



# **Peanut Genome Initiative**

## **Peanut Genome Project**

### **2014-2015 Research Accomplishment Report to the U.S. Peanut Industry**

**July 31, 2015**

**Peanut Genome Project  
Research Accomplishment Report to the U.S. Peanut Industry  
July, 2015**

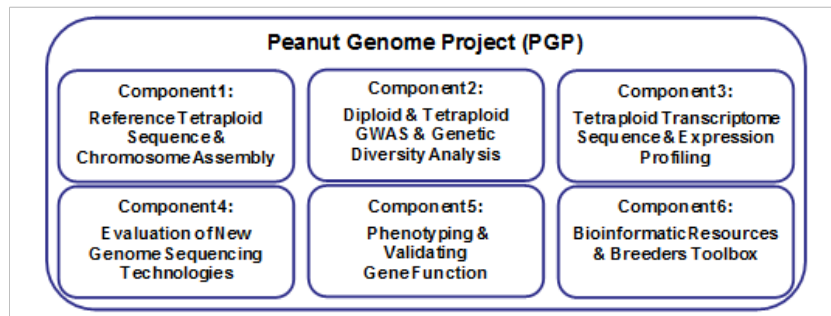
*Table of Contents*

<b>Executive Summary</b>	<b>3</b>
<b>Introduction</b>	<b>8</b>
<b>Component 1: Whole Genome Sequencing</b>	<b>9</b>
<b>Component 2: High-density Genetic Maps &amp; Gene Markers</b>	<b>12</b>
<b>Component 3: Expressed Gene Sequences</b>	<b>14</b>
<b>Component 4: Evaluation of New Genome Sequencing Technologies</b>	<b>16</b>
<b>Component 5: Phenotyping Genetic Resources</b>	<b>19</b>
<b>Component 6: Making it all Useful Through PeanutBase.org</b>	<b>21</b>
<b>Appendices:</b>	
<b>Exhibit 1: What materials were used for DNA sequencing?</b>	<b>24</b>
<b>Exhibit 2: Who has invested in this project?</b>	<b>25</b>
<b>Exhibit 3: Members of the Peanut Genome Consortium</b>	<b>26</b>
<b>Exhibit 4: Terms and Definitions</b>	<b>27</b>

## 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.



**How Key Research Accomplishments in 2014-2015 Tie Together & Help Define the Future Direction of Peanut Breeding.** The six Components of the Peanut Genome Project generate research findings that enable three avenues of investigation: 1) *Generating Detailed Maps* of the peanut genome, 2) development of *Tools for Marker Assisted Selection*, and 3) *Application of Markers and Maps in Breeding* programs. The following shows how the outstanding 2014/15 research accomplishments summarized in this Executive Summary and presented in more detail in the full report fit in the overall performance of the Peanut Genome Project.

**Generating Detailed Genome Maps.** Because cultivated peanut is a very young crop, the two wild parent genomes present in cultivated peanut are very similar to each other since there has not been enough evolutionary time to accumulate genetic differences that help distinguish the two genomes. Maps of the cultivated peanut genome are needed to show the location of all genes and to discover useful DNA markers for those genes. However, assembling a more detailed map of the cultivated peanut genome depends on ability to distinguish the pieces of DNA that came from each of the two wild parent genomes.

- About 70% of the DNA in each wild parent genome is very similar. A new method was used to sort a mixed pool of DNA pieces to their original genome with 98% accuracy. This approach in conjunction with specialized software that determine the order, length and orientation of fragments helps ensure accurate assembly of high definition maps of the genomes of the two wild parents and cultivated peanut (**See Component 1**).
- As parts of the genome map are assembled, the connected sequences may be visualized as a series of small 'Gene Islands'. A new technology called 'Moleculo' was used to bridge between the small islands, thus creating larger islands which finally will be joined into a continent, a complete cultivated peanut genome map (**See Component 4**).

## 2014-2015 Peanut Genomic Research Accomplishments

**Tools for Marker Assisted Selection.** As more DNA markers are discovered in peanut, a better way is needed to screen hundreds of lines in a breeding population simultaneously with thousands of markers.

- A device called a 'chip' was created that holds 60,000 DNA fragments each in individual wells. DNA fragments from a breeding line that stick in each well can be traced to a point on the cultivated peanut genome map. Phenotyping shows positions on that map are important in achieving the breeding objectives. (See Component 2).
- For the first time in plant genomics history, our scientists demonstrated an accurate and innovative way to track the part of a gene sequence that helps locate the best markers for the gene. (See Component 3).

**Application of Markers & Maps in Breeding.** Marker Assisted Selection has been shown to reduce the time needed to add a new trait to a current cultivated variety. Good markers have been found for resistance to late leaf spot (LLS), early leaf spot (ELS), tomato spotted wilt virus (TSWV), root knot nematode (RKN), and high oleic acid. More useful markers are being developed through genome mapping for traits such as cylindrocladium black rot (CBR), white mold (WM), peanut rust and drought tolerance. 'Chip' technology facilitates new breeding strategies for stacking all these traits in improved varieties for each market type and geographic production area. Many breeders are beginning to use these markers in their breeding programs. Their work will create new varieties that help reduce the cost of production, enhance peanut quality and ensure an adequate/safe supply of peanut products. Continued work on markers and maps will help expand breeding objectives to address critical needs of producers, shellers, manufacturers and consumers as presented during APRES-2015 and as will be documented in the next Strategic Plan for Peanut Genome Research during 2016 to 2020.

- Markers identified the location of three different major genes each for resistance to TSWV, ELS, and LLS resistance in the cultivated peanut genome. In addition, it was found that genes for LLS resistance may be linked to peanut rust resistance (See Component 5).
- The training exercises and genomic resources in the 'Breeders Toolbox' in the website 'PeanutBase' were accessed by 5,797 users since July 2014 (See Component 6).

### Publications.

- The first major manuscript from the Peanut Genome Project outlining the genome of the cultivated peanut parents has been submitted to the journal, Nature Genetics for publication. This paper will be the cornerstone upon which many future publications will be based, and will strengthen the competitive position of peanut researchers for future funding from outside sources.
- The book 'Peanuts: Genetics, Processing & Utilization' was accepted for publication in January 2016 by AOCs Press/Elsevier Press. This work presents the first comprehensive view of the peanut value-chain in nearly three decades, and will receive international attention.

## 2014-2015 Peanut Genomic Research Accomplishments

**Key Research Accomplishments for 2014-2015.** The following summarizes most important research accomplishments of each component of the Peanut Genome Project in layman's terms. The remainder of the report (beyond the Executive Summary) provides more detailed descriptions of all 2014/15 research contributions toward strategic goals of the six components.

**Component 1: Whole Genome Sequencing.** Assemblies of the two genomes of the respective parents of cultivated peanut serve as a guide for assembling the genome of the cultivated peanut. However, the current assembly of the cultivated peanut genome resembles small islands without connecting bridges. We now have a means to connect the small islands to create large islands until finally we have the whole cultivated peanut genome assembled or one large island.

KEY ACCOMPLISHMENT –

- State-of-the-art DNA sequencing technology was used by Hudson Alpha to fine tune the current genome assemblies of the cultivated peanut and its parents (wild species *A. duranensis* and *A. ipaensis*). Specialized software was used to determine the order, length and orientation of fragments of the genomes of the two wild parents and the current cultivated peanut with 98% accuracy.

Dr. Scott Jackson, University of Georgia, and chair of the project technical team remarked, "Dicing up and then reassembling the peanut genome correctly is as difficult as climbing a sheer rock face with few handholds. However, we are making our way to the top. When we get there, we will have a clear view of where to find genes with tools that will accelerate development of superior peanut varieties."

**Component 2: High-Density Genetic Maps and Gene Markers.** The key genes of individual lines in a breeding population may be distinguished by thousands of DNA markers know as SNP's (Single Nucleotide Polymorphisms). New methods were tested to find validated DNA markers or SNPs that pointed to locations in the parental genomes that contained key genes involved in crop productivity, protection or improved quality. Validated DNA markers have been positioned on an international genetic roadmap to help breeders locate genes for agronomic traits. This work has enabled a new and revolutionary breeding strategy that should accelerate the development of superior peanut varieties in a timely manner.

KEY ACCOMPLISHMENT –

- New genomic approaches proved useful in processing over 55 million DNA markers or SNPs found in the DNA from 6 peanut lines representing all market types. The number was culled to 60,000 validated SNPs that provide high definition coverage of each peanut chromosome. The genome origin of 96% of those SNPs was determined. These DNA markers or SNP's were placed on a "Gene Chip" or "SNP Chip" that enables breeders to select several different traits simultaneous, instead of the current limitation of one-at-a-time. Each of the 60,000 DNA fragments is placed in an individual well on the "chips". Useful gene markers are found when a DNA fragment for a specific trait in a breeding line sticks to a chip-fragment in a 'well'. Special software associates the 'matches' with a position on a chromosome, and phenotypic data associates the position with a trait. This tells the breeder if the breeding line contains desired genes, and helps eliminate lines without having to grow them out in multi-year evaluation trials.

## 2014-2015 Peanut Genomic Research Accomplishments

Team member Dr. Rajeev Varshney, Director of Research at ICRISAT, says, “The Affymetrix Company has pioneered the development of “SNP chip” technology. A ‘Peanut Gene Chip’ will expedite characterization of genetic diversity in peanut germplasm and identification of the best lines in breeding populations. This resource will enable accurate and cost effective selection of new elite varieties”.

**Component 3: Expressed Gene Sequences.** Genes may be turned on or off in response to an environmental stimulus such as drought stress or attack by a pathogen. A great deal was learned about how trait expression is regulated in different peanut organs (leaves, roots, flowers, pods, and seeds) at different stages of plant development of cultivated peanuts. This knowledge not only helps identify gene sequences that control a given trait, but also helps develop specific markers for the most important genes that influence expression of a plant’s physical or chemical traits.

### KEY ACCOMPLISHMENT

- Another way to develop useful DNA markers or SNP’s is to find the component inside a gene that is responsible for dominant or recessive traits like resistance or susceptibility. For the first time in plant genomics research, our scientists demonstrated which portion of a gene sequence came from parents that were either resistant or susceptible to root knot nematode. This ability helps pinpoint where to look for 'perfect' markers that directly identify the gene that gives resistance. 'Perfect markers' for traits like oleic acid already are in use by breeders.

Dr. Peggy Ozias-Akins, University of Georgia, who is one of the leaders of this component, says, “Gene expression is the basis for all the differences we see in peanuts, including disease resistance and quality attributes. Knowing how genes respond to stressful as well as favorable conditions helps identify genes that can be nudged to maximize peanut productivity and to breed a better peanut”.

**Component 4: Evaluation of New Genome Sequencing Technologies.** Our scientists routinely test the most up to date technologies to ensure the best sequencing and assembly methods are used in the Peanut Genome Project. Many vendors of those new technologies work with our research team members to demonstrate and validate their products. Always having access to the best new technologies will help simplify and improve accuracy of assemblies of the cultivated genome.

### KEY ACCOMPLISHMENT –

- The reassembly of pieces of DNA from 30,000+ genes in peanut is a step by step process that may be visualized as a series of small ‘Gene Islands’. A new technology called ‘MolecuLo’ was used to combine the small islands, thus creating larger islands. We now have a means to continue to connect the large islands until finally we have the whole cultivated peanut genome assembled or one large island.

“Each year, new technologies for genome sequencing and assembly are released that replace older methods. Our researchers have the reputation and connections to get free tests of technologies to see if they work on the peanut genome”, said Dr. Rich Wilson, technical consultant to the Peanut Foundation.

## 2014-2015 Peanut Genomic Research Accomplishments

**Component 5: Phenotyping Genetic Resources.** Phenotyping is the art of associating measurable and heritable characteristics with a DNA sequence that contains a gene that controls a trait. Phenotyping is necessary to create an accurate genome assembly. Phenotyping also helps identify and accelerate the selection of cultivars with resistance to CBR, late leaf spot, early leaf spot, aflatoxin contamination, white mold, TSWV; pod filling, yield, grade, drought tolerance, oil composition.

KEY ACCOMPLISHMENT –

- Three genes for TSWV resistance, three genes for Early Leafspot and three genes for Late Leaf Spot have been identified with the new array of DNA markers. In some cases, like Late Leaf Spot and Peanut Rust resistance, the genes appear to be close together possibly linked, which means that selection for one indirectly selects for the other or the breeder gets two for the price of one.

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. Although the technology of gene marker assisted selection in peanut is in its infancy, it already has had a tremendous impact on my breeding program”.

**Component 6: Making it All Useful through PeanutBase.** The ‘PeanutBase’ website contains the genetic and genomic information assembled in the Peanut Genome Project, including a Breeder’s Toolbox to assist peanut breeders in selecting gene markers and germplasm to be used in breeding programs. These tools include: genome browsers for the ancestral parents; gene finders; several large sets of genetic markers (available for download and for searching and browsing); and many mapped genetic traits available for search. Peanut genes also may be compared to counterpart genes from other legume genomes (including soybean, common bean, chickpea, and pigeon pea) to leverage genetic knowledge in those crops.

KEY ACCOMPLISHMENT –

- Training exercises were developed to demonstrate how PeanutBase can be used by breeders to facilitate marker-assisted-selection. One of the search exercises started from a trait (Root Not Nematode resistance), explored the genetic and genomic regions where a gene might be found, identified DNA markers or SNP’s for the gene, started a search for candidate genes (to develop perfect gene markers), and provided information to facilitate Marker Assisted Selection. Similar scenarios may be developed for Late Leaf Spot, Leaf Rust, High-Oleic Acid, and eventually all important peanut traits. To date, over 5797 users have viewed 117,199 pages in 13,570 sessions on PeanutBase since July 1, 2014.

“Resources generated by the PGI are useful when they are accessible to breeders and researchers. In the PeanutBase website we are working to integrate all of the information in the peanut genomics project to help researchers discover the basis for valuable traits, and use this knowledge to make faster breeding progress,” said Dr. Steven Cannon, Research Geneticist at USDA-ARS in Ames, IA.

**Overall Benefits to Breeders** -This project has made great strides in 2014-2015. 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”.

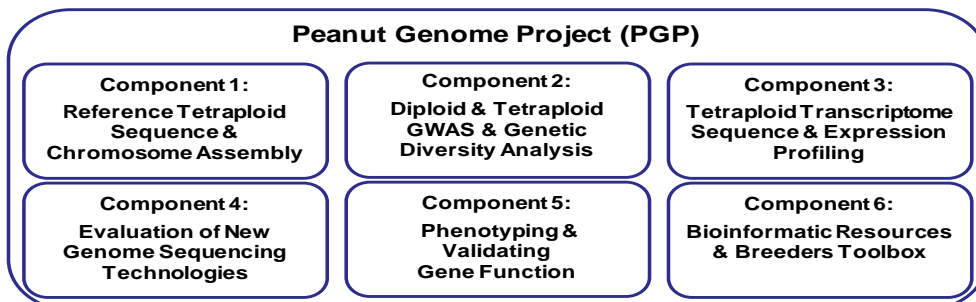
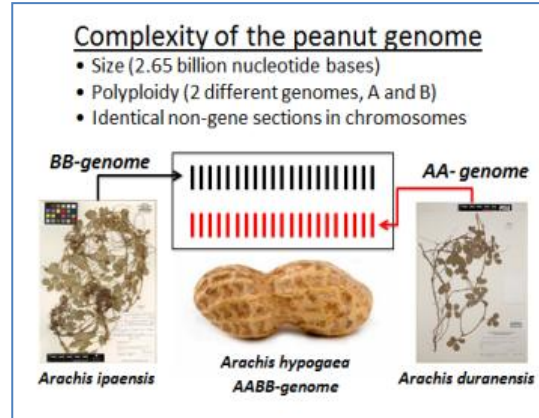


## Peanut Genome Project

### Research Technical Accomplishment Report to the U.S. Peanut Industry July, 2015

#### 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



**First peanut genome sequenced**  
THE INTERNATIONAL PEANUT GENOME INITIATIVE RELEASES FIRST PEANUT GENOME SEQUENCES

April 2, 2014



A. duranensis (A-genome)      A. ipaensis (B-genome)

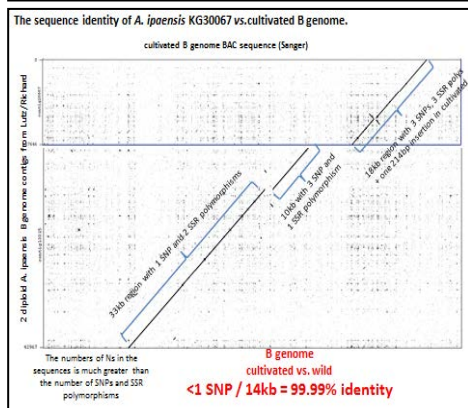
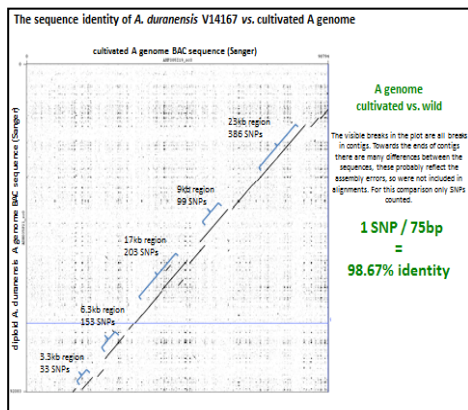
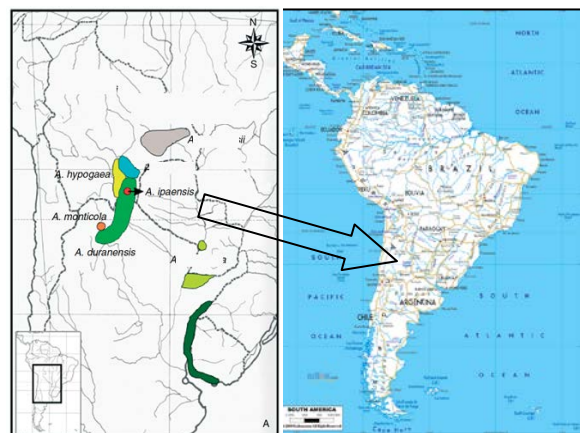
Remember when you used to consult a road atlas before taking a trip? Researchers now have an extremely high quality roadmap to find gene locations on peanut chromosomes.

On April 2, 2014, The Peanut Foundation (<http://www.peanutfoundation.org/>) announced the first peanut (groundnut) genome sequence is available at (<http://peanutbase.org/files/genomes/>). This resource will help expedite efforts to develop enhanced peanut cultivars in the U.S., China, India, South American and Africa.

### What has been learned from these genome sequences so far?

**Whole genome sequences have helped trace the location and origin of cultivated peanut.**

- Direct evidence showed that the A- and B-subgenomes of cultivated peanut (*Arachis hypogaea*) came from two diploid ancestors: *A. duranensis* (which is abundant in the wild) and *A. ipaensis* which has only been found in a single location in regions now in Argentina.



- DNA sequences of these two diploids were 98% identical to each other, and very similar to the A- and B-subgenomes of cultivated peanut.
- Gene markers could distinguish the A- or B-genome of the progenitor diploid species, and gene markers in a diploid progenitor species genome were found in the counterpart cultivated peanut subgenome.
- Diploid genomes contribute equally to the genes in the reference tetraploid genome (cv Tifrunner)
- The *A. ipaensis* B-genome is a near perfect match for the B-genome portion of the cultivated peanut genome and likely is a remnant of the very same population that contributed the B genome to cultivated peanut

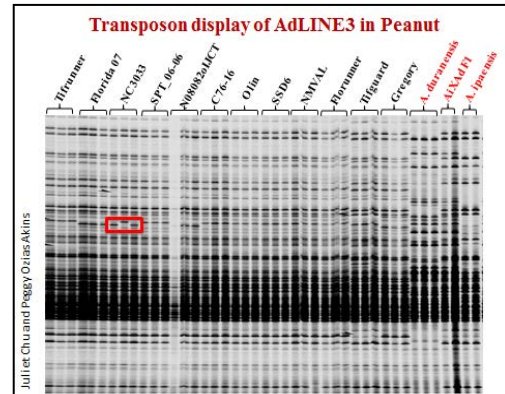
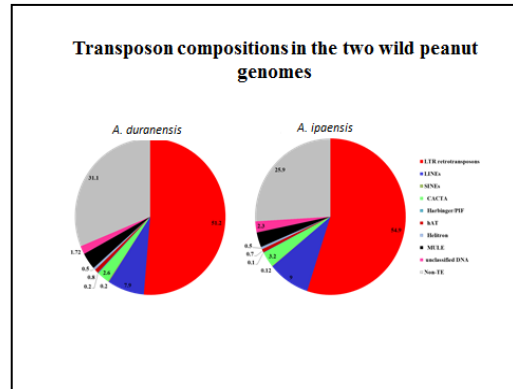
## 2014-2015 Peanut Genomic Research Accomplishments

around 10-20,000 years ago.

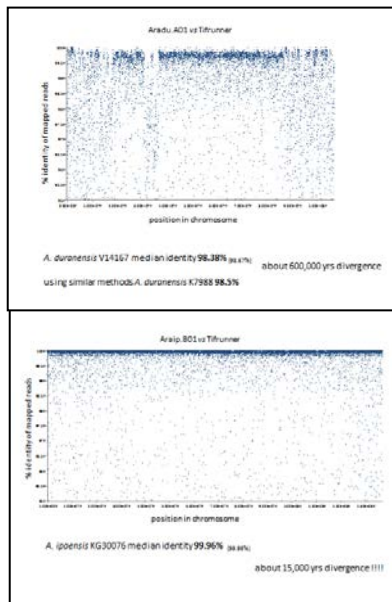
- There is high probability that *A. ipaensis* is the first actual progenitor genome of any crop to be sequenced.

### Understanding peanut chromosome structure helps create strategies for finding key genes.

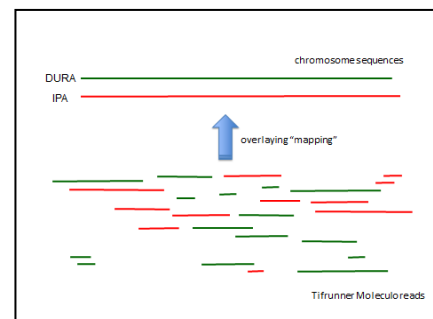
- About 70% of the cultivated peanut genome is composed of structures called ‘transposable elements’ or ‘jumping-genes’.
- The cultivated peanut genome has a greater proportion of transposable elements than any other legume.
- There are many types of transposable elements (such as: LTR retrotransposons, LINE, SINE).
- The function of these structures is unknown, but they seldom contain active genes. There is one example where a transposable element (called) MITE caused a good result, when it ‘jumped’ (inserted itself) into a DNA sequence that inactivated the FAD2-B gene (this generated a stable recessive allele that contributes to a high-oleic phenotype).
- In general, when sequenced the thousands of identical DNA fragments from transposable elements swamp the assembly, making it difficult to find active genes.
- A transposon library was created as a first step toward effective approaches to remove or assemble the ‘repeating’ sequences in diploid and tetraploid peanut genome assemblies.



### A picture of the ‘Gene Rich Regions’ in the peanut genome is coming into focus.

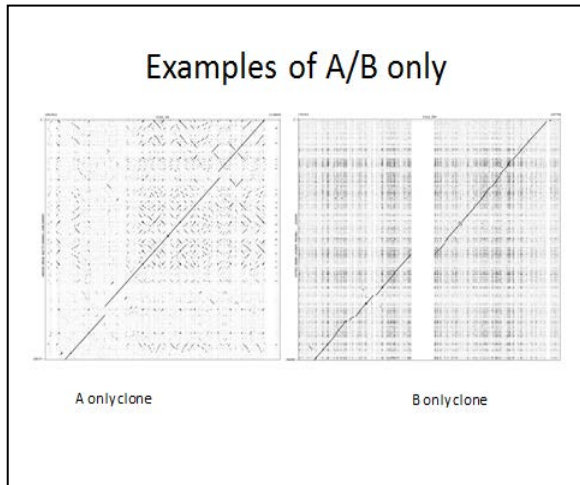
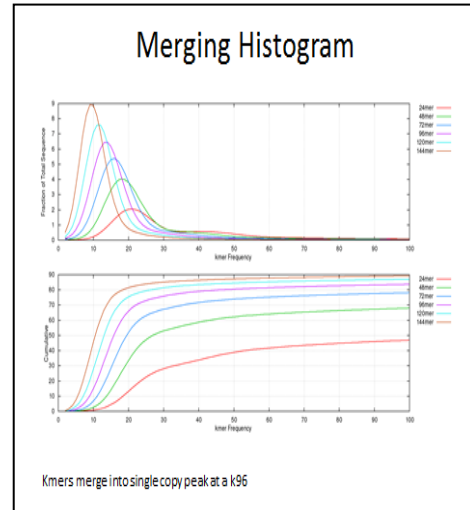


- The 30,000+ genes in peanut are highly conserved among A- and B-genomes, and may be visualized as ‘Gene Islands’ in a sea of transposable elements (repetitive DNA).
- ‘Moleculo’ reads (see Componet-4) were mapped to the two diploid genome assemblies to generate a ‘tiling path’ that should direct assembly of the tetraploid genome.
- A tiling path is a map that defines the minimal number and order of DNA fragments (contigs) needed to construct a chromosome.



**Moving forward on the tetraploid (cultivated) peanut genome assembly**

- Hudson Alpha was contracted to sequence and assemble 6 PCR-free libraries for Tifrunner, *A. duranensis*, and *A. ipaensis* (2 x 400 bp, 2 x 600 bp, 2 x 800 bp) with Rapid HiSeq technology.
- Kmers merged into a single copy peak at k96.
- Single shear results showed very low AT bias, which indicated that clear separation of A- and B-genomes was possible.
- Assembly of tetraploid data was initiated at 96X coverage of the subject libraries
- ABySS (a de novo, parallel, paired-end sequence assembler designed for short reads) assembled 2.7 Gb (Contig L50, 10.4 kb; Scaffold L50, 14.2kb) at 42X fragment coverage.



- These data were validated by SSPACE (a program that determines the order, distance and orientation of contigs in scaffolds).
  - Unique non-overlapping 100mers from the tetraploid libraries were aligned to A- and B-genomes.
  - The proportion of 100mers perfectly matched to each genome was: 15% (A), 26% (B), 0.4% (A and B). 57% of scaffolds could be called A- or B. Only 1.5% of scaffolds were unassigned.
  - Excellent separation of tetraploid subgenomes and respective alignment with counterpart diploid genomes was achieved
- Next steps include: completing the diploid sequences, developing nexerra pairs with longer reads to fill gaps in tetraploid scaffolds, and assembling tetraploid data sets with programs from ALLPATHS-LG (for Kmer assembly) and MERACULOUS (a whole genome assembler developed by JGI that uses Illumina sequence to assemble large-sized genomes on commodity clusters).

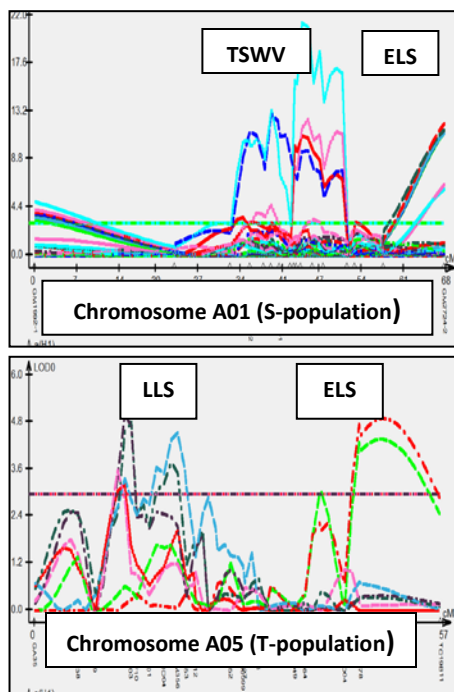
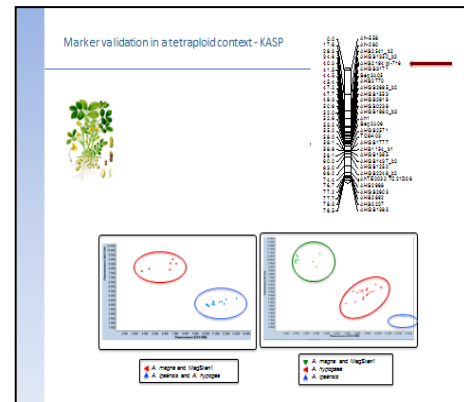
## Component 2: High-Density Genetic Maps & Gene Markers

The genotypes of individual progeny plants in breeding populations may be distinguished by thousands of DNA markers (SNPs or Single Nucleotide Polymorphisms). However, only a handful may be useful in a breeding program. New methods were tested to find validated SNPs that marked QTL of A- and B- genomes that harbored key genes involved in crop productivity, protection or improved quality. It was discovered that validated markers in diploid progenitor species also could help pinpoint their counterparts on chromosomes of the cultivated (tetraploid) genome. Validated DNA markers have been positioned on an international reference genetic map to help breeders locate genes for agronomic traits. This road-map of DNA markers is being used to test the utility of a new and revolutionary breeding strategy called “Genotyping By Sequencing (GBS)”, a new breeding method that should accelerate the development of superior peanut varieties in a timely manner.

### What has been learned from QTL discovery and mapping so far?

#### Discovering that genes for genetic disease resistance may be clustered on the same chromosome.

- Homozygous and heterozygous alleles (genes) in a segregating population may be distinguished with KASP technology.
- KASP (Kompetitive Allele Specific PCR) is a fluorescence-based genotyping technology that identifies specific alleles.
- This technology enables the location of a specific allele (gene copy) to a particular chromosome, and also to the A- or B-genome in DNA from diploid or tetraploid peanuts.



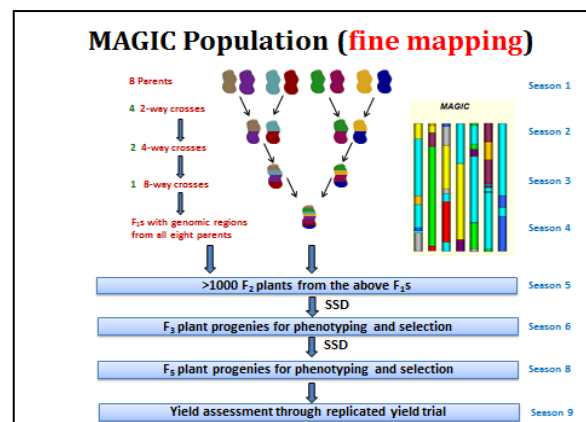
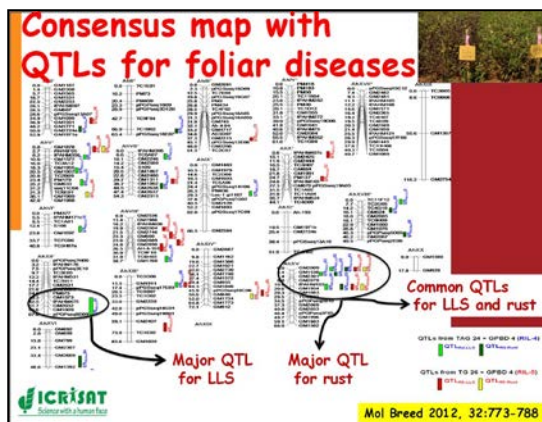
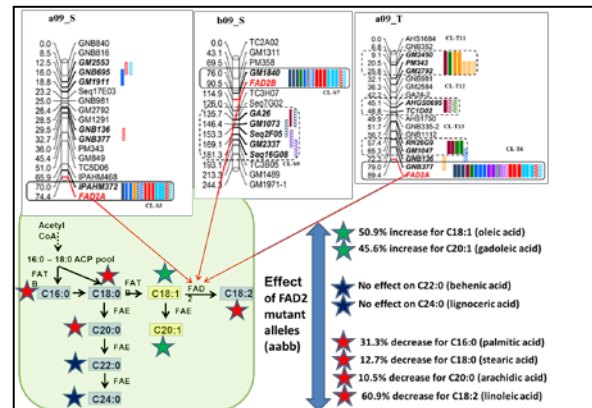
- Genes for a given trait do not always locate on the same chromosome in different breeding populations because of how DNA is exchanged between parents
  - For example, S population: SunOleic 97R x NC94022. SunOleic 97R is susceptible to TSWV, ELS, LLS, high O/L. NC94022 is resistant to TSWV, LLS, ELS, low O/L
  - Another example: T population: Tifrunner x GT-C20. Tifrunner is resistant to TSWV, ELS and LLS; GT-C20 is susceptible to TSWV, ELS, LLS.
- Multiple markers in a region indicate several copies of a gene for a resistance trait (like TSWV or LLS) may co-locate on the same QTL.
- Clusters of genes for different disease resistance traits may cluster on the same chromosome (TSWV & ELS in S-population) or (LLS & ELS in T-population)
- The S-population exhibits gene-rich regions for disease resistance traits on the A01 and B03 chromosomes in both diploids and tetraploid genomes; The T-population has gene-rich regions for disease resistance on A05 and A06



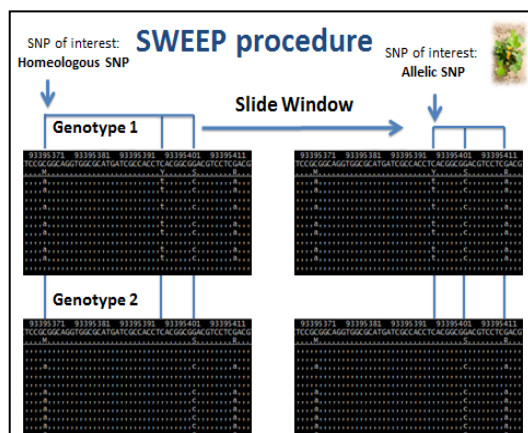
## 2014-2015 Peanut Genomic Research Accomplishments

chromosomes in the diploid and tetraploid genome.

- The genetic regulation of quality traits such as oleic acid concentration in peanut oil may be easier to handle in a breeding population.
- Recessive FAD2A and FAD2B genes mediate the high oleic acid trait.
- Double recessive FAD2 genes also cause an increase in gadoleic acid; decreases in palmitic, stearic, arachidic and linoleic acids
- FAD2A alleles are found on A09 in both the S- and T-populations
- FAD2B alleles are found on the B09 chromosome in the S-population



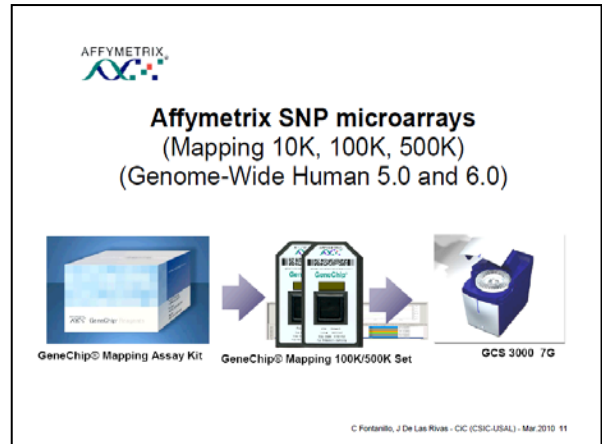
- Consensus maps were developed that show the most probable locations of QTL for specific traits on each peanut chromosome (based on all available mapping populations)
- A breeding method called MAGIC (Multi-parents Advanced Generation Inter-Crossing) is being used to pyramid genes from each mapping population parent into a single mapping population that will test the accuracy of the consensus maps
- However, Marker Assisted Selection is limited when gene markers for each trait are assayed one-at-a-time among the progeny of a breeding generation.
- GBS methods help improve selection efficiency, but resequencing is expensive because the thousands of SNPs found in each line must be validated, and only a few SNPs are relevant to the selected traits



- New genomic approaches were initiated to improve the accuracy of SNP validation (Universal Network Enabled Analysis-UNEAK and Sliding Window Extraction of Explicit Polymorphisms-SWEEP)
- These methods were tested with a Reference panel that included over 60,000 SNPs from 34 tetraploid and 24 diploid genomes..
- Another test was run on leaf transcriptome sequences from 6 genotypes (representing all market types) that generated over 55 million SNPs
- 80,000 were retained after initial SWEEP filtering

## 2014-2015 Peanut Genomic Research Accomplishments

- 29,000 SNPs after maximum quality filtering
- Achieved 85% overall accuracy in SNP calling (validated by Sanger sequencing); 96% mapped accurately to the B-genome; 67% mapped accurately to the A-genome
- SWEEP and UNEAK provide accurate ID of allelic SNPs for high confidence application of GBS
- SNP sets based on SWEEP and UNEAK data were provided to AFFYMETRICS Inc to develop a 60K SNP chip
- Affymetrix pioneered the development of SNP chip technology, and has many commercial products
- This technology will expedite characterization of genetic diversity in peanut germplasm and breeding populations
- The initial products are now available for use
- Further modifications to this technology will enable accurate and cost effective identification of copy numbers for specific alleles

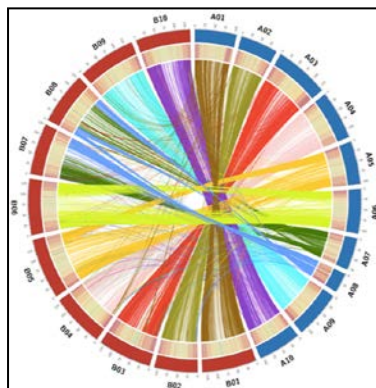


### Component 3: Expressed Gene Sequences

This research helps characterize the sequence of genes that are made (expressed) in the cultivar Tifrunner 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. .

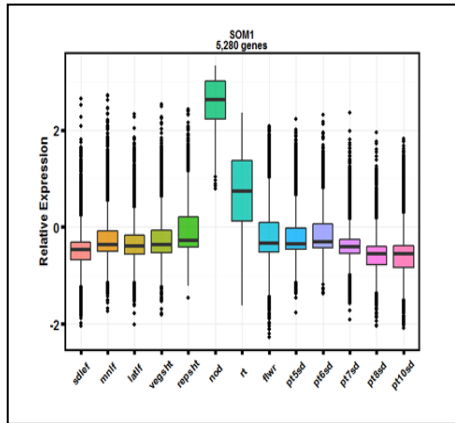
### What has been learned from gene expression analysis so far?

- 58 RNA libraries from 22 organs/ developmental stage (2-3 reps each plus control treatments) of the cv Tifrunner. These libraries represented 17000 expressed genes were mapped to the A-or B-genome.

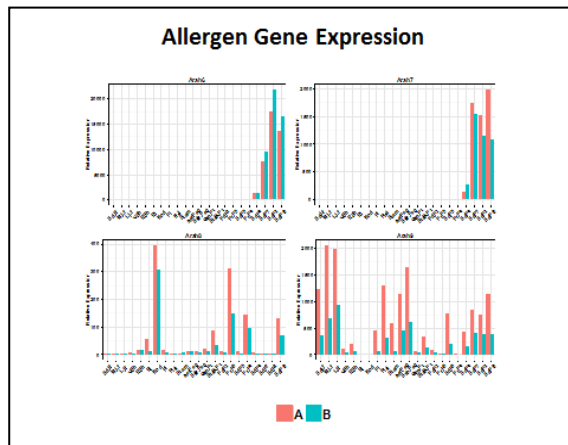
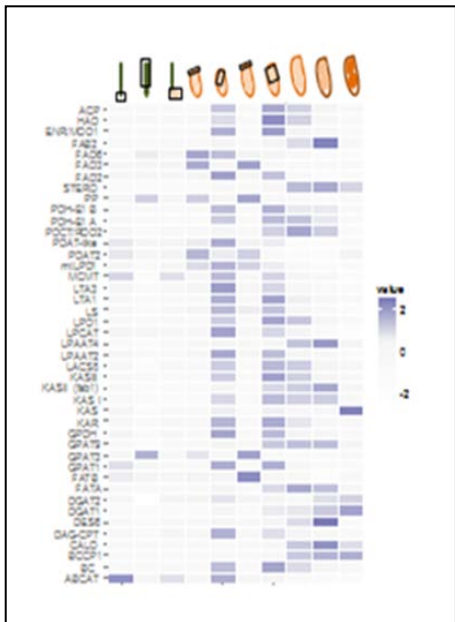


No	Tissue	Stage	sample collection	RNA extract	Agilent check	Sequenced	id	Flower	Fully open, morning of anthesis; stigma and ovary	yes	yes	yes
1	Leaf	10 d post-emergence; leaflets partially open	SOY:0000252	yes	yes	yes	yes	8: Flower	Fully open, morning of anthesis; anthers	yes	yes	yes
2	Leaf	Growth stage Bootie R1 – first flower; leaflets partially open, from mainstem (n)	yes	yes	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
3	Leaf	Growth stage Bootie R1 – first flower; leaflets partially open, from laterals (n+1)	yes	yes	yes	yes	yes	10: Gynophore tip (pod)	At pod swelling (Pattee stage 1)	yes	yes	yes
4a	Vegetative shoot (5 mm max)	Growth stage Bootie R1 – first flower, from mainstem (n)	yes	yes	yes	yes	yes	11: Gynophore "stalk"	At pod swelling (Pattee stage 1)	yes	yes	yes
5a	Reproductive shoot (5 mm max)	Growth stage Bootie R1 – first flower, from laterals (n+1)	yes	yes	yes	yes	yes	12: Pod	Pericarp very watery, embryo very small and not easily removed (Pattee stage 3/4)	yes	yes	yes
6	Root structures	10 d post-emergence	SOY:0000252	yes	yes	yes	yes	14: Pericarp	Pericarp soft, not as watery, inner pericarp without cracks (Pattee stage 5)	yes	yes	yes
7	Nodules	25 d post-emergence	SOY:0001301	yes	yes	yes	yes	15: Seed	embryo flat, white or just turning pink at one end (Pattee stage 5)	yes	yes	yes
8a	Flower	Fully open, morning of anthesis; wings, banner, hyparhium, keel; SOY:0001277	yes	yes	yes	yes	yes	16: Pericarp	Inner pericarp tissue beginning to show cracks or cottony (Pattee stage 6/7)	yes	yes	yes
								17: Seed	Torpedos shaped; generally pink at embryonic-axis end of kernels (Pattee stage 6)	yes	yes	yes
								18: Seed	Torpedos to round shaped; embryonic axis end of kernel pink; other end white to light pink (Pattee stage 7)	yes	yes	yes
								19: Seed	Round, light pink all over (Pattee stage 8)	yes	yes	yes
								20: Seed	Large, generally dark pink all over; seed coat beginning to dry out (Pattee stage 10)	yes	yes	yes

## 2014-2015 Peanut Genomic Research Accomplishments




- In response to experimental treatments, the rate at which a gene is activated or deactivated can be visualized by the level of transcription, either up-regulated or down-regulated
- Different patterns of gene expression are found among plant organs (ie leaves, seed, roots) at difference stages of plant development.
- For example, the arah1, arah2, and arah3 (arah 6 and arah7 as well) storage proteins (considered antigenic) are expressed only in seed, whereas arah 5 (including arah8 and arah9) are expressed in all organs



- Another example, genes that encode enzymes in the oil synthetic pathway are expressed a different stages of seed development (ie. high expression of the FAD2 gene which govns oleic acid synthesis occurs early in seed development).


### Nematode RNAseq





- Parental Genotypes:  
Tifguard (Resistant, A09 introgression)  
Gregory (Susceptible)
- Recombinant RILs:  
46 (Resistant)  
48 (Partially resistant)
- Inoculated and Control
- Time points:  
0 (20 day-old roots)  
3 days after infection  
7 days after infection


### Tifguard Introgression Mapping

- Mapped the reads to a Tifrunner reference transcriptome, used SWEEP to call SNPs
- Mapped 73 SNPs to A09

**Tifguard - R**  


**Gregory - S**  


**48 - Partially R** 2 Mb  


**46 - R** 100 kb  


- Gene expression profiling can be used to explain why some peanuts are only partially resistant to a disease.



## 2014-2015 Peanut Genomic Research Accomplishments

- For example: Differential expression profiles of genes associated with Root Knot Nematode resistance were observed at 3 dates after infection of the cv Tifguard, Gregory, and a resistant and a partially resistant RIL from Tifguard x Gregory
- Expressed gene sequences were mapped to a Tifrunner reference transcriptom. SWEEP was used to call SNPs. 73 SNPs were mapped to chromosome A09.
- A regions of the A09 that carry introgressions from Tifguard were mapped in RIL46 (resistant) and RIL48 (partially resistant).
- The RIL46 expression data suggests the location of a novel RKN resistance gene on A09.

### 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 has been learned about technologies for sequencing/assembling peanut genomes?

- The main impediment to assembly of the tetraploid peanut genome is a relatively low number of validated sequence and structural variants (unique markers) that can be used to distinguish the A- and B-genomes in Tifrunner. This situation may be attributed to the relatively short time that cultivated peanut has been in existence (i.e. cultivated peanut is too young in the evolutionary sense to have accumulated an optimal number of variants needed to facilitate genome assembly).
- Thus far the Peanut Genome Project (PGP) has generated a number of libraries of DNA fragments (from Tifrunner) that range in size (length) from 170b to 40kb; about 60% are less than 500 b and about 0.2% are longer than 10 kb. DNA fragments within each library vary in length; those that contain a common sequence can be aligned computationally into monolithic stacks or scaffolds. A single progenitor species genome has an estimated 600,000 scaffolds. Unfortunately, all scaffolds in the entire population of libraries do not overlap and may emulate an ocean filled with numerous islands of various sizes.
- The goal of genome assembly is to link as many of the scaffolds as possible in proper order and place. However, the occurrence of unique distinguishing structural variations (such as a SNP) within and among scaffolds from short DNA fragments often is rare. A greater proportion of scaffolds with longer 'read' length (i.e. greater than 20 kb) would help span the gaps, bridge between sets of shorter length scaffolds and facilitate chromosome assembly.

Several new technologies for sequencing and assembling peanut genomes have been evaluated since the launch of the PGP. Many have promise, but also have limitations. For example:

- BAC Sequencing Strategy as proposed by BGI employs pooling multiple BAC libraries to guide the assembly of tetraploid chromosome models. Although the BGI assembly protocol was optimized, it is still difficult to assemble BACs from pooled peanut BAC samples. Results suggest a 2+ BAC pool was less promising compared to single-BAC x BAC sequencing. Therefore, a BAC x BAC sequencing strategy for Tifrunner using two BAC libraries (**250 bp, 500 bp**) with an average sequencing depth of 50X for each BAC pool was considered. However, BAC x BAC is a relatively expensive approach. Implementation has been postponed pending evaluation of alternative strategies.
- BioNano technology features IrysView data analysis software that recreates a whole genome consensus map of the original genome. This approach requires highly contiguous gap-free assembly

## 2014-2015 Peanut Genomic Research Accomplishments

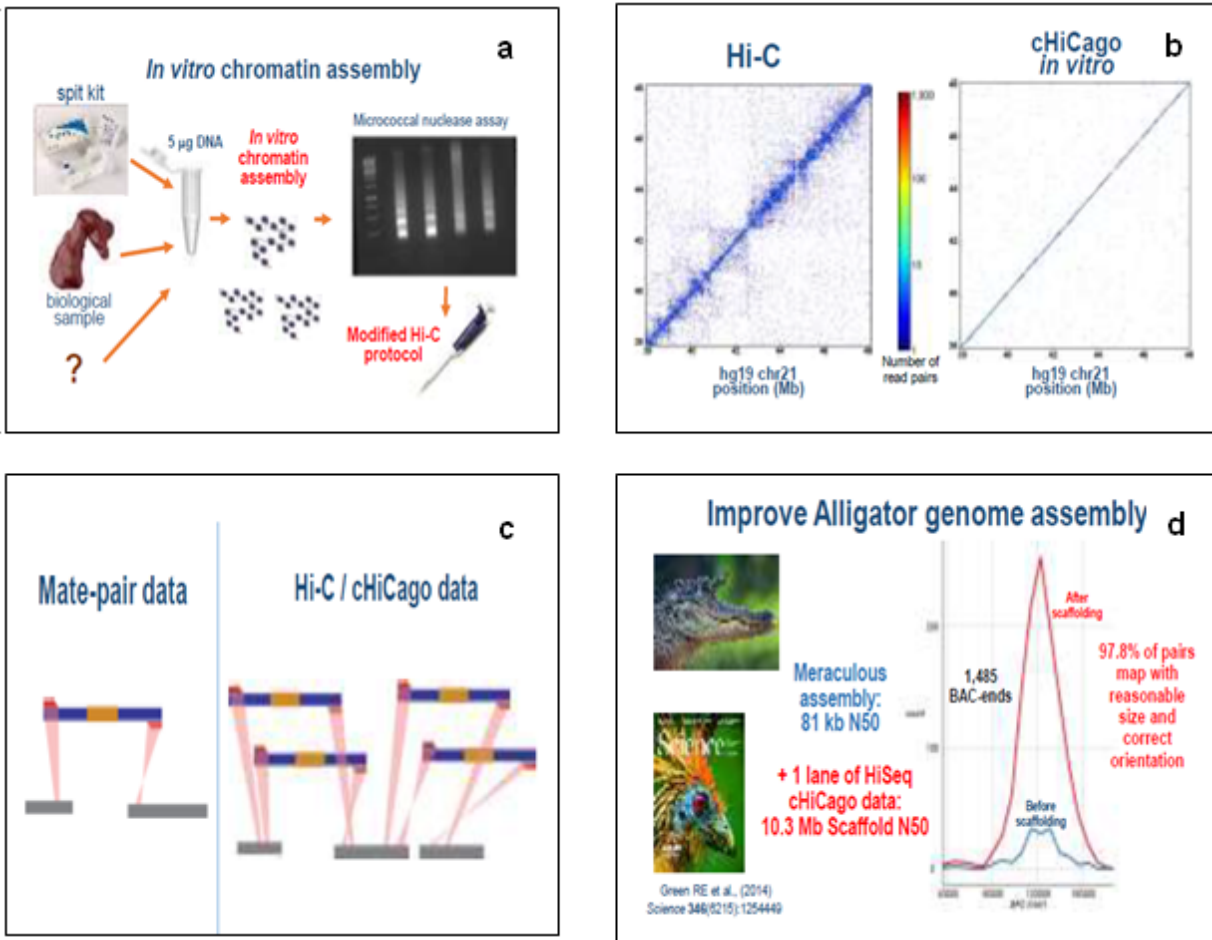
data (it can map neighboring contigs over **40 kb** in length). The peanut genome assemblies currently are not suitable for this method.

- Oxford Nanopore Sequencing: This new technology, which is still limited to an early access (beta) program, can potentially generate long reads of up to **80 kb**. Beta-tests on high quality viral DNA currently shows many quality problems. 70% of the sequence data do not align with the reference genome. Error rates ranged from 20 to 30% in the high quality parts of the Nanopore reads. The technology has been improved but currently is not useable for big genomes. It will be reevaluated when data of control samples demonstrate significant improvements.
- PacBio technology presents a real time sequencing protocol for very long reads from single DNA molecules. New chemistry is being tested on plant genomes (by Froenicke). Error rates were about 15% on **15-20 kb** reads. Although throughput recently has more than doubled to 1.3 Gb with good quality DNA, the new chemistry protocol has not been tested for peanut. Because of the error rate PacBio data is useful only at a minimum genome coverage of 10X, preferably 20X.
- Moleculo. Extremely good progress has been made with Moleculo™ technology by Illumina on diploid genomes. A-genome analysis at 4X coverage produced high quality long reads (mean, **3.7 kb**, longest, **22 kb**); and B-genome analysis at 6X coverage yielded reads with mean length of **4.1 kb** (longest, **20 kb**). These reads can differentiate A- and B-genomes with very low error; but the A-T rich content of the peanut genome is not handled well by Moleculo, causing gaps in the assembly. Illumina is working to overcome this limitation, and has developed data sets (0.5X coverage) of A-T rich regions in the tetraploid peanut genome. Results are pending.
- Spiral Genetics proposed Anchored Assembly (AA) technology as an analysis pipeline to detect and map variations that often are missed by standard analysis algorithms. AA uses direct de novo read overlap assemblies to detect and characterize SNPs, indels and structural variants. The pipeline developed for peanut used existing Illumina™ Hi-Seq™ data for tetraploid assembly based on reference diploid assemblies. However, only 3 million variants were detected; 90% were SNPs; and the A-genome presented 10-fold more SNPs than the B-genome which left major gaps in the Tifrunner chromosomal assemblies. No further consideration has been given to this approach.
- AnyTag assembler/technology. This assembler is being evaluated because of its unique ability to fill overlaps on both sides of a read. The latest version is not public yet, but is being tested on tetraploids (fish) with good results. This method would need 3-4 Tifrunner libraries (**400bp, 600bp, 800bp, 1000bp**). The libraries generated for the assembly (2x250 PE) should also be compatible with the clustering approach proposed by Hudson Alpha. Further evaluation of this approach is pending.
- Hudson Alpha, a clustering approach that sequences **400bp, 600bp and 800bp** libraries from both diploid ancestors and Tifrunner to 20X at 2x250 bp read lengths in rapid mode on an Illumina Hi-Seq platform. Clustering methods will partition **250 bp paired end** Hi-Seq reads from Tifrunner to respective A- and B-subgenomes and verified with diploid ancestor (A- or B-genome) reads. This will allow assembly of Tifrunner A and B genomes in separate runs ([See Component 1](#)).
- NimbleGen. SeqCap EZ Library technology will be evaluated for capture of the whole peanut exome (gene rich region) in a single extract. Exome capture should enhance discovery of coding variants (markers) – it is not carried out for genome assembly purposes. TPF funded this work for 2015 to generate the first exome data for peanut.

Another genome assembly technology that has drawn PGC attention was showcased by Dovetail Genomics LLC (<http://www.dovetailgenomics.com>) at PAG XXIII. Dovetail Genomics has developed a modified Hi-C protocol (cHiCago) for in vitro long-range mate-pair assembly of genome scaffolds. The Hi-C protocol uses proximity ligation and massively parallel sequencing to probe the three-dimensional architecture of chromosomes within the nucleus; the cHiCago protocol applies the same principle to isolated very long DNA fragments (up to **150 kb**) *in vitro*, thus removing most of the noise from the analysis (please see figure b below). Interacting regions are captured as paired-end reads. Genome-wide chromatin interaction data sets generated

## 2014-2015 Peanut Genomic Research Accomplishments

by these protocols provide information about the grouping and linear organization of sequences along entire chromosomes. A computational method generates de novo genome assemblies by: 1) clustering contigs or scaffolds to chromosome groups; 2) ordering contigs or scaffolds within each chromosome group; and 3) assigning relative orientations to individual contigs or scaffolds. (Reference: J.Burton, A.Adey, R.Patwardhan, R. Qiu, J. Kitzman, J. Shendure. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology. November 2013; doi:10.1038/nbt.2727). The following pictorial comparison of results from Hi-C and the cHiCago protocols is reprinted from: An in vitro Long-Range Mate-Pair Protocol for Genome Scaffolding, by Richard E. Green, University of California, Santa Cruz & Dovetail Genomics, LLC.



The quality of results from the cHiCago protocol depend on DNA fragment size and purity. 5.5 µg DNA is the minimum amount required for sequencing a single Hi-seq lane (rapid mode) with PE100 reads. The cHiCago protocol zips up (assembles into ordered scaffolds) mammalian and reptile genomes very efficiently and with high accuracy (scaffold N50 = 10+ Mb). The fragments used to span gaps in an assembly have sizes up to **150 kb (nearly 40-times longer than the average length of Moleculo reads for diploid peanut)**. Thus the cHiCago protocol could help validate the order of the progenitor diploid assemblies, and help interconnect hundreds of thousands of scaffolds that are present in the tetraploid peanut assembly. Although Dovetail Genomics does not have finished data for plants yet due to problems with DNA fragment size and purity, recent improvements in plant DNA isolation protocol have generated 150 kb DNA fragments that work for several plant species.

### Component 5: Phenotyping Genetic Resources

Phenotyping is the art of associating measurable and heritable characteristics with a DNA sequence that harbors or encodes a gene that mediates the trait. Accurate genome assembly is not possible without a multitude of these intrinsic connections. Massive peanut germplasm collections in India, China and the U.S. plus a substantial number of breeding and mapping populations provide millions of potentially useful markers that facilitate gene discovery, tracking genes among progeny of a breeding population. Phenotyping utilizes these genomic tools to ‘connect the dots’ to make practical application of the information and knowledge that is gained from genome sequences, and accelerate the selection of elite cultivars in a timely manner.

#### What has been learned from the resident arsenal of phenotypic resources?

- Major peanut germplasm collections are maintained in India (ICRISAT), China (CAAS) and the U.S. (USDA-ARS). The U.S. collection contains about 10,000 accessions (including 700 wild species).
- A massive effort was initiated to characterize phenotypes and associated genotypes of all accessions in the U.S. peanut germplasm collection
- This effort includes creation of digital profiles of the visual and analytically determined phenotypic characteristics of each line



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
N080820JCT	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	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

- Genotypes are being documented by GBS or new technology to characterize genetic diversity.



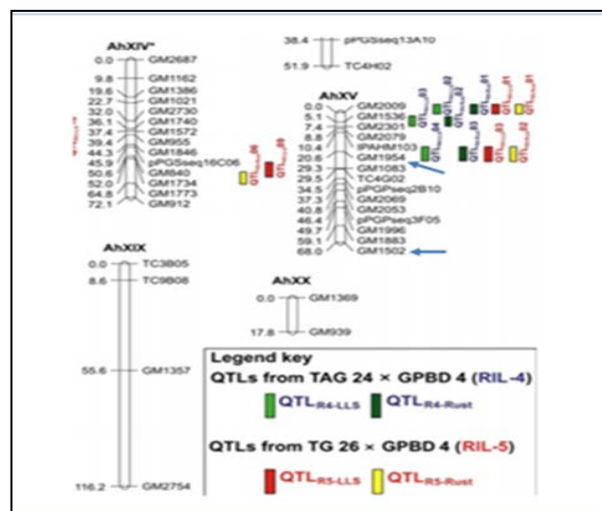
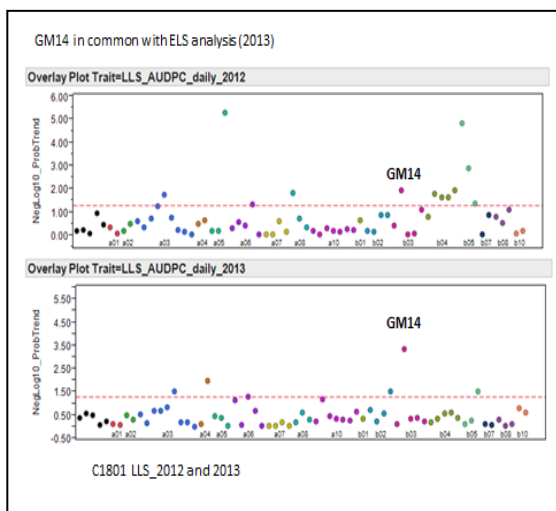
## 2014-2015 Peanut Genomic Research Accomplishments

The development, maintenance and characterization of the peanut phenotyping resources is a monumental team effort. The 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 N08082oIJCT (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.

RIL populations were increased and phenotyped in 2015. For example, special attention is given to the C1799 and C1801 RIL populations. Birdsong Inc. graciously stored all seed for these populations in -18C freezers. A seed-inventory is posted on [www.PeanutBioscience.com/](http://www.PeanutBioscience.com/).

- C1799: a RIL population of 375 lines from Tifrunner x NC3033. 286 lines were genotyped with 105 polymorphic SSR markers. 165 lines were advanced. QTL on chromosomes A04, A06, and A07 accounted for about 45% of the phenotypic variation for TSWV resistance.
- C1801: a RIL populations of 381 lines derived from Florida 07 x SPT-06-06. was genotyped with 61 polymorphic SSR. 161 lines were advanced. QTL on chromosomes A03, B05, and B06 accounted for 42% of the phenotypic variation for ELS resistance. QTL on chromosomes A01, A04, and B06 accounted for 40+% of the phenotypic variation for LLS resistance.

**Genome wide tracking of QTL for LLS resistance reveals the location of resistance genes on multiple chromosomes, and because of gene clustering at a QTL selection for LLS resistance may also be complemented by resistance to other diseases such as Peanut Rust**



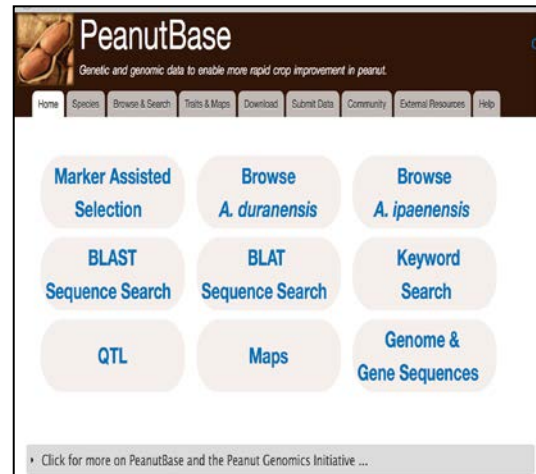
## Component 6: Making it All Useful through PeanutBase

Peanut genomic research has the potential to create exabytes (1 trillion trillion bytes) of data. PeanutBase.org was created to provide a secure internet-based home for all data generated by the peanut genome initiative; and to present a platform for genomic analyses of those data in the practical form of a ‘Breeders Tool Box’.

### What is in PeanutBase.org?

PeanutBase features include:

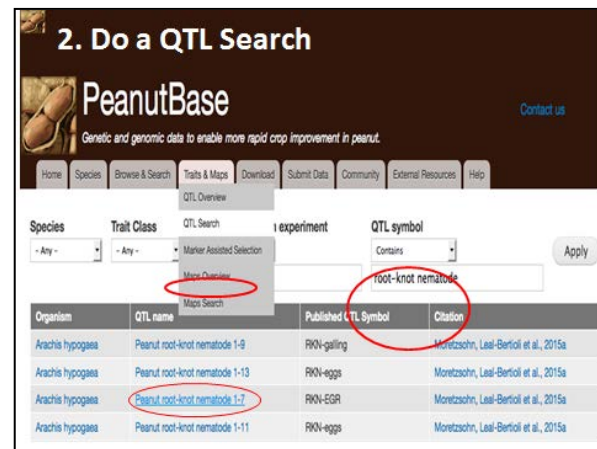
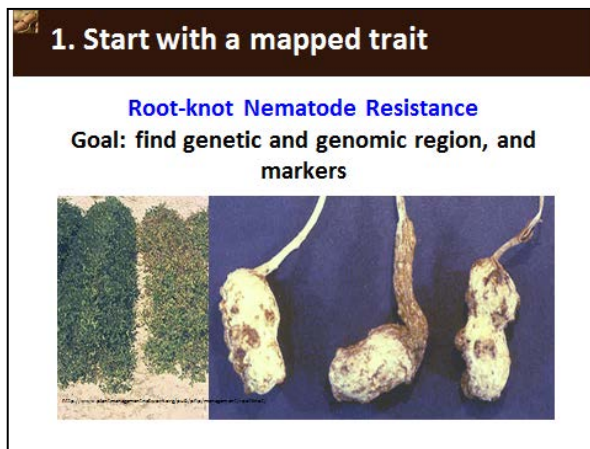
- The complete genome sequences of the two closest wild relatives of cultivated peanut
- Links to genetic and genomic resources for peanut, legumes and other external sources
- Biological information on peanut and relatives to provide context
- Map, trait, and QTL information with methods for data collection & analysis
- Ability to query and view QTL positions on genetic maps in various formats.
- High powered genome browsers which are very responsive, scalable, well integrated
- “Virtual/physical” maps that are tied to QTL
- A collection of peanut phenotype descriptors with relation to other ontologies
- Tools for using haplotype, accession, & phenotype data for breeding
- Gene family & orthology tools, including an Atlas of gene functional information
- Training, outreach, and coordination tutorials
- Sequence and key-word search tools



PeanutBase also leverages 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 can PeanutBase help Breeders Create Superior Peanut Varieties?

The following shows how PeanutBase can help breeders find and use markers for specific traits. The example is based on markers for Root Knot Nematode resistance.



# 2014-2015 Peanut Genomic Research Accomplishments

### 3. Search QTL data

Home Species Browse & Search Traits & Maps Download

QTL Overview  
[QTL Details](#)  
 Map Positions  
 Measurements  
 Trait  
 Alignments  
 Relationships  
 Synonyms

**QTL Details**

Nearest Marker	Leg1Gm
Flanking Marker Low	
Flanking Marker High	
LOD	6
Likelihood ratio	
Marker R2	11.9
Total R2	
Additivity	0.5775

### 4. Use the marker in a keyword search

Home Species Browse & Search Traits & Maps Download

Search genom BLAST  
 Enter terms to search BLAT  
 of chromosomes or Maps Search  
 Search term: Leg1  
 Examples: lipoxigenase  
 Species: A. duranensis

QTL Search  
 Publication Search  
 Feature Search  
 Keyword Search

marker r  
 in go fro  
 Aradu.F1

QTL Overview  
[QTL Details](#)  
 Map Positions  
 Measurements  
 Trait  
 Alignments  
 Relationships  
 Synonyms

**QTL Details**

Nearest Marker	Leg1Gm
Flanking Marker Low	
Flanking Marker High	
LOD	6
Likelihood ratio	
Marker R2	11.9
Total R2	
Additivity	0.5775

### 5. Find the genome / chromosome

Home Species Browse & Search Traits & Maps Download Score Data Community

Search genom BLAST  
 Enter terms to search BLAT  
 of chromosomes or Maps Search  
 Search term: Leg1  
 Examples: lipoxigenase  
 Species: A. duranensis

QTL Search  
 Publication Search  
 Feature Search  
 Keyword Search

marker r  
 in go fro  
 Aradu.F1

Genomes  
 Sequence Search  
 BLAST  
 BLAT  
 Maps Search  
 QTL Search  
 Publication Search  
 Feature Search  
 Keyword Search

**File Help**

PeanutBase - *Arachis duranensis* 1.0: 878 bp from Aradu.A08:112,848,858..112,850,  
 Browser Select Tracks Snapshots Custom Tracks Preferences

Search Landmarks or Region Search Set 1

Data Source  
 PeanutBase - *Arachis duranensis* 1.0

The following 1 regions match your request.

Name	Type	Description	Position
Leg1	SNP	LG A08:chr 1:700	Aradu.A08:112848858..112850855

### 6. From marker .. to genome sequence

The marker's genomic region on Chr A.09

Remember: the genomic region may be large!

### 7. From QTL ... to genetic map

The marker's genetic region on Chr A.09

The genetic and genomic regions are both valuable; check both.

click on arrows to zoom and identify other markers

### 8. Marker Assisted Selection (MAS)

PeanutBase  
 Search for genes and traits on the world's largest peanut genome database

Rust, Puccinia

Markers for the trait

- GWAM103 • GM1536 • GM2021 • GM2079
- TE 360 • TE 498
- SSR\_G03A045 • SSR\_H0115758

Informative for learners

TE 360 TE 498  
 RP GP RB GB M RP GP RB GB

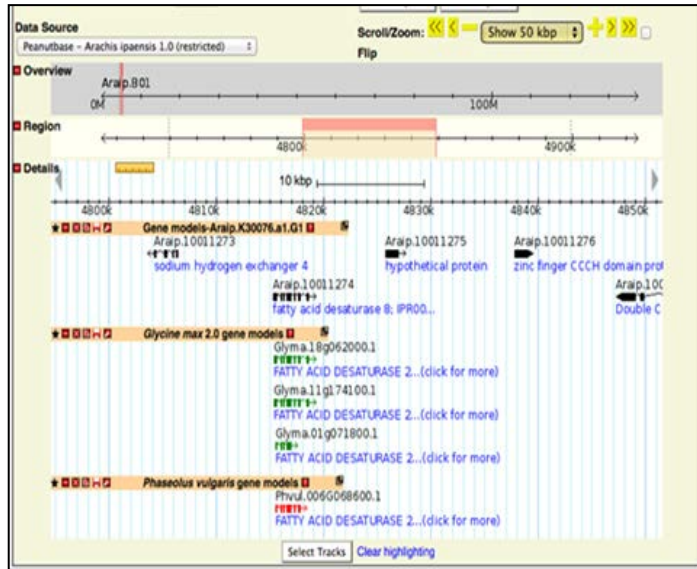
This exercise started from a trait (RKN resistance), explored the genetic and genomic regions where a gene might be found, identified gene markers for the QTL, started a search for candidate genes (to develop perfect gene markers), and provided information to facilitate Marker Assisted Selection.

Similar scenarios may be developed for Late Leaf Spot, Leaf Rust, High-Oleic Acid, and eventually all important peanut traits.



### PeanutBase reaches out to other genome databases

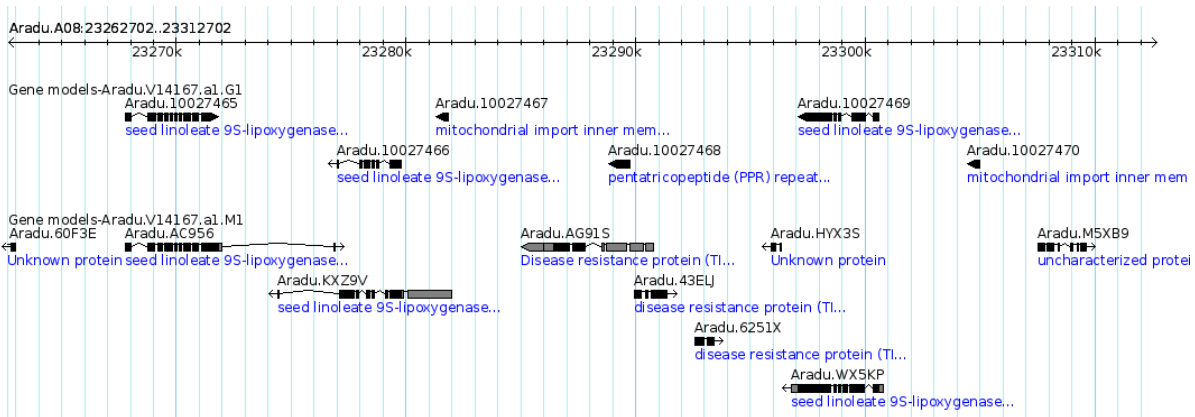
Genes for a given peanut trait in peanut also may occur in other legume species such as soybean and common bean. Often times those genes have common structural features. PeanutBase can access the other genome databases to capitalize on those similarities in various ways. As shown below, the function and location of a gene that regulates oleic acid concentration in peanut seed was validated with sequences of the FAD2 gene from soybean and common bean.



- Example shows the location of a gene that encodes a fatty acid desaturase, FAD2, a gene that mediates oleic acid concentration in peanut seed
- PeanutBase shows a whole length map of chromosome 01, B-subgenome, A. hypogaea
- QTL with FAD2 may contain other genes, as indicated
- The ID of FAD2 in peanut was confirmed by sequence comparison to FAD2 genes from Soybean (3-different copies) & Common bean

### New features are being added to PeanutBase

Most recently, a new gene expression resource has become available at PeanutBase. It contains differential transcriptome data from 22 tissues from Tifrunner (See Component 3), with reads mapped onto the diploid genomes. The tracks will be made public fairly soon when published. The example below shows some of the differential expression mapped to A08 in *A. duranensis*:



### PeanutBase use statistics

From July 2014 to July 2015, 5797 users conducted 13,579 sessions and viewed 117,199 pages. 42% of the users accessed PeanutBase for the first-time.

## 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*

## 2014-2015 Peanut Genomic Research Accomplishments

### **Exhibit 2: Sponsors who provide financial support for the Peanut Genome Project**

#### **U.S. Peanut Sheller Companies:**

- American Peanut Sheller's Assoc. Birdsong Peanuts
  - Damascus Peanut Company
  - Golden Peanut Company
  - McCleskey Mills
  - Snyder's/Lance
  - Tifton Peanut Company
  - Williston Peanuts
- Southwestern Peanut Sheller's – Birdsong Peanuts
  - Clint Williams Company
  - Golden Peanut Company
  - Wilco Peanut Company
- Virginia Carolina Sheller's 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

#### **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)

#### **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](#)

National Center Genome Resources  
Peanut Company 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 (Co-chair)**  
**Peggy Ozias-Akins, UGA (Co-chair)**  
**Richard Michelmore, UC-Davis (Co-chair)**  
**Rajeev Varshney, ICRISAT (India)**  
**Howard Valentine, TPF**  
**Raymond Schnell, MARS, Inc.**  
**Victor Nwosu, MARS, Inc.**  
**Corley Holbrook, USDA-ARS**  
**Baozhu Guo, USDA-ARS**  
**Brian Scheffler, USDA-ARS**  
**Steven Cannon, USDA-ARS**  
**Andrew Farmer, NCGR**  
**Tom Stalker, NCSU**

**Xin Liu, BGI**  
**Lutz Froenicke, UC-Davis**  
**Haile Desmae, ICRISAT-WCA**  
**Boshou Liao, CAAS, (China)**  
**David Bertoli, U Brazila (Brazil)**  
**Soraya Bertoli, EMBRAPA**  
**Xingyou Zhang, HAAS (China)**  
**Xingjun Wang, SAAS (China)**  
**Mark Burow, TAMU**  
**Graeme Wright (PCA (Australia)**  
**Sachiko Isobe, KDRI (Japan)**  
**Ran Hovav, ARS TVC (Israel)**  
**Richard Wilson, OBC**

#### **Ex Officio**

**Roy Scott, USDA-ARS-ONP**  
**Maricio Lopez, President, EMBRAPA**  
**Jean-Marcel Ribuat, Director, GCP**  
**Jeff Ehlers, Program Officer, Gates Foundation**  
**David Hoisington, Director Global Programs, UGA**

**Howard Shapiro, Chief Agricultrual Officer, MARS, Inc**  
**Xun Xu, Deputy Director, BGI-Shenzhen**  
**Steve Brown, Executive Director, TPF**  
**T. Radhakrishnon, Director, DGR, ICAR**  
**Fuhe Luo, Vice Chairman CPPCC & CAPD CC**

## **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:

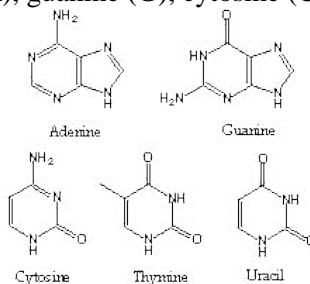
- Function(s) of the protein.
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
- Secondary structure.
- Quaternary structure. 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 recombine 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.

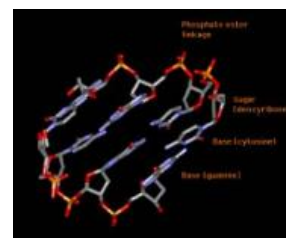
**Conserved sequence:** A sequence within DNA or protein that is consistent across species or has remained unchanged within the species over its evolutionary period.

**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.

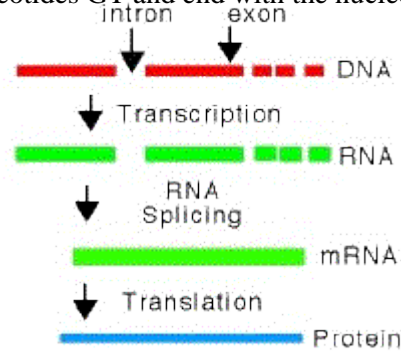
**Genotyping:** The use of markers to organize the genetic information found in individual DNA samples and to measure the variation between such samples.

**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

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

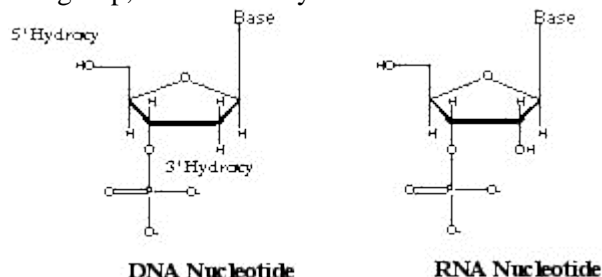
**Motifs:** A pattern of DNA sequence that is similar for genes of similar function. Also a pattern for protein primary structure (sequence motifs) and tertiary structure that is the same across proteins of similar families.

**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.ncbi.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>).

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



**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 copied, 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 is 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.

**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.

**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.