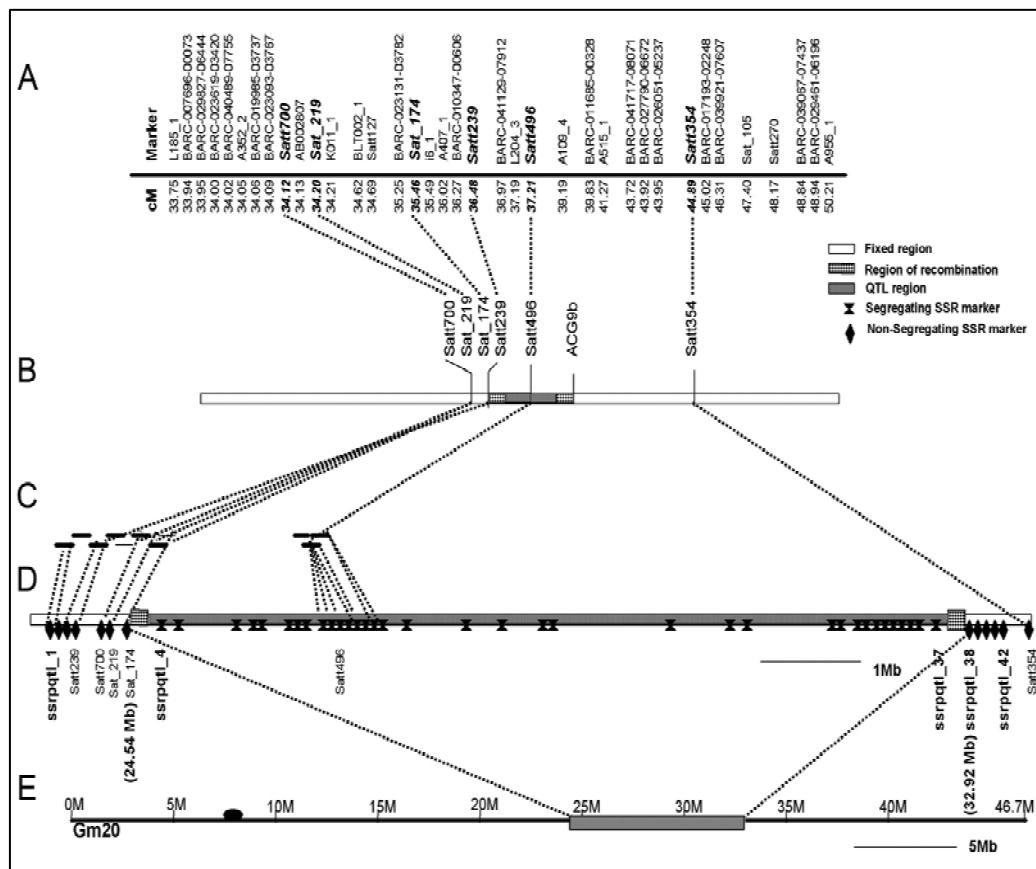


Guidance for Developing a DRAFT

International Peanut Genome Initiative Strategic Plan for 2012 to 2016

Characterization of the Peanut Genome

v 1.0



Introduction

Successful Research Initiatives usually follow a process that helps keep their activities in a favorable spotlight with stakeholders and investors. The process has at least three basic steps:

1) establishment of forums for communication like AAGB; 2) development of a Strategic Plan that provides a framework for implementation of research priorities that address stakeholder needs; and 3) internal and external evaluations to reassure stakeholders that programmatic milestones are being met. The latter step often involves documentation of deliverables and accomplishments that are relevant to objectives of the Strategic Plan. Taken together, these steps promote program transparency & accountability; attributes that will help attract and maintain the support needed to sequence and characterize *Arachis* genomes.

Keeping the Spotlight on IPGI

- Global forums for communication between stakeholders and research communities
- Strategic Plans to define and guide program implementation
- Credible processes to ensure program relevance, high-quality research, and external assessment of progress

Scope of the Plan

Improve understanding of peanut genome organization & structure

Develop genomic tools and characterize genes for traits that meet stakeholder needs

Apply technology to improve crop productivity, stress management & crop quality

As we prepare for AAGB-2011, it is time to revisit research goals, priorities, and expectations for the next 5-year IPGI Strategic Plan (2012 to 2016). The scope of the new plan will: 1) strive for a better understanding of the peanut genome structure, 2) develop useful genomic tools and technologies that help solve major problems, and 3) direct the application of technology toward enhanced crop productivity, increased crop protection, and improved product quality/safety.

The new plan will feature five research areas. These areas are the featured topics of plenary sessions at AAGB-2011. Plenary session talks should help generate discussion in facilitated breakout sessions on the same topic. In the breakout sessions, facilitators will help capture stakeholder input for developing each section of the new plan (Note: the Gene Mapping & Gene Sequencing areas are combined for logistical convenience in the Breakout sessions). Please consult the current IPGI Strategic Plan on the *PeanutBioscience* website as a reference. It may be desirable to carry some aspects of that plan over to the new plan.

Concurrent Breakout Sessions

<u>Allelic Diversity & Germplasm Resources</u> (new sources of genetic variation)	T. Stalker, J. Valls
<u>Genetic Mapping & Gene Discovery</u> (QTL validation, gene function)	P. Ozias-Akins, R. Varshney
<u>Genome Sequence & Structure</u> (genome assembly, bioinformatics)	S. Jackson, D. Bertoli
<u>Product Quality & Safety</u> (aflatoxin, flavor, nutrition)	V. Nwosu, F. Waliyar
<u>Crop Improvement</u> (breeding, yield, stresses, biotech)	C. Holbrook, K. Sharma

Strategic Research Plan Format

- **Component:** describes a general area of research needed to achieve the mission

Goal: a statement defining the desired outcome from a general area of research

Performance Measure: a research Objective that is relevant to the Goal (why the work is important & what is needed to get the job done)

Milestones & Deliverables (outputs expected over the 5-year program cycle)

After everyone in a Breakout session gets acquainted, elect a rapporteur who will speak for the group in the concluding general session of the AAGB-2011 program. Facilitators will guide discussion and record input.

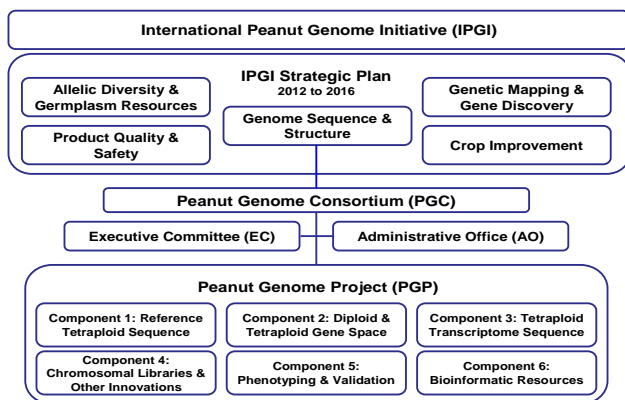
The first discussion item should be the selection of one or more Goals that define a desired outcome from the research area. Facilitators

should have language in mind to expedite deliberations and preserve time for:

- Developing a list of research Objectives (Performance Measures) that defines research needs for solving relevant problems.
- Prioritizing the list of Objectives, considering why the objective is needed, and the best way to achieve it.
- Making a list of Anticipated Products (deliverables or milestones) for each Objective over a five year period.

Session facilitators and volunteers will serve on the writing team. Drafts of the new plan will be distributed to all IGPI members for comment and ratification via www.PeanutBioscience.com/ by July 9, 2011.

As you may know, discussion of ways to implement the *Genome Sequencing & Structural Characterization* area of the new Strategic Plan was initiated in December 2010 at a PGI meeting in Atlanta GA. Those early deliberations explored the feasibility of organizing an international collaborative effort to generate a high-quality draft chromosomal scale map of the peanut genome. The *PeanutBioscience* website will post the proceedings of meetings on this subject. These documents chronicle the establishment of the Peanut Genome Consortium (PGC), a coalition of international scientists and stakeholders that will guide and administer research conducted in the Peanut Genome Project (PGP) as an extended program of the IPGI. PGP goals are: 1) development of a high quality chromosome scale draft of a tetraploid (cultivated species) reference genome sequence, 2) high throughput genome and transcriptome characterization of gene space in tetraploid, amphidiploid and diploid (progenitor) species, 3) validation of phenotypic trait association with genetic haplotypes, and 4) bioinformatic resources for data curation and analysis.



Current thinking on implementation of the six components of the PGP is presented in **Exhibit 1** for discussion in the breakout session, and as an example of format recommended for sections of the new Strategic Plan.

Additional input for the PGP is welcome, as are nominations for new contributing members of the PGC. The PGC Policy & Procedure manual will be posted on the *PeanutBioscience* website. Plans for official launch of the PGP will be made after IPGI ratification of the *Genome Sequencing & Structural Characterization* section of the new Strategic Plan at AAGB-2011.

Thank you for your help in this important task. Please contact Rich Wilson (919-906-6937 or rfwilson@mindspring.com) if there are questions or if conflicts arise.

Exhibit 1

Table of Contents

International Peanut Genome Initiative Strategic Plan for 2012 to 2016

Executive Summary	X
Introduction	X
Strategic Research Goals & Performance Measures	
Allelic Diversity & Germplasm Resources.....	X
Genome Sequencing & Structural Characterization.....	X
Genetic Trait Mapping & Gene Discovery.....	X
Product Quality & Safety	X
Crop Improvement	X
Milestones	X
Collaborators	X

Strategic Research Goals & Performance Measures

Proposed Example

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 2: Genome sequence annotation, assembly & characterization in *Arachis* species

Performance Measures:

PM 2.1 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 hypogaea* 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 hypogaea* cv. Tifrunner, with sequence depth of at least 50x with overall coverage of at least 300x
- DNA sequences of individual BACs of *Arachis hypogaea* cv. GT-C20, with sequence depth of at least 60x
- DNA sequences of individual BACs from each of up to 100 RILs from *Arachis hypogaea* 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 heterozygosity, 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.
- OTHER

PM 2.2 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)
- ADD appropriate Chinese germplasm and approach
- OTHER

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.
- OTHER

PM 2.3 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.
- OTHER

PM 2.4 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 2.1)

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.
- OTHER

PM 2.5 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
- OTHER

PM 2.6 Development of bioinformatic resources for peanut genome data (PGP Component 6)

Advances in sequencing technology often reduce operational costs per base pair, but significantly attenuate the amount of genomic and associated data that can be generated. Although data storage limitations may be overcome, individual laboratories can become overwhelmed by the complexity of genomic analyses. Bioinformatic resources are needed to improve ability to compile, analyze, and interpret genomic data in a useful and timely manner. Bioinformatic resources developed by the National Center for Genome Resources (NCGR) include the Legume Information System (LIS). LIS was created to integrate genomic information from *Arabidopsis thaliana*, *Medicago truncatula* and *Lotus japonicus* to important agronomic legumes, e.g., soybean, alfalfa, pea, dry beans and peanut. Since its inception, LIS development has been guided by ideas and suggestions from the legume research communities. Based on user needs, the power of this information resource has been amplified for identification of candidate genes, unique genes, and evolutionary relationships among genes for crop improvement. LIS also is a foundation for the Comparative Legume Biology (CLB) Program which expands LIS capacity for novel data analysis and visualization tools. With the emergence of high throughput biotechnologies and bioinformatics, comparative biology enables more sophisticated analyses of genomic sequences, genomic maps, micro-arrays, protein arrays, metabolic arrays, genetic regulatory networks, biochemical and whole organism phenotypes. For these reasons, NCGR will 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, genetics 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
- OTHER

Collaborators

David Bertioli

Instituto de Ciencias Biologia
University of Brasilia
davidbertioli@unb.br
david.bertioli@pq.cnpq.br

Soraya Bertioli

Embrapa Recursos Genéticos e Biotecnologia
SAIN Parque rural W-5 Norte
Caixa Postal 02372
70849-970 – Brasília, DF – Brasil
(61) 3448 4735 Fax: (61) 3340 3658
soraya@cenargen.embrapa.br

Lutz Froenicke

Lutz.Fr@gmail.com

Baozhu Guo

USDA-ARS, P.O. Box 748
Tifton, GA 31793-0748
229-387-2375
Baozhu.Guo@ars.usda.gov

Scott A. Jackson

Department of Agronomy
224 WSLR, Lilly Hall of Life Sciences
Purdue University
915 W. State Street, West Lafayette, Indiana 47907
765-496-3621; FAX: 765-496-7255
sjackson@purdue.edu

Gregory D. May

President/Chief Operating Officer NCGR
2935 Rodeo Park Drive East
Santa Fe, New Mexico 87505
gdm@ncgr.org

Richard Michelmore

Director, UC Davis Genome Center
4311 Genome and Biomedical Sciences Facility
451 Health Sciences Drive, Davis, CA 95616
rwmichelmore@ucdavis.edu
(530) 752-1729

C. Victor Nwosu

Masterfoods USA
800 High St., Hackettstown, NJ 07840
908-850-7545
victor.nwosu@effem.com

Peggy Ozias-Akins

NESPAL, University of Georgia-Tifton
P.O. Box 748, Tifton GA 31793
229-391-2594
pozias@uga.edu

Brian Scheffler

USDA ARS JWDSRC
141 Experiment Station Rd.
Stoneville, MS, 38776-0038
brian.scheffler@ars.usda.gov
(662) 686-5454; Fax: (662) 686-5372

Howard-Yana Shapiro, PhD

Director, Plant Science
MARS Incorporated
6885 Elm Street, McLean VA 22101-3883
703-821-4900; Fax 703-448-9678
Howard.Shapiro@effem.com

Howard Valentine

Peanut Foundation / American Peanut Council
Alexandria, VA 22314
1500 King Street, Suite 301
706-579-1755
hvalentine@peanutsusa.com

Rajeev K. Varshney

Centre of Excellence in Genomics
Bldg 300, ICRISAT, Patancheru, 502 324
Greater Hyderabad, A.P., India
+91-40-3071-3305; r.k.varshney@cgiar.org

Richard F. Wilson, PhD

The Peanut Foundation/
Oilseeds & Bioscience Consulting
5517 Hickory Leaf Drive
Raleigh, North Carolina 27606-9502
919-834-8695; 919-906-6937 (mobile)
rfwilson@mindspring.com
richard.wilson@ars.usda.gov (ARS business)

Xu Xun, VP Bioinformatics Center

BGI Shenzhen
+86-18688745853; xuxun@genomics.cn

Xingyou Zhang, Vice President

Henan Academy of Agricultural Sciences at
Zhengzhou
haasz@sohu.com; haasz@126.com

Xingjun Wang

Shandong Academy of Ag. Sci. High Tech Center
xingjunw@hotmail.com

Tom Stalker

Crop Science Department
North Carolina State University
Greenhouse Unit 3-Campus Box 7629
Raleigh, NC 27695-7629
(919) 513-2929; Fax: 919-515-5657
tom_stalker@ncsu.edu

Roy Scott (ex officio) PENDING

USDA, Agricultural Research Service
Office of National Programs
Room 4-2218, GWCC 5601 Sunnyside Ave
Beltsville MD 20705-5139
Phone: (301) 504-4670; Fax: 301-504-6191
Roy.Scott@ars.usda.gov

David Hoisington (ex officio) PENDING

Deputy Director General for Research
International Crops Research Institute for the Semi-
Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh, 502 324, India
d.hoisington@cgiar.org