nature genetics

Article

https://doi.org/10.1038/s41588-024-01660-7

A genomic variation map provides insights into peanut diversity in China and associations with 28 agronomic traits

Received: 24 April 2022

Accepted: 9 January 2024

Published online: 20 February 2024

Check for updates

Qing Lu[1,4], Lu Huang[1,4], Hao Liu[1,4], Vanika Garg[2], Sunil S. Gangurde[3], Haifen Li[1], Annapurna Chitikineni[2], Dandan Guo[1], Manish K. Pandey[3], Shaoxiong Li[1], Haiyan Liu[1], Runfeng Wang[1], Quanqing Deng[1], Puxuan Du[1], Rajeev K. Varshney[2], Xuanqiang Liang[1], Yanbin Hong[1] & Xiaoping Chen[1]

Peanut (*Arachis hypogaea* L.) is an important allotetraploid oil and food legume crop. China is one of the world's largest peanut producers and consumers. However, genomic variations underlying the migration and divergence of peanuts in China remain unclear. Here we reported a genome-wide variation map based on the resequencing of 390 peanut accessions, suggesting that peanuts might have been introduced into southern and northern China separately, forming two cultivation centers. Selective sweep analysis highlights asymmetric selection between the two subgenomes during peanut improvement. A classical pedigree from South China offers a context for the examination of the impact of artificial selection on peanut genome. Genome-wide association studies identified 22,309 significant associations with 28 agronomic traits, including candidate genes for plant architecture and oil biosynthesis. Our findings shed light on peanut migration and diversity in China and provide valuable genomic resources for peanut improvement.

Peanut (*Arachis hypogaea* L.) is an important oil and food legume worldwide, offering nutritional elements and economic value to address malnutrition and poverty[1]. In China, peanut is an important source of vegetable oil for its residents and a major cash crop for increasing farmer income and lifting them out of poverty. It is distributed across southern and northern regions, with a total production of 18.05 million tons, accounting for 33.64% of the world's production (FAOSTAT, 2020). Peanut was possibly derived from the hybridization between its two progenitors, *Arachis duranensis* and *Arachis ipaensis*, in the southwestern Mato Grosso do Sul region of Brazil (South America) and was domesticated 6,000 years ago and then widely dispersed in post-Columbian times[2–8]. Historical records suggest that peanut was introduced into China via three possible paths as follows: Africa–India–South China (SC), South America–Indonesia–SC (through the Indian Ocean) and Mexico–Philippines–SC (through the Acapulco–Manila galleon trade route)[7]. China, being outside the world's propagation centers, is among the countries that have benefited from the use of imported germplasm. The usage of introduced peanuts, including landraces and breeding lines, has brought great diversity to improved varieties in both northern and southern regions of China. Distinct peanut market types have been formed between southern China (Spanish type, small pod) and northern China (Virginia type, large pod), creating

[1]Crops Research Institute, Guangdong Academy of Agricultural Sciences, Guangdong Provincial Key Laboratory of Crop Genetic Improvement, South China Peanut Sub-Centre of National Centre of Oilseed Crops Improvement, Guangzhou, China. [2]WA State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, Western Australia, Australia. [3]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India. [4]These authors contributed equally: Qing Lu, Lu Huang, Hao Liu. ✉e-mail: luqing@gdaas.cn; rajeev.varshney@murdoch.edu.au; liangxuanqiang@gdaas.cn; hongyanbin@gdaas.cn; chenxiaoping@gdaas.cn

two major Chinese transmission centers. Exploring the genetic diversity of germplasm for cultivar improvement is a priority in enhancing the yield capacity for peanuts. However, the genomic variations underlying the phenotypic diversity owning to natural and artificial selections have not yet been fully investigated, and only a few germplasm have been used for improving agronomic traits[9].

Diverse consumption needs have always been an important driving force for improving agronomic traits in peanuts. A few genetic loci have been discovered for several traits with traditional linkage and association mapping[10–17]. Resequencing of germplasm is an integral part of identifying the genomic variations and is essential to accelerate genomic research. A genome-wide association study (GWAS) is appropriate for identifying genomic variations associated with agronomic traits using a natural population[18]. Substantial progress has been made in the genome sequencing of diploid crops, such as rice[19,20], maize[21], soybean[22], chickpea[23,24], castor bean[25] and millet[26], as well as in polyploid crops, including cotton[27,28], rapeseed[29,30] and wheat[31]. However, GWAS-based genetic dissection with high throughput genotyping data in peanuts is limited[32,33].

The availability of the high-quality genomes of cultivated peanut[6,7,34] and its two progenitors[5,35] has facilitated sequencing-based trait mapping and marker development for important traits. Here we resequenced a total of 390 peanut accessions to evaluate the genome-wide diversity of landraces and breeding lines in southern and northern regions of China, compared to accessions from regions outside of China and conducted GWAS for 28 agronomic traits. To further explore the genomic signatures of breeder-driven selection, we resequenced 11 elite varieties, including 2 landraces and 9 breeding lines, from a classical pedigree in the breeding programs in southern China. Our results provide important insights into the transmission of peanuts after introduction into China, the genetic diversity and genomic variation underlying peanut agronomic traits, and valuable genomic resources and candidate genes applicable for peanut improvement.

Results

Genome-wide variation map for peanut

We resequenced 390 worldwide peanut accessions, including 261 (66.92%) from northern and southern regions of China, and 129 (33.08%) from regions outside of China (Fig. 1a,b and Supplementary Table 1). A total of 1.29×10^{13} bases of raw data were generated, with an average sequencing depth of 10.95× (8.70–16.92×) and genome coverage of 97.21% (96.04–97.69%; Supplementary Table 1). The sequencing data were aligned to the *A. hypogaea* cv. Fuhuasheng reference genome[34], resulting in the identification of 8,803,668 single-nucleotide polymorphisms (SNPs) and 2,137,963 insertions and deletions (InDels; Supplementary Tables 2 and 3 and Supplementary Fig. 1). Of these SNPs, 13.24% were present in the coding regions. Of the identified InDels, 8,452 and 5,414 were annotated as frameshift and nonframeshift variations, respectively (Supplementary Table 2 and Supplementary Fig. 2). We observed a higher number of SNPs and density in the B subgenome (Bt; 4.99 million SNPs and 3.70 per kb) as compared to the A subgenome (At; 3.66 million SNPs and 3.15 per kb). However, the InDel densities were similar between the two subgenomes (0.82 and 0.85 InDels per kb in Bt and At, respectively; Supplementary Table 3 and Supplementary Fig. 3).

Genomic diversity and phylogeography of peanut in China

Phylogenetic and population structure analyses showed that all accessions were clustered into three groups ($K = 3$), namely non-China (NonC)-0, North China (NC) and SC (Fig. 1c,d and Supplementary Fig. 4a). Principal component analysis (PCA) indicated that all accessions were clustered into three main groups, roughly corresponding to their phylogenetic classifications and respective geographic distributions (Fig. 1a–e and Supplementary Fig. 4b). In addition, all accessions from gene banks outside of China were split into three subclades, referred to as NonC-0, NonC-1 and NonC-2 (Supplementary Fig. 4c,d). The NonC-0 accessions were not well separated into substructures at $K = 3$ and 4 (Fig. 1c), indicating that the NonC-0 gene pool might be monophyletic. Furthermore, the NonC-1 accessions, mainly from North America (especially the United States and Mexico), were clustered closely with the NC group, primarily from northern China with a center in Shandong and Henan provinces. The NonC-2 accessions from South and Southeast Asia were integrated into the SC group, mainly collected from southern China centered on the Pearl River Delta dominated by Guangdong and Guangxi provinces (Fig. 1b,c). Moreover, the fixation index (F_{ST}) values between NonC-1 and NC (0.08) and between NonC-2 and SC (0.07) were the lowest, followed by that between NC and SC (0.13; Fig. 1f). These results suggested that contributions of the NonC-1 and NonC-2 gene pools to NC and SC genotypes, respectively, are evident. Admixture due to genetic introgression also showed clear evidence of genetic heterogeneity in NC and SC genotypes (when $K = 4$; Fig. 1c), indicating frequent gene exchanges between China's two main peanut production areas during the long-term breeding process.

To investigate the migration routes of Chinese peanuts introduced from abroad, phylogenetic relationship and gene flow were examined using Treemix[36], which implements composite likelihood to infer probable population splits. The results showed that the direction of gene flow for NC peanut accessions was mainly from North America (Supplementary Fig. 5a), whereas SC accessions were mainly from South and Southeast Asia (Supplementary Fig. 5b,c). These findings were consistent with the result of DIYABC[37] analysis (when scenario = 5; Supplementary Fig. 5d,e). This graph model inferred that peanut was introduced into China through two major propagation routes. In northern China, peanut was introduced from North America, while those in southern China migrated from South and Southeast Asia (Supplementary Fig. 5f). These migrations resulted in the emergence of two major peanut cultivation centers in China. One is the Pearl River Delta, from which peanuts spread to southwestern and southeastern China, as well as Southeast Asia. The other is the NC Plain, from which peanuts spread to northeastern China (Fig. 1b and Supplementary Fig. 5f). The NC and SC peanut cultivation centers reflected the different production systems and climate environments in the respective regions, representing the large-pod (like Virginia type) and small-pod (like Spanish type) market types, respectively. These results indicated that the two peanut cultivation centers have a crucial role in the spread of peanuts in China. Nucleotide diversity (π) analysis showed that NonC-0 had the highest π values (2.14×10^{-4}; Fig. 1f), suggesting that genotypes in the NonC-0 can be used to expand the genetic base for future peanut improvement in China. The mean linkage disequilibrium (LD) decay distance of the whole genome was 92.3 kb (decaying to r^2 of approximately 0.16); however, for the At and Bt subgenomes, it was 104.2 kb and 59.2 kb, respectively (Fig. 1g). Moreover, the LD decay distance was greatest for NC, followed by SC and NonC (Supplementary Fig. 4e).

Genome-wide association analysis for agronomic traits

The genome-wide variation map enabled GWAS for 28 agronomic traits in four peanut-growing seasons in 2017 and 2018 in Guangzhou, China (Supplementary Table 4 and Supplementary Fig. 6). Correlation coefficients among these traits were calculated using the best linear unbiased prediction (BLUP) combined phenotypes (Supplementary Fig. 7). GWAS analysis was performed for 28 traits, resulting in a total of 22,309 significant associations, including 17,803 unique significant SNPs (Supplementary Fig. 8 and Supplementary Tables 5 and 6). Moreover, 791 significant SNPs associated with six traits were detected more than once in different environments or using BLUPs (Supplementary Table 7). These repetitive associations were mainly observed in four major regions on chromosomes A03 (17.4–95.0 Mb), A09 (5.8–8.9 Mb), A10 (16.1–16.2 Mb) and B09 (158.9–159.0 Mb; Supplementary Fig. 8).

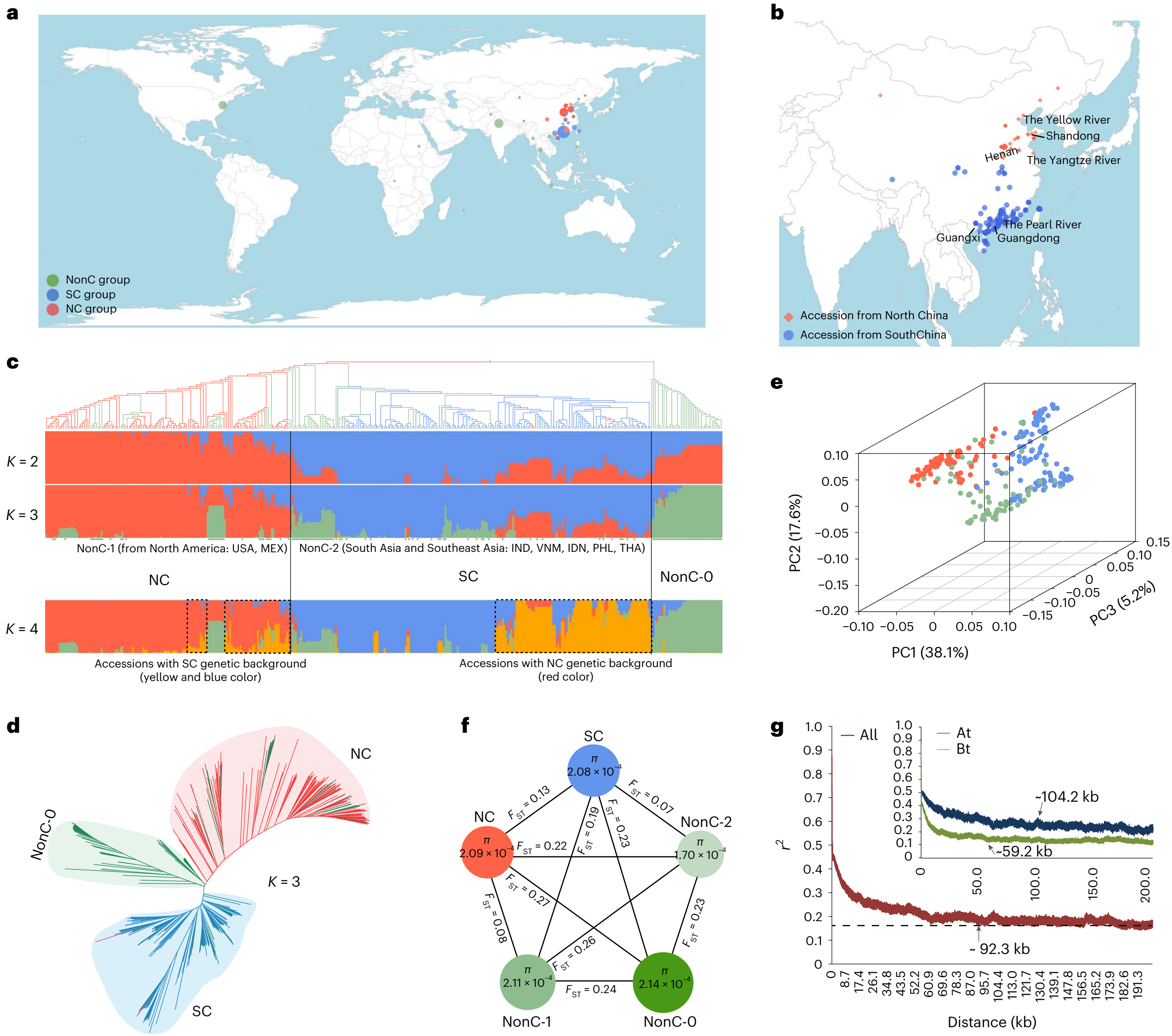

Fig. 1 | Geographic distribution, population structure and LD decay of 390 peanut accessions. a, Worldwide distribution of 390 peanut accessions. The size of the plot represents the sample size. **b**, Geographic distribution of accessions from China. **c**, Population structure and phylogenetic analysis. **d**, Phylogenetic tree of all accessions. **e**, Three-dimensional PCA plot of the first three principal components. **f**, Genetic diversity (π) and population differentiation (F_{ST}) across the three subpopulations. **g**, LD decay estimation of all accessions, subgenome A (At) and subgenome B (Bt). The maps in **a** and **b** were generated using the map data function in ggplot2 packages in R (v4.2.0).

Two loci that were strongly associated with oleic acid (OA) metabolism were found on A09 and B09, which were harbored by *FAD2A* (LOC112712140) and *FAD2B* (LOC112776164), respectively (Supplementary Fig. 8), which were known to regulate the conversion from OA to linoleic acid (LA)[38]. Another major region on A10 for seed coat color (SCC) was found near *J3K16K* (named *AhTc1*), which has been reported to control peanut purple testa color[39] (Supplementary Fig. 8). Regarding the quality-related traits, the ratio of oleic and linoleic (ROLA) was associated with 7,141 SNPs, followed by SCC (4,586), LA (2,497), OA (1,317) and palmitic acid (PA; 1,057). In terms of the yield-related traits, the highest number of SNPs was associated with seed number per pod (SNPP; 3,118), followed by seed length (SL; 341). For the plant-type-related traits, main stem height (MSH) and first branch length (FBL) had almost the same number of associated SNPs (494 and 452, respectively; Supplementary Table 6).

To explore gene linkage and pleiotropy, the newly detected SNPs and known genes associated with different traits were mapped on each chromosome (Fig. 2a). We identified 1,654 SNPs that were associated with at least two traits (Supplementary Table 8). Importantly, 74% (1,224) of these SNPs were associated with at least two of four seed quality-related traits (OA, LA, ROLA and PA). Two loci, close to *FAD2A* and *FAD2B*, were also substantially associated with PA, suggesting that they might be involved in OA, LA and PA metabolisms (Fig. 2b and Supplementary Fig. 8). In addition, we found that 24 SNPs on A05 (8,627,257–13,086,683 bp) and 58 SNPs on B06 (142,951,711–149,177,818 bp) were substantially associated with at least two traits of pod width (PW), pod thickness (PT), hundred pod weight (HPW), SL and hundred seed weight (HSW; Fig. 2c), which was in agreement with the phenotypic correlation analysis (Supplementary Fig. 7). Moreover, 15 SNPs on A09 (8,042,653–8,613,448 bp) were substantially associated with

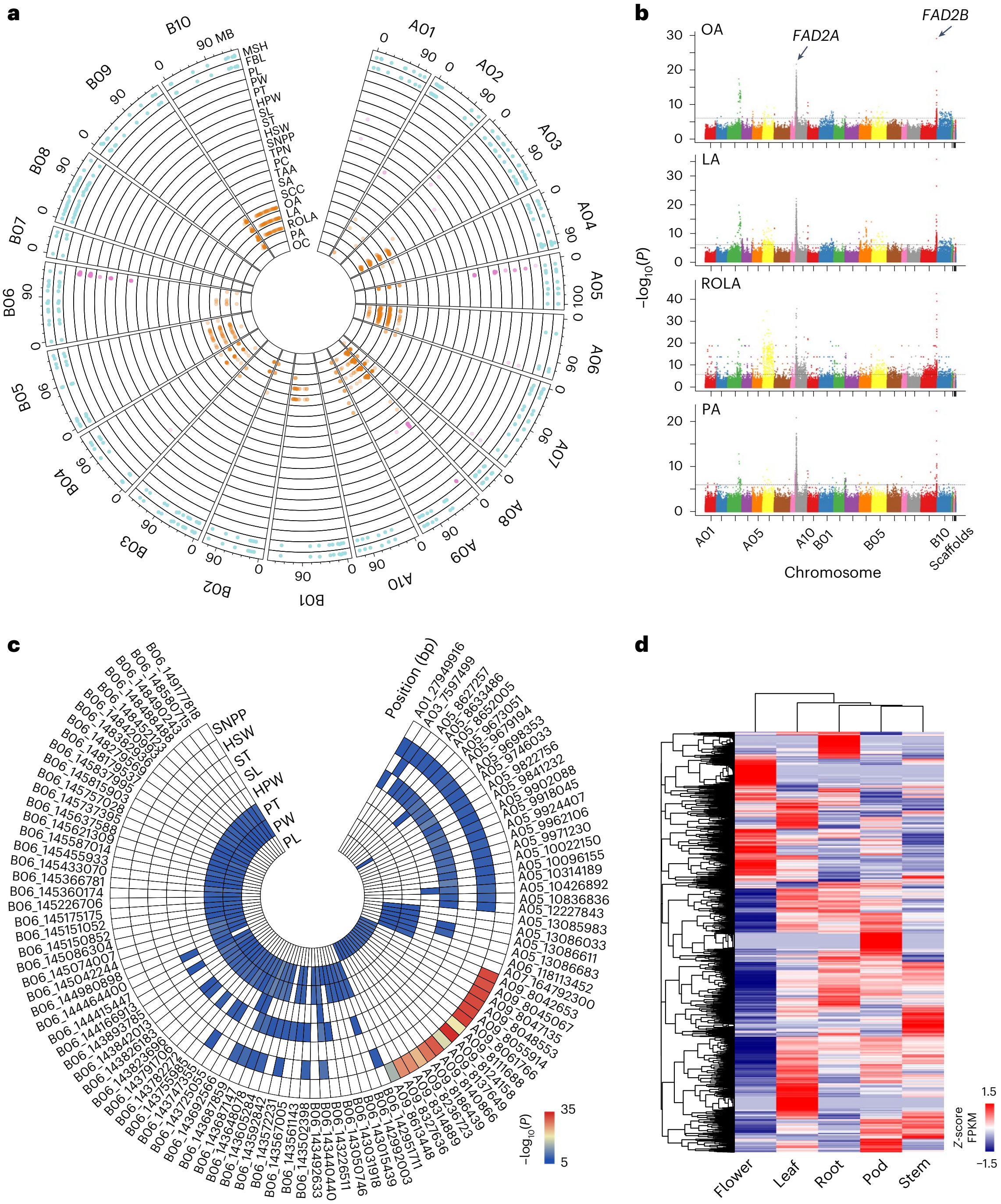

Fig. 2 | Summary of gene linkage or pleiotropy. a, Circular plot of all identified pleiotropic significant SNPs for different traits. **b**, Manhattan plot of GWAS for OA, LA, ROLA and PA. **c**, Circular heatmap plot of the pleiotropy among yield-related traits. **d**, Gene expression patterns in the roots, leaves, stems, flowers and pods according to our previous data[34]. The Bonferroni-corrected genome-wide significance thresholds ($P = 1 \times 10^{-6}$) were used in **b** (horizontal-dashed lines) and **c**.

both pod length (PL) and SNPP, which were also found to exhibit a positive correlation ($r = 0.30$; Fig. 2c and Supplementary Fig. 7). For MSH and FBL traits, 309 pleiotropic associations were detected throughout the genome (Supplementary Fig. 9 and Supplementary Table 8). These findings will help to identify potential pleiotropic genes that may be strong determinants of phenotypic variations under natural and artificial selection in peanuts. Subsequently, a total of 16,018 nonredundant genes were identified within an approximate average LD decay region (Supplementary Tables 9 and 10). Compared to the Bt (7,196), the At contained more predicted genes (8,645; Supplementary Table 10). Most of these genes showed distinct tissue-specific expression patterns (Fig. 2d).

Selective sweeps and asymmetric subgenome selection

Crop improvement is the outcome of continuous artificial selection for high yield and quality. Consequently, many agronomic traits have been dramatically improved following successive artificial selection, resulting in decreased genetic diversity of peanut germplasm. In this study, significant phenotypic variations of the yield- and quality-related traits were observed among the NonC, NC and SC subpopulations (Supplementary Fig. 10). To examine the genetic variations occurring along the history of artificial selection, genome-wide selective signals were detected among the three subpopulations by the cross-population composite likelihood ratio (XP-CLR)[40]. Using the top 5% XP-CLR values, 5,827 (NonC versus NC), 5,825 (NonC versus SC)

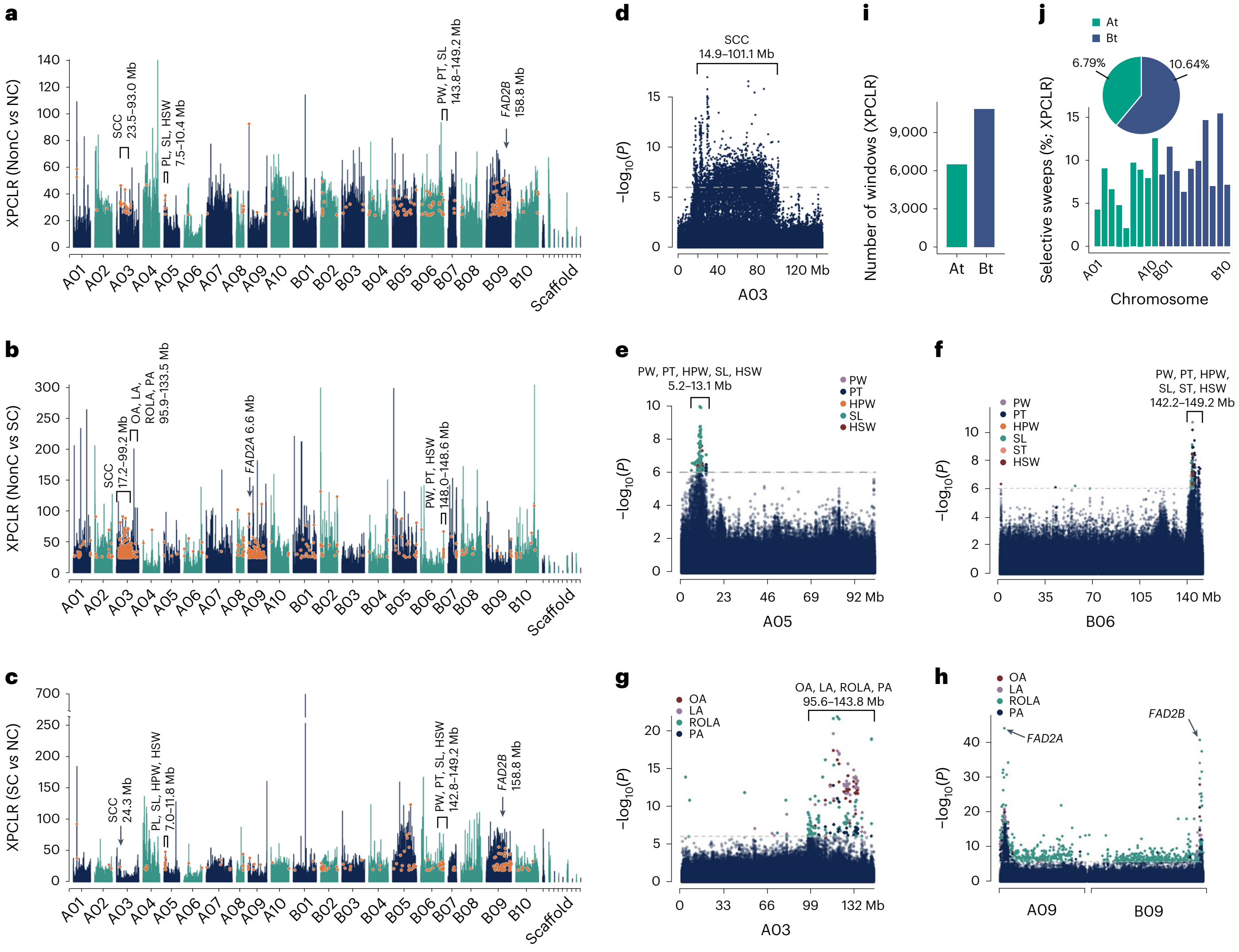

Fig. 3 | Genome-wide screening of selective sweep regions and GWAS signals. **a**–**c**, Whole-genome screening of selective sweep regions and the overlapping GWAS signals. **d**, Manhattan plot of GWAS for SCC on chromosome A03. **e**,**f**, Manhattan plot of GWAS for yield-related traits on A05 and B06. **g**,**h**, Manhattan plot of GWAS for quality-related traits on A03, and A09 and B09. The horizontal-dashed lines in **d**–**h** represent the significant threshold ($P = 1 \times 10^{-6}$, Bonferroni correction). **i**, Total number of selective sweep windows identified in subgenome A (At) and subgenome B (Bt). **j**, Proportion of selective sweep regions on 20 chromosomes and two subgenomes.

and 5,788 (NC versus SC) selective sweeps were identified (Fig. 3a–c and Supplementary Table 11). Among them, 32 selective sweep windows overlapped, indicating that these regions might contain multiple key genes related to artificial selection (Supplementary Fig. 11).

Furthermore, we combined selective and associated signals for yield and quality traits to explore their relationships. The analysis identified 50 selective signals overlapping with previously reported QTLs for various traits (Supplementary Table 11). Moreover, 2,204 substantially associated SNPs located in 1,016 selective sweep regions were identified (Fig. 3a–c and Supplementary Table 12). Several SNPs associated with SCC on chromosome A03 (14.9–101.1 Mb) were found to fit into selective signals among the three subpopulations (Fig. 3a–d). Selective signals were consistently detected on chromosomes A05 (7.0–12.0 Mb) and B06 (142.0–149.0 Mb; Fig. 3a–c), where significant genome-wide associations for yield-related traits were also identified (Fig. 3e,f). Furthermore, the *π* ratio was higher for chromosome B06 than other chromosomes, suggesting that major genes related to yield traits were potentially located in the selective sweep regions (Supplementary Fig. 12a–c). Strong selective sweep signals were observed on chromosome A03 (95.9–133.5 Mb), which overlapped with GWAS signals for oil quality traits such as OA, LA, ROLA and PA (Fig. 3b,g). Multiple selective signals were found on chromosomes A09 and B09, which overlapped with *FAD2A* and *FAD2B*, respectively (Fig. 3a–c,h). Overall, the overlapping signals identified in this study were mainly associated with yield and oil quality traits, consistent with the improvement in high yield and quality in breeding programs.

We examined the selective signals at the subgenome and chromosome levels and found that more selective signals were detected in Bt (10,873) than in At (6,516; Fig. 3i), resulting in longer selective sweep regions in Bt (143.6 Mb, accounting for 10.64% of Bt) as compared to At (78.7 Mb, accounting for 6.78% of At; Fig. 3j). At the chromosome level, B09 had the highest (15.4%) proportion of selective sweep regions, while A06 had the lowest (1.3%; Fig. 3j). These results were supported by calculating the *π* ratio among the three subpopulations (Supplementary Fig. 12d,e). These findings suggested that the subgenomes of cultivated peanuts have been asymmetrically modified by natural or artificial selection.

Pedigree-based genomic signatures of artificial selection

After being introduced into northern and southern regions of China, peanuts in both regions have independently undergone

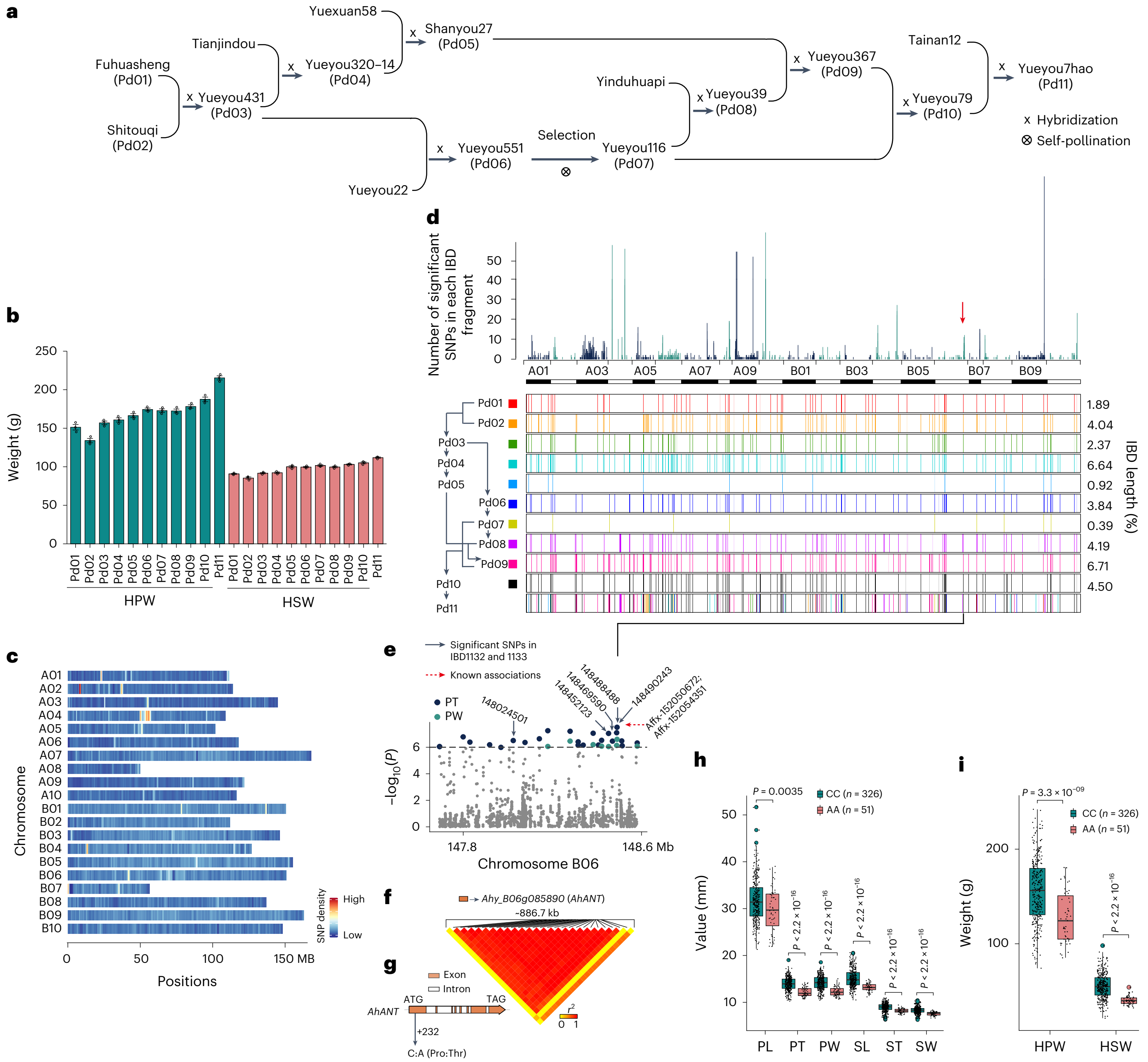

Fig. 4 | Pedigree of Yueyou7hao and GWAS for yield-related traits. a, Pedigree of Yueyou7hao. **b**, HPW and HSW of each variety. Data are given as mean ± s.e.; $n = 3$ biologically independent samples. **c**, SNP density of 20 chromosomes. **d**, Genome flow of Yueyou7hao. The top image presents the number of significant SNPs in IBD fragments. The red arrow indicates the position of the core IBD fragment overlapping the GWAS signal on chromosome B06. The bottom image presents the IBD fragments identified in the pedigree. **e**, Manhattan plot of the GWAS signals for PT and PW, overlapping with the core IBD fragments on chromosome B06. Five solid black arrows indicate the significant SNPs in the core IBD1132 and IBD1133. The red dashed arrow indicates two known associations for seed weight and pod weight[41]. The horizontal-dashed line represents the significant threshold ($P = 1 \times 10^{-6}$, Bonferroni correction). **f**, LD heatmap and the candidate gene *Ahy_B06g085890* (*AhANT*) in the target region. **g**, Gene structure and a nonsynonymous SNP of *AhANT*. **h**,**i**, Box plots for pod and seed size (**h**), and pod and seed weight (**i**) according to the genotype of the nonsynonymous SNP (C/A) in *AhANT*. The number of accessions with CC and AA genotypes is 326 and 51, respectively. Centerline, median; box lower and upper edges, the 25% and 75% quartiles, respectively; whiskers, 1.5× IQR; colored dots, outliers. *P* values were calculated by two-tailed Student's *t* test (**h** and **i**). IQR, interquartile range.

natural and strong artificial selection in breeding programs aimed at improving important agronomic traits, particularly yield. More than half a century of breeder-driven selection has resulted in significant changes in pod yield and related traits, with distinct genomic modifications and phenotypes targeted by different breeders. Here we exemplify the impact of artificial selection on genomic variations with the pedigree of Yueyou7hao, a widely planted peanut cultivar in southern China during the last two decades (Fig. 4a,b). Of the Yueyou7hao pedigree, 11 elite varieties, including 2 landraces and 9 breeding lines, were selected for whole-genome resequencing, identifying a total of 2,394,915 high-quality SNPs, with an average density of 0.92 SNPs per kb (Fig. 4c and Supplementary Tables 13 and 14).

By employing identity-by-decent (IBD) to trace key genomic regions, we identified 16,743 IBD fragments derived from the ten backbone parents. Yueyou116 (Pd07) and Yueyou367 (Pd09) inherited the lowest and highest numbers of foreground IBD fragments, accounting

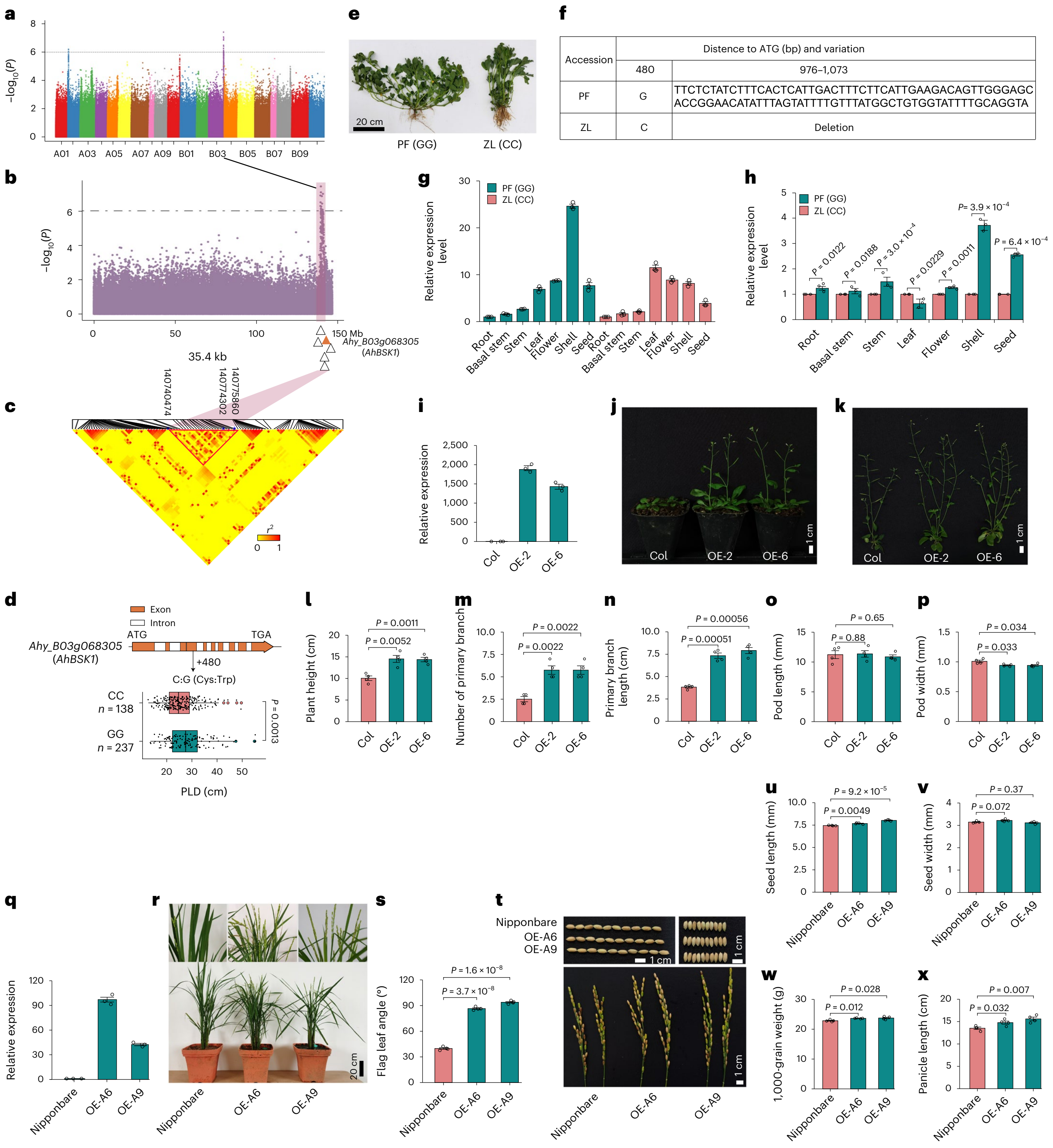

Fig. 5 | GWAS for branching habit and *AhBSK1* identification. **a**,**b**, Manhattan plots for PLD-associated SNPs. The horizontal-dashed lines represent the significant threshold ($P = 1 \times 10^{-6}$, Bonferroni correction). **c**, LD heatmap and the candidate gene *AhBSK1*. **d**, Gene structure and a nonsynonymous SNP of *AhBSK1*. The lower box plot showed a significant difference in PLD between GG (237 accessions) and CC (138 accessions) genotypes. Centerline, median; box left and right edges, the 25% and 75% quartiles, respectively; whiskers, 1.5× IQR; and colored dots, outliers. **e**, Phenotypes of PF and ZL accessions. **f**, SNP and InDel in *AhBSK1* CDS region. **g**,**h**, Relative expression (**g**) and its comparison (**h**) of *AhBSK1* in PF and ZL accessions. **i**, Relative expression of *AhBSK1* in OE and wild-type (Col) *Arabidopsis*. **j**,**k**, The OE of *AhBSK1* resulted in earlier flowering (**j**) and increased biomass (**k**). **l**–**p**, Comparison of the plant height, number of branches, branch length, PL and PW between the OE and Col plants. **q**, Relative expression of *AhBSK1* in Nipponbare and the OE plants. **r**,**s**, The OE of *AhBSK1* had an early heading date (**r**) and a drooping flag leaf (**s**). **t**, Spike and seed sizes of Nipponbare and the OE plants. **u**–**x**, Comparison of the SL, seed width, 1,000-grain weight and panicle length between Nipponbare and the OE plants. Scale bars in **e** and **r**, 20 cm; scale bars in **j**, **k** and **t**, 1 cm. Data in **g**–**i** and **q** are given as mean ± s.e.; *n* = 3 biologically independent samples. Data in **l**–**p**, **s** and **u**–**x** are given as mean ± s.e.; *n* = 4 biologically independent samples. *P* values were calculated by two-tailed Student's *t* test (**h**, **l**–**p**, **s** and **u**–**x**).

for 0.39% and 6.71% of the genome, respectively (Fig. 4d and Supplementary Table 15). The traceable IBD fragments and genes were more abundant in At than Bt, indicating that At contributed more to the genomic content of Yueyou7hao (Supplementary Fig. 13). Furthermore, these IBD fragments overlapped with a total of 8,873 significant SNPs associated with yield- and quality-related traits, accounting for almost half (49.8%) of the total number of significant nonredundant associations (17,083; Fig. 4d and Supplementary Table 15). Of particular importance were 2,064 core traceable IBD fragments that were repeatedly identified in at least two backbone accessions, covering 3,528 significant SNPs involving 8,436 genes (Supplementary Table 16). These core fragments included a series of adjacent fragments on B06 that overlapped with multiple significant SNPs associated with plant-type-, yield-, and quality-related traits (Supplementary Table 16). Moreover, two important core fragments, IBD1132 (147,953,751–148,066,862 bp) and IBD1133 (148,427,293–148,508,442 bp), located at the end of B06, included up to five significant SNPs associated with PT and PW (Supplementary Table 16). The GWAS-based analysis showed that these five SNPs were located in an 886.7-kb LD block (B06: 147,693,870–148,580,715 bp) overlapping two genetic variations associated with seed weight and pod weight identified in a previous study[41] (Fig. 4e,f). This region contained 16 candidate genes with exonic variations, one of which (*Ahy_B06g085890*, named *AhANT*) is highly homologous to *ANT* (At4g37750), which regulates the hormone signaling pathway controlling seed size in *Arabidopsis*[42,43]. The structure of the encoded protein demonstrated that AhANT includes two YRG elements and one RAYD element, which are two key conserved elements in the AP2 domain[44] (Supplementary Fig. 14). Importantly, a nonsynonymous SNP (C/A) detected in *AhANT* resulted in an amino acid change from proline to threonine (Fig. 4g). The pod- and seed-related trait values were substantially higher for the accessions with the CC genotype than the AA genotype (Fig. 4h,i).

Candidate gene for branching habit

Branching habit is an important agronomic trait in peanuts assessed by plant lateral branch dispersion (PLD; Supplementary Figs. 6 and 15). In this study, 13 significant SNPs for PLD were identified on chromosome B03 (Fig. 5a). LD block analyses revealed six genes in a 35.4-kb block (140,740,474–140,775,860 bp), which contains the most significantly associated SNP (140,774,302 bp; $P = 1 \times 10^{-7}$; Fig. 5b,c and Supplementary Table 17). Gene annotation showed that *Ahy_B03g068305* (named *AhBSK1*) encodes a homolog of brassinosteroid (BR) signaling kinase, which might be involved in regulating the plant architecture, as previously reported in *Setaria italica*[45] and *Arabidopsis*[46]. Exon variation revealed that *AhBSK1* contains one nonsynonymous SNP (C/G), resulting in the conversion of cysteine to tryptophan. Moreover, genotype-based association showed that the GG genotype was mainly found in accessions with a larger PLD (Fig. 5d), which was further confirmed in the runner-type accession PF (GG genotype) and the erect-type accession ZL (CC genotype). In addition, an InDel was found in the coding sequence region (Fig. 5e,f and Supplementary Fig. 16). Previous RNA sequencing (RNA-seq) data[34] showed that *AhBSK1* was highly expressed in the pod, leaf and stem (Supplementary Table 17). qRT–PCR analysis validated that *AhBSK1* was highly expressed in the shell and leaf (Fig. 5g), and substantially highly expressed in runner-type varieties (GG genotype; Fig. 5h and Supplementary Figs. 17 and 18). Because *AhBSK1* may function in BR signaling, the quantitative detection of metabolites related to the BR pathway showed significant differences in different tissues between the PF and ZL varieties (Supplementary Fig. 19). The 24-epi-brassinolide (24-epiBL; a synthetic highly active BL analog) treatments substantially increased the seedling height for both PF and ZL varieties but increased the lateral bud growth only for PF at the seedling stage (Supplementary Fig. 20a–c). However, these effects were not found at the stages of flowering and pod-setting (Supplementary Fig. 20d–f). These findings suggested that BRs may be important for controlling branching habits during the early growth stage in peanuts.

Subsequently, the *AhBSK1* with GG genotype was overexpressed in *Arabidopsis* (Fig. 5i). The overexpressing transgenic plants flowered earlier, had larger biomass, and showed higher plant height, more branches and longer branch length (Fig. 5j–n). However, the transgenic plants had slightly smaller pods than the wild type (Fig. 5o,p). In addition, the overexpression (OE) of *AhBSK1* (GG genotype) in *Oryza sativa* L. spp. Japonica cv. Nipponbare increased the flag leaf angle, seed size and panicle length (Fig. 5q–x and Supplementary Fig. 21). Overall, we concluded that *AhBSK1* is a key candidate gene involved in the regulation of peanut-branching habits.

Candidate gene for oil biosynthesis

Improving oil content and quality are important goals for peanut breeders. Thus, GWAS was also performed to identify candidate genes for oil content and quality. A total of 41, 43, 19 and 1 significant SNP on chromosome A08 were identified for PA, LA, OA and OC, respectively (Supplementary Table 18). Among them, one pleiotropic SNP (44,514,436 bp) was associated with LA, OA and OC, and another SNP (44,411,216 bp) was associated with PA (Fig. 6a and Supplementary Table 18). The XP-CLR values indicated that these SNPs underwent positive selection (Fig. 6b). LD block analysis helped us focus on a 74.2-kb region (44,466,905–44,541,087 bp), containing six candidate genes (Fig. 6c). Three of them contained nonsynonymous SNPs, but only *Ahy_A08g040760* (named *AhWRI1*) was highly homologous to wrinkled1 (Supplementary Tables 19 and 20), which was involved in lipid biosynthesis[47–51]. Phylogenetic analysis showed that *AhWRI1* was a close relative to homologs in oil crops such as soybean and rapeseed (Supplementary Fig. 22). The nonsynonymous SNP (G/T) of *AhWRI1* resulted in an amino acid change from arginine to methionine (Fig. 6d). The GG genotype was substantially associated with an increased OC and decreased LA and PA (Fig. 6e). The expression pattern of *AhWRI1* in seed was found to be similar to some of the genes involved in de novo fatty acid (FA) synthesis, elongation and triacylglycerol synthesis[34] (Fig. 6f and Supplementary Table 21). RNA-seq data showed that *AhWRI1* was highly expressed in pods and seeds, especially during the seed-filling stages (Supplementary Table 19 and Supplementary Fig. 23). Two varieties, GH4238 (TT genotype) and CY1016 (GG genotype), with significant differences in OC, OA, LA and PA (Fig. 6g, h), were selected for differential expression analysis. The qRT–PCR revealed that *AhWRI1* was highly expressed in the seed, while other predicted genes showed no significant expression (Fig. 6i and Supplementary Fig. 24), and its expression levels in the shell and seed were substantially higher in CY1016 than that in GH4238 (Fig. 6j). Moreover, the expression of *AhWRI1* was gradually increased during seed development with the highest expression level at seed-filling stages, consistent with the RNA-seq analyses (Fig. 6k and Supplementary Table 19). In addition, *AhWRI1* showed higher expression levels in randomly selected high-oil-content varieties than in low-oil-content varieties (Supplementary Fig. 25).

Transgenic *Arabidopsis* plants with overexpressed *AhWRI1* (GG genotype) showed larger rosette leaves, early flowering, larger pods and longer seeds than the wild type (Fig. 6l–q). Particularly, the seed oil content was about 4.7% higher for the transgenic plants than wild-type plants (Fig. 6r). The transmission electron microscopy images indicated that the transgenic plants had a lower oil body density and a greater oil body volume than the wild type (Fig. 6m,s,t). Hence, we speculated that *AhWRI1* might control oil content by regulating the oil body volume. Moreover, the *AhWRI1*-OE showed different levels of FA composition (Fig. 6u), with a higher unsaturated FA level than wild type (Fig. 6v). However, there was no significant difference in rosette leaf, pod and seed size and oil content between the transgenic plants of TT genotype and the wild type (Supplementary Fig. 26). Therefore, we concluded that *AhWRI1* is a new candidate gene underlying peanut oil content and quality.

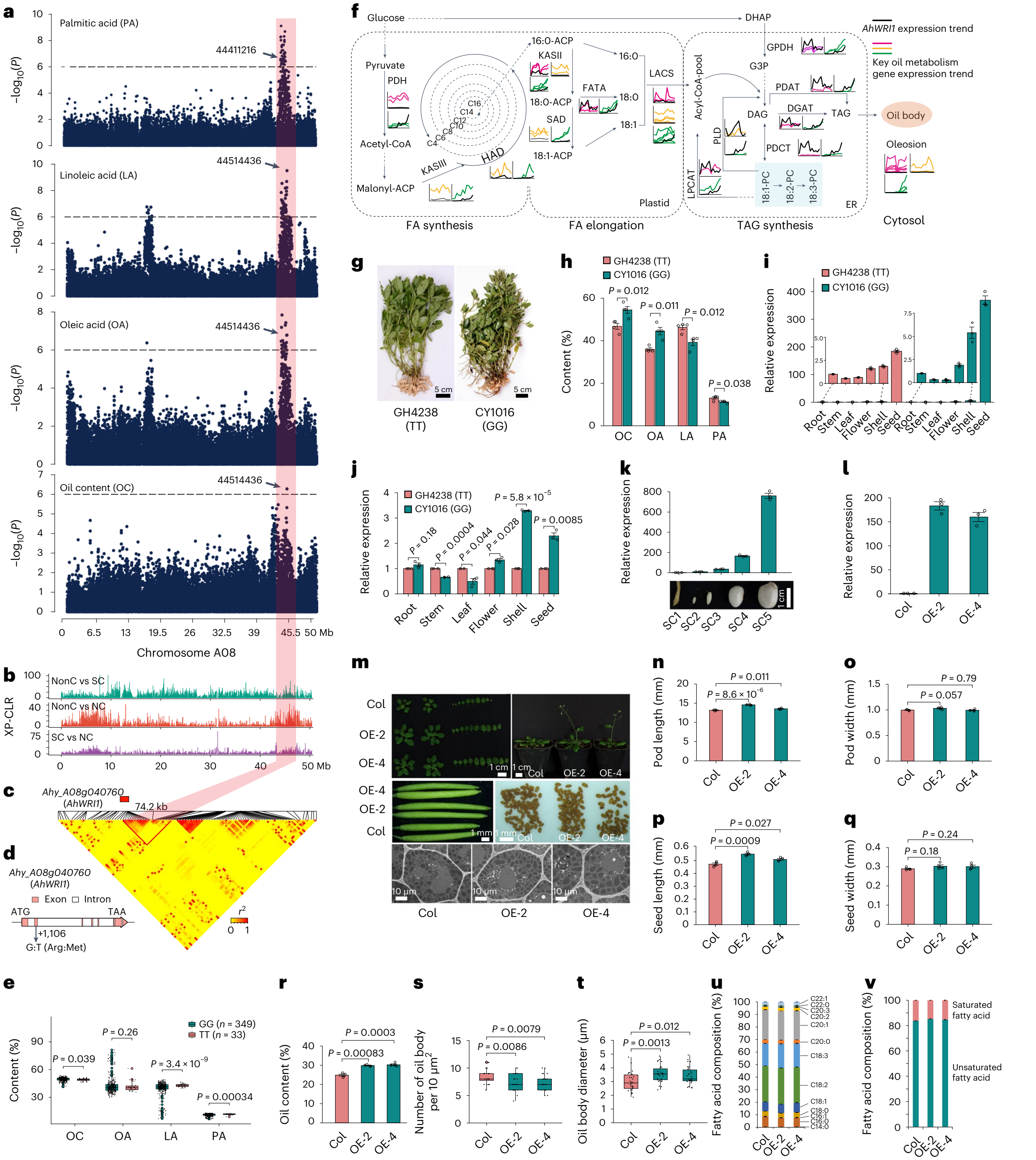

Fig. 6 | GWAS for oil traits and *AhWRI1* identification. a, Manhattan plots for SNPs associated with oil traits on chromosome A08. The dashed lines represent the significant threshold ($P = 1 \times 10^{-6}$, Bonferroni correction). **b**, Selective sweep on chromosome A08. **c**, LD heatmap and the candidate gene *AhWRI1* (red box). **d**, Gene structure and a nonsynonymous SNP of *AhWRI1*. **e**, Box plot of oil traits between GG (349 accessions) and TT (33 accessions) genotypes. **f**, Expression tendency of *AhWRI1* and key oil metabolism genes. **g**,**h**, Phenotypes of GH4238 and CY1016 accessions and their oil traits. Scale bars, 5 cm. **i**,**j**, Relative expression (**i**) and its comparison (**j**) of *AhWRI1* in GH4238 and CY1016 accessions. **k**, Relative expression of *AhWRI1* in developing seeds. Scale bar, 1 cm. **l**, Relative expression of *AhWRI1* in OE and wild-type (Col) *Arabidopsis*. **m**, Plant (top), pod and seed (middle) and oil body (bottom) sizes of the OE and Col. Scale bars in plant, pod and seed, 1 cm; scale bars in cells, 20 μm. **n**–**t**, Statistical analyses of pod and seed sizes (**n**–**q**), oil content (**r**), and oil body density (**s**) and size (**t**) in the OE and Col. **u**,**v**, FA compositions, and unsaturated and saturated FA compositions of the OE and Col. In box plots (**e**, **s** and **t**), centerline, median; box lower and upper edges, the 25% and 75% quartiles, respectively; whiskers, 1.5× IQR; and colored dots, outliers. Sample sizes in **s** and **t** are $n = 10$ cells examined over three independent experiments. Data in **h**–**l** are given as mean ± s.e.; $n = 3$ biologically independent samples. Data in **h**, **n**–**r**, **u** and **v** are given as mean ± s.e.; $n = 4$ biologically independent samples. *P* values were calculated by two-tailed Student's *t* test (**e**, **h**, **j** and **n**–**t**).

Discussion

The consistent growth in the global population has led to a rapid increase in global food demand, posing a challenge to global food security[52]. Peanut is an important crop in developing countries of Asia and Africa, particularly in the semi-arid tropics of the world. It is a source of OA-rich oil, protein, dietary fiber and various vitamins, which can help to combat malnutrition in developing countries. Due to continuous artificial directional selection, the diversity of crops has been drastically reduced, and breeders are engaged in bringing the wild alleles into the cultivated gene pool to diversify the genetic base for crop improvement. Therefore, it is particularly important to determine genetic variations and identify new alleles associated with the peanut phenotype through a comprehensive large-scale genomic analysis. In peanuts, huge genomic resources have been developed, including genome sequences of wild progenitor species *A. ipaensis* and *A. duranensis*[5,35] as well as cultivars Tifrunner[6], Shitouqi[7] and Fuhuasheng[34]. Although some studies have been conducted for genetic diversity analysis in peanuts, the resolution was limited[53–56]. Our study reported a genomic variation analysis in a global collection of 390 peanut germplasm. We performed GWAS for 28 component traits, which identified multiple selective signals relevant to crop improvement, thousands of significant associations, and several candidate genes related to key agronomic traits. In this study, the *AhANT* was identified to be associated with seed and pod weight on chromosome B06. The *ANT*-like genes control organ cell number and size throughout shoot development and negatively regulate salt tolerance in *Arabidopsis*[42]. An important candidate gene *AhBSK1*, BR-signaling kinase family[57], encoding a serine/threonine-protein kinase, was associated with peanut-branching habits. The *AhWRI1*, encoding an ethylene-responsive transcription factor, was identified as being involved in oil biosynthesis in peanut and other crops[58,59]. As an effective transgenic system has not been established in peanuts, these genes for key agronomic traits are still considered candidate genes, although the heterologous expression of these genes improved corresponding traits in *Arabidopsis* or rice. Future studies, including functional genomics methods (for instance, transformation and gene editing), will need to be conducted to verify the biological effects of these genes in peanuts.

In summary, this study involving the large-scale resequencing of peanut accessions generated a substantial amount of new genomic data and identified multiple candidate genes applicable to peanut molecular breeding to accelerate crop improvement.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01660-7.

References

1. Akram, N. A., Shafiq, F. & Ashraf, M. Peanut (*Arachis hypogaea* L.): a prospective legume crop to offer multiple health benefits under changing climate. *Compr. Rev. Food Sci. Food Saf.* **17**, 1325–1338 (2018).
2. Fávero, A. P., Simpson, C. E., Valls, J. M. & Velo, N. A. Study of evolution of cultivated peanut through crossability studies among *Arachis ipaënsis*, *A. duranensis*, and *A. hypogaea*. *Crop Sci.* **46**, 1546–1552 (2006).
3. 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).
4. Samoluk, S. S. et al. First insight into divergence, representation and chromosome distribution of reverse transcriptase fragments from L1 retrotransposons in peanut and wild relative species. *Genetica* **143**, 113–125 (2015).
5. Bertioli, 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).
6. Bertioli, D. J. et al. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* **51**, 877–884 (2019).
7. 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).
8. Yin, D. et al. Genome of an allotetraploid wild peanut *Arachis monticola*: a de novo assembly. *GigaScience.* **7**, giy066 (2018).
9. Pandey, M. K. et al. Advances in *Arachis* genomics for peanut improvement. *Biotechnol. Adv.* **30**, 639–651 (2012).
10. Li, L. et al. GWAS and bulked segregant analysis reveal the loci controlling growth habit-related traits in cultivated peanut (*Arachis hypogaea* L.). *BMC Genomics* **23**, 403 (2022).
11. Li, L. et al. Construction of high-density genetic map and mapping quantitative trait loci for growth habit-related traits of peanut (*Arachis hypogaea* L.). *Front. Plant Sci.* **10**, 745 (2019).
12. Luo, H. et al. Next-generation sequencing identified genomic region and diagnostic markers for resistance to bacterial wilt on chromosome B02 in peanut (*Arachis hypogaea* L.). *Plant Biotechnol. J.* **17**, 2356–2369 (2019).
13. Zhao, K. et al. *PSW1*, an LRR receptor kinase, regulates pod size in peanut. *Plant Biotechnol. J.* **21**, 2113–2124 (2023).
14. Han, S. et al. *AhNPR3* regulates the expression of WRKY and PR genes, and mediates the immune response of the peanut (*Arachis hypogaea* L.). *Plant J.* **110**, 735–747 (2022).
15. Lu, Q. et al. Consensus map integration and QTL meta-analysis narrowed a locus for yield traits to 0.7 cM and refined a region for late leaf spot resistance traits to 0.38 cM on linkage group A05 in peanut (*Arachis hypogaea* L.). *BMC Genomics* **19**, 887 (2018).
16. Luo, H. et al. Discovery of genomic regions and candidate genes controlling shelling percentage using QTL-seq approach in cultivated peanut (*Arachis hypogaea* L.). *Plant Biotechnol. J.* **17**, 1248–1260 (2019).
17. Yang, Y. et al. Genetic analysis and exploration of major effect QTLs underlying oil content in peanut. *Theor. Appl. Genet.* **136**, 97 (2023).
18. Zhu, C., Gore, M., Buckler, E. S. & Yu, J. Status and prospects of association mapping in plants. *Plant Genome* **1**, 5–20 (2008).
19. Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
20. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
21. Tian, F. et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
22. Zhou, Z. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
23. Varshney, R. K. et al. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat. Genet.* **51**, 857–864 (2019).
24. Varshney, R. K. et al. A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* **599**, 622–627 (2021).
25. Fan, W. et al. Sequencing of Chinese castor lines reveals genetic signatures of selection and yield-associated loci. *Nat. Commun.* **10**, 3418 (2019).
26. Jia, G. et al. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45**, 957–961 (2013).
27. Fang, L. et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098 (2017).

28. Ma, Z. et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* **50**, 803–813 (2018).
29. Kang, L. et al. Genomic insights into the origin, domestication and diversification of *Brassica juncea*. *Nat. Genet.* **53**, 1392–1402 (2021).
30. Lu, K. et al. Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat. Commun.* **10**, 1154 (2019).
31. Guo, J. et al. Association of yield-related traits in founder genotypes and derivatives of common wheat (*Triticum aestivum* L.). *BMC Plant Biol.* **18**, 38 (2018).
32. Zhang, X. et al. Genome-wide association study of major agronomic traits related to domestication in peanut. *Front. Plant Sci.* **8**, 1611 (2017).
33. Liu, Y. et al. Genomic insights into the genetic signatures of selection and seed trait loci in cultivated peanut. *J. Adv. Res.* **42**, 237–248 (2022).
34. Chen, X. et al. Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Mol. Plant* **12**, 920–934 (2019).
35. Chen, X. et al. Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc. Natl Acad. Sci. USA* **113**, 6785–6790 (2016).
36. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
37. Collin, F. D. et al. Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Mol. Ecol. Resour.* **21**, 2598–2613 (2021).
38. Pandey, M. K. et al. Identification of QTLs associated with oil content and mapping *FAD2* genes and their relative contribution to oil quality in peanut (*Arachis hypogaea* L.). *BMC Genet.* **15**, 133 (2014).
39. Zhao, Y. et al. Whole-genome resequencing-based QTL-seq identified *AhTc1* gene encoding a R2R3-MYB transcription factor controlling peanut purple testa colour. *Plant Biotechnol. J.* **18**, 96–105 (2020).
40. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
41. 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).
42. Meng, L. S., Wang, Z. B., Yao, S. Q. & Liu, A. The *ARF2-ANT-COR15A* gene cascade regulates ABA-signaling-mediated resistance of large seeds to drought in *Arabidopsis*. *J. Cell Sci.* **128**, 3922–3932 (2015).
43. Schruff, M. C. et al. The *AUXIN RESPONSE FACTOR 2* gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs. *Development* **133**, 251–261 (2006).
44. Okamuro, J. K., Caster, B., Villarroel, R., Van Montagu, M. & Jofuku, K. D. The AP2 domain of *APETALA2* defines a large new family of DNA binding proteins in Arabidopsis. *Proc. Natl Acad. Sci. USA* **94**, 7076–7081 (1997).
45. Zhao, M. et al. DROOPY LEAF1 controls leaf architecture by orchestrating early brassinosteroid signaling. *Proc. Natl Acad. Sci. USA* **117**, 21766–21774 (2020).
46. Sreeramulu, S. et al. BSKs are partially redundant positive regulators of brassinosteroid signaling in *Arabidopsis*. *Plant J.* **74**, 905–919 (2013).
47. Kong, Q., Yuan, L. & Ma, W. WRINKLED1, a 'Master Regulator' in transcriptional control of plant oil biosynthesis. *Plants (Basel)* **8**, 238 (2019).
48. Li, Q. et al. Wrinkled1 accelerates flowering and regulates lipid homeostasis between oil accumulation and membrane lipid anabolism in *Brassica napus*. *Front. Plant Sci.* **6**, 1015 (2015).
49. Liu, J. et al. Increasing seed mass and oil content in transgenic *Arabidopsis* by the overexpression of *wri1*-like gene from *Brassica napus*. *Plant Physiol. Biochem.* **48**, 9–15 (2010).
50. Chen, B. et al. Multiple *GmWRI1s* are redundantly involved in seed filling and nodulation by regulating plastidic glycolysis, lipid biosynthesis and hormone signalling in soybean (*Glycine max*). *Plant Biotechnol. J.* **18**, 155–171 (2020).
51. Pouvreau, B. et al. Duplicate maize *Wrinkled1* transcription factors activate target genes involved in seed oil biosynthesis. *Plant Physiol.* **156**, 674–686 (2011).
52. Tyczewska, A., Woźniak, E., Gracz, J., Kuczyński, J. & Twardowski, T. Towards food security: current state and future prospects of agrobiotechnology. *Trends Biotechnol.* **36**, 1219–1229 (2018).
53. Moretzsohn, M. et al. Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome. *BMC Plant Biol.* **4**, 11 (2004).
54. Ferguson, M. E., Bramel, P. J. & Chandra, S. Gene diversity among botanical varieties in peanut (*Arachis hypogaea* L.). *Crop Sci.* **44**, 1847–1854 (2004).
55. Khera, P. et al. Single nucleotide polymorphism-based genetic diversity in the reference set of peanut (*Arachis spp.*) by developing and applying cost-effective kompetitive allele specific polymerase chain reaction genotyping assays. *Plant Genome* **6**, (2013).
56. Wang, H. et al. Analysis of genetic diversity and population structure of peanut cultivars and breeding lines from China, India and the US using simple sequence repeat markers. *J. Integr. Plant Biol.* **58**, 452–465 (2016).
57. Shi, H. et al. BR-SIGNALING KINASE1 physically associates with FLAGELLIN SENSING2 and regulates plant innate immunity in *Arabidopsis*. *Plant Cell* **25**, 1143–1157 (2013).
58. Qu, J. et al. Dissecting functions of *KATANIN* and *WRINKLED1* in cotton fiber development by virus-induced gene silencing. *Plant Physiol.* **160**, 738–748 (2012).
59. Liu, Z. J. et al. Over-expression of transcription factor *GhWRI1* in upland cotton. *Biol. Plant.* **62**, 335–342 (2018).

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

Methods

Plant materials and sequencing

All 390 analyzed accessions were collected from major global peanut-growing countries, including India, the USA and China, and conserved at the Crop Research Institute, Guangdong Academy of Agriculture Sciences, Guangzhou, China. To ensure the genetic purity of each accession, we first cultivated all accessions in 2016 (early and late growing seasons) and then harvested pods of each accession from individual plants. During the two growing seasons in 2017 and 2018, the accessions were cultivated at the experimental station of Guangdong Academy of Agricultural Sciences, Guangzhou, China. Tender leaves were collected from individual seedlings and immediately frozen in liquid nitrogen for the subsequent DNA extraction. Total genomic DNA (1.5 µg per sample) was extracted using a CTAB method. Whole-genome sequencing libraries were constructed using the TruseqNano DNA HT Sample Preparation Kit (Illumina), after which index codes were added to attribute sequences to specific samples. The libraries were sequenced using the Illumina HiSeq X Ten platform. and a total of 1.29×10^{13} bases were obtained, with a 150-bp read length.

Phenotyping

The 390 accessions were cultivated in four natural environments during the early and late growing seasons of 2017 and 2018 in Guangzhou, China. Three replicates of all accessions were grown in a randomized complete block design. In each plot (6 columns × 6 rows), plants were separated by about 10 cm. A total of 28 agronomic traits, including plant type, yield, quality and disease resistance-related traits, were systematically characterized and scored at maturity. Details of the methods used to measure each trait are summarized in Supplementary Table 4 and Supplementary Fig. 6.

Sequence quality check and filtering

The quality of raw data was assessed using FastQC (v0.11.9; http://www.bioinformatics.babraham.ac.uk/projects/fastqc). To obtain high-quality sequencing data, we filtered the raw sequencing data using Trimmomatic (v0.36)[60] to eliminate the following: (1) reads with ≥10% unidentified nucleotides; (2) reads with >10 adapter nucleotides (≤10% mismatches were allowed); and (3) reads with >50% bases having a Phred quality score of <5. Consequently, clean reads comprising about 1.29×10^{13} bases were retained for subsequent analyses.

Variant detection and annotation

Mapping. All clean reads for each accession were mapped to the cultivated peanut reference genome, *A. hypogaea* L. cv. Fuhuasheng[34], using the command 'mem -t 4 -k 32 -M' of the Burrows-Wheeler Aligner (v0.7.8 -r455)[61]. To minimize the mismatches generated by the PCR amplification before sequencing, SAMtools (v0.1.19)[62] was used to remove duplicated and low-quality (MQ < 30) reads and to convert the mapping results into a BAM format. The SAMtools program was also used to determine the sequencing depth of each site.

Variant calling. SNPs were identified using the Genome Analysis Toolkit (GATK; v2.4-7-g5e89f01)[63]. The genome-wide SNPs were called at the population level. The SNP confidence score of GATK was set as >30, and stand_call_conf was set as 30. To exclude SNP calling errors caused by incorrect mapping or InDels, only high-quality SNPs (depth ≥ 4, MAF ≥ 0.05, miss ≤ 0.2) were retained for subsequent analyses. The InDel calling procedure was similar to that used for identifying SNPs, but the Unified Genotyper parameter '–glm indel only' was applied. Moreover, only InDels ≥ 2 bp and ≤ 50 bp were considered.

Functional annotation. The identified SNPs were annotated according to the cultivated peanut reference genome using the ANNOVAR package (v2018 Apr16)[64]. On the basis of the annotated reference genome, the SNPs were localized to exonic regions (overlapping a coding exon), splicing sites (within 2 bp of a splicing junction), 5′ and 3′ UTRs, intronic regions (overlapping an intron), upstream and downstream regions (within a 1-kb region upstream or downstream of the transcription start site) and intergenic regions. The SNPs in coding exons that did not cause amino acid changes were considered to be synonymous SNPs. The remaining SNPs were classified as nonsynonymous SNPs. Additionally, mutations resulting in the introduction or loss of a stop codon were classified as stop gain or stop loss mutations, respectively. The InDels in exonic regions were classified based on whether they were frameshift (3 bp insertion or deletion) mutations.

Population structure and linkage disequilibrium analyses

A neighbor-joining phylogenetic tree was constructed according to the *P*-distance using TreeBest (v1.9.2)[65], with a bootstrap value of 1,000 to elucidate phylogenetic relationships from a genome-wide perspective. The population structure was analyzed using the expectation–maximization algorithm in ADMIXTURE (v1.3.0)[66], with the ancestry-specifying *K* ranging from 2 to 8 and 10,000 iterations per run. PCA was conducted using the parameter '–make –grm' of the GCTA software (v1.93.0)[67]. The first three principal components were calculated using the parameter '–pca3'. Linkage disequilibrium was calculated using PopLDdecay (v3.30)[68]. The squared correlation coefficient (r^2) for the pairwise analysis of SNPs was calculated to evaluate the LD level of the whole genome and the two subgenomes. The pattern of gene flow was explored using Treemix (v1.13)[36]. To reduce the bias inference, the accessions with a genetic component larger than 0.6 and phylogenetic cluster matched were used to construct the gene flow admixture tree. Admixture trees were constructed with $m = 0–7$ migration events. The *k* parameter (number of SNPs for resampled block) was set as 500. Demographic history analysis was performed using DIYABC Random Forest (v1.2.1)[37], which adopted the Random Forest approach for efficient scenario discrimination at a lower computational burden. Based on the clustering results and prior knowledge of peanut history, we designed eight scenarios to explore the routes of peanut introduction into China. In the 'Training set simulation' module of DIYABC Random Forest, training datasets were generated using 2,000 simulations for each scenario, and the prior distribution was set as uniform. Then the training datasets were subjected to the 'Random Forest analyses' module which constructed 1,000 trees to predict the best scenario and estimate the posterior probability.

Genome-wide selective signals

The XP-CLR values calculated by the XPCLR program (v1.1.2 (ref. 40); https://github.com/hardingnj/xpclr) were used to screen for genome-wide selective signals, using the mean likelihood score of a 40-kb sliding window with a 20-kb step length. Three comparisons (NonC versus SC, NonC versus NC and SC versus NC) of the XP-CLR values were performed to evaluate the genome selection level. The top 5% of XP-CLR values were considered to identify selective sweep regions. To confirm the selective signals, transformed genetic diversity (π) ratios ($\log_2(\pi$ ratio)) for the three comparisons were calculated using VCFtools (v0.1.15)[69] for each 40 kb sliding window. The top 5% $\log_2(\pi$ ratio) values were obtained as the candidate outliers of the selective sweeps.

GWAS

In total, 28 sets of phenotypic and BLUP values were used to perform a large-scale GWAS involving 2,564,993 SNPs filtered at MAF > 0.05. The BLUP values of each trait in different environments were obtained using the BLUP algorithm of the lme4 package (https://cran.r-project.org/web/packages/lme4/) in R (v4.0.2). The association analysis was performed using the Efficient Mixed-Model Association eXpedited (EMMAX) program[70]. The whole-genome significance threshold was set as $-\log_{10}(P) \geq 6$ according to Bonferroni correction[71]. The significant associated SNPs were thoroughly analyzed as follows. First, the

GWAS-associated signals were detected according to the threshold, and the LD block, which was defined based on the pairwise LD correlation (r^2) ≥ 8, was used to estimate the candidate association regions. The LD block was calculated using the LDheatmap package[72] in R (v4.0.2). Second, the nonsynonymous significant SNPs were further analyzed using the annotated variations. Third, each candidate gene containing a nonsynonymous significant SNP was subjected to functional annotation and homology analyses. The homologs in the model species *Arabidopsis* were identified via a sequence comparison using TBtools (v2.028)[73]. Finally, the high-confidence candidate genes were characterized by analyzing their expression and functions in transgenic *Arabidopsis* or Nipponbare plants.

Candidate gene expression and validation using qRT–PCR

Four RNA-seq datasets, two published[34,74] and two unpublished (provided by H. Li and X.C.), were used for preliminary verification of the expression levels of candidate genes. A qRT–PCR analysis was performed to confirm the candidate gene expression level. Total RNA was extracted with a Plant RNA Extraction Kit (TIANGEN, DP432) and reverse transcribed into cDNA using the PrimeScript-RT Reagent Kit (Takara, KR116) according to the manufacturer's instructions. The qRT–PCR assay was performed in triplicate using SYBR Green Master Mix (Yeasen, 11203ES). The target gene expression level was calculated according to the comparative $2^{-\Delta\Delta Ct}$ method[75]. Primers (Supplementary Table 22) were designed using Primer 3 (v4.1.0)[76].

Gene cloning and plant transformation

The full-length open-reading frame of each selected gene was cloned by PCR using the cDNA reverse transcribed from the total RNA isolated from seedlings as the template. The amplified PCR fragment was inserted into the pGEOEP35s-H-GFP vector for the subsequent expression under the control of the cauliflower mosaic virus 35S promoter. The recombinant plasmid was introduced into *Agrobacterium tumefaciens* strain GV3101 and was used to transform *Arabidopsis* (Columbia type) by the floral dipping method[77]. After selecting according to hygromycin B resistance, the T_4 generation transgenic plants were used for the phenotypic analysis of the candidate gene. Pod and seed sizes were measured using an anatomical microscope with a ×1 objective lens and a ×10 eye lens (Mshot). The primers used for gene cloning are listed in Supplementary Table 22. The phylogenetic tree was constructed using MEGA (v7.0) under the neighbor-joining method based on a tree file produced by the CLUSTALW model[78].

IBD analysis

All 11 backbone varieties were sequenced using the Illumina HiSeq X Ten platform. After the quality control and filtering step, a total of 2,394,915 SNPs were identified. To dissect the genetic components of the backbone parents, a method involving sliding windows and SNP ratios was used to detect IBD regions[27]. A window size of 200 SNPs and a step size of 20 SNPs were used to perform the genome-wide scans. To detect IBD fragments in the pedigree, the SNP ratio between Yueyou7hao and individual older cultivars was calculated. An SNP ratio of ≥ 95% in a window was used to identify an inheritable IBD fragment in the pedigree.

Statistical analysis

The statistical analyses were performed in R (v4.2.0). A two-tailed Student's *t* test was conducted in R package ggsignif[79] to compare the difference in gene expression levels, phenotypic values, metabolite contents and oil contents between two groups of samples.

Geographic map generation

The geographical location of the collection sites of all accessions in this study is marked on the map using ggplot2 (ref. 80) package in R (v4.2.0). In Fig. 2b, the Yellow River, the Yangtze River and the Pearl River are manually added to the map according to the public knowledge of China map.

Reporting summary

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

Data availability

All the 390 genomic sequence data for GWAS analysis have been deposited in the National Center for Biotechnology Information (NCBI) database under BioProject number PRJNA776707. All the 11 varieties of genomic sequence data for IBD analysis have been deposited in the NCBI database under BioProject number PRJNA1031811. The SNP and InDel genotypes have been deposited in Zenodo[81] (https://doi.org/10.5281/zenodo.10054109). The published transcriptomic datasets for candidate gene expression analysis can be downloaded from the NCBI Sequence Read Archive under accession numbers SRP167797 and SRP033292 mentioned in the corresponding original literature. Source data are provided with this paper.

Code availability

Custom scripts for calculating the coverage of the aligned sequence are available at Zenodo[82] (https://doi.org/10.5281/zenodo.10023694).

References

60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
62. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
64. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
65. Vilella, A. J. et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
66. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
67. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
68. 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).
69. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
70. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
71. Moran, M. D. Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos* **100**, 403–405 (2003).
72. Shin, J. H., Blay, S., McNeney, B. & Graham, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* **16**, Code Snippet 3 (2006).
73. Chen, C. et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **13**, 1194–1202 (2020).

74. Chen, X. et al. Transcriptome-wide sequencing provides insights into geocarpy in peanut (*Arachis hypogaea* L.). *Plant Biotechnol. J.* **14**, 1215–1224 (2016).
75. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res.* **29**, e45 (2001).
76. Untergasser, A. et al. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
77. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
78. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
79. Ahlmann-Eltze, C. & Patil, I. ggsignif: R package for displaying significance brackets for 'ggplot2'. *PsyArXiv* https://doi.org/10.31234/osf.io/7awm6 (2021).
80. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag: 2016).
81. Lu, Q. SNPs and InDels identified in 390 peanut accessions. *Zenodo* https://doi.org/10.5281/zenodo.10054109 (2023).
82. Lu, Q. An in-house Perl script used for the calculation of the coverage of aligned sequences (1.0). *Zenodo* https://doi.org/10.5281/zenodo.10023694 (2023).

Acknowledgements

This research was partially supported by the Open Competition Program of Top 10 Critical Priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan in Guangdong Province (2022SDZG05 to X.C.), the National Natural Science Foundation of China (32301869 to L.H. and 32172051 to Q.L.), the China Agriculture Research System of MOF and MARA (CARS-13 to X.L.), the Guangdong Provincial Key Research and Development Program-Modern Seed Industry (2020B020219003 to X.C. and 2022B0202060004 to Y.H.), the Guangdong Basic and Applied Basic Research Foundation (2023A1515010098 and 2021A1515010811 to Q.L.), the Guangdong Provincial Department of Science and Technology Project-International Scientific and Technological Cooperation (20200503 to Y.H.), the Special Support Program of Guangdong Province (2021TX06N789 to X.C.), the Agricultural Competitive Industry Discipline Team Building Project of Guangdong Academy of Agricultural Sciences (202104TD to X.C.), the Special Fund for Scientific Innovation Strategy-Construction of High Level Academy of Agriculture Science (R2020PY-JX004 to Q.L., R2020PY-JG005 to X.C. and R2021PY-QY003 to Hao Liu), the Open Fund of Guangdong Provincial Key Laboratory of Crop Genetic Improvement (202101 to Hao Liu and 202201 to Q.L.) and Start-Up grant to R.K.V.

Author contributions

Q.L., X.C., Y.H., X.L. and R.K.V. conceived and designed the study. Q.L., Hao Liu, H. Li, D.G., L.H. and S.L. performed data analysis. Haiyan Liu and R.W. prepared the samples. Q.D. and P.D. measured the agronomic traits. Q.L. wrote the manuscript. R.K.V., V.G., A.C., M.K.P. and S.S.G. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01660-7.

Correspondence and requests for materials should be addressed to Qing Lu, Rajeev K. Varshney, Xuanqiang Liang, Yanbin Hong or Xiaoping Chen.

Peer review information Hon-Ming Lam, Eric von Wettberg, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

nature portfolio

Corresponding author(s): Qing Lu, Rajeev K. Varshney, Xuanqiang Liang, Yanbin Hong & Xiaoping Chen

Last updated by author(s): Dec 13, 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	The resequenced data was generated from Illumina HiSeq X ten sequencing platform.
Data analysis	All software used in the present study were described in detail on the section of Online Methods. Software are listed as follow: FastQC (version 0.11.9) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc); Trimmomatic (version 0.36); Burrows–Wheeler Aligner (version 0.7.8 -r455); SAMtools (version 0.1.19); GATK (version 2.4-7-g5e89f01); ANNOVAR (version 2018 Apr16); TreeBest (version 1.9.2); ADMIXTURE (version 1.3.0); GCTA (version 1.93.0); PopLDdecay (version 3.30); Treemix (version 1.13); DIYABC Random Forest (version 1.2.1); XPCLR program (version 1.1.2) (https://github.com/hardingnj/xpclr); VCFtools (version 0.1.15); EMMAX (version 20120210); R (version 4.0.2);TBtools (version 2.028); Primer 3 (version 4.1.0); MEGA (version 7.0). Custom scripts are available at Zenodo (https://doi.org/10.5281/zenodo.10023694).

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The resequencing data of 390 peanut accessions have been deposited in the National Center for Biotechnology Information (NCBI) database under accession number: PRJNA776707. The resequencing data of 11 varieties for IBD analysis have been deposited in the NCBI database under accession number: PRJNA1031811. The SNP and InDel genotypes have been deposited in Zenodo: https://doi.org/10.5281/zenodo.10054109. The published transcriptomic data sets for candidate gene expression analysis can be downloaded from the NCBI Sequence Read Archive under accession numbers SRP167797 and SRP033292.

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　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental 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

We used 390 peanut accessions for genetic diversity, population structure and genome-wide association study analyses. These accessions collected in this study cover majority of the peanut-planting regions across world. We used a pedigree containing 11 historically famous backbone varieties from South China to elucidate the impact of artificial selection on peanut genome.

Data exclusions

No data was excluded.

Replication

Three biologically independent samples were performed in qRT-PCR analyses. Four biologically independent samples were performed in the estimation of phenotypic analyses in transgenic lines. Phenotypic evaluation of agronomic traits of 390 accessions were preformed with three biological replicates per accession. The details of biological replication in each experiment are shown in the figure legends.

Randomization

A randomized complete block design was used in planting for phenotype data collection in four growing seasons.

Blinding

All accessions were only labeled by numbers when planting and data collection.

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 and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

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