Comparative and Evolutionary Analysis of Major Peanut Allergen Gene Families

Milind B. Ratnaparkhe[1,2,†], Tae-Ho Lee[1,†‡], Xu Tan[1], Xiyin Wang[1,3], Jingping Li[1], Changsoo Kim[1], Lisa K. Rainville[1], Cornelia Lemke[1], Rosana O. Compton[1], Jon Robertson[1], Maria Gallo[4], David J. Bertioli[5], and Andrew H. Paterson[1,*]

[1]Plant Genome Mapping Laboratory, University of Georgia

[2]Directorate of Soybean Research, Indian Council of Agriculture Research (ICAR), Indore, (M.P.), India

[3]Center for Genomics and Computational Biology, School of Life Sciences, School of Sciences, Hebei United University, Tangshan, Hebei, China

[4]Department of Molecular Biosciences and Bioengineering, University of Hawaii at Mānoa

[5]University of Brasília, Campus Universitário Darcy Ribeiro, DF, Brazil

*Corresponding author: E-mail: paterson@plantbio.uga.edu.

[†]These authors contributed equally to this work.

[‡]Present address: Genomics Division, National Academy of Agricultural Science, Rural Development Administration, Suwon, South Korea

Accepted: August 25, 2014

Abstract

Peanut (*Arachis hypogaea* L.) causes one of the most serious food allergies. Peanut seed proteins, *Arah1*, *Arah2*, and *Arah3*, are considered to be among the most important peanut allergens. To gain insights into genome organization and evolution of allergen-encoding genes, approximately 617 kb from the genome of cultivated peanut and 215 kb from a wild relative were sequenced including three *Arah1*, one *Arah2*, eight *Arah3*, and two *Arah6* gene family members. To assign polarity to differences between homoeologous regions in peanut, we used as outgroups the single orthologous regions in *Medicago*, *Lotus*, common bean, chickpea, and pigeonpea, which diverged from peanut about 50 Ma and have not undergone subsequent polyploidy. These regions were also compared with orthologs in many additional dicot plant species to help clarify the timing of evolutionary events. The lack of conservation of allergenic epitopes between species, and the fact that many different proteins can be allergenic, makes the identification of allergens across species by comparative studies difficult. The peanut allergen genes are interspersed with low-copy genes and transposable elements. Phylogenetic analyses revealed lineage-specific expansion and loss of low-copy genes between species and homoeologs. *Arah1* syntenic regions are conserved in soybean, pigeonpea, tomato, grape, *Lotus,* and *Arabidopsis,* whereas *Arah3* syntenic regions show genome rearrangements. We infer that tandem and segmental duplications led to the establishment of the *Arah3* gene family. Our analysis indicates differences in conserved motifs in allergen proteins and in the promoter regions of the allergen-encoding genes. Phylogenetic analysis and genomic organization studies provide new insights into the evolution of the major peanut allergen-encoding genes.

Key words: *Arachis hypogaea* L., allergens, gene synteny, genome organization, homologs, evolution.

Introduction

Cultivated peanut (*Arachis hypogaea* L.) is a valuable oilseed and food crop, grown on 25.5 million ha with a total global production of approximately 35.5 million tons (FAO 2009), and thus ranks among the top five oilseed crops in the world alongside soybean, cottonseed, rapeseed, and sunflower. However, allergy caused by ingesting peanut seeds is one of the most serious and life-threatening food sensitivities, particularly among children (Sampson 1996; Burks 2003). Additionally, peanut allergy is one of the most common and serious of the immediate hypersensitivity reactions to foods in terms of persistence and severity of reaction. Unlike the clinical symptoms of other food allergies, reactions to peanuts are rarely outgrown; therefore, most diagnosed children will have the disease for a lifetime.

To date, 11 peanut allergens (*Arah1–11*) have been identified (http://www.allergen.org/Allergen.aspx, last accessed

March 23, 2014). Among them, *Arah1, Arah2* and *Arah3*, which belong to the peanut seed storage protein classes conarachin, conglutin and arachin, respectively, are classified as the major peanut allergens which can be recognized by more than 50% of peanut-allergic patients (Burks et al. 1998; Rabjohn et al. 1999; Barre et al. 2005). Specifically, *Arah1* and *Arah2* are recognized by 70–90% of patients with peanut allergy (Burks et al. 1995; Stanley et al. 1997; Clarke et al. 1998), and *Arah3* is recognized by serum IgE from approximately 44–54% of different patient populations with a history of peanut sensitivity (Kleber-Janke et al. 1999; Rabjohn et al. 1999). Other minor allergens have been described which less frequently trigger the synthesis of appreciable amounts of specific IgE in sensitized individuals. Allergy to plant-derived foods is a highly complex disorder with clinical manifestations ranging from mild oral, gastrointestinal, and cutaneous symptoms to life-threatening systemic conditions. This heterogeneity in clinical manifestations has been attributed to different properties of allergenic molecules. Currently, the only effective treatment for food allergy is avoidance of the food. For peanut allergic individuals, total avoidance is difficult as peanuts are increasingly being used as an economical protein source in processed foods (Kang and Gallo 2007; Kang et al. 2007; Jin et al. 2009). Because of the significance of the allergic reaction and the widening use of peanuts as protein extenders in processed foods, there is increasing interest in defining the allergenic proteins and exploring ways to decrease risk to the peanut-sensitive individual. The majority of cases of fatal food-induced anaphylaxis involves ingestion of peanuts.

Cultivated peanut, an allotetraploid ($2n = 4 \times = 40$), probably originated from a single hybridization event between two wild diploids with A and B genomes and spontaneous chromosome duplication. Conservation of genome macrostructure (macrosynteny) has been reported between the respective subgenomes of peanut (Burow et al. 2001; Moretzsohn et al. 2009) and between peanut and other legumes including soybean (*Glycine max* [Gm]), *Medicago truncatula* (Mt), and *Lotus japonicus* (Lj) (Zhu et al. 2005; Hougaard et al. 2008; Bertioli et al. 2009). Genomic research in peanut benefits greatly from comparative approaches, leveraging knowledge of other plant genomes to contribute to a richer understanding of gene and genome functions and evolution. For example, Ratnaparkhe et al. (2011) provided new insights into the function and evolution of peanut nucleotide binding site–leucine-rich repeat (NBS-LRR) genes, utilizing gene synteny with related legumes and other model plants.

The objectives of this study were to isolate, sequence and determine the structure, organization and evolution of *Arah1*, *Arah2*, *Arah3* and *Arah6* allergen-encoding genes in peanut. We sequenced approximately 617 kb from cultivated peanut (cv. Florunner UF-439-16-1003-2) and approximately 215 kb from a wild species (*Arachis duranensis*; A genome) to gain insights into *Arah1*, *Arah2*, *Arah3*, and *Arah6* organization and evolution. To assign polarity to differences between homoeologous regions in peanut, we used as outgroups the single orthologous regions in *Medicago* and *Lotus*, which diverged from peanut approximately 50 Ma and have not undergone subsequent polyploidy. We also compared these regions with orthologs in soybean, poplar, *Vitis*, *Arabidopsis,* and tomato to help clarify the timing of evolutionary events. We analyzed the *Arah1*, *Arah2*, *Arah3,* and *Arah6* homologs in terms of genomic organization, gene duplication, selection pressure, and evolutionary relationships to homologous genes in other species.

Materials and Methods

Screen of Bacterial Artificial Chromosome Clones and Sequencing

Bacterial artificial chromosome (BAC) library of *A. duranensis* (A Genome) and peanut BAC library (Yuksel and Paterson 2005) were used for screening. Overlapping oligonucleotide (overgo) screens of BAC libraries were performed as described by Bowers et al. (2005). The BAC hit scores were converted to BAC addresses with an in-house script and analyzed with BACman software (bacman.sourceforge.net, last accessed April 17, 2014) to assign each BAC to specific overgos. Subclones from each BAC clone were picked using a QBOT (Genetix, New Milton, UK). BAC clone sequencing was performed by shotgun fragmentation and Sanger and/or 454 sequencing. Sanger sequencing was done by ABI3730xl in the Plant Genome Mapping Lab, whereas sequencing by 454 was performed by GATC Biotech AG (Konstanz, Germany). Purified DNA was sheared to fragments of about 700 bp and ligated to Roche A- and B-adapters. The samples were separated by electrophoresis through a 2% agarose gel with TAE buffer. DNA between 700 and 900 bp was excised and column purified. The resulting DNA libraries were immobilized onto DNA capture beads and amplified using emulsion polymerase chain reaction according to the manufacturer's recommendations. The emulsion was then chemically broken and the beads carrying the amplified DNA library were recovered and washed by filtration. Samples were sequenced on a Genome Sequencing FLX Pico-Titer plate device with GS FLX Titanium XLR70 chemistry. Sequence data were produced in Standard Flowgram Format for each read with basecalls and per-base quality scores.

Assembling BAC Sequences from Reads

Read sequences were called from the chromatogram files generated by the ABI3730xl using Phred (Ewing and Green 1998; Ewing et al. 1998). Vector and primer sequence were screened out by Cross_match software included in Phrap assembly software package (http://www.phrap.org/phred-phrapconsed.html, last accessed January 20, 2014). The contigs were assembled by Phrap assembler with default parameters. Besides Phrap, we assembled the reads by Arachne

(Jaffe et al. 2003) and CAP3 (Huang and Madan 1999) in order to help the determination of direction and order of contigs. The direction and order of contigs in a BAC were determined by Consed (Gordon 2003) based on the paired-end information of sequenced clones as well as comparing with results using other assemblers. Finally, the contigs were concatenated based on the determined direction and order to one supercontig with 100 Ns at each junction. Assembly of 454 sequence reads was performed by the GS FLX System Software GS De Novo Assembler (Newbler) using default parameters.

Gene Annotation

Genes in the BAC sequences were predicted by FGENESH (Salamov and Solovyev 2000) at http://www.softberry.com (last accessed January 25, 2014) with *Medicago* gene model and default parameters. To determine putative gene function, predicted protein sequences translated from predicted genes were BLASTed by BLASTp against the GenBank nr protein database with 1e-10 *E* value cutoff. When no significant similarity was found for a predicted protein, we annotated the protein as "Unknown function." In addition, to see the relationships of the predicted genes with those in other species, sequences were used as queries in BLAST similarity searches against databases of expressed sequence tags (ESTs) from nine Fabaceae species including *A. hypogaea* (peanut) (Wu et al. 2013), *Cicer arietinum* (chickpea), *Glycine max* (soybean), *Lotus japonicus*, *Medicago* spp., *Phaseolus vulgaris* (common bean), *Cajanus cajan* (pigeonpea), *Pisum sativum* (pea), and *Vigna unguiculata* (cowpea) (downloaded from the National Center for Biotechnology Information's dbEST).

Determining and Drawing Multiple Alignment of Colinear Blocks

The detection of blocks of colinear genes harbored on peanut BACs was done using MCscanX (Wang et al. 2012). The genome sequences of plants used for the detection were downloaded from various resources: *Medicago*: *M. truncatula* sequencing resources database (http://www.medicago.org/, last accessed January 26, 2014), *Arabidopsis*: TAIR (ftp://ftp.arabidopsis.org/, last accessed March 15, 2014), *Vitis*: Genoscope (http://www.genoscope.cns.fr/, last accessed March 17, 2014), and Phytozome (www.phytozome.net, last accessed March 25, 2014). The default number of homologous genes required to call a colinear block in a genome-wide search is normally 5. However, as our search was focused on small regions of a genome sampled by BACs, with less risk of false positive associations, we used three, considering colinear blocks at lower levels of synteny. Homologous genes were determined by BLASTp with an *E* value cutoff of 1e-5. Given the colinear blocks detected by MCscanX, multiple alignment figures were drawn by in-house programs using genePlotR (Guy et al. 2010).

Identification and Classification of Long Terminal Repeat-Retrotransposons

A combination of structural analyses and sequence homology comparisons was used to identify retrotransposons. Intact long terminal repeat (LTR) elements were identified by using LTR_STRUC, an LTR-retrotransposon mining program (McCarthy and McDonald 2003), and by homology using methods previously described (Ma and Bennetzen 2004; Ma et al. 2004). For all intact retroelements with two LTRs, the LTR sequences were aligned by Clustalw (Chenna et al. 2003) with default parameters. Pairwise sequence divergence was calculated using the Kimura 2 parameter model using MEGA 4.0 (Tamura et al. 2007). The time of insertion was calculated using the synonymous substitution rate per site per year of 1.3×10^{-8} (Ma et al. 2004) and the equation $T = D/2t$ where T is the time of insertion; D, divergence; and *t*, mutation rate per nucleotide site per year.

Phylogenetic Analysis of Allergen-Encoding Genes

Both peptide and coding sequences (CDS) alignments were used to construct phylogenetic trees of the gene regions encoding conserved peptide motifs for *Arah1*, *Arah2*, *Arah3,* and *Arah6* genes. Various approaches were tested, and the results obtained using maximum-likelihood methods implemented in PHYML were used for the definitive analysis (Guindon et al. 2005). Bootstrapping tests of trees were performed using 100 sampling repetitions. The constructed trees were compared with one another and the best supported trees were used for interpretation and as input to run PAML (Yang 1997), software to estimate selection pressure along tree branches. For the identification of "positive" (diversifying) selection, we used model M8 with two Bayesian approaches (Naive Empirical Bayesian analysis and Bayes Empirical Bayesian or BEB) (Yang et al. 2005). We also estimated the evolutionary distance between the genes in each family with the Nei–Gojobori model (Nei and Gojobori 1986).

Gene Conversion Inference

Aligned allergen gene CDS were used for analysis of gene conversion using the GENECONV software as described in Sawyer (1989).

Promoter Analysis

The regions approximately 2 kb upstream of the allergen-encoding genes start codons were analyzed for promoter motifs. The source of sequences was Phytozome v7.0, except for *Lotus japonicus*, which was from PlantGDB (http://www.plantgdb.org/LjGDB/, last accessed January, 27 2014). Due to the sequence divergence of the 5′-regions, conserved motifs were manually inspected and compared with motifs detected in previously published studies (Viquez et al. 2001, 2003, 2004; Fu et al. 2010).

Results

Sequencing, Annotation, and Analysis of the Allergen-Encoding Gene Sequences

A total of nine BACs were sequenced, seven from cultivated peanut (cv. Florunner) and two from the diploid wild species *A. duranensis*, the probable A genome donor of cultivated peanut. These BACs were selected based on strong hybridization to overgo probes for three major peanut allergen-encoding genes *Arah1*, *Arah2*, and *Arah3* as described in supplementary table S1, Supplementary Material online. Sequencing of these nine BACs yielded 833,098 bp with an average GC content of 0.34. Seven BACs from cultivated peanut harbor 617,208 bp of cloned DNA, whereas the two BACs from the diploid species harbor 215,890 bp of cloned DNA. Assembly results of the BAC sequences are shown in supplementary table S1, Supplementary Material online. Genes predicted by FGENESH were supported by varying amount of secondary evidences. We sought supporting evidence for the predicted genes from searches against the Pfam databases, predicted proteins, and ESTs. Some predicted genes were well supported by peanut ESTs and therefore could be assigned putative functions. Other genes had none of this supporting evidence and were therefore annotated as putative proteins. Predicted gene sequences were blasted against ESTs from nine Fabaceae species. EST sequence comparisons of *Arah1*, *Arah2*, *Arah3*, *Arah6,* and low copy genes with other legumes are given in supplementary table S2, Supplementary Material online. Among the 14 allergen-encoding genes, three are *Arah1*, one *Arah2*, eight *Arah3*, and two *Arah6*. Sequences of *Arah1*, *Arah2*, *Arah3,* and *Arah6* genes and proteins are provided in supplementary tables S3 and S4, Supplementary Material online, respectively.

Identification and Characterization of Peanut Allergen-Encoding Genes

Location of *Arah1*, *Arah2*, *Arah3,* and *Arah6* allergen-encoding genes on peanut BACs is shown in figure 1. Three BAC clones were obtained containing the peanut allergen-encoding gene *Arah1*. Two BACs are from the *A. hypogaea* library and one from the *A. duranensis* library. One of the sequences from the *A. hypogaea* library was almost identical to the *A. duranensis* clone and thus A and B genome representatives from the *A. hypogaea* library could be assigned. In our study, two *Arah1* genes (AHF85I19.07 and AHF417E07.26) were identified from the cultivated peanut whereas one (ADH35P21.18) was identified from A genome wild species. BLASTP analysis of AHF417E07.26 shows 100% protein identity with *Arah1_P41b* (GenBank: AAB00861). *Arah1* AHF85I19.07 is truncated protein with several amino acids deleted from its N and C terminal ends (supplementary fig. S1, Supplementary Material online). Protein sequence similarities of *Arah1* homologs are given in supplementary table S5, Supplementary Material online. Protein sequence similarities of peanut *Arah1* with homologs from other plant species range between 27.27% and 76.92%. Figure 2 shows phylogenetic tree of *Arah1* homologs. The *Arah1* phylogenetic tree consists of three major clades (Clades 1–3) with all peanut *Arah1* grouped in clade 2. There are two subgroups in clade 2 representing A and B subgenomes. Phylogenetic analysis indicates that ADH 35P21.18 (A Genome) is orthologous to AHF85I19.07 derived from cultivated peanut (clade 2, fig. 2). Clade 1 and clade 3 contain *Arah1* homologs from pigeonpea, soybean, *Medicago*, *Arabidopsis*, *Lotus*, poplar, and tomato.

A single BAC contained one *Arah2* gene (AHF221B03.31) and two *Arah6* genes (AHF221B03.3 and AHF221B03.29) as shown in figure 1. The *Arah2* peanut allergen is recognized by serum IgE from 90% of peanut-allergic patients, thus establishing the importance of this protein in the etiology of the disease. BLASTP analysis of *Arah2* protein (AHF221B03.31) indicated that it has 100% identity (*E* value 0.0) with previously identified *Arah2.02* sequence (GenBank accession number: ABQ96215). *Arah2* has two isoforms, *Arah2.01* and *Arah2.02,* which has a duplication of 12 amino acids containing an extra copy of DPYSPS (supplementary fig. S2, Supplementary Material online). The *Arah2* gene (AHF221B03.31) belongs to the B-subgenome (putative donor *Arachis ipaensis* $2n = 2 \times = 20$, BB). Supplementary table S6, Supplementary Material online, shows protein sequence similarities of *Arah2* homologs from peanut, soybean, *Medicago,* and *Arabidopsis*. Protein sequence similarities of peanut *Arah2* with soybean and *Medicago* homologs range between 32.86% and 41.60%, whereas similarities with *Arabidopsis* homologs range between 25.95% and 34.78%. Figure 3 shows the phylogenetic tree of peanut allergens *Arah2* homologs. Peanut *Arah2* homologs are grouped in clade 1, whereas clade 2 has *Arah2* homologs of soybean, *Medicago* and *Arabidopsis*.

We identified two *Arah6* genes on BAC AHF221B03 (fig. 1). *Arah6* is a minor allergen and is recognized by about 40–50% of patient population. BLASTP analysis of *Arah6* protein AHF221B03.3 showed 100% sequence identity with previously identified *Arah6* (Swiss-Prot: Q647G9.1), whereas AHF221B03.29 has 26 amino acid insertion at the C-terminal end (supplementary fig. S3, Supplementary Material online). Figure 4 shows the phylogenetic tree of peanut *Arah6* allergens. All the peanut *Arah6* homologs are grouped in clade 1, whereas clade 2 has *Arah6* homologs of *Medicago* and soybean. Protein sequence similarities of *Arah6* homologs are given in supplementary table S7, Supplementary Material online. Protein sequence similarities of peanut *Arah6* with soybean homologs range between 23.47% and 33.60%, whereas similarities with *Medicago* homologs range between 26.73% and 45.71%.

We identified eight *Arah3* allergen-encoding genes distributed on four BACs (AHF44O16, AHF259D10, AHF74J10, and

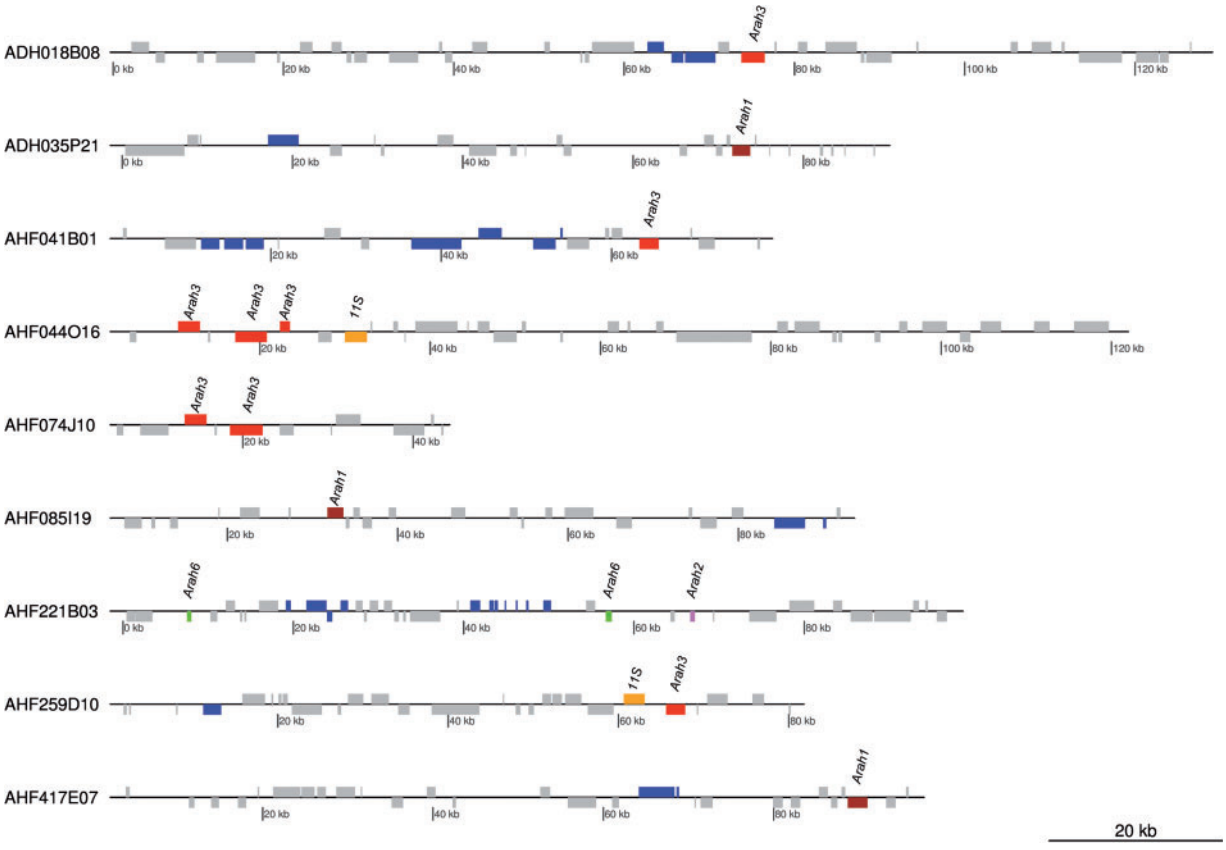

Fig. 1.—Position of genes on sequenced peanut BACs. Allergen-encoding genes are indicated in red, maroon, pink, and green boxes: *Arah1* (AHF85I19, AHF417E07, ADH35P21), *Arah2* (AHF221B03), *Arah3* (AHF74J10, AHF44O16, AHF41B01, ADH18B08, AHF259D10), and *Arah6* (AHF221B03). Annotation of these BACs revealed repetitive sequences showing significant similarities to retroviruses, peanut allergen-encoding genes (three *Arah1*, AHF85I19.07, AHF417E07.26, ADH35P21.18, one *Arah2*, AHF221B03.31, eight *Arah3*, AHF74J10.3, AHF44O16.2, AHF41B01.16, ADH18B08.22, AHF259D10.23, AHF44O16.05, AHF44O16.4, 74J10.5, two *Arah6*, AHF221B03.3 and AHF221B03.29), two 11S arachin storage protein genes (AHF44016.07 and AHF259d10.22), and other low copy genes. Box color indicates genes and retroelements (red, maroon, pink, and green: allergens, blue: retroelements, yellow: 11S seed storage proteins, gray boxes and lines: low copy genes). Numbers refers to the position of kilobasepairs (kb) on the BAC sequence.

AHF41B01) from cultivated peanut and one BAC (ADH 18B08) from the diploid A genome (fig. 1). *Arah3* is a legumin-like seed storage protein that has high sequence similarity to glycinin, the major 11S globulin seed storage protein family in soybean. Most BACs include only one or two allergen-encoding genes; however, BAC AHF44O16 has three genes (AHF44O16.02, AHF44O16.04, AHF44O16.05) clustered within 30 kb (fig. 1). The presence of two cupin domains and their respective motifs indicates that these proteins are members of the cupin superfamily. Supplementary figure S4, Supplementary Material online, shows multiple sequence alignment of *Arah3* proteins from peanut BACs and their comparison with previously identified *Arah3* proteins. AHF44O16.5 is a truncated protein with 345 amino acids and has only one cupin domain. Present on the same BAC are two full length proteins (AHF44O16.2 and AHF 44O16.4) with 528 and 382 amino acids. BLASTP analysis of AHF44O16.04 showed only 87% identity with the arachin *Ahy-3* (UniProtKB/Swiss-Prot: Q647H2.1). These results indicate that AHF44O16.04 is potentially new *Arah3* protein. We identified BAC AHF259D10 which is in a region homoeologous to BAC AHF44O16. Interestingly BAC AHF259D10 has only one allergen-encoding gene (fig. 1), suggesting that the three *Arah3* genes on BAC AHF44O16 are due to single-gene duplications after the divergence of A and B genomes. We also sequenced another BAC AHF74J10 that has two allergen-encoding genes (74J10.03 and 74J10.05). AHF74J10.3 has 528 amino acids and showed 99% protein sequence similarity with *Gly1* (GenBank accession number: AAG01363.1), with five amino acid differences at the N-terminal end. Protein sequence similarity between the peanut *Arah3* homologs ranges between 53.12% and 100% (supplementary table S8, Supplementary Material online). Figure 5 shows the phylogenetic tree of peanut allergen *Arah3* including protein sequences present on sequenced BACs and previously identified proteins. Phylogenetic tree of *Arah3* consists of two major clades (fig. 5). The first clade contains AHF74J10.3, AHF44O16.2, AHF41B01.16, and ADH18B08.22 along with

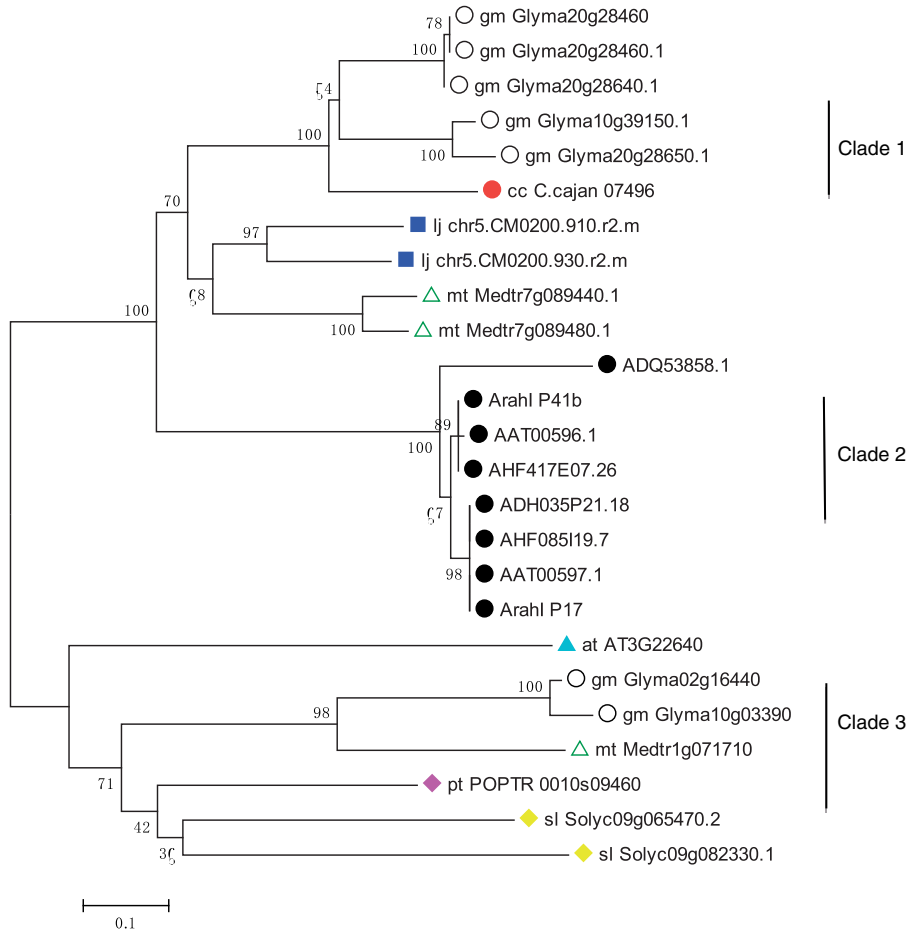

Fig. 2.—Phylogenetic tree of *Arah1* peanut allergens, including protein sequences from sequenced BACs, previously identified homologs from peanut and other species. Peanut allergens are shown in black circles (clade 2). Other species are indicated as soybean (GM) white circles, pigeonpea (CC) red circles, *Medicago* (MT) green triangle, *Arabidopsis* (AT) aqua triangle, *Lotus japonicus* (Lj) blue box and poplar (PT) pink box. Two *Arah1* genes (AHF85I19.07 and AHF417E07.26) were identified from the cultivated peanut whereas one (ADH35P21.18) was identified from an A genome wild species. Phylogenetic analysis indicates that ADH35P21.18 is orthologous to AHF85I19.07. *Arah1_P17* and *Arah1_P41b* both represent *Arah1* allergen and were previously derived from peanut A and B subgenomes, respectively. Genebank accession numbers: *Arah1_P71* (AAA60336.1)/UniProtKB/Swiss-Prot: P43237.1 and *Arah1_P41B* (AAB00861.1/Swiss-Prot: P43238.1).

the other closely related *Arah3* protein sequences from peanut. Clade 1 also contains AHF74J10.5 and AHF44O16.4 and groups with previously identified *Arah3* homologs. Clade 2 contains soybean, pigeonpea, *Lotus, Medicago,* and *Arabidopsis* homologs of *Arah3*.

Phylogenetic Analysis of *Arah1* and *Arah3* Homologs from Soybean, *Medicago, Lotus*, and *Arabidopsis*

Globulins (7S and 11S) comprise the majority of the total protein in many seeds that are consumed by humans. Vicilins (7S globulins) and legumins (11S globulins) share similar folds and belong to the cupin superfamily of proteins. The cupin domain is widely distributed in plant proteins and many other seed storage legumins have been characterized as potent allergens. To identify the homologs of *Arah1* and *Arah3* in model plants, we performed BLASTP searches using the databases of the predicted proteins from soybean, *Medicago*, *Lotus,* and *Arabidopsis*. Using *Arah1*, *Arah3* and related proteins for a BLASTP search, we identified homologous sequences from soybean, *Medicago*, *Lotus,* and *Arabidopsis*. To investigate the molecular evolution of *Arah1, Arah3* and their homologs in soybean, *Lotus*, *Medicago* and *Arabidopsis*, we constructed a phylogenetic tree using the amino acid sequences (fig. 6). Phylogenetic relationships among the *Arah1* and *Arah3* proteins were consistent with the expected relationships between different species. Four subgroups are evident in the phylogenetic tree (fig. 6). *Arah1*- and *Arah3*-related proteins were clustered in different subgroups (subgroups 4 and 1, respectively), whereas the other seed storage proteins formed two subgroups (subgroups 2 and 3). Subgroup 4 contains protein sequences of the β-conglycinin family. Peanut *Arah1* proteins are

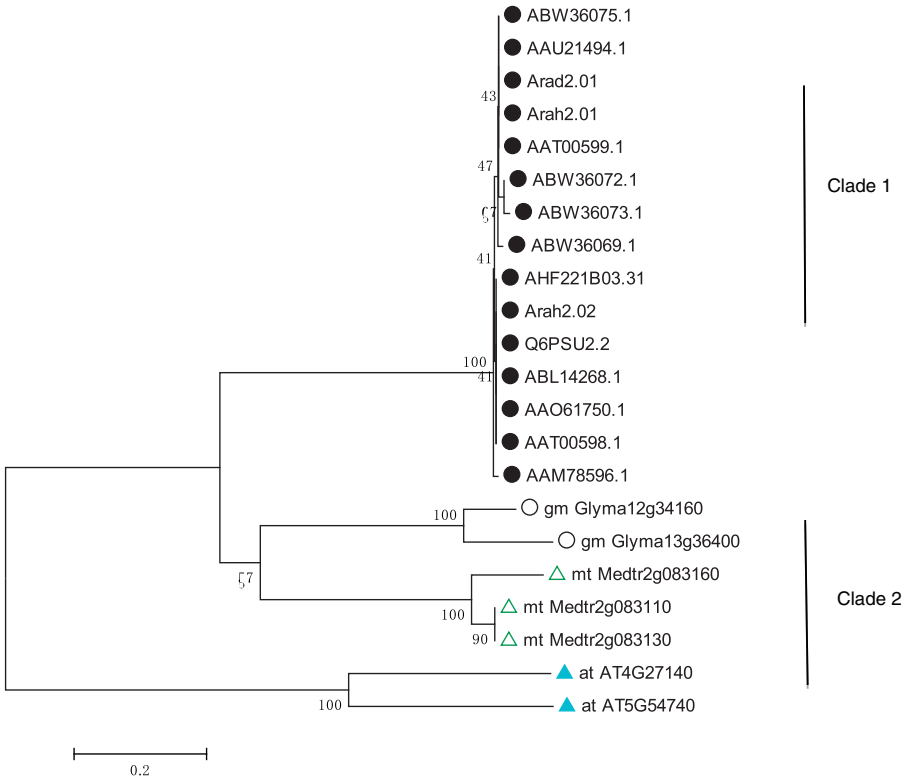

Fig. 3.—Phylogenetic tree of *Arah2* homologs. Peanut homologs are shown in black circles (clade 1). Previously identified *Arah2* protein sequences are indicated by the Gene Bank accession numbers. In the phylogenetic tree, *Arah2* derived from A and B genome (*Arah2.01* and *Arah2.02*) are grouped in two different subgroups (clade 1). Gene Bank accession number of *Arah2* homologs are: *Arah2.01* (ACN62248.1), *Arah2.02* (AY158467.1), *Ara d 2.01* (*Arachis duranensis*: ABQ96212.1). Other species are indicated as soybean (GM) white circles, *Medicago* (MT) green triangles, and *Arabidopsis* (AT) as aqua triangles.

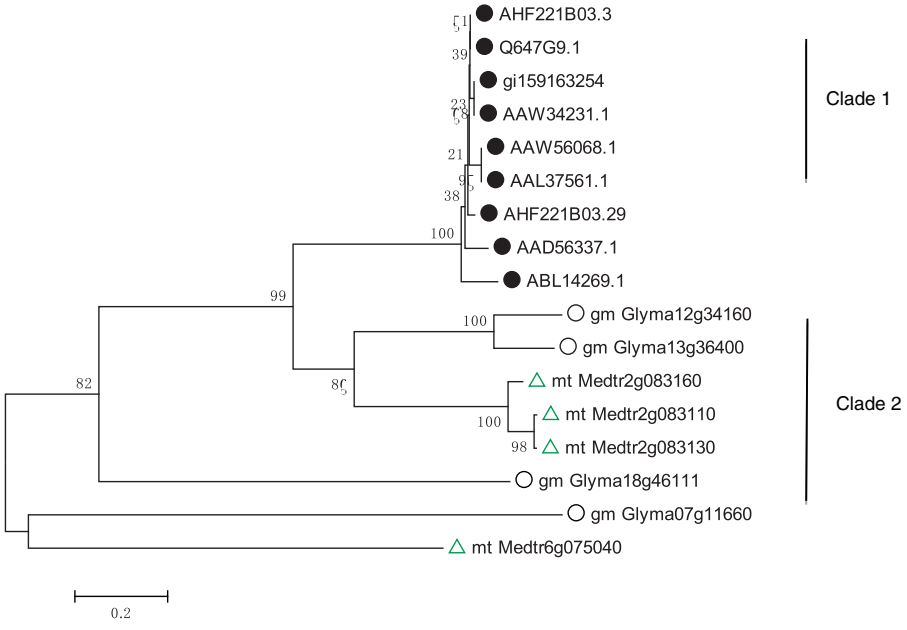

Fig. 4.—Phylogenetic tree of *Arah6* homologs. Two *Arah6* proteins from the sequenced peanut BAC are grouped together in the phylogenetic tree in clade 1 along with previously identified *Arah6* (Genebank accession number: AAL37561.1 and AAW34231.1). Two *Arah6* homologs (AHF221B03.3 and AHF221B03.3.29) are present on the same BAC AHF221B03. Other species are indicated as soybean (GM) white circles and *Medicago* (MT) green triangles.

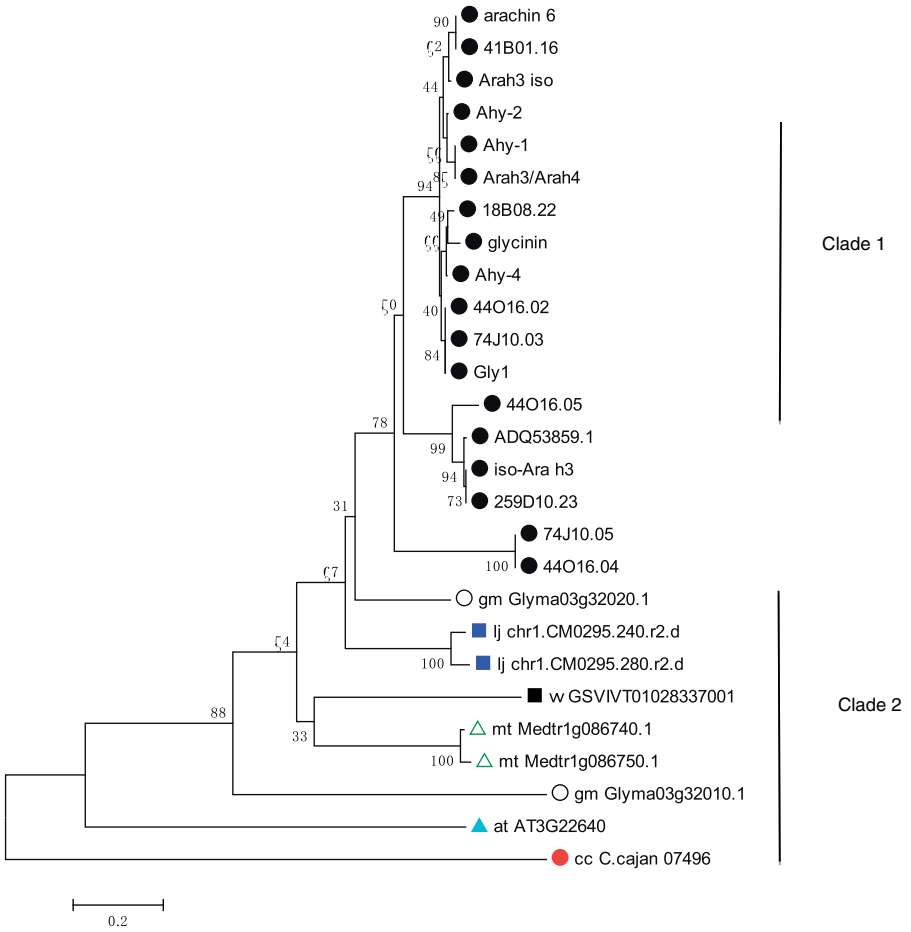

Fig. 5.—Phylogenetic tree of *Arah3 homologs.* All the peanut *Arah3* homologs (shown in black circles) are grouped in clade 1, whereas clade 2 contains *Arah3* homologs from other species. Other species are indicated as soybean (GM), pigeonpea (CC), *Medicago* (MT), *Arabidopsis* (AT), Vitis vinifera (Vv), and *Lotus japonicus* (Lj). Eight peanut *Arah3* allergen-encoding genes were distributed on four BACs (AHF44O16, AHF259D10, AHF74J10, and AHF41B01) from cultivated peanut and one BAC (ADH18B08) from the diploid A genome. Genebank accession numbers of peanut *Arah3* homologs are: *Gly1*: AAG01363.1, *Ahy1*: AAU21490.1, *Ahy2*: AAU21491.1, *Ahy-4*: AAW56067.1, *Arah3 iso*: ACH91862.1, *Arah3/arah4*: AAM46958.1, *arachin6*: ABL14270.1, *Iso-arah3*: ABI17154.1, *Arah3*: ADQ53859.1, *Glycinin*: AAC63045.1.

grouped with the α, α′ and *β*-subunit of β-conglycinin protein family in soybean (fig. 6).

Arah3 Homologs Are Absent from *Prunus persica, Cucumis sativus*, and *Medicago*

To find homologs of *Arah1* and *Arah3* in the other sequenced genomes, the *Arah1* and *Arah3* sequences identified above were used as the basis for BLASTP searches against the genomes of *P. persica* and *Cu. sativus*. Supplementary figures S5–S7, Supplementary Material online, show phylogenetic analysis of *Arah3* homologs with *P. persica*, *Cu. sativus,* and *Medicago,* respectively. Phylogenetic analysis clearly shows that *Arah3* homologs were absent in *P. persica, Cu. sativus,* and *Medicago* genomes or substantially diverged (supplementary figs. S5–S7, Supplementary Material online). The combined phylogenetic tree of *Arah3* homologs of peanut and other plants sequences indicates that the *Arah3* homologs of peanut, soybean, and *Lotus* form a separate group (clade 1, fig. 7). There clearly has been recent expansion of *Arah3* clusters in peanut as evidenced by terminal branches with multiple closely related proteins. Sequences of *Arah1* and *Arah3* homologs from *Arabidopsis thaliana, G. max*, *M. truncatula*, *Lotus japonicas, P. persica,* and *Cu. sativus* are provided in supplementary table S9, Supplementary Material online. Protein sequence similarities of *Arah1* and *Arah3* homologs of peanut with *G. max*, *M. truncatula*, *Lotus japonicus*, *Arabidopsis thaliana*, *P. persica,* and *Cu. sativus* are given in supplementary table S10, Supplementary Material online.

Identification of Orthologous Regions and Gene Synteny of Peanut with Other Plant Species

Important questions in genome evolution, particularly about the evolution of gene families and genome structure, can be addressed effectively by analysis of contiguous blocks of DNA

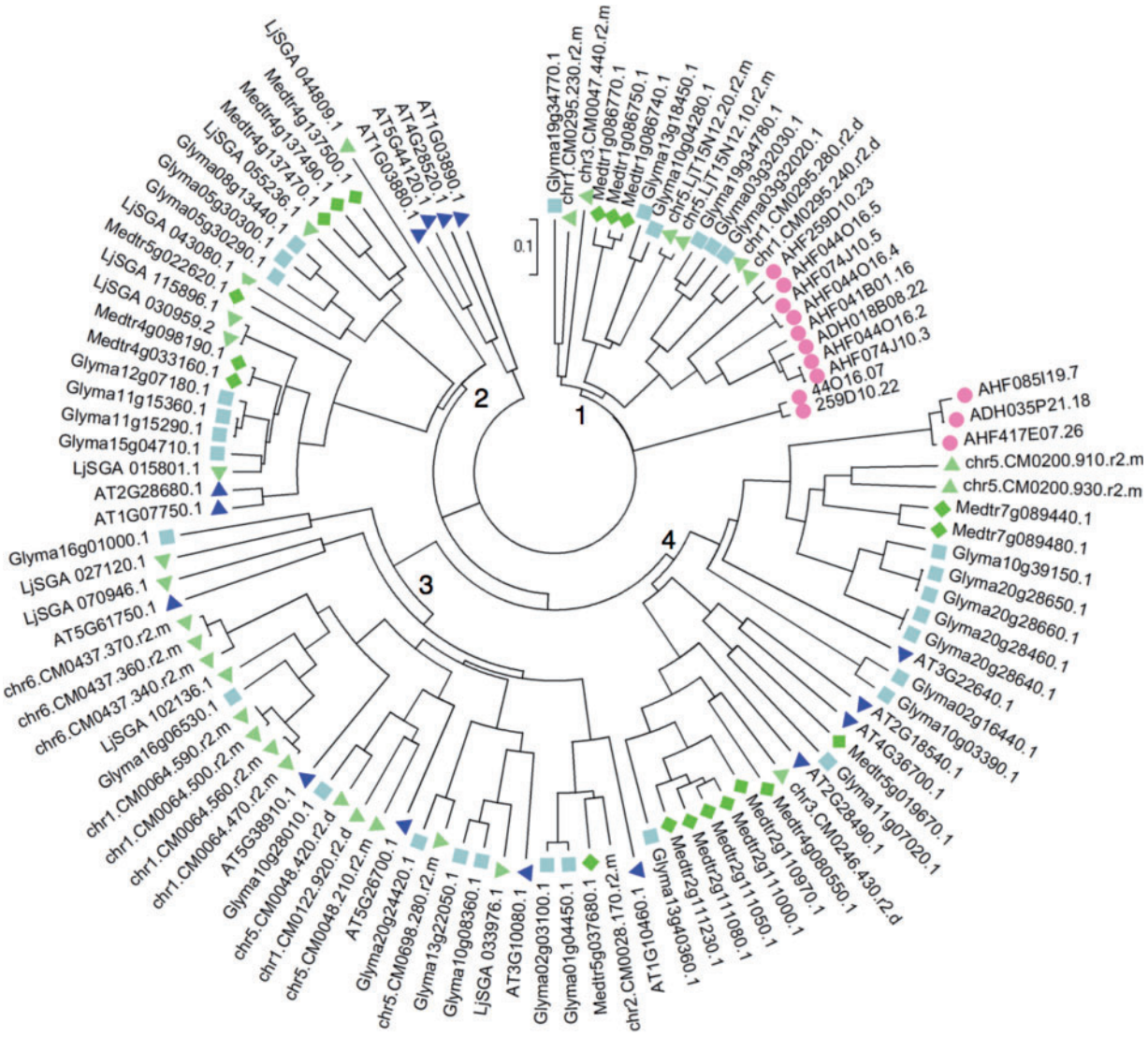

Fig. 6.—Phylogenetic analysis of *Arah1*, *Arah3,* and other seed storage proteins from peanut, soybean, *Medicago,* and *Arabidopsis*. The tree is constructed by using the PHYML program using amino acid sequences and re-editing with MEGA. The tree most strongly supported by bootstrapping (1,000 repetitions) is displayed. The phylogenetic tree was constructed by using sequences of *Arah1*, *Arah3,* and the related 7S and 11S proteins. Pink circles indicate *Arah1* and *Arah3* homologs from peanut, whereas other color indicates homologs from different species (blue: *Arabidopsis*, aqua: soybean, light green: *Medicago*). *Arah1*- and *Arah3*-related proteins were clustered in different subgroups (subgroups 4 and 1, respectively), whereas the other seed storage proteins formed two subgroups (subgroups 2 and 3). Soybean homologs (*Gy1*–*Gy7*) correspond to *Gy1*:Glyma03g32030.1, *Gy2*:Glyma03g32020.1, *Gy3*:Glyma19g34780.1, *Gy4*:Glyma10g04280.1, *Gy5*:Glyma13g18450.1, *Gy6*:Glyma03g32010, *Gy7*:Glyma19g34770, respectively. Peanut *Arah3* homologs are grouped together with glycine *Gy1*, *Gy2,* and *Gy3* (subgroup 1). *Arah1* homologous sequences were distributed on chromosomes 2, 10, 11, 13 and 20 in soybean; chromosomes 2, 4, 5 and 7 in *Medicago*; chromosomes 2, 3, 4 in *Arabidopsis*, and chromosomes 3 and 5 on *Lotus*. Sequences of *Arah1* and *Arah3* homologs from *Arabidopsis thaliana* (Ath)*, Glycine max* (Glyma), *Medicago truncatula* (Medtr), and *Lotus japonicas* (Lja) are provided in supplementary table S9, Supplementary Material online.

sequence from multiple species. We investigated the level and structure of microcolinearity between peanut and the largely sequenced genomes of *G. max* (Gm), *M. truncatula* (Mt), *Ca. cajan* (Cc), *Phaseolus vulgaris* (Pv), *Ci. arietinum* (Ca), *Vitis vinifera* (Vv), *Solanum lycopersicum* (Sl), poplar, and *Arabidopsis thaliana* (At). Although homologous sequences in peanut, soybean, common bean, chickpea, poplar, *M. truncatula*, and *Arabidopsis* are thought to have been separated from *V. vinifera* for at least 111 Myr (Wang et al. 2009), we observed conserved microsynteny among related species. We examined genome rearrangements at the microscale level, whether they are shared or lineage specific, to dissect the fine structure of conserved colinear genes. In peanut, differences in gene content were observed between the diploid

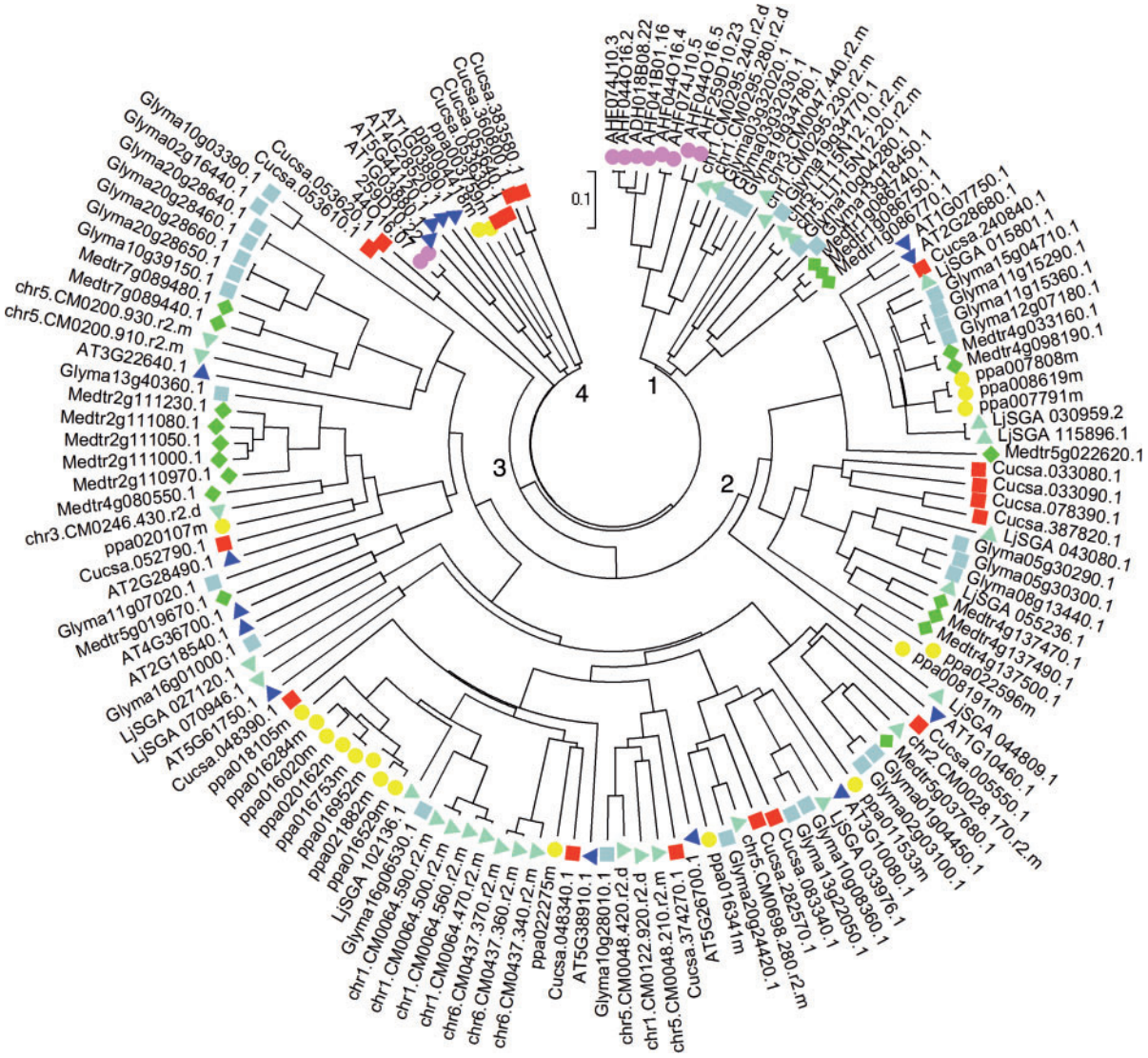

Fig. 7.—*Arah3* homologs are absent in *Prunus persica, Cucumis sativus*, and *Medicago* genomes, and have recently expanded in copy number in peanut. *Arah3*-related protein sequences from peanut (shown in pink circles), soybean, and *Lotus* are grouped together indicating that these proteins are conserved and were present in the most recent common ancestor of these taxa. There has been recent expansion of *Arah3* clusters in peanut as evidenced by terminal branches with multiple closely related proteins. Species abbreviations and color code: *Glycine max* (Glyma): Aqua box, *Medicago truncatula* (Medtr): green box, *Lotus japonicus* (Lja): acid green triangle, *Arabidopsis thaliana* (Ath): Blue triangle, *Prunus persica* (Ppa): yellow circle, and *Cucumis sativus* (Cucsa): red box. Sequences of *Arah1* and *Arah3* homologs from all the species are provided in supplementary table S9, Supplementary Material online.

A genome BAC (ADH35P21) and the tetraploid peanut BAC (AHF85I19) derived from the A genome (region between genes 5–7, 13–14, and 17–25 [fig. 8]). A retroelement was present in the syntenic region of the A genome and cultivated peanut (AB genome) (Gene No. 4, AHF85I19 and ADH35P21, respectively). Minor differences in gene content were also observed between the homoeologous BACs in peanut from A and B genomes (AHF85I19 and AHF417E07), although several genes were in the same order and orientation between the two homoeologous regions. Nucleotide substitution rates at silent sites (Ks) (Yang and Nielsen 2000) for low-copy genes spread over the entire aligned region suggested an approximate divergence time of A and B genome of 3.5 Ma as previously reported by Bertioli et al. (2013). We detected appreciable conserved synteny between peanut, *Lotus*, pigeonpea, chickpea, common bean, and soybean. Two orthologous regions were identified in the soybean genome (chromosomes 10 and 20) and one each in the pigeonpea, chickpea, common bean, and *Lotus* genomes. As more sequenced species are added to the analysis, we begin to see the actual changes that distinguish one genome from another. Gene colinearity was also observed between peanut, *Vitis*,

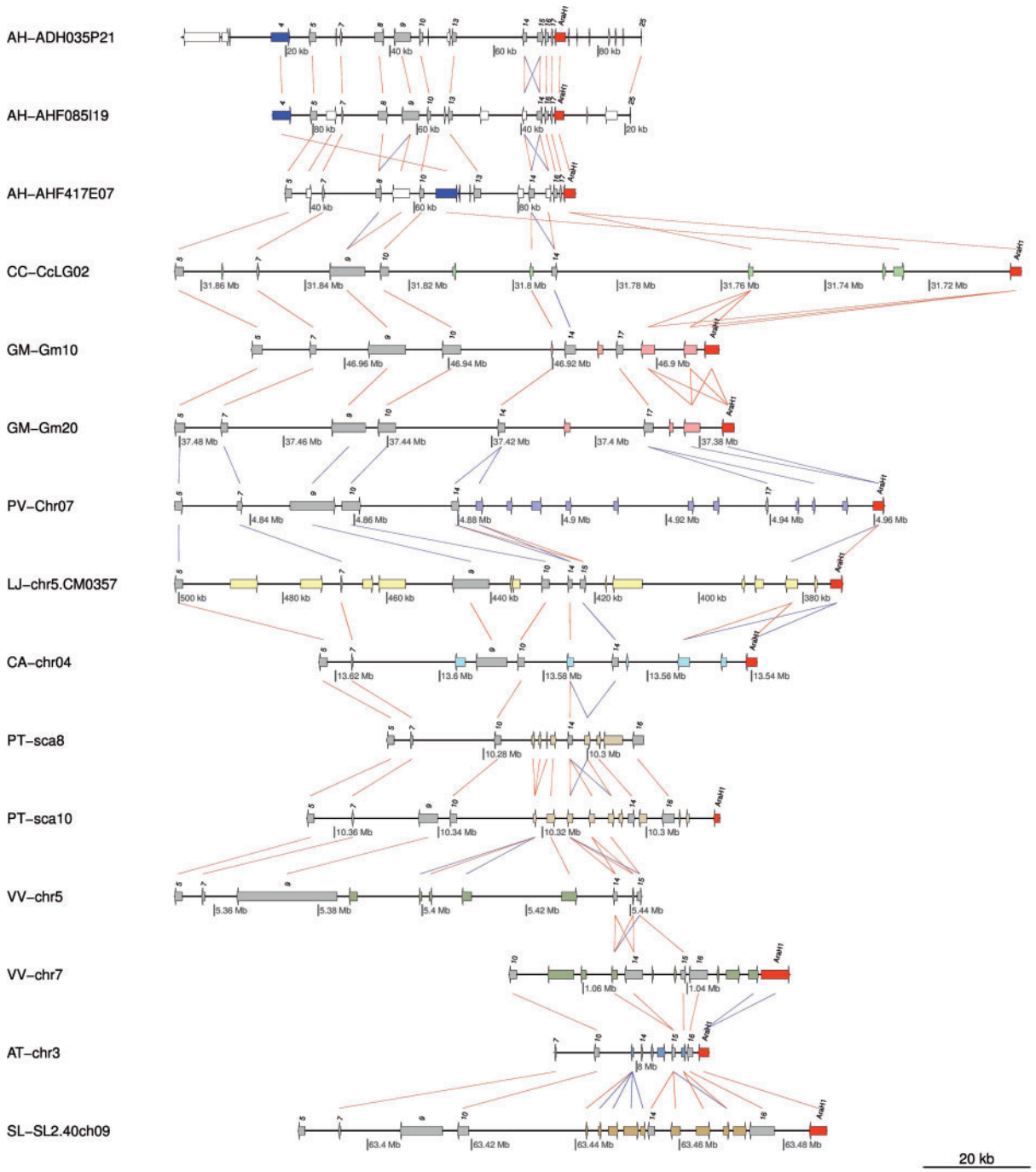

Fig. 8.—Alignment of *Arah1* BACs (ADH35P21, AHF85I19, and AHF417E07) with inferred syntenic regions from soybean (GM), pigeonpea (CC), *Medicago* (MT), *Lotus* (LJ), *Phaseolus vulgaris* (Pv), chickpea (Ca), poplar (PT), *Arabidopsis* (AT), *Vitis* (VV), and tomato (SL). Genes are shown with directional blocks, with specific colors displaying their sources. Gene synteny is shown with lines, and the color scheme displays different homologous relationships. Numbers 1–25 indicate gene order and correspond to the genes on peanut BAC ADH35P21. Boxes indicate allergen genes (red) or retroelements (blue). In peanut, differences in gene content were observed between the diploid A genome BAC (ADH35P21) and the tetraploid peanut BAC (AHF85I19) derived from A genome. Genes are conserved between the species; however, differences in gene content, gene order, and orientation were observed along with fragmented synteny between different species.

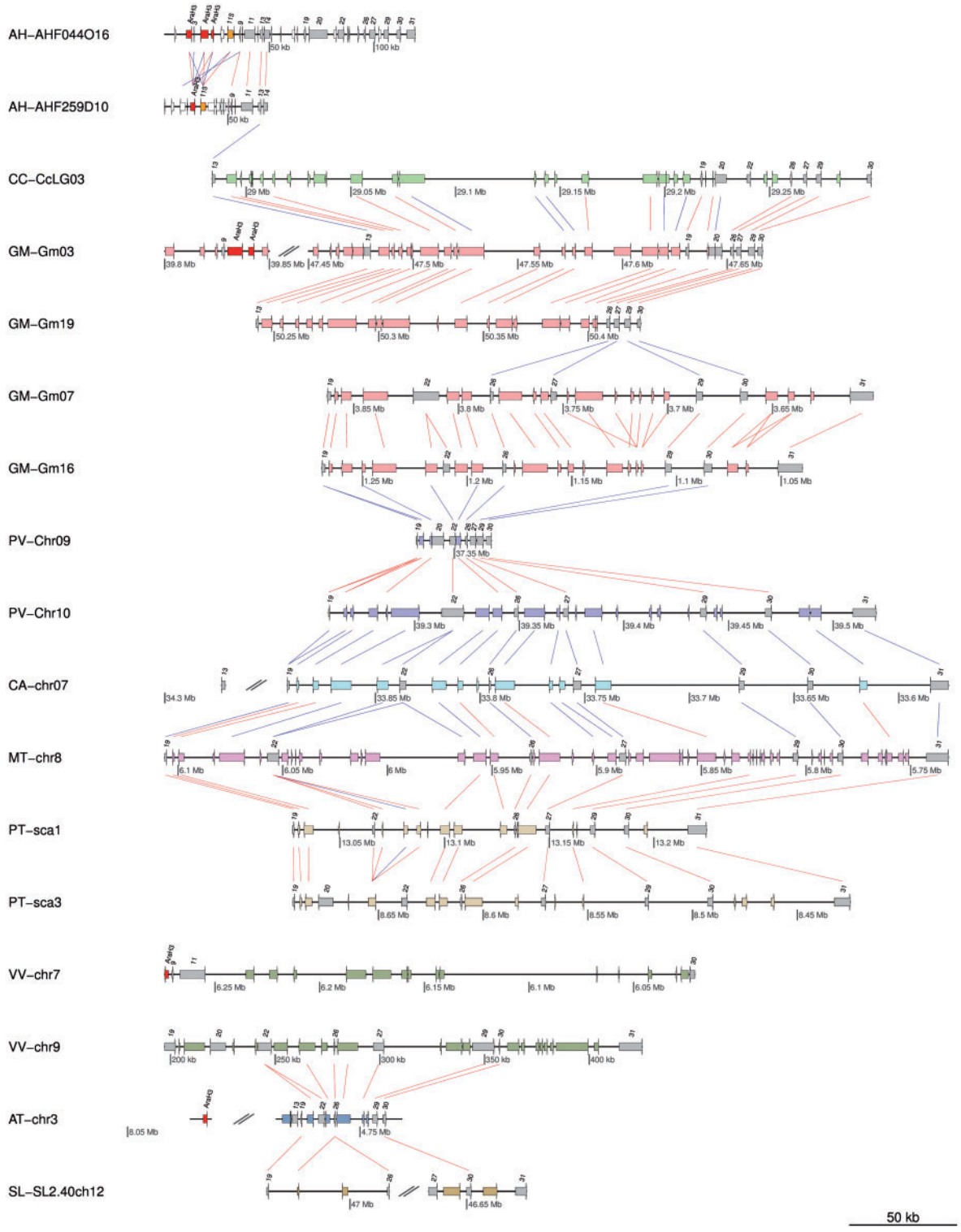

Fig. 9.—Alignment of *Arah3* BACs (AHF44O16, AHF259D10) with inferred syntenic regions from soybean (GM), pigeonpea (CC), *Medicago* (MT), *Phaseolus vulgaris* (Pv), chickpea (Ca), poplar (PT), *Arabidopsis* (AT), *Vitis* (VV), and tomato (SL). Genes are shown with directional blocks, with specific colors displaying their sources. Numbers 1–31 indicate gene order corresponding to the genes present on peanut BAC AHF44O16. Gene synteny is shown with lines, and the color scheme displays different homologous relationships. Comparative gene synteny indicates that the position of the *Arah3* gene is not well conserved in syntenic regions and *Arah3* orthologs are not seen in close proximity to the low copy genes in other plants, as observed in *Arah1*. Orthologs of *Arah3* were identified only in soybean, *Vitis*, and *Arabidopsis*. Red box indicates allergen-encoding genes.

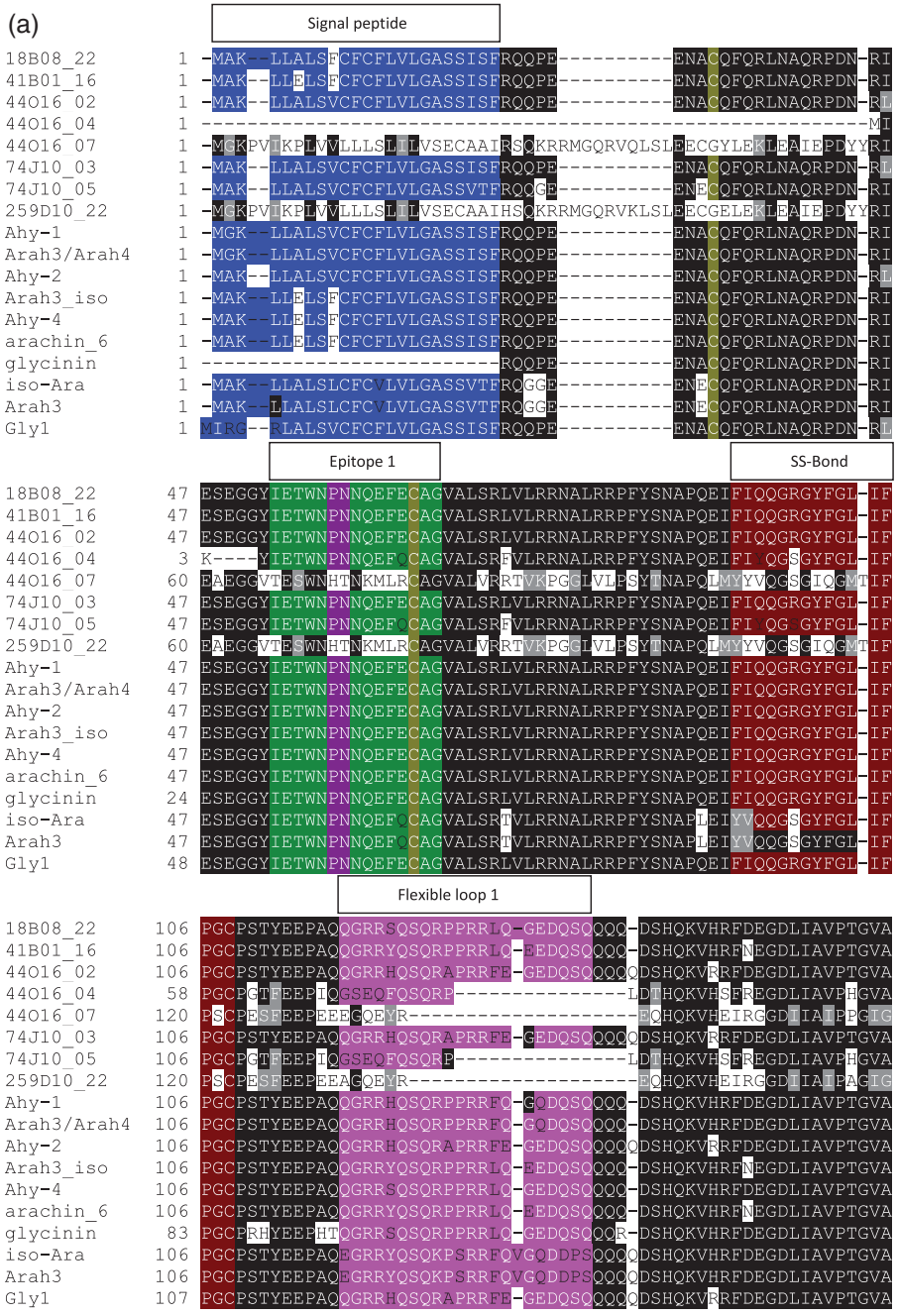

Fig. 10.—Multiple sequence alignment of *Arah3* and *Arah2* sequences from peanut. (*a*) Amino acid sequences of sixteen *Arah3* and two 11S proteins were aligned to find conserved epitopes. Epitopes 1 and 2 are conserved among the 16 *Arah3* sequences however absent from two 11S storage protein genes (44O16.04 and 259D10.22). Epitope 3 is shared by all the *Arah3* and two 11S storage proteins with minor differences in amino acid residues. Epitope 4 is not conserved and is deleted in several *Arah3* genes. Genebank accession numbers of *Arah3* homologs are: *Gly1*: AAG01363.1, *Ahy1*: AAU21490.1, *Ahy2*: AAU21491.1, *Ahy-4*: AAW56067.1, *Arah3_iso*: ACH91862.1, *Arah3/arah4*: AAM46958.1, *arachin_6*: ABL14270.1, *Iso-ara*: ABI17154.1, *Arah3*: ADQ53859.1, *Glycinin*: AAC63045.1. (*b*) Multiple sequence alignment of *Arah2* and *Arah6* sequences from peanut. Amino acid sequences of 16 allergen proteins related to *Arah2* and *Arah6* were also aligned to identify the conserved epitopes. Epitope 10 is shared by both *Arah2* and *Arah6,* whereas epitopes (1–9) are conserved only in *Arah2*. Genebank accessions numbers: *Cong_ARAHY*: AAU21495.1, *Conglutin*: AAW56068.1, *Conglutin2*: AF366561.1, *Conglutin8*: ABL14269.1, *Arah6*: AAD56337.1, *Arah6_2*: gi159163254, *Arah6_3*: AAW34231.1, *Conglutin_d*: ABQ96212.1, *Arah2isoform*: AAM78596.1, *2Sprotein1*: GI:57669905, *Conglutin_d2*: ABW36075.1, *Ara*: AY848699.1, *Conglutin d3*: ABW36069.1. Epitopes, conserved motifs, and amino acids are highlighted.

(continued)

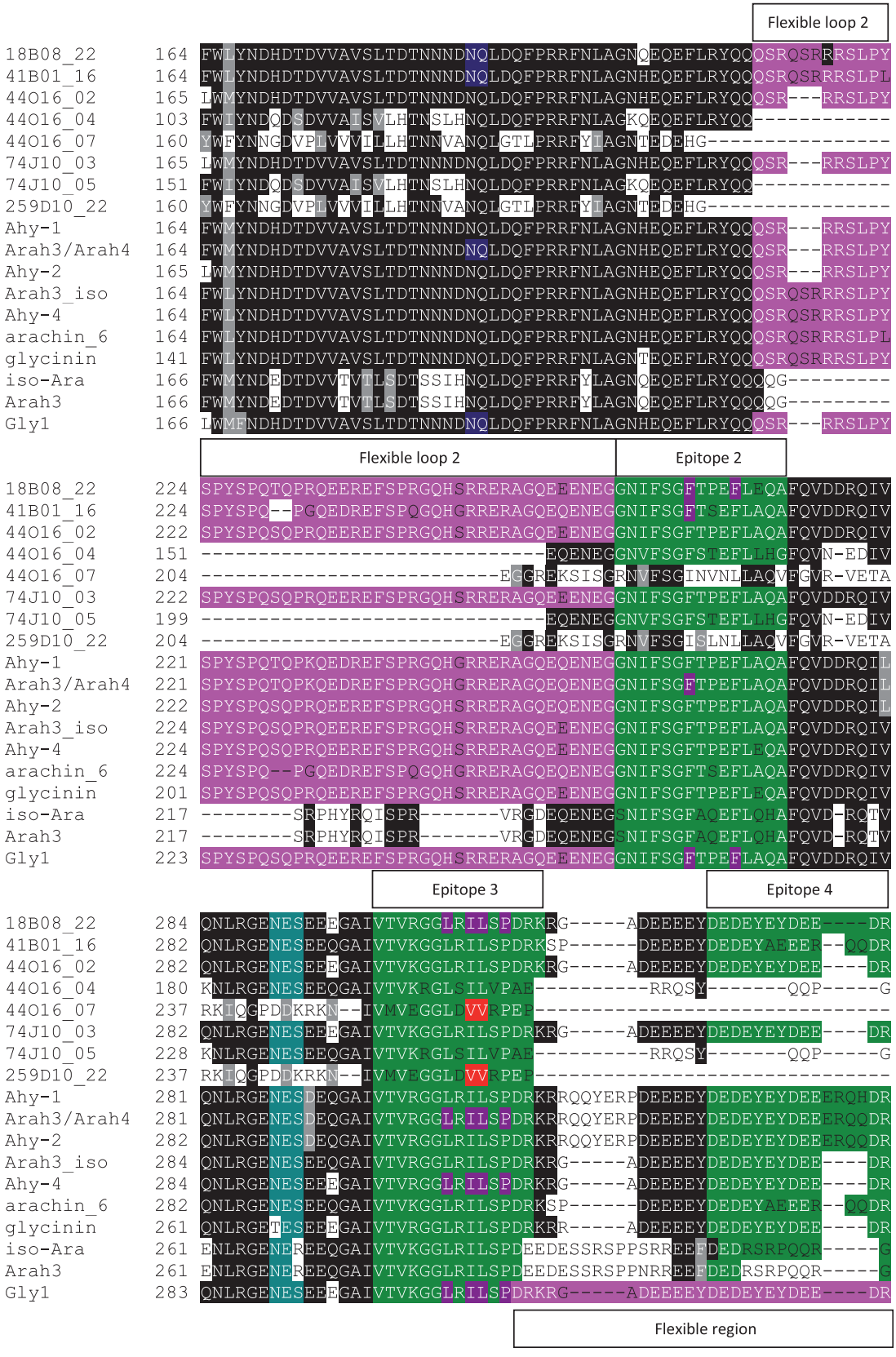

Fig. 10.—Continued.

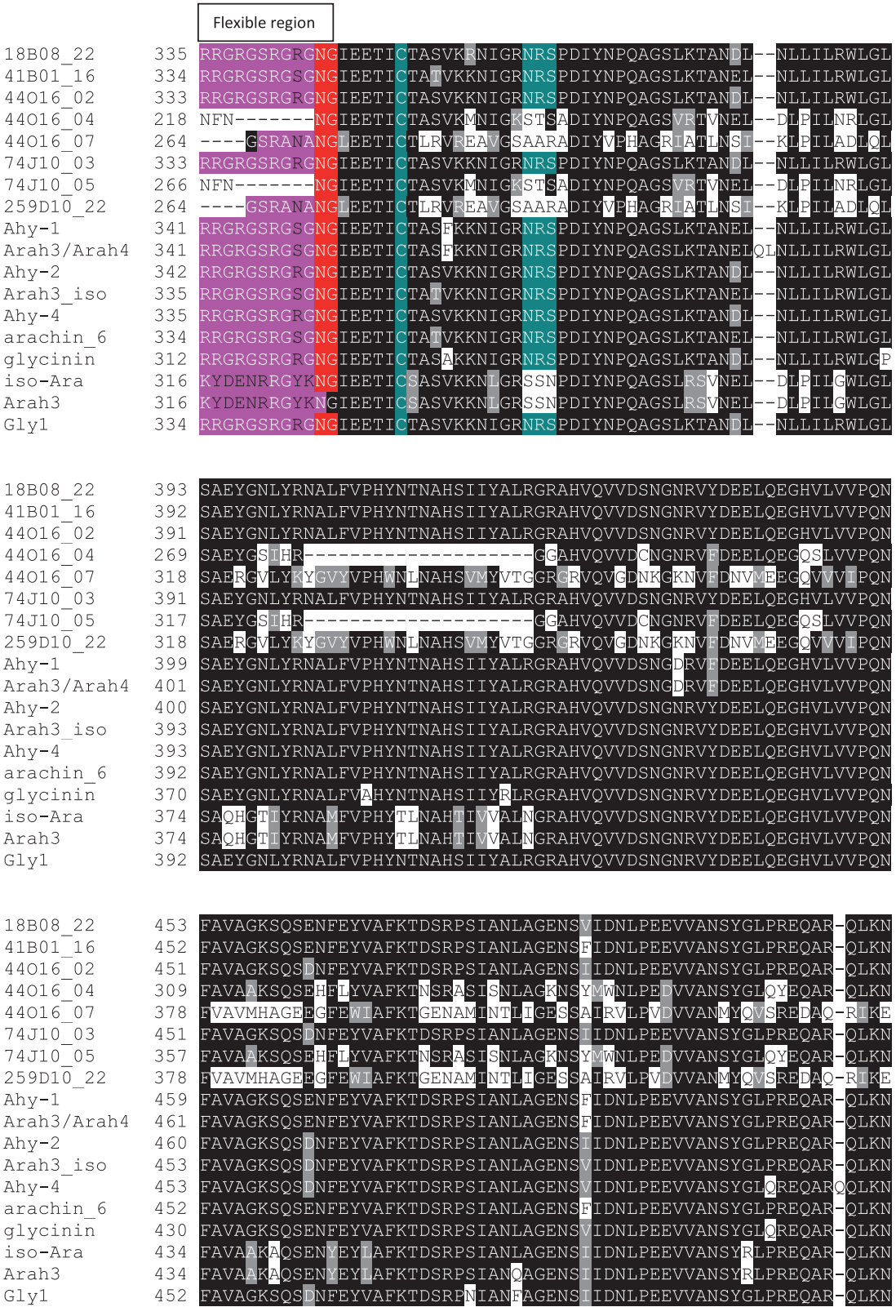

Fig. 10.—Continued.

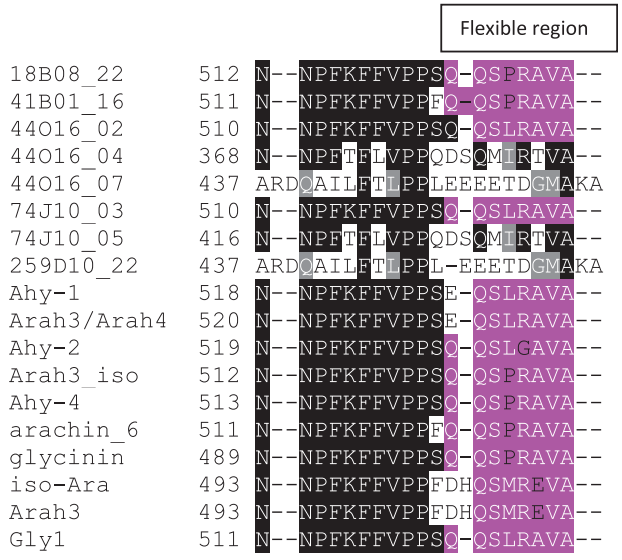

Fig. 10.—Continued.

Arabidopsis, tomato, and poplar (fig. 8). Although several genes that are conserved between species have been identified, differences in gene content, gene order, and orientation were also observed along with the fragmented synteny between different species. Our result indicates that microsynteny can be used to identify orthologous regions between peanut and other species, but to varying degrees among species and genomic regions.

Arah1 Syntenic Regions Are Widely Conserved in Eudicots Whereas *Arah3* Regions Show Genome Rearrangements

Arah1 orthologs are highly conserved in eudicots including soybean, pigeonpea, *Medicago*, *Lotus*, poplar, *Vitis*, *Arabidopsis,* and tomato. The physical location of *Arah1* is also conserved as nearby genes (Gene No. 7, 9, 10, and 14) are located in close proximity to *Arah1* homologs in pigeonpea, soybean, chickpea, common bean, *Lotus*, *Arabidopsis, Vitis*, and tomato (fig. 8). In soybean, homologs of *Arah1* are found on chromosomes 10 and 20, but absent from corresponding homoeologous chromosomes 1 and 2. In contrast to *Arah1* orthologs, the positions of the *Arah3* genes are not well conserved in syntenic regions and *Arah3* orthologs are not seen in close proximity to low copy genes in other plants (fig. 9). Orthologs of *Arah3* were identified only in soybean, *Arabidopsis,* and *Vitis*, being absent from the syntenic regions in other species studied. There also appears to have been gene loss of *Arah3* homologs in soybean and other species, or at least a failure to expand. The soybean genes most similar to *Arah3* are located on chromosome 3 and approximately 7.7 Mb from the aligned regions. Chromosome rearrangements in this region might be responsible for the different positions of *Arah3*-related genes.

We found one *Arah3* gene (ADH18B08.22) on the diploid A genome of peanut on BAC ADH18B08, which is absent from all syntenic regions in other studied crops except pigeonpea (supplementary fig. S8, Supplementary Material online). A retroelement was also found near the *Arah3* gene on BAC ADH18B08, but not in other legumes, suggesting that it was inserted in peanut after its divergence from common ancestors. We identified four intact retroelements on peanut BACs AHF41B01, AHF85I19, AHF221B03, and AHF259D10 (supplementary table S11, Supplementary Material online) that show high similarity with previously reported retroelements (Bertioli et al. 2013). BAC AHF41B01 has a "MATITA" family retroelement inserted about 1.1 Ma, whereas another "MATITA" element on BAC AHF85I19 inserted about 1.4 Ma. BAC221B03 has a "PIPA" family retroelement inserted at 1.7 Ma. The oldest element was on BAC AHF259D10, inserted at 3.7 Ma, and related to the "MICO" family. The activity of these transposons since divergence of the A and B genomes has been a very significant driver of the erosion of genome sequence similarity. Some of these transposon insertions appear to be very recent, albeit noting that the accumulation of retrotransposons in the region could have been ancient with periodic turnover of the specific elements present. In addition to complete retrotransposons, pseudogenes from different classes of transposable elements and retroviruses were present, although they could not be completely characterized. More detailed analysis of these retroelements is discussed by Bertioli et al. (2013).

Selection Pressure on *Arah1* and *Arah3* Gene Evolution

In parallel with its genomic surroundings, *Arah1* gene sequences have been evolving in a very conservative manner. GENECONV showed no evidence of conversion between *Arah1* genes. PAML showed no evidence of positive selection at the whole gene sequence scale under a free parameter model, or on each codon site by using NSsites model to do BEB analysis. For *Arah3*, we also did not find evidence for gene conversion, but a small region may have been under positive selection found by using NSsites model to do BEB analysis. One site in the region may have been significantly positively selected (supplementary table S12, Supplementary Material online).

Conserved Epitopes and Motifs in Peanut Allergen Proteins

Amino acid sequences of all sixteen *Arah3* and two 11S proteins were aligned to find conserved epitopes (fig. 10*a*). A previous study using a recombinant *Arah3* discovered four IgE epitopes containing critical amino acid residues. The number of patients sera recognizing epitope 3 is significantly larger than the number recognizing epitopes 1, 2, and 4

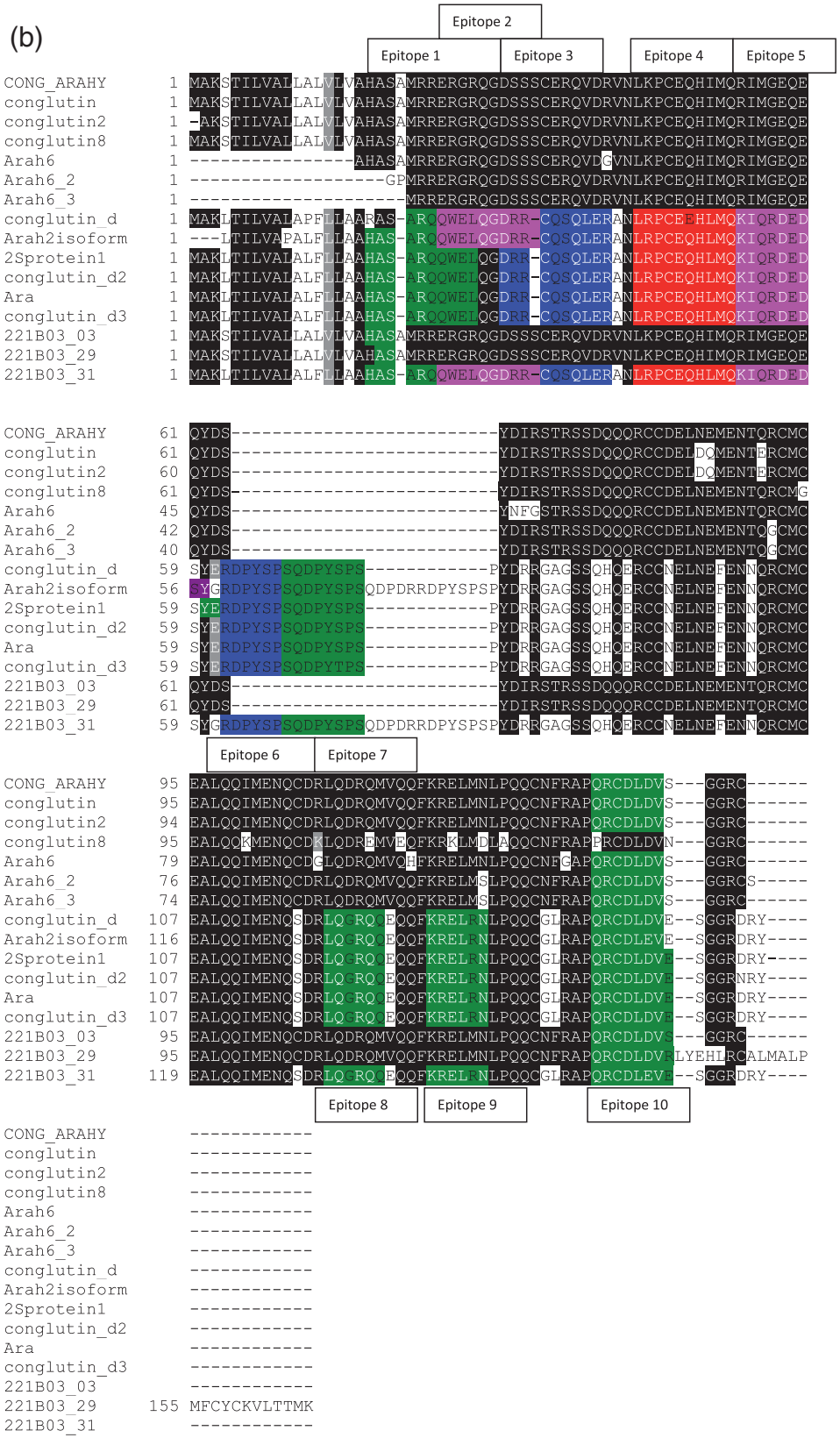

Fig. 10.—Continued.

(Rabjohn et al. 1999). We found that epitopes 1 and 2 are conserved among the 16 *Arah3* sequences but absent from two 11S storage protein genes (44O16.07 and 259D10.22). Epitope 3 is shared by all the *Arah3* and two 11S storage proteins with minor differences in amino acid residues. Epitope 4 is not conserved and is deleted in several *Arah3*. According to the identified epitope 4, it is composed of 15 amino acid residues (DEDEYEYDEEDRRRG), recognized by serum IgE from 38% of peanut patients tested and the amino acid residues, particularly the sixth residue, glutamic acid (E), are critical for IgE-binding of epitope 4. Deletions and amino acid variations were also observed in signal peptides, flexible loops and flexible regions of *Arah3* and related protein sequences. Mutation analysis of the epitopes reveals that single or critical amino acid changes within these peptides can lead to a reduction or loss of IgE-binding ability, and consequently can affect the allergen's immunogenicity (Rabjohn et al. 1999). *Arah3* and related proteins belong to the 11S globulin storage protein family that is characterized by three common features. The first one is that they contain an acidic and basic chain separated by a conserved Asn–Gly (N–G) peptide bond. Second, the formation of intra and interdisulfide bonds is observed due to four conserved cysteine residue. Third, an Asn–Gln (N–Q) peptide bond is present that serves as a potential proteolytic cleavage site (fig. 10*a*). Sixteen allergen proteins related to *Arah2* and *Arah6* were also aligned to identify the conserved epitopes. Epitope 10 is shared by both *Arah2* and *Arah6*, whereas epitopes 1–9 are conserved only in *Arah2* (fig. 10*b*).

Motifs in the Promoter Regions Are Less Conserved in *Arah1* Than *Arah3* Genes

We have analyzed the conserved motifs of promoter regions for all the *Arah1* and *Arah3* genes in sequenced BACs and their homologs in soybean and Medicago. Several *cis*-elements are highly conserved within families, such as the legumin-box (Baumlein et al. 1986), RY element, G-box, and TATA-box. Other motifs are less consistent in comparison with previous studies (Viquez et al. 2004; Fu et al. 2010), and our study as well, due to higher degrees of divergence and short motif lengths. Multiple sequence alignments of *Arah1* and *Arah3* gene promoters are shown in supplementary figure S9, Supplementary Material online. In general, motifs in promoter regions of the *Arah1* family in peanut, soybean, and Medicago are less conserved than the *Arah3* promoter regions, especially near the start codon (supplementary fig. S9, Supplementary Material online). Sequence divergence of the upstream regions was too high for reliable phylogenetic analysis. An exception is the TATA-box (5′-TAT AAA-3′) which is conserved in all promoter regions, as expected, although of slightly different positions, −47 to approximately 94 in the *Arah1* gene family and −55 to approximately 77 in *Arah3*. The three key motifs, the legumin-box, RY-element, and G-box, were conserved across all *Arah3* genes included in this study. These three *cis*-elements are also conserved across *Arah1* and *Arah3* promoters, though to a lesser extent.

Discussion

These findings address multiple questions regarding the evolution of the peanut allergen genes and genome rearrangements in the syntenic regions. Comparative analysis of three *Arah1*, one *Arah2*, eight *Arah3*, and two *Arah6* genes significantly extended our knowledge of peanut allergen genes in general, and allowed useful comparisons with other species. On the basis of syntenic alignments, we established the orthology of most sequences analyzed and we concluded that all of the syntenic regions shared a common ancestor.

Segmental and Tandem Duplications Led to the Establishment of the *Arah3* Gene Family

Our analyses provide insights into chromosomal rearrangements related to *Arah1* and *Arah3* genes. *Arah1* syntenic regions are widely conserved in eudicots including soybean, pigeonpea, common bean, chickpea *Lotus*, poplar, *Arabidopsis, Vitis,* and tomato. *Arah3* genes in the syntenic regions are present only in soybean, *Vitis* and *Arabidopsis,* and genome restructuring progressively broke down syntenic relationships between these species over evolutionary time. These results are consistent with prior analysis of glycinin gene duplication in soybean (Beilinson et al. 2002; Li and Zhang 2011). We infer that segmental duplication and tandem duplications led to the establishment of the *Arah3* gene family.

Gene Content and Order Are Well Conserved between Homoeologous Regions in Peanut

By comparing the sequences of peanut homoeologous regions with the orthologous region of other legumes such as soybean, pigeonpea, common bean, chickpea *Medicago,* and *Lotus*, we were able to hypothesize the polarity of most changes. Gene content and order are well conserved between homoeologous regions in peanut; however, some differences were also observed. These differences are mainly due to the insertion of retrotransposons in the peanut genome. Similar observations were also reported by Bertioli et al. (2013), where repetitive DNA played a key role in the structural divergence of the A and B genomes.

In the peanut homoeologous regions that we analyzed, the majority of low-copy gene duplicates has been maintained over an estimated 3.5 Myr since divergence of the A and B genomes. The conservation of most homoeologous gene duplicates implies that both copies continue to function and are under selection. The evolution of new functions or the partitioning of existing functions between gene copies are

potential outcomes for any duplicate gene pair and may lead to the selection-driven preservation of both copies (Lynch and Force 2000). Gene duplication in polyploids may be a foundation for divergence of expression networks (Blanc and Wolfe 2004), and duplicate gene preservation may be influenced by dosage balance (Birchler and Veitia 2007). Testing the roles of these evolutionary models for the peanut genome will require extensive data on the expression of homoeologous genes.

By comparing the peanut sequences with orthologous sequences in many other eudicots, we were able to gain insights into the tempo and mode of divergence of homoeologous peanut sequences. Although the genomic regions in peanut, soybean, *M. truncatula*, *Arabidopsis,* and poplar have been separated from *V. vinifera* for an estimated 111 Myr, we observed conserved microsynteny in many cases. However, differences in gene content, gene order, and orientation were also observed along with fragmentation of synteny. Synteny at the microlevel between peanut and model legumes will be useful for understanding peanut genome structure.

Arah1 and *Arah3* Gene Families Are Generally Evolving in a Conservative Manner

Allergic reactions provoked by the ingestion of legume seed-derived foods, particularly peanut, soybean, lentil, and chickpea, are common. However, sharp differences in the legume species mainly consumed, as well as their clinical prevalence and cross reactivity, are observed in different geographic areas. Moreover, mechanisms underlying peanut allergy are not fully understood yet, and peanut allergy management most often relies on a patient's compliance to avoid suspected foods. Therefore, recent efforts aim at the identification of peanut allergen-encoding genes and their comparative analysis with proteins from different species. To compare the allergen proteins from different species, we identified homologous genes of *Arah1* and *Arah3* from soybean, *Medicago*, *Lotus, Arabidopsis,* and several other species. *Arah1* is a major allergen that causes IgE-mediated sensitization in up to 95% of patients with peanut allergy. The protein is a 63 kDa glycoprotein that is present in peanuts as a highly structured, stable trimer. This allergen is a 7S seed storage glycoprotein or vicilin, which comprises 12–16% of the total protein content in peanut extracts (Rabjohn et al. 1999). The *Arah3* are 11S seed storage proteins, known to have multiple isoforms in numerous species. Each isoform of the 11S globulin is generally encoded by a single gene. Phylogenetic analysis of the cupin domain of our sequenced peanut allergens and their nearest homologs from other plant species suggested the evolutionary history of these genes, with evidence of gene duplication occurring in peanut after speciation. Our study indicates that *Arah1* and *Arah3* gene families are generally evolving in a very conservative manner, implying that they experience functional constraints.

Structural characterization of peanut allergen-encoding genes indicates differences in conserved motifs in both *Arah1* and *Arah3* proteins. We found that epitopes 1–3 are conserved, whereas epitope 4 is deleted in several *Arah3*-related proteins. Deletions and amino acid variations were also observed in signal peptides, flexible loops, and flexible regions of *Arah3* and related protein sequences. Mutation analysis of the epitopes reveals that single or critical amino acid changes within these peptides can lead to a reduction or loss of IgE binding ability, and consequently can affect the allergen's immunogenicity (Rabjohn et al. 1999). Protein isoforms can show small variations in amino acid sequences and in post-translational processing, thus potentially can be distinguished by molecular weight, isoelectric point, and peptide signature. Such variation may affect allergenicity as has been demonstrated in *Arah1* and *Arah3*. In a successful example, an *Arah3* isoform displays potentially decreased allergenicity (Kang and Gallo 2007). Single amino acid changes in characterized allergen epitopes can have a dramatic effect on IgE recognition (Burks et al. 1999). We have also analyzed the conserved motifs of promoter regions for all the *Arah1* and *Arah3* genes in sequenced BACs and their homologs from soybean and Medicago, finding *Arah1* promoter regions less conserved than those of *Arah3* in these legumes.

In conclusion, we have characterized the genomic regions containing *Arah1*, *Arah2*, *Arah3,* and *Arah6* allergen-encoding genes in peanut and their homologs in other species. By exploiting resources for diverse legumes and other eudicots, we gained insights into both the evolution of allergen-encoding genes and genome rearrangements. Phylogenetic analysis, genomic organization studies, and analysis of conserved epitopes and promoters provide new insights into the evolution of major peanut allergen-encoding genes. Such knowledge may contribute to the development of molecule-based diagnosis and allergen-specific therapies.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S12 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgment

This work was supported by the CGIAR Generation Challenge Program (TL-1 project) and the Georgia Peanut Commission (PID# 340).

Literature Cited

Barre A, Borges JP, Rouge P. 2005. Molecular modelling of the major peanut allergen Arah1 and other homotrimeric allergens of the cupin superfamily: a structural basis for their IgE-binding cross-reactivity. Biochimie. 87:499–506.

Baumlein H, Wobus U, Pustell J, Kafatos FC. 1986. The legumin gene family—structure of A B-type gene of Vicia-Faba and a possible legumin gene specific regulatory element. Nucleic Acids Res. 14: 2707–2720.
Beilinson V, et al. 2002. Genomic organization of glycinin genes in soybean. Theor Appl Genet. 104:1132–1140.
Bertioli DJ, et al. 2009. An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics 10:45.
Bertioli DJ, et al. 2013. The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodeling intergenic sequence space since its evolutionary divergence from the B genome. Ann Bot. 112(3):545–559.
Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell 19:395–402.
Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16: 1679–1691.
Bowers JE, et al. 2005. Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc Natl Acad Sci U S A. 102:13206–13211.
Burks AW, Cockrell G, Stanley JS, Helm RM, Bannon GA. 1995. Recombinant peanut allergen Ara-H-I expression and IGE binding in patients with peanut hypersensitivity. J Clin Invest. 96: 1715–1721.
Burks AW, King N, Bannon GA. 1999. Modification of a major peanut allergen leads to loss of IgE binding. Int Arch Allergy Immunol. 118: 313–314.
Burks W. 2003. Peanut allergy: a growing phenomenon. J Clin Invest. 111: 950–952.
Burks W, Sampson HA, Bannon G. 1998. Peanut allergens. Allergy 53: 725–730.
Burow MD, Simpson CE, Starr JL, Paterson AH. 2001. Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.): broadening the gene pool of a monophyletic polyploid species. Genetics 159:823–837.
Chenna R, et al. 2003. Multiple sequence alignment with the Clustal series of programs. Nucl Acids Res. 31:3497–3500.
Clarke MC, et al. 1998. Serological characteristics of peanut allergy. Clin Exp Allergy. 28:1251–1257.
Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186–194.
Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175–185.
FAO. 2009. Food and Agricultural Organization of the United Nation, FAO Statistical Database [cited 2013 Aug 10]. Available from: http://faostat.fao.org/faostat/collections?subset=agriculture.
Fu GH, et al. 2010. Epigenetic regulation of peanut allergen gene Arah3 in developing embryos. Planta 231:1049–1060.
Gordon D. 2003. Viewing and editing assembled sequences using Consed. Curr Protoc Bioinformatics, Chapter 11:Unit11.2.
Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucl Acids Res. 33:W557–W559.
Guy L, Kultima JR, Andersson SGE. 2010 . genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334–2335.
Hougaard BK, et al. 2008. Legume anchor markers link syntenic regions between *Phaseolus vulgaris*, *Lotus japonicus*, *Medicago truncatula* and *Arachis*. Genetics 179:2299–2312.
Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.
Jaffe DB, et al. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 13:91–96.
Jin T, Guo F, Chen YW, Howard A, Zhang YZ. 2009. Crystal structure of Arah3, a major allergen in peanut. Mol Immunol. 46:1796–1804.
Kang IH, Gallo M. 2007. Cloning and characterization of a novel peanut allergen Arah3 isoform displaying potentially decreased allergenicity. Plant Sci. 172:345–353.
Kang IH, Srivastava P, Ozias-Akins P, Gallo M. 2007. Temporal and spatial expression of the major allergens in developing and germinating peanut seed. Plant Physiol. 144:836–845.
Kleber-Janke T, Crameri R, Appenzeller U, Schlaak M, Beker WM. 1999. Selective cloning of peanut allergens, including profilin and 2S albumins, by phage display technology. Int Arch Allergy Immunol. 119: 265–274.
Li C, Zhang YM. 2011. Molecular evolution of glycinin and beta-conglycinin gene families in soybean (*Glycine max* L. Merr.). Heredity 106: 633–641.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.
Ma JX, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A. 101:12404–12410.
Ma JX, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14:860–869.
McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362–367.
Moretzsohn MC, et al. 2009. A linkage map for the B-genome of *Arachis* (Fabaceae) and its synteny to the A-genome. BMC Plant Biol. 9:40.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
Rabjohn P, et al. 1999. Molecular cloning and epitope analysis of the peanut allergen Arah3. J Clin Invest. 103:535–542.
Ratnaparkhe MB, et al. 2011. Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny. New Phytol. 192: 164–178.
Salamov AA, Solovyev VV. 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res. 10:516–522.
Sampson HA. 1996. Managing peanut allergy—demands aggressive intervention in prevention and treatment. Br Med J. 312:1050–1051.
Sawyer S. 1989. Statistical tests for detecting gene conversion. Mol Biol Evol. 6:526–538.
Stanley JS, et al. 1997. Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Arah2. Arch Biochem Biophys. 342:244–253.
Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol. 24: 1596–1599.
Viquez OM, Konan KN, Dodo HW. 2003. Structure and organization of the genomic clone of a major peanut allergen gene, Arah1. Mol Immunol. 40:565–571.
Viquez OM, Konan KN, Dodo HW. 2004. Genomic organization of peanut allergen gene, Arah3. Mol Immunol. 41:1235–1240.
Viquez OM, Summer CG, Dodo HW. 2001. Isolation and molecular characterization of the first genomic clone of a major peanut allergen, Arah2. J Allergy Clin Immunol. 107:713–717.
Wang H, et al. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A. 106:3853–3858.
Wang Y, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and co linearity. Nucl Acids Res. 40(7):e49.
Wu N, et al. 2013. De novo next-generation sequencing, assembling and annotation of *Arachis hypogaea* L. Spanish botanical type whole plant transcriptome. Theor Appl Genet. 126:1145–1149.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22: 1107–1118.
Yuksel B, Paterson AH. 2005. Construction and characterization of a peanut HindIII BAC library. Theor Appl Genet. 111(4):630–639.
Zhu HY, Choi HK, Cook DR, Shoemaker RC. 2005. Bridging model and crop legumes through comparative genomics. Plant Physiol. 137: 1189–1196.

Associate editor: Michael Purugganan