9 Sequencing Ancestor Diploid Genomes for Enhanced Genome Understanding and Peanut Improvement

Spurthi N. Nayak, Manish K. Pandey, Scott A. Jackson, Xuanqiang Liang and Rajeev K. Varshney

Abstract

Cultivated peanut (*Arachis hypogaea*) is an allotetraploid with closely related subgenomes of a total size of ~2.7 Gb. To understand the genome of the cultivated peanut, it is prerequisite to know the genome organization of its diploid progenitors, A-genome—*Arachis duranensis* and B-genome—*A. ipaensis*. Two genome sequencing projects conducted sequencing and analysis of the genomes of diploid ancestors: (1) International Peanut Genome Initiative (IPGI) reported the sequencing of both A- and B-genomes; while (2) Diploid Progenitor Peanut Arachis Genome Sequencing Consortium (DPPAGSC) reported the sequencing of A-genome. IPGI study showed that these genomes are similar to cultivated peanut's A- and B-subgenomes and used them to identify candidate disease resistance genes, to guide tetraploid transcript assemblies and to detect genetic exchange between cultivated peanut's subgenomes thus providing evidence about direct descendant of the B subgenome in cultivated peanut. The DPPAGSC study, on the other hand, provided new insights into geocarpy, oil biosynthesis, and allergens in addition to providing information about evolution and polyploidization. These genome sequencing efforts have improved the understanding about the complex peanut genome and genome architecture which will play a very important role in peanut applied genomics and breeding.

S.N. Nayak · M.K. Pandey · R.K. Varshney (✉)
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India
e-mail: r.k.varshney@cgiar.org

S.A. Jackson
Center for Applied Genetic Technologies, University of Georgia (UGA), Athens, USA

X. Liang
Crop Research Institute, Guangdong Academy of Agricultural Sciences (GAAS), Guangzhou, China

R.K. Varshney
University of Western Australia (UWA), Crawley, Australia

R.K. Varshney et al. (eds.), *The Peanut Genome*,
Compendium of Plant Genomes, DOI 10.1007/978-3-319-63935-2_9

9.1 Introduction

Since the availability of first plant genome *Arabidopsis thaliana* in 2000, genomes of several plant species have been sequenced (Michael and Jackson 2013). With advancements in sequencing technologies and genome assembly methodologies over the past decade, genome sequencing is now not limited to only model plant species or small genomes. Several crop plants, plantation crops, vegetables, fruits and even the wild progenitors of important crop species have been sequenced and many are in progress. With the advent of next-generation sequencing (NGS) technologies, there is a rapid increase in sequenced plant genomes due to the exponential decrease in cost and time in generating sequencing data (Varshney et al. 2009; Schatz et al. 2012). Rice was the first sequenced crop genome and had a major impact on accelerating rice genetics research and breeding applications (Jackson 2016). The genome sequencing projects for most crops have been possible due to international collaborations and both formal and informal consortia.

Most of the sequenced plants have been diploids while the sequencing of polyploids and large sized genomes has been less frequent. Polyploid genomes increase the genome complexity and therefore, pose a serious challenge towards the development of high-quality assemblies of pseudomolecules and genomes. Hence as a basis for the polyploid genome sequencing, where available, the diploid progenitors have been sequenced for several polyploid plant species like cotton (Wang et al. 2012), wheat (Ling et al. 2013, Jia et al. 2013, Marcussen et al. 2014), and capsicum (Qin et al. 2014). Polyploidy or whole-genome duplication (WGD) has been proposed to be a major evolutionary force in plants, especially in angiosperms (see Soltis et al. 2014). Cultivated peanut is an allotetraploid with total genome size of ~2.7 Gb. The peanut subgenomes are closely related (Nielen et al. 2012; Moretzsohn et al. 2013). However, the A and B subgenomes appear to have undergone relatively few changes since polyploidization as evidenced by genomic in situ hybridization (GISH) which clearly distinguished A and B chromosomes without much mosaics (Ramos et al. 2006; Seijo et al. 2007). The genome size of *A. hypogaea* is close to the sum of those for *A. duranensis* (1.25 Gb) and *A. ipaensis* (1.56 Gb), indicating that there has been no large change in genome size since polyploidy (Samoluk et al. 2015). In addition, progenies derived from crosses between cultivated peanut and an artificially induced allotetraploid (*A. ipaensis* K30076 × *A. duranensis* V14167) ($2n = 4x = 40$) were fertile and phenotypically normal with low segregation distortion (Foncéka et al. 2009). These observations strongly support the close relationships between the diploid genomes of the progenitors and the corresponding subgenomes of *A. hypogaea* (Fávero et al. 2006). Hence sequencing of diploid progenitors was a logical choice as it not only provides ease of tetraploid assembly, but also provides a deeper understanding of *Arachis* biology, evolution and any genomic change following polyploid formation.

9.2 Sequencing of Progenitor Diploid Genomes of Cultivated Peanut

Sequencing of the peanut A-genome progenitor, *A. duranensis* V14167, and the B-genome progenitor, *A. ipaensis* K30076, was completed by the International Peanut Genome Initiative (IPGI, http://www.peanutbioscience.com/peanutgenomeinitiative.html) and published in *Nature Genetics* (Bertioli et al. 2016) (Table 9.1). In another effort, the A-genome progenitor, *A. duranensis* PI475845 was sequenced by China-ICRISAT-UGA co-led initiative (Diploid Progenitor Peanut A-Genome Sequencing Consortium, DPPAGSC, http://ceg.icrisat.org/dppga/Manuscript.html) and published in *Proceedings of National Academy of Sciences of the United States of America* (Chen et al. 2016) (Table 9.1). The genotype V14167 (A-genome, *A. duranensis*) originated from Argentina while the other two genotypes, PI 475845 (A-genome, *A. duranensis*) and K30076 (B-genome, *A. ipaensis*), originated in Bolivia.

9.3 Strategies and Tools for Sequencing

9.3.1 Sequencing Platform

The Illumina HiSeq 2000/2500 platforms were used to generate sequence data in the peanut genome projects. Illumina captures template DNA that has been ligated to specific adapters in a flow cell, a glass enclosure similar in size to a microscope slide, with a dense lawn of primers. The template is then amplified into clusters of identical molecules, or polonies, and sequenced in cycles using DNA polymerase. Terminator dNTPs in the reaction are labeled with different fluorescent labels and detection is by optical fluorescence. As only terminators are used, only one base can be incorporated in one cluster in every cycle. After the reaction is imaged in four different fluorescence levels, the dye and terminator group is cleaved off and another round of dye-labeled terminators is added. The total number of cycles determine the length of the read. While generating peanut genome sequences, the read lengths ranged from 90–150 bp.

9.3.2 Sequence Data Generation

The sequence data were generated using paired-end sequencing insert libraries with insert sizes of 250 bp, 500 bp, 2, 5, 10 and 20 kb using standard protocols provided by Illumina (San Diego, USA). The sequencing yielded in 325.73 Gb of raw data reads for *A. duranensis* and 416.59 Gb for *A. ipaensis* under the IPGI project whereas 229.94 Gb raw data was obtained from *A. duranensis* for the DPPAGSC project.

9.3.3 Quality Filtering

Reads with more than 5% Ns or with polyadenylated termini; reads from the short-insert libraries (170–800 bp) with 20 or more bases having quality scores ≤ 7; reads from the large-insert libraries (2–40 kb) with 40 or more bases having quality score ≤ 7; reads with adaptor contamination (more than 10 bp aligned to the adaptor sequence when allowing ≤ 3 bp of mismatches); reads with read 1 and read 2 having ≥ 10 bp overlapping (allowing 10% mismatches; except for the 250-bp insert library, where the paired reads should overlap); reads identical to each other at both ends that might have been caused by PCR duplication; and reads where the quality of the bases at the head or tail was ≤ 7 were discarded in US-led initiative.

Under DPPAGSC project, the reads of short-insert libraries were trimmed of four low-quality bases at both ends, and reads of long-insert libraries were trimmed of three low-quality bases; duplicated reads from long-insert libraries were filtered out; the reads with 10 or more Ns (no sequenced bases) and low-quality bases were also filtered out from individual reads in all lanes.

9.3.4 *k*-mer Analysis

k-mers were extracted from sequences generated from the short-insert libraries, and the frequencies were calculated and plotted. Genome sizes were estimated by dividing the total numbers of *k*-mers by the depths of the major peaks.

Table 9.1 Summary of genome sequencing efforts for diploid progenitor species

Progenitor species	Genome	Genotype sequenced	Assembly size (Gb)	Genes predicted	Lead consortium
A. duranensis	A	V14167	1.21	36,734	USA-led IPGI
A. duranensis	A	PI475845	1.05	50,324	China-led DPPAGSC
A. ipaensis	B	K30076	1.51	41,840	USA-led IPGI

9.3.5 Error Correction

k-mers were used to correct for errors. For sequencing with high depth, the *k*-mers without any sequencing errors should appear multiple times in the read data set, whereas error-containing *k*-mers should have low frequencies. Sequencing errors in the 17-mers with frequencies lower than three in the clean data for the 250- and 500-bp insert libraries were corrected.

9.4 Tools and Technology Used in Genome Assembly

Under the IPGI project, COPE (Liu et al. 2012) was used to join paired-end reads from the 250-bp insert library into single longer reads of ~250 bp. Genome assembly was performed using SOAPdenovo version 2.05 (Li et al. 2010), with parameters –K 81 –R. Gaps were filled using KGF and Gapcloser version 1.10 (Luo et al. 2012). Finally, SSPACE (Boetzer et al. 2011) was used to further link the scaffolds where connections were supported by more than five paired reads. For assembling genome under DPPAGSC project, 159.07Gb filtered reads were further used for genome assembling. SOAPdenovo2 (version 2.04.4) with optimized parameters (pregraph -K 79 -p 16 -d 5; scaff -F -b 1.5) was used to construct contigs and original scaffolds. Newbler and SOAPdenovo were used with parameters -K 79 -p 16 -d 5. The gaps were closed with GapCloser, scaffolds were reconstructed using Haplomerger (Huang et al. 2012). The paired-end information was subsequently applied to link contigs into scaffolds in a stepwise manner. Several intra-scaffold gaps were filled by local assembly using the reads in a read-pair, where one end uniquely mapped to a contig, whereas the other end was located within a gap. Subsequently, SSPACE (version 2.0; using core parameters "-k 6 -T 4 -g 2") was used to link the SOAPdenovo2 scaffolds.

Under IPGI project, ultradense genetic maps were generated through genotyping-by-sequencing (GBS) of two diploid recombinant inbred line (RIL) populations. SNPs within scaffolds were used to validate the assemblies and confirmed their high quality. Based on the presence of diagnostic population-wide switches in SNP genotypic data occurring at the point of misjoin, 190 of 1297 initial scaffolds of *A. duranensis* and 49 of 353 initial scaffolds of *A. ipaensis* were identified as chimeric. These chimeric scaffolds were split and used for remapping. Thus, approximate chromosomal placements were obtained for 1692 and 459 genetically verified scaffolds, respectively. Conventional linkage maps along with the syntenic inferences were used to refine the ordering of scaffolds within the initial genetic bins. Generally, agreement was good for maps in euchromatic arms and poorer in pericentromeric regions. Overall, 96.0 and 99.2% of the sequence in contigs $\geq$ 10,000 bp in length, represented by 1692 and 459 scaffolds, could be ordered into 10 chromosomal pseudomolecules per genome of 1025 and 1338 Mb for *A. duranensis* and *A. ipaensis*, respectively. The pseudomolecules were named as Aradu.A01–Aradu.A10 (GCA_000817695.1) and Araip.B01–Araip.B10 (GCA_000816755.1). The pseudomolecules mostly showed one-to-one equivalence between the A- and B-genomes and were numbered according to previously published linkage maps (Shirasawa et al. 2013, Gautami et al. 2012, Moretzsohn et al. 2005, 2009). They represent 82% and 86% of the genomes, respectively, when considering genome size estimates based on flow cytometry, or 95 and 98% of the genomes when using estimates derived from *k*-mer frequencies with $k = 17$. Comparisons of the chromosomal pseudomolecules with 14 BAC sequences from *A. duranensis* and 6 BAC sequences from *A. ipaensis* showed collinearity of contigs and high sequence identity ($\geq$ 99%). This information was used to improve the genome assembly to pseudomolecule level under IPGI whereas, DPPAGSC has assembly that contained 8173 scaffolds.

9.4.1 Production of Moleculo Synthetic Long Reads

In IPGI project, the TruSeq synthetic long-read sequencing libraries (McCoy et al 2014) were

generated by Moleculo and Illumina as part of beta tests of this technology. Fifteen libraries were generated for *A. duranensis* K7988, and each library was sequenced on a HiSeq 2500 lane; the PE100 reads were assembled into 1.5 million TruSeq (Moleculo) synthetic long reads, providing approximately 5X genome coverage with a mean read length of 3684 bases and an N50 of 4344 bases. Twelve libraries were used for *A. ipaensis* K30076 to yield approximately 2 million Moleculo reads with mean length of 4054 bases and an N50 length of 5152 bases, providing ~6X genome coverage. Thirteen libraries were used for *A. hypogaea* cv. Tifrunner, which produced 1263,111 Moleculo reads with a mean length of 4547 bases and an N50 length of 6137 bases, providing 2.3X genome coverage. These reads were used for genome comparisons and were not incorporated in the diploid genome assemblies.

9.4.2 Linkage Maps and Identification of Misjoins

Conventional molecular marker maps from diploid A- and B-genomes and cultivated peanut × induced allotetraploid recombinant inbred lines (RIL) populations were used to find the order of the scaffolds from peanut assembly. Genetic maps generated from genotyping-by-sequencing data for diploid A- and B-genome RIL populations were used in identification of chimeric scaffolds. RILs from the diploid A- and B-genome populations were shotgun sequenced to 1X genome coverage with paired-end 100-bp reads on a HiSeq 2500 sequencer. The parents were sequenced at 20X genome coverage. Parental-homozygous SNPs were identified by alignments to the scaffolds of the *A. duranensis* and *A. ipaensis* genome assemblies as well as local realignment and probabilistic variant calling in CLC Genomics Workbench (CLC Bio). Filtering in CLC Workbench resulted in about 3 million high-quality homozygous-parental SNPs for both A- and B-genome mapping population parents. The coordinates of these SNPs were converted into BED format, and the alignment data at the SNP coordinates were extracted with SAMtools mpileup60. From the low-coverage sequencing data, groups of 20 consecutive SNPs were haplotyped with a set of custom Python scripts. Genotype calls were inspected visually and by a hidden Markov model (HMM) script (courtesy of Ian Korf, University of California, Davis) to identify population-wide switches in genotype calls corresponding to scaffold misjoins. Scaffolds not displaying recombination for an individual RIL were haplotyped. Linkage groups were identified from the haplotyping data using MadMapper and Carthagene, applying logarithm of odds (LOD) score thresholds of 8 and distance thresholds of 50 cM; genetic maps were generated with Carthagene using the lkh traveling salesman algorithm and flips, polish and annealing optimizations. Additional scaffolds (indicated in the data files) were added to genetic bins in two rounds of binning with a custom Python script. Misjoined scaffolds were split at breakpoint locations identified by flanking GBS SNP locations, at the "upstream SNP" and the "downstream SNP", delineating the switches in genotype calls, and intervening sequence was excluded from the pseudomolecule assembly.

9.4.3 Generation of Chromosomal Pseudomolecules

Under the IPGI project, scaffolds less than 10 kb in length were removed (they are available in the full assembly scaffold files at PeanutBase: Adur1.split6.fa and Aipa2 s.split7.fa, http://peanutbase.org/files/genomes/Arachis_ipaensis/assembly/). Sequences were subjected to RepeatMasker using *Arachis* repeat libraries available at PeanutBase (mobile-elements-AA051914.fasta and mobile-elements-BB051914.fasta). Pseudomolecules were given initial chromosomal placements and orderings according to the GBS maps. Placement was arbitrary within blocks with the same centiMorgan value. Scaffold orientation and placement were refined according to the different genetic maps such as the tetraploid AB-genome map, the diploid

A-genome map (for the *A. duranensis* assembly), the diploid B-genome map (for the *A. ipaensis* assembly) and finally the tetraploid AB-genome consensus map (Shirasawa et al. 2013). Markers were located on the scaffolds using BLAST and ePCR (electronic PCR) with high similarity parameters (taking the top hits only, with placement by BLAST (*e* value < 1 × 10 − 10) given preference over ePCR where both were available). Markers placing scaffolds on linkage groups other than the one assigned by the GBS data were dropped.

Where allowed by map data, scaffold positions and orientations were adjusted using synteny between the two *Arachis* species and, where necessary (generally within pericentromeric regions), synteny with *G. max* and *Proteus vulgaris*; the presence of telomeric repeats near chromosome ends; information from repeat-masked paired-end sequences from 42,000 BAC clones of *A. duranensis* V14167 (FI321525–FI281689) and Moleculo sequence reads from *A. ipaensis* and *A. duranensis*. Apparent inversions were visually inspected and confirmed. Scaffolds with either <5000 non-N bases or <20,000 bp in length and with <10,000 non-N bases were removed. Pseudomolecules were generated with 10,000 Ns separating the scaffold sequences and were oriented and numbered in accordance with previously published maps (Shirasawa et al. 2013; Gautami et al. 2012; Moretzsohn et al. 2005 and 2009). The scaffolds were thus assigned to pseudomolecules under IPGI project. Due to unavailability of proper information on linkage mapping on A-genome, the genome assembly was made at scaffold level under DPPAGSC project.

9.4.4 Gene Prediction and Annotation

Under IPGI project, genome assemblies were masked with RepeatMasker using the repeat libraries developed for the two diploid species and annotated for gene models using the MAKER-P pipeline (Campbell et al. 2014). *Arachis*-specific models for the ab initio gene predictor SNAP were trained using high-scoring gene models from a first iteration of the pipeline and then used in the final annotation pass; no training was done for the other ab initio predictors included in the pipeline. RNA sequencing de novo assemblies for *A. hypogaea* and the diploid *Arachis* species were supplied as transcript evidence along with available EST and mRNA data sets from NCBI for these same species. Further evidence was supplied by proteomes derived from the annotations for *G. max*, *P. vulgaris,* and *Medicago truncatula* as represented in Phytozome v. 10. Default MAKER-P parameters were used for all other options, with the exception of disabling splice isoform prediction (alt_splice = 0) and forcing start and stop codons into every gene (always_complete = 1). The resulting MAKER-P gene models were post-processed to exclude from the main annotation files gene models with relatively poor support (annotation evidence distance scores of ≥0.75) or with significant BLASTN homology to identified mobile-elements (HSP (high-scoring segment pair) coverage over ≥50% of the transcript sequence at ≥80% identity and *e* value ≤1 × 10 − 10). Provisional functional assignments for the gene models were produced using InterProScan and BLASTP against annotated proteins from *Arabidopsis thaliana*, *G. max,* and *M. truncatula*, with outputs processed using AHRD (https://github.com/groupschoof/AHRD), for lexical analysis and selection of the best functional descriptor of each gene product.

Under DPPAGSC project, to annotate the *A. duranensis* genome, an automated genome annotation pipeline MAKER was used that aligns and filters EST and protein homology evidence and produces de novo gene prediction, infers 5′ and 3′ UTR, and integrates these data to generate final downstream gene models with quality control statistics. Several iterative runs of MAKER were used to produce the final gene set. In total, 50,324 gene models for *A. duranensis* were predicted. All predicted protein sequences were functionally annotated using the BLAST + (version 2.2.27) with a threshold E-value of 1e-5 against a variety of protein and nucleotide databases, including the NCBI

nucleotide (NT), the non-redundant protein (NR), the Conserved Domain Database (CDD), the UniProtKB (www.uniprot.org), Pfam and the Gene Ontology (GO). The *A. duranensis* genes were also mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps of KEGG databases. To infer functions for the predicted genes, InterProScan was used to search the predicted genes against the protein signature from InterPro with default parameters. Fifteen gene sets from legumes, oilseed crops and other plant species were used for comparative analysis. A Cytoscape plugin BiNGO was used for enrichment analysis with hypergeometric test and Benjamini multiple testing correction at a significance level of 0.01.

9.5 Assembly of Diploid Genomes

The total assembly sizes were 1.21 and 1.51 Gb for *A. duranensis* and *A. ipaensis*, respectively, from the data generated from the seven paired-end libraries corresponded to an estimated 154X and 163X base-pair coverage from IPGI (Table 9.1). The assembly size of *A. duranensis* obtained by DPPAGSC is 1.05 Gb with 57.14X read depth (Table 9.1).

The approximate chromosomal/pseudomolecule placements were obtained by using ultradense genetic maps in case of IPGI project. SNPs within scaffolds were used to validate the assemblies and confirmed their high quality; 190 of 1297 initial scaffolds of *A. duranensis* and 49 of 353 initial scaffolds of *A. ipaensis* were identified as chimeric, on the basis of the presence of diagnostic population-wide switches in genotype calls occurring at the point of misjoin. Overall, 96.0 and 99.2% of the sequence in contigs $\geq$ 10,000 bp in length, represented by 1692 and 459 scaffolds, could be ordered into 10 chromosomal pseudomolecules per genome of 1025 and 1338 Mb for *A. duranensis* and *A. ipaensis*, respectively (Aradu.A01–Aradu.A10 and Araip.B01–Araip.B10; GenBank, assembly accessions GCA_000817695.1 and GCA_000816755.1. The pseudomolecules mostly showed one-to-one equivalence between the A- and B-genomes and were numbered according to previously published linkage maps (Shirasawa et al. 2013; Gautami et al. 2012; Moretzsohn et al. 2005, 2009). Comparisons of the chromosomal pseudomolecules with 14 BAC sequences from *A. duranensis* and 6 BAC sequences from *A. ipaensis* showed collinearity of contigs and high sequence identity ($\geq$ 99%).

Whereas in China-led initiative, PCR amplification of randomly selected regions, sequence-depth distribution, and expressed sequence tag validation indicated the high quality of the assembled genome with 8173 scaffolds. K-mer analysis indicated *A. duranensis* genome size of 1.38 Gb that is consistent with previous report (Temsch and Greilhuber 2000). However 50,324 protein coding gene models were predicted using transcriptome sequences (Table 9.1). When compared with the gene sets of legumes, oilseeds, and other plant species, *A duranensis* showed highest similarity to legumes with gene numbers comparable with *Medicago truncatula* (50,894), lower than soybean (tetraploid Glycine max, 56,044), and higher than other legumes.

9.5.1 Repetitive Sequences

Under the IPGI project, the transposable elements accounted for 61.7 and 68.5% of the *A. duranensis* and *A. ipaensis* genomes respectively with long terminal repeat (LTR) comprise of more than 50% of each genome. This observation was similar in DPPAGSC study as well where about 59.77% of the *A. duranensis* genome appeared to have transposable elements with ~40% LTR retrotransposons. These observations were comparable with the estimated repetitive content (64%) for cultivated peanut using renaturation kinetics in the past (Dhillon and Rake 1980). The DNA transposons constituted about 10% of the genome under IPGI project whereas they were about 5.19% in case of *A. duranensis* under DPPAGSC project. The long interspersed nuclear elements (LINEs) were about 7.8 and 11.7% in *A. duranensis* and *A. ipaensis* genomes respectively (US-led initiative) and only about 1.26% of the *A. duranensis* genome (China-led initiative). Besides, under

DPPAGSC project, a total of 105,003 simple sequence repeats (SSRs) were identified in *A. duranensis*. Furthermore, resequencing of two other A-genome genotypes and four B-genome genotypes allowed the discovery of ~8 million SNPs and other structural variations.

9.5.2 Gene Annotation and Analysis of Gene Duplications

Under IPGI project, transcript assemblies were constructed using sequences expressed in diverse tissues of *A. duranensis* V14167, *A. ipaensis* K30076, and *A. hypogaea* cv. Tifrunner (16,439,433, 21,406,315, and 2,064,268,316 paired-end reads for each species, respectively. Using these assemblies and representative characterized transposon sequences, 36,734 and 41,840 high-quality non-transposable element genes for *A. duranensis* and *A. ipaensis*, respectively were generated (Table 9.1). The elevated gene numbers in *A. ipaensis* appear to originate from more local duplications, which can be seen in counts of genomically "close" paralogous genes. Considering similar genes within a ten-gene window, there were 25% more in *A. ipaensis* than in *A. duranensis* (7825 vs. 6241). Gene families known to occur in clusters such as those encoding NB-ARC, leucine-rich repeat (LRR), pentatricopeptide-repeat, kinase, WD40-repeat, and kinesin proteins had large differential counts between the two genomes. These differences were also apparent with wider inspection. In a set of 9236 gene families with members in *A. ipaensis* or *A. duranensis*, or both, 2879 families had more members in *A. ipaensis*, 1983 had more members in *A. duranensis* and 4374 had the same number of members in both species.

Under DPPAGSC project, about 50,324 protein coding gene models were predicted using transcriptome sequences in *A. duranensis*. When compared with the gene sets of legumes, oilseeds, and other plant species, *A duranensis* showed highest similarity to legumes with gene numbers comparable with *Medicago truncatula* (50,894), lower than soybean (tetraploid *Glycine max*, 56,044), and higher than other legumes. Of the 50,324 gene models, ~90% matched entries in publically available databases. Approximately 10.9% (5494) of gene models with no homology to known proteins were supported by transcriptome data and may be peanut-specific. A total of 5251 putative *A. duranensis* transcription factor genes in 57 families, 10.4% of the predicted *A. duranensis* genes, slightly higher than soybean, and much higher than most plant species were analyzed. Certain TFs like B3, E2F/DP, FAR1, GeBP, HSF, NAC, S1Fa-like, and STAT were dominant in *A. duranensis.* Families such as ARR-B, CAMTA, DBB, MIKC, and NF-YA, were sparser in *A. duranensis* than in most plants. Expansion and contraction of TF families may reflect regulatory differences in biological functions of *A. duranensis*. In this study, 816 *Arachis* microRNAs (miRNAs), 913 transfer RNAs (tRNAs), 115 ribosomal RNAs (rRNAs), and 202 small nucleolar RNAs (snRNAs) were also annotated. A total of 64 target genes were predicted after aligning 15 new miRNAs to gene models.

9.5.3 Gene Evolution and Genome Duplication

The IPGI project analyses suggest that the *Arachis* lineages have been accumulating mutations relatively quickly since the divergence of the Dalbergioid clade ~58 million years ago. Modal *K*S values (synonymous substitutions per synonymous site) for paralogs are approximately 0.95 for *A. ipaensis* and 0.90 for *A. duranensis*, more similar to that the Ks value for *Medicago* paralogs of ~0.95) than to those of *Lotus* (~0.65), *Glycine* (~0.65) or *Phaseolus* (~0.80). Average rates of change for *Arachis* genes were estimated at $8.12 \times 10 - 9$ *KS*/year. *Arachis* has accumulated silent changes at a rate ~1.4 times faster than that in *G. max*. On the basis of average rates of change for *Arachis* of 8.12×10^{-9} *KS*/year, it was estimated that *A. duranensis* and *A. ipaensis* diverged ~2.16 million years ago.

Under the DPPAGSC project, the genome duplication of *A. duranensis* was compared with

that of *Medicago* and soybean. Collinear genes from *Medicago*, soybean (*Glycine max*), and grape (*Vitis vinifera*) were used to analyze related evolutionary events. The Ks distribution of peanut homologs shows a prominent peak around Ks = 0.5, overlapping the peak of soybean duplicated genes resulting from a pan-legume tetraploidization previously inferred to be ~60 Mya (Young et al. 2011). Adding the pan-eudicot γ-hexaploidy (~130 Mya) and polyploidy producing tetraploid peanut by joining the *Arachis* A and B subgenomes, estimated to have diverged 3.5 Mya (Nielen et al. 2012), the *Arachis* lineage has been affected by at least three polyploidizations since the origin of eudicots, with a collective 12X paleoduplication depth.

In addition, the gene conversion among the subgenomes was discussed in the DPPAGSC study, where there is unidirectional homeologous exchanges between genes from different subgenomes can overwrite one progenitor allele with additional copies of the other (Paterson et al. 2012; Wang et al. 2012). Implicated as a possible contributor to the transgressive properties of polyploids relative to their progenitors, extensive gene conversion was inferred to have occurred about 7500–12,500 years ago since formation of the Neolithic species *Brassica napus* (Chalhoub et al. 2014). By performing a three-way comparison of the synthetic tetraploid ISATGR 184 and its progenitor lines, ICG 8123 and ICG 8206, evidence of extensive gene conversion was observed between subgenomes in the ~ three seed-to-seed generations since its formation by human hands. The vast majority (~93%) of alleles have been converted to homozygosity for the A-genome allele in ISATGR 184, an asymmetry resembling those found in cotton and canola (Young et al. 2011, Chalhoub et al. 2014). ISATGR 1212, a reciprocal cross between the same parental lines as ISATGR 184, shared Bt to At bias of conversion but had far fewer converted sites than ISATGR 184 ($\chi 2 << 0.001$), perhaps indicating a contribution of germ-line types to genomic variation in the offspring.

9.6 Synteny with Allied and Model Genomes

IPGI study provided the syntenic relations between A and B subgenomes and their sequence comparison with tetraploid peanut. Most pseudomolecules had symmetrically positioned pericentromeres that was in accordance with cytogenetic observations (Robledo and Seijo 2010; Robledo et al. 2009). Most pseudomolecules showed a one-to-one correspondence between the two species: pairs 02, 03, 04, and 10 were collinear; pairs 05, 06, and 09 were each differentiated by a large inversion in one arm of one of the pseudomolecules; and the pseudomolecules in pair 01 were differentiated by large inversions of both arms. In contrast, chromosomes 07 and 08 have undergone complex rearrangements that transported repeat-rich DNA to A07 and gene-rich DNA to A08. As a result, A07 has only one normal (upper) euchromatic arm and A08 is abnormally small, with low repetitive content. In accordance with cytogenetic observations (Seijo et al. 2007; Nielen et al. 2010), A08 could be assigned as the characteristic small "A chromosome" (cytogenetic chromosome A09).

All *A. ipaensis* pseudomolecules were larger than their *A. duranensis* counterparts. This is partly because of a greater frequency of local duplications and higher transposon content in *A. ipaensis*. In chromosomes without inversions, there were characteristic density gradients for genes, repetitive DNA and methylation (with gene densities increasing and densities of repetitive DNA and methylation decreasing toward chromosome ends). However, in regions that had undergone large rearrangements, in *A. duranensis*, these gradients were disrupted. From these observations, we concluded that most major rearrangements occurred in the A-genome lineage. Size differences between homeologous chromosomes that were differentiated by large rearrangements tended to be greater than those between collinear ones. Because the *A. duranensis* chromosomes that have undergone inversions are smaller than expected, it is evident that,

in this dynamic, on balance, the elimination of DNA has predominated over its accumulation. Comparisons with *Phaseolus vulgaris* L., which shared a common ancestor with *Arachis* about 58 million years ago, showed syntenous chromosomal segments. In some cases, there was almost a one-to-one correspondence between chromosomes (for example, B01 and Pv03, B05 and Pv02, B06 and Pv01, and B08 and Pv05).

Sequence comparison to tetraploid peanut showed fundamentally one-to-one correspondences between the diploid chromosomal pseudomolecules and cultivated peanut linkage groups. Of the marker sequences from three maps (Shirasawa et al. 2012; Zhou et al. 2014), 83, 83, and 94% were assigned by sequence similarity searches to the expected diploid chromosomal pseudomolecules. For more detailed genome-wide comparisons, about 5.74 Gb (2X coverage) of long-sequence Moleculo reads from *A. hypogaea* cv. Tifrunner were generated and mapped the reads to the combined diploid pseudomolecules. The corrected median identities between the *A. hypogaea* Moleculo reads and the pseudomolecules of *A. duranensis* and *A. ipaensis* were 98.36 and 99.96%, respectively. When visualized as plots along the chromosomal pseudomolecules, the diploid A-genome chromosomes were distinctly less similar to *A. hypogaea* sequences than the B-genome chromosomes.

9.7 Trait Understanding

DPPAGSC project also provided insights into some unique traits found in peanut-like fructification, oil biosynthesis, and allergens. A unique characteristic of peanut is the peg/gynophore, a specialized organ that grows downwards upon fertilization, driving the developing pod into the soil. Fruit development in other plants is controlled in light; on the contrary there is subterranean fructification in peanut. A total of 151 genes related to "gravitropism" were found during pod development. Five TF families related to photomorphogenesis were identified in very large numbers in *A. duranensis*, namely S1Fa-like, FAR1, HSF, NAC, and STAT. S1Fa-like TFs containing a small peptide (70 aa) with a nuclear localization and DNA binding domain were more highly expressed in roots and etiolated seedlings than green leaves. The FAR1 TF family plays an important role in modulating phyA-signaling homeostasis in higher plants (Lin et al. 2007). Importantly, phyA localized in the cytosol of dark-grown seedlings acts primarily as a far-red sensor, which regulates the transition from skotomorphogenesis to photomorphogenesis (Whitelam and Halliday 2008). PhyB, exhibiting a fast and strong but incomplete dark conversion in some cases, is the main light receptor responsible for the shade-avoidance response in mature plants (Medzihradszky et al. 2013) and shows evidence of positive selection in *A. duranensis* suggesting a role in skotomorphogenesis.

Oil biosynthesis is one more important trait of peanut, where better understanding of this trait will be very helpful to breed confectionary suitable peanut. Considering the importance of peanut as an oil crop, annotations of 67 gene models were searched for their similarity with the genes involved in fatty acid biosynthesis and triacylglycerol (TAG), that represent the oleic and linoleic acids (Moore and Knauft 1989). FAD2 encoding δ-12 oleic acid desaturase, the key enzyme controlling the high oleate trait, was highly expressed in seed filling but less during desiccation. Genes encoding key enzymes in the TAG pathway were expressed at diverse levels at different developmental stages. Multiple copies or isoforms of some key genes were detected in the *A. duranensis* genome like glycerol-3-phosphate acyltransferase and diacylglycerol acyltransferase, which catalyze the first and final steps in the TAG pathway. Information on copy number and expression diversity of these metabolic genes is important for improvement of oil quality parameters in peanut, such as a high oleic to linoleic acid ratio (O/L).

Peanut allergy is one of the most serious life-threatening food sensitivities prevalent among a section of world population, particularly

among children. Comparison with known allergenic proteins from peanut and other crops identified 21 candidate allergen-encoding genes in *A. duranensis*, of which nine have already been reported in peanut and others are homologs from other crops. Understanding of allergen-encoding genes in peanut can be utilized to produce allergy-free peanuts either by genomics-enabled breeding or by cis-genic approaches.

9.8 Genome Dominance

Whole-genome duplications have occurred in many eukaryotic lineages, particularly in plants. Following most ancient tetraploidies, the two subgenomes are distinguishable, because the dominant subgenome tends to have more genes than the other subgenome. Additionally, among retained pairs, the gene on the dominant subgenome tends to be expressed more than its recessive homeolog (Woodhouse et al. 2014).

The most thorough study of the location and number of rDNAs was conducted by Seijo and collaborators (2004) using fluorescent in situ hybridization (FISH). The study showed, as previously mentioned, that the number, size, and distribution of rDNA clusters in *A. hypogaea* are virtually equivalent to the sum of those present in *A. duranensis* and *A. ipaënsis*. A single pair of 5S sites is present on each of the A and B chromosome complements, and two pairs of 18S-25S sites on the A chromosomes and three pairs on the B. The only exception to this equivalence is that in both of the diploid species, 18S-25S sites bear a thread-like constriction indicating intense transcriptional activity (forming the SAT chromosome; Fernandez and Krapovickas 1994). However, in the allotetraploid the constrictions are observed only on the A-genome. This indicates that the transcriptional activity of the B-genome rDNAs has been silenced, a common event in polyploids called nucleolar dominance (Cermeno et al. 1984; Preuss and Pikaard 2007).

9.9 Conclusion

The IPGI project has used the genome information to identify candidate pest- and disease-resistance genes, to reduce collapse in tetraploid transcriptome assemblies and to show the impact of recombination between subgenomes in cultivated peanut. Besides providing basic knowledge about the A-genome (*A. duranensis)* progenitor, the DPPAGSC project also provided a major source of candidate genes for fructification, oil biosynthesis, and allergens, expanding knowledge of understudied areas of plant biology and human health impacts of plants. The availability of these genomes will lead to further advances in knowledge of genetic changes since the very recent polyploidization event that gave rise to cultivated peanut and to the production of better tools for molecular breeding and crop improvement.

References

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S, Ren L et al (2016) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet 48(4):438–446

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27(4):578–579

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, Ware D et al (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol 164(2):513–524

Cermeno MC, Orellana J, Santos JL, Lacadena JR (1984) Nucleolar activity and competition (amphiplasty) in the genus *Aegilops*. Heredity 53(3):603–611

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, Corréa M et al (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345(6199):950–953

Chen X, Li H, Pandey MK, Yang Q, Wang X, Garg V, Li H, Chi X, Doddamani D, Hong Y, Upadhyaya H et al (2016) Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into

geocarpy, oil biosynthesis, and allergens. Proc Natl Acad Sci 113(24):6785–6790

Dhillon SS, Rake AV (1980) Miksche JP. Reassociation kinetics and cytophotometric characterization of peanut (*Arachis hypogaea* L.) DNA. Plant Physiol 65 (6):1121–1127

Fávero AP, Simpson CE, Valls JF, Vello NA (2006) Study of the evolution of cultivated peanut through crossability studies. Crop Sci 46(4):1546–1552

Fernández A, Krapovickas A (1994) Cromosomas Y Evolucion En" *Arachis* (Leguminosae)". *Bonplandia* 187–220

Foncéka D, Hodo-Abalo T, Rivallan R, Faye I, Sall MN, Ndoye O, Fávero AP, Bertioli DJ, Glaszmann JC, Courtois B, Rami JF (2009) Genetic mapping of wild introgressions into cultivated peanut: a way toward enlarging the genetic basis of a recent allotetraploid. BMC Plant Biol 9(1):103

Gautami B, Foncéka D, Pandey MK, Moretzsohn MC, Sujay V, Qin H, Hong Y, Faye I, Chen X, BhanuPrakash A, Shah TM et al (2012) An international reference consensus genetic map with 897 marker loci based on 11 mapping populations for tetraploid groundnut (*Arachis hypogaea* L.). PLoS One 7(7):e41213

Huang S, Chen Z, Huang G, Yu T, Yang P, Li J, Fu Y, Yuan S, Chen S, Xu A (2012) HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res 22(8):1581–1588

Jackson SA (2016) Rice: The first crop genome. Rice. doi:10.1186/s12284-016-0087-4

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R et al (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496 (7443):91–95

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20(2):265–272

Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H (2007) Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. Science 318(5854):1302–1305

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C et al (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496(7443):87–90

Liu B, Yuan J, Yiu SM, Li Z, Xie Y, Chen Y, Shi Y, Zhang H, Li Y, Lam TW, Luo R (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics 28 (22):2870–2874

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 1(1):1

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M International wheat genome sequencing consortium, et al. (2014) Ancient hybridizations among the ancestral genomes of bread wheat. Science 345

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS (2014) Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One 9 (9):e106689

Medzihradszky M, Bindics J, Adam E, Viczian A, Klement E, Lorrain S, Gyula P, Merai Z, Fankhauser C, Medzihradszky KF, Kunkel T, Schafer E, Nagy F (2013) Phosphorylation of phytochrome B inhibits light-induced signaling via accelerated dark reversion in Arabidopsis. The Plant Cell 25(2):535–544

Michael TP, Jackson S (2013) The first 50 plant genomes. The Plant Genome, 6(2)

Moretzsohn MC, Leoi L, Proite K, Guimaraes PM, Leal-Bertioli SC, Gimenes MA, Martins WS, Valls JF, Grattapaglia D, Bertioli DJ (2005) A microsatellite-based, gene-rich linkage map for the AA genome of Arachis (Fabaceae). Theor Appl Genet 111(6):1060–1071

Moretzsohn MC, Barbosa AV, Alves-Freitas DM, Teixeira C, Leal-Bertioli SC, Guimarães PM, Pereira RW, Lopes CR, Cavallari MM, Valls JF, Bertioli DJ (2009) A linkage map for the B-genome of *Arachis* (Fabaceae) and its synteny to the A-genome. BMC Plant Biol 9(1):40

Moore KM, Knauft DA (1989) The inheritance of high oleic acid in peanut. J Heredity 80:252–253

Moretzsohn MC, Gouvea EG, Inglis PW, Leal-Bertioli SC, Valls JF, Bertioli DJ (2013) A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. Ann Bot 111 (1):113–126

Nielen S, Campos-Fonseca F, Leal-Bertioli S, Guimarães P, Seijo G, Town C, Arrial R, Bertioli D (2010) FIDEL—a retrovirus-like retrotransposon and its distinct evolutionary histories in the A-and B-genome components of cultivated peanut. Chromosome Res 18 (2):227–246

Nielen S, Vidigal BS, Leal-Bertioli SC, Ratnaparkhe M, Paterson AH, Garsmeur O, D'Hont A, Guimaraes PM, Bertioli DJ (2012) Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the Arachis A-B genome divergence. Mol Genet Genomics 287(1):21–38

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492(7429):423–427

Preuss S, Pikaard CS (2007) rRNA gene silencing and nucleolar dominance: insights into a chromosome-scale epigenetic on/off switch. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1769(5):383–92

Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, Cheng J, Zhao S, Xu M, Luo Y, Yang Y et al (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci 111 (14):5135–5140

Ramos ML, Fleming G, Chu Y, Akiyama Y, Gallo M, Ozias-Akins P (2006) Chromosomal and phylogenetic context for conglutin genes in *Arachis* based on genomic sequence. Mol Genet Genomics 275(6):578–592

Robledo G, Seijo G (2010) Species relationships among the wild B genome of Arachis species (section Arachis) based on FISH mapping of rDNA loci and heterochromatin detection: a new proposal for genome arrangement. Theor Appl Genet 121(6):1033–1046

Robledo G, Lavia GI, Seijo G (2009) Species relations among wild *Arachis* species with the A genome as revealed by FISH mapping of rDNA loci and heterochromatin detection. Theor Appl Genet 118 (7):1295–1307

Samoluk SS, Chalup L, Robledo G, Seijo JG (2015) Genome sizes in diploid and allopolyploid Arachis L. species (section *Arachis*). Genet Resour Crop Evol 62 (5):747–763

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13(4):1

Seijo JG, Lavia GI, Fernandez A, Kaprovickas A, Ducasse D, Moscone EA (2004) Physical mapping of the 5S and 18S-25S RRNA gene by FISH as evidence that Arachis duranensis and A. ipaensis are the wild diploid progenitors of *A. hypogaea* (Leguminosae). Am J Bot 91:1294–1303

Seijo G, Lavia GI, Fernández A, Krapovickas A, Ducasse DA, Bertioli DJ, Moscone EA (2007) Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. Am J Bot 94 (12):1963–1971

Shirasawa K, Koilkonda P, Aoki K, Hirakawa H, Tabata S, Watanabe M, Hasegawa M, Kiyoshima H, Suzuki S, Kuwata C, Naito Y (2012) In silico polymorphism analysis for the development of simple sequence repeat and transposon markers and construction of linkage map in cultivated peanut. BMC Plant Biol 12(1):1

Shirasawa KE, Bertioli DJ, Varshney RK, Moretzsohn MC, Leal-Bertioli SC, Thudi MA, Pandey MK, Rami JF, Foncéka DA, Gowda MV, Qin HO et al (2013) Integrated consensus map of cultivated peanut and wild relatives reveals structures of the A and B genomes of Arachis and divergence of the legume genomes. DNA Res 20(2):173–184

Soltis DE, Visger CJ, Soltis PS (2014) The polyploidy revolution then and now: Stebbins revisited. Am J Bot 101(7):1057–1078

Temsch EM, Greilhuber J (2000) Genome size variation in *Arachis hypogaea* and *A. monticola* re-evaluated. Genome 43(3):449–451

Varshney RK, Nayak SN, Jackson S, May G (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27(9):522–530

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C et al (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44(10):1098–1103

Whitelam GC, Halliday KJ (2008) Annual plant reviews, light and plant development. John Wiley & Sons

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc Natl Acad Sci 111(14):5283–5288

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 480(7378):520–524

Zhou X, Xia Y, Ren X, Chen Y, Huang L, Huang S, Liao B, Lei Y, Yan L, Jiang H (2014) Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). BMC Genom 15(1):1