Sequencing of Cultivated Peanut, *Arachis hypogaea*, Yields Insights into Genome Evolution and Oil Improvement

Xiaoping Chen1,12,*, Qing Lu1,12, Hao Liu1,12, Jianan Zhang2,12, Yanbin Hong1,12, Haofa Lan3, Haifen Li1, Jinxeng Wang1, Haiyan Liu1, Shaohong Li1, Manish K. Pandey5, Zhikang Zhang4, Guiyuan Zhou1, Jigao Yu4, Guoqiang Zhang6, Xiangyu Liu1, Shijie Wen1, Fanbo Meng4, Shanlin Yu7, Xiyin Wang4, Kadambot H.M. Siddique8, Zhong-Jian Liu9,10,*, Andrew H. Paterson11,*, Rajeev K. Varshney5,*, and Xuanqiang Liang1,*

1South China Peanut Sub-center of National Center of Oilseed Crops Improvement, Guangdong Key Laboratory for Crops Genetic Improvement, Crops Research Institute, Guangdong Academy of Agricultural Sciences (GAAS), Guangzhou, China
2National Foxtail Millet Improvement Center, Minor Cereal Crops Laboratory of Hebei Province, Institute of Millet Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, China
3MolBreeding Biotechnology Co., Ltd., Shijiazhuang, China
4School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, China
5Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India
6Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Center of China and Orchid Conservation and Research Center of Shenzhen, Shenzhen, China
7Shandong Peanut Research Institute, Shandong Academy of Agricultural Sciences, Qingdao, China
8UWA Institute of Agriculture, The University of Western Australia, Crawley, Australia
9Key Laboratory of National Forestry and Grassland Administration for Orchard Conservation and Utilization at College of Landscape Architecture, Fujian Agriculture and Forestry University, Fuzhou 350002, China
10Fujian Colleges and Universities Engineering Research Institute of Conservation and Utilization of Natural Bioresources, College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China
11Plant Genome Mapping Laboratory, University of Georgia, Athens, USA
12These authors contributed equally to this work.

*Correspondence: Xiaoping Chen (chenxiaoping@gdaas.cn), Zhong-Jian Liu (zjliu@fafu.edu.cn), Andrew H. Paterson (paterson@uga.edu), Rajeev K. Varshney (r.k.varshney@cgiar.org), Xuanqiang Liang (liangxuanqiang@gdaas.cn)

https://doi.org/10.1016/j.molp.2019.03.005

ABSTRACT

Cultivated peanut (*Arachis hypogaea*) is an allotetraploid crop planted in Asia, Africa, and America for edible oil and protein. To explore the origins and consequences of tetraploidy, we sequenced the allotetraploid *A. hypogaea* genome and compared it with the related diploid *Arachis duranensis* and *Arachis ipaensis* genomes. We annotated 39 888 A-subgenome genes and 41 526 B-subgenome genes in allotetraploid peanut. The *A. hypogaea* subgenomes have evolved asymmetrically, with the B subgenome resembling the ancestral state and the A subgenome undergoing more gene disruption, loss, conversion, and transposable element proliferation, and having reduced gene expression during seed development despite lacking genome-wide expression dominance. Genomic and transcriptomic analyses identified more than 2 500 oil metabolism-related genes and revealed that most of them show altered expression early in seed development while their expression ceases during desiccation, presenting a comprehensive map of peanut lipid biosynthesis. The availability of these genomic resources will facilitate a better understanding of the complex genome architecture, agronomically and economically important genes, and genetic improvement of peanut.

Key words: cultivated peanut, *de novo* sequencing, comparative genomics, genome evolution, oil metabolism

INTRODUCTION

Cultivated peanut (Arachis hypogaea), belonging to the Fabaceae or Leguminosae family, is a New World crop that was disseminated to Europe, Africa, Asia, and the Pacific Islands by early explorers (Hammons, 1973). China and India together account for more than 50% of the world’s total peanut production (FAOSTAT, 2017). Cultivated peanut is an allotetraploid (AABB, 2n = 4x = 40) thought to be derived from hybridization between the diploids A. duranensis (A genome) and A. ipaensis (B genome) (Smartt et al., 1978; Seijo et al., 2007; Robledo et al., 2009), which have recently been sequenced (Bertioli et al., 2016; Chen et al., 2016; Lu et al., 2018). It is necessary to sequence the allotetraploid species to fully understand peanut evolution and trait biology (e.g., oil synthesis).

Evidence from a number of sources suggests that peanut was domesticated at least 3500 years ago and cultivated and selected ever since (Singh and Simpson, 1994; Simpson et al., 2001; Dillehay et al., 2007; Grabiele et al., 2012). Peanut domestication has resulted in highly modified plant architecture and seed size, and striking changes in yield, but a lack of genetic diversity (Milla et al., 2005). Although A. duranensis and A. ipaensis are the putative donor species for the A and B chromosome groups, respectively, tetraploid peanut species differ greatly with respect to plant morphology as well as economic characteristics, including oil content, protein content, and disease resistance. Peanut oil, composed mainly of triacylglycerol (TAG), is obtained from pressing the kernel cotyledons and provides nutrients required for development (A and B chromosome groups, respectively, tetraploid peanut cv. Fuhuasheng, a mid-twentieth century landrace from North China (Supplemental Figure 1), by performing whole-genome shotgun sequencing using Illumina HiSeq and PacBio technologies combined with BioNano genome mapping, and organized the assembled sequences into chromosomes using high-density genetic maps (Supplemental Figure 2, Supplemental Information, and Methods). We generated 700 Gb (~260X genome equivalents) of high-quality Illumina and Chromium data (Supplemental Table 1), which were assembled using DenovoMAGiC2 (NRGene, Nes Ziona, Israel), yielding a 2.53-Gb assembly containing 491 scaffolds with a contig N50 of 47.91 kb and a scaffold N50 of 31.82 Mb (Table 1 and Supplemental Table 6). A genetic linkage map constructed using an F2 population of 108 individuals derived from a cross between Fuhuasheng and Shitoqiu, another mid-twentieth century landrace from South China (Supplemental Tables 7 and 8), permitted us to assign >98.31% of the assembled sequences (and 98% of the gene content) to chromosomal locations; 77 scaffolds (from 806 kb to 160 Mb in size) were organized into 20 chromosomal regions.

RESULTS

Genome Sequencing and Assembly

We sequenced the genome of the allotetraploid peanut (A. hypogaea) cultivar Fuhuasheng, a mid-twentieth century landrace from North China (Supplemental Figure 1), by performing whole-genome shotgun sequencing using Illumina HiSeq and PacBio technologies combined with BioNano genome mapping, and organized the assembled sequences into chromosomes using high-density genetic maps (Supplemental Figure 2, Supplemental Information, and Methods). We generated 700 Gb (~260X genome equivalents) of high-quality Illumina and Chromium data (Supplemental Table 1), which were assembled using DenovoMAGiC2 (NRGene, Nes Ziona, Israel), yielding a 2.53-Gb assembly containing 491 scaffolds with a contig N50 of 47.91 kb and a scaffold N50 of 31.82 Mb (Table 1 and Supplemental Table 6). A genetic linkage map constructed using an F2 population of 108 individuals derived from a cross between Fuhuasheng and Shitoqiu, another mid-twentieth century landrace from South China (Supplemental Tables 7 and 8), permitted us to assign >98.31% of the assembled sequences (and 98% of the gene content) to chromosomal locations; 77 scaffolds (from 806 kb to 160 Mb in size) were organized into 20 chromosomal regions.
Molecular Plant

pseudomolecules (Supplemental Figure 3), with nine unplaced scaffolds (289 kb to 14 Mb). This final assembly spanned ~96.7% (~2.64 Gb) of the estimated allotetraploid genome (Supplemental Figure 4 and Supplemental Table 9), and 1.16 Gb (44 scaffolds) and 1.35 Gb (33 scaffolds) were assigned to the \(A\) and \(B\) subgenomes (the subscript “t” indicates tetraploid), respectively (Supplemental Table 10); these sizes are close to those of the diploid progenitors, \(A\). \(duranensis\) and \(A\). \(ipaensis\) (Bertioli et al., 2016). With few gaps and high coverage, this assembly provides a high-quality reference with high physical resolution for whole-genome analyses of allotetraploid peanut.

Assessment of the Assembly Quality

The completeness and accuracy of the assembled genome was assessed using various approaches. Sequencing data from a 250 bp paired-end (PE) library were properly mapped onto the genome assembly, and the mean insert size was 234 bp (STD = 28), which is close to the expected library insert size (250 bp) (Supplemental Information and Supplemental Figure 5). Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al., 2015) and Core Eukaryotic Gene Mapping Approach (CEGMA) analyses (Parra et al., 2007) were performed, and >95.5% of BUSCOs and KOGs were found in the genome assembly (Supplemental Tables 11 and 12). The correlation between the number of full-length long terminal repeat (LTR) retrotransposons and genome size (Supplemental Figure 6) supported the completeness of the genome assembly (Paterson et al., 2009; Avni et al., 2017; Mascher et al., 2017). Approximately 78% of the RNA sequencing (RNA-seq) data from roots, stems, flowers, leaves, and pods matched the genome assembly (Supplemental Figure 7 and Supplemental Table 13). The accuracy of the assembly was assessed using bacterial artificial chromosomes (BACs) retrieved from GenBank, and 99% of BACs aligned properly (Supplemental Figure 8 and Supplemental Table 14).

Gene Content and Repetitive Nature of the \(A\). \(hypogaea\) Genome

We predicted 108 604 gene models in the \(A\). \(hypogaea\) genome and annotated 83 087 genes with high confidence (HC) by combining ab initio prediction, homologous protein data searches, and transcriptome alignment (Table 1 and Supplemental Table 15).

Comparative Genomic and Phylogenetic Analyses

Among the 83 087 HC \(A\). \(hypogaea\) genes, ~98% were homologous with those of other plant species, covering ~99% of genes in \(A\)- and \(B\)-progenitor genomes (Supplemental Tables 21 and 22). We identified 22 110 orthologous gene groups in 18 diverse plant species using OrthoMCL (Li et al., 2003), including 6367 commonly shared gene families and 1946 peanut-specific families consisting of 6926 genes, which was the largest number of species-specific gene families (Figure 2A; Supplemental Figure 14; Supplemental Tables 23 and 24). A total of 15 071 gene families were common to \(A\). \(hypogaea\) and its two progenitors (Supplemental Figure 15). In addition, a total of 10 064 gene families were common to five leguminous species (Supplemental Figure 16), while 9370 gene families were shared between \(A\). \(hypogaea\) and other distantly related plant species (Supplemental Figure 17). A species tree based on single-copy orthologous genes indicated that \(A\). \(hypogaea\) and its progenitors form a single clade not including any other legume species, which is consistent with the phylogenetic placement of these species (Figure 2B and Supplemental Figure 18). We compared the two diploid genomes and classified the genes/families into different classes, finding 22 699 gene families shared by \(A\). \(duranensis\) and \(A\). \(ipaensis\) and 1668 \(A\)-genome-specific and 2758 \(B\)-genome-specific gene families (Supplemental Tables 25). Among the genes in the shared class, 15 827 were retained in the \(A\) and \(B\) subgenomes. In addition, we also found that 1984 gene families were present in the tetraploid but in neither wild diploid genome.

Molecular Evolutionary History of the Allotetraploid \(A\). \(hypogaea\)

The evolutionary relationships between \(A\). \(hypogaea\) and representative \(Arachis\) (\(A\). \(duranensis\) and \(A\). \(ipaensis\)), legume (soybean (tRNAs), and 25 299 small nuclear RNAs (snRNAs) (Table 1 and Supplemental Table 18).

A total of 5161 putative transcription factor (TF) genes from 58 families were identified, representing 6.21% of HC genes, a higher percentage than that in \(A\). \(duranensis\) and \(A\). \(ipaensis\), but slightly lower than that in soybean (Supplemental Table 19). Strikingly, the FAR1 TF families were expanded in \(A\). \(hypogaea\) (Supplemental Figure 10) and its wild progenitors (Chen et al., 2016; Lu et al., 2018). This \(Arachis\)-specific expansion may be related to geocarpy, a prominent feature in the \(Arachis\) genus, considering the important role of the FAR1 TF family in modulating phyA-signaling homeostasis and of phyB in regulation of skotomorphogenesis and photomorphogenesis in higher plants (Medzhiradszky et al., 2013).

We annotated 54.34% of the \(A\). \(hypogaea\) genome as repeat regions (Table 1 and Supplemental Table 20), which is comparable with the percentage observed in pigeonpea (51.6%) (Varshney et al., 2011). LTR retrotransposons account for 52.3% of the \(A\). \(hypogaea\) genome, with one major burst of amplification occurring around 1–2 million years ago (Mya) (Supplemental Figure 11) and with Gypsy repeats being most abundant, followed by Copia (Supplemental Figure 12 and Supplemental Table 20). Most \(A\). \(hypogaea\) transposable element sequences had a divergence rate of ~20% (Supplemental Figure 13).

Approximately 99% of the genes in the shared class, 15 827 were retained in the \(A\) and \(B\) subgenomes. In addition, we also found that 1984 gene families were present in the tetraploid but in neither wild diploid genome.
and Medicago truncatula), and eudicot (grape and Theobroma cacao) species were evaluated by measuring the synonymous nucleotide substitution rate ($K_s$) of orthologous gene pairs. The distribution of these rates suggests that A. hypogaea experienced the core eudicot paleohexaploidy event shared with grape and T. cacao (Tang et al., 2008), a more recent pan-legume duplication event with legume species, such as soybean and M. truncatula (Young et al., 2011), as well as one duplication shared by the closely related Arachis species before tetraploidization, consistent with previous reports (Bertioli et al., 2016; Chen et al., 2016). This suggests that there were at least three whole-genome duplication (WGD) events in the evolutionary history of A. hypogaea together with the production of a tetraploid by the joining of the A_t and B_t subgenomes (Figure 3A). The origin of modern cultivated peanut A. hypogaea (AABB) was proposed to be the result of an initial hybridization of A. duranensis (AA) and A. ipaensis (BB) followed by chromosome doubling (Seijo et al., 2007; Grabiele et al., 2012). The A_t and B_t chromosome sets of A. hypogaea therefore represent the descendants of the two diploid progenitors,

Figure 1. Overview of Arachis hypogaea Genome.
From the outer edge inward, circles represent (1) the 20 chromosomal pseudomolecules, (2) gene density, (3) long terminal repeat density, (4) positions of oil synthesis genes, and gene expression levels in the (5) root, (6) stem, (7) pod, (8) leaf, and (9) flower. Central colored lines represent syntenic links between the A_t and B_t subgenomes.
confirming the allotetraploid hypothesis. Analysis of synonymous divergence suggests that the A1 and B1 subgenomes diverged from each other around 2 Mya ($K_s$ peak at 0.03), similar to the divergence time between the A- and B-progenitor genomes (Figure 3A and 3B). We estimated that the A. duranensis-A1 divergence occurred around 0.25 Mya ($K_s$ peak at 0.004) and that the A. ipaensis-B1 divergence occurred around 0.18 Mya ($K_s$ peak at 0.003); this is inconsistent with the previous estimation (Bertioli et al., 2016) and thus constrains the allotetraploid event to <0.18 Mya considering exchange between the two subgenomes and the inflation of $K_s$ estimation (Figure 3C). The newly formed polyploid peanut may have occasionally outcrossed to other A-genome diploids, decreasing the divergence between the A1 subgenome in A. hypogaea and its original donor genomes. Uneven distributions of $K_s$ values were observed between the subgenomes and their suspected progenitor genomes (Supplemental Figure 19).

Comparison of the peanut genomes with the seven ancestral protochromosomes derived from grape (Jaillon et al., 2007) suggested that paleopolyploidy was commonly shared at orthologous loci from the ancestor to A. hypogaea and its progenitors, A. duranensis and A. ipaensis (Figure 3D). The modern genomes of Arachis included at least four to seven ancestral chromosomal fragments with chromosome 02 containing four fragments from both the A and B (sub)genomes. Different fragments between the progenitor genomes and the two subgenomes were observed in chromosomes 07 and 10. In the remaining chromosomes the same number of ancestral chromosomes were retained between the subgenomes and the progenitor genomes.

Synteny analysis provided a robust and precise sequence framework for understanding A. hypogaea genome evolution, and revealed a high number of syntenic blocks between A. hypogaea and its progenitors without large chromosome rearrangements (Figure 3E). Additionally we identified syntenic disruptions between A. hypogaea and its wild progenitors, especially on chromosomes 07 and 08, implying possible complex rearrangements after tetraploidyization (Figures 1 and 3E). There are more syntenic blocks between A. duranensis-B, than between A. ipaensis-A, while larger fragment exchanges are observed in chromosome A07 (Supplemental Figure 20).

Asymmetric Evolution of the Two Subgenomes of A. hypogaea

The non-synonymous ($K_s$) and synonymous substitution rates ($K_a$) were calculated by comparing genes in the A1 and B1 subgenomes with those in their corresponding A- and B-progenitor genomes (Figure 4A). The different $K_s$ and $K_a$ rates suggest that the A1 gene sets might have evolved faster than the B1 gene sets. The assembled A1 subgenome (1159 Mb) was larger than its corresponding A-progenitor genome (1068 Mb), while the assembled B1 subgenome (1349 Mb) was almost equal in size to its B-progenitor genome (1349 Mb) (Bertioli et al., 2016) (Figure 4B). In addition to WGD, mobile element proliferation contributes to the evolution of plant genome size. Analysis of genome composition demonstrated that transposable elements, especially those in the Gypsy lineage, were the main contributors to differences in genome size, with a higher proportion of transposable elements (TEs) in the A1 subgenome (52.11%) than in the A-progenitor genome (42.07%) (Figure 4B and Supplemental Table 26). Peaks in LTR retrotransposons are footprints of these insertion events, demonstrating a major burst of amplification in all four genome sets around 1–2 Mya, and revealing an additional burst only in the A1 subgenome, which contributed to a larger genome size than that of the suspected progenitor (Figure 4C). Strikingly, this A1-specific activation of TEs occurred around ~0.2 Mya before/during allopolyploid formation, indicating that the two subgenomes independently asymmetrically evolved, implying the possibility of another wild A-genome diploid as the donor for the A1 subgenome of A. hypogaea, or multiple hybridization events of the B progenitor, A. ipaensis, with different varieties of A. duranensis, to form the present-day cultivated peanut (Zhang et al., 2016). Asymmetric evolution was also reflected by the genomic signature of selection; there were 694 positively selected genes (PSGs), with significantly more PSGs in the A1 subgenome (395 PSGs) than in the B1 subgenome (299 PSGs, $P < 0.01$, Fisher’s exact test; Figure 4D). Interestingly, 335 (85%) and 239 (80%) PSGs were A. duranensis-specific and A. ipaensis-specific genes, respectively, implying that specific genes have undergone more stringent positive selection.
Gene Loss and Conversion

Gene loss was not significantly different between homoeologous subgenomes in the allotetraploid peanut, with 187 (185 genes only present in $A$. duranensis) and 171 (169 genes only present in $A$. ipaensis) genes lost in the $A_t$ and $B_t$ subgenomes, respectively ($P > 0.1$, Fisher’s exact test, Figure 4D), implying that those genes shared by the two diploids were more conserved and rarely lost during natural evolution. Like some other polyploids (Schnable et al., 2011; Zhang et al., 2015), more than 29,301 genes were disrupted by frameshifts or premature stop codons in $A$. hypogaea compared with their orthologous genes. Particularly, there were significantly more disrupted genes in the $A_t$ subgenome (14,839) than in the $B_t$ subgenome (14,462) ($P < 0.01$, Fisher’s exact test, Figure 4D). Among the 29,301 disrupted genes, 8,603 were shared by the two diploids, and 6,236 and 5,723 were $A$. duranensis- and $A$. ipaensis-specific genes, respectively. The recent origin of allotetraploid peanut may be reflected in the higher number disrupted genes than
lost genes, implying future gene loss (Schnable et al., 2011; Zhang et al., 2015). However, approximately 48%–57% gene translocation/loss rates were identified in the A_t and B_t subgenomes if large chromosomal segment exchanges were also considered (Supplemental Tables 27 and 28). Multiple consecutive genes were found to be lost on a few chromosomal segments (Supplemental Figure 21).

Extensive gene conversion, a possible contributor to the transgressive properties of polyploids relative to their progenitors, has occurred as recently as ~12 500 years ago (Chalhoub et al., 2014). By performing a quartet comparison between the four related (sub)genomes from tetraploid A. hypogaea and its two suspected progenitors, as many as 66.73% of alleles were found to be non-reciprocal exchanges between A_t and B_t homeologues at the single-nucleotide scale (Figure 4E and Supplemental Table 29). There are 3747 A_t genes and 640 B_t genes harboring at least two conversion sites. Reciprocal exchanges between the A_t and B_t subgenomes account for 25.3% of these sites, with A_t genes converted to B_t alleles at more than five times the rate (21.27%) of the conversion of B_t genes to A_t alleles (4.03%), which is opposite to the results from a comparison of a synthetic tetraploid peanut line and its parents, in which conversions from the B_t to the A_t subgenome was far more common (>60% B2A and ~4% A2B) (Chen et al., 2016). The contrary results suggest that DNA sequence changes in the allotetraploid progeny of artificial crosses are completely different from those in natural allotetraploids, implying the difficulty of mimicking the speciation processes of natural evolution. Analysis of the expression of those genes with allelic changes indicated that genes with different allelic changes had different average expression levels, but the numbers of expressed genes were similar between the A_t and B_t subgenomes in different tissues and developmental stages (Supplemental Figure 22).

Analysis of Homoeologous Genes in Allotetraploid A. hypogaea

Polyploidy has played a prominent role in shaping peanut genomic architecture, with an array of evolutionary processes acting on duplicate genes. Of the A. hypogaea HC genes, we identified 19 961 and 20 206 orthologous groups from the A_t and B_t subgenomes, respectively, in A. hypogaea and the suspected A- and B-progenitor genomes. Of these, 12 951 A_t genes correspond 1:1 with A. duranensis genes, and 13 286 B_t genes correspond 1:1 with A. ipaensis genes (Supplemental Tables 30 and 31). We found that 15 827 orthologous gene groups between A. duranensis and A. ipaensis were conserved in A. hypogaea. On the basis of orthologous groups and the best reciprocal BLAST matches between the A_t and B_t subgenomes, we identified 16 403 gene pairs (referred to as homoeologous duads) consisting of 32 806 genes in A. hypogaea that had a 1:1 correspondence across the two homoeologous subgenomes. The biological functions of these homoeologous duads were explored by performing Gene Ontology (GO) analysis, and the homoeologous duads were assigned to a total of 347 biological process GO categories, including 306 and 280 for A_t and B_t homeologs, respectively (Supplemental Table 32). Furthermore, GO enrichment analysis suggested no significant functional divergence in the A_t and B_t homoeologous genes, with 67 A_t-preferred categories and 41 B_t-preferred (P = 0.0153; Figure 5A and 5B).

Unequal expression of homoeologous genes in allopolyploids can be an important feature and consequence of polyploidization (Grover et al., 2012; Wu et al., 2018), although little divergence in gene function and genome-wide expression dominance were observed between the two homoeologous Arachis subgenomes.
In total, \( \sim 57\% \) of HC genes were expressed (fragments per kilobase of exon model per million mapped reads [FPKM] \( \geq 1 \)) in roots, stems, leaves, flowers, and multiple pod developmental stages (Supplemental Figure 23 and Supplemental Table 33), consistent with findings in hexaploid bread wheat (Pfeifer et al., 2014). The \( A_t \) and \( B_t \) subgenomes contribute about equally to the number of expressed genes (28.4% and 28.5%, respectively). Of the homeologs, 23,427 were expressed, and 2% and 2.1% were

**Figure 5. Analysis of Homoeologous *A. hypogaea* Genes.**

(A) Significantly enriched biological process Gene Ontology (GO) categories (green, \( A_t \) subgenome; purple, \( B_t \) subgenome). Color intensity reflects significance of enrichment, with darker colors corresponding to lower \( P \) values. Circle radii depict the size of aggregated GO terms.

(B) Ratio of gene numbers in each enriched GO biological process category: green, \( A_t \)-preferred GO enriched categories; purple, \( B_t \)-preferred GO enriched categories; and black, equivalent GO enrichment in \( A_t \) and \( B_t \) genes. Detailed GO analysis information is provided in Supplemental Table 32.

(C) Numbers of expressed genes and average expression levels in different tissues and seed developmental stages. The left \( y \) axis represents the number of expressed genes shown in bars. In each group of bars, the left and right bar represents the number of expressed \( A_t \) and \( B_t \) genes, respectively. The right \( y \) axis represents the average expression levels shown in the line chart.

(D) The number of homoeologous genes expressed (orange outside arc) or not expressed (gray outside arc), and their distribution in the \( A_t \) and \( B_t \) subgenomes.

(E) Box plot of \( \log_2(\text{AFPKM}/\text{BFPKM}) \) values for co-expression of homoeologous gene pairs. The central line in each box plot indicates the median. The red line represents an equal ratio, \( \log_2(1) \).
Figure 6. Evolution and Expression of Oil Biosynthesis-Related Genes.

(A) Allelic changes between A and B genes related to oil metabolism.

(B) Venn diagram showing shared and unique gene families among five representative oilseed crops.

(C) Identification of four temporal expression patterns of oil metabolism-related genes across peanut seed development using StepMiner: one-step-up (K1, expression level transition from low to high in two consecutive developmental stages), one-step-down (K2, transition from high to low),

(legend continued on next page)
A<sub>r</sub> and B<sub>r</sub>-preferred homoeologs, respectively, which is a non-significant difference (Figure 5D and Supplemental Table 34). Most peanut homoeologous duads (of A and B genome copies) showed balanced expression patterns, with only 6.2% (2029) considered to have biased expression in different tissues and developmental stages; there was a slight preference toward the B<sub>r</sub> subgenome with 1027 B<sub>r</sub>- and 1002 A<sub>r</sub>-biased homoeologs (Figure 4D). We found a similar distribution for A<sub>r</sub> and B<sub>r</sub>-biased homoeologs with approximately 15% of biased genes present only in A. <i>hypogaea</i>, and ~75% shared by the two diploids A. <i>duranensis</i> and A. <i>ipaensis</i>. When homoeologous gene duads were both expressed, the average expression level of the B<sub>r</sub> copy was similar to that of the A<sub>r</sub> copy in roots, stems, leaves, flowers, and whole pods, but slightly higher across seed developmental stages (Figure 5E).

**Analysis of Oil Metabolism Genes and Biosynthesis Pathway**

Peanut oil, which is mainly composed of TAG, provides nutrients required for human health. In the <i>A. hypogaea</i> genome, a total of 2559 genes were found to be involved in fatty acid carbon flux and lipid storage on the basis of sequence identity, pathway membership, and enzyme code (Supplemental Table 35). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis showed that 1918 (~75%) of the 2559 genes were classified into metabolism process with lipid metabolism highly represented, in agreement with the GO annotations (Supplemental Figure 24). The two homoeologous subgenomes contributed almost equally to the number of enriched biological process GO categories (Supplemental Figure 23). Distribution of oil-related gene loci along the 20 <i>A. hypogaea</i> pseudomolecules was uneven, tending to cluster near distal chromosome regions (Figure 1). Analysis of subcellular localization indicated that ~91% of oil metabolism genes were located in four organelles including the nucleus, chloroplast, cytoplasm, and plastid where fatty acids are synthesized (Supplemental Figure 26). About 93% of single-nucleotide sites matched the sequences of their diploid progenitors, with 2.24% representing an exchange from A<sub>r</sub> to B<sub>r</sub> and 1.73% representing the reverse (Figure 6A). The K<sub>r</sub>/K<sub>s</sub> ratios for oil metabolism genes suggested that only 12 (0.5%) are under positive selection including a gene encoding pyruvate dehydrogenase (PDH), the catalytic enzyme in the first step of de novo fatty acid biosynthesis (Supplemental Figure 27; Supplemental Tables 36 and 37). These results suggest little direct selection on oil metabolism genes, with conservation of oil metabolism in oilseed plants. This was also reflected by a comparison of orthologous groups between the three <i>Arachis</i> species, which revealed almost no specific orthologous groups (Supplemental Figure 28). To gain insight into differences in the genic repertoire of peanut oil metabolism genes, we classified gene families in five oilseed plants, namely peanut, soybean, sunflower, cotton, and rape, and found that 2263 (~90%) of 2559 peanut oil metabolism genes were shared with at least one oilseed species, despite at least 50 Mya of divergence from peanut (Figure 6B).

Analysis of genome-wide expression profiles using hierarchical clustering showed 1767 (~70%) of 2559 oil metabolism genes to be expressed during peanut seed development, with three major clusters representing relatively high expression at early (cluster K2 including P0 and P1), intermediate (cluster K3 including P2SD to P7SD), and late (cluster K1 including P8SD, P9SD, and P10SD) stages; 593 oil metabolism genes were consistently co-expressed during all seed developmental stages (Supplemental Figure 29 and Supplemental Table 38). We also identified 26 A<sub>r</sub>-biased and 34 B<sub>r</sub>-biased oil metabolism genes with biased expression between the two subgenomes (Supplemental Table 35). The number of expressed genes gradually increased, from 1170 (P0) to 1404 (P4SD), and then gradually decreased to 879 (P10SD) (Supplemental Figure 30). To further characterize the temporal expression patterns of these genes throughout peanut seed development, we used StepMiner (Sahoo et al., 2007) to identify four typical temporal expression patterns with one or two transition points involving 1031 oil metabolism genes (Figure 6C and Supplemental Table 39). Most of oil metabolism genes increase their expression at P2SD before seed filling, but decrease/cease expression at the desiccation stage (P10SD) (Figure 6C and Supplemental Figure 31).

Considering the importance of TAG in peanut, which mainly corresponds to oleic and linoleic acids and constitutes ~80% of peanut oil (Moore and Knauft, 1989), we next manually examined the presence of 267 genes encoding 34 crucial lipid biosynthesis enzymes, including those involved in de novo fatty acid synthesis, elongation, and TAG assembly (Figure 6D and Supplemental Table 40), clustering near the ends of chromosomes (Supplemental Figure 33). Most members in a few enzyme-encoding gene families (MCMT, FATB, CK, CCT, DAGTA, and FAD2) showed low expression during seed development (Figure 6D). The genome-based phylogeny allowed us to characterize oil biosynthesis genes from genomic and evolutionary angles. We investigated the oil biosynthesis gene repertoires of <i>A. hypogaea</i> in comparison with the A and B progenitors, and identified a PDH gene showing evidence of positive selection (Figure 6D and Supplemental Figure 27) and one KASI and two ER genes lost in the B<sub>r</sub> subgenome (Figure 6D and Supplemental Figure 33).

A protein–protein interaction network based on 267 lipid genes involved in TAG assembly according to their GO assignment was predicted using Cytoscape (www.cytoscape.org) (Supplemental Figure 29). Two-step-up/down (K3, transition from low to high and then back down over a series of developmental stages), and two-step-down/up (K4, transition from high to low and then back up). The number of genes in each cluster is indicated in parentheses. The scale color bar is shown above.

**D** Lipid biosynthesis pathway including de novo fatty acid synthesis and elongation, and TAG synthesis. A total of 267 genes were placed in the pathway, including 212 expressed during peanut seed development. One PDH-encoding gene (ArGene057704, in red) was found to be under positive selection. A 21 bp insertion in this gene is shown. Detailed alignment information for this gene is provided in Supplemental Figure 27. Genes encoding ER and KASI enzymes lost from the B<sub>r</sub> subgenome are indicated in red. Numbers of A<sub>r</sub> and B<sub>r</sub> genes encoding enzymes are shown in parentheses (the first is the number for A<sub>r</sub>, the second is the number for B<sub>r</sub>). Beside the enzymes are expression heatmaps of enzyme-encoding genes, with rows representing genes and columns representing 11 seed developmental stages (from left to right: P0, P1, P2SD, P3SD, P4SD, P5SD, P6SD, P7SD, P8SD, P9SD, and P10SD).
Molecular Plant

Figure 34 and Supplemental Table 41. Depending upon the degree of correlation, an enclosed circular protein-protein interaction network was constructed based on 83 core genes, of which 38 were extrapolated to directly interact with at least 30 target proteins. Interestingly, six enzymes executing the function of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) contained the maximum number of protein interaction pairings, implying that GAPDH family genes occupy a central position in the TAG formation pathway.

Conclusion

Formation of allotetraploid peanut appears to have been more complex than a single hybridization. Asymmetrical evolution has occurred in the two peanut subgenomes, with the B subgenome more consistently resembling the ancestral condition and the A subgenome undergoing more TE amplification, gene loss and conversion, and rearrangement, suggesting that A. duranensis might not be the single/direct A-genome donor as previously expected, or that multiple hybridizations of A. ipaensis with several varieties of A. duranensis contributed to the formation of the allotetraploid. There is stage- and individual-dependent but no global subgenome dominance between the two A. hypogaea subgenomes despite more sequence and structural variations in the A1 subgenome. Genome-wide expression analysis suggested that genes encoding key enzymes in the lipid biosynthesis pathway were expressed at diverse levels at different peanut seed developmental stages. We also found evidence of positive selection and loss of lipid biosynthesis genes. This will affect the improvement of oil traits and contribute to edible oil security. The extensive datasets and analyses presented in this study provide a framework that facilitates the development of strategies to improve peanut by manipulating individual or multiple homoeologs.

METHODS

Plant Materials

Detailed information about the landrace, Fuhuasheng, used in this project is provided in Supplemental Information. In brief, Fuhuasheng was collected by a farmer in 1944 in Yantai of Shandong province, North China. The genotype is suitable for high-quality genome sequence due to its wide utilization as a parent in breeding programs. Approximately 80% of cultivars developed in China during the past half century were directly or indirectly derived from Fuhuasheng.

DNA Extraction and Sequencing

Genomic DNA was isolated from the leaves of a peanut cultivar (cv. Fuhuasheng) using a previously described method (Doyle and Doyle, 1990) and used to construct libraries. Five size-selected genomic DNA libraries ranging from 470 bp to 10 kb were constructed. One PE library was made using DNA fragments ~470 bp in size with no PCR amplification (PCR-free). This no-PCR library was used to produce reads of approximately 265-520 bp in length. These reads were selected to produce an overlap of the fragments, which were sequenced on the HiSeq2500 v2 in Rapid mode as 2 × 265 bp reads. One 800 bp genomic library was prepared using the TruSeq DNA Sample Preparation Kit version 2 with no PCR amplification (PCR-free) according to the manufacturer’s protocol (Illumina, San Diego, CA). To increase sequence diversity and genome coverage, we constructed three mate-pair (MP) libraries with 2–5, 5–7, and 7–10 kb jumps using the Illumina Nextera Mate-Pair Sample Preparation Kit (Illumina). The 800 bp shotgun library was sequenced on an Illumina HiSeq2500 platform as 2 × 160 bp reads (using Illumina v4 chemistry), while the MP libraries were sequenced on the HiSeq4000 platform as 2 × 150 bp reads.

In addition, high molecular weight DNA was prepared, and the quality of the DNA samples was verified by pulsed-field gel electrophoresis. DNA fragments longer than 50 kb were used to construct one Gemcode library using the Chromium instrument (10X Genomics, Pleasanton, CA). This library was sequenced on the HiSeqX platform to produce 2 × 150 bp reads. Construction and sequencing of PE and MP libraries were conducted at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign.

Genome Size Estimation

The distribution of K-mer frequencies was used to estimate the genome size according to the formula: Genome size = k-mer_num/peak_depth, where the k-mer_num is the total number of k-mers in the sequence and the peak_depth is the k-mer depth value obtained from the distribution map. In this study, the modified K-mer number is 71 254 450 241 and an obvious repeat peak was observed at 27. Consequently, the peanut genome size was estimated to be ~2.64 Gb.

Genome Assembly

De novo genome assembly was conducted using the DeNovoMAGIC2 software platform (NRGene, Nes Ziona, Israel), which is a DeBrujink-graph-based assembler, designed to efficiently extract the underlying information in the raw reads to solve the complexity in the DeBriujin graph due to genome polyploidy, heterozygosity, and repetitive-ness. This task is accomplished using accurate-read-based traveling in the graph that iteratively connects consecutive phased contigs over local repeats to generate long phased scaffolds (Lu et al., 2015; Hirsch et al., 2016; Avni et al., 2017; Luo et al., 2017; Zhao et al., 2017). The additional raw Chromium 10X data were utilized to phase polyploidy/heterozygosity, support scaffold validation, and further elongate the phased scaffolds. This resulted in a first assembly with a total size of ~2.53 Gb. PacBio sequencing data was obtained to fill gaps and elongate the outputted scaffolds using the PBJelly2 pipeline (English et al., 2012). The scaffolds were ordered into super-scaffolds using SSPACE-LongRead (Boetzer and Pirovano, 2014) with a second set of PacBio sequencing data. To further improve the quality of the genome assembly, we assembled a BioNano map using the IrysView v2 software package (BioNano Genomics, CA, USA). The genome assembled using the BioNano approach spanned ~2.55 Gb and had a larger scaffold N50 value of ~56.57 Mb, and the longest scaffold was ~160 Mb. The detailed assembly procedure is provided in Supplemental Information.

Pseudomolecule Chromosome Construction

Genetic linkage maps were constructed for anchoring the improved scaffolds to 20 chromosomes using 108 F2 individuals derived from the cross of “Fuhuasheng” and “Shitouqi," which have been widely used as parents in China. Genotyping was performed using a custom-designed 37K SNP Panel (our unpublished data). JoinMap4.0 was used to construct the genetic linkage map with the default parameter set (Stam, 1993). Linkage group identification was performed using a logarithm of odds score of 10, and the scaffold order was determined using the ALLMAPS tool (Tang et al., 2015). Finally, a complete set of 20 pseudochromosomes of A. hypogaea cv. Fuhuasheng was obtained, with chromosomes 1–10 corresponding to the A1 subgenome A and chromosomes 11–20 to the B1 subgenome.
Gene Prediction and Functional Annotation

To annotate the A. hypogaea genome, we used de novo gene prediction, a homology-based strategy, and RNA-seq data to predict gene structures, and integrated these results into a final gene model using the automated genome annotation pipeline MAKER (Cantarel et al., 2008). (1) Protein sequences of nine genomes, including M. truncatula, chickpea, soybean, common bean, Vigna radiate, Vigna angularis, A. duranensis, A. ipaensis, and Arabidopsis thaliana were aligned to the A. hypogaea genome to perform homology-based gene prediction. (2) For transcript evidence, high-quality transcripts from iso-seq were polished using Illumina RNA-seq reads and aligned to the genome using GMAP (Wu et al., 2016). A total of 372,851 transcripts were identified using the pbttranscript model of SMRTLink with the following parameters: -c 0.9 -i 0.95. Moreover, a set of 276,968 transcripts were obtained using HISAT2 (Kim et al., 2015) and StringTie (Pertea et al., 2015) to assemble the Illumina RNA-seq data. (3) Integration for gene prediction was performed using AUGUSTUS software (Stanke et al., 2006). All the predicted protein sequences were aligned to the non-redundant protein, GO, KEGG, and UniProtKB databases using BLASTP with a threshold E value of 1E-20. All predictions were filtered using the pbttranscript model of SMRTLink with the following parameters: -c 0.9 -i 0.95. The protein identity between the predicted and original gene was >40% with a coverage >20%, the gene loss was considered a false-positive event and was filtered out from the gene loss list.

Annotiation of Repetitive DNA

Repetitive sequences were detected and classified by performing homology searches using RepeatMasker-open-4.0.7 (http://www.repeatmasker.org) against the RepeatMasker combined database: Dfam Consensus-20170127 (Hubley et al., 2016) and RepBase-20170127 (http://www.girinst.org/). Full-length LTR retrotransposons were identified using LTRharvest (Ellingshaus et al., 2008) and clustered using CD-HIT (Li and Godzik, 2006) with 90% sequence similarity and 90% coverage of the shorter sequence. The following parameter settings were used for LTRharvest: -overlaps best -seed 30 -minlenl 100 -maxlenl 2000 -mindistl 3000 -maxdistl 25000 -similar 85 -mintsd 20 -motif tga -motifms f1 -vic 60 -xdrop 5 -mat 2 -mins 2 -ins 3 -del 3. The LTRharvest output was annotated for PfamA domains (Pfam31.0, http://pfam.xfam.org/)(Finn et al., 2016) with PfamScan. The sequence divergence rate was calculated between the identified TE families in the peanut genome and the consensus sequence in the TE library (Repbase, http://www.girinst.org/repbase). Insertion ages of the LTRs were calculated by measuring the divergence of the 5’ and 3’ regions of the LTRs, with identity at the time of transposition and using a mutation rate of 1.3 × 10^-8 mutations per site per year.

Identification of Homologous and Orthologous Gene Sets

An OrthoMCL clustering program was employed to detect orthologous gene families in the A. hypogaea genome and 17 other plant species. A total of 1946 homologous groups containing 6926 genes specific to the A. hypogaea genome were identified, and a total of 4135 single-copy orthologous groups were identified as positively selected according to the Fisher’s test (probability > 0.01). A few genes under positive selection were aligned to their orthologues using PRANK (Loytynoja, 2014) and the alignments were visualized using PRANKSTER.

Molecular Plant

Synthetic and K_a Analysis

Synthetic blocks were identified using MCSVScanX with default parameters (Wang et al., 2012). Gene CDSs were used as queries in searches against the genomes of other plant species to find the best matching pairs. Each aligned block represented a conserved orthologous pair derived from the common ancestor. K_a (the number of synonymous substitutions per synonymous site) values of the homologs within collinear blocks were calculated using the Nei-Gojobori method. Genes with an FDR < 0.05 and at least one amino acid site possessing a high probability of being positively selected (Bayes probability > 0.95%) were considered positively selected. These genes were considered to be positively selected according to the Fisher’s test (P < 0.01, FDR < 0.05). A few genes under positive selection were aligned to their orthologues using PAGOL (Loytynoja, 2014) and the alignments were visualized using PRANKSTER.

Divergence Time

The divergence times of the two subgenomes of A. hypogaea and their wild progenitors (A. duranensis and A. ipaensis) were estimated based on synonymous substitution rates (K_a), which were calculated between all three Arachis species. The formula t = K_a/2r, where r is the neutral substitution rate, was used to estimate the divergence time between sister species. A neutral substitution rate of 8.12 × 10^-8 was used in the current study (Bertioli et al., 2016).

Phylogenetic Tree Construction and Evolution Rate Estimation

Along with the two subgenomes of A. hypogaea, 18 species were used to build gene families. These species included 13 eudicots (A. duranensis, A. ipaensis, Ricinus communis, Lotus japonicus, M. truncatula, Glycine hypogaea, A. hypogaea).

Molecular Plant 12, 920–934, July 2019 © The Author 2019.
Molecular Plant

max, Cajanus cajan, Crocus sativus, Malus domestica, T. cacao, A. thaliana, Vitis vinifera, Solanum lycopersicum), three monocots (Zea mays, Orizya sativa, Musa acuminate), and an outgroup (Amborella trichopoda). Orthologous groups were identified using OrthoFinder (v2.2.3) (Emms and Kelly, 2015), and 123 single-copy orthologous genes were used to build an ML tree using FastTree (v2.1.9) (Price et al., 2010). This ML tree was converted to an ultrametric time-scaled phylogenetic tree by r8s (Sanderson, 2003) using the calibrated times from the TimeTree (Kumar et al., 2017) website. Changes in gene family size along the phylogenetic tree were analyzed by CAPE (v4.1) (De Bie et al., 2006). Evolutionary rates were estimated using the codeml program in PAML under the free-ratio “branch” model that allows distinct evolutionary rates on each branch (Yang, 2007). The phylogenetic tree was reconstructed using the maximum-likelihood algorithm implemented in MEGA X (Kumar et al., 2018).

Expression Bias of Homoeologs

Protein-coding genes from the A t and B t subgenomes of A. hypogaea were employed as queries in a BLAST search against each other. The best reciprocal hits with > 80% of identity, an E-value cutoff of ≤1E–30, and an alignment accounting for > 80% of the shorter sequence were obtained as gene pairs between A t and B t subgenomes. To investigate the expression bias of these paired homoeologs from the two subgenomes, we calculated the FPKM values of the homoeologs in the root, stem leaf, flower, whole pod, and 11 seed developmental stages. A t > B t indicated biased expression of the A homoeolog and B t > A t indicated biased expression of the B homoeolog.

ACCESSION NUMBERS

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SDMP00000000. The version described in this paper is version SDMP01000000. Sequence data for A. hypogaea transcribeatns are available in the NCBI Sequence Read Archive under accession numbers SRP167797 and SRP033292.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at Molecular Plant Online.

FUNDING

This project was supported by the National Natural Science Foundation of China (31501246, 31711841, 31810401), the National Natural Science Foundation of Guangdong Province (2017A030311007), the Modern Agroindustry Technology Research System (CARS-14), the Science and Technology Planning Project of Guangdong Province (2015B020231006, 2015A020209051, 2016B020201003, 2016LM3161, 2016LM3164, 2014A020208060 and S201320012647), the International Science & Technology Cooperation Program of Guangdong Province (2013B050800021), the Agricultural Science and Technology Program of Guangdong (2013B020301014), and the teamwork projects funded Guangdong Natural Science Foundation of Guangdong Province (no. 2017A030312004).

AUTHOR CONTRIBUTIONS


ACKNOWLEDGMENTS

No conflict of interest declared.
Genome Evolution of Cultivated Peanut


