Peanut Genome Initiative

Peanut Genome Project

2013-2014 Research Accomplishment Report to the U.S. Peanut Industry

July 31, 2014

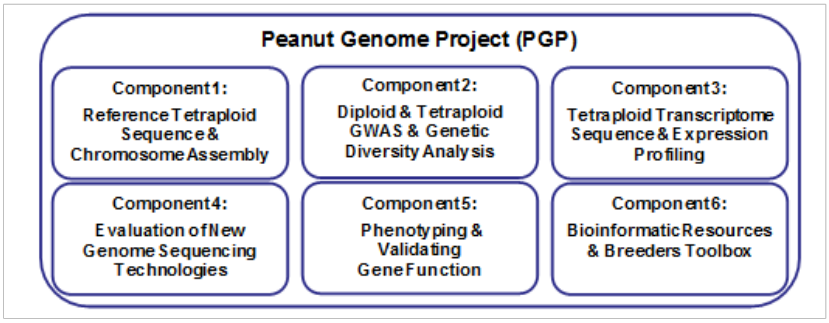
Peanut Genome Project
Research Accomplishment Report to the U.S. Peanut Industry
July, 2014

Table of Contents

Executive Summary

THE INDUSTRY CHALLENGE: *One of the biggest challenges for the U.S. peanut industry is the ability to compete with other crops for production. Most growers today are focused naturally on dollar value per acre and peanuts have often been uncompetitive in regards to yield and production costs as compared to crops such as cotton and corn. As an industry, the best way to compete is to enhance our peanut varieties for disease resistance and yield potential. This can best be done through genomics. We have to maximize yield while minimizing inputs in order to sustain and compete with other crops. The industry is committed to peanut consumption growth through marketing efforts to promote the nutritional aspects of peanuts. As we grow consumption, we must grow our yield potential to sustain our industry. Genomics is the key to a sustainable future for peanuts.*

The Peanut Genome Project (PGP) features six interactive research components in its strategic plan and is ahead of the timeline for deliverables from every component at the current writing.



Most Important 2013-2014 Accomplishment

Public release of the first chromosomal-scale assemblies for two peanut species (*Arachis duranensis* and *Arachis ipaensis*) that are believed to be the ancestral progenitors of cultivated peanuts we plant today (*Arachis hypogaea*). Both of these ancestral genomes have been duplicated within the genome of cultivated peanut. Thus, the known structure of those genomes provides a platform for assembling the *A. hypogaea* genome. This information is found at: www.PeanutBase.org/

The following summarizes key accomplishments of each component in layman's terms. The remainder of the report provides technical details of the accomplishments and goals of the six components.

Component 1 - Whole Genome Sequencing (Reference Tetraploid Sequence & Chromosomal Assembly), is being done using the variety Tifrunner and the two ancestral wild species that individually represent the A- and B- genomes of cultivated peanut. When finished, these three genome assemblies will establish the baseline for comparisons to identify needed traits in other varieties and breeding lines.

- KEY ACCOMPLISHMENT - We completed the assembly of each of the 10 chromosomes in the two ancestor species that contribute the A- and B-genomes to cultivated peanut. What makes this work extraordinary is the fact that the size of each ancestor genome is about the same as soybean. It took 5 years to assemble the soybean genome; the PGP has achieved twice as much in only two years.

Dr. Scott Jackson, University of Georgia, and chair of the project technical team remarked, "We're making good progress toward achieving a genome sequence for cultivated peanut, in fact, better progress than I anticipated given the challenges facing us. We are already seeing people begin to use the data to develop additional markers for breeding and to associate traits with the genome sequence."

Component 2, High-Density Genetic Maps and Gene Markers (**Diploid & Tetraploid Genome Wide Association Studies-GWAS & Genetic Diversity Analysis**), has made similar progress toward combining 11 individual genetic maps into one high-definition map, which is similar to combining geographic maps of

States into an atlas of the U.S. Detail of this composite genetic map was enhanced significantly by collaborative efforts between laboratories at Tifton GA, Lubbock TX and Davis CA to detect additional differences in the genetic code. This work was completed on peanuts sequences that represents 90% of the genetic traits in cultivated peanut, which include: 150 highly inbred lines from Tifrunner and other varieties developed at Tifton GA; 99 pure core lines from the Chinese germplasm collection at Wuhan, China; 108 (purified at Auburn University) mini-core lines from the U.S. germplasm collection at Griffin GA; 300 hybrid lines from EMBRAPA in Brasilia, Brazil representing the ancestral genomes; and 575 accessions from the ICRISAT germplasm collection at Hyderabad, India.

- KEY ACCOMPLISHMENT - Results so far have totaled about 200,000 unique sequence variations across the entire peanut genome. About 90% of these variations in the genetic code will be used to 'anchor' or assemble the Tifrunner genome; about 10% help define regions of a chromosome that contain genes for a desired trait, and to further improve genetic maps for plant breeding.
- Our team in India used these genetic maps to locate some of the genes for late leaf spot and rust resistance, and then selected breeding lines that are available to U.S. breeders.

"Improving peanut varieties to be more drought-, insect- and disease-resistant through molecular breeding can help farmers in developed nations produce more peanuts with fewer pesticides and other chemicals and help farmers in developing nations feed their families and build more secure livelihoods," said team member Rajeev Varshney of ICRISAT in India.

Component 3, Expressed Gene Sequences (Tetraploid Transcriptome Sequence & Expression Profile), we now know that sequence variations in the ancestral A-genome can be used to identify the same variation in the cultivated A-genome; and likewise for variations in the ancestral and cultivated B-genomes. This means that sequence variations in the ancestral genomes can be used to expedite the assembly of the cultivated genome, and will help locate the position of genes in the Tifrunner genome.

- KEY ACCOMPLISHMENT - About 25,000 unique gene-coding sequences have been identified so far and a name (function) was determined for about 50% of them; including genes for rust and late leaf spot resistance.

Dr. Peggy Ozias-Akins, University of Georgia, who is one of the leaders of this component, says, "Cataloging all the genes in the peanut genome, along with when and where in the plant body they are expressed, gives us the power to guide selection for new gene combinations to breed a better peanut".

Component 4, Evaluation of New Genome Sequencing Technologies ensures that the PGP is using the most up to date sequencing and assembly methods. Many vendors of new technology are working with PGP members to demonstrate and validate their products. These new technologies will help simplify and improve accuracy of assemblies of the cultivated genome. In addition to first-hand experience, a panel of world renowned experts in crop genome assembly was convened on March 24 and July 10, 2014 by The Peanut Foundation to help evaluate which technologies are the best for peanuts.

- KEY ACCOMPLISHMENT - The experts commended the PGP for a superior level of achievement in each of the six research components. This affirmed that the PGP is on the right track
- The experts recommended the PGP consider two optional technologies for generating about 100-fold longer pieces of DNA to simplify and improve assembly of the Tifrunner chromosomes

"Achieving a perfect peanut genome assembly is akin to landing a man on the moon. Most people thought it would be too difficult and too expensive. New technologies have made the challenge

attainable, and will allow us to better understand and optimize the key traits of the peanut for superior variety development," said Dr. Rich Wilson, technical consultant to the Peanut Foundation.

Component 5, Phenotyping Genetic Resources (Phenotyping & Validating Gene Function) is focused on validation of molecular markers to determine which markers work best in selecting important traits in peanut. This work is labor intensive and requires a team effort to associate phenotypic differences with genotypic variation for gene markers. The PGP team is now evaluating over 16 structured populations at Tifton, Marianna, Raleigh, Dawson, Griffin and Citra for resistance to CBR, late leaf spot, early leaf spot, aflatoxin contamination, white mold, TSWV; pod filling, yield, grade, drought tolerance, oil composition.

- KEY ACCOMPLISHMENT - Initial results validated genetic map positions of gene markers for: resistance to late leaf spot, resistance to early leaf spot, resistance to TSWV, drought tolerance, heat tolerance, branching habit, percent oil, and fatty acid composition.
- In addition to useful markers, lines from these populations most likely will be released as parents or varieties outright.

Corley Holbrook, USDA-ARS, who is leading this effort, recently stated "I believe that now is the time to use the recent advances in plant genomic technology to advance the science of peanut breeding and genetics. The use of gene markers and sequencing has already had a tremendous impact on my breeding program. We have recently submitted our first cultivar developed using markers for approval, and anticipate several more in the next few years. These efforts would be greatly expanded if plant breeders had access to more genetic markers".

Component 6, Bioinformatics Resources & Breeders Toolbox, is an effort lead by USDA-ARS in Ames, Iowa and the National Center for Genome Resources (NCGR) in Santa Fe, NM. 'PeanutBase' our web-site contains a Breeder's Toolbox to assist peanut breeders in selecting gene markers and varieties (whether wild or cultivated) to be used in their breeding programs. The system now features:

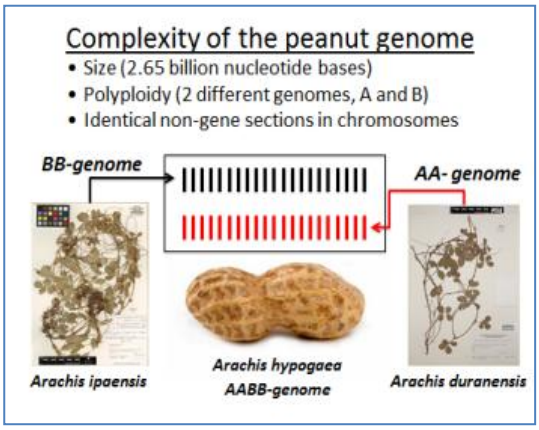
- KEY ACCOMPLISHMENT - A gene finder for 392 traits and markers
- Published genotypic & phenotypic data sets from breeding programs for key traits
- Web-browsers that help locate peanut genes based on data from soybean and other legumes
- Training sessions for PeanutBase use, one of which will available for all breeders at the international peanut genomics meeting in Savannah GA on November 11-14, 2014.

Mark Burow, peanut breeder at Texas A&M says, "Web-based genome libraries and databases will help breeders find markers that can be used in their breeding populations. The Breeder's Toolbox will allow breeders to merge genomics and phenotypic information to use in marker-assisted breeding for faster development of new varieties".

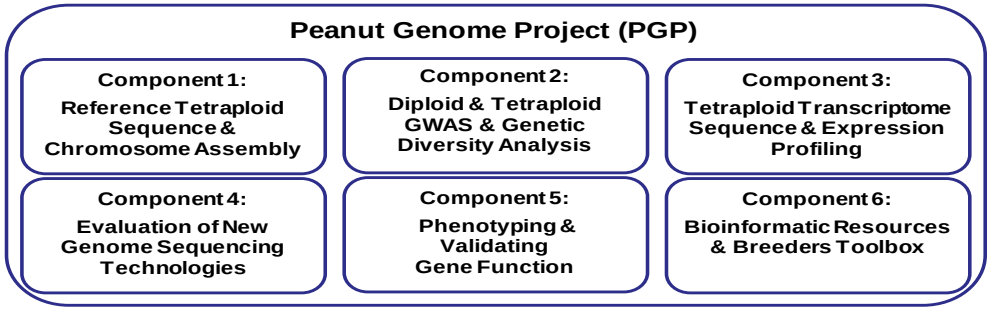
SUMMARY -*This project has made great strides in 2013-2014. Howard Shapiro, Director of Research for Mars Chocolate, who has been involved in more than 90+ plant genomic projects states, "This project has really made great progress in its second year. I expected a lot with the team of researchers assembled but they have exceeded my expectations". Even though we are ahead of schedule on every component, we still understand the urgency of completing this work as soon as possible. Research information already is being released as discovered, and many breeders are putting "marker assisted selection" to use in their breeding programs. With your continuing financial support and the emergence of even better technologies, we are certain our industry will become more competitive.*

Peanut Genome Project
Research Technical Accomplishment Report to the U.S. Peanut Industry
July, 2014

Introduction



Scott Jackson from the University of Georgia and Co-Chair of the Peanut Genome Consortium (PGC) says, "The peanut genome project (PGP) has released the first high-quality chromosomal scale draft of peanut genome assemblies. This extraordinary achievement establishes a very sound foundation for looking deeply inside the peanut at the DNA level to discover genes that control crop productivity and quality." This is a major step toward tackling the challenges presented by the cultivated peanut genome which is very large, twice the size of soybean and equal in size to the human genome. Great size makes the puzzle harder to solve. The peanut has complex structure; it contains two different genomes derived from the wild species that now have been sequenced. In addition, the peanut genome contains large sections of DNA in which nucleotide sequences repeat themselves many times. These 'repeating elements' are like spacers between gene-rich regions of the genome, but when broken into short fragments during sequencing pose problems in fitting the correct order and length when assembling a chromosome. Putting all the pieces together again in the right order requires an elegant strategy, and a team of world-class experts in genomics and peanut biology. Completion of the two wild specie genome assemblies in such a short time, is evidence that the best and brightest people are working on this project.

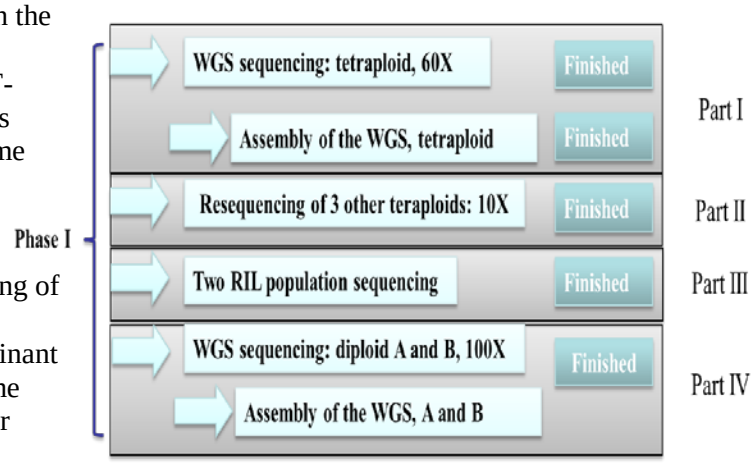


Scientists working on the peanut genome project bring a great deal of experience from other crop genome sequencing projects, and are among the best in the world, with research partners from several countries in Asia, North and South America, and Africa. The U.S. is represented by scientists at University of California-Davis; the University of Georgia at Athens and Tifton; USDA ARS at Tifton GA, Griffin GA, Ames IA and Stoneville MS; NC State University; and NCGR at Santa Fe NM. Each U.S. scientist and their international collaborators have a very specific role within the PGP Action Plan. The research contributions of each PGP member are vital to the overall mission of developing useful genetic tools that will accelerate the breeding programs for traits such as disease resistance and drought tolerance; traits that are difficult to achieve with conventional breeding strategies. PGP members are pioneers, clearing new ground with each deliberate step. This report chronicles individual responsibilities, the current state of the genome, and the strategies to move toward completion of the cultivated peanut genome sequence.

Appendices: For convenience a description of the plant materials used to establish reference genome sequences for wild and cultivated peanut is shown in Exhibit 1; a list of sponsors who provide financial support for the PGP is presented in Exhibit 2; Peanut Genome Consortium members are listed in Exhibit 3; a glossary of 'genomic' terms & definitions is presented in Exhibit 4.

Component 1: Whole Genome Sequencing.

BGI-Shenzhen, China a collaborating partner in the PGP is responsible for developing the whole genome shotgun sequences from Tifrunner, GT-C20, SunOleic 97R, NC91022, RIL populations from the T and S populations, and whole genome shotgun sequences from *A. duranensis* (AA-genome) and *A. ipaensis* (BB-genome). Phase I of the contract with BGI was conducted in four parts: 1) whole genome shot-gun sequencing of the tetraploid species, 2) resequencing of three tetraploid strains, 3) sequencing of two recombinant inbred line populations (RILs), 4) whole genome shot-gun sequencing of the two diploid ancestor species. All four parts have been finished.

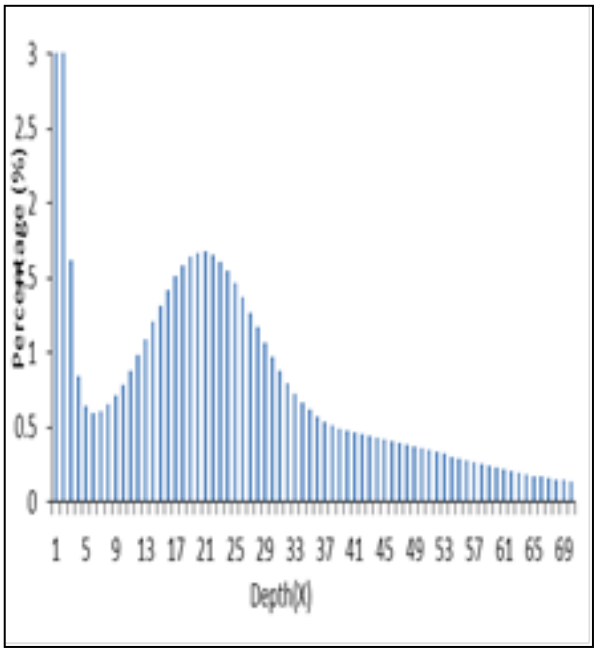


What has been learned from the sequence so far?

Whole genome shot-gun sequencing and assembly of the tetraploid genome

Eight (8) Illumina paired end sequencing libraries were created with insert sizes of 170bp, 250bp, 500bp, 800bp, 2 Kb, 5 Kb, 10 Kb, 20 Kb and 40Kb. About 327.8 Gb sequence data were generated from these libraries and 63.9% (111 Gb) were high quality. Sequence depth was estimated at over 100X coverage.

Insert Size	Reads Length	Total Data(Gb)	Sequence Depth(X)	Physical Depth(X
170bp	98	98.2	35.06	30.41
500bp	98	96.3	34.38	87.69
800bp	98	73.6	26.29	107.3
2kb	86	19.1	6.82	79.25
5kb	86	7.6	2.71	78.84
10kb	86	12.4	4.43	257.5
20kb	88	2.3	0.83	93.91
40kb	88	3.3	1.17	266.9



Kmer analysis revealed the genome size of the tetraploid species was found at the frequency of 21, and the number of kmers was 55.68 Gb. Thus the genome size was estimated to be (kmer number)/(kmer depth) = 55.68/21 ≈ 2.65 Gb.

Kmer	#Kmer	Peak depth	Genome size	Data	Reads	Depth
17	55,684,048,320	21	2,651,621,348	66,549,228,480	679,073,760	25.10

These sequences were assembled in to contigs and scaffolds using SOAPdenovo (v2.04). The resulting contig N50 (sequence length longer than 50% of the contigs) was 9.8 kb, and the scaffold N50 (sequence length longer than 50% of the scaffolds) was 81.6 kb. The total assembled sequence was 2.1 Gb without gaps (Ns) and 2.4 Gb with gaps (inserted by Ns). An updated version of SOAPdenovo was used to improve the scaffold N50 to ~100 kb. Overall, the whole genome shot-gun strategy generated more than 100 fold coverage, but only assembled the genome to contig N50 less than 10 kb and scaffold N50 less than 100 kb.

Statistics for the Tifrunner genome using WGS strategy.

	Contig		Scaffold	
	Size(bp)	**Number**	**Size(bp)**	**Number**
N90	1,768	246,198	12,036	33,839
N80	3,666	165,045	31,765	22,290
N70	5,610	118,877	48,373	16,287
N60	7,630	86,747	64,590	12,026
N50	**9,816**	62,422	**81,674**	8,733
Longest	112,663	----	720,356	----
Total Size	2,106,851,179	----	2,391,019,650	----
Total (>=100bp)	----	536,262	----	212,147
Total (>=2kb)	----	232,152	----	60,226

Resequencing of the other tetraploid stains

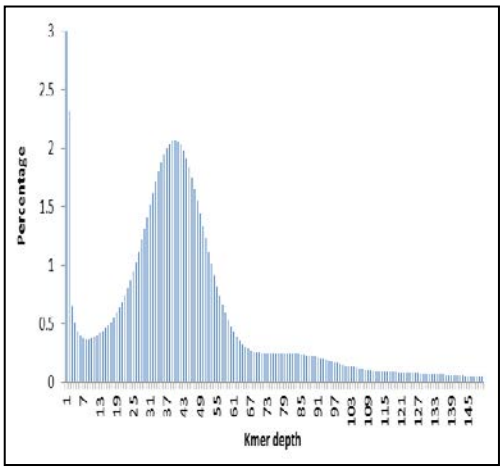
The genomes of GT-C20, SunOleic_97R and NC94022 were fragmented and one paired-end 500 bp insert library was constructed for each line. Illumina Hiseq2000 was used to generate 124 Gb raw data (102 Gb clean data) in total. The sequencing depth of each strain was about 12- fold. In addition, one paired-end 500 bp insert library was constructed from 261 lines of two RIL populations, and sequenced each of these samples was sequenced to ~3 fold (~9 Gb clean data). In total, 3,260.75 Gb raw data were generated. After filtering (the same as for the whole genome sequencing data), 2,706.49 Gb clean data was obtained for each sample, the average sequencing depth was ~3.7-fold.

Whole genome shot-gun sequencing of the diploid species *A. duranensis*

Seven (7) Illumina paired end insert libraries were created (250bp to 40Kb). About 216 Gb data equaled about 154-fold coverage of the genome.. **Kmer analysis revealed *A. duranensis genome size* ≈ 1.08Gb**

Sequencing libraries and data for *A. duranensis*.

Insert Size	Reads Length	Total Data(Gb)	Sequence Depth(X)	Physical Depth(X)
250bp	149	97.22	69.44	58.26
500bp	99	66.57	47.55	120.08
2kb	88	17.43	12.45	141.51
5kb	87	15.47	11.05	317.51
10kb	87	13.92	9.95	571.59
20kb	87	3.66	2.61	300.09
40kb	85	1.91	1.36	320.66
Total		216.18	154.42	1,829.70



Kmer	#Kmer (G)	Peak depth	Genome size (Gb)	Data (bases)	Read number	Depth
17	**43.04**	**40**	**1.076**	**48,228,377,779**	**324,140,127**	**44.82**

Contigs and scaffolds were assembled using the SOAPdenovo (v 2.05). The final assembly had contig N50 of 19.6 kb, and scaffold N50 of 1.08 Mb. The length of the assembled genome was 1.07 Gb without gaps, which was 99% of the genome size estimated by the kmer analysis. This genome assembly was anchored onto chromosomes. After the genome assembly, gene annotation revealed ~68% of the genome sequence was composed of repeat elements and the assembly contained 37,955 protein coding genes.

Statistics of the *A. duranensis* genome.

	Contig		Scaffold	
	Size(bp)	**Number**	**Size(bp)**	**Number**
N90	535	133,457	5,337	4,657
N80	3,341	50,141	246,140	931
N70	8,703	30,829	500,080	590
N60	14,057	21,258	772,611	397
N50	19,596	14,839	1,079,637	267
Longest	221,145	----	9,325,046	----
Total Size	1,067,715,028	----	1,205,530,677	----
Total (>=100bp)	----	774,676	----	635,165
Totalr(>=2kb)	----	64,435	----	7,854

Repeat content in the *A. duranensis* genome

	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (kb)	**%**	**Length (kb)**	**%**	**Length (kb)**	**%**	**Length (kb)**	**%**
DNA	14,206	1.17	19,561	1.62	41,051	3.39	52,675	4.35
LINE	7,915	0.65	14,530	1.20	16,746	1.38	25,502	2.11
LTR	106,294	8.78	150,273	12.41	416,978	34.4	524,520	43.33
SINE	17	0.00	0	0.00	2,895	0.24	2,912	0.24
Other	35	0.00	0	0.00	0	0.00	35	0.00
Unknown	24	0.00	16	0.00	292,063	24.1	292,098	24.13
Total	130,291	10.7	184,365	15.23	742,090	61.3	812,380	67.10

Gene annotation of the *A. duranensis* genome.

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	*Augustus*	43,940	2,145.18	974.32	4.41	220.87	343.23
	GlimmerHMM	47,209	1,813.40	810.49	3.65	222.17	378.75
Homolog	*G. max*	29,560	3,061.61	1,150.56	4.83	238.27	499.13
	L. japonicus	34,281	2,058.86	865.38	3.57	242.08	463.52
	M. truncatula	27,656	3,022.48	1,074.48	4.37	245.66	577.38
	P. vulgaris	27,511	3,057.28	1,175.56	4.94	237.97	477.60
EST		314,548	6581.91	936.79	3.41	265.30	6336.86
GLEAN		38,632	2,265.29	1,021.83	4.29	238.02	377.60
RNAseq		65,928	3,119.19	931.85	4.76	195.66	343.07
Final set		37,955	2,273.64	1,030.20	4.30	239.38	376.39

Whole genome shot-gun sequencing of the diploid species *A. ipaensis*

Seven (7) Illumina paired end insert libraries were created (250bp to 40Kb). 254 Gb data accounted for ~182-fold coverage of the genome. **Kmer analysis of the *A. ipaensis* genome indicated a genome size of 1.36 Gb.** The final assembly had contig N50 of 19.7 kb, and scaffold N50 of 6.19 Mb. The length of the assembled genome was 1.39 Gb without gaps. This assembly was anchored onto chromosomes. After gene annotation, ~72% of the genome contained repeat elements plus 42,250 protein coding genes.

Statistics of the *A. ipaensis* genome.

	Contig		Scaffold	
	Size(bp)	**Number**	**Size(bp)**	**Number**
N90	1,298	100,545	22,609	845
N80	5,917	55,888	1,923,111	201
N70	10,212	38,251	3,228,270	141
N60	14,727	26,977	4,578,890	101
N50	19,749	18,841	6,188,345	72
Longest	250,973	----	36,812,236	----
Total Size	1,387,433,434	----	1,511,094,544	----
Total (>=100bp)	----	889,664	----	759,441
Total (>=2kb)	----	88,566	----	8,382

Repeat content in the *A. ipaensis* genome.

	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (kb)	**%**	**Length (kb)**	**%**	**Length (kb)**	**%**	**Length (kb)**	**%**
DNA	16,532	1.09	23,638	1.56	49,215	3.26	61,796	4.09
LINE	9,483	0.63	19,397	1.28	19,328	1.28	28,497	1.89
LTR	151,251	10.01	211,249	13.98	590,923	39.11	747,298	49.46
SINE	20	0.00	0	0.00	6,291	0.42	6,311	0.42
Other	18	0.00	0	0.00	0	0.00	18	0.00
Unknow	36	0.00	21,809	0.00	400,260	26.49	400,305	26.49
Total	179,968	11.91	254,284	16.83	1,012,980	67.04	1,088,223	72.02

Gene annotation result of the *A. ipaensis* genome.

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	*Augustus*	51,887	2,041.90	904.04	4.19	215.90	356.99
	GlimmerHMM	54,551	1,705.03	752.11	3.48	215.99	383.90
Homolog	*G. max*	30,181	3,242.84	1,145.71	4.80	238.88	552.45
	L. japonicus	35,808	2,135.83	843.06	3.48	242.04	520.62
	M. truncatula	28,288	3,281.24	1,062.46	4.30	246.98	671.99
	P. vulgaris	28,056	3,277.20	1,170.51	4.90	238.72	539.72
EST		243,045	4,670.61	1,292.54	4.32	299.82	243,045
GLEAN		42,860	2,183.45	966.75	4.10	235.87	392.66
RNAseq		62,237	2,921.22	886.73	4.52	196.38	356.79
Final set		42,250	2,176.82	967.51	4.09	236.48	391.20

Component 2: High-Density Genetic Maps of Gene Space

Thousands of gene markers were identified in wild and cultivated peanuts from international germplasm collections and breeding populations. A- and B- genome gene markers found in diploid progenitor species were useful in pinpointing their counterparts on chromosomes of the culitvated (tetraploid) genome. These markers were used to compile an international reference genetic map to help breeders locate genes for agronomic traits. Research showed that markers associated with gene sequences in wild species could be used to identify the corresponding gene in cultivated species. These added resources will enable accelerated superior peanut variety development by a new breeding method called, Genotyping by Sequencing (GBS).

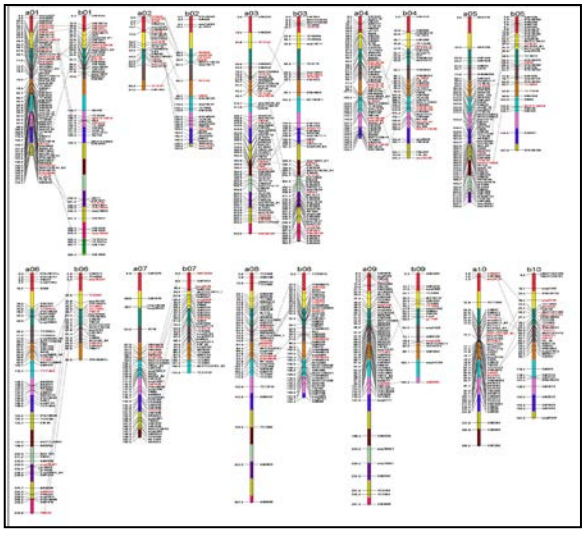
Discovering Gene Markers from Interntional Germplasm Collections

DNA sequence differences revealed 1300 unique SNP (single nucleotide polymorphism) markers among representative accessions of the ICRISAT diversity panel, Chinese mini-core collection and the U.S. mini core collection. These markers helped distinguish genetic variability for seed and pod characteristics. Examples of genetic variation are shown from the U.S. peanut mini-core collection. (A) PI 292950 (A. hypogaea L. subsp. Hypogaea var. hypogaea); (B) PI 288146 (A. hypogaea L. subsp. fastigiata var. vulgaris); (C) PI 196635 (A. hypogaea L. subsp. hypogaea var. hypogaea); (D) PI 493631 (A. hypogaea L. subsp. fastigiata var. fastigiata); (E) PI 478850 (A.hypogaea L. subsp. fastigiata var. fastigiata); (F) PI 337406 (A. hypogaea L. subsp. fastigiata var. fastigiata); (G) PI 288210 (A. hypogaea L. subsp. fastigiata var. vulgaris); (H) PI 372305 (A. hypogaea L. subsp. hypogaea var. hypogaea); (I) PI 502111 (A. hypogaea L. subsp. fastigiata var. peruviana).

See: Charles Y. Chen, Noelle A. Barkley, Ming L. Wang, C. Corley Holbrook, and Phat M. Dang Registration of Purified Accessions for the U.S. Peanut Mini-Core Germplasm Collection. Journal of Plant Registrations doi: 10.3198/jpr2013.01.0003crg

Using Gene Markers to Locate QTL (gene locations) on Genetic Linkage Maps

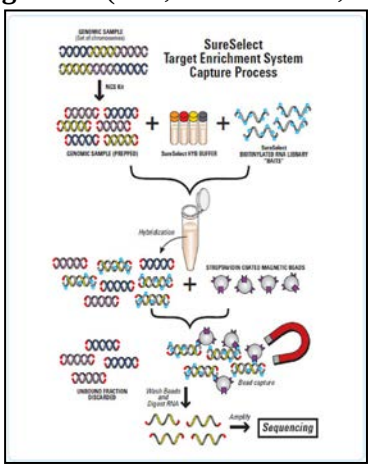
Over 32 genetic maps of various peanut breeding populations have been published since 1993. These maps often were specific to the parents of the population. A meta-analysis was conducted with populations segregating for (afflatoxin) PAC, drought tolerance, LLS, rust, protein & oil concentration, TSWV and yield. Results were compiled in an **International Reference Consensus Genetic Map for cultivated peanut**. This map gives a high-density view of marker locations on A- and B-chromosomes in cultivated peanut, andt also shows connections between chromosomes (linkage groups) of the wild and cultivated species.



See: Gautami B, Fonce´ka D, Pandey MK, Moretzsohn MC, Sujay V, et al. (2012) An International Reference Consensus Genetic Map with 897 Marker Loci Based on 11 Mapping Populations for Tetraploid Groundnut (Arachis hypogaea L.). PLoS ONE 7(7): e41213. doi:10.1371/journal.pone.0041213

Sorting Out Useful Markers from Gene-Rich and Non-Coding Regions of the Genome

DNA markers from peanut resoruces often are associated with regions of the genome that contain few genes. The non-protein coding regions of the genome contain DNA **with repeating sequences associated with long terminal repeat (LTR) retrotransposons that may account for 53% of the cultivated peanut genome** (Gao, Chavarro etal, UGA). The function of transposable elements in a DNA sequence may create mutations or alter genome size. Therefore methods such as the 'Sure Select Target Enrichment System Capture Process' were deployed to isolate the gene-rich (exome) regions to disginguish the most useful set of gene-markers for genes that code functional proteins (enzymes) in wild and cultivated species. See Component 3 for more detail.

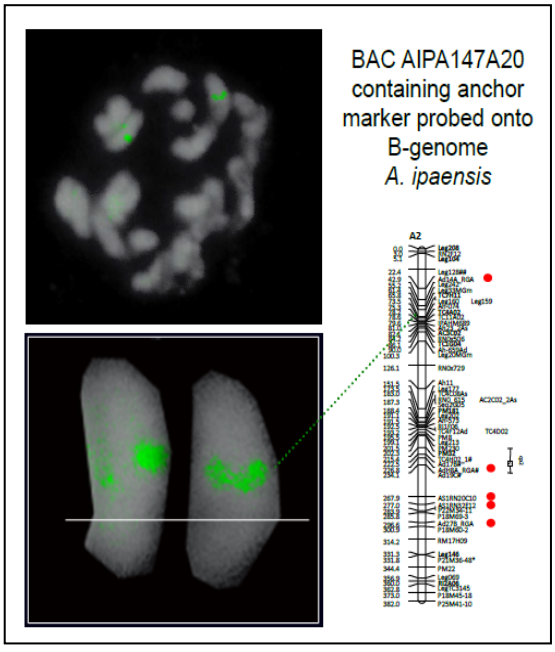


Transposon	Arachis duranensis		Arachis ipaensis	
	Coverage (bp)	Content (%)	Coverage (bp)	Content (%)
LTR(copia+gypsy)	431856291	40.4	582169147	40.26
LTR(TRIM)	2064016	0.19	4630608	0.32
LINEs	78926879	7.39	165352780	11.43
DNA/MULE	6450346	0.60	6679814	0.46
DNA/CACTA	1690540	0.16	5613777	0.39
DNA/Helitron	1835579	0.17	1953051	0.14
DNA/Habinger	311078	0.03	530455	0.04
DNA/hAT	2841876	0.27	3577067	0.25
Total	525976605	49.26	770506699	53.28

Laying a Foundation for Genotype by Sequencing (GBS) Breeding Methods for Peanut

Gao, Chavarro etal at UGA found a high rate of alignment of specific gene markers resident in diploid species with the corresponding gene in cultivated genotypes. This finding demonstrates how information from genetic maps can be traced to a specific chromosome of a given genome, and thusexpands the utility of markers for the wild species to trait associations in cultivated peanut; and expedites deliver of useful markers to breeders for application in GBS methods, thereby improving the efficiency of varital development..

							A. duranensis		A. ipaensis	
SeedLot	Cultivar	Species	Ploidy	Total number of reads	Total HQ reads	Total HQ reads %	Total aligned	Overall Alignment rate	Total aligned	Overall Alignment rate
PI 565460	NC3033	A. hypogaea	tetraploid	1,114,234	1,084,244	97%	1,001,964	92.41%	952,709	87.87%
PI 659502	Bailey	A. hypogaea	tetraploid	845,949	827,364	98%	773,994	93.55%	742,403	89.73%
SPT06-6	SPT06-6	A. hypogaea	tetraploid	1,157,616	1,134,173	98%	1,062,698	93.70%	1,018,136	89.77%
C76-16	C76-16	A. hypogaea	tetraploid	2,328,393	2,272,538	98%	1,642,956	72.30%	1,603,529	70.56%
PI 631176	Olin	A. hypogaea	tetraploid	936,162	912,700	97%	850,777	93.22%	827,791	90.70%
PI 565452	New Mexico Valencia A	A. hypogaea	tetraploid	1,088,763	1,062,763	98%	964,756	90.78%	937,673	88.23%
PI 644011	Tifrunner	A. hypogaea	tetraploid	2,071,568	2,028,389	98%	1,856,204	91.51%	1,833,942	90.41%
PI 576638	SSD 6	A. hypogaea	tetraploid	609,186	596,953	98%	542,793	90.93%	532,505	89.20%
PI 565448	Florunner	A. hypogaea	tetraploid	751,725	739,146	98%	692,683	93.71%	680,825	92.11%
PI 652938	Florida 07	A. hypogaea	tetraploid	953,429	933,469	98%	868,602	93.05%	847,997	90.84%
PI 262133		A. duranensis	diploid	1,307,906	1,279,096	98%	1,232,021	96.32%	1,113,339	87.04%
PI 468321		A. duranensis	diploid	1,867,809	1,821,359	98%	1,711,523	93.97%	1,576,996	86.58%
PI 468322		A. ipaensis	diploid	1,457,057	1,426,396	98%	1,300,857	91.20%	1,392,918	97.65%
Mean (Mb - %)				1.27	1.24	97.78%	1.12	91.28%	1.08	88.51%



Mining SSRs and identification of genes underlying QTL genomic regions for foliar disease resistance

Scientists at ICRISAT in collaboration with USDA-ARS and Catholic University Brasilia used two diploid genome assemblies for SSR mining and annotation of the target QTL genomic regions controlling two important foliar fungal diseases (rust and late leaf spot).

- SSR mining from AA and BB genome: Ten pseudo-molecules each of A. duranensis (A-genome) and A. ipaensis (B-genome) were examined for SSR identification and a total of 86,398 and 116,806 SSR containing sequences were identified, respectively.
- Non-redundant and genome specific SSRs were also identified. A total 15,782 primer pairs from existing (published and unpublished) SSRs were used via an ICRISAT developed Python program to mine non-redundant SSRs in both the genome assemblies. A total of 77,501 and 107,678 primer pairs were found non-redundant for A- and B-genome assemblies, respectively.
- SSR primers (A-genome) were BLAST searched against B-genome primers and vice versa. The A-genome SSR primers showed 37,476 hits against the B-genome whereas B-genome primers showed 54,979 hits with A-genome. These were filtered on the basis of location of the SSR and its repeat motif (allowing for polymorphism) using a Python script to arrive at a final list of SSRs that are common between the genomes. Apart from these common SSR primers, the number of SSR primers that did not map to the other genome were considered to be unique to that genome. Finally, a total of 76,984 A-genome specific SSRs and 106,979 B-genome specific SSRs were identified.

Gene annotation of QTL genomic region controlling foliar fungal diseases: Markers for rust resistance (GM2079, GM2009, IPAHM103, GM1536 and GM2301) and late leaf spot resistance (GM1009, Seq8D09 and GM1573) QTL were exploited to identify significant genes linked to resistance.

- *Rust resistance*: Sequences of markers associated with rust resistance were blast searched against A- and B-genomes.
    - The genomic coordinates of the best hits for these sequences spanned over 2.0 Mb of sequence in A- genome. GENSCAN1.0 predicted 323 CDS/peptides and annotation revealed 130 sequences enriched with at least one gene ontology (GO) id in this 2.0 Mb sequence. Many of these genes are known to have ostensible role against pathogen attack. Ethylene-induced calmodulin-binding transcription activator 4 identified in this region is known to regulate ethylene-induced senescence by directly binding to the ETHYLENE INSENSITIVE3 (EIN3) and NON-RACE-SPECIFIC-DISEASE RESISTANCE1 (NRDR1) promoter region. Cysteine proteinase inhibitors, involved against plant defense, were also found among other proteins. Besides these, kinases, phosphatases and leucine rich repeats (LRRs) were also identified, like CBL-interacting serine/threonine-protein kinase, CBL-interacting serine/threonine-protein kinase, serine/threonine-protein kinase and Serine/threonine-protein phosphatase. In addition, TIR-NBS-LRR type disease resistance protein, also known to be involvement in ATP binding, DNA binding, RNA binding, RNA-directed DNA polymerase activity and defense response, was identified.
    - Analysis for rust resistance in B-genome using marker sequences fetched approximately a 772 Kb sequence on chromosome 3 based on the e-values and alignment length. GENSCAN identified 139 CDS/peptide sequences and 71 could be assigned an uniprot id (including 39 from SwissProt). Important proteins with significant role were identified e.g., pentatricopeptides, ornithine aminotransferase, cullin-4, type I 1,4,5-triphosphate 5 phosphatase, probable 9-cis-epoxycarotenoid dioxygenase, purple acid phosphatase (PAP) etc. An important protein, receptor-like protein kinase 'FERONIA' was identified which is known to suppress the abscisic acid signaling, brassinosteroid mediated and ethylene activated signaling, defense response to fungus, protein kinase activity, serine/threonine kinase activity and protein autophosphorylation.

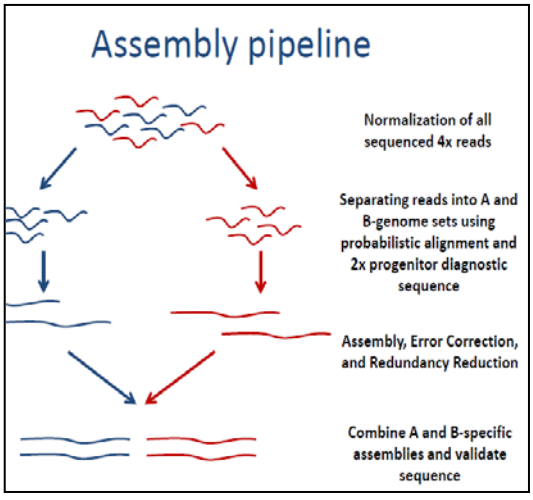
- *Late leaf spot resistance*: The marker sequences associated with late leaf spot resistance were also subjected to blastp searches against A-genome in order to identify the genes responsible for imparting LLS resistance. However, only one marker (Seq8D09) could hit the expected genomic region with best hit as well as had highest confidence from marker-trait validation results. Five important proteins namely, transporter 1, chloroplastic group IIA intron splicing facilitator, retrovirus-related polymerase polyprotein, pumilio homolog 4 and the probable s-acyltransferase 15 were annotated. The s-acyltransferase could be of great interest as its gene ontology reveals cysteine S-palmitoyltransferase activity and zinc ion binding ability as its molecular functions. Palmitoylation aids in regulation of stress responses, disease resistance, hormone signaling and cytoskeletal organization. Also, its ability to bind zinc prompts that this protein could have zinc finger domains and involved in gene transcription by coordinating with one or more zinc ions through cysteine residues. Similar analysis for B-genome identified 15 CDS/peptides on B09. Interestingly, we found the same s-acyltransferase 15 protein with cysteine S-palmitoyltransferase activity and zinc ion binding in this region as was found in A-genome LLS flanking region.

Component 3: Expressed Gene Sequences

This research helps characterize the sequence of genes that are made (expressed) in the cultivar Tifurnner during plant development. Different sets of genes are active or inactive in various plant organs at different growth stages, and under different environmental conditions. Knowledge of where, when and how genes that mediate a given trait is expressed enables construction of gene markers for specific alleles, which are essential tools for the 'Breeders Toolbox'. This information also will help distinguish genes that come from the A- or B-genome in cultivated peanuts, and provides the basis for developing an Atlas that will eventually catalog the sequence and function of all genes in the peanut genome. .

What was sequenced and how useful is the data?

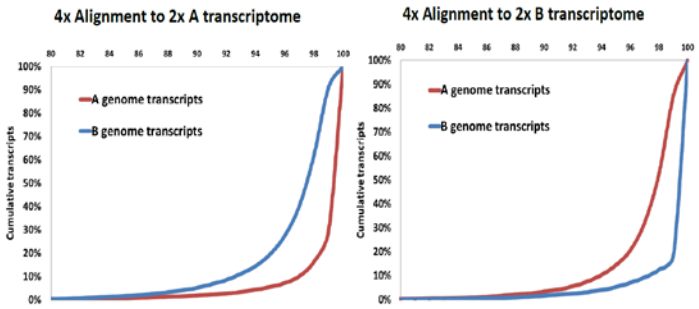
Scientists at the Universtiy of Georgia-Tifton and Athens, USDA ARS at Tifton GA and Stoneville MS, and the Universtiy of California-Davis developed and sequenced **24 RNA libraries from 8 organs at 3 states of growth of the cv Tifrunner**. These libraries represent the transcriptome (all RNA that encode proteins) of culitvated peanut. An assembly pipeline was built that separated SNP markers that were transcribed from the A- and B-genomes. **The transcripts divided equally between the two genomes in all libraries.**



No	Tissue	Stage	sample collection	RNA extraction	Agilent check	Sequence ad
1	Leaf	10 d post-emergence; leaflets partially open SOY:0000252	yes	yes	yes	yes
2	Leaf	Growth stage Boote R1 – first flower; leaflets partially open, from mainstem (n)	yes	yes	yes	yes
3	Leaf	Growth stage Boote R1 – first flower; leaflets partially open, from laterals (n+1)	yes	yes	yes	yes
4a	Vegetative shoot (5 mm max)	Growth stage Boote R1 – first flower, from mainstem (n)	yes	yes	yes	yes
5a	Reproductive shoot (5 mm max)	Growth stage Boote R1 – first flower, from laterals (n+1)	yes	yes	yes	yes
6	Root structures	10 d post-emergence SOY:0000252	yes	yes	yes	yes
7	Nodules	25 d post-emergence SOY:0001301	yes	yes	yes	yes
8a	Flower	Fully open, morning of anthesis; wings, banner, hyparithium, keel; SOY:0001277	yes	yes	yes	yes
8b	Flower	Fully open, morning of anthesis; stigma and ovary	yes	yes	yes	
8c	Flower	Fully open, morning of anthesis; anthers	yes	yes	yes	
9	Gynophore tip – 5 mm (=mostly ovary and zone of cell division)	From elongating peg prior to soil penetration	yes	yes	yes	
10	Gynophore tip – 5 mm (=mostly ovary and zone of cell division)	From elongating peg of approximately same length as #9, but 24 h after soil penetration	yes	yes	yes	
11	Gynophore tip (pod)	At pod swelling (Pattee stage 1)	yes	yes	yes	
12	Gynophore "stalk"	At pod swelling (Pattee stage 1)	yes	yes	yes	
13	Pod	Pericarp very watery, embryo very small and not easily removed (Pattee stage 3/4)	yes	yes	yes	
14	Pericarp	Pericarp soft, not as watery, inner pericarp without cracks (Pattee stage 5)	yes	yes	yes	
15	Seed	embryo flat, white or just turning pink at one end (Pattee stage 5)	yes	yes	yes	init
16	Pericarp	Inner pericarp tissue beginning to show cracks or cottony (Pattee stage 6/7)	yes	yes	yes	
17	Seed	Torpedo shaped; generally pink at embryonic-axis end of kernels (Pattee stage 6)	yes	yes	yes	init
18	Seed	Torpedo to round shaped; embryonic axis end of kernel pink; other end white to light pink (Pattee stage 7)	yes	yes	yes	init
19	Seed	Round, light pink all over (Pattee stage 8)	yes	yes	yes	init
20	Seed	Large, generally dark pink allover; seed coat beginning to dry out (Pattee stage 10)	yes	yes	yes	init

Another pipeline was built to remove incomplete transcripts and transcripts from the 'repeatitive' non-protein coding region of the genome. **3855 transcripts were aligned with repetitive elements in the B-genome and 2618 transcripts were associated to A-genome repetitve sequences.** Those transcripts were filtered out for future use.

Diploid A-genome transcripts from gene-rich (exome) regions aligned perfectly with the tetraploid A-genome; likewise for diploid and tetraplolid B-genome transcripts.



300 SNP markers based on diploid parent polymorphism in transcripts were validated by genomic data (KASP system) and Sanger sequencing. **These SNP identified distinguishable mutations within genes of interest in a diploid species and may be used to identify the same gene in cultivated genoyptes.**

Gene annotation of QTL genomic region controlling foliar fungal diseases: ICRISAT scientists identified significant genes linked to rust and late leaf spot resistance.

a. Candidate Genes for Rust resistance:

A-genome

- Ethylene-induced calmodulin-binding transcription activator 4 regulated ethylene-induced senescence by directly binding to the *ETHYLENE INSENSITIVE3* (*EIN3*) and NON-RACE-SPECIFIC-DISEASE RESISTANCE1 (NRDR1) promoter region.
- Cysteine proteinase inhibitors, involved against plant defense.
- CBL-interacting serine/threonine-protein kinase and Serine/threonine-protein phosphatase.
- TIR-NBS-LRR type disease resistance protein, involved in ATP binding, DNA binding, RNA binding
- RNA-directed DNA polymerase activity and defense response.

B-genome

- Important proteins were identified e.g., pentatricopeptides, ornithine aminotransferase, cullin-4, type I 1,4,5-triphosphate 5 phosphatase, probable 9-cis-epoxycarotenoid dioxygenase, purple acid phosphatase (PAP).
- A receptor-like protein kinase 'FERONIA' was identified which is known to suppress the abscisic acid signaling, brassinosteroid mediated and ethylene activated signaling, defense response to fungus, protein kinase activity, serine/threonine kinase activity and protein autophosphorylation.

b. Late leaf spot resistance:

A-genome

- Five important proteins namely, transporter 1, chloroplastic group IIA intron splicing facilitator, retrovirus-related polymerase polyprotein, pumilio homolog 4 and the probable s-acyltransferase 15 were annotated.

B-genome

- 15 CDS/peptides on B09.
- the same s-acyltransferase 15 protein with cysteine S-palmitoyltransferase activity and zinc ion binding in this region as was found in A-genome LLS flanking region.

Component 4: Evaluation of New Genome Sequencing Technologies

Research findings have shown that more than one DNA sequencing technology will be needed to properly assemble the peanut genome. There are many options that not only ensure high quality results but also help reduce project costs.

What Methods are being Considered to Assemble the Reference Tetraploid Genome?

Although a BAC x BAC approach was not needed for assembly of the diploid A- and B-genomes, the increased complexity of the tetraploid genome structure may require this process, which is labor intensive and expensive. Scientists at the University of California-Davis and BGI evaluated a preliminary test of a BAC x BAC approach for the cultivated peanut genome.

BGI conducted a pooling test to find out if BAC-to-BAC sequencing can deliver assembly statistics with contig N50 greater than 20 kb and scaffold N50 greater than 2000 kb. 100 BACs from the cv Tifrunner genome.(average insert size 110kb) were constructed. Three pooling methods were tested, 1) 100 single BAC pools, 2) 50 two BAC pools and 3) 25 four BAC pools. The pools were index sequenced to at least 100x depth of the pooled BACs in one 500 bp insert size library. QC measures were taken. Each pool was assembled using filtered data with *SOAPdenovo* software. Muti-kmers were used to get the best scaffold N50 quality. The 500 bp insert size library for each pool was sequenced with Hiseq 2000 to generate ~ 7.8Gb raw data. The reads were assigned to each BAC by the index. Mean raw data statistics for each treatment are shown below.

Pooling	Sequenced pool number	Insert size(bp)	Total reads(Mb)	Average pool reads(Mb)	Average BAC reads(Mb)
One	100	500	3,588	35.8	35.8
Two	50	500	1,974	39.4	19.7
Four	25	500	2,258	90.3	22.5

Pooling	Assembled pool number	Average scaffold N50 per pool	Average scaffold N90 per pool	Average total length per pool
One	100	15,251	2,404	119,879
Two	50	14,081	1,947	203,002
Four	25	10,220	1,172	420,195

Based on results of pooling experiments, BGI found that single BAC pools gave best results scaffold N50 = 15kb but assembly was difficult. A BAC x BAC sequencing strategy was proposed for Tifrunner using one pool of two BAC libraries (250 bp, 500 bp) with an average sequencing depth of 50X for each BAC.

BAC sequencing experiments conducted at UC-Davis found contig N50 = 50kb (single BAC) and 13.5kb (4 BAC pool). Assembly was in all cases very sensitive to the number of reads. About 100x coverage provided optimal results for most BACs when assembled individually. This indicated it would not be possible to get exact copy number estimates from each BAC when pooled. .

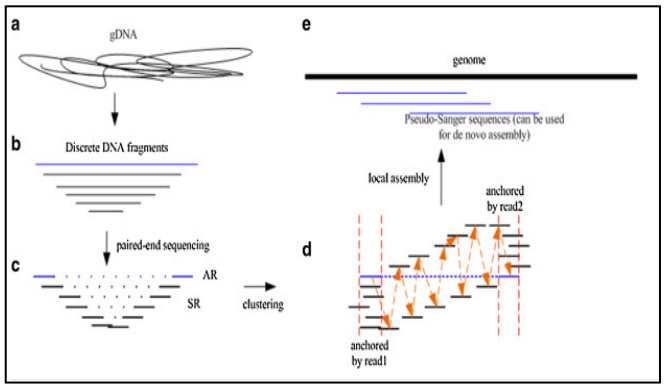
Other approaches to generate better homolog discrimination via longer sequences or other types of long range information for the tetraploid and diploid genomes included consideration of:

- BioNano technology. UC-Davis has this equipment and experience with sequencing wheat genomes. However, current Arachis scaffold and conting sizes may not be long enough.

- PacBio technology. A new system designed for heterozygous diploid genome assemblies (similar to homozygous tetraploid genomes) will soon be available, but has not been tested on plant genomes; therefore high risk.

- Illumina technology. Illumina is expected to release a new kit for 2x250bp on Hi-Seq. HudsonAlpha and USDA-ARS (Scheffler) are the only known labs with Hi-Seq equipment already carrying out 2x250 Hiseq reads with plant DNA.

- AnyTag assembler technology. This assembler has unique ability to fill overlaps on both sides of a read. The latest version is not public yet, but according to the authors, is tested on tetraploids with good results. This method would require 3-4 Tifrunner libraries (400bp, 600bp, 800bp, 1000bp). The libraries generated for this assembly (2x250 PE) should also be compatible with the clustering approach and thus should not entail additional expenses in addition to the library prep costs. http://www.biomedcentral.com/1471-2164/14/711



- Clustering approach. Would separate the 250 bp paired end Hiseq reads into the subgenomes by clustering to 250 bp paired end Hiseq reads of the diploid ancestors. Should generate good data regardless of assembler used. Overall sequencing depth should not exceed 100X for all libraries.

Given concerns raised by results of the BAC x BAC pooling and assembly experiments, the PGC is exploring the efficacy of Hi-Seq and new assembler technologies for improving contiguity of tetraploid and diploid Arachis genome assemblies. Vendors of alternative approaches will be make presentations to the PGC for consideration during the APRES meeting in July, 2014.

Component 5: Phenotyping Genetic Resources

This research is essential for making the peanut genome sequence and genomic tools useful to breeders because it makes the connection between genes, gene markers, genetic maps, and agronomic traits in peanut. The peanut genome initiative is ahead of many other crop genome projects because of the attention that is being given to phenotyping.

What is being done to associate SNP markers with important traits in cultivated peanut?

Attributes for Parents of 16 RIL Populations

Parent	Common or Unique Parent	Market Class	Oleic Acid	TSWV	Early Leaf Spot	Late Leaf Spot	White Mold	Sclerotium	CBR
Tifrunner	Common	Runner	L	R	MR	MR	S	U	U
Florida-07	Common	Runner	H	R	S	S	MR	U	U
N08082olJCT	Unique	Virginia	H	MR	MS	U	U	MR	MR
C76-16	Unique	Runner	L	MR	U	U	U	U	U
NC3033	Unique	Virginia	L	HS	MR	HS	R	U	HR
NM Valencia A	Unique	Valencia	L	S	S	S	HS	HS	U
OLin	Unique	Spanish	H	MS	S	S	U	R	U
SSD6	Unique	Exotic	L	HR	U	U	U	U	U
SPT 06-6	Unique	Exotic	L	U	HR	HR	U	U	U
Florunner	Unique	Runner	L	HS	S	S	S	S	S

RIL Population Phenotyping in Progress

RIL Population	Trait	PIs
Florida-07 x SPT-06-06	•Late leaf spot resis •Early leaf spot resis •TSWV	•P. Ozias-Akins, C. Holbrook, A. Culbreath, S. Jackson •T. Isleib •A. Culbreath
Tifrunner x NC 3033	•Pod fill •Drought tolerance •Late leaf spot resis •White mold resistance •TSWV •CBR resis	•R. Hovav, P. Ozias-Akins, S. Jackson •T. Sinclair •A. Culbreath, P. Ozias-Akins, C. Holbrook •T. Brenneman, B. Tillman, N. Dufault, J. Wang, C. Holbrook •A. Culbreath •T. Brenneman
Florida-07 x NC 3033	•CBR resis	•T. Brenneman
Florida-07 x C76-16	•Preharvest aflatoxin contamination	•P. Ozias-Akins, C. Holbrook, S. Jackson
Tifrunner x C76-16	•Drought tolerance	•C. Chen

Associating SNP markers that define a genotype (at a chromosomal location, within a QTL) with a trait is called 'phenotyping'. This is the area of research that ties all the genomics to practical peanut improvement. Work is led by Corley Holbrook (USDA-ARS, Tifton, GA) with partners at the University of Georgia-Tifton and Athens; Volcani Institute-Israel; University of Florida, Auburn University, ICRISAT, Hyderabad India, Tuskegee University and NPRI. Sixteen inbred mapping populations have been created with parents that maximize genetic diversity for practical breeding objectives. Two modern runner cultivars (Tifrunner and Florida-07) were selected as common parents because runner cultivars account for about 80% of the production in the US. In addition, eight unique 'donor' parents were selected to supply diversity across market classes and are donors of favorable genes for enhancing drought tolerance and resistance to most important diseases of peanut in the U.S. The eight unique parents are N08082olJCT (a Bailey derived high oleic breeding line), C76-16, NC 3033, SPT 06-06, SSD 6 (PI 576638), OLin, New Mexico Valencia A, and Florunner. The 16 populations were advanced in two sets due to the massive requirement for field plot space. A standardized system for evaluating phenotypes has been developed. Seed increase has begun to provide the community with material for extensive phenotyping. In-depth phenotyping is in progress for the five populations shown above. Linking SNP-derived genotypes (mapped QTL) with phenotypic traits segregating in these populations will establish useful markers that can be deployed by breeding programs. Selected progeny of these populations also may serve as valuable parents for the development of improved cultivars.

16 RIL populations were increased in 2014. Four populations were phenotyped (C1799, C1801, C1798 and C1803). Birdsong Inc graciously stored all seed for these populations in -18C freezers. A seed-inventory was updated by Holbrook and posted on www.PeanutBioscience.com/. :

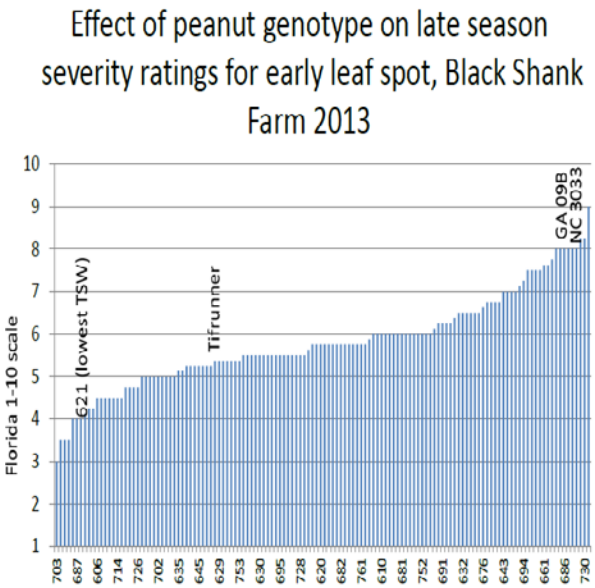
- C1799: a F6:8 RIL population of 286 lines from Tifrunner x NC3033. The population was genotyped with 121 polymorphic SSR primer sets. Transgressive segregation was observed for early leaf spot and TSWV resistance/tolerance. Phenotypic data was taken on these RIL. Significant differences were found for: pod weight, kernel weight, pod volume, pod filling and pod density, maturity, seed color and transpiration (indicator of drought tolerance).

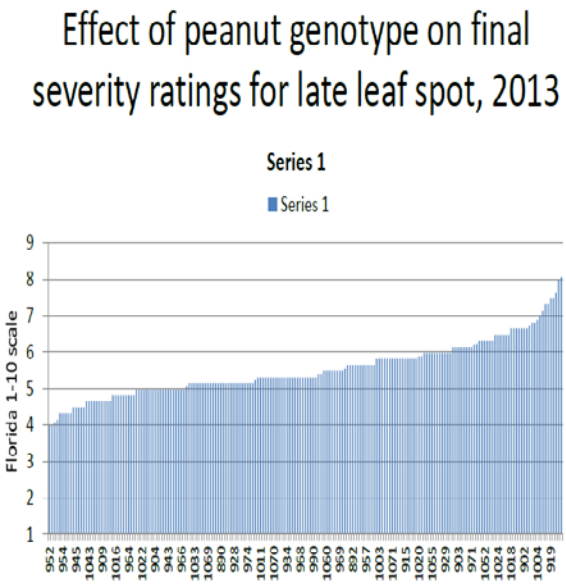
- C1799 RILs also were genotyped for resistance to white mold (S. rolfsii). Lines were genotyped by Ozias- Akins with 380 polymorphic SSR markers newly developed from transcript sequences, and 2000 publically available SSRs with high polymorphism.

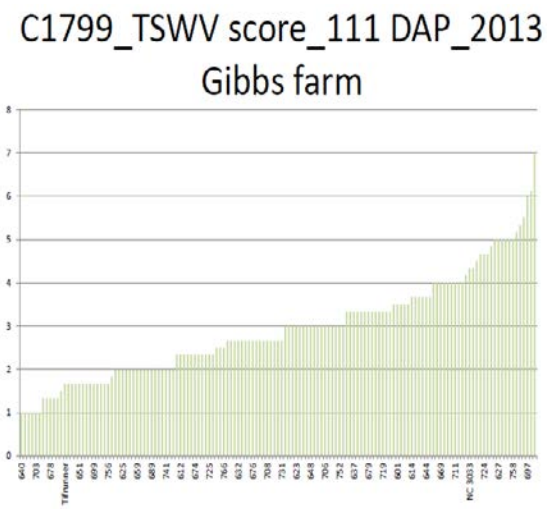
- C1801 (Florida 07 x SPT-06-06) was genotyped with 79 polymorphic SSR. Segregates were evident for early leaf spot resistance. Initial QTL maps were developed for ELS and TSWV.

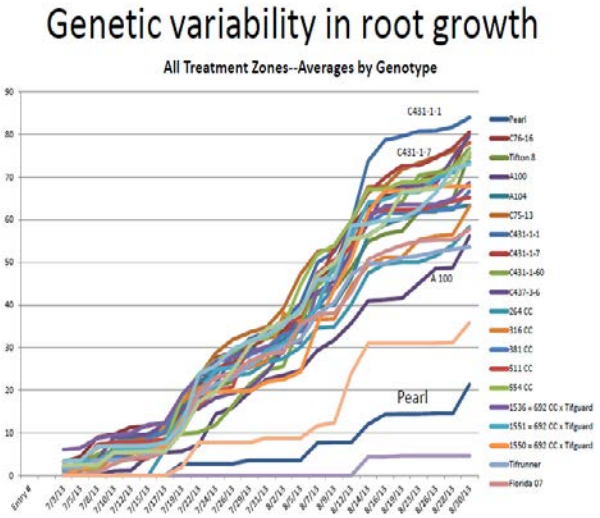
Remaining populations will be phenotyped and mapped in 2014 for CBR, LLS, TSWV, WM & ELS.In addition, USDA-ARS at Dawson GA will provide access to X-ray grading system that is being considered as an alternative standard for peanut grading. Initial tests will be conducted with the 1799 and 1801 populations.

Examples of Genetic Variation in RIL Populations for Important Traits in Peanut









Component 6: Making it All Useful through Bioinformatic Resources

This research creates a secure internet-based home for all data generated by the peanut genome initiative. This website features tools that were instrumental in constructing a high quality draft assembly of each chromosome in the two progenitor diploid species, and software to help make these data useful to breeders.

What is in PeanutBase.org?

USDA-ARS scientists at Ames IA and NCGR in Santa Fe NM are building an electronic 'Peanut Genetic & Genomic Toolbox' at PeanutBase.org/. This website will be modeled after and connected to SoyBase and the Legume Information System. Features of the website will include:

- A convenient way to access datasets such as: maps, transcriptome, SNPs, RNA-seq, Genome assembly, annotations, etc.
- Links to research community, genetic and genomic resources for peanut, legumes and other external sources
- Biological information on peanut and relatives to provide context
- Map, trait, and QTL information with methods for data collection & database loading integrated with other resources
- Ability to query and view QTL positions on genetic maps in various formats .
- Genome browsers served from a 48-core machine which are very responsive, scalable, well integrated
- Integration of multiple maps by placing sequence-based markers onto the genome sequence(s), producing a "virtual/physical" map. These will be tied to QTLs, and to the browsers.
- A collection of peanut phenotype descriptors with relation to other ontologies
- Tools for using haplotype, accession, & phenotype data for breeding
- Gene family & orthology tools, and gene functional information; from a project to catalog and describe the phenotypes of genes with characterized mutations, using ontologies, from Arabidopsis, soybean, maize, rice, Medicago, tomato; map these into all gene families
- Training, outreach, and coordination
- Sequence and key-word search tools

This website will help PGP members leverage information in other external genome databases. Gene function (noted by a change in phenotype due to gene mutations) is conserved across a gene family, so function identified in other species will often be applicable across different species (e.g. in peanut).

How will all of these data and tools be stored and made available to breeders?

The PeanutBase.org website was substantially extended in 2014. Beyond the genetic maps, research links, and collections of genetic traits assembled in 2013, the website now has these resources:

- The complete genome sequences of the two closest wild relatives of cultivated peanut – which, together, essentially comprise the genetic sequence for the 20 chromosomes of cultivated peanut. (There are small but important differences between the genome sequences of the wild relatives and cultivated peanut, which the PGI researchers are focusing on in the coming year).
- The genome sequences are available in several formats for researchers: for download, for searching by sequence, and for browsing using an interactive genome browser.
- The ~40,000 predicted genes from each sequenced genome are available to researchers, along with their predicted gene functions for many genes (based on functions in soybean, common bean, and other model plant species that have "gone before").
- Genetic maps (12 maps in electronic form) curated from the literature.

- Genetic locations of traits of interest ("Quantitative Trait Loci, or QTL), ranging from disease resistance to architecture and abiotic stress characteristics) curated from the literature. There are currently 392 genetically mapped traits available at the website.
- Links to resources in the peanut research community, to genetic and genomic resources for peanut, and to crop relatives of peanut.
- Data templates available for breeders and research groups to contribute and integrate their own data with the other resources at the website.

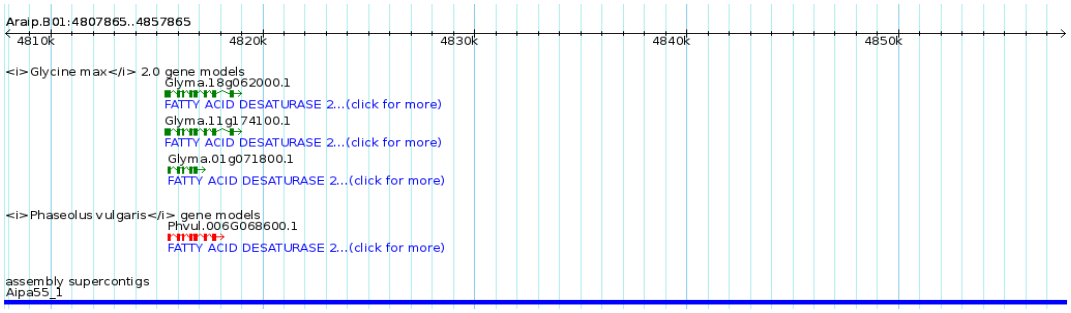
Ongoing work, continuing through 2014 and into 2015, includes description of important varieties with useful traits such as disease resistance and drought tolerance, and a collection of genetic markers for these and other difficult-to-score traits such as nematode resistance and resilience against various stresses. The group is also contributing to the efforts to sequence a cultivated variety of peanut, to re-sequence and analyze a diverse breeding collection, to improve tools to access genetic data, and to interact with breeders and researchers to identify needs and solutions.

PeanutBase is a Portal to Other Crop Genomes

We now know that SNP markers from Arachis diploid species can be used to identify gene loci in cultivated peanut. The same concept may apply for markers from other crops. As shown below, the PeanutBase browser can locate the chromosomal position of a gene in peanut from gene models in other crops (such as soybean and common bean).

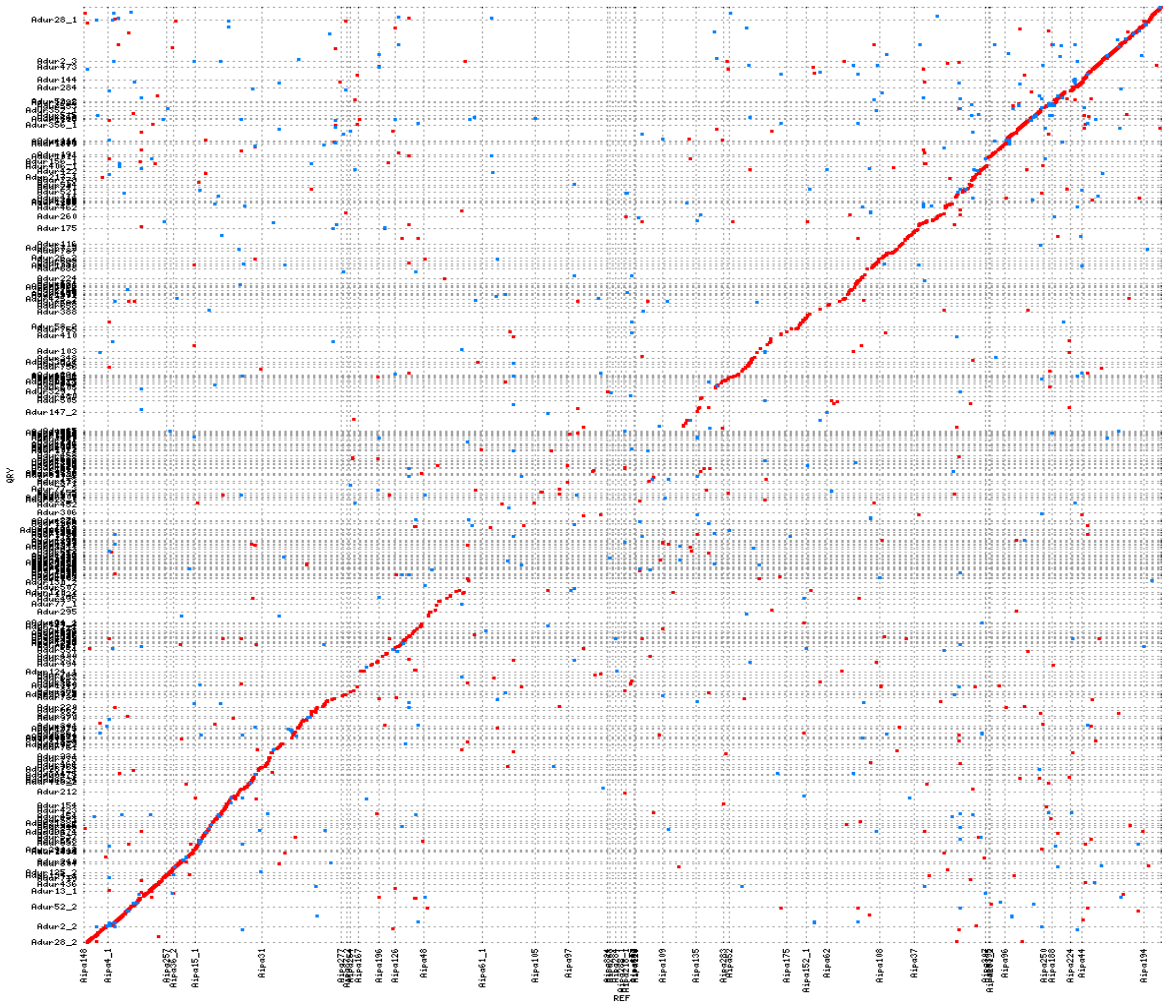
Screen shot of gene positions on a portion of chromosome 2 of the A-genome

Screen shop of gene positions on a portion of chromosome 1 of the B-genome



PeanutBase Tools Enabled Assembly of Wild Specie Chromosomes

This website is not only the repository for the first chromosomal scale draft sequence assemblies for the progenitor diploid species of cultivated peanut, but also has tools that help ensure the highest quality control that gene order in an assembly is correct. The following example shows a comparison of gene order between chromosome- 05 from A- and B-genome diploid species. The non-connected are in the middle of the diagonal line shows the 'non-protein coding region' within chromosome 05..



Appendices

Exhibit 1: What materials were used for DNA sequencing?

Cultivated peanut (Arachis hypogea)

- The cv Tifrunner was selected as the best representative modern variety to provide a 'reference' standard for characterization of genome structure in other cultivated peanut germplasm. Tifrunner exhibits: normal-oleic, TSWV resistance, early leaf spot traits.
- Three other varieties with important attributes compared to Tifrunner
    - GT-C20, a Spanish-type Chinese cultivar (low oleic, susceptible to TSWV and early leaf spot, resistance to aflatoxin contamination)
    - SunOleic 97R (high oleic, susceptible to TSWV, early and late leaf spots)
    - NC94022 (low oleic, high resistance to TSWV, early and late leaf spots)
- Two recombinant inbred line (RIL) populations to help distinguish unique genetic markers that derive from each parent
    - Tifrunner x GT-C20 (T-population with 113 RILs sequenced)
    - SunOleic 97R x NC94022 (S-population with 137 RILs sequenced)
- 192 phenotyped RILs segregating for drought tolerance and foliar diseases (ICRISAT).
- 325 accessions with known genotype and phenotype diversity from the ICRISAT germplasm collection
- 99 accessions from the Chinese mini-core germplasm collection representing genetic diversity in Chinese peanuts
- 112 accessions from the USDA mini-core germplasm collection representing genetic diversity in U.S. peanuts

Wild peanuts (to help assign DNA fragments to A and B genomes in cultivated peanut; and to capture and transfer desirable traits from wild to cultivated peanut)

1. AA-genome progenitors: *A. duranensis, A. stenosperma*
2. BB-genome progenitors: *A. ipaensis, A. magna*
3. AA-genome RIL populations from *A. duranensis x A. stenosperma*
4. BB-genome RIL populations from *A. ipaensis, A. magna*
5. AABB-genome (synthetic) RIL populations from (*A. duranensis x A. stenosperma*) x (*A. ipaensis x A. magna*)
6. AABB-genome (synthetic) x (*A. hypogea*) RIL populations from [(*A. duranensis x A. stenosperma*) x (*A. ipaensis x A. magna*)] x *A hypogea*

Exhibit 2: Sponors who provide financial support for the Peanut Genome Project

U.S. Peanut Sheller Companies:
• American Peanut Shellers Assoc. Birdsong Peanuts
– Damascus Peanut Company
– Golden Peanut Company
– McCleskey Mills
– Snyder's/Lance
– Tifton Peanut Company
– Williston Peanuts
• Southwestern Peanut Shellers – Birdsong Peanuts
– Clint Williams Company
– Golden Peanut Company
– Wilco Peanut Company
• Virginia Carolina Shellers Assoc. Birdsong Peanuts
– Golden Peanut Company
– Peanut Processors
– Severn Peanut Company
• American Peanut Growers Group
• Brooks Peanut Company
• Sessions Company
• Tifton Quality Growers

Food Manufacturing Companies:
• Algood Food Company
• American Blanching
• Arway Confections, Inc.
• Diamond Foods, Inc.
• E.J. Cox
• Hampton Farms
• The Hershey Company
• J.B. Sanfilippo
• Jimbo's Jumbo's
• J.M. Smucker
• Kraft – Planters
• Mars Chocolate
• Old Home Foods
• Pardoe's Perky Peanuts
• Peanut Butter & Company
• The Peanut Shop of Williamsburg
• Producers Peanut Company

US Peanut Producer Organizations:
• National Peanut Board
• Florida Peanut Producers Association
• Texas Peanut Producers Association
• Georgia Peanut Commission

Allied Sector Companies:
• B.A.G.
• Bayer CropScience
• Chips Group
• Concordia, LLC
• Dothan Warehouse
• Early Trucking
• Georgia Federal-State Inspection Service
• Hofler Brokerage
• International Service Group
• JLA USA
• Jack Wynn & Company
• J.R. James Brokerage
• Lewis M. Carter
• Kelly Manufacturing Company
• Lovatt & Rushing
• Mazur & Hockman
• M.C. McNeill & Co. LLC
• National Peanut Brokers Assn.
• National Peanut Buying Points Assn.
• Nolin Steel
• O'Connor & Company
• Olam International Limited
• RCB Nuts
• Reed Marketing, LLC
• Satake USA, Inc.
• SGL International, LLC
• Southern Ag Carriers

International Collaborators
BGI-Americas
Henan Academy of Agricultural Sciences
Chinese Academy of Agricultural Sciences
Shandong Academy of Agricultural Sciences

Scientific and Technical Contributions to the Peanut Genome Project are provided by:

Auburn University
BGI-Americas
Catholic University-Brasilia
Chinese Academy of Agricultural Sciences
EMBRAPA
Generation Challenge-Gates Foundation
Henan Academy of Agricultural Sciences
ICRISAT (India, West & Central Africa)
Indian Council of Agricultural Research (ICAR)
Kazusa DNA Research Institute (Japan)
National Center Genome Resources
Peanut Corporation of Australia
Shandong Academy of Agricultural Sciences
North Carolina State University
Texas A & M University
University of California-Davis
University of Florida
University of Georgia
USDA-Agricultural Research Service
Volcani Center (Israel)

Exhibit 3: Members of the Peanut Genome Consortium

Scott Jackson, UGA (Chairperson)
Peggy Ozias-Akins, UGA (Co-Chair)
Richard Michelmore, UC-Davis (Co-Chair)
Rajeev Varshney , ICRISAT (India)
Howard Valentine, TPF (Administrator)
Howard Shapiro, MARS, Inc
Victor Nwosu, MARS, Inc
Baozhu Guo, USDA/ARS
Corley Holbrook, USDA/ARS
Brian Scheffler, USDA/ARS
Steven Cannon, USDA/ARS
Boshou Liao, CAAS (China)
Andrew Farmer, NCGR
David Bertioli, U Brasilia
Soraya Bertioli, EMBRAPA (Brazil)
Xingyou Zhang, HAAS (China)
Xun Xu, BGI (China)
Xingjun Wang, SAAS (China)
Mark Burow TAMU
Farid Waliyar, ICRISAT (W. Africa)
Graeme Wright, PCA (Australia)
Sachiko Isobe, KDRI (Japan)
Ran Hovav, ARO/VC (Israel)
Tom Stalker, NCSU
Richard Wilson, OBC (Raleigh)
Lutz Froenicke, UC-Davis

Ex Officio

Roy Scott, USDA/ARS/ONP
Maricio Lopez, President EMBRAPA
Jean-Marcel Ribuat, Director, GCP
Luo Fuhe, Vchairman, CPPCC
David Hoisington, UGA
Swapan Datta, DDG, ICAR

Exhibit 4: Terms and Definitions

Abridged from Http://www.panzea.org/infor/faq.html, and http://www.netsci.org/Science/Bioinform/terms.html

Allele: Different forms of a gene which occupy the same position on the chromosome.

Allotetraploid: A cell containing two pairs of different chromosomes (i.e. Peanut)

Autotetraploid: A cell containing two pairs of the same chromosomes (i.e. Soybean)

Amplification: The process of repeatedly making copies of the same piece of DNA.

Annotation: Text fields of information about a biosequence which are added to a sequence databases. Annotation (the elucidation and description of biologically relevant features in the sequence) consists of the description of the following items:

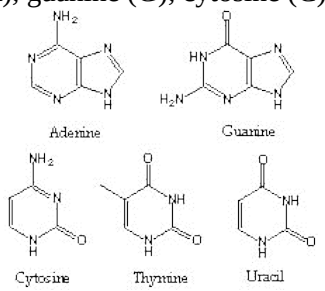
- Function(s) of the protein.
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
- Secondary structure.
- Quaternary structure. For example homodimer, heterotrimer, etc.
- Similarities to other proteins.
- Disease(s) associated with deficiency(s) in the protein.
- Sequence conflicts, variants, etc.

Assembly: The process of placing fragments of DNA that have been sequenced into their correct position within the chromosome.

Association Mapping: As in QTL mapping, the goal of association mapping is to find a statistical association between genetic markers and a quantitative trait. However, in association mapping, the genetic markers usually must lie relatively close to a candidate gene. The goal is to identify the actual genes affecting that trait, rather than just (relatively large) chromosomal segments. QTL mapping is performed in a genetically defined population. Association mapping is performed at the population level within a set of unrelated or distantly-related individuals sampled from a population. Association mapping relies on linkage disequilibrium (LD) between the candidate gene markers and the polymorphism in that gene causes the differences in the phenotypic trait.

Bacterial artificial chromosome (BAC): A long sequencing vector which is created from a bacterial chromosome by splicing a DNA fragment from another species. Once the foreign DNA has been cloned into the host bacteria, many copies of the new chromosome can be made.

Base: One of five molecules which are assembled, along with a ribose and a phosphate, to form nucleotides (Figure 1). Adenine (A), guanine (G), cytosine (C), and thymine (T) are found in DNA while RNA is made from adenine (A), guanine (G), cytosine (C), and uracil (U).



Base pair (BP): The complementary bases on opposite strands of DNA which are held together by hydrogen bonding. The atomic structure of these bases preselect the pairing of adenine with thymine and the pairing of guanine with cytosine (or uracil in RNA).

Bioinformatics: An absolute definition of bioinformatics has not been agreed upon. The first level, however, can be defined as the design and application of methods for the collection, organization, indexing, storage, and analysis of biological sequences (both nucleic acids [DNA and RNA] and proteins). The next stage of bioinformatics is the derivation of knowledge concerning the pathways, functions, and interactions of these genes (functional genomics) and proteins (proteomics). Bioinformatics is also referred to as computational biology.

Candidate Genes: The distinction between "random" and "candidate" genes is of great importance. By random genes we refer to genes without any known function of the proteins (or RNAs) that they encode. They may be selected from a random set of expressed DNA sequences (DNA sequences that are copied, or transcribed, into RNA) at a time in cell development. Candidate genes refer to genes of known or suspected function or traits of interest.

Cell: The smallest functional structural unit of living matter. Cells are classed as either procaryotic and eucaryotic.

CentiMorgan (cM): The unit of measurement for distance and recombinate frequency on a genetic map. Formally, the length (number of bases) that have a 1% probability of participating in mixing of genes. For humans, the average length of a cM is one million base pairs (or 1 megabase, Mb).

cDNA (complementary DNA): An artificial piece of **DNA** that is synthesized from an mRNA (messenger RNA) template and is created using reverse transcriptase. The single stranded form of cDNA is frequently used as a probe in the preparation of a physical map of a genome. cDNA is preferred for sequence analysis because the introns found in DNA are removed in translation from DNA ----> mRNA ----> cDNA.

Chromosome: A collection of DNA and protein which organizes the human genome. Each human cell contains 23 sets of chromosomes; 22 pairs of autosomes (non sex determining chromosomes) and one pair of sex determining chromosomes. The human genome within the 23 sets of chromosomes is made of approximately 30,000 genes which are built from over 3 billion base pairs. While eukaryotic chromosomes are complex sets of proteins and DNA, prokaryotic chromosomal DNA is circular with the entire genome on a single chromosome.

Cloning: The technique used to produce copies of a piece of DNA. A DNA fragment that contains a gene of interest is inserted into the genome of a virus or plasmid which is then allowed to replicate.

Cloning vector: A piece of DNA from any foreign body which is grafted into a host DNA strand that can then self replicate. Vectors are used to introduce foreign DNA into host cells for the purpose of manufacturing large quantities of the new DNA or the protein that the DNA expresses.

Coding region: The portion of a genome that is translated to RNA which in turn codes protein (also see exon).

Codon: The set of three nucleotides along a strand of mRNA that determine (or code) the amino acid placement during protein synthesis. The number of possible arrangements of these three nucleotides (or triplet codes) available for protein synthesis is $(4 \text{ bases})^3 = 64$. Thus, each amino acid can be coded by up to 6 different triplet codes. Three triplet codes (UAA, UAG, UGA) specify the end of the protein. In the example below, three codons are shown.

--- UCA CGU CAU ---
Ser ------ Arg ------- His

Complementarity: The sequence-specific or shape-specific recognition that occurs when two or more molecules bind together. DNA forms double stranded helixes because the complementary orientation of the bases in each strand facilitate the formation of the hydrogen bonds which hold the strands together.

Computational biology: See bioinformatics

Consensus sequence: The most commonly occurring amino acid or nucleotide at each position of an aligned series of proteins or polynucleotides.

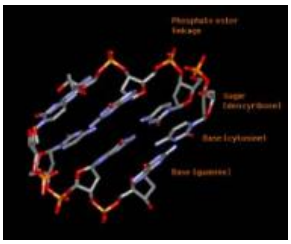
Consensus map: The location of all consensus sequences in a series of multiply aligned proteins or polynucleotides.

Conserved sequence: A sequence within DNA or protein that is consistent across species or has remained unchanged within the species over its evolutionary period.

Contig maps: The representation of the structure of contiguous regions of the genome (contigs) by specifying overlap relationships among a set of clones.

Contigs: A series of cloning vectors which are ordered in such a way as to have each sequence overlap that of its neighbors. The result is that the assembly of the series provides a contiguous part of a genome.

Diploid: A cell containing two sets of chromosomes.



DNA (deoxyribonucleic acid): A double stranded molecule made of a linear assembly of nucleotides. DNA holds the genetic code for an organism in the arrangement of the bases. The double strand of DNA results from the hydrogen bonds formed between bases when two polynucleotide chains, identical, but running in opposite directions, associate.

DNA polymerase: The enzyme which assembles DNA into a double helix by adding complementary bases to a single strand of DNA. Linkages are formed by adding nucleotides at the 5' hydroxyl group to the phosphate group located on the 3' hydroxyl.

EMBL: The European Molecular Biology Laboratory (http://www.embl-heidelberg.de) which is located in Heidelberg Germany.

EMBL Nucleotide Sequence Database: Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications. The database is produced in collaboration with GenBank and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis.

Endonuclease: An enzyme that cleaves at internal locations within a nucleotide sequence. The enzyme's site of action is generally a sequence of 8 bases. For *E. coli*, treatment with a restriction endonuclease will lead to around 70 fragments. Cleavage of human DNA leads to around 50,000 fragments.

Enzyme: A protein which catalyzes (or speeds the rate of reaction for) biochemical processes, but which does not alter the nature or direction of the reaction.

EST (Expressed Sequence Tag): A partial sequence of a cDNA clone that can be used to identify sites in a gene.

Eukaryote: An organism whose genomic DNA is organized as multiple chromosomes within a separate organelle -- the cell nucleus.

Exon: The region of DNA which encodes proteins. These regions are usually found scattered throughout a given strand of DNA. During transcription of DNA to RNA, the separate exons are joined to form a continuous coding region.

Exonuclease: An enzyme which cleaves nucleotides sequentially starting at the free end of the linear chain of DNA.

FASTA: An alignment program for protein sequences created by Pearson and Lipman in 1988. The program is one of the many heuristic algorithms proposed to speed up sequence comparison. The basic idea is to add a fast prescreen step to locate the highly matching segments between two sequences, and then extend these matching segments to local alignments using more rigorous algorithms such as Smith-Waterman.

Fingerprinting: The process of identifying overlapping regions at the ends of DNA fragments.

FISH: Fluorescence in situ hybridization. A method used to pinpoint the location of a DNA sequence on a chromosome.

Frameshift: Genetic mutation which shifts the reading frame used to translate mRNA (see reading frame).

Functional genomics: The development and application of experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics.

Gene: A section of DNA at a specific position on a particular chromosome that specifies the amino acid sequence for a protein.

Gene expression profiling: Determining specifically which genes are “switched on,” with precise definition of the phenotypic trait.

Gene mapping: Determining the relative physical locations of genes on a chromosome. Useful for plant and animal breeding.

GenBank: The NIH genetic sequence database. An annotated collection of all publicly available DNA sequences which is located at http://www.ncbi.nlm.nih.gov. There are approximately 2,162,000,000 bases in 3,044,000 sequence records as of December 1998. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Gene expression: The conversion of the information encoded in a gene to messenger RNA which is in turn converted to protein.

Genetic map (Linkage Map): The linear order of genes on a chromosome of a species. Genetic maps are created by observing the recombination of tagged genetic segments (STSs) during meiosis. The map shows the position of known genes and markers relative to each other, but does not show the specific physical points on the chromosomes.

Genetic mutation: An inheritable alteration in DNA or RNA which results in a change in the structure, sequence, or function of a gene.

Genetic polymorphism: The occurrence of one or more different alleles at the same locus in a one percent or greater of a specific population.

Genome: The total genetic material of a given organism.

Genomics: The mapping, sequencing, and analysis of an organism's genome.

Genomic library: A collection of biomolecules made from DNA fragments of a genome that represent the genetic information of an organism that can be propagated and then systematically screened for particular properties. The DNA may be derived from the genomic DNA of an organism or from DNA copies made from messenger RNA molecules. A computer-based collection of genetic information from these biomolecules can be a virtual genomic library.

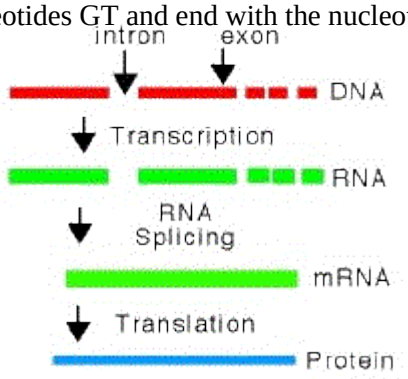
Genotyping: The use of markers to organize the genetic information found in individual DNA samples and to measure the variation between such samples.

Haploid: A cell containing only one set of chromosomes.

Hexaploid: A cell containing three sets of the same chromosomes (i.e. Wheat)

Hybridization: The formation of a double stranded DNA, RNA, or DNA/RNA from two complementary oligonucleotide strands.

Intron: The portion of a DNA sequence which interrupts the protein coding sequences of the gene. Most introns begin with the nucleotides GT and end with the nucleotides AG.



***In vitro*:** Outside a living organism, usually in a test tube.

***In vivo*:** Inside a living organism.

Kilobase (kb): A length of DNA equal to 1,000 nucleotides.

Linkage analysis: The process used to study genotype variations between affected and healthy individuals wherein specific regions of the genome that may be inherited with, or "linked" to, disease are determined.

Linkage Disequilibrium (LD): In population genetics, LD is the association of alleles at two or more loci on same or different chromosome that is greater than random association. Populations where combinations of alleles or genotypes can be found in the expected proportions are said to be in linkage equilibrium.

Linkage map: A map which displays the relative positions of genetic loci on a chromosome.

Loci: The location of a gene or other marker on the surface of a chromosome. The use of locus is sometimes restricted to mean regions of DNA that are expressed.

Mapping: The process of determining the positions of genes and the distances between them on a chromosome. This is accomplished by identifying unique genome markers (ESTs, STSs, etc.) and localizing these to specific sites on the chromosome. There are three types of DNA maps: physical maps, genetic maps, and cytogenetic maps. The types of markers identified differentiate the map produced.

Marker: A physical location on a chromosome which can be reliably monitored during replication and inheritance. Markers on the Human Transcript Map are all STSs.

Microarray: DNA which has been anchored to a chip as an array of microscopic dots, each one of which represents a gene. Messenger RNA which encodes for known proteins is added and will hybridize with its complementary DNA on the chip. The result will be a fluorescent signal indicating that the specific gene has been activated.

Microsatellite: a specific sequence of DNA bases or nucleotides which contains mono, di, tri, or tetra tandem repeats. For example

GGGGGGGG is a (G)8
ACACACAC is referred to as a (AC)4
ATCATCACTACTACT would be referred to as (ATC)5
ATCTATCT would be referred to as (ATCT)2

Microsatallites also are called simple sequence repeats (SSR), short tandem repeats (STR), or variable number tandem repeats (VNTR).

Motifs: A pattern of DNA sequence that is similar for genes of similar function. Also a pattern for protein primary structure (sequence motifs) and tertiary structure that is the same across proteins of similar families.

mRNA (messenger RNA): RNA that is used as the template for protein synthesis. The first codon in a messenger RNA sequence is almost always AUG

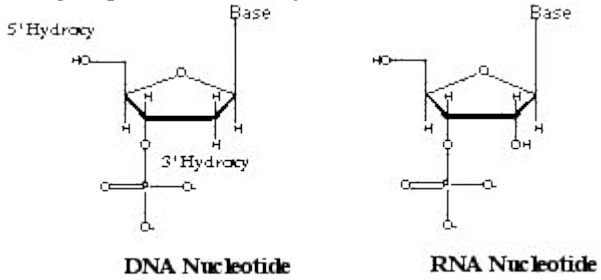
NCBI: The National Center for Biotechnology Information (http://www.ncgi.nlm.nih.gov), a division of the NIH, is the home of the BLAST and Entrez servers.

NCGR: The National Center for Genome Resources (http://www.ncgr.org).

NHGRI: The National Human Genome Research Institute of the NIH (http://www.nhgri.nih.gov)

Northern Blot: An electrophoresis-based technique which is used to find mRNA sequences that are complementary to a piece of DNA called a probe.

Nucleotide (nt): A molecule which contains three components: a sugar (deoxyribose in DNA, ribose in RNA), a phosphate group, and a heterocyclic base.



DNA Nucleotide RNA Nucleotide

Oligos (Oligonucleotides): A chain of nucleotides.

Pairwise alignment: In the first step, two sequences are padded by gaps so that they are the same length and so that they display the maximum similarity on a residue to residue basis. An optimal Pairwise Alignment is an alignment which has the maximum amount of similarity with the minimum number of residue 'substitutions'.

PCR (polymerase chain reaction; in vitro DNA amplification): The laboratory technique for duplicating (or replicating) DNA using the bacterium Thermus aquaticus, a heat stable bacterium from the hot springs of Yellowstone. As with the polymerase reaction that occurs in cells, there are three stages of a PCR process: separation of the DNA double helix, addition of the primer to the section of the DNA strand which is to be copies, and synthesis of the new DNA. Since PCR is run in

a single reaction vessel, the reactor contains all of the components necessary for replication: the target DNA, nucleotides, the primer, and the bacterial DNA polymerase. PCR is initiated by heating the reaction vessel to 90° which causes the DNA chains to separate. The temperature is lowered to 55° to allow the primers to bind to the section of the DNA that they were designed to recognize. Replication is then initiated by heating the vessel to 75°. The process is repeated until the quantity of new DNA desired in obtained. Thirty cycles of PCR can produce over 1 million copies of a target DNA.

Physical map: The physical locations (and order) on chromosomes of identifiable areas of DNA sequences such as restriction sites, genes, coding regions, etc. Physical maps are used when searching for disease genes by positional cloning strategies and for DNA sequencing.

Polymerase: The process of copying DNA in each chromosome during cell division. In the first step the two DNA chains of the double helix unwind and separate into separate strands. Each strand then serves as a template for the DNA polymerase to make a copy of each strand starting at the 3' end of the chain.

Polymorphic marker: A length of DNA that displays population-based variability so that its inheritance can be followed.

Polymorphism: Individual differences in DNA. Single nucleotide polymorphism (the difference of one nucleotide in a DNA strand) is currently of interest to a number of companies.

Quantitative trait locus (QTL): A locus, or location, on a chromosome for genes that govern a measurable trait with continuous variation, such as height, weight, or color intensity. The presence of a QTL is inferred from genetic mapping, where the total variation is partitioned into components linked to a number of discrete chromosome regions.

QTL mapping: QTLs are detected through QTL mapping populations produced from crossing two inbred lines. The first offspring generation (the F1) is uniformly heterozygous. However, in the second generation (the F2) the parental alleles segregate and most chromosomes recombine. Genes and genetic markers that are close together on a chromosome will tend to co-segregate in the F2 (the same allele combinations that occurred in one of the parents will tend to occur together in the offspring). The closer together are two markers or genes on a chromosome, the less likely the parental alleles at the two loci will be split up in the F2 as a result of recombination. This will lead to a statistical association between a gene segregating for alleles that have a measurable difference in their effect on a quantitative trait and segregating alleles at closely linked marker loci. QTLs can thus be localized to specific chromosomal segments if the trait is measured in all the F2 offspring and if all of these offspring are genotyped at hundreds of genetic markers covering the whole genome.

Reading frame (also open reading frame): The stretch of triplet sequence of DNA that encodes a protein. The reading frame is designated by the initiation or start codon and is terminated by a stop codon. As an example, the sequence CAGAUGAGGUCAGGCAUA potentially can be translated as follows:

Position 1 CAGAUGAGGUCAGGCAUA
gln met arg ser Gly ile

Position 2 C AGAUGAGGUCAGGCAUA
arg trp gly Gln ala

Position 3 CA GAUGAGGUCAGGCAUA
asp glu val Arg his

DNA (through RNA) uses a triplet code to specify the amino acid for a given protein. As can be seen above, a given strand of DNA has three possible starting points (position [or reading frame]

one, two, or three). Since both strands of DNA can be translated into RNA and then into protein, a sequence of double helical DNA can specify six different reading frames.

Recombinant Inbred Lines (RIL): RILs are the highly inbred progeny of a segregating population or QTL mapping resource. Two parental inbred lines are crossed, the F1 are intermated (or selfed) to form an F2 generation. Numerous individuals from the segregating F2 generation then serve as the founders of RILs. Each subsequent generation of a given RIL is formed by selfing in the previous generation and with single seed descent. In this manner each RIL, after several generations, will contain two identical copies of each chromosome, with most of them being recombinant.

Resolution: The amount of information (or molecular detail) that is available on a physical map.

Scaffold: A series of contigs that are in the correct order, but are not connected in one continuous length.

Sequencing: Determining the order of nucleotides in a gene or the order of amino acids in a protein.

Sequence tagged sites (STS): The unique occurrence of a short, specific length of DNA within a genome whose location and sequence are known and that can be detected by a specific PCR. An STS is used to orient and identify mapping data for the construction of physical genome maps.

Shotgun method: A method that uses enzymes to cut DNA into hundreds (or thousands) of random bits which are then reassembled by computer so it looks like the original genome. The Human Genome Project shotgun approach is applied to cloned DNA fragments that already have been mapped so that it is known exactly where they are located on the genome, making assembly easier and much less prone to error.

Single nucleotide polymorphism (SNP): The most common type of DNA sequence variation. An SNP is a change in a single base pair at a particular position along the DNA strand. When an SNP occurs, the gene's function may change, as seen in the development of bacterial resistance to antibiotics or of cancer in humans.

Transcriptome: The complete collection of RNA molecules transcribed (or processed) from the DNA of a cell.

Transcription: The process of copying a strand of DNA to yield a complementary strand of RNA

Translation: The process of sequentially converting the codons on mRNA into amino acids which are then linked to form a protein.

Western Blot: An electrophoresis-based technique used to find proteins based on their ability to bind to specific antibodies.