this family, which are *MCMI* in *Saccharomyces cerevisiae* [2], *AGAMOUS* in *Arabidopsis thaliana* [3], *DEFICENS* in *Antirrhinum majus* [4] and *SRF4* in *Homo sapiens* [5]. Based on protein domain structure, the MADS-box genes are divided into two categories: type I and type II. The type I MADS-box genes can be further classified into Mα, Mβ, Mγ, Mδ subclasses. Type II lineage, also known as MIKC type, has a special MIKC structure, which is composed of an N-terminal MADS domain, the I (intervening) and K (keratin-like) regions and a variable C-terminal transcriptional activation domain [6]. Type MIKC were further divided into two subgroups, $MIKC^C$ and MIKC*, according to their MIKC structural features [7].

The MADS-box gene family is known to have functions in many significant physiological and developmental processes, such as the regulation of floral organ specificity [3,4], control of flowering signals and initiation [8,9], fruit development [10], meristem identify specification [11], and seed development [12]. For example, Wheat *VERNALIZATION1* (*VRN1*) is a key regulator of flowering time and floral meristem determination [13] The MADS-box gene *FLOWERING LOCUS C* (*FLC*) controls the vernalization pathway in

Arabidopsis [14]. Apple *MdDAM1* plays a role in bud dormancy and growth cessation in autumn [15]. Although MADS-box genes are well-known for their roles in the flower developmental process and participating in the classical ABC flower development model, some of them have been validated to function on root and leaf morphogenesis [16,17]. To date, the MADS-box proteins have been characterized in various kinds of plants, including *Arabidopsis* [18], *Populus trichocarpa* [19], *pineapple* [20], *Saccharum spontaneum* [21], *Erigeron breviscapus* [22]. However, little is known regarding the MADS-box gene family in *Theobroma cacao*.

Theobroma cacao is an economically important tropical tree, native to South America, which is planted in large quantities for its fruits (cacao pods), where its beans were used as the raw material for making chocolate, coco butter, cosmetics and confectionery [23]. Additionally, some studies have proposed that an ingredient found in coco might exert cardiovascular benefits [24]. Research into the sequencing and assembling genome of *Theobroma cacao* was carried out in 2010 [25], leading to the genome-wide identification and analysis of important gene families such as the NAC domain transcription factor family [26], WRKY transcription factor family [27], and GPX family [28]. The metabolome and transcriptome profiling of the *Theobroma cacao* pods was completed [29]. In this scenario, we conducted a bioinformatics analysis of MADS-box members of *Theobroma cacao* at the gene level. We identified 69 MADS-box gene members, investigated their phylogenetic relationship, classified them, and analyzed gene structures, motifs, and chromosome location. Moreover, subcellular localization and cis-acting elements were also performed. Our results may provide a basis for further functional studies of coco tree genes and references for subsequent research into molecular mechanisms.

2. Materials and Methods

2.1. Identification of MADS-Box Genes in Theobroma cacao

Theobroma cacao genome sequences and annotation files were provided by Ensembl Plants (http://plants.ensembl.org/index.html, accessed on 16 April 2021. The hidden Markov model (HMM) profile of the MADS-domain was retrieved from the Pfam database (release 34.0; http://pfam.xfam.org/, accessed on 16 April 2021) with the accession number 'PF00319' [30].

MADS-box proteins in *Theobroma cacao* were searched using the following two approaches. First, the downloaded HMM profile was employed using the HMMER v3.3.2 program to search proteins containing the MADS-domain. Secondly, to avoid missing candidates, we constructed a new HMM model with proteins with e-value $< 1 \times 10^{-20}$, and ClustalW (version 2.1) was used for multiple sequence alignments [31]. The new model was used to search all *Theobroma cacao* protein sequences using HMMER (version 3.3.2), with a cut-off e-value of 0.05. Additionally, the predicted proteins were invalidated by conducting protein domain searches on the SMART program (http://smart.embl-heidelberg.de/, accessed on 19 June 2021) and NCBI Conserved Domain Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, accessed on 19 June 2021) to confirm the presence of the MADS-domain in all candidate proteins.

2.2. Phylogenetic Analysis and Classification of MADS-Box Genes

To understand the phylogenetic relationship and to classify the MADS-box genes, a rooted neighbor-joining (NJ) phylogenetic tree for *Theobroma cacao* (*TcMADS*) and *Arabidopsis* MADS-box proteins was constructed using MEGA X software (version 10.2.2) [32]. The *TcMADS* gene family was classified according to their phylogenetic relations with corresponding *Arabidopsis* MADS-box members. *Arabidopsis* MADS-box protein sequences were downloaded from TAIR (https://www.arabidopsis.org/, accessed on 24 July 2021) with the accession numbers reported by Parenicová et al [18]. All protein sequences were aligned by Muscle with the default parameters [33]. The Neighbor-Joining method was used, with the following parameters: 1000 replications for bootstrap method, Poission

model, Pairwise deletion. Additionally, an individual phylogenetic tree of *TcMADS* genes was built with the same method and beautified by ggtree [34].

2.3. Conserved Motif and Gene Structure Analysis

Online program MEME (https://meme-suite.org/meme/tools/meme, accessed on 29 July 2021) was applied to analyze the conserved motifs in the MADS-box protein with the following settings: maximum number of motifs 10, minimum motif width 6, maximum motif width 50, number of repetitions any [35]. The intron–exon structure information was contained in the *Theobroma cacao* gtf file downloaded from Ensembl Plants. Conserved motif and gene structure were both visualized by TBtools software (version 0.665).

The online tools ProtParam (https://web.expasy.org/protparam/, accessed on 10 August 2021) and Compute pI/Mw (https://web.expasy.org/compute_pi/, accessed on 10 August 2021) was employed to analyze physicochemical properties including theoretical isoelectric points (PI), average molecular weight (MW), instability index and aliphatic index. Number of amino acids (aa) and open reading frame (ORF) lengths were both found with the ORFfinder website (https://www.ncbi.nlm.nih.gov/orffinder/, accessed on 11 August 2021). The BUSCA program (https://busca.biocomp.unibo.it/, accessed on 4 August 2021) was used to predict TcMADS proteins' subcellular localization (SL).

2.4. Chromosomal Localization and Gene Duplication

The locational information on the chromosomes and chromosome length of *TcMADS* genes was acquired from Ensembl Plant. All identified genes were mapped to 10 chromosomes with MG2C (http://mg2c.iask.in/mg2c_v2.1/, accessed on 25 June 2021) according to their chromosomal positions and relative distance. *TcMADS* gene potential duplication was confirmed based on major criteria as follows: (a) sequence alignment length cover > 75% of longer sequence, and (b) the similarity of the aligned region > 75% [36]. Bio-Linux was used to screened tandem repeat sequences. The TcMADS protein sequences were aligned by MAFFT (version 7.481), and then multiple protein alignment were confirmed and the corresponding DNA sequences were sorted into codon alignments [37], which were used to calculate the Ka/Ks ratios using KaKs calculator Toolbox 2.0 (version 2.0).

2.5. Analysis of Cis-Acting Element in MADS-Box Genes' Promoters

The upstream sequences (2 kb) of *TcMADS* genes' CDS were retrieved from the *Theobroma cacao* genome by TBtools software according to gene ID, and then submitted to PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/, accessed on 5 August 2021) to identify four cis-acting elements, including light-responsive elements, wound-responsive elements, gibberellin-responsive elements, and auxin-responsive elements, after filtering and screening. The variety and quantity of cis-acting elements upstream each gene was found with TBtools.

3. Results

3.1. Identification of MADS-Box Genes in Theobroma cacao

To identify the MADS-box genes in *Theobroma cacao*, two HMM analyses were performed: after removing duplicates, a total of 68 putative MADS proteins were obtained by first HMMER searches, using the MADS domain profile as a query, in the coco tree protein database. For the second HMM analysis, we selected proteins which e-value > 0.05 as candidate members, choosing the longest transcript for each screened gene, and thus generating 69 *MADS-box* genes after confirming MADS domain by SMART and NCBI Conserved Domain Search Service (Supplementary File S1). These 69 MADS-box genes were sequentially renamed from *TcMADS1* to *TcMADS69* based on their chromosomal location and subjected to further analyses. Detailed characteristics, including number of amino acids (aa), average molecular weight (MW), theoretical pI, instability index, and aliphatic index about *TcMADS* genes, are listed in Table 1. The statistical results showed that the protein length varied, ranging from 78 (*TcMADS23*) to 600 (*TcMADS7*) amino

acids, with an average length of amino acids, and the molecular weights varied from 66752.75 Da (*TcMADS23*) to 8995.45 Da (*TcMADS7*). Additionally, thirteen MADS-box proteins were acidic, with pI values less than 6.5; 52 were alkaline, with pI values greater than 7.5; four were neutral, with a pI are between 6.5 and 7.5. The instability index analysis indicated that most of the TcMADS proteins were unstable, with an instability index greater than 40, except for *TcMADS12*, *TcMADS37*, *TcMADS57*, *TcMADS67*, *TcMADS1*, *TcMADS55*, *TcMADS9*, *TcMADS47*. The subcellular localization prediction of *TcMADS* genes was analyzed by BUSCA tools. From the analysis results, most *TcMADS* genes appeared to mainly be located in the nucleus (63.77%) and chloroplast (34.78%), with only *TcMADS11* found in the endomembrane system.

Table 1. Detailed information regarding *MADS-box* gene family in *Theobroma cacao*.

Gene Name	Gene ID	Physicochemical Characteristics					SL	ORF
		PI	MW (Da)	Length (aa)	Instability Index	Aliphatic Index		
TcMADS1	TCM_000239	9.48	46,436.9	406	39.09	77.32	nucleus	1221
TcMADS2	TCM_000266	9.59	26,095.8	224	46.52	87.1	chloroplast	675
TcMADS3	TCM_000725	9.91	36,097.68	315	58.23	83.87	nucleus	948
TcMADS4	TCM_000878	6.85	26,140.89	237	41.11	78.23	chloroplast	714
TcMADS5	TCM_000931	9.12	29,034.14	250	45.33	84.56	chloroplast	753
TcMADS6	TCM_000992	6.55	25,353.79	224	53.32	85.76	nucleus	675
TcMADS7	TCM_001181	5.86	66,752.75	600	48.76	77.52	nucleus	1803
TcMADS8	TCM_001182	5.43	37,831.38	337	62.2	73.77	nucleus	1014
TcMADS9	TCM_001335	9.19	38,279.15	338	38.49	68.11	nucleus	1017
TcMADS10	TCM_001841	9.85	31,109.18	269	62.98	6.17	chloroplast	810
TcMADS11	TCM_005456	8.91	30,163.08	262	53.63	84.89	endomembrane system	789
TcMADS12	TCM_005458	9.08	27,810.84	243	38.49	82.3	nucleus	732
TcMADS13	TCM_005818	8.51	42,280.27	375	42.75	101.09	nucleus	1128
TcMADS14	TCM_006323	9.42	28,732.02	254	52.22	62.24	nucleus	765
TcMADS15	TCM_006324	9.15	27,699.77	243	43.97	85.56	nucleus	732
TcMADS16	TCM_006325	5.42	24,228.12	218	45.89	55.96	nucleus	657
TcMADS17	TCM_007324	9.52	20,330.45	174	48.32	100.29	chloroplast	525
TcMADS18	TCM_007378	8.96	24,754.12	219	51.13	85.11	chloroplast	660
TcMADS19	TCM_007713	7.74	27,574.3	233	66.54	80.73	chloroplast	702
TcMADS20	TCM_007787	9.12	36,022.57	310	46.31	92.13	nucleus	933
TcMADS21	TCM_008703	8.5	29,428.29	258	46.99	82.79	nucleus	777
TcMADS22	TCM_008716	5.92	22,921.21	199	56.26	89.65	nucleus	600
TcMADS23	TCM_008973	9.39	8995.45	78	51.89	96.15	chloroplast	237
TcMADS24	TCM_011475	8.97	28,016.08	241	57.36	80.17	nucleus	726
TcMADS25	TCM_011478	6.61	27,447.96	240	58.38	73.62	nucleus	723
TcMADS26	TCM_011687	6.33	39,385.48	351	47.73	78.66	nucleus	1056
TcMADS27	TCM_012489	6.85	23,710.15	210	40.52	71.57	nucleus	633
TcMADS28	TCM_014051	6.13	41,766.39	370	51.46	77.22	chloroplast	1113
TcMADS29	TCM_014337	8.79	46,247.61	407	52.14	88.87	nucleus	1224
TcMADS30	TCM_014345	9.06	27,737.47	239	50.17	76.78	nucleus	720
TcMADS31	TCM_014661	9.83	24,429.07	210	55.07	86.33	chloroplast	633
TcMADS32	TCM_015044	8.82	27,249.02	236	53.76	86.78	nucleus	711
TcMADS33	TCM_015049	9.88	27,106.38	237	47.23	90.13	chloroplast	714
TcMADS34	TCM_015674	5.47	27,657.45	238	69.7	87.65	nucleus	717
TcMADS35	TCM_016147	9.24	24,830.39	215	63.31	73.95	nucleus	648
TcMADS36	TCM_017242	8.51	24,320.82	209	47.08	86.75	nucleus	630
TcMADS37	TCM_018979	9.07	27,740.63	243	34.06	85.14	nucleus	732
TcMADS38	TCM_018981	8.77	28,366.14	248	62.86	80.24	nucleus	747
TcMADS39	TCM_019362	8.23	30,811.58	267	58.24	70.15	nucleus	804
TcMADS40	TCM_021050	9.7	17,902.07	155	45.29	89.94	chloroplast	468
TcMADS41	TCM_022993	9.51	40,637.63	356	53.16	76.71	chloroplast	1071
TcMADS42	TCM_023006	9.2	38,815.19	354	58.36	70.54	nucleus	1065
TcMADS43	TCM_023041	8.93	38,451.37	354	59.01	69.49	nucleus	1065

Table 1. *Cont.*

Gene Name	Gene ID	Physicochemical Characteristics					SL	ORF
		PI	MW (Da)	Length (aa)	Instability Index	Aliphatic Index		
TcMADS44	TCM_024579	6.04	28,901.4	252	61.39	85.48	nucleus	759
TcMADS45	TCM_025670	8.86	26,594.65	233	64.66	74.12	chloroplast	702
TcMADS46	TCM_025671	9.37	26,384.43	233	63.02	70.39	nucleus	702
TcMADS47	TCM_025674	9.64	16,948.73	150	36.03	79.4	nucleus	453
TcMADS48	TCM_025676	10.29	11,744.9	103	56.26	68.25	chloroplast	312
TcMADS49	TCM_026842	9.26	30,499	273	46.06	71.87	nucleus	822
TcMADS50	TCM_026845	9.47	23,787.52	207	49.15	78.74	chloroplast	624
TcMADS51	TCM_029234	4.91	19,812.05	174	42.71	76.21	nucleus	525
TcMADS52	TCM_029518	9.52	20,156.2	182	44.45	79.84	chloroplast	549
TcMADS53	TCM_029519	9.64	49,499.24	437	50.15	68.56	nucleus	1314
TcMADS54	TCM_029596	9.68	34,012.76	294	66.03	74.05	nucleus	885
TcMADS55	TCM_032402	9.25	13,764.45	132	34.38	60.68	nucleus	399
TcMADS56	TCM_032403	7.74	19,620.34	172	50.98	74.24	chloroplast	519
TcMADS57	TCM_034148	9.08	26,051.7	225	33.21	88.36	chloroplast	678
TcMADS58	TCM_034501	7.64	25,504.16	227	55.81	91.06	nucleus	684
TcMADS59	TCM_034549	9.62	21,512.84	184	50.3	88.42	chloroplast	555
TcMADS60	TCM_034757	5.45	37,593.54	333	59.52	84.83	nucleus	1002
TcMADS61	TCM_034970	5.26	38,476.16	337	60.18	82.43	nucleus	1014
TcMADS62	TCM_035212	8.87	24,432.87	209	61.75	78.42	nucleus	630
TcMADS63	TCM_036473	9.34	18,742.04	162	41.19	99.38	chloroplast	489
TcMADS64	TCM_036541	9.43	25,550.24	222	61.61	91.4	chloroplast	669
TcMADS65	TCM_036568	9.52	23,170.78	203	56.62	87	nucleus	612
TcMADS66	TCM_037394	9.62	21,551.01	186	42.9	92.31	chloroplast	561
TcMADS67	TCM_040735	9.76	24,088.7	214	39.18	87.01	chloroplast	645
TcMADS68	TCM_042799	4.84	26,046.39	233	57.5	68.28	nucleus	702
TcMADS69	TCM_042848	4.81	26,121.69	233	55.41	75.41	nucleus	702

3.2. Phylogenetic Analysis and Classification of the MADS-Box Gene

To understand the phylogenetic relationship among MADS-box genes in the coco tree and group them into the established subfamilies, we employed MEGA X to construct a rooted neighbor-joining phylogenetic tree based on the amino acid sequence alignment of 69 proteins from *Theobroma cacao* and 96 from *Arabidopsis* (Figure 1) [15], which also allowed for inferences to be made about the possible function of these genes based on *Arabidopsis* gene function research. According to the general MADS-box gene classification in *Arabidopsis*, the *TcMADS* genes were grouped into two types: type I and type II. Then, based on the phylogenetic relationships, the type I MADS-box genes were further subdivided into more detailed subfamilies: Mα (11), Mβ (2), Mγ (14), Mδ (9). The Mβ group has the minimum number of members, 2, while the corresponding group members in *Arabidopsis* contains 16, which indicates that genes were lost over the development of evolution. and the remaining 32 members were classified as MIKC type II. It is notable that *TcMADS39* is not classified into any of these subfamilies; therefore, we group it as UN.

3.3. Conserved Motif and Structure Analysis

To gain insights into the structural diversity and similarity of MADS-box genes in coco tree, we analyzed the intron–exon arrangements and conserved motifs according to their phylogenetic relations. As shown in Figure 2A, we first constructed an individual phylogenetic tree using an NJ method similar to that of the species tree described above, and then mapped their intron–exon structure (Figure 2B). A very striking distribution of introns in the *Arabidopsis* MADS-box genes was previously reported: the MICK subfamily of *TcMADS* genes contained multiple introns, as did the Mδ group, whereas the remaining three subfamilies (Mα, Mβ and Mγ) usually had no introns, or only one or two introns. The reason the Mα, Mβ and Mγ groups contain fewer introns might be a differential tendency to lose or acquire introns or a reverse-transcribed origin for the ancestors of the three subfamilies [18]. In our study, the number of introns in *TcMADS* genes ranged from one (*TcMADS14*, *TcMADS15*, *TcMADS4*, *TcMADS63*, *TcMADS47*, *TcMADS49*) to eighteen (*TcMADS7*). Furthermore, closely related genes have a similar gene structure, differing

only in the length of exons and introns. The shortest *TcMADS* gene was just 237 bp in length (*TcMADS23*), while the longest gene was *TcMADS7*, with a length of 1803 bp.

Figure 1. Phylogenetic tree of MADS-box genes in *Arabidopsis* and *Theobroma cacao*. The MADS-box genes are indicated with light pink and light green shade for *Arabidopsis* and *Theobroma cacao*, respectively. In second (narrow) ring from outside, the size of the area represented by the two colors shows the proportion of genes from two species in each group. The subgroups are marked by colorful background and circles.

To further study the characteristics of the MADS-box gene family and the conserved motifs that are shared among different subfamilies in *Theobroma cacao*, Multiple Expectation Maximization for Motif Elicitation program was used to identify the conserved motifs. A total of 10 conservative motifs were predicted and named from Motif 1 to Motif 10 (Figure 2C). Among these motifs, Motif 1 was prevalent in all genes; it is worth noting that there were only two Motif 1s in *TcMADS53*. Motif 2 was also present in almost *TcMADS* genes. Motif 3, Motif 8, Motif 5, Motif 9 and Motif 10 were only observed in the Mγ subfamily, which indicated that they might be unique to the Mγ group. Generally, *TcMADS* genes of the same subfamily had similar motifs; we speculated that they might have a similar biological function.

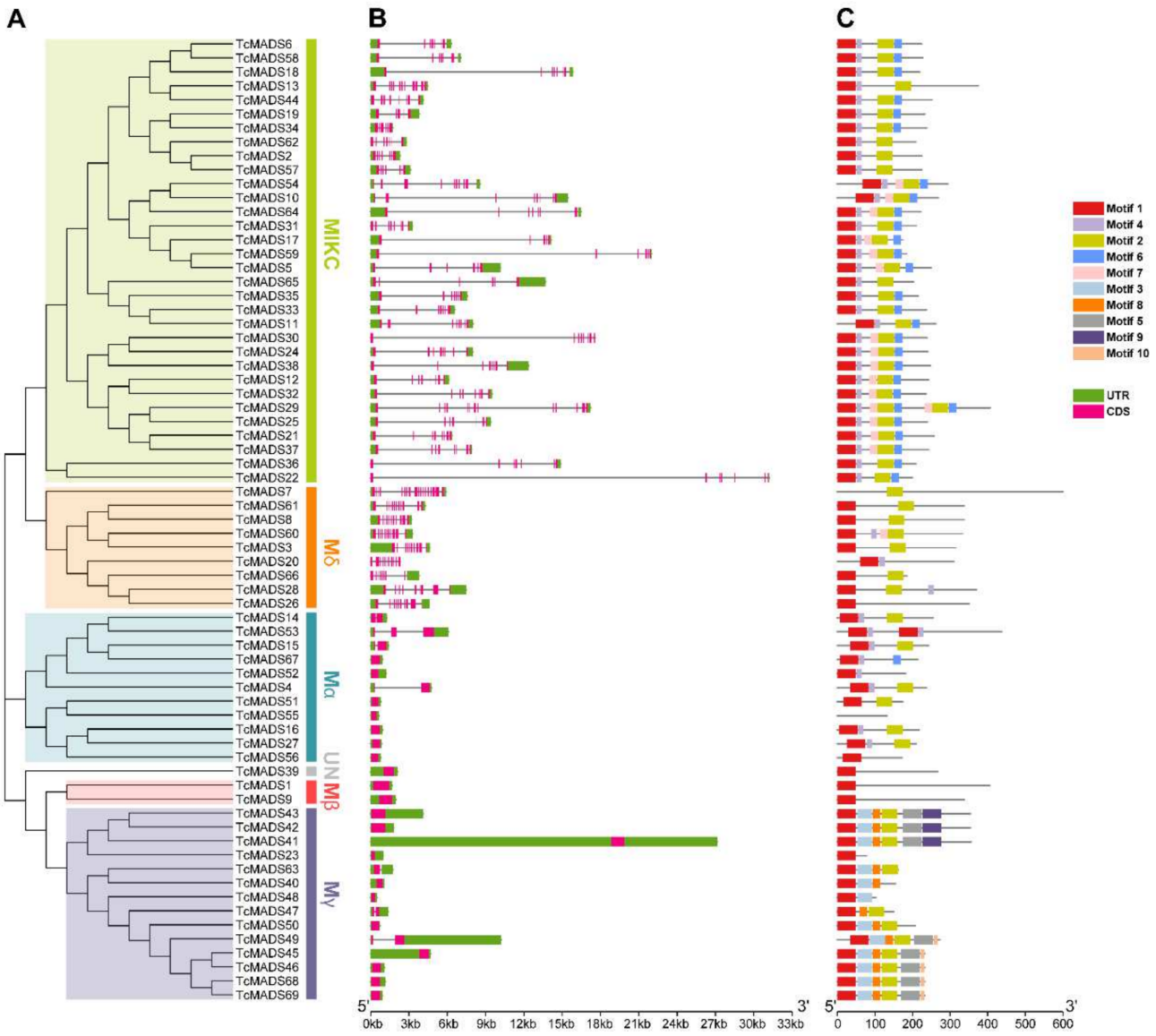

Figure 2. Phylogenetic relationship, gene structure and conserved motifs of the *TcMADS* genes. (**A**) An unrooted NJ tree (left side of the figure) obtained using the MEGA X based on coco tree MADS-box protein sequences. (**B**) The exon–intro structures of *Theobroma cacao* MADS-box genes (central of the figure) were displayed by TBtools software. (**C**) Conserved motif composition of the TcMADS proteins (right side). Detailed information on the ten motifs is provided in Supplementary Table S1.

3.4. Genome Distribution and Gene Evolution Analysis of TcMADS Genes

According to the location information acquired from genome annotation file downloaded in Ensembl Plants database, 69 *TcMADS* genes were evenly distributed on 10 chromosomals (Figure 3A) and renamed based on their position on the chromosome. A higher abundance of MADS-box genes (18.84%) of coco tree was observed on chromosome (Chr) I and II, whereas ChrVII, ChrXI, ChrX had only two MADS-box genes (2.90%). As shown in Figure 3B, a chromosomal bias was observed in the distribution of Mγ subfamily, which was mainly confined to ChrV. ChrIII and ChrVIII were both contained nine *TcMADS* genes. The other MADS-box genes of *Theobroma cacao* were located as follows: 5, 10 and 4 on ChrIV, ChrV and ChrVI, respectively.

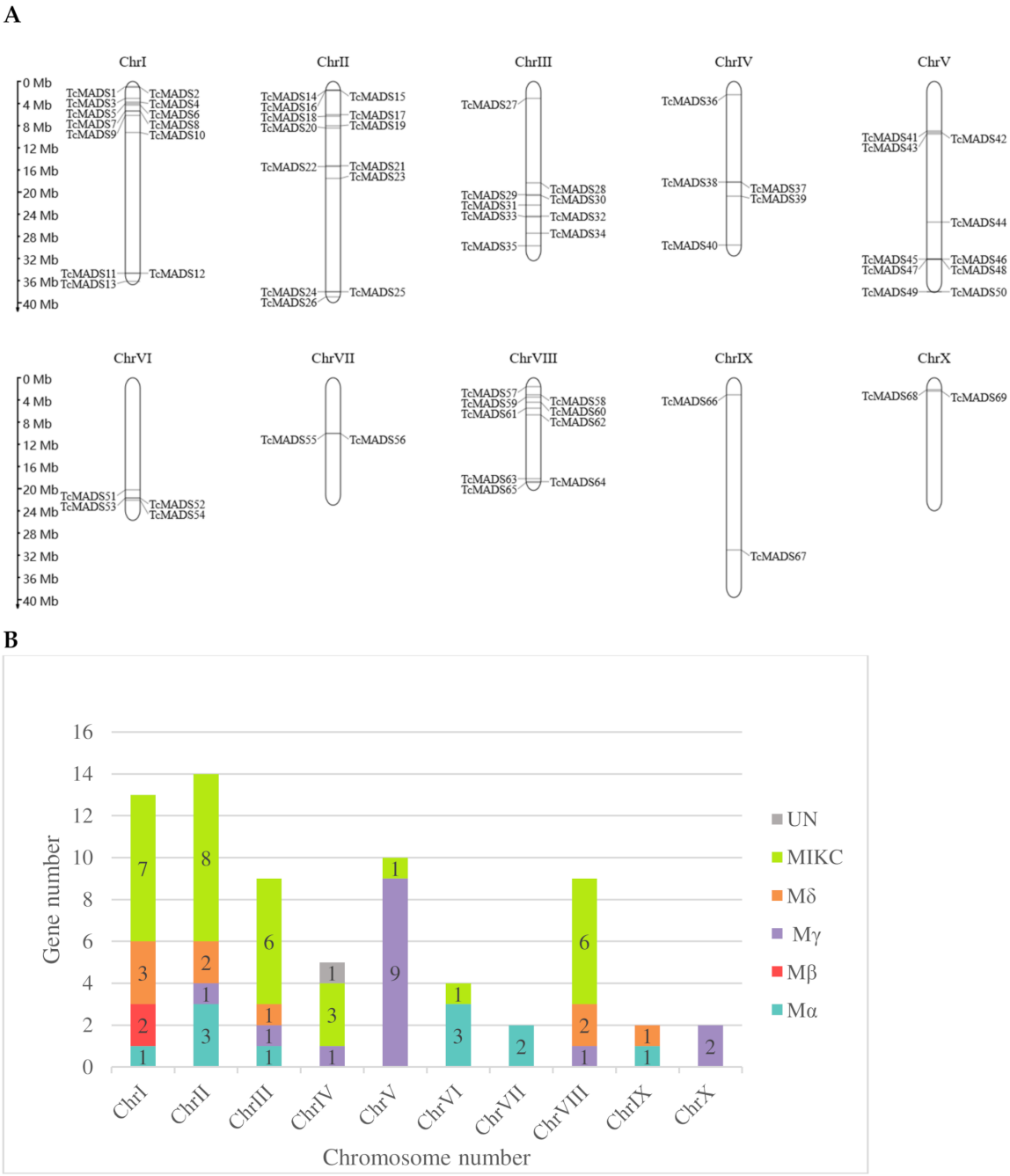

Figure 3. (**A**) Physical distribution of *TcMADS* genes among 10 chromosomes. (**B**) Number of TcMADS subfamily on each chromosome.

Some of the MADS-box genes distribution showed a relatively high density on chromosomes. We screened tandem duplicated gene pairs among sixty-nine *TcMADS* genes. The analysis showed that three genes (*TcMADS41*, *TcMADS42*, *TcMADS43*) on ChrV are duplications of each other, and two gene pairs were also found on ChrV (*TcMADS49*&*TcMADS50*, *TcMADS45*&*TcMADS46*) and one pair on ChrII (*TcMADS68*&*TcMADS69*). Additionally, the substitution ratio of non-synonymous (Ka) to synonymous (Ks) mutations (Ka/Ks) of above six pairs were calculated. As shown in Table 2, Ka/Ks values of *TcMADS43*&*TcMADS42* and *TcMADS43*&*TcMADS41* > 1, which means that these genes were positively selected over the course of evolution and the new protein functions could be beneficial to the survival and reproduction of the coco tree. The remaining four gene pairs had Ka/Ks < 1, indicating that these duplicated gene pairs evolved under purifying selection.

Table 2. Tandem duplicated gene pairs and their Ka, Ks, Ka/Ks values.

Tandem Duplicated Gene Pairs	Chromosome	Ka	Ks	Ka/Ks
TcMADS43&*TcMADS42*	ChrV	0.108301	0.103758	1.04379
TcMADS43&*TcMADS41*	ChrV	0.213254	0.170408	1.25143
TcMADS42&*TcMADS41*	ChrV	0.215878	0.236528	0.912695
TcMADS49&*TcMADS50*	ChrV	0.098915	0.237876	0.415826
TcMADS45&*TcMADS46*	ChrV	0.078058	0.111367	0.700902
TcMADS68&*TcMADS69*	ChrII	0.035449	0.086453	0.410042

3.5. Analysis of Putative Promoter Regions in TcMADS Genes

The cis-regulatory elements serve as a molecular switch by binding to transcription factors, which are associated with gene transcription initiation and transcription activity. To explore the putative functions of *TcMADS* genes, we extracted and examined the 2k bp sequences upstream the transcription start site. Four types of cis-acting elements were present in the promoter regions when submitted to PlantCARE Online program, including a light-responsive element, wound-responsive element, gibberellin-responsive element and auxin-responsive element. These were identified in our study, indicating that *TcMADS* genes are closely related to abiotic stress response. The distribution of these cis-acting elements on the promoters is shown in Supplementary Figure S1. Light-responsive elements were present in almost all promoter regions of MADS-box genes, with an especially large number in *TcMADS28*, *TcMADS46*, *TcMADS62*, *TcMADS65*.

4. Discussion

MADS-box proteins are major transcription factors involved in almost every biological process, and a surprising number of them have been systematically identified and analyzed in a variety of species. Although many MADS-box genes have been shown to have conserved functions in flower development and fruit ripening [38–41], some MADS-box genes have acquired novel functions in specific species during evolution [42]. To date, no detailed analysis of MADS-box genes has been performed in *Theobroma coco*. A better understanding of this family in terms of their member feature, structure characteristics can provide new ideas for further functional analysis. Compared with previous studies, the number of this family member varies in different species, with 107 in *Arabidopsis* [18], 105 in *populus trichocarpa* [19], 48 in *pineapple* [20], 182 in *Saccharum spontaneum* [21], 44 in *Erigeron breviscapus* [19,22], 64 in *Salix suchowensis* [43]. A total of 69 MADS-box proteins were identified from the coco tree in this study (Table 1), which is less than that in *Arabidopsis*. One possible explanation for this is that MADS-box genes coco tree may have a higher gene loss rate compared to that of *Arabidopsis*, indicating an important role of gene duplication over the course of evolution in various species [44]. These 69 *TcMADS* genes were renamed (*TcMADS1-TcMADS69*) based on their chromosomal location and further classified two types according to their phylogenetic relationship with *Arabidopsis*: type I including subclass Mα (11 genes), Mβ (2 genes), Mγ (14 genes), Mδ (9 genes) and type II MIKC (32 genes). The remaining *TcMADS* gene was classified as group UN. We found that most MADS-box genes belong to the MIKC subfamily, and *Theobroma coco* had a comparable number of Mδ and Mγ genes but fewer Mα, Mβ and MIKC genes than *Arabidopsis*, meaning that *Arabidopsis* may undergo more gene duplication events than *Theobroma cacao*. The structures of two types of MADS-box genes were obviously different, and MIKC subgroup genes were more conservative compared with other groups. Additionally, Type II genes usually have multiple introns, whereas most Mα, Mβ and Mγ members have fewer or no introns, indicating that these genes may experience more intron loss during gene family diversification. Previous studies proposed that the number of gene introns correlates with the expression level: the fewer the introns, the higher the pression [45,46]. The same pattern of intron–exon structures in type I and type II exists among diverse species including watermelon [47], *Brachypodium distachyon* [48], rice [49], and lettuce [50].

Overall, genes within the same group are structurally different from other genes; therefore, we speculated that there may be a complicated gene structural evolution in *TcMADS* genes.

Phylogenomic analyses shows that gene and genome duplication events usually contribute to the diversification of the MADS-box transcription factor and play significant roles in shaping the regulatory networks involved in key phenotypic characters [51]. In this study, six tandem-duplicated gene pairs were identified, which all belong to the Mδ subfamily. As ubiquitous genetic components, promoters drive gene transcription and precisely, temporarily and spatially control gene in response to developmental and environmental signals [52]. The cis-acting elements located upstream of the transcription start sites play a vital biological role in regulating gene expression during growth and development [53]. A promoter analysis indicated that *TcMADS* genes are involved in diverse stress and hormone responses, making it possible to study individual gene function.

5. Conclusions

In this study, a systematic analysis was conducted of the Theobroma MADS-box gene family. Based on the *Theobroma cacao* genome data, we used HMM profiles to identify 69 MADS-box genes. Sixty-nine MADS-box genes were distributed across 10 chromosomes and phylogenetically classified into six subfamilies, which showed high similarity in terms of gene structure and conserved motifs within the same subfamily. Furthermore, cis-acting elements analysis show that *TcMADS* genes may be involved in diverse stress responses. In summary, these results provided more information about MADS-box genes and establish a foundation for future study of MADS-box genes in *Theobroma cacao*.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12111799/s1, File S1: MADS-box gene sequences identified in *Theobroma cacao* in this study; Table S1: Sequence logs of ten conserved motifs in *Theobroma cacao*; Figure S1: The predicted cis-regulatory elements in promoters of *TcMADS* genes.

Author Contributions: Conceptualization, Q.Z.; methodology, Q.Z.; software, Q.Z. and S.H.; formal analysis, S.H. and Z.S.; investigation, Q.Z.; resources, Q.Z. and J.C.; writing—original draft preparation, Q.Z.; writing—review and editing, Y.G. and D.L.; visualization, Q.Z. and J.M.; supervision, Q.Z.; project administration, R.W. and Y.G.; funding acquisition, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (31370669).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available on reasonable request.

Acknowledgments: We are thankful to Ang Dong (Center for Computational Biology, College of Biological Sciences and Technology, Beijing Forestry University) and Shiya Shen (Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, College of Biological Sciences and Technology Beijing Forestry University) for the kindly technical support. This work was supported by a grant from the National Natural Science Foundation of China (31370669).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Bodt, S.; Raes, J.; Van de Peer, Y.; Theissen, G. And then there were many: MADS goes genomic. *Trends Plant Sci.* **2003**, *8*, 475–483. [CrossRef] [PubMed]
2. Passmore, S.; Elble, R.; Tye, B.K. A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. *Genes Dev.* **1989**, *3*, 921–935. [CrossRef]
3. Yanofsky, M.F.; Ma, H.; Bowman, J.L.; Drews, G.N.; Feldmann, K.A.; Meyerowitz, E.M. The protein encoded by the *Arabidopsis* homeotic gene agamous resembles transcription factors. *Nature* **1990**, *346*, 35–39. [CrossRef]
4. Sommer, H.; Beltrán, J.P.; Huijser, P.; Pape, H.; Lönnig, W.E.; Saedler, H.; Schwarz-Sommer, Z. *Deficiens*, a homeotic gene involved in the control of flower morphogenesis in *Antirrhinum majus*: The protein shows homology to transcription factors. *EMBO J.* **1990**, *9*, 605–613. [CrossRef] [PubMed]

5. Norman, C.; Runswick, M.; Pollock, R.; Treisman, R. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. *Cell* **1988**, *55*, 989–1003. [CrossRef]
6. Kaufmann, K.; Melzer, R.; Theissen, G. MIKC-type MADS-domain proteins: Structural modularity, protein interactions and network evolution in land plants. *Gene* **2005**, *347*, 183–198. [CrossRef]
7. Henschel, K.; Kofuji, R.; Hasebe, M.; Saedler, H.; Münster, T.; Theissen, G. Two ancient classes of MIKC-type MADS-box genes are present in the *moss Physcomitrella patens*. *Mol. Biol. Evol.* **2002**, *19*, 801–814. [CrossRef] [PubMed]
8. Li, D.; Liu, C.; Shen, L.; Wu, Y.; Chen, H.; Robertson, M.; Helliwell, C.A.; Ito, T.; Meyerowitz, E.; Yu, H. A repressor complex governs the integration of flowering signals in *Arabidopsis*. *Dev. Cell* **2008**, *15*, 110–120. [CrossRef]
9. Deng, W.; Ying, H.; Helliwell, C.A.; Taylor, J.M.; Peacock, W.J.; Dennis, E.S. *FLOWERING LOCUS C (FLC)* regulates development pathways throughout the life cycle of *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6680–6885. [CrossRef] [PubMed]
10. Ferrándiz, C.; Liljegren, S.J.; Yanofsky, M.F. Negative regulation of the *SHATTERPROOF* genes by *FRUITFULL* during *Arabidopsis* fruit development. *Science* **2000**, *289*, 436–438. [CrossRef] [PubMed]
11. Ferrándiz, C.; Gu, Q.; Martienssen, R.; Yanofsky, M.F. Redundant regulation of meristem identity and plant architecture by *FRUITFULL*, *APETALA1* and *CAULIFLOWER*. *Development* **2000**, *127*, 725–734. [CrossRef] [PubMed]
12. Köhler, C.; Hennig, L.; Spillane, C.; Pien, S.; Gruissem, W.; Grossniklaus, U. The Polycomb-group protein MEDEA regulates seed development by controlling expression of the *MADS-box* gene *PHERES1*. *Genes Dev.* **2003**, *17*, 1540–1553. [CrossRef] [PubMed]
13. Li, C.; Lin, H.; Chen, A.; Lau, M.; Jernstedt, J.; Dubcovsky, J. Wheat *VRN1*, *FUL2* and *FUL3* play critical and redundant roles in spikelet development and spike determinacy. *Development* **2019**, *146*, dev175398. [CrossRef] [PubMed]
14. Michaels, S.D.; Amasino, R.M. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell.* **1999**, *11*, 949–956. [CrossRef] [PubMed]
15. Moser, M.; Asquini, E.; Miolli, G.V.; Weigl, K.; Hanke, M.V.; Flachowsky, H.; Si-Ammour, A. The *MADS-box* gene *MdDAM1* controls growth cessation and bud dormancy in Apple. *Front. Plant Sci.* **2020**, *11*, 1003. [CrossRef] [PubMed]
16. Gan, Y.; Filleur, S.; Rahman, A.; Gotensparre, S.; Forde, B.G. Nutritional regulation of *ANR1* and other root-expressed MADS-box genes in *Arabidopsis thaliana*. *Planta* **2005**, *222*, 730–742. [CrossRef]
17. Kutter, C.; Schöb, H.; Stadler, M.; Meins, F.J.; Si-Ammour, A. MicroRNA-mediated regulation of stomatal development in *Arabidopsis*. *Plant Cell.* **2007**, *19*, 2417–2429. [CrossRef]
18. Parenicová, L.; de Folter, S.; Kieffer, M.; Horner, D.S.; Favalli, C.; Busscher, J.; Cook, H.E.; Ingram, R.M.; Kater, M.M.; Davies, B.; et al. Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: New openings to the MADS world. *Plant Cell.* **2003**, *15*, 1538–1551. [CrossRef]
19. Leseberg, C.H.; Li, A.; Kang, H.; Duvall, M.; Mao, L. Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* **2006**, *378*, 84–94. [CrossRef]
20. Zhang, X.; Fatima, M.; Zhou, P.; Ma, Q.; Ming, R. Analysis of MADS-box genes revealed modified flowering gene network and diurnal expression in pineapple. *BMC Genom.* **2020**, *21*, 8. [CrossRef] [PubMed]
21. Fatima, M.; Zhang, X.; Lin, J.; Zhou, P.; Zhou, D.; Ming, R. Expression profiling of MADS-box gene family revealed its role in vegetative development and stem ripening in *S. spontaneum*. *Sci. Rep.* **2020**, *10*, 20536. [CrossRef]
22. Tang, W.; Tu, Y.; Cheng, X.; Zhang, L.; Meng, H.; Zhao, X.; Zhang, W.; He, B. Genome-wide identification and expression profile of the MADS-box gene family in *Erigeron breviscapus*. *PLoS ONE* **2019**, *14*, e0226599. [CrossRef]
23. Mustiga, G.M.; Gezan, S.A.; Phillips-Mora, W.; Arciniegas-Leal, A.; Mata-Quirós, A.; Motamayor, J.C. Phenotypic description of *Theobroma cacao L.* for yield and vigor traits from 34 hybrid families in Costa Rica based on the genetic basis of the parental population. *Front. Plant Sci.* **2018**, *9*, 808. [CrossRef] [PubMed]
24. Corti, R.; Flammer, A.J.; Hollenberg, N.K.; Lüscher, T.F. Cocoa and cardiovascular health. *Circulation* **2009**, *119*, 1433–1441. [CrossRef]
25. Argout, X.; Salse, J.; Aury, J.M.; Guiltinan, M.J.; Droc, G.; Gouzy, J.; Allegre, M.; Chaparro, C.; Legavre, T.; Maximova, S.N. The genome of *Theobroma cacao*. *Nat. Genet.* **2011**, *3*, 101–108. [CrossRef]
26. Shen, S.; Zhang, Q.; Shi, Y.; Sun, Z.; Zhang, Q.; Hou, S.; Wu, R.; Jiang, L.; Zhao, X.; Guo, Y. Genome-wide analysis of the NAC Domain transcription factor gene family in *Theobroma cacao*. *Genes* **2019**, *11*, 35. [CrossRef] [PubMed]
27. Silva Monteiro de Almeida, D.; Oliveira Jordão do Amaral, D.; Del-Bem, L.E.; Bronze Dos Santos, E.; Santana Silva, R.J.; Peres Gramacho, K.; Vincentz, M.; Micheli, F. Genome-wide identification and characterization of cacao WRKY transcription factors and analysis of their expression in response to witches' broom disease. *PLoS ONE* **2017**, *12*, e0187346.
28. Martins Alves, A.M.; Pereira Menezes Reis, S.; Peres Gramacho, K.; Micheli, F. The glutathione peroxidase family of *Theobroma cacao*: Involvement in the oxidative stress during witches' broom disease. *Int. J. Biol. Macromol.* **2020**, *164*, 3698–3708. [CrossRef] [PubMed]
29. Li, F.; Wu, B.; Yan, L.; Qin, X.; Lai, J. Metabolome and transcriptome profiling of *Theobroma cacao* provides insights into the molecular basis of pod color variation. *J. Plant Res.* **2021**, *134*, 1323–1334. [CrossRef]
30. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [CrossRef]
31. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [CrossRef] [PubMed]

32. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
33. Edgar, R.C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **2004**, *5*, 113. [CrossRef] [PubMed]
34. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinform.* **2020**, *69*, e96. [CrossRef] [PubMed]
35. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef] [PubMed]
36. Yang, S.; Zhang, X.; Yue, J.X.; Tian, D.; Chen, J.Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genom.* **2008**, *280*, 187–198. [CrossRef]
37. Suyama, M.; Torrents, D.; Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **2006**, *34*, W609–W612. [CrossRef]
38. Qi, X.; Liu, C.; Song, L.; Li, M. *PaMADS7*, a MADS-box transcription factor, regulates sweet cherry fruit ripening and softening. *Plant Sci.* **2020**, *301*, 110634. [CrossRef]
39. Liu, J.H.; Xu, B.Y.; Zhang, J.; Jin, Z.Q. The interaction of MADS-box transcription factors and manipulating fruit development and ripening. *Yi Chuan* **2010**, *32*, 893–902. [PubMed]
40. Vrebalov, J.; Ruezinsky, D.; Padmanabhan, V.; White, R.; Medrano, D.; Drake, R.; Schuch, W.; Giovannoni, J. A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. *Science* **2002**, *296*, 343–346. [CrossRef]
41. Fujisawa, M.; Nakano, T.; Shima, Y.; Ito, Y. A large-scale identification of direct targets of the tomato MADS box transcription factor *RIPENING INHIBITOR* reveals the regulation of fruit ripening. *Plant Cell* **2013**, *25*, 371–386. [CrossRef]
42. Smaczniak, C.; Immink, R.G.; Angenent, G.C.; Kaufmann, K. Developmental and evolutionary diversity of plant MADS-domain factors: Insights from recent studies. *Development* **2012**, *139*, 3081–3098. [CrossRef]
43. Qu, Y.; Bi, C.; He, B.; Ye, N.; Yin, T.; Xu, L.A. Genome-wide identification and characterization of the MADS-box gene family in *Salix suchowensis*. *PeerJ* **2019**, *7*, e8019. [CrossRef]
44. Airoldi, C.A.; Davies, B. Gene duplication and the evolution of plant MADS-box transcription factors. *J. Genet. Genom.* **2012**, *39*, 157–165. [CrossRef]
45. Chung, B.Y.; Simons, C.; Firth, A.E.; Brown, C.M.; Hellens, R.P. Effect of 5'UTR introns on gene expression in *Arabidopsis thaliana*. *BMC Genom.* **2006**, *7*, 120. [CrossRef]
46. Jeffares, D.C.; Penkett, C.J.; Bähler, J. Rapidly regulated genes are intron poor. *Trends Genet.* **2008**, *24*, 375–378. [CrossRef]
47. Wang, P.; Wang, S.; Chen, Y.; Xu, X.; Guang, X.; Zhang, Y. Genome-wide analysis of the MADS-box gene family in Watermelon. *Comput. Biol. Chem.* **2019**, *80*, 341–350. [CrossRef] [PubMed]
48. Wei, B.; Zhang, R.Z.; Guo, J.J.; Liu, D.M.; Li, A.L.; Fan, R.C.; Mao, L.; Zhang, X.Q. Genome-wide analysis of the MADS-box gene family in *Brachypodium distachyon*. *PLoS ONE* **2014**, *9*, e84781.
49. Arora, R.; Agarwal, P.; Ray, S.; Singh, A.K.; Singh, V.P.; Tyagi, A.K.; Kapoor, S. MADS-box gene family in rice: Genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genom.* **2007**, *8*, 242. [CrossRef] [PubMed]
50. Ning, K.; Han, Y.; Chen, Z.; Luo, C.; Wang, S.; Zhang, W.; Li, L.; Zhang, X.; Fan, S.; Wang, Q. Genome-wide analysis of MADS-box family genes during flower development in lettuce. *Plant Cell Environ.* **2019**, *42*, 1868–1881. [CrossRef]
51. Shan, H.; Zahn, L.; Guindon, S.; Wall, P.K.; Kong, H.; Ma, H.; DePamphilis, C.W.; Leebens-Mack, J. Evolution of plant MADS box transcription factors: Evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Mol. Biol. Evol.* **2009**, *26*, 2229–2244. [CrossRef] [PubMed]
52. Hernandez-Garcia, C.M.; Finer, J.J. Identification and validation of promoters and cis-acting regulatory elements. *Plant Sci.* **2014**, *217–218*, 109–119. [CrossRef] [PubMed]
53. Ho, C.L.; Geisler, M. Genome-wide computational identification of biologically significant cis-regulatory elements and associated transcription factors from rice. *Plants* **2019**, *8*, 441. [CrossRef] [PubMed]