


OPEN

The ankyrin repeat gene family in *Capsicum* spp: Genome-wide survey, characterization and gene expression profile

Carlos Lopez-Ortiz^{1,5}, Yadira Peña-García^{1,5}, Purushothaman Natarajan^{1,2}, Menuka Bhandari¹, Venkata Abburi¹, Sudip Kumar Dutta^{1,3}, Lav Yadav¹, John Stommel⁴, Padma Nimmakayala ^{1*} & Umesh K. Reddy^{1*}

The ankyrin (ANK) repeat protein family is largely distributed across plants and has been found to participate in multiple processes such as plant growth and development, hormone response, response to biotic and abiotic stresses. It is considered as one of the major markers of capsaicin content in pepper fruits. In this study, we performed a genome-wide identification and expression analysis of genes encoding ANK proteins in three *Capsicum* species: *Capsicum baccatum*, *Capsicum annuum* and *Capsicum chinense*. We identified a total of 87, 85 and 96 ANK genes in *C. baccatum*, *C. annuum* and *C. chinense* genomes, respectively. Next, we performed a comprehensive bioinformatics analysis of the *Capsicum* ANK gene family including gene chromosomal localization, *Cis*-elements, conserved motif identification, intron/exon structural patterns and gene ontology classification as well as profile expression. Phylogenetic and domain organization analysis grouped the *Capsicum* ANK gene family into ten subfamilies distributed across all 12 pepper chromosomes at different densities. Analysis of the expression of ANK genes in leaf and pepper fruits suggested that the ANKs have specific expression patterns at various developmental stages in placenta tissue. Our results provide valuable information for further studies of the evolution, classification and putative functions of ANK genes in pepper.

The ankyrin (ANK) repeat domain is one of the most common conserved protein domains widely distributed in different organisms ranging from viruses to humans¹. This protein domain was identified for the first time in the yeast cell-cycle regulators Swi6 and Cdc10² and in the *Drosophila melanogaster* signaling protein Notch³. The name of these proteins was assigned after the cytoskeletal ankyrin protein was found to contain 24 copies of this sequence⁴. ANK repeats contain a ~33-residue motif repeated in tandem and consisting of two antiparallel α -helices connected by a series of inverting β -hairpin motifs. Indeed, the ANK domain is better characterized by a folding structure rather than functional requirements^{5,6}. Although the ANK domain proteins are not known to have a specific function, they play important roles in several biological activities and have been identified in abundant different proteins with diverse functions⁷. For instance, these repeats can mediate protein–protein interactions^{8,9} and may serve as molecular chaperones¹⁰.

ANK proteins reported in plants are classified into 13 subfamilies based on different domains that have been identified by genome structure studies and gene expression profiles⁶. ANK domain-containing proteins of plants are usually involved in crucial physiological and developmental processes such as signaling and growth¹¹, plastid differentiation¹², embryogenesis¹³, chloroplast biogenesis¹⁴, formation of grana¹⁵, leaf morphogenesis¹⁶, pollen germination and polarized pollen tube growth^{17,18}. Additionally, the ANK repeat-containing proteins play important roles in the response to both biotic and abiotic stresses. These proteins have been observed to participate in

¹Department of Biology, Gus R. Douglass Institute, West Virginia State University, Institute, West Virginia, United States of America. ²Department of Genetic Engineering, School of Bioengineering, SRM Institute of Science and Technology, Kattankulathur, 603203, India. ³ICAR RC NEH Region, Mizoram Centre, Kolasib, Mizoram, India. ⁴Genetic Improvement of Fruits and Vegetables Laboratory (USDA, ARS), Beltsville, MD, 20705, USA. ⁵These authors contributed equally: Carlos Lopez-Ortiz and Yadira Peña-García. *email: padma@wvstateu.edu; ureddy@wvstateu.edu

drought tolerance¹⁹, ABA-mediated regulation of reactive oxygen species levels under salt-stress²⁰, and several plant diseases²¹, including those generated by fungus such as rice blast²².

The release of genomic data and the development of bioinformatics analyses have led to comprehensive research on the identification and characterization of the ANK gene family in plants such as *Arabidopsis*²³, rice¹⁸, tomato²⁴, maize²⁵, *Physcomitrella patens*²⁶ and soybean²⁷. The number of ANK repeats, genes, and proteins in plants varies considerably across diverse plant species. In *Arabidopsis thaliana*, 509 ANK repeats encoded by 105 genes were reported, whereas rice contains 175 ANK repeat genes^{18,23}.

Pepper (*Capsicum* spp.) is a member of the Solanaceae family and is closely related to potato, tomato, egg-plant, tobacco and petunia. Pepper represents an economically important horticultural crop worldwide because of its wide variety of uses, as a food, coloring agent, and spice and in pharmaceuticals, cosmetics and ornamental products as well as for its nutrimental value^{28,29}. Despite the importance of pepper, genome-wide studies remain limited. The recent whole-genome sequencing of pepper²⁸ provided an excellent tool for genome-wide analysis for the identification and characterization of entire gene families present in this crop^{30–32}. Recently, the ANK repeat domain was identified as one of the major markers linked to capsaicinoid synthesis in *Capsicum annuum*³³ and *Capsicum chinense*³⁴.

The present study aimed to analyze the gene locus and chromosome localization, protein length, number of ANK repeats, molecular weight (MW), isoelectric points (pI), gene structure and phylogenetic relationship of ANK genes in three *Capsicum* species: *Capsicum baccatum*, *C. annuum* and *C. chinense*. We also surveyed the expression patterns of ANK genes of *C. annuum* and *C. chinense*. These analyses will contribute to a better understanding of the evolution, function and future insights for research of the ANK gene family in *Capsicum* species and will also provide a robust database for the *Capsicum* research community.

Results

Genome-wide identification of ANK proteins in pepper. We used the conserved amino acid sequence of the ankyrin domain (Accession no. PF00023) to search the three pepper genomes in the PGP, with the HMM profile used as a query and identified 268 genes potentially encoding ANK proteins across the three genome databases: 87 in *C. baccatum* (*CbANK*), 85 in *C. annuum* (*CaANK*) and 96 in *C. chinense* (*CcANK*). Each of these ANK protein sequences were verified by SMART and Pfam analyses. For convenience, in this study we provide a simplified nomenclature for each identified ANK gene. We designated the acronyms *CbANK1* to *CbANK87* for *C. baccatum*, *CaANK1* to *CaANK85* for *C. annuum* and *CcANK1* to *CcANK96* for *C. chinense*, based on the order of appearance on chromosomes 1 to 12. Length, MW and pI of ANK proteins were deduced from their protein sequences and are in Tables S1–S3. The protein length of *CbANKs* ranged from 102 to 842 residues, *CaANKs* from 117 to 958 residues, and *CcANKs* from 81 to 760 residues. Hence, we identified 1541 ANK domains within these 268 proteins across all three *Capsicum* species. The number of ANK repeats varied greatly, from 1 to 19 repeats per protein; however, the proteins with 2 to 5 ANK repeats were the most common among all the species (Fig. S1).

Among the 268 ANK genes identified, 243 (92.4%) were physically mapped and unevenly distributed across the 12 chromosomes of pepper at different densities, whereas the other 25 genes were located on scaffolds (Fig. 1). Among all chromosomes and species, chromosome 5 of *C. chinense* had the highest number of ANK genes, 39 (~41%), followed by *C. baccatum*, 22 (25.2%), and *C. annuum*, 20 (23.52%). Chromosome 1 of *C. baccatum* and *C. annuum* had 15 ANK genes, and chromosomes 8 and 1 of *C. baccatum* and *C. annuum* had 11 ANK genes. Among all species, chromosome 7 contained the lowest number of ANK genes, only one.

The distribution pattern of ANK genes on individual chromosomes also indicated physical regions with a relatively higher accumulation of multiple ANK gene clusters, such as chromosome 5 at the lower end of the arm for all species, with higher density in *C. chinense*. The distribution and density of ANK gene clusters differed among the three genomes. For example, chromosome 1 of *C. baccatum* and *C. annuum* had a high-density cluster, which was not in the same density in *C. chinense*. Another cluster was found in the lower arm of chromosome 3 but only in *C. annuum* and just few genes were found in *C. baccatum* and *C. chinense*. In chromosome 8, ANK genes were distributed in the upper part of *C. baccatum* and *C. annuum* but not *C. chinense*. Similarly, one ANK gene was present in the upper end of chromosome 12 for *C. baccatum* but was absent in the other two pepper species.

Identification of conserved motifs of ANK genes in *Capsicum* and domain analysis. To analyze the function and domain distribution of the putative *Capsicum* ANK genes, Pfam and Hmmer platforms were used for protein analysis, to allow for identifying the major domains. As a result, the 268 ANK genes were classified into 10 subfamilies based on their domain composition: ANK-U, ANK-TM, ANK-PK, ANK-ZnF, ANK-BTB, ANK-BPA, ANK-ACBP, ANK-GPCR, ANK-IQ and ANK-O (Table 1). Among these genes, 42 ANK domain-containing proteins were found in *C. baccatum* and 45 in *C. annuum*, whereas in *C. chinense*, 62 proteins with a unique ANK domain (ANK-U subfamily) were reported. The transmembrane domain (ANK-TM) was the second most abundant subfamily, containing 33, 29, and 26 proteins in *C. baccatum*, *C. annuum*, and *C. chinense*, respectively. The ANK-ZnF (zinc-finger) subfamily contained two members in each species, which were classified into the subgroup ANK-CCCH based on the type of zinc finger. In the same way, *C. annuum* and *C. baccatum* featured the tramtrack and bricabrac domains, two members of the ANK-BTB-containing broad complex, but only one (*CcANK23*) was identified in *C. chinense*. One gene belonging to the ANK-ACBP (Acyl-CoA-binding protein) subfamily and one more from the ANK-GPCR subfamily, which contains a GPCR-chaperone-1, were found in all *Capsicum* species analyzed. *C. baccatum* and *C. chinense* but not *C. annuum* had a member of the protein kinase domain family, ANK-PK, containing serine/threonine or tyrosine kinase. *C. annuum* and *C. baccatum* shared one gene in common that belongs to the ANK-BPA family, containing the BAR, PH and ArfGap domains. Only *C. annuum* featured the ANK-IQ subfamily, containing the calmodulin-binding domain with only one member. The remaining proteins were grouped into the ANK-O subfamily, containing different domains including the motile-sperm, bromodomain, STII, UreD, G-patch, bVFLR1 and Myb DNA binding domains.

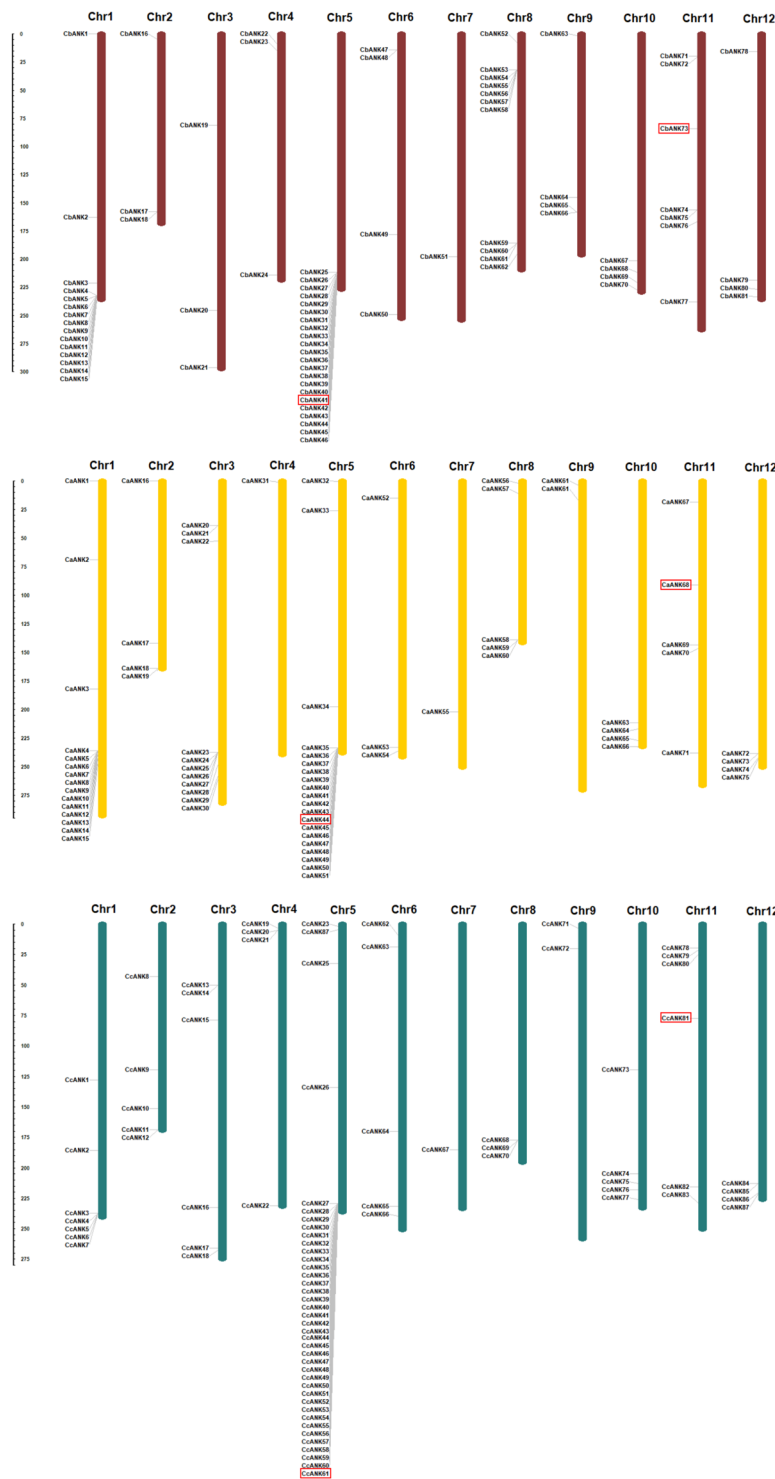


Figure 1. Chromosomal locations of ankyrin (ANK) proteins in the pepper *Capsicum baccatum* (red), *C. annuum* (yellow) and *C. chinense* (green). Chromosome numbers are represented at the top of each chromosome. The left panel scale indicates the chromosome length in Mbp. Orthologs genes of CA05g18080 and CA11g09160 are represented by red boxes.

The conserved motifs of ANK proteins were analyzed with the MEME server. The number of conserved motifs in each *Capsicum* ANK protein ranged from 2 to 9 (Fig. S2). Moreover, the length of the motifs ranged from 15 to 50 amino acids (Fig. S3, Table S4). Motif 5 was conserved in the ANK-U subfamily among the three *Capsicum* species, whereas motifs 8, 9 and 10 were identified in ANK-O and ANK-U subfamilies. Additionally, motifs 1, 2 and 3 were distinctively detected in all *Capsicum* ANK genes forming the configurations of ANK domain. Motifs 4 and 7 were evenly distributed in *C. annuum* and motif 6 in *C. baccatum*. The information

Specie	ANK Subfamily												# of ANK proteins	# of Proteins	% of ANK proteins	Reference
	U	TM	ZnF	BTB	ACBP	GPCR	PK	BPA	IQ	RF	TPR	O				
<i>A. thaliana</i>	18	40	6	7	2	0	7	4	4	5	1	11	105	25,498	0.41	23
<i>O. sativa</i>	73	37	7	6	0	0	4	3	4	9	22	10	175	35,825	0.49	18
<i>Z. mays</i>	30	15	3	2	0	0	4	2	1	9	2	3	71	39,591	0.18	25
<i>P. patens</i>	21	0	3	3	3	2	9	3	0	6	0	4	54	35,398	0.15	26
<i>E. siliculosus</i>	178	0	7	0	0	0	56	3	8	3	2	82	339	16,256	2.08	38
<i>G. max</i>	48	30	13	3	2	4	12	7	13	2	1	27	162	55,897	0.29	27
<i>S. lycopersicum</i>	26	25	8	7	0	4	9	4	7	7	4	26	130	33,952	0.38	24
<i>C. baccatum</i>	42	33	2	2	1	1	1	1	0	0	0	4	87	35,874	0.24	
<i>C. annuum</i>	45	29	2	2	1	1	0	1	1	0	0	3	85	35,884	0.24	
<i>C. chinense</i>	62	26	2	1	1	1	1	0	0	0	0	2	96	35,009	0.27	

Table 1. Comparative analysis of ankyrin (ANK) repeat proteins between *Capsicum* and other plant species. U ankyrin repeat; TM transmembrane; ZnF, zinc finger; BTB, Broad-Complex, Tramtrack and Bric a brac; ACBP, Acyl CoA binding protein domain; GPCR, G protein-coupled receptors; PK, protein kinase domain; BPA, BAR domain, Pleckstrin homology domain and ArfGTPase-activating domain; IQ, Short calmodulin-binding motif containing conserved Ile and Gln residues; RF, RING finger; TPR, tetratricopeptide repeats; O, others domains.

obtained from ScanProsite analysis revealed that the function of most of the motifs was related to not only ANK domain-containing proteins but also ACCELERATED CELL DEATH 6-like protein (*ACD6*).

Phylogenetic tree and gene structure of *Capsicum* ANK genes. To gain insight into the evolution of ANK genes and infer their function based on homologs present in other plant species, the protein sequences of *Capsicum* species and 105 full-length ANK protein sequences from *Arabidopsis* were aligned by using MEGAX. We expanded the maximum-likelihood method with 1000 bootstrap replication to construct an unrooted phylogenetic tree (Fig. 2). In total, 373 ANK proteins from the four species were clearly divided into six major groups (groups I to VI). Most of the members in the same groups shared one or more domains outside of the ANK domain, which further supported the subfamily definition described above. For example, the ANK-U subfamily was distributed throughout all the groups; similarly, the ANK-TM subfamily was present in most of the groups except for groups III and IV. The ANK-IQ, ANK-GPCR and ANK-ACBP subfamilies were placed in group III, and the subfamilies ANK-ZnF, ANK-PK and ANK-BTB were in group IV. Finally, the ANK-BPA subfamily was placed in group V.

Next, we compared the structural diversity between *Capsicum* ANK genes in terms of exon/intron arrangement of the coding and genome sequences by using the GSDS tool for generating gene structure schematic diagrams. Structure analyses of the ANK genes revealed that the positions, length and number of introns varied across all species and subfamilies. The number of introns present ranged from 0 to 18 in *C. baccatum* and *C. annuum* genes and from 0 to 11 in *C. chinense* genes. Detailed gene structure of the *Capsicum* ANK genes is in Fig. 3. *CaANK44* showed three introns, with only one in *CbANK41* and *CcANK61*. In total, 19, 23 and 30 genes contained only one exon in *C. baccatum*, *C. annuum* and *C. chinense*, respectively. *C. baccatum* and *C. annuum* shared one gene with 19 exons, and the maximum number of exons in *C. chinense*, 12, was found in *CcANK68*. The most closely related ANK genes in the same subfamily shared a similar gene structure in terms of intron number and intron-exon length.

We determined the protein structures of CA05g18080 orthologs (*CbANK41*, *CaANK44*, *CcANK61*) for each *Capsicum* species. The protein structure of *Capsicum* ANK proteins was modelled at >90% confidence using the alignment of hidden Markov models via an HMM-HMM search at the Phyre2 server³⁵. Overall, ANKs appear to have very similar structures, with an α -solenoid fold architecture and secondary structure predominantly consisted of α -helices (Fig. S4) according to the classification defined by Kajava (2012)³⁶. Hence, all predicted protein structures are considered highly reliable, which offers a preliminary basis for understanding the molecular function of *Capsicum* ANK proteins.

Prediction of *Capsicum* ANK gene promoter elements. To identify putative *cis*-elements in *Capsicum* ANK promoters, we analyzed 1500-bp DNA sequences upstream of the start codon (ATG) at the Plant *Cis*-acting Regulatory DNA Elements (PLACE) website. The analysis identified 124 common *cis*-elements among all *Capsicum* ANK genes (Table S4). Furthermore, we revealed four common *cis*-regulatory elements — WBOXATNPR1, ASF1MOTIFCAMV, GCCCORE and SEBFCONSSTPR10A — which are involved in response to plant hormones, including auxin and salicylic acid (SA), as well as disease resistance. These results agree with the known role of ANK genes in plant resistance to biotic and abiotic stresses. Among the 124 *Cis*-elements, 15 were related to auxin, ABA and SA, and another six to gibberellic acid (GA). We also identified IBOX, -10PEHVPSBD, TBOXATGAPB, INRNTPSADB and GT1CONSENSUS, which have been found required for transcriptional regulation by light. Other *Cis*-elements present, MYBCORE, MYCATERD1, MYCATRD22, MYBATRD22 and MYB2AT, are related to water stress and dehydration. Finally, 11 were associated with binding protein site function. Overall, most of the predicted *Cis*-elements play a role in the response to signaling hormones and are involved in different stresses.

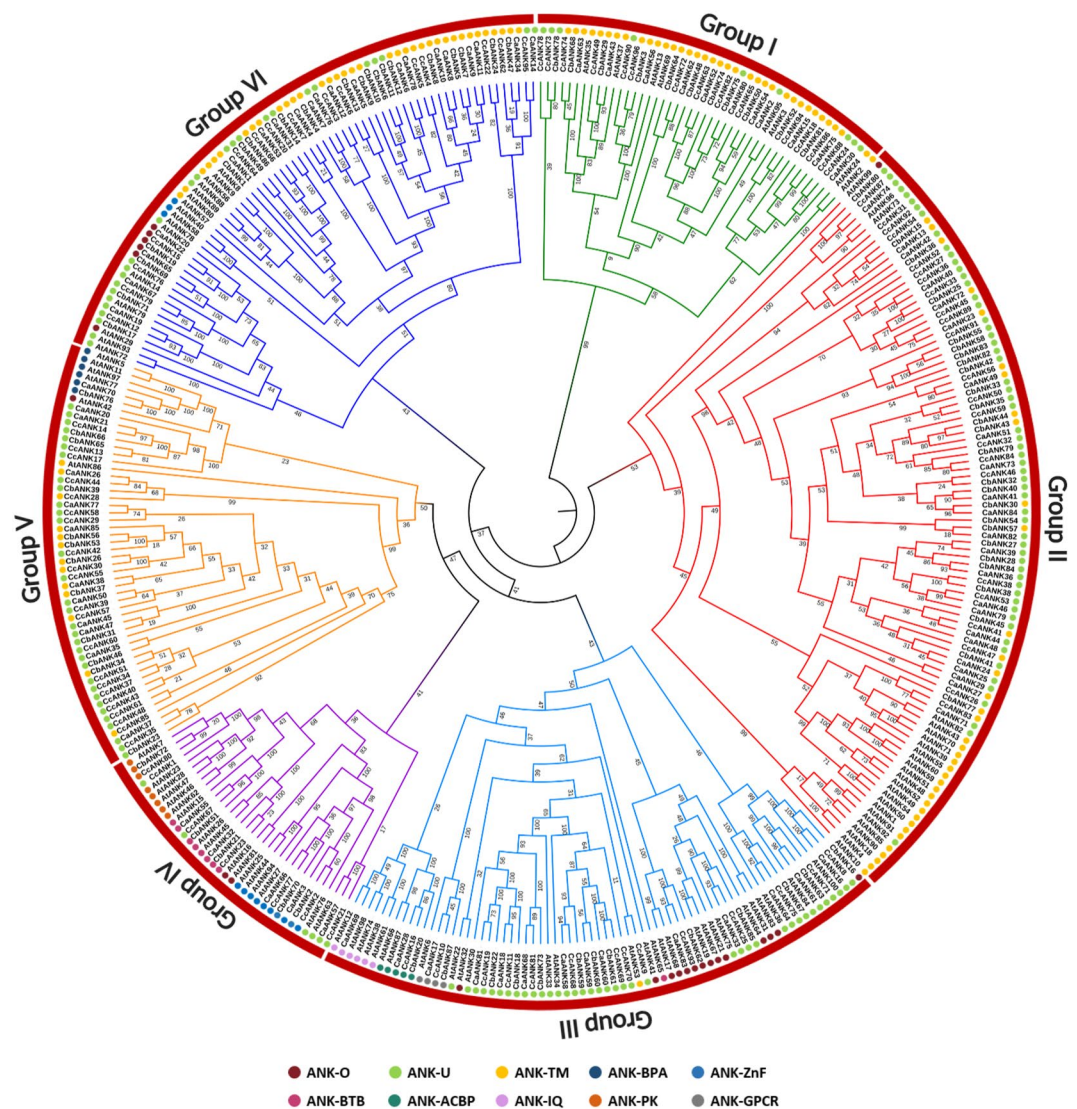


Figure 2. Phylogenetic relationships of *Capsicum* species and *Arabidopsis* ANK proteins. Phylogenetic analysis by the maximum-likelihood method with 1000 bootstrap replicates of the 268 and 105 ANK proteins identified from *Capsicum* species and *Arabidopsis*, respectively. ANK subfamilies are represented by different colored dots.

Syntenic *Capsicum* ANK paralog pairs. To examine the impact of duplications on the ANK gene family, we analyzed tandem and segmental duplication events. Syntenic paralog pairs were identified between and within the three *Capsicum* genomes and the syntenic relationships of ANK genes among *Capsicum* species and *Arabidopsis* genomes were visualized by generating a circular plot (Fig. S5). We found 23 pairs of ANK syntenic paralogs across all species (Table 2); 16 were intra-species and the remaining were inter-species. Among the seven inter-species pairs, four segmental duplications were intra-chromosomal, located on chromosomes 5 and 1. Most of the duplicated paralog pairs belong to the same subfamily, with exception of the segmental duplication *CbANK72-CcANK1*. The paralogs of this last pair were identified in chromosomes 11 and 1 and belong to the ANK-PK and ANK-U subfamilies, respectively. The last two duplication pairs were found in scaffold positions.

We further identified the rate of synonymous per non-synonymous site (Ks) and non-synonymous substitutions per non-synonymous site (Ka) values to explore the selective pressures on these paralog pairs and understand the expansion of this gene family in pepper. The value of the Ka/Ks ratio represents the type of selection pressure on the gene and evolutionary rate. Ka/Ks ~ 0 indicates that the selection is neutral, Ka/Ks < 1 indicates purifying selection and Ka/Ks > 1 indicates positive selection³⁷. The Ka/Ks (ω) ratios for segmental duplications ranged from 0.15 to 1.05. In total, 22 of 23 paralog pairs were under purifying selection, with ω ratios < 1. These ratios suggest that the ANK gene family in these pepper species evolved by the removal of deleterious alleles. This type of selection preserves the long-term stability of genes over the course of evolution³⁷. The ω ratio for *CbANK84-CbANK28* was > 1, which indicates positive selection. Along with the selective pressures, we estimated the duplication time of *Capsicum* ANK paralog pairs by using a relative Ks measure as a proxy for time, and it spanned from 0.05 to 5.9 million years ago (MYA), with an average duplication time of ~1.1 MYA.

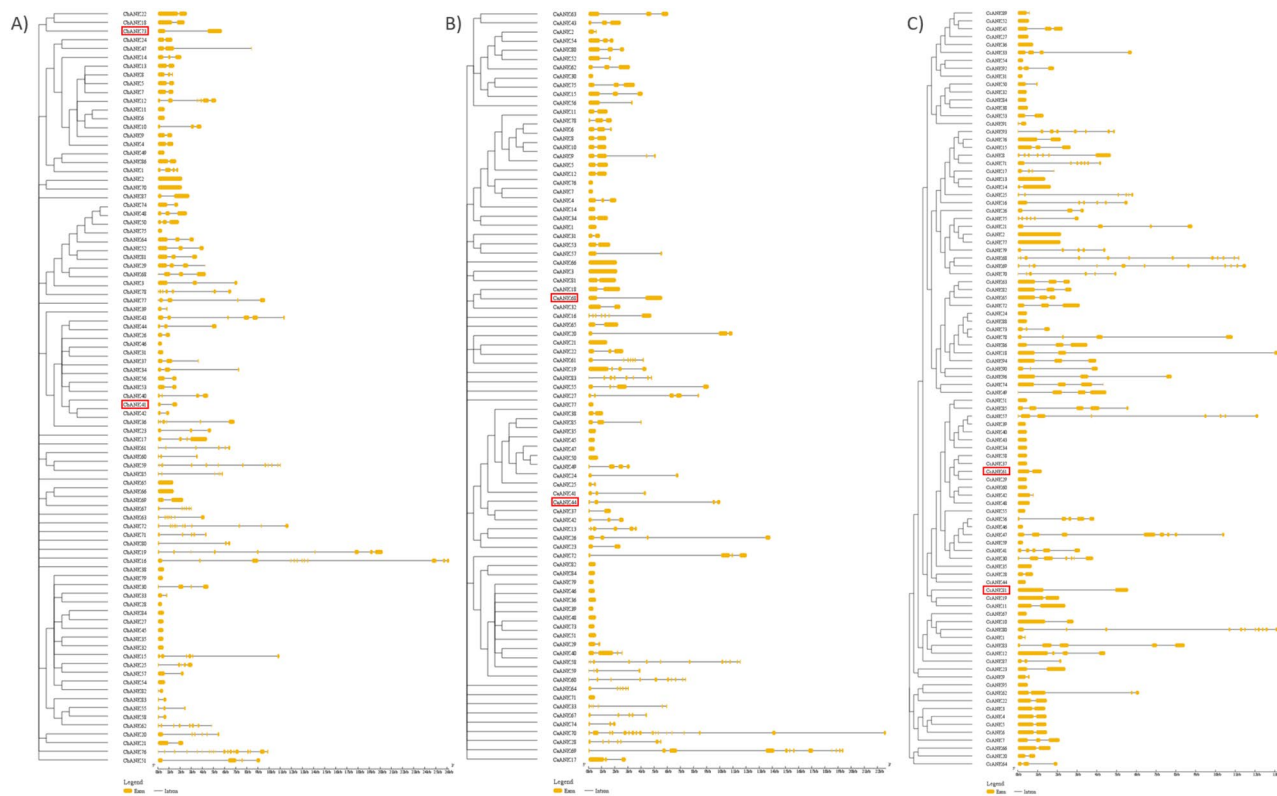


Figure 3. Gene structure analysis of the ANK proteins in *Capsicum* species: (A) *C. baccatum*, (B) *C. annum* and (C) *C. chinense*. The exons and introns are represented by the yellow boxes and black lines, respectively. The scale bar in the bottom represents the gene length in kb. Orthologs genes of CA05g18080 and CA11g09160 are represented by red boxes.

Syntenic paralog pairs	S-Sites	N-Sites	Ka	Ks	Ka/Ks	Selection pressure	Duplication time (MYA)
<i>CbANK3-CcANK96</i>	371.91	1197.09	0.00	0.02	0.15	Purifying selection	0.11
<i>CaANK25-CaANK24</i>	96.22	320.78	0.28	0.33	0.85	Purifying selection	1.67
<i>CaANK82-CbANK57</i>	130.25	424.76	0.03	0.06	0.54	Purifying selection	0.30
<i>CbANK54-CbANK57</i>	151.41	499.59	0.02	0.02	0.73	Purifying selection	0.12
<i>CaANK36-CbANK28</i>	87.41	293.59	0.02	0.05	0.36	Purifying selection	0.24
<i>CbANK84-CbANK28</i>	88.09	292.91	0.02	0.02	1.05	Positive selection	0.12
<i>CbANK35-CbANK33</i>	69.45	245.55	0.14	0.24	0.60	Purifying selection	1.20
<i>CbANK38-CbANK36</i>	120.81	401.19	1.07	1.16	0.92	Purifying selection	5.87
<i>CaANK77-CaANK38</i>	80.09	306.91	0.03	0.08	0.38	Purifying selection	0.40
<i>CcANK37-CcANK43</i>	100.22	364.78	0.03	0.13	0.21	Purifying selection	0.66
<i>CcANK60-CcANK48</i>	102.73	371.27	0.07	0.14	0.51	Purifying selection	0.70
<i>CaANK45-CaANK47</i>	102.56	386.44	0.01	0.01	0.53	Purifying selection	0.05
<i>CbANK31-CaANK47</i>	102.70	386.31	0.01	0.02	0.26	Purifying selection	0.10
<i>CcANK55-CcANK42</i>	78.25	302.75	0.08	0.11	0.76	Purifying selection	0.55
<i>CaANK50-CcANK40</i>	97.15	361.85	0.08	0.21	0.38	Purifying selection	1.07
<i>CcANK39-CcANK40</i>	87.98	320.02	0.09	0.19	0.46	Purifying selection	0.97
<i>CbANK56-CbANK53</i>	228.17	797.83	0.04	0.09	0.44	Purifying selection	0.45
<i>CbANK8-CaANK10</i>	227.16	714.84	0.01	0.01	0.95	Purifying selection	0.07
<i>CbANK7-CbANK5</i>	287.24	921.76	0.01	0.01	0.93	Purifying selection	0.07
<i>CaANK78-CaANK6</i>	289.95	931.06	0.01	0.01	0.39	Purifying selection	0.07
<i>CbANK11-CbANK6</i>	139.63	481.37	0.03	0.04	0.77	Purifying selection	0.22
<i>CbANK72-CcANK1</i>	65.47	228.53	0.26	1.00	0.27	Purifying selection	5.02
<i>CcANK80-CcANK1</i>	64.38	226.62	0.27	1.03	0.26	Purifying selection	5.20

Table 2. Ka-Ks calculation for each pair of syntenic Capsicum ANK paralogs. S-Sites, number of synonymous sites; N-Sites, number of non-synonymous sites; Ka, non-synonymous substitution rate; Ks, synonymous substitution; MYA, million years ago.

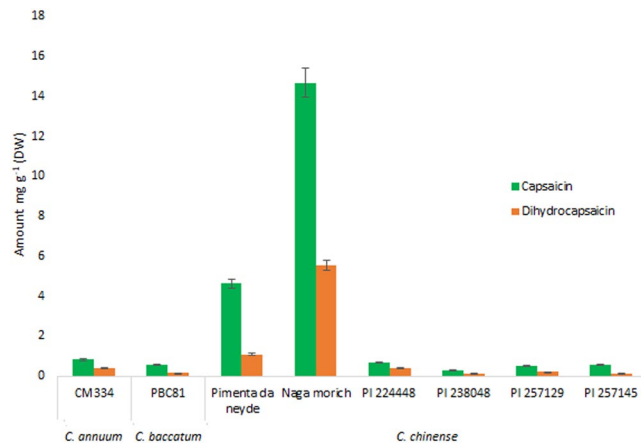


Figure 4. Capsaicin and dihydrocapsaicin content in pepper. Capsaicin and dihydrocapsaicin levels in pepper powder from dried green fruit (16 days post-anthesis [dpa]). Data are mean \pm SD; $n = 3$.

GO annotation of *Capsicum* ANK genes. GO analysis performed with Blast2Go suggested the putative participation of ANK genes in multiple biological processes, molecular functions, and cellular component (Fig. S6). For instance, most of the ANK genes are likely related to stress response, followed by signal transduction, and response to endogenous stimulus across all the *Capsicum* species. As well, in each of the three *Capsicum* species, we identified two ANK genes involved in transport, one in response to biotic stimulus, and one more in response to external stimulus. Analysis of the molecular functions predicted that the main roles of the ANK genes in all species were related to binding, transferase and kinase activities. Similarly, cellular component analysis revealed that 197 ANK genes were predicted in cell membrane, 9 were intracellular and 6 were in chloroplasts. These results provide useful information for future gene characterization studies in pepper.

Capsaicinoid content in pepper. To further explore the potential functions of ANK genes related with capsaicinoid content in pepper fruits, we firstly determined the capsaicin and dihydrocapsaicin amount in fruits at 16 dpa (days post-anthesis) from the three *Capsicum* species analyzed (Fig. 4). The highest content of capsaicin and dihydrocapsaicin was in *C. chinense* cv. Naga morich — 14.67 and 5.54 mg g⁻¹ dry weight (DW) tissue, respectively — followed by *C. chinense* cv. Pimenta da neyde — 4.62 and 1.08 mg g⁻¹ DW tissue, respectively. In contrast, for the remaining *C. chinense* varieties analyzed (PI 224448, PI 238048, PI 257129 and PI 257145), the content ranged from 0.28 to 0.68 mg g⁻¹ DW for capsaicin and 0.13 to 0.4 mg g⁻¹ DW for dihydrocapsaicin. Contrary to *C. chinense* fruits, for *C. annuum* cv. CM334, the content of capsaicin and dihydrocapsaicin was 0.823 and 0.393 mg g⁻¹ DW tissue, respectively. The lowest content across all species was for *C. baccatum*, 0.55 and 0.15 mg g⁻¹ for capsaicin and dihydrocapsaicin, respectively.

Ankyrin genes associated with capsaicin in pepper. Previous studies identified ANK genes that are involved in the biosynthesis of capsaicinoids and in the pungency modulation of pepper. For instance, in a genome-wide association study, Nimmakayala *et al.*³³, reported two significant single nucleotide polymorphisms (SNPs) associated with capsaicin content, S5_227837931 and S11_83930015, were located in the locus CA05g18080, which codes for an ankyrin-like protein, and CA11g09160, which encodes a protein with acyl-transferase activity, respectively. As shown in Fig. 5, the SNP positions S5_227837931 and S11_83930015, located in chromosome 5 and 11 respectively, have high allelic effect for capsaicin content and fruit weight; however, in both SNP positions, the non-pungent and high fruit weight cultivars contained a G allele, whereas the pungent and less fruit weight cultivars possessed A alleles. Likewise, Park *et al.*³⁴, reported a major quantitative trait loci located on the locus CC.CCv1.2.scaffold31.147 of *C. chinense*, which encodes for an ankyrin-repeat-containing protein. This quantitative trait locus was suggested to play an important role in capsaicinoid biosynthesis of pungent pepper pericarps.

Expression profile of ANK genes in *C. annuum* and *C. chinense*. We identified orthologs by a BLASTN strategy with two previously identified capsaicinoid markers (i.e., CA05g18080 and CA11g09160), which were obtained from the Sol Genomics Network database. The resulting orthologs were *CbANK41*, *CaANK44* and *CcANK61* for CA05g18080 and *CbANK73*, *CaANK68* and *CcANK81* for CA11g09160. To investigate the expression profile of individual *CaANK* genes across different tissues, including leaf and placenta (6, 16 and 25 dpa), we used publicly available RNA-seq data for *C. annuum* cv. CM334²⁸. An expression heat map was used to visualize the *CaANK* tissue-specific expression patterns (Fig. 6). Overall, 81 out of 85 genes were expressed in at least one of the tested tissues (Fig. 6A). However, only 57 genes were expressed in all assessed tissues and at various expression levels. Moreover, some genes exhibited unique expression profiles in a specific tissue. For instance, we found 4, 3, 2 and 1 tissue-specific *CaANKs* in leaf, 6-dpa, 16-dpa and 25-dpa placenta, respectively (Fig. 6B). *CaANK44* was highly expressed in placenta tissue at 16-dpa but moderately at 25-dpa and almost not expressed in leaf tissue (Fig. 6A). However, *CaANK68* showed a high expression pattern in leaf tissue versus placenta tissue at the three different stages.

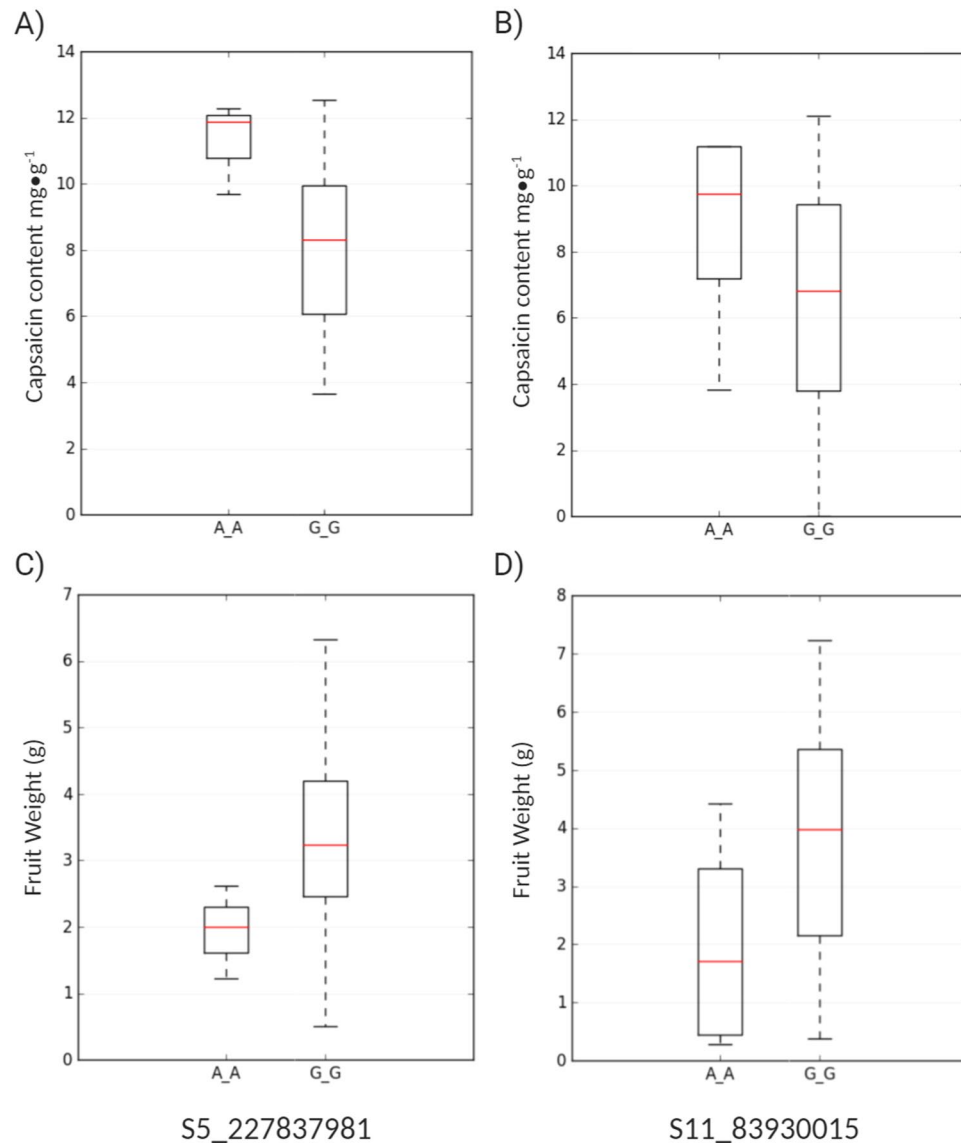


Figure 5. Allelic effect of significantly associated single nucleotide polymorphic (SNP) markers for capsaicin content in *C. annuum*. Boxplot shows the effect of SNP marker in locus (A) CA05g18080 on chromosome 05 and (B) CA1109160 on chromosome 11. Y-axis represents the values for capsaicin levels ($\text{mg}\cdot\text{g}^{-1}$) in pepper powder from dried green fruit (A,B) and total fruit weight (C,D).

We analyzed the gene expression of different ANK subfamilies. A member of the ANK-BTB subfamily (*CaANK32*) was highly expressed in leaf, whereas another member of the same subfamily (*CaANK55*) showed high expression in placenta tissue at 16-dpa. Likewise, *CaANK28* and *CaANK17*, members of the ANK-ACBP and ANK-GPCR subfamilies, respectively, showed similar expression patterns and were highly expressed in placenta at 16-dpa. Members of the ANK-ZnF subfamily (*CaANK3* and *CaANK66*) were highly expressed in placenta tissue at 25 dpa. *CaANK70*, which belongs to the ANK-BPA subfamily, was highly expressed in the same tissue and stage. Members of the ANK-U and ANK-O subfamilies were evenly expressed across all tissues and stages.

We analyzed the expression profile of *CcANKs* genes on the basis of their RPKM values from RNA-seq data of different *C. chinense* varieties, generating a hierarchical cluster and the expression profile of genes in placenta tissue at 16-dpa based on the log values of each gene (Fig. 7). Among 96 *CcANK* genes, 72 were expressed in at least one *C. chinense* cultivar (Fig. 7A) and 36 were expressed across all six cultivars (Fig. 7B). Four *CcANK* genes (*CcANK1*, *CcANK13*, *CcANK35*, *CcANK52*) were exclusively expressed in PI257145, whereas *CcANK33*, *CcANK55* and *CcANK78* were only found in PI224448. For PI257129, only *CcANK29* was uniquely expressed, whereas *CcANK95* and *CcANK91* were exclusively expressed in Pimenta da neyde. *CcANK61*, previously described as a major marker for capsaicin and dihydrocapsaicin content, was mostly expressed in PI257129 and with moderate expression in Naga morich, whereas *CcANK81* was highly expressed in this last variety. *CcANK65*, another candidate marker associated with capsaicin content in *C. chinense*, was highly expressed in PI257145 but

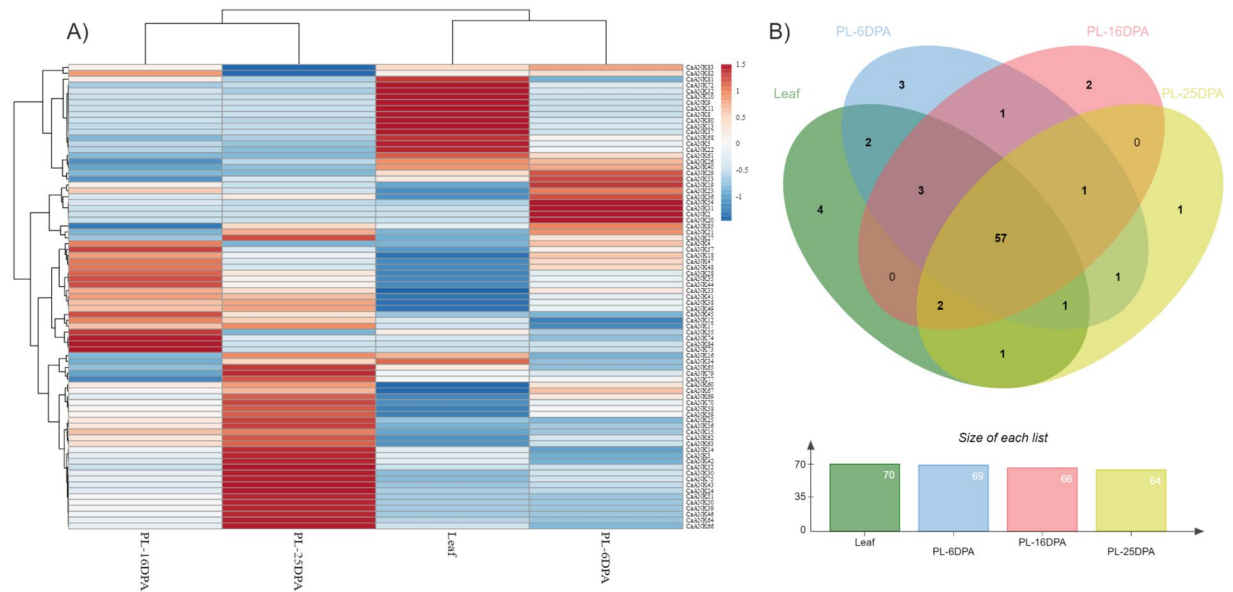


Figure 6. Expression patterns of *CaANK* genes in leaf and placenta tissue of *C. annuum* var CM344. **(A)** Heat map of expression profiles (in log₂-based RPKM) from leaf and placenta tissue (6, 16, 25 days post-anthesis [dpa]). The expression levels are represented by the color: red, upregulated, and blue, downregulated. **(B)** Venn diagram analysis of the tissue expression of *CaANK* genes. PL, placenta.

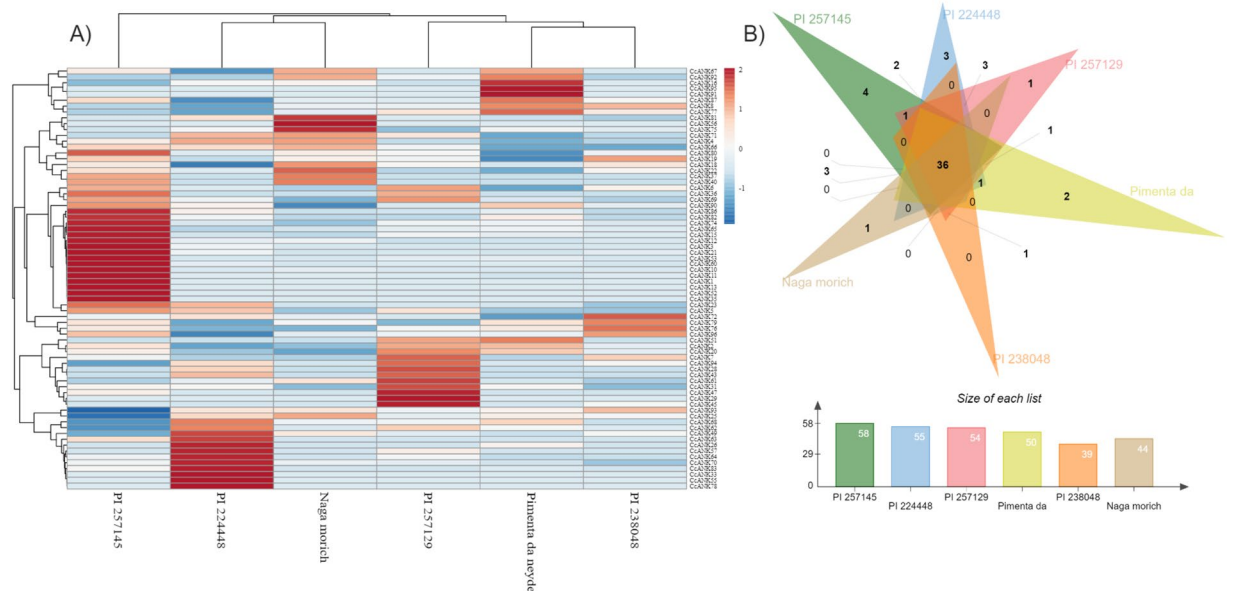


Figure 7. Expression patterns of *CcANK* genes in placenta tissue from six varieties of *C. chinense*. **(A)** Heat map of expression profiles (in log₂-based RPKM) from placenta tissue at 16 dpa for six *C. chinense* cultivars. The expression levels are represented by the color: red, upregulated, and blue, downregulated. **(B)** Venn diagram analysis of the cultivar expression of *CcANK* genes. PL, placenta.

with almost no expression in the other cultivars. In order of subfamilies, mainly ANK-U and ANK-O were unevenly expressed across all six cultivars, whereas the ANK-ZnF and ACBP subfamilies showed high expression in the Pimenta de neyde variety. Another member of the ANK-ZnF subfamily, *CcANK2*, was moderately expressed in PI257129. *CcANK23* showed moderate expression in PI224448 and PI257145. *CcANK10* and *CcANK80*, members of the ANK-GPCR and ANK-PK subfamilies, respectively, were highly expressed in PI257145.

Although the RNA-seq data provided relevant information about the expression pattern of ANK genes, there might be possible bias in the analysis. To confirm this information, we determined the preferential expression of CA05g18080, CA11g09160 and CC.CCv1.2.scaffold31.147 ortholog genes by using real-time PCR analysis of ANK genes in leaf and placenta tissue from the three *Capsicum* species at 16 and 25 dpa by using gene-specific primers (Fig. 8). The results showed a predominant expression of CA05g18080 orthologs in placenta at 16 dpa in

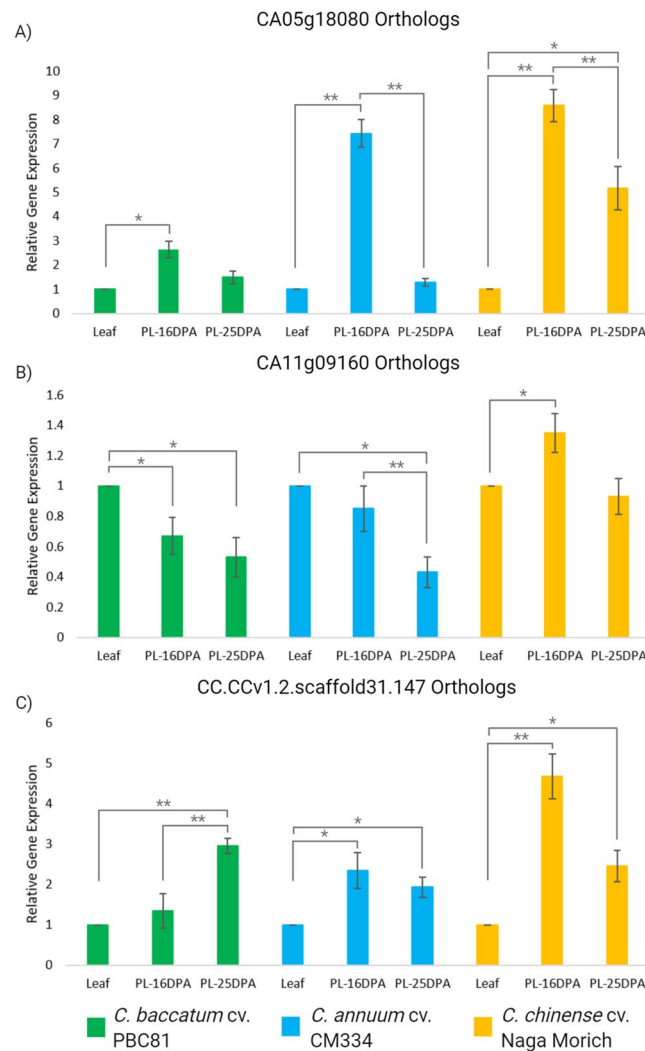


Figure 8. RT-qPCR analysis of mRNA expression of (A) CA05g18080, (B) CA11g09160 and (C) CC.CCv1.2.scaffold31.147 orthologs in leaf and placenta tissues at 16 and 25 dpa (days post-anthesis) across *Capsicum* species. Leaf was used as a calibrator for relative expression. Data are mean \pm SD; $n = 3$, with significant differences as $p \leq 0.05$ (*) and $p \leq 0.001$ (**). PL, placenta.

C. chinense, followed by *C. annuum* at the same stage in placenta tissue with significant difference at $p \leq 0.001$ compared to leaf and 25 dpa in both species. Similarly, the CA05g18080 ortholog had high expression in placenta tissue from *C. chinense* at 25 dpa. However, genes from *C. baccatum* showed the lowest expression levels, although the placental expression at 25 dpa was higher than in *C. annuum* at the same stage but not significant difference was found related to leaf and 16 dpa (Fig. 8A). In addition, the CA11g09160 ortholog from *C. chinense* showed high transcript abundance in placental tissues at 16 dpa, with lower expression at 25 dpa but these results were not significant. Contrary to *C. chinense*, the expression of this gene in *C. baccatum* and *C. annuum* was lowest in placenta tissues compared to leaf with significant difference at $p \leq 0.05$ (Fig. 8B). Similarly, the orthologs of CC.CCv1.2.scaffold31.147 were significantly expressed in *C. chinense* at 16 dpa ($p \leq 0.001$), followed by *C. annuum* ($p \leq 0.05$) compared to leaf, while there was not significant difference in *C. baccatum* at the same stage. The expression in placenta tissue at 25 dpa was high in *C. baccatum* showing significant difference at $p \leq 0.001$ in comparison to leaf and 16 dpa, nevertheless this expression was lower when comparing with *C. chinense* and *C. annuum* (Fig. 8C).

Discussion

ANK proteins are known to play important roles in various developmental processes of different plant species such as *Arabidopsis*, rice, maize, soybean and tomato^{18,23–25,27}. Although the role of ANK genes in plants was previously suggested, no systematic studies of the ANK gene family in *Capsicum* spp. have been performed; however, the recent availability of the pepper genome sequence has provided material support for the identification and characterization of varied gene families.

In this study, we identified 87 ANK genes in *C. baccatum*, 85 in *C. annuum* and 96 in *C. chinense*. All these genes were distributed among the 12 chromosomes at different densities across the three species. ANK repeat-containing proteins generally represent from 0.1% to 0.5% of the total proteins in most of the plants studied; however, this content can reach as much as 2% in plants such as *Ectocarpus siliculosus*³⁸. In pepper, this percentage of the annotated protein-coding genes ranged from 0.24% in *C. baccatum* and *C. annuum* to 0.27% in *C. chinense* (Table 1). Thus, the number of ANK genes is lower in pepper than other plant species such as *Arabidopsis*, rice, and tomato^{18,23,24}, which contain 0.41%, 0.49% and 0.38% of ANK genes, respectively. Given that the genome size of pepper is 3,500 Mbp³⁹, nearly 26 times larger than the *Arabidopsis* genome (135 Mbp)⁴⁰, about 8 times larger than the rice genome (430 Mbp)⁴¹, and almost 4 times the size of the tomato genome (950 Mbp)⁴², it is surprising that the number of ANK genes is much lower in pepper than in other plants. This paradox has also been observed in maize, with only 71 ANK genes reported in a genome size of 2,300 Mbp²⁵.

In comparison to other gene families such as ABC transporters, the number of ANK genes in pepper is low⁴³. However, this abundance is high as compared with other families such as the ARF³¹, DHN³⁰, DIR³² and GRAS⁴⁴ families, with 19, 7, 24 and 50 members, respectively. In general, most ANK proteins have 2 to 6 repeats, with a maximum of 34 repeats, which was reported in a *Giardia lamblia* protein⁴⁵. In this study, we found that *C. chinense* contained proteins with 14 and 18 ANK repeats, whereas *C. annuum* proteins had 17 repeats (Fig. S1).

The predicted proteins were classified into 10 subfamilies on the basis of their domain compositions, and a combined phylogenetic tree was constructed with the aligned *Capsicum* and *Arabidopsis* ANK protein sequences. Most of the members within the same group shared one or more different domains in addition to the ANK domain. Gene structure analysis showed that pepper ANK subgroup members contained a similar organization in terms of the number and length of introns and exons. For instance, CA05g18080 orthologs in *C. baccatum* and *C. chinense* had two exons, whereas *C. annuum* had four exons. Similarly, CA11g09160 orthologs contained two exons in all three *Capsicum* species.

Although other plant species contain more ANK genes^{23,24} than do pepper species, *Capsicum* species contain more ANK proteins with an exclusively ANK domain (48–64%), classified in the ANK-U subfamily. The second largest group after the ANK-U subfamily was the ANK-TM subfamily. The orthologs of CA05g18080 and CA11g01960 were classified in the ANK-U subfamily, and the locus CC.CCv1.2.scaffold31.147 (*CcANK65*), homolog of At5g02620, was a member of the ANK-TM subfamily. Members of the ANK-TM family such as *AKT1*, which contains five ANK repeats and several transmembrane domains, play a key role in nitrogen fixation in root nodules of *Lotus japonicus*⁴⁶ and in root K⁺ uptake^{47,48}.

The *Capsicum* ANK genes included the ANK-BTB subfamily, which are known to be involved in plant morphogenesis and serve as a protein–protein interaction motif in several transcription factors^{49,50}. The BLADE-ON-PETIOLE1 (*BOP1*) gene, which encodes for a BTB/POZ with an ANK repeat domain architecture, regulates the expression of the class I KNOX gene and modulates meristematic activity in leaves¹⁶. All *Capsicum* species analyzed had one gene belonging to the ANK-ACBP (Acyl-CoA-binding protein) subfamily and one more from the ANK-GPCR subfamily, which contains a GPCR-chaperone-1. This last subfamily was previously reported in tomato and soybean, but we have no evidence demonstrating its presence in model plants such as *Arabidopsis*, rice or maize. We did not identify ANK proteins having the ring finger domain (ANK-RF subfamily) or tetratricopeptide repeat domain (ANK-TPR) in any of the three species, even though these domains have been identified at different densities in tomato, *Arabidopsis*, rice, maize and soybean^{18,23–25,27}. Another identified subfamily was ANK-PK, with protein kinase activity. Previous studies described that the interaction between the receptor-like kinase complex and an ANK repeat domain can activate downstream defense signaling components, which suggests a role for ANK proteins in biotic stress⁵¹. For instance, Chinchilla *et al.*⁵², reported that the expression of an ANK protein kinase in alfalfa was induced by an osmotic effect.

Motif analysis with MEME revealed that the motifs *ACD6* and *ANK* were highly related. Earlier studies reported that the expression of *ACD6*, a protein containing a transmembrane region and a putative ankyrin repeat, was directly regulated by SA, which is highly required in response to plant pathogens and diseases^{53,54}. Likewise, the gain of function of the *acd6-1 Arabidopsis* mutant was influenced by SA and light to activate an immune response against pathogens during infection⁵⁴. These findings support plant ANK proteins as playing a crucial role in biotic stress response, mainly against pathogens and diseases. Nodzson *et al.*⁵⁵, found that the expression of some ANK genes in plants can be affected by auxin, ABA and SA/jasmonic acid, which principally mediate the responses of plants to biotic and abiotic stress. The promoter elements found in *Capsicum* ANK genes were related to auxin, ABA and SA, and to GA, which has been reported to increase the growth and yield of *C. annuum* under greenhouse conditions⁵⁶. Seong *et al.*⁵⁷, demonstrated that the expression of *CaKRI* (Ankyrin Repeat-Containing Zinc Finger Protein) in *C. annuum* was strongly induced by SA and an ethylene regulator. One group of 11 *Cis*-elements was associated with binding protein site function, which supports that ANK proteins are evolutionarily conserved protein domains involved in mediating protein–protein interactions in different molecular process. Additionally, five *Cis*-elements were related to dehydration and water stress. Some ANK proteins may be involved in transpiration and adaptation to water deprivation stress, which agrees with the above results. Disruption of *AKT1* in *Arabidopsis* mutants increased stomatal closure during water deficit, thereby enhancing drought tolerance⁵⁸.

The number of ANK genes is higher in *C. chinense* than other pepper species, so this species may be more resistant to drought stress while containing a higher amount of capsaicinoids^{59,60}. Indeed, hot pepper cultivars with high capsaicinoid content are less sensitive to drought stress than are low and medium pungent cultivars⁶¹. Nevertheless, this feature cannot be generalized because many factors affect capsaicinoid content, including genotypic variation⁶². For instance, the different *C. chinense* varieties used in this study contained different capsaicinoid content (Fig. 4), which confirms that the pungency level is cultivar-dependent in part^{63,64}.

Gene duplication events are important for the rapid expansion and evolution of gene families and are crucial for the origin of new gene functions⁶⁵. Among the different types of duplication, segmental and tandem

duplication are the most common involved in gene family expansion⁶⁶. In our study, the average duplication time of syntenic *Capsicum* ANK paralogs pairs was ~1.1 MYA, which is close to the estimated lineage-divergence times of *C. baccatum* and a progenitor of the other two pepper species (~1.7 MYA). However, the duplication time between *C. annuum* and *C. chinense* was reported as 1.14 MYA⁶⁷. We identified a total of 41 *Capsicum* ANK genes in 23 pairs of syntenic paralogs across all the species. The selection pressure type was measured according to the ratio of non-synonymous to synonymous substitutions ω ($=K_a/K_s$). The K_a/K_s ratios of 22 paralog pairs were <1 , so these paralogs were under purifying selection pressures. Furthermore, the ω ratio of one paralog pair was >1 , representing positive selection and fast evolutionary rate. These findings are similar to those for other gene families such as BURP in Medicago and ACD in tomato, which contain a few or even no paralog pairs undergoing positive selection^{68,69}.

Most of the *Capsicum* ANK products were predicted to be membrane, intracellular and chloroplast proteins. ANK repeats have been found in proteins with different functions, including mitochondrial proteins, cell cycle regulation, cytoskeleton interactions, signal transduction, disease resistance and stress responses^{1,70}. We found that *Capsicum* ANK genes are mainly involved in response to stress and endogenous stimulus and are involved in signal transduction in all *Capsicum* species. The response to endogenous stimulus as a biological process supports the participation of ANK genes in stress tolerance, as was shown in other species. For example, the ankyrin repeat-containing XA21 binding protein 3 (XB3), which forms a protein complex with receptor-like kinase (XA21), confers resistance to bacterial blight caused by *Xanthomonas oryzae* in rice⁵¹. Similarly, the *OsPIANK1* gene positively regulates rice basal defense against blast fungus disease²². The ANK repeat–receptor-like protein complex mediates the plant immune response, but AKR2 plays a role in plant reactive oxygen species scavenging and metabolism¹⁰, and the OXIDATIVE STRESS 2 protein may be an activator in a stress response pathway⁷¹.

CaANK genes exhibited stage-specific expression, whereas *CcANK* genes showed a genotype-specific expression pattern, which suggests that the expression profile of these genes differs across *Capsicum* species. Furthermore, different patterns in the expression between orthologs of the Capsaicinoid markers previously mentioned demonstrated that the expression of *Capsicum* ANK genes was extensive during different development stages and may be species-specific for each of the ANK genes in *Capsicum* species. Overall, the identification of the putative *Capsicum* ANK genes, classification of these genes, construction of a phylogenetic tree, analysis of gene structure and more detailed knowledge of the expression profile of the *Capsicum* ANK genes may provide clues regarding the function of this gene family. Our findings provide new insights into the *Capsicum* ANK gene family, which helps improve our understanding of the possible role of these genes in aspects such as plant and fruit development, response against stress, and capsaicinoid content in pepper fruits.

Materials and Methods

Plant materials. *C. baccatum* cv. PBC81, *C. annuum* cv. CM334 and six varieties of *C. chinense* (i.e., Pimenta da neyde, Naga morich, PI 224448, PI 238048, PI 257129 and PI 257145) were grown in triplicate samples in an experimental field at West Virginia State University, adapting a row-to-plant spacing of 100 × 30 cm. Leaf and fruits at 6, 16 and 25 days post-anthesis (dpa) were collected from all cultivars and stored at -80°C . Quantitative analysis of capsaicin and dihydrocapsaicin content in green pepper fruits (16 dpa) involved using the 1200 series HPLC system (Agilent Technologies, Santa Clara, CA) as described⁷².

Identification of ANK genes in pepper. To identify all candidate members of the ANK gene family in pepper genomes, we downloaded the previously reported ANK protein sequences from *Arabidopsis*²³ from the TAIR database (<https://www.Arabidopsis.org/>) and used them as a query in a local BLASTP with E-value cut-off of $1e-3$ against the proteomes of the three *Capsicum* species downloaded from the Pepper Genome Platform (PGP) (<http://passport.pepper.snu.ac.kr/?t=PGENOME>)²⁸. The typical ANK repeat domains PF00023 and SM00248 were searched across all the output genes by using the Pfam web server (<http://Pfam.sanger.ac.uk/>)⁷³ and SMART database (http://smart.embl-heidelberg.de/smart/set_mode.cgi?NORMAL=1)⁷⁴ respectively, with default cut-off parameters. Genes with E-value $> 1E-05$ and redundant genes were deleted. Analysis of the domain structure for all the peptide sequences of ANK genes selected involved using the latest Hidden Markov Model (HMM) with the HMMER3.0 software⁷⁵. Protein size, MW and theoretical pI of each ANK gene were predicted by using the proteome database and sequence analysis tools on the ExpASY Proteomics Server (<http://expasy.org/>)⁷⁶.

MEME motif analysis. Conserved motifs of all ANK proteins from the three *Capsicum* species were identified by using the motif investigation software Multiple Em for Motif Elicitation (MEME) (<http://meme-suite.org/tools/meme>)⁷⁷. The analysis was performed with maximum number of motifs 10 and optimum width of motif from 6 to 50. To identify motif function, discovered MEME motifs were searched in the ExpASY-PROSITE database by using the ScanProsite tool (<https://prosite.expasy.org/scanprosite/>)⁷⁸. Prediction of the tertiary structure and homologs of ANK domain-containing proteins associated with capsaicin content in pepper for the three *Capsicum* species involved using the online server Phyre2 (Protein Homology/Analog Y Recognition Engine; <http://www.sbg.bio.ic.ac.uk/phyre2>) under the “intensive” mode as described by Kelley *et al.*⁷⁹.

Chromosomal location and gene structures of ANK genes. The chromosomal position of individual ANK genes was extracted from the PGP and the genes were physically mapped on each of the 12 chromosomes by using MapInspect. Structural analysis of ANK genes was generated by using the Gene Structure Display Server (GSDS: <http://gsds.cbi.pku.edu.cn/>)⁸⁰. For *Cis*-element analysis, all ANK gene promoter sequences (1,500 bp upstream of the initiation codon “ATG”) were extracted from the pepper genomes. Then, the *cis*-regulatory elements of promoters for each gene were identified by using PLACE: A database of plant *cis*-acting regulatory DNA elements (<http://www.dna.affrc.go.jp/PLACE/>)⁸¹.

Sequence alignment and phylogenetic analysis. The full-length ANK protein sequences from *Arabidopsis* and *Capsicum* species were aligned by using ClustalW as described⁸². The alignment file was then used to construct a phylogenetic tree based on the maximum-likelihood method of the MEGAX (Molecular Evolutionary Genetics Analysis) software⁸³. The phylogenetic tree of the ANK proteins was displayed by using the interactive Tree Of Life platform (iTOL; <http://itol.embl.de/index.shtml>) after bootstrap analysis with 1000 replicates⁸⁴.

Gene synteny analysis of ANK proteins. The syntenic ANK paralog pairs across the *Capsicum* species were identified by searching gene duplication with the following criteria: (1) genes with >70% coverage of the alignment length; (2) genes with >70% identity in the aligned region; and (3) a minimum of two duplication events considered for strongly connected genes⁸⁵. For each paralog pair, the non-synonymous substitution rate (Ka), synonymous substitution rate (Ks) and ω ($=Ka/Ks$) of paralog pairs were estimated by using KaKs_Calculator 2.0⁸⁶. The duplication date of paralog pairs was estimated by the formula $T = Ks/2\lambda$, assuming a clock-like rate (λ) of 6.96 synonymous substitutions per 10^{-9} years⁸⁷. The syntenic plot was generated by using the BioCircos package in R.

cDNA library construction and RNA-seq of *C. chinense* green fruits. Green fruits (16 dpa) from six different cultivars of *C. chinense* were used for whole-transcriptome sequencing. Total RNA was isolated from the pooled tissues of three biological replicates for each cultivar by using the plant RNA mini spin kit (Macherey-Nagel). Total mRNA was isolated, fragmented, and reverse-transcribed into cDNA. Double-stranded cDNA was then purified by using 1.8x Agencourt AMP XP beads. Sequencing libraries were constructed by using the NEBNext Ultra II RNA Library Prep Kit according to the manufacturer's protocol. To confirm accuracy before sequencing, the insert size and integrity of the libraries were analyzed with an Agilent 2100 Bioanalyzer (Invitrogen), and the Qubit 4 Fluorometer (Invitrogen) was used for library quantification. The RNA sequencing library from each sample was sequenced in the Illumina NextSeq. 500 platform to produce paired-end reads.

Sequencing analysis and functional annotation. High-quality reads were obtained by removing the adapter sequence with Cutadapt and low-quality reads (Phred score $QV < 30$) with the Sickle program^{88,89}. All cleaned reads were mapped to the *C. chinense* reference genome version 1.2 by using the mem algorithm of the BWA tool to generate a SAM alignment. The read count table for genes from *C. chinense* was created for all samples by using the SAM alignment and HTSeq R package^{90,91}. The gene expression based on the read counts were normalized by calculating the reads per kilobase of transcript per million mapped reads (RPKM) values. The RPKM values for each gene were calculated by using the read count table, total number of reads, and gene length (kb). Found genes were functionally annotated to analyze their gene ontology (GO) by using the Blast2GO application (<http://www.blast2go.com>)⁹².

Expression pattern of ANK genes in *C. annuum* and *C. chinense* and qRT-PCR validation. RNA sequencing (RNA-seq) gene expression data of leaf and placenta tissues (6, 16, 25 dpa) from *C. annuum* cv. CM334 were retrieved from the RNA-seq data published by Kim *et al.*²⁸. The RPKM expression values were used to generate a heatmap for ANK genes from *C. annuum* and *C. chinense* by using the ClustVis web tool (<https://biit.cs.ut.ee/clustvis/>)⁹³. RNA-seq data were further validated by qRT-PCR analysis of samples from different tissues (i.e., leaf, and placenta tissues at 6, 16 and 25 dpa). Synthesis of cDNA involved 1 μ g total RNA with oligo dT primers and SuperScript IV reverse transcriptase. The resulting cDNA was subjected to RT-qPCR analysis with a total volume of 20 μ L containing 1 μ L cDNA template, 2 μ L forward and reverse primers (10 μ M), 10 μ L SYBR Green PCR Master (ROX) (Roche, Shanghai) and 7 μ L sterile distilled water on a StepOne Plus Real-Time PCR System. The two step RT-qPCR program began at 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. The reactions were performed in three biological replicates with three technical replications to compute the average Ct values. The relative expression of specific genes was quantified with the $2^{-\Delta\Delta Ct}$ calculation method⁹⁴. Relative gene expression of target genes was normalized to that of the endogenous control β -tubulin⁹⁵ and leaf tissue as a calibrator. Student's t-test was used to identify statistical significance between samples.

Data availability

The raw Illumina mRNA-seq reads generated and/or analyzed during the current study are available in the Sequence Read Archive repository (NCBI-SRA) under the following accession numbers PRJNA526219 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA526219>) and PRJNA562491 (<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA562491>).

Received: 22 September 2019; Accepted: 20 February 2020;

Published online: 04 March 2020

References

- Sedgwick, S. G. & Smerdon, S. J. The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends in Biochemical Sciences* **24**, 311–316, [https://doi.org/10.1016/S0968-0004\(99\)01426-7](https://doi.org/10.1016/S0968-0004(99)01426-7) (1999).
- Breeden, L. & Nasmyth, K. Cell cycle control of the yeast *HO* gene: Cis- and Trans-acting regulators. *Cell* **48**, 389–397, [https://doi.org/10.1016/0092-8674\(87\)90190-5](https://doi.org/10.1016/0092-8674(87)90190-5) (1987).
- Breeden, L. & Nasmyth, K. Similarity between cell-cycle genes of budding yeast and fission yeast and the Notch gene of *Drosophila*. *Nature* **329**, 651–654, <https://doi.org/10.1038/329651a0> (1987).
- Lux, S. E., John, K. M. & Bennett, V. Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. *Nature* **344**, 36–42, <https://doi.org/10.1038/344036a0> (1990).
- Chakrabarty, B. & Parekh, N. Identifying tandem Ankyrin repeats in protein structures. *BMC bioinformatics* **15**, 6599, <https://doi.org/10.1186/s12859-014-0440-9> (2014).

6. Vo, K. T. X. *et al.* Molecular insights into the function of ankyrin proteins in plants. *Journal of Plant Biology* **58**, 271–284, <https://doi.org/10.1007/s12374-015-0228-0> (2015).
7. Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z. Y. The ankyrin repeat as molecular architecture for protein recognition. *Protein science: a publication of the Protein Society* **13**, 1435–1448, <https://doi.org/10.1110/ps.03554604> (2004).
8. Michaely, P. & Bennett, V. The ANK repeat: a ubiquitous motif involved in macromolecular recognition. *Trends in Cell Biology* **2**, 127–129, [https://doi.org/10.1016/0962-8924\(92\)90084-Z](https://doi.org/10.1016/0962-8924(92)90084-Z) (1992).
9. Li, J., Mahajan, A. & Tsai, M.-D. Ankyrin Repeat: A Unique Motif Mediating Protein–Protein Interactions. *Biochemistry* **45**, 15168–15178, <https://doi.org/10.1021/bi062188q> (2006).
10. Shen, G. *et al.* Ankyrin repeat-containing protein 2A Is an essential molecular chaperone for peroxisomal membrane-bound Ascorbate Peroxidase3 in *Arabidopsis*. *The Plant Cell* **22**, 811–831, <https://doi.org/10.1105/tpc.109.065979> (2010).
11. Yuan, X. *et al.* Global analysis of ankyrin repeat domain C3HC4-type RING finger gene family in plants. *PLoS one* **8**, e58003, <https://doi.org/10.1371/journal.pone.0058003> (2013).
12. Garcion, C. *et al.* AKRP and EMB506 are two ankyrin repeat proteins essential for plastid differentiation and plant development in *Arabidopsis*. *The Plant journal: for cell and molecular biology* **48**, 895–906, <https://doi.org/10.1111/j.1365-313X.2006.02922.x> (2006).
13. Albert, S. *et al.* The EMB 506 gene encodes a novel ankyrin repeat containing protein that is essential for the normal development of *Arabidopsis* embryos. *The Plant Journal* **17**, 169–179, <https://doi.org/10.1046/j.1365-313X.1999.00361.x> (1999).
14. Bae, W. *et al.* AKR2A-mediated import of chloroplast outer membrane proteins is essential for chloroplast biogenesis. *Nature cell biology* **10**, 220–227, <https://doi.org/10.1038/ncb1683> (2008).
15. Cui, Y. L. *et al.* The *GDC1* gene encodes a novel ankyrin domain-containing protein that is essential for grana formation in *Arabidopsis*. *Plant physiology* **155**, 130–141, <https://doi.org/10.1104/pp.110.165589> (2011).
16. Ha, C. M., Jun, J. H., Nam, H. G. & Fletcher, J. C. BLADE-ON-PETIOLE1 Encodes a BTB/POZ Domain Protein Required for Leaf Morphogenesis in *Arabidopsis thaliana*. *Plant and Cell Physiology* **45**, 1361–1370, <https://doi.org/10.1093/pcp/pch201> (2004).
17. Huang, J. *et al.* An ankyrin repeat-containing protein, characterized as a ubiquitin ligase, is closely associated with membrane-enclosed organelles and required for pollen germination and pollen tube growth in lily. *Plant physiology* **140**, 1374–1383, <https://doi.org/10.1104/pp.105.074922> (2006).
18. Huang, J. *et al.* The ankyrin repeat gene family in rice: genome-wide identification, classification and expression profiling. *Plant molecular biology* **71**, 207–226, <https://doi.org/10.1007/s11103-009-9518-6> (2009).
19. Sakamoto, H., Nakagawara, Y. & Oguri, S. The Expression of a Novel Gene Encoding an Ankyrin-Repeat Protein, DRA1, is Regulated by Drought-Responsive Alternative Splicing. *International Journal of Biotechnology and Bioengineering* **7** (2013).
20. Sakamoto, H., Matsuda, O. & Iba, K. ITN1, a novel gene encoding an ankyrin-repeat protein that affects the ABA-mediated production of reactive oxygen species and is involved in salt-stress tolerance in *Arabidopsis thaliana*. *The Plant journal: for cell and molecular biology* **56**, 411–422, <https://doi.org/10.1111/j.1365-313X.2008.03614.x> (2008).
21. Lu, H., Liu, Y. & Greenberg, J. T. Structure-function analysis of the plasma membrane-localized *Arabidopsis* defense component ACD6. *The Plant journal: for cell and molecular biology* **44**, 798–809, <https://doi.org/10.1111/j.1365-313X.2005.02567.x> (2005).
22. Mou, S. *et al.* Functional Analysis and Expression Characterization of Rice Ankyrin Repeat-Containing Protein, OsPIANK1, in Basal Defense against *Magnaporthe oryzae* Attack. *PLoS one* **8**, e59699, <https://doi.org/10.1371/journal.pone.0059699> (2013).
23. Becerra, C., Jahrmann, T., Puigdomenech, P. & Vicient, C. M. Ankyrin repeat-containing proteins in *Arabidopsis*: characterization of a novel and abundant group of genes coding ankyrin-transmembrane proteins. *Gene* **340**, 111–121, <https://doi.org/10.1016/j.gene.2004.06.006> (2004).
24. Yuan, X. *et al.* Superfamily of ankyrin repeat proteins in tomato. *Gene* **523**, 126–136, <https://doi.org/10.1016/j.gene.2013.03.122> (2013).
25. Jiang, H. *et al.* Genome-wide identification and expression profiling of ankyrin-repeat gene family in maize. *Development genes and evolution* **223**, 303–318, <https://doi.org/10.1007/s00427-013-0447-7> (2013).
26. Mahmood, N. & Tamanna, N. Analyses of *Physcomitrella patens* Ankyrin Repeat Proteins by Computational Approach. *Molecular biology international* **2016**, 9156735, <https://doi.org/10.1155/2016/9156735> (2016).
27. Zhang, D. *et al.* Genome-wide characterization of the ankyrin repeats gene family under salt stress in soybean. *The Science of the total environment* **568**, 899–909, <https://doi.org/10.1016/j.scitotenv.2016.06.078> (2016).
28. Kim, D. H. *et al.* An ankyrin repeat domain of AKR2 drives chloroplast targeting through coincident binding of two chloroplast lipids. *Developmental cell* **30**, 598–609, <https://doi.org/10.1016/j.devcel.2014.07.026> (2014).
29. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 5135–5140, <https://doi.org/10.1073/pnas.1400975111> (2014).
30. Jing, H. *et al.* Genome-Wide Identification, Expression Diversification of Dehydrin Gene Family and Characterization of CaDHN3 in Pepper (*Capsicum annuum* L.). *PLoS one* **11**, e0161073, <https://doi.org/10.1371/journal.pone.0161073> (2016).
31. Zhang, H., Cao, N., Dong, C. & Shang, Q. Genome-wide Identification and Expression of ARF Gene Family during Adventitious Root Development in Hot Pepper (*Capsicum annuum*). *Horticultural Plant Journal* **3**, 151–164, <https://doi.org/10.1016/j.hpj.2017.07.001> (2017).
32. Khan, A. *et al.* Genome-wide analysis of dirigent gene family in pepper (*Capsicum annuum* L.) and characterization of CaDIR7 in biotic and abiotic stresses. *Scientific reports* **8**, 5500, <https://doi.org/10.1038/s41598-018-23761-0> (2018).
33. Nimmakayala, P. *et al.* Genome-wide Diversity and Association Mapping for Capsaicinoids and Fruit Weight in *Capsicum annuum* L. *Scientific reports* **6**, 38081, <https://doi.org/10.1038/srep38081> (2016).
34. Park, M. *et al.* A major QTL and candidate genes for capsaicinoid biosynthesis in the pericarp of *Capsicum chinense* revealed using QTL-seq and RNA-seq. *THEORETICAL AND APPLIED GENETICS* **132**, 515–529, <https://doi.org/10.1007/s00122-018-3238-8> (2019).
35. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960 (2004).
36. Kajava, A. V. Tandem repeats in proteins: from sequence to structure. *Journal of structural biology* **179**, 279–288, <https://doi.org/10.1016/j.jsb.2011.08.009> (2012).
37. Koonin, E. V. & Rogozin, I. B. Getting positive about selection. *Genome biology* **4**, 331, <https://doi.org/10.1186/gb-2003-4-8-331> (2003).
38. Mahmood, N. *et al.* In silico analysis reveals the presence of a large number of Ankyrin repeat containing proteins in *Ectocarpus siliculosus*. *Interdisciplinary sciences, computational life sciences* **4**, 291–295, <https://doi.org/10.1007/s12539-012-0134-9> (2012).
39. Hulse-Kemp, A. M. *et al.* Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Horticulture research* **5**, 4, <https://doi.org/10.1038/s41438-017-0011-0> (2018).
40. The Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815, <https://doi.org/10.1038/35048692> (2000).
41. Eckardt, N. A. Sequencing the Rice Genome. *The Plant Cell* **12**, 2011–2017, <https://doi.org/10.1105/tpc.12.11.2011> (2000).
42. Tomato Genome, C. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641, <https://doi.org/10.1038/nature11119> (2012).
43. Lopez-Ortiz, C. *et al.* Genome-wide identification and gene expression pattern of ABC transporter gene family in *Capsicum* spp. *PLoS one* **14**, e0215901, <https://doi.org/10.1371/journal.pone.0215901> (2019).

44. Liu, B., Sun, Y., Xue, J., Jia, X. & Li, R. Genome-wide characterization and expression analysis of GRAS gene family in pepper (*Capsicum annuum* L.). *PeerJ* **6**, e4796, <https://doi.org/10.7717/peerj.4796> (2018).
45. Elmendorf, H. G., Rohrer, S. C., Houry, R. S., Boutenot, R. E. & Nash, T. E. Examination of a novel head-stalk protein family in *Giardia lamblia* characterised by the pairing of ankyrin repeats and coiled-coil domains. *International journal for parasitology* **35**, 1001–1011 (2005).
46. Kumagai, H. *et al.* A novel ankyrin-repeat membrane protein, IGN1, is required for persistence of nitrogen-fixing symbiosis in root nodules of *Lotus japonicus*. *Plant physiology* **143**, 1293–1305, <https://doi.org/10.1104/pp.106.095356> (2007).
47. Rubio, F., Nieves-Cordones, M., Alemán, F. & Martínez, V. Relative contribution of *AtHAK5* and *AtAKT1* to K⁺ uptake in the high-affinity range of concentrations. *Physiologia plantarum* **134**, 598–608 (2008).
48. Aleman, F., Nieves-Cordones, M., Martínez, V. & Rubio, F. Root K⁽⁺⁾ acquisition in plants: the *Arabidopsis thaliana* model. *Plant & cell physiology* **52**, 1603–1612, <https://doi.org/10.1093/pcp/pcr096> (2011).
49. Ha, C. M. *et al.* The BLADE-ON-PETIOLE 1 gene controls leaf pattern formation through the modulation of meristematic activity in *Arabidopsis*. *Development* **130**, 161–172, <https://doi.org/10.1242/dev.00196> (2003).
50. Guan, P. *et al.* Global evolution and expression analysis of BTB-containing ankyrin repeat genes in plants. *Archives of Biological Sciences* **70**, 249–258, <https://doi.org/10.2298/abs170306042g> (2018).
51. Wang, Y. S. *et al.* Rice XA21 binding protein 3 is a ubiquitin ligase required for full Xa21-mediated disease resistance. *Plant Cell* **18**, 3635–3646, <https://doi.org/10.1105/tpc.106.046730> (2006).
52. Chinchilla, D. *et al.* Ankyrin protein kinases: a novel type of plant kinase gene whose expression is induced by osmotic stress in alfalfa. *Plant molecular biology* **51**, 555–566, <https://doi.org/10.1023/a:1022337221225> (2003).
53. Parra, R. G., Espada, R., Verstraete, N. & Ferreira, D. U. Structural and Energetic Characterization of the Ankyrin Repeat Protein Family. *PLoS computational biology* **11**, e1004659, <https://doi.org/10.1371/journal.pcbi.1004659> (2015).
54. Sharma, M. & Pandey, G. K. Expansion and Function of Repeat Domain Proteins During Stress and Development in Plants. *Frontiers in plant science* **6**, 1218, <https://doi.org/10.3389/fpls.2015.01218> (2015).
55. Nodzon, L. A. *et al.* The ubiquitin ligase XBAT32 regulates lateral root development in *Arabidopsis*. *The Plant journal: for cell and molecular biology* **40**, 996–1006, <https://doi.org/10.1111/j.1365-313X.2004.02266.x> (2004).
56. Batlang, U. Benzyladenine plus Gibberellins (GA4 + 7) Increase Fruit Size and Yield in Greenhouse-Grown Hot Pepper (*Capsicum annuum* L.). *Journal of Biological Sciences* **8**, 659–662 (2008).
57. Seong, E. *et al.* Characterization of a Stress-Responsive Ankyrin Repeat-Containing Zinc Finger Protein of *Capsicum annuum* (CaKR1). *Journal of Biochemistry and Molecular Biology* **40**, 952–958 (2007).
58. Nieves-Cordones, M., Caballero, F., Martínez, V. & Rubio, F. Disruption of the *Arabidopsis thaliana* inward-rectifier K⁺ channel AKT1 improves plant responses to water stress. *Plant & cell physiology* **53**, 423–432, <https://doi.org/10.1093/pcp/pcr194> (2012).
59. Estrada, B., Pomar, F., Diaz, J., Merino, F. & Bernal, M. A. Pungency level in fruits of the Padrón pepper with different water supply. *Scientia Horticulturae* **81**, 385–396, [https://doi.org/10.1016/S0304-4238\(99\)00029-1](https://doi.org/10.1016/S0304-4238(99)00029-1) (1999).
60. Okunlola, G. O., Olatunji, O. A., Akinwale, R. O., Tariq, A. & Adelusi, A. A. Physiological response of the three most cultivated pepper species (*Capsicum* spp.) in Africa to drought stress imposed at three stages of growth and development. *Scientia Horticulturae* **224**, 198–205, <https://doi.org/10.1016/j.scienta.2017.06.020> (2017).
61. Phimchan, P., Chanthai, S., Bosland, P. W. & Techawongstien, S. Enzymatic changes in phenylalanine ammonia-lyase, cinnamic-4-hydroxylase, capsaicin synthase, and peroxidase activities in *capsicum* under drought stress. *Journal of agricultural and food chemistry* **62**, 7057–7062, <https://doi.org/10.1021/jf4051717> (2014).
62. Phimchan, P., Techawongstien, S., Chanthai, S. & Bosland, P. W. Impact of Drought Stress on the Accumulation of Capsaicinoids in *Capsicum* Cultivars with Different Initial Capsaicinoid Levels. **47**, 1204, <https://doi.org/10.21273/hortsci.47.9.1204> (2012).
63. Zewdie, Y. & Bosland, P. W. Evaluation of genotype, environment, and genotype-by-environment interaction for capsaicinoids in *Capsicum annuum* L. *Euphytica* **111**, 185–190, <https://doi.org/10.1023/a:1003837314929> (2000).
64. Gurung, T., Techawongstien, S., Suriharn, B. & Techawongstien, S. Impact of Environments on the Accumulation of Capsaicinoids in *Capsicum* spp. **46**, 1576, <https://doi.org/10.21273/hortsci.46.12.1576> (2011).
65. Roth, C. *et al.* Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *Journal of experimental zoology. Part B, Molecular and developmental evolution* **308**, 58–73, <https://doi.org/10.1002/jez.b.21124> (2007).
66. Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D. & May, G. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC plant biology* **4**, 10, <https://doi.org/10.1186/1471-2229-4-10> (2004).
67. Kim, S. *et al.* New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome biology* **18**, 210, <https://doi.org/10.1186/s13059-017-1341-9> (2017).
68. Li, Y. *et al.* Identification and Expression Analysis of BURP Domain-Containing Genes in *Medicago truncatula*. *Frontiers in plant science* **7**, 485, <https://doi.org/10.3389/fpls.2016.00485> (2016).
69. Paul, A., Rao, S. & Mathur, S. The alpha-Crystallin Domain Containing Genes: Identification, Phylogeny and Expression Profiling in Abiotic Stress, Phytohormone Response and Development in Tomato (*Solanum lycopersicum*). *Frontiers in plant science* **7**, 426, <https://doi.org/10.3389/fpls.2016.00426> (2016).
70. Yan, J., Wang, J. & Zhang, H. An ankyrin repeat-containing protein plays a role in both disease resistance and antioxidation metabolism. *The Plant Journal* **29**, 193–202, <https://doi.org/10.1046/j.0960-7412.2001.01205.x> (2002).
71. Blanvillain, R., Wei, S., Wei, P., Kim, J. H. & Ow, D. W. Stress tolerance to stress escape in plants: role of the OXS2 zinc-finger transcription factor family. *The EMBO Journal* **30**, 3812–3822, <https://doi.org/10.1038/emboj.2011.270> (2011).
72. Nimmakayala, P. *et al.* Linkage disequilibrium and population-structure analysis among *Capsicum annuum* L. cultivars for use in association mapping. *Molecular genetics and genomics* **289**, 513–521 (2014).
73. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42**, D222–230, <https://doi.org/10.1093/nar/gkt1223> (2014).
74. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic acids research* **46**, D493–D496, <https://doi.org/10.1093/nar/gkx922> (2018).
75. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic acids research* **46**, W200–W204, <https://doi.org/10.1093/nar/gky448> (2018).
76. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research* **31**, 3784–3788, <https://doi.org/10.1093/nar/gkg563> (2003).
77. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202–208, <https://doi.org/10.1093/nar/gkp335> (2009).
78. de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research* **34**, W362–365, <https://doi.org/10.1093/nar/gkl124> (2006).
79. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845, <https://doi.org/10.1038/nprot.2015.053> (2015).
80. Hu, B. *et al.* GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* **31**, 1296–1297, <https://doi.org/10.1093/bioinformatics/btu817> (2015).
81. Higo, K., Ugawa, Y., Iwamoto, M. & Korenaga, T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic acids research* **27**, 297–300, <https://doi.org/10.1093/nar/27.1.297> (1999).

82. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools. *Nucleic acids research* **25**, 4876–4882, <https://doi.org/10.1093/nar/25.24.4876> (1997).
83. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution* **35**, 1547–1549, <https://doi.org/10.1093/molbev/msy096> (2018).
84. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128, <https://doi.org/10.1093/bioinformatics/btl529> (2007).
85. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature genetics* **31**, 205 (2002).
86. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genomics, Proteomics & Bioinformatics* **8**, 77–80, [https://doi.org/10.1016/s1672-0229\(10\)60008-3](https://doi.org/10.1016/s1672-0229(10)60008-3) (2010).
87. Moniz de Sá, M. & Drouin, G. Phylogeny and substitution rates of angiosperm actin genes. *Molecular biology and evolution* **13**, 1198–1212, <https://doi.org/10.1093/oxfordjournals.molbev.a025685> (1996).
88. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3, <https://doi.org/10.14806/ej.17.1.200> (2011).
89. Joshi, N. & Fass, J. Sickle-A windowed adaptive trimming tool for FASTQ file using quality. *Online publication*, <https://github.com/najoshi/sickle> [Google Scholar] (2011).
90. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595, <https://doi.org/10.1093/bioinformatics/btp698> (2010).
91. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169, <https://doi.org/10.1093/bioinformatics/btu638> (2015).
92. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676, <https://doi.org/10.1093/bioinformatics/bti610> (2005).
93. Metsalu, T. & Vilo, J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic acids research* **43**, W566–570, <https://doi.org/10.1093/nar/gkv468> (2015).
94. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{(-Delta Delta C(T))} Method. *Methods* **25**, 402–408, <https://doi.org/10.1006/meth.2001.1262> (2001).
95. Wan, H. *et al.* Identification of reference genes for reverse transcription quantitative real-time PCR normalization in pepper (*Capsicum annuum* L.). *Biochemical and biophysical research communications* **416**, 24–30, <https://doi.org/10.1016/j.bbrc.2011.10.105> (2011).

Acknowledgements

This study was supported by the National Institute of Food and Agriculture USDA-NIFA (grant nos. 2017-38821-26434 and 1008515) and Department of Defense award (Agreement Number W911NF-16-1-0423) for next-generation sequencing instrument.

Author contributions

Conceptualization, U.K.R., P.N., C.L.; Formal Analysis, C.L., Y.P.; Funding Acquisition, U.K.R., P.N.; Investigation, C.L., Y.P., P.N.A., M.B., V.A., S.D., L.Y.; Writing – original draft C.L., Y.P.; Writing – review and editing, C.L., Y.P., J.S., P.N. and U.K.R. All authors agree on the accuracy and integrity of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-61057-4>.

Correspondence and requests for materials should be addressed to P.N. or U.K.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020