# DNA sequencing sheds light on the evolutionary history of peanut and identifies genes associated with phenotypic diversification

**Zheng Zheng** ( zhengzheng@hnagri.org.cn )
Henan Academy of Agricultural Sciences https://orcid.org/0000-0002-2001-5336

**Ziqi Sun**
Henan Provincial Key Laboratory for Genetic Improvement of Oil Crops, Industrial Crops Research Institute, Henan Academy of Agricultural Sciences

**Feiyan Qi**
Henan Academy of Agricultural Sciences

**Yuanjin Fang**
Henan Academy of Agricultural Sciences

**Ke Lin**
Wageningen University

**Stefano Pavan**
University of Bari Aldo Moro

**Bingyan Huang**
Henan Academy of Agricultural Sciences

**Wenzhao Dong**
Henan Academy of Agricultural Sciences

**Pei Du**
Henan Academy of Agricultural Sciences

**Mengdi Tian**
Henan Academy of Agricultural Sciences

**Lei Shi**
Henan Academy of Agricultural Sciences

**Jing Xu**
Henan Academy of Agricultural Sciences

**Suoyi Han**
Henan Academy of Agricultural Sciences

**Hua Liu**
Henan Academy of Agricultural Sciences

**Li Qin**
Henan Academy of Agricultural Sciences

**Zhongxin Zhang**
Henan Academy of Agricultural Sciences
**Xiaodong Dai**
Henan Academy of Agricultural Sciences
**Lijuan Miao**
Henan Academy of Agricultural Sciences
**Ruifang Zhao**
Henan Academy of Agricultural Sciences
**Juan Wang**
Henan Academy of Agricultural Sciences
**Yuling Bai**
Wagningen University and Research, Netherlands
**Richard Visser**
Wageningen University & Research https://orcid.org/0000-0002-0213-4016
**Xinyou Zhang**
Henan Academy of Agricultural Sciences https://orcid.org/0000-0002-1942-997X

**Article**

**Keywords:**

**Posted Date:** July 5th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1776558/v1

# Running title:

DNA sequencing sheds light on the evolutionary history of peanut and identifies genes associated with phenotypic diversification

Zheng Zheng [*#1,2,3], Ziqi Sun[*1,2,3], Feiyan Qi[*1,2,3], Yuanjin Fang[*1,2,3], Ke Lin[*1,2,3], Stefano Pavan[*5], Bingyan Huang[1,2,3], Wenzhao Dong[1,2,3], Pei du[1,2,3], Mengdi Tian[1,2,3] , Lei Shi[1,2,3], Jing Xu[1,2,3], Suoyi Han[1,2,3], Hua Liu[1,2,3], Li Qin[1,2,3], Zhongxin Zhang[1,2,3], Xiaodong Dai[1,2,3], Lijuan Miao[1,2,3], Ruifang Zhao[1,2,3], Juan Wang[1,2,3], Yuling Bai[6], Richard GF Visser[6], Xinyou Zhang[#1,2,3,4]

[#] Corresponding authors

[*] These authors made equal contributions

[1] Institute of Crops Molecular Breeding, Henan Academy of Agricultural Sciences, Zhengzhou, China;

[2] Henan Provincial Key Laboratory for Genetic Improvement of Oil Crops, China;

[3] National Centre for Plant Breeding, China;

[4]The Shennong Laboratory, Zhengzhou, China

[5] Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy;

[6]Plant Breeding, Wageningen University and Research, Wageningen, The Netherlands.

[*] Correspondence:

Zheng Zheng, Tel:+86-371-65721637, Email: zhengzheng@hnagri.org.cn

Xinyou Zhang, Tel: +86-371-65729560, Email: haasz@126.com;

# Abstract

Cultivated peanut (*Arachis hypogaea* L.) is one of the most widely grown oilseed crops worldwide, however the events leading to its origin and diversification are not fully understood. Here, by combining chloroplast and whole genome sequence data from a large germplasm collection, we show that the two *A. hypogaea* subspecies (*hypogaea* and *fastigiata*) likely arose from distinct allopolyploidization and domestication events. Peanut genetic clusters were then differentiated in relation to dissemination routes and breeding efforts. A combination of linkage mapping and genome-wide association studies allowed us to characterize genes and genomic regions related to main peanut morpho-agronomic traits, namely inflorescence architecture, inner integument color, growth habit, pod/seed weight, and oil content. Together, our findings shed light on peanut evolutionary history and provide an important genomic framework resource for the genetic improvement of this crop.


# Introduction

Cultivated peanut or groundnut (*Arachis hypogaea* L.) is a sustainable and affordable source of edible oil and proteins, which globally yields 54 million tons from a cultivated area of 32 million ha (http://www.fao.org/faostat, 2020). Its allotetraploid nature (genome AABB, size ~2.7 Gb) is thought to arise from the polyploidization of an interspecific hybrid between two of the 81 wild species currently described in the genus *Arachis*, *A. duranensis* Krapov. and W.C. Gregory (genome AA, size ~1.25 Gb, female parent) and *A. ipaënsis* Krapov. and W.C. Gregory (genome BB, size ~1.56 Gb, male parent) (Seijo et al. 2007; Carvalho et al. 2020).

*A. hypogaea* is commonly assumed to be domesticated from the wild tetraploid progenitor *A. monticola*, most probably in a region now encompassing part of Southern Bolivia and Northern Argentina (Krapovickas, 1968; Seijo et al. 2007; Yin et al. 2018; Zhuang et al. 2019). The first archaeological evidence of peanut cultivation traces back to 7,600 years ago (Dillehay et al. 2007). In the 16$^{th}$ century, peanut cultivation diffused from South America to other regions of the world through the Portuguese and the Spanish explorers (Stalker & Wilson, 2016a). Nowadays, peanut is grown in more than 100 countries, with China being the first for production and India the first for cultivated area.

*A. hypogaea* is a self-pollinating species characterized by low levels of genetic variation, as the result of a series of domestication bottlenecks (Varshney et al. 2009; Mallikarjuna & Varshney, 2014); nonetheless it displays large morphological variation. The absence or presence of flowers on the main axis, and the flowering pattern, alternate or sequential, are at the basis of the classification of *A. hypogaea* in two subspecies, *A. hypogaea* subsp. *hypogaea* (*Ahh*), and *A. hypogaea* subsp. *fastigiata* (*Ahf*) (Krapovickas and Gregory 1994). Additional traits led to the distinction of two botanical varieties within *Ahh* (var. *hypogaea* and var. *hirsuta*) and four within *Ahf*

(var. *fastigiata,* var. *vulgaris*, var. *aequatoriana* and var. *peruviana*) (Krapovickas and Gregory 1994). Breeding resulted in hybridization among these taxa and thus irregular morphologies. Today, a widely used peanut classification is in accordance with five main market types (Virginia, Runner, Peruvian Runner, Valencia, and Spanish) (Stalker & Wilson, 2016b). Analysis of genetic structure resulted in clustering patterns approximately in accordance with both classifications (Zheng et al. 2018; Otyama et al. 2019).

Recently, the International Peanut Genome Initiative (IPGI) and two other research groups announced the release of cultivated peanut genome assemblies (Bertioli et al. 2019, Chen et al. 2019 and Zhuang et al. 2019), thus paving the way to in-depth exploration of peanut genetic diversity. Here, we performed chloroplast and whole genome sequencing of global peanut germplasm panels, aiming to define the peanut genetic structure and evolutionary history. Mapping approaches based on a genome-wide association study (GWAS) and recombinant inbred line (RIL) populations derived from bi-parental crosses were used to identify genomic regions and genes underlying key traits associated with peanut diversification, domestication and breeding.

# Results

### Sequencing and genotyping

Chloroplast *de-novo* sequencing was performed on 36 wild *Arachis* accessions (34 diploid *Arachis* accessions and two tetraploid species *A. monticola*) and a selection of 77 cultivated accessions that, based on the USDA taxonomic descriptors (Pittmann, 1995), could be unambiguously assigned to *A. hypogaea* subspecies and botanical varieties (Supplementary Tables 1 and 2). The length of the assembled chloroplast genomes ranged between 156,258 and 160,366 bp (Supplementary Table 3). In total,1,884 polymorphisms were found between the 113 assembled chloroplast genomes. Most of the polymorphic sites occurred between wild and cultivated peanuts, whereas only 14 polymorphisms were found within *A. hypogaea* (Supplementary Table 4). Eight additional polymorphic sites were found in a panel including, besides *A*. *hypogaea*, accessions representing six wild species of the AA genome section (Supplementary Table 4). Sanger sequencing and Kompetitive allele specific PCR (KASP) assays (Semagn et al. 2014) allowed the validation of four randomly chosen chloroplast polymorphisms detected by *de-novo* sequencing (Supplementary Fig. 1 and Supplementary Table 5).

Whole genome resequencing (WGR) was performed on two *A. monticola* and 353 *A. hypogaea* accessions originating from different countries (Fig. 1a and Supplementary Tables 1 and 2), resulting in 155.17 billion reads and 14.12 terabase pairs (Tb) of clean data. Alignment against the peanut cv. *Tifrunner* genome assembly (Bertioli et al. 2019) resulted in unique mapped reads associated with 29.00x mean depth and 88.12% genome coverage (Supplementary Table 6). In total, 864,179 SNPs and 71,052 InDels were obtained after quality control. About 40% of the variants were located on the first 10 chromosomes (corresponding to the A sub-genome), resulting

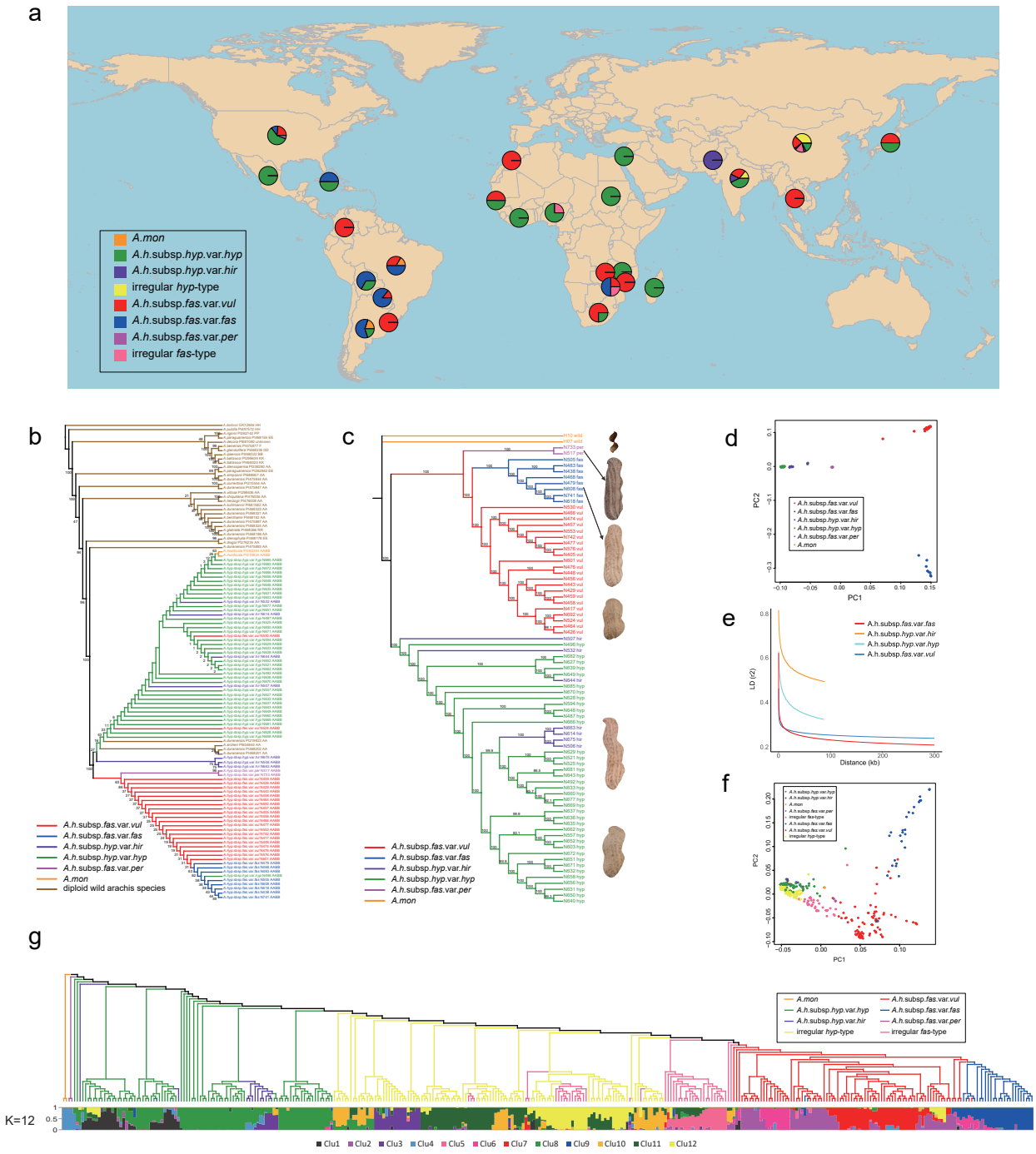


**Fig.1 Genetic structure of peanut.** a) Geographic distribution of 355 *Arachis* accessions re-sequenced in this study. The color proportion of the circle is proportional to the number of different types of accessions; b) Chloroplast phylogeny obtained by de-novo sequencing of 36 wild *Arachis* species and 77 primitive landraces assigned to *A. hypogaea* subspecies and botanical varieties; c-d) Results of phylogenesis, parametric clustering and principal components analysis (PCA) from whole genome resequencing of the same tetraploid accessions described in b; e) Extent of linkage disequilibrium (LD) decay in different *A. hypogaea* botanical varieties; f-g) PCA and parametric clustering of the 355 *Arachis* accessions re-sequenced in this study.

on average in one variant every 3.0 Kb, while 60% of the variants were located on the last 10 chromosomes (the B sub-genome), resulting on average in one variant every 2.6 Kb. The application of the KASP assay to a panel of 30 SNP loci and 10,650 data points resulted in the validation of 97.5% of the SNP calls (Supplementary Tables 7 and 8). Raw WGR data are made available at the Sequence Read Archive (SRA) database (PRJNA 605106) to serve as a public genomic resource for the scientific community (https://dataview.ncbi.nlm.nih.gov/object/PRJNA605106?reviewer=aaa4c4fbdbtthpmc20j22gqpgc).

**The peanut evolutionary history and genetic structure**

To gain insight into the origin of cultivated peanut, we first carried out a phylogenetic analysis based on chloroplast sequence data. Germplasm assigned to *Ahh* and *Ahf* grouped in well differentiated monophyletic clades (Fig. 1b), except for three accessions (N496, N524, and N530) which, according to available pedigree notes, originated from hybridization between these two taxa (Supplementary Fig. 2). Sub-clades could also be defined separating, within *Ahf*, the three botanical varieties *fastigiata*, *vulgaris*, and *peruviana* (Fig. 1b). One *A. duranensis* accession (PI 475883) was the last wild accession to diverge before the node separating *Ahh* from *Ahf* (Fig. 1b), in accordance with previous studies suggesting *A. duranensis* as the donor of the *A. hypogaea* maternal genome (Grabiele et al. 2012). Remarkably, three *A. duranensis* accessions (PI219823, PI468201 and PI468202) were found within the same phylogenetic clade of *Ahh* (Fig. 1b), together with one *A. archeri* accession (PI604844) which was previously shown to be most likely a misclassified *A. duranensis* accession by genomic *in situ* hybridization (GISH) (Du et al. 2019). Together, the identification of *A. duranensis* chloroplast genomes more closely related to *Ahh* than *Ahf*, and the clear-cut phenotypic and genetic differentiation between *Ahh* and *Ahf*, strongly indicate that the peanut subspecies *Ahh* and *Ahf* arose from different allopolyploidization events from *A. duranensis* and *A. ipaensis*, and independent domestication.

Although *A. monticola* is thought to be the wild progenitor of cultivated peanut, the two *A. monticola* accessions considered in this study also diverged after the split between the two *A. hypogaea* subspecies, as they clustered with *Ahh* (Fig. 1b). This suggests that the two accessions of *A. monticola* considered in this study might indeed represent feral forms originating from the hybridization of *Ahh*. Further studies, considering more accessions classified as *A. monticola*, might clarify the position of this species in the peanut evolutionary history.

The same tetraploid accessions used for the chloroplast phylogenesis were also subjected to genetic structure analysis with WGR data. Parametric modelling, PCA, and maximum likelihood phylogenesis provided further support for the clear-cut differentiation between the two *A. hypogaea* subspecies and, within *Ahf*, the botanical varieties *fastigiata*, *vulgaris*, and *peruviana* (Fig. 1c-d). Linkage disequilibrium (LD) decay significantly varied within *A. hypogaea*, as it was slower in var. *hirsuta* and *hypogaea* than in var. *fastigiata* and *vulgaris*. (Fig. 1e). This is consistent with the lower level of genetic diversity found in var. *hirsuta* and *hypogaea* (Supplementary Fig. 3).

In order to identify genomic regions which are highly divergent between the peanut subspecies *Ahh* and *Ahf,* thus contributing to their diversification, we performed haplotype analysis using 79 selected landraces which cover five varieties of the two sub-species. Specific haplotypes were clearly found distinguishing the botanical varieties (Supplementary Fig. 4).

The effect of recent breeding history on the peanut genetic structure was investigated using the whole panel of 355 accessions sequenced in this study, also including cultivars derived from hybridization breeding programs. Parametric modeling, PCA, and hierarchical clustering (Fig. 1f-g and Supplementary Table 9) defined additional levels of population stratification. In more detail, within var. *hypogaea*, one cluster was associated with several Chinese landraces (Cls8), and one (Cls1) with American varieties or derivatives. Within var. *vulgaris*, distinct clusters were found for African landraces (Cls6), Chinese landraces (Cls2), and cultivars from southern China (Cls7). Cls9 was found mainly for var. *fastigiata*. Finally, five clusters (Cls3, Cls5, Cls10, Cls11, and Cls12) were found for irregular type peanuts, originating from hybridization between the two *A. hypogaea* subspecies, with Cls3 and Cls5 being morphologically more similar to *Ahh* and *Ahf*, respectively.

## Genes associated with divergence between peanut subspecies

Different evolutionary histories of the peanut subspecies *Ahh* and *Ahf* were accompanied by the fixation of contrasting phenotypes for several traits, including the flowering pattern, the number of branches, the growth habit, and the color of the inner seed coat integument (tegmen). The flowering pattern, sequential in *Ahf* and alternate in *Ahh* (Fig. 2a-b) is thought to have a major role in the adaptation to different ecosystems. Mapping by two recombinant inbred line (RIL) populations identified a major locus controlling the flowering pattern at the end of chromosome 12 (Fig. 2c-d and Supplementary Table 10). Notably, this region contains a gene of the phosphatidylethanolamine binding protein (*PEBP*) family, named *AhTFL1*, which, based on phylogenetic reconstruction, was deemed as the putative orthologue of *AtTFL1*, involved in the control of inflorescence architecture in *Arabidopsis* (Severin et al. 2010, Dhanasekar et al. 2015, Krylova et al. 2020) (Fig. 2e-f and Supplementary Table 11). A genome-wide association study (GWAS) confirmed the presence of a strong signal ($-\log_{10}$(p-value) =27.31) for a marker close (14.4 Kb) to the *AhTFL1* gene at the terminal region of chromosome 12. However, it also highlighted an association at the end of chromosome 2 (Fig. 2g and Supplementary Table 12). This last result is likely due to an assembly error involving homoeologous regions of chromosomes 2 and 12, which has been announced to be fixed in the upcoming release of the peanut (cv. Tifrunner) genome assembly V.2 (https://peanutbase.org/peanut_genome_v1_v2).

*AhTFL1* sequencing in the GWAS population revealed the occurrence of three mutations (a MITE insertion, a 1492 bp deletion and a 1 bp deletion) (Supplementary Fig. 5) were predominantly present in subsp. *fastigiata* and fully co-segregating with the sequential flowering pattern (Fig. 2h-i). GWAS for the total number of branches (TNB) resulted in the strongest signal co-localizing with *AhTFL1*, indicating that it may have a pleiotropic effect on this trait (Fig. 2j and Supplementary

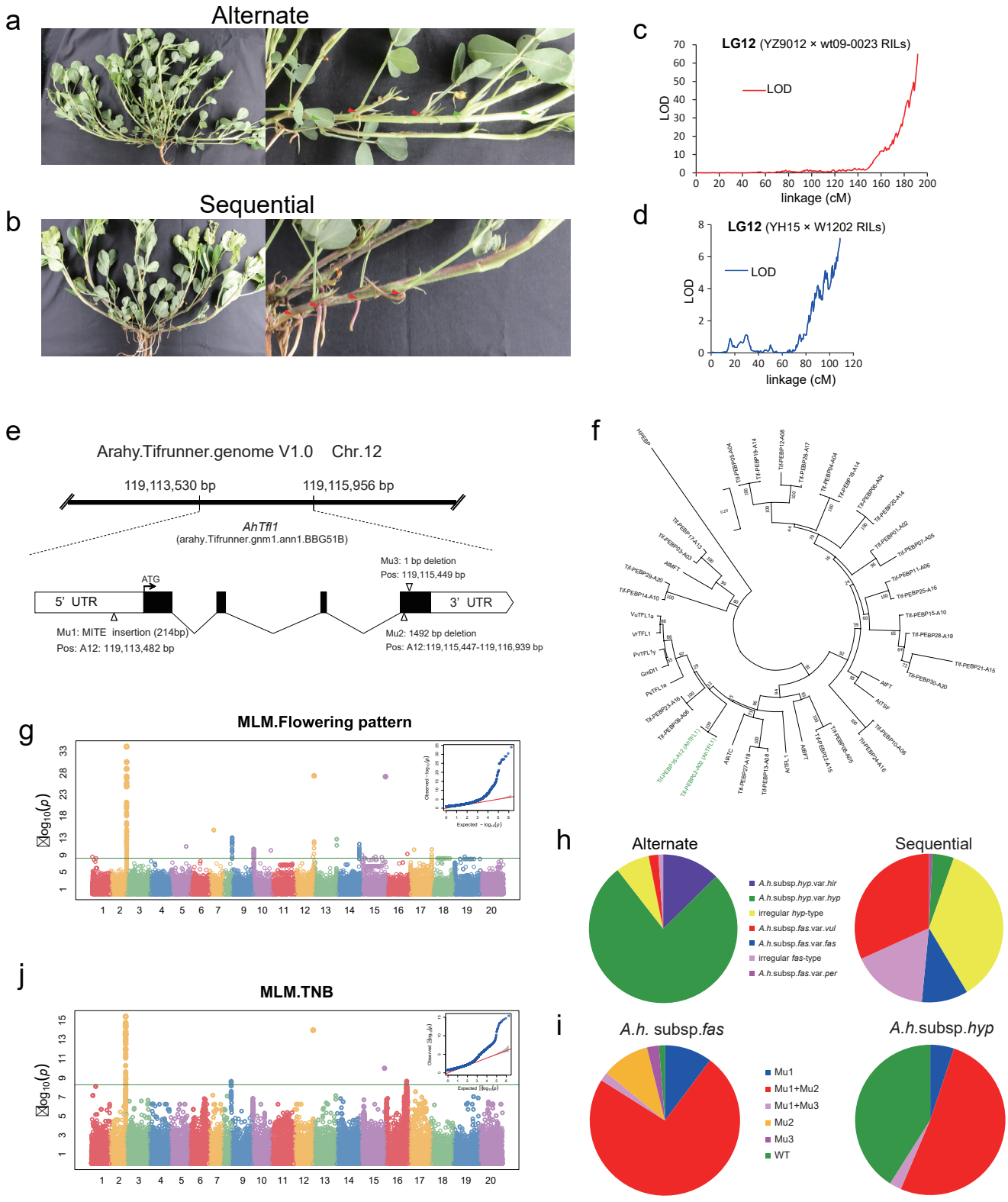


**Fig.2 Genetic control of peanut flowering pattern and total number of branches.** a) Example of a plant with alternate pattern and; b) sequential pattern; c-d) Chromosome 12 LOD score graphs obtained by composite interval mapping with the YZ9012 × wt09-0023 recombinant inbred line (RIL) population (c) and the YH15 × W1202 RIL population (d); e) Structure of the *AhTFL1* gene and features of the three mutations (Mu) found in the GWAS population; f) Phylogenetic relationships among *Arachis* and *Arabidopsis TFL* homologs. *AhTFL1* and *AtTFL1* are highlighted with a green text; g and j) GWAS Manhattan plots and quantile-quantile (Q-Q) plots for the flowering pattern (g) and total number of branches (j). The horizontal line in each Manhattan plot indicates the threshold for significant association ($p<0.05$) after a Bonferroni correction on the number of markers; h-i) The distribution of 353 accessions according to flower type (h) and mutation types (i).

Table 13).

Another trait displaying divergent phenotypes between the two peanut subspecies is the color of the seed coat inner integument (tegmen), which is invariably yellow in *Ahh* (Fig. 3a) and white in *Ahf* (Fig. 3b). Both GWAS and RIL-based mapping highlighted strong association between tegmen color and a genomic region on chromosome 5 (Fig. 3c-e and Supplementary Table 14). The strongest GWAS signal ($-\log_{10}$(p-value) =22.17) was 68.96 Kb from a gene, named *AhLAC*, encoding a laccase-like protein (Fig. 3f). Notably, this gene is the putative ortholog of the *Arabidopsis* gene *AtLAC15* (also referred to as *TRANSPARENT TESTA 10* or *AtTT10*) (Fig. 3g and Supplementary Table 15), which was shown to influence the color of the seed coat and seed dormancy through its enzymatic role in the oxidative polymerization of flavonoids (Pourcel et al. 2005). *AhLAC* sequencing in the GWAS population revealed the occurrence of two mutations (a MITE insertion and a 1 bp insertion) (Supplementary Fig. 6). Heterologous overexpression of *AhLAC* partially complemented the Arabidopsis *Attt10* loss-of-function mutant, thus providing evidence of functional conservation (Fig. 3h).

## Genetic dissection of main peanut economic traits

The peanut growth habit (erect or prostrate) (Fig. 4a-b) strongly conditions cultivation practices (Butzler et al. 1998). Genetic mapping using two RIL populations resulted in a strong signal for a genomic region on chromosome 15 (Fig. 4c-d and Supplementary Table 10), in accordance with a previous study (Kayam et al. 2017). We found that this region harbours a homolog of the MADS box family of transcription factors, previously associated with plant growth habit (Rosin et al. 2003) (Fig. 4e), which, based on phylogenetic analysis, was deemed as the putative orthologue of the gene in *Arabidopsis* (Fig. 4f and Supplementary Table 16). At least one of two mutations (a 2 bp insertion in the first exon and a 1870 bp deletion in the first intron) was found to co-segregate with the erect phenotype (Fig. 4e and Supplementary Fig. 7).

Pod and kernel dimensions, together with oil content of the kernel, are key peanut commercial traits. RIL-based mapping indicated that kernel weight, kernel length, and pod weight, are genetically correlated. The identification of QTLs on chromosomes 5 and 16 is in accordance with previous studies (Luo et al. 2018 and Gangurde et al. 2019). GWAS confirmed marker-trait associations on chromosomes 5 and 16 ($-\log_{10}$(p-value) =13.05 and 15.91, respectively), however a signal on chromosome 6 was also found (Fig. 5a-c and Supplementary Table17), possibly indicating assembly errors in correspondence of homoeologous regions of chromosomes 6 and 16. Finally, GWAS for oil content highlighted a main signal on chromosome 8 ($-\log_{10}$(p-value) = 8.94) (Supplementary Fig. 8 and Supplementary Table 18), in correspondence with a previously mapped QTL (Liu et al. 2020).

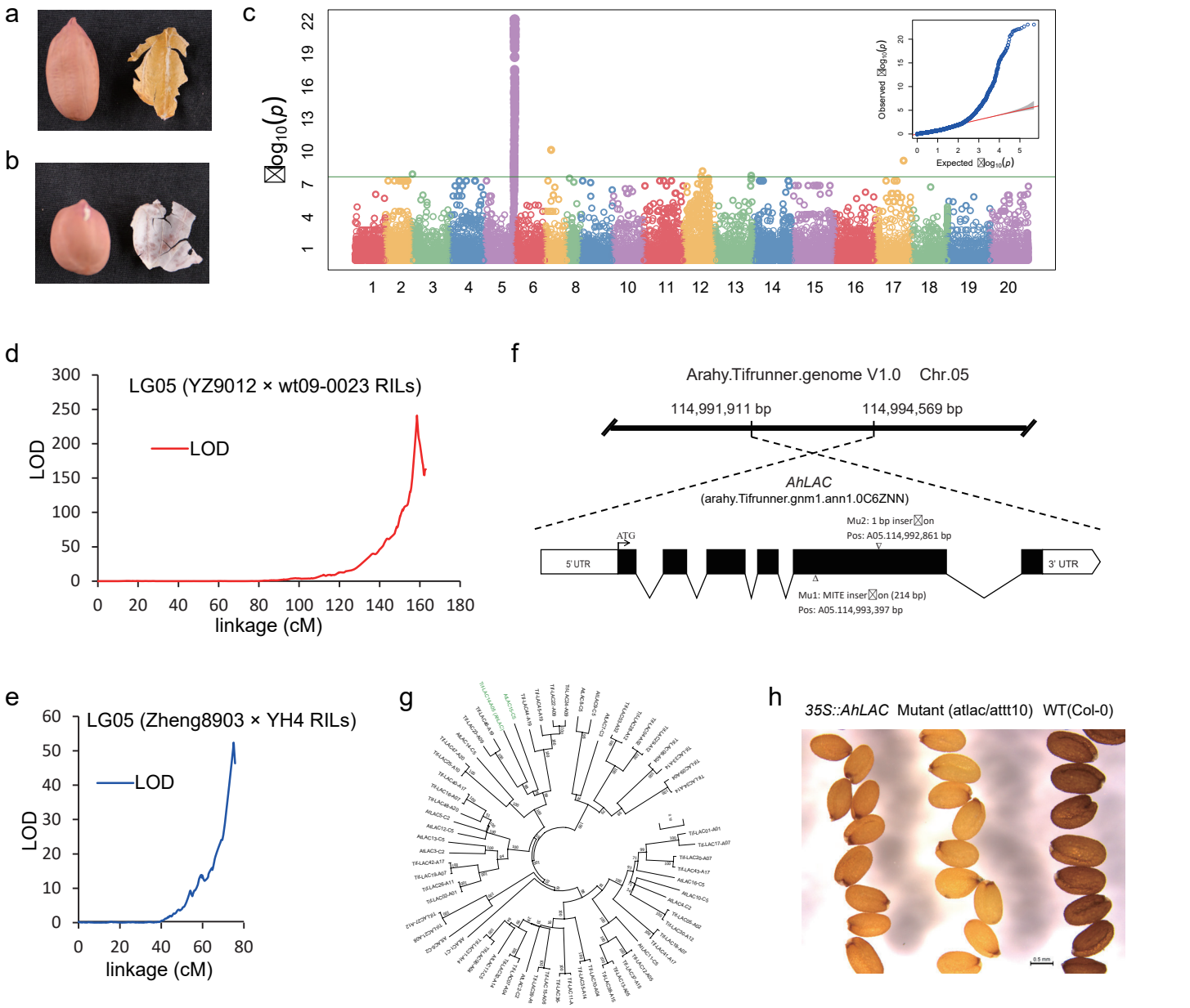


**Fig.3 Genetic control of the inner integument color.** a) Yellow; b) White; c) GWAS Manhattan plot and quantile-quantile (Q-Q) plot. The horizontal line in each Manhattan plot indicates the threshold for significant association ($p<0.05$) after a Bonferroni correction on the number of markers; d-e) Chromosome 5 LOD score graphs obtained by composite interval mapping with the YZ9012 × wt09-0023 recombinant inbred line (RIL) population (d) and the Zheng8903 × YH4 RIL population (e); f) Structure of the *AhLAC* gene and features of the two mutations (Mu) found in the GWAS population g) Phylogenetic relationships among *Arachis and Arabidopsis LAC* homologs. *AhLAC* and *Arabidopsis TRANSPARENT TESTA 10 (AtTT10)* are highlighted with a green text; h) complementation of the Arabidopsis atlac/attt10 mutant with peanut *AhLAC*. The phenotype of the wild type Col-0 accession is also shown.

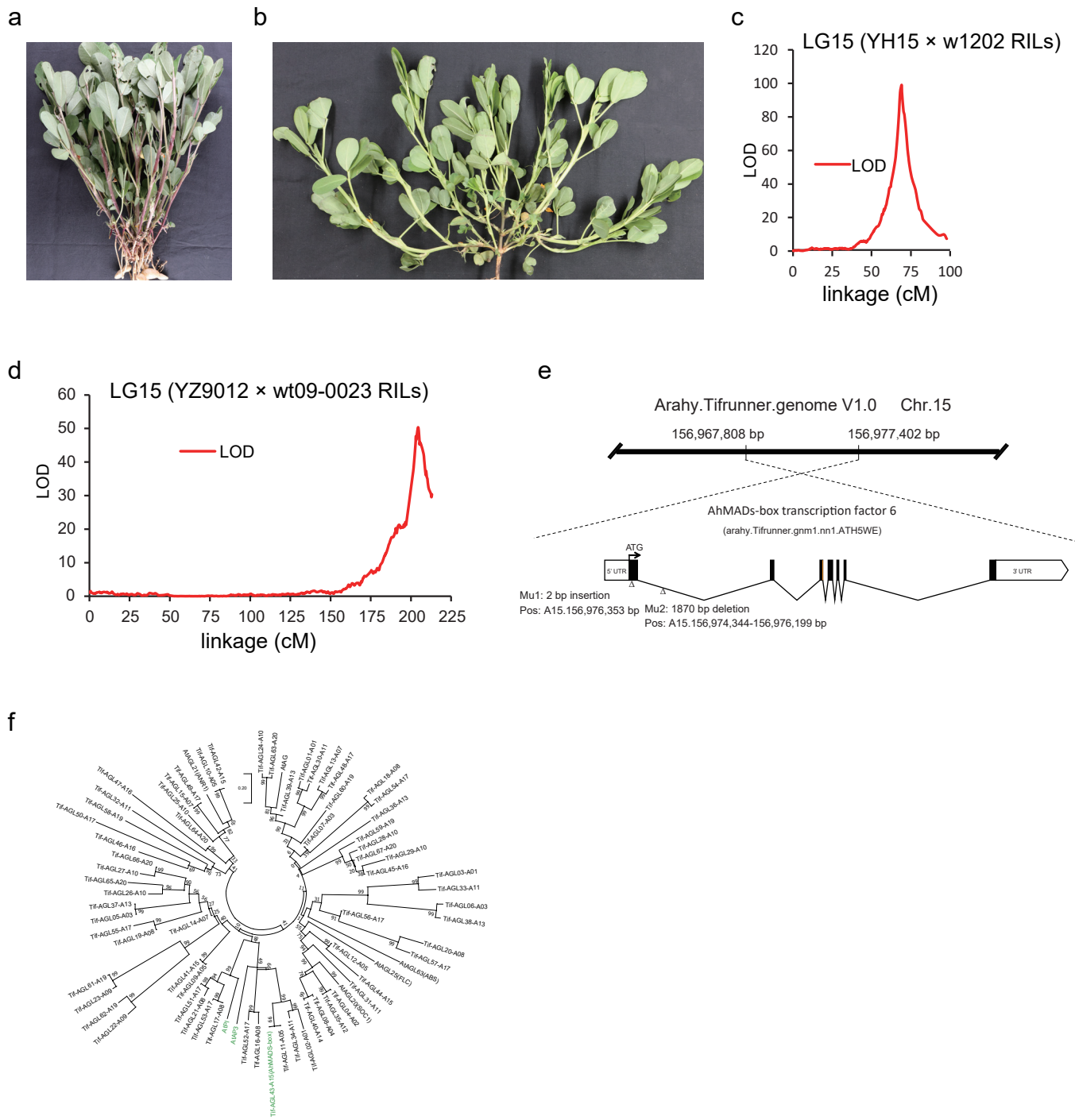


**Fig.4 Genetic control of the growth habit.** a) Erect; b) Prostrate; c-d) Chromosome 15 LOD score graphs obtained by composite interval mapping with the YH15 × w1202 recombinant inbred line (RIL) population and the YZ9012 × wt09-0023 RIL population; e) structure of the *AhMADS-box trancription factor 6* gene and features of the two mutations (Mu) found in the GWAS population; f) Phylogenetic relationships among *Arachis* and *Arabidopsis AGL* homologs. *AtPI*, *AtAP3* and MADS-box are highlighted with a green text; g-h) The distribution of 353 accessions according to growth habit (g) and mutation types (h).

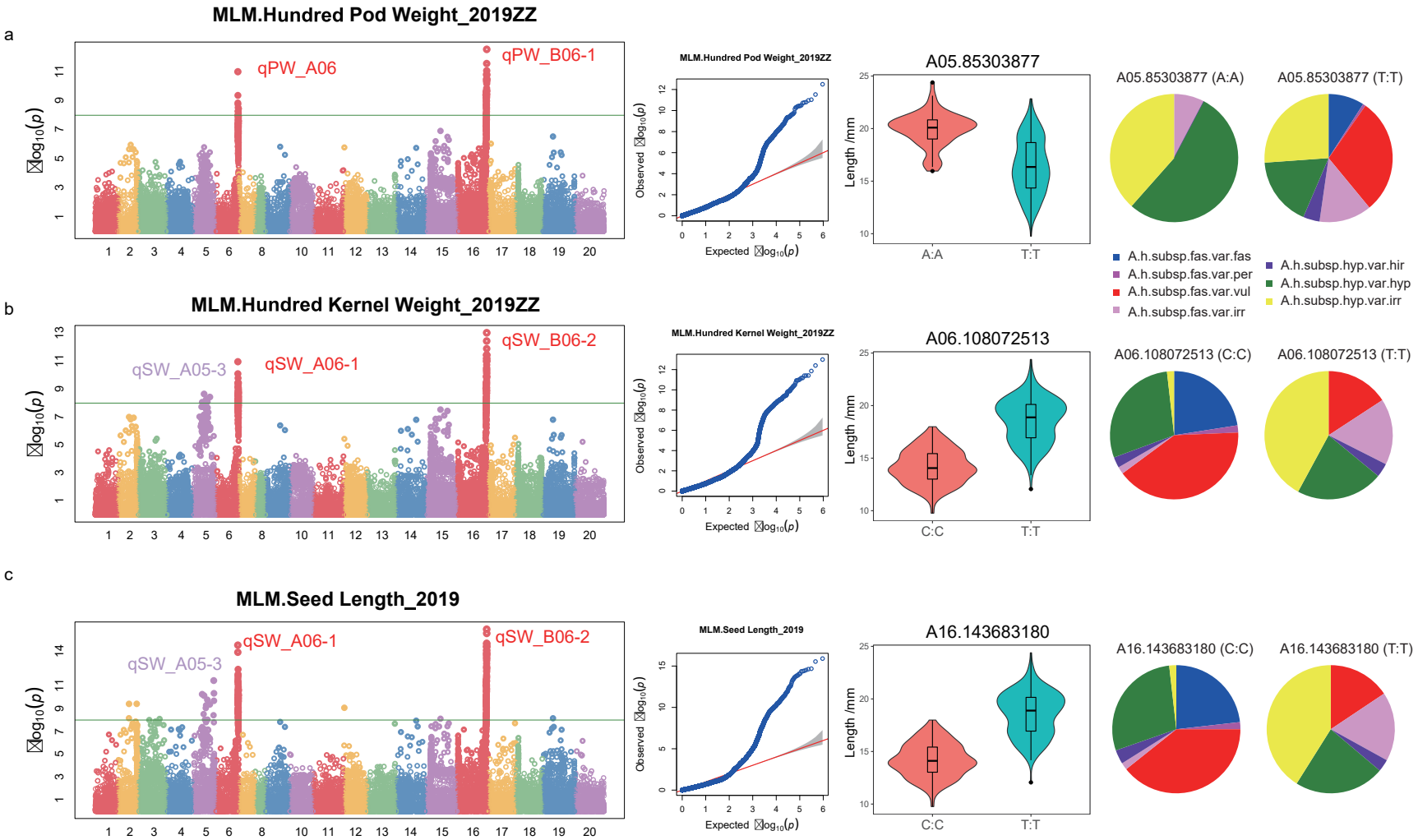


**Fig.5 GWAS for pod weight (a), kernel weight (b) and seed length (c).** For each trait, the Manhattan plot, the quantile-quantile (Q-Q plot), and the violin plot describing the phenotypic effect of the leading SNP, are reported.

# DISCUSSION

Extensive DNA sequencing allowed the fine-scale reconstruction of the peanut evolutionary history and genetic structure. Chloroplast and genomic phylogenesis provided solid indication that the two subspecies *Ahh* and *Ahf* are the result of distinct polyploidization and domestication events. A combination of biparental mapping and GWAS provided insights into the genetic basis of phenotypic divergence between *Ahh* and *Ahf,* and the control of several economically important traits.

As shown for *Arabidopsis*, mutations at genes of the *PEBP* gene family likely had a major role in determining the change of the flowering pattern from alternate to sequential. Tegmen pigmentation, due to the accumulation of oxidized polymeric forms of flavonoids, positively correlates with dormancy in many plant species, including legumes (Hradilová et al. 2019; Smykal et al. 2014). Thus, the selection of white tegument in *Ahf*, due to mutation of the laccase *AhLAC*, might have contributed to loss of seed dormancy, still occurring in *Ahh*. These traits were selected with strong pressure due to these characters making the crop acquire ecological adaptive abilities and making it suitable for different agricultural practices. For example, the spreading growth habit type is more preferred for machine harvesting, whereas the erect type is more suitable for dense planting. The white color of the inner seed coat integument showing no seed dormancy is more easily acceptable in warm ecosystems or for yearly double peanut cropping. In addition, continuous flowering often coincides with early maturing which is more suitable for areas with a shorter growing season.

Together, our findings shed light on the evolutionary history of peanut and the genetic control of some economically important traits. In addition, data reported in this study provide an important genomic resource for further and faster genetic improvement of this crop.

# METHODS

### Plant material and DNA extraction

The germplasm panel used in this study included 34 wild diploid accessions, two accessions of wild tetraploid *A. monticola*, 353 accessions of cultivated tetraploid *A. hypogaea*, and three previously described RIL populations (Supplementary Tables 1 and 2; Liu et al. 2020, Sun et al. 2021, Qi et al., 2022). Genomic DNA extraction was performed on the whole germplasm set using the Plant Genomic DNA Kit (Tiangen Biotech (Beijing) Co., Ltd, China).

### Chloroplast de-novo sequencing and variant identification

The 113 samples of chloroplast genomes were *de-novo* assembled using the pipeline of the GetOrganelle toolkit (Jin et al. 2020). All assembled accessions got repeat_pattern1 and repeat_pattern2 in circular sequences, while, some of the wild species got repeat_pattern3. The

chloroplast genomes and repeat_pattern1 that consist of two equimolar isomeric sequences and with the same direction of the small single-copy (SSC) regions were used for making alignments with the MAFFT program for pairwise comparisons. The SNP and INDEL (variants) between the chloroplast genomes were counted by using the MEGA program with the Chlorophycean Mitochondrial code set.

**Genomic re-sequencing and variant identification**

Paired-end (PE) DNA libraries with inserts of approximately 300 bp were constructed and sequenced using the Illumina HiSeq Xten (Illumina, Inc., San Diego, CA, USA) platform with PE151. Raw data were cut with an average coverage of 20x per sample for further analysis. The high-quality reads which passed the quality check and filtering were aligned to the genome of cultivated peanut *Arachis hypogaea* cv. Tifrunner version 1 using minimap2 (v2.10) (Li 2018) software with the command '-ax sr -t 25 -K 5G'. BAM alignment files were then generated with sambamba (v0.6.8) (Tarasov et al. 2015) by removing potential PCR duplications.

SNP and INDEL calling were performed with the Genome Analysis Toolkit (GATK, version v4.0.12.0) (Poplin et al. 2018) with the HaplotypeCaller method. Detected SNPs matching any of the following conditions were filtered out: QualByDepth < 2.0, FisherStrand > 60.0, RMSMappingQuality < 40.0, MappingQualityRankSumTest < -12.5 and ReadPosRankSumTest < -8.0. The conditions used to filter out INDELs are: QualByDepth < 2.0, FisherStrand > 200.0 and ReadPosRankSumTest < -20.0. After applying the aforementioned filtering conditions, we obtained variationSet1. To further exclude variant calling errors, all variations with missing rate > 0.05 (any alleles having less than five reads supporting them were marked as missing), minor allele frequency < 0.01 and the number of heterozygous alleles > 10 were filtered out using vcftools (v 0.1.19) (Danecek et al. 2011) and bcftools (v 1.10.2) (Danecek et al. 2021) which resulted in variationSet2.

**Population genetics analysis**

After clumping the remaining variants in variationSet2 using PLINK (v1.90b6.9) (Purcell et al. 2007) with "--clump-p1 1 --clump-p2 1 --clump-r2 0.5", variations (variationSet3) were retained for phylogenetic tree constructions. The maximum likelihood tree was constructed with IQ-TREE (v 1.6.12) (Minh et al. 2020) using the optimal model (GTR + F + ASC + R5) as determined by the Bayesian information criterion. Population structure was quantified using ADMIXTURE (v 1.30) (Alexander and Lange, 2011) with k between 1 and 20. The program smartpca from the Eigenstrat package (v 7.2.1) (Patterson et al. 2006) was used to calculate eigenvectors of variationSet2. The first two eigenvectors for each individual were plotted and colour coded by their sub species type. The percentage variation explained by PCA axes 1 and 2 are indicated in the axis titles. Allelic differentiation between populations was measured by nucleotide diversity (π) of each sub species

group using vcftools (v0.1.19) with a 100-kb window and a step size of 10 kb for each sub species on variationSet2. LD decay was calculated for all pairs of variations from variationSet2 within 100 kb using PopLDdecay (v3.31) (Zhang et al. 2019) with parameters '-MaxDist 500 -Het 0.05 -Miss 0.05'.

**Population genetics analysis on selected landraces**

We selected 77 landraces representing five population clusters of different varieties including *var. hypogeae*, *var. hirsuta*, *var. fastigiata*, *var. vulgaris*, *var. peruviana* and two wild tetraploid accessions. The chloroplast genomes were *de-novo* assembled for a set of 79 primitive cultivated peanut accessions as well as 34 diploid wild relatives including different genomic sections. The 113 chloroplast genomes were configured and of which the small single-copy (SSC) regions were aligned in the same direction to construct the evolutionary tree. To call haplotype blocks in 79 selected landraces we used the R package HaploBlocker (v1.5.18) (Pook et al. 2019) with adaptive mode and different sub species as subgroups on variationSet2. All 79 samples were clustered with the binary matrix output from haplotype blocks using ade4 in R (v 1.7-16) (Bougeard et al. 2018) on the first 10 chromosomes (sub genome A) and the second 10 chromosomes (sub genome B) separately.

**Genome-Wide Association Study**

GWAS was carried out on the 353 cultivated peanuts from variationSet2. Univariate GWAS method (MLM) (Yu et al. 2006) implemented in R was employed to evaluate trait-SNP associations for the target traits (Supplementary Table 19). In addition, multivariate GWAS methods (MLMM, FarmCPU and BLINK) implemented in the R package GAPIT (v 3.0), together with Generalized linear Mixed Model methods (GLMMs) in GMMAT (v 1.3.1) (Lipka et al. 2012) were employed to evaluate the MLM results. The 353 accessions were phenotyped in a randomized complete block design with two replicates in seven environments (2017: Yuanyang (2017YY); 2018: Yuanyang (2018YY), Xinyang (2018XY), Weifang (2018WF)); 2019: Zhengzhou (2019ZZ), Shangqiu (2019SQ), Weifang (2019WF)). Flowering pattern, total number of branches (TNB), the color of the inner integument, growth habit and oil content measured by gas chromatograph (GC) were investigated in one environment (2019ZZ), while hundred kernel weight (HKW), hundred pod weight (HPW) and the seed length were investigated in all seven environments. In the GAPIT analysis, we accounted for population structure from admixture analysis (Q) as the kingship matrix and the first two principal components (PCs) were also used as covariates to correct population structure due to subpopulations existing in the datasets. For each trait, the mean of seed length, the mean of pod length, the mean of seed width, the mean of pod width and flowering pattern were used as covariates as well separately in the analysis. The genome-wide significant thresholds of the GWAS were set as 0.05/n (n is the number of markers). The Manhattan plots and QQ plots for GWAS were visualized using the R package 'rMVP' (v 1.0.6) (Yin et al. 2021).

# Acknowledgements

We would like to thank the financial support from the Special Project for National Supercomputing Zhengzhou Center Innovation Ecosystem Construction (201400210600), Henan Provincial R&D Projects of Inter-regional Cooperation for Local Scientific and Technological Development Guided by Central Government (YDZX20214100004191), Major Science and Technology Projects of Henan Province (201300111000), the earmarked fund for CARS-13, Henan Provincial Agriculture Research System, China (S2012-5), the Thousand Top Talent Youth in Zhongyuan (ZYQR201912171).

# Author contributions

Z.Z designed the experiments and wrote the manuscript; Z.S. and F.Q. prepared the DNA, performed field experiments and analyzed the candidate genes; Y.F., K.L. analyzed genetic variation and GWAS; S.P. assisted in manuscript preparation; B.H. and W.D. provided help to design the experiments; P. D. provided the wild accessions; M.T., L.S., J.X., S.H., H.L., L.Q., Z.Z., X.D., L.M., R.Z., J.W. provided the help in laboratory and field experiments; Y.B. and R.G.F.V. revised the manuscript and offered suggestions; XZ conceived and facilitated the project, and constructed the RIL populations and revised the manuscript. All authors read and approved the final manuscript.

# Competing interests

The authors declare no competing interests.

# References

1. 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).
2. Carvalho, P. A. S. V. et al. Presence of resveratrol in wild *Arachis* species adds new value to this overlooked genetic resource. *Sci Rep* **10**, 12787 (2020).
3. Krapovickas, A. Origen, variabilidad y diffusion del Mani (*Arachis hypogaea*). *Actas Y Memorias Cong. Inter. Americanistas* 2517-2534 (1968).
4. Yin, D. et al. Genome of an allotetraploid wild peanut *Arachis* monticola: a de novo assembly. *Gigascience* **7**, giy066 (2018).
5. Zhuang, W. et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* **51**, 865-876 (2019).
6. 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).
7. Stalker, H. T. & Wilson, R. F. *Peanuts: Origin and Early History of the Peanut Ch.1*. (Academic Press and AOCS Press, London, 2016a).
8. Varshney, R. K. et al. The first SSR-based genetic linkage map for cultivated groundnut (*Arachis hypogaea* L.). *Theor Appl Genet* **118**, 729-39 (2009).

9. Mallikarjuna, N. & Varshney, R. K. *Genetics, Genomics and Breeding of Peanuts: Molecular markers, genetic maps and QTLs for molecular breeding in peanut Ch.5*. (CRC Press, Boca Raton, 2014).
10. Krapovickas, A. & Gregory, W. C. Taxonomy of the genus *Arachis* (Leguminosae). *Bonplandia* **8**, 1-186 (1994).
11. Stalker, H.T. & Wilson, R.F. *Peanuts: Overview of the peanut industry supply chain Ch.9*. (Academic Press and AOCS Press, London, 2016b).
12. Zheng, Z. et al. Genetic Diversity, Population structure, and botanical variety of 320 global peanut accessions revealed through Tunable Genotyping-by-Sequencing. *Scientific Reports* **8**, 14500 (2018).
13. Otyama, P. I. et al. Evaluation of linkage disequilibrium, population structure, and genetic diversity in the U.S. peanut mini core collection. *BMC Genomics* **20**, 481 (2019).
14. Bertioli, D. J. et al. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat Genet* **51**, 877-884 (2019).
15. Chen, X. et al. Sequencing of Cultivated Peanut, *Arachis hypogaea*, Yields Insights into Genome Evolution and Oil Improvement. *Mol Plant* **12**, 920-934 (2019).
16. Pittman, R.N. *United States peanut descriptors*. (US Government printing office, Washington, 1995).
17. 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).
18. Stalker, H. T. Utilizing wild species for peanut improvement. *Crop Sci* **57**, 1102-1120 (2017).
19. 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**, 1295-307 (2009).
20. Stalker, H. T. A new species in section *Arachis* of peanuts with a D genome. *Amer J Bot* **78**, 630-637 (1991).
21. Valls, J. F. M. & Simpson, C. E. Chapter 1. Taxonomy, natural distribution, and attributes of *Arachis*. *Biology and agronomy of forage Arachis*, 1-18 (1994).
22. Shandong Peanut Research Institute. *Peanut Varieties of China*. (Agriculture press, Beijing, 1987)
23. Yu SL. *Chinese Peanut Varieties and Their Pedigree*. (Shanghai Science and Technology Press, Shanghai, 2008)
24. Banks, D. J. and Kirby, J. S. Registration of Pronto peanut (Reg No. 28). *Crop Sci* **23**, 184 (1983).
25. Oil Crops Research Institute, Chinese Academy of Agricultural Sciences. *Directory of Peanut Variety Resources in China (Continued)*. (Agriculture press, Beijing, 1993)
26. Shandong Peanut Research Institute. *Directory of Peanut Variety Resources in China*. (Yantai: Penglai County Printing Factory, Yantai, 1978)
27. Bailey, W. K. & Hammons, R. O. Registration of Chico peanut germplasm (Reg. No. GP 2). *Crop Sci* **15**, 105 (1975).
28. Belamkar, V. et al. A first insight into population structure and linkage disequilibrium in the US peanut minicore collection. *Genetica* **139**, 411-429 (2011).
29. Alyr, M. H. et al. Fine-Mapping of a wild genomic region involved in pod and seed size reduction on chromosome A07 in peanut (*Arachis hypogaea* L.). *Genes (Basel)* **11**, 1402 (2020).
30. Shrestha, A., Srinivasan, R., Sundaraj, S., Culbreath, A. K., Riley, D. G. Second generation peanut genotypes resistant to thrips-transmitted tomato spotted wilt virus exhibit tolerance rather than true resistance and differentially affect thrips fitness. *J Econ Entomol* **106**, 587-96 (2013).
31. Semagn, K., Babu, R., Hearne, S., Olsen, M. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol Breed* **33**, 1-14 (2014).
32. Grabiele, M., Chalup, L., Robledo, R., 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).
33. Du, P. et al. Development of an oligonucleotide dye solution facilitates high throughput and cost-efficient chromosome identification in peanut. *Plant Methods* **15**, 69 (2019).
34. Severin, A. J. et al. RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol* **10**, 160 (2010).
35. Dhanasekar, P., Reddy, K. S. A novel mutation in TFL1 homolog affecting determinacy in cowpea (*Vigna unguiculata*). *Mol Genet Genomics* **290**, 55-65 (2015).
36. Krylova, E. A., Khlestkina, E. K., Burlyaeva, M. O., Vishnyakova, M. A. Determinate growth habit of grain legumes: role in domestication and selection, genetic control. *Ecological genetics* **18**, 43-58 (2020).
37. Pourcel, L. et al. TRANSPARENT TESTA10 encodes a laccase-like enzyme involved in oxidative

polymerization of flavonoids in *Arabidopsis* seed coat. *Plant Cell* **17**, 2966-2980 (2005).
38. Butzler, T. M., Bailey, J., Beute, M. K. Integrated management of sclerotinia blight in peanut: utilizing canopy morphology, mechanical pruning, and fungicide timing. *Plant Dis* **82**, 1312-1318 (1998).
39. Kayam, G. et al. Fine-Mapping the branching habit trait in cultivated peanut by combining bulked segregant analysis and high-throughput sequencing. *Front Plant Sci* **8**, 467 (2017).
40. Rosin, F. M., Hart, J. K., Onckelen, H. V., Hannapel, D. J. Suppression of a vegetative MADS box gene of potato activates axillary meristem development. *Plant Physiology* **131**, 1613-1622 (2003).
41. Luo, H. et al. Chromosomes A07 and A05 associated with stable and major QTLs for pod weight and size in cultivated peanut (*Arachis hypogaea* L.). *Theor Appl Genet* **131**, 267-282 (2018).
42. Gangurde, S. S. et al. Nested-association mapping (NAM)-based genetic dissection uncovers candidate genes for seed and pod weights in peanut (*Arachis hypogaea*). *Plant Biotechnol J* **18**, 1457-1471 (2020).
43. Liu N.et al. High-resolution mapping of a major and consensus quantitative trait locus for oil content to a ~ 0.8-Mb region on chromosome A08 in peanut (*Arachis hypogaea* L.). *Theor Appl Genet*. **133**, 37-49 (2020).
44. Hradilová, I. P. et al. Variation in wild pea (Pisum sativum subsp. elatius) seed dormancy and its relationship to the environment and seed coat traits. *PeerJ* **7**, e6263 (2019).
45. Smýkal, P., Vernoud, V., Blair, M. W., Soukup, A., Thompson, R. D. The role of the testa during development and in establishment of dormancy of the legume seed. *Front Plant Sci* **5**, 351 (2014).
46. Liu, H. et al. QTL mapping of web blotch resistance in peanut by high-throughput genome-wide sequencing. *BMC Plant Biol* **20**, 249 (2020).
47. Sun, Z. et al. QTL mapping of quality traits in peanut using whole-genome resequencing. *The crop journal* **10**, 177-184 (2022).
48. Qi, F. et al. QTL identification, fine mapping, and marker development for breeding peanut (*Arachis hypogaea* L.) resistant to bacterial wilt. *Theor Appl Genet* **135**, 1319-1330 (2022).
49. Jin J. J. et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* **21**, 241 (2020).
50. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
51. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034 (2015).
52. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, doi: https://doi.org/10.1101/201178 (2018).
53. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
54. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
55. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80-92 (2012).
56. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
57. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* **37**, 1530-1534 (2020).
58. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
59. Patterson, N., Price, A. L., Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006)
60. Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., Yang, T. L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786-1788 (2019).
61. Pook, T., et al. HaploBlocker: creation of subgroup-specific haplotype blocks and libraries. *Genetics* **212**, 1045-1061 (2019).
62. Bougeard, S. & Dray, S. Supervised multiblock analysis in R with the ade4 package. *J Stat Softw* **86**, 1-17 (2018).
63. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-208 (2006).
64. Lipka, A.E. et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397-2399 (2012).
65. Yin, L., et al. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide association study. *Genomics Proteomics Bioinformatics* **19**, 619-628 (2021).

## Figure legends

**Fig.1: Genetic structure of peanut.** a) Geographic distribution of 355 *Arachis* accessions re-sequenced in this study. The color proportion of the circle is proportional to the number of different types of accessions; b) Chloroplast phylogeny obtained by *de-novo* sequencing of 36 wild *Arachis* species and 77 primitive landraces assigned to *A. hypogaea* subspecies and botanical varieties; c-d) Results of phylogenesis, parametric clustering and principal components analysis (PCA) from whole genome resequencing of the same tetraploid accessions described in b; e) Extent of linkage disequilibrium (LD) decay in different *A. hypogaea* botanical varieties; f-g) PCA and parametric clustering of the 355 *Arachis* accessions re-sequenced in this study.

**Fig.2 Genetic control of peanut flowering pattern and total number of branches.** a) Example of a plant with alternate pattern and; b) sequential pattern; c-d) Chromosome 12 LOD score graphs obtained by composite interval mapping with the YZ9012 × wt09-0023 recombinant inbred line (RIL) population (c) and the YH15 × W1202 RIL population (d); e) Structure of the *AhTFL1* gene and features of the three mutations (Mu) found in the GWAS population; f) Phylogenetic relationships among *Arachis* and *Arabidopsis TFL* homologs. *AhTFL1* and *AtTFL1* are highlighted with a green text; g and j) GWAS Manhattan plots and quantile-quantile (Q-Q) plots for the flowering pattern (g) and total number of branches (j). The horizontal line in each Manhattan plot indicates the threshold for significant association ($p<0.05$) after a Bonferroni correction on the number of markers; h-i) The distribution of 353 accessions according to flower type (h) and mutation types (i).

**Fig.3 Genetic control of the inner integument color.** a) Yellow; b) White; c) GWAS Manhattan plot and quantile-quantile (Q-Q) plot. The horizontal line in each Manhattan plot indicates the threshold for significant association ($p<0.05$) after a Bonferroni correction on the number of markers; d-e) Chromosome 5 LOD score graphs obtained by composite interval mapping with the YZ9012 × wt09-0023 recombinant inbred line (RIL) population (d) and the Zheng8903 × YH4 RIL population (e); f) Structure of the *AhLAC* gene and features of the two mutations (Mu) found in the GWAS population g) Phylogenetic relationships among *Arachis* and *Arabidopsis LAC* homologs. *AhLAC* and Arabidopsis *TRANSPARENT TESTA 10* (*AtTT10*) are highlighted with a green text; h) complementation of the Arabidopsis *atlac/attt10* mutant with peanut *AhLAC*. The phenotype of the wild type Col-0 accession is also shown.

**Fig.4 Genetic control of the growth habit.** a) Erect; b) Prostrate; c-d) Chromosome 15 LOD score graphs obtained by composite interval mapping with the YH15 × w1202 recombinant inbred line (RIL) population and the YZ9012 × wt09-0023 RIL population; e) structure of the *AhMADS-box trancription factor 6* gene and features of the two mutations (Mu) found in the GWAS population; f) Phylogenetic relationships among *Arachis* and *Arabidopsis AGL* homologs. *AtPI, AtAP3* and MADS-box are highlighted with a green text; g-h) The distribution of 353 accessions according to growth habit (g) and mutation types (h).

**Fig.5 GWAS for pod weight (a), kernel weight (b) and seed length (c).** For each trait, the Manhattan plot, the quantile-quantile (Q-Q plot), and the violin plot describing the phenotypic effect of the leading SNP, are reported.

**Supplementary Figure Legends**

**Supplementary Figure S1. Validation of chloroplast DNA polymorphisms.** Four randomly chosen variations were confirmed by Sanger sequencing (a-d). Three of them were also validated by Kompetitive allele specific PCR (KASP) assays (e-g).

**Supplementary Figure S2. Pedigree information of the accession N524.** The red and blue lines indicate the female and male parent, respectively. The accession N524 inherited the chloroplast genome from N744.

**Supplementary Figure S3. Nucleotide diversity (Pi) for different botanical types of peanut.** The A and B sub-genomes were analyzed separately. Sub-A for AA genome (a) and Sub-B for BB genome (b).

**Supplementary Figure S4. Haplotype blocks for different peanut botanical types.** Graphical block structure representation of 20 chromosomes. Haplotypes were computed with adaptive mode using window sizes of 5, 10, 20 and 50 markers and target coverage of 90%.

**Supplementary Figure S5. Features of the mutations (Mu) identified for the gene *AhTFL1*.** a) Sequence alignment showing the 214 bp MITE insertion of the mutation type 1 (Mu 1); b) Integrative Genomics Viewer (IGV) image of the genomic region showing paired-end reads mapped on the candidate gene with a 1492 bp deletion described as mutation type 2 (Mu 2); c) the paired-end reads mapped on the candidate gene with a 1 bp deletion (with a C missing) described as mutation type 3 (Mu 3).

**Supplementary Figure S6. Features of the mutations (Mu) identified for the gene *AhLAC.*** a) Sequence alignment showing the 214 bp insertion of the mutation type 1 (Mu 1); b) Integrative Genomics Viewer (IGV) image of the genomic region showing paired-end reads mapped on the candidate gene with a 1 bp insertion (purple I) described as mutation type 2 (Mu 2).

**Supplementary Figure S7. Features of the mutations (Mu) identified for the gene *AhMADs-box transcription factor 6*.** a) Integrative Genomics Viewer (IGV) image of the genomic region showing paired-end reads mapped on the candidate gene with a 2 bp insertion (purple 2) described as mutation type 1 (Mu 1); b) Integrative Genomics Viewer (IGV) image of the genomic region showing paired-end reads mapped on the candidate gene with a 1870 bp deletion described as mutation type 2 (Mu 2).

**Supplementary Figure S8. GWAS for seed oil content.** Manhattan plot (left) and quantile-quantile (Q-Q) plot (right). The horizontal line in the Manhattan plot indicates the significance threshold for association ($p<0.05$) after a Bonferroni correction on the number of markers.

**Supplementary Table Legends**

Supplementary Table 1. Features of the 36 wild *Arachis accessions* subjected to de novo chloroplast sequencing

Supplementary Table 2. Features of the 353 *Arachis hypogaea* subjected to whole genome resequencing. A selection of 77 accessions also subjected to chloroplast de novo sequencing is marked with an asterisk

Supplementary Table 3. Length and GC count associated with 113 de novo assembled chloroplast *Arachis genomes*

Supplementary Table 4. Polymorphic sites identified in a germplasm panel including 77 tetraploid cultivated species and six wild species of the AA genome section

Supplementary Table 5. Genotyping of 77 tetraploid *A. hypogaea* accessions with three Kompetitive allele specific PCR (KASP) assays designed on chloroplast polymorphisms

Supplementary Table 6. Statistics from whole genome resequencing and alignment

Supplementary Table 7. Results from genotyping of genomic SNP loci with 30 Kompetitive allele specific PCR (KASP) assays

Supplementary Table 8. Accuracy of genomic SNP validation with 30 Kompetitive allele specific PCR (KASP) assays

Supplementary Table 9. The 12 clusters calculated by ADMIXTURE according to CV error

Supplementary Table 10. QTLs identified for the flowering pattern, inner integument color, and growth habit, using three recombinant inbred line (RIL) populations

Supplementary Table 11. Information on phosphatidylethanolamine-binding protein (PEBP) homologous gene from different species

Supplementary Table 12. GWAS results for flowering pattern

Supplementary Table 13. GWAS results for total number of branches

Supplementary Table 14. GWAS results for inner integument color

Supplementary Table 15 The information of laccase (LAC) homologous gene from different species

Supplmentary Table 16. Information on MADS-Box homologous gene from different species

Supplementary Table 17. GWAS results for hundred kernel weight, hundred pod weight and seed length

Supplementary Table 18. GWAS results for oil content

Supplementary Table 19.  Phenotypes for GWAS

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigs.pdf
- SupplementaryTables.xlsx