The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

David J. Bertioli, Jerry Jenkins, Josh Clevenger, Olga Dudchenko, Dongying Gao, Guillermo Seijo, Soraya C. M. Leal-Bertioli, Longhui Ren, Andrew D. Farmer, Manish K. Pandey, Sergio S. Samoluk, Brian Abernathy, Gaurav Agarwal, Carolina Ballén-Taborda, Connor Cameron, Jacqueline Campbell, Carolina Chavarro, Annapurna Chitikineni, Ye Chu, Sudhansu Dash, Moaine El Baidouri, Baozhu Guo, Wei Huang, Kyung Do Kim, Waldid Korani, Sophie Lanciano, Christopher G. Lui, Marie Mirouze, Márcio C. Moretzsohn, Melanie Pham, Jin Hee Shin, Kenta Shirasawa, Senjuti Sinharoy, Avinash Sreedasyam, Nathan T. Weeks, Xinyou Zhang, Zheng Zheng, Ziqi Sun, Lutz Froenicke, Josh Clevenger, Olga Dudchenko, C. Corley Holbrook, Jane Grimwood, Peggy Ozias-Akins, Jeremy Schmutz.

Like many other crops, the cultivated peanut (*Arachis hypogaea* L.) is of hybrid origin and has a polyploid genome that contains essentially complete sets of chromosomes from two ancestral species. Here we report the genome sequence of peanut and show that after its polyploid origin, the genome has evolved through mobile-element activity, deletions and by the flow of genetic information between corresponding ancestral chromosomes (that is, homeologous recombination). Uniformity of patterns of homeologous recombination at the ends of chromosomes favors a single origin for cultivated peanut and its wild counterpart *A. monticola*. However, through much of the genome, homeologous recombination has created diversity. Using new polyploid hybrids made from the ancestral species, we show how this can generate phenotypic changes such as spontaneous changes in the color of the flowers. We suggest that diversity generated by these genetic mechanisms helped to favor the domestication of the polyploid *A. hypogaea* over other diploid *Arachis* species cultivated by humans.

The domestication of plants, thousands of years ago, increased food supply and allowed the formation of large, complex human societies. Out of many thousands of wild species, only a few became domesticated crops and they now provide most of the food consumed by humans. It has long been noted that many of these crops are polyploid: their nuclei have more than two sets of chromosomes that are often derived from different species. Although it has been surprisingly difficult to rigorously demonstrate, it has long been thought that domestication may favor polyploids.

Peanut (also called groundnut; *Arachis hypogaea* L.) is an important food crop (annual production of ~44 million tons based on
FAOSTAT data for 2016 (http://www.fao.org/faostat/en/#home). Whereas almost all related species in the genus Arachis are diploid (two sets of ten chromosomes; mostly 2n = 2x = 20 chromosomes), A. hypogaea is polyploid11. The seeds of all of these species are an attractive food, and several have been cultivated for thousands of years (Supplementary Note 1). Indeed, the action of humans was key to the formation of A. hypogaea itself. About 9,400 years ago (estimated by nucleotide divergence), the human transport of the ‘B’ genome species, A. ipaensis Krapov. & W.C. Greg., into the range of the ‘A’ genome species A. duranensis Krapov. & W.C. Greg. enabled their hybridization and the formation of A. hypogaea. It has two sets of chromosome pairs, one from each of the ancestral species: a type of polyploid termed allotetraploid (AABB-type genome; 2n = 4x = 40 chromosomes; genome size of ~2.7 Gb).

The origin of A. hypogaea was associated with a particularly severe population bottleneck3-5. This could, in principle, have reduced the variability on which, over generations, human selection could act. However, A. hypogaea evolved, becoming completely dependent on cultivation and morphologically very diverse. Two subspecies (hypogaea and fastigia) and six botanical varieties (hypogaea, hisrita, fastigia, vulgaris, aequatoriana and peruviana) are recognized4. Different grain colors and sizes, pod shapes and growth habits distinguish thousands of landraces and cultivars5 (see also United States Department of Agriculture (USDA) Germplasm Resources Information Network (https://www.ars-grin.gov)). It seems notable that, in spite of the higher genetic diversity of the diploid species22, and their cultivation starting earlier (Supplementary Note 1), it was the derived allotetraploid, A. hypogaea, that underwent the transformation to become the crop of worldwide importance.

Some time ago, while planning to sequence and assemble the peanut genome, we realized that it would not be possible using the short-read data (~100–200bp DNA) that were generated by the only technology that was economically feasible at the time; such sequences were too short to reliably resolve the very similar A and B genomes, which frequently have more than 98% DNA identity between corresponding genes6-8. This level of similarity is due to the progenitor species that gave rise to the two subgenomes having diverged only around 2.2 million years ago (refs. 6,4,14). Therefore, as a foundation for understanding the genome of cultivated peanut, we first sequenced the genomes of both the diploid ancestral species. These diploid genomes afforded new insights into peanut genetics. Notably, it was possible to infer that some chromosome ends of A. hypogaea had changed from the expected AABB structure to AAAA or BBBB, implying a particular complexity in peanut genetics6,15-18.

Here, using the much longer-read data obtained with PacBio technology9, and scaffolding using Hi-C10,11, a method used for determining the conformation of DNA in the nucleus, we report the complete chromosome-scale genome sequence of A. hypogaea cv. Tifrunner, a runner-type peanut. We also characterize the genomes of a diverse selection of cultivated peanuts, together with its wild counterpart, A. monticola Krapov. & Rigoni, and induced allotetraploid hybrids derived from the ancestral species. We are able to visualize, in considerable detail, the products of variable deletions from, and genomic recombination between, the A and B subgenomes. It seems likely that these variations in genome structure generated phenotypic variation on which selection could act, and helped to favor A. hypogaea over its diploid relatives during the process of domestication.

Results
Sequencing and assembly of the peanut genome. Arachis hypogaea cv. Tifrunner12, a runner-type peanut (registration number CV-93, PI 644011) was sequenced using whole-genome shotgun sequencing. Twenty chromosome sequences were produced (for assembly metrics see Supplementary Tables 1 and 2). They were numbered Arathy.01–Arathy.20, where the A subgenome is represented as Arathy.01–Arathy.10 and the B subgenome as Arathy.11–Arathy.20. The chromosome sequences contain 99.3% of the assembled sequence and are 2.54 Gb, 93% of the size estimated by flow cytometry21.

Chromosome architecture. The chromosomes of A. hypogaea cv. Tifrunner largely reflect their ancestral structures; the homologous chromosomes mostly have a one-to-one correspondence: Arathy.02/12, 03/13, 04/14 and 10/20 are almost completely collinear; 06/16 and 09/19 are differentiated by a large inversion in one arm; 05/15 are differentiated by two large inversions; and 01/11 are differentiated by three large inversions. Chromosomes 17/18 have undergone reciprocal translocations relative to 07/08 (Supplementary Figs. 1–12). Gene densities are highest in distal chromosome regions (Supplementary Fig. 13). Gene counts are 11% higher in the B subgenome, with 35,110 predicted genes, compared to 31,359 genes in the A subgenome. Long terminal repeat (LTR) retrotransposons are highly abundant in pericentromeric regions, whereas DNA transposons are more frequent in euchromatic arms (Supplementary Fig. 14). Other transposable elements, together with approximately 3,300 pararetrovirus sequences account for 74% of the assembled genome sequence (Supplementary Tables 3 and 4). Notably, this compares to 64% repetitive content estimated by readassociation kinetics22, indicating the high quality and relative lack of collapse of repeats in this long read-based assembly. The chloroplast genome of A. hypogaea and a chloroplastic plasmid were inherited from A. duranensis (Supplementary Fig. 15).

DNA methylation and small RNAs. Genic methylation patterns were typical for plants, with lower methylation in transcribed regions and characteristic dips in methylation at transcription start and end sites (Supplementary Fig. 16). Genome-wide methylation per cytosine content was higher in pericentromeric regions than chromosome arms (Supplementary Fig. 17). Methylation was lower in the A subgenome than the B subgenome; with 76.0% and 80.5% methylation at CG sites, 61.7% and 65.1% methylation at CHG sites (where H is an A, T or C) and 5.14% and 5.51% methylation at CHH sites (Supplementary Table 15). Greater densities of DNA sequences corresponding to small RNAs were found in proximal, repetitive-rich regions of chromosomes (Supplementary Fig. 19). However, greater densities of DNA sequences that corresponded to uniquely mapping small RNAs were found in gene-rich chromosomal regions (Supplementary Fig. 20). Within genes, the B subgenome was enriched relative to the A subgenome for DNA sequences that corresponded to small RNAs (Supplementary Fig. 18b).

Comparison of gene expression in subgenomes. The expression of homeologous gene pairs (dataset 1a in ref. 21) from the A and B subgenomes of Tifrunner was investigated in diverse tissues and developmental stages (dataset 1bc in ref. 21). As has been reported in other recent polyploids26,27, overall, the number of homeologous gene pairs with expression biased towards the A subgenome was not significantly different from the number biased towards the B subgenome (P = 0.2, two-sided binomial test; n = 3,648 and 3,759 for A and B, respectively). However, when tissues were considered separately, all but one had slightly more B than A subgenome-biased genes from homeologous pairs. In three reproductive tissues and in roots this difference was significant (P < 0.05, one-sided binomial test; Supplementary Fig. 21; dataset 1 in ref. 21).

Broadly, homeologous pairs with the highest asymmetry in expression (log(expression ratios) > 3, Benjamini–Hochberg-adjusted P < 0.05, Wald test; Supplementary Fig. 22) were more commonly involved in oxidation–reduction processes, pollen recognition, lipid and chitin metabolic processes and response to biotic stimulus (Supplementary Fig. 23a; dataset 1c in ref. 21).
Taking the example of the subterranean peg tip (a unique reproductive structure in peanut), the A subgenome-biased homeologous pairs were enriched for genes involved in mannose metabolic processes, nitrate assimilation and cell wall assembly, whereas the B subgenome-biased homeologous pairs were enriched for genes involved in the response to biotic stimulus, sucrose transport and glucan metabolic processes. In the maturing pericarp (Pattee stage 6), the A subgenome-biased homeologous pairs were enriched for genes involved in phosphorylation signal transduction, carbohydrate metabolism and cell wall biogenesis, whereas B subgenome-biased homeologous pairs were enriched for genes involved in inorganic ion transport and response to biotic stimulus (Supplementary Fig. 23a, ref. 25). Additionally, we identified homeologous gene pairs with biased homeologous pairs were enriched for genes involved in mannose metabolic processes. In the maturing pericarp (Pattee stage 6), the A subgenome-biased homeologous pairs were enriched for genes involved in the response to biotic stimulus, sucrose transport and glucan metabolic processes. In the maturing pericarp (Pattee stage 6), the A subgenome-biased homeologous pairs were enriched for genes involved in phosphorylation signal transduction, carbohydrate metabolism and cell wall biogenesis, whereas B subgenome-biased homeologous pairs were enriched for genes involved in inorganic ion transport and response to biotic stimulus (Supplementary Fig. 23a, ref. 25) and, as might be expected, the consistently asymmetrically expressed homeologous pairs were mainly enriched for functions associated with fundamental biological processes such as organelle organization, molecular transport and protein complex biogenesis (dataset 1c in ref. 25).

Changes following polyploidy. Genetic exchange between subgenomes and deletions. For allotetraploids, chromosome associations during meiosis and genetic exchange are mostly limited to corresponding chromosomes within the same subgenome (that is, homologous chromosomes); however, as has been characterized in other plants such as Brassica26,27,28,29, these may also occur at lower frequency between corresponding chromosomes from the other subgenome (that is, homeologous chromosomes)26,27,28,29. We investigated genetic exchange between the subgenomes and deletions in more than 200 diverse genotypes comprising the wild tetraploid peanut (A. monticola), landraces and cultivars of A. hypogaea, and new allotetraploid hybrids made from the ancestral species (dataset 2 in ref. 25). Two different approaches were used: observation of mapping densities of short-read whole-genome sequences onto the combined sequenced diploid ancestral species genomes, and analysis of the...
short-read whole-genome sequences for single-nucleotide polymorphisms (SNPs) that consistently differentiate representatives of A and B genome diploid species5,9,30–32 (Supplementary Fig. 24). (It should be noted that, except for the assembled reference genotype of Tifrunner, these methods are not capable of detecting genome changes that result from balanced homeologous exchanges or chromosome rearrangements.)

Genetic exchange between ancestral genomes could be inferred towards the ends of collinear pairs of homeologous chromosomes. In these regions, the genome structure was not the expected AABB, but may be better described as AAAA or BBBB, that is, ‘tetrasomic’ conformations. The abrupt junctions of these segments signify that they may have occurred by crossover (Figs. 1 and 2 and Supplementary Figs. 1–12 and 25; datasets 3–5 in ref. 25). In Tifrunner, 14.8 Mb of the A genome has been transferred, in blocks, into B chromosomes, and 3.1 Mb of the B genome has been transferred, in blocks, into A chromosomes (Supplementary Tables 6 and 7). Most of these tetrasomic regions are at the very distal ends of chromosomes—for example, the lower regions of Arahy.02/Arahy.12, Arahy.04/Arahy.14, Arahy.06/Arahy.16 and the upper regions of Arahy.05/Arahy.15—and these were present in all of the A. hypogaea and A. monticola genotypes surveyed (but not in induced allotetraploids derived from the same diploid ancestral A. hypogaea and A. ipaensis (normalized values, in blue and red respectively; distances scaled to Arahy.14). Where mapping densities cluster around the expected value of one, the genome composition is AABB. Where mapping densities on one genome increase to approximately two and on the other genome decrease to near zero, the genome composition is better described as AAAA or BBBB. Mapping densities decrease to around zero on one genome and remain around one on the other indicate a deletion (common in b). d, A panel of 68 representative diverse genotypes in a region of Arahy.04 and Arahy.14 in which hypervariability has been created by differential recombination between subgenomes. The panel represents a heat map of log_2-transformed ratios of mapping densities on the B and A genomes; blue represents AAAA and red represents BBBB, Tif., PI and Chib. are the genotypes represented in a–c, respectively (for full visualizations, see dataset 4b in ref. 25).

The signals of disperse genetic exchange were also detectable through the bodies of chromosomes. Overall, this dispersed genetic exchange has had a greater total effect than the transfer of chromosome segments. In Tifrunner, almost twice as many B alleles have been transferred to A chromosomes than vice versa (Supplementary Table 6; dataset 3 in ref. 25). In addition, variable deletions were frequent in proximal chromosome regions (Fig. 2; dataset 4c in ref. 25). Notably, a large deletion (around 10 Mb) was common on Arahy.14 of botanical varieties fastigata and vulgaris (e.g., Fig. 2b).

In Tifrunner, genome deletions have disproportionally affected some gene families. The genes most frequently lost were members of the serine/threonine-protein phosphatase (around 89 genes) and FAR1-related families (around 83 genes). Genes in these families...
Furthermore, significantly higher DNA identity between Arahy.07 and five *A. duranensis* accessions (including the closest ones to the A subgenome ancestor; see below) is observed when compared to others (Supplementary Tables 8 and 9). Similarly, for Arahy.05, markedly higher identities to three *A. duranensis* accessions may indicate the presence of the inversion in some representatives of *A. duranensis*, possibly including the ancestral A subgenome donor (Supplementary Table 8 and 9).

We previously reported that inversions move repeat-rich DNA to more distal chromosome regions where DNA is lost by recombination, thus reducing genome size (although regions moved to more proximal positions gain DNA, this effect is smaller)\(^{3,7}\). Following this pattern, the inverted region in Arahy.05 has shrunk relative to *A. duranensis* V14167 (tetraploid size/diploid size = 0.89; Supplementary Table 10). We found that removal of LTR retrotransposons is the predominant cause of this reduction (Supplementary Fig. 29). Furthermore, the presence of repeats in *A. duranensis*, at the ends of the regions, which are missing in *A. hypogaea*, clearly implicates unequal intrasubgenome recombination in about 20% of cases (107 out of 502 regions). By contrast, there is little difference in relative sizes of the inversions on Arahy.07 and Arahy.11.

**Observations of independent polyploidy events.** We used allotetraploids derived by colchicine treatment of hybrids of the peanut's ancestral diploid species\(^{38}\) to investigate genome changes that followed independent polyploidy events. We studied 37 different lineages from two independent induced polyploidy events. Genetic exchange between subgenomes occurred in large blocks and interspersed alleles along chromosome segments; these events seem at least partly stochastic, and were different between different lineages and from *A. hypogaea*. Spontaneous changes in flower color in some lineages (Fig. 3) could be ascribed to genetic exchange between subgenomes; the A genome region that confers the yellow flower color had been replaced by the homeologous B genome region that confers orange flower color (dataset 6 in ref. \(^{25}\)). This provides a simple demonstration of phenotypic change as a consequence of genetic exchange between subgenomes.

A closer representative of the A subgenome ancestor. Because their seeds develop underground, wild *Arachis* populations are unusually static over time\(^{48}\). In addition, they typically have very high rates of self-pollination. This, and a serendipitous collection by pioneering botanical collectors, enabled our previous discovery that the sequenced *A. ipaensis* K30076 was very likely a descendant of the same population that donated the B subgenome to *A. hypogaea*. Here we endeavored to identify the extant *A. duranensis* population that is closest to the A subgenome donor. We characterized 55 accessions, representative of all known major populations of *A. duranensis*, by sequencing DNA enriched for genomic regions (using exome capture methods). A selection of these accessions was chosen for whole-genome re-sequencing. The *A. duranensis* accessions that were most similar to the Tifrunner A subgenome were from Rio Seco (Argentina), a location previously indicated as the likely origin of the A subgenome ancestor on the basis of chloroplast and ribosomal DNA haplotypes\(^{49}\) (Fig. 4 and Supplementary Tables 8, 9 and 11). However, in some cases, the ranking of similarity changed by chromosome (especially for Arahy.05), possibly reflecting variations in chromosomal arrangements in different accessions of *A. duranensis* (as discussed above; Supplementary Table 9). Comparisons of the Tifrunner A subgenome with the whole-genome sequences of *A. duranensis* accessions indicated median DNA identities of 99.76% for the Rio Seco accessions (KGBSPSc 30065, PI 468201 and KGBSPSc 30067, PI 468202); 99.61% for the sequenced V14167 (ref. \(^{25}\)); and 98.23% for PI 475845 from the northern range of the species and with a partially assembled genome\(^{49}\) (Supplementary Table 8 and Supplementary Fig. 30; dataset 7 in ref. \(^{25}\)).
Polyploidy has long been recognized as an important feature of almost all flowering plants. Following each polyploidy event, over tens of millions of years, deletions, divergence of duplicated genes and rearrangements return the genome to a diploid state. The recurrence of these ‘wondrous cycles’ is thought to have played an important part in diversification and adaptation during plant evolution\(^{35-41}\). It has also long been recognized that many crop plants are recent polyploids; and, although the matter has generated decades of debate, it does seem that polyploids are favored for domestication\(^{12}\).

We consider the evidence that polyploid *A. hypogaea* was favored for domestication over its diploid relatives very persuasive. Archaeological remains and remnant populations of *Arachis* species far from their natural distributions, and the existence of a diploid domesticated species (*A. villosa* or *A. carici*) testify to widespread and large-scale cultivation of at least four diploid species (Supplementary Note 1). Indeed, the hybridization that gave rise to *A. hypogaea* was only possible because of human transport of *A. ipaensis* into the range of *A. duranensis*. It seems important that, in spite of higher genetic diversity of the diploid species and their cultivation having started earlier, it was—in fact—the allotetraploid *A. hypogaea* that became the crop of worldwide importance.

Following trends seen in many plants, *Arachis* allotetraploids are larger than their diploid progenitors. The tetraploids also have different transpiration characteristics\(^{35-41}\) and produce more photosynthetic pigments\(^{35-41}\). These traits—or other ploidy-related changes—may have been advantageous; however, contrary to common expectations, the seeds of the allotetraploid ancestor of peanut seem likely to have been similar size to those of its diploid progenitors\(^{35-41}\). The increased number of alleles associated with being a ‘fixed hybrid’ would have increased heterosis and therefore probably adaptability. However, the extreme genetic bottleneck that accompanied the polyploid origin may have been expected to reduce variability on which artificial selection could act. We investigated genome changes after polyploidy that could have generated variation. We found no evidence for widespread mobilization of transposable elements (Supplementary Fig. 27). However, we could identify some mobile element insertion polymorphisms and some of these are likely to have influenced gene activity (Supplementary Fig. 28). In addition, variable deletions, especially in proximal chromosome regions, have occurred since polyploidy and these also must have generated variation. However, it was a different genetic phenomenon, associated with harboring full chromosome complements from two species, that most drew our attention: genetic exchange between subgenomes\(^{36,42-49}\).

We identified two patterns of homeologous recombination. One involves the transfer of chromosome segments between distal collinear regions of chromosomes mostly resulting in tetrasomic genome structures (AAAA and, to a lesser extent, BBBB; Figs. 1 and 2 and Supplementary Figs. 1–12, 25). The other involves transfer of dispersed alleles that has occurred throughout the chromosomes; it is strongly biased, with much more transfer of alleles from B subgenome to A subgenome (Supplementary Tables 6 and 7). Overall, the genetic flux seems to have caused a greater erosion of similarity of the A subgenome to its progenitor *A. duranensis* than of the B subgenome to its progenitor *A. ipaensis* (even though the distal regions of the B chromosomes are more invaded by segments of the A genome than vice versa). Collections from Rio Seco were the closest representatives of the A subgenome ancestor, although several accessions from Salta (including the sequenced V14167 (ref. 25)) showed quite similar degrees of identity (Fig. 4, Supplementary Tables 8, 9 and 11 and Supplementary Fig. 30).

On the whole-genome scale, the effects of homeologous recombination appear similar in diverse peanut accessions. Most of the tetrasomic structures were present in all *A. hypogaea* and *A. monticola* analyzed; furthermore, fingerprint-like fine-scale patterns of interspersed homeologous alleles within the distal tetrasomic regions were also found to be uniform (Fig. 1; datasets 4a and 5 in ref. 25).

**Fig. 4 | Similarity of *A. duranensis* from different locations to the A subgenome of Tifrunner.** Genomic DNAs of 55 accessions, representing all known major populations of *A. duranensis*, were compared to the A subgenome of Tifrunner. Similarity is strongly influenced by hydrographic basins. Accessions with the highest similarity (in red) are concentrated around Rio Seco, a tributary of the Rio San Francisco; next in similarity (in orange) are accessions concentrated around Jujuy, a region that drains into the Rio San Francisco and the Lerma valleys; followed by accessions from the Rio Juramento (in light brown), a region that receives water from the Lerma valley. Following these in similarity are accessions from the endorheic basins that occasionally drain in the Bermejo River (northwest Argentina and south Bolivia) (in yellow) followed by accessions in the basins of the Rio Pilcomayo (in light green) and finally accessions from the Rio Parapeti basin, Isozog Swamps and West Paraguay sand dunes (in dark green). Outliers to this general pattern are likely to represent populations that have resulted from the occasional human movement of seeds among basins (most of these movements are likely to have occurred long ago). The maps were generated using Natural Earth.

The A subgenome chromosomes are, in general, less similar to their *A. duranensis* counterparts than the B subgenome chromosomes are to their *A. ipaensis* counterparts. This is consistent with the greater flow of alleles from the B subgenome into the A subgenome than vice versa (as described above, see also a previously published study\(^7\)).

**Discussion**

A genome sequence is a landmark for the research of the biology of a crop. It provides a catalog of gene content, with chromosomal context and a unified framework for biological investigations and cross-species comparisons. In the case of peanut, a polyploid of recent hybrid origin, the previous sequencing of very close representatives of its diploid ancestors provides the opportunity to investigate more generally applicable principles regarding the genetics of polyploidy and its importance to crop domestication.

Polyploidy has long been recognized as an important feature of plant evolution; it has occurred multiple times during the evolution of almost all flowering plants. Following each polyploidy event, over tens of millions of years, deletions, divergence of duplicated genes and rearrangements return the genome to a diploid state. The recurrence of these ‘wondrous cycles’ is thought to have played an important part in diversification and adaptation during plant evolution\(^{35-41}\). It has also long been recognized that many crop plants are recent polyploids; and, although the matter has generated decades of debate, it does seem that polyploids are favored for domestication\(^{12}\).
By contrast, homeologous recombination patterns in allotetraploid hybrids were completely distinct (Fig. 1; dataset 4a,b in ref. 25). This emphasizes the close relationship of A. hypogaea and A. monticola, and favors a single polyploid origin of both species. However, when observed on a finer scale in other genome regions, it becomes apparent that homeologous recombination in A. hypogaea has generated new diversity (Fig. 2). Some tetrasomic regions differ in different accessions of A. hypogaea; in certain genome regions some peanut accessions have an AAAA genome structure, whereas others have BBBB (Fig. 2 and Supplementary Fig. 25). Our observation for flower color, although a simple trait, provides a proof-of-principle link between homeologous recombination and generation of phenotypic diversity (Fig. 3; dataset 6 in ref. 25).

In summary, we determined the genome sequence of one reference peanut cultivar, and surveyed the genome structures of a diverse sample of landraces and cultivars. The genome structure of peanut is segmental allotetraploid (as defined by Stebbins26). We suggest that genetic deletions and exchange between the subgenomes generated variation that helped to favor the domestication of A. hypogaea over its diploid relatives. These results highlight a possible wider importance of these genetic mechanisms in accounting for the higher than expected frequency of polyploids in domesticated plants.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0405-z.

Received: 30 July 2018; Accepted: 28 March 2019;
Published online: 1 May 2019

References
1. Hilt, K. W. Polyploidy and the evolution of domesticated plants. Am. J. Bot. 80, 1494–1499 (1993).
2. Salman-Minkov, A., Sabath, N. & Mayrose, I. Whole-genome duplication as a mechanism for DNA sequence evolution. Mol. Phylogenet. Evol. 43, 469 (2008).
3. Husted, L. Cytological studies of the peanut Arachis. II. Chromosome number, morphology and behavior, and their application to the problem of the origin of the cultivated forms. Cytologia 7, 396–423 (1936).
4. Bernstein, A. & Krapovickas, A. Chromosomies y evolución en Arachis (Leguminosae). Bol. Cídasia 8, 187–220 (1994).
5. Krapovickas, A. & Gregory, W. C. Taxonomy of the genus Arachis (Leguminosae). Bol. Cídasia 16, 1–205 (1997).
6. Bertoldi, D. J. et al. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. Nat. Genet. 48, 438–446 (2016).
7. Kochert, G. et al. AFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, Arachis hypogaea (Leguminosae). Am. J. Bot. 83, 1282–1291 (1996).
8. 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).
9. 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).
10. Bertoldi, D. J. et al. An overview of peanut and its wild relatives. Plant Genet. Resour. 9, 134–149 (2011).
11. Krapovickas, A., Vanni, R. O., Pietrarelli, J. R., Williams, D. E. & Simpson, C. E. Las Razas de Maní de Bolivia. Bol. Cídasia 18, 95–189 (2009).
12. 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).
13. Bertoldi, 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).
14. Nielsen, S. et al. Mátisa, 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).
45. Leal-Bertioli, S. C. et al. Phenotypic effects of allotetraploidization of wild 
Arachis and their implications for peanut domestication. Am. J. Bot. 104, 
379–388 (2017).

46. Stebbins, G. L. Types of polyploids: their classification and significance. Adv. 
Genet. 1, 403–429 (1947).

Acknowledgements

We thank G. Birdsong, V. Nwosu, J. Elder, D. Smyth, H. Valentine, F. Luu, D. Hosington, 
H. Shapiro, D. Ward, S. Knapp, R. Wilson and S. Brown for their support of, and work 
for, the Peanut Initiative. Major financial contributors for this work were from 
Mars-Wrigley Confectionary, US peanut sheller associations, the National Peanut Board 
and other industry groups. A full list can be downloaded at https://peanutbase.org/IGPI. 
D.J.B. thanks the Georgia Peanut Commission and the Georgia Research Alliance for 
support. The genome sequencing was funded from grant 04-852-14 from The Peanut 
Foundation to J.S. and B.E.S., characterization of diverse genotypes was funded from 
grant 04-805-17 to D.J.B. The work conducted by the US Department of Energy Joint 
Genome Institute is supported by the Office of Science of the US Department of Energy 
under contract number DE-AC02-05CH11231. The work done at the DNA Technologies 
and Expression Analysis Cores at the UC Davis Genome Center was supported by NIH 
Shared Instrumentation Grant 1S10OD010786-01. We thank the US National Science 
Foundation for support from grant number 1339194 to S.A.J. This research was funded 
in part by the US Department of Agriculture Agricultural Research Service, projects 
5030-21000-069-00-D, 6048-21000-028-00-D, 6048-21000-029-00-D, 6066-21310-005-00-D 
and NIFA Award no. 2018-67013-28139. We also grateful for funding granted 
to X.Z. and Z.Z. from the Henan Province Open Cooperation Project of Science and 
Technology (172106000007), the Henan Science and Technology Major Project of the 
Ministry of Science and Technology of China (16110111000), the China Agriculture 
Research System (CARS-13), and Henan Agriculture Research System (S2012-5). We 
thank the Indian Council of Agricultural Research, National Agricultural Science Funds, 
Government of India and the CGIAR. S.S. was 
supported by the Ramalingwaswami Re-entry Grant (BT/RLF/Re-entry/41/2013) from 
the Ministry of Science and Technology. We thank D. Kudrna at the University of 
Arizona for high-molecular-weight DNA extractions, S. Simpson of USDA ARS GBRU 
for valuable support with PacBio sequencing and the USDA National Plant Germplasm 
System for Arachis seeds.

Author contributions

Project planning and coordination: J.S., S.A.J., B.E.S., S.B.C., P.O.-A., D.J.B., C.C.H., J.G., 
E.K.S.C., R.K.V., R.M., S.C.M.L.-B., L.F. and A.D.F. Production of base genetic material 
(including mapping populations and allotetraploid hybrids): C.C.H., Y.C., P.O.-A., 
S.C.M.L.-B., D.J.B. and M.C.M. Tifrunner genome sequencing (including BACs and 
quality control): J.I., A.S., J.S., J.G. and B.E.S. Hi-C libraries and sequencing: O.D., C.G.L., 
M.K.P. and E.L.A. Genome assembly: J.J., J.S., O.D. and D.J.B. Transcriptome assembly; 
gene, mobile element and repeat annotation: D.G., J. Campbell, C. Cameron, S.D., A.D.F., 
N.T.W., P.O.-A., D.J.B. and S.A.J. Comparison of gene expression in subgenomes: A.S. 
and J.S. Extra chromosomal circular DNAs: M.M. and S.L. Small RNAs and methylation: 
K.D.K., J.H.S., M.E.R., S.A.J., S.C.M.L.-B. and D.J.B. Structural analysis of Tifrunner 
genome: L.R., S.B.C., E.K.S.C., D.G., D.J.B., J. Clevenger and B.A. Data preparation and 
data basing, visualizations and expression analysis: S.B.C., E.K.S.C. and W.H. Genotyping 
and linkage mapping: C.B.-T., C. Chavarro, Y.C., J. Clevenger, S.C.M.L.-B., G.A., B.G., 
P.O.-A., S.A.J. and D.J.B. Analysis of homoeologous recombination: B.A., J. Clevenger, 
S.C.M.L.-B. and D.J.B. Diverse tetraploid samples and data: S.C.M.L.-B., R.K., Z.Z., Z.S., 
A.C., M.K.P., R.K.V., K.S., P.O.-A., S.A.J. D.J.B. and C. Chavarro. Analysis of variations in 
tetraploid genome structure: B.A., J. Clevenger, W.K. and D.J.B. Curating and sequencing 
of A. duranensis accessions: S.S.S., G.S., S.C.M.L.-B. and D.J.B. Exome capture and 
analysis: L.F., R.M., S.S.S., S.S.S., G.S., J. Clevenger, S.C.M.L.-B. and D.J.B. Biogeography: 
G.S., S.S.S. and D.J.B. Manuscript: D.J.B., J.S., B.E.S., S.B.C, M.M., J.J., P.O.-A., J. Campbell, 
J. Clevenger, S.C.M.L.-B., R.M., D.G., M.E., S.S.S., L.R. and S.A.J.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/ 
s41588-019-0405-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.J.B., S.A.J. or J.S.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in 
published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Open Access This article is licensed under a Creative Commons 
Attribution 4.0 International License, which permits use, sharing, adap-
tation, distribution and reproduction in any medium or format, as long 
as you give appropriate credit to the original author(s) and the source, provide a link to 
the Creative Commons license, and indicate if changes were made. The images or other 
third party material in this article are included in the article’s Creative Commons license, 
unless indicated otherwise in a credit line to the material. If material is not included in 
the article’s Creative Commons license and your intended use is not permitted by statu-
tory regulation or exceeds the permitted use, you will need to obtain permission directly 
from the copyright holder. To view a copy of this license, visit http://creativecommons. 
org/licenses/by/4.0/.
Methods
Plant material for genome sequencing. To generate the reference genome, we used A. hypogaea cv. Tifrunner—a runner-type peanut adapted for the southeast of the United States (CV-93, PI 640111). Phenotypically, Tifrunner would be classified as A. hypogaea ssp. hypogaea, although, like almost all runner peanuts that are currently grown in the southeast United States, it has cultivars termed ‘Spanish’ in its pedigree (Supplementary Fig. 31). Plants grown in an isolation plot in 2005 were genotyped with 146 simple sequence repeat DNA markers positioned on 20 chromosomes. A line with no detected heterozygosity was used as the founder of the peanut genome project stock.

Sequencing of the reference tetraploid genome. We sequenced Arachis hypogaea cv. Tifrunner using a whole-genome shotgun sequencing strategy and standard sequencing protocols. Illumina and PacBio reads were produced at USDA ARS GBRI and the HudsonAlpha Institute. Illumina reads were produced using the Illumina HiSeq platform and the PacBio reads were generated on the RSII platform. Two 800-bp insert 2x250 Illumina fragment libraries were obtained for a total of 63.09x coverage. Before use, all Illumina reads were screened for mitochondria, chloroplast and PhiX contamination. Reads composed of >95% simple sequences were removed. Illumina reads that were <75 bp after trimming for adapter and quality (Q < 20) were removed. An additional deduplication step was performed on the Illumina mate pairs that identifies and retains only one copy of each PCR duplicate. These two Illumina libraries were used in the final polishing of homozygous SNPs and insertions and deletions (indels) in the consensus sequence. For the PacBio sequencing, high-molecular weight DNA was isolated at the Agricultural Genetic Genomics Laboratory (https://www.genetics.arizona.edu), a total of 301 chips (P6C4 chemistry) were sequenced with a total yield of 207.2 Gb (76.7x41) and after error correction a total of 130.27 Gb (48.25x) was used in the assembly (Supplementary Tables 12 and 13).

Genome assembly and construction of pseudomolecule chromosomes. The 17,747,748 PacBio reads (76.7x sequence coverage) were assembled using MECAT. This produced 7,692 contigs with an N50 of 696.6k, 4,778 larger than 100 kb and a total genome size of 2,502.6 Mb (Supplementary Table 14). The resulting assembly was polished using Quiver. Three genetic maps (see below; dataset 8 in ref. 1) were used to identify potential misjoined regions in the MECAT assembly. Synteny with A. duranensis and A. ipaensis diploidal references was then used to pinpoint breakpoints. A total of 856 potential misjoined regions were identified and broken.

Hi-C scaffolding. The broken assembly was then scaffolded with Hi-C data using the 3D-DNA pipeline. We prepared two in situ Hi-C libraries as previously described and sequenced them (library 1: 62,762,161 of PE85 and 114,895,839 of PE150 reads; library 2: 228,896,977 of PE150; Supplementary Table 12). The Hi-C reads were aligned to the broken assembly using the Juicer pipeline. The 3D-DNA pipeline was run with the following parameters: --editor-saturation-centile 10 —editor-coarse-resolution 100000 —editor-coarse-region 400000 —editor-repeat-coverage 50. The results were polished using the Juicertools. Assembly Tools—an assembly-specific module in the Juicebox visualization system—was performed on the Illumina mate pairs that identifies and retains only one copy of each PCR duplicate. These two Illumina libraries were used in the final polishing of homozygous SNPs and insertions and deletions (indels) in the consensus sequence. For the PacBio sequencing, high-molecular weight DNA was isolated at the Agricultural Genetic Genomics Laboratory (https://www.genetics.arizona.edu), a total of 301 chips (P6C4 chemistry) were sequenced with a total yield of 207.2 Gb (76.7x41) and after error correction a total of 130.27 Gb (48.25x) was used in the assembly (Supplementary Tables 12 and 13).

Mapping populations, genotyping and linkage maps. The A. hypogaea cv. Tifrunner × A. hypogaea GT-C20 population was composed of 91 F1 individuals derived by single-seed descent and was used for mapping. Whole-genome sequencing and marker calling as previously described. Jmap 5.0 was used for genetic map construction after selecting markers without segregation distortion (p value; P > 0.05; 1:1 ratio of alleles), using the Kosambi mapping function and a minimum logarithm of odds score for linkage of 10. The A. hypogaea cv. Runner IAC 886 × (A. ipaensis K30076 × A. duranensis V141617) population consists of 89 F1 individuals that were derived by single-seed descent. The linkage map was constructed using SNPs that were done using the Affymetrix genotyping array. Maps were constructed using the Kosambi function in Mapdisto version 2.0; 20% of missing data was allowed, with a minimum logarithm of odds score of 20 and a maximum recombination frequency of 0.30.

Identification of repetitive DNA. Mobile elements were identified using a number of homology and de novo structural pattern-finding algorithms and manual curation; see Supplementary Note 2.

Structural comparisons of chromosomes. Structural comparisons between chromosomes were generated and visualized using the MU/Miner suite of alignment tools.

Assembly of transcripts and gene annotation. A transcriptome assembly to support annotation was generated from more than 6.4 billion cleaned sequence reads from A. hypogaea ssp. hypogaea genotypes (Supplementary Table 16). Libraries were constructed and 100- or 125-bp paired-end sequences generated following recommendations of the manufacturer (Illumina). Assembly was carried out with Trinity using the tetraploid genome as a guide. Read redundancy was first reduced with Trinity in silico normalization, with --max-coverage 100, giving 97 million normalized reads. The normalized reads were aligned to the ‘TifRunner genome assembly using gmap and then assembled using Trinitymaseq version 2.5.0, with maximum intron size of 10,000, and k-mer minimum coverage of 3. After filtering transcript assemblies using Kallisto transcripts per million of 1.5; 90,519 assembled transcripts were retained, with an average size of 1,340 nucleotides.

The A. hypogaea cv. Tifrunner genome was annotated using the MAKER pipeline version 2.31.9 (specifically, the dockerized image maker-2.31.9-r274 run under singularity 2.4). The genome sequence was hard-masked for ‘complex repeats’ (for example, transposable elements) using RepeatMasker and a library of repeat sequences identified in A. duranensis and A. ipaensis. The A. hypogaea (this manuscript). Simple repeats were soft-masked by MAKER, allowing them to be accessible for gene annotation in some cases. Ab initio gene prediction methods used within MAKER included SNAP version 2006-07-28 and AUGUSTUS version 3.2.3. Arachis-specific model parameters for the ab initio predictors were obtained initially from gene model calls made against chromosomes Arah3:03 and Arah3:11 (manually assigning contributions from the two diploid progenitor species) by using only the highest-confidence gene models produced in a first iteration of the pipeline (annotated edit distance ≤ 0.25); this subset was used to train the predictors for the model parameters used in subsequent iterations of the full annotation process (four iterations in total). Protein sequences used as queries for homology-based predictions consisted of the Uniprot Fabaceae protein set (retrieved December 2017). Nucleotide sequences used as queries for homology-based predictions consisted of the two transcriptome assemblies generated from A. hypogaea Tifrunner: the genome-guided transcriptome assembly described above, and the 22-tissue transcriptome assembly that has been described previously. Functional virtual assignments for the gene models were produced using InterProScan and BLASTP against annotated proteins from Arabidopsis thaliana, Glycine max and Medicago 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.

Comparison of gene expression in subgenomes. Paired-end sequencing data from expressed RNA was quality trimmed (Q ≥ 25) and reads shorter than 50 bp of the pools was performed using HGAP3 (version 2.3.0) followed by consensus sequence calling with Quiver (version 2.1). Vectors were identified and trimmed, clonal rectification and repolished with Quiver to obtain the final BAC contigs. A range of variants were detected in the comparison of the BAC clone contigs and the genome assembly. Two of the BAC contigs were excluded, because they aligned to highly repetitive pericentromeric regions, and 46 of the contigs were excluded based on length (≤20kb), leaving 175 contig alignments for analysis. Of these, a total of 79 alignments were of high quality (<0.1% bp error; Supplementary Fig. 3C); dot plots were generated using Gepard. The next 86 BAC contigs indicate a higher error rate, which was mainly due to their placement in more repetitive regions (Supplementary Fig. 34). The final ten BAC contigs indicate putative overlaps on adjacent contigs within a chromosome (Supplementary Fig. 35). The overall bp error rate (including marked gap bases) in the BAC clone contigs is 1 error per 33,510 bp (431 disrepectant bp out of 14,442,956).

Comparison of gene expression in subgenomes. Paired-end sequencing data from expressed RNA was quality trimmed (Q ≥ 25) and reads shorter than 50 bp
after trimming were discarded. Sequences were then aligned to the \textit{A. hypogaea} cv. Tifrunner genome and counts of reads uniquely mapping to annotated genes were obtained using STAR \textsuperscript{10} version 2.5.3a. Outliers among the individual experimental samples were verified based on the Pearson correlation coefficient, \( r \geq 0.85 \). Fragments per kilobase of exon per million fragments mapped values were calculated for each gene by normalizing the read count data to both the length of the gene and the total number of mapped reads in the sample and considered as the metric for estimating gene expression levels. Normalized count data was obtained using the relative log expression (RLE) method in DESeq2 \textsuperscript{20} (version 1.14.1). Genes with low expression were filtered out, by requiring 2 \( \geq \) RLE-normalized counts in at least two samples for each gene.

High-confidence homeologous gene pairs were initially identified by their reciprocal highest scores in similarity searches (BLAT) of all annotated genes in each Tifrunner subgenome versus the other. We also applied the criteria of a minimum of 100 bp identity and 85% of fragment length and only considered gene pairs that reside on homeologous chromosomes and established reciprocal translocations (dataset 1 in ref. \textsuperscript{11}). We performed differential expression analysis between the genes in homeologous pairs for each tissue and pod developmental stage using DESeq2 \textsuperscript{20} (version 1.14.1) with log2-transformed expression ratio \( 2 \pm 1 \) and Benjamin–Hochberg-adjusted \( P \leq 0.05 \) as the statistical cut-off for asymmetrically expressed genes. We used Gene Ontology for functional analysis of asymmetrically expressed homeologous gene pairs. To determine overrepresented Gene Ontology categories across biological processes, cellular component and molecular function domains, topGO \textsuperscript{70}, \textsuperscript{71}, an R Bioconductor package was used. Enrichment of Gene Ontology terms could be calculated using Fisher's exact test with \( P \leq 0.05 \) considered as significant. Statistical analyses and visualizations were performed using the R version 3.4.1 statistical software (R Development Core Team 2011).

\textbf{DNA methylation.} Genomic DNA was isolated from whole young unexpanded leaves using the DNeasy Plant Mini Kit (Qiagen). Metachromic of the A and B subgenomes of \textit{A. hypogaea} and \textit{A. duranensis} were constructed as previously described \textsuperscript{12}. In brief, approximately 1 \( \mu \)g of genomic DNA spiked with about 10 ng of unmethylated lambda DNA was sonicated to around 200 bp using a Covaris S-2. Size selection was performed using magnetic purification beads. The End-IT DNA End Repair Kit (Epiction) was used to perform end repair on the fragmented DNA. A-tails were added to blunt-end fragments using Klenow 3'→5' exo- nuclease and dA-Tailing Buffer (New England Biolabs). Methylated NEXTllex DNA adapters (Bio Scientific) were then ligated onto the DNA using T4 DNA ligase (New England Biolabs). Bisulfite conversion was done using the MethylCode Bisulfite Conversion Kit (Invitrogen). Finally, eight rounds of PCR using Kapa HiFi Uracil and Hotstart DNA polymerase (Kapa Biosystems) was used to amplify the libraries. Between each reaction, magnetic purification beads were used to clean up the DNA. Libraries were sequenced on an Illumina HiSeq 2500.

Quality-trimmed reads were aligned to the \textit{A. hypogaea} cv. Tifrunner genome using Bismark \textsuperscript{9} version 0.7.0. Multiple mapped reads and clonal reads that corresponded to potential bias from PCR amplification were discarded. The first and last 5 bp of each read where masked before methylation calling to remove biases in methylation levels introduced during the end-repairing step of library preparation. Cytosine methylation levels were calculated using the binomial distribution as previously described \textsuperscript{12}. The bisulfite non-conversion rate was calculated by mapping the unmapped reads to the unmethylated lambda genome. Only cytosines covered by at least three reads in at least one of the two replicates were retained and the two replicates were then merged for further analysis.

\textbf{Small RNAs.} Low-molecular-weight RNAs were separated from total cellular RNAs extracted using Direct-zol RNA MiniPrep (Zymo Research) as previously described \textsuperscript{13}. Libraries were prepared with TruSeq Small RNA Library Preparation Kit (Illumina) and sequenced using NextSeq (Illumina).

Small RNA reads from three replicates were trimmed for adapters and quality using cutadapt \textsuperscript{14} and merged into a single non-redundant small RNA library. Small RNA reads of 21-, 22- and 24-nucleotide length were then mapped to the \textit{A. hypogaea} cv. Tifrunner genome using Bowtie2. For each read, all alignments where reported using the -a option in Bowtie2 \textsuperscript{15}. Only perfectly matched reads were kept for further analysis. Unique small RNA reads were defined as reads that perfectly aligned to a single location in the reference genome.

\textbf{Diverse genotypes for analysis of genome structures.} A diverse panel of more than 200 tetraploid genotypes that represent the wild \textit{A. monticola}, all six botanical varieties of \textit{A. hypogaea}, and a number of wild species and introgressed allotetraploid hybrids of peanut's ancestral species \textit{A. duranensis} (V14167) and \textit{A. ipaensis} (K30076) were sequenced using Illumina short (100–250 bp) paired-end sequencing (dataset 2 in ref. \textsuperscript{11}).

\textbf{Investigating genetic exchange between subgenomes.} Genetic exchange between the subgenomes was inferred by different methods: observing mapping densities of Illumina whole-genome sequences onto the combined sequenced diploid ancestral species genomes; and by analysis of SNPs that consistently differentiate representatives of A and B genome diploid species.
read coverage. This strategy controls for differences in covered sites among the accessions.

Statistical analysis. For a description of the statistical analyses, see Supplementary Note 3.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The datasets generated during and/or analyzed for this study are available in the public repository of the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) and/or the open access internet sites PeanutBase (https://peanutbase.org). Genome assemblies and annotations, identified transposable elements, transcript assemblies, base methylation states and map data are available at PeanutBase (https://peanutbase.org/peanut_genome). A. hypogaea cv. Tifrunner sequence reads are archived in the NCBI under BioProject accession number PRJNA19393, the genome assembly has GenBank accession numbers CM009801–CM009820. Small RNA sequences are deposited in the NCBI Sequence Read Archive (SRA) under accession numbers SAMN06658954, SAMN06658955 and SAMN06658956. Datasets 1–9, as cited in manuscript, are deposited at https://doi.org/10.25739/hh3s-wx74, Cyverse (http://datacommons.cyverse.org/browse/plant/home/shared/commons_repo/curated/Bertioli_Arachis_genome_supplement_TVDM_Mar2019) and PeanutBase (https://peanutbase.org/data/public/Arachis_hypogaea/Tifrunner.esm.TVDM/). Whole-genome sequencing data of diverse accessions are deposited in the NCBI under BioProject accession numbers PRJNA525866, PRJNA511155 and PRJNA490832.

References
47. Xiao, C. L. et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods 14, 1072–1074 (2017).
48. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569 (2013).
49. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).
50. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3, 95–98 (2016).
51. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3, 99–101 (2016).
52. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at bioRxiv https://doi.org/10.1101/254797 (2018).
53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
54. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
55. Krumsvik, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23, 1026–1028 (2007).
56. Agarwal, G. et al. High-density genetic map using whole-genome resequencing for fine mapping and candidate gene discovery for disease resistance in peanut. Plant Biotechnol. J. 16, 1954–1967 (2018).
57. Pandey, M. K. et al. Development and evaluation of a high density genotyping ‘Axiom Arachis’ array with 58 K SNPs for accelerating genetics and breeding in groundnut. Sci. Rep. 7, 40577 (2017).
58. Lorieux, M. MapDisto: fast and efficient computation of genetic linkage maps. Mol. Breed. 30, 1231–1235 (2012).
59. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).
60. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881 (2010).
61. 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).
62. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).
63. 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).
64. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
65. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34, W435–W439 (2006).
66. Cleveinger, J., Chu, Y., Scheffler, B. & Ozias-Akins, P. A developmental transcriptome map for allotetraploid Arachis hypogaea. Front. Plant Sci. 7, 1446 (2016).
67. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).
68. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421–430 (2009).
69. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
70. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).
71. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 530 (2014).
72. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607 (2006).
73. Alexa, A. & Rahnenführer, J. topGO: enrichment analysis for gene ontology. R version 2.22.0 (2016); https://www.r-project.org.
74. Urch, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library preparation for base-resolution whole-genome bisulfiite sequencing. Nat. Protoc. 10, 475–483 (2015).
75. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572 (2011).
76. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–322 (2009).
77. Lu, C., Meyers, B. C. & Green, P. J. Construction of small RNA cDNA libraries for deep sequencing. Methods 43, 110–117 (2007).
78. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).
79. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 10, 1571–1572 (2011).
80. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–322 (2009).
81. Lu, C., Meyers, B. C. & Green, P. J. Construction of small RNA cDNA libraries for deep sequencing. Methods 43, 110–117 (2007).
82. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17, 10 (2011).
83. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
84. 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).

NATURE GENETICS | www.nature.com/naturegenetics
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a □ Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☑ The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of all covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

☐ Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

All commercial DNA and RNA sequencing platforms used in this study are fully described.

Data analysis

2018 versions of, Microsoft Office Excel, MECAT, Quiver, Juicebox, GATK, Joinmap 5.0, MapDisto 2.0, Axiom Analysis Suite Software, Joinmap 4.1, Blast, SINE-finder, MITE-hunter, RepeatMasker, Trinity, Kallisto, MAKER 2.31.9, SNAP, AUGUSTUS, InterProScan, Mummer, Bowtie2 v2.2, DAGchainer, Integrative Genomics Viewer, Samtools, Perl, Biopython, Unix. As cited in manuscript

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analyzed for this study are available in Supplementary Data, the public repository of the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov), the open access internet site PeanutBase (https://peanutbase.org/). Genome assemblies and annotations, identified transposable elements, transcript assemblies, base methylation states and map data are available at PeanutBase (https://peanutbase.org/peanut_genome). Arachis hypogaea cv. Tifrunner sequence reads are archived in NCBI under BioProject PRJNA419393, the genome assembly has GenBank accession numbers CM009801–CM009820. Small RNA sequences are deposited with NCBI Sequence Read Archives SAMN06658954, SAMN06658955, SAMN06658956. Whole genome sequence data of diverse accessions are deposited in the Sequence Read Archive of NCBI, Bioproject IDs: PRJNA525866, PRJNA511155 and PRJNA490832.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | Genomes from individual plants were analyzed. |
|-------------|---------------------------------------------|
| Data exclusions | No data exclusions. Sequencing data was quality filtered, as described in manuscript. |
| Replication | Genomes from individual plants were analyzed, replication not applicable. |
| Randomization | Randomization is not relevant to our study. Taxonomic classifications were used for ordering data presentations. |
| Blinding | Blinding was not applicable for this study. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ Antibodies |
| ☒ Eukaryotic cell lines |
| ☒ Palaeontology |
| ☒ Animals and other organisms |
| ☒ Human research participants |
| ☒ Clinical data |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ ChIP-seq |
| ☒ Flow cytometry |
| ☒ MRI-based neuroimaging |

Animals and other organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

| Laboratory animals | No laboratory animals were used in the study |
|--------------------|---------------------------------------------|
| Wild animals       | No wild animals were used in the study |
| Field-collected samples | Samples were from greenhouse grown plants |
| Ethics oversight   | No ethical approval was required for studying greenhouse grown plants. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.