Electronic Journal of Biotechnology ISSN: 0717-3458
Vol.8 No.3, Issue of December 15, 2005

Received January 20, 2005 / Accepted May 19, 2005
DOI: 10.2225/vol8-issue3-fulltext-3
RESEARCH ARTICLE

Analysis of genomic sequences from peanut (*Arachis hypogaea*)

Jayashree B.
Bioinformatics and Computational Biology Unit
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh 502 324, India
Tel: 91 40 30713071
Fax: 91 40 30713075
E-mail: b.jayashree@cgiar.org

Morag Ferguson
International Institute for Tropical Agriculture (IITA)
c/o ILRI, P.O. Box 30709
Nairobi, Kenya
Tel: 254 20 422 3000
Fax: 254 20 422 3001
E-mail: m.ferguson@ilri.exch.cgiar.org

Dan Ilut
Department of Plant Biology
Cornell University
Ithaca, NY 14853-430, USA
E-mail: dcil@cornell.edu

Jeff Doyle
Department of Plant Biology
Cornell University
Ithaca, NY 14853-430, USA
Tel: 1 607 255 7972
Fax: 1 607 255 7979
E-mail: jjd5@cornell.edu

Jonathan H. Crouch*
M.S. Swaminathan Applied Genomics Lab.
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh 502 324, India
Tel: 52 55 5804 7574
Fax: 52 55 58047558
E-mail: j.crouch@cgiar.org

Financial support: Legume genomics, bioinformatics and molecular breeding research at ICRISAT has benefited from unrestricted grants from the governments of UK, Japan and European Union, and from the Generation Challenge Program.

Keywords: *Arachis hypogaea*, codon usage, gene-based markers, peanut, SSR markers.

Abbreviations:
BLAST: Basic Local Alignment Search Tool
EST: expressed sequence tag
GI: gene indices
HSP: high scoring segment pair
PCR: polymerase chain reaction
SSR: simple sequence repeat
TIGR: The Institute for Genomic Research

Peanut is an important legume crop across the world. However, in contrast to most legume crops, groundnut lacks taxonomic proximity to any major model genome. A relatively large number of genomic sequences were generated from groundnut as part of a microsatellite marker development project. In the current study, a total of 1312 sequences were analyzed of which 448 contained microsatellite motifs. All sequences (GenBank Accessions: BZ999351-CC000573) were analyzed after clustering for possible similarity with publicly available

*Corresponding author

sequences from *Arabidopsis*, *Lotus*, soybean and *Medicago*. At least 39% of the sequences analyzed had significant BLAST similarities with sequences from the four databases searched, of which nearly half (47%) found significant similarity with *Lotus japonicus* sequences. Over one quarter (26.7%) of sequences found similarity with *Arabidopsis thaliana*, while the remainder aligned with publicly available sequences from the legumes soybean and *Medicago truncatula*. At least 17% of microsatellite containing sequences could be assigned an identity. The codon usage pattern for *Arachis hypogaea* most closely resembles that of *L. japonicus* reflecting the similarly high sequence similarity observed in BLAST searches at the protein level. The implications of these findings for the taxonomy, and comparative genomics of groundnut and its legume family relatives are discussed.

Peanut (*Arachis hypogaea*), also known as groundnut, is an important legume crop across the Americas, Africa and Asia, where it is grown for local consumption and international trade, as a food and oil product. Although there is high level of morphological diversity among varieties of *A. hypogaea*, this has not been generally reflected in the level of detectable genetic diversity (Tombs, 1963; Savoy 1976; Grieshammer and Wynne, 1990; Halward et al. 1991; Kochert et al. 1991; Halward et al. 1992; Paik-Ro et al. 1992; Lacks and Stalker, 1993; Bianchi-Hall et al. 1994; Lanham et al. 1994). Groundnut is thought to have evolved relatively recently through a single hybridization event, most likely between the unreduced gametes of two diploid species representing the A and B genomes (Kochert et al. 1996). It is postulated that the resultant amphidiploid plant was then reproductively isolated from diploid wild relatives leading to a very narrow genetic base. Genetic maps have been reported for the genomes of both diploid (Halward et al. 1993) and amphidiploid (Burow et al. 2001) *Arachis*. However, as a consequence of the low level of genetic variation amongst cultivated groundnut, the first and as yet only reported genetic linkage map of amphidiploid groundnut had to be based on an interspecific cross (Burow et al. 2001). Although this is an important milestone for groundnut genomics, this map may be of limited value for molecular breeding due to the very different recombination patterns compared with the intraspecific crosses that form the basis of groundnut breeding. More recently, considerable efforts have been made to develop simple sequence repeat (SSR) markers, which generally detect higher levels of polymorphism within species than other assays.

Most of the major legume crops and model legume species are concentrated in a very limited region of the legume phylogenetic tree: viz. the sister clades of "halogalegina" (*Lotus* and *Medicago*) and "phaseoloid/millettioids" (soybean and *Phaseolus*). In contrast, *Arachis* falls in the "dalbergioid" clade, which includes no other major crops and is differentiated from all others by possession of a unique 50 kb inversion (Doyle and Luckow, 2003). Members of this clade are known to be diverse and relationships amongst members of this group and with members of other clades are still somewhat unclear. It is

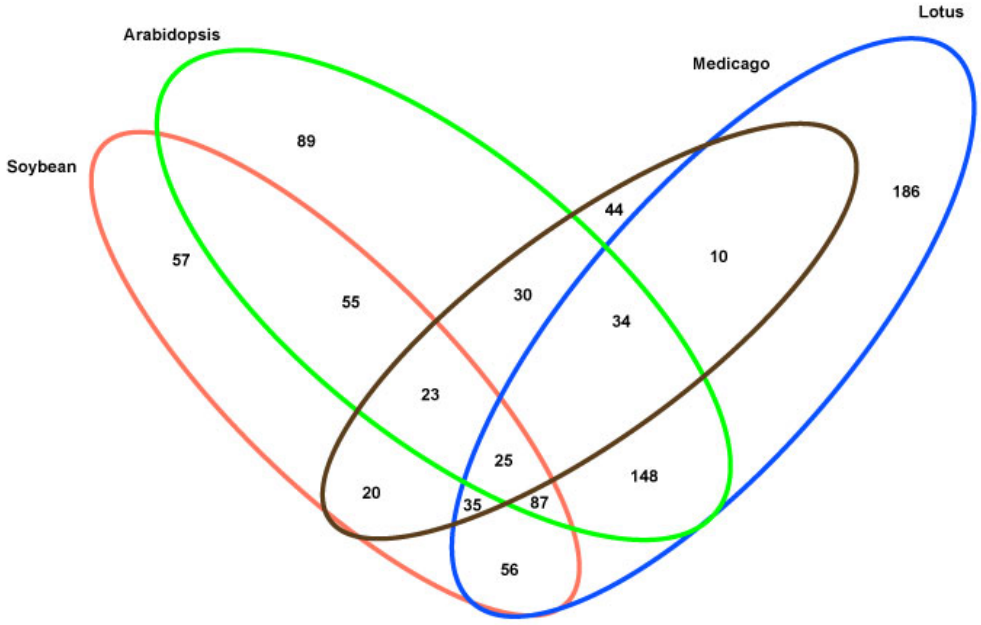

Figure 1. Sequences from *A. hypogaea*. Venn diagram summarizing the distribution of BLAST matches between *A. hypogaea* sequences and four dicot model species: *A. thaliana*, *L. japonicus*, soybean and *M. truncatula*. The maximum number of hits was with sequences from the *L. japonicus* Gene Indices (581). Only 25 sequences found matches in all the four TIGR Gene Indices searched. Numbers indicate the number of sequences sharing similarity in tBLASTx searches.

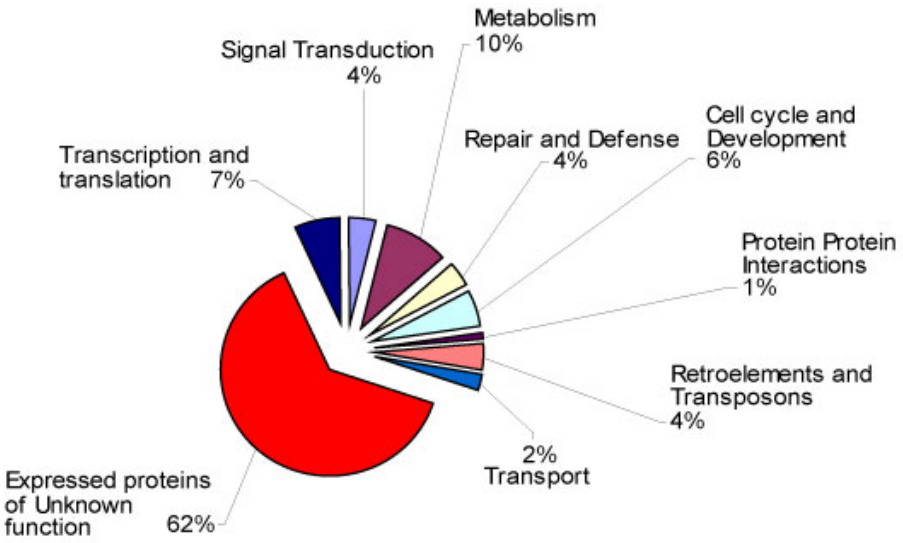

Figure 2. Functional assignments for the 475 non-redundant sequences that could be putatively annotated from tBLASTx search results. Percentages indicate number of non-redundant sequences in each functional category.

believed that a large majority (>90%) of plant sequences have close homologs within most other plant genomes (Bennetzen, 2000). Thus, comparisons of DNA sequences from taxa whose genomes have been partially or fully sequenced can to a certain extent provide estimates of relationships, such as the comparisons between *Arabidopsis* genomic sequences and rice ESTs (Eckardt, 2001), wherein it was concluded that *Arabidopsis* could not be used as a model system for monocots.

Thus, having generated 1312 new genomic sequences from groundnut (Ferguson et al. 2004) we took the opportunity to compare these sequences with major crops and model genomes, with the intention of better understanding the relationship between the genomes of groundnut*, Lotus,* soybean, *Medicago* and *Arabidopsis.* This approach was also expected to provide functional annotations for some of the new SSR markers that we have reported previously (Ferguson et al. 2004), which can then be used for the development of candidate gene-based markers. This is consistent with previous reports indicating that microsatellites are highly abundant in protein coding regions (Cardle et al. 2001). Since the methylation sensitive restriction enzymes PstI and Sau3A1/BamH1 were used during genomic library construction we expected the library to be enriched for low-copy sequences, which could find homologs during database searches. We report here, putatively annotated SSR markers, sequences that may be used as gene-based markers and the choice of codons amongst the dicots studied.

MATERIALS AND METHODS

Sequence generation and analysis

Genomic libraries were sequenced as reported elsewhere (Ferguson et al. 2004). Two peanut genomic DNA libraries were constructed, one following PstI digestion of the whole genome and the other after Sau3AI/BamHI digestion. Both libraries were probed using γ^{P32}-dATP labeled oligonucleotide probes: $(GT)_{15}$, $(GA)_{15}$, $(AAC)_{10}$, $(ATC)_{10}$, $(AGT)_{10}$, $(ATT)_{10}$, $(CAC)_{10}$, $(CTT)_{10}$ and $(CTG)_{10}$. The PstI library was also probed with $(ATCC)_{10}$, $(GATA)_{10}$ and $(AAAT)_{10}$. Filters were screened with genomic DNA to eliminate those clones containing repetitive DNA. Sequencing was performed using an ABI3700 automated DNA Analyzer (Applied Biosystems, Foster City CA). Low quality sequences were eliminated using Sequencher (Gene Codes, Ann Arbor, MI). The sequences are now available with Genbank: Accessions: BZ999351-CC000573.

Vector sequences were trimmed using Sequencher v.4 software and low quality bases (quality score <20) were trimmed from both ends of sequences. A total of 1312 sequences were selected for the analysis of which 448 contained SSR motifs. All sequences containing interspersed or simple repeats were masked with the RepeatMasker software (http://www.repeatmasker.org/). The sequences were assembled using cap3 sequence assembly software (Huang and Madan, 1999) with default parameters. The singletons and consensus sequences were screened using tBLASTx against *Arabidopsis*, soybean, *Medicago truncatula* and *Lotus japonicus* TIGR Gene Indices. An in-house script was used to retrieve best matches from the outputs of tBLASTx text files, retaining the following information: gene identifier, description, BLAST score, percent identity, alignment length, reading frame, and database name from the top-scoring alignments from each database. A transformation of e-values was carried out to pick up the best e-value across hits from all four databases. The e-value transformation was carried out

Table 1. *A. hypogaea* sequences with putative annotation to transcription factors, proteins involved in signal transduction and stress response, which could be used as markers in peanut.

Genbank ID	Putative annotation-tBLASTx description	nr ID	TIGR Gene Indices
CC000540	Disease resistance protein-like	Q9FN83	*Arabidopsis thaliana* GI
CC000350	Subtilisin-like serine proteinase	JC7519	*Arabidopsis thaliana* GI
BZ999841	Retinitis pigmentosa GTPase	Q9HD28	*Arabidopsis thaliana* GI
CC000556	Multi drug resistance associated protein (MRP)	Q9LK62	*Arabidopsis thaliana* GI
CC000353	Zinc finger like protein	AAM61245.1	*Arabidopsis thaliana* GI
CC000219	Transcription factor, partial	O04238	*Lotus japonicus* GI
BZ999844	Probable WRKY transcription factor 56	Q8VWQ4	*Arabidopsis thaliana* GI
CC000464	Similarity to helix-loop-helix DNA binding protein	GP/11994233	*Medicago truncatula* GI
CC000500	Scarecrow like protein 14	T51232	*Medicago truncatula* GI
CC000445	BZIP family	AAB86455.2	*Medicago truncatula* GI
CC000493	AtMYB84 transcription factor, complete	Q9M2Y9	*Arabidopsis thaliana* GI
BZ999621	Sensory box histidine kinase response regulator	Q886T6	*Glycine max* GI
BZ999779	EF-hand calcium binding protein-like At5g04170	Q9FYE4	*Lotus japonicus* GI
BZ999724	Protein kinase 3	Q43466	*Lotus japonicus* GI
BZ999955	Type-A response regulator	Q71BZ2	*Glycine max* GI
BZ999368	Similar to protein phosphatase 2AB kappa subunit	Q61621	*Glycine max* GI
CC000278	Heat shock protein 105kDa	Q92598	*Arabidopsis thaliana* GI
CC000424	Sucrose transport protein	Q9XHL6	*Lotus japonicus* GI

using the following formula: E = (m*n)/(2^S') where m = length of query sequence, n = reference target database length, S'= the normalized HSP score, based on the method of Altschul et al. 1997. The resultant best hit was then used to annotate the sequence if there was a minimum of 30% identity over at least 20% of the protein sequence (Quackenbush et al. 2000; Quackenbush et al. 2001). Such matches were considered to provide a reasonable putative annotation. Tandem repeats were searched for using RepeatFinder (http://tandem.bu.edu/trf/trf.advanced.submit.html). Alignment parameters (2,3,5) were used for match, mismatch and indels, and a maximum period size of 10 was applied.

The EMBOSS tool 'cusp' was used to calculate the amino acids and codon usage frequencies of *Arachis* sequences (http://bioweb.pasteur.fr/seqanal/interfaces/cusp.html). Previously reported codon usage tables for *Arabidopsis*, soybean, *Medicago* and *Lotus* were obtained from http://www.kazusa.or.jp/codon/ (Nakamura et al. 2000). Relative Synonymous Codon Usage values (RSCU) were calculated using the formula $RSCU_{ij} = X_{ij}/1/n_i\Sigma\ X_{ij}$, where X_{ij} is the frequency of occurrence of the *j*th codon for the *i*th amino acid, and n_i is the number of codons for the *i*th amino acid. The degree of preference for a particular codon was examined. In the absence of any codon usage bias, the RSCU value would be 1.00. A codon that is used less/more

Table 2. *A. hypogaea* sequences containing micro satellites (simple sequence repeats-SSRs) and their putative similarities to known plant proteins.

Genbank Id	SSRs	Putative annotation tBLASTx-description	TIGR Gene Indices
BZ999634	TTA(9),TAT(6)	Phytochelatin synthase 1	*Arabidopsis thaliana* GI
CC000258	TTA(17),TGG(10)	AgCP5892	*Arabidopsis thaliana* GI
BZ999605	TCCCTTC(3),CTCTT(2)	Sterol 24-C-methyltransferase	*Lotus japonicus* GI
CC000294	TC(6)	Chitin-bindinglectin 1 precursor	*Glycine max* GI
CC000393	TC(13)	Abnormal spindle-like protein	*Arabidopsis thaliana* GI
CC000429	TC(12)	Structural maintenance of chromosomes (SMC)-like protein	*Arabidopsis thaliana* GI
CC000326	TAGGGT(3)	Chloroplast nucleoid DNA binding protein	*Medicago truncatula* GI
BZ999561	TAA(8)	Pherophorin-dz1protein precursor	*Lotus japonicus* GI
CC000313	GA(17)	RAB2A, complete	*Lotus japonicus* GI
CC000325	GA(10)	PDR-like ABC transporter	*Glycine max* GI
CC000331	CT(19),CA(9)	Squalene synthase, complete	*Lotus japonicus* GI
CC000442	ACC(5),GAA(6), GCCGCA(3)	3-hydroxy-3-methylglutaryl-coenzyme A	*Medicago truncatula* GI
CC000357	CAG(6)	BEL1-like homeodomain 1	*Glycine max* GI
CC000353	CA(9)	n-calpain-1 large subunit	*Arabidopsis thaliana* GI
BZ999495	ATTCTC(2)	Transcription factor	*Lotus japonicus* GI
BZ999565	ATGGCT(2)	Cysteine-tRNA ligase	*Lotus japonicus* GI
CC000387	AT(21),TTGTGTG(2), TG(6)	Calcium/calmodulin-dependent	*Lotus japonicus* GI
BZ999564	AT(16)	Eukaryotic translation initiation factor 4G	*Arabidopsis thaliana* GI
CC000431	AT(10),AG(12)	Farnesyltransferase beta subunit	*Arabidopsis thaliana* GI
BZ999563	AG(7)	Uricase (Nod-35) complete	*Medicago truncatula* GI
CC000312	AG(42)	Inositol polyphosphate5-phosphatase	*Glycine max* GI
CC000286	AAT(8)	Cellulase F19H22.110 - *Arabidopsis thaliana*	*Arabidopsis thaliana* GI

frequently than expected will have a value of less/more than 1.00.

RESULTS AND DISCUSSION

A total of 1312 sequences were analyzed of which 448 contained microsatellite motifs. The average length of sequences in this dataset was 542 bp. When assembled using cap3, 135 sequences assembled into 56 contiguous

sequences (generally comprising 2-4 sequences), while 1177 sequences remained as singletons.

In silico comparative legume genomics

Comparative genomics enables the application of data from one species to investigations of taxonomically disparate species, as has been achieved for example in the annotation of wheat sequences using information from *E. coli*, human, *A. thaliana* and rice (Heslop-Harrison, 2000). Thus, large-scale genome sequencing projects have been partially justified on the basis of the premise that knowledge of the whole sequence of *Arabidopsis*, for example, will help in the isolation from crop plants of agronomically important genes that bear homology to *Arabidopsis* genes. While sequence-based comparisons using BLASTn have been used to identify similar sequences from other taxa (rice ESTs compared with *A. thaliana* genomic sequences for instance), it has been suggested that BLASTn may not always be reliable when dealing with divergent, partially sequenced genomes (Devos et al. 1999). A comparison at the amino-acid level is considered more effective. About 38.5% (475 singletons and contigs from a total of 1233 sequences) of the *Arachis* sequences had significant similarity with sequences in public databases at the amino acid level. All contigs and singletons were searched against *Arabidopsis*, soybean, *Lotus* and *Medicago* Gene Indices. Of the sequences that found matches in any of the four databases, the maximum number of matches was with sequences from the *L. japonicus* Gene Indices (581 hits), followed by 491 hits against the *Arabidopsis* Gene Indices, 358 hits against the soybean Gene Indices and only 221 hits against *Medicago* Gene Indices (Figure 1). A total of 325 *Arachis* sequences found matches with both the *Lotus* and *Arabidopsis* data bases, while 203 sequences found matches with both the *Lotus* and soybean databases. Finally 116 *Arachis* sequences found hits with all three databases (*Lotus*, *Arabidopsis* and soybean), while only 25 sequences found matches in all four databases searched. These numbers depict the numbers of *Arachis* sequences finding matches in each of the four databases searched; they do not indicate numbers after e-value transformation and selection of the best match across databases.

The TIGR Gene Indices integrate data from international EST sequencing and gene research projects, and their process of clustering ESTs into contigs or TCs (tentative consensus sequences) reduces redundancy levels. As expected, the *Arabidopsis* Gene Indices dataset size is far larger than the other databases that were screened. Though there are a considerable number of sequences from the model legumes *Medicag*o, soybean and *Lotus* in the public domain, the size of these databases is still smaller than *Arabidopsis,* which is almost 5 times the size of the *Lotus* dataset. Thus, when carrying out searches against several databases it is important to account for the differences in sizes amongst target databases, so that abundance in target data set does not bias the distribution of the best hits. The process of transforming e-values by calculating an equivalent e-value for a reference database using a formula based on Altschul et al. 1997 (see Materials and Methods) helps to overcome the effect of variation in database size, when comparing searches against several databases. Thus, the best match from across the four species was used for the process of putative annotation. Furthermore, parameters such as "percent identity" and length of alignment of the query sequence with subject were used to determine whether the alignment could be considered for annotation. Using this process 475 of the 1233 non-redundant sequences were assigned a putative identity. Of the 475 sequences that have been annotated, 222 were assigned their putative annotations from *L. japonicus* sequences (46.7%), 127 (26.7%) from *Arabidopsis*, 89 from soybean and 37 from *Medicag*o. The large disparity between the frequency of matches with *Medicago* versus *Lotus* and soybean suggests that genome relationships amongst the legumes are highly complex. Moreover, the genomic relationships between legumes and *Arabidopsis* may also be more complex than previously expected.

The low degree of sequence redundancy as evidenced from the clustering process (56 contigs and 1177 singletons) suggests that the use of methylation sensitive restriction enzymes has enriched for low copy number sequences. However, the annotation of only 38.5% of the sequences, suggests that *Arachis* is relatively diverged from all the other species studied here.

Comparative studies have revealed that (a) the evolution of essential gene coding regions has proceeded relatively slowly, as a result of which reproductively isolated taxa have retained recognizable intragenic regions and (b) a large majority (>90%) of plant sequences have close homologs within most other plant genomes (Bennetzen, 2000; Paterson et al. 2000). Comparisons of DNA sequences from taxa whose genomes have been fully sequenced can to a certain extent provide estimates of relationships, such as the comparisons between *Arabidopsis* genomic sequences and rice EST's (Eckardt, 2001), wherein it was concluded that *Arabidopsis* could not serve as a model system for monocots. While the present study is limited to sequence comparisons, which cannot by themselves explain relationships at the species level, they do provide valuable practical insight into the similarities, differences and relative distances amongst the legumes and beyond. Thus, it appears from these preliminary comparisons (Figure 1) that *L. japonicus* could be the most appropriate model species for peanut comparative genomics and genome evolution studies.

Assigning putative identities

Putative functional annotations for the sequences were obtained from the most significant match obtained from database searches. The putative annotations were grouped under nine general categories based on similarity of biochemical function, which is summarized in Figure 2. A large number of sequences had similarity with expressed

Table 3. Relative Synonymous Codon Usage values (RSCU) in *A. hypogaea, L. japonicus, A. thaliana, G. max and M. truncatula.* Values given are RSCU (relative synonymous codon usage).

Amino acid Name	Codon triplet	*Arachis hypogaea*	*Lotus japonicus*	*Arabidopsis thaliana*	*Glycine max*	*Medicago truncatula*
Ala	GCA	1.3*	0.8	1.0	1.3	1.4
	GCC	1.0	1.2	0.6	0.8	0.5
	GCG	0.3	0.3	0.5	0.4	0.2
	GCT	1.3*	1.5*	1.7*	1.5*	1.8*
Cys	TGC	1.2*	1.0*	0.8	0.8	0.6
	TGT	0.7	0.9*	1.2*	1.2*	1.3*
Asp	GAC	0.8	0.9*	0.6	0.7	0.5
	GAT	1.1*	1.0*	1.4*	1.3*	1.4*
Glu	GAA	0.9*	0.9*	1.1*	1.3*	1.1*
	GAG	1.0*	1.0*	0.9	0.7	0.8
Phe	TTC	1.2*	1.2*	0.9*	0.6	0.8
	TTT	0.7	0.7	1*	1.4*	1.1*
Gly	GGA	1.3*	1.1*	1.5*	1.4*	1.5*
	GGC	0.9	1.1*	0.5	0.7	0.4
	GGG	0.5	0.5	0.6	0.8	0.4
	GGT	1.1	1.1*	1.4	1.1	1.5*
His	CAC	1.0*	1.1*	0.8	0.7	0.6
	CAT	0.9*	0.8	1.2*	1.3*	1.3*
Val	GTA	0.4	0.2	0.6	0.9	0.6
	GTC	0.7	1.0	0.8	0.6	0.5
	GTG	1.2	1.1	1.0	0.9	0.9
	GTT	1.6*	1.5*	1.6*	1.5*	1.9*
Ile	ATA	0.6	0.4	0.7	1.0	0.8
	ATC	1.2*	1.4*	1.0	0.6	0.7
	ATT	1.1*	1.1	1.2*	1.3*	1.4*

Lys	AAA	0.7	0.9	0.9	0.6	1.0*
	AAG	1.2*	1.1*	1*	1.3*	0.9*
Leu	CTA	0.4	0.4	0.6	0.7	0.6
	CTC	1.4	1.7*	1.0	0.7	0.6
	CTG	0.6	0.7	0.6	0.5	0.4
	CTT	1.7*	1.4	1.5*	1.4*	1.7*
	TTA	0.4	0.3	0.8	1.2	0.9
	TTG	1.3	1.1	1.3	1.3	1.5
Asn	AAC	1.3*	1.2*	0.9*	1.3*	0.7
	AAT	0.6	0.7	1.0*	0.6	1.2*
Pro	CCA	1.4*	1.0	1.3	1.4*	1.6*
	CCC	0.6	1.0	0.4	0.8	0.4
	CCG	0.5	0.6	0.7	0.4	0.3
	CCT	1.3	1.3*	1.5*	1.4*	1.6*
Trp	TGG	1	1	1	1	1
Gln	CAA	1.0*	1.0*	1.1*	1.4*	1.3*
	CAG	0.9*	0.9*	0.8	0.6	0.6
Arg	AGA	1.6*	1.4*	2.1*	2.1*	2.2*
	AGG	1.7*	1.4*	1.2	1.5	1.4
	CGA	0.4	0.7	0.7	0.7	0.5
	CGC	0.8	1.1	0.4	0.5	0.4
	CGG	0.4	0.4	0.5	0.5	0.3
	CGT	0.8	0.8	1.0	0.7	1.0
Ser	AGC	1.3*	0.8	0.7	0.7	0.6
	AGT	0.8	0.5	0.9	1.0	1.0
	TCA	1.2*	1.2	1.2	1.3	1.6*
	TCC	1.0	1.5*	0.7	0.9	0.7
	TCG	0.4	0.3	0.6	0.4	0.3
	TCT	1.1	1.5*	1.7*	1.4*	1.6*

Thr	ACA	1.0	1.0	1.2	1.2	1.5*
	ACC	1.3*	1.5*	0.8	0.8	0.7
	ACG	0.3	0.4	0.6	0.4	0.2
	ACT	1.2	0.9	1.4*	1.4*	1.5*
Tyr	TAC	1.0*	1.2*	0.9*	0.6	0.8
	TAT	0.9	0.7	1.0*	1.4*	1.1*
TER	TAG	0.7	0	0.6	0.7	0.6
	TGA	0.9	0	1.3	0.9	1.2
	TAA	1.3	0	1.0	1.3	1.1
Total number of codons	19531	9071	25973879	74581	101743	

*most frequently used codons

proteins of unknown function. Amongst those that could be assigned a functional classification, 10% (47/475) were grouped under the category 'Metabolism' that includes amino acid, carbohydrate and lipid metabolism. The functional category transcription factors had 34 sequences assigned, while there were a small number of sequences that were transposons or retroelements (20), and about 18 sequences had putative annotations with proteins involved in repair/defence, including chaperones, disease resistance protein etc. (Table 1). A total of 76 sequences containing SSRs had significant matches, most of them with proteins of unknown function. Only those corresponding to orthologous sequences annotated for known proteins are presented in Table 2.

Codon usage

Patterns of codon usage were also assessed in the *A. hypogaea* sequences, and compared with those of the other three legumes and *Arabidopsis*. Codon usage can be surprisingly biased in different species. It is well known that the pattern of synonymous codons used varies depending on the organism involved, and each type of genome has a particular codon usage profile (Grantham et al. 1981). Knowing an organism's preferred codon usage is of direct practical relevance in minimizing degeneracy of PCR primers, while trends in codon usage can also be used in molecular phylogenetic reconstruction. The codon usage table for *A. hypogaea* was compared with previously reported codon usage tables for *Arabidopsis*, soybean, *Medicago* and *Lotus*. We have studied the computed relative synonymous codon usage value (RSCU) of each codon, the RSCU being the observed frequency of that codon divided by the frequency expected if usage of synonymous codons was uniform. The overall codon usage for *A. hypogaea* is presented in Table 3 along with that of *Arabidopsis,* soybean, *Medicago* and *Lotus* for comparison. These data illustrate similar codon usage patterns amongst these dicots, with 8 of the 18 frequently used codons being common amongst all the five plant species. The codon usage pattern for *A. hypogaea* most closely resembles that of *L. japonicus,* with 15 of the 18 most frequently used codons in *A. hypogaea* being the most commonly used in *L. japonicus*. *A. hypogaea* and *L. japonicus* differ from the other dicots in their usage of codons for two-fold degenerate amino acids cysteine, glutamine, phenylalanine, histidine, tyrosine and the three-fold degenerate isoleucine. *A. hypogaea* and *L. japonicus* differ from each other only in the use of codons for six-fold degenerate amino acids leucine and serine, and the four-fold degenerate proline. Amongst the other two legumes and *Arabidopsis*, soybean and *Arabidopsis* are closest, all 18 frequently used codons in *Arabidopsis* are also the most frequently used codons in soybean, while 17 of the 18 frequently used codons are common between *Arabidopsis* and *Medicago*. This is consistent with comparative data at the structural level, where comparisons of *Arabidopsis* with soybean report significant synteny along the entire length of *Arabidopsis* chromosome 1 and soybean linkage group A2 and blocks of synteny were found among other chromosomes as well (Grant et al. 2002). Soybean and *Arabidopsis* are estimated to have diverged from a common ancestor 92 million years ago. In contrast, a comparative study between *Medicago truncatula* and *Arabidopsis* concluded low levels of microsynteny (Zhu et al. 2003). Finally, there is a bias towards C/G in the third position in *Arachis* and *Lotus* whereas soybean, *Medicago* and *Arabidopsis* tend to choose codons ending with A/T in the third position.

Codon usage is broadly similar in closely related species and tends to diverge with increasing phylogenetic distance. As codon usage divergence is correlated with evolutionary distance (Grantham et al. 1981; Long and Gillespie, 1991), it has been suggested that codon usage can help unravel the evolutionary relationships between species (Goldman and Yang, 1994; Nesti et al. 1995). The very similar pattern of codon usage in *A. hypogaea* and *L. japonicus* relative to the other three dicots, reflects the sequence similarity observed during the BLAST searches at the protein level, although the coding strategy will vary at the level of the individual gene. Thus, codon usage profiles also support the genomic relationship between the dicot species investigated in this study.

CONCLUDING REMARKS

A primary objective of this analysis was to infer biological function from the conservation of homologous DNA sequences between plant species. This approach was expected to provide functional annotations for new SSR markers and facilitate the development of gene-based markers in *A. hypogaea*. This analysis also has the potential for contributing to the understanding of genome evolution from the DNA/protein sequence perspective. Only 38.5% of the sequences found putative annotations, this may be due to the fact that not all *Arachis* sequences generated in this study are coding sequences, and non-coding sequences are known to evolve much faster than coding regions do. This is highly relevant given that the target legume sequences used in searches are transcribed sequences (TIGR Gene Indices). Several of the microsatellite-containing sequences have been assigned a putative identity and at least some of these will now be useful as gene-based markers in peanut research and breeding. The remaining sequences with putative identity assigned can also serve as candidate genes. In particular those sequences putatively identified to be involved with disease resistance/defence and retroelement-containing sequences can serve as candidate gene-based markers in diversity analysis and trait mapping studies. Combining expression studies with computational analysis using publicly available tools and curation should help in assigning an identity to the sequences with unknown function, as has recently been shown with "hypothetical" genes from the *H. influenzae* dataset (Kolker et al. 2004). All the *Arachis* sequences studied here are available at GenBank (Accessions: BZ999351-CC000573) and could be used in comparative and functional genomic studies.

In this study we have found that *L. japonicus* has considerable sequence similarity with *A. hypogaea*. To evaluate codon choice trends for each plant species and to help explain the observed similarity, the codon usage tables for all the five-plant species were studied. Though the overall coding strategy in *Arachis* appears similar to that of other dicots, there is reason to believe that based on comparison with those legume species with large-scale public repositories, *Arachis* is closer to *Lotus*, from which *Medicago* is more distant and soybean the most distant. While the data presented here is from the complete protein coding sequences available in GenBank, and not from individual genes, they still provide an overall picture of the relationships between the five dicots and in particular the differences amongst the legumes. The observed similarities correlate with relationships proposed from the sequence analysis data. This data can be a useful consideration when studying the levels of gene expression in these plants. We anticipate that these sequences and their annotations will prove useful for those interested in plant comparative genomics in general, and for legume researchers and molecular breeders in particular.

REFERENCES

ALTSCHUL, S.F.; MADDEN, T.L.; SCHÄFFER, A.A.; ZHANG, J.; ZHANG, Z.; MILLER, W. and LIPMAN, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 1997, vol. 25, no. 17, p. 3389-3402.

BENNETZEN, J.L. Comparative sequence analysis of plant nuclear genomes. *Plant Cell*, 2000, vol.12, no. 7, p.1021-1030.

BIANCHI-HALL, C.; KEYS, R. and STALKER H. Use of protein profiles to characterize peanut cultivars. *Peanut Science*, 1994, vol. 21, no. 2, p.152-159.

BUROW, M.D.; SIMPSON, C.E.; STARR, J.L. and PATERSON, A.H. Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (Arachis hypogaea L.): Broadening the gene pool of a monophyletic polyploid species. *Genetics*, 2001, vol. 159, no. 2, p. 823-837.

CARDLE, L.; RAMSAY, L.; MILBOURNE, D.; MACAULAY, M.; MARSHALL, D. and WAUGH, R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics*, 2001, vol. 156, no.2, p. 847-854.

DOYLE, J.J. and LUCKOW, A.M. The rest of the Iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiology*, 2003, vol. 131, no. 3, p. 900-910.

DEVOS, K.M.; BEALES, J.; NAGAMURA, Y. and SASAKI, T. Arabidopsis-Rice: will colinearity allow gene prediction across the eudicot-monocot divide? *Genome Research*, 1999, vol. 9, no. 9, p. 825-829.

ECKARDT, N.A. Everything in its place: Conservation of gene order among distantly related plant species. *Plant Cell*, 2001, vol. 13, no. 4, p. 723-725.

FERGUSON, M.E.; BUROW, M.D.; SCHULZE, S.R.; BRAMEL, P.J.; PATERSON, A.H.; KRESOVICH, S. and MITCHELL, S. Microsatellite identification and

characterization in peanut (A. hypogaea L.). *Theoretical and Applied Genetics*, 2004, vol. 108, no. 6, p.1064-1070.

GOLDMAN, N. and YANG, Z.H. A codon based model of nucleotide substitution for protein coding DNA sequences. *Molecular Biology and Evolution*, 1994, vol. 11, no. 5, p. 725-736.

GRANT, D.; CREGAN, P. and SHOEMAKER, R. C. Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proceedings of the National Academy of Sciences*, USA, 2002, vol. 97, no. 8, p. 4168-4173.

GRANTHAM, R.; GAUTIER, C.; GUOY, M.; JACOBZONE, M. and MERCIER, R. Codon catalogue usage is a genome strategy for genome expressivity. *Nucleic Acids Research*, 1981, vol. 9, no. 1, p. r43-r75.

GRIESHAMMER, U. and WYNNE, J.C. Isozyme variability in mature seeds of U.S. peanut cultivars and collections. *Peanut Science*, 1990, vol. 17, no. 2, p. 72-75.

HALWARD, T; STALKER, H.T. and KOCHERT, G. Development of an RFLP linkage map in peanut species. *Theoretical and Applied Genetics*, 1993, vol. 87, no. 3, p. 379-384.

HALWARD, T.; STALKER, T.; LARUE, E. and KOCHERT, G. Use of single-primer DNA amplifications in genetic studies of peanut (Arachis hypogaea L.). *Plant Molecular Biology*, 1992, vol. 18, no. 2, p. 315-325.

HALWARD, T.M.; STALKER, H.T.; LARUE, E.A. and KOCHERT. G. Genetic variation detectable with molecular markers among unadapted germ-plasm resources of cultivated peanut and related wild species. *Genome*, 1991, vol. 34, no. 6, p. 1013-1020.

HESLOP-HARRISON, J.S. Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. *Plant Cell*, 2000, vol.12, no. 5, p. 617-636.

HUANG, X. and MADAN, A. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 1999, vol. 9, no. 9, p. 868-877.

KOCHERT, G.; HALWARD, T.; BRANCH, W.D. and SIMPSON, C.E. RFLP variability in peanut (Arachis hypogaea L.) cultivars and wild species. *Theoretical and Applied Genetics*, 1991, vol. 81, no. 5, p. 565-570.

KOCHERT, G.; STALKER, H. T.; GIMENES, M.; GALGARO, L.; LOPES, C. R.; and MOORE K. RFLP and cytological evidence on the origin and evolution of allotetraploid domesticated peanut, Arachis hypogaea (Leguminosae). *American Journal of Botany*, 1996, vol. 83, no. 10, p. 1282-1291.

KOLKER, E.; MAKAROVA, K.S.; SHABALINA. S.; PICONE, A.F.; PURVINE, S.; HOLZMAN, T.; CHERNY, T.; ARMBRUSTER, D.; MUNSON, R.S. Jr; KOLESOV, G.; FRISHMAN, D. and GALPERIN, M.Y. Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. *Nucleic Acids Research*, 2004, vol. 32, no. 8, p. 2353-2361.

LACKS. G. and STALKER, H. Isozyme analyses of Arachis species and interspecific hybrids. *Peanut Science*, 1993, vol. 20, p. 76-81.

LANHAM, P.; FORSTER, B.; MCNICOL, P.; MOSS, J. and POWELL, W. Seed storage protein variation in Arachis species. *Genome*, 1994, vol. 37, no. 3, p. 487-496.

LONG, M.Y. and GILLESPIE, J.H. Codon usage divergence of homologous vertebrate genes and codon usage clock. *Journal of Molecular Evolution*, 1991, vol. 32, p. 6-15.

NAKAMURA, Y.; GOJOBORI, T. and IKEMURA, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 2000, vol. 28, no. 1, p. 292.

NESTI, C.; POLI, G.; CHICCA, M.; AMBROSINO, P. and SCAPOLI, C. and BARRAI, I. Phylogeny inferred from codon usage pattern in 31 organisms. Computer *Applications for the Biosciences*, 1995, vol. 11, no. 2, p. 167-171.

PATERSON, A.H.; BOWERS, J.E.; BUROW, M.D.; DRAYE, X.; ELSIK, C.G.; JIANG, C.-X.; KATSAR, C.S.; LAN, T.-H.; LIN, Y.-R.; MING, R. and WRIGHT, R.J. Comparative genomics of plant chromosomes. *Plant Cell*, 2000, vol. 12, no. 9, p. 1523-1540.

PAIK-RO, O.G.; SMITH, R.L. and KNAUFT, D.A. Restriction fragment length polymorphism evaluation of six peanut species within the Arachis section. *Theoretical and Applied Genetics*, 1992, vol. 84, no. 1-2, p. 201-208.

QUACKENBUSH, J.; CHO, J.; LEE, D.; LIANG, F.; HOLT, I.; KARAMYCHEVA, S.; PARVIZI, B.; PERTEA, G.; SULTANA, R. and WHITE, J. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, 2001, vol. 29, no. 1, p. 159-164.

QUACKENBUSH, J.; LIANG, F.; HOLT, I.; PERTEA, G. and UPTON, J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, 2000, vol. 28 no.1, p. 141-145.

SAVOY, C. Peanut (Arachis hypogaea L.). Seed protein characterization and genotype sample classification using polyacrylamide gel electrophoresis. *Biochemistry Biophysics Research Communication*, 1976, vol. 68, no. 3, p. 886-893.

TOMBS, M.P. Variant forms of Arachis. *Nature*, 1963, vol. 200, p. 1321-1322.

ZHU, H.; KIM, D-J.; BAEK, J-M.; CHOI, H-K.; ELLIS, L.C.; KUESTER, H.; MCCOMBIE, W.R.; PENG, H-M. and COOK, D.R. Syntenic relationships between Medicago truncatula and Arabidopsis reveal extensive divergence of genome organization. *Plant Physiology*, 2003, vol. 131, no. 3, p. 1018-1026.