CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

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

Xiaoping Chen[1,12,*], Qing Lu[1,12], Hao Liu[1,12], Jianan Zhang[2,12], Yanbin Hong[1,12], Haofa Lan[3], Haifen Li[1], Jinpeng Wang[4], Haiyan Liu[1], Shaoxiong Li[1], Manish K. Pandey[5], Zhikang Zhang[4], Guiyuan Zhou[1], Jigao Yu[4], Guoqiang Zhang[6], Jiaqing Yuan[4], Xingyu Li[1], Shijie Wen[1], Fanbo Meng[4], Shanlin Yu[7], Xiyin Wang[4], Kadambot H.M. Siddique[8], Zhong-Jian Liu[9,10,*], Andrew H. Paterson[11,*], Rajeev K. Varshney[5,*] and Xuanqiang Liang[1,*]

[1]South 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

[2]National Foxtail Millet Improvement Center, Minor Cereal Crops Laboratory of Hebei Province, Institute of Millet Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, China

[3]MolBreeding Biotechnology Co., Ltd., Shijiazhuang, China

[4]School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, China

[5]Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

[6]Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Center of China and Orchid Conservation and Research Center of Shenzhen, Shenzhen, China

[7]Shandong Peanut Research Institute, Shandong Academy of Agricultural Sciences, Qingdao, China

[8]UWA Institute of Agriculture, The University of Western Australia, Crawley, Australia

[9]Key Laboratory of National Forestry and Grassland Administration for Orchid Conservation and Utilization at College of Landscape Architecture, Fujian Agriculture and Forestry University, Fuzhou 350002, China

[10]Fujian Colleges and Universities Engineering Research Institute of Conservation and Utilization of Natural Bioresources, College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China

[11]Plant Genome Mapping Laboratory, University of Georgia, Athens, USA

[12]These 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

Chen X., Lu Q., Liu H., Zhang J., Hong Y., Lan H., Li H., Wang J., Liu H., Li S., Pandey M.K., Zhang Z., Zhou G., Yu J., Zhang G., Yuan J., Li X., Wen S., Meng F., Yu S., Wang X., Siddique K.H.M., Liu Z.-J., Paterson A.H., Varshney R.K., and Liang X. (2019). Sequencing of Cultivated Peanut, *Arachis hypogaea*, Yields Insights into Genome Evolution and Oil Improvement. Mol. Plant. **12**, 920–934.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

Category	Number	Size (Mb)	N50 (Mb)	Longest (Mb)	Gap (%)
Scaffolds (HiSeq)	491	2532	31.82	132.98	1.59
Scaffolds (HiSeq + PacBio)	486	2578	32.67	134.59	0.40
Scaffolds (HiSeq + PacBio + BioNano)	86 (77[a])	2552	56.57	160.08	1.04
HC gene models	83 087	355 (13.92%)[b]			
miRNA	241	0.03 (<0.01%)			
rRNA	3511	1.16 (0.04%)			
tRNA	2239	0.17 (<0.01%)			
snRNA	25 299	2.71 (0.11%)			
Repeat sequences	–	1387 (54.34%)			

Table 1. Statistic Summary of *A. hypogaea* Genome Assembly and Annotation.

[a] Anchored scaffolds.

[b] Percentage of the assembly indicated in parentheses.

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 human health. Approximately 80% of peanut TAGs consist of monounsaturated oleic acid (C18:1) and polyunsaturated linoleic acid (C18:2). One of the most important vegetable oils worldwide, peanut oil does not contribute to the *trans*-isomer content of foods, but has been shown to lower low-density lipoprotein cholesterol levels in the blood (Knauft and Ozias-Akins, 1995). The fusion of two diploid progenitors isolated peanut reproductively from other wild species, partly resulting in the paucity of genetic diversity. A whole-genome sequence of cultivated peanut, together with the recently-sequenced genomes of its two wild progenitors, might overcome these difficulties (Bertioli et al., 2016; Chen et al., 2016; Lu et al., 2018).

Here we report a genome assembly of the allotetraploid peanut cv. Fuhuasheng, a widely used parent from which ~70% of Chinese peanut cultivars released during the past half century have been derived. We used the peanut genome to assess phylogenetic relationships with other legume and oilseed crops, and to compare transcriptome data among different organs (root, stem, leaf, and seed) and seed developmental stages. This assembly was compared with the genomes of its two suspected progenitors for understanding the possible paths for genome evolution and species divergence. This *A. hypogaea* genome assembly provides a high-quality chromosome-scale reference for analysis of the evolution and biology of agronomic traits.

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 2). To reduce fragmentation, we used PacBio sequencing data for self-correction and assembly, which allowed us to improve the genome assembly and capture ~2.58 Gb in 486 scaffolds with a contig N50 of 211 kb (Supplemental Tables 3 and 4). Super-scaffolding using BioNano genome map data (Supplemental Table 5) yielded a high-quality assembly of 2.55 Gb comprising 86 scaffolds with an N50 of 56.57 Mb (Table 1 and Supplemental Table 6). A genetic linkage map constructed using an F_2 population of 108 individuals derived from a cross between Fuhuasheng and Shitouqi, 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 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_t and B_t 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). The number of genes we identified is comparable with those in other polyploid species such as upland cotton, oilseed rape, and bread wheat, which have 76 943, 101 040, and 124 201 gene models, respectively (Chalhoub et al., 2014; International Wheat Genome Sequencing Consortium, 2014; Li et al., 2015). Of the 83 087 HC genes, 81 414 (97.98%) were assigned to a chromosomal location, including 39 888 in the A_t subgenome and 41 526 in the B_t subgenome (Supplemental Table 10); these genes were unevenly distributed along the chromosomes with a distinct preference for the ends (Figure 1). The average gene length (4275 bp), coding sequence length (226 bp with 4.23 exons), and intron length (578 bp) were similar to those of other plant species (Supplemental Table 16). The average GC content was 36.33% (Supplemental Table 15), consistent with that of the two wild relatives, but different from that of other plant species (Supplemental Figure 9). Approximately 99% of HC genes matched entries in at least one publicly available database (Supplemental Table 17). We also annotated 241 microRNAs (miRNAs), 3511 ribosomal RNAs (rRNAs), 2239 transfer RNAs (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 (Medzihradszky 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).

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_t and B_t 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

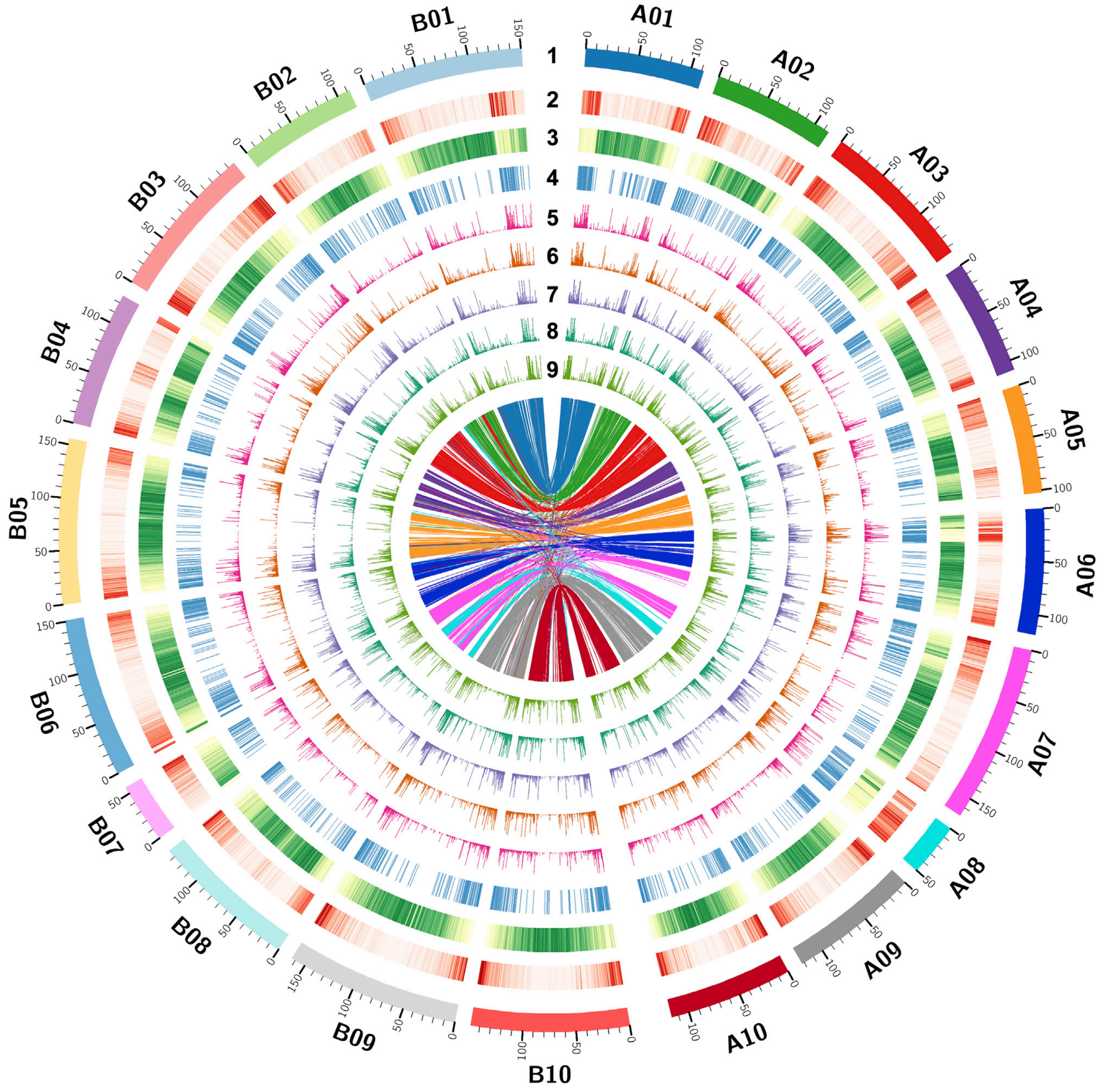

Figure 1. Overview of *Arachis hypogaea* Genome.
From the outer edge inward, circles represent (1) the 20 chromosomal pseudomolecules, (2) gene density, (3) long terminal repeat density, (4) positions of oil synthesis genes, and gene expression levels in the (5) root, (6) stem, (7) pod, (8) leaf, and (9) flower. Central colored lines represent syntenic links between the A_t and B_t subgenomes.

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,

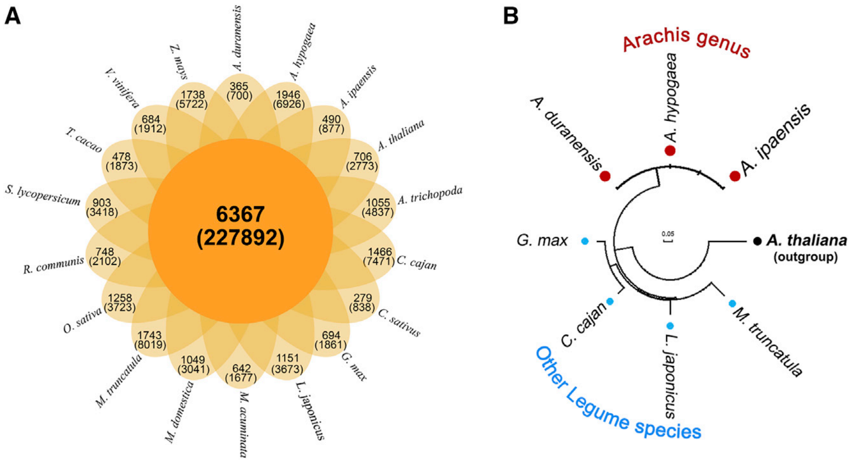

Figure 2. Orthologous Gene Families and Phylogenetic Tree.

(A) Venn diagram showing shared and unique gene families among *A. hypogaea* and other plant species.

(B) Topology of a phylogenetic tree for *A. hypogaea*, wild relatives, and other legumes with *Arabidopsis* as the outgroup. The *Arachis* genus was placed on an individual clade separate from other legume species. A detailed phylogenetic tree for 18 species is provided in Supplemental Figure 18.

confirming the allotetraploid hypothesis. Analysis of synonymous divergence suggests that the A_t and B_t 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*-A_t divergence occurred around 0.25 Mya (K_s peak at 0.004) and that the *A. ipaensis*-B_t 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 A_t 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 tetraploidization (Figures 1 and 3E). There are more syntenic blocks between *A. duranensis*-B_t than between *A. ipaensis*-A_t, while larger fragment exchanges are observed in chromosome A07 (Supplemental Figure 20). Interestingly, comparison of *A. hypogaea* and *Arachis monticola*, a wild tetraploid *Arachis* species, revealed a high level of synteny between the A subgenome of *A. hypogaea* and the B subgenome of *A. monticola*, and a similar result was observed for the other two subgenomes.

Asymmetric Evolution of the Two Subgenomes of *A. hypogaea*

The non-synonymous (K_a) and synonymous substitution rates (K_s) were calculated by comparing genes in the A_t and B_t subgenomes with those in their corresponding A- and B-progenitor genomes (Figure 4A). The different K_a and K_s rates suggest that the A_t gene sets might have evolved faster than the B_t gene sets. The assembled A_t subgenome (1159 Mb) was larger than its corresponding A-progenitor genome (1068 Mb), while the assembled B_t 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 A_t 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 A_t subgenome, which contributed to a larger genome size than that of the suspected progenitor (Figure 4C). Strikingly, this A_t-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 A_t 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 A_t subgenome (395 PSGs) than in the B_t 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.

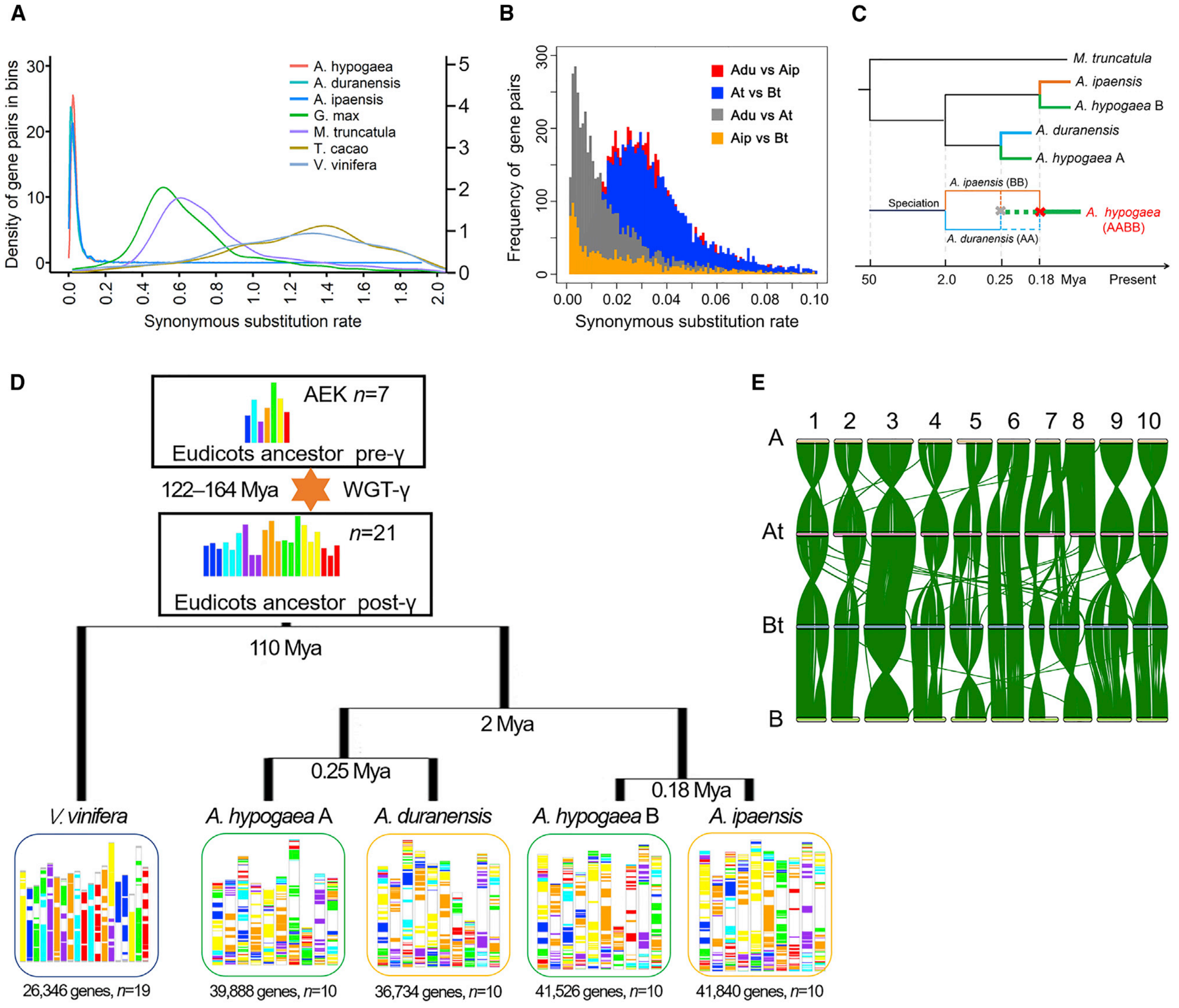

Figure 3. *A. hypogaea* Genome Evolution.

(A) Distribution of synonymous nucleotide substitutions (K_s) for orthologous gene pairs in *A. hypogaea* and its suspected progenitors (*A. duranensis* and *A. ipaensis*), legumes (soybean and *M. truncatula*), and eudicots (*T. cacao* and grape). The scale on the left *y* axis is for the three *Arachis* species, and that on the right *y* axis is for the other four species.

(B) Distribution of K_s values for orthologous genes between the A_t subgenome, B_t subgenome and their suspected progenitor genomes.

(C) Schematic diagram of a phylogenetic tree illustrating different epochs in the evolutionary history of *Arachis* (*M. truncatula* as the outgroup). We date the speciation of *A. duranensis* and *A. ipaensis* at around 2 Mya. The hybridization occurred <0.18 Mya.

(D) Evolutionary scenario of cultivated peanut and its suspected wild progenitors descended from the ancestral eudicot karyotype (AEK, $n = 7$) of eudicot protochromosomes. Colored blocks within modern chromosomes illustrated at the bottom represent the chromosomal regions originating from the seven ancestral chromosomes (top). Numbers denote the predicted divergence times (million years, Mya).

(E) Syntenic analysis between the two homoeologous subgenomes (indicated by A_t and B_t) of *A. hypogaea*, the A-progenitor genome (indicated by A) and the B-progenitor genome (indicated by B). The subscript "t" indicates tetraploid.

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, 8603 were shared by the two diploids, and 6236 and 5723 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

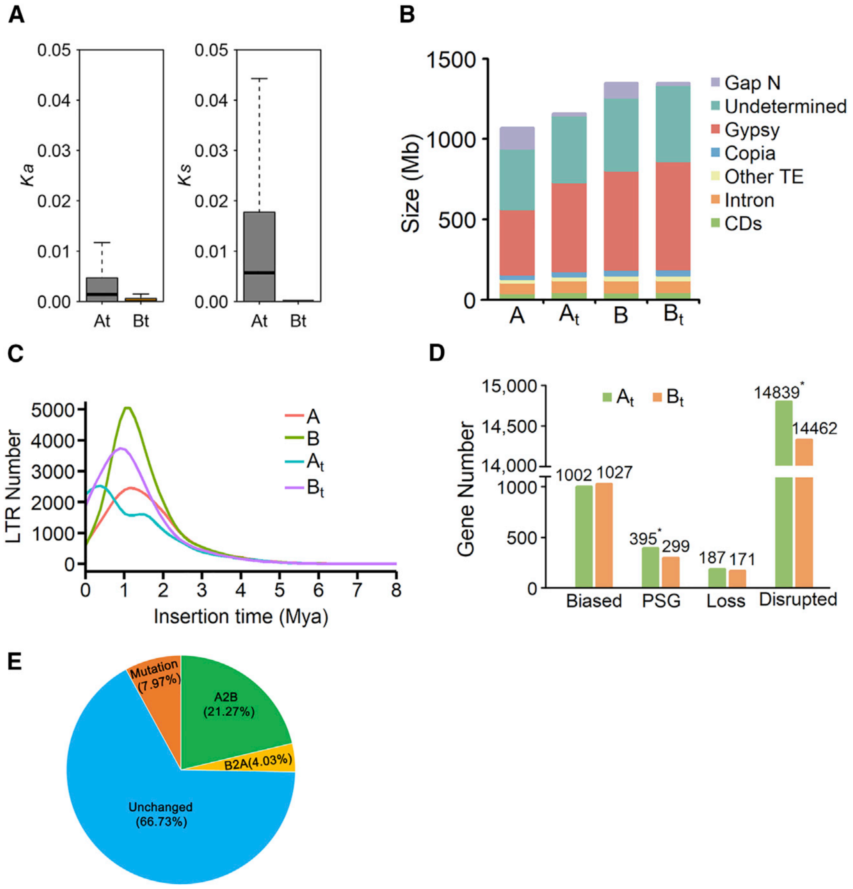

Figure 4. Asymmetric Evolution of *A. hypogaea* Homoeologous Subgenomes.

(A) Comparison of K_a and K_s distributions in the A_t and B_t subgenomes.

(B) Composition of the *A. duranensis* genome (A), *A. ipaensis* genome (B), and the two *A. hypogaea* subgenomes (A_t and B_t). The A_t subgenome has a substantially increased level of RNA transposons compared with its progenitor.

(C) Dating of LTR retrotransposon insertions in the A_t and B_t subgenomes of *A. hypogaea* and the suspected A- and B-progenitor genomes.

(D) Comparison of the number of expression-biased genes, positively selected genes (PSG), lost genes, and disrupted genes in the two subgenomes. *$P < 0.01$ according to Fisher's exact test.

(E) Allelic changes at the single-nucleotide scale between the A_t and B_t subgenomes of *A. hypogaea*.

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 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 homoeologs, 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

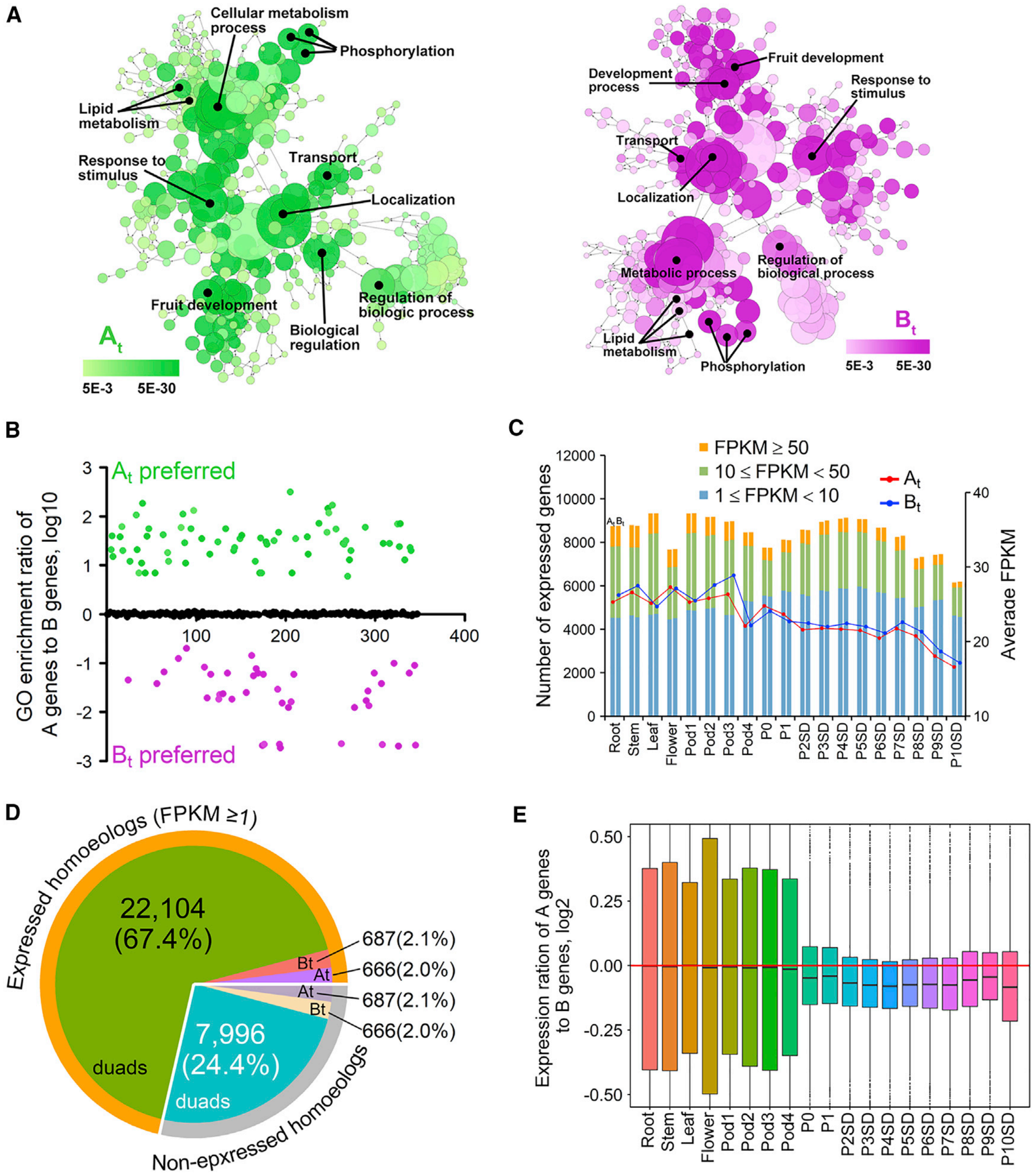

Figure 5. Analysis of Homoeologous *A. hypogaea* Genes.

(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 $\log_2(A_{FPKM}/B_{FPKM})$ 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 5C). In total, ~57% of HC genes were expressed (fragments per kilobase of exon model per million mapped reads [FPKM] ≥ 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, 23 457 were expressed, and 2% and 2.1% were

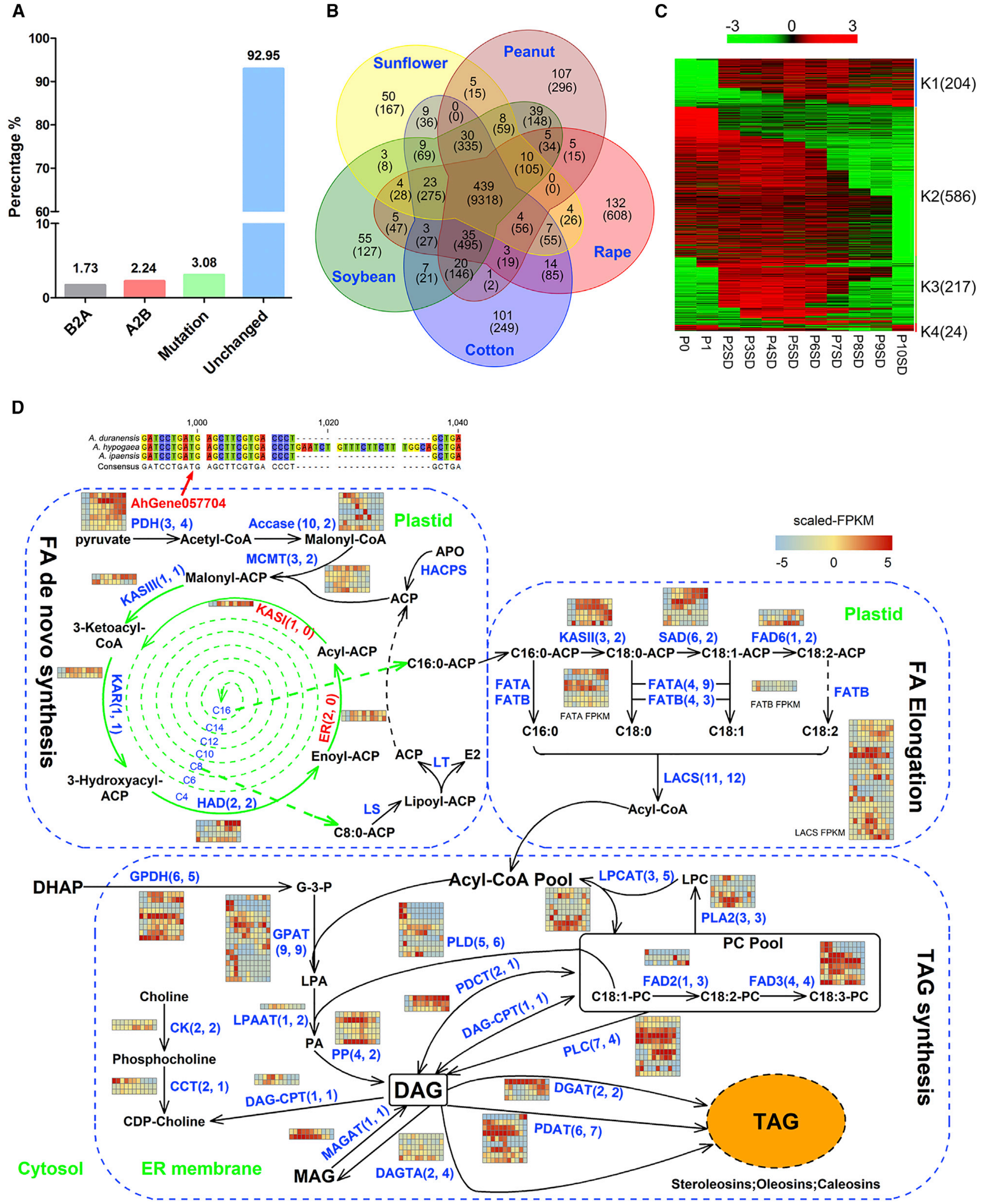

Figure 6. Evolution and Expression of Oil Biosynthesis-Related Genes.

(A) Allelic changes between A_t and B_t 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_t- and B_t-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_t subgenome with 1027 B_t- and 1002 A_t-biased homoeologs (Figure 4D). We found a similar distribution for A_t- and B_t-biased homoeologs with approximately 15% of biased genes present only in *A. hypogaea*, and ~75% shared by the two diploids *A. duranensis* and *A. ipaensis*. When homoeologous gene duads were both expressed, the average expression level of the B_t copy was similar to that of the A_t 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 *A. hypogaea* 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 25). Distribution of oil-related gene loci along the 20 *A. hypogaea* 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_t to B_t and 1.73% representing the reverse (Figure 6A). The K_a/K_s 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 *Arachis* 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_t-biased and 34 B_t-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 befoe 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 32). 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 *A. hypogaea* 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_t 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

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 (AhGene057704, 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_t subgenome are indicated in red. Numbers of A_t and B_t genes encoding enzymes are shown in parentheses (the first is the number for A_t, the second is the number for B_t). 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).

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 glycerol-3-phosphate acyltransferase (GPAT) contained the maximum number of protein interaction pairings, implying that GPAT 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_t subgenome more consistently resembling the ancestral condition and the A_t 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 A_t 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 assembly because it is highly homozygous as a result of the many generations of self-fertilization that occurred during the domestication process. We selected this landrace for *de novo* sequencing because of 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. The 10X Chromium library construction and sequencing were conducted at HudsonAlpha Institute for Biotechnology, Huntsville, AL.

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 DeBruijn-graph-based assembler, designed to efficiently extract the underlying information in the raw reads to solve the complexity in the DeBruijn graph due to genome polyploidy, heterozygosity, and repetitiveness. 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 A_t subgenome A and chromosomes 11–20 to the B_t 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 pbtranscript 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−10. A gene detected in at least one database was considered to be an HC gene.

Annotation 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 (Ellinghaus 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 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000 -maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -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 TEs 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 Homoeologous 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 *A. hypogaea* were identified, and a total of 4135 single-copy orthologous genes sets were found between the *A. hypogaea* genome and 17 other plant species. Protein-coding genes from the A_t and B_t subgenomes were used as queries in BLAST searches against each other. Gene pairs that were the best reciprocal BLAST hits between the two subgenomes were extracted. 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 that had a 1:1 correspondence across the two homoeologous subgenomes.

Transcription Factor Annotation

To detect known TFs in the *A. hypogaea* genome, we used the Plant Transcription Factor Database (PlantTFDB) (http://planttfdb.cbi.pku.edu.cn/) to identify TFs in other plant species. The predicted gene sets were then used as queries in searches against the database. Finally, a total of 5161 putative TFs, belonging to 58 families and representing 4.75% of the predicted protein-coding genes, were identified.

Positively Selected Genes and Gene Loss

The branch of phylogenetic tree corresponding to the *A. hypogaea* genome was used as the foreground branch and the other branches were used as background branches to detect PSGs under the branch-site model (Zhang et al., 2005) incorporated in PAML (Yang, 2007). The null model assumed that the K_a/K_s values for all codons in all branches must be ≤ 1, whereas the alternative model was $K_a/K_s > 1$. A maximum-likelihood ratio was then used to compare the two models. Next, the *P* values calculated by performing a chi-square test (df = 1) were adjusted for multiple testing using the false discovery rate (FDR) method. Genes with an FDR < 0.05 and at least one amino acid site possessing a high probability of being positively selected (Bayes probability >95%) were considered positively selected. These genes were identified as 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 PRANK (Loytynoja, 2014) and the alignments were visualized using PRANKSTER.

Gene losses were identified from the synteny table using the flanking gene strategy. The primary steps were as follows. (1) Given three flanking genes M, K, and N in order, if all the three genes are present in the two progenitors and subgenome A, when the M and N genes are present in the B subgenome without the K gene, then the K gene is considered as a gene loss event in the B subgenome. A similar process was used to identify gene loss events in the A subgenome. (2) Given the potential for false-positive identifications, the intergenic sequence between the M and K genes was extracted, and pseudogenes in this region were predicted using the B protein sequence and GeneWise software (Birney and Durbin, 2000). If 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.

Samples for RNA-Seq and Gene Expression

RNA-seq was performed using the Illumina Hiseq X10 platform to obtain a comprehensive transcriptome profile. Samples for RNA-seq included the root, stem, flower, leaf, and whole pod (including four developmental stages) (Supplemental Figure 7 and Supplemental Table 13). In total, ~270.48 Gb clean of RNA-seq data were generated, and the average percentage of mapped reads was as high as 77.86% (Supplemental Table 13). RNA-seq reads were remapped to the reference genome assembly using Bowtie2 (Langmead and Salzberg, 2012) with the following parameters: –no-mixed –no-discordant –gbar 1000 –end-to-end -k 200 -q -X 800, and the FPKM was calculated to evaluate the expression level of each gene using the RSEM tool (Li and Dewey, 2011).

Syntenic and K_s Analysis

Syntenic blocks were identified using MCScanX with default parameters (Wang et al., 2012). Gene CDS were used as queries in searches against the genomes of other plant species to find the best matching pairs. Each aligned block represented an orthologous pair derived from the common ancestor. K_s (the number of synonymous substitutions per synonymous site) values of the homologs within collinear blocks were calculated using the Nei-Gojobori approach implemented in PAML (Yang, 2007), and the median of K_s values was considered to be representative of the collinear blocks.

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_s), which were calculated between all three *Arachis* species. The formula $t = K_s/2r$, where r is the neutral substitution rate, was used to estimate the divergence time between two subgenomes. A neutral substitution rate of 8.12×10^{-9} 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* *max*, *Cajanus cajan*, *Crocus sativus*, *Malus domestica*, *T. cacao*, *A. thaliana*, *Vitis vinifera*, *Solanum lycopersicum*), three monocots (*Zea mays*, *Oryza sativa*, *Musa acuminata*), and an outgroup (*Amborella trichopoda*). Orthogroups 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 CAFE (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 $\geq$80% of identity, an E-value cutoff of $\leq$1E$-$30, and an alignment accounting for $\geq$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* transcriptome analyses 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, 31771841, 31801401), the 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 S2013020012647), 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

Conceptualization, X.C., Z.-J. L., A.H.P., R.K.V., and X. Liang; Methodology, Y.H., H. Li, Haiyan Liu, S.L., G. Zhou, G. Zhang, X. Li, and S.W.; Investigation, X.C., Q.L., Hao Liu, J.Z., Y.H., M.K.P., X.W., H. Li, J.W., H. Lan, Haiyan Liu, S.L., G. Zhou, G. Zhang, X. Li, S.W., S.Y., and Z.-J. L.; Formal Analysis, X.C., Q.L., Hao Liu, J.Z., Y.H., M.K.P., Z.Z., X.W., H. Li, J.W., H. Lan, Haiyan Liu, S.L., G. Zhou, J. Yu, G. Zhang, J. Yuan, X. Li, S.W., F.M., S.Y., Z.-J. L., X.W., and K.H.M.S.; Writing – Original Draft, X.C., Q.L., Hao Liu, and J.Z.; Writing – Review & Editing, Z.-J. L., A.H.P., R.K.V., and X. Liang; Supervision, X.C., Z.-J. L., A.H.P., R.K.V., and X. Liang.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: January 25, 2019
Revised: February 21, 2019
Accepted: March 10, 2019
Published: March 19, 2019

REFERENCES

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science **357**:93–97.

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

Birney, E., and Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. Genome Res. **10**:547–548.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics **15**:211.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18**:188–196.

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

Chen, X., Li, H., Pandey, M.K., Yang, Q., Wang, X., Garg, V., Li, H., Chi, X., Doddamani, D., Hong, Y., 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. U S A **113**:6785–6790.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics **22**:1269–1271.

Dillehay, T.D., Rossen, J., Andres, T.C., and Williams, D.E. (2007). Preceramic adoption of peanut, squash, and cotton in northern Peru. Science **316**:1890–1893.

Doyle, J.J., and Doyle, J.L. (1990). Isolation of plant DNA from fresh tissue. Focus (Gico-BRL) **12**:13–15.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9**:18.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. **16**:157.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One **7**:e47768.

FAOSTAT. (2017). Production Crops Data. http://www.fao.org/faostat/en/#data/QC.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. **44**:D279–D285.

Grabiele, M., Chalup, L., Robledo, G., and Seijo, G. (2012). Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. Plant Syst. Evol. **298**:1151–1165.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol. **196**:966–971.

Hammons, R.O. (1973). Early history and origin of the peanut. In Peanuts—Cultures and Uses, A.J. St. Angelo, F.S. Arant, M.H. Bass, G.A. Buchanan, W.Y. Cobb, F.R. Cox, J.M. Davison, J.W. Dickens, U.L. Diener, and K.H. Garren, et al., eds. (Stillwater, OK: American Peanut Research and Education Society), pp. 17–45.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., et al. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell **28**:2700–2714.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. **44**:D81–D89.

International Wheat Genome Sequencing Consortium. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science **345**:1251788.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al.; Consortium for Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**:463–467.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360.

Knauft, D., and Ozias-Akins, P. (1995). Recent methods for gerrnplasm enhancement and breeding. In Advances in Peanut Science, H.E. Pattee and H.T. Stalker, eds. (Stillwater: APRES), pp. 54–94.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. **35**:1547–1549.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. **34**:1812–1819.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**:357–359.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**:323.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., Wu, J., et al. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat. Biotechnol. **33**:524–530.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**:1658–1659.

Loytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol. Biol. **1079**:155–170.

Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Y., Li, Y., Semagn, K., Zhang, X., et al. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. Nat. Commun. **6**:6914.

Lu, Q., Li, H., Hong, Y., Zhang, G., Wen, S., Li, X., Zhou, G., Li, S., Liu, H., Liu, H., et al. (2018). Genome sequencing and analysis of the peanut B-genome progenitor (Arachis ipaensis). Front. Plant Sci. **9**:604.

Luo, M.C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., Deal, K.R., Huo, N., Zhu, T., Wang, L., Wang, Y., et al. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature **551**:498–502.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature **544**:427–433.

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

Milla, S.R., Isleib, T.G., and Stalker, H.T. (2005). Taxonomic relationshipsamong *Arachis* sect. *Arachis* species as revealed by AFLP markers. Genome **48**:1–11.

Moore, K.M., and Knauft, D.A. (1989). The inheritance of high oleic acid in peanut. J. Hered. **80**:252–253.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061–1067.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**:551–556.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33**:290–295.

Pfeifer, M., Kugler, K.G., Sandve, S.R., Zhan, B., Rudi, H., Hvidsten, T.R., International Wheat Genome Sequencing, C., Mayer, K.F., and Olsen, O.A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. Science **345**:1250091.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One **5**:e9490.

Robledo, G., Lavia, G.I., and 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**:1295–1307.

Sahoo, D., Dill, D.L., Tibshirani, R., and Plevritis, S.K. (2007). Extracting binary signals from microarray time-course data. Nucleic Acids Res. **35**:3705–3712.

Sanderson, M.J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**:301–302.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. U S A **108**:4069–4074.

Seijo, G., Lavia, G.I., Fernandez, A., Krapovickas, A., Ducasse, D.A., Bertioli, D.J., and Moscone, E.A. (2007). Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. Am. J. Bot. **94**:1963–1971.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Simpson, C.E., Krapovickas, A., and Valls, J.F.M. (2001). History of *Arachis* included evidence of *Arachis hypogaea* progenitors. Peanut Sci. **28**:78–80.

Singh, A.K., and Simpson, C.E. (1994). Biosystematics and genetic resources. In The Groundnut Crop: A Scientific Basis for Improvement, J. Smartt, ed. (London: Chapman and Hall), pp. 96–137.

Smartt, J., Gregory, W.C., and Gregory, M.P. (1978). The genomes of *Arachis hypogaea*. 1. Cytogenetic studies of putative genome donors. Euphytica **27**:665–675.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J. **3**:739–744.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. **34**:W435–W439.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. **18**:1944–1954.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., Schnable, P.S., Lyons, E., and Lu, J. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. **16**:3.

Varshney, R.K., Chen, W., Li, Y., Bharti, A.K., Saxena, R.K., Schlueter, J.A., Donoghue, M.T., Azam, S., Fan, G., Whaley, A.M., et al. (2011). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat. Biotechnol. **30**:83–89.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49.

Wu, J., Lin, L., Xu, M., Chen, P., Liu, D., Sun, Q., Ran, L., and Wang, Y. (2018). Homoeolog expression bias and expression level dominance in resynthesized allopolyploid Brassica napus. BMC Genomics **19**:586.

Wu, T.D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M.J. (2016). GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol. Biol. **1418**:283–334.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591.

Young, N.D., Debelle, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A., Mayer, K.F., Gouzy, J., Schoof, H., et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature **480**:520–524.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. **22**:2472–2479.

Zhang, L., Yang, X., Tian, L., Chen, L., and Yu, W. (2016). Identification of peanut (*Arachis hypogaea*) chromosomes using a fluorescence in situ hybridization system reveals multiple hybridization events during tetraploid peanut formation. New Phytol. **211**:1424–1439.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C.A., Scheffler, B.E., Stelly, D.M., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. **33**:531–537.

Zhao, G., Zou, C., Li, K., Wang, K., Li, T., Gao, L., Zhang, X., Wang, H., Yang, Z., Liu, X., et al. (2017). The *Aegilops tauschii* genome reveals multiple impacts of transposons. Nat. Plants **3**:946–955.