Title: Genotyping an Array of Peanut Germplasm, including the U.S. Peanut Minicore Collection, Southwest Varieties, and a Mapping Population using SNP-based Markers.

Mark Burow - Texas Tech University, Dept. of Plant and Soil Science, Lubbock TX 79409
Thea A. Wilkins - 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

1. Non-technical Project Summary

Summary and Benefits to US Peanut Industry:
Using technology in use by the Peanut Genome Initiative for sequencing the peanut genome, we have developed partial gene sequences of over 35,000 genes of four southwest-developed peanut varieties, and have found between 4,300 and 7,000 gene differences between different pairings of these varieties. We have also developed a large F2 population made between two A-genome diploid species, for the purpose of mapping these single nucleotide polymorphisms (SNPs) in peanut, as a step towards their use in marker-assisted breeding.

In this experiment, we have found 40 to 100 times the number of differences between varieties that the current microsatellite-based technology has to offer. This demonstrates the power of the new genome sequencing technology, and promises to make marker-assisted breeding more effective, and speed up the development of new varieties for the southwest peanut growing region and its specific needs.

2. Results.

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.

a. Identify SNPs from peanut varieties, minicore accessions, and wild species.
We extracted total RNA from leaf, root and pod tissues of greenhouse grown plants of four peanut varieties developed by Southwest US breeding programs, namely Tamrun OL07 (runner), OLin (Spanish), New Mexico Valencia C (Valencia), and Jupiter (Virginia). Total RNA was used for generating the sequence of genes being expressed in these tissues using the next generation Illumina sequencing technology being used for the peanut genome sequencing project. This was sent to the National Center for Genome Resources in Santa Fe, NM for sequencing.

Sequences were aligned to the reference of 46,813 genes (contigs) which was generated by combining publicly-available expressed gene DNA sequences (ESTs) and the 454 sequence (Transcriptome Shotgun Assembly) of Arachis hypogaea cv. Tifrunner, removing the redundant contigs. Single nucleotide polymorphisms (SNPs) and insertion/deletion mutations (indels) were analyzed with NCGR's Alpheus software, using SNP calling filters >10% and >2 uniquely aligning reads with average quality scores ≥ 20. From 38 million to 44 million partial gene
sequences were obtained from each variety (Table 1). A total of 65\% of the sequences obtained were aligned to reference of which 55\% were uniquely aligned and approximately 0.3 million SNPs or indels were found in 36,102 genes (contigs) found from the sequences of the four southwestern peanut varieties in comparison to Tifrunner.

<table>
<thead>
<tr>
<th>Variety</th>
<th>Total no. of reads</th>
<th>No. of SNPs specific to each cultivar</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLin</td>
<td>38,335,246</td>
<td>15,298</td>
</tr>
<tr>
<td>Jupiter</td>
<td>43,494,034</td>
<td>12,552</td>
</tr>
<tr>
<td>NmValC</td>
<td>43,327,345</td>
<td>13,981</td>
</tr>
<tr>
<td>TamRun_OLO7</td>
<td>41,601,127</td>
<td>11,198</td>
</tr>
</tbody>
</table>

Table 1. Results of sequencing of four southwest peanut varieties. The number of SNPs specific to each variety is the number of gene variations found between that variety and all the other three varieties combined.

The number of SNPs was far in excess of the number expected, and it was found that a far-more reasonable number was obtained solely by comparing the sequences of the four southwest peanut varieties to each other (Fig. 1). A low of 4370 gene sequence differences were found between New Mexico Valencia C and Tamrun OLO7, and a high of 6922 differences were found between OLin and Tamrun OLO7. Results were presented at the international Plant and Animal Genome meeting in San Diego in January 2012.

![Venn diagram showing gene sequence differences among four peanut varieties](image)

Figure 1. Number of gene sequence differences discovered among different comparisons of the four southwest varieties. The value of 179,852 polymorphisms was found between the set of four southwestern peanut varieties compared to the Tifrunner sequence.
The reasons for the unreasonably high number of differences compared to Tifrunner remains to be seen. We suggest three possible explanations: (a) the 454 technology used for sequencing of Tifrunner was more prone to error, or a high enough stringency was not used to eliminate errors, (b) DNA libraries were prepared by different laboratories, and it is known that this can lead to major differences in results, or (c) there were major differences in gene expression, and many of the differences are due to differences in the environments in Georgia and Texas.

*Significance of the results.* These differences remain to be confirmed, but this is far in excess of the ca. 100 SSR marker differences found between any two varieties found by using the entire set of ca. 5000 SSR primer pairs. That is to say, in one experiment, we have found 40 to 100 times the number of differences that the current microsatellite (SSR, single-sequence repeat) mapping technology has to offer. This demonstrates the power of the new genome sequencing technology. In fact, researchers using SSR technology have struggled to find enough differences to allow mapping and marker-assisted selection of crossed between peanut varieties. It is possible, but just barely, and after great effort.

Also, the differences in the results between the Georgia and Texas data demonstrate that we need additional sequencing of materials useful for the southwest in the same laboratories as the current work was done in.

b. **Train one graduate student in analysis of Solexa sequencing results.**
Ratan Chopra (M.S. student, Dept. of Plant and Soil Science, Texas Tech University) was sent to NCGR in Santa Fe for an internship in the summer of 2011. He learned some aspects of how to analyze the large amounts of data (millions of sequences) using the new software available.

c. **Develop a large set (hundreds to a thousand) of SNPs from Solexa sequences that can be used for marker analysis.** This will be done using quality assessment guidelines to ensure the highest quality sequences are used.

We have developed a short list of 50 genes’ SNPs to test to confirm our sequencing results among the four southwester varieties. We have ordered the DNA primers needed to test five genes’ SNPs so far, and expect to test these soon before expanding our work.

After we have sequencing results from additional accessions (see (d) below), have confirmed our sequencing results, understand better which SNPs are the most reliable, and which are likely to be useful (not due to the presence of many (>4) copies of the gene), we will develop the larger set of SNPs to test.

We envision developing at least two sets: (1) One or more for mapping a diploid A genome population, a diploid B-genome population, and the TxAG-6 x Florunner or newer introgression population, and (2) one or more for use with the peanut minicore collection, and for specific breeding populations that we have. The reason for two sets is that we expect that many of the SNPs useful for wild species may not find differences among peanut varieties because of the much smaller difference between them. A different set or sets will be needed for peanut varieties and the minicore collection.

d. **Continue developing an A genome mapping population to allow initial mapping of markers in**
a diploid, wild species cross where the genetics are simpler. Approximately 1,500 F2 seed were obtained from an *A. cardenasi* x *A. duranensis* diploid wild species cross. *A. cardenasi* is one of the parents of the TxAG-6 amphidiploid developed previously by C. Simpson, and was the donor parent of the root-knot nematode gene in the varieties COAN and NemaTAM. *A. duranensis* is thought to be one of the two species that hybridized to form the *A. hypogaea* species that all our peanut varieties are a member of.

This population will be used for the initial mapping of SNP-based markers, and will allow future high-resolution mapping of peanut genes. This is needed, because the presence of duplicate gene copies in the cultivated peanut species means that there will be interference from the genes derived from the other genome, and peanut researchers are still trying to find a method for scoring SNP-based markers in the presence of this interference.

e. Sequence DNA of additional accessions to expand the applicability of SNP analysis to peanut mapping and breeding populations.

We have collected and extracted leaf and root tip tissues from 10 additional peanut accessions (one each runner, Virginia, Spanish, and Valencia, and two each Peruviana, Aequatoriana, and Hirsuta botanical types) from the U.S. peanut minicore collection or type specimens. Pod collection is expected to begin soon. Sequencing will be performed by NCGR and/or the Texas A&M Borlaug Center.

We have also collected and extracted leaf and root tip tissues of three A and three B genome wild species, and expect to collect pods soon also. These will be useful for mapping in diploids, understanding how to perform mapping and marker-assisted selection in cultivated peanut, and for breeding using Texas A&M wild species introgression populations.