GENETIC DIVERSITY IN WILD AND CULTIVATED PEANUT

by

SAMEER KHANAL

(Under the Direction of S.J. KNAPP)

ABSTRACT

Narrow genetic diversity and a deficiency of polymorphic DNA markers have hindered genetic mapping and the application of genomics and molecular breeding approaches in cultivated peanut (*Arachis hypogaea* L.). We developed and mined genome survey sequences (GSSs) and a peanut EST database, assessed the frequency of polymorphic SSRs, and initiated the development of several hundred SSR markers with the goal of breaking the DNA marker bottleneck in cultivated peanut. Primers were designed and tested for a broad spectrum of SSR motifs and repeat lengths, and 97 GSS-based SSR, 59 EST-SSR, and 612 previously reported SSR markers were screened among diploid and tetraploid germplasm accessions. Population structures were resolved and allelic diversities were estimated among wild and cultivated peanut. The frequency of polymorphic SSRs is sufficient for developing a critical mass of DNA markers for genetic mapping and downstream molecular breeding applications in peanut.

INDEX WORDS: *Arachis hypogaea*, EST, GSS, peanut, SSR

GENETIC DIVERSITY IN WILD AND CULTIVATED PEANUT

by

SAMEER KHANAL

B.S., Tribhuvan University, Nepal, 2003

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2008

GENETIC DIVERSITY IN WILD AND CULTIVATED PEANUT

by

SAMEER KHANAL

Major Professor: Steven J. Knapp

Committee: Albert K. Culbreath
E. Charles Brummer

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

DEDICATION

I would like to dedicate this thesis to my parents. A special thanks goes to my brother Subhash; you are always an inspiration.

ACKNOWLEDGEMENTS

I would like to thank Dr. Steven J. Knapp for his support and mentoring – you have been considerate throughout the period. I would also like to thank all my friends, colleagues, and instructors for their unconditional support – two years at this educational institution has been a great learning experience.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Purpose of the Study

Historically, peanut (*Arachis hypogaea* L.) has lagged behind crops of equal or lesser economic significance in the development of genomics resources and the application of molecular breeding solutions to the problems of agricultural and economic importance (He *et al.* 2003; Gepts *et al.* 2005). The outlook in peanut has been bleak - genetic mapping has not been accomplished in intraspecific *A. hypogaea* populations, phenotypic and quantitative trait loci (QTL) have not been genetically mapped, and the tools needed for routinely applying forward genetics approaches and marker-assisted selection (MAS) have not been developed. The lack of progress in peanut, an autogamous tetraploid (AABB; 2n = 4x = 40), partly stems from limited nucleotide diversity in the elite gene pool, a product of speciation, domestication, and breeding bottlenecks (Knauft and Gorbet 1989; Halward *et al.* 1991; Kochert *et al.* 1991; Isleib and Wynne 1992; Young *et al.* 1996).

Diploid wild species, on the other hand, are shown to be highly polymorphic and are potential sources of several crop-friendly traits. Also, the utility of wild diploids for genetic mapping in peanut has been known for many years (Halward *et al.* 1991, 1992; Kochert *et al.* 1991). Diploid linkage maps are crucial in light of the fact that mapping in cultivated peanut has been severely hampered by a dearth of adequate molecular polymorphisms in tetraploids; loci mapped in diploids could help prepare a genetic map for *A. hypogaea*. Nevertheless, progress has been slow; only two diploid and a tetraploid interspecific maps have been produced (Halward *et*

al. 1993; Burow *et al.* 2001; Moretzsohn *et al.* 2005). These linkage maps are sparse and available polymorphic markers grossly underrepresented the critical volume required for the construction of saturated linkage maps.

One of the goals of this research project is to gain an understanding of the abundance and polymorphisms of SSRs in a peanut EST database. This study has practical significance, in-that the information will be used to develop predictions for large-scale EST-SSR marker development. Another goal of this study is to decipher the population structure of the tetraploid and diploid peanut genotypes such that the most polymorphic set of individuals are identified as diploid and tetraploid models for molecular mapping. Further, this research will supply the peanut industry with additional DNA markers and mapping resources critical for applying genomic solutions to peanut breeding and cultivar development problems. Additionally, utility of methylation-filtering (MF) approach for reducing genomic complexity and gene space sequencing of diploid and tetraploid peanut will be assessed, and genomic sequences will be used for SSR marker development. From the peanut EST sequences and genome survey sequences, more than hundred polymorphic SSR markers will be developed. Also, several hundred polymorphic markers will be identified in different intraspecific diploid mapping populations.

Economic Importance of Peanut

Cultivated peanut or groundnut, *A. hypogaea* L., belongs to genus *Arachis* L. under the family of legumes viz. Fabaceae. The genus harbors 80 accepted species assembled into 9 sections, including section *Arachis* representing the cultivated tetraploid, together with 31 wild diploids and a wild tetraploid species, *A. monticola* (Krapovickas and Gregory 1994; Valls and Simpson 2005). Endemic to South America, cultivated peanut is now a globally important crop,

widely grown in tropical, sub-tropical, and warm-temperate regions of the world (Stalker 1997). International Crops Research Institute for the Semi-Arid Tropics (ICRISAT, www.icrisat.org) and the United Nations Food and Agricultural Organization (UN-FAO, www.fao.org) identify peanut as the third most important source of vegetable protein, fourth most important source of vegetable oil, and twelfth most important food crop with 47.7 million tones in-shell groundnut harvested worldwide from 22.2 million hectares in 2006 (FAOSTAT, http://faostat.fao.org). Further, among food legumes, it is second only to soybean and equal to common bean in national and global economic importance. Among the wild relatives, *A. glabrata* Benth. and *A. pintoi* Krapov. & W. C. Greg. are pasture crops of commercial importance at small scale.

Genomics Resources and Molecular Breeding in Cultivated Peanut

Quantifying nucleotide diversity and tagging genomes at those sites are important forward genetic approaches for biological studies. As such, DNA markers are crucial in strengthening scientific knowledge base and accelerating crop improvement endeavors. Conspicuous differences flagged by DNA markers have helped develop genetic linkage maps in several crop species which, in turn, facilitated targeted gene cloning and marker assisted selection. Further, they have been frequently used in functional genetic studies, comparative genomics, and molecular evolutionary studies. However, despite its huge economic significance, peanut has historically lagged behind crops of equal or lesser importance in the development of genomics resources and application of molecular breeding solutions to problems of economic importance (He *et al.* 2003; Gepts *et al.* 2005). The development of DNA markers for genetic mapping, forward genetic analyses, and molecular breeding has been impeded in peanut by a scarcity of DNA polymorphisms among elite lines (Grieshammer and Wynne 1990; Halward *et al.* 1991, 1992, 1993; Kochert *et al.* 1991; Lacks and Stalker 1993; Stalker and Simpson 1995).

Despite abundant morphological variation, polymorphism studies with several classes of molecular markers have reported lower levels of nucleotide diversity in cultivated peanut (Grieshammer and Wynne 1990; Halward *et al.* 1991, 1992; Kochert *et al.* 1991; Paik-Ro *et al.* 1992; Stalker *et al.* 1994; He and Prakash 1997). The genetic narrowness has been attributed to speciation, domestication, and breeding bottlenecks (Halward *et al.* 1991; Krapovickas and Gregory 1994; Young *et al.* 1996). *A. hypogaea* is hypothesized to have been founded by the reproductive isolation of an amphidiploid produced by spontaneous doubling of the chromosomes of an interspecific hybrid (Halward *et al.* 1991; Krapovickas and Gregory 1994; Young *et al.* 1996). Speciation, domestication, and breeding bottlenecks together with varietal homogeneity rapidly eroded genetic variation, which was further exhausted by rearing susceptible genotypes under chemical intensive production systems (Stalker 1997).

Scarcity of DNA polymorphism has impeded genetic mapping in elite × elite crosses and necessitated genetic mapping in interspecific diploid and tetraploid populations. Since the advent and use of DNA markers (Botstein *et al.* 1980), genetic mapping has only been done in two interspecific diploid hybrids and an interspecific tetraploid hybrid in peanut (Halward *et al.* 1993; Burow *et al.* 2001; Moretzsohn *et al.* 2005), and no QTL have been mapped. The problem can be partly solved through the development of diploid models and comparative mapping in highly polymorphic wild diploid and less polymorphic cultivated tetraploid species. Wild *Arachis* spp. show greater allelic diversities and are also the potential sources of desirable traits in cultivated peanut.

Importance of Wild *Arachis* Germplasm

Diversity assessment for several agronomic traits has suggested higher levels of variation in wild *Arachis* germplasm pool. The wild peanut species show extensive variability in seed

storage protein profiles (Cherry 1977; Bianchi-Hall *et al.* 1993), resistance to several biotic (Lynch 1990; Yang *et al.*, 1993; Holbrook and Stalker 2003; Kalyani *et al.* 2007) and abiotic stresses (Holbrook and Stalker 2003; Rao *et al.* 2003; Nautiyal *et al.* 2008). Although introgression of agronomic traits from these wild relatives into the cultivated genotypes is a promising prospect for the development of superior cultivars, large scale incorporation strategies (Simmonds 1993) could help broaden the genetic base of the elite gene pool. The feasibility of introgression has been facilitated by cross compatibility of several of the closest wild relatives constituting the secondary gene pool for cultivated peanut (Singh and Simpson 1994; Stalker 1997; Holbrook and Stalker 2003).

A large number of wild introgression lines have been reported in recent years. In 2002, several introgression lines derived from an *A. hypogaea* (PI 261942) X *A. cardenasii* (PI 262141) interspecific cross were released (Stalker *et al.* 2002a; Stalker *et al.* 2002b; Stalker and Lynch 2002). These lines are frequently crossed to cultivated peanut for the transfer of the traits as resistant to insect pests and diseases (Anderson *et al.* 2006). Similarly, four elite germplasm lines with wild introgressions of rust and leaf spot resistance from *A. duranensis*, *A. batizocoi*, *A. villosa*, and *A. stenosperma* were reported by Singh *et al.* (2003). Transfer of resistance to leaf spots and some insect pests has also been reported from *A. kempff-mercadoi* (Mallikarjuna *et al.* 2004), adding up to the list of several species within genus *Arachis* shown to be amenable for gene transfer through wide hybridization. Further, these wild species are potential sources of multiple traits of interest in cultivated peanuts. However, to date, very few elite cultivars with wild introgressions have been released (Simpson and Starr 2001; Simpson *et al.* 2003). Some excellent wild introgression lines and varieties have been found inferior to their check cultivars possibly due to linkage drag (Isleib *et al.* 2006; Mondal *et al.* 2007, Holbrook *et al.* 2008). As

such, identification and tagging of genes of interest along with marker-assisted selection provides better selection efficiency and opens up possibilities for gene pyramiding. A critical mass of molecular markers mapped onto saturated linkage maps is an undisputed genomics asset which facilitates targeted gene cloning and marker assisted selection besides other downstream applications as functional genetic studies, comparative genomics, and molecular evolutionary studies.

The utility of wild diploids for genetic mapping in peanut has been known for many years (Halward *et al.* 1991, 1992; Kochert *et al.* 1991). Diploid linkage maps are crucial in light of the fact that mapping in cultivated peanut has been severely hampered by a dearth of adequate molecular polymorphisms in tetraploids; loci mapped in diploids could help prepare a genetic map for *A. hypogaea*. Nevertheless, progress has been slow; only two interspecific diploid maps have been produced (Halward *et al.* 1993; Moretzsohn *et al.* 2005). Prior to the publication of the first public map for *Arachis* (Moretzsohn *et al.* 2005), two maps had been published, a diploid RFLP map produced from an interspecific (*A. duranensis* × *A. stenosperma*) hybrid (Halward *et al.* 1993) and a tetraploid RFLP map produced from an interspecific hybrid between *A. hypogaea* and a 'synthetic' amphidiploid [*A. batizocoi* × (*A. cardenasii* × *A. diogoi*)] (Burow *et al.* 2001). These linkage maps are sparse and available polymorphic markers grossly underrepresented the critical volume required for the construction of saturated linkage maps.

Simple Sequence Repeats as a Marker Class of Choice

Although various molecular marker systems viz. Isozymes (Lu and Pickersgill 1993; Stalker *et al.* 1994), RFLPs (Kochert *et al.* 1991; Halward *et al.* 1993; Burow *et al.* 2001), RAPDs (Halward *et al.* 1991; Halward *et al.* 1992; Hilu and Stalker 1995; Raina *et al.* 2001), AFLPs (He and Prakash 2001; Gimenes *et al.* 2002; Herselman 2003), and simple sequence

repeats (SSRs) have been developed and assessed for polymorphisms, SSRs (microsatellites) were identified as the most informative marker system in *Arachis* spp. (Hopkins *et al.* 1999; Raina *et al.* 2001; Gimenes *et al.* 2007).

SSRs have become a marker class of choice in many crop species because they are co-dominant and highly mutable, multiallelic and locus specific, reproducible, and adaptable to high-throughput genotyping platforms (Powell *et al.* 1996a, 1996b). There are tamdemly repeated arrays of one to six nucleotides found interspersed in the genomes of the organisms. Their identification involves mining DNA sequence databases (either random genomic, SSR enriched genomic or cDNA libraries) for the repeat tracts.

So far, about 1000 SSR markers have been developed for peanut. Conventional genomic approaches involving SSR enrichment procedures have been frequently used for marker development (Hopkins *et al.* 1999; Palmieri *et al.* 2002, 2005; He *et al.* 2003; Ferguson *et al.* 2004; Moretzsohn *et al.* 2005; Bravo *et al.* 2006; Gimenes *et al.* 2007; Cuc *et al.* 2008). Since genomic approaches are expensive and time-consuming (Karagyozov *et al.* 1993; Tang *et al.* 2002; He *et al.* 2003), relatively simple and inexpensive approaches as EST database mining (Luo *et al.* 2005; Moretzsohn *et al.* 2005; Proite *et al.* 2007) and inter-generic transfers of SSRs (He *et al.* 2006; Mace *et al.* 2008; Sanders *et al.* 2008) have emerged as alternative marker development approaches in recent years.

Genomics Initiatives and Magic Peanut Database (PeanutDB)

Our efforts in the development of *Arachis* genomics resources closely follow initiatives in peanut envisioned by cross-legume advances in genomics conference (http://catg.ucdavis.edu, Gepts *et al.* 2005). With an aim of breaking genomics gridlock and application of molecular breeding approaches, we followed massively parallel DNA sequencing and marker development

strategy (especially SSRs) in *Arachis* species. First, we produced and mined methylation-filtered (MF) and unfiltered (UF) genome survey sequences (GSSs) from *A. duranensis*, *A. batizocoi*, and *A. hypogaea* for SSRs. The development and characterization of 97 SSR markers from these sequences is discussed in the second chapter of this thesis. Second, we assessed 58 diploid *Arachis* accessions for allelic diversities. Also, we screened 709 previously reported SSR markers against parents of several diploid and tetraploid mapping populations. SSR diversity and large scale marker screening studies in the diploids are discussed in the fourth chapter of this thesis. Third, we produced Sanger and 454 ESTs from normalized and non-normalized leaf and developing-seed cDNA libraries of *A. hypogaea* (Table 1). We also developed a pipeline for building and annotating transcript assemblies and developed a groundnut EST database (PeanutDB, www.peanut.uga.edu, unpublished). In the third chapter, we report mining and characterizing groundnut EST database and results of a pilot study on EST-SSR screening against 28 tetraploid peanut genotypes.

Table 1.1. Groundnut ESTs and transcript assembly statistics.

ESTs			Unigenes		
Sanger	454	Total	Singletons	Contigs	Total
71,448	304,215	375,663	63,218	37,914	101,132

We are currently screening more than 2000 EST-SSRs for length polymorphisms and expect at least 20% and 50% of these markers polymorphic, respectively, in tetraploid and in diploid mapping populations.

We are concentrating our resources on *A. duranensis* and *A. batizocoi* as A- and B-genome diploid models for cultivated tetraploids. We also included *A. stenosperma, A. diogoi,*

and A. kuhlmanii in some of our studies. Based on phylogenetic studies, these species are close to *A. hypogaea* and are representatives of secondary gene pool for *A. hypogaea* (Holbrook and Stalker 2003).

A. duranensis is one of the prime candidates for the A-genome progenitor and is represented by 39 (GRIN) to 73 (ICRISAT) germplasm accessions in public seed banks, and is highly polymorphic. However, source of B-genome has been elusive. *A. ipaensis* and *A. batizocoi* are the prime candidates, although *A. batizocoi* has been discounted in some analysis (Kochert *et al.* 1991, 1996; Raina and Mukai 1999a, 1999b; Jung *et al.* 2003; Seijo *et al.* 2004; Milla *et al.* 2005). *A. ipaensis* was eliminated as a candidate for B-genome mapping because public seed banks presently only hold one (or two) germplasm accessions. *A. batizocoi*, on the other hand, is represented by 13 (GRIN) to 26 (ICRISAT) germplasm accessions in public seed banks, and is highly polymorphic. *A. stenosperma*, on the other hand, is shown to have resistance to several cultivated peanut pathogens and is cross-fertile with *A. duranensis*, thereby making it a potential candidate for A-genome donor for synthetic allotetraploids (Proite *et al.* 2007).

Summary and Goals

Although cultivated peanut (*Arachis hypogaea* L.) is globally an important food crops, it is lacking in genomics resources and molecular breeding solutions to address the pertinent production problems. Despite abundant morphological variation, lower level of molecular polymorphisms is one of the major crop improvement constraints in cultivated peanut. Polymorphism studies with several classes of molecular markers have reported lower levels of nucleotide diversity. However, SSRs are shown to be the most informative marker system in peanut and database mining for EST-SSRs has emerged as an efficient marker development approach. The problems from apparent paucity of molecular polymorphisms in cultivated peanut

could be partly solved through the development of diploid models and comparative mapping in highly polymorphic wild diploid and less polymorphic cultivated tetraploid species. Marker resources are critical for the purpose. Several hundred polymorphic SSRs have been publicly reported in *Arachis* but these grossly under-represent the critical volume required for the construction of saturated linkage maps in *Arachis* species. Therefore, we identified the following goals for this thesis

1. Gain an understanding of molecular genetic diversity in the A- and B-genome diploid and AB-genome tetraploid peanut species.
2. Select parents for developing intraspecific A- and B-genome diploid mapping populations to lay the groundwork for developing diploid models for peanut genomics.
3. Significantly increase the supply of simple sequence repeat (SSR) markers for genotyping applications in diploid and tetraploid peanut.
4. Significantly increase the supply of DNA markers for transcribed loci for comparative mapping and synteny analysis with other legumes.
5. Assess the utility of methylation-filtering approaches for efficient gene space sequencing in diploid and tetraploid peanut.

References

Anderson, W.F., C.C. Holbrook, and P. Timper. 2006. Registration of root-knot nematode resistant peanut germplasm lines NR 0812 and NR 0817. Crop Sci. 46:481-482.

Bianchi-Hall, C.M., R.D. Keys, H.T. Stalker, and J.P. Murphy. 1993. Diversity of seed storage protein-patterns in wild peanut (*Arachis*, Fabaceae) species. Plant Systematics and Evolution 186:1-15.

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

Bravo, J.P., A.A. Hoshino, C. Angelici, C.R. Lopes, and M.A. Gimenes. 2006. Transferability and use of microsatellite markers for the genetic analysis of the germplasm of some Arachis section species of the genus Arachis. Genetics and Molecular Biology 29:516-524.

Burow, M.D., C.E. Simpson, J.L. Starr, and A.H. Paterson. 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.

Cherry, J.P. 1977. Potential sources of peanut seed proteins and oil in the genus *Arachis*. J. Agric. Food Chem. 25:186-193.

Cuc, L., E. Mace, J. Crouch, V. Quang, T. Long, and R. Varshney. 2008. Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (Arachis hypogaea). Bmc Plant Biology 8:55.

Ferguson, M., M. Burow, S. Schulze, P. Bramel, A. Paterson, S. Kresovich, and S. Mitchell. 2004. Microsatellite identification and characterization in peanut (A. hypogaea L.). TAG Theoretical and Applied Genetics 108:1064-1070.

Gepts, P., W.D. Beavis, E.C. Brummer, R.C. Shoemaker, H.T. Stalker, N.F. Weeden, and N.D. Young. 2005. Legumes as a model plant family. Plant Physiology 137:1228-1235.

Gimenes, M.A., A.A. Hoshino, A.V.G. Barbosa, D.A. Palmieri, and C.R. Lopes. 2007. Characterization and transferability of microsatellite markers of the cultivated peanut (Arachis hypogaea). Bmc Plant Biology 7.

Gimenes, M.A., C.R. Lopes, and J.F.M. Valls. 2002. Genetic relationships among Arachis species based on AFLP. Genetics and Molecular Biology 25:349-353.

Grieshammer, U., and J.C. Wynne. 1990. Isozyme variability in mature seeds of U.S. peanut cultivars and collections. Peanut Sci. 18:72-75.

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

Halward, T., H.T. Stalker, E. LaRue, G. Kochert. 1992. Use of single-primer amplifications in genetic studies of peanut (Arachis hypogaea L.). Plant Mol. Biol. 18:315-325.

Halward, T., H.T. Stalker, and G. Kochert. 1993. Development of an RFLP linkage map in diploid peanut species. Theoretical and Applied Genetics 87:379-384.

Han, Z.G., C.B. Wang, X.L. Song, W.Z. Guo, J.Y. Gou, C.H. Li, X.Y. Chen, and T.Z. Zhang. 2006. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theoretical and Applied Genetics 112:430-439.

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

He, G.H., and C. Prakash. 2001. Evaluation of genetic relationships among botanical varieties of cultivated peanut (Arachis hypogaea L.) using AFLP markers. Genetic Resources and Crop Evolution 48:347-352.

He, G., F.E. Woullard, I. Morong, and B.Z. Guo. 2006. Transferability of soybean SSR markers in peanut (*Arachis hypogaea* L.). Peanut Science. 33:22-28.

He, G., R. Meng, M. Newman, G. Gao, R. Pittman, and C.S. Prakash. 2003. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biology 3:3.

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

Hilu, K.W., and H.T. Stalker. 1995. Genetic relationships between peanut and wild species of Arachis sect Arachis (Fabaceae): Evidence from RAPDs. Plant Systematics and Evolution 198:167-178.

Holbrook, C.C., and H.T. Stalker. 2003. Peanut breeding and genetic resources, pp. 297-356. In J. Janick (ed.) Plant Breeding Reviews 22, John Wiley & Sons, Inc.

Holbrook, C.C., P. Timper, W. Dong, C.K. Kvien, and A.K. Culbreath. 2008. Development of near-isogenic peanut lines with and without resistance to the peanut root-knot nematode. Crop Sci. 48:194-198.

Hopkins, M.S., A.M. Casa, T. Wang, S.E. Mitchell, R.E. Dean, G.D. Kochert, and S. Kresovich. 1999. Discovery and characterization of polymorphic simple sequence repeats (SSRs) in peanut. Crop Science 39:1243-1247.

Isleib, T.G., P.W. Rice, R.W. Mozingo II, S.C. Copeland, J.B. Graeber, B.B. Shew, D.L. Smith, H.A. Melouk, and H.T. Stalker. 2006. Crop Sci. 46:2329-2330.

Jung, S., P.L. Tate, R. Horn, G. Kochert, K. Moore, and A.G. Abbott. 2003. The phylogenetic relationship of possible progenitors of the cultivated peanut. J. of Heredity 94(4):334-340.

Karagyozov, L., I.D. Kalcheva, and V.M. Chapman. 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research 21:3911-3912.

Knauft, D.A., and D.W. Gorbet. 1989. Genetic diversity among peanut cultivars. 29:1417-1422.

Kochert, G., T. Halward, W.D. Branch, and C.E. Simpson. 1991. RFLP variability in peanut (*Arachis hypogaea* L.) cultivars and wild-species. Theoretical and Applied Genetics 81:565-570.

Krapovickas, A., and W.C. Gregory. 1994. Taxonomia del genero *Arachis* (Leguminosae). Bonplandia 8: 1-186.

Lacks, G.D. and H.T. Stalker. 1993. Isozyme analysis of *Arachis* species and interspecific hybrids. Peanut Sci. 20:76-81.

Lu, J., and B. Pickersgill. 1993. Isozyme variation and species relationships in peanut and its wild relatives (*Arachis* L. - Leguminosae). Theoretical and Applied Genetics 85:550-560.

Luo, M., P. Dang, B.Z. Guo, G. He, C.C. Holbrook, M.G. Bausher, and R.D. Lee. 2005. Generation of Expressed Sequence Tags (ESTs) for Gene Discovery and Marker Development in Cultivated Peanut. Crop Sci 45:346-353.

Lynch, R.E. 1990. Resistance in peanut to major arthropod pests. Florida Entomologist 73:422-445.

Mace, E.S., R.K. Varshney, V. Mahalakshmi, K. Seetha, A. Gafoor, Y. Leeladevi, and J.H. Crouch. 2008. 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 Science 174:51-60.

Mallikarjuna, N., S. Pande, D.R. Jadhav, D.C. Sastri, and J.N. Rao. 2004. Introgression of disease resistance genes from *Arachis kempff-mercadoi* into cultivated groundnut. Plant Breeding 123:573-576.

Milla, S.R., T.G. Isleib, and H.T. Stalker. 2005. Taxonomic relationship among *Arachis* sect. *Arachis* species as revealed by AFLP markers. Genome 48:1-11.

Mondal, S., A.M. Badigannavar, G.S.S. Murty. 2007. RAPD markers linked to a rust resistance gene in cultivated groundnut (*Arachis hypogaea* L.). Euphytica 159:233-239.

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

Nautiyal, P.C., K. Rajgopal, P.V. Zala, D.S. Pujari, M. Basu, B.A. Dhadhal, and B.M. Nandre. 2008. Evaluation of wild Arachis species for abiotic stress tolerance: I. Thermal stress and leaf water relations. Euphytica 159:43-57.

Paik-Ro, O.G., R.L. Smith, and D.A. Knauft. 1992. Restriction fragment lenth polymorphism evaluation of six peanut species within the *Arachis* section. Theoret. Appl. Genet. 84:201-208.

Palmieri, D.A., A.A. Hoshino, J.P. Bravo, C.R. Lopes, and M.A. Gimenes. 2002. Isolation and characterization of microsatellite loci from the forage species Arachis pintoi (Genus Arachis). Molecular Ecology Notes 2:551-553.

Palmieri, D.A., B. M. D, C. R. A, G. M. A, and L. C. R. 2005. Novel polymorphic microsatellite markers in section Caulorrhizae (Arachis, Fabaceae). Molecular Ecology Notes 5:77-79.

Powell, W., G.C. Machray, and J. Provan. 1996a. Polymorphism revealed by simple sequence repeats. Trends in Plant Science 1:215-222.

Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996b. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 2:225-238.

Proite, K., S.C.M. Leal-Bertioli, D.J. Bertioli, M.C. Moretzsohn, F.R. da Silva, N.F. Martins, and P.M. Guimaraes. 2007. ESTs from a wild *Arachis* species for gene discovery and marker development. Bmc Plant Biology 7.

Raina, S.N., and Y. Mukai. 1999a. Genomic in situ hybridization in *Arachis* (*Fabaceae*) identifies the diploid wild progenitors of cultivated (*A. hypogaea*) and related wild (*A. monticola*) peanut. Pl. Syst. and Evol. 214:251-262.

Raina, S.N., and Y. Mukai. 1999b. Detection of a variable number of 18S-5.8S-26S and 5S ribosomal DNA loci by florescent in situ hybridization in diploid and tetraploid *Arachis* species. Genome 42:52-59.

Raina, S.N., V. Rani, T. Kojima, Y. Ogihara, K.P. Singh, and R.M. Devarumath. 2001. RAPD and ISSR fingerprints as useful genetic markers for analysis of genetic diversity, varietal identification, and phylogenetic relationships in peanut (Arachis hypogaea) cultivars and wild species. Genome 44:763-772.

Rao, N.K., L.J. Reddy, and P.J. Bramel. 2003. Potential of wild species for genetic enhancement of some semi-arid food crops. Genetic Resources and Crop Evolution 50:707-721.

Sanders, F.E., H. Guohao, L. Gong, M. Egnin, and D. Morley. 2008. Transferability of soybean (*Glycine max*) SSR markers in peanut genomic DNA (*Arachis hypogaea* L.). In Vitro Cell. Dev. Biol.-Plant 44:356-357.

Seijo, J.G., G.I. Lavia, A. Fernandez, A. Krapovickas, D. Ducasse, and E.A. Moscone. 2004. Physical mapping of the 5S and 18S-25S rRNA genes by FISH as evidence that Arachis duranensis and A. ipaensis are the wild diploid progenitors of A. hypogaea (Leguminosae). Am. J. Bot. 91:1294-1303.

Simmonds, N.W. 1993. Introgression and incorporation strategies for the use of crop genetic resources. Biol. Rev. 68:539-562.

Simpson, C.E., and J.L. Starr. 2001. Registration of 'COAN" peanut. Crop Sci. 41:918.

Simpson, C.E., J.L. Starr, G.T. Church, M.D. Burow, and A.H. Paterson. 2003. Registration of 'NemaTAM' peanut. 43:1561.

Singh, A.K., and C.E. Simpson. 1994. Biosystematics and genetic resources. The ground crop: a scientific basis for improvement. J.Smartt. London, Chapman & Hall: 96-137.

Singh A.K., S.L. Dwivedi, S. Pande, J.P. Moss, S.N. Nigam, and D.C. Sastri. 2003. Registration of rust and late leaf spot resistant peanut germplasm lines. Crop Sci. 43:440-441.

Stalker, H.T. 1997. Peanut (*Arachis hypogaea* L.). Field Crops Research 53:205-217.

Stalker, H.T., and C.E. Simpson. 1995. Genetic resources in Arachis, pp. 14-53. In H.E. Pattee and H.T. Stalker (eds.) Advances in peanut science. Am. Peanut Res. Educ. Soc., Stillwater, OK.

Stalker, H.T., T.D. Phillips, J.P. Murphy, and T.M. Jones. 1994. Variation of isozyme patterns among *Arachis* species. Theoretical and Applied Genetics 87:746-755.

Stalker and Lynch. 2002. Registration of four insect-resistant peanut germplasm lines. Crop Sci. 42:313-314.

Stalker, H.T., M.K. Beute, B.B. Shew, and K.R. Barker. 2002a. Registration of two root-knot nematode-resistant peanut germplasm lines. Crop Sci. 42:312-313.

Stalker, H.T., M.K. Beute, B.B. Shew, and T.G. Isleib. 2002b. Registration of five leaf spot-resistant peanut germplasm lines. Crop Sci. 42:314-316.

Tang, S., J.K. Yu, M.B. Slabaugh, D.K. Shintani, and S.J. Knapp. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105:1124-1136.

Valls, J.F.M, and C.E. Simpson. 2005. New species of Arachis L. (Leguminosae) from Brazil, Paraguay and Bolivia. Bonplandia 14:35-64.

Yang, G., K.E. Espelie, J.W. Todd, A.K. Culbreath, R.N. Pittman, and J.W. Demski. 1993. Cuticular lipids from wild and cultivated peanuts and the relative resistance of these peanut species to fall armyworm and thrips. Journal of Agricultural and Food Chemistry 41:814-818.

Young, N.D., N.F. Weeden, and G. Kochert. 1996. Genome mapping in legumes (Family Fabaceae), pp. 211-227. In A.H. Paterson (ed.) Genome mapping in plants. R.G. Landes Co., Austin, TX.

CHAPTER 2

MINING A- AND B-GENOME DIPLOID AND AB-GENOME TETRAPLOID GENOME SURVEY SEQUENCES FOR SIMPLE SEQUENCE REPEATS IN PEANUT[1]

[1]Khanal, S., S. Tang, W. Ma, and S.J. Knapp. To be submitted to *Theor Appl Genet.*

Abstract

Deficiency of polymorphic DNA markers has restricted genetic mapping and the application of genomics and molecular breeding approaches in peanut (*Arachis hypogaea* L.). Genomic survey sequences (GSSs) were produced from unfiltered and methylation-filtered genomic DNA libraries of *A. duranensis* (AA), *A. batizocoi* (BB), and *A. hypogaea* (AABB), and 97 simple sequence repeat markers were developed from 960 SSRs identified among 9,517 GSSs with a density of one SSR/4.7 kb. The markers were screened for polymorphisms among four A- and B-genome diploid and eight cultivated tetraploid germplasm accessions. Dinucleotide repeat motifs (80.3%) were more abundant than trinucleotide (16.8%) and tetranucleotide (2.7%) repeat motifs. In tetraploid peanuts, SSRs longer than 26 bp (H = 0.44) were almost four times more polymorphic than SSRs shorter than 26 bp (H =0.12). Of the 97 SSR markers, 90 (93%) were polymorphic among wild diploid ecotypes (H = 0.57) and 38 (40%) were polymorphic among tetraploid germplasm accessions (H = 0.24). Transferability of SSRs was moderately high (70% to 100%) among the three source species. The application of methylation-filtering in A- and B-genome diploid and AB-genome tetraploid species of peanut reduced the genomic complexity and effectively enriched for genic DNA sequences.

Introduction

The cultivated peanut or groundnut (*Arachis hypogaea* L.) is a leguminous crop, taxonomically placed into genus *Arachis* with 80 other accepted species in 9 sections (Krapovickas and Gregory 1994; Valls and Simpson 2005). Section *Arachis* includes cultivated tetraploid peanut (*A. hypogaea*), 31 diploid species, and a wild or feral tetraploid species, *A. monticola* Krapov. & Rigoni. Cultivated peanut, which is endemic to South America is a

globally important crop widely grown in tropical, sub-tropical, and warm-temperate regions of the world (Stalker 1997).

Despite global economic importance, peanut has historically lagged behind in the development of genomics resources and application of molecular breeding solutions to problems of economic importance (He *et al.* 2003; Gepts *et al.* 2005). While numerous DNA markers and genetic diversity analyses are reported, only three sparse linkage maps have been published (Halward *et al.* 1993; Burow *et al.* 2001; Moretzsohn *et al.* 2005) and no QTL have been mapped. A critical scarcity of DNA polymorphisms among elite lines has severely hampered the development of DNA markers for genetic mapping, forward genetic analyses, and molecular breeding (Grieshammer and Wynne 1990; Halward *et al.* 1991, 1992, 1993; Kochert *et al.* 1991; Lacks and Stalker 1993; Stalker and Simpson 1995). The genetic narrowness has been attributed to speciation, domestication, and breeding bottlenecks (Krapovickas and Gregory 1994; Halward *et al.* 1991; Young *et al.* 1996).

Among DNA markers, SSRs have been identified as the most informative marker system in *Arachis* (Hopkins *et al.* 1999; Raina *et al.* 2001; Gimenes *et al.* 2007). SSRs are tamdemly repeated arrays of one to six nucleotides found interspersed in the genomes of the organisms. Their identification involves mining DNA sequences (either random genomic, SSR enriched genomic or cDNA libraries) for the repeat tracts. Thus far, about 800 SSR markers have been developed for peanut. Conventional genomic approaches involving SSR enrichment procedures have been frequently used for marker development (Hopkins *et al.* 1999; Palmieri *et al.* 2002; He *et al.* 2003; Ferguson *et al.* 2004; Moretzsohn *et al.* 2005; Palmieri *et al.* 2005; Bravo *et al.* 2006; Gimenes *et al.* 2007; Cuc *et al.* 2008). Although expensive and time-consuming (Karagyozov *et al.* 1993; Tang *et al.* 2002), these approaches yield higher frequencies of polymorphic SSRs

compared to the EST-based approaches (Ferguson *et al.* 2004; Luo *et al.* 2005; Moretzsohn *et al.* 2005; Cuc *et al.* 2008). Cuc *et al.* (2008) found 44.2% of genomic SSRs to be polymorphic in a study on a set of 32 cultivated peanut accessions. Likewise, 57.3% of the genomic-based markers were shown to be polymorphic in a diverse array of 24 cultivated peanut accessions by Ferguson *et al.* (2004). Expressed sequence tag-derived SSRs (EST-SSRs), on the other hand, deliver lower frequencies of polymorphic markers compared to the genomic-based SSRs (Luo *et al.* 2005; Moretzsohn *et al.* 2005). Moretzsohn *et al.* (2005) reported about 55% of genomic-derived and about 48% of EST-derived amplifiable markers polymorphic between *A. duranensis* and *A. stenosperma*, the parents of an interspecific A-genome diploid mapping population. Only 9% of the EST-SSRs were polymorphic among six accessions of cultivated peanut, compared to 41% of the genomic-SSRs (Moretzsohn *et al.* 2005). These observations furnished the assertion that genome sequence-based SSRs are comparatively more polymorphic than the EST-SSRs. Therefore, we produced and mined methylation-filtered (MF) and unfiltered (UF) genome survey sequences (GSSs) from *A. duranensis*, *A. batizocoi*, and *A. hypogaea* for SSRs. Here, we report the development and characterization of 97 SSR markers from GSS sequences. Also, we assessed the utility of methylation-filtering approaches for reducing genomic complexity and efficient genespace sequencing in diploid and tetraploid peanut. Efficiency of MF approaches for gene enrichment and genespace sequencing will be discussed in context of *Arachis* species.

Materials and Methods

Development and Mining of Genome Survey Sequences (GSSs) for SSRs

A total of 9,517 unique genome survey sequences (GSS) from methylation filtered (MF) and unfiltered (U) genomic libraries of *Arachis duranensis* Krapov. and W. C. Greg., *Arachis batizocoi* Krapov. and W. C. Greg. and *Arachis hypogaea* L. were obtained from Orion

Genomics (Saint Louis, Missouri) (Whitelaw *et al.* 2003; Bedell *et al.* 2005). Sequences were blasted against several plant protein and EST databases and were annotated at hits below 1e-5. Web based simple sequence repeat identification tool, SSRIT (Temnykh *et al.* 2001) was used for mining GSS for SSR tracts. Parameters were set for detecting perfectly repeated di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 5 repeats for each motif type. Information on repeat motif, repeat number, and SSR start- and end positions within the respective ESTs were extracted from the SSRIT output.

Primer Design

Sequences were screened for their quality and those lacking adequate SSR flanking regions were discarded. Sequences with a minimum of 9 repeats for dinucleotides, 6 repeats for trinucleotides, and 5 repeats for tetranucleotides were selected and *Primer3* (Rozen and Skaletsky 2000), a primer prediction package, was used for designing primers. Primer design parameters were set as follows: primer length from 20 to 27 with 21 base pairs as optimum, and amplification size of 100 to 500 base pairs.

Plant Materials, DNA Extraction and Marker Development

A total of 12 genotypes including 4 wild diploid and 8 cultivated tetraploid accessions from genus *Arachis* were selected (Table 2.1). *A. duranensis* represents A-genome species, *A. batizocoi* represents B-genome species, and *A. hypogaea* represents AB-genome species.

DNA was extracted from leaf tissues collected from the greenhouse grown plants. Modified CTAB method was used for DNA isolation (Murray and Thompson 1980) and Synergy HT Multi-Mode Microplate reader (BioTek Instruments, Inc.; Winooski, VT, USA) was used for DNA quantification.

Unlabeled primers were ordered and were initially screened for amplification against a bulk DNA from a total of 8 genotypes, including 4 tetraploid and 4 diploid accessions. PCR products were resolved on agarose gels and amplification information was recorded. Primers showing satisfactory amplification were used for screening the panel of 12 genotypes. PCR reactions were carried out in 96 well plates, following Tang *et al.* (2002) and genotypes were resolved in SSCP gels.

Haplotypes were scored and all statistical analysis were performed in excel spreadsheets. Polymorphism information content (PIC) values were calculated according to the formula $1-\sum p_i^2$, following Anderson *et al.* (1993).

Results and Discussion

Gene space sequencing of *Arachis* using methylation-filtration (MF) approach

In order to assess the utility of MF to reduce the genomic complexity by genic enrichment and to produce DNA for efficient genespace sequencing, MF and UF libraries for *A. duranensis*, *A. batizocoi*, and *A. hypogaea* were generated. A total of 6,528 UF and 2,989 MF non-redundant genomic DNA sequences were developed (Figure 2.1; GenBank Acc. No. DX5O5857-DX517373). The gene enrichment or filter power (FP) for MF for each species was calculated comparing percent hit from MF and UF sequences with curated *Arabidopsis* database at BLAST e values ranging from 1e-5 to 1e-20 (Bedell *et al.* 2005). Median FPs achieved by MF were 4.4, 14.6, and 5.5 for *A. duranensis*, *A. batizocoi*, and *A. hypogaea* respectively. Based on FPs and estimated genome sizes of 1,240 and 2,813 Mb for *A. duranensis* and *A. hypogaea*, enrichment predicted gene space of 279 and 478 Mb for the respective species. Significantly high gene enrichment (14.6) and a very low estimate of genespace of 65 Mb in *A. batizocoi* (951 Mb genome size) necessitated further investigation. When compared to other legumes,

enrichment following MF observed in all three *Arachis* species were higher than those observed in cowpea (4.1), soybean (3.2) and *Phaseolus* (2.4) (Timko *et al.* 2008).

Rate of gene discovery for the individual species were estimated using empirically derived results from the Orion Sorghum GeneThresher project and simulations conducted on finished *Arabidopsis* sequence. The analysis predicted that ca. 95% and greater than 99% of *Arachis* genes would be tagged by [1x] and [2x] raw sequence coverage each *Arachis* species. Assuming 650 bp reads, 478 Mb genespace, and 85% sequencing success, ca. 865,158 successful GeneThresher reads were required for [1x] coverage of the genespace in *A. hypogaea*. Similarly ca. 504,977 and ca. 325,792 sequences were estimated for [1x] coverage of genespaces of *A. duranensis* and *A. batizocoi* respectively. These results suggested feasibility of MF sequencing strategy for reducing the genomic complexity in larger genomes like *Arachis* where hypermethylated fraction of the genome is overrepresented in the genomic libraries. MF showed effective genespace enrichment and reduced representation of hypermethylated fraction of the genome in the sequence library.

Characteristics of SSRs Derived from the GSS Sequences of *Arachis* species

With MF and UF sequences, we performed BLAST against EST, protein and repeat databases and thirty-three per cent of the total sequences showed significant blast hits against putatively annotated sequences. MF sequences showed reduced representation of repetitive fraction of the genome; with an exception of *A. duranensis*, greater number of MF sequences showed significant homology with putative genic sequences (Table 2.2). In *A. duranensis*, although repeat fraction was greatly reduced with MF, sequences showing significant hits with putative genes were comparatively lower for MF (15%) than that for UF sequences (19%). On the contrary, BLAST against highly-curated *Arabidopsis* database showed reduced

representation of repetitive sequences and concurrent enrichment for the genic sequences in the methylation-filtered sequences of *A. duranensis*. Although comparisons against cross species and highly-curated databases are expected to exclude mutually unrepresented genes, comparisons with comprehensive databases are more error prone; especially at higher e values, hits against poorly annotated genes or pseudogenes might produce larger number of false positive results.

We mined 9,517 GSSs representing ca. 5.5 Mb of the *Arachis* genome and identified a total of 1,168 perfectly repeated di-, tri-, tetranucleotide motifs interspersed in 960 unique sequences; 10% of the total GSSs contained SSRs. This corresponds to the overall SSR density of about 3,380 bp per Mb and SSR frequency of approximately one SSR per 4.7 kb (0.2/kb) of the genomic sequences. Genomic derived sequences seem to be more frequent in SSRs compared to expressed sequence tags where 7.3% of the total unigenes had SSRs (unpublished). However, it is not logical to directly compare these GSSs with cDNA based sequences since 20% of these sequences represent known proteins or ESTs; preponderance of genic sequences can skew our estimates. In general, the frequencies observed in *Arachis* sequences are comparable to several other dicotyledonous species and are higher than those reported for a large number of plant species including *Arabidopsis*, *Glycine*, and *Medicago* (Mahalakshmi *et al.* 2002, Kumpatla and Mukhopadhyay 2005). Considering the higher rates of mutations in SSR sites, interrogating *Arachis* SSRs for polymorphisms could yield a greater frequency of productive markers.

Among repeat motifs, dinucleotides were predominant (80.3%), followed by trinucleotides (16.8%), and tetranucleotides (2.7%) (Figure 2.2). One of our earlier studies showed that the genic regions also had a preponderance of dinucleotide repeat motifs (53.27%) compared to trinucleotides (42.84%) in *Arachis* sequences; distribution study, however, suggested that dinucleotide repeats were more abundant in the untranslated regions (UTRs) while

trinucleotides were more frequent in the exonic regions (unpublished). More than 50% of all the motifs were repeated exactly 5 times, which is consistent with EST-SSR study. Average length of SSR was about 16 bp with almost 88% of SSRs shorter than 22 bp. The highest number of repeat for dinucleotide motif was 71 with an average of 7.26; almost 88% of dinucleotides were repeated less than 10 times. Also, more than 88% of trinucleotide motifs were repeated less than 10 times with the maximum number of repeats of 36 and a mean of 6.98. Similarly, the highest repeat numbers for tetranucleotide motifs was 17 with corresponding mean of 6.4.

Development of Polymorphic Genomic SSR Markers in *Arachis* species

A total of 248 SSRs had a minimum of five repeats; however, SSR markers could only be developed for 153 SSRs, and 97 of the 153 SSR markers amplified alleles across species and accessions and produced high-quality genotypes. *A. duranensis and A. batizocoi* contributed 21 markers each and *A. hypogaea* contributed 55 markers. Almost 93% (90 out of 97) of the markers were polymorphic in wild diploid genotypes with mean polymorphic information content (*PIC*) of 0.57. Very high frequency of polymorphic markers (about 60%) was obtained between *A. duranensis* accessions. It suggests the feasibility of developing intraspecific *A. duranensis* maps and using the species as a A-genome model for peanut genomics. On the other hand, only about 25% of the markers were polymorphic between accessions and forty percent of the markers (38 out of 97) were polymorphic among accesssions of *A. hypogaea*. Average PIC was 0.24 among the tetraploids. These observations suggested a lack of allelic variation among the tetraploids. However, compared to our study on EST-SSRs among 28 tetraploid genotypes, the frequency and average PIC of GSS-SSRs was much higher; only 32% of the amplifiable EST-SSR markers (19 out of 59) were polymorphic with an average heterozygosity of 0.11. Therefore, GSS-derived SSRs are more polymorphic than EST-SSRs in tetraploid peanut.

Cross-Species Transferability of Genomic SSRs in *Arachis*

Moderate (60-70%) to higher (>90%) levels of cross species transferability of microsatellite markers have been reported among *Arachis* species (Hopkins *et al.* 1999; Moretzsohn *et al.* 2004, 2005; Gimenes *et al.* 2007). Out of a total of 21 *A. duranensis* based markers, 7 (33%) failed to amplify in *A. batizocoi,* but all the markers were amplifiable in at least one of the *A. hypogaea* accessions. Similarly, almost all the *A. batizocoi* based markers were transferable to *A. duranensis* (20 out of 21). Although 100% of the markers from the diploids were transferable to the tetraploid peanut, only about 82% and 70% of *A. hypogaea* based markers were transferable to *A. duranensis* and *A. batizocoi*, respectively. In general, we observed moderately high level of cross species transferability of genomic microsatellite markers among *Arachis* species. However, compared to the EST-SSR study, GSS-based SSRs showed higher frequencies of null alleles, suggesting higher levels of sequence conservation in the genic regions. Similar observations have been reported in several interspecific and generic studies in *Arachis* (Gao *et al.* 2003; Varshney *et al.* 2005; Mace *et al.* 2008).

Appreciably high frequencies of soybean SSR markers are transferable to peanut (He *et al.* 2006, Sanders *et al.* 2008). This is particularly encouraging in context of initiatives in peanut envisioned by cross-legume advances in genomics conference (http://catg.ucdavis.edu, Gepts *et al.* 2005); comparative genomics approaches are crucial to extend understanding from reference legume genomes to the non-reference ones.

SSR Characteristics and Polymorphisms

Out of 97 markers, 55 (56.7%) were found in genic regions and 42 (43.3%) were found in non-genic regions of the genome; 38.1% of the non-genic markers were polymorphic while 41.8% of the genic markers were polymorphic in cultivated accessions. In wild species, the

average PIC for non-genic markers was higher (H = 0.59) than that for the genic markers (H = 0.55). However in the tetraploids, average PIC scores were comparable with averages of 0.24 and 0.23 for SSRs located in the genic and non-genic sequences, which were greater than those observed in EST-SSRs (0.11) (unpublished). Among repeat motifs, dinucleotide repeats were more polymorphic (PIC = 0.62) than trinucleotide (PIC = 0.55) and the tetranucleotide (PIC = 0.51) repeat motifs across the panel of 12 genotypes. Albeit lower in average polymorphism values, similar observation was recorded for individual repeat motifs in the tetraploid subset. We also tested the assertion that the probability of polymorphism increases with increasing length of repeats (Cho *et al.* 2000; La Rota *et al.* 2005; Temnykh *et al.* 2001). In tetraploids, we found a significant positive correlation between the length of SSRs and corresponding average heterozygosity values (r = 0.51, $p < 0.001$) (Figure 2.3); SSRs longer than 26 bp were nearly four-fold more polymorphic (PIC = 0.44) than SSRs shorter than 26 bp (PIC = 0.12).

Conclusion

In this study, genome survey sequences (GSSs) were developed, mined, and characterized for simple sequence repeats (SSRs). Also, we assessed the utility of methylation-filtering (MF) for gene enrichment and genespace sequencing. MF was shown to enrich for the genic sequences and reduce the representation of hypermethylated repetitive sequences in the genomic libraries of *A. duranensis*, *A. batizocoi*, and *A. hypogaea*. GSSs were shown to be a rich source of polymorphic SSR markers in *Arachis* species, and a total of 97 SSR markers were developed, 93 of which were polymorphic among the A-genome and B-genome diploids, and 40 were polymorphic among the tetraploids. These markers add to the repertoire of DNA marker resources needed for molecular breeding and genomic applications in peanut. Remarkably large

number (about 60%) of polymorphic markers between accessions suggested the feasibility of developing *A. duranensis* as an A-genome diploid model.

References

Anderson, J.A., G.A. Churchill, J.E. Autrique, S.D. Tanksley, and M.E. Sorrells. 1993. Optimizing parental selection for genetic linkage maps. Genome 36:181-186.

Bedell, J.A., M.A. Budiman, A. Nunberg, R.W. Citek, D. Robbins, J. Jones, E. Flick, *et al*. 2005. Sorghum genome sequencing by methylation filtration. PLoS Biol. 3(1):e13.

Bravo, J.P., A.A. Hoshino, C. Angelici, C.R. Lopes, and M.A. Gimenes. 2006. Transferability and use of microsatellite markers for the genetic analysis of the germplasm of some Arachis section species of the genus Arachis. Genetics and Molecular Biology 29:516-524.

Burow, M.D., C.E. Simpson, J.L. Starr, and A.H. Paterson. 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.

Cho, Y.G., T. Ishii, S. Temnykh, X. Chen, L. Lipovich, S.R. McCouch, W.D. Park, N. Ayres, and S. Cartinhour. 2000. Diversity of microsatellites derived from genomic libraries and Genbank sequences in rice (*Oryza sativa* L.). Theor. Appl. Genet. 100:713-722.

Cuc, L., E. Mace, J. Crouch, V. Quang, T. Long, and R. Varshney. 2008. Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (Arachis hypogaea). Bmc Plant Biology 8:55.

Ferguson, M., M. Burow, S. Schulze, P. Bramel, A. Paterson, S. Kresovich, and S. Mitchell. 2004. Microsatellite identification and characterization in peanut (A. hypogaea L.). TAG Theoretical and Applied Genetics 108:1064-1070.

Gao, L.F., J.F. Tang, H.W. Li, and J.Z. Jia. 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. Molecular Breeding 12:245-261.

Gepts, P., W.D. Beavis, E.C. Brummer, R.C. Shoemaker, H.T. Stalker, N.F. Weeden, and N.D. Young. 2005. Legumes as a model plant family. Plant Physiology 137:1228-1235.

Gimenes, M.A., A.A. Hoshino, A.V.G. Barbosa, D.A. Palmieri, and C.R. Lopes. 2007. Characterization and transferability of microsatellite markers of the cultivated peanut (Arachis hypogaea). Bmc Plant Biology 7.

Grieshammer, U., and J.C. Wynne. 1990. Isozyme variability in mature seeds of U.S. peanut cultivars and collections. Peanut Sci. 18:72-75.

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

Halward, T., H.T. Stalker, E. LaRue, G. Kochert. 1992. Use of single-primer amplifications in genetic studies of peanut (Arachis hypogaea L.). Plant Mol. Biol. 18:315-325.

Halward, T., H.T. Stalker, and G. Kochert. 1993. Development of an RFLP linkage map in diploid peanut species. Theoretical and Applied Genetics 87:379-384.

He, G., F.E. Woullard, I. Morong, and B.Z. Guo. 2006. Transferability of soybean SSR markers in peanut (*Arachis hypogaea* L.). Peanut Science. 33:22-28.

He, G., R. Meng, M. Newman, G. Gao, R. Pittman, and C.S. Prakash. 2003. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biology 3:3.

Hopkins, M.S., A.M. Casa, T. Wang, S.E. Mitchell, R.E. Dean, G.D. Kochert, and S. Kresovich. 1999. Discovery and characterization of polymorphic simple sequence repeats (SSRs) in peanut. Crop Science 39:1243-1247.

Karagyozov, L., I.D. Kalcheva, and V.M. Chapman. 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research 21:3911-3912.

Kochert, G., T. Halward, W.D. Branch, and C.E. Simpson. 1991. RFLP variability in peanut (*Arachis hypogaea* L.) cultivars and wild-species. Theoretical and Applied Genetics 81:565-570.

Krapovickas, A., and W.C. Gregory. 1994. Taxonomia del genero *Arachis* (Leguminosae). Bonplandia 8: 1-186.

Kumpatla, S.P., and S. Mukhopadhyay. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985-998.

Lacks, G.D. and H.T. Stalker. 1993. Isozyme analysis of *Arachis* species and interspecific hybrids. Peanut Sci. 20:76-81.

La Rota, M., R.V. Kantety, J.K. Yu, and M.E. Sorrells. 2005. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. Bmc Genomics 6.

Luo, M., P. Dang, B.Z. Guo, G. He, C.C. Holbrook, M.G. Bausher, and R.D. Lee. 2005. Generation of Expressed Sequence Tags (ESTs) for Gene Discovery and Marker Development in Cultivated Peanut. Crop Sci 45:346-353.

Mace, E.S., R.K. Varshney, V. Mahalakshmi, K. Seetha, A. Gafoor, Y. Leeladevi, and J.H. Crouch. 2008. 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 Science 174:51-60.

Mahalakshmi, V., P. Aparna, S. Ramadevi, and R. Ortiz. 2002. Genomic sequence derived simple sequence repeats markers a case study with Medicago spp. Electronic Journal of Biotechnology 5(3):233-242.

Moretzsohn, M., M. Hopkins, S. Mitchell, S. Kresovich, J. Valls, and M. Ferreira. 2004. Genetic diversity of peanut (Arachis hypogaea L.) and its wild relatives based on the analysis of hypervariable regions of the genome. BMC Plant Biology 4:11.

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

Murray, M.G., and W.F. Thompson. 1980. Rapid isolation of high molecular-weight plant DNA. Nucleic Acids Research 8:4321-4325.

Palmieri, D.A., A.A. Hoshino, J.P. Bravo, C.R. Lopes, and M.A. Gimenes. 2002. Isolation and characterization of microsatellite loci from the forage species Arachis pintoi (Genus Arachis). Molecular Ecology Notes 2:551-553.

Palmieri, D.A., B. M. D, C. R. A, G. M. A, and L. C. R. 2005. Novel polymorphic microsatellite markers in section Caulorrhizae (Arachis, Fabaceae). Molecular Ecology Notes 5:77-79.

Raina, S.N., V. Rani, T. Kojima, Y. Ogihara, K.P. Singh, and R.M. Devarumath. 2001. RAPD and ISSR fingerprints as useful genetic markers for analysis of genetic diversity, varietal identification, and phylogenetic relationships in peanut (*Arachis hypogaea*) cultivars and wild species. Genome 44:763-772.

Rozen, S., and H.J. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers, pp. 365-386. In S. Krawetz, and S. Misener (eds.) Bioinformatics Methods and Protocols: Methods in molecular Biology. Humana Press, Totowa, NJ.

Stalker, H.T. 1997. Peanut (*Arachis hypogaea* L.). Field Crops Research 53:205-217.

Stalker, H.T., and C.E. Simpson. 1995. Genetic resources in Arachis, pp. 14-53. In H.E. Pattee and H.T. Stalker (eds.) Advances in peanut science. Am. Peanut Res. Educ. Soc., Stillwater, OK.

Tang, S., J.K. Yu, M.B. Slabaugh, D.K. Shintani, and S.J. Knapp. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105:1124-1136.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Research 11:1441-1452.

Timko, M.P., P.J. Rushton, T.W. Laudeman, M.T. Bokowiec, E. Chipumuro, F. Cheung, C.D. Town, and X. Chen. 2008. Sequencing and analysis of the gene-rich space of cowpea. BMC Genomics 9:103.

Valls, J.F.M, and C.E. Simpson. 2005. New species of Arachis L. (Leguminosae) from Brazil, Paraguay and Bolivia. Bonplandia 14:35-64.

Varshney, R.K., A. Graner, and M.E. Sorrells. 2005. Genic microsatellite markers in plants: features and applications. Trends in Biotechnology 23:48-55.

Whitelaw, C.A., W.B. Barbazuk, G. Pertea, A.P. Chan, F. Cheung, Y. Lee, L. Zheng, *et al*. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302:2118-2120.

Young, N.D., N.F. Weeden, and G. Kochert. 1996. Genome mapping in legumes (Family Fabaceae), pp. 211-227. In A.H. Paterson (ed.) Genome mapping in plants. R.G. Landes Co., Austin, TX.

CHAPTER 3

DISCOVERY AND CHARACTERIZATION OF SIMPLE SEQUENCE REPEATS (SSRs) FROM TETRAPLOID PEANUT EXPRESSED SEQUENCE TAG (EST) DATABASE[1]

[1]Khanal, S., Y. Guo, E. Nagy, S. Tang, and S.J. Knapp. To be submitted to *BMC Genomics.*

Abstract

Narrow genetic diversity and a deficiency of polymorphic DNA markers have hindered genetic mapping and the application of translational genomics approaches and marker-assisted selection in groundnut (*Arachis hypogaea* L.). We developed and mined a groundnut expressed sequence tag (EST) database for simple sequence repeats (SSRs), assessed the frequency of polymorphic SSRs in ESTs, and developed 2,054 EST-SSR markers. The latter could play a significant role in breaking the DNA marker bottleneck in groundnut. We assembled 71,448 long-read (Sanger) ESTs and 304,215 short-read (454) ESTs into 101,132 unigenes, identified 7,413 perfect repeats, designed and tested primers for 80 EST-SSRs sampled from broad spectrum of motifs and repeat lengths, and screened 59 EST-SSR markers for polymorphisms among 28 tetraploid germplasm accessions (primarily Runner, Virginia, Spanish, and Valencia cultivars) and four diploid germplasm accessions (the parents of *A. duranensis* and *A. batizocoi* mapping populations). *Arachis* transcribed sequences showed moderately high frequency of SSRs (one/5 kb). Dinucleotide repeat motifs (53.3%) were more abundant than trinucleotide (42.8%) and tetranucleotide (2.52%) repeat motifs. Among repeat classes, $(AG)_n$ and $(AAG)_n$ were the most frequent dinucleotide and trinucleotide repeat classes, respectively. SSRs longer than 26 bp (Polymorphic Information Content = PIC = 0.26) were significantly more polymorphic than SSRs shorter than 26 bp (PIC = 0.10) among tetraploid germplasm accessions. Of the 59 EST-SSR markers, 51 (86.4%) were polymorphic among diploid and tetraploid accessions, 20 (33.8%) were polymorphic among tetraploid accessions (PIC = 0.12), and 48 (81.1%) were polymorphic in two diploid mapping populations. ESTs are a rich source of SSRs in peanut, and the frequency of polymorphic EST-SSRs seems to be more than sufficient for

developing a critical mass of DNA markers for genomics and molecular breeding applications in cultivated peanut.

Introduction

Quantifying nucleotide diversity and tagging genomes at those sites are important forward genetic approaches for biological studies. As such, DNA markers are crucial in strengthening scientific knowledge base and accelerating crop improvement endeavors. Of several classes of DNA markers, simple sequence repeats (SSRs, also microsatellites) have become a marker class of choice because they are co-dominant and highly mutable, multiallelic and locus specific, reproducible, and adaptable to high-throughput genotyping platforms (Powell *et al.*1996a; Powell *et al.*1996b).

SSRs are tamdemly repeated arrays of one to six nucleotides found interspersed in the genomes of the organisms. Their identification involves mining DNA sequence databases (either random genomic, SSR enriched genomic or cDNA libraries) for the repeat tracts. In recent years, developing and mining EST databases for microsatellites has emerged as a prominent marker development strategy in several plant species including *Arachis* (Temnykh *et al.*2001; Kantety *et al.*2002; La Rota *et al.*2005; Luo *et al.*2005; Moretzsohn *et al.*2005; Han *et al.*2006; Proite *et al.*2007).

Arachis hypogaea L., the cultivated peanut, belongs to genus *Arachis* under the family of legumes viz. *Fabaceae*. It is the twelfth most important food crop (FAOSTAT, http://faostat.fao.org), extensively grown in tropical, sub-tropical, and warm-temperate regions of the world (Stalker 1997). Despite abundant morphological variation, polymorphism studies with several classes of molecular markers have reported lower levels of nucleotide diversity within the species (Grieshammer and Wynne 1990; Halward *et al.* 1991, 1992; Kochert *et al.*1991;

Paik-Ro *et al.*1992; Stalker *et al.*1994; He and Prakash 1997). Relative paucity of nucleotide diversity in cultivated peanut has been attributed to a genetic bottleneck inherent with interspecific hybridization leading to speciation due to reproductive isolation of an amphidiploid (Hopkins *et al.* 1999). Domestication, selection and varietal homogeneity rapidly eroded the genetic variation which was further exhausted by rearing susceptible genotypes under chemical intensive production systems (Stalker 1997). Also in developed countries, rigid demands for specific kernel/pod characteristics have made it difficult to use more diverse lines in production (personal communication A. K. Culbreath). Wild *Arachis* spp., on the other hand, evolved under natural selection, consequently retaining greater amount of genetic diversity and represent potential sources of desirable traits in cultivated peanut. Introgressions of these traits offer molecular breeding solutions to the pertinent problems of cultivated peanut (Kochert *et al.* 1991). However, lack of adequate genomics resources has severely impeded application of forward genetic approaches in *Arachis* spp., and quantifying and exploiting nucleotide diversity appear crucial fundamental steps towards accelerating crop improvement endeavors. As such, a critical mass of molecular markers mapped onto saturated linkage maps is an undisputed genomics asset which facilitates targeted gene cloning and marker assisted selection besides other downstream applications as functional genetic studies, comparative genomics, and molecular evolutionary studies.

Although various molecular marker systems viz. Isozymes (Lu and Pickersgill 1993; Stalker *et al.* 1994), RFLPs (Kochert *et al.* 1991; Halward *et al.* 1993; Burow *et al.* 2001), RAPDs (Halward *et al.* 1991; Halward *et al.* 1992; Hilu and Stalker 1995; Raina *et al.* 2001), AFLPs (He and Prakash 2001; Gimenes *et al.* 2002; Herselman 2003), and SSRs have been developed, only three linkage maps have been published for *Arachis* spp. Besides, cultivated

tetraploid based linkage maps have not been reported yet. The published linkage maps include two RFLP based diploid (Halward *et al.* 1993) and tetraploid (Burow *et al.* 2001) maps and one SSR based diploid (Moretzsohn *et al.* 2005) map; the only tetraploid linkage map is based on a backcross population between a synthetic amphidiploid and a tetraploid cultivated *Arachis hypogaea* (Burow *et al.* 2001).

SSRs have been identified as the most informative marker system in *Arachis* spp. (Hopkins *et al.* 1999; Raina *et al.* 2001; Gimenes *et al.* 2007). Conventional genomic approaches involving SSR enrichment procedures have been frequently used for marker development (Hopkins *et al.* 1999; Palmieri *et al.* 2002, 2005; He *et al.* 2003; Ferguson *et al.* 2004; Moretzsohn *et al.* 2005; Bravo *et al.* 2006; Gimenes *et al.* 2007; Cuc *et al.* 2008). These approaches are expensive and time-consuming (Karagyozov *et al.* 1993; Tang *et al.* 2002; He *et al.* 2003). Recently, relatively simple and inexpensive approaches as EST database mining (Luo *et al.* 2005; Moretzsohn *et al.* 2005; Proite *et al.* 2007) and inter-generic transfers of SSRs (He *et al.* 2006; Mace *et al.* 2008) have been reported. To date, several hundred polymorphic SSRs have been developed and reported for *Arachis* species. In an earlier study, we screened 709 previously reported SSR markers against parents of several diploid and tetraploid mapping populations (unpublished). Our results suggested a lack of critical mass of polymorphic markers for developing saturated linkage maps in *Arachis hypogaea*. However, development and mining of peanut EST database offered prospect for a large scale marker development. In this study, we report (1) the discovery and characteristics of 7,413 SSRs from peanut EST database, (2) the development of 59 EST-SSR markers and their polymorphisms, (3) the polymorphisms offered by different repeat motifs, repeat lengths and repeat locations of SSRs, and (4) the population structure of a diverse panel of 28 tetraploid peanut genotypes.

Materials and Methods

Mining Peanut EST Database for SSRs

The EST sequence data were downloaded in fasta format from PeanutDB (http://www.peanut.uga.edu, unpublished). This database harbors a total of 71,448 long-read (Sanger) ESTs and 304,215 short-read (454) ESTs assembled into 101,132 unigenes. Sequences were mined for SSRs using web-based SSR identification tool, SSRIT (http://www.gramene.org/db.searches/ssrtool; Temnykh *et al.* 2001). Parameters were set for detecting perfectly repeated di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 5 repeats for each motif type. Information on repeat motif, repeat number, and SSR start- and end positions within the respective ESTs were extracted from the SSRIT output. The grouping of SSR motifs into respective repeat classes was performed following the method of Jurka and Pethiyagoda (1995), where $(AG)_n$ is equivalent to $(GA)_n$, $(TC)_n$, and $(CT)_n$, while $(TTC)_n$ is equivalent to $(TCT)_n$, $(CTT)_n$, $(AAG)_n$, $(AGA)_n$, and $(GAA)_n$ in different reading frames or on a complementary strand. Accordingly, frequencies of different repeat classes for di-, and trinucleotide repeat motifs were analyzed.

SSR distribution observation was based on the length of a sequence, SSR start and SSR end positions, start and end of the open reading frame (exons), and the reading frame itself. Blast hits against Mt2.0, a sequence assembly released by *Medicago* Genome Sequence Consortium (MGSC, http://www.medicago.org/genome) and corresponding reading frames were recorded for each sequence. Start and end positions of open reading frames were predicted using OrfPredictor (https://fungalgenome.concordia.ca/tools/OrfPredictor.html; Min *et al.* 2005). EST-SSRs selected for SSR distribution study included long-read (Sanger) sequences showing significant

blast hit with a cutoff value of E^{-30} and with reading frames at least 30 bp from their start and the end.

Primer Design

A subset of 80 EST-SSRs representing a broad spectrum of repeat motifs, repeat lengths and repeat locations was selected. Major selection criteria included design of primers based on stringent parameters and higher Phred scores at primer landing sites. Primer3 (http://frodo.wi.mit.edu; Rozen and Skaletsky 2000), a primer prediction package was used for designing primers. Primer design parameters were set as follows: primer length from 19 to 23 with 21 base pairs as optimum, amplification size of 100 to 400 base pairs, annealing temperatures from 59°C to 63°C with a maximum difference of 3°C, and GC contents from 25% to 45%. Forward primers were labeled with one of the three fluorophores viz. 6FAM, HEX, and TAMRA.

Plant Materials, DNA Extraction and Marker Development

A total of 32 genotypes including 28 tetraploid and 4 diploid accessions from genus *Arachis* were used for EST-SSR marker amplification and length polymorphisms (Table 3.1). Tetraploid peanut genotypes represent four botanical market classes (Runner, Virginia, Valencia and Spanish), historical accessions and exotic species. Diploid genotypes represent two accessions each of *A. duranensis* and *A. batizocoi*, the parents of 'A' and 'B' genome intra-specific mapping populations. Leaf tissues collected from the greenhouse grown plants were used for DNA extraction. Modified CTAB (cetyltrimethylammonium bromide) method was used for DNA isolation (Murray and Thompson 1980) and Synergy HT Multi-Mode Microplate reader (BioTek Instruments, Inc.; Winooski, VT, USA) was used for DNA quantification.

Primers were initially screened for amplification against bulk DNA from 6 genotypes. PCR products were resolved on agarose gels and sizes of amplification were recorded. Markers showing amplification products within the range of genotyping (≤ 500 bp) were screened for polymorphisms against 32 peanut genotypes. PCR reactions essentially followed the methodology as described by Tang *et al.*(2002). Genotypes were determined using the ABI 3730 DNA analyzer and GeneMapper Software Version 4 (Applied Biosystems, Foster City, CA).

Marker Characteristics, Statistical Analysis, and Population Structure

Polymorphic markers were identified by the presence or absence of bands amplified by the individual markers across the genotype panel. Band frequency, total number of polymorphic bands and average number of bands per marker were recorded. For multilocus genotype data of polyploid species, data is often entered as a binary matrix and treated as dominant markers since an estimate of the exact number of copies of individual alleles is difficult (Mengoni *et al.* 2000; Gupta *et al.* 2003). Therefore, input was prepared as a rectangular binary data matrix wherein 1 represented the presence and 0 represented the absence of a particular fragment size or a null allele. Major band frequency, gene diversity and polymorphism information content (PIC) for individual markers were obtained using software tool *PowerMarker V3.25* (Liu and Muse 2005).

Model-based clustering software package *structure 2.2* (Pritchard *et al.* 2000; Falush *et al.* 2003, 2007) was used for deciphering the population structure of 28 tetraploid peanut genotypes. It implements Bayesian clustering algorithm to probabilistically assign individuals to populations and is extensively used because of its versatility in analyzing molecular data generated from different marker systems. The input matrix included 145 loci data corresponding to 59 markers screened against 28 peanut genotypes. Default settings of ancestry with admixture and allele frequencies correlated models were used for running the program. Several simulations

at different burning lengths and MCMS replications together with several independent runs at different population sizes (designated as K) were performed to obtain appropriate burnin length and MCMC replications. Inferences on an appropriate population size were drawn basically following instructions from the *structure* manual and Evanno *et al.* (2005).

Microsat (Minch *et al.* 1997) was used for the construction of distance matrices based on the proportion of shared bands ($D = 1 - p_s$). The matrices were imported into *Phylip v3.67* (Felsenstein 1989) for the construction of neighbor-joining trees and the trees were edited and colored using software program *TreeDyn 198.3* (Chevenet *et al.* 2006). Principle coordinate analysis (PCoA) was performed using Microsoft excel based software *GenAlEx 6.1* (Peakall and Smoush 2006).

Results

Characteristics of *A. hypogaea* EST-SSRs

We mined 101,132 unigenes representing ca. 37 Mb of the *A. hypogaea* genome and identified a total of 7,413 perfectly repeated di-, tri-, tetra-, penta-, and hexanucleotide motifs interspersed in 6,371 unigenes; 7.3% of the total unigenes contained SSRs. This corresponds to the overall SSR density of 3189.9 bp per Mb and SSR frequency of approximately one SSR per 5 kb (0.2/kb) of genic sequences. The abundance of SSRs in peanut ESTs is comparable to several other dicotyledonous species and higher than those reported for large number of plant species (Kumpatla and Mukhopadhyay 2005). The average lengths of unigenes containing more than one SSR (mean = 792.7) were significantly higher ($P < 0.01$) than those containing single SSR (mean = 590.7); 856 unigenes had multiple SSRs.

Among repeat motifs, dinucleotides were predominant (53.27%), followed by trinucleotides (42.84%) and tetranucleotides (2.52%); frequencies of penta- (0.7%) and

hexanucleotide repeat motifs (0.66%) were low (Figure 3.1). SSR distribution study showed over-representation of untranslated regions (UTRs) in SSR containing EST sequences of *Arachis*; more than 76% of the SSRs were located in the UTRs. Dinucleotide repeat motifs were predominantly distributed in the UTRs while trinucleotide repeat motifs were more frequent in exons compared to both the UTRs (Figure 3.2). Average number of repeats for dimers in exons was shorter than those in the UTRs. In general, microsatellites in the exonic regions were relatively shorter in length compared to those in the UTRs.

More than 50% of all the motifs were repeated exactly 5 times. Average length of SSR was about 16 bp with almost 88% of SSRs shorter than 22 bp. The highest number of repeat for dinucleotide motif was 70 with an average of 6.84; almost 88% of dinucleotides were repeated less than 10 times. Also, more than 88% of trinucleotide motifs were repeated less than 8 times with the maximum number of repeats of 27 and a mean of 6. Similarly, the highest repeat numbers for tetra-, penta-, and hexanucleotides were 16, 7, and 23, respectively, with corresponding means of 5.8, 5.2, and 5.9. Repeat length distribution revealed sharp decrease in the frequencies of repeat motifs with increasing SSR lengths (Figure 3.3).

Of four different classes of dinucleotide repeat motifs, $(AG)_n$ was the most frequent class (61.02%) while $(CG)_n$ repeat class was the least frequent one (Figure 3.4.). $(AG)_n$ have been shown as the most abundant repeat class both in cereal and leguminous species (Jayashree *et al.* 2006), as well as in 48 out of 49 dicotyledonous species studied by Kumpatla and Mukhopadhyay (2005). Out of 10 different repeat motif classes, $(AAG)_n$ was the most frequent trinucleotide repeat class (35.2%) while $(ACG)_n$ was the least frequent one (Figure 3.5).

Polymorphism of EST-SSR Markers in Tetraploid *Arachis* spp.

Out of a total of 80 markers, 59 were found suitable for genotyping and subsequent analysis. Almost 34% (20 out of 59) of the markers were polymorphic in tetraploid peanut. However, only about 10% of the markers were polymorphic between any two tetraploid accessions. We obtained a mean PIC of 0.11 and an average gene diversity (GD) of 0.12 (averaged for all the markers). Band frequency ranged from 0.11 to 1 with a mean of 0.77 and an average of 2.45 bands per maker. 58 markers produced a total of 145 bands (fragments), 62 (42.75%) of which were polymorphic.

We also tested the assertion that the probability of polymorphism increases with increasing length of repeats (Cho *et al.* 2000; Temnykh *et al.* 2001; La Rota *et al.* 2005). Although there was a weak positive correlation between the length of SSR and corresponding PIC values ($r = 0.24$; $p < 0.02$) (Figure 3.6), SSRs longer than 26 bp were three-fold more polymorphic (PIC = 0.18) than SSRs shorter than 26 bp (PIC = 0.06). Among repeat motifs, dinucleotides and trinucleotides showed equal levels of polymorphisms (PIC = 0.14). Also, the location of a repeat motif within the genic regions did not seem to affect polymorphism as SSRs in exons (CDS) and in UTRs were equally polymorphic (PIC = 0.23).

Cross-Species Transferability of EST-SSRs and Allelic Diversity in Diploid *Arachis* spp.

Although *A. hypogaea* EST sequences were used for designing primers flanking SSRs, the transferability of these markers was almost perfect with insignificant number of null alleles recorded for the diploid accessions. Also, very high rates of polymorphisms were observed among the diploid mapping populations with thirty-four (57.6%) markers polymorphic between the two *A. duranensis* accessions (PI 475887 and PI 497483) and 42 (71.2%) markers polymorphic between the two *A. batizocoi* accessions (Grif 15031 and PI 468326). Collectively,

48 (81.3%) markers were polymorphic among the four diploid accessions with an average PIC of 0.50.

Population Structure and Cluster Analysis in Tetraploid *Arachis* Species

Structure simulations gave an appropriate burnin length and MCMS replication of 100,000 each. The most appropriate population size repeatedly observed across different simulations and independent runs was two (Figure 3.7). Individuals were grouped into two populations with 55% of the genotypes in one and 45% of the individuals in the other population.

One population was primarily comprised of runner cultivars with minimal admixture from the other population. The other population was primarily comprised of Spanish cultivars. Different genotypes identified as Runners were allocated to a population with individuals showing very little admixture from the other population. Genotypes identified as Virginia and its derivatives, however, showed admixture in different proportions. NC-12C (PI 596406) and Indio (PI 121067) did not show much admixture while exotic introgression derivatives GP-NC WS-14 (PI 619178) and NC94022 showed higher levels of admixture. Similarly, Valencia cultivars also showed different proportions of admixture. Among Spanish cultivars, GT-C20 showed about 50% of admixture while the other three varieties showed very limited admixture. Figure 3.7 presents graphical display of structure results for 28 tetraploid genotypes.

Neighbor-joining tree (Figure 3.8) and principal coordinate analysis (PCoA) corroborated the structure observations. In principal coordinate analysis, cultivated accessions belonging to Runner, Virginia, and Valencia were separated from Spanish and most of the other genotypes along the principal coordinate one (Figure 3.9). 56.62% of the total variation was explained by the first two axes in PCoA with 36.39% and 20.23% contribution from the first and the second components respectively.

Discussion

The Abundance, Distribution, and Characteristics of *Arachis* EST-SSRs

In recent years, mining EST databases for microsatellites has emerged as a prominent marker development strategy. Rapid advancement in genetic technologies and concurrent decrease in expenses for routine genomics applications has facilitated *in silico* EST database mining for SSRs. Genomic approaches for identifying and developing microsatellite markers are resource intensive (Karagyozov *et al.* 1993; Tang *et al.* 2002; He *et al.* 2003). Further, larger volumes of microsatellite markers are developed through EST database mining, which far supersedes genomic approaches (Morgante *et al.* 2002). Therefore, the use of EST derived SSRs in genomics and molecular breeding applications in several plant species has surged in recent years.

EST-SSRs identified in this study offer possibility of unblocking the deadlock due to the lack of adequate genomics resources for molecular breeding applications in cultivated peanut. Higher frequency of SSRs observed in this study is suggestive of abundance of SSRs in single or a low-copy region of the genome and is in agreement with Morgante *et al.* (2002). However, in absence of comparable studies where *Arachis* genomic sequences are mined for SSRs, we cannot assert that high-copy regions, or specifically the intergenic regions, are poor in SSRs. Also, when compared to methylation-filtered (MF) and unfiltered (UF) genome survey sequences (GSSs), SSR frequency was found to be greater for GSSs (10%) compared to that of ESTs (unpublished); the former, however, represented both the genic and non-genic regions thereby complicating straightforward comparisons between the two. Kumpatla and Mukhopadhyay (2005) observed the frequency of SSRs in a range of 2.65% to 10.62% in 18 dicotyledonous species (with > 10,000 EST sequences); interestingly, our observation on SSR frequency (7.3%) is higher than

those for several other species including *Arabidopsis thaliana* (5.04%) (Kumpatla and Mukhopadhyay 2005).

Among repeat motifs, we found that dinucleotides are more frequent than trinucleotides in the genic sequences of *Arachis.* However, prevalence of trinucleotide repeats and lower frequencies of other motif types in open reading frames have been shown in SSR distribution studies (Morgante *et al.* 2002; Lawson and Zhang 2006). Our observation is also in contrast to the earlier reports in *Arachis* where trinucleotides are shown to be more frequent than the dinucleotide repeat motifs (Luo *et al.* 2005; Moretzsohn *et al.* 2005). However, preponderance of dimeric repeats in EST sequences has been reported in a large number of plant species including *Arachis* (Kumpatla and Mukhopadhyay 2005; Proite *et al.* 2007). Considering the mode of slippage-mediated mutations, it is unlikely that a large proportion of the dinucleotides would be present in the coding regions; mutations in trinucleotides or their multiples would only be tolerated in the coding regions, since they do not disturb the open reading frame (Katti *et al.* 2001). Discrepancies observed in various studies could be explained by the degree of representation of dinucleotide rich untranslated regions (UTRs) in the genic sequences used in different studies. Our study showed that dinucleotide repeats were mostly located in the UTRs while trinucleotide repeats were predominant in the coding regions, attributing the preponderance of dimeric repeats in *Arachis* EST database to the over-representation of dinucleotide rich UTRs.

Number of repetitions of a microsatellite in the genic sequence of *Arachis* has been suggested to be much smaller than that in its genomic sequence (Moretzsohn *et al.* 2005). However, comparison between this study (ESTs) and a study on GSSs (unpublished) did not suggest repeat number variations between two sources of SSRs (mean differences not significant

at $P < 0.05$). More than 50% of all the motifs were repeated 5 times and about 88% of the SSRs were shorter than 22 bp. The length distribution of repeat motifs indicated that lower number of motifs were more frequent than the higher number of repeats. Similar observations have been reported in different taxonomic groups of plants (Berube *et al.* 2007; Katti *et al.* 2001). Katti *et al.* (2001) cites downward mutation bias, short persistence time, and higher frequency of contraction mutations for longer alleles as the reasons for relative paucity of longer microsatellites. Because more than 88% of di- and trinucleotide repeat motifs were repeated less than 10 times, we suspect certain *in vivo* mechanisms conferring physical limit to SSR expansion in the genic regions.

In *Arachis* ESTs, $(AG)_n$ repeat class is the most frequent dimeric group followed by $(AT)_n$, while $(AAG)_n$ repeat class is the predominant trimeric group followed by $(AGT)_n$. $(CG)_n$ among dinucleotides, and $(ACG)_n$ among trinucleotides are the least frequent groups of respective repeat motif types. Interestingly, $(ACG)_n$ group is the most frequent of all the trinucleotides in EST sequences of some conifers (Berube *et al.* 2007) and CCG repeat group, one of the least frequent groups in our study, has been reported as the most frequent one in monocotyledonous species (Crane 2007; Varshney *et al.* 2002). Relative paucity of CCG trimeric groups could be attributed to the lower G+C content of dicotyledonous species compared to the monocotyledonous species (Morgante *et al.* 2002). Among dinucleotides, $(AG)_n$ has been reported as the most frequent motif class in majority of plant species except for some conifers and *Gossypium*, where $(AT)_n$ is the predominant one (Kumpatla and Mukhopadhyay 2005; Berube *et al.* 2007; Crane 2007). $(CG)_n$ repeat group, on the other hand, is exclusively the least frequent of all dinucleotide repeat motifs across different taxonomic groups of plants (Crane 2007).

Polymorphisms of *Arachis* EST-SSRs

Despite conspicuous morphological differences (Upadhyaya *et al.* 2001, 2003), levels of molecular variation among tetraploid peanut are low (Grieshammer and Wynne 1990; Halward *et al.* 1991; Kochert *et al.* 1991; Halward *et al.* 1992; Paik-Ro *et al.* 1992; Stalker *et al.* 1994; He and Prakash 1997). This apparent paucity of nucleotide diversity in tetraploid peanut has been attributed to the reproductive isolation, domestication, and genetic bottlenecks following speciation of an amphidiploid (Hopkins *et al.* 1999). Although average polymorphisms from SSR markers are not very high, they are hitherto the most informative marker system in cultivated peanut (Hopkins *et al.* 1999; Gimenes *et al.* 2007). Our observations on average PIC and mean gene diversity were similar to earlier reports in tetraploid *Arachis* (Moretzsohn *et al.* 2005; Gimenes *et al.* 2007); albeit, the frequency of polymorphic SSRs among tetraploid accessions was comparatively high. However, only 10% of the markers were actually polymorphic between two tetraploid accessions.

Genomic-based SSRs are shown to be comparatively more polymorphic than the EST-based SSRs. In *Arachis* spp., genomic-based approaches have yielded higher frequencies of polymorphic SSRs compared to the EST-based approaches (Ferguson *et al.* 2004; Luo *et al.* 2005; Moretzsohn *et al* .2005; Cuc *et al.* 2008). Cuc *et al.* (2008) found 44.2% of genomic SSRs to be polymorphic in a study on a set of 32 cultivated peanut accessions. Likewise, 57.3% of the genomic-based markers were shown to be polymorphic in a diverse array of 24 cultivated peanut accessions by Ferguson *et al.* (2004). EST-SSR based studies, on the other hand, have shown to deliver lower frequencies of polymorphic markers compared to the genomic-based SSRs (Luo *et al.* 2005; Moretzsohn *et al.* 2005). Moretzsohn *et al.* (2005) reported about 55% of genomic-derived and about 48% of EST-derived amplifiable markers polymorphic between *A. duranensis*

and *A. stenosperma*, the parents of the A-genome diploid mapping population. Further, compared to 41% of the genomic-derived amplifiable markers, only 9% of the EST-derived SSRs were found polymorphic in six accessions of cultivated peanut. These observations furnished the assertion that genome sequence-based SSRs are more prolific compared to EST-based SSRs. However, moderate to higher levels of polymorphisms of EST-SSRs compared to those derived from genomic DNA libraries have been reported in several species (reviewed by Kumpatla and Mukhopadhyay 2005). In an study where methylation filtered (MF) and unfiltered (UF) genome survey sequences were used for marker development, 41.81% of the genic SSRs (annotated) and 38.09% of the non-genic SSRs were polymorphic in tetraploid accessions; average polymorphism for the genic SSRs (0.24), was also comparable to that for the non-genic SSRs (0.23) (unpublished). In this study, we observed lower frequency of polymorphic markers from the EST-based SSRs when compared to the earlier reports from genomic-derived SSRs. However, a comparison of our result with similar studies concerning EST-SSRs suggested that the frequency of polymorphic markers is higher than what has been estimated earlier. These results suggest that the development of a critical mass of polymorphic SSR markers for routine genetic and molecular breeding applications in tetraploid peanut is possible with EST mining approaches.

Cross-Species Transferability of *Arachis* EST-SSRs

Higher levels of transferability of EST-derived SSRs have been reported in several interspecific and generic studies in *Arachis* (Gao *et al.* 2003; Varshney *et al.* 2005; Mace *et al.* 2008). EST-SSRs are shown to be more transferable than genomic SSRs. In our study almost all the markers developed from *Arachis hypogaea* sequences were amplified in both the A- and B-genomes. In another study, methylation filtered (MF) and unfiltered (UF) genome survey

sequences from *A. duranensis* (A-genome), *A. batizocoi* (B-genome) and *A. hypogaea* (AB-genome) were mined for SSRs and cross-species transferability of the markers was studied (unpublished). About 70% to 80% of the markers developed from *A. hypogaea* were amplified in the diploid genomes, while all the markers were transferable to the tetraploid spp. Other studies involving SSR enrichment procedures also showed similar levels of transferability (Hopkins *et al.* 1999; Moretzsohn *et al.* 2004, 2005; Gimenes *et al.* 2007). These observations suggest that genic regions between *Arachis* spp. have higher levels of conservation of nucleotide sequences. As such, the paucity of genomics resources in several peanut species could be offset by using EST-SSR markers developed from tetraploid peanut and *vice versa*. Further, the reports of appreciably high frequencies of soybean SSR markers transferable to peanut (He *et al.* 2006, Sanders *et al.* 2008) are particularly encouraging in context of initiatives in peanut envisioned by cross-legume advances in genomics conference (http://catg.ucdavis.edu, Gepts *et al.* 2005); comparative genomics approaches are crucial to extend understanding from reference legume genomes to the non-reference ones. Further, transferable microsatellites can be used to evaluate genetic variability and to decipher evolutionary relationships between different legumes as well as among *Arachis* species.

Genetic Diversity in Tetraploid Peanut Germplasm

Mace *et al.* (2006) studied SSR diversity among 46 genotypes, representing six botanical groups of peanut and suggested that the variety designations might not truly reflect gross genetic diversity in cultivated peanut. We also observed that the botanical market classes were not clustered together, and most of the genotypes showed admixed in different proportions. This is reasonable as most of the cultivars were actually developed from individuals belonging to different populations. For example, both Georgia Green and NemaTAM have Florunner in their

immediate ancestry and Florunner, in turn, shares its parentage in Early Runner and a Spanish variety, Florispan. As such, it is difficult to assert that 19 polymorphic markers were insufficient to allocate genotypes from the same botanical class into the same population cluster. Growth habit and other phenotypic differences could be due to domestication and induced selection. Nucleotide diversity at a limited number of loci could bring about greater differences at the phenotypic levels which, however, might not represent the major differences between the genotypes (Kochert *et al.* 1991). Even the differences at protein profiles of peanut seeds might not reflect true nucleotide diversity; differentially expressed genes or the product of post-translational modifications can bring about such changes (Liang *et al.* 2006, Kottapalli *et al.* 2008). Therefore, to decipher the magnitude of genetic diversity and to understand evolutionary relationship among different cultivated peanut accessions, a large number of polymorphic markers showed be screened against several representatives of each botanical variety.

Conclusion

In this study, expressed sequence tags were developed, mined, and characterized for simple sequence repeats. More than 7,000 SSRs were identified in the peanut EST database. From the pilot study, ESTs were shown as a rich source of polymorphic SSRs in peanut, and the frequency of polymorphic EST-SSRs seemed to be more than sufficient for developing a critical mass of DNA markers for genomics and molecular breeding applications in cultivated peanut. Therefore, we designed and tested 2,054 EST-SSRs for genotyping utility and polymorphisms among diploid and tetraploid lines (unpublished). These markers will offset the DNA marker bottleneck for peanut improvement.

References

Berube, Y., J. Zhuang, D. Rungis, S. Ralph, J. Bohlmann, and K. Ritland. 2007. Characterization of EST SSRs in loblolly pine and spruce. Tree Genetics & Genomes 3:251-259.

Bravo, J.P., A.A. Hoshino, C. Angelici, C.R. Lopes, and M.A. Gimenes. 2006. Transferability and use of microsatellite markers for the genetic analysis of the germplasm of some Arachis section species of the genus Arachis. Genetics and Molecular Biology 29:516-524.

Burow, M.D., C.E. Simpson, J.L. Starr, and A.H. Paterson. 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.

Chevenet, F., C. Brun, A.L. Banuls, B. Jacq, and R. Christen. 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. Bmc Bioinformatics 7.

Cho, Y.G., T. Ishii, S. Temnykh, X. Chen, L. Lipovich, S.R. McCouch, W.D. Park, N. Ayres, and S. Cartinhour. 2000. Diversity of microsatellites derived from genomic libraries and Genbank sequences in rice (*Oryza sativa* L.). Theor. Appl. Genet. 100:713-722.

Crane, C.F. 2007. Patterned sequence in the transcriptome of vascular plants. Bmc Genomics 8.

Cuc, L., E. Mace, J. Crouch, V. Quang, T. Long, and R. Varshney. 2008. Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (Arachis hypogaea). Bmc Plant Biology 8:55.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14(8):2611-2620.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure: extensions to linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Falush, D., M. Stephens, and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes 7:574-578.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164-166.

Ferguson, M., M. Burow, S. Schulze, P. Bramel, A. Paterson, S. Kresovich, and S. Mitchell. 2004. Microsatellite identification and characterization in peanut (A. hypogaea L.). TAG Theoretical and Applied Genetics 108:1064-1070.

Gao, L.F., J.F. Tang, H.W. Li, and J.Z. Jia. 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. Molecular Breeding 12:245-261.

Gepts, P., W.D. Beavis, E.C. Brummer, R.C. Shoemaker, H.T. Stalker, N.F. Weeden, and N.D. Young. 2005. Legumes as a model plant family. Plant Physiology 137:1228-1235.

Gimenes, M.A., A.A. Hoshino, A.V.G. Barbosa, D.A. Palmieri, and C.R. Lopes. 2007. Characterization and transferability of microsatellite markers of the cultivated peanut (Arachis hypogaea). Bmc Plant Biology 7.

Gimenes, M.A., C.R. Lopes, and J.F.M. Valls. 2002. Genetic relationships among Arachis species based on AFLP. Genetics and Molecular Biology 25:349-353.

Grieshammer, U., and J.C. Wynne. 1990. Isozyme variability in mature seeds of U.S. peanut cultivars and collections. Peanut Sci. 18:72-75.

Gupta, P.K., S. Rustgi, S. Sharma, R. Singh, N. Kumar, H.S. Balyan. 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Gen Genomics 270:315-323.

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

Halward, T., H.T. Stalker, E. LaRue, G. Kochert. 1992. Use of single-primer amplifications in genetic studies of peanut (Arachis hypogaea L.). Plant Mol. Biol. 18:315-325.

Halward, T., H.T. Stalker, and G. Kochert. 1993. Development of an RFLP linkage map in diploid peanut species. Theoretical and Applied Genetics 87:379-384.

Han, Z.G., C.B. Wang, X.L. Song, W.Z. Guo, J.Y. Gou, C.H. Li, X.Y. Chen, and T.Z. Zhang. 2006. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theoretical and Applied Genetics 112:430-439.

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

He, G.H., and C. Prakash. 2001. Evaluation of genetic relationships among botanical varieties of cultivated peanut (Arachis hypogaea L.) using AFLP markers. Genetic Resources and Crop Evolution 48:347-352.

He, G., R. Meng, M. Newman, G. Gao, R. Pittman, and C.S. Prakash. 2003. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biology 3:3.

He, G., F.E. Woullard, I. Morong, and B.Z. Guo. 2006. Transferability of soybean SSR markers in peanut (*Arachis hypogaea* L.). Peanut Science. 33:22-28.

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

Hilu, K.W., and H.T. Stalker. 1995. Genetic relationships between peanut and wild species of Arachis sect Arachis (Fabaceae): Evidence from RAPDs. Plant Systematics and Evolution 198:167-178.

Hopkins, M.S., A.M. Casa, T. Wang, S.E. Mitchell, R.E. Dean, G.D. Kochert, and S. Kresovich. 1999. Discovery and characterization of polymorphic simple sequence repeats (SSRs) in peanut. Crop Science 39:1243-1247.

Jayashree, B. R. Punna, P. Prasad, K. Bantte, C.T. Hash, S. Chandra, D.A. Hoisington, and R.K. Varshney. 2006. In Silico Biol. 6:607-620.

Jurka, J., and C. Pethiyagoda. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. 40(2): 120-126.

Kantety, R.V., M. La Rota, D.E. Matthews, and M.E. Sorrells. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Molecular Biology 48:501-510.

Karagyozov, L., I.D. Kalcheva, and V.M. Chapman. 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research 21:3911-3912.

Katti, M.V., P.K. Ranjekar, and V.S. Gupta. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution 18:1161-1167.

Kochert, G., T. Halward, W.D. Branch, and C.E. Simpson. 1991. RFLP variability in peanut (*Arachis hypogaea* L.) cultivars and wild-species. Theoretical and Applied Genetics 81:565-570.

Kottapalli, K.R., P. Payton, R. Rakwal, G.K. Agrawal, J. Shibato, M. Burow, and N. Puppala. 2008. Proteomics analysis of mature seed of four peanut cultivars using two-dimensional gel electrophoresis reveals distinct differential expression of storage, anti-nutritional, and allergenic proteins. Pl. Sci. 175:321-329.

Kumpatla, S.P., and S. Mukhopadhyay. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985-998.

La Rota, M., R.V. Kantety, J.K. Yu, and M.E. Sorrells. 2005. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. Bmc Genomics 6.

Lawson, M., and L. Zhang. 2006. Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biology 7:R14.

Liang, X.Q., M. Luo, C.C. Holbrook, and B.Z. Guo. Storage protein profiles in Spanish and runner marker type peanuts and potential markers. BMC Plant Biol. 6:24.

Liu, K.J., and S.V. Muse. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128-2129.

Lu, J., and B. Pickersgill. 1993. Isozyme variation and species relationships in peanut and its wild relatives (*Arachis* L. - Leguminosae). Theoretical and Applied Genetics 85:550-560.

Luo, M., P. Dang, B.Z. Guo, G. He, C.C. Holbrook, M.G. Bausher, and R.D. Lee. 2005. Generation of Expressed Sequence Tags (ESTs) for Gene Discovery and Marker Development in Cultivated Peanut. Crop Sci 45:346-353.

Mace, E.S., R.K. Varshney, V. Mahalakshmi, K. Seetha, A. Gafoor, Y. Leeladevi, and J.H. Crouch. 2008. 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 Science 174:51-60.

Min, X.J., G. Butler, R. Storms, and A. Tsang. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research 33:W677-W680.

Minch, E., A. Ruiz-Linares, D. Goldstein, M. Feldman & L.L. Cavalli-Sforza, 1997. MICROSAT: A Computer Program for Calculating Various Statistics on Microsatellite Allele Data, ver. 1.5d. Stanford University, Stanford, CA. Available at: http://hpgl.stanford.edu/projects/microsat/.

Mengoni A., A. Gori, M. Bazzicalupo. 2000. Use of RAPD and microsatellite (SSR) variation to assess genetic relationships among populations of tetraploid alfalfa, Medicago sativa. Plant Breeding 119:311-317.

Moretzsohn, M., M. Hopkins, S. Mitchell, S. Kresovich, J. Valls, and M. Ferreira. 2004. Genetic diversity of peanut (Arachis hypogaea L.) and its wild relatives based on the analysis of hypervariable regions of the genome. BMC Plant Biology 4:11.

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

Morgante, M., M. Hanafey, and W. Powell. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genetics 30:194.

Murray, M.G., and W.F. Thompson. 1980. Rapid isolation of high molecular-weight plant DNA. Nucleic Acids Research 8:4321-4325.

Paik-Ro, O.G., R.L. Smith, and D.A. Knauft. 1992. Restriction fragment lenth polymorphism evaluation of six peanut species within the *Arachis* section. Theoret. Appl. Genet. 84:201-208.

Palmieri, D.A., A.A. Hoshino, J.P. Bravo, C.R. Lopes, and M.A. Gimenes. 2002. Isolation and characterization of microsatellite loci from the forage species Arachis pintoi (Genus Arachis). Molecular Ecology Notes 2:551-553.

Palmieri, D.A., B. M. D, C. R. A, G. M. A, and L. C. R. 2005. Novel polymorphic microsatellite markers in section Caulorrhizae (*Arachis, Fabaceae*). Molecular Ecology Notes 5:77-79.

Peakall, R., Smouse, P.E., 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6:288-295.

Powell, W., G.C. Machray, and J. Provan. 1996a. Polymorphism revealed by simple sequence repeats. Trends in Plant Science 1:215-222.

Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996b. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 2:225-238.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Proite, K., S.C.M. Leal-Bertioli, D.J. Bertioli, M.C. Moretzsohn, F.R. da Silva, N.F. Martins, and P.M. Guimaraes. 2007. ESTs from a wild *Arachis* species for gene discovery and marker development. Bmc Plant Biology 7.

Raina, S.N., V. Rani, T. Kojima, Y. Ogihara, K.P. Singh, and R.M. Devarumath. 2001. RAPD and ISSR fingerprints as useful genetic markers for analysis of genetic diversity, varietal

identification, and phylogenetic relationships in peanut (Arachis hypogaea) cultivars and wild species. Genome 44:763-772.

Rozen, S., and H.J. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers, pp. 365-386. In S. Krawetz, and S. Misener (eds.) Bioinformatics Methods and Protocols: Methods in molecular Biology. Humana Press, Totowa, NJ.

Sanders, F.E., H. Guohao, L. Gong, M. Egnin, and D. Morley. 2008. Transferability of soybean (*Glycine max*) SSR markers in peanut genomic DNA (*Arachis hypogaea* L.). In Vitro Cell. Dev. Biol.-Plant 44:356-357.

Stalker, H.T., T.D. Phillips, J.P. Murphy, and T.M. Jones. 1994. Variation of isozyme patterns among *Arachis* species. Theoretical and Applied Genetics 87:746-755.

Stalker, H.T. 1997. Peanut (*Arachis hypogaea* L.). Field Crops Research 53:205-217.

Tang, S., J.K. Yu, M.B. Slabaugh, D.K. Shintani, and S.J. Knapp. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105:1124-1136.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Research 11:1441-1452.

Upadhyaya, H.D., M.E. Ferguson, and P.J. Bramel. 2001. Status of the Arachis germplasm collection at ICRISAT. Peanut Science 28: 89–96.

Upadhyaya H.D., R. Ortiz, P.J. Bramel, and S. Singh. 2003. Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. Genetic Resources and Crop Evolution 50: 139–148.

Varshney, R.K., T. Thiel, N. Stein, P. Langridge, and A. Graner. 2002. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular & Molecular Biology Letters 7:537-546.

Varshney, R.K., A. Graner, and M.E. Sorrells. 2005. Genic microsatellite markers in plants: features and applications. Trends in Biotechnology 23:48-55.

CHAPTER 4

SSR DIVERSITY IN A- AND B-GENOME DIPLOID PEANUT SPECIES[1]

[1]Khanal, S., W. Ma, and S.J. Knapp. To be submitted to *Theor Appl Genet*.

Abstract

Simple sequence repeat (SSR) variation in 60 accessions belonging to three wild diploid peanut species viz. *A. duranensis*, *A. batizocoi*, and *A. stenosperma* were studied with 32 previously mapped microsatellite markers. The objectives of this study were to gain an understanding of molecular genetic diversity in the A-genome progenitor (*A. duranensis*), and A- (*A. stenosperma*) and B-genome (*A. batizocoi*) diploid relatives of cultivated (tetraploid) peanut, and to select parents for developing intraspecific A- and B-genome diploid mapping populations to lay the groundwork for developing diploid models for peanut genomics. Further, a total of 612 previously reported and 97 recently developed SSR markers were screened against a panel of 12 diploid genotypes representing the parents of A- and B-genome mapping populations. SSR diversity study revealed a large amount of within population variation (71.1%). Also, higher levels of intraspecific allelic diversities (range: 3.02 to 9.2 alleles per marker; mean: 10.55) and mean heterozygosities (range: 0.30 to 0.67; *H*: 0.72) were observed. *Structure* and principal coordinate analysis (PCoA) showed major sub-population clusters in *A. duranensis*, suggestive of high species diversity within the A-genome progenitor. In the large scale SSR screening study, out of a total of 709 SSR markers, 553 showed amplification and 95.3% of the amplifiable markers were polymorphic among the 12 diploid accessions. The genetic diversities within the diploids were large enough to construct high-density A- and B-genome intraspecific maps of *Arachis* species.

Introduction

The genus *Arachis* L. belongs to the family of legumes (*Fabaceae*) and harbors 80 accepted species assembled into 9 sections (Krapovickas and Gregory 1994; Valls and Simpson 2005). Cultivated peanut or groundnut, *A. hypogaea* L., belongs to section *Arachis*; the section

also represents 31 diploid and a tetraploid wild species, *A. monticola*. Endemic to South America, cultivated peanut is now a globally important crop widely grown in tropical, sub-tropical and warm-temperate regions of the world (Stalker 1997). It is an allotetraploid (2n = 4x = 40) species which might have evolved through speciation due to reproductive isolation of an amphidiploid (Hopkins *et al.* 1999). Polyploidization coupled with domestication probably lead to severe breeding bottleneck which greatly reduced genetic diversity and contributed to genetic vulnerability of the cultivated groundnut varieties. Lower levels of genetic polymorphisms among cultivated peanuts have also undermined the application of molecular breeding approaches in crop improvement; genomics resources for peanut lag behind several other crop species of equal or even lesser economic importance. As such, wild peanut species offer two-fold opportunities: first, untangling genetic complexity of tetraploid genome and secondly, introducing diversity in the cultivated gene pool.

Diversity assessment for several agronomic traits has suggested higher levels of variation in wild *Arachis* germplasm pool. The wild peanut species show variation in seed storage protein profiles (Bianchi-Hall *et al.* 1993), resistance to several insect-pest and diseases (Lynch 1990; Yang *et al.* 1993; Holbrook and Stalker 2003; Kalyani *et al.* 2007), and abiotic stresses (Holbrook and Stalker 2003; Rao *et al.* 2003; Nautiyal *et al.* 2008). Besides the development of superior cultivars, introgressions of agronomic traits from these wild relatives into the cultivated genotypes also broaden the genetic base of the elite gene pool. The feasibility of introgression has been facilitated by cross compatibility of several of the closest wild relatives constituting the secondary gene pool for cultivated peanut (Singh and Simpson 1994; Stalker 1997; Holbrook and Stalker 2003).

A large number of wild introgression lines have been reported in recent years. In 2002, several introgression lines derived from an *A. hypogaea* (PI 261942) X *A. cardenasii* (PI 262141) interspecific cross were released (Stalker *et al.* 2002a; Stalker *et al.* 2002b; Stalker and Lynch 2002). These lines are frequently crossed to cultivated peanut for the transfer of the traits as resistant to insect pests and diseases (Anderson *et al.* 2006). Similarly, four elite germplasm lines with wild introgressions of rust and leaf spot resistance from *A. duranensis*, *A. batizocoi*, *A. villosa*, and *A. stenosperma* were reported by Singh *et al.* (2003). Transfer of resistance to leaf spots and some insect pests has also been reported from *A. kempff-mercadoi* (Mallikarjuna *et al.* 2004), adding up to the list of several species within genus *Arachis* shown to be amenable for gene transfer through wide hybridization. Further, these wild species are potential sources of multiple traits of interest in cultivated peanuts. However, to date, very few elite cultivars with wild introgressions have been released (Simpson and Starr 2001; Simpson *et al.* 2003). Some excellent wild introgression lines have been found inferior to their check cultivars possibly due to linkage drag (Isleib *et al.* 2006; Mondal *et al.* 2007). As such, identification and tagging of genes of interest along with marker-assisted selection provides better selection efficiency and opens up possibilities for gene pyramiding.

Diploid linkage maps are crucial in light of the fact that mapping in cultivated peanut has been severely hampered by a dearth of adequate molecular polymorphisms in tetraploids; loci mapped in diploids could help prepare a genetic map for *A. hypogaea*. Such linkage maps facilitate crop improvement through introgression breeding and transgenic approaches. However, morphological attributes (Singh *et al.* 1996), protein profiles (Bianchi-Hall *et al.* 1993; Singh *et al.* 1994, 1996), and isozyme analysis (Lu and Pickersgill 1993) suggested lower levels of intraspecific variability in several *Arachis* species. Further, because of the paucity of DNA based

markers, only three sparse linkage maps have been reported: first, a diploid RFLP map produced from an interspecific (*A. duranensis x A. stenosperma*) hybrid (Halward *et al.* 1993); second, a tetraploid RFLP map produced from a complex cross between A. hypogaea and a synthetic amphidiploid [*A. batizocoi x* (*A. cardenasii x A. diogoi)*] (Burow *et al.* 2001); and third, a diploid SSR map produced from an interspecific (*A. duranensis x A. stenosperma*) hybrid (Moretzsohn *et al.* 2005) hybrid. Although, available polymorphic markers grossly underrepresented the critical volume of markers required for the development of saturated linkage maps, SSRs were found to be the most informative marker system in *Arachis* (Hopkins *et al.* 1999).

To date, several hundred polymorphic SSRs have been publicly reported. But since the makers were developed and screened for polymorphisms against panels consisting of different genotypes, there is a need for rescreening those against the parents of the mapping populations. Further, the construction of highly saturated linkage maps requires more number of polymorphic markers. Also, besides identifying and developing the most polymorphic crosses, identifying minimum number of genotypes capturing major amount of genetic diversity is necessary for efficient marker screening. With an aim of breaking genomics gridlock and application of molecular breeding approaches, we started massively parallel DNA sequencing and marker development strategy (especially SSRs) in *Arachis* species (unpublished). As a part of building molecular breeding tools, we measured the amount of genetic diversity in diploid peanut accessions and identified the most polymorphic intraspecific crosses for constructing highly saturated linkage maps. Further, we screened all the publicly available SSR markers against a panel of 12 genotypes constituting intraspecific mapping population parents of A- and B-genome diploids developed at different research institutions. In this study, we report (1) the simple sequence repeat (SSR) diversity in 58 accessions belonging to three wild diploid *Arachis* spp.

viz. *A. duranensis*, *A. batizocoi* and *A. stenosperma*, (2) the population structure of diploid wild peanut accessions, and (3) the development of polymorphic SSR markers in A- and B-genome mapping populations. Twenty-seven markers used for SSR diversity study with 58 diploid accessions are from Moretzsohn *et al.* (2005) and the accessions belong to *A. duranensis*, a putative A-genome donor to tetraploid *Arachis* spp., *A. batizocoi*, a much diverse surrogate to putative B-genome donor *A. ipaensis*, and *A. stenosperma*, a frequently used source of traits of interest in cultivated peanut. The markers used to screen the parents of A- and B-genome mapping populations are based on publicly available SSR marker information.

Materials and Methods

Plant Materials and Marker Resources for the SSR Diversity Study

A total of 37 *A. duranensis*, 9 *A. batizocoi*, and 14 *A. stenesperma* accessions obtained from USDA (Griffin) and North Carolina State University were used for the diversity analysis (Table 4.1). These accessions represent a large portion of the collected germplasm for the individual species. Complete information including accession names, identification numbers, country of origin, and geographic coordinates are provided as a supplementary data file (Supplementary Table 4.1). For this study we used thirty-two previously mapped SSR markers representing 11 linkage groups reported by Moretzsohn *et al.* (2005) (Table 4.2).

Marker Resources for Screening Parents of the Mapping Populations

A total of 12 diploid genotypes consisting one each of *A. kuhlmanii* and A. diogoi, 7 of *A. duranensis*, and 3 of *A. batizocoi* accessions obtained from USDA (Griffin) and North Carolina State University were selected for large scale marker screening (Table 4.3). These accessions represent the parents of A- and B-genome mapping populations developed at different research institutions. A total of 612 previously reported and 97 recently developed SSR markers were

screened for polymorphisms among the diploid accessions. Information regarding marker identities, original references, primers flanking SSRs, annealing temperatures (Ta), product lengths, and fluorescent labels are provided as a supplementary data file (Supplementary Table 4.2).

DNA Extraction and Marker Development

Leaf tissues from greenhouse grown plants were used for DNA extraction. Modified CTAB (cetyltrimethylammonium bromide) method was used for DNA isolation (Murray and Thompson 1980) and Synergy HT Multi-Mode Microplate reader (BioTek Instruments, Inc.; Winooski, VT, USA) was used for DNA quantification. Forward primers were labeled with one of the three fluorescent dyes viz. 6-FAM, HEX, or TAMRA and PCR reactions essentially followed the methodology as described by Tang *et al.* (2002). Genotypes were determined using the ABI 3730 DNA analyzer and GeneMapper Software Version 4 (Applied Biosystems, Foster City, CA).

Statistical Analysis

Rectangular binary data matrices were prepared wherein 1 represented the presence and 0 represented the absence of a particular fragment in a given accession. Band frequency, total number of polymorphic bands and average number of bands per marker were recorded. Major allele frequency, gene diversity and polymorphism information content (PIC) for individual markers were obtained using software tool *PowerMarker V3.25* (Liu and Muse 2005).

Inferences on population structure of 58 diploid accessions were drawn using model-based clustering software package *structure 2.2* (Pritchard *et al.* 2000, 2002; Falush *et al.* 2007). The binary data matrix was used as an input which necessitated processing the data as haploid. Thus, a matrix with a total of 285 bands (identified as loci) corresponding to 27 markers was run

under ancestry with no admixture and allele frequencies correlated models as operating settings for *structure*. Simulations were run with 10,000 to 100,000 burning lengths and MCMC replications each and population sizes of 1 to 6. Inferences on an appropriate population size were drawn basically following instructions from the *structure* manual and Evanno *et al.* (2005).

Microsat (Minch *et al.* 1997) was used for the construction of distance matrices based on the proportion of shared bands ($D = 1 - p_s$). The matrices were imported into *Phylip v3.67* (Felsenstein 1989) for the construction of neighbor-joining trees. Principle coordinate analysis (PCoA) and analysis of molecular variance (AMOVA) were performed using Microsoft excel based software *GenAlEx 6.1* (Peakall and Smoush 2006).

Results and Discussion

Allelic Diversities in 60 Diploid *Arachis* accessions

Simple sequence repeats are the most useful markers in unraveling intra- and interspecific diversity in *Arachis* spp. Morphological and other molecular marker systems have suggested lower levels of intraspecific variations even among the wild *Arachis* spp. It was especially true for *A. duranensis* and *A. batizocoi* accessions which were collected from a narrow geographic range essentially with uniform phytogeographic conditions (Singh *et al.* 1996). SSR markers in our study, however, were able to detect appreciably high molecular polymorphisms among different *Arachis* species. Average genetic diversities among *A. duranensis* and *A. batizocoi* accessions were very high. As such, identifying the most polymorphic crosses could facilitate construction of highly saturated linkage maps for A- and B-genome intraspecific populations.

Out of 32 previously mapped markers, 27 showed amplification against 60 accessions used in this study. Two of the 60 accessions viz. DUR38 (PI 468321) and BAT3 (Grif 15031) were excluded from further analysis since they failed to conform to their species identity and

were represented in clusters belonging to other species, suggesting misclassification (Figure 4.3). In case of BAT3, its genetic similarity with DUR3 was much higher than that even among other *A. duranensis* accessions. Further, an interspecies between BAT3 and another *A. batizocoi* accession (BAT8) failed to produce pegs, substantiating our suspicion; cytological studies showed meiotic irregularities and lower pollen fertility of F_1 (unpublished). However, greenhouse grown BAT3 genotypes closely resembled other *A. batizocoi* accessions and phenotypic characteristics also were in complete contrast to those of *A. duranensis*. Since intraspecific *A. batizocoi* hybrids have been reported showing extensive chromosomal rearrangements and lower rates of fertility (Stalker *et al.* 1991), it necessitates further investigation into the accession. Another excluded species, DUR38 (PI 468321), although was reclassified from *A. ipaensis,* clustered together with *A. stenosperma* accessions, suggesting misclassification. Another *A. duranensis* accession, DUR37 (PI 468324), on the other hand, was reclassified from *A. batizocoi*. In this study it clustered together with the other *A. duranensis* accessions and an interspecies between DUR37 and an *A. batizocoi* accession failed to produce pegs and showed meiotic irregularities (unpublished), essentially suggesting correct reclassification.

In 58 accessions, a total of 285 alleles were scored with an average of 10.55 alleles per marker. Null alleles accounted for 5.4% of the total alleles, limiting the average to 9.8 bands per marker. Except for the null alleles, each marker amplified single locus in most of the accessions. Average allele frequency was 0.093 and the major allele frequency ranged from 0.12 to 0.85 with an average of 0.37. Gene diversity ranged from 0.27 to 0.94 with a mean of 0.74 and PIC ranged from 0.25 to 0.94 with a mean of 0.72. Marker statistics for individual species are summarized in Table 4.4. Analysis of molecular variance (AMOVA) apportioned the total variation into among

population and within population variations. Results showed that 71% of the total genetic diversity resided within three *Arachis* species, whereas only, 29% was attributable to the differences between the species (Table 4.5).

All the markers were polymorphic in *A. duranensis* subset of 36 accessions. A total of 249 alleles were recorded with an average of 9.2 alleles per marker. Null alleles accounted for 3.9% of the total alleles, limiting the average to 8.5 bands per marker. Average allele frequency was 0.11 and the major allele frequency ranged from 0.11 to 0.97 with an average of 0.41. Mean gene diversity and average PIC were 0.69 and 0.67 respectively.

In *A. batizocoi* subset with 8 accessions, all 27 markers were polymorphic and a total of 104 alleles were scored with an average of 3.8 alleles per marker. The frequency of null alleles was high (13.88%), consequently limiting the number of fragments per marker to 3.4. Average allele frequency was 0.11 and the major allele frequency ranged from 0.25 to 0.87 with an average of 0.53. Mean gene diversity and average PIC were 0.59 and 0.54 respectively.

19 out of 27 markers were found polymorphic between 14 accessions of *A. stenosperma.* A total of 82 alleles were scored with an average of 3.03 alleles per marker. Null alleles accounted for 4.6% of the total alleles, limiting the average to 2.4 bands per marker. Average allele frequency was 0.33 and the major allele frequency ranged from 0.14 to 1.00 with an average of 0.75. Mean gene diversity and average PIC were 0.32 and 0.30 respectively.

Population Structure and Cluster Analysis in Diploid Peanut Species

Structure simulations gave an appropriate burnin length and MCMS replication of 100,000 each. The most appropriate population size repeatedly observed across different simulations and independent runs was three (Figure 4.1). The populations, however, were not clustered according to their taxonomic groupings. Interestingly, *A. duranensis* accessions showed

sub-population structure while *A. batizocoi* accessions were grouped together with one of the *A. duranensis* populations (Figure 4.1). *Structure* simulation on 36 accessions of *A. duranensis* suggested that the sub-population structure was real.

Principle coordinate analysis (PCoA) corroborated the *structure* results in that *A. duranensis* showed two population clusters and *A. batizocoi* was very close to one of them (Figure 4.2). Also, one of the clusters contained *A. duranensis* accessions which represented a sub-population in *structure* analysis. 54.85% of the total variation was explained by the first two axes in PCoA with 33.71% and 21.14% contribution from the first and the second components respectively. NJ tree clearly indicated that populations from different *Arachis* taxa form distinct clades and that diversity among 36 *A. duranensis* accessions was high, as suggested by individuals grouped into several distinct clusters (Figure 4.3). *A. stenosperma* accessions, on the other hand, formed a tight cluster, suggesting lower levels of diversity among 14 accessions used in this study. Although, eight *A. batizocoi* accessions were clustered together, wider spread of the cluster suggested appreciable amount of diversity within the species.

Geospatial studies of the collection sites have suggested a narrow distributional range for *A. duranensis* accessions (Jarvis *et al.* 2003). All these accessions we used in this study were collected from adjacent Argentinian, Bolivian and Paraguan provinces. Neighbor-joining tree clustered individuals relative to their collection sites such that individuals from same geographic regions were grouped together. Similar results have been reported from morphological (Stalker 1990), RFLP (Stalker *et al.* 1995), RAPD (Hilu and Stalker 1995) and AFLP (Milla *et al.* 2005) analysis. However, s*tructure* analysis suggested two major sub-populations within *A. duranensis*. One population cluster represented individuals predominantly collected from lower altitudinal range (< 900 masl) while the other cluster represented individuals from higher altitudinal range

(> 900 masl). The high altitudinal sub-cluster was also evident from neighbor-joining tree and PCA plot. Although genotypes from higher elevations were representatives of a particular region, nevertheless, the extent of population differentiation within a narrow geographic range is suggestive of huge geographic influence on divergence within the species.

Large Scale Marker Screening and Cluster Analysis

Out of a total of 709 SSR markers, 532 showed amplification against 12 diploid accessions used in this study; 95.3% of the amplifiable markers were polymorphic in the diploid accessions. A total of 2,951 alleles were scored with an average of 5.5 alleles per marker. Null alleles accounted for 12.6% of the total alleles, limiting the average to 5 bands per marker. Except for the null alleles, most of the markers amplified single locus in the diploid accessions. Average allele frequency was 0.17 and the major allele frequency ranged from 0.08 to 1.00 with an average of 0.44. Gene diversity ranged from 0.00 to 0.92 with a mean of 0.67 and PIC ranged from 0.00 to 0.91 with a mean of 0.64. Observations on the individual species or the parents of the mapping populations are summarized in Table 4.6.

Of 512 amplifiable markers in *A. duranensis* accessions, 89% (457) were polymorphic. We recorded a total of 1,667 alleles with an average of 3.2 alleles per marker. Null alleles accounted for 8.5% of the total alleles, limiting the average to 2.9 bands per marker. Average allele frequency was 0.26 and the major allele frequency ranged from 0.14 to 1.00 with an average of 0.54. Gene diversity ranged from 0.00 to 0.86 with a mean of 0.56 and PIC ranged from 0.00 to 0.84 with a mean of 0.52.

In *A. batizocoi* subset with 3 accessions, 389 (81.9%) out of 475 amplifiable markers were polymorphic. A total of 873 alleles were scored with an average of 1.8 alleles per marker. Null alleles accounted for 9.27% of the total alleles, limiting the average to 1.4 fragments per

marker. Average allele frequency was 0.46 and the major allele frequency ranged from 0.33 to 1 with an average of 0.61. Gene diversity ranged from 0.00 to 0.67 with a mean of 0.44 and PIC ranged from 0.00 to 0.59 with a mean of 0.37.

Between *A. kuhlmanii* and *A. diogoi*, a total of 474 markers showed amplification and 71.3% (338) of those were polymorphic. A total of 728 alleles were scored with an average of 1.5 alleles per marker. Null alleles accounted for 4.14% of the total alleles, limiting the average to 1.4 bands per marker. Average allele frequency was 0.58 and the major allele frequency ranged from 0.5 to 1.00 with an average of 0.65. Gene diversity ranged from 0.00 to 0.50 with a mean of 0.35 and PIC ranged from 0.00 to 0.37 with a mean of 0.26.

Figure 4.3 shows the neighbor-joining tree constructed from the genetic distances estimated from 556 SSR markers among 12 diploid accessions, representing parents of A- and B-genome *Arachis* mapping populations.

Conclusion

In this study, 32 mapped SSR markers were screened among 37 *A. duranensis*, 9 *A. batizocoi*, and 14 *A. stenosperma* accessions. Further, a total of 612 previously reported and 97 recently developed SSR markers were screened against a panel of 12 diploid genotypes representing the parents of A- and B-genome mapping populations. Sufficiently higher levels of intraspecific allelic diversities, average heterozygosities, and number of polymorphic markers suggested the feasibility of high-density A-and B-genome intraspecific maps in *Arachis* species. Parents for intraspecific mapping populations were identified, populations were created, and mapping of A- and B-genome diploid models for tetraploid peanut were initiated.

References

Anderson, W.F., C.C. Holbrook, and P. Timper. 2006. Registration of root-knot nematode resistant peanut germplasm lines NR 0812 and NR 0817. Crop Sci. 46:481-482.

Bianchi-Hall, C.M., R.D. Keys, H.T. Stalker, and J.P. Murphy. 1993. Diversity of seed storage protein-patterns in wild peanut (*Arachis*, Fabaceae) species. Plant Systematics and Evolution 186:1-15.

Burow, M.D., C.E. Simpson, J.L. Starr, and A.H. Paterson. 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.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14(8):2611-2620.

Falush, D., M. Stephens, and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes 7:574-578.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164-166.

Halward, T., H.T. Stalker, and G. Kochert. 1993. Development of an RFLP linkage map in

Hilu, K.W., and H.T. Stalker. 1995. Genetic relationships between peanut and wild species of Arachis sect Arachis (Fabaceae): Evidence from RAPDs. Plant Systematics and Evolution 198:167-178.

Holbrook, C.C., and H.T. Stalker. 2003. Peanut breeding and genetic resources, pp. 297-356. In J. Janick (ed.) Plant Breeding Reviews 22, John Wiley & Sons, Inc.

Hopkins, M.S., A.M. Casa, T. Wang, S.E. Mitchell, R.E. Dean, G.D. Kochert, and S. Kresovich. 1999. Discovery and characterization of polymorphic simple sequence repeats (SSRs) in peanut. Crop Science 39:1243-1247.

Isleib, T.G., P.W. Rice, R.W. Mozingo II, S.C. Copeland, J.B. Graeber, B.B. Shew, D.L. Smith, H.A. Melouk, and H.T. Stalker. 2006. Registration of N96076L peanut germplasm. Crop Sci. 46:2329-2330.

Jarvis, A., M.E. Ferguson, D.E. William, L. Guarino, P.G. Jones, H.T. Stalker, J.F.M. Valls, R.N. Pittman, C.E. Simpson, and P. Bramel. 2003. Biogeography of wild Arachis: assessing conservation status and setting future priorities. Crop Sci. 43.1100-1108.

Kalyani, G., A.S. Reddy, P.L. Kumar, R. Rao, R. Aruna, F. Waliyar, and S.N. Nigam. 2007. Sources of resistance to Tobacco streak virus in wild arachis (Fabaceae : Papilionoidae) germplasm. Plant Disease 91:1585-1590.

Krapovickas, A., and W.C. Gregory. 1994. Taxonomia del genero *Arachis* (Leguminosae). Bonplandia 8: 1-186.

Liu, K.J., and S.V. Muse. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128-2129.

Lu, J., and B. Pickersgill. 1993. Isozyme variation and species relationships in peanut and its

Lynch, R.E. 1990. Resistance in peanut to major arthropod pests. Florida Entomologist 73:422-445.

Mallikarjuna, N., S. Pande, D.R. Jadhav, D.C. Sastri, and J.N. Rao. 2004. Introgression of disease resistance genes from *Arachis kempff-mercadoi* into cultivated groundnut. Plant Breeding 123:573-576.

Milla, S.R., T.G. Isleib, H.T. Stalker, and G.J. Scoles. 2005. Taxonomic relationships among *Arachis* sect. *Arachis* species as revealed by AFLP markers. Genome 48:1-11.

Minch, E., A. Ruiz-Linares, D. Goldstein, M. Feldman & L.L. Cavalli-Sforza, 1997. MICROSAT: A Computer Program for Calculating Various Statistics on Microsatellite Allele Data, ver. 1.5d. Stanford University, Stanford, CA. Available at: http://hpgl.stanford.edu/projects/microsat/.

Mondal, S., A.M. Badigannavar, G.S.S. Murty. 2007. RAPD markers linked to a rust resistance gene in cultivated groundnut (*Arachis hypogaea* L.). Euphytica 159:233-239.

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

Murray, M.G., and W.F. Thompson. 1980. Rapid isolation of high molecular-weight plant DNA. Nucleic Acids Research 8:4321-4325.

Nautiyal, P.C., K. Rajgopal, P.V. Zala, D.S. Pujari, M. Basu, B.A. Dhadhal, and B.M. Nandre. 2008. Evaluation of wild Arachis species for abiotic stress tolerance: I. Thermal stress and leaf water relations. Euphytica 159:43-57.

Peakall, R., Smouse, P.E., 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6:288-295.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Pritchard, J., D. Falush, and M. Stephens. 2002. Inference of population structure in recently admixed populations. American Journal of Human Genetics 71:177-177.

Rao, N.K., L.J. Reddy, and P.J. Bramel. 2003. Potential of wild species for genetic enhancement of some semi-arid food crops. Genetic Resources and Crop Evolution 50:707-721.

Simpson, C.E., and J.L. Starr. 2001. Registration of 'COAN" peanut. Crop Sci. 41:918.

Simpson, C.E., J.L. Starr, G.T. Church, M.D. Burow, and A.H. Paterson. 2003. Registration of 'NemaTAM' peanut. 43:1561.

Singh, A.K., and C.E. Simpson. 1994. Biosystematics and genetic resources. The ground crop: a scientific basis for improvement. J.Smartt. London, Chapman & Hall: 96-137.

Singh, A.K., P. Subrahmanyam, and S. Gurtu. 1996. Variation in a wild groundnut species, *Arachis duranensis* Krapov & W.C. Gregory. Genetic Resources and Crop Evolution 43:135-142.

Singh A.K., S.L. Dwivedi, S. Pande, J.P. Moss, S.N. Nigam, and D.C. Sastri. 2003. Registration of rust and late leaf spot resistant peanut germplasm lines. Crop Sci. 43:440-441.

Stalker, H.T. 1990. A morphological appraisal of wild species in section *Arachis* of peanuts. Peanut Sci. 17:117-122.

Stalker, H.T., J.S. Dhesi, and G. Kochert. 1995. Genetic diversity within the species Arachis duranensis Krapov & W.C. Gregory, a possible progenitor of cultivated peanut. Genome 38:1201-1212.

Stalker, H.T. 1997. Peanut (*Arachis hypogaea* L.). Field Crops Research 53:205-217.

Stalker and Lynch. 2002. Registration of four insect-resistant peanut germplasm lines. Crop Sci. 42:313-314.

Stalker, H.T., M.K. Beute, B.B. Shew, and K.R. Barker. 2002a. Registration of two root-knot nematode-resistant peanut germplasm lines. Crop Sci. 42:312-313.

Stalker, H.T., M.K. Beute, B.B. Shew, and T.G. Isleib. 2002b. Registration of five leaf spot-resistant peanut germplasm lines. Crop Sci. 42:314-316.

Tang, S., J.K. Yu, M.B. Slabaugh, D.K. Shintani, and S.J. Knapp. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105:1124-1136.

Valls, J.F.M, and C.E. Simpson. 2005. New species of Arachis L. (Leguminosae) from Brazil, Paraguay and Bolivia. Bonplandia 14:35-64.

Yang, G., K.E. Espelie, J.W. Todd, A.K. Culbreath, R.N. Pittman, and J.W. Demski. 1993. Cuticular lipids from wild and cultivated peanuts and the relative resistance of these peanut species to fall armyworm and thrips. Journal of Agricultural and Food Chemistry 41:814-818.

Table 2.1. *Arachis* germplasm screened for SSR marker amplification and length polymorphisms.

Scientific Name	Population or Cultivar Name	PI Number
A. duranensis Krapov. and W. C. Greg.	36036	PI 475887
A. duranensis Krapov. and W. C. Greg.	38901	PI 497483
A. duranensis Krapov. and W. C. Greg.	30078	PI 468324
A. batizocoi Krapov. and W. C. Greg.	30080	PI 468326
A. hypogaea var. *aequatoriana* Krapov. and W. C. Greg.		PI 497630
A. hypogaea var. *fastigiata* (Waldron) Krapov. & W. C. Greg.		PI 497471
A. hypogaea var. *hirsuta* J. Kohler		PI 576613
A. hypogaea var. *hypogaea*	Tifrunner	PI 644011
A. hypogaea var. *peruviana* Krapov. and W. C. Greg.		PI 502045
A. hypogaea var. *fastigiata (Waldron)* Krapov. & W. C. Greg.	GT-C20	-
A. hypogaea L.	GT-C9	-
A. hypogaea L.	A100	-

Table 2.2. Annotation statistics of methylation-filtered (MF) and unfiltered (UF) GSSs from *Arachis*.

***Arachis* species**	**Library**	**Number of Sequences**	**Number of Putative Genes**	**Number of Putative Repeats**	**Total Annotated Sequences**
A. duranensis	MF	1,344	200 (15%)	42 (3%)	242 (18%)
	UF	2,400	455 (19%)	611 (25%)	1,066 (44%)
A. batizocoi	MF	1,248	380 (30%)	44 (4%)	424 (34%)
	UF	2,400	365 (15%)	285 (12%)	650 (27%)
A. hypogaea	MF	1,248	525 (42%)	64 (5%)	589 (47%)
	UF	2,304	333 (14%)	356 (15%)	689 (29%)
Total		10,944	2,168 (20%)	1,402 (13%)	3,660 (33%)

Table 3.1. *Arachis* germplasm screened for EST-SSR marker amplification and length polymorphisms among 28 tetraploids and 4 diploid peanut accessions.

S.N.	Population or Cultivar Name or Collector's ID	PI Number	Scientific Name	Description[a]
1	Tifrunner	PI 644011	*A. hypogaea* ssp. *hypogaea*	Runner
2	Georgia Green	PI 587093	*A. hypogaea*	Runner
3	Georgia 06G	PI 644220	*A. hypogaea*	Runner
4	Florunner	PI 565448	*A. hypogaea*	Runner
5	Basse	PI 229553	*A. hypogaea*	Germplasm
6	Dixie Giant	PI 290676	*A. hypogaea*	Germplasm
7	GA207-3-4	NA[b]	*A. hypogaea*	Germplasm
8	2201	PI 203396	*A. hypogaea*	Runner
9	Mani = GP-10	PI 109839	*A. hypogaea*	Runner
10	NC12C	PI 596406	*A. hypogaea* ssp. *hypogaea*	Virginia
11	Indio = Strain AH 1119-A	PI 121067	*A. hypogaea*	Virginia
12	FAV 70	PI 337396	*A. hypogaea*	Valencia
13	New Mexico Valencia A	PI 565452	*A. hypogaea* ssp. *fastigiata* var. *fastigiata*	Valencia
14	US 608-1	PI 497288	*A. hypogaea*	Valencia
15	Georgia Valencia	PI 617040	*A. hypogaea*	Valencia
16	GT-C20	NA	*A. hypogaea* ssp. *fastigiata* var. *vulgaris*	Spanish
17	Chico	PI 565455	*A. hypogaea*	Spanish
18	31-A	PI 161317	*A. hypogaea*	Spanish
19	5001	PI 121070	*A. hypogaea*	Spanish
20	NC94022	NA	*A. hypogaea*	Germplasm
21	Overo Chiquitano	PI 313949	*A. hypogaea*	Runner
22	Sun Oleic 97R	PI 596800	*A. hypogaea*	Runner
23	NemaTAM	PI 631175	*A. hypogaea*	Cultivar
24	SSD 6	PI 576638	*A. hypogaea* ssp. *hypogaea* var. *hirsuta*	Landrace
25	GP-NC WS 14	PI 619178	Interspecific Hybrid[c]	Exotic
26	FST-1	PI 476052	*A. hypogaea* ssp. *fastigiata* var.	Landrace
27	AEQ-2	PI 497630	*A. hypogaea* ssp. *fastigiata* var. *peruviana*	Landrace
28	GKBSPSc 30062	PI 468196	*A. monticola*	Wild
29	36036	PI 475887	*A. duranensis*	Wild
30	38901	PI 497483	*A. duranensis*	Wild
31	1504-W	Grif 15031	*A. batizocoi*	Wild
32	30080	PI 468326	*A. batizocoi*	Wild

[a]Description of genotypes either based on botanical market class or cultivation status

[b]Indicates that the plant introduction number is not available

[c]GP-NC WS 14 = NC 6//PI270806/GP-NC WS 4 (Stalker and Lynch 2002)

Table 4.1. Diploid *Arachis* accessions genotyped with 32 mapped SSR markers. The markers were reported by Moretzsohn et al. 2005.

Serial Number	PI Number	UGA Alias	Genome	Scientific Name
1	Grif 15035	DUR1	AA	*A. duranensis* Krapov. and W. C. Greg.
2	Grif 15036	DUR2		
3	Grif 15037	DUR3		
4	Grif 15038	DUR4		
5	Grif 15039	DUR5		
6	PI 262133	DUR6		
7	PI 468197	DUR7		
8	PI 468201	DUR9		
9	PI 468202	DUR10		
10	PI 468203	DUR11		
11	PI 468319	DUR12		
12	PI 468320	DUR13		
13	PI 468323	DUR14		
14	PI 468372	DUR15		
15	PI 475844	DUR16		
16	PI 475845	DUR17		
17	PI 475846	DUR18		
18	PI 475847	DUR19		
19	PI 475882	DUR20		
20	PI 475883	DUR21		
21	PI 475884	DUR22		
22	PI 475885	DUR23		
23	PI 475886	DUR24		
24	PI 475887	DUR25		
25	PI 497262	DUR26		
26	PI 497263	DUR27		
27	PI 497264	DUR28		
28	PI 497265	DUR29		
29	PI 497266	DUR30		
30	PI 497267	DUR31		
31	PI 497268	DUR32		
32	PI 497269	DUR33		
33	PI 497270	DUR34		
34	PI 497483	DUR35		
35	PI 497484	DUR36		
36	PI 468324	DUR37		
37	PI 468321	DUR38		

Table 4.1 continued from the previous page…

Serial Number	PI Number	UGA Alias	Genome	Scientific Name
38	Grif 14248	BAT1	BB	*A. batizocoi* Krapov. and W. C. Greg.
39	Grif 15030	BAT2		
40	Grif 15031	BAT3		
41	Grif 15033	BAT5		
42	PI 468325	BAT7		
43	PI 468326	BAT8		
44	PI 468327	BAT9		
45	PI 468328	BAT10		
46	PI 468329	BAT11		
47	Grif 7480	STP1	AA	*A. stenosperma* Krapov. and W. C. Greg.
48	Grif 7482	STP2		
49	Grif 7731	STP4		
50	PI 338279	STP5		
51	PI 497578	STP7		
52	PI 497579	STP8		
53	PI 497580	STP9		
54	PI 497581	STP10		
55	PI 591350	STP11		
56	PI 591351	STP12		
57	PI 591352	STP13		
58	PI 591359	STP14		
59	PI 599180	STP15		
60	PI 599185	STP16		

Table 4.2. The 32 mapped SSR markers genotyped among 60 diploid *Arachis* accessions. The markers were reported by Moretzsohn et al. 2005.

Marker	Forward primer	Reverse primer	Linkage Group	Expected Size
GI876	ctttggggactgactgcata	acaccataagcacaaacagaca	1	190
GI919	gctacacgctgagttgagtgag	gacgaggatgaggacgagag		350
TC3H02	atggtgagctcgacgctagt	ctctccgccatccatgtaat		300
TC9B08	ggttgggttgagaacaagg	accctcaccactaactccatta		N/A[a]
RN0615[#]	gggcatttaagggacaatgg	cccgacccacacatttaacata	2	410
RN2F12	gcgagaaggcgagaaacat	accgcaaccaccacaatc		340
TC11A02	aatcggaatggcaagagaca	agagcaaagggcgaatctatg		260
TC4F12	ggtgaatgacagatgctcca	gatctttccgccattttctc		220
TC1E06	accgttacgaacgctttgtc	tccctctcatacgacaccct	3	150
TC3E02	caaaccgaaggaggaacttg	tgaaagataggtttcggtgga		250
TC4E10	ccattttctcctcgaaccaa	acgtcatcttccctcctcct		300
TC7E04[#]	gaaggaccccatctattcaaa	tccgatttctctctctctctc		300
GI832	gccactttattctaagcactcc	aagagaccacacgctcaca	4	200
TC1H04	cattacttcctaggtttgttttc	atggcgtgacaacggaac		270
TC7G10	ccagccatgcactcatagaata	aatggggttcacaagagagaga		150
TC1D02	gatccaaaatctcgccttga	gctgctctgcacaacaagaa	5	N/A
TC3A12	gcccatatcaagctccaaaa	tagccagcgaaggactcaat		N/A
TC6E01	ctccctcgcttcctctttct	acgcattaaccacacaccaa		N/A
TC1A08	aaggggttaagggcatgact	ccacaaatgggtcgtcgat	6	N/A
TC3H07	caatgggaggcaaatcaagt	gccaaatggttccttctcaa		N/A
TC7C06[#]	ggcaggggaataaaactactaact	ttttccttccttctcctttgtc		100
TC4G10	ttcggtcatgtttgtccaga	ctcgagtgctcacccttcat	7	230
TC6G09	ggaggttgcatgcatcatagt	tcattgaacgtatttgaaagctc		140
TC9H08	gccaaaggggaccataaac	tccatcttccatctcatccac		N/A
TC6H03	tcacaatcagagctccaacaa	caggttcaccaggaacgagt	8	N/A
TC9F04	cctaaacaacgacaaacactca	aagcacaacacagaaccctaaa		150
GI1107	gatacatctcatccgttcgtg	ccgtccgaccacatacaa	9	320
ML4D02	gggtcaagtcattttctatgct	gccatcacacatcatccat		130
GI936	tcacagatccatagactttaaca	ccggtgtggattcatagtagag	10	180
RN0681[#]	atgatccctttctagcag	tttcacaaacacaagacaa		320
TC3B05[#]	ggagaaaacgcattggaact	tttgtcccgttgggaatagt	11	260
TC7A02	ccttggcttacacgacttcct	cgaaaacgacactatgaaactgc		380

[#] Markers failed in our study

[a] Expected size not available

Table 4.3. The 12 diploid *Arachis* genotypes screened for polymorphisms with 709 SSR markers.

Scientific Name	Cultivar or Collector ID	Genome	PI Number	Seed Source
A. kuhlmannii Krapov. and W. C. Greg.	7369	AA	Grif 7571	NCSU
A. diogoi Hoehne	10602		PI 276235	
A. duranensis Krapov. and W. C. Greg.	30073		PI 468319	
A. duranensis Krapov. and W. C. Greg.	36005		PI 475885	
A. duranensis Krapov. and W. C. Greg.	30064		PI 468200	
A. duranensis Krapov. and W. C. Greg.	30061		PI 468198	
A. duranensis Krapov. and W. C. Greg.	36036		PI 475887	USDA Griffin
A. duranensis Krapov. and W. C. Greg.	38901		PI 497483	
A. duranensis Krapov. and W. C. Greg.	30078		PI 468324	
A. batizocoi Krapov. and W. C. Greg.	9484	BB	PI 298639	NCSU
A. batizocoi Krapov. and W. C. Greg.	30081		PI 468327	
A. batizocoi Krapov. and W. C. Greg.	30080		PI 468326	USDA Griffin

Table 4.4. Polymorphisms of the 27 SSR markers screened among 36 *A. duranensis*, 8 *A. batizocoi*, and 14 *A. stenosperma* accessions.

***Arachis* species**	**No. of Polymorphic Markers**	**Mean Common Allele Freq.**	**No. of Unique Bands/Marker**	**Freq. of Null Alleles (%)**	**Av. Gene Diversity**	**Mean PIC**
A. duranensis	27	0.41	8.5	3.9	0.69	0.67
A. stenosperma	19	0.75	2.4	4.6	0.32	0.30
A. batizocoi	27	0.53	3.4	13.88	0.59	0.54

Table 4.5. Analysis of molecular variance among and within *n* accessions of three diploid species: *A. duranensis* (*n* = 36), *A. batizocoi* (n = 8), and *A. stenosperma* (n = 14). *Probability, P(rand>=data), for PhiPT is based on permutation across the full data set.

Source	df	SS	MS	Est. Var.	%
Among Pops	2	244.260	122.130	6.758	29%
Within Pops	55	923.861	16.797	16.797	71%
Total	57	1168.121		23.556	100%

Stat	Value	P(rand >= data)
PhiPT*	0.287	0.010

Table 4.6. Polymorphisms of 556 SSR markers among *A. duranensis*, *A. batizocoi*, *A. kuhlmanii*, and *A. diogoi* germplasm accessions.

***Arachis* Species**	**No. of Polymorphic Markers**	**Mean Maj. Allele Freq.**	**No. of Unique Bands/Marker**	**Freq. of Null Alleles (%)**	**Av. Gene Diversity**	**Mean PIC**
Among *A. duranensis*	457	0.54	2.9	8.5	0.56	0.52
Between *A. kuhlmanii and A. diogoi*	338	0.65	1.4	4.1	0.35	0.26
Among *A. batizocoi*	389	0.61	1.4	9.2	0.44	0.37

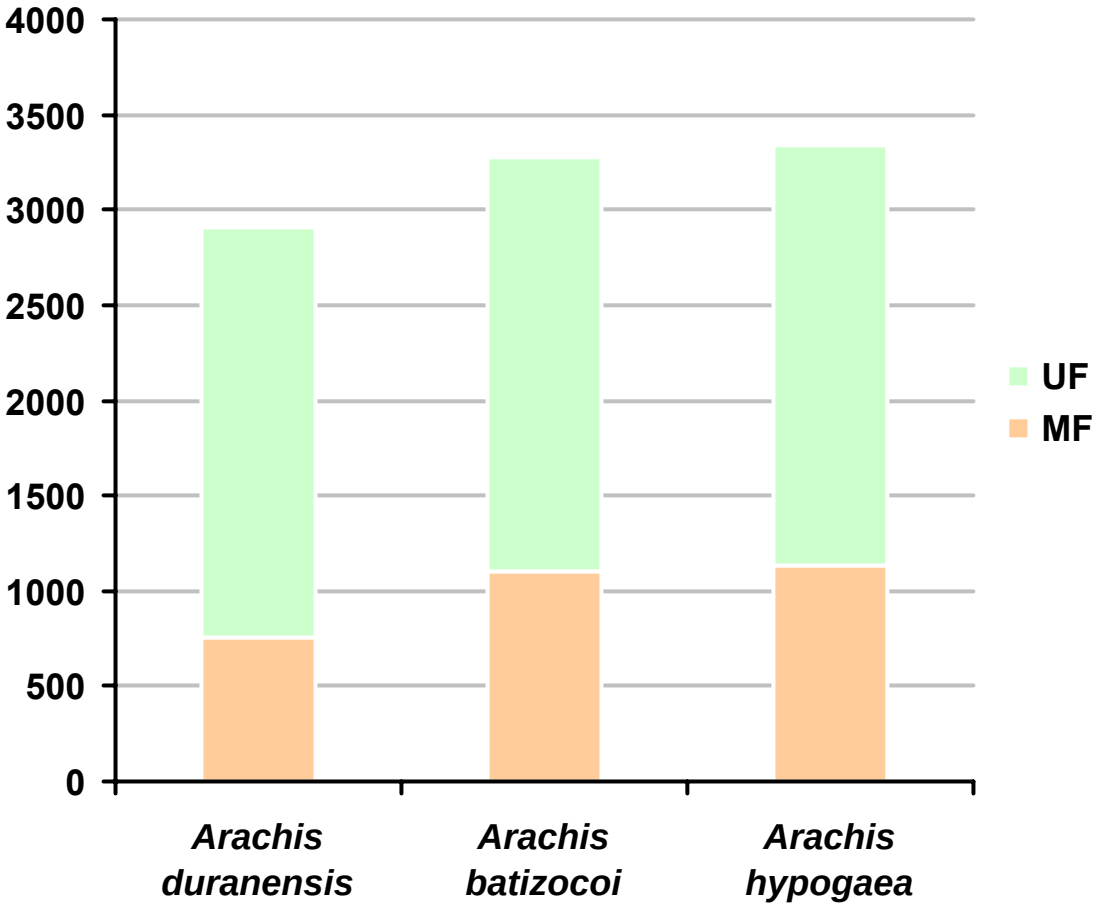


Figure 2.1. Number of genome survey sequences (GSSs) from methylation-filtered (MF) and unfiltered (UF) genomic libraries of peanut.

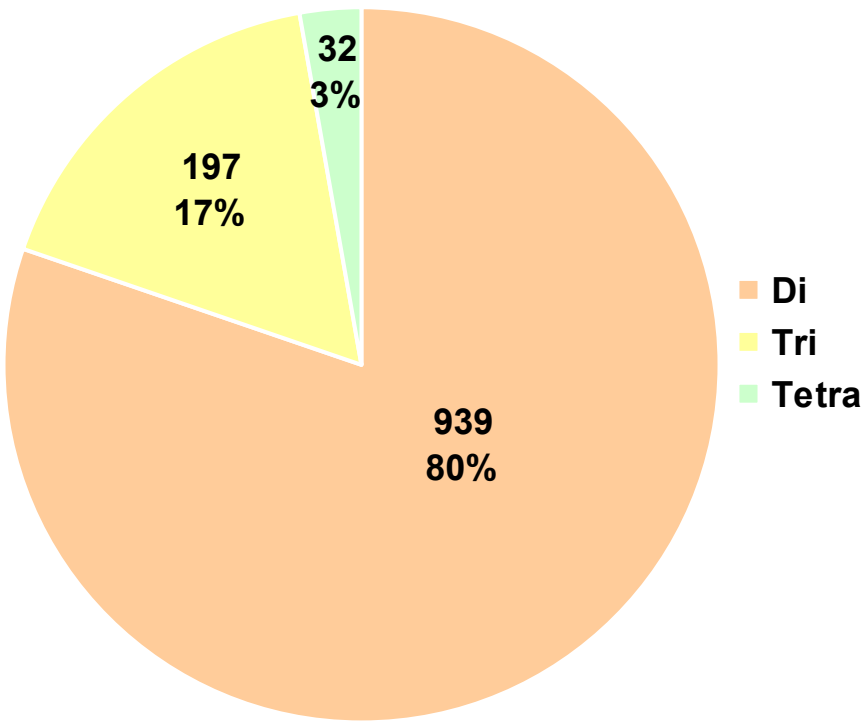


Figure 2.2. Abundance of di-, tri-, and tetranucleotide repeats among 9,517 genomic survey sequences of *Arachis*.

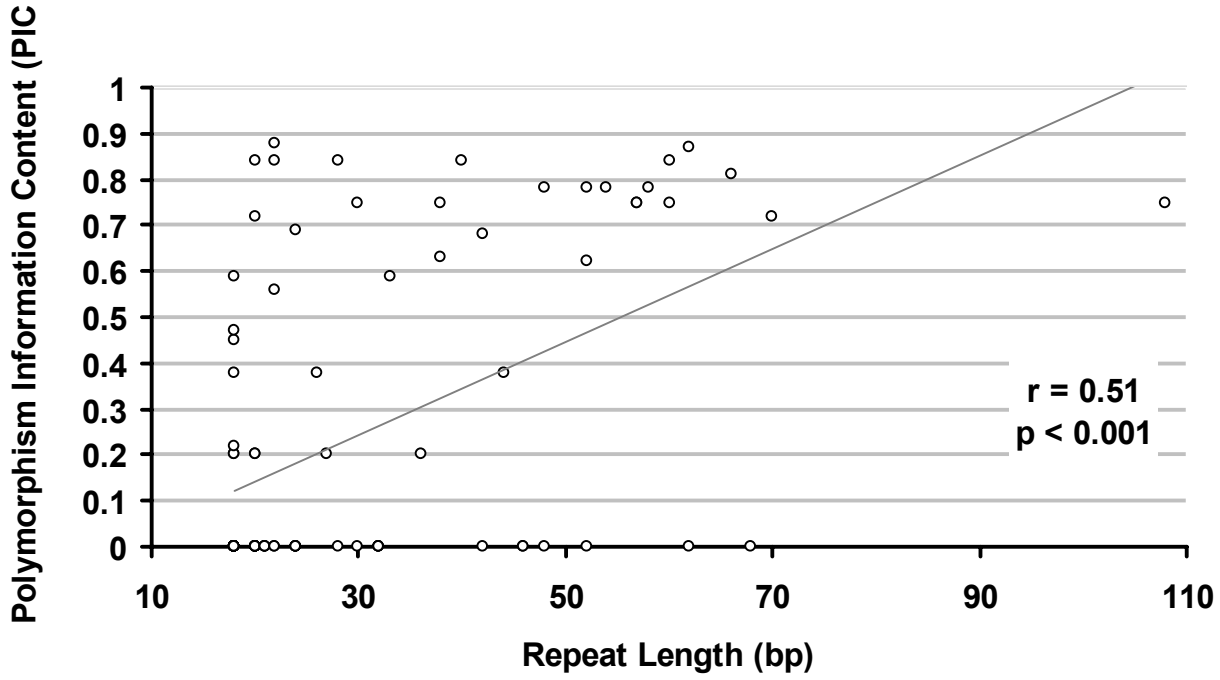


Figure 2.3. Relationship between the simple sequence repeat length (bp) and heterozygosity for 97 SSR markers among eight peanut germplasm accessions.

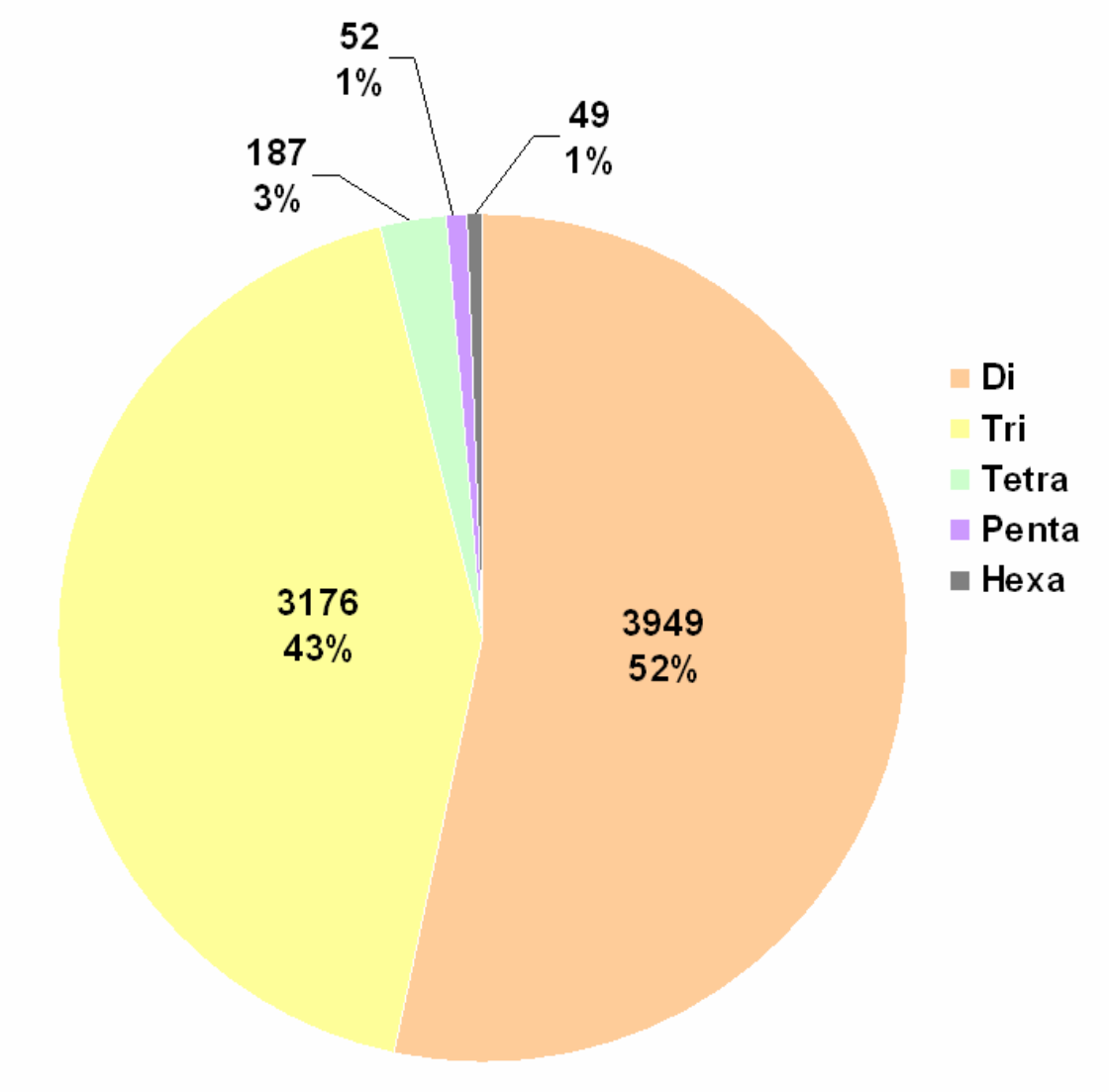


Figure 3.1. Abundance of di-, tri-, tetra-, penta-, and hexanucleotide repeats among 101,132 unigenes in a peanut EST database.

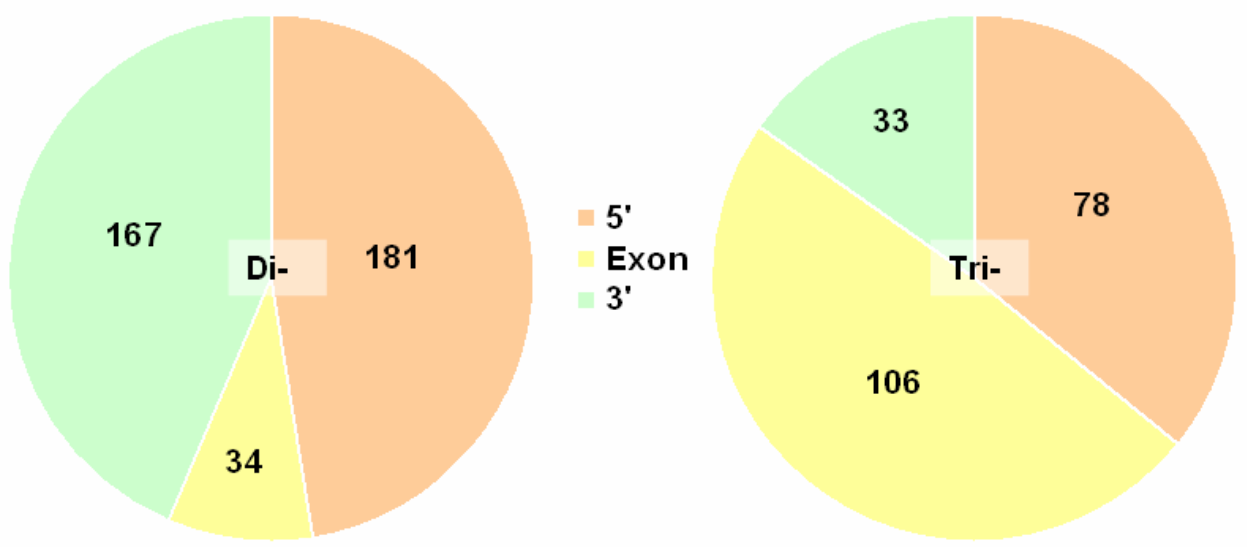


Figure 3.2. Distribution of di- and trinucleotide repeats in 5’ or 3’ untranslated regions and exons of peanut transcripts.

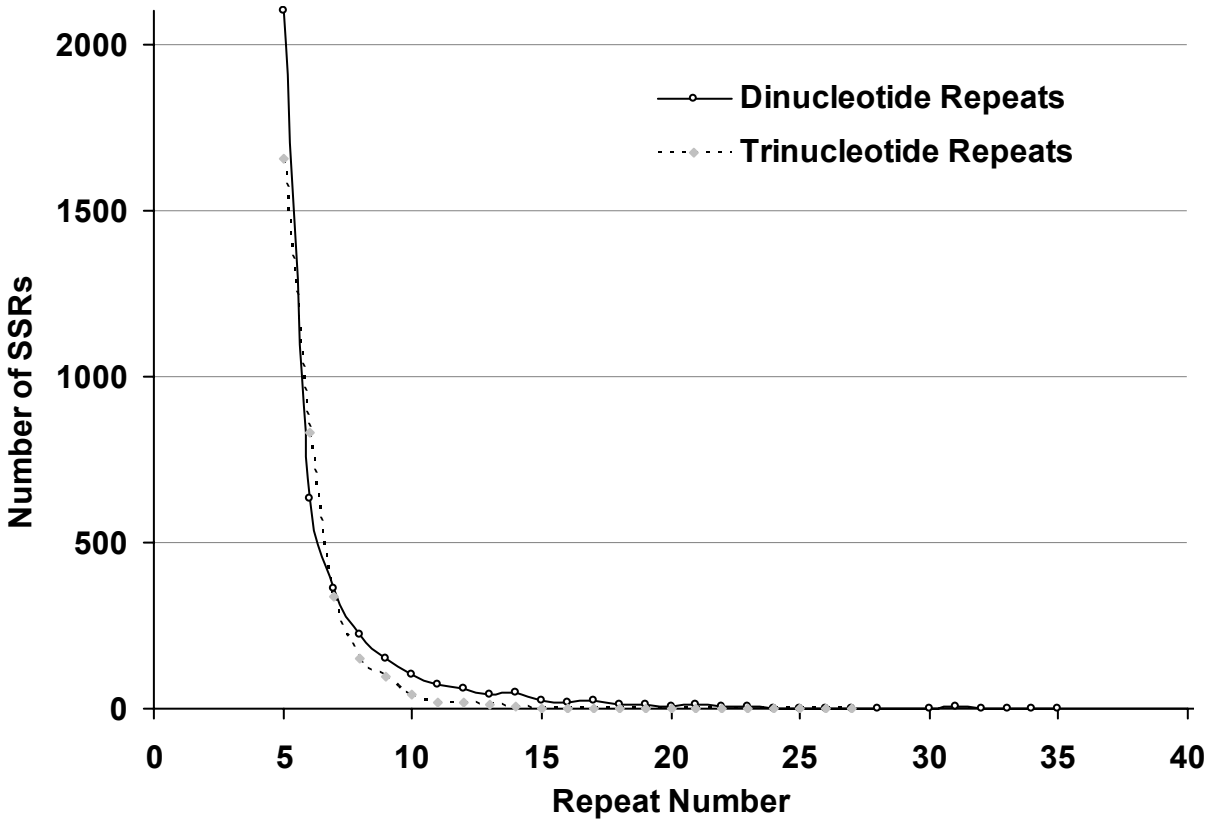


Figure 3.3. Distribution of number of repeat units among 3,949 dinucleotide and 3,176 trinucleotide repeats identified in the peanut EST database.

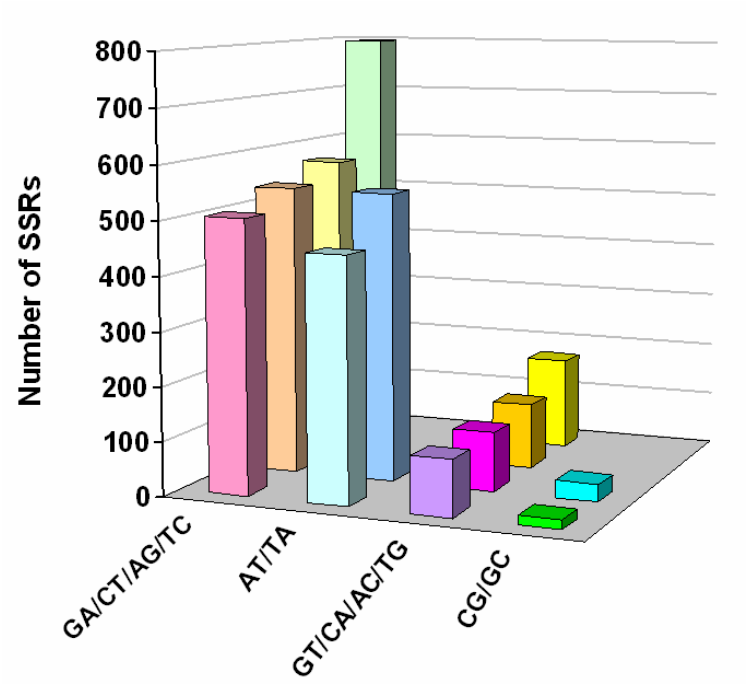


Figure 3.4. Frequencies of four dinucleotide repeat classes in *A. hypogaea* ESTs. A total of 3,949 dinucleotide repeat motifs were identified in 101,132 uniscripts from *A. hypogaea*.

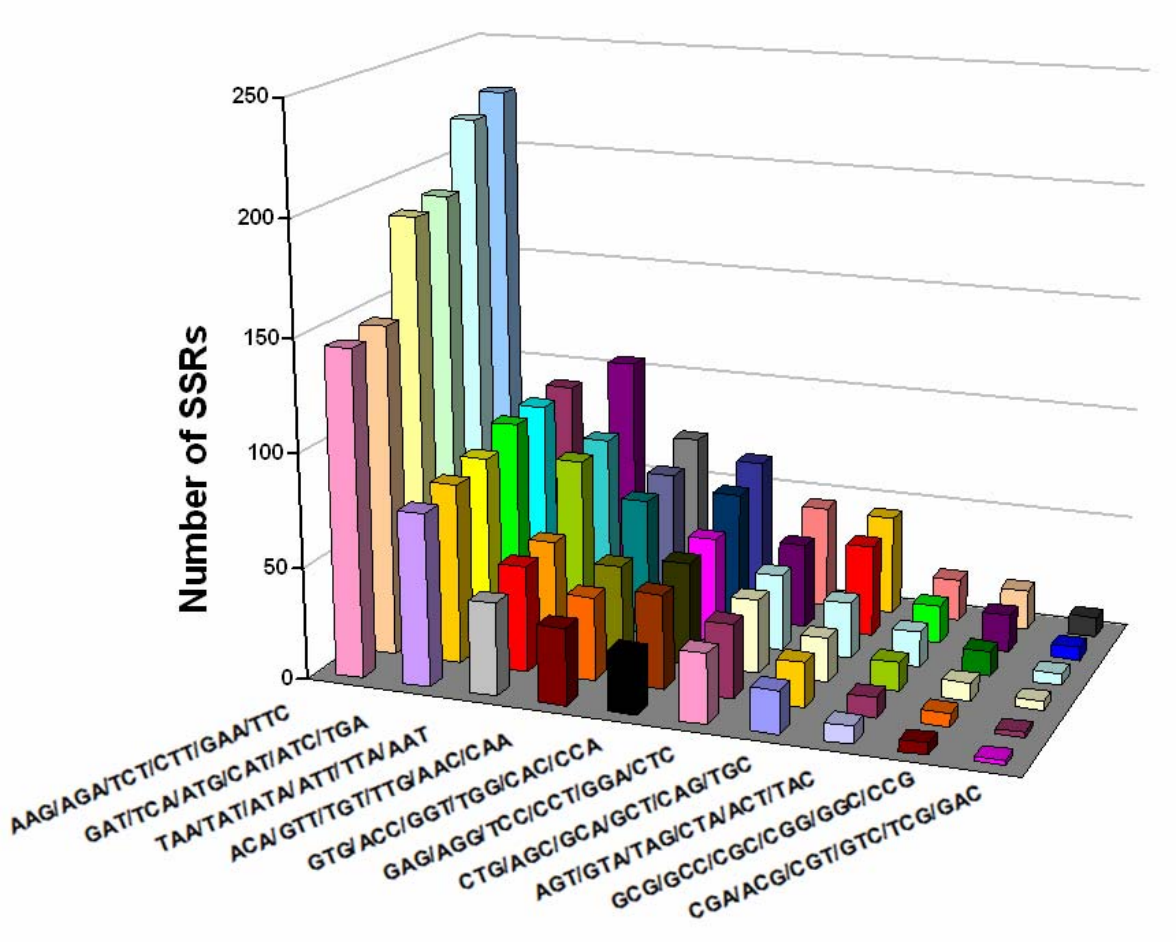


Figure 3.5. Frequencies of ten repeat motifs among 3,176 trinucleotide repeats in the peanut EST database.

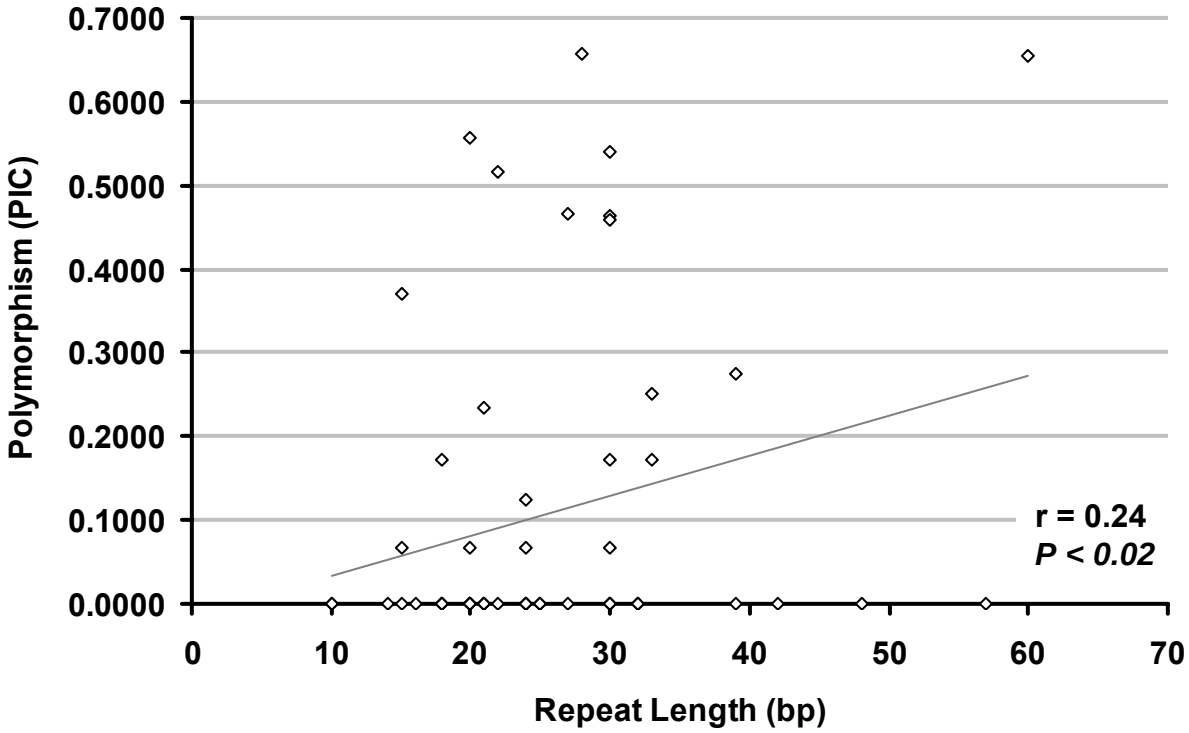
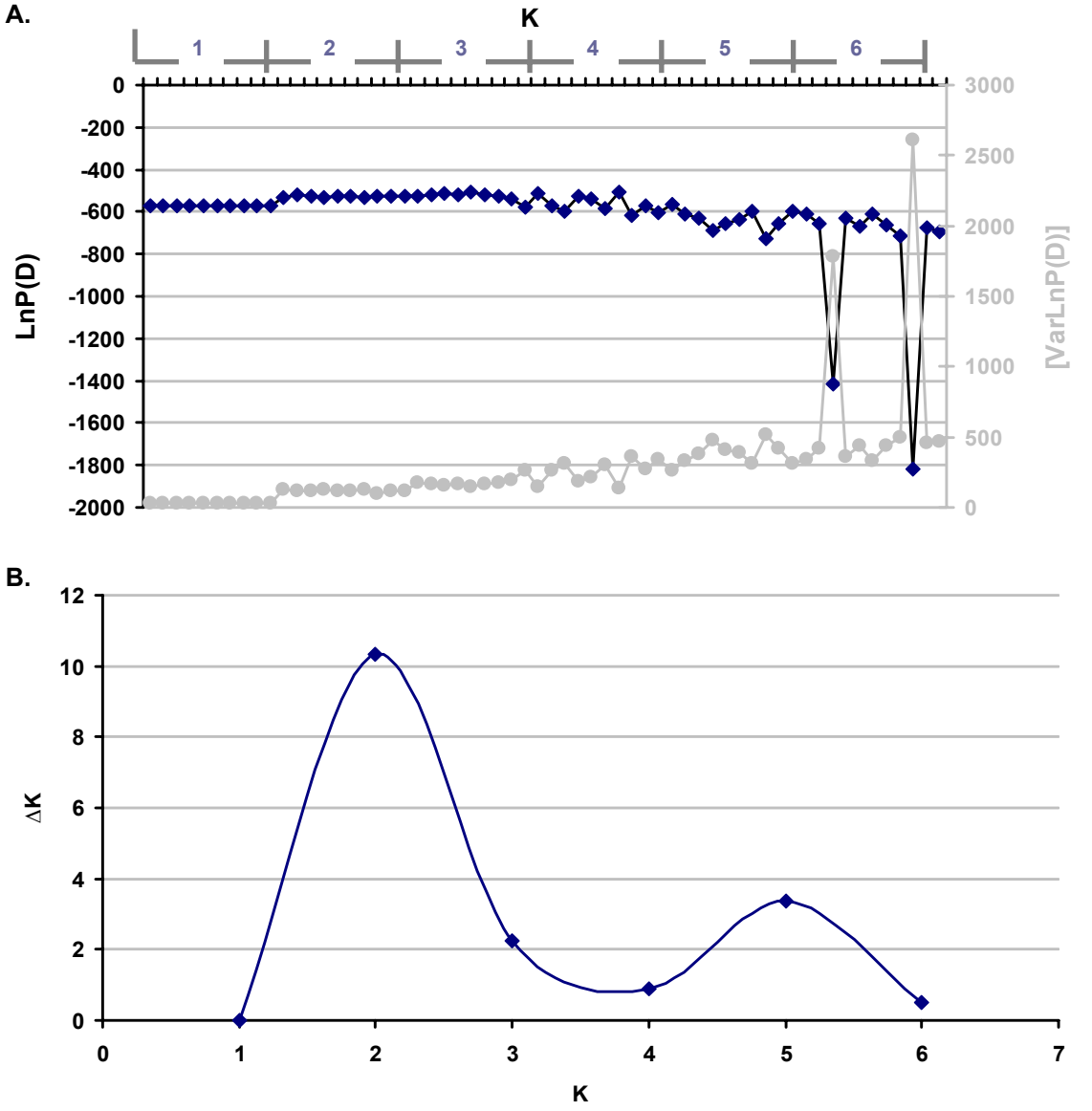


Figure 3.6. Relationship between the simple sequence repeat length (bp) and polymorphism for 59 SSR markers among 28 tetraploid peanut accessions.

Figure 3.7. Inferences on the population structure of tetraploid *Arachis* species. Population genetic software *structure* was used for the analysis. 59 simple sequence repeat (SSR) markers were used to screen a total of 28 tetraploid *Arachis* spp. (A) Estimation of appropriate population size (K) from *structure* simulation summary. The dark and light colored lines join the data points for LnP(D) and Var[LnP(D)], respectively, with marks representing the data points for ten iterations each for population sizes from 1 to 6. (B) Second order rate of change of the likelihood function with respect to K (ΔK) showing a highest ΔK = 2 (see Evanno *et al.* 2005). (C) Population structure of 28 tetraploid peanut accessions; two colors represent two populations in which the 58 accessions (individual bars) are grouped. Both colors representing a genotype suggest admixture.


A.
K
1
2
3
4
5
6
0
-200
-400
-600
-800
-1000
-1200
-1400
-1600
-1800
-2000
LnP(D)
3000
2500
2000
1500
1000
500
0
[VarLnP(D)]
B.
12
10
8
6
4
2
0
ΔK
0
1
2
3
4
5
6
7
K
C.
Tifrunner
GA Green
GA 06G
Florunner
Basse
Dixie Giant
GA207-3-4
2201
Mani
NC12C
Indio
FAV 70
N. M. Valencia A
US 608-1
GA Valencia
GT-C20
Chico
31-A
5001
NC94022
Overo Chiquitano
Sun Oleic 97R
NemaTAM
SSD 6
GP-NC WS 14
FST-1
AEQ-2
A. monticola
1.0
.80
.60
.40
.20
.00

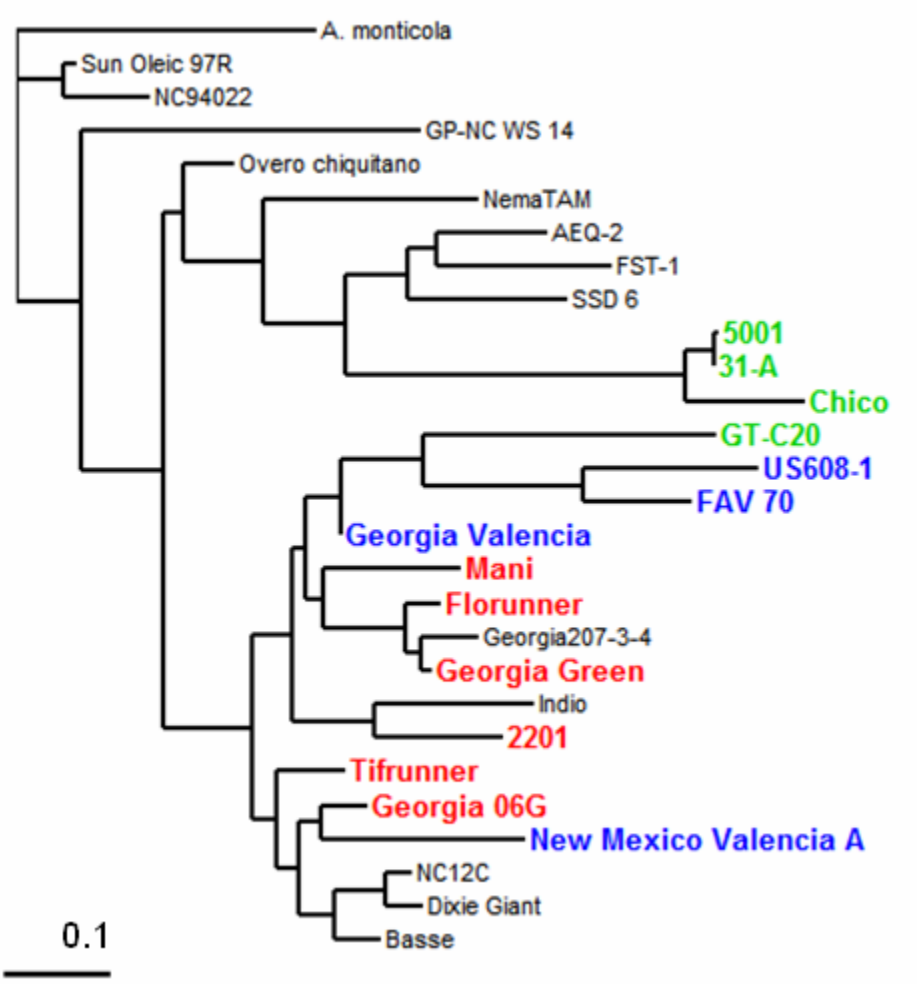


Figure 3.8. Neighbor-joining tree produced from genetic distances estimated from 59 EST-SSR markers among 28 tetraploid peanuts. Runners are shown in red, valencia in blue and Spanish in Green.

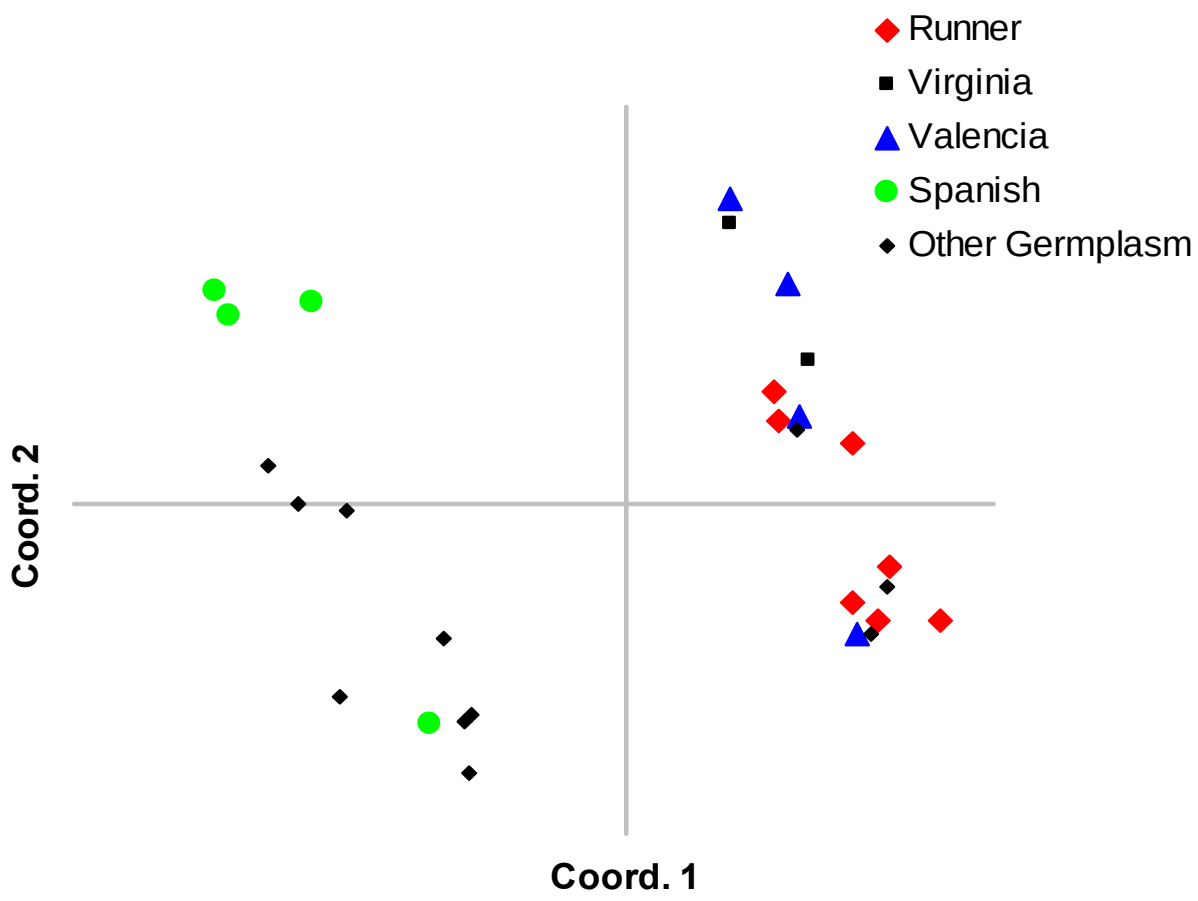
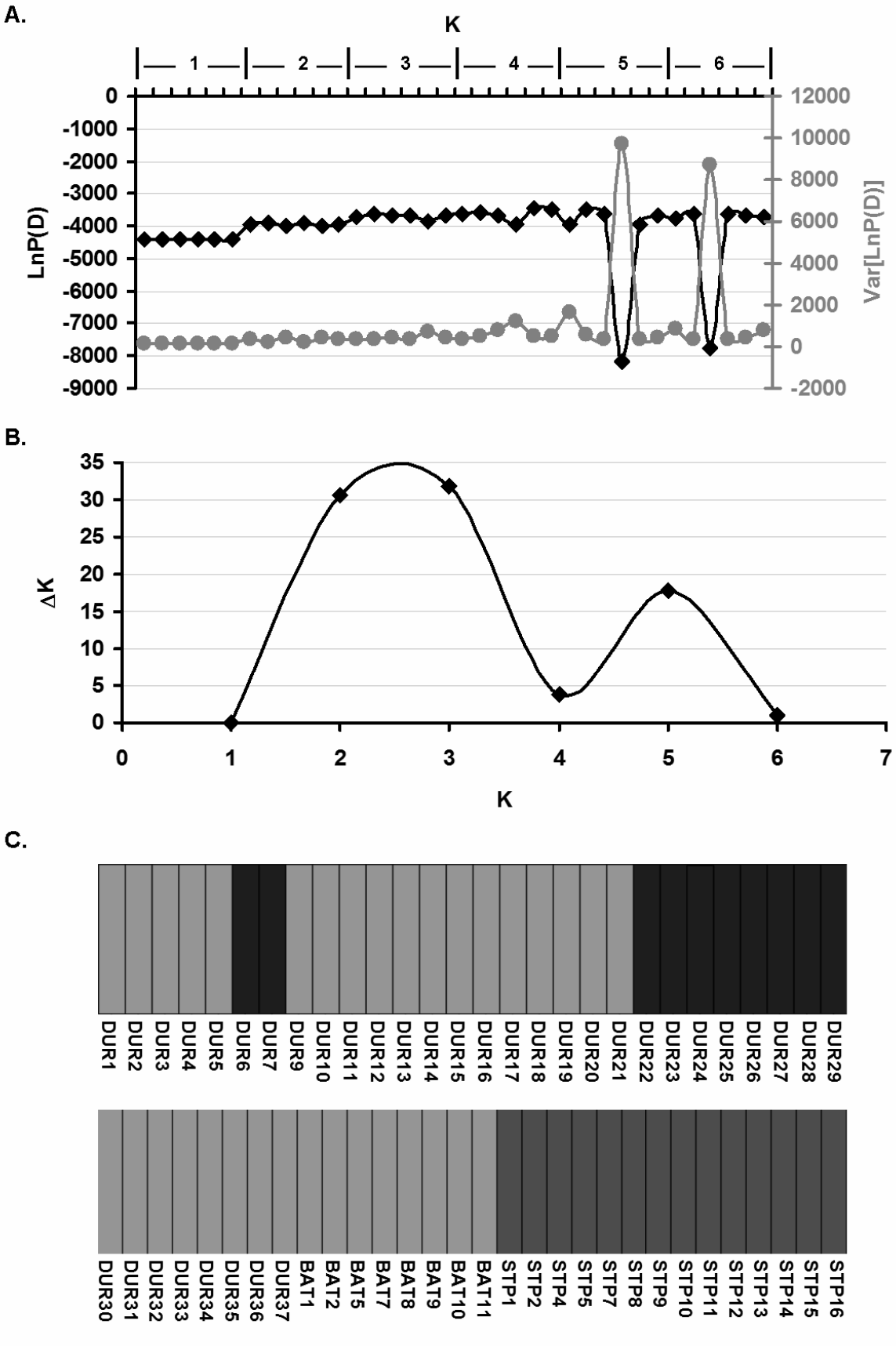


Figure 3.9. Principal coordinate analysis of mean genetic distance matrix estimated from 59 EST-SSRs genotyped in 28 tetraploid *Arachis* genotypes. The first two principle coordinates explained 36.39% and 20.23% of the total variance respectively.

Figure 4.1. Inferences on the population structure of diploid *Arachis* species. Population genetic software *structure* was used for the analysis. Twenty-seven simple sequence repeat (SSR) markers were used to screen a total of 58 wild diploids belonging to three different species. (A) Decrease in LnP(D) with concurrent increase in its variance, Var[LnP(D)], above the population size (K) of 3 (each K value iterated over 6 runs). LnP(D) is an estimate of posterior probability of the data for a given population size (K). No admixture model and correlated allele frequencies option were used for simulation with burning lengths and MCMC replications of 10,000 each. (B) Second order rate of change of the likelihood function with respect to K (ΔK) showing a highest ΔK = 3 (see Evanno *et al.* 2005). (C) Population structure of 58 diploid peanut accessions; three colors represent three populations in which the 58 accessions (individual bars) are grouped.


A.
K
1
2
3
4
5
6
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
-8000
-9000
LnP(D)
12000
10000
8000
6000
4000
2000
0
-2000
Var[LnP(D)]
B.
35
30
25
20
15
10
5
0
ΔK
0
1
2
3
4
5
6
7
K
C.
DUR1
DUR2
DUR3
DUR4
DUR5
DUR6
DUR7
DUR9
DUR10
DUR11
DUR12
DUR13
DUR14
DUR15
DUR16
DUR17
DUR18
DUR19
DUR20
DUR21
DUR22
DUR23
DUR24
DUR25
DUR26
DUR27
DUR28
DUR29
DUR30
DUR31
DUR32
DUR33
DUR34
DUR35
DUR36
DUR37
BAT1
BAT2
BAT5
BAT7
BAT8
BAT9
BAT10
BAT11
STP1
STP2
STP4
STP5
STP7
STP8
STP9
STP10
STP11
STP12
STP13
STP14
STP15
STP16

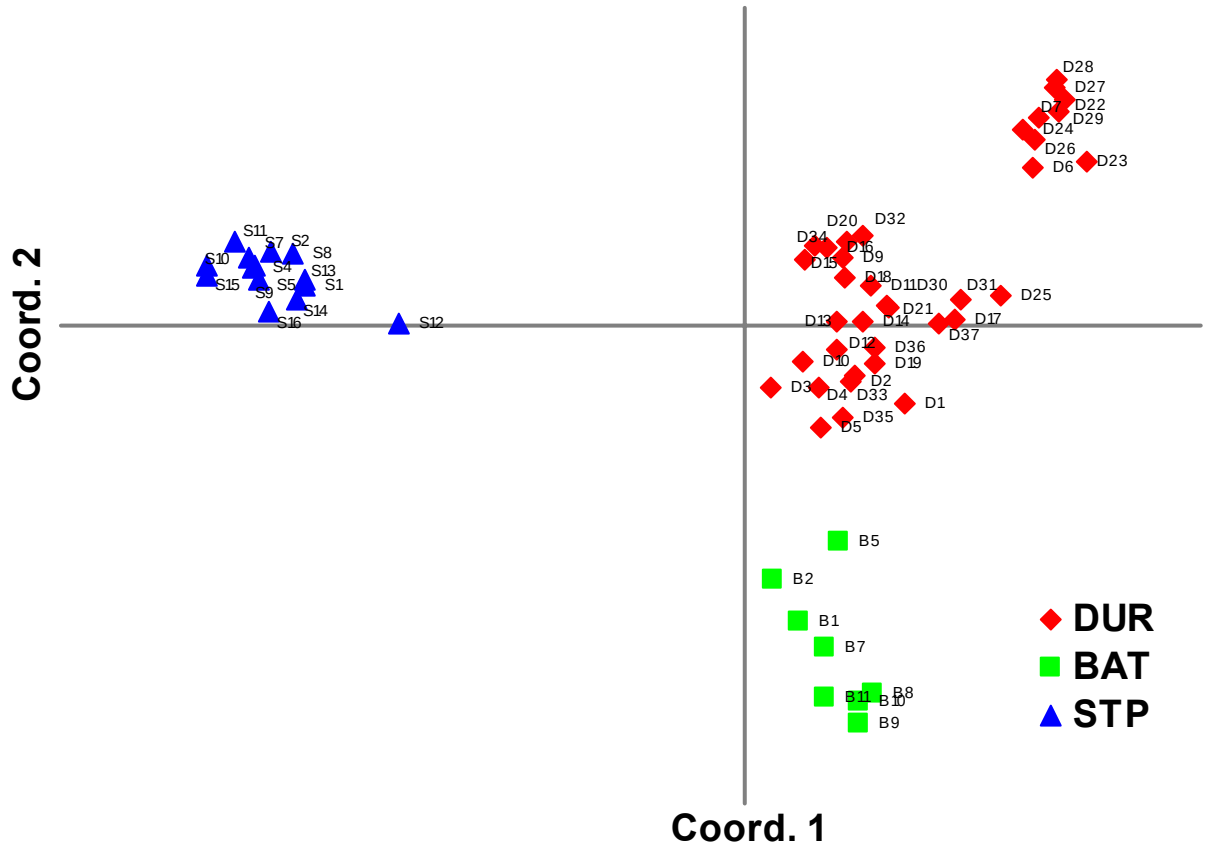


Figure 4.2. Principle coordinate analysis of mean genetic distance matrix estimated from 27 mapped SSR markers genotyped in 58 wild diploid accessions including 36 *A. duranensis*, 8 *A. batizocoi*, and 14 *A. stenosperma*. The first two principle coordinates explained 33.71% and 21.14% of the total variation respectively.

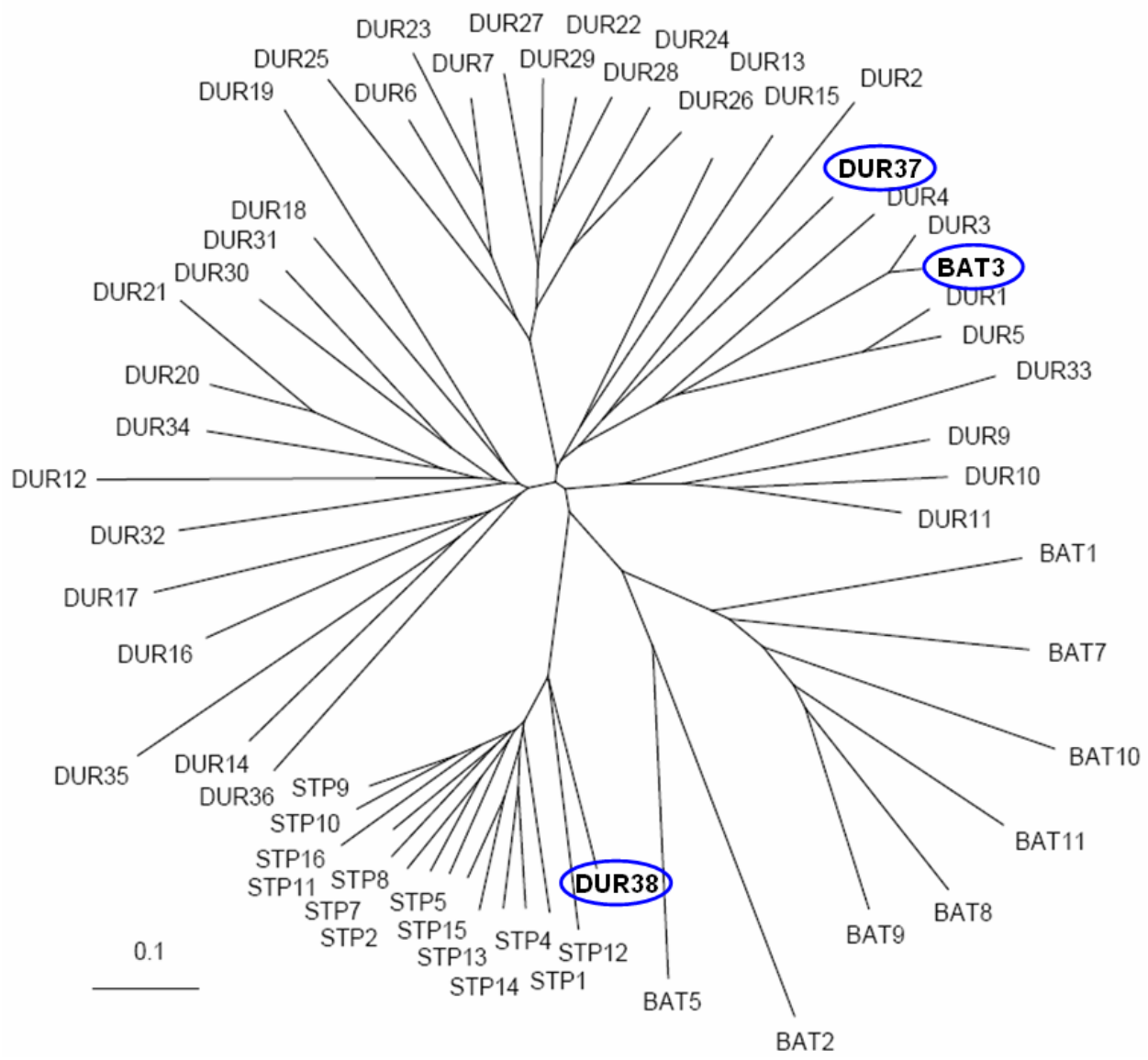


Figure 4.3. Neighbor-joining tree produced from the genetic distances estimated from 27 SSR markers screened among 60 diploid *Arachis* accessions including 37 *A. duranensis* (DUR), 9 *A. batizocoi* (BAT), and 14 *A. stenosperma* (STP). Marked accessions were reclassified [DUR37 (PI 468324) from *A. batizocoi* and DUR38 (PI 468321) from *A. ipaensis*] and/or suggest misclassification [DUR38 and BAT3 (Grif 15031)].

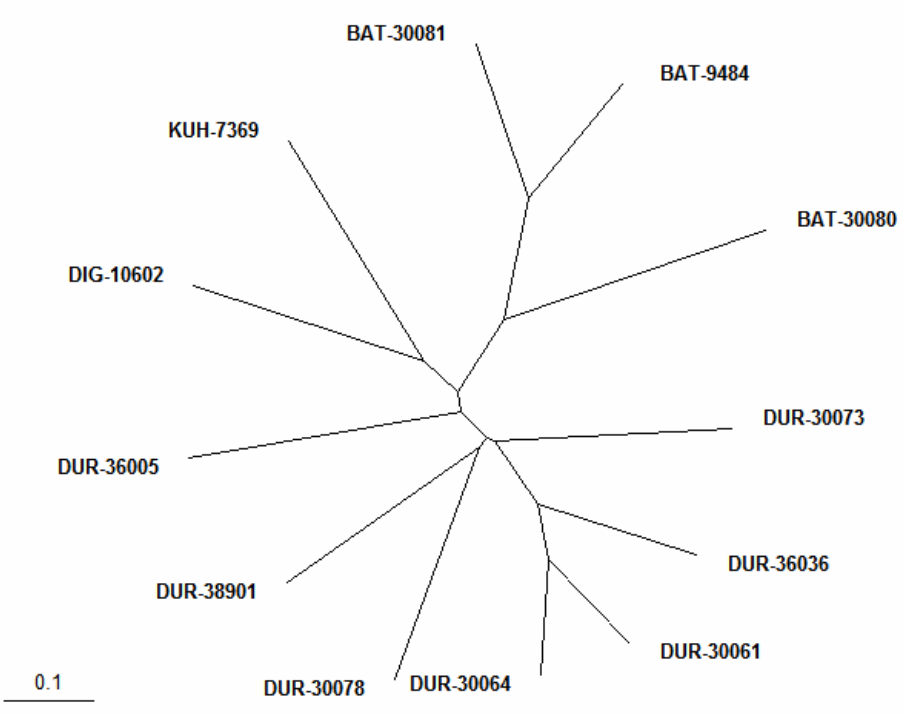


Figure 4.4. NJ tree constructed from genetic distance matrix estimated from 556 SSR markers screened among 7 *A. duranensis*, 3 *A. batizocoi*, 1 *A. diogoi*, and 1 *A. kuhlmanii*.

CHAPTER 5

SUMMARY

We followed a massively parallel DNA sequencing and SSR development approach to offset the gridlock posed by a lack of DNA marker resources for genetic mapping, and genomics and molecular breeding applications in peanut. Expressed sequence tags (ESTs) and genome survey sequences (GSSs) were developed, mined, and characterized for simple sequence repeats. High frequencies of SSRs were observed in GSSs (one/4.7kb) and ESTs (one/5kb); also, both were rich sources of polymorphic SSRs. Dinucleotide repeats were more abundant than trinucleotide and tetranucleotide repeats both in the genic and non-genic regions of peanut genome, and within the genic regions, they were more prevalent in UTRs than in exons. Transcriptome-based SSRs were more transferable than the genomic-based SSRs, while genomic-based SSRs were more polymorphic than the transcriptome-based SSRs in peanut. Also, longer SSRs (>26 bp) were significantly more polymorphic than shorter ones (<26 bp). Based on the observations from our pilot study, we designed and tested 2,054 EST-SSR markers among the parents of diploid and tetraploid mapping populations. These markers would be crucial in comparative mapping and synteny analysis among legumes. We also developed a large number of informative markers (532) by screening 612 previously reported and 97 newly developed SSRs among the parents of several diploid mapping populations. Molecular genetic diversity among A- and B-genome diploid, and AB-genome tetraploid peanut suggested prospects for the construction of intraspecific diploid and tetraploid linkage maps; albeit, the frequency of polymorphic markers for a tetraploid mapping population was considerably lower

than that for the intraspecific diploid mapping populations. Therefore, we identified and developed intraspecific A- and B-genome diploid mapping populations as diploid models for peanut genomics. The frequencies of polymorphic markers in these populations seemed to be more than sufficient for developing a critical mass of DNA markers for genomics and molecular breeding applications in cultivated peanut (unpublished). Further, we assessed the utility of methylation-filtering approaches for gene enrichment and genespace sequencing in A- and B-genome diploids, and AB-genome tetraploid *Arachis* spp. At least 4-fold reduction in sampled genome size suggested feasibility of MF sequencing strategy for reducing the genomic complexity in larger genomes like *Arachis* where hypermethylated fraction of the genome is overrepresented in the genomic libraries. MF showed effective genespace enrichment and reduced representation of hypermethylated fraction of the genome in the sequence library.