

The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut

David John Bertioli^{1,2,27}, Steven B Cannon^{3,27}, Lutz Froenicke^{4,5,27}, Guodong Huang^{6,27}, Andrew D Farmer⁷, Ethalinda K S Cannon⁸, Xin Liu⁶, Dongying Gao², Josh Clevenger⁹, Sudhansu Dash⁷, Longhui Ren¹⁰, Márcio C Moretzsohn¹¹, Kenta Shirasawa¹², Wei Huang¹³, Bruna Vidigal^{1,11}, Brian Abernathy², Ye Chu¹⁴, Chad E Niederhuth¹⁵, Pooja Umale⁷, Ana Cláudia G Araújo¹¹, Alexander Kozik⁴, Kyung Do Kim², Mark D Burow^{16,17}, Rajeev K Varshney¹⁸, Xingjun Wang¹⁹, Xinyou Zhang²⁰, Noelle Barkley^{21,22}, Patrícia M Guimarães¹¹, Sachiko Isobe¹², Baozhu Guo²³, Boshou Liao²⁴, H Thomas Stalker²⁵, Robert J Schmitz¹⁵, Brian E Scheffler²⁶, Soraya C M Leal-Bertioli^{2,11}, Xu Xun⁶, Scott A Jackson², Richard Michelmore^{4,5} & Peggy Ozias-Akins^{9,14}

Cultivated peanut (*Arachis hypogaea*) is an allotetraploid with closely related subgenomes of a total size of ~2.7 Gb. This makes the assembly of chromosomal pseudomolecules very challenging. As a foundation to understanding the genome of cultivated peanut, we report the genome sequences of its diploid ancestors (*Arachis duranensis* and *Arachis ipaensis*). We show that these genomes are similar to cultivated peanut's A and B subgenomes and use them to identify candidate disease resistance genes, to guide tetraploid transcript assemblies and to detect genetic exchange between cultivated peanut's subgenomes. On the basis of remarkably high DNA identity of the *A. ipaensis* genome and the B subgenome of cultivated peanut and biogeographic evidence, we conclude that *A. ipaensis* may be a direct descendant of the same population that contributed the B subgenome to cultivated peanut.

Peanut (also called groundnut; *A. hypogaea* L.) is a grain legume and oilseed, which is widely cultivated in tropical and subtropical regions (annual production of ~46 million tons). It has a key role in human nutrition. In Africa and Asia, more peanut is grown than any other grain legume (including soy bean) (FAOSTAT 2015; see URLs). The *Arachis* genus is endemic to South America and is composed mostly of diploid species ($2n = 2x = 20$). *A. hypogaea* is an allotetraploid (AABB-type genome; $2n = 4x = 40$), probably derived from a single recent hybridization event between two diploid species and polyploidization^{1–6}. Chromosomes are of mostly similar size and are metacentric, but strong chromosomal centromeric banding and one pair of small chromosomes distinguish the A from the B subgenome. Cytogenetic, phylogeographic and molecular evidence indicate *A. duranensis* Krapov. & W.C. Greg. and *A. ipaensis* Krapov. & W.C. Greg. as the donors of the A and B subgenomes, respectively^{3,5,7–11}.

The peanut subgenomes are closely related^{5,12}. This, together with a total genome size of ~2.7 Gb and an estimated repetitive content of 64% (ref. 13), makes the assembly of the peanut genome sequence very challenging. However, the A and B subgenomes appear to have undergone relatively few changes since polyploidization: genomic *in situ* hybridization (GISH), using genomic DNA from the diploid species as probes, clearly distinguishes A and B chromosomes and does not show large A-B mosaics^{7,8}. Also, the genome size of *A. hypogaea* is close to the sum of those for *A. duranensis* and *A. ipaensis* (1.25 and 1.56 Gb, respectively¹⁴), indicating that there has been no large change in genome size since polyploidy. Most notably, observations of progeny derived from crosses between cultivated peanut and an artificially induced allotetraploid *A. ipaensis* K30076 × *A. duranensis* V14167 ($2n = 4x = 40$)¹⁵ strongly support the close relationships between the diploid genomes and the corresponding

¹Institute of Biological Sciences, University of Brasília, Brasília, Brazil. ²Center for Applied Genetic Technologies, University of Georgia, Athens, Georgia, USA. ³Corn Insects and Crop Genetics Research Unit, US Department of Agriculture–Agricultural Research Service, Ames, Iowa, USA. ⁴Genome Center, University of California, Davis, Davis, California, USA. ⁵Department of Plant Sciences, University of California, Davis, Davis, California, USA. ⁶BGI-Shenzhen, Shenzhen, China. ⁷National Center for Genome Resources, Santa Fe, New Mexico, USA. ⁸Department of Computer Science, Iowa State University, Ames, Iowa, USA. ⁹Institute of Plant Breeding, Genetics and Genomics, University of Georgia, Tifton, Georgia, USA. ¹⁰Interdepartmental Genetics Graduate Program, Iowa State University, Ames, Iowa, USA. ¹¹Embrapa Genetic Resources and Biotechnology, Brasília, Brazil. ¹²Kazusa DNA Research Institute, Department of Frontier Research, Kisarazu, Japan. ¹³Department of Agronomy, Iowa State University, Ames, Iowa, USA. ¹⁴Department of Horticulture, University of Georgia, Tifton, Georgia, USA. ¹⁵Department of Genetics, University of Georgia, Athens, Georgia, USA. ¹⁶Texas A&M AgriLife Research, Lubbock, Texas, USA. ¹⁷Department of Plant and Soil Science, Texas Tech University, Lubbock, Texas, USA. ¹⁸International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India. ¹⁹Shandong Academy of Agricultural Sciences, Biotechnology Research Center, Jinan, China. ²⁰Henan Academy of Agricultural Sciences, Zhengzhou, China. ²¹Plant Genetic Resources Conservation Unit, US Department of Agriculture–Agricultural Research Service, Griffin, Georgia, USA. ²²International Potato Center, Lima, Peru. ²³Crop Protection and Management Research Unit, US Department of Agriculture–Agricultural Research Service, Tifton, Georgia, USA. ²⁴Chinese Academy of Agricultural Sciences, Oil Crops Research Institute, Wuhan, China. ²⁵Department of Crop Science, North Carolina State University, Raleigh, North Carolina, USA. ²⁶Middle Southern Area Genomics Laboratory, US Department of Agriculture–Agricultural Research Service, Stoneville, Mississippi, USA. ²⁷These authors contributed equally to this work. Correspondence should be addressed to D.J.B. (djbertioli@gmail.com).

Received 30 July 2015; accepted 29 January 2016; published online 22 February 2016; doi:10.1038/ng.3517

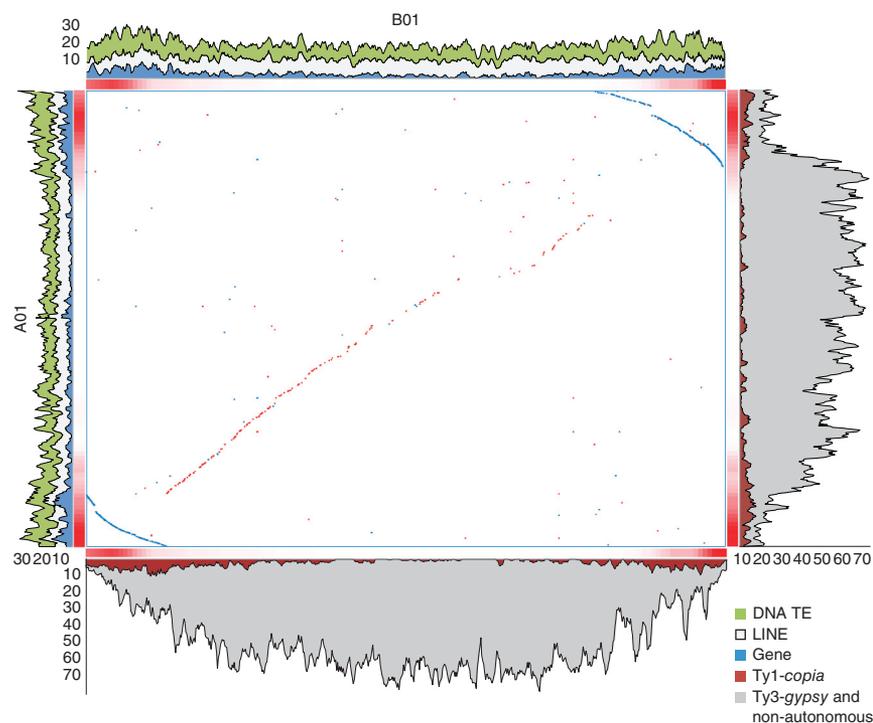
subgenomes of *A. hypogaea*. Progeny are vigorous, phenotypically normal and fertile and showed lower segregation distortion^{16,17} than has been observed for some populations derived from *A. hypogaea* intraspecific crosses^{18–21}. Therefore, as a first step to characterizing the genome of cultivated peanut, we sequenced and analyzed the genomes of the two diploid ancestors of cultivated peanut.

RESULTS

Sequencing and assembly of the diploid A and B genomes

Considering that *A. duranensis* V14167 and *A. ipaensis* K30076 are likely good representatives of the ancestral species of *A. hypogaea*, we sequenced their genomes. After filtering, the data generated from the seven paired-end libraries corresponded to an estimated 154× and 163× base-pair coverage for *A. duranensis* and *A. ipaensis*, respectively (Supplementary Tables 1–6). The total assembly sizes were 1,211 and 1,512 Mb for *A. duranensis* and *A. ipaensis*, respectively, of which 1,081 and 1,371 Mb were represented in scaffolds of 10 kb or greater in size (Supplementary Table 7). Ultradense genetic maps were generated through genotyping by sequencing (GBS) of two diploid recombinant inbred line (RIL) populations (Supplementary Data Set 1). SNPs within scaffolds were used to validate the assemblies and confirmed their high quality; 190 of 1,297 initial scaffolds of *A. duranensis* and 49 of 353 initial scaffolds of *A. ipaensis* were identified as chimeric, on the basis of the presence of diagnostic population-wide switches in genotype calls occurring at the point of misjoin. Chimeric scaffolds were split, and their components were remapped. Thus, approximate chromosomal placements were obtained for 1,692 and 459 genetically verified scaffolds, respectively. Conventional molecular marker maps (Supplementary Data Set 2) and syntenic inferences were then used to refine the ordering of scaffolds within the initial genetic bins. Generally, agreement was good for maps in euchromatic arms and poorer in pericentromeric regions (although one map²² showed large inversions in two linkage groups in comparison to the other maps; Supplementary Data Set 2). Overall, 96.0% and 99.2% of the sequence in contigs $\geq 10,000$ bp in length, represented by 1,692 and 459 scaffolds, could be ordered into 10 chromosomal pseudomolecules per genome of 1,025 and 1,338 Mb for *A. duranensis* and *A. ipaensis*, respectively (Aradu.A01–Aradu.A10 and Araip.B01–Araip.B10; GenBank, assembly accessions GCA_000817695.1 and GCA_000816755.1; Supplementary Table 8). The pseudomolecules mostly showed one-to-one equivalence between the A and B genomes (Figs. 1 and 2, and Supplementary Figs. 1–12)

Figure 1 Structural overview and comparison of chromosomal pseudomolecules A01 and B01. The distributions of genes and mobile elements are represented as stacked areas. High frequency of genetic recombination (represented by red on a white-red heat scale) is confined to distal regions. In the dot-plot comparison, note how inverted chromosome regions form arcs, indicating that, over the evolutionary time since the divergence of the two species, accumulation of DNA has been greater in more central regions of the chromosomes and elimination of DNA has been more frequent in distal regions. Genes, DNA transposable elements (TEs) and Ty1-*copia* elements are more frequent in more distal regions. Ty3-*gypsy* elements are more frequent in pericentromeric regions.



and were numbered according to previously published linkage maps^{17,19,23,24}. They represent 82% and 86% of the genomes, respectively, when considering genome size estimates based on flow cytometry^{14,25}, or 95% and 98% of the genomes when using estimates derived from *k*-mer frequencies with *k* = 17 (Supplementary Figs. 13 and 14). Comparisons of the chromosomal pseudomolecules with 14 BAC sequences from *A. duranensis* and 6 BAC sequences from *A. ipaensis* showed collinearity of contigs and high sequence identity ($\geq 99\%$) (Supplementary Fig. 15a–l and Supplementary Table 9).

Characterization of transposons

We identified transposable elements that contributed 61.7% and 68.5% of the *A. duranensis* and *A. ipaensis* genomes, respectively (Supplementary Tables 10 and 11; PeanutBase). These values are compatible with the 64% repetitive content estimated for cultivated peanut using renaturation kinetics¹³. Most transposon families were shared by the two genomes, and, for abundant families, macroscale positioning in the two genomes seemed similar. However, because of transposon activity since the divergence of the two genomes, microscale positioning and relative abundance differed (data not shown). A few Ty3-*gypsy* and non-autonomous retrotransposon families were very abundant, forming dense accumulations in pericentromeres (Fig. 1 and Supplementary Figs. 16 and 17). These included the previously described autonomous/non-autonomous pairs FIDEL/Feral and Pipoka/Pipa, the non-autonomous Gordo^{26,27}, and the newly observed Apolo and Polo. Overall, long terminal repeat (LTR) retrotransposons comprised more than half of each genome. In contrast, DNA transposons constituted about 10%. Notably, 7.8% and 11.7% of the genomes could be attributed to long interspersed nuclear elements (LINEs) for *A. duranensis* and *A. ipaensis*, respectively. These are the highest coverages for LINEs thus far reported for plant genomes.

Gene annotation and analysis of gene duplications

Transcript assemblies were constructed using sequences expressed in diverse tissues of *A. duranensis* V14167, *A. ipaensis* K30076 and

Figure 2 Circos diagram depicting the relationships of the chromosomal pseudomolecules of *A. duranensis* and *A. ipaensis*. Blue color represents the density of genes, and brown color represents the density of Ty3-gypsy elements and non-autonomous LTR retrotransposons. The scale for the gray bars is in megabases.

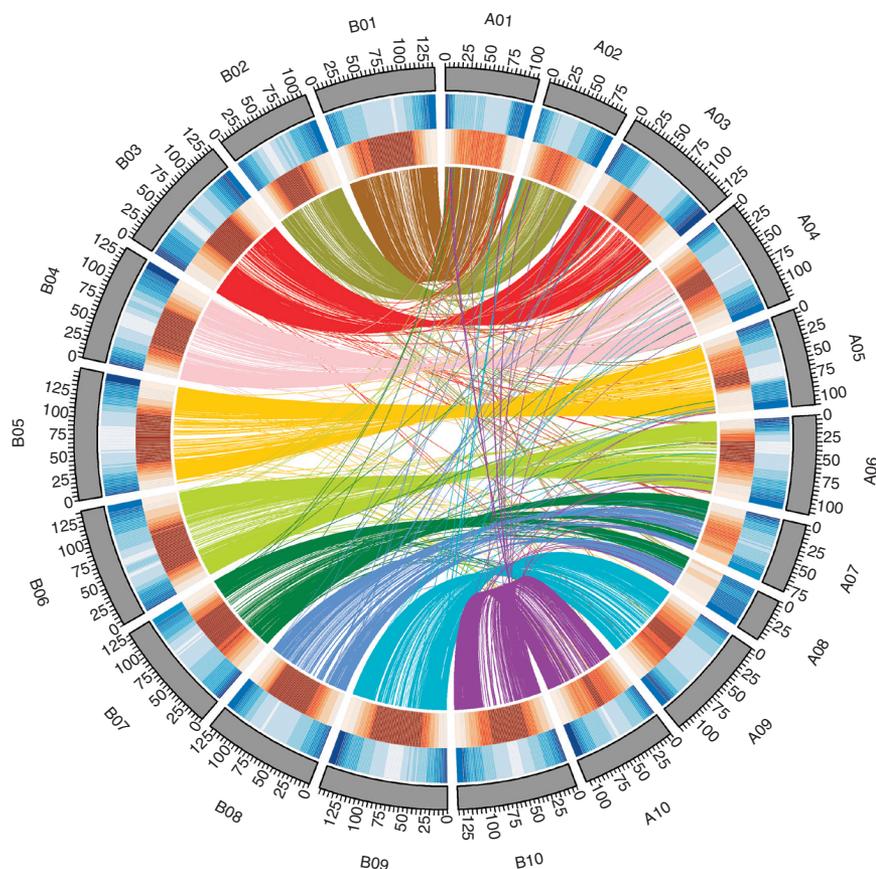
A. hypogaea cv. Tifrunner²⁸ (16,439,433, 21,406,315 and 2,064,268,316 paired-end reads for each species, respectively; details below and in **Supplementary Tables 12** and **13**). Using these assemblies and representative characterized transposon sequences, MAKER2 (ref. 29) delineated 36,734 and 41,840 high-quality non-transposable element genes for *A. duranensis* and *A. ipaensis*, respectively (PeanutBase). The elevated gene numbers in *A. ipaensis* appear to originate from more local duplications, which can be seen in counts of genomically 'close' paralogous genes. Considering similar genes within a ten-gene window, there were 25% more in *A. ipaensis* than in *A. duranensis* (7,825 versus 6,241). Gene families known to occur in clusters such as those encoding NB-ARC, leucine-rich repeat (LRR), pentatricopeptide-repeat, kinase, WD40-repeat and kinesin proteins had large differential counts between the two genomes. These differences were also apparent with wider inspection. In a set of 9,236 gene families with members in *A. ipaensis* or *A. duranensis*, or both, 2,879 families had more members in *A. ipaensis*, 1,983 had more members in *A. duranensis* and 4,374 had the same number of members in both species (**Supplementary Data Sets 3–5**).

DNA methylation

Analysis of DNA methylation by whole-genome bisulfite sequencing using MethylC-seq³⁰ generated 189,653,337 and 277,101,705 uniquely aligned reads, giving ~8.6× and 10.0× coverage per strand for *A. duranensis* and *A. ipaensis*, respectively. Genome-wide methylation per cytosine content³¹ was similar for *A. duranensis* and *A. ipaensis*, with 73% and 75% methylation at CG sites, 57% and 60% methylation at CHG sites (where H is an A, T or C), and 8% and 6% methylation at CHH sites, respectively. The genic methylation patterns were typical for plants and provide independent verification of gene annotation^{32,33} (**Supplementary Figs. 18** and **19**; Gene Expression Omnibus (GEO), [GSE71357](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71357)).

Disease resistances and NB-LRR-encoding genes

Nucleotide-binding-leucine-rich repeat (NB-LRR)-encoding genes are of particular interest because they confer resistance against pests and diseases. We identified 345 and 397 of these genes in the *A. duranensis* and *A. ipaensis* genotypes, respectively (**Supplementary Data Set 6**). The largest clusters were on distal regions of chromosomal pseudomolecule 02, the lower arms of chromosomal pseudomolecule 04 and the upper arms of chromosomal pseudomolecule 09 (**Supplementary Fig. 20**). The genome assemblies allowed us to associate quantitative trait loci (QTLs) with candidate genes (**Supplementary Data Set 7**). A strong, consistent QTL for resistance to root-knot nematode (*Meloidogyne arenaria* (Neal.) Chitwood) identified on A02



of *Arachis stenosperma* V10309 Krapov. & W.C. Greg. (ref. 34) resides in a cluster of 38 NB-LRR-encoding genes covering 6.1 Mb. Another source of nematode resistance already widely used in the United States originates from an introgression of the A-genome species *Arachis cardenasii* Krapov. & W.C. Greg. (ref. 35). This segment resides in the upper distal 7.6 Mb of chromosome A09 (ref. 36) and contains many NB-LRR-encoding genes. A major QTL conferring reduction in lesion number, size and sporulation of rust was identified in *Arachis magna* K30097 Krapov., W.C. Greg. & C.E. Simpson³⁷. The closest linked marker (Ah280; Araip.B08, 126,645,511) maps close to an NB-LRR-encoding gene (Araip.RV63R). Another QTL for rust resistance has previously been identified in peanut varieties that have the wild A-genome species *A. cardenasii* in their pedigree³⁸. Markers mapped this QTL to an introgressed chromosome segment at the lower end of A03 (Aradu.A03, 131,305,113–133,690,542) where an NB-LRR-encoding gene resides in *A. duranensis* (Aradu.Z87JB). The genes harbored on these genome segments from *A. stenosperma*, *A. magna* and *A. cardenasii* provide good pest and disease resistance and warrant further investigation.

Gene evolution in *A. ipaensis* and *A. duranensis* and species divergence

Analyses suggest that the *Arachis* lineages have been accumulating mutations relatively quickly since the divergence of the Dalbergioid clade ~58 million years ago. Modal K_S paralog values (synonymous substitutions per synonymous site) are approximately 0.95 for *A. ipaensis* and 0.90 for *A. duranensis*, more similar to that for *Medicago* (paralogous K_S value of ~0.95) than to those of *Lotus* (~0.65), *Glycine* (~0.65) or *Phaseolus* (~0.80). Average rates of change for *Arachis* genes were estimated at 8.12×10^{-9} K_S /year. On the basis of this and the peak in the frequency of K_S values between *A. duranensis* and *A. ipaensis* being 0.035,

Figure 3 Mutations and genome duplications. Frequency distributions are shown of values of synonymous substitutions (K_S) for paralogous and orthologous genes in comparisons of *A. duranensis* (Ad), *A. ipaensis* (Ai) and *Glycine max* (Gm). Peaks in the *G. max*–*G. max* comparison represent the *Glycine* whole-genome duplication (WGD) at $K_S = 0.10$ (~10 million years ago) and the early papilionoid WGD at $K_S = 0.65$ (58 million years ago). The same early papilionoid WGD also affected the *Arachis* lineage, so the shift in the *A. duranensis*–*A. duranensis* and *A. ipaensis*–*A. ipaensis* peaks (at $K_S = 0.90$ and 0.95 , respectively) indicates that *Arachis* has accumulated silent changes at a rate ~1.4 times faster than that in *G. max*. On the basis of average rates of change for *Arachis* of 8.12×10^{-9} K_S /year, we estimate that *A. duranensis* and *A. ipaensis* diverged ~2.16 million years ago.

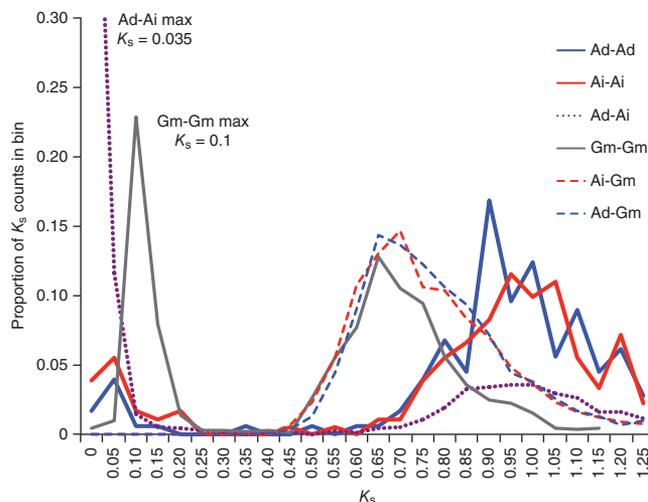
the divergence of the two species was estimated as occurring ~2.16 million years ago (Fig. 3 and Supplementary Figs. 21 and 22).

Analysis of chromosomal structure and synteny

In accordance with cytogenetic observations^{9,10}, most pseudomolecules had symmetrically positioned pericentromeres. Most pseudomolecules showed a one-to-one correspondence between the two species: pairs 02, 03, 04 and 10 were collinear; pairs 05, 06 and 09 were each differentiated by a large inversion in one arm of one of the pseudomolecules; and the pseudomolecules in pair 01 were differentiated by large inversions of both arms (Figs. 1 and 2, and Supplementary Figs. 1–12). In contrast, chromosomes 07 and 08 have undergone complex rearrangements that transported repeat-rich DNA to A07 and gene-rich DNA to A08. As a result, A07 has only one normal (upper) euchromatic arm and A08 is abnormally small, with low repetitive content (Fig. 4 and Supplementary Table 11). In accordance with cytogenetic observations^{8,26}, A08 could be assigned as the characteristic small ‘A chromosome’ (cytogenetic chromosome A09; Supplementary Fig. 23).

All *A. ipaensis* pseudomolecules were larger than their *A. duranensis* counterparts (Supplementary Table 8). This is partly because of a greater frequency of local duplications and higher transposon content in *A. ipaensis*. In dot plots of collinear chromosomes, slopes formed by orthologous genes were similar in both euchromatic and pericentromeric regions, with *A. duranensis* regions being ~80–90% the length of the corresponding regions in *A. ipaensis* (Supplementary Figs. 2–4 and 12). In contrast, in the dot plots, chromosomal regions differentiated by inversions showed distinct arcs (Fig. 1 and Supplementary Figs. 1, 5, 6 and 11). These arcs are due to changes in rates of DNA loss and gain^{39,40} in regions that switch from distal to pericentromeric contexts, or vice versa, when inverted (Fig. 5). In chromosomes without inversions, there were characteristic density gradients for genes, repetitive DNA and methylation (with gene densities increasing and densities of repetitive DNA and methylation decreasing toward chromosome ends). However, in regions that had undergone large rearrangements, in *A. duranensis*, these gradients were disrupted (Supplementary Figs. 16, 17 and 24–27).

Figure 4 Schematic showing the rearrangements between chromosomes 7 and 8. These rearrangements gave rise to the small, repeat-poor chromosome, represented by pseudomolecule Aradu.A08 (equivalent to cytogenetic A09), which is characteristic of A genomes, and Aradu.A07, which has only one normal euchromatic arm (the upper one). Syntenic chromosomal segments are represented by blocks of the same color. The Ty3-gypsy and non-autonomous retroelement distributions are represented in gray. Note the low repetitive content of Aradu.A08 and the ‘knob’ of repeat-rich DNA in the upper distal region. This unusual composition seems likely to account for the distinct chromatin condensation of this chromosome pair (Supplementary Fig. 23).



From these observations, we concluded that the major rearrangements have all occurred in the A-genome lineage. Size differences between homeologous chromosomes that were differentiated by large rearrangements tended to be greater than those between collinear ones ($r(6) = 0.65$, $P < 0.05$; Supplementary Table 14). Because the *A. duranensis* chromosomes that have undergone inversions are smaller than expected, it is evident that, in this dynamic, on balance, the elimination of DNA has predominated over its accumulation.

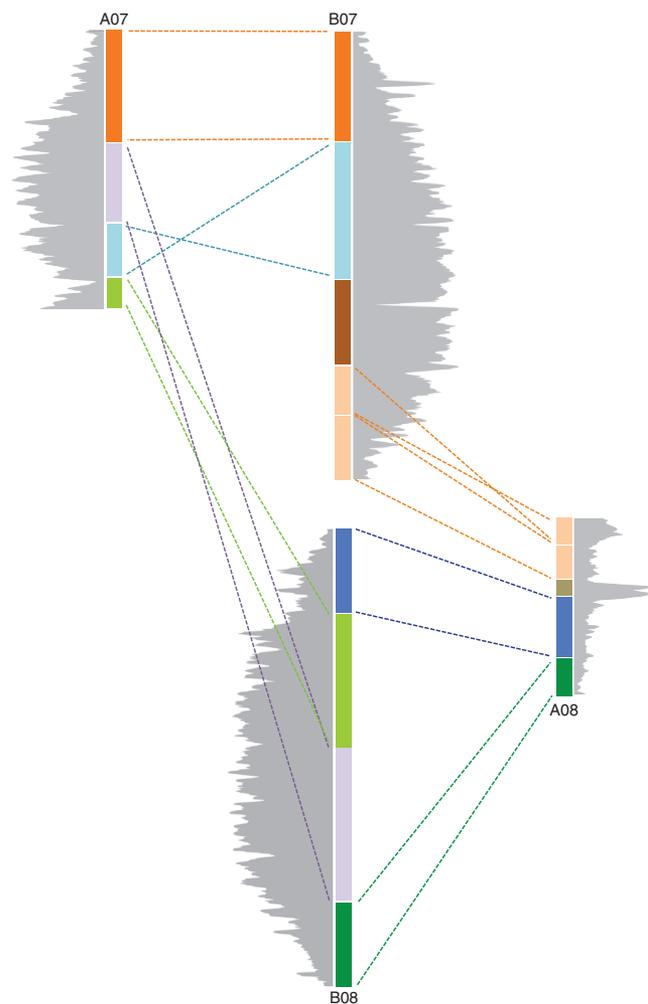
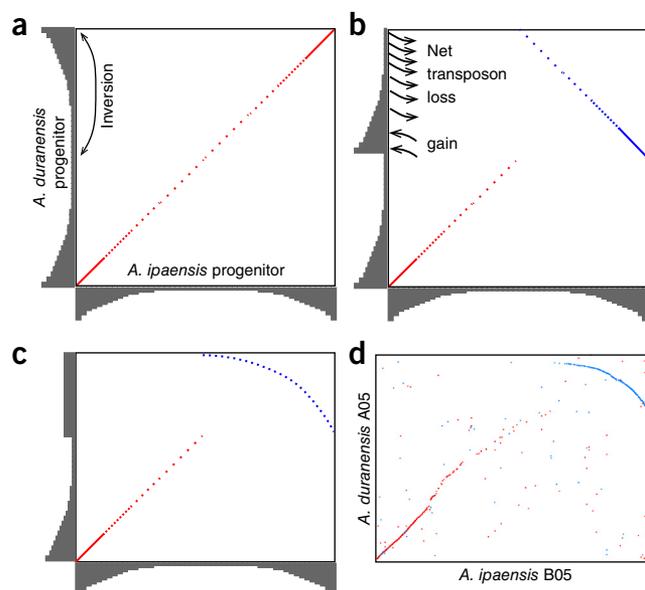


Figure 5 Model for the formation of the arcs in dot plots of genome regions that have been inverted since the divergence of the A and B genomes. Gene densities are shown in gray. (a) The inversion transports repeat-rich, gene-poor DNA to the distal chromosomal region and repeat-poor, gene-rich DNA to the more central region. (b) In the distal region, the inverted segment then loses DNA by recombination-driven deletion, and the more central region gains DNA. (c) Thus, the characteristic arc and atypical gene, repetitive DNA and methylation density patterns are formed. The presence of these atypical patterns indicates that all major genome rearrangements occurred in the A-genome lineage (Supplementary Figs. 16, 24 and 26). (d) An example dot plot comparing A05 and B05 that shows the characteristic arc.

Comparisons with *Phaseolus vulgaris* L., which shared a common ancestor with *Arachis* about 58 million years ago, showed syntenous chromosomal segments. In some cases, although the dot plots were highly distorted, there was almost a one-to-one correspondence between chromosomes (for example, B01 and Pv03, B05 and Pv02, B06 and Pv01, and B08 and Pv05; Supplementary Figs. 28–31).

Sequence comparisons to tetraploid cultivated peanut

Comparisons showed fundamentally one-to-one correspondences between the diploid chromosomal pseudomolecules and cultivated peanut linkage groups. Of the marker sequences from three maps^{21,41}, 83%, 83% and 94% were assigned by sequence similarity searches to the expected diploid chromosomal pseudomolecules (Supplementary Table 15a–c and Supplementary Data Set 8). For more detailed genome-wide comparisons, we produced 5.74 Gb (2× coverage) of long-sequence Moleculo reads from *A. hypogaea* cv. Tifrunner and mapped the reads to the combined diploid pseudomolecules. The corrected median identities between the *A. hypogaea* Moleculo reads and the pseudomolecules of *A. duranensis* and *A. ipaensis* were 98.36% and 99.96%, respectively (Supplementary Data Set 6). When visualized as plots along the chromosomal pseudomolecules, the diploid A-genome chromosomes were distinctly less similar to *A. hypogaea*



sequences than the B-genome chromosomes (Fig. 6, Supplementary Fig. 32a–t and Supplementary Data Set 9).

We found distinct signals of genetic recombination between the A and B subgenomes of *A. hypogaea*, and, as expected, these signals were more frequent in regions of the homeologous chromosome pairs that were collinear. This recombination erodes the similarities between the tetraploid subgenomes and their corresponding diploid genomes. We observed a significant tendency for *A. hypogaea* Moleculo reads that mapped to collinear A-genome pseudomolecules to have, on average, lower sequence identity than reads that mapped to pseudomolecules with inversions (Kruskal-Wallis test, $P < 0.0001$; Supplementary Tables 16 and 17, and Supplementary Data Set 9). This trend was much weaker for the B subgenome on a

Figure 6 Example graphs comparing DNA sequences from cultivated peanut with chromosomal pseudomolecules of *A. duranensis* and *A. ipaensis*. (a,b) Graphs show mapping of Moleculo DNA sequence reads from the tetraploid *A. hypogaea* cv. Tifrunner along the diploid chromosomal pseudomolecules Aradu.A05 (a) and Araip.B05 (b). Dark blue dots represent percentage identity of reads in tiling paths, and red dots represent density of Moleculo bases mapping in windows of 0.5 Mb (normalized to a value of 1 for the expected number). Note how the percentage identities of mapped reads for Aradu.A05 are, contrary to expectation, more consistent in the pericentromeric regions than in distal ones. This may reflect that sequence similarity between *A. duranensis* and the A subgenome of *A. hypogaea* has been eroded by recombination between the A and B subgenomes in the tetraploid by, for example, gene conversion. In contrast, mapping on Araip.B05 is much more consistent and generally very high identity, except for the upper distal 6.1 Mb, where identities fall dramatically (blue arrow). Also note how deviations in expected mapping density (indicated by red arrows) show that this region, in the tetraploid genome of *A. hypogaea* cv. Tifrunner, has undergone tetrasomic recombination and has changed from the expected genome formula of AABB to AAAA.

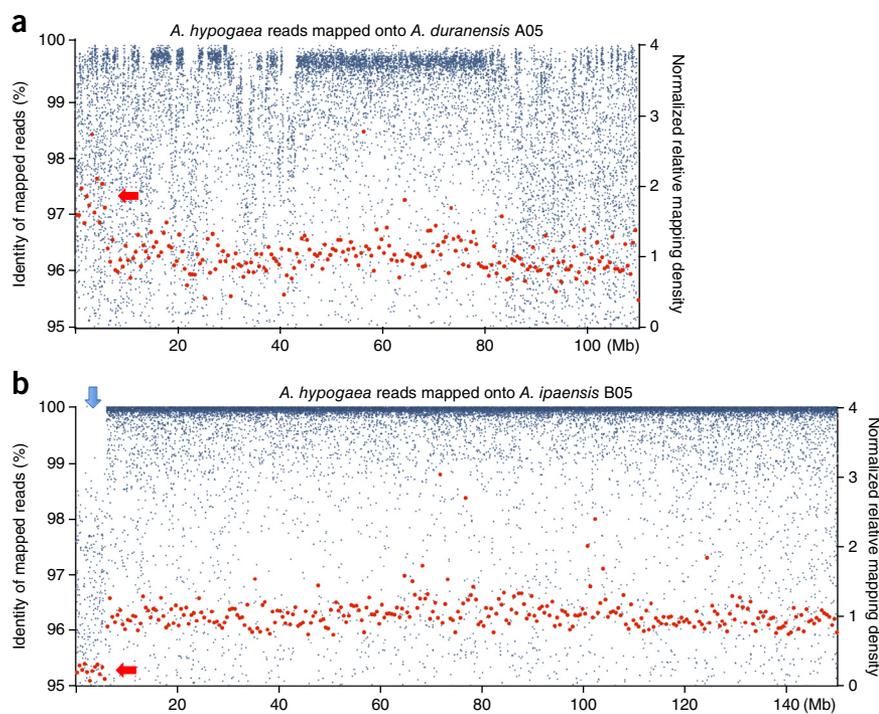
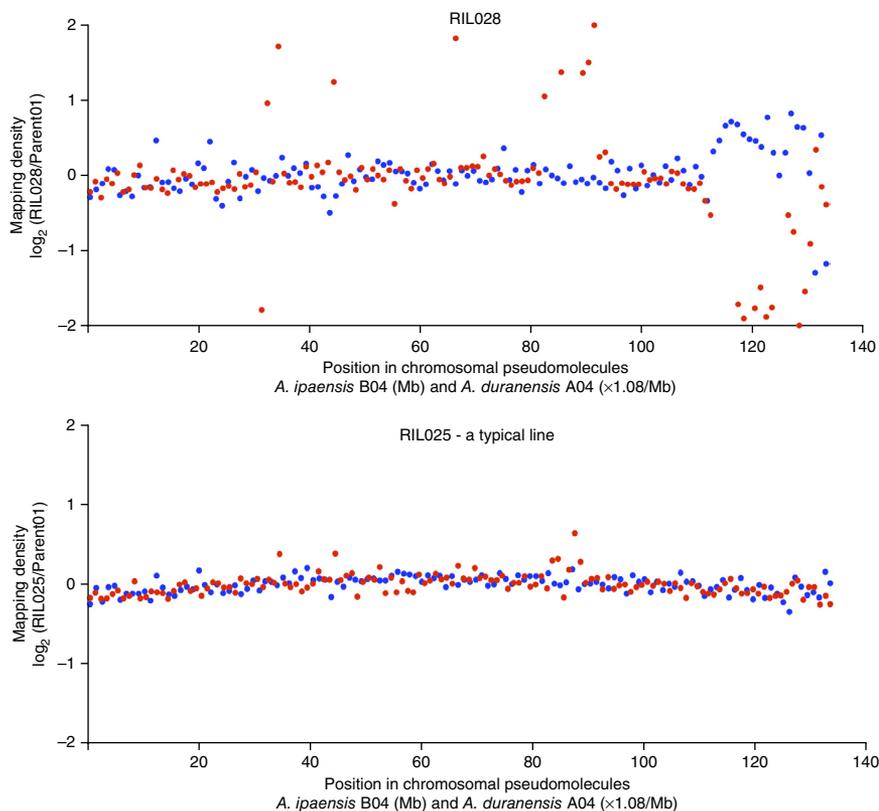


Figure 7 Identification of genetic exchange between subgenomes in cultivated peanut. The top graph depicts the result of recombination between A04 and B04 in RIL028, and, for comparison, the bottom graph shows RIL025, a typical line where this type of recombination has not occurred (lines described in Zhou *et al.*⁴¹). The y axis shows \log_2 -transformed ratios of densities of mapping for restriction site-associated sequence reads along the diploid chromosomal pseudomolecules divided by the mapping densities of a parental line. The x axis shows the positions of mapping, in 1-Mb windows, along Araip.B04 (red dots) and Aradu.A04 (blue dots), the latter with distances scaled so the homeologous chromosomal pseudomolecules are directly comparable. In RIL028, the relative dosage of the subgenomes has changed greatly in the lower chromosome arms. This indicates that a new event of genetic exchange between the A and B subgenomes occurred.



whole-chromosome scale but was clearly visible at the ends of some of the collinear B-subgenome chromosome arms, where percentage identities dropped dramatically. An example of this is indicated by the blue arrow in **Figure 6b**. In this case, the *A. hypogaea* B subgenome has become nullisomic and the A subgenome has become tetrasomic (producing a genome composition of AAAA instead of the expected AABB). This was confirmed by an inverse symmetry in the total number of bases mapping to the A and B genomes (**Fig. 6a,b**, red arrows). This phenomenon, a degree of tetrasomic genetic behavior, has been observed in the progeny of interploidy crosses involving wild species⁴² and recently in the cultivated \times induced allotetraploid RILs used in this study⁴³, but this is the first time, to our knowledge, that it has been observed in pure cultivated peanut. The event depicted in **Figure 6** affects approximately the top 6 Mb, about half the euchromatic arms of *A. hypogaea* cv. Tifrunner chromosomes 05. Smaller similar events covered the bottom \sim 1 Mb of A02 and B02, the bottom \sim 0.4 Mb of A03 and B03, the bottom \sim 2 Mb of A06 and B06, and the top \sim 0.5 Mb of A09 and B09. Because of lower sequence identities, A-subgenome nullisomes were more difficult to detect; nevertheless, the bottom \sim 3 Mb of chromosomes 04 appeared to be nullisomic for the A subgenome and tetrasomic for the B subgenome (**Supplementary Fig. 32** and **Supplementary Data Set 9**).

Although we recognize that genetic exchange between the A and B subgenomes will tend to inflate the values calculated (especially for the A genomes), we estimated the dates of evolutionary divergence of the sequenced diploid genomes and the corresponding subgenomes of *A. hypogaea*. To estimate the genome-wide *Arachis* mutation rate, we mapped *A. ipaensis* Moleculo reads against the *A. duranensis* pseudomolecules. This gave a corrected median DNA identity of 93.11% (a value compatible with previous comparisons using BAC sequences²⁷). Considering the date of divergence of the A and B genomes as 2.16 million years ago (from K_S values), this gives an *Arachis* genome-wide mutation rate of 1.6×10^{-8} mutations per base per year (within the range of $1-2 \times 10^{-8}$ calculated for other plants⁴⁴). This mutation rate and the divergence of the most conserved chromosomes (presumably the ones that have undergone the least recombination between subgenomes; A01 and B07) gives an estimate of the divergence time of *A. duranensis* V14167 from the

A subgenome of *A. hypogaea* as \sim 247,000 years and for the divergence time of *A. ipaensis* from the B subgenome of *A. hypogaea* as a remarkably recent \sim 9,400 years.

We used the chromosomal pseudomolecules to investigate the frequency of recombination between A and B subgenomes in 166 cultivated peanut RILs described in a previous study⁴¹. To do this, we calculated the mapping densities of restriction site-associated sequence reads from these RILs and their parental lines along the chromosomal pseudomolecules. Mostly, the relative dosage of mapping on the A and B genomes was equal and the same as in the parents, but for one line (RIL028) the relative dosage was dramatically altered for two homeologous chromosomal regions (**Fig. 7**): 104–112 Mb on Aradu.A04 and 112–126 Mb on Araip.B04. In these regions, mapping to Araip.B04 almost disappeared and mapping to Aradu.A04 dramatically increased in density. This clear signal indicates that genetic exchange occurred between the A and B subgenomes in regions of the cultivated peanut genome that had balanced dosage in the parental lines. This seems most likely to have occurred by tetrasomic recombination, but gene conversion after the formation of an unresolved Holiday junction is also possible.

Diploid genome-guided tetraploid transcriptome assembly

Assemblies of transcribed sequences from tetraploid cultivated peanut are challenging because reads from genes on the A and B subgenomes are erroneously assembled together, resulting in chimeric sequences. We used the diploid genomes to minimize this collapse and produced tetraploid transcript assemblies. We assessed four assembly software approaches in three different ways: *de novo* assembly; parsing into A- and B-genome sets followed by separate assembly; and parsing followed by genome-guided assembly using the combined pseudomolecules. Results were compared by measuring the percentage of assembled transcripts that mapped back to the pseudomolecules without mismatches. Higher percentages indicate less

Figure 8 The approximate known distributions of *A. duranensis* and *A. magna*, the location of the single known occurrence of *A. ipaensis* and the center of diversity for the most primitive type of cultivated peanut, *A. hypogaea* subsp. *hypogaea* var. *hypogaea*. *A. ipaensis* is only known to be from a single location and is biologically conspecific with *A. magna*, which occurs far to the north and at lower altitude. The isolated occurrence and estimated divergence of the *A. ipaensis* genome from the B genome of *A. hypogaea*, only ~9,400 years ago, indicate that *A. ipaensis* was probably taken to its present location from the north by prehistoric inhabitants of the region. *A. hypogaea* was formed by hybridization of *A. ipaensis* with *A. duranensis* and polyploidization. The figure was generated using Natural Earth.

collapse, as collapsed transcripts can only map with mismatches. This analysis showed that the *de novo* assemblies were the least accurate, with the percentage of mismatch-free mapping ranging from 32.17 to 39.82%, followed by the parsed set assemblies (40.07 to 55.8%). Finally, the genome-guided assembly was the most accurate with 65.87% mismatch-free mapping (Supplementary Fig. 33). Using this workflow, together with filtering for transposable elements, low-expression transcripts and redundancy, we obtained 183,062 assembled *A. hypogaea* transcripts, of which we could tentatively assign 88,643 (48.42%) to the A subgenome and 94,419 (51.58%) to the B subgenome.

DISCUSSION

The peoples of South America have cultivated *Arachis* species since prehistoric times, but only the allotetraploid *A. hypogaea* was completely transformed by domestication to become a crop of global importance⁴⁵. As a foundation to investigate peanut's genome, we sequenced its diploid progenitors *A. duranensis* and *A. ipaensis*. Comparisons of the chromosomal pseudomolecules of the diploid species with *A. hypogaea* show high levels of similarity, but, most notably, the cultivated peanut B subgenome is nearly identical to the genome of *A. ipaensis*. This similarity suggests a remarkable story for this particular *Arachis* population dating back to the time of the earliest human occupation of South America.

Arachis species have an unusual reproductive biology; although the flowers develop above ground, a special 'peg' structure (gynophore) pushes the young pod underground, where development is completed⁴⁶. The seeds are protected and have privileged access to water at the beginning of the rainy season. However, they are not usually dispersed and germinate within an area of roughly 1 m in diameter covered by the mother plant. Therefore, populations are quite static over long periods of time: over a thousand years, they can usually move only about 1 km. Rarely, water-driven soil erosion will disperse seeds downhill. This pattern of dispersal, coupled with a high rate of self-pollination, has led to species distributions that consist of patchy, often highly homozygous populations distributed over areas defined by major river systems⁴⁷.

Both sequenced accessions were collected in the most likely geographic region for the origin of cultivated peanut (Fig. 8). Whereas *A. duranensis* is represented by numerous, genetically diverse populations in the region, *A. ipaensis* is only known to be from a single location (from where it may now have disappeared; G. Seijo, personal communication). This site of occurrence of *A. ipaensis* is incompatible with natural seed dispersion: the closest relative of *A. ipaensis*, the biologically conspecific *A. magna*, occurs in numerous natural populations more than 500 km to the north and more than 200 m lower in altitude⁴⁵. Therefore, it seems most likely that humans transported the seed that founded this population, and several lines of evidence indicate that this same population was involved in the formation of *A. hypogaea*. *A. ipaensis* is the only B-genome *Arachis* species ever found within the range of *A. duranensis*. Its site of occurrence is



immediately adjacent to the center of diversity for the most primitive of the botanical varieties of cultivated peanut (*A. hypogaea* subsp. *hypogaea* var. *hypogaea*), and, most notably, it has extremely high DNA similarity with the B subgenome of *A. hypogaea*. We estimate that they diverged ~9,400 years ago, at a time when the region was becoming populated by early inhabitants⁴⁸. The earliest archeological records of *Arachis* cultivation are about 7,800 years old from Peru, far from the most likely region of origin for *A. hypogaea* or any natural *Arachis* species distribution⁴⁹. It seems likely that *Arachis* cultivation would have started within the native range well before then^{11,50}. The date of polyploidization is uncertain, but the earliest identifiable remains of *A. hypogaea* date from ~3,500–4,500 years ago⁵¹. For most plants, following polyploidization, sequence identity between the diploid progenitors and the polyploid subgenomes would have been dispersed by genetic recombination in subsequent generations. However, in the case of the B genomes, it persisted, perhaps owing to extreme genetic bottlenecks and reproductive isolation in both species (*A. ipaensis* and *A. hypogaea*). We think that this is a unique find among crop plants, which was possible because of the peculiar biology of the genus and the remarkable work of botanical collectors.

Building on their tractability as genetic systems, we sequenced peanut's diploid ancestors. We used them to identify candidate pest and disease resistance genes, to reduce collapse in tetraploid transcriptome assemblies and to show the impact of recombination between subgenomes in cultivated peanut. The availability of these genomes will lead to further advances in knowledge of genetic changes since the very recent polyploidization event that gave rise to cultivated peanut and to the production of better tools for molecular breeding and crop improvement.

URLs. Food and Agriculture Organization Corporate Statistical Database (FAOSTAT), <http://faostat3.fao.org/home/>; MadMapper, <http://www.atgc.org/XLinkage/MadMapper/>; RepeatMasker, <http://www.repeatmasker.org/>; FastQC, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>; Trim Galore! v0.3.5, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; FASTX-Toolkit, http://hannonlab.cshl.edu/fastx_toolkit/; PeanutBase, *Arachis* genome sequences and repeat libraries, <http://peanutbase.org/download>; Natural Earth maps, <http://www.naturalearthdata.com/>.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Accession codes. Genome assemblies and annotations, identified transposable elements, transcript assemblies and map data are available at <http://www.peanutbase.org/download>. Genome assemblies have also been deposited in GenBank under assembly accessions [GCA_000817695.1](#) and [GCA_000816755.1](#). MethylC-seq data are available under Gene Expression Omnibus (GEO) accession [GSE71357](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank G. Birdsong, V. Nwosu, J. Elder, D. Smyth, F. Luo, D. Hoisington, H. Shapiro, H. Valentine, R. Wilson and D. Ward for their support of and work for the Peanut Genome Initiative. We thank J.F.M. Valls, C. Simpson and G. Seijo for valuable discussions. Major financial contributors for this work were from Mars, US peanut sheller associations, the National Peanut Board and other industry groups. A full list can be downloaded at <http://peanutbioscience.com/peanutgenomeproject.html>. We also thank the generous support of the Agriculture and Food Research Initiative for grant 2012-85117-19435 from the US Department of Agriculture, the US National Science Foundation for grant NSF-MCB-1339194 to R.J.S., EMBRAPA (Brazil) for project 02.11.08.006.00 and the National High-Technology Research and Development Program (China) for grant 2012AA02A701.

AUTHOR CONTRIBUTIONS

Project planning: D.J.B., S.B.C., L.F., X.L., S.I., B.G., B.L., X.Z., M.D.B., R.K.V., X.W., N.B., H.T.S., B.E.S., S.C.M.L.-B., X.X., S.A.J., R.M. and P.O.-A. Cultivation of accessions and lines, preparation of DNA: M.C.M., S.C.M.L.-B., L.F. and D.J.B. Production of Illumina libraries, sequencing and genome assembly: G.H., X.L. and X.X. RIL sequencing, production of GBS maps and genetic validation of scaffolds, Moleculo sequencing: L.F. and R.M. Construction and curation of conventional genetic maps: M.C.M., S.C.M.L.-B., P.O.-A. and D.J.B. Ordering of scaffolds into chromosomal pseudomolecules: S.B.C., E.K.S.C. and S.D. BAC sequencing and comparison with pseudomolecules: P.M.G. and D.J.B. Analysis of mobile elements: D.G., B.V., A.C.G.A., B.A., D.J.B. and S.A.J. Tissue sampling, transcriptome sequencing and assembly: J.C., B.E.S., Y.C., S.C.M.L.-B., L.F., S.A.J. and P.O.-A. Gene annotation: A.D.F., S.D., P.U. and L.R. Methylation analysis: C.E.N., K.D.K., S.A.J. and R.J.S. Analysis of disease resistance and candidate genes: S.C.M.L.-B., A.K., M.C.M., D.J.B., S.B.C., E.K.S.C. and R.M. Analysis of synteny and gene and genome evolution: E.K.S.C., L.R., W.H., D.J.B. and S.B.C. Comparisons of diploid and tetraploid genomes and biogeography: B.A., K.S., M.C.M. and D.J.B. Analysis of genetic recombination in tetraploid RILs: K.S. and D.J.B. Writing of the manuscript: D.J.B., S.B.C., S.C.M.L.-B., P.O.-A., B.E.S., C.E.N., J.C., D.G., K.S., X.L., G.H., L.F., R.M., R.J.S. and S.A.J. All authors approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material.

To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>.

- Husted, L. Cytological studies of the peanut *Arachis*. 2. Chromosome number, morphology and behavior, and their application to the problem of the origin of the cultivated forms. *Cytologia (Tokyo)* **7**, 396–423 (1936).
- Halward, T., Stalker, T., LaRue, E. & Kochert, G. Use of single-primer DNA amplifications in genetic studies of peanut (*Arachis hypogaea* L.). *Plant Mol. Biol.* **18**, 315–325 (1992).
- Kochert, G. *et al.* RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). *Am. J. Bot.* **83**, 1282–1291 (1996).
- Cuc, L.M. *et al.* Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (*Arachis hypogaea*). *BMC Plant Biol.* **8**, 55 (2008).

- Moretzsohn, M.C. *et al.* A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Ann. Bot.* **111**, 113–126 (2013).
- Kochert, G., Halward, T., Branch, W.D. & Simpson, C.E. RFLP variability in peanut (*Arachis hypogaea* L.) cultivars and wild species. *Theor. Appl. Genet.* **81**, 565–570 (1991).
- Ramos, M.L. *et al.* Chromosomal and phylogenetic context for conglutin genes in *Arachis* based on genomic sequence. *Mol. Genet. Genomics* **275**, 578–592 (2006).
- Seijo, G. *et al.* Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. *Am. J. Bot.* **94**, 1963–1971 (2007).
- Robledo, G. & Seijo, G. Species relationships among the wild B genome of *Arachis* species (section *Arachis*) based on FISH mapping of rDNA loci and heterochromatin detection: a new proposal for genome arrangement. *Theor. Appl. Genet.* **121**, 1033–1046 (2010).
- Robledo, G., Lavia, G.I. & Seijo, G. 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 (2009).
- Grabiele, M., Chalup, L., Robledo, G. & Seijo, G. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Syst. Evol.* **298**, 1151–1165 (2012).
- Nielen, S. *et al.* Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the *Arachis* A-B genome divergence. *Mol. Genet. Genomics* **287**, 21–38 (2012).
- Dhillon, S.S., Rake, A.V. & Miksche, J.P. Reassociation kinetics and cytophotometric characterization of peanut (*Arachis hypogaea* L.) DNA. *Plant Physiol.* **65**, 1121–1127 (1980).
- Samoluk, S.S. *et al.* First insight into divergence, representation and chromosome distribution of reverse transcriptase fragments from L1 retrotransposons in peanut and wild relative species. *Genetica* **143**, 113–125 (2015).
- Fávero, A.P., Simpson, C.E., Valls, F.M.J. & Velo, N.A. Study of evolution of cultivated peanut through crossability studies among *Arachis ipaensis*, *A. duranensis* and *A. hypogaea*. *Crop Sci.* **46**, 1546–1552 (2006).
- Foncéca, D. *et al.* Genetic mapping of wild introgressions into cultivated peanut: a way toward enlarging the genetic basis of a recent allotetraploid. *BMC Plant Biol.* **9**, 103 (2009).
- Shirasawa, K. *et al.* Integrated consensus map of cultivated peanut and wild relatives reveals structures of the A and B genomes of *Arachis* and divergence of the legume genomes. *DNA Res.* **20**, 173–184 (2013).
- Hong, Y. *et al.* A SSR-based composite genetic linkage map for the cultivated peanut (*Arachis hypogaea* L.) genome. *BMC Plant Biol.* **10**, 17 (2010).
- Gautami, B. *et al.* An international reference consensus genetic map with 897 marker loci based on 11 mapping populations for tetraploid groundnut (*Arachis hypogaea* L.). *PLoS One* **7**, e41213 (2012).
- Qin, H. *et al.* An integrated genetic linkage map of cultivated peanut (*Arachis hypogaea* L.) constructed from two RIL populations. *Theor. Appl. Genet.* **124**, 653–664 (2012).
- Shirasawa, K. *et al.* *In silico* polymorphism analysis for the development of simple sequence repeat and transposon markers and construction of linkage map in cultivated peanut. *BMC Plant Biol.* **12**, 80 (2012).
- Nagy, E.D. *et al.* A high-density genetic map of *Arachis duranensis*, a diploid ancestor of cultivated peanut. *BMC Genomics* **13**, 469 (2012).
- Moretzsohn, M.C. *et al.* A microsatellite-based, gene-rich linkage map for the AA genome of *Arachis* (Fabaceae). *Theor. Appl. Genet.* **111**, 1060–1071 (2005).
- Moretzsohn, M.C. *et al.* A linkage map for the B-genome of *Arachis* (Fabaceae) and its synteny to the A-genome. *BMC Plant Biol.* **9**, 40 (2009).
- Samoluk, S.S., Chalup, L., Robledo, G. & Seijo, J.G. Genome sizes in diploid and allopolyploid *Arachis* L. species (section *Arachis*). *Genet. Resour. Crop Evol.* **62**, 747–763 (2015).
- Nielen, S. *et al.* FIDEL—a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut. *Chromosome Res.* **18**, 227–246 (2010).
- Bertioli, D.J. *et al.* The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodelling intergenic sequence space since its evolutionary divergence from the B genome. *Ann. Bot.* **112**, 545–559 (2013).
- Holbrook, C.C. & Culbreath, A.K. Registration of 'Tifrunner' peanut. *J. Plant Registr.* **1**, 124 (2007).
- Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
- Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
- Schultz, M.D., Schmitz, R.J. & Ecker, J.R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**, 583–585 (2012).
- Zemach, A., McDaniel, I.E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
- Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. USA* **107**, 8689–8694 (2010).
- Leal-Bertioli, S.C.M. *et al.* Genetic mapping of resistance to *Meloidogyne arenaria* in *Arachis stenosperma*: a new source of nematode resistance for peanut. *G3 (Bethesda)* **6**, 377–390 (2016).

35. Burow, M.D., Simpson, C.E., Paterson, A.H. & Starr, J.L. Identification of peanut (*Arachis hypogaea*) RAPD markers diagnostic of root-knot nematode (*Meloidogyne arenaria* (Neal) Chitwood) resistance. *Mol. Breed.* **2**, 369–379 (1996).
36. Nagy, E. *et al.* Recombination is suppressed in an alien introgression in peanut harboring *Rma*, a dominant root-knot nematode resistance gene. *Mol. Breed.* **26**, 357–370 (2010).
37. Leal-Bertioli, S.C. *et al.* Identification of QTLs for rust resistance in the peanut wild species *Arachis magna* and the development of KASP markers for marker assisted selection. *G3 (Bethesda)* **5**, 1403–1413 (2015).
38. Sujay, V. *et al.* Quantitative trait locus analysis and construction of consensus genetic map for foliar disease resistance based on two recombinant inbred line populations in cultivated groundnut (*Arachis hypogaea* L.). *Mol. Breed.* **30**, 773–788 (2012).
39. Bennetzen, J.L., Ma, J. & Devos, K.M. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95**, 127–132 (2005).
40. Tian, Z. *et al.* Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* **19**, 2221–2230 (2009).
41. Zhou, X. *et al.* Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics* **15**, 351 (2014).
42. Garcia, G.M., Stalker, H.T., Shroeder, E. & Kochert, G. Identification of RAPD, SCAR, and RFLP markers tightly linked to nematode resistance genes introgressed from *Arachis cardenasii* into *Arachis hypogaea*. *Genome* **39**, 836–845 (1996).
43. Leal-Bertioli, S. *et al.* Tetrasomic recombination is surprisingly frequent in allotetraploid *Arachis*. *Genetics* **199**, 1093–1105 (2015).
44. Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012); corrigendum **46**, 1039–1040 (2014).
45. Krapovickas, A. & Gregory, W.C. Taxonomy of the genus *Arachis* (Leguminosae). *Bonplandia* **16** (Supl.), 1–205 (2007) [transl.].
46. Smith, B. *Arachis hypogaea*, aerial flower and subterranean fruit. *Am. J. Bot.* **37**, 802–850 (1950).
47. Bertioli, D.J. *et al.* An overview of peanut and its wild relatives. *Plant Genet. Resour.; Characterization Util.* **9**, 134–149 (2011).
48. Dillehay, T.D. *et al.* Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science* **320**, 784–786 (2008).
49. Dillehay, T.D., Rossen, J., Andres, T.C. & Williams, D.E. Preceramic adoption of peanut, squash, and cotton in northern Peru. *Science* **316**, 1890–1893 (2007).
50. Simpson, C.E., Krapovickas, A. & Valls, J.F.M. History of *Arachis* including evidence of *A. hypogaea* L. progenitors. *Peanut Sci.* **28**, 78–80 (2001).
51. Bonavia, D. *Precerámico Peruano. Los Gavilanes. Mar, Desierto y Oasis en La Historia del Hombre* (Corporación Financiera de Desarrollo and Instituto Arqueológico Alemán, 1982).

ONLINE METHODS

Species accessions for genome sequencing. Stock seeds were from the Brazilian *Arachis* germplasm collection, maintained at Embrapa Genetic Resources and Biotechnology (Brasília, Brazil). Plants were maintained in pollinator-proof greenhouses. To sequence the A genome, *A. duranensis* V14167 was used: this yellow-flowered accession was collected by J.F.M. Valls, L. Novara and A. Etcheverryin in 1997 from Ruta Nacional 51, Estación Alvarado, near Tabacalera, Argentina, 24° 50' 18.4" S, 65° 27' 28.9" W, at an elevation of 1,206 m. To sequence the B genome, *A. ipaensis* K30076 was used. This accession, collected by A. Krapovickas, W.C. Gregory, D.J. Banks, J.R. Pietrarello, A. Schinini and C.E. Simpson in 1977, is the only available accession in germplasm collections worldwide. It originated from the same collection site as the holotype of this species, the species' only known site of occurrence, ~30 km north of Villa Montes, Bolivia. The site was originally recorded with the techniques available in 1977 as 21° S, 63° 25' W, at an elevation of 650 m (ref. 45) but, considering the topology of the region and the description of the site of occurrence, was more probably near Camatindi about 21° 00' 01" S, 63° 23' 37" W, at an elevation of ~600 m, or in or near Tigüipa (J.F.M. Valls and C. Simpson, personal communication).

Genome sequencing and assembly. *Sequence generation.* Illumina HiSeq 2000 paired-end sequencing libraries with insert sizes of 250 bp, 500 bp, 2 kb, 5 kb, 10 kb and 20 kb were constructed following the manufacturer's instructions. Libraries with 40-kb inserts were constructed using a fosmid-based method⁵². In total, 14 and 19 sequencing libraries were constructed for *A. duranensis* and *A. ipaensis*, respectively, from which we generated 325.73 Gb of raw data reads for *A. duranensis* and 416.59 Gb of raw data reads for *A. ipaensis*, with read lengths from 90–150 bp (**Supplementary Tables 1–5**).

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

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

Error correction. We also used *k*-mers to correct for errors. For sequencing with high depth, the *k*-mers without any sequencing errors should appear multiple times in the read data set, whereas error-containing *k*-mers should have low frequencies. We corrected sequencing errors in the 17-mers with frequencies lower than 3 in the clean data for the 250-bp and 500-bp insert libraries

Genome assembly. COPE⁵³ was used to join paired-end reads from the 250-bp insert library into single longer reads of ~250 bp. Genome assembly was performed using SOAPdenovo version 2.05 (ref. 54), with parameters --K 81 --R. Gaps were filled using KGF and Gapcloser⁵⁵ (version 1.10). Finally, SSPACE⁵⁶ was used to further link the scaffolds where connections were supported by more than five paired reads.

Production of Molecu synthetic long reads. TruSeq synthetic long-read sequencing libraries⁵⁷ were generated by Molecu and Illumina as part of beta tests of this technology. Fifteen libraries were generated for *A. duranensis* K7988, and each was sequenced on a HiSeq 2500 lane; the PE100 reads were assembled into 1.5 million TruSeq (Molecu) synthetic long reads, providing approximately 5 \times genome coverage with a mean read length of 3,684 bases and an N50 of 4,344 bases. Twelve libraries were used for *A. ipaensis* K30076 to yield approximately 2 million Molecu reads with mean length of 4,054 bases and an N50 length of 5,152 bases, providing ~6 \times genome coverage. Thirteen libraries were used for *A. hypogaea* cv. Tifrunner, which produced 1,263,111 Molecu reads with a mean length of 4,547 bases and an N50 length of 6,137

bases, providing 2.3 \times genome coverage. These reads were used for genome comparisons and were not incorporated in the diploid genome assemblies.

Mapping populations. Three RIL mapping populations derived by single-seed descent were used, a diploid A-genome population composed of 90 F₂ individuals derived from *A. duranensis* K7988 and *A. stenosperma* V10309, a diploid B-genome RIL population composed of 94 F₆ individuals derived from a cross between *A. ipaensis* KG30076 and *A. magna* KG30097, and a tetraploid AB RIL population composed of 89 F₆ individuals derived from a cross between *A. hypogaea* cv. Runner IAC 886 (ref. 58) and a colchicine-induced tetraploid *A. ipaensis* K30076 \times *A. duranensis* V14167 ($2n = 4x = 40$)¹⁵. Populations were developed and maintained in pollinator-proof greenhouses.

Linkage maps and identification of misjoins. *Conventional molecular marker maps from diploid A and B genomes and cultivated peanut \times induced allotetraploid RIL populations.* Linkage maps were constructed using Mapmaker Macintosh 2.0 (ref. 59) or as previously published. For details, see the **Supplementary Note**.

Genetic maps generated from genotyping-by-sequencing data for diploid A- and B-genome RIL populations and identification of chimeric scaffolds. Recombinant inbred lines from the diploid A- and B-genome populations were shotgun sequenced to 1 \times genome coverage with paired-end 100-bp reads on a HiSeq 2500 sequencer. The parents were sequenced at 20 \times genome coverage. Parental-homozygous SNPs were identified by alignments to the scaffolds of the *A. duranensis* and *A. ipaensis* genome assemblies as well as local realignment and probabilistic variant calling in CLC Genomics Workbench (CLC Bio). Filtering in CLC Workbench resulted in about 3 million high-quality homozygous-parental SNPs for both A- and B-genome mapping population parents. The coordinates of these SNPs were converted into BED format, and the alignment data at the SNP coordinates were extracted with SAMtools mpileup⁶⁰. From the low-coverage sequencing data, groups of 20 consecutive SNPs were haplotyped with a set of custom Python scripts. Genotype calls were inspected visually and by a hidden Markov model (HMM) script (courtesy of I. Korf, University of California, Davis) to identify population-wide switches in genotype calls corresponding to scaffold misjoins. Scaffolds not displaying recombination for an individual RIL were haplotyped. Linkage groups were identified from the haplotyping data using MadMapper and Carthagene⁶¹, applying logarithm of odds (LOD) score thresholds of 8 and distance thresholds of 50 cM; genetic maps were generated with Carthagene using the lkh traveling salesman algorithm and flips, polish and annealing optimizations. Additional scaffolds (indicated in the data files) were added to genetic bins in two rounds of binning with a custom Python script. Misjoined scaffolds were split at breakpoint locations identified by flanking GBS SNP locations, at the 'upstream SNP' and the 'downstream SNP', delineating the switches in genotype calls, and intervening sequence was excluded from the pseudomolecule assembly.

Generation of chromosomal pseudomolecules. Scaffolds less than 10 kb in length were removed (they are available in the full assembly scaffold files at PeanutBase: Adur1.split6.fa and Aipa2s.split7.fa). Sequences were subjected to RepeatMasker using *Arachis* repeat libraries available at PeanutBase (mobile-elements-AA051914.fasta and mobile-elements-BB051914.fasta). Pseudomolecules were given initial chromosomal placements and orderings according to the GBS maps. Placement was arbitrary within blocks with the same centiMorgan value. Scaffold orientation and placement were refined according to the conventional maps using, in order of priority, the tetraploid AB-genome map, the diploid F₂ A-genome Nagy map²² (for the *A. duranensis* assembly), the diploid B-genome map²⁴ (for the *A. ipaensis* assembly) (**Supplementary Data Set 2**) and finally the tetraploid AB-genome Shirasawa map¹⁷. Markers were located on the scaffolds using BLAST and ePCR (electronic PCR) with high similarity parameters (taking the top hits only, with placement by BLAST (e value $< 1 \times 10^{-10}$) given preference over ePCR where both were available). Markers placing scaffolds on linkage groups other than the one assigned by the GBS data were dropped.

Where allowed by map data, scaffold positions and orientations were adjusted using synteny between the two *Arachis* species and, where necessary (generally within pericentromeric regions), synteny with *G. max* and

Proteus vulgaris; the presence of telomeric repeats near chromosome ends; information from repeat-masked paired-end sequences from 42,000 BAC clones of *A. duranensis* V14167 (FI321525–FI281689) and Moleclo sequence reads from *A. ipaensis* and *A. duranensis*. Apparent inversions were visually inspected and confirmed. Scaffolds with either <5,000 non-N bases or <20,000 bp in length and with <10,000 non-N bases were removed. Pseudomolecules were generated with 10,000 Ns separating the scaffold sequences and were oriented and numbered in accordance with previously published maps^{17,19,23,24}.

Characterization of transposons. Mobile elements were identified using a number of homology and *de novo* structural pattern finding algorithms and manual curation. For details, see the **Supplementary Note**.

Estimation of transposon coverages. All annotated transposons were combined together and used as a library to screen the diploid genomes (pseudomolecules and unplaced scaffolds) using RepeatMasker with default parameters except with -nolow and -norna to not mask low-complexity sequences and rDNA, respectively. The output files were summarized using a custom Perl script, and regions masked by more than one sequence in the repeat library were recognized and counted only once. Base-pair counts for the diploid genomes excluded gaps.

Gene prediction and annotation. Genome assemblies were masked with RepeatMasker using the repeat libraries developed for the two diploid species and annotated for gene models using the MAKER-P pipeline⁶². *Arachis*-specific models for the *ab initio* gene predictor SNAP⁶³ were trained using high-scoring gene models from a first iteration of the pipeline and then used in the final annotation pass; no training was done for the other *ab initio* predictors included in the pipeline. RNA sequencing *de novo* assemblies for *A. hypogaea* and the diploid *Arachis* species were supplied as transcript evidence along with available EST and mRNA data sets from NCBI for these same species. Further evidence was supplied by proteomes derived from the annotations for *G. max*, *P. vulgaris* and *Medicago truncatula* as represented in Phytozome v. 10 (ref. 64). Default MAKER-P parameters were used for all other options, with the exception of disabling splice isoform prediction (alt_splice = 0) and forcing start and stop codons into every gene (always_complete = 1). The resulting MAKER-P gene models were post-processed to exclude from the main annotation files gene models with relatively poor support (annotation evidence distance scores of ≥ 0.75) or with significant BLASTN⁶⁴ homology to identified mobile elements (HSP (high-scoring segment pair) coverage over $\geq 50\%$ of the transcript sequence at $\geq 80\%$ identity and e value $\leq 1 \times 10^{-10}$). Provisional functional assignments for the gene models were produced using InterProScan⁶⁵ and BLASTP⁶⁶ against annotated proteins from *Arabidopsis thaliana*, *G. max* and *M. truncatula*, with outputs processed using AHRD (<https://github.com/groupschoof/AHRD>), for lexical analysis and selection of the best functional descriptor of each gene product.

Analysis of gene duplications. The protein sequences of precalculated gene families were downloaded from Phytozome 10. Multiple-sequence alignments were built for each gene family using Muscle⁶⁷. HMMs⁶⁸ were built from each gene family alignment and were searched against the protein sequences of *A. duranensis* and *A. ipaensis* using HMMER. Genes were assigned family IDs using their best hits. Local gene duplication was defined as genes from the same gene family within ten successive genes and was calculated by a sliding window with a window size of ten genes and a step of one gene. After the calculation, only the number of locally duplicated genes was recorded.

DNA methylation analysis. See the **Supplementary Note**.

Disease resistances and NB-LRR-encoding genes. We used two complementary approaches to identify R-gene candidates: HMM scans (HMMER 3.1b1; ref. 68) against the Pfam protein domains TIR, NB-ARC and nine LRRs plus the *Arabidopsis* NB-ARC domain (NBS_712) to provide information about domain composition and a BLASTP search⁶⁹ against two consensus sequences for the TIR and non-TIR classes of the NB-ARC-encoding genes to assist in distinguishing the two classes (NBS-TIR and NBS-CC)⁷⁰. Results were compiled in Microsoft Excel for further analysis, and candidates that matched

with an expectation value better than 1×10^{-10} were considered significant (**Supplementary Data Set 6**).

Gene evolution in *A. ipaensis* and *A. duranensis* and species divergence. All-by-all synteny and K_S comparisons were made between *A. duranensis*, *A. ipaensis*, *G. max*, *Lotus japonicus*, *M. truncatula* and *P. vulgaris*. Synteny blocks were identified within and between these species using DAGChainer, on the basis of gene alignments. K_S values for aligned genes were calculated using the codeml method from the PAML package⁷¹. Median values were then taken for each synteny block. On the basis of known WGD and speciation information, the structure of a phylogenetic tree for species of interest was constructed, with branch lengths derived from modal K_S values from the K_S plots. Internal branch lengths were calculated from a system of equations based on modal values from all species comparisons.

Analysis of chromosomal structure and synteny. Structural comparisons between *A. ipaensis* and *A. duranensis* were made using visual interpretations of dot plots created using mummer and mummerplot from the MUMmer suite of alignment tools⁷². Gene density plots were created using CViT⁷³. Synteny comparisons were also made with other legume genomes (*G. max*, *L. japonicus*, *M. truncatula* and *P. vulgaris*), using MUMmer and DAGChainer⁷⁴.

Sequence comparisons to tetraploid cultivated peanut. Moleclo reads were mapped against the diploid chromosomal pseudomolecules using nucmer maxgap = 500 -mincluster = 100. Show-coords, a nucmer utility, was run on the resulting nucmer delta files to produce alignment files. A single 'best' alignment to the diploid pseudomolecules was selected for each *A. hypogaea* cv. Tifrunner Moleclo read. Alignment selection was based primarily on length and secondarily on identity. Show-tiling, another nucmer utility, was used to produce the tiling path of Moleclo reads. Output files were further processed using in-house scripts and Microsoft Excel (**Supplementary Data Set 9**).

Analysis of genetic exchange in cultivated peanut RILs. Paired-end sequence reads of restriction site-associated DNA sequencing for 166 RILs and their parents were obtained from the Sequence Read Archive (SRR1236437 and SRR1236438)⁴¹. The data were divided into 168 subsets of individuals (2 parents and 166 RILs) on the basis of index tags. Low-quality sequences (quality value of <10) and adaptors (AGATCGGAAGAGC) were trimmed with PRINSEQ⁷⁵ and fastx_clipper in the FASTX-Toolkit. The filtered reads were mapped onto the two *Arachis* diploid genomes with Bowtie 2 (ref. 76) (parameters of --minins 0 --maxins 1000). The resultant SAM files excluding reads mapped at multiple loci on the reference were converted to BAM files with SAMtools⁶⁰. Depth of coverage in 1-Mb bins was calculated from the BAM files with the GenomeCoverageBed option in BEDtools⁷⁷ (using parameter -d). Biases with depth of coverage among the RILs due to different numbers of mapped reads were corrected by converting the depth of coverage into the percentage relative to the total number of mapped reads in each line. Log₂-transformed ratios of the corrected values in each RIL to that in the parent were calculated and plotted with R (ref. 76) and Excel.

Generation and assembly of transcribed sequences. Details of the tissues sampled can be found in **Supplementary Table 12** and the **Supplementary Note**.

Sequencing of cDNAs. Libraries were constructed with Illumina TruSeq RNA Sample Preparation v2 kit (tissues listed in **Supplementary Tables 12** and **13**). Three biological replicates were used for diploid tissue, and five biological replicates were used for *A. hypogaea* cv. Tifrunner tissues (for some seed stages, only two biological replicates were used). For diploids, RNA from biological replicates was pooled before generating the libraries. For tetraploids, libraries constructed individually for each biological replicate were combined in equimolar pools for sequencing. Libraries were sequenced on an Illumina HiSeq 2500 instrument with a total of 209 cycles of TruSeq Rapid SBS kit v1 (Illumina) chemistry. For the diploids, to obtain longer reads to improve transcriptome assemblies, size-selected libraries were sequenced using an Illumina MiSeq instrument with v3 chemistry, and additional paired-end sequencing data were generated on Illumina NextSeq500.

Quality control and k-mer normalization. Total raw reads were trimmed 10 bp from the 5' end and 2 bp from the 3' end after inspection of nucleotide bias using FastQC. Trimmed reads were then aligned to a compiled

set of rRNA sequences using Bowtie, allowing two mismatches per 25-bp seed, and mapped reads were discarded. rRNA contamination varied from as low as 0.39% to as high as 51%, but most libraries had between 1–2% rRNA contamination. A total of 2,064,268,316 paired-end reads (4,128,536,632 total 100-nt reads) were subjected to *k*-mer normalization using the Trinity package⁷⁸. A script was used to randomly discard reads with mean *k*-mer coverage of more than 20.

Transcriptome assemblies. Adaptor and quality trimming was performed using Trim Galore! v0.3.5. Transcripts were assembled using the genome-guided pipeline from Trinity⁷⁹. For *A. duranensis* and *A. ipaensis*, reads were mapped to their respective genomes using GSNAP^{80,81}. For *A. hypogaea* cv. Tifrunner, a diploid genome-guided tetraploid assembly was carried out: total reads were mapped to an *in silico* tetraploid genome (a concatenate of the chromosomal pseudomolecules of *A. duranensis* and *A. ipaensis*). Once the reads were mapped, the SAM files were run through the genome-guided pipeline. Briefly, loci of reads were extracted into separate directories, where they were then assembled on a locus-by-locus basis. For the *A. hypogaea* cv. Tifrunner assembly, information on whether loci were guided by *A. duranensis* or *A. ipaensis* was retained so that transcripts were annotated as being either 'A' or 'B'.

Expression-based filtering of the final assembly of tetraploid transcripts. Total reads were mapped to the transcript assembly from the 58 libraries using Bowtie, allowing two mismatches in a 25-bp seed. Fragments per kilobase per million reads mapped (FPKM) values were estimated using RSEM⁸² for each library. When reads map to multiple transcripts, RSEM fractionates the read count among the transcripts such that read counts are not integers. Transcripts were filtered out that had FPKM <1 for all 58 libraries using filter_fasta_by_rsem_values.pl from the Trinity package and were deemed lacking in minimum read coverage evidence to be supported. Expression-based filtered transcripts were tested for redundancy using a custom script to retain locus information from the assembly in the transcript names. Filtering was performed using an intra-subgenome self-BLAST. Transcripts with 90% or greater coverage and 100% identity were filtered out, leaving the longer transcript. Transcripts were aligned to the annotated repetitive sequences from *A. duranensis* and *A. ipaensis* using GMAP⁸², with the following parameters: -n 4 where -n controls the number of paths.

Estimation of the accuracy of transcriptome assembly of *A. hypogaea* cv. Tifrunner using diagnostic sequences. *A. duranensis* and *A. ipaensis* pseudomolecules were fragmented into 100-bp fragments that were mapped to their opposite genome using Bowtie. The SAM files were filtered for fragments that mapped uniquely and completely (no clipping) with a maximum of only one mismatch to the opposite genome. These fragments were collected as diagnostic sequences. To test the accuracy of the assembled transcripts, diagnostic sequences were mapped to the transcript assembly using GSNAP with the following parameters: -n 1 -m 0 -A sam --nofails. Fragments diagnostic for A and mapping to A-derived transcripts were counted as correct, and those mapping to B-derived transcripts were counted as ambiguous. This testing was also done for B diagnostic transcripts.

Diploid protein comparisons. Assembled transcripts were compared to the combined *A. duranensis* and *A. ipaensis* predicted protein models using BLASTX (*e* value < 1×10^{-20}), and the best hit was taken for each. Using AWK, sets of hits were filtered for the following criteria: >80% amino acid identity and >70% coverage of the protein model; >80% amino acid identity and >80% coverage of the protein model; 90% amino acid identity and >80% coverage of the protein model; and >90% amino acid identity and >90% coverage of the protein model.

52. Zhang, G. *et al.* Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* **30**, 549–554 (2012).
53. Liu, B. *et al.* COPE: an accurate *k*-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics* **28**, 2870–2874 (2012).
54. Li, R. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
55. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
56. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
57. McCoy, R.C. *et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**, e106689 (2014).
58. Godoy, I.J. *et al.* *Cultivares de Amendoim: Novas Opções para o Mercado de Confeitaria* (Campinas: Instituto Agrônômico, 2003).
59. Lander, E.S. *et al.* MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181 (1987).
60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Schiex, T. & Gaspin, C. CARTHAGENE: constructing and joining maximum likelihood genetic maps. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 258–267 (1997).
62. Campbell, M.S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).
63. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
64. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421–430 (2009).
65. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
66. Hallab, A. *Protein Function Prediction Using Phylogenomics, Domain Architecture Analysis, Data Integration, and Lexical Scoring*. PhD thesis, Univ. Bonn (2015).
67. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
68. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
69. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
70. Ameline-Torregrosa, C. *et al.* Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. *Plant Physiol.* **146**, 5–21 (2008).
71. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
72. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
73. Cannon, E.K. & Cannon, S.B. Chromosome visualization tool: a whole genome viewer. *Int. J. Plant Genomics* **2011**, 373875 (2011).
74. Haas, B.J., Delcher, A.L., Wortman, J.R. & Salzberg, S.L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646 (2004).
75. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
76. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
77. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
78. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
79. Haas, B.J. *et al.* *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
80. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
81. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
82. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).