The Peanut Genome Project Action Plan

Performance Measures & Deliverables

Genome Sequencing & Structural Characterization

The Peanut Genome Consortium (PGC), a coalition of international scientists and stakeholders, will guide and implement research conducted in the Peanut Genome Project (PGP) as an integral program of the IPGI. PGP goals are: 1) a high quality chromosome scale draft of a tetraploid (cultivated species) reference genome sequence, 2) high throughput genome and transcriptome characterization of tetraploid and diploid (progenitor species) genetic resources, 3) phenotypic trait association with genetic haplotypes, and 4) interactive bioinformatic resources for data curation and analysis. The outcome of these efforts will enable molecular breeding approaches for enhancing peanut yielding ability, resistance to diseases and insects, tolerance to environmental stresses, and improved quality traits that promote peanut crop competitiveness and grower's profitability in an environmentally sustainable manner.

Goal: Genome sequence annotation, assembly & characterization in Arachis species

Performance Measures:

PM 1 Generation of a high quality reference genome sequence of cultivated peanut anchored to chromosomal linkage groups (PGP Component 1)

The Beijing Genome Institute (BGI), a collaborating partner in the PGC, will develop a reference genome sequence of peanut (*Arachis hypogaea L. Fabaceae*). The PGC will provide BGI with genomic DNA from *Arachis hypogea* cv. Tifrunner, cv. GT-C20 and 100 lines of a RIL population developed from Tifrunner x GT-C20. BGI will perform deep sequencing of cv. Tifrunner using a combination of whole genome shotgun and BAC-by-BAC approaches. BGI will use sequences generated by BGI, genetic/ haplotype data from progenitor species generated under PM 2.2, and RNA-seq data generated under PM 2.3 to assemble and annotate a reference genome. BGI will simultaneously transfer results to all PGC members via the National Center Genome Resources (NCGR) in Santa Fe, New Mexico, U.S.A

Anticipated Products:

- DNA sequences of individual BACs of a 6x coverage library from *Arachis hypogea* cv. Tifrunner, with sequence depth of at least 50x with overall coverage of at least 300x
- DNA sequences of individual BACs of *Arachis hypogea* cv. GT-C20, with sequence depth of at least 60x
- DNA sequences of individual BACs from each of up to 100 RILs from *Arachis hypogea* cv. Tifrunner x GT-C20, with a sequence depth of 1x
- A high-resolution genome assembly of each of the four haplotypes present in the two constituent genomes of the allotetraploid with contig parameters of: N50 >20 Kb, scaffold N50 > 300 Kb, and single base error rate <1/100,000. The total number of scaffolds should be less than 10,000.
- Raw data sufficient to allow genome assembly (FASTA files with base quality scores or FASTQ files before and after trimming); paired-end and mate-pair information; depth of base pair coverage at all positions within the pseudo-molecules, scaffolds and non-scaffolded contigs; sequences of individual BACs; sequences of scaffolds and contigs in a separate FASTA file from the pseudomolecule
- Physical and genetic coordinates of the scaffolds and contigs in chromosomal linkage groups based on analysis of the segregation data from the RILs.
- Advanced bioinformatics analyses by BGI including:
  - Genome assembly, statistics and gene prediction.
  - Basic genome information (size, GC content, average heterozygousity, repeat information)

- Results of genome sequencing and assembly (sequence image analysis, base calling and sequence analysis; sequencing data summary; contig size/number, scaffold size/number from N50 to N90)
- Data on genome assemblies: (euchromatic and gene region coverage with sequencing depth)
- Genome annotation results (repeat analysis and annotation; annotated protein-coding genes including gene structure prediction and gene function annotation; non-coding RNA gene annotation including microRNA, tRNA, rRNA and other ncRNA; annotation of transposons and tandem repeats)
- Comparative genomics and evolution analysis (chromosome structure variation detection of specific genome regions; specific gene detection; fast evolutionary gene detection; synteny blocks; and gene family analysis)
- Validated genome assembly with a linear order of the contigs in chromosomal linkage groups.

PM 2 Genome mapping and allelic analysis through Genome-Wide-Association-Studies (PGP Component 2)

The international peanut research community has created sets of genetic resources that facilitate both the generation of high resolution genetic maps through mapping by sequencing as well as the mapping of important agricultural traits through genome wide association studies (GWAS). These resources include:

- Diploid RIL mapping populations for both genomes of the progenitor species (AA: *A. duranensis x A. stenosperma*; BB: *A. ipaensis x A. magna*). Analysis of diploid RIL populations will enable the generation of high resolution genetic maps without the potential complications generated by ploidy in cultivated *Arachis* (Univesidade de Brasilia, EMBRAPA). A physical map of the *A. duranensis* is being generated (University of Georgia)
- A tetraploid mapping population derived from *A. hypogaea* IAC Runner x a synthetic (AABB) amphidiploid. This mapping population presents a polymorphism-rich model that integrates and enhances access to the high degree of allelic variation between diploid species and cultivated peanut. (Univesidade de Brasilia, EMBRAPA).
- A diversity panel of 300 accessions representing the diversity of the international peanut germplasm collection. This material has been genotyped based on SSR data and phenotyped for several agronomic traits (ICRISAT).
- RILs segregating for drought tolerance and foliar diseases (ICRISAT).
- USDA Mini-Core (112) and Core (750) accessions representing genetic diversity in the USDA Peanut Germplasm Collection (USDA-ARS, Griffin GA)
- RILs Chinese peanut germplasm mini-core collection (Oil Crops Research Institute, Wuhan, China)

Genetic mapping through sequencing and analysis of diploid and amphidiploid RIL populations at UC-Davis will capture gene space in parental lines and each RIL of populations from the Univesidade de Brasilia & EMBRAPA. Analysis of RIL data will be used to generate an ultra-dense, gene-based genetic map for each population. Genetic mapping and GWAS through low-coverage sequencing of the diversity panel (ICRISAT) at UC-Davis, and parallel analyses of the Mini-Core collection (USDA) at USDA-ARS, Stoneville, MS and University of Georgia) will capture gene-space and sequence variation in cultivated peanut germplasm. Analysis of SNPs will reveal the level of linkage disequilibrium in these germplasm and help refine the genetic maps generated from Tifrunner x GT-C20 RILs (PM 2.1) and the populations described above. These analyses provide the foundation for efficient QTL mapping and the generation of a peanut haplotype map in conjunction with the reference sequence.

Anticipated Products:

- High resolution genome maps of A and B genomes of the *Arachis* ancestors and the amphidiploid synthetic hybrid.
- SNP maps correlated with the variation captured in the diversity panels and germplasm collections.
- GWAS studies of the agricultural traits phenotyped on the ICRISAT panel.
- The sequence assemblies will be distributed through the NCGR after QC analysis.

PM 3 Catalog expressed genes and profile gene expression in cultivated peanut (PGP Component 3)

Other genome projects suggest the number of protein encoding genes in crop species may exceed 40,000. Genome sequencing reveals all of the genes present within an organism, but does not reveal which of those genes are active in different metabolic pathways, tissues, or stages of development. Until recently, analysis of cDNA libraries of expressed gene sequences (ESTs) was limited to a gene-by-gene approach. New high-through-put sequencing platforms (such as RNA-seq) provide a rapid and sensitive means to survey gene expression and create a comprehensive peanut gene expression atlas that catalogs gene activity in different tissues and treatments. Such an atlas would be a valuable resource for the study of peanut gene function. RNA-Seq (whole transcriptome shotgun sequencing) deploys high-throughput sequencing technology to discern how individual alleles are expressed, detect post-translational mutations, and discover other functional aspects of gene expression profiles. RNA-seq provides a comprehensive and accurate measurement of gene expression that complements cDNA characterization by Sanger sequencing, SAGE and MPSS methods. RNA-seq will be used to catalog expressed genes, validate gene predictions and profile gene expression in *Arachis hypogea* cv. Tifrunner tissues (leaf, apical meristem, stem, root, flower, gynophore, pericarp, seed) across multiple developmental stages and under challenge with various stresses. This information will add definitive context to the annotation of the reference peanut genome sequence.

Anticipated Products:

- A standardized methodology for submitting data towards annotation of the whole peanut genome.
- Expression profiles of genes that mediate resistance to diseases and pests, such as: tomato spotted wilt virus (TSWV), leaf spot (early - *Cercospora arachidicola*; late - *Cercosporidium personatum*), rust (*Puccinia arachidis*), white mold (*Sclerotium rolfsii*), nematode (*Meloidogyne arenaria*), and pre-harvest aflatoxin contamination (*Aspergillus flavus*)
- Expression profiles of genes that mediate tolerance to abiotic stresses, such as: drought, temperature (cold, heat), and nutrient deficiency
- A peanut gene atlas which includes a comprehensive list of all expressed soybean genes, alternative splice products, the identification of co-regulated genes and gene networks.

PM 4 Evaluation of emerging technologies for genome sequencing and characterization (PGP Component 4)

Technological advances in genome sequencing and characterization will be considered as they become available. Although implementation of this priority will be necessarily delayed, two potential opportunities that may be considered are: 1) direct sequencing using the Pacific Biosciences platform for single molecule real-time analysis; and 2) analysis of chromosome specific libraries. Direct sequencing technology is potentially a very powerful complement to the short reads generated by Illumina methods. Strobe sequencing in particular could be useful for scaffolding contigs and assigning haplotypes in heterozygous and tetraploid genomes. If individual peanut chromosomes can be separated by microfluidic techniques, the DNA should be suitable for whole genome amplification (WGA) and small insert paired-end library sequencing. Analysis of chromosome specific libraries would complement the BAC-based, whole-genome sequencing approach proposed for Component 1 should allow for the assignment of homeologous sequences (PM 1)

.

Anticipated Products:

- Evaluation of the utility of the Pacific Biosciences platform by UC-Davis.
- Pending a collaboration with Stanford University, UC-Davis will evaluate microfluidic methods that separate and amplify individual chromosomes from single peanut cells of the cv.Tifrunner.

PM 5 Phenotypic validation of gene predictions (PGP Component 5)

Many DNA markers revealed by GWAS of genomic haplotypes plus RNA-seq and Sanger analysis of transcriptomes among germplasm resources noted in PM2.2 and PM2.3 will facilitate more efficient QTL mapping and the generation of a peanut haplotype map in conjunction with the reference sequence. The identification of candidate genes within QTL will reveal potentially superior DNA markers which must be validated to enable effective marker-assisted-selection for specific traits. Several mapping populations for important protection traits and improved quality traits have been developed between *Arachis hypogea* cv. Tifrunner and other parents. Accurate and timely phenotypic association with candidate genes will help

refine the genetic maps generated from Tifrunner x GT-C20 RILs (PM 2.1) and enable pre-breeding with perfect and flanking markers for optimal detection of specific alleles.

Anticipated Products:

- DNA markers that contribute to the assembly and annotation of the peanut genome
- DNA markers that can be used in pre-breeding for disease and pest resistance including TSWV, Early & Late Leaf Spot, CBR, nematodes, PAC, drought.
- DNA markers that can be used in pre-breeding for quality traits including seed fatty acid composition, flavor quality, nutritional benefits, and other seed composition traits
- DNA markers for peanut yielding ability and other agronomic traits

PM 6 Development of bioinformatic resources for peanut genome data (PGP Component 6)

The value of the peanut genome sequence and related genetic and genomic resources will be greatly enhanced if the data can be integrated and made accessible using state of the art on-line genomic research tools. While it is crucial to provide, at a minimum, unimpeded access to the intermediate products of a genomics project (the sequence assemblies, the gene models and coordinates, the gene read counts, etc.), these raw data sets in themselves are relatively limited value to end users such as breeders and molecular biologists. Bioinformatic resources are needed to improve ability to compile, analyze, and interpret genomic data in a useful and timely manner. A natural framework for the peanut genomic data would be the combination of the Legume Information System and a peanut version of SoyBase (ie. PeanutBase). These web resources have been collaboratively developed by NCGR and USDA-ARS at Ames IA:. LIS (http://comparative-legumes.org) will provide a framework for integrating new legume species; a set of search and visualization tools; map and comparative (syntenic) information; and gene family and evolutionary information. SoyBase (http://soybase.org) would provide detailed species about one legume species: genome browser; metabolic maps; trait, QTL, map, and phenotype data; trait and gene and developmental ontology information; expression information; and gene descriptive information. LIS and SoyBase staffs will work closely with data generators, curators and biologists from the peanut research community to serve as the custodian, curator and distribution agent for peanut genomic data generated by all members of the PGC.

Anticipated Products:

- A PGP Informatics Steering Committee to address current and future informatics needs.
- An International PGC Annotation Group to interface with BGI for peanut genome annotation and the establishment of a controlled vocabulary nomenclature.
- Community standards for expression, protein and metabolite profiling platforms and data.
- A peanut genomic database that facilitates navigation from maps to genes to traits
- An integrated database including available genetic stocks, mutants and germplasm collections
- A HapMap browser that connects the sequence to polymorphisms for traits of interest
- Ability to map RNA-seq and Sanger reads from expression data onto QTL data
- Integration of genome sequence with physical, genetic and transcriptome maps
- Molecular tools for the identification of candidate genes underlying QTLs.
- Integration of plant trait and phenotypic data with genetic maps and other genetic data.
- Workshops/jamborees for community-driven annotation updates at the gene family level
- Integrated mutagenesis, knockout, expression data for gene function annotation.
- Annotation resources for transposons, repeats, sRNAs, and conserved non-coding elements
- Integrated annotation among *Arachis* genome resources.
- Ability to discover evidence of synteny, orthologous genes, expression/co-expression levels, and regulatory networks in a comparative context.
- A plan for long-term curation of the peanut genome sequence, updates on annotation, correction of assembly errors and incorporation of other relevant data

Collaborators

David Bertioli
University of Brasilia

Soraya Bertioli
EMBRAPA, Brasília, Brazil

Mark D Burow
Texas Agric. Expt. Station, Lubbock, TX

Steven Cannon
USDA, ARS, Ames IA

Lutz Froenicke
UC-Davis Genome Center

Baozhu Guo
USDA-ARS, Tifton, GA

Corley Holbrook
USDA ARS Tifton GA

Ran Hovav
The Volcani Center, Bet-Dagan, Israel

Sachiko Isobe
Kazusa DNA Research Institute
Kisarazu, Chiba JAPAN ...

Scott A. Jackson
University of Georgia, Athens GA

Gregory D. May
Pioneer HiBred International

Richard Michelmore
Director, UC Davis Genome Center

C. Victor Nwosu
Masterfoods USA

Peggy Ozias-Akins
NESPAL, University of Georgia-Tifton GA

Brian Scheffler
USDA ARS JWDSRC, Stoneville, MS

Howard-Yana Shapiro
Director, Plant Science, MARS Incorporated

Tom Stalker
North Carolina State University, Raleigh, NC

Howard Valentine,
The Peanut Foundation

Rajeev K. Varshney
ICRISAT, Patancheru, Hyderabad, India

Farid Waliyar
Director, ICRISAT West and Central Africa
Niamey, Niger

Xingjun Wang
Shandong Academy Agric. Sci, Shandong, China

Richard F. Wilson
Oilseeds & Bioscience Consulting, Raleigh, NC

Graeme Wright
Peanut Company of Australia, Kingaroy, Australia

Xun Xu, VP Bioinformatics Center
BGI Shenzhen, China

Xingyou Zhang, Vice President
Henan Academy of Agricultural Sciences
Zhengzhou, China

Pedro Arraes (ex officio)
President, EMBRAPA
Brasilia, DF, BRAZIL

David Hoisington (ex officio)
Deputy Director General for Research
ICRISAT, Patancheru, Andhra Pradesh, India

Chairman Luo Fuhe (ex officio)
Vice Chairman of the 11th National Committee of the Chinese People's Political Consultative Conference (CPPCC) and
Executive Vice-Chairman of the China Association for Promoting Democracy (CAPD)
Beijing China

Jean-Marcel Ribaut (ex officio)
Director, The GENERATION Challenge Program
CIMMYT, Mexico DF, Mexico

Roy Scott (ex officio)
USDA, ARS, Office of National Programs