Plant Breeding
doi:10.1111/j.1439-0523.2009.01638.x

High level of natural variation in a groundnut (*Arachis hypogaea* L.) germplasm collection assayed by selected informative SSR markers

R. K. Varshney[1], T. Mahendar[1,2], R. Aruna[1], S. N. Nigam[1], K. Neelima[1,3], V. Vadez[1] and D. A. Hoisington[1]

[1]Centre of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Hyderabad 502324, Andhra Pradesh, India, E-mail: r.k.varshney@cgiar.org; [2]Department of Plant Sciences, University of Hyderabad, Hyderabad 500046, Andhra Pradesh, India; [3]Department of Biotechnology, Jawaharlal Nehru Technological University (JNTU), Hyderabad 500072, Andhra Pradesh, India

With 1 figure and 4 tables

Received September 14, 2008/Accepted January 5, 2009
Communicated by S. Abbo

Abstract

The ability to identify genetic variation is indispensable for effective management and use of genetic resources in crop breeding. Genetic variation among 189 groundnut (*Arachis hypogaea* L.) accessions comprising landraces, cultivars, a mutant, advanced breeding lines and others (unknown genetic background) representing 29 countries and 10 geographical regions was assessed at 25 microsatellite or simple sequence repeat loci. A high number of alleles (265) were detected in the range of 3 (Ah1TC6G09) to 20 (Ah1TC11H06) with an average of 10.6 alleles per locus. The polymorphism information content value at these loci varied from 0.38 (Ah1TC6G09) to 0.88 (Ah1TC11H06) with an average of 0.70. A total of 59 unique alleles and 127 rare alleles were detected at almost all the loci assayed. Cluster analysis grouped 189 accessions into four clusters. In general, genotypes of South America and South Asia showed high level of diversity. Extraordinary level of natural genetic variation reported here provides opportunities to the groundnut community to make better decisions and define suitable strategies for harnessing the genetic variation in groundnut breeding.

Key words: heterozygosity — molecular markers — microsatellite — peanut — PIC value — stepwise mutation model

Cultivated groundnut (*Arachis hypogaea* L.) is cleistogamous, annual allotetraploid (AABB genome, 2n = 4*x* = 40), tropical grain legume with a genome size of 3×10^9 bp, while most of its wild relatives are diploid (AA or BB genomes, 2n = 2*x* = 20) (Krapovickas and Gregory 1994). Cultivated groundnut has two subspecies *hypogaea* and *fastigiata*, which, in turn, have two (*hypogaea* and *hirsuta*) and four (*fastigiata*, *vulgaris*, *aequatoriana* and *peruviana*) botanical varieties, respectively. It is the 13th most important food crop and fourth most important oilseed crop of the world. It is grown extensively in Africa, Asia and the Americas with a total annual global area of nearly 25.2 million ha with a total production of 35.9 million metric tons (FAOSTAT 2006).

The seeds of groundnut apart from being rich source of oil (44–55%), protein (20–50%) and carbohydrate (10–20%) are also nutritional source of vitamin E, niacin, falacin, calcium, phosphorus, magnesium, zinc, iron, riboflavin, thiamine and potassium. In addition to drought stress, foliar diseases such as early leaf spot (ELS) (*Cercospora arachidicola*), late leaf spot (LLS) (*Phaeoisariopsis personata*) and rust (*Puccinia arachidis*) are generally considered the major constraints for groundnut production especially in semi-arid tropic environments. In Africa, groundnut rosette disease is also an important production constraint. Because of these reasons, the groundnut yield in semi-arid regions is low, averaging about 800 kg per ha which is less than one-third of the potential yield of 3000 kg per ha (Krapovickas 1968). It is important to introduce genetic variability in breeding programmes for resistance/tolerance to biotic/abiotic stresses as well as for yield.

Evidences from molecular genetic variation studies on cultivated and wild groundnut species suggest that cultivated groundnut has originated from a single hybridization event of the wild species *Arachis duranensis* Krapovickas & W. C. Gregory (AA genome) and *Arachis ipaënsis* Krapovickas & W. C. Gregory (BB genome) followed by chromosome duplication (Kochert et al. 1996). Abundant germplasm resources of both AA and BB genome (wild) species are available to groundnut breeders, although there is only one available accession of *A. ipaënsis*. Because of difficulties in crossing of diploid and tetraploid species, most groundnut breeding programmes have traditionally relied on crossing of elite breeding lines for developing improved cultivars (Knauft and Gorbet 1989). As a result, the genetic base of tetraploid (domesticated) groundnut has been relatively narrow. Nevertheless based on variation for morpho-agronomic traits and geographical locations and knowledge of breeding values over the last 32 years, a germplasm collection comprising mainly tetraploid groundnut lines has become available with groundnut breeding team at ICRISAT. Although this collection has been very useful for introducing genetic variability in breeding programme at ICRISAT, no information, however, is available on genetic variation at molecular level.

Molecular marker technologies are playing an increasingly important role in conservation and use of plant genetic resources in plant breeding programmes (see Varshney et al. 2007). In case of groundnut, different kinds of molecular markers including restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) were used in the past to assess the diversity and understand the relationships in various groundnut germplasm collections (He and Prakash 1997, Galgaro et al. 1998, Subramanian et al. 2000, Dwivedi et al. 2001, 2003, Herselman 2003). Majority of these

studies, however, revealed low level of DNA polymorphism especially in cultivated germplasm collection. This may be attributed to the low level of genetic variation that existed in the germplasm collection, due to the origin of groundnut, or poor discriminatory power of marker systems such as RFLPs, RAPDs and AFLPs.

Microsatellite or simple sequence repeat (SSR) markers however have shown higher information content and because of some other features such as high reproducibility and co-dominant nature, these markers have been considered as the markers of choice in crop breeding (Gupta and Varshney 2000). In recent years, as a result of considerable efforts of several research groups at international level, several hundred SSR markers have become available in groundnut (for a review see Varshney et al. 2007). These SSR markers have been found very useful to detect genetic diversity in groundnut germplasm including cultivated/tetraploid genotypes (Mace et al. 2006, Tang et al. 2007, Cuc et al. 2008).

In order to understand the genetic variation in the germplasm collection comprising of 189 accessions, as mentioned above and to harness its potential in groundnut breeding programme in systematic and efficient manner, the present study surveys the genetic variation at molecular level by using a set of 21 highly polymorphic and informative SSR markers. The study indicates the availability of large genetic variation in the germplasm collection which can be helpful to select diverse parental lines at DNA level for developing populations for mapping and understanding the genetics of the traits of interest to groundnut community.

Materials and Methods

Plant material: A total of 185 groundnut accessions comprising landraces (LR; 31), cultivars (C; 29), a mutant line (MT; 1), advanced breeding lines (ABL; 65) and others (OTH; 59) that represented 29 countries and 10 geographical regions were analysed. Four genotypes ('ICG 156', 'ICG 2738', 'ICG 13941' and 'ICG 13942') for which allele sizes for all the markers examined were known from a separate study (Upadhyaya et al. 2006) were used along with 185 accessions to ensure quality in the genotyping data. Details on country of origin, geographical regions and biological status of different groundnut accessions are given in ESM1.

DNA isolation: A single seed from each selected groundnut accessions was sown in green house and leaf tissues were harvested from 10- to 15-day-old seedlings for DNA isolation. A high-throughput DNA isolation protocol, as mentioned in Cuc et al. (2008) was adopted to isolate DNA from the leaf tissues. DNA quantification and quality check were done on 0.8% agarose gel.

Polymerase chain reaction: Twenty one most informative and polymorphic SSR markers, also used in a separate study (Upadhyaya et al. 2006), were used for analysing molecular diversity. These SSR markers represent different linkage groups of groundnut and exhibit high polymorphism information content (PIC) value. These markers have come from two sources; the primer sequence information for pPGPSeq and Ah1 series markers are available in Ferguson et al. (2004) and Moretzsohn et al. (2005), respectively. The forward primer of these markers was labelled with one of four fluorescence dyes: 6-FAM, VIC, NED, PET (Applied Biosystems, Foster City, CA, USA).

Polymerase chain reactions were performed in 10 μl reaction volume in an ABI 9700 thermal cycler (Applied Biosystems) using 384-well PCR plates (Applied Biosystems). Reaction volume consisted of 2 pmol of primer, 10 mM $MgCl_2$, 2 mM dNTPs, 0.1 U of *Taq* DNA polymerase (Qiagen, Hilden, Germany), 1× PCR buffer (Qiagen) and 5 ng DNA template. A touch down PCR amplification profile with 3 min of initial denaturation cycle, followed by first five cycles of 94°C for 20 s, 60°C for 20 s and 72°C for 30 s, with 1°C decrease in temperature per cycle, then 30 cycles of 94°C for 20 s with constant annealing temperature (56°C) for 20 s and 72°C for 30 s, followed by a final extension for 20 min at 72°C. The amplified products were tested on 1.2% agarose gel to check for the amplification.

SSR fragment analysis: After confirming the PCR amplification on 1.2% agarose gel, five post-PCR multiplex sets were constructed based on the allele size range estimates and the type of forward primer label of the markers. Markers that had different labels and allele size ranges were considered for a set. However, markers with the same label separated by more than 50 bp were also considered for a set to accommodate 21 markers in five mutiplexes. For post-PCR multiplexing, 1.5 μl PCR product of each of 6-FAM, VIC, NED and PET-labelled products were pooled (according to above mentioned criteria) and mixed with 7 μl of Hi-Di formamide (Applied Biosystems), 0.25 μl of the LIZ-500 size standard (Applied Biosystems) and 1.5 μl of distilled water. The pooled PCR amplicons were denatured and size fractioned using capillary electrophoresis on an ABI 3700 automatic DNA sequencer (Applied Biosystems). Allele sizing of the electrophoretic data thus obtained was done using GENSCAN 3.1 software (Applied Biosystems) and GENOTYPER 3.1.

Data analysis: The allelic data obtained on 189 groundnut accessions for all 21 SSR markers were binned using AlleloBin programme (that automates the process of assigning allele sizes into their appropriate allele 'bins' and developed at ICRISAT). Major allele frequency, gene diversity, unique alleles, rare alleles, shared allele frequencies (SAF) and PIC values for all 25 loci were computed using PowerMarker V3.25 (Liu and Muse 2005). The PIC values were based on Botstein et al. (1980):

$$\mathrm{PIC} = 1 - \left[\sum_{i=1}^{n} P_i^2\right] - \left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 2P_i^2P_j^2\right],$$

where P_i and P_j are the frequencies of ith and jth allele.

Allele frequencies and allele sizes obtained for all SSR markers were subjected for the strict one-step stepwise mutation model (SMM; Ohta and Kimura 1973), the infinite allele model (IAM; Kimura and Crow 1964) and two phase mutation model (Di Rienzo et al. 1994) for microsatellite allele distribution in the population.

Pair-wise dissimilarities among all 189 accessions were computed with simple matching coefficient and the dissimilarity matrix thus generated was further, used to construct a neighbour-joining (NJ) tree using DARwin 5.0.128 (Perrier et al. 2003). ARLEQUIN ver 3.01 was used to compute the analysis of molecular variance (AMOVA) among and within germplasm of different geographical regions, biological status and botanical variety. Allelic richness was surveyed using FSTAT 2.9.3.2 (Goudet 2001).

Results

SSR polymorphism

Out of 21 SSR markers, 17 SSR markers produced one locus per marker, while four markers namely Ah1TC6E01, Ah1TC1A02, pPGPseq1B09 and pPGPseq15C12 yielded two different loci per marker. Thus in total, allelic data were obtained for a total of 25 loci amplified by 21 SSR markers. A total of 265 alleles were generated at 25 SSR loci in the germplasm analysed with a mean of 10.6 alleles per locus. However, the number of alleles per locus ranged from 3 (Ah1TC6G09) to 20 (Ah1TC11H06). The amplicon sizes across all the loci and genotypes ranged from 129 to 351 bp. PCR amplicons obtained in the four control genotypes ('ICG 156', 'ICG 2738', 'ICG 13941' and 'ICG 13942') for all the SSR markers

showed the same allele sizes across all the PCR reactions performed. In order to ensure the good quality data, low-height peaks (< 50% of the average peak across the germplasm) were excluded from the dataset. Eventually for each SSR locus, up to 10% missing data were recorded. Of 21 SSR markers used for genotyping the germplasm, 12 (57%) were di-nucleotide repeat markers, six (29%) tri-nucleotide and three were compound microsatellites. The major allele frequency ranged from 0.19 to 0.68 with an average of 0.37 (Table 1).

Allele features and diversity

The replication slippage mechanism is responsible for the hypervariability of SSRs that would lead to larger average allele sizes in 'derived' groups. The SSR loci, at which diversity was surveyed, could be classified into three main groups based on their frequencies in the germplasm analysed (ESM 2): (i) SSR loci (pPGPseq1B09a and Ah1TC3E02) following the SMM very strictly, (ii) SSR loci (pPGPseq17E03, Ah1TC6E01a, Ah1TC1A02a and pPGPseq1B09b) showing little deviation but in agreement with SMM model, and (iii) SSR loci (the remaining 19 loci) following IAM.

Unique alleles

Unique alleles are genotype and marker specific alleles, i.e. allele produced by a marker and present only in one genotype and absent in all other analysed germplasm. Of 25 SSR loci analysed, 23 (92%) loci detected at least one unique allele. In total, 59 unique alleles were detected at 23 SSR loci across all 189 groundnut germplasm lines analysed (Table 1). Majority of such unique alleles (25, i.e. 42.4%) were present in the Asian germplasm. The number of unique alleles observed in breeding material, cultivars, landraces and others (genotypes of unknown biological status) were 18, 15, 14 and 12, respectively. Among 14 landraces that possessed unique alleles, four were of African origin.

The microsatellite loci pPGPseq18C05, pPGPseq2D12B, Ah1TC11A04 and Ah1TC6E01 detected one unique allele each in 'ICG 1675', a South American landrace. Similarly, pPGPseq5D05 and pPGPseq8E03 detected one unique allele each in 'ICG 461', a South American landrace. The highest number of unique alleles (8 of 59 alleles) was detected at SSR locus pPGPseq1B09b. Interestingly seven of the eight unique alleles detected at this locus occurred in Southeast Asian germplasm ('ICG 11337', 'TMV 2', 'ICG 13941', 'ICG 13942', 'ICG 3933', 'ICG 4240' and 'ICG 2738'). Similarly, the SSR locus pPGPseq15C12a could detect one unique allele each in 'ICG 30363', 'ICR 48', 'ICG 6565', 'ICG 3520' and 'ICG 2738'. One unique allele each in 'ICG 3106' (unknown origin) and 'ICG 32' (Southeast Asian) lines was detected at SSR locus Ah1TC6E01b.

Rare alleles

The alleles with frequency < 0.05 in the total sample are considered as rare alleles. Rare alleles accounted for 51.69%, while most abundant (frequency > 0.30) and intermediate (0.05 < frequency < 0.3) alleles constituted 13.58% and 34.71%, respectively. A total of 127 (48%) rare alleles were observed at 25 SSR loci across all the germplasm examined. The number of rare alleles detected at each locus is presented in Table 1. The number of rare alleles ranged from 1 (pPGPseq17E03, Ah1TC6G09, Ah1TC6H03, Ah1TC3E02) to 12 (Ah1TC11H06) (Table 1). No significant difference has been observed in the number of rare alleles detected based on the biological status of the germplasm. For instance, 31, 32

Table 1: Molecular diversity estimates in the germplasm collection based on 25 SSR loci

	Marker ID[1]	Marker locus ID	Repeat unit	Allele size range (bp)	No. alleles	Major allele frequency	PIC value	No. unique alleles	No. rare alleles
1	pPGPSeq13E09	pPGPSeq13E09	$(TAA)_{16}$	211–238	10	0.32	0.80	1	3
2	pPGPSeq15C12	pPGPSeq15C12a	$(TAA)_{28}$	195–291	10	0.32	0.75	2	5
3		pPGPSeq15C12b		255–291	11	0.34	0.78	1	4
4	pPGPSeq17E03	pPGPSeq17E3	$(CTT)_{15}$	175–196	4	0.44	0.55	1	1
5	pPGPSeq18C05	pPGPSeq18C5	$(TAA)_{23}$	266–314	12	0.27	0.77	5	7
6	pPGPSeq19B01	pPGPSeq19B01	$(GA)_{15}$	161–239	18	0.24	0.87	4	10
7	pPGPSeq1B09	pPGPSeq1B09a	$(GA)_{19}$	238–268	10	0.47	0.58	5	7
8		pPGPSeq1B09b		244–284	12	0.68	0.48	8	8
9	pPGPSeq2D12B	pPGPSeq2D12B	$(TAA)_{16}$	253–310	16	0.35	0.76	4	9
10	pPGPSeq 5D05	pPGPSeq5D05	$(GA)_{32}$	226–274	14	0.32	0.77	3	9
11	pPGPSeq7H06	pPGPSeq7H06	$(CTT)_{12}$	228–297	6	0.54	0.52	1	2
12	pPGPSeq8E12	pPGPSeq8E12	$(TTG)_6(TAA)_{15}$	192–294	13	0.33	0.79	5	6
13	Ah1TC11A04	Ah1TC11A04	$(CT)_{16}(CT)_{33}$	167–179	6	0.40	0.64	1	2
14	Ah1TC11H06	Ah1TC11H06	$(AG)_{34}$	174–224	20	0.20	0.88	2	12
15	Ah1TC1A02	Ah1TC1A02a	$(TC)_{35}$	211–229	9	0.38	0.70	2	5
16		Ah1TC1A02b		221–251	7	0.30	0.77	2	6
17	Ah1TC1E01	Ah1TC1E01	$(GA)_{29}$	203–229	11	0.54	0.63	1	6
18	Ah1TC3E02	Ah1TC3E02	$(CT)_{26}(CA)_7(CA)_5$	251–257	4	0.59	0.40	–	1
19	Ah1TC4F12	Ah1TC4F12	$(CT)_{23}$	211–239	14	0.19	0.87	1	6
20	Ah1TC6E01	Ah1TC6E01a	$(GA)_{22}$	158–166	5	0.44	0.53	1	2
21		Ah1TC6E01b		167–189	12	0.24	0.82	2	5
22	Ah1TC6G09	Ah1TC6G09	$(CT)_{18}$	129–135	3	0.57	0.38	–	1
23	Ah1TC6H03	Ah1TC6H03	$(AG)_{21}$	203–233	8	0.38	0.71	1	1
24	Ah1TC7H11	Ah1TC7H11	$(AG)_{18}$	319–351	17	0.20	0.85	5	6
25	Ah1TC9F10	Ah1TC9F10	$(AG)_{31}$	248–272	13	0.27	0.84	1	3
		Mean			10.60	0.37	0.70	2.56	5.08

[1]Sequence information for the markers 1–12 is available in Ferguson et al. (2004) while 13–25 in Moretzsohn et al. (2005).

and 20 rare alleles were detected in cases of cultivars, advanced breeding lines and landraces, respectively, while 44 rare alleles were observed in the accessions having no information on biological status.

Shared allele frequency

The SAF of each genotype with all other germplasm analysed are presented in ESM 3. The range of SAF of germplasm from Southeast Asia was low (38.10–66.67%). Interestingly, North American germplasm were found to show a broad range of SAF (38.1–90.48%) with the other germplasm collections analysed. A broad range of SAF were observed in germplasm from different geographical regions e.g. Southeast Asia (38.10–95.24%), South America (42.86–80.95%), Eastern Africa (47.62–95.24%), Southern Africa (47.62–85.71%) and West Africa (52.38–95.24%), respectively.

Number of alleles and allelic richness

Two allelic diversity parameters namely number of alleles and allelic richness were analysed in the germplasm collection as per geographical region, biological status and botanical type (Table 2). In terms of geographical regions, average number of alleles ranged from 3.76 (West Africa) to 8.84 (South Asia) while overall allelic richness varied from 0.95 (West Africa) to 3.59 (South Asia). Similarly, in terms of biological status, average number of alleles were present in the range of 6.76 (LR) to 7.68 (ABL) and allelic richness varied from 2.79 (C) to 3.59 (ABL). As per botanical types, the genotypes representing Virginia runner type had the lowest number of alleles (4.40) as well as allelic richness (1.83) while the Spanish bunch germplasm showed the highest number of alleles (9.28) as well as allelic richness (4.05). A significant correlation was observed between allelic richness and average number of alleles in all the classes i.e. geographical regions ($r^2 = 0.92$; $P < 0.05$), biological status ($r^2 = 0.53$; $P < 0.05$) and botanical types ($r^2 = 0.99$; $P < 0.05$).

Polymorphism information content

The PIC values over the 25 SSR marker loci ranged between 0.38 (Ah1TC6G09) to 0.88 (Ah1TC11H06) with an average of 0.70 (Table 1). About 88% SSR loci (22) revealed PIC values >0.5 and remaining three SSR loci (pPGPseq1B09a, Ah1TC3E02 and Ah1TC6G09) have shown PIC values <0.50. An attempt was made to understand the relationship of number of repeat units and number of alleles with the PIC values of SSR loci assayed (data not shown). The di-nucleotide repeats, in general, exhibited higher PIC values than the tri- and compound microsatellite types, as evident from Table 1. However, no specific relationship was observed with either number of repeat units or allele numbers with PIC values.

Molecular variance among the germplasm

Analysis of molecular variance was performed on the dataset in order to partition the total genetic variation within and between three parameters: (i) within and between geographical regions, (ii) within and between botanical varieties, and (iii) within and between biological status of the germplasm. The AMOVA revealed that only 1.5% of the total variation observed was accounted for between geographical regions, whereas the majority of variation (98.5%) was observed among individuals within each geographical region (Table 3). Similarly, upon partitioning the total genetic variation between and within the botanical varieties (Table 3), very negligible amount (0.1%) of the total variation was accounted for between different botanical varieties. Finally, 0.28% of the variation was accounted for between five different biological status of the genotypes included in this study. In total, 99.9% and 99.72% genetic variation was accounted for differences among individuals within each botanical type and biological status groups, respectively.

Table 2: Allelic diversity in different groups of germplasm collection as per geographical regions, botanical types and biological status

	Average number of alleles	Allelic richness	Correlation coefficient between average allele number and allele richness (r^2*)
Geographical region			
North America	5.76	2.23	0.92
South America	4.56	1.74	
Eastern Africa	4.40	2.13	
Southern Africa	4.08	1.54	
West Africa	3.76	0.95	
East Asia	4.40	1.81	
South Asia	8.84	3.59	
Southeast Asia	3.84	1.28	
Others	5.52	2.72	
Biological status			
Landraces	6.76	3.22	0.53
Cultivars	7.12	2.79	
Advanced breeding lines	7.68	3.59	
Others	7.44	3.20	
Botanical type			
Spanish bunch	9.28	4.05	0.99
Virginia bunch	7.16	3.13	
Virginia runner	4.40	1.83	
Valencia	6.40	2.46	
Others	3.84	1.37	

*Significant at $P = 0.05$.

Table 3: Analysis of molecular variance (AMOVA) for 189 genotypes analysed as three groups based on geographical regions, botanical type and biological status

Source of variation	df	Sum of squares	Variance components	% Variation
Geographical regions				
Among geographical regions	8	164.87	0.122 Va	1.49
Among individuals within geographical regions	181	2932.07	8.099 Vb	98.51
Biological status				
Among biological status groups	4	60.14	0.22 Va	0.28
Among individuals within biological status groups	184	3015.80	8.19 Vb	99.72
Botanical types				
Among botanical type groups	3	51.33	0.00792 Va	0.10
Among individuals within botanical types	185	3031.95	8.19 Vb	99.90

df, degrees of freedom.

Genetic relationships among genotypes

Based on genotyping data for 265 alleles obtained at 25 SSR loci, a NJ tree was constructed using DARwin5. The dendrogram classified the germplasm into four major clusters (Cl I, Cl II, Cl III and Cl IV) (Fig. 1; High resolution image is available at http://www.icrisat.org/gt-bt/PBR1.htm). The cluster I (Cl I), being major cluster with 93 genotypes was further divided into three subclusters, i.e. Cl Ia (40 genotypes), Cl Ib (25 genotypes) and Cl Ic (28 genotypes) (Table 4). The genotypic composition of each cluster is presented in the ESM 4. While comparing the clusters or subclusters within different groups of germplasm, no specific/significant trend of grouping of genotypes was observed. However, a loose clustering of genotypes as per their geographical origin and botanical type was observed. For example, germplasm from South America (83.3%), West Africa (71.4%), East Asia (45.4%), South Asia (37.4%), Southeast Asia (57.1%), Southern Africa (50%) and North America (41.2%) were grouped in Cl IIb, Cl Ia, Cl III, Cl Ia, Cl III, Cl Ia and Cl III, respectively (Fig. 1). For instance, majority of genotypes coming from North America (12 of 20, 60%), Central America (13 of 14, 92.8%) and South Asia (47.4%) were grouped in the cluster Cl I. Similarly >30% landraces (11 of 31) and cultivars (13 of 34) were grouped in the cluster Cl III. In terms of botanical types, 50% of Spanish bunch type (51 of 102) were grouped in the cluster Cl I. Interestingly, about 30% of all the botanical types, i.e. Spanish bunch (30 of 102), Virginia bunch (14 of 42), Virginia runner (three of nine) and Valencia (11 of 29) were grouped in the cluster III.

Discussion

Features and trends in SSR diversity

In total, 25 SSR loci were obtained by using 21 SSR markers that can be attributed to tetraploid nature of groundnut (Krapovickas and Gregory 1994, Cuc et al. 2008). Amplification of homoeoloci is a common feature in allopolyploid species like common wheat (Varshney et al. 2000), brassica (Parkin et al. 2003). However, these homoeoloci can be visualized only when they vary in the amplicon length.

All 25 microsatellite loci were found polymorphic in the germplasm examined. Earlier diversity studies in tetraploid groundnut germplasm based on RFLPs (Halward et al. 1991), RAPDs (Bhagwat et al. 1997), DAFs (He and Prakash 1997) and AFLPs (He and Prakash 1997) showed no or very low level of genetic polymorphism. However, recent SSR marker-based genetic diversity studies revealed relatively moderate level of polymorphism in cultivated groundnuts (Gimenes et al. 2007, Mace et al. 2006, 2007, Kottapalli et al. 2007, Cuc et al. 2008, Varshney et al. 2008). As compared to these studies, higher number of alleles (average 10.64 alleles per locus) and a higher PIC values were observed at SSR loci assayed in the present study. This may be attributed to use of a larger number of accessions (189) as earlier mentioned studies employed lower number of genotypes (generally <80). Two additional factors may be attributed to high level of diversity recorded in the present study: (i) SSR markers used in the present study are a selected set of highly informative and polymorphic SSR markers, used for genotyping a global composite collection of groundnut (Upadhyaya et al. 2006); and (ii) genotypes present in the germplasm collection are highly diverse as they have come from different geographical regions, represent different botanical types as well as different biological status.

Apart from allele features, we investigated whether the polymorphism of SSR loci could be affected by any other factors including different repeat units, SSR type, repeat numbers and total SSR length. The degree of polymorphism observed in microsatellite is usually correlated with the length of the repeat unit (Gupta and Varshney 2000). However, no such correlation was observed between: (i) PIC values and length of repeat unit ($r^2 = 0.15$, $P < 0.001$) and (ii) PIC value and number of alleles ($r^2 = 0.17$, $P < 0.001$) (data not shown). No correlation in PIC values with number of repeat units and allele numbers have been reported in several earlier studies. For instance, Ni et al. (2002) reported strong correlations for PIC values with number of repeat units and allele numbers in rice, however Stachel et al. (2000) and Prasad et al. (2000) did not observe such correlation in wheat. In case of groundnut also, some weak correlations have been observed by Ferguson et al. (2004) and Cuc et al. (2008), however, a strong correlation was reported by Moretzsohn et al. (2005).

Allele frequencies and informativeness

Out of 25 SSR loci analysed on the set of 189 groundnut genotypes, 23 loci detected 57 unique alleles and 25 loci detected 127 rare alleles. While the unique alleles observed in a particular genotype are useful for identification of the genotype, the rare alleles present in a particular germplasm pool provide the specificity of that germplasm pool. For instance, unique alleles identified in the Southeast Asian genotypes ('ICG 11337', 'TMV 2', 'ICG 13941', 'ICG 13942', 'ICG 156', 'ICG 3933', 'ICG 4240', 'ICG 2738' and 'ICG 156') and a North American genotype ('ICG 4240') at the SSR locus pPGPSeq1B09b suggests the utility of unique alleles of the

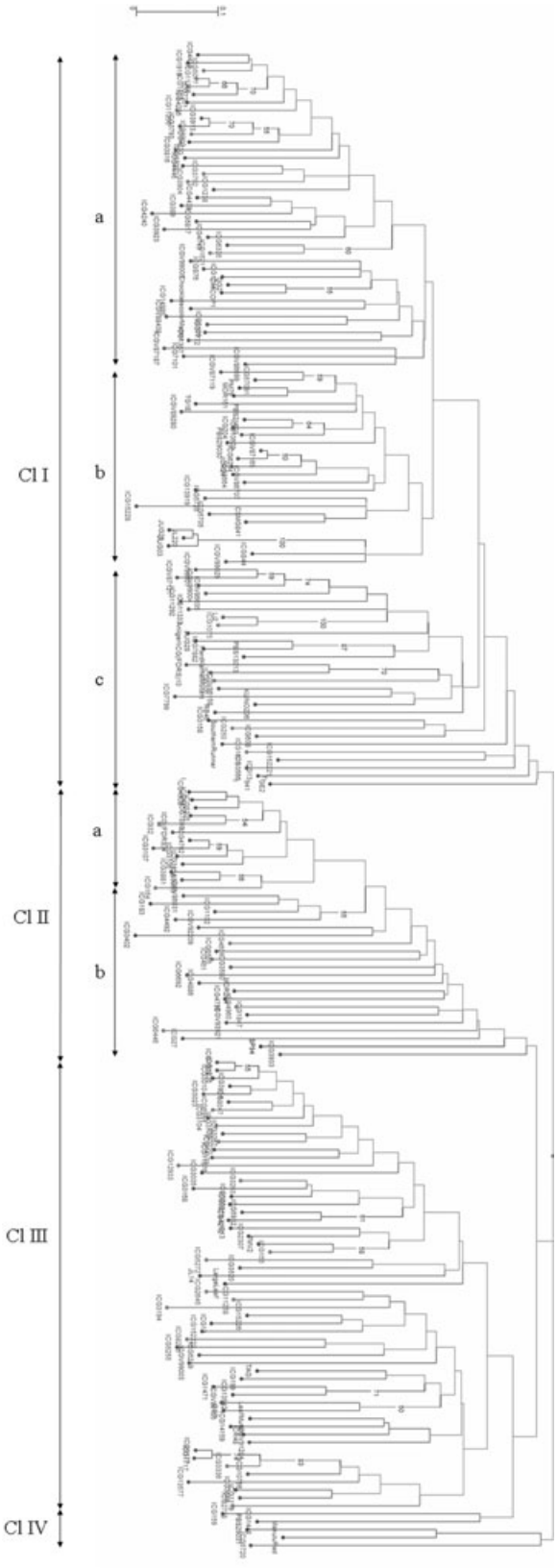

Fig. 1: Genetic relationships among 189 groundnut lines based on NJ tree constructed by using 286 alleles detected at 25 SSR loci with help of DARwin5 program. Details about different clusters are given in Table 4 and ESM 4. High resolution image is available at http://www.icrisat.org/gt-bt/PBR1.htm

pPGPSeq1B09 locus as diagnostic allele/marker to identify these germplasm lines. The highest percentage (54.45%) of unique alleles was observed in accessions from Southeast Asia. Interestingly 31.5% of unique alleles were present in advanced breeding lines, 26.5% in cultivars and 21% in landraces and others. Occurrence of unique alleles that are specific to single accessions was reported in several previous studies also (e.g. Li et al. 2002, Bantte and Prasanna 2003, Kottapalli et al. 2007). Occurrence of rare allele, however, is a feature of the germplasm pool in which they are observed; it also indicates the higher information content of the marker that detects the rare alleles. Indeed, the presence of many rare alleles may be explained by the relatively high rate of mutation at SSR loci (Henderson and Petes 1992). Occurrence of both the unique as well as rare alleles in a given genotype also indicates the diverse nature of the genotype. Sometimes these alleles may be diagnostic for particular regions of the genome specific to a particular trait. The genotypes carrying the unique as well as rare alleles may prove useful for introducing the diversity in the applied breeding programmes.

Directional evolution of SSR loci

As compared to other molecular markers like RFLP, RAPD and AFLP, SSRs are the fastest-evolving DNA sequences with high mutation rates, 10^{-2}–10^{-3} per locus per gamete per generation (Weber and Wong 1993). This is the reason that SSR markers as compared to other type of molecular markers have been found useful and superior for assessing the molecular diversity in groundnut. While analysing the SSR allele distribution in the germplasm, five SSR loci (two strictly and three with some deviation) followed SMM; remaining 92% SSR loci were found in accordance with IAM. Assuming that SSR loci of a particular motif share common mutational mechanism the SMM tries to better account for the actual mutational process that occurs at SSR loci. However, at majority of SSR loci, mutations appear to be single-steps, in some rare cases (about 1/30–1/50) mutations occur in two-steps with asymmetric probabilities. Therefore, the one-step symmetric SMM may not be appropriate for studying the population dynamics at all SSR loci. The IAM refers to a model where each new mutation is different i.e., there are an infinite number of states that an allele can mutate, hence each mutation is assumed to be unique. Violation of the SMM process at majority of SSR loci indicates that most variation at groundnut SSR loci may involve not only the number of repeat units, but also different kinds of interruptions within a tandem-repeat array (Garza et al. 1995, Estoup et al. 1995, Angers and Bernatchez 1997) as well as nucleotide substitutions and insertions/deletions (indels) in regions flanking the repeat motif (Girmaldi and Crouau-Roy 1997).

Genetic diversity and relationships

The tree dendrogram prepared based on the genetic diversity data showed grouping of the genotypes into four clusters and the landraces, advanced breeding lines and cultivars were found interspersed among all the four clusters. This indicates the broad genetic base and diverse nature of the germplasm collection examined. Further, relatively low level of SAF (average 11.2%), the high level of genetic dissimilarity (average 0.73) and high PIC value of SSR loci (average 0.82) also supports this hypothesis. AMOVA analysis indicated that

Table 4: Information on number of genotypes distributed among different clusters

	Cl Ia[1]	Cl Ib	Cl Ic	Cl IIa	Cl IIb	Cl III	Cl IV	Total
Geographical regions								
North America	4 (20.0)	3 (15.0)	5 (25.0)	–	2 (10)	6 (30.0)	–	20
South America	3 (25.0)	2 (16.7)	2 (16.7)	2 (16.7)	3 (25.0)	–	–	12
Central America	12 (85.8)	–	1 (7.1)	–	1 (7.1)	–	–	14
Eastern Africa	3 (20.0)	1 (6.7)	1 (6.7)	1 (6.7)	2 (13.4)	7 (53.3)	–	15
Southern Africa	–	–	1 (10.0)	2 (20.0)	2 (20.0)	4 (40.0)	1 (10.0)	10
West Africa	4 (50.0)	–	1 (12.5)	–	1 (12.5)	2 (25.0)	–	8
East Asia	3 (30.0)	1 (10.0)	1 (10.0)	–	–	5 (50.0)	–	10
South Asia	7 (9.0)	16 (20.5)	14 (17.9)	5 (6.4)	9 (11.5)	26 (33.3)	1 (1.3)	78
Southeast Asia	–	2 (20.0)	2 (20.0)	1 (10.0)	–	5 (50.0)	–	10
Middle East	–	–	–	–	–	–	1 (100.0)	1
Others/unknown	4 (36.4)	–	–	2 (18.2)	1 (9.0)	4 (36.4)	–	11
Biological status								
Landraces	8 (25.8)	3 (9.7)	2 (6.5)	3 (9.7)	4 (12.9)	11 (35.5)	–	31
Cultivars	6 (17.7)	5 (14.7)	6 (17.7)	1 (2.9)	1 (2.9)	13 (38.2)	2 (5.9)	34
Advanced breeding lines	15 (23.4)	8 (12.5)	10 (15.6)	5 (7.8)	10 (15.6)	16 (25.0)	–	64
Mutant	–	–	–	–	–	1 (100.0)	–	1
Others/unknown	11 (18.6)	9 (15.3)	10 (16.9)	4 (6.8)	6 (10.1)	18 (35.6)	1 (1.7)	59
Botanical type								
Spanish bunch	17 (16.7)	16 (15.7)	18 (17.6)	7 (6.9)	11 (10.7)	30 (29.4)	3 (2.9)	102
Virginia bunch	15 (35.7)	4 (9.5)	5 (11.9)	2 (4.8)	2 (4.7)	14 (33.3)	–	42
Virginia runner	2 (22.2)	1 (11.1)	1 (11.1)	1 (11.1)	1 (11.1)	3 (33.3)	–	9
Valencia	6 (20.7)	3 (10.3)	1 (3.4)	2 (6.9)	6 (20.6)	11 (37.9)	–	29
Others/unknown	–	1 (14.3)	3 (42.8)	1 (14.3)	1 (14.2)	1 (14.3)	–	7

[1]Per cent accessions in each cluster/sub-cluster are indicated in parenthesis.

majority (>95%) of genetic variation observed in germplasm is due to variation in individuals instead of from a specific group. In general, genotypes of South America and South Asia showed higher level of diversity as compared to other geographical regions. While the genotypes representing South America showed least SAF (61.1%), higher allelic richness was observed in South Asia germplasm that can be attributed to analysis of relatively larger number of genotypes from South Asia (83) as compared to South America (12).

In general, no significant grouping of genotypes as per geographical region was observed that further indicates higher level of genetic diversity in the germplasm. The cluster I comprising of majority of genotypes from North America, Central America and South Asia seems to be a diverse and important cluster, as large amount of genetic diversity in groundnut has been reported from accessions of the above mentioned geographical locations (see Varshney et al. 2007). It is also interesting to note that genotypes originated from South America and South Asia in general were found dispersed in different clusters what indicates their diverse nature.

In terms of classification of genotypes as per botanical types, clear demarcated groups containing genotypes of a particular botanical type could not be obtained as it was in case of study of Kottapalli et al. (2007). This can be explained by use of larger number of SSR markers (73) and less number of genotypes (112 accessions) used by Kottapalli et al. (2007) as compared to this study where 21 SSR markers have been used to examine diversity in 189 accessions. Distribution of Spanish bunch type accessions across the clusters showed the higher diversity in Spanish types as compared to other types, though genotypes used were more in number for Spanish types. The cluster Cl III seems to be quite diverse cluster in terms of botanical types as it contains about 30% accessions from all four botanical types. Likewise, the cluster Cl III contains about 30% of landraces and cultivars. It is also interesting to note that the genotypes 'JUG 28', 'JUG 03' and 'JUG 13' from Gujarat were found to be clustered together in the cluster Ia with 100 boot-strap values indicating high similarity among these accessions. All these three genotypes originated from the same cross 'ICGS 76' × 'CSMG 84-1'.

A low level of DNA polymorphism has been a major constraint, in the past, in developing genetic maps and undertaking molecular breeding in groundnut (e.g. Varshney et al. 2008). Further, non-availability of sources with high levels of resistance in cultivated gene pools of groundnut for several diseases, e.g. ELS, LLS and the difficulties of crossing cultivated species with wild species are other barriers that hampered the development of appropriate mapping populations in groundnut (Varshney et al. 2007). However, the present study revealed a high level of genetic variation present in the germplasm analysed. Percent dissimilarities between the pairs of accessions, in several cases, were very high and such pairs of accessions can be used in breeding programmes. For instance, 'ICG 164', an Indian landrace, resistant to ELS and groundnut rosette virus is 92% dissimilar with 'ICG 1102' (breeding material from India) and 'ICG 1123' (a cultivar from China), both susceptible to ELS, and can be used as potential parents in developing cultivars resistant to ELS. Similarly, 'ICG 27' an Indian landrace showed high dissimilarity (96%) with 'CG 7' (ICRISAT-bred cultivar).

Thus, the information generated from this study may be used to select diverse lines (at molecular level also) for creating segregating mapping populations that will be useful to develop the dense molecular genetic maps as well as quantitative trait locus (QTL) analysis to understand the genetic basis of complex traits. Such molecular diversity studies in long term will allow mapping of genes and QTLs for marker-assisted introgression of the traits into elite breeding lines so that the potential of molecular or genomics-assisted breeding could be realized in groundnut (Varshney et al. 2005).

Acknowledgements

The present study was carried out by financial support from Generation Challenge Programme (http://www.generationcp.org) and National Fund for Basic and Strategic Research in Agriculture of Indian Council of Agricultural Research.

References

Angers, B., and L. Bernatchez, 1997: Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Mol. Biol. Evol. **14,** 230—238.

Bantte, K., and B. M. Prasanna, 2003: Simple sequence repeat polymorphism in quality protein maize (QPM) lines. Euphytica **129,** 337—344.

Bhagwat, A., T. G. Krishna, and C. R. Bhatia, 1997: RAPD analysis of induced mutants of groundnut (*Arachis hypogaea* L.). J. Genet. **76,** 1—8.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis, 1980: Construction of genetic linkage maps in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32,** 314—331.

Cuc, L. M., E. S. Mace, J. H. Crouch, V. D. Quang, T. D. Long, and R. K. Varshney, 2008: Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (*Arachis hypogaea*). BMC Plant Biol. **8,** 55.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer, 1994: Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl Acad. Sci. USA **91,** 3166—3170.

Dwivedi, S. L., S. Gurtu, S. Chandra, W. Yuejin, and S. N. Nigam, 2001: Assessment of genetic diversity among selected groundnut germplasm - I: RAPD analysis. Plant Breeding **120,** 345—349.

Dwivedi, S. L., J. H. Crouch, S. N. Nigam, M. E. Ferguson, and A. H. Paterson, 2003: Molecular breeding of groundnut for enhanced productivity and food security in the semi-arid tropics: opportunities and challenges. Adv. Agron. **80,** 154—222.

Estoup, A., L. Garnery, M. Solingnac, and J. M. Cournuet, 1995: Microsatellite variation in honey *Beeapis mellifera* L. populations; hirarchial genetic structure and test of the infinite allele and stepwise mutation models. Genetics **140,** 679—695.

FAOSTAT, 2006: FAO Production Yearbook, Vol. 60. FAOSTAT, Rome, Italy.

Ferguson, M. E., M. D. Burow, S. R. Schultze, P. J. Bramel, A. H. Paterson, S. Kresovich, and S. Mitchell, 2004: Microsatellite identification and characterization in peanut (*A. hypogaea* L.). Theor. Appl. Genet. **108,** 1064—1070.

Galgaro, L., C. R. Lopes, M. Gimenes, J. F. M. Valls, and G. Kochert, 1998: Genetic variation between several species of sections Extranervosae, Caulorrhizae, Heteranthae and Triseminatae (genus Arachis) estimated by DNA polymorphism. Genome **41,** 445—454.

Garza, J. C., M. Slatkin, and N. B. Freimer, 1995: Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size. Mol. Biol. Evol. **12,** 594—630.

Gimenes, M. A., A. A. Hosino, A. V. G. Barbosa, D. A. Palmieri, and C. R. Lopes, 2007: Characterization and transferability of microsatellite markers of cultivated peanut (*Arachis hypogaea*). BMC Plant Biol. **7,** 9.

Girmaldi, M. C., and B. Crouau-Roy, 1997: Microsatellite allelic homoplasy due to variable flanking sequences. J. Mol. Evol. **44,** 336—340.

Goudet, J., 2001. Available at: http://www2.unil.ch/popgen/softwares/fstat.htm (last accessed 19 February 2008).

Gupta, P. K., and R. K. Varshney, 2000: The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica **113,** 163—185.

Halward, T. M., H. T. Stalker, E. A. Larue, and G. Kochert, 1991: Genetic variation detectable with molecular markers among unadapted germplasm resources of cultivated peanut and related wild species. Genome **34,** 1013—1020.

He, G., and C. S. Prakash, 1997: Identification of polymorphic DNA markers in cultivated peanut (*Arachis hypogaea* L.). Euphytica **97,** 143—149.

Henderson, S. T., and T. D. Petes, 1992: Instability of simple sequence DNA in *Saccharomyces cervisieae*. Mol. Cell. Biol. **12,** 2749—2757.

Herselman, L., 2003: Genetic variation among Southern African cultivated peanut (*A. hypogaea* L.) genotypes as revealed by AFLP analysis. Euphytica **133,** 319—327.

Kimura, M., and J. Crow, 1964: The Number of alleles that can be maintained in a finite population. Genetics **49,** 725—738.

Knauft, D. A., and D. W. Gorbet, 1989: Genetic diversity among peanut cultivars. Crop Sci. **29,** 1417—1422.

Kochert, G., H. T. Stalker, M. Gimenes, L. Galgaro, C. R. Lopes, and K. Moore, 1996: RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut *Arachis hypogaea* (Leguminosae). Am. J. Bot. **83,** 1282—1291.

Kottapalli, K., M. Burow, G. B. Burow, J. J. Burke, and N. Puppala, 2007: Molecular characterization of the US peanut mini core collection using microsatellite markers. Crop Sci. **47,** 1718—1727.

Krapovickas, A., 1968: Origen, variabilidad y diffusion de1 mani (*Aruchis hypogueu* L.). Actas Memorias del XXXVII Congresso Internacional de Americanistas **2,** 517—534.

Krapovickas, A., and W. C. Gregory, 1994: Taxonomia del *Arachis* (*Leguminosae*). Bonplandia **VIII,** 1—187.

Li, Y., J. Du, T. Wang, Y. Shi, and J. Jia, 2002: Genetic diversity and relationships among Chinese maize inbred lines revealed by SSR markers. Maydica **47,** 93—101.

Liu, K., and S. Muse, 2005: PowerMarker: New Genetic Data Analysis. Software, Version 2.7. Available at: http://www.powermarker.net (last accessed 23 February 2008).

Mace, E. S., D. T. Phong, H. D. Upadhyaya, S. Chandra, and J. H. Crouch, 2006: SSR analysis of cultivated groundnut (*Arachis hypogaea* L.) germplasm resistant to rust and late leaf spot diseases. Euphytica **152,** 317—330.

Mace, E. S., R. K. Varshney, V. Mahalakshmi, K. Seetha, A. Gafoor, Y. Leeladevi, and J. H. Crouch, 2007: *In silico* development of simple sequence repeat markers within the aeschynomenoid/ dalbergoid and genistoid clades of the Leguminosae family and their transferability to *Arachis hypogaea*, groundnut. Plant Sci. **174,** 51—60.

Moretzsohn, M. C., L. Leoi, K. Proite, P. M. Guimarães, S. C. Leal-Bertioli, M. A. Gimenes, W. S. Martins, J. F. M. Valls, D. Grattapaglia, and D. J. Bertioli, 2005: A microsatellite-based, gene-rich linkage map for the AA genome of *Arachis* (Fabaceae). Theor. Appl. Genet. **111,** 1060—1071.

Ni, J., P. M. Colowit, and D. J. Mackill, 2002: Evaluation of genetic diversity in rice subspecies using microsatellite markers. Crop Sci. **42,** 601—607.

Ohta, T., and M. Kimura, 1973: A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22,** 201—204.

Parkin, I. A., P. Sharpe, and D. J. Lydiate, 2003: Patterns of genome duplication with in the *Brassica napus* genome. Genome **46,** 291—303.

Perrier, X., A. Flori, and F. Bonnot, 2003: Data analysis methods. In: P. Hamon, M. Seguin, X. Perrier, and J. C. Glaszmann (eds), Genetic Diversity of Cultivated Tropical Plants, 43—76. CIRAD/ Science, Montpellier, France.

Prasad, M., R. K. Varshney, J. K. Roy, H. S. Balyam, and P. K. Gupta, 2000: The use of microsatellites for detecting DNA polymorphism, genotype identification and genetic diversity in wheat. Theor. Appl. Genet. **100,** 594—602.

Stachel, M., T. Lelly, H. Grausgruber, and J. Vollmann, 2000: Application of microsatellites in wheat (*Triticum aestivum* L.) for studying genetic differentiation caused by selection for adaptation and use. Theor. Appl. Genet. **100,** 242—248.

Subramanian, V., S. Gurtu, R. C. N. Rao, and S. N. Nigam, 2000: Identification of DNA polymorphism in cultivated groundnut using random amplified polymorphic DNA (RAPD) assay. Genome **43,** 656—660.

Tang, R., G. Gao, L. He, Z. Han, S. Shan, R. Zhong, C. Zhou, J. Jiang, Y. Li, and W. Zhuang, 2007: Genetic diversity in cultivated groundnut based on SSR Markers. J. Genet. Genomics **34,** 449—459.

Upadhyaya, H. D., R. Bhattacherjee, D. Hoisington, S. Chandra, R. K. Varshney, S. Singh, M. C. Moretzsohn, S. Leal-Bertoli, P. Guimereas, and D. Bertioli, 2006: Molecular Characterization of Groundnut (*Arachis hypogaea* L.) Composite Collection. Aposter presented in the annual meetings of Generation Challenge Programme, 12th–16th Sep 2006, Sao Paulo, Brazil.

Varshney, R. K., A. Kumar, H. S. Balyan, J. K. Roy, M. Prasad, and P. K. Gupta, 2000: Characterization of microsatellites and development of chromosome specific STMS markers in bread wheat. Plant Mol. Biol. Rep. **18,** 5—16.

Varshney, R. K., A. Graner, and M. E. Sorrells, 2005: Genomics-assisted breeding for crop improvement. Trends Plant Sci. **10,** 621—630.

Varshney, R. K., D. A. Hoisington, H. D. Upadhyaya, P. M. Gaur, S. N. Nigam, K. Saxena, V. Vadez, N. K. Sethy, S. Bhatia, R. Aruna, M. V. C. Gowda, and N. K. Singh, 2007: Genomic assisted crop improvement vol II: genomics applications in crops. In: R. K. Varshney, and R. Tuberosa (eds), Molecular Genetics and Breeding of Grain Legume Crops for the Semi-arid Tropics, 207—242. Springer, The Netherlands.

Varshney, R. K., D. J. Bertioli, M. C. Moretzsohn, V. Vadez, L. Krishnamurthy, A. Rupakula, S. N. Nigam, B. J. Moss, K. Seetha, K. Ravi, G. He, S. J. Knapp, and D. A. Hoisington, 2008: The first SSR-based genetic linkage map for cultivated groundnut (*Arachis hypogaea* L.). Theor. Appl. Genet. **118,** 729—739.

Weber, J. L., and C. Wong, 1993: Mutation of human short tandem repeats. Hum. Mol. Genet. **2,** 1123—1128.