Plant Omics Journal POJ

POJ 7(4):260-270 (2014) ISSN:1836-3644

Genome-wide *in silico* identification and comparison of Growth Regulating Factor (*GRF*) genes in Cucurbitaceae family

Mehmet Cengiz Baloglu*

Kastamonu University, Faculty of Engineering and Architecture, Department of Genetics and Bioengineering, 37100, Kastamonu, Turkey

***Corresponding author: mcbaloglu@gmail.com**

Abstract

Growth-regulating factor (*GRF*) genes play a regulatory role for plant growth and development. The recently available cucumber, melon and watermelon genomes provide an opportunity to conduct a comprehensive overview of the *GRF* gene family. In the present study, identification and analysis of the *GRF* gene family was conducted using bioinformatics methods. Totally, 24 potential *GRF* genes were identified in cucumber, melon and watermelon. Cucumber and watermelon *GRF* gene members were physically mapped to their corresponding chromosomes. All *GRF* genes contain an intron whose number ranging from 2 to 3. Phylogenetic analysis categorized the cucurbit GRF proteins into 2 distinct classes. GRF proteins of cucurbits and Arabidopsis were clustered together in a joined tree and grouped into the same cluster with high bootstrap values. WRC and QLQ motifs, specific for GRF proteins, were found in all predicted GRF proteins. Gene Ontology analysis showed that majority of the GRFs was predicted to function in response to biological regulation and binding activity. In addition, predicted GRF proteins were localized in the nucleus. These results provide information about the relationship between evolution and functional divergence in the GRF family. We assume that systematic characterization of these *GRF* genes will enable researcher to open new insights for further exploration into the functions of this significant gene family in Cucurbitaceae family members.

Keywords: growth-regulating factor; *Cucumis sativus*; *Cucumis melo*; *Citrullus lanatus*; genome-wide analysis; bioinformatics; phylogenetic relationships.
Abbreviations: GRF_Growth-regulating factor; Transcription factors_TFs; QLQ_Gln, Leu, Gln; WRC_Trp, Arg, Cys; GRF-Interacting Factor_GIF; BLAST_Basic Local Alignment Search Tool; MEME_Motif Elucidation; Hidden Markov model_HMM.

Introduction

Cucurbitaceae family, commonly known as cucurbits has long played important role in economic affairs of human being. The genera included in the family are used for food and medicinal purposes. The family includes such economically and nutritionally important food plants as pumpkin, cucumber, gherkin, watermelon, melon, squash, chayote, and gourd. Totally, 192 million tons of vegetables, fruits, and seeds belonging to cucurbits were annually produced in nine million hectares of land (http://faostat.fao.org). For the study of sex determination, fruit ripening and plant vascular biology, cucurbit family is considered as a model system (Huang et al., 2009). Cucumber is the first cucurbit family plant whose genome sequence published in 2009 (Huang et al., 2009). Then, the other cucurbit family members, melon and watermelon genome sequences have been announced in 2012 (Garcia-Mas et al., 2012) and 2013 (Guo et al., 2013), respectively. Moreover, a variation map of the cucumber genome at single-base resolution was generated by performing deep resequencing of all 115 lines with the wild cucumber genome which was compared to the genome of cultivated cucumber (Qi et al., 2013). These genomic resources enable new insights for understanding the genetic basis of domestication and diversity of these important family members. In addition, the genome sequences of cucurbit family provide an exhaustive phylogenic comparison of the cucumber genome with melon and watermelon. Transcription factors (TFs) are one of the most important control mechanisms for regulation of gene expression. They were classified based on their DNA-binding and multimerization domains (Yamasaki et al., 2008). Among them, growth-regulating factor (GRF) TF family genes are plant-specific and distributed in all plant genomes (Kim et al., 2003). Generally, 2 conserved regions, the QLQ (Gln, Leu, Gln) and WRC (Trp, Arg, Cys) domains, are found in GRF family proteins (Kim et al., 2003; Zhang et al., 2008; Filiz et al., 2014). The QLQ domain resembles to a protein–protein interaction domain of SWI2/SNF2 which is found in the SWI/SNF chromatin-remodeling complex in yeast (Treich et al., 1995). The WRC domain has two functional features; a functional nuclear localization signal and a putative zinc finger motif (Van der Knaap et al., 2000). Recent study showed that GRF-Interacting Factor (GIF) interacted with the QLQ domain of GRF in Arabidopsis and forms a functional complex with GRF which involved in regulation of cell proliferation (Kim and Kende, 2004). GRF TFs genes are generally expressed in growing and developing tissues and may act as a transcription activator and a repressor. In Arabidopsis, GRF1/2/3 genes function as a positive regulator of cell proliferation associated with size and shape of leaf organ (Kim and Kende, 2004). Also, same group showed requirement of GRF4 for both leaf cell proliferation and the embryonic development of cotyledons and the shoot apical meristem (Kim and Lee, 2006). In a recent study, GRF7 repressed a broad range of osmotic

stress–responsive genes to prevent growth inhibition under normal conditions (Kim et al., 2012). However, there is a little information about the genome wide survey and expression patterns of this gene family. The work from Arabidopsis (Kim et al., 2003), rice (Choi et al., 2004), maize (Zhang et al., 2008) and a grass species, *Brachypodium distachyon* (Filiz et al., 2014) are a few examples of genome wide survey studies. On the other hand, a draft of the *Cucumis sativus*, *Cucumis melo* and *Citrullus lanatus* genome sequences were reported recently (Huang et al., 2009; Garcia-Mas et al., 2012; Guo et al., 2013). However, to our knowledge, no GRF TF genes have been identified and isolated in Cucurbitaceae family so far. Moreover, there is a limited data related with genome-wide identification and their characterizations in the genomes of cucumber, melon and watermelon (Baloglu et al., 2014). So, it is important for us to identification of cucumber, melon and watermelon GRF gene family members and comparison of the identified GRF in these three species. In this study, we firstly identified GRF family in three species based on its complete genome sequence analysis by comparing with the members from Arabidopsis, grape, lycophyte (*Selaginella moellendorffii*), maize, papaya, poplar, rice, (*Oryza sativa* subsp. indica and japonica) and sorghum. Subsequently, we have identified genomic distribution, gene structure, conserved motifs and putative biological functions of GRFs. These results will provide valuable information about the relationship between evolution and functional divergence in the GRF family among the cucurbits.

Results

Genome-wide identification of the GRF gene family in the cucumber, melon and watermelon genomes

For identification of *GRF* TF genes in three species, both BLAST and profile hidden Markov model (HMM) searches were performed. Totally 113 GRF proteins from nine plant genomes including Arabidopsis (18), grape (7), lycophyte (*Selaginella moellendorffii*) (4), maize (30), papaya (6), poplar (9), rice, (*Oryza sativa* subsp. indica and japonica) (32) and sorghum (7) were used as query sequences. Then, all predicted GRF proteins were subjected to domain searches to verify presence of QLQ and WRC motifs. By removal of different transcripts of the same gene, we identified 8 putative *GRF* genes for each species (Table 1). For convenience, the *GRF* genes were named from their scientific names, in other words, CsaGRF, CmeGRF and ClaGRF are used for genes from *Cucumis sativus* (cucumber), *Cucumis melo* (melon) and *Citrullus lanatus* (watermelon), respectively. They were also numbered 1 to 8 based on their order on the chromosomes. The *GRF* genes showed significant difference in the size and sequences of their encoded proteins, and their physicochemical properties. Protein length of all three species GRFs varied from 315 to 672 amino acids. In addition, GRF protein sequences had large variations in isoelectric point (pI) values (ranging from 7.04 to 9.75) and molecular weight (ranging from 35.78 kDa to 71.21 kDa). The details of other parameters of GRF protein sequences were summarized in Table 1.

Chromosomal distribution and structure of GRFs

Cucumber and watermelon *GRF* gene members were physically mapped to chromosomes of cucumber and watermelon, respectively. Because of a lack of genome browser function in Melonomics database, we did not achieve mapping of *CmeGRF* genes on melon chromosomes. They were indicated in different scaffolds. In cucumber, chromosome 2 and 3 contain the highest number of *CsaGRF*s, while *ClaGRF* genes were equally distributed on watermelon chromosomes (Fig. 1). The exact position (in bp) of each *GRF* genes for three species is shown in Table 1. Distribution pattern of the *ClaGRF* genes on individual chromosomes also indicated certain physical regions. For example, *ClaGRF* genes located on chromosomes 1, 3, 7, 8, 10, 11 and chromosomes 5 appear to be congregate at the lower end and upper end of the arms, respectively (Fig. 1B). In addition, it was observed that there are no tandem or segmental duplication events among the all *GRF* genes for each species. Based on exon and intron structure analysis, we have detected all *GRF* genes contain an intron. The number of introns in their open reading frames varied from 2 to 3. They were distributed into different classes of the GRF family (Fig. 2).

Phylogenetic classification of GRFs and identification of domain conservation

In order to understand the evolutionary significance of GRF proteins in Cucurbitaceae family and Arabidopsis, 24 GRF proteins from three species and 9 GRF proteins from Arabidopsis were used for construction of phylogenetic tree based on the Neighbor-Joining (NJ) method. The phylogenetic analysis categorized all the GRFs into two discrete groups (Class I and II) comprising of 12 proteins for each classes (Fig. 3). A high bootstrap value (100%) was observed for both groups, which reflects derivation of statistically reliable pairs of possible homologous. To confirm protein homology between cucumber GRFs and melon, watermelon GRFs, BLASTP search was also performed. Cucumber and melon showed above 97% protein homology, while GRF protein homology between cucumber and watermelon varied from 73% to 90 %. It seems that cucumber GRF proteins are close to melon GRFs when compared to watermelon GRF proteins. Additionally, reliability of the phylogeny was further evidenced by parameters like motif compositions. Totally, 5 distinct motifs were detected in all GRF proteins (Table 2, Fig. 4). Motif I and motif II contain WRC and QLQ domains, respectively. Sequence alignment of all GRF proteins was also performed to verify presence of QLQ and WRC domains (Fig. 5). Search and alignment outputs showed the presence of QLQ and WRC motifs in all 8 GRFs for each species. Only CsaGRF-06 and CmeGRF-02 proteins do not have QLQ motif. Besides these two motifs, two other known GRF domains, FFD and TQL, were observed in 24 GRF proteins. Further, one unidentified conserved motifs were found. It was observed that a majority of the members, predicted to have higher protein homology rate, clustered together in phylogenic tree.

Gene ontology annotation

The GO slim analysis was performed using Blast2Go and indicated the putative participation of 24 GRF proteins in diverse biological processes (Fig. 6, Table 3). Total 5 categories of biological processes were defined. Majority of the GRFs were predicted to function in response to biological regulation [34 (~39%)], followed by metabolic process [22 (~25%)]. Molecular function prediction showed that about 83% of the GRF were evidenced to participation of binding activity which may function as assisting protein-protein interactions. Cellular localization prediction indicated that

Table 1. A catalog of 24 Cucurbitaceae family GRF proteins.

ID	Phytozome Identfier	Chromosome	Start position (bp)	End Position (bp)	Protein length (aa)	pI	Molecular weight (kDa)	Instability index	Intron	Phylogeny Group
Cucumis sativus										
CsaGRF-01	Csa012939	1	19492834	19493998	333	9,38	37.82	54.05	2	Class I
CsaGRF-02	Csa009129	2	5885636	5886764	319	8,95	36.81	56.47	2	Class I
CsaGRF-03	Csa011522	2	15375567	15378637	664	7,04	71.21	53.13	3	Class II
CsaGRF-04	Csa001246	2	21781115	21785472	336	8,29	36.93	55.11	2	Class I
CsaGRF-05	Csa000291	3	7016145	7019112	502	9,59	55.04	52.44	3	Class II
CsaGRF-06	Csa018803	3	22199564	22200808	351	7,13	35.78	66.95	2	Class I
CsaGRF-07	Csa005092	3	25563914	25565707	416	8,74	45.08	48.68	3	Class II
CsaGRF-08	Csa012136	6	21890406	21893382	572	7,2	62.05	56.66	3	Class II
Cucumis melo										
CmeGRF-01	MELO3C004650P1	scaffold00003	8907989	8909113	315	8,94	36.33	60.53	2	Class I
CmeGRF-02	MELO3C006174P1	scaffold00006	1479047	1480369	351	9,75	38.34	48.84	1	Class II
CmeGRF-03	MELO3C007656P1	scaffold00007	4405919	4408890	572	7,62	62.12	54.11	3	Class II
CmeGRF-04	MELO3C009444P1	scaffold00011	2247200	2249020	405	8,59	44.38	51.09	3	Class II
CmeGRF-05	MELO3C010786P1	scaffold00014	815613	818693	672	7,04	71.04	54.80	3	Class II
CmeGRF-06	MELO3C015513P1	scaffold00025	2454092	2455260	334	9,34	37.88	53.76	2	Class I
CmeGRF-07	MELO3C024739P1	scaffold00070	433323	437629	344	8,29	37.74	53.14	2	Class I
CmeGRF-08	MELO3C025804P1	scaffold00083	353626	355045	342	7,17	38.20	61.88	3	Class I
Citrullus lanatus										
ClaGRF-01	Cla014050	1	27669694	27672658	571	8,04	62.02	55.90	3	Class II
ClaGRF-02	Cla006802	2	9598025	9602432	334	8,29	36.72	56.60	2	Class I
ClaGRF-03	Cla011169	3	25922407	25923548	327	9,15	37.36	52.72	2	Class I
ClaGRF-04	Cla021292	5	1911214	1914635	501	9,57	55.08	53.39	3	Class II
ClaGRF-05	Cla012578	7	24177658	24179043	350	7,59	39.05	63.77	2	Class I
ClaGRF-06	Cla022646	8	25341218	25344253	651	7,31	69.43	54.16	3	Class II
ClaGRF-07	Cla017716	10	25528058	25530475	440	9,08	47.67	56.68	4	Class II
ClaGRF-08	Cla016859	11	25481111	25482330	333	8,31	38.17	50.99	2	Class I

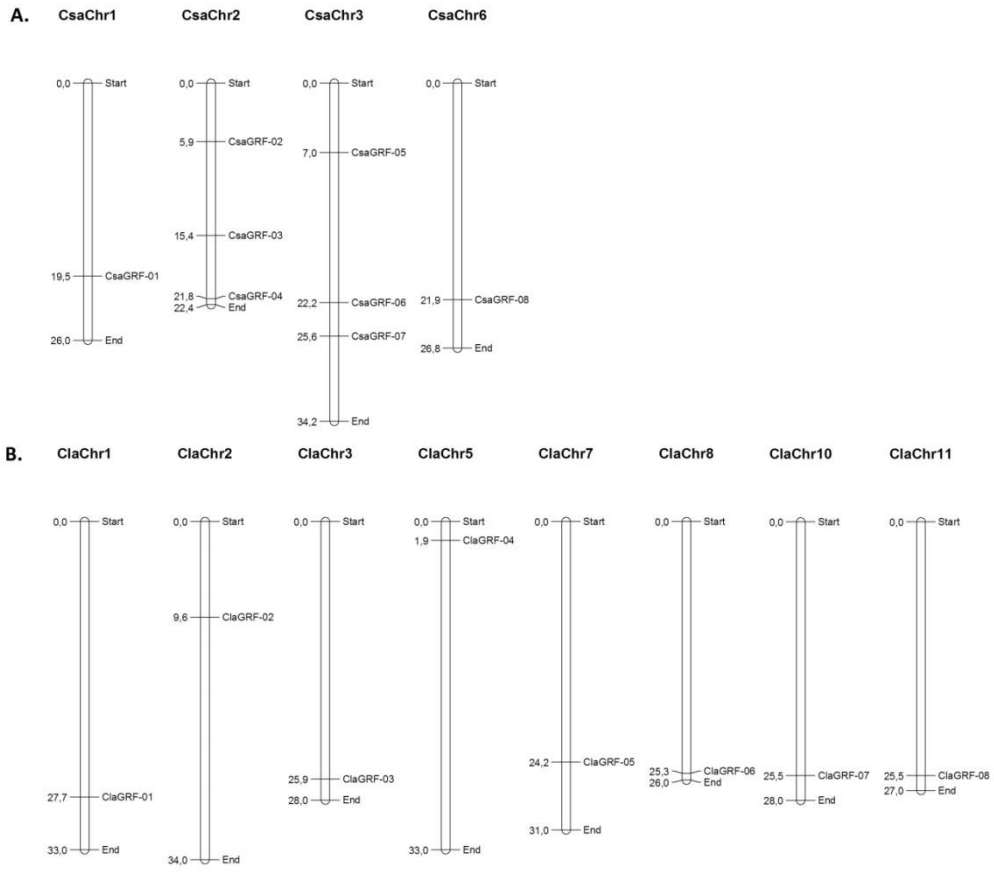

Fig 1. Distribution of 16 *GRF* genes onto cucumber and watermelon chromosomes. Graphical (scaled) representation of physical locations for each **A)** *CsaGRF* genes on cucumber chromosomes and **B)** *ClaGRF* genes on watermelon chromosomes. *CmeGRF* genes were not indicated on melon chromosomes because of a lack of genome browser function in Melonomics database. Chromosomal distances are given in Mb.

majority of GRF proteins (~65%) are localized in the nucleus (Fig. 6).

Discussion

In this study, protein sequences of GRF from nine different plant species were used for prediction of Cucurbitaceae family GRF proteins. Finally, we identified and characterized 24 putative *GRF* genes in the cucumber, melon and watermelon genomes. Over the last decade, the *GRF* gene family has been previously identified and described on a genome-wide level in different plants. First genome wide identification study for *GRF* gene family was performed in Arabidopsis which contains 9 *GRF* genes (Kim et al., 2003). Then, a total of 12 and 14 *GRF* gene family members have been identified in rice and maize genomes, respectively (Choi et al., 2004; Zhang et al., 2008). In a recent study, characterization of 10 *GRF* genes from Brachypodium was announced (Filiz et al., 2014). We also found similar gene numbers in all three species with 8 genes for each. Genome size of Cucurbitaceae family members varied from 367 Mb (cucumber) to 425 Mb (watermelon). When compared to genome size of Cucurbitaceae family members with other species, Brachypodium (300 Mb) and maize (2.5 Gb) genomes size are the smallest and largest ones, respectively (Vain, 2011). It is commonly suggested that exon-intron organization is important for understanding evolutionary and functional relationships (Hu and Liu, 2011). Also, exon or intron gain/loss events provide structural divergence and functional differentiation (Xu et al., 2012). Exon-intron organization of *GRF* genes from Arabidopsis, rice and Brachypodium showed a difference each other. In Arabidopsis, 7 genes contain 3 introns while 2 genes have 2 introns (Kim et al., 2003). Total 2 or 4 introns were found in rice *GRF* genes (Choi et al., 2004). In Brachypodium, intron numbers varied from 1 to 3 (Filiz et al., 2014). Exon-intron structure of the 24 *GRF* genes was also investigated to obtain some insight into their gene structures and to compare other plant *GRF* genes. Similar findings were observed to the previous studies. Averagely, 2 or 3 introns were detected in Cucurbitaceae *GRF* gene family members, which were similar to Arabidopsis. It can be deduced that *GRF* genes from cucumber, melon and watermelon may regard as a similar history among the dicots, which reveals that *GRF* genes in cucurbits genomes were adequately conserved in dicots. Phylogenetic analysis revealed that Cucurbitaceae GRF proteins were closely clustered with high bootstrap values. For example, the highest bootstrap value was observed in members of the Class II, ranging from 97% to 100%. In addition, similar results were obtained in Class I GRF proteins including, ClaGRF-03, CsaGRF-01, CmeGRF-06 and ClaGRF-08, CsaGRF-02, CmeGRF-01. Phylogeny-

Table 2. Conserved motifs identified in Cucurbitaceae family GRF proteins by MEME software.

Motif No.	Sites	e-value	Amino acid sequence composition of motif	Width (aa)	Domain
Motif 1	24	6.7e-1029	K[IM]DPEPGRCRRTDGKKWRCS[KR][DE]A[YV][PA]D[QS]KYCERHM[HN]RG[RK][NH]RSRKPVEG[QS]T[TG]	50	WRC
Motif 2	22	1.1e-543	R[SFG]PFT[AV][AS]QWQELE[HL]QALI[FY]KY[ML]V[AS][GN]VP[VI]P[PS][ED]LL[FI][PS]IK[KR]SL[LE]	41	QLQ
Motif 3	21	1.0e-186	S[QT]QED[DA]E[DPS]Q[KH][PT][LV][HR][HQ]FF[DE][DE]W[PS][PK][KS]Q[RS]DS[WS]LD	29	FFD
Motif 4	21	7.8e-114	[SR]S[NL][TE][TP]S[SV][SD][TA]T[QR]LS[IM]SIPMA[SD]S	21	TQL
Motif 5	18	3.3e-090	P[QP][AP]VGW[GN][SY][FL]QMG[FS][GS][RG]	15	NA

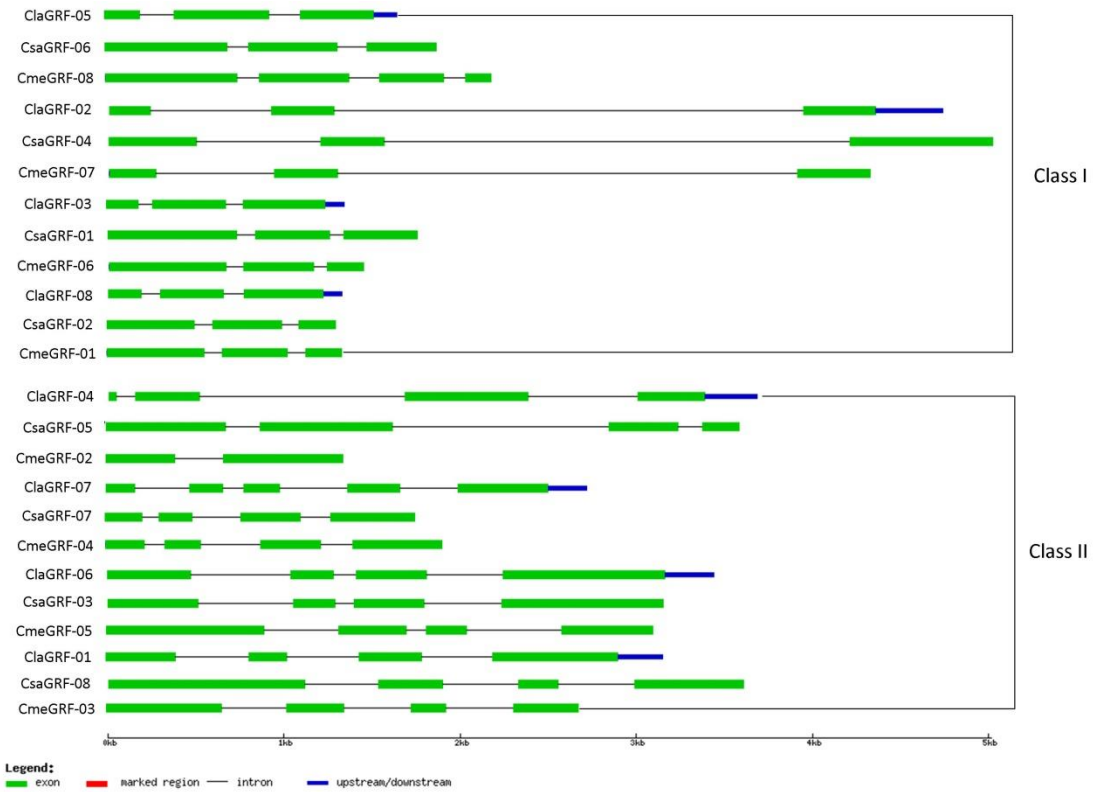

Fig 2. Exon-intron organization of 2 classes of Cucurbitaceae family *GRF* genes. The GRF family was classified according to Fig. 4. Exons and introns are represented by green boxes and black lines, respectively.

Table 3. Blast2Go annotation details of GRF protein sequences.

Gene name	Matched sequences	Query coverage	e value	Similarity	GO: BIOLOGICAL PROCESS	GO: MOLECULAR FUNCTION	GO: CELLULAR COMPONENT
CsaGRF-01	XP_004144026: growth-regulating factor 5-like [Cucumis sativus]	100%	0	100%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-02	XP_004141176: growth-regulating factor 5-like [Cucumis sativus]	100%	0	100%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-03	XP_004171939: growth-regulating factor 6-like [Cucumis sativus]	94%	0	100%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-04	XP_004151910: growth-regulating factor 3-like [Cucumis sativus]	100%	0	100%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-05	XP_004133902: growth-regulating factor 9-like [Cucumis sativus]	97%	0	100%	Regulation of transcription, DNA-dependent, leaf development	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-06	XP_004148588: growth-regulating factor 5-like [Cucumis sativus]	100%	1,7 e-175	99%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-07	XP_004138010: growth-regulating factor 4-like [Cucumis sativus]	100%	0	100%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CsaGRF-08	XP_004143409: growth-regulating factor 1-like [Cucumis sativus]	100%	0	100%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CmeGRF-01	XP_004141176: growth-regulating factor 5-like [Cucumis sativus]	100%	0	94%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CmeGRF-02	XP_004133902: growth-regulating factor 9-like [Cucumis sativus]	94%	0	95%	Response to fructose stimulus, hydrogen peroxide catabolic process, water transport, response to salt stress, glycolysis, response to temperature stimulus	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides, calcium ion binding\ protein phosphorylated amino acid binding	Cytoplasm
CmeGRF-03	XP_004143409: growth-regulating factor 1-like [Cucumis sativus]	100%	0	99%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CmeGRF-04	XP_004138010: growth-regulating factor 4-like [Cucumis sativus]	100%	0	90%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus

CmeGRF-05	XP_004143032: growth-regulating factor 6-like [Cucumis sativus]	95%	0	96%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CmeGRF-06	XP_004144026: growth-regulating factor 5-like [Cucumis sativus]	100%	0	98%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CmeGRF-07	XP_004151910: growth-regulating factor 3-like [Cucumis sativus]	100%	0	97%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
CmeGRF-08	XP_004148588: growth-regulating factor 5-like [Cucumis sativus]	100%	0	94%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-01	XP_004143409: growth-regulating factor 1-like [Cucumis sativus]	100%	0	98%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-02	XP_004151910: growth-regulating factor 3-like [Cucumis sativus]	100%	0	96%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-03	XP_004144026: growth-regulating factor 5-like [Cucumis sativus]	100%	3,4 e-172	93%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-04	XP_004133902: growth-regulating factor 9-like [Cucumis sativus]	97%	0	92%	Regulation of transcription, DNA-dependent, leaf development	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-05	XP_004148588: growth-regulating factor 5-like [Cucumis sativus]	100%	0	91%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-06	XP_004143032: growth-regulating factor 6-like [Cucumis sativus]	97%	0	92%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-07	XP_007202117: hypothetical protein [Prunus persica]	91%	3,1e-82	61%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus
ClaGRF-08	XP_004141176: growth-regulating factor 5-like [Cucumis sativus]	100%	2,0e-154	83%	Regulation of transcription, DNA-dependent	ATP binding, Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Nucleus

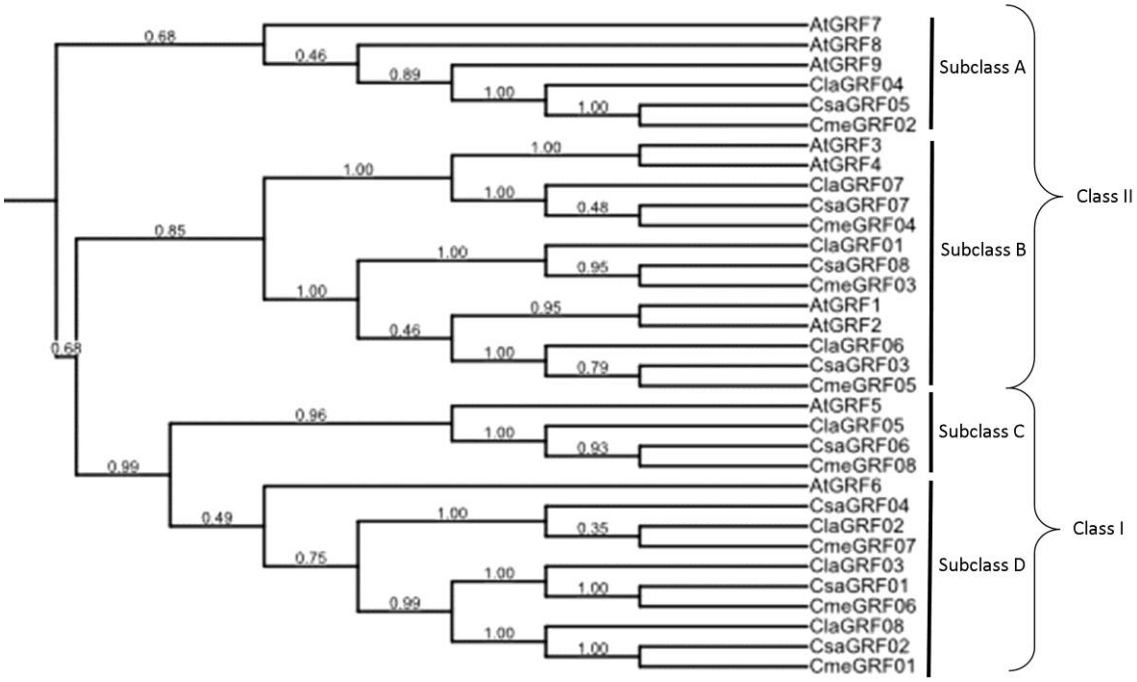

Fig 3. Phylogenetic relationships of cucumber, melon, watermelon and Arabidopsis GRF proteins. The sequences were aligned by ClustalW at MEGA5 and the unrooted phylogenetic tree was deduced by neighbor-joining method. The proteins were classified into two distinct clusters with four subclasses. Bootstrap values were indicated on the phylogenetic tree.

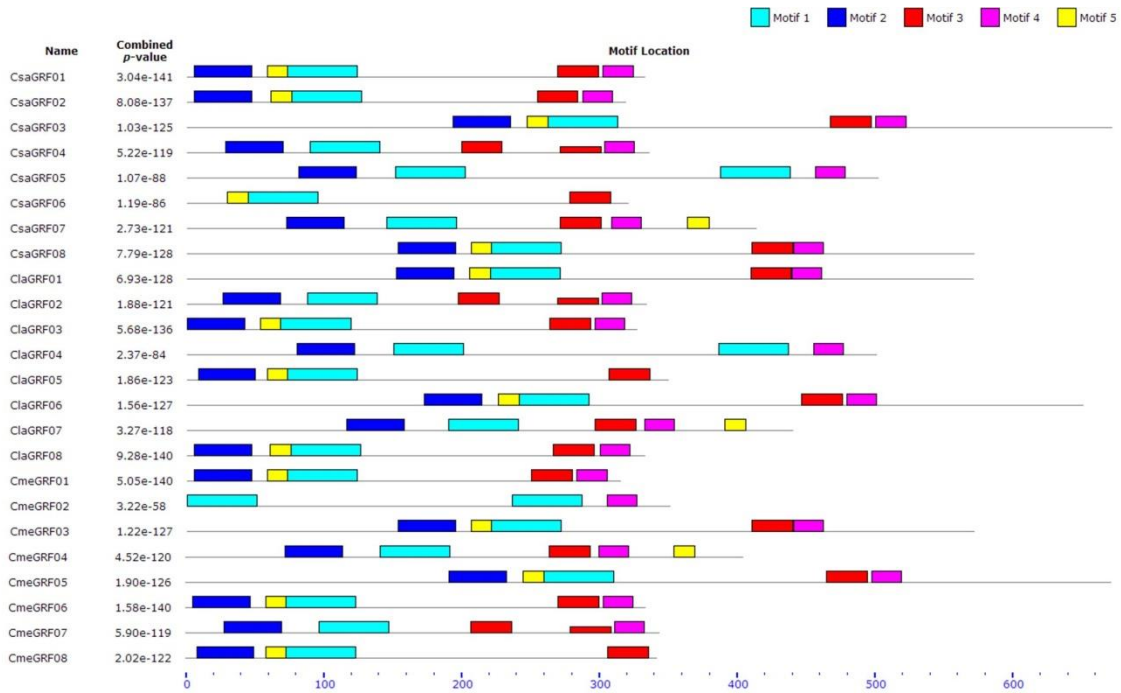

Fig 4. Variation in motif clades for the GRF proteins. The MEME motifs are shown as different-colored boxes at the N-terminal and C-terminal region for the transcription regulatory region.

based function prediction has been applied for prediction of GRF proteins in other species like Arabidopsis, rice, maize and Brachypodium. Filiz et al. (2014) reported a joined tree for these species. They found that Brachypodium GRF proteins were more closely clustered with maize and rice than Arabidopsis. So, they concluded that Brachypodium *GRF* genes resembles to monocot *GRF* genes other than dicot ones. Therefore, we constructed the joined phylogenetic tree using GRF proteins from dicots including cucumber, melon, watermelon and Arabidopsis (Fig. 3). According to phylogenetic analysis, classes (Class I and II) and subclasses (Subclasses A-B-C-D) were separated from their clusters which have also been observed by Filiz et al. (2014). High bootstrap values of 68%, 85% and 96% were observed in Subclass A, Subclass B and Subclass C, respectively. Compared to other Subclasses, only Subclass D showed low bootstrap value which was 49 %. This could be explained by the fact that cucurbits *GRF* genes were more similar to the *GRF* genes of Arabidopsis other than monocot ones. This could be arisen from gene structures and conserved domains similarity of *GRF* genes between cucurbits and Arabidopsis. We have also performed motif analysis to increase reliability of the phylogenic tree. We have detected 4 conserved motifs (QLQ, WRC, FFD, and TQL), which is consistent with earlier studies. In detail, QLQ, WRC, FFD, and TQL motifs were identified in maize (Zhang et al., 2008) and in Brachypodium (Filiz et al., 2014), while rice GRF proteins include QLQ, WRC, and TQL motifs, not FFD. The WRC domain, named after the conserved Trp-Arg-Cys motif, contains two distinctive features: a putative nuclear localization signal and a zinc-finger motif (C3H). It is suggested that the WRC domain functions transcriptional regulation by DNA binding (Van der Knaap et al., 2000). The QLQ domain is named after the conserved Gln, Leu, Gln motif. The QLQ domain is found at the N-terminus of SWI2/SNF2 protein, which has been shown to be involved in protein-protein interactions. This domain has thus been postulated to be involved in mediating protein interactions (Treich et al., 1995). Gene Ontology annotation provides information related with gene classification based on their biological processes, cellular components, and molecular functions (Conesa et al., 2005). In order to evaluate putative function of cucurbit GRF proteins, we have also performed GO slim analysis using Blast2Go. The biological processes of the all three species GRF proteins were found in the same GO groups. They belong to "regulation of transcription" and "DNA-dependent" groups. Additionally, their molecular function predictions indicated that "ATP binding" and "hydrolase activity, acting on acid anhydrides, in phosphorus containing anhydrides". GO groups were observed in all Cucurbitaceae GRF proteins. Except for CmeGRF-02 protein, all cucurbits GRF proteins were localized in the "nucleus" according to prediction of cellular component analysis. CmeGRF-02 protein was only GRF protein which was found in cytoplasm. In addition, its biological process and molecular function predictions showed "response to stimulus" and "binding" activities, respectively (Table 3). Our findings are consistent with previous studies. Filiz et al. (2014) also observed similar biological roles and functions of Brachypodium GRFs. According to functional studies on Arabidopsis, GRF proteins in Arabidopsis were highly related with cell proliferation associated with size and shape of leaf organ (Kim and Kende 2004) and acting as a transcription activator and a repressor (Kim et al., 2012). In our study, our findings are based on gene annotations of the Cucurbitaceae GRF proteins. We assume that the characterization of cucumber, melon and watermelon GRF proteins will be useful for the further functional identifications of these proteins.

Materials and Methods

Sequence retrieval and identification of GRF family from Cucumis sativus, Cucumis melo and Citrullus lanatus

In order to identify putative GRF proteins from cucumber (CsaGRF), melon (CmeGRF) and watermelon (CleGRF), different approaches were used. Initially, 113 amino acid sequences encoding GRF transcription factors from nine plants (*Arabidopsis thaliana, Carica papaya, Oryza sativa subsp. japonica, Oryza sativa subsp. indica, Populus trichocarpa, Selaginella moellendorffii, Sorghum bicolor, Vitis vinifera* and *Zea mays*) were retrieved from plant transcription factor database 3.0 (plntfdb.bio.uni-potsdam.de/) (Zhang et al., 2011). These sequences were used for identification of peptides from cucumber, melon and watermelon by performing a BLASTP search at Phytozome v9.1 database (www. phytozome.net/) (Goodstein et al., 2012), Melonomics database (https://melonomics.net/) (Garcia-Mas et al., 2012) and, Cucurbit Genomics database (http://www.icugi.org/cgi-bin/ICuGI/index.cgi) (Guo et al., 2013), respectively. Moreover, The Hidden Markov Model (HMM) profiles of the QLQ and WRC domains in the Pfam database (http://pfam.sanger.ac.uk/) were searched against the Phytozome, Melonomics and Cucurbit Genomics databases for three species. All hits with expected values less than 1.0 were retrieved and redundant sequences were removed using the decrease redundancy tool (web.expasy.org/decrease_redundancy). Each non-redundant sequence was checked for the presence of the conserved QLQ and WRC domains by SMART (http://smart.emblheidelberg. de/) (Letunic et al., 2012) and Pfam (http://pfam.sanger.ac. uk/) (Punta et al., 2012) proteomics servers to verify the conserved domains of GRF proteins.

Chromosomal location and gene structure prediction

Chromosomal distributions of GRF genes were performed by searching their physical positions of their corresponding locus numbers in the Phytozome database for cucumber, the Cucurbit Genomics database for watermelon. The genes were plotted separately onto the seven cucumber chromosomes and twelve watermelon chromosomes according to their ascending order of physical position (bp), from the short arm telomere to the long arm telomere and finally displayed using MapChart (Voorrips, 2002). A structural analysis (such as exon and intron numbers and locations as well as conserved domain locations) of GRF genes were determined using Gene structure display server (gsds.cbi.pku.edu.cn/) (Guo et al., 2007) through comparison of their full-length cDNA or predicted coding sequence (CDS) with their corresponding genomic sequence. Some physicochemical characteristics including the number of amino acids, molecular weight, instability index, and theoretical isoelectric point (pI) were computed using ProtParam tool (http://www.expasy.org /tools/protparam.html).

Sequence alignment, phylogenetic analysis and identification of conserved motifs

The amino acid sequences were imported into MEGA5 (Tamura et al., 2011) and multiple sequence alignments were performed using ClustalW with a gap open and gap extension

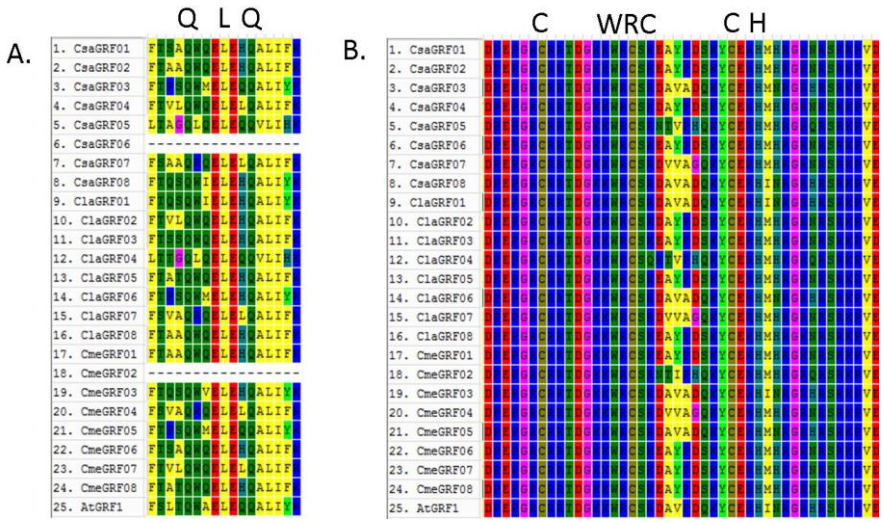

Fig 5. Comparison of the amino acid sequences of Cucurbitaceae family GRF proteins. **A)** The QLQ domains of cucumber, melon and watermelon GRF and AtGRF1 proteins. **B)** The WRC domains of Cucurbitaceae family GRFs and AtGRF1 protein with the Cys3His zinc-finger motif.

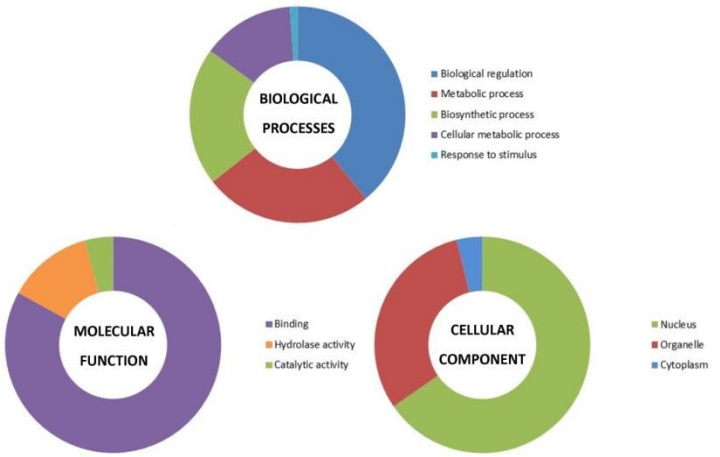

Fig 6. Gene Ontology (GO) distributions for the GRF proteins. The Blast2Go program defines the gene ontology under three categories, biological processes, molecular functions and cellular component.

penalties of 10 and 0.1, respectively (Thompson et al., 1997). The alignment file was then used to construct an unrooted phylogenetic tree based on the neighbor-joining method (Saitou and Nei, 1987) and after bootstrap analysis for 1000 replicates, the tree was displayed using Interactive tree of life (iTOL; http://itol.embl.de/index.shtml) (Letunic and Bork, 2011). Protein sequence motifs were identified using the multiple EM for motif elicitation (MEME); (http://meme.nbcr.net/meme3/meme.html) (Bailey and Elkan, 1994). The analysis was performed by keeping number of repetitions, any; maximum number of motifs, 20; and optimum width of the motif, ≤50. Discovered MEME motifs (≤1E-30) were searched in the InterPro database with InterProScan (Quevillon et al., 2005).

Comparative analysis of GRF proteins between cucumber and melon, watermelon

For deriving orthologous relationship among the cucumber and two other species amino acid sequences of GRF, BLASTP search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was performed against peptide sequences of melon and watermelon. Hits with E-value ≤ 1e-10 and at least 70% identify were considered significant.

Gene ontology (GO) annotation

The functional annotation of GRF sequences and the analysis of annotation data were performed using Blast2GO (http://www.blast2go.com) (Conesa and Götz, 2008). The amino acid sequences of GRFs were imported into Blast2GO program. Firstly, BLASTp against the non-redundant protein database of NCBI were performed. Then, mapping and retrieval of GO terms were carried out. Finally, annotation of GO terms associated with each query was determined to relate the sequences to known protein function. The program provides the output defining three categories of GO classification namely biological processes, cellular components, and molecular functions.

Conclusion

In the present study, the identification and bioinformatics analysis of cucumber, melon and watermelon *GRF* genes at

the whole genome level were conducted. A total of 24 *GRF* genes were identified. Gene structure, phylogenetic relationship, and sequence characteristics were investigated. As a conclusion, *in silico* genome wide survey has contributed to the understanding of gene structure and evolutionary relationship of the cucurbits *GRF* gene family. Hence, this report would be useful for the cucurbit research community for the discovery of new GRF members in other Cucurbitaceae family members and selection of these candidate genes for functional and cloning studies.

References

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Paper presented at Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California, 28–36.

Baloglu MC, Eldem V, Hajyzadeh M, Unver T (2014) Genome-wide analysis of the bZIP transcription factors in cucumber. Plos One. 9(4): e96014.

Choi D, Kim JH, Kende H (2004) Whole genome analysis of the *OsGRF* gene family encoding plant-specific putative transcription activators in rice (*Oryza sativa* L.). Plant Cell Physiol. 45: 897–904.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21: 3674–3676.

Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008: 619832.

Filiz E, Koç İ, Tombuloğlu H (2014) Genome-wide identification and analysis of growth regulating factor genes in *Brachypodium distachyon*: *in silico* approaches. Turk J Biol. 38: 296-306.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Hénaff E, Câmara F, Cozzuto L, Lowy E, Alioto T et al. (2012) The genome of melon (*Cucumis melo* L.). Proc Natl Acad Sci USA. 109 (29): 11872-11877.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar D (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–D1186.

Guo AY, Zhu QH, Chen X, Luo JC (2007) GSDS: a gene structure display server. Yi Chuan. 29: 1023–1026.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z et al. (2013) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet. 45: 51–58.

Hu LF, Liu SQ (2011) Genome-wide identification and phylogenetic analysis of the ERF gene family in cucumbers. Genet Mol Biol. 34: 624–633.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P et al. (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet. 41: 1275–1281.

Kim HJ, Lee BH (2006) Growth-regulating factor4 of *Arabidopsis thaliana* is required for development of leaves, cotyledons, and shoot apical meristem. J Plant Biol. 49: 463-468.

Kim JH, Choi D, Kende H (2003) The AtGRF family of putative transcription factors is involved in leaf and cotyledon growth in Arabidopsis. Plant J. 36: 94–104.

Kim JH, Kende H (2004) A transcriptional coactivator, AtGIF1, is involved in regulating leaf growth and morphology in Arabidopsis. Proc Natl Acad Sci USA. 101: 13374–13379.

Kim JS, Mizoi J, Kidokoro S, Maruyama K, Nakajima J, Nakashima K, Mitsuda N, Takiguchi Y, Ohme-Takagi M, Kondou Y et al. (2012) Arabidopsis growth-regulating factor7 functions as a transcriptional repressor of abscisic acid and osmotic stress–responsive genes, including DREB2A. Plant Cell. 24: 3393–3405.

Letunic I, Bork P (2011) Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39: W475–8.

Letunic I, Doerks T, Bork P (2012) Smart 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 40: D302–D305.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J et al. (2012) The Pfam protein families database. Nucleic Acids Res. 40: D290-D301.

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X et al (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet. 45: 1510–1515.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res. 33: W116–W120.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4: 406–425.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) Mega5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28: 2731–2739.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The Clustal_x windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25: 4876–4882.

Treich I, Cairns BR, de los Santos T, Brewster E, Carlson M (1995) SNF11, a new component of the yeast SNF-SWI complex that interacts with a conserved region of SNF2. Mol Cell Biol. 15: 4240–4248.

Vain P (2011) Brachypodium as a model system for grass research. J Cereal Sci. 54: 1–7.

Van der Knaap E, Kim JH, Kende H (2000) A novel gibberellin induced gene from rice and its potential regulatory role in stem growth. Plant Physiol. 122: 695–704.

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered. 93: 77–78.

Xu G, Guo C, Shan H, Kong H (2012) Divergence of duplicate genes in exon–intron structure. Proc Natl Acad Sci USA. 109: 1187–1192.

Yamasaki K, Kigawa T, Inoue M, Watanabe S, Tateno M, Seki M, Shinozaki K, Yokoyama S (2008) Structures and evolutionary origins of plant-specific transcription factor DNA-binding domains. Plant Physiol Bioch. 46: 394–401.

Zhang DF, Li B, Jia GQ, Zhang TF, Dai JR, Li JS, Wang SC (2008) Isolation and characterization of genes encoding GRF transcription factors and GIF transcriptional coactivators in Maize (*Zea mays* L.). Plant Sci. 175: 809–817.

Zhang H, Jin JP, Tang L, Zhao Y, Gu XC (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. Nucleic Acids Res. 39: D1114–D1117.