



Plant Archives

Journal homepage: <http://www.plantarchives.org>

DOI Url : <https://doi.org/10.51470/PLANTARCHIVES.2026.v26.no.2.194>

GENOME-WIDE IDENTIFICATION AND COMPREHENSIVE *IN-SILICO* CHARACTERIZATION OF *Ara h 2* AND *Ara h 6* ALLERGEN GENES IN CULTIVATED GROUNDNUT (*Arachis hypogaea* L.)

Sampurna Bhattacharya^{1*}, Niravkumar Ashokbhai Delvadiya², Riddhi Rajyaguru²,
Rukam Singh Tomar², L. J. Raval¹ and G. U. Kulkarni¹

¹Department of Genetics and Plant Breeding, Junagadh Agricultural University, Junagadh - 362 001, Gujarat, India.

²Crop Improvement Unit, ICAR-Indian Institute of Groundnut Research, Junagadh - 362 015, Gujarat, India.

*Corresponding author E-mail : sampurna1298@gmail.com

(Date of Receiving-13-07-2026; Date of Revision-08-09-2026; Date of Acceptance-18-09-2026)

ABSTRACT

Groundnut (*Arachis hypogaea* L.) is an important food and oilseed crop, but its seeds contain several major allergens, among which *Ara h 2* and *Ara h 6* are prominent 2S albumin proteins associated with strong allergenic activity. Despite extensive biochemical and immunological characterization, the genome-level organization and sequence diversity of their allergen encoding genes in cultivated groundnut remain comparatively less explored. The present study aimed to identify and comprehensively characterize *Ara h 2* and *Ara h 6* genes in the chromosome-scale Tifrunner genome using an integrated *in-silico* approach. Reference-guided BLASTP analysis, HMM-based family screening, genome annotation, InterPro analysis, MCScanX based synteny assessment, gene-structure analysis, physicochemical characterization, MEME motif discovery, and MUSCLE multiple sequence alignment was employed. Five loci were identified, comprising two *Ara h 2* genes and three *Ara h 6* genes, localized to chromosomes 8 (A08) and 18 (B08). Syntenic analysis supported the homeologous relationship of the two *Ara h 2* copies and indicated a corresponding A08/B08 organization for *Ara h 6*, with an additional duplicated copy, *Ara h 6-3*, on B08. All five genes exhibited a single-exon organization and encoded compact proteins of 145-172 amino acids with conserved 2S albumin-related domain features. Motif analysis identified three broadly conserved motifs, while additional motifs were restricted to *Ara h 2* copies. Multiple sequence alignment revealed high within-group conservation, including 98.8% identity between *Ara h 2-1* and *Ara h 2-2* and 97.9-100% identity among the *Ara h 6* copies, together with 64.3-67.1% identity between the two allergen groups. Overall, the study establishes a genome-to-protein framework for the organization and sequence diversity of *Ara h 2* and *Ara h 6* in cultivated groundnut and provides a useful reference for future investigation of allergen isoforms and copy-specific variation.

Key words : *Arachis hypogaea*, Tifrunner, Groundnut allergy, *Ara h 2*, *Ara h 6*, 2S albumin, Allergen, Homeology, *In-silico* characterization, Conserved domain.

Introduction

Groundnut (*Arachis hypogaea* L.) is an economically important legume cultivated worldwide for its protein rich seeds and edible oil. Despite its nutritional and agricultural importance, it is also a major source of food allergy, causing reactions ranging from mild symptoms to severe systemic responses, including anaphylaxis. Its allergenic activity is associated with

diverse seed storage proteins belonging to several superfamilies, including 2S albumins, cupins and other allergen families. The WHO/IUIS Allergen Nomenclature provides a standardized framework for identifying and classifying these allergenic proteins and their variants. Among these, 2S albumins have received particular attention because of their abundance, structural stability, and strong IgE-reactive properties.

Among the major groundnut allergens, Ara h 2 and Ara h 6 are particularly important, belonging to the 2S albumin/conglutinin group of the prolamin superfamily and sharing substantial sequence and structural similarity. Together, they account for a large proportion of the effector activity of groundnut extracts in functional allergenicity assays, supporting their classification among the most potent allergens (Kulis *et al.*, 2012). Their clinical importance is further supported by component resolved diagnostic studies showing strong diagnostic value of IgE responses to Ara h 2 and Ara h 6 in groundnut allergy. Although Ara h 2 is generally considered the dominant allergen and may exhibit greater effector activity, Ara h 6 independently contributes to allergic responses and shows substantial immunological cross-reactivity with Ara h 2 (Klemans *et al.*, 2014; Hemmings *et al.*, 2020). Thus, both allergens need to be considered to fully represent the diversity of the 2S albumin allergen group in groundnut.

At the molecular level, Ara h 2 and Ara h 6 exhibit characteristic features of 2S albumin seed storage proteins. Their mature proteins contain multiple cysteine residues that form disulfide bonds, contributing to their compact and structurally constrained allergen cores. Structural studies have shown that their protease resistant regions adopt closely related helical folds, while the disulfide stabilized cores contribute to their structural persistence. Both allergens are relatively resistant to proteolytic digestion, and digestion resistant fragments can retain IgE-binding and allergenic activity (Lehmann *et al.*, 2006; Hazebrouck *et al.*, 2012). These properties may contribute to their persistence during gastrointestinal digestion and retention of allergen associated structural features under some processing conditions. However, their response to thermal processing depends on the processing conditions and surrounding food matrix rather than reflecting uniform resistance to heating (de Jongh *et al.*, 2020).

The genomic organization of cultivated groundnut provides an important framework for understanding copy level diversity. *Arachis hypogaea* is a segmental allotetraploid with A and B subgenomes derived from diploid ancestors related to *Arachis duranensis* and *Arachis ipaensis*. The chromosome scale of Tifrunner genome comprises 20 chromosomes, with Arah1.01-Arah1.10 belonging to the A subgenome and Arah1.11-Arah1.20 to the B subgenome (Gundaraniya *et al.*, 2023). Although the homeologous chromosomes retain substantial structural correspondence, chromosome specific rearrangements and other structural differences are also present (Bertioli *et al.*, 2019). Thus, closely related genes may occur as homeologous A/B

counterparts, while local duplications can further increase gene copy number. The increasing availability of genome sequences and other genomic resources has further enabled data mining and candidate gene identification in crop species (Jaiswal *et al.*, 2018; Satyavathi *et al.*, 2019). For allergen genes, distinguishing homeologs from tandem duplicates and other paralogs is therefore important, as sequence similarity alone cannot establish their genomic relationships.

Genome and transcriptome scale approaches have also been increasingly applied to resolve complex molecular processes underlying important traits in crop species (Hamid *et al.*, 2019). Recent transcriptome based studies in groundnut have demonstrated the utility of genome scale approaches for systematic gene discovery and molecular characterization (Rathod *et al.*, 2020). Although, Ara h 2 and Ara h 6 have been extensively studied through immunological, biochemical, and structural approaches, their genome level organization in cultivated groundnut remains less explored. Genome scale studies in groundnut associated systems have demonstrated the value of computational genome analysis for resolving gene content and functional characteristics (Iqbal *et al.*, 2017). An integrated analysis linking allergen gene identification with chromosomal distribution, homeology and duplication, gene structure, protein physicochemical properties, conserved domains and motifs, and sequence conservation can provide a more complete understanding of their genomic and molecular organization. Such an approach is particularly relevant in the allotetraploid groundnut genome, where homeologous and duplicated copies may contribute to allergen sequence diversity.

Therefore, the present study aimed to identify and comprehensively characterize Ara h 2 and Ara h 6 genes in the groundnut Tifrunner genome using an integrated *in-silico* approach. The analysis encompassed genomewide identification, chromosomal localization, homeology, duplication and synteny assessment, gene structure analysis, protein physicochemical and amino acid composition analysis, conserved domain and motif characterization, and multiple sequence alignment. By integrating these genomic and protein level analyses, the study sought to resolve the organization of Ara h 2 and Ara h 6 copies and identify their conserved and copy specific sequence features.

Materials and Methods

Genome and annotation resources

The chromosome scale genome assembly of cultivated groundnut (*Arachis hypogaea* L.) cv. Tifrunner was used as the genomic reference for the

present study. The Tifrunner genome comprises 20 chromosomes representing the A and B subgenomes of cultivated groundnut and provides a chromosome level framework for genome-wide analysis of allergen related genes. The genomic resources used in this study included the Tifrunner genome FASTA file, genome annotation file, coding sequence (CDS) FASTA file, and predicted protein FASTA file. These resources were used to examine the genomic location, gene organization, and predicted protein characteristics of *Ara h 2* and *Ara h 6* related sequences. Known *Ara h 2* and *Ara h 6* protein sequences were used as reference sequences for the identification of corresponding sequences within the Tifrunner predicted protein dataset. Reference allergen identities and accession information were cross checked against NCBI, UniProt, and the WHO/IUIS Allergen Nomenclature Database to ensure consistent identification of *Ara h 2* and *Ara h 6* related proteins.

Identification of *Ara h 2* and *Ara h 6* genes

A stepwise sequence based approach was employed to identify *Ara h 2* and *Ara h 6* related proteins from the Tifrunner predicted protein dataset. Known *Ara h 2* and *Ara h 6* protein sequences were used as reference sequences for similarity based searching against the Tifrunner protein FASTA file using NCBI BLASTP. The resulting matches were examined on the basis of sequence similarity and alignment characteristics to identify proteins showing substantial similarity to the reference allergens. BLASTP was used as the initial similarity searching step because it enables the identification of homologous proteins within a target sequence database based on local sequence similarity. The resulting candidate proteins were subsequently examined using profile hidden Markov model (HMM) based analysis with HMMER and the corresponding Pfam protein family model to support family level classification. In the present analysis, the Pfam 2S albumin related model PF00234 was used as an additional family level screening criterion. Candidate sequences were then evaluated using the combined evidence from BLASTP similarity, reference protein identity, genome annotation, and protein family/domain information. The selected protein sequences were traced back to their corresponding gene and transcript models in the Tifrunner genome annotation to establish their genomic identities and coordinates.

Chromosomal localization and genomic organization

The chromosomal positions of the identified *Ara h 2* and *Ara h 6* genes were obtained from the Tifrunner genome annotation by extracting chromosome number,

genomic coordinates and strand information for each locus. Based on chromosome assignment, genes located on chromosomes 1-10 were classified as belonging to the A subgenome, whereas genes located on chromosomes 11-20 were assigned to the B subgenome. Chromosomal distribution of the allergen genes was visualized using the Gene Location Visualize module implemented in TBtools-II (Chen *et al.*, 2023). Chromosome length information and gene coordinates were supplied as input to generate a chromosome scale map illustrating the physical distribution of *Ara h 2* and *Ara h 6* loci within the cultivated groundnut genome.

Homeology, duplication and synteny analysis

To investigate evolutionary relationships among the identified allergen genes, local collinearity and syntenic relationships were analyzed using the MCScanX package integrated within TBtools-II (Wang *et al.*, 2012). The genomic regions surrounding the *Ara h 2* and *Ara h 6* loci were examined to identify conserved neighboring genes and collinear blocks between the A and B subgenomes. Particular attention was given to chromosomes A08 and B08, where the identified allergen genes were located. Conserved anchor genes flanking the allergen loci were used to evaluate homeologous relationships between subgenomic regions. The presence of additional copies was further assessed to distinguish homeologous gene pairs from potential duplication derived paralogs. Syntenic relationships were visualized using TBtools-II (Chen *et al.*, 2023) and cross checked with R software (Chambers, 2008).

Gene structure analysis

The gene structures of the identified *Ara h 2* and *Ara h 6* loci were characterized using the transcript and coding sequence (CDS) information retrieved from the Tifrunner genome annotation. For each gene, the genomic coordinates of the corresponding mRNA and CDS features were examined to determine transcript length, CDS length, exon number, and the relative positions of the coding regions within the transcripts. The five identified genes were subsequently compared to assess similarities and differences in their transcript organization. For visualization, the relevant annotation features were converted into a GFF3 compatible format suitable for gene structure analysis. Gene structures were visualized using the Gene Structure Display Server (GSDS) (Hu *et al.*, 2015), with CDS regions represented separately from untranslated regions (UTRs). Because UTR features were not explicitly provided as independent features in the original Tifrunner annotation used in this study, the displayed 5' and 3' UTR regions were inferred from the

boundaries of the annotated mRNA and CDS features. Thus, the UTRs shown in the gene structure diagram represent coordinate derived regions rather than independently annotated UTR features.

Protein physicochemical characterization and amino-acid composition

The physicochemical properties of the five identified *Ara h 2* and *Ara h 6* proteins were characterized using the ExPASy ProtParam tool (Gasteiger *et al.*, 2003). The predicted protein sequences were submitted individually to ProtParam to calculate protein length, molecular weight, theoretical isoelectric point (pI), instability index, aliphatic index, and grand average of hydropathy (GRAVY). The percentage of cysteine residues was also recorded because cysteine mediated disulfide bonds are an important structural feature of 2S albumin proteins. ProtParam-derived values were used to compare the physicochemical characteristics of the identified *Ara h 2* and *Ara h 6* proteins and to assess similarities and sequence level differences among the copies. The amino acid composition of each protein sequence was determined from the corresponding amino acid counts and expressed as the percentage contribution of each of the 20 standard amino acids. The resulting composition matrix was organized and visualized as a heatmap using R version 4.3.2 with the ggplot2 (Wickham and Sievert, 2009) and tidyr (Boehmke, 2016) packages. The heatmap was generated to facilitate comparison of residue composition among the five proteins and to identify amino acids showing relatively conserved or variable representation across *Ara h 2* and *Ara h 6* protein sequences.

Conserved domain and protein family analysis

The domain architecture of the five identified *Ara h 2* and *Ara h 6* proteins was characterized using the InterPro database, which integrates protein signatures from multiple member databases to provide information on protein families, domains, conserved regions, and other sequence features (Blum *et al.*, 2021). Each protein sequence was analyzed individually, and the resulting annotations were examined with particular attention to the 2S seed-storage protein/conglutin family related domain signatures, and other predicted sequence regions relevant to the structural organization of the proteins. The amino acid coordinates of the identified domains and regions were recorded for each protein. Signal peptide subregions, including the N, H, and C regions, together with predicted non-cytoplasmic regions and coil regions, were also recorded where available. Overlapping InterPro signatures were retained when they represented distinct

database annotations of the same sequence region rather than being treated as independent, non-overlapping structural domains. The final domain architecture of each protein was reconstructed from the InterPro-derived coordinates and visualized using R (Chambers, 2008), allowing comparison of conserved and variable regions among the five *Ara h 2* and *Ara h 6* proteins.

Conserved motif analysis

Conserved sequence motifs among the five *Ara h 2* and *Ara h 6* proteins were identified using the MEME Suite (version 5.5.9) (Bailey *et al.*, 2015). The five protein sequences were analyzed using the classical MEME motif discovery algorithm with protein selected as the sequence type and the “any number of repetitions” (anr) model enabled. A maximum of 10 motifs was initially requested. The analysis returned five motif patterns before no additional starting points were identified. For each detected motif, the consensus sequence, motif width, occurrence frequency, amino acid positions and associated significance statistics were obtained from the MEME output. Motif occurrence coordinates were then mapped onto the corresponding protein sequences and visualized to compare motif organization among the *Ara h 2* and *Ara h 6* proteins. The resulting motif architecture was used to distinguish broadly conserved sequence patterns from motifs restricted to particular protein copies. Although five motifs were detected by MEME, the motifs were not considered equally informative for comparative interpretation. Motifs showing consistent occurrence across the five proteins and strong MEME significance were considered the principal conserved motifs, whereas motifs detected at only a small number of sites and showing weak overall significance were treated as minor or copy specific sequence features. This approach was used to avoid interpreting low confidence or sparsely distributed motifs as universally conserved characteristics of the *Ara h 2/Ara h 6* group.

Multiple sequence alignment and sequence conservation analysis

The amino acid sequences of the five identified *Ara h 2* and *Ara h 6* proteins were subjected to multiple sequence alignment using MUSCLE through the EMBL-EBI web interface (Edgar, 2004). The resulting alignment was subsequently examined and formatted using Jalview, which was used to organize the sequence labels, apply Clustal based residue colouring, and display sequence conservation and consensus information (Waterhouse *et al.*, 2009). The final alignment was examined to identify conserved and variable amino acid positions, sequence insertions and deletions and regions of greater sequence

divergence among the five proteins. The alignment was also used to evaluate the extent of sequence conservation between the homeologous and duplicated copies identified in the Tifrunner genome.

Results and Discussion

Genome-wide identification of *Ara h 2* and *Ara h 6* loci

The integrated sequence search and annotation based workflow identified five *Ara h 2* and *Ara h 6* related loci in the Tifrunner genome, comprising two *Ara h 2* genes and three *Ara h 6* genes. The *Ara h 2* loci were designated *Ara h 2-1* (LOC112707245) and *Ara h 2-2* (LOC112771110), whereas the three *Ara h 6* loci were designated *Ara h 6-1* (LOC112707246), *Ara h 6-2* (LOC112771111), and *Ara h 6-3* (LOC112771582). Their corresponding transcript and protein accessions were consistently linked to the respective gene models in the Tifrunner annotation, providing a defined set of five loci for subsequent genomic and protein level characterization. The identified genes were restricted to chromosomes 8 and 18. *Ara h 2-1* and *Ara h 6-1* were located on chromosome 8 at 35,990,900-35,991,697 bp and 36,038,508-36,039,244 bp, respectively. On chromosome 18, *Ara h 2-2* was located at 15,347,067-15,347,887 bp, while *Ara h 6-2* and *Ara h 6-3* occurred at 15,355,204-15,355,972 bp and 15,404,235-15,404,974 bp, respectively. All five retained genes were located on the positive strand and contained a single exon. Transcript lengths ranged from 737 to 821 bp, with CDS lengths of 438-519 bp, corresponding to predicted proteins of 145-172 amino acids (Fig. 1, Table 1).

The identification of two *Ara h 2* and three *Ara h 6* loci is consistent with previous reports of multiple conglutin copies in cultivated groundnut and supports the presence of duplicated allergen genes within the allotetraploid genome (Ramos *et al.*, 2006). Their localization to A08 and B08 is particularly relevant because these chromosomes represent corresponding A and B subgenome regions that retain substantial genomic collinearity. The presence of multiple closely related copies also provides a genomic basis for examining sequence conservation and copy specific variation, which is particularly relevant for *Ara h 2* and *Ara h 6* because naturally occurring sequence variants have been reported for both allergen groups (Bertioli *et al.*, 2019).

Chromosomal distribution of the identified genes

The five identified loci were confined to two chromosomes *i.e.*; *Ara h 2-1* and *Ara h 6-1* on chromosome 8 (A08) and *Ara h 2-2*, *Ara h 6-2* and *Ara h 6-3* on chromosome 18 (B08) (Fig. 1). Thus, both

Table 1 : Genomic, transcript, and protein characteristics of the identified *Ara h 2* and *Ara h 6* loci in the Tifrunner genome.

Allergen	Copy	Gene ID	Transcript Accession	Protein Accession	Annotation	Chr	Start	End	Strand	Gene Length (bp)	CDS Length (bp)	Protein Length	Exon Count
<i>Ara h 2</i>	Ah2-1	LOC112707245	NM_001376217.1	NP_001363146.1	Conglutin-7-like	8	35990900	35991697	+	798	483	160 aa	1
<i>Ara h 2</i>	Ah2-2	LOC112771110	NM_001405411.1	NP_001392340.1	Conglutin-7-like	18	15347067	15347887	+	821	519	172 aa	1
<i>Ara h 6</i>	Ah6-1	LOC112707246	XM_025758916.2	XP_025614701.1	Conglutin-like	8	36038508	36039244	+	737	438	145 aa	1
<i>Ara h 6</i>	Ah6-2	LOC112771111	NM_001376212.1	NP_001363141.1	Conglutin	18	15355204	15355972	+	769	438	145 aa	1
<i>Ara h 6</i>	Ah6-3	LOC112771582	XM_025816367.3	XP_025672152.1	Conglutin-like	18	15404235	15404974	+	740	438	145 aa	1

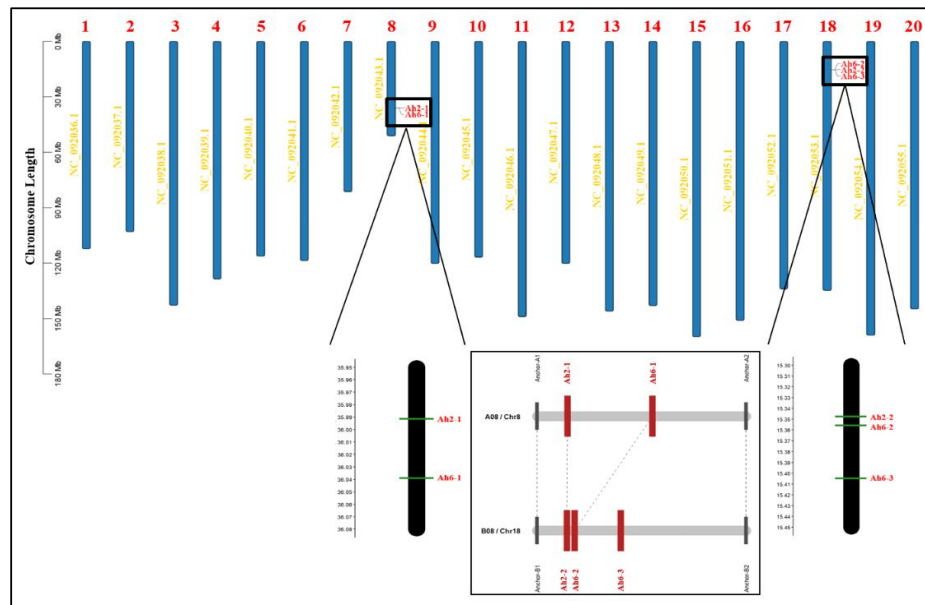


Fig. 1 : Chromosomal distribution & syntenic organization of *Ara h 2* and *Ara h 6* loci in the Tifrunner genome.

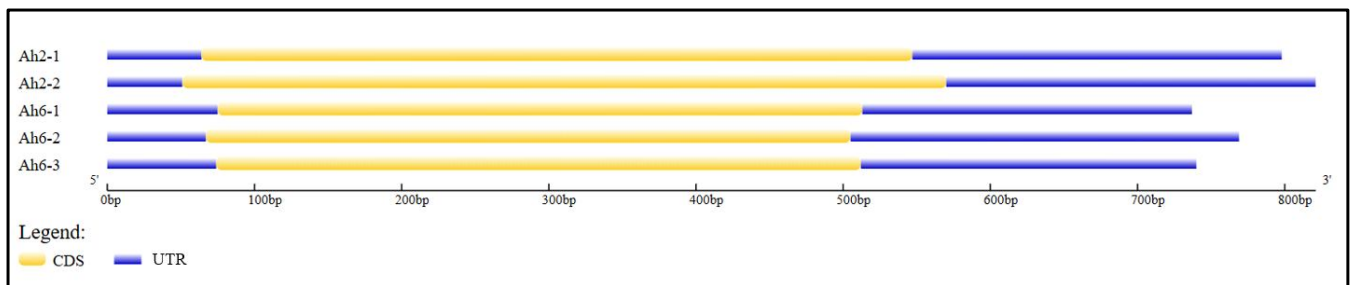


Fig. 2 : Gene structure of the identified *Ara h 2* and *Ara h 6* loci showing CDS and inferred UTR regions.

allergen groups were restricted to corresponding A and B sub genome regions rather than being dispersed across the genome. This distribution provided the basis for examining homeology and local duplication between the A08 and B08 regions.

The A08/B08 distribution is consistent with the allotetraploid organization of cultivated groundnut and its largely corresponding A and B sub genome chromosome structure (Bertioli *et al.*, 2019). The concentration of *Ara h 2* and *Ara h 6* loci in these corresponding regions also agrees with earlier genomic characterization of conglutinin genes in *Arachis* (Ramos *et al.*, 2006).

Syntenic organization and duplication of *Ara h 2* and *Ara h 6* loci

The A08 and B08 regions showed conserved local gene organization around the identified allergen loci (Fig. 1). *Ara h 2-1* and *Ara h 2-2* were located within corresponding syntenic blocks and were supported by conserved neighboring anchor genes, providing strong evidence for their homeologous relationship. In contrast, *Ara h 6-1* on A08 corresponded to the *Ara h 6* region on B08, where *Ara h 6-2* and *Ara h 6-3* occurred as closely related copies. The local organization therefore supports

Ara h 6-1/Ara h 6-2 as corresponding A/B copies, while *Ara h 6-3* represents an additional B08 duplicate.

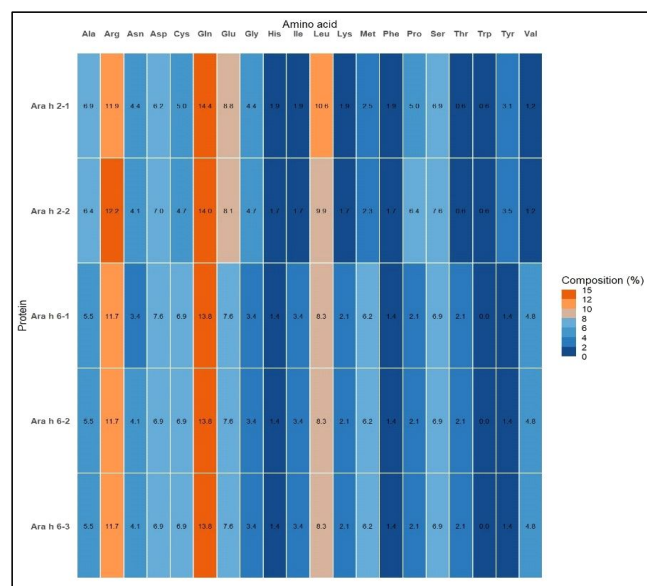
This organization closely agrees with the earlier genomic study of Ramos *et al.* (2006), which identified two *Ara h 2* homeologs and three *Ara h 6* copies, including one A genome and two B genome *Ara h 6* copies and proposed B genome duplication associated with tandem organization. The present synteny analysis therefore provides genome-level support for this organization in Tifrunner, while avoiding the assumption that all three *Ara h 6* copies are homeologs.

Gene structure of *Ara h 2* and *Ara h 6* copies

All five identified allergen genes exhibited a single exon organization, with no introns detected in the analyzed gene models (Fig. 2). *Ara h 2-1* had a gene span of 798 bp and a 483 bp CDS, whereas *Ara h 2-2* spanned 821 bp with a 519 bp CDS. The three *Ara h 6* genes showed gene spans of 737, 769, and 740 bp, respectively, with each containing a 438 bp CDS. The 52 and 32 UTR regions shown in Fig. 2 were inferred from the annotated mRNA and CDS boundaries rather than from independent UTR annotations.

Table 2 : Predicted physicochemical properties of the identified *Ara h 2* and *Ara h 6* proteins.

Protein	Length (aa)	MW (Da)	Theoretical pI	Instability index	Aliphatic index	Grand average of hydropathicity (GRAVY)	Cys (%)
Ara h 2-1	160	18757.85	5.96	72.41	59.25	-1.163	5.00%
Ara h 2-2	172	20114.26	5.96	76.37	55.12	-1.242	4.70%
Ara h 6-1	145	16921.19	5.73	62.31	65.24	-0.821	6.90%
Ara h 6-2	145	16920.2	6.13	60.47	65.24	-0.821	6.90%
Ara h 6-3	145	16920.2	6.13	60.47	65.24	-0.821	6.90%

**Fig. 3 :** Amino acid composition of the identified *Ara h 2* and *Ara h 6* proteins.

The conserved single exon organization across all five loci indicates that their structural differences are not associated with intron gain or loss. Instead, variation among the copies is more evident in their sequence and transcript/CDS lengths. The intronless organization is consistent with the earlier genomic characterization of *Ara h 2*, in which comparison of genomic and cDNA sequences also demonstrated the absence of an intron (Viquez *et al.*, 2001).

Protein physicochemical characteristics and amino acid composition

The predicted physicochemical characteristics of the five proteins are summarized in Table 2. Protein lengths ranged from 145 to 172 amino acids, with *Ara h 2-2* being the longest (172 aa), followed by *Ara h 2-1* (160 aa), while all three *Ara h 6* proteins were 145 aa. The predicted molecular masses ranged from 16.92 to 20.11 kDa, with *Ara h 2-2* showing the highest molecular weight (20.11 kDa). The theoretical pI values were relatively similar, ranging from 5.73 to 6.13, indicating mildly acidic to near neutral theoretical charge characteristics. All five proteins

showed negative GRAVY values, ranging from -1.242 to -0.821, indicating an overall hydrophilic tendency based on the Kyte-Doolittle scale. The *Ara h 6* proteins showed slightly less negative GRAVY values than the *Ara h 2* proteins. Cysteine content ranged from 4.7 to 6.9%, with all three *Ara h 6* proteins containing 6.9% cysteine. The ProtParam instability indices ranged from 60.47 to 76.37, exceeding the conventional value of 40 for all five proteins. However, this index is a computational prediction and should not be interpreted as evidence of experimental protein instability. The amino acid composition showed a broadly conserved pattern among the five proteins (Fig. 3). Glutamine was the most abundant residue in all sequences (13.8-14.4%), followed by arginine (11.7-12.2%), while several residues, including tryptophan, histidine, and isoleucine, occurred at relatively low frequencies. *Ara h 2-1* and *Ara h 2-2* showed closely related compositional profiles, whereas *Ara h 6-1*, *Ara h 6-2*, and *Ara h 6-3* displayed nearly identical compositions. The three *Ara h 6* proteins also shared the same length and highly similar physicochemical characteristics.

The predicted molecular characteristics are consistent with the compact nature of groundnut *Ara h 2* and *Ara h 6* as 2S albumin type seed storage allergens. Their relatively high cysteine content is particularly relevant because disulfide bonds contribute to the structural stability and protease resistance of these proteins. Experimental studies have demonstrated that the conserved disulfide network is important for maintaining the stable core of *Ara h 2* and *Ara h 6* (Lehmann *et al.*, 2006; Hazebrouck *et al.*, 2012). The highly similar composition and physicochemical properties of the three *Ara h 6* copies further support their close sequence relationship, while the modest differences between *Ara h 2* and *Ara h 6* are consistent with their divergence as distinct but structurally related allergen groups.

Conserved domain architecture of *Ara h 2* and *Ara h 6*

InterPro analysis confirmed that all five proteins were

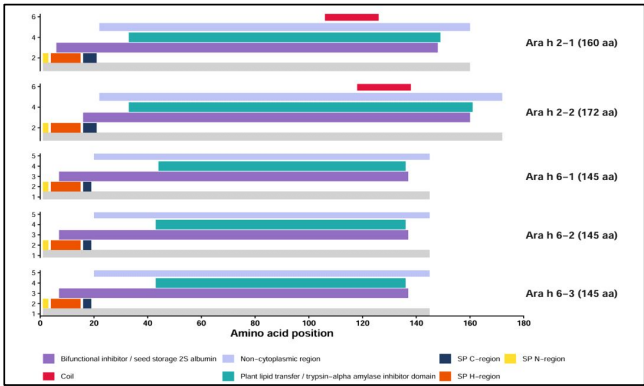


Fig. 4 : Conserved domain architecture and predicted sequence features of Ara h 2 and Ara h 6 proteins.

associated with the 2S seed storage protein/Conglutin family, supporting their classification as 2S albumin related proteins. Additional overlapping signatures corresponding to bifunctional inhibitor/seed storage 2S albumin and plant lipid transfer/trypsin α -amylase inhibitor related regions were detected across the sequences (Fig. 4). Because InterPro integrates signatures from multiple protein databases, these overlapping annotations were considered as complementary sequence features rather than independent biological domains (Blum *et al.*, 2021). The overall domain organization was conserved among the five proteins. N terminal signal peptide regions were detected at residues 1-21 in *Ara h 2* and 1-19 in *Ara h 6*, with the corresponding N, H, and C regions occupying the expected N terminal positions. The bifunctional inhibitor/2S albumin related signature extended across most of the mature proteins, while the lipid transfer/trypsin α -amylase inhibitor related signature occupied a partially overlapping region. A predicted non-cytoplasmic region covered most of each protein, whereas a coil feature was detected only in *Ara h 2*-1 and *Ara h 2*-2.

The conserved domain organization across all five sequences supports their shared classification as 2S albumin-type groundnut allergens. This is consistent with experimental studies showing that *Ara h 2* and *Ara h 6* possess highly similar 2S albumin folds stabilized by conserved disulfide bonds (Lehmann *et al.*, 2006). The additional InterPro features should therefore be interpreted primarily as overlapping sequence signatures within a conserved protein framework, rather than as evidence for functionally distinct domains.

Conserved motif organization

MEME analysis identified five motif patterns among the five proteins (Fig. 5, Table 3). Motifs 1-3 showed the most consistent distribution across the dataset. Motif 1 (EEQLDRLNLRRCCEQHAKQRJM) occurred 20 times, with four occurrences in each protein, while Motif 2

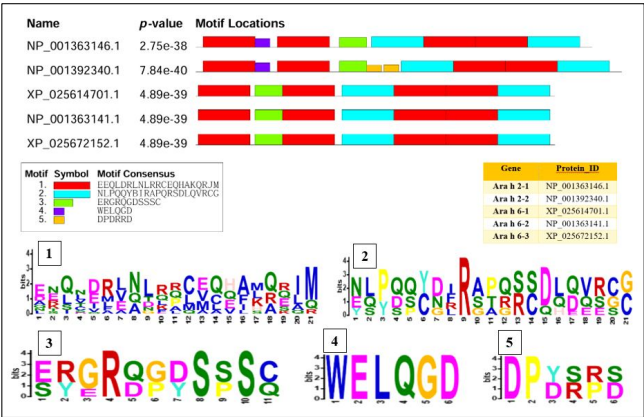


Fig. 5 : Conserved motif distribution and sequence logos of the identified Ara h 2 and Ara h 6 proteins.

(NLPQQYBIRAPQRSDLQVRCG) occurred 10 times, with two occurrences in each protein. Motif 3 (ERGRQGDSSSC) occurred once in each of the five proteins. Together, these three motifs represented the principal conserved motif pattern shared by both allergen groups. Motif 4 (WELQGD) was restricted to the two *Ara h 2* proteins, whereas Motif 5 (DPDRRD) occurred only in *Ara h 2*-2. Because of their restricted occurrence, these two motifs were retained in the visualization but were not considered broadly conserved motifs. Their distribution may therefore represent *Ara h 2* specific or copy specific sequence features.

The conserved occurrence of Motifs 1-3 across both allergen groups is consistent with the substantial sequence relationship between *Ara h 2* and *Ara h 6*, which share approximately 60% sequence identity and related structural characteristics (Mueller *et al.*, 2014; Hemmings *et al.*, 2020). The restricted occurrence of Motifs 4 and 5 further suggests that, despite their shared conserved framework, individual copies retain localized sequence differences.

Multiple sequence alignment reveals strong within-group conservation

The MUSCLE alignment revealed high sequence conservation within the *Ara h 2* and *Ara h 6* groups, while also showing clear sequence differences between the two allergen groups (Fig. 6). *Ara h 2*-1 and *Ara h 2*-2 shared approximately 98.8% sequence identity, with *Ara h 2*-2 containing an approximately 12 residue insertion that accounted for its longer 172 aa sequence. The three *Ara h 6* proteins showed even greater conservation. *Ara h 6*-2 and *Ara h 6*-3 were identical, whereas *Ara h 6*-1 shared approximately 97.9% sequence identity with the other two copies. Between the two allergen groups, pairwise identity ranged from approximately 64.3 to 67.1%, indicating substantial sequence relatedness despite

Table 3 : MEME-predicted conserved motif sequences, positions, and significance values of the identified *Ara h 2* and *Ara h 6* proteins.

Gene	Protein ID	Motif	Motif Sequence	Start aa	Endaa	p value
Ara h 2-1	NP_001363146.1	1	EEQLDRLNLRCEQHAKQRJM	4	24	9.32e-11
		4	WELQGD	25	30	7.17e-09
		1	EEQLDRLNLRCEQHAKQRJM	34	54	7.90e-14
		3	ERGRQGDSSSC	59	69	7.01e-10
		2	NLPQQYBIRAPQRSDDLQVRCG	72	92	7.95e-16
		1	EEQLDRLNLRCEQHAKQRJM	93	113	1.19e-14
		1	EEQLDRLNLRCEQHAKQRJM	114	134	1.80e-11
		2	NLPQQYBIRAPQRSDDLQVRCG	135	155	1.71e-18
Ara h 2-2	NP_001392340.1	1	EEQLDRLNLRCEQHAKQRJM	4	24	9.32e-11
		4	WELQGD	25	30	7.17e-09
		1	EEQLDRLNLRCEQHAKQRJM	34	54	7.90e-14
		3	ERGRQGDSSSC	59	69	1.06e-11
		5	DPDRRD	70	75	8.16e-07
		5	DPDRRD	77	82	3.21e-07
		2	NLPQQYBIRAPQRSDDLQVRCG	84	104	7.95e-16
		1	EEQLDRLNLRCEQHAKQRJM	105	125	1.19e-14
		1	EEQLDRLNLRCEQHAKQRJM	126	146	1.80e-11
		2	NLPQQYBIRAPQRSDDLQVRCG	147	167	4.80e-18
Ara h 6-1	XP_025614701.1	1	EEQLDRLNLRCEQHAKQRJM	2	22	1.19e-14
		3	ERGRQGDSSSC	25	35	2.15e-12
		1	EEQLDRLNLRCEQHAKQRJM	36	56	2.06e-19
		2	NLPQQYBIRAPQRSDDLQVRCG	60	80	6.48e-17
		1	EEQLDRLNLRCEQHAKQRJM	81	101	1.93e-12
		1	EEQLDRLNLRCEQHAKQRJM	102	122	6.37e-14
		2	NLPQQYBIRAPQRSDDLQVRCG	123	143	2.29e-21
Ara h 6-2	NP_001363141.1	1	EEQLDRLNLRCEQHAKQRJM	2	22	1.19e-14
		3	ERGRQGDSSSC	25	35	2.15e-12
		1	EEQLDRLNLRCEQHAKQRJM	36	56	2.06e-19
		2	NLPQQYBIRAPQRSDDLQVRCG	60	80	6.48e-17
		1	EEQLDRLNLRCEQHAKQRJM	81	101	5.89e-15
		1	EEQLDRLNLRCEQHAKQRJM	102	122	6.37e-14
		2	NLPQQYBIRAPQRSDDLQVRCG	123	143	2.29e-21
Ara h 6-3	XP_025672152.1	1	EEQLDRLNLRCEQHAKQRJM	2	22	1.19e-14
		3	ERGRQGDSSSC	25	35	2.15e-12
		1	EEQLDRLNLRCEQHAKQRJM	36	56	2.06e-19
		2	NLPQQYBIRAPQRSDDLQVRCG	60	80	6.48e-17
		1	EEQLDRLNLRCEQHAKQRJM	81	101	5.89e-15
		1	EEQLDRLNLRCEQHAKQRJM	102	122	6.37e-14
		2	NLPQQYBIRAPQRSDDLQVRCG	123	143	2.29e-21

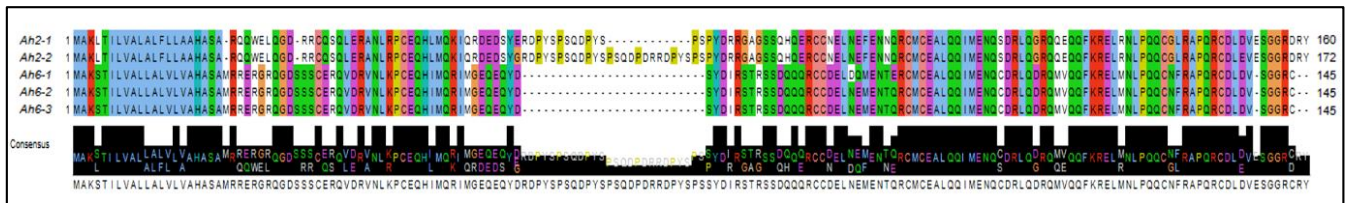


Fig. 6 : Multiple sequence alignment and conservation profile of the identified Ara h 2 and Ara h 6 proteins.

their distinction as separate allergen groups. Eight aligned positions were occupied by cysteine in all five sequences, highlighting the strong conservation of cysteine residues within the shared protein framework.

The high within group sequence identity supports the close relationship among the identified copies, particularly the near identical Ara h 6 sequences. The approximately 64-67% identity between Ara h 2 and Ara h 6 is consistent with previous studies reporting approximately 59-60% sequence identity between these two allergen groups (Hemmings *et al.*, 2020; Bublin and Breiteneder, 2014). The 12 residue insertion observed in Ara h 2-2 also agrees with experimentally characterized *Ara h 2* isoforms, in which a longer isoform contains an additional 12 amino acids (Chatel *et al.*, 2003; Li *et al.*, 2010). The conserved cysteine positions further support the shared 2S albumin/prolamin structural framework, in which cysteine residues contribute to disulfide-bond stabilization (Lehmann *et al.*, 2006; Bublin and Breiteneder, 2014).

Conclusion

The present study provides an integrated *in-silico* characterization of the *Ara h 2* and *Ara h 6* allergen gene in cultivated groundnut cv. Tifrunner, identifying two *Ara h 2* and three *Ara h 6* loci distributed across the corresponding A08 and B08 genomic regions. Synteny and duplication analyses supported the homeologous relationship of the *Ara h 2* copies and indicated a corresponding A/B organization for *Ara h 6*, with an additional duplicated B08 copy represented by *Ara h 6-3*. Despite differences in transcript and protein length, all five loci retained a conserved single exon organization and showed broadly similar 2S albumin related protein features, including conserved domain architecture, motif patterns, and cysteine-rich regions. Multiple sequence alignment further demonstrated strong within group conservation, particularly among the three *Ara h 6* copies, while retaining sequence differences between *Ara h 2* and *Ara h 6*. Overall, these findings establish a genome to protein framework for understanding the organization and sequence diversity of these two major groundnut allergens and provide a useful reference for future studies investigating allergen isoforms, copy specific variation, and their potential relevance to groundnut allergen

characterization.

Acknowledgement

This work forms a preliminary component of the Ph.D. research of the corresponding author. The corresponding author sincerely acknowledges the Knowledge Consortium of Gujarat (Education Department, Government of Gujarat) for providing financial support through the SHODH - ScHeme of Developing High Quality Research scholarship, which facilitated the pursuit of the Ph.D. degree programme. The author also gratefully acknowledges Junagadh Agricultural University, Junagadh, Gujarat, and ICAR-Indian Institute of Groundnut Research, Junagadh, Gujarat for their institutional support and the research environment provided during the course of the Ph.D. programme.

References

- Bailey, T.L., Johnson J., Grant C.E. and Noble W.S. (2015). The MEME suite. *Nucl. Acids Res.*, **43**(W1), W39-W49.
- Bertioli, D.J., Jenkins J., Clevenger J., Dudchenko O., Gao D., Seijo G., Leal Bertioli S.C.M., Ren L., Farmer A.D., Pandey M.K., Samoluk S.S., Abernathy B., Agarwal G., Ballén-Taborda C., Cameron C., Campbell J., Chavarro C., Chitikineni A., Chu Y., Dash S., El Baidouri M., Guo B., Huang W., Kim K.D., Korani W., Lanciano S., Lui C.G., Mirouze M., Moretzsohn M.C., Pham M., Shin J.H., Shirasawa K., Sinharoy S., Sreedasyam A., Weeks N.T., Zhang X., Zheng Z., Sun Z., Froenicke L., Aiden E.L., Micheltore R., Varshney R.K., Holbrook C.C., Cannon E.K.S., Scheffler B.E., Grimwood J., Ozias-Akins P., Cannon S.B., Jackson S.A. and Schmutz J. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nature Genetics*, **51**(5), 877-884.
- Blum, M., Chang H.Y., Chuguransky S., Grego T., Kandasamy S., Mitchell A., Nuka G., Paysan-Lafosse T., Qureshi M., Raj S., Richardson L., Salazar G.A., Williams L., Bork P., Bridge A., Gough J., Haft D.H., Letunic I., Marchler-Bauer A., Mi H., Natale D.A., Necci M., Orengo C.A., Pandurangan A.P., Rivoire C., Sigrist C.J.A., Sillitoe I., Thanki N., Thomas P.D., Tosatto S.C.E., Wu C.H., Bateman A. and Finn R.D. (2021). The InterPro protein families and domains database: 20 years on. *Nucl. Acids Res.*, **49**(D1), D344-D354.
- Boehmke, B.C. (2016). Reshaping your data with tidy. In : *Data wrangling with R* (pp. 211-218). Cham: Springer International Publishing.

- Bublin, M. and Breiteneder H. (2014). Cross-reactivity of peanut allergens. *Curr. Allergy and Asthma Rep.*, **14**(4), 426.
- Chambers, J. (2008). *Software for data analysis: programming with R*. New York, NY: Springer New York.
- Chatel, J.M., Bernard H. and Orson F.M. (2003). Isolation and characterization of two complete *Ara h 2* isoforms cDNA. *Int. Arch. Allergy Immunol.*, **131**(1), 14-18.
- Chen, C., Wu Y., Li J., Wang X., Zeng Z., Xu J., Liu Y., Feng J., Chen H., He Y. and Xia R. (2023). TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. *Mole. Plant*, **16**(11), 1733-1742.
- de Jongh, H.H., de Jong G.A., Apostolovic D., Taylor S.L., Baumert J.L. and Koppelman S.J. (2020). Effect of heat treatment on the conformational stability of intact and cleaved forms of the peanut allergen *Ara h 6* in relation to its IgE-binding potency. *Food Chem.*, **326**, 127027.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, **32**(5), 1792-1797.
- Gasteiger, E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D. and Bairoch A. (2003). ExpASY: the proteomics server for in-depth protein knowledge and analysis. *Nucl. Acids Res.*, **31**(13), 3784-3788.
- Gundaraniya, S.A., Ambalam P.S., Budhwar R., Padhiyar S.M. and Tomar R.S. (2023). Transcriptome analysis provides insights into the stress response in cultivated peanut (*Arachis hypogaea* L.) subjected to drought-stress. *Mole. Biol. Rep.*, **50**(8), 6691-6701.
- Hamid, R., Marashi H., Tomar R.S., Shafaroudi S.M. and Sabara P.H. (2019). Transcriptome analysis identified aberrant gene expression in pollen developmental pathways leading to CGMS in cotton (*Gossypium hirsutum* L.). *PloS One*, **14**(6), e0218381.
- Hazebrouck, S., Guillon B., Drumare M.F., Paty E., Wal J.M. and Bernard H. (2012). Trypsin resistance of the major peanut allergen *Ara h 6* and allergenicity of the digestion products are abolished after selective disruption of disulfide bonds. *Mole. Nutr. Food Res.*, **56**(4), 548-557.
- Hemmings, O., Du Toit G., Radulovic S., Lack G. and Santos A.F. (2020). *Ara h 2* is the dominant peanut allergen despite similarities with *Ara h 6*. *J. Allergy Clin. Immunol.*, **146**(3), 621-630.
- Hu, B., Jin J., Guo A.Y., Zhang H., Luo J. and Gao G. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*, **31**(8), 1296-1297.
- Iquebal, M.A., Tomar R.S., Parakhia M.V., Singla D., Jaiswal S., Rathod V.M., Padhiyar S.M., Kumar N., Rai A. and Kumar D. (2017). Draft whole genome sequence of groundnut stem rot fungus *Athelia rolfsii* revealing genetic architect of its pathogenicity and virulence. *Scientific Reports*, **7**(1), 5299.
- Jaiswal, S., Antala T.J., Mandavia M.K., Chopra M., Jasrotia R.S., Tomar R.S., Kheni J., Angadi U.B., Iquebal M.A., Golakia B.A., Rai A. and Kumar D. (2018). Transcriptomic signature of drought response in pearl millet (*Pennisetum glaucum* (L.) and development of web-genomic resources. *Scientific Reports*, **8**(1), 3382.
- Klemans, R.J.B., Knol E.F., Bruijnzeel Koomen C.A.F.M. and Knulst A.C. (2014). The diagnostic accuracy of specific IgE to *Ara h 6* in adults is as good as *Ara h 2*. *Allergy*, **69**(8), 1112-1114.
- Kulis, M., Chen X., Lew J., Wang Q., Patel O.P., Zhuang Y., Murray K.S., Duncan M.W., Porterfield H.S., Burks A.W. and Dreskin S.C. (2012). The 2S albumin allergens of *Arachis hypogaea*, *Ara h 2* and *Ara h 6*, are the major elicitors of anaphylaxis and can effectively desensitize peanut allergic mice. *Clin. Exp. Allergy*, **42**(2), 326-336.
- Lehmann, K., Schweimer K., Reese G., Randow S., Suhr M., Becker W.M., Vieths S. and Rösch P. (2006). Structure and stability of 2S albumin-type peanut allergens: implications for the severity of peanut allergic reactions. *Biochem. J.*, **395**(3), 463-472.
- Li, J., Shefcheck K., Callahan J. and Fenselau C. (2010). Primary sequence and site selective hydroxylation of prolines in isoforms of a major peanut allergen protein *Ara h 2*. *Protein Science*, **19**(1), 174-182.
- Mueller, G.A., Maleki S.J. and Pedersen L.C. (2014). The molecular basis of peanut allergy. *Curr. Allergy Asthma Rep.*, **14**(5), 429.
- Ramos, M.L., Fleming G., Chu Y., Akiyama Y., Gallo M. and Ozias-Akins P. (2006). Chromosomal and phylogenetic context for conglutin genes in *Arachis* based on genomic sequence. *Mole. Gen. Genom.*, **275**(6), 578-592.
- Rathod, V., Hamid R., Tomar R.S., Patel R., Padhiyar S., Kheni J., Thirumalaisamy P.P. and Munshi N.S. (2020). Comparative RNA-Seq profiling of a resistant and susceptible peanut (*Arachis hypogaea*) genotypes in response to leaf rust infection caused by *Puccinia arachidis*. *3 Biotech*, **10**(6), 284.
- Satyavathi, C.T., Solanki R.K., Kakani R.K., Bharadwaj C., Singhal T., Padaria J., Khandelwal V., Srivastava R., Tomar R.S. and Iqbal M.A. (2019). Genomics assisted breeding for abiotic stress tolerance in millets. In : *Genomics Assisted Breeding of Crops for Abiotic Stress Tolerance*, Vol. II (pp. 241-255). Cham: Springer International Publishing.
- Viquez, O.M., Summer C.G. and Dodo H.W. (2001). Isolation and molecular characterization of the first genomic clone of a major peanut allergen, *Ara h 2*. *J. Allergy Clin. Immunol.*, **107**(4), 713-717.
- Wang, Y., Tang H., DeBarry J.D., Tan X., Li J., Wang X., Lee T., Jin H., Marler B., Guo H., Kissinger J.C. and Paterson A.H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucl. Acids Res.*, **40**(7), e49-e49.
- Waterhouse, A.M., Procter J.B., Martin D.M., Clamp M. and Barton G.J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189-1191.
- Wickham, H. and Sievert C. (2009). *ggplot2: elegant graphics for data analysis* (Vol. 10, pp. 978-0). New York: springer.