THE PEANUT FOUNDATION

Peanut Genome Initiative

2012-2013 Research Accomplishment Report to the U.S. Peanut Industry

July 31, 2013

Peanut Genome Initiative
2013 Research Accomplishment Report to the U.S. Peanut
July, 2013

Table of Contents

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 naturally focused on dollar value per acre, and peanuts have often been less competitive in yield and production costs as compared to crops such as cotton and corn. As an industry, the best way to compete is by enhancing peanut varieties for disease resistance and yield potential. This can best be accomplished by mapping and assembling the peanut genome to identify genes that confer these desired traits. The goal of genomics is to maximize yield while minimizing inputs in order to sustain production and better compete with other crops. In addition, the genomics effort will enhance the nutritional aspects of peanuts which will increase consumption through marketing. As we grow consumption, we must grow our yield potential to sustain our industry. Genomics is the key to a sustainable future for peanuts.*

Executive Summary

Even though the peanut genome is about the same size as the human genome, it took the Human Genome Project (HGP) 15 years at a cost of over $3 billion to accomplish. By comparison, the Peanut Genome Initiative (PGI) is a 5 year project at a cost of $6 million. The PGI includes 6 components in its strategic plan and is ahead of the timeline in every component at the current writing. The integration of these six research components will create approaches that allow us to better understand and optimize the key traits of the peanut accelerating development of superior seed varieties in a much shorter timeframe.

Following is a summary of the key accomplishments of each component in layman's terms. The remainder of the paper provides technical details of the accomplishments and goals of the six components of the PGI team.

PGI Accomplishments Summary

Component 1 - Whole Genome Sequencing is being done using the variety Tifrunner. When assembled, this sequence will be a baseline for future assembly of other varieties and wild species genomes and will allow comparisons to varieties with desired traits.

- KEY ACCOMPLISHMENT - The two wild parents of the cultivated peanut have been sequenced and essentially assembled. This will aid in the assembly of the more complex Tifrunner cultivar. The information is also being prepared for publication in late 2013.
- Over 150 other lines derived from crosses with Tifrunner and other varieties with desired traits have been sequenced and are being assembled.
- The sequencing of the genome of the Tifrunner is complete and assembly has begun.
- Sequencing is now ready to begin on an additional 1,082 peanut lines:
    - 99 core lines from the Chinese germplasm collection have been purified and are ready for sequencing.
    - 108 mini-core lines from the U.S. germplasm collection have been purified and are ready for sequencing.

- 300 experimental lines representing the genomes of the wild parents of modern cultivars are ready for sequencing in late 2013.

- 575 experimental lines from the ICRISAT Indian germplasm collection.

These 1,082 lines once sequenced will represent 90% of the total genetic variability for all traits that exists in peanut. Dr. Scott Jackson, University of Georgia, and chair of the project technical team remarked, “We're making good progress toward achieving a genome sequence for cultivated peanut, in fact, better progress than I anticipated given the challenges facing us. We are already seeing people begin to use the data to develop additional markers for breeding and to associate traits with the genome sequence.”

<u>Component 2</u> - <u>High-Density Maps, Gene Space</u> have made similar progress. These maps will assist with the assembly of the Tifrunner genome.

- KEY ACCOMPLISHMENT - Over 11,000 markers have been discovered this year and field work is underway to associate these markers with needed traits. This compares to only 5,000 markers identified over the last 5 years.

- 11 different Genetic Maps have been combined into a single consensus for traits which will assist researchers in knowing where to place the sequenced genes in the peanut genome.

Dr. Tom Stalker of NCSU feels, “the populations developed under this component will lay the foundation for molecular marker associations with traits important to all phases of the peanut industry. Selection for most of these traits is difficult using conventional methodologies, and linking markers to traits will enable peanut breeders to more efficiently make selections and, more importantly, to efficiently combine traits into a single breeding line that will lead to commercialization”.

<u>Component 3</u> - <u>Expressed Gene Sequences</u> will identify the genes that are most likely to confer the desired traits.

- KEY ACCOMPLISHMENT – Researchers have determined that the cultivated peanut genome contains over 30,000 unique genes and a size of 2.65 Gb (gigabases = a billion molecules).

- The two wild parents (identified now as A genome and B genome) have over 25,000 unique genes and a size of 1.1 Gb (A genome) and 1.36 Gb (B genome).

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

<u>Component 4</u> - <u>Evaluating State of the Art Sequencing Technologies</u> has helped the PGI group make many discoveries that are reducing the cost and time for sequencing.

- KEY ACCOMPLISHMENT-Using a new technology called Moleculo has allowed the rapid sequencing and assembly of the two wild parents to the cultivated peanut.

- Evaluation of two additional technologies will assist in building the consensus map which will be important to assembling the entire genome of the Tifrunner. Like a road map this places key genes along the reference Tifrunner genome.

"Achieving the peanut genome sequence is akin to landing a man on the moon. Most people thought it would be too difficult and too expensive. New technologies have been developed that help make the challenge attainable," said Dr. Rich Wilson, technical consultant to the Peanut Foundation.

Component 5 - Phenotyping Genetic Resources involves evaluating hundreds of populations of peanuts around the world for both physical and quality traits as well as disease resistant traits over the necessary 3 seasons to adequately challenge the populations to all environmental stresses.

- KEY ACCOMPLISHMENT-Field work began in April with 8 populations that have been planted repeatedly for 7 seasons giving us plants that have segregated for the traits we are seeking. This will give us the ability to associate the markers discovered in components 2 and 3 with the resistant and quality traits we need.

- Another 8 populations of segregating populations are in the final year of seed increases and will be planted in 2014 for field evaluation.

- In India, researchers with ICRISAT have completed 6 years of field evaluations for drought tolerance and that data is being uploaded to the data center at USDA.

- A standardized system for evaluating these plants has been developed and is now available as a software package for the field and lab evaluators.

Corley Holbrook, USDA-ARS, who is leading this effort, recently talked about how this association of gene markers to traits will affect his breeding program, stating "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 marker assisted selection (MAS) in peanut is in its infancy, it has already had a tremendous impact on my breeding program".

Component 6 - Bioinformatics Resources is a database for the collection, storage and easy usage of this data by breeders. This effort is being led by USDA-ARS in Ames, Iowa and the National Center for Genome Resources (NCGR) in Santa Fe, NM. The web database will contain all the data and a Breeder's Toolbox to assist peanut breeders in selecting gene markers and varieties (whether wild or cultivated) to be used in their breeding programs.

- KEY ACCOMPLISHMENT-The system is already live and ready to accept the data from all the other components as it is generated.

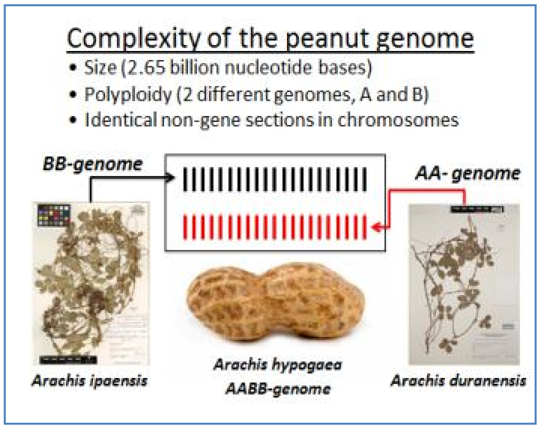
Mark Burow, peanut breeder at Texas A&M says, "Web-based genome libraries and databases of markers will help breeders find specificmarkers that can be used in their breeding populations. The Breeder's Toolbox will allow breeders to merge genomics and phenotypic information to use in marker-assisted breeding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

In conclusion, this project has made great strides in 2012-2013. Howard Shapiro, Director of Research for Mars Chocolate, who has been involved in more than 90 plant genomic projects states, "This project has really made great progress in its first year. I expected a lot with the team of researchers assembled but they have exceeded my expectations".

Even though we are ahead of schedule on every component, we still understand the urgency of completing this work as soon as possible. With your continuing financial support and the emergence of even better technologies, we are certain our industry will become more competitive.

Peanut Genome Initiative
2013 Research Technical Accomplishment Report to the U.S. Peanut Industry
July, 2013

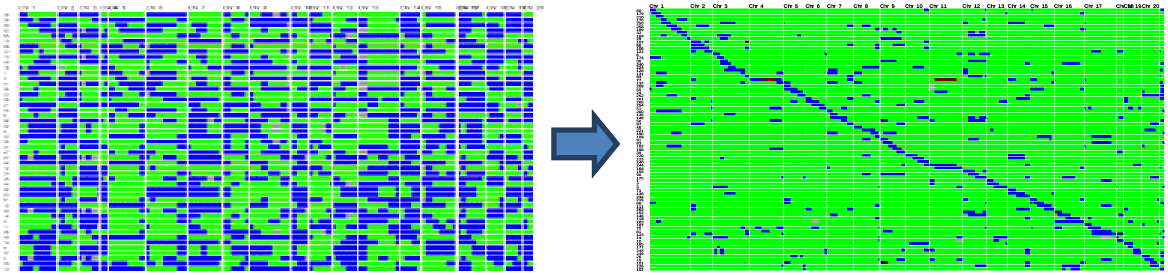
Introduction



Complexity of the peanut genome

- Size (2.65 billion nucleotide bases)
- Polyploidy (2 different genomes, A and B)
- Identical non-gene sections in chromosomes

Scott Jackson (University of Georgia and Senior Co-Chair of the Peanut Genome Consortium) says, “The peanut genome initiative is an ambitious endeavor to look deeply inside the cultivated peanut at the DNA level and discover all the genes that control crop productivity and quality.” This quest is not a simple matter. The cultivated peanut genome 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 wild species, a merger that occurred about 4000 years ago. The challenge is to correctly place sequences in the right genome. 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. For example, the panels below show maps of DNA sequence fragments before and after the successful assembly of a chromosome.



Scientists in the Peanut Genome Initiative 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 detailed and technical description of the process and technologies used in DNA sequencing is presented in Exhibit 1; and a glossary of ‘genomic’ terms & definitions is presented in Exhibit 2.

Component 1: Whole Genome Sequencing.

This research established the variety Tifrunner as the U.S. industry standard for the reference peanut genome which will be used as a baseline comparison for genetic differences in sequence structure for specific agronomic traits in selected cultivated and wild peanut germplasm. Toward this goal the following questions were answered in 2013.

<u>What materials were selected for DNA sequencing and why</u>?

Cultivated peanut (Arachis hypogea)

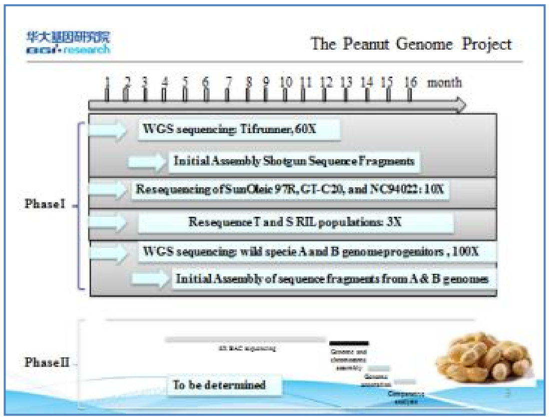
- The cv Tifrunner was selected as the best representative modern variety to provide a 'reference' standard for characterization of genome structure in other cultivated peanut germplasm. Tifrunner exhibits: normal-oleic, TSWV resistance, early leaf spot traits.
- Three other varieties with important 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*

<u>What has been learned from the sequence so far?</u>

BGI-Shenzhen, China a collaborating partner in the PGP is responsible for developing the whole genome shotgun sequences from Tifrunner, GT-C20, SunOleic 97R, NC91022, RIL populations from the T and S populations, and whole genome shotgun sequences from *A. duranensis* (AA-genome) and *A. ipaensis* (BB-genome). This work has been completed. An initial assembly (a basic map of all DNA fragments) from



The Peanut Genome Project

WGS sequencing: Tifrunner, 60X

Initial Assembly Shotgun Sequence Fragments

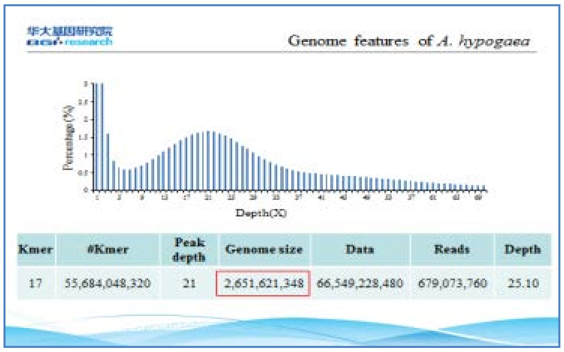
Resequencing of SunOleic 97R, GT-C20, and NC94022: 10X

Resequence T and S RIL populations: 3X

WGS sequencing: wild specie A and B genomeprogenitors , 100X

Initial Assembly of sequence fragments from A & B genomes

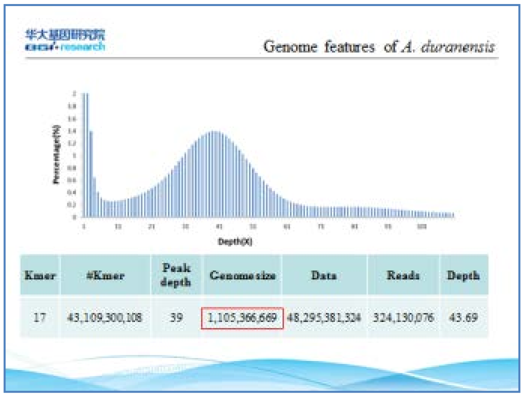
To be determined


Genome features of *A. hypogaea*

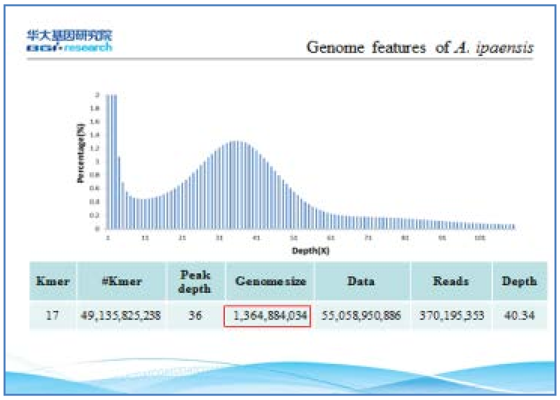
Kmer	#Kmer	Peak depth	Genome size	Data	Reads	Depth
17	55,684,048,320	21	2,651,621,348	66,549,228,480	679,073,760	25.10

Tifrunner and the AA- and BB- wild species genomes also has been constructed, but these assemblies are far from finished. This work constitutes Phase I of the BGI contract. These data are being evaluated to determine the best strategies and technologies for developing a high quality, annotated chromosome-based assembly of the cultivated peanut genome. Thus far Phase I results show the genome size of the Tifrunner (*A. hypogea*) genome is about 2.65 Gb (based on 312.7 Gb data at 111.7X coverage). There was no obvious sign of heterozygosity (a good thing, meaning the DNA samples were taken from genetically pure lines), but the fat tail in the Figure (on the right) indicates that this genome may contain a high content of repeating sequences (a potential problem). The initial assembly produced 232000 contigs that were longer than 2 kb and about 60000 scaffolds. This is a good start, but far from good enough. Several new DNA sequencing technologies including: BAC x BAC, Moleculo™, Long-Fragment-Reads, and Pacific Biosciences™ will be evaluated to determine the best approach for Phase II of the BGI contract. Detailed information on these technologies may be found in the Appendices to this report.

DNA sequence characterization of the wild species genomes is another approach to accrue information that may help assemble the cultivated peanut (*A. hypogea*) genome. Phase I results show the size of the AA-genome (*A. duranensis*) is 1.1 Gb (based on 216 Gb data at 154.4X coverage). As indicated in the Figure (lower left) there was no obvious heterozygosity peak, but the fat tail indicates that this genome may also contain high repeat content. The initial AA-genome assembly exhibited 53000 contigs and 7400 scaffolds greater than 2 kb in length. The genome size of the BB-genome (*A. ipaensis*) was estimated to be 1.36 Gb (based on 168.8 Gb data at 120.6X coverage. There was no obvious heterozygosity peak, but again the fat tail indicated that this genome might contain high repeat content (see Figure lower right). The initial BB-genome assembly had 127000 contigs and 13000 scaffolds greater than 2 kb in length. Although repeating sequences may complicate assembly of each of the wild diploid genomes, the final product should be quite good, and useful in helping to distinguish AA- from BB-genome sequences in cultivated peanut.


Genome features of *A. duranensis*

Kmer	#Kmer	Peak depth	Genome size	Data	Reads	Depth
17	43,109,300,108	39	1,105,366,669	48,295,381,324	324,130,076	43.69


Genome features of *A. ipaensis*

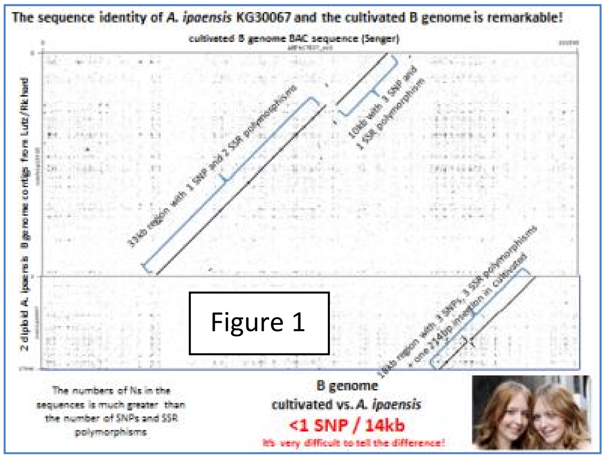
Kmer	#Kmer	Peak depth	Genome size	Data	Reads	Depth
17	49,135,825,238	36	1,364,884,034	55,058,950,886	370,195,353	40.34

Component 2: High-Density Genetic Maps, Gene Space.

This research enables the identifcation of thousands of gene markers in wild and cultivated peanuts. A- and B- genome gene markers found in diploid progenitor species are needed to identify their counterparts in the culitvated (tetraploid) genome assembly. Gene markers for agronomic traits help breeders accelerate superior peanut variety development. Toward these goals the following critical questions were answered in 2013.

<u>How well do the AA- and BB-genome assemblies match with their counterparts in cultivated peanut?</u>

Scientists at the Catholic University Brasilia, University of California-Davis, University of Georgia-Athens & Tifton, NC State University, Kazusa DNA Research Institute-Japan, and EMBRAPA have answered this quesiton by using SNP analysis to map DNA fragments from a B-genome BAC of cultivated peanut, against two diploid contigs from the BB-genome (*A. ipaensis*). As shown below in Figure 1, the sequence identity between the two genomes is remarkably similar, having only 1 SNP mutation every 14 kb of sequence. This indicates that the complete BB-genome sequence from *A. ipaensis* should be extremely useful in filtering out the B-genome of cultivated peanut. In a similar comparison of wild and cultivated AA-genome BACs (Figure 2), the sequence identity between *A. duranensis* and the cultivated A-genome is less striking, at 1 SNP mutation/75 bp, good enough to still distinguished the species of origin.

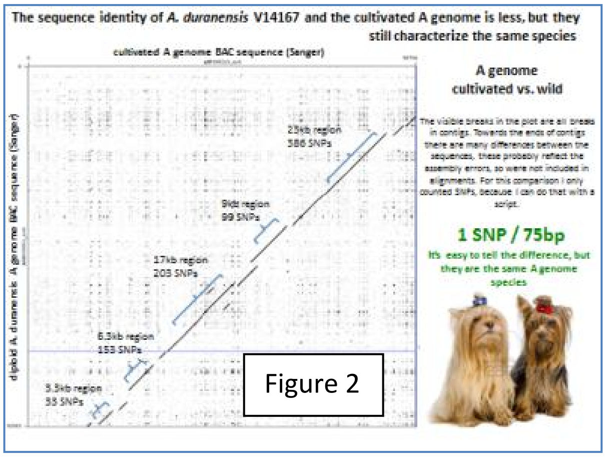


The sequence identity of *A. ipaensis* KG30067 and the cultivated B genome is remarkable!

The numbers of Ns in the sequences is much greater than the number of SNPs and SSR polymorphisms

B genome
cultivated vs. *A. ipaensis*
<1 SNP / 14kb
It's very difficult to tell the difference!

Figure 1



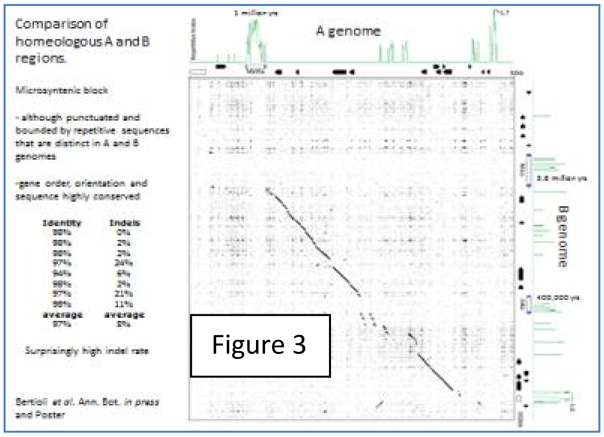
The sequence identity of *A. duranensis* V14167 and the cultivated A genome is less, but they still characterize the same species

A genome
cultivated vs. wild

The visible breaks in the plot are all breaks in contigs. Towards the ends of contigs there are many differences between the sequences, these probably reflect the assembly errors, so were not included in alignments. For this comparison I only counted SNPs, because I can do that with a script.

1 SNP / 75bp
It's easy to tell the difference, but they are the same A genome species

Figure 2



Comparison of homeologous A and B regions.

Microsyntenic block

- although punctuated and bounded by repetitive sequences that are distinct in A and B genomes

-gene order, orientation and sequence highly conserved

Identity	Indels
98%	0%
98%	2%
98%	2%
97%	24%
94%	6%
98%	2%
97%	21%
98%	11%
average 97%	average 8%

Surprisingly high indel rate

Bertioli *et al.* Ann. Bot. *in press* and Poster

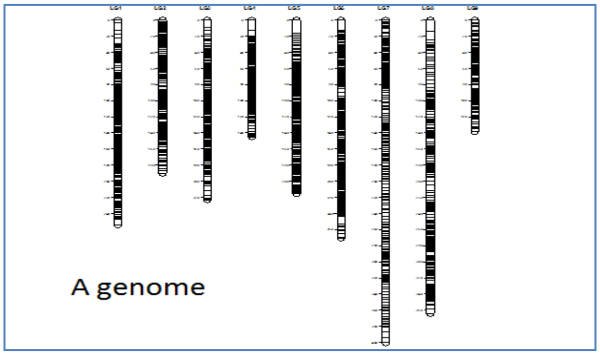
Figure 3

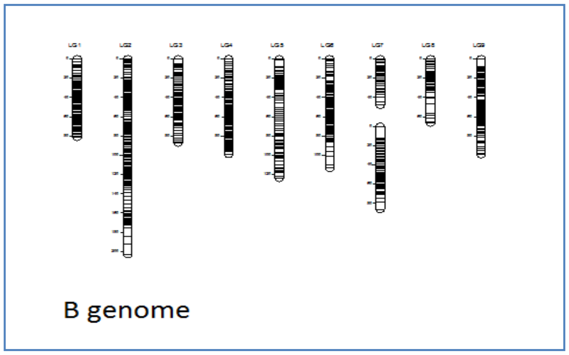
In addition, 30 to 37% of SNPs from AA-genome RIL populations (*A. duranensis x A. stenosperma*) were equivalent to SNPs in AABB-genome populations from [(*A. duranensis x A. stenosperma*) x (*A. ipaensis x A. magna*)] x *A hypogea*; and a comparison of homologous A and B regions indicated that gene order, orientation and sequence were highly conserved between those genomes. Thus, more SNPs will enable the utility of SNP assays in developing high density genetic maps because a majority of SNPs appear to be unique to a given sub-genome. This will be a powerful tool for cultivated genome assembly.

What portion of the AA- and BB-genomes actually contain genes?

Scientists at the University of California-Davis, Catholic University Brasilia, EMBRAPA, and ICRISAT have used a technique known as ‘genotyping by sequencing’ to develop ultra-dense genetic maps of the AA- (*A. duranensis*) and BB- (*A. ipaensis*) genomes with SNPs discovered in gene-rich regions on each chromosome. This work was facilitated by mate-pair data from 5 kb and 10 kb insert libraries from *A. duranensis*-V14167 (provided by collaborators at BGI). The assembly analysis revealed that the ‘gene-space’ occupied 83.5% of the BB-genome and 65.2% of the AA-genome.

A genotype-by-sequencing pipeline was built that was capable of mapping 99% of the filtered contigs (>10 kb) with 2,000,000 high-quality SNPs from *A. duranensis*-V14167. The largest scaffold without a SNP was 46 kb. Overall, the mean point mutation density was 2.1 high-quality SNPs/kb. At this stage 6130 SNP scaffolds have been positioned on a high-density genetic map of the AA-genome (below-left), covering a total of 1930 cM or 40% of the genome. In a similar analysis, 5941 BB-genome scafflolds were used to map 40% or 1115 cM of the BB-genome (below-right). These maps of DNA markers within the gene-space of the wild peanut progenitors will not only facilitate the tetraploid genome assembly by anchoring genomic scaffolds together, but also pinpoint the location of ~~a~~ useful genes in breeding populations.



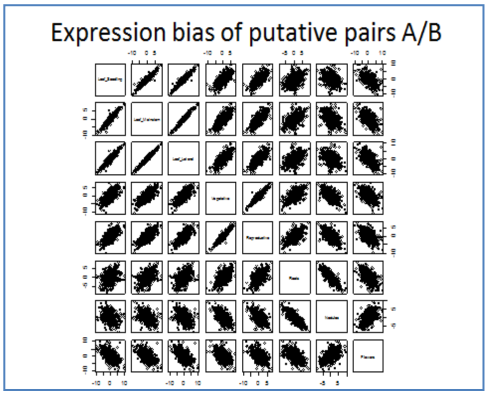


Component 3: Expressed Gene Sequences

This research helps characterize the sequence of genes that are made (expressed) in the cultivar Tifurnner during plant development, in various plant organs, and under different environmental conditions to help identify genes that mediate a given trait. Such knowledge 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. Toward these goals, the following critical questions were answered in 2013.

<u>What genes will be targeted?</u>

Scientists at the Universtiy of Georgia-Tifton and Athens, USDA ARS at Tifton GA and Stoneville MS, and the Universtiy of California-Davis have plans to target genes in cultivated peanut that mediate resistance to diseases and pests such as: tomato spotted wilt virus (TSWV), leaf spot [early (*Cercospora arachidicola*) and late (*Cercosporidium personatum*)], rust (*Puccinia arachidis*); white mold (*Sclerotium rolfsii*); nematode (*Meloidogyne arenaria*); and pre-harvest aflatoxin contamination (*Aspergillus flavus*). In addition, plans include genes that mediate tolerance to abiotic stresses (drought, temperature, nutrient deficiency). Eventually, their goal is to construct a peanut gene atlas which includes a comprehensive list of all expressed genes, alternatively spliced products, co-regulated genes and gene networks. A method called RNA-Seq will be used to discover expressed genes that are turned on or off in the presence of a given treatment. Data will be analyzed from replicated sequencing of RNA taken from eight plant organs (leaf at three stages of plant development; vegetative, pods/seed, roots, nodules and flowers) of the cv Tifrunner (at this point work is focussed on generating base-line data for the control treatment). Eventually differential gene expression profiles will be developed from Tifrunner plants exposed to a given treatment to zero-in on genes that are associated with a given trait. Currently, twenty-four libraries (8 organs, 3 reps, control treatment) were sequenced by Ilumina Hi-Seq, and the transriptomes were assembled with Trinity™ or CLCBio™ software. These data were compared to similar analyses of *A. duranensis* and *A. ipaenea* transcriptomes. Within the Tifrunner transcriptome assembly, ca. 7500 expressed genes had only two transcripts. Each of the pairs could be assigned to the A- or B-genome based on Blast hits to a complementary *A. duranensis* or *A. ipaenea* transcript (as indicated by the posititive or negative slopes in the expression panels below). This suggests it will be possible to associate alleles for genes in Tifrunner with the A- or B-genome of cultivated peanuts (Figure on left).

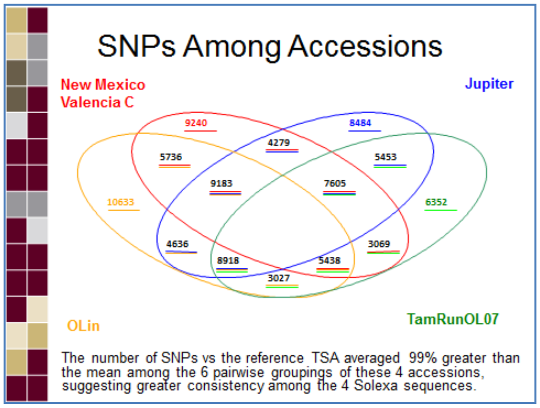


If the sequence of a gene of interest is known (derived from genome databases), a method called TILLInG may be used to find mutations in those genes. Scientist at the University of Georga-Tifon have taken this concept a step farther by deploying 'TILLInG by Sequencing' platforms for sequencing wild-type and mutant amplicons (PCR amplified DNA) for targeted genes (experimental selections are: lipoxygenase, phospholipase D1, Ara h 1.01, Ara h 1.02, Ara h 2.01, Ara h 2.02, fatty acid desaturase 2 from the A-genome, and fatty acid desaturase 2 from the B-genome). Twenty alleleic mutations were identified for an overall mutation rate ~1 SNP/967 kb; and nucleotide substitutions (knockouts) were found in one gene each for Ara h 1, Ara h 2, and FAD2. A total of 13 new mutants plus 6 known mutants were identified and validated using either Cleaved Amplified Polymorphic Sequences (CAPS), Sanger sequencing or Single Strand Conformation

Polymorphism (SSCP). Observed expression of the phenotype of plants endowed with these mutations confers knowledge of gene function.

Are SNP makers in cultivated peanut the same in all market types?

U.S. peanut production involves four U.S. market types. However, little is known about the SNP distribution among market types. The subject question was addressed by scientists at Texas Tech University-Lubbock, National Center for Genome Resources-Santa Fe NM, USDA-ARS-Lubbock TX, Texas A&M AgriLife Research-Stephenville TX, and Texas A&M AgriLife Research-Lubbock TX. Tissue was collected from four Southwestern U.S. peanut cultivars: OLin (Spanish), TamrunOL07 (runner), Jupiter (Virginia) and New Mexico Valencia C (Valencia). Subsequently, 10 wild species and eight additonal cultivars also were sampled. RNA was isolated from leaf, root and pod tissues and sequenced by Illumina GAIIx at NCGR. These data were compared to reference Tifrunner transcriptome sequences provided by the University of Georgia. Approximately 40000 contigs were found in each market type, with up to 10000 SNP compared to the Tifrunner reference. As shown (Figure-right) certain combinations of SNPs can be used to distinguish the genotype of market-types of cultivated peanut.


SNPs Among Accessions

The number of SNPs vs the reference TSA averaged 99% greater than the mean among the 6 pairwise groupings of these 4 accessions, suggesting greater consistency among the 4 Solexa sequences.

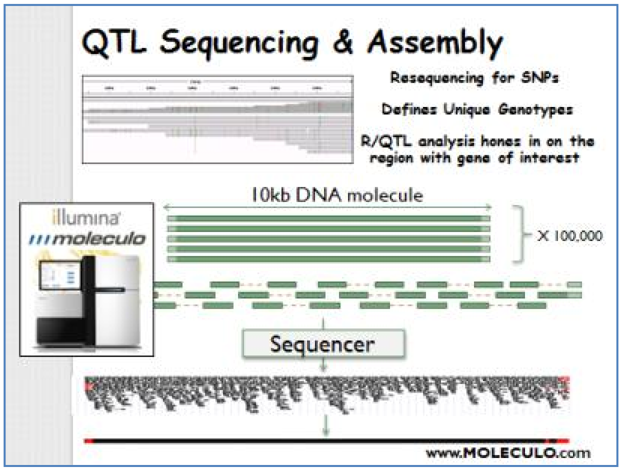
Denovo assembly with Trinity™ software was used to identify genes and transcripts in all 22 accessions by Hi-Seq. Results showed that all types of cultivated peanut contained 30000 to 32000 unique (A and B-genome copies were considered as the same) genes in tetraploid accessions, with 25000 to 28000 unique genes in the wild species. Work will continue to identify SNPs between pairs of accessions, determine function, and develop useful SNPs for genetic mapping of several populations.

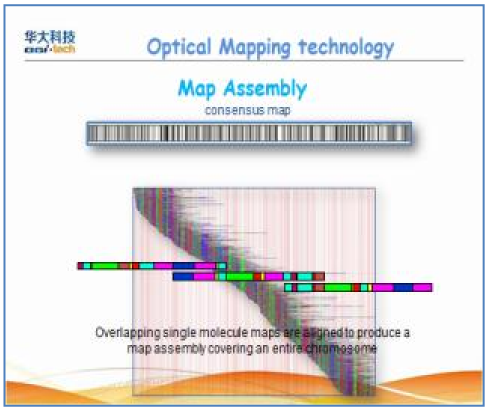
Component 4: Evaluation of State-of-Art Sequencing Technologies

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. Toward this goal, the following critical questions were answered in 2013.

What is being done to improve genetic maps and genome assembly?

Although a BAC x BAC approach is a common method to aid high-quality genome assembly in many complex genomes, the process is labor intensive and expensive. Therefore scientists at the University of California-Davis and BGI are exploring alternative methods that may have advantages to BAC x BAC. At this time, data is available from the evaluation of Moleculo™ technology. Illumina-Moleculo™ is an innovative technology for generating extremely accurate long reads from short reads regardless of genome complexity. Moleculo technology begins by fragmenting genomic DNA to approximately 10 kb fragments. These fragments are clonally amplified, sheared, and marked with a unique barcode. They are then sequenced with Illumina Hi-Seq technology. The short sequence reads originating from each molecule are assembled separately. The end result is a full sequence of all the fragments. Essentially, short-read data is reconstructed into long reads with near perfect quality. Current results from studies of the BB-genome improve contig assembly (>10 kb) by 3.5-fold compared to controls. This covers 83.5% of the genome. However, many gaps still remain to be filled between contigs. These gaps are attributed to the apparent inability of the current Moleculo assembler to recognize DNA-repeats that are rich in A and T nucleotides. Work is in progress to correct this deficiency, and to examine complementary strategies such as Pac Bio™, RAD-Seq, LFR, and optical mapping technologies as ways to move forward.



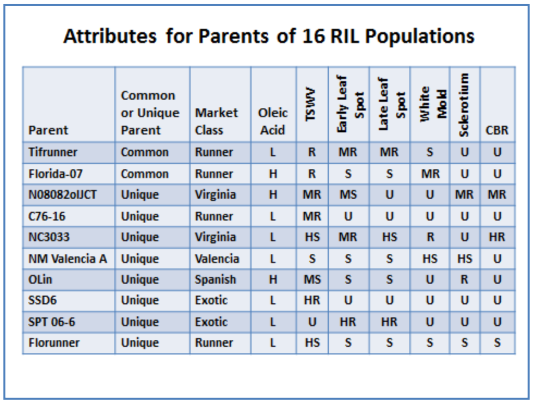

华大科技 BGI-tech
Optical Mapping technology
Map Assembly
consensus map
Overlapping single molecule maps are aligned to produce a map assembly covering an entire chromosome

Optical mapping conducted at BGI is a technique for constructing ordered, genome-wide, high-resolution restriction maps from single, stained molecules of DNA, called "optical maps". By mapping the location of restriction enzyme sites along the unknown DNA of an organism, the spectrum of resulting DNA fragments collectively serve as a unique "fingerprint" or "barcode" for that sequence. Alignment of all the bar-coded fragments produces a consensus map that should cover the entire genome. More tests of these technologies are needed, but it is clear that options are available to overcome problems with the complex structure of the cultivated peanut genome.

Component 5: Phenotyping Genetic Resources

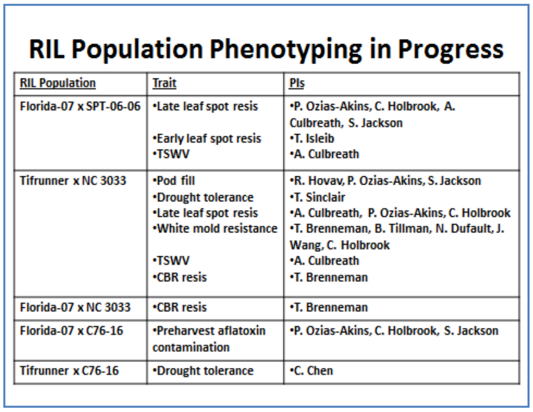
This research is essential for making the peanut genome sequence and genomic tools useful to breeders because it makes the connection between genes, gene markers, genetic maps, and agronomic traits in peanut. The peanut genome initiative is ahead of many other crop genome projects because of the attention that is being given to phenotyping. Toward that goal, the following critical question was answered in 2013.

<u>What is being done to associate SNP markers with important traits in cultivated peanut?</u>



Attributes for Parents of 16 RIL Populations

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



RIL Population Phenotyping in Progress

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

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

Component 6: Bioinformatic Resources

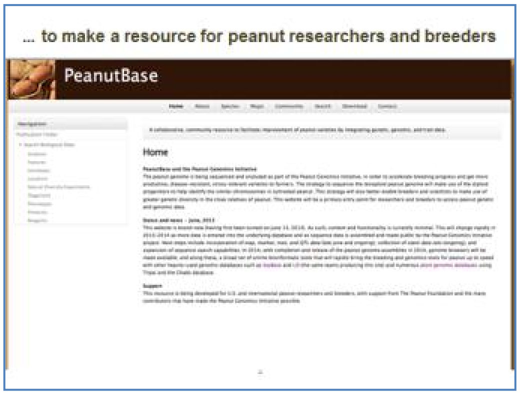
This research creates a secure internet-based home for all data generated by the peanut genome initiative. This website will also feature software to help make these data useful to breeders. Toward that goal, the following question was answered in 2013.

<u>How will all of these data and tools be stored and made available to breeders?</u>

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

- A convenient way to access datasets such as: maps, transcriptome, SNPs, RNA-seq, Genome assembly, annotations, etc.
- Links to research community, genetic and genomic resources for peanut, legumes and other external sources
- Biological information on peanut and relatives to provide context
- Map, trait, and QTL information with methods for data collection & database loading integrated with other resources
- Ability to query and view QTL positions on genetic maps in various formats (this section now has 199 QTL, 12,500 markers).
- Genome browsers served from a 48-core machine which are very responsive, scalable, well integrated
- Integration of multiple maps by placing sequence-based markers onto the genome sequence(s), producing a "virtual/physical" map. These will be tied to QTLs, and to the browsers.
- A collection of peanut phenotype descriptors with relation to other ontologies
- Tools for using haplotype, accession, & phenotype data for breeding
- Gene family & orthology tools, and gene functional information; from a project to catalog and describe the phenotypes of genes with characterized mutations, using ontologies, from Arabidopsis, soybean, maize, rice, Medicago, tomato; map these into all gene families
- Training, outreach, and coordination
- Sequence and key-word search tools

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



… to make a resource for peanut researchers and breeders

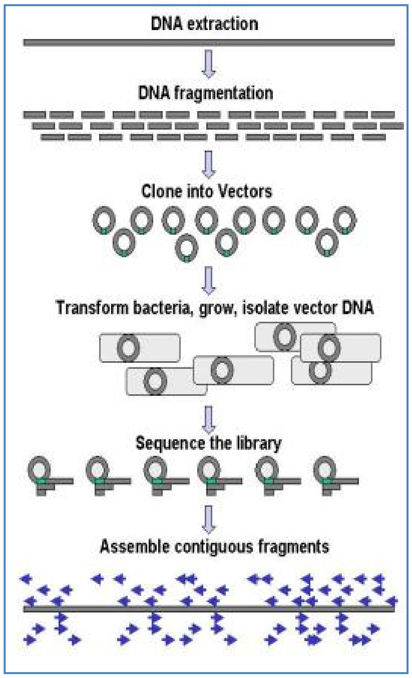
PeanutBase

Appendices

Exhibit 1: Genome Sequencing Technologies

Abridged from: Pareek, Smoczynski, and Tretyn (2011) J Appl Genetics 52:413–435; and www.Wikipedia.ort/wiki/Genome_Sequencers

Introduction: Advances continue to be made in DNA sequencing technology and bioinformatic tools at an astonishing rate. High-throughput next generation sequencing (HT-NGS) technologies can produce over 100 times more data compared to capillary sequencers based on the Sanger method. Since 2005, HT-NGS technologies have continued to improve and enable whole genome genotyping, genome wide structural variation, de novo assembling and re-assembling of genomes, mutation detection and association to genetic traits of interest. This brief overview is by no means exhaustive, but provides background for the dynamic development of this technology that should convey that genome sequencing will become a routine step in genetic improvement of biological organisms.


DNA extraction
DNA fragmentation
Clone into Vectors
Transform bacteria, grow, isolate vector DNA
Sequence the library
Assemble contiguous fragments

Traditional Sanger Sequencing: Genomic DNA is fragmented into random pieces and cloned to form a bacterial library. DNA from individual bacterial clones (BAC) are sequenced and assembled by overlapping DNA regions. Large-scale sequencing such as shot-gun methods for chromosomes commonly entail cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA (sequences longer than 1000 base pairs up to entire chromosomes) may then be cloned into a DNA vector and amplified in a bacterial host such as Escherichia coli. Randomly sequenced DNA fragments from BACs are reassembled on the basis of their overlapping regions into longer contiguous sequences (contigs) and in comparison to a reference (previously sequenced standard) genome. De novo sequencing applies when there is no previously known sequence. Selection of a sequencing technology depends on the complexity of the genome (level of ploidy and sequence repeat content). Most sequencing approaches use an in vitro cloning step to amplify individual DNA molecules above a detection threshold. Individual DNA molecules are linked to primer-coated beads in aqueous droplets within an oil phase. A polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in sequencing with 454 (Life Sciences) and SOLiD sequencing (Applied Biosystems, now Life Technologies). Bridge PCR is a method for in vitro clonal amplification where fragments are amplified upon primers attached to a solid surface to form DNA clusters. This method is used in Illumina Genome Analyzer sequencers.

Comparison of next-generation sequencing methods

Method	**SMRT sequencing (Pacific Bio)**	**Ion semiconductor (Ion Torrent)**	**Pyrosequencing (454)**	**Sequencing by synthesis (Illumina HiSeq)**	**Sequencing by ligation (SOLiD)**	**Chain termination (Sanger)**
Read length	2900 bp	200 bp	700 bp	50 to 250 bp	50 bp	400 to 900 bp
Reads per run	35–75,000	up to 5 million	1 million	up to 3 billion	1.4 billion	N/A
Time per run	30 min to 2 hrs	2 hrs	24 hrs	1 to 10 days	1 to 2 weeks	20 min to 3 hrs
Cost / Mbase	$2	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Long read length. Fast.	Less expensive equipment. Fast.	Long read size. Fast.	High sequence yield	Low cost per base.	Long individual reads.
Disadvantages	Low yield, Equipment expense	Homopolymer errors.	Expensive runs. Homopolymer errors.	Equipment expensive.	Slow method	Expense, impractical for larger projects

However, genomes with large non-coding (repeats) regions are difficult to assemble from short reads. New methods of tagging ends of DNA with unique barcodes has helped overcome this problem, and should help achieve genome sequences for crops like wheat, sunflower, canola, palm, peanut and cotton.

Reference:
First generation DNA sequencers. The first automated DNA sequencers that were commercialized by Applied Biosystems (ABI), the European Molecular Biology Laboratory (EMBL) and Pharmacia-Amersham. These sequencing methods were based on concepts introduced in 1975 by Sanger (an enzymatic technique based on chain-terminating dideoxynucleotide analogues); and Maxam & Gilbert (terminally labeled DNA fragments chemically cleaved at specific bases were separated by gel electrophoresis). In 1996, ABI introduced the first commercial DNA sequencer (ABI Prism 310) that utilized slab gel electrophoresis. In 1998, slab gels were replaced with 96 capillaries with polymer matrix (ABI Prism 3700). These technologies were used to sequence the human genome at an estimated cost of $2.7 billion. Modifications of the basic ‘dideoxy’ method continued until about year 2005.

Next-Generation Sequencing Technologies (NGS): In 2005, 454 Life Sciences commercialized the first NGS platform (GS 20) which combined single-molecule emulsion PCR with pyrosequencing. Roche Applied Science acquired 454 Life Sciences and marketed the GS FLX titanium platform with fused fiber-optic bundle flow cells. Pyrosequencing depends on detection of pyrophosphate released on nucleotide incorporation, rather than chain termination with dideoxynucleotides. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead. Reactions are carried out in picoliter-volume wells containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA with a sensitive CCD camera. After each reaction the nucleotide and the dye were removed, and the cycle was repeated to determine the next base in the sequence.

Second generation HT-NGS platforms: The second generation high-throughput (HT-NGS) platforms can generate 500 million bases of raw sequence (Roche) to billions of bases in a single run (Illumina, SOLiD). These novel methods rely on parallel, cyclic interrogation of sequences from spatially separated clonal amplicons (26 µm oil-aqueous emulsion bead in Roche pyrosequencing chemistry, 1 µm clonal bead in SOLiD: sequencing by sequential ligation of oligonucleotide probes, and clonal bridges in Illumina: sequencing by reversible dye terminators.

Illumina (Solexa™): A method based on reversible dye-terminators technology, and engineered polymerases. The In 2004, Solexa acquired the company Manteia Predictive Medicine in order to gain a massively parallel sequencing technology based on "DNA Clusters", which involves the clonal amplification of DNA on a surface. The cluster technology was co-acquired with Lynx Therapeutics of California. Solexa Ltd. later merged with Lynx to form Solexa Inc.

Illumina MiSeq™, HiSeq™, TruSeq™: In this method, DNA molecules and primers are first attached on a slide and amplified with polymerase to form clonal DNA clusters. Four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labeled nucleotides then the dye along with the terminal 3' blocker is chemically removed from the DNA to begin the next cycle. Unlike pyrosequencing, the DNA chains are extended one nucleotide at a time and image acquisition can be performed at any moment, allowing formation of very large arrays of DNA clusters to be captured in sequential images. Decoupling the enzymatic reaction and the image capture allows for optimal throughput. MiSeq output may reach 20 gigabases with 2x400 base-pair reads. HiSeq output may achieve 600 gigabases using 2x100 base pair reads. TruSeq is a PCR-free sample prep method for sequencing cDNA or gDNA fragments by HiSeq. PCR amplification revolutionized DNA analysis, but often introduces base sequence errors or doesn’t have the same affinity for all sequences which can generate false-positive reads. In TruSeq adapter bases added to blunt end of each fragment for ligation to sequencing base adapters that cover the full complement of primer hybridization sites, thus eliminating need for additional PCR. .

Life Technologies SOLiD™: Although currently discontinued, SOLiD technology employed sequencing by ligation. A pool of all possible oligonucleotides of a fixed length are labeled according to sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting beads, each containing single copies of the same DNA molecule, are deposited on a glass slide. Sequence output and lengths are comparable to Illumina systems.

Third to Nth generation HT-NGS platforms: These technologies strive to reduce sequencing errors and improve sequence quality from complex genomes. They range from generation of sequence information directly from a single DNA molecule (**sequencing-by synthesis**) which eliminates PCR amplification steps to sophisticated methods for tagging unique sequences of any length.

Heliscope™. Introduced in 2007, the principle relies on true single molecule sequencing (tSMS) technology which begins with DNA library preparation through DNA shearing and addition of poli-(A) tail to generated DNA fragments, followed by hybridization of DNA fragments to the poli-(T) oligonucleotides which are attached to the flow cell and simultaneously sequenced in parallel reactions. The sequencing cycle consists of DNA extension with one, out of four fluorescently labeled nucleotides, followed by nucleotide detection. The subsequent chemical cleavage of fluorophores allows the next cycle of DNA elongation to begin with another fluorescently labeled nucleotide. Heliscope sequencer output is up to 28 Gb in a single sequencing run in 8 days with maximal read length of 55 bases. However the company is reportedly bankrupt at this time and this technology was discontinued two years ago...

SMRT™: Single Molecule Real Time sequences single molecules in real time by synthesis on a sequencing chip containing thousands of zero-mode waveguides (ZMWs). The sequencing reaction is performed by a single DNA polymerase molecule attached to each ZMW. During the reaction, the DNA fragment is elongated by DNA polymerase with dNTPs that are fluorescently labeled with a different fluorophore at the terminal phosphate moiety. The DNA sequence is determined by fluorescence nucleotide detection before nucleotide incorporation. The fluorescence is stopped by phosphodiester bond formation and release of the fluorophore. The SMRT sequencer is a product of **Pacific Biosciences**. SMRT analyzers are capable of obtaining 100 Gb per hour with reads up to 2900 bp in a single run.

Nanopore™: DNA sequencing with Nanopore technology relies on the conversion of electrical signals generated by nucleotides passing through a nanopore covalently attached to a cyclodextrin binding site for nucleotides. The principle of this technique is based on the modulation of the ionic current through the pore as a DNA molecule traverses it, revealing characteristics and parameters (diameter, length and conformation) of the molecule. During the sequencing process the ionic current that passes through the nanopore is blocked by the nucleotide previously cleaved by exonuclease from a DNA strand that interacts with cyclodextrin. The time period of current block is characteristic for each base and enables the DNA sequence to be determined. Although no products are available at this time, Oxford Nanopore Technologies Ltd may produce disposable micro-sequences powered by multicore processors and customized sequence clusters (MiniION™) which can sequence whole genomes in minutes, have the size of a flash-drive.

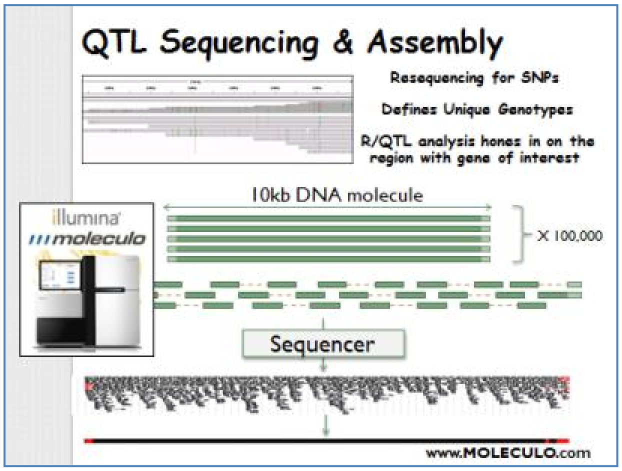
Ion Torrent™: Ion Torrent (Life Technologies) using standard sequencing chemistry, but with a novel, semiconductor based detection system. This method of sequencing is based on the detection of hydrogen ions that are released during the polymerization of DNA, as opposed to the optical methods used in other sequencing systems. A microwell containing a template DNA strand to be sequenced is flooded with a single type of nucleotide. If the introduced nucleotide is complementary to the leading template nucleotide it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence multiple nucleotides will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.

RNASeq™: HT-NGS is also finding application in the study of differentially expressed miRNA binding sites. However, the most frequent application of RNA-Seq on Illumina, SOLiD or 454 platforms in whole transcriptome shotgun sequencing (WTSS). HT-NGS applications include sequencing cDNA to characterize: differentially expressed genes, gene alleles, spliced transcrips, non-coding RNAs, post-transcriptional mutations and gene fusions.

RADSeq™: Restriction site Associated DNA is a method for sampling genomes of multiple individuals in a population by HT-NGS technology. It involves cutting a genome with at least one restriction enzyme and sequencing the ends of the fragments. Each fragment is tagged with a unique identifier sequence which sorts sequences derived from pooled fragments. This technique helps identify genotypic as well as phenotypic differences in genomes from diverse or related populations of biological organisms.

Complete Genomics (now part of BGI) developed ‘Long Fragment Reads’ (LFR) which is achieved by the stochastic separation of corresponding long parental DNA fragments into physically distinct pools followed by subsequent fragmentation to generate shorter sequencing templates. The same principles are used in aliquoting fosmid clones. As the fraction of the genome in each pool decreases to less than a haploid genome, the statistical likelihood of having a corresponding fragment from both parental chromosomes in the same pool is markedly diminishes. The end result is a roughly 95% overall chance that a sequences in each of the 384 reaction wells will be unique.

Illumina-Moleculo™: An innovative technology for generating extremely accurate long reads from short reads regardless of genome complexity. Moleculo technology begins by fragmenting genomic DNA to approximately 10 kb fragments. These fragments are clonally amplified, sheared, and marked with a unique barcode. They are then sequenced with Illumina technology. The short sequence reads originating from each molecule are assembled separately. The end result can be a full sequence of all the fragments. Essentially, short-read data is reconstructed into long reads with near perfect quality.



QTL Sequencing & Assembly

Resequencing for SNPs

Defines Unique Genotypes

R/QTL analysis hones in on the region with gene of interest

Tile-Seq™: is similar to Moleculo technology for assembling long DNA sequences up to 3000 bp. However, uniquely bar coded (tagged) amplicons are treated with an exonuclease and additional tags are added at intervals to the other end of the molecule. This creates a population of tagged different length fragments of an amplicon. Sequences are assembled from the overlapping reads.

Exhibit 2: Terms and Definitions in Bioinformatics

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:

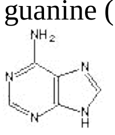
- Function(s) of the protein.
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
- Secondary structure.
- Quaternary structure.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

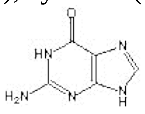
Association Mapping: As in QTL mapping, the goal of association mapping is to find a statistical association between genetic markers and a quantitative trait. However, in association mapping, the genetic markers usually must lie relatively close to a candidate gene. The goal is to identify the actual 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.

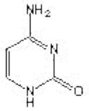
Base: One of five molecules which are assembled, along with a ribose and a phosphate, to form nucleotides (Figure 1). Adenine (A), guanine (G), cytosine (C), and thymine (T) are found in DNA while RNA is made from adenine (A), guanine (G), cytosine (C), and uracil (U).



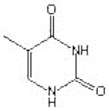
Adenine



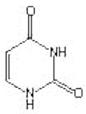
Guanine



Cytosine



Thymine



Uracil

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

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

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

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

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

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

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

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

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

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

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

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

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

Computational biology: See bioinformatics

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

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

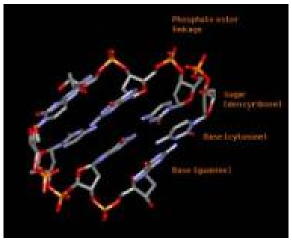
Conserved sequence: A sequence within DNA or protein that is consistent across species or has remained unchanged within the species over its evolutionary period.

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

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

Diploid: A cell containing two sets of chromosomes.

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Genome: The total genetic material of a given organism.

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

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

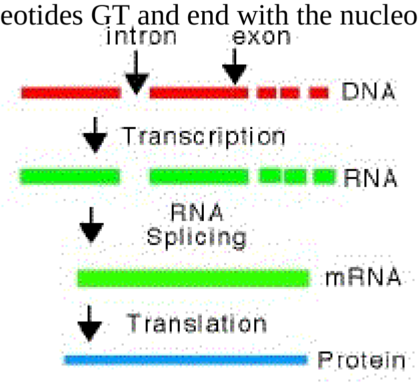
Genotyping: The use of markers to organize the genetic information found in individual DNA samples and to measure the variation between such samples.

Haploid: A cell containing only one set of chromosomes.

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

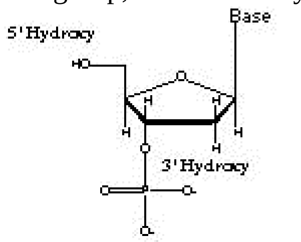
NCBI: The National Center for Biotechnology Information (http://www.ncgi.nlm.nih.gov), a division of the NIH, is the home of the BLAST and Entrez servers.

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

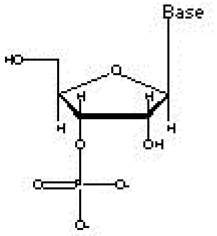
NHGRI: The National Human Genome Research Institute of the NIH (http://www.nhgri.nih.gov)

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

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



DNA Nucleotide



RNA Nucleotide

Oligos (Oligonucleotides): A chain of nucleotides.

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

PCR (polymerase chain reaction; in vitro DNA amplification): The laboratory technique for duplicating (or replicating) DNA using the bacterium Thermus aquaticus, a heat stable bacterium from the hot springs of Yellowstone. As with the polymerase reaction that occurs in cells, there are three stages of a PCR process: separation of the DNA double helix, addition of the primer to the section of the DNA strand which is to be copies, and synthesis of the new DNA. Since PCR is run in a single reaction vessel, the reactor contains all of the components necessary for replication: the target DNA, nucleotides, the primer, and the bacterial DNA polymerase. PCR is initiated by heating the reaction vessel to 90° which causes the DNA chains to separate. The temperature is lowered to 55° to allow the primers to bind to the section of the DNA that they were designed to recognize. Replication is then initiated by heating the vessel to 75°. The process is repeated until the quantity of new DNA desired in obtained. Thirty cycles of PCR can produce over 1 million copies of a target DNA.

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

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

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

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

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

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

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

Position 1	CAGAUGAGGUCAGGCAUA	
	gln met arg ser Gly ile	
Position 2	C	AGAUGAGGUCAGGCAUA
		arg trp gly Gln ala
Position 3	CA	GAUGAGGUCAGGCAUA
		asp glu val Arg his

DNA (through RNA) uses a triplet code to specify the amino acid for a given protein. As can be seen above, a given strand of DNA has three possible starting points (position [or reading frame] one, two, or three). Since both strands of DNA can be translated into RNA and then into protein, a sequence of double helical DNA can specify six different reading frames.

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

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

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

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

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

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

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

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

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

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

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

Yeast artificial chromosome (YAC): An artificial chromosome containing a yeast centromere, two telomeric sequences, and a marker. The YAC is constructed by cloning very large genomic fragments (up to one million bases) from another species into yeast vectors that can replicate in yeast.