Project Title: Development and Use of SNP-Based Markers for Eventual Development of a Peanut DNA Chip for Marker-Assisted Breeding.

Project Investigators:
Mark D. Burow, Thea A. Wilkins, and Ratan Chopra -Texas Tech University, Dept. of Plant and Soil Science, Lubbock TX 79409
Charles E. Simpson -Texas AgriLife Research, 1229 US Hwy 281, Stephenville, TX 76401

Cooperating Personnel:
G. Burow, USDA-ARS-CSRL - 3810 4th Street Lubbock, TX 79415
A. Farmer, J. Mudge, I. Lindquist, G. D. May - National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505.

Lay Summary:
Analysis of gene sequences among 4 southwest peanut cultivars indicated from 3,000 to 6,500 genes had sequence differences among them. Comparison to the standard Tifrunner gene sequence assembly used to develop the aligned new sequences suggested that results were not consistent if the Tifrunner sequence was included. We have sequenced genes from leaf, root, and pod tissue of 18 new peanut accessions, for a total of 22 accessions, 12 of them in the cultivated species, and 10 being wild species. Assembly of the short sequence reads appears to be more useful than alignment against the Tifrunner sequence. An average gene read length of approximately 1500 base pairs or longer has been obtained for most accessions so far, and it is estimated that there are from 34,000 to 41,000 genes in diploid wild species, and from 66,000 to 68,000 genes in the cultivated species. The goal is to identify a very common type of difference among peanut genes (single-nucleotide polymorphisms, SNPs) in a broad array of peanut germplasm, and then develop a new type of DNA marker that can be miniaturized (as computer chips have become ever smaller and more powerful) and used for much faster DNA marker-assisted breeding of peanut.

I. Technical Abstract

Alignment of Solexa (2x54) transcriptome reads of 4 peanut cultivars against a custom reference comprised of the Tifrunner 454 TSA and published ESTs produced approximately 40K contigs, with from 6.4 to 10.6K SNPs between any of these four accessions and the reference. However, the number of SNPs between Tamrun OL07 and the reference was almost triple the number of SNPs between Tamrun OL07 and OLin, suggesting that the number of SNPs compared to the reference was too high. We have since had RNASeq transcriptome (2x50) sequences made for pod, root, and shoot tissue of 18 additional peanut accessions, 8 of them in the cultivated species, and 10 being wild species. Assembly of the short sequence reads appears to be more useful than alignment against the Tifrunner sequence. An average read length of approximately 1500 base pairs or longer has been obtained for most accessions so far, compared to from 500 to 750bp for the alignment against the reference. Based on preliminary transcriptome assemblies, it is estimated that there are from 34,000 to 41,000 genes in diploid wild species, and from 66,000 to 68,000 genes in the cultivated species. Attempted validation of SNPs using KASP primers revealed that the 3 primer sets made in accord with established rules were validated, but that adherence to proper guidelines will be needed for SNP analysis to succeed. An A-genome mapping population developed from a cross between *A. cardenasi*i and *A. duranensis* has been expanded to almost 2,500 seed. The long-term goal is to identify and use SNPs as a rapid and new type of DNA marker to accelerate DNA marker-assisted breeding of peanut.

Project Title: Development and Use of SNP-Based Markers for Eventual Development of a Peanut DNA Chip for Marker-Assisted Breeding.

Project Investigators:
Mark D. Burow, Thea A. Wilkins, and Ratan Chopra -Texas Tech University, Dept. of Plant and Soil Science, Lubbock TX 79409
Charles E. Simpson -Texas AgriLife Research, 1229 US Hwy 281, Stephenville, TX 76401

Lay Summary:
Analysis of gene sequences among 4 southwest peanut cultivars indicated from 3,000 to 6,500 genes had sequence differences among them. Comparison to the standard Tifrunner gene sequence assembly used to develop the aligned new sequences suggested that results were not consistent if the Tifrunner sequence was included. We have sequenced genes from leaf, root, and pod tissue of 18 new peanut accessions, for a total of 22 accessions, 12 of them in the cultivated species, and 10 being wild species. Assembly of the short sequence reads appears to be more useful than alignment against the Tifrunner sequence. An average gene read length of approximately 1500 base pairs or longer has been obtained for most accessions so far, and it is estimated that there are from 34,000 to 41,000 genes in diploid wild species, and from 66,000 to 68,000 genes in the cultivated species. The goal is to identify a very common type of difference among peanut genes (single-nucleotide polymorphisms, SNPs) in a broad array of peanut germplasm, and then develop a new type of DNA marker that can be miniaturized (as computer chips have become ever smaller and more powerful) and used for much faster DNA marker-assisted breeding of peanut.

II. Main Body:
Project Title: Development and Use of SNP-Based Markers for Eventual Development of a Peanut DNA Chip for Marker-Assisted Breeding.

Project Investigators:
Mark D. Burow, Thea A. Wilkins, and Ratan Chopra -Texas Tech University, Dept. of Plant and Soil Science, Lubbock TX 79409
Charles E. Simpson -Texas AgriLife Research, 1229 US Hwy 281, Stephenville, TX 76401

Cooperating Personnel:
G. Burow, USDA-ARS-CSRL - 3810 4th Street Lubbock, TX 79415
A. Farmer, J. Mudge, I. Lindquist, G. D. May - National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505.

Objectives:
1.Catalog SNPs from a set of 4 accessions, and work on developing SNPs from additional accessions.
2. Validate SNPs in a small set of genes.
3. Develop a set of 1536 SNPs from Solexa sequences that can be used for SNP analysis
4. Continue developing an A genome mapping population.

Procedures:
1. Develop and Catalog SNPs. We completed the sequencing of the 4 southwestern varieties, planted materials to collect tissue for the remainder of the sequencing work. RNA was extracted from leaf, root, and developing pod tissue, and sent to NCGR for library synthesis and sequencing. Root, shoot, and pod tissues from 18 additional accessions were collected from at least 9 plants per accession, and RNA was extracted. RNA was sent to NCGR for library synthesis and sequencing on an Illumina HiSeq 2000 (2 x 50).

2. Validate SNPs in a small set of genes. We selected 6 genes and attempted to validate the SNPs by real-time PCR. KASP-based primers were synthesized and validated on a Roche

Lightcycler 480. This was done using the original 4 accessions sequenced, and then in larger sets of 24 and 48 accessions.

3. Develop a set of 1536 SNPs from Solexa sequences that can be used for SNP analysis. We had >10,000 genes with differences among accessions; however, as the number of SNPs between our accessions and the Tifrunner TSA was too large, and we concluded that there must have been some errors, and we decided to wait on this step until the new sequence data were available and would allow performing this step with confidence that it was being done accurately.

4. Continue develop an A genome mapping population, as well as tetraploid populations. Seeds of the F_1 generation had been planted, and harvest of F_2 seeds was continued.

Results and Discussion.

This phase of the project had several goals, with the aim of finding a large number of differences between genes of four different southwest peanut varieties, expanding this to a wider number of peanut accessions once we had succeeded with the four varieties, and then testing these differences and methods of using the differences for marker-assisted breeding.

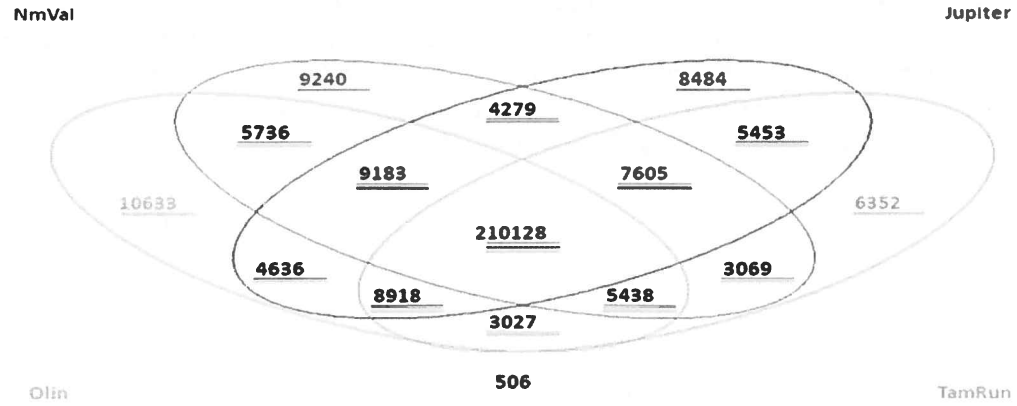
1. Sequence additional accessions to expand the applicability of SNP analysis to peanut mapping and breeding populations. RNA isolation from leaf, root and pod of four U.S. market types namely Jupiter, New Mexico Valencia C, TamrunOL07, and OLin was performed in 2011 using Trizol method followed by Qiagen clean up. Sequencing of the above mentioned genotypes on GAIIx Analyzer produced on average of 40 million reads per accessions. A custom reference of 43,102 contigs >200bp was generated by pooling the Arachis EST database and Arachis hypogaea TSA at NCBI by removing redundancy. Alignment of the sequences from four genotypes was performed against the custom reference using Alpheus, and on average 39K contigs were aligned to reference (see Table 1). From 6300 to over 10,000 SNPs were found in comparision to the reference.

Table 1. Results from Sequencing of four cultivars.

Genotype	Raw reads	Filtered reads	No. of contigs aligned	Number SNPs compared to Custom reference
Jupiter	54,428,708	43,494,034	39,803	8,484
New Mexico Valencia C	56,27,7178	43,327,345	40,267	9,240
Tamrun OL07	49,941,892	41,601,127	39,539	10,633
OLin	45,179,053	38,335,246	39,619	6,352

SNPs among the four cultivars that we had sequenced were identified using allele frequency >10% and unique reads >2. This estimate was revised after a re-analysis of the data, and it was found that approximately 4-6.5 K SNPs were identified in pair-wise combination of accessions (Figure 1).

Figure 1. Revised number of SNPs distinguishing cultivars.



A major concern was that the number of SNPs between any of the four cultivars that we had sequenced and the reference was greater than the number of SNPs among the four cultivars that we had sequenced. This was unexpected; we had assumed that Tamrun OL07 and Tifrunner would be the most similar, and either OLin or New Mexico Valencia C would be the most different from Tifrunner. But, instead, the number of SNPs between Tamrun OL07 and the reference was 10,655, and the number of SNPs distinguishing Tamrun OL07 and OLin was only 3,027. This suggested that the accuracy of the reference was not as good as we had expected.

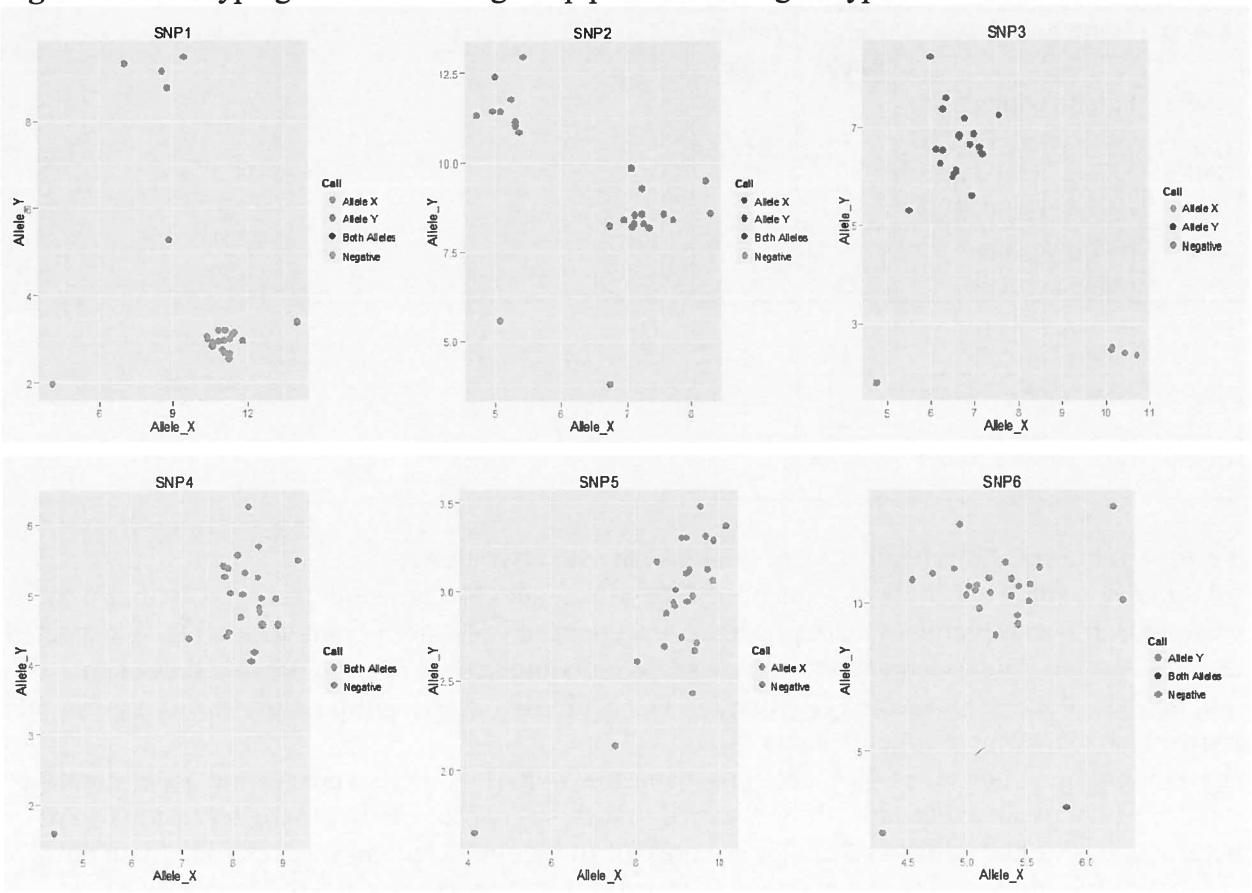
2. Testing of SNP-based markers. We have had synthesized 6 sets of SNP-based primers as an initial validation of sequencing results. A subset of 6 SNPs on array of 48 genotypes including wild species was analyzed on Roche Lightcycler 480 using KASP chemistry. Two primers were made for the FAD2 A/ and B genes, two for pathogenesis-related proteins, and two for heat shock proteins. The first three primers sets were made according to established rules for design of primers. The last three were made in violation of several rules (see Comments section), to determine how strict adherence needed to be for successful results. In the case of SNP4, this was based on a high number of reads, but no evidence for the presence of a variant allele missing in any of the genotypes. In the case of SNPs 5 and 6, the number of reads was too low to distinguish with certainty whether there were other alleles present.

Table 2. SNPs used for testing of KASP primers.

Name	Gene annotated	Ref. Allele	Variant Allele
SNP1	omega 6 fatty acid desaturase FAD2B	-	A
SNP2	omega 6 fatty acid desaturase FAD2A	G	A
SNP3	pathogenesis-related protein	G	C
SNP4	pathogenesis-related protein	A	T
SNP5	heat shock protein	G	T
SNP6	heat shock protein	A	G

Testing of the six SNPs against 48 accessions revealed two things.
(a) Against a small set of the 4 sequenced accessions, and 24 accessions, the genotyping went as expected. It was possible to distinguish the 4 sequenced accessions from each other as expected, as based on the Solexa sequences. Likewise, two distinguishable genotypes could be seen among the 24 accessions tested **(Figure 2)**. For SNPs 4-6, it was not possible to distinguish the 4 sequenced genotypes from each other.
(b) Among the larger set of 48 accessions, even for SNPs 1-3, there were several accessions that did not fit the expected groups. It is assumed that this is because either these accessions were heterozygous or had additional alleles not present in the four sequenced accessions (data not shown). This suggested the need to obtain transcriptome sequence data of a larger set of accessions.

Figure 2. Genotyping of 6 SNPs using Kasp primers on 24 genotypes.



3. Expansion of the Set of Sequenced Accessions.

We collected and extracted leaf, root tip, and pod tissues from 11 additional peanut accessions (one each runner, Virginia, Spanish, and Valencia, and two each each Peruviana, Aequatoriana, and hirsuta botanical types) from the U.S. peanut minicore collection or type specimens, plus one Florunner component line. Pod samples were collected at the yellow, brown, and black stages of maturity. In addition, we added 8 diploid wild species accessions of A, B, and K genome, plus the synthetic amphidiploid TxAG-6, plus *A. monticola*, the only other tetraploid species in section Arachis. It was hoped that including several diploid accessions would assist in separating the homoeologous gene sequences.

From *A. hypogaea,* an additional accession each of runner, Virginia, Spanish, and Valencia market types were selected from the minicore collection, except for the Spanish accession BSS56, which was selected instead because its phenotype was so different from other Spanish accessions. One accession each of Peruviana, Aequatoriana, and hirsuta botanical types were sequenced; the second accession of each of these three botanical types was not sequenced due to limitations on funding. A list of genotypes is provided in Table 3.

Table 3. Accessions planted for RNA extraction and DNA sequencing. The original four cultivars are included in the list.

Species	Subspecies	Market Type	Genotype	Origin
A. hypogaea	hypogaea	runner	Tamrun OL07	Texas
A. hypogaea	hypogaea	runner	PI290538/COC224	India
A. hypogaea	hypogaea	runner	Florunner	Florida
A. hypogaea	hypogaea	Virginia	Jupiter	Oklahoma
A. hypogaea	hypogaea	Virginia	PI268868/COC367	Sudan
A. hypogaea	hypogaea	hirsuta	PI648241/En113	Ecuador
A. hypogaea	fastigiata	Spanish	OLin	Texas
A. hypogaea	fastigiata	Spanish	BSS56	West Africa
A. hypogaea	fastigiata	Valencia	New Mexico Valencia C	New Mexico
A. hypogaea	fastigiata	Valencia	PI158854/COC559	China
A. hypogaea	fastigiata	peruviana	PI502111/COC155A	Peru
A. hypogaea	fastigiata	equatoriana	PI648242/En115	Peru
Species	**Diploid/Tetraploid**	**Genome**	**Genotype**	
A. cardenasii	diploid	A	GKP10017	
A. diogoi	diploid	A	GKP10602	
A. duranensis	diploid	A	K7988	
A. duranensis	diploid	A	K38901	
A. ipaensis	diploid	B	K30076	
A. magna	diploid	B	K30097	
A. batizocoi	diploid	K	K9484	
A. cruziana	diploid	K	K36024	
A. monticola	tetraploid	AB	K30062	
synthetic amphidiploid	tetraploid	AB	TxAG-6	

RNA was isolated from leaf, root, and pod tissue of all 18 additional genotypes, from at least 9 plants per tissue type in each accession. The exception was *A. ipaënsis*, for which only 3 plants survived to maturity.

RNA was sent to NCGR for sequencing. Sequencing has been performed on these using high throughput multiplexing on HiSeq2000 analyzer at NCGR. Tetraploids were pooled and run on one lane, and diploids were pooled and run on another lane. Raw sequencing results have been received back.

We have been working on de novo assembly of sequences. This was done because the polymorphism rate compared to the TSA used previously was considered to be too high, indicating possible errors in sequence in the TSA. Therefore, we attempted to perform de novo assembly instead, to see if this would be more satisfactory.

Several programs are being used for de novo assembly. These include SOAP transdenovo, AByss, and Trinity. For SOAP andAByss, the kmer length is being varied, but the best results so far have come from Trinity with a kmer length of 25, using a minimum match of 95% and a minimum coverage of 2. Programs are being run on a Linux server.

The number of reads and aligned reads for data analyzed so far are given in Table 4.

Table 4. De novo assembly data from new and old sequence data.

Genotype	Raw reads	Aligned reads	N50 (bp)
A.duranensis (K38901)	16,774,125	15,073,362	1,401
A.ipaënsis (K30076)	16,206,929	14,723,818	799
OLin	41,601,127	38,461,894	1687
Jupiter	38,335,246	39,612,783	1641
TamRunOL07	43,494,034	41,786,565	1535

It was clear that de novo assembly was useful. The mean read length (N50) for the previous sequence data was increased from a mean length of from 500 to 800bp in the alignment vs. the custom TSA to over 1500 bp by de novo assembly.

In contrast to the previous alignment against the Tifrunner TSA, where approximately 40,000 contigs were aligned for the tetraploid, the de novo assembly is identifying a larger number of contigs (genes). In the diploid *A. duranensis*, a range of from 34K to 41K has been identified so far. In the tetraploid Tamrun OL07, the number of contigs is approximately double the number in the diploid, ranging from 66K to 68K. These values for Tamrun OL07 are much higher than in the alignment against the Tifrunner TSA, suggesting that A and B genome sequences in the Tifrunner TSA were collapsed on each other, and it was not possible to distinguish homoeologs from each other.

Assembly of the sequences of the remaining 17 sequences will be performed in 2013.

4. Expansion of the mapping population.
The F_2 mapping population of *A. duranensis* K38901 x *A. cardenasii* 10017 has been increased from 1500 to almost 2500 seeds. A subset of this is ready for planting in 2013.

Publications:
Chopra, R, G. Burow, A. Farmer, J. Mudge, I. Lindquist, G. D. May, M. S. Gomez, Z. Xin, C. Simpson, N. Puppala, K. D. Chamberlin, T. Wilkins, and M. D. Burow. Genome-wide Assessment of SNPs in Peanut Using Illumina (Solexa) Sequencing. Plant and Animal Genome XX, January 2012.