nature genetics

Article

https://doi.org/10.1038/s41588-025-02170-w

Pangenome analysis reveals structural variation associated with seed size and weight traits in peanut

Received: 12 September 2024

Accepted: 17 March 2025

Published online: 28 April 2025

Check for updates

Kunkun Zhao[1,5], Hongzhang Xue[2,5], Guowei Li[3,5], Annapurna Chitikineni[4], Yi Fan[1], Zenghui Cao[1], Xiaorui Dong[2], Huimin Lu[2], Kai Zhao[1], Lin Zhang[1], Ding Qiu[1], Rui Ren[1], Fangping Gong[1], Zhongfeng Li[1], Xingli Ma[1], Shubo Wan[3], Rajeev K. Varshney[4] ✉, Chaochun Wei[2] ✉ & Dongmei Yin[1] ✉

Peanut (*Arachis hypogaea* L.) is an important oilseed and food legume crop, with seed size and weight being critical traits for domestication and breeding. However, genomic rearrangements like structural variations (SVs) underlying seed size and weight remain unclear. Here we present a comprehensive pangenome analysis utilizing eight high-quality genomes (two diploid wild, two tetraploid wild and four tetraploid cultivated peanuts) and resequencing data of 269 accessions with diverse seed sizes. We identified 22,222 core or soft-core, 22,232 distributed and 5,643 private gene families. The frequency of SVs in subgenome A is higher than in subgenome B. We identified 1,335 domestication-related SVs and 190 SVs associated with seed size or weight. Notably, a 275-bp deletion in gene *AhARF2-2* results in loss of interaction with AhIAA13 and TOPLESS, reducing the inhibitory effect on *AhGRF5* and promoting seed expansion. This high-quality pangenome serves as a fundamental resource for the genetic enhancement of peanuts and other legume crops.

Peanut (*Arachis hypogaea* L.) is an important legume crop for vegetable oil and protein[1]. Cultivated peanut is an allotetraploid species that originated from hybridization of two wild diploid progenitors, *Arachis duranensis* (AA) and *Arachis ipaensis* (BB)[2,3]. Pod size is a critical quantitative agronomic trait that impacts peanut yield. During the domestication process, peanuts were selectively bred from wild accessions with small pods to modern varieties characterized by larger pods. Despite limited exploration of causal genes in peanuts[4–11], the molecular mechanisms regulating pod size remain largely unresolved. In addition, currently available genomes still contain numerous gaps and high-quality genomic resources for peanuts are lacking because of the complexities of assembling polyploid genomes[2,12–16]. To optimize the exploration of genomic variations, it is important to have access to high-quality genomic resources across various populations of peanuts.

Different individual genomes of a crop species can vary greatly[17–19]. A pangenome, which amalgamates genomics data from diverse individuals or populations, offers an holistic perspective for a more profound understanding of the genome structural and functional diversity of a species or population[18–21]. The pangenome of a crop can be much larger than its reference genome[19]. Developing a pangenome can identify missing or unannotated genes in the current reference genome, aiding in the discover of genes linked to yield, disease resistance and adaptability, and thus enhancing crop improvement.

In practice, the utilization of genetically diverse germplasm resources for varietal improvement is critical for enhancing peanut

[1]College of Agronomy, Henan Agricultural University, Zhengzhou, China. [2]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [3]Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Ji'nan, China. [4]WA State Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, Western Australia, Australia. [5]These authors contributed equally: Kunkun Zhao, Hongzhang Xue, Guowei Li. ✉e-mail: rajeev.varshney@murdoch.edu.au; ccwei@sjtu.edu.cn; yindm@henau.edu.cn

yield[22]. However, the use of germplasm resources to improve traits is still limited, and genomic variations affecting phenotypic diversity remain underexplored. Recently, resequencing of germplasm resources was imperative for the identification of genomic variations and fundamental to the acceleration of genomic research[23,24]. With the resequencing data of a large number of accessions, genome-wide association studies (GWAS) can be conducted to identify genomic variations associated with specific traits in natural populations. This approach has been extensively utilized in various gramineous and leguminous crops[25–34]. However, research of the mechanism and application after GWAS in peanuts remains limited[23,24,35,36], especially for large-sized variations such as SVs.

In this study, we generated a comprehensive pangenome for both wild and cultivated peanuts, encompassing two diploid wild species, two tetraploid wild species and four tetraploid cultivated species. We assessed the genome-wide diversity of 269 peanut accessions, including wild species, landraces and improved species, and conducted SV-GWAS to investigate agronomic traits related to yield. Our analysis of the peanut pangenome showed significant genomic variations from breeding and highlighted trait-related variations affecting seed size and weight, providing a valuable resource for future improvements.

Results

Genetic diversity of 269 accessions with diverse seed sizes

To explore genetic diversity and population structure in peanut, we collected ~8 TB of resequencing data from 269 accessions with various seed sizes. These accessions come mainly from four populations (32 diploid wild species, 8 tetraploid wild species, 155 tetraploid cultivated landrace, 67 tetraploid improved accessions and 7 unclassified accessions) that are geographically widespread (Fig. 1a). Two accessions were included from previous studies[12,15], and 267 accessions were newly sequenced (Supplementary Table 1). After mapping reads and calling variants to the reference genome (Methods), we obtained 5,989,854 (AA: 5,048,032; BB: 941,822) high-quality SNPs (minor allele frequency >0.05 and a missing rate <0.5). The phylogenetic tree showed that peanut accessions have diverged from wild and cultivated accessions (Fig. 1b). During the domestication of cultivated accessions from wild accessions, seed size and weight were increased significantly (Fig. 1c). Principal component analysis (PCA) of SNPs in both subgenomes A and B showed clear separation of wild and cultivated accessions. A large genetic distance was found between the A subgenomes of AA wild diploids and AABB allotetraploids, whereas high genetic diversity existed in the B subgenome of landrace peanuts (Fig. 1d,e), perhaps due to the limited availability of BB diploid accessions.

AA wild diploid accessions had higher genetic diversity (π; 8.57×10^{-4}) than BB wild diploid accessions (5.69×10^{-5}). However, the allotetraploid accessions had higher values of π in the B subgenome than in the A subgenome (AABB wild: $9.62 \times 10^{-5} > 7.48 \times 10^{-5}$; AABB landrace: $1.46 \times 10^{-4} > 1.17 \times 10^{-4}$; AABB improved: $1.26 \times 10^{-4} > 1.06 \times 10^{-4}$). Of the four populations, AABB landrace accessions had the highest value of π in the B subgenome, which was consistent with the conclusion from PCA (Fig. 1f,g). Population fixation statistics (F_{ST}) values between AA wild diploids and other allotetraploids were high (0.739–0.866), whereas F_{ST} values between BB wild diploids and other allotetraploids were low (0.153–0.201) (Fig. 1f,g). The findings indicate an asymmetrical domestication process between subgenomes A and B. ADMIXTURE population structure analysis revealed peanuts possess ancestry proportions that derived from a combination of AABB wild and landrace accessions, with an estimated cluster number $K = 9$ (Fig. 1h and Supplementary Fig. 1). All tetraploid wild accessions (Amon_N249, A7, US10 and so on) and several landrace accessions (SA080, DL183, SA117 and so on) were clustered into one group and shared more than 35% proportions with their ancestry BB accessions (H104, US38, US37, US36), suggesting that cultivated accessions likely originated from wild allotetraploid accessions.

Generation of a high-quality pangenome with diverse pod sizes

We selected the genomes of eight representative peanut accessions that demonstrate a range of diverse pod sizes for the construction of a high-quality peanut pangenome: two previously published genomes and six newly published genomes, including one AA wild diploid, two AABB wild allotetraploids, one AABB landrace allotetraploid and two AABB improved allotetraploids (Fig. 2a,b). Cultivated peanuts (*A. hypogaea* (Ahy)) originated from tetraploid ancestors (*Arachis monticola* (Amon)) nearly 4,000 years ago, and these were derived from hybridization between two wild diploids (*A. duranensis* (Adu) and *A. ipaensis* (Aip)) approximately 11,000 years ago[37] (Fig. 2c). The genomes of the six new sequenced accessions were assembled de novo based on the combination of Nanopore ultra-long, PacBio HiFi and Hi-C technologies, mainly using NextDenovo and HiC-Pro[38,39] (Methods). These chromosome-level genome assemblies were more contiguous and complete than the previously published ones, with scaffold N50 values of 134–148 Mb, contig N50 values of 25–79 Mb, long terminal repeat (LTR) Assembly Index values of 21.27–29.37 and BUSCO values of 98.6–99.4% (Supplementary Tables 2 and 3), which indicated the high quality of these genomes. Peanut genomes had a repeat sequence content of nearly 85%, with repeats mainly including retroelements (64–67%) and DNA transposons (10–12%) (Supplementary Fig. 2 and Supplementary Table 4). In total 39,736–77,824 protein-coding genes were annotated (Supplementary Table 5).

We constructed a pangenome of peanut using protein-coding genes. The pangenome was composed of 50,097 gene orthogroups (family): 17,137 core, 5,085 soft-core, 22,232 distributed and 5,643 private. Clustering of the gene family by presence–absence variations (PAVs) showed that the newly de novo assembled genomes had significantly more gene families (~37,000) than the two public genomes (~30,000; Wilcoxon rank sum test (WRST) $P < 0.001$) (Fig. 2d and Supplementary Fig. 3a). RNA sequencing (RNA-seq) for five tissues showed that 1,086 genes from 1,021 novel families had strong expression evidence in at least one tissue (transcript coverage ≥95% and fragments per kilobase of transcript per million fragments mapped fragments (FPKM) ≥ 2; WRST $P < 0.001$) (Supplementary Fig. 3b,c) indicating that a large proportion of the novel genes were real. Of the peanut pangenome, nearly 26,493–37,829 gene families were present in each peanut genome (Supplementary Table 5). The numbers of novel gene families and samples followed a power law distribution ($R^2 = 0.987$, $k = 10{,}150$, $a = 1.2$) (Fig. 2e), indicating that the constructed peanut pangenome was closed. Of the 50,097 gene families, 34.2% were core, 10.2% were soft-core, 44.4% were distributed and 11.3% were private (Fig. 2f). All genes were located near the two telomeres of each chromosome, except for the private genes, which showed peaks on chr. 1 and chr. 7 (Supplementary Fig. 4a, b). Core and private genes in the A subgenome, regardless of the type of gene family, were longer than those in the B subgenome (WRST $P < 0.001$ (core) and $P < 0.05$ (private)) (Fig. 2g). The core and soft-core families had longer genes, transcripts, exons and coding sequences (CDSs) than the distributed and private families (WRST, gene, transcript, exon and CDS length, $P < 0.001$) (Supplementary Fig. 5a–d). The core and soft-core family genes in the B subgenome had higher K_a/K_s values than those in A subgenome (WRST, missense and synonymous rates, $P < 0.001$) (Fig. 2h). The genes of core and soft-core families had lower missense and synonymous rates than those of distributed and private families (WRST $P < 0.001$) (Supplementary Fig. 5e–g) and higher expression levels in all tissues investigated (WRST $P < 0.001$) (Supplementary Fig. 5h–l and Supplementary Table 6). The genes of core and soft-core families in both the A and B subgenomes were enriched in glycosphingolipid biosynthesis and signaling pathways regulating the pluripotency of stem cells, whereas the genes of private gene families were enriched in mannose-type *O*-glycan biosynthesis. The genes of soft-core families in the A subgenome were uniquely enriched in aflatoxin biosynthesis and

https://doi.org/10.1038/s41588-025-02170-w

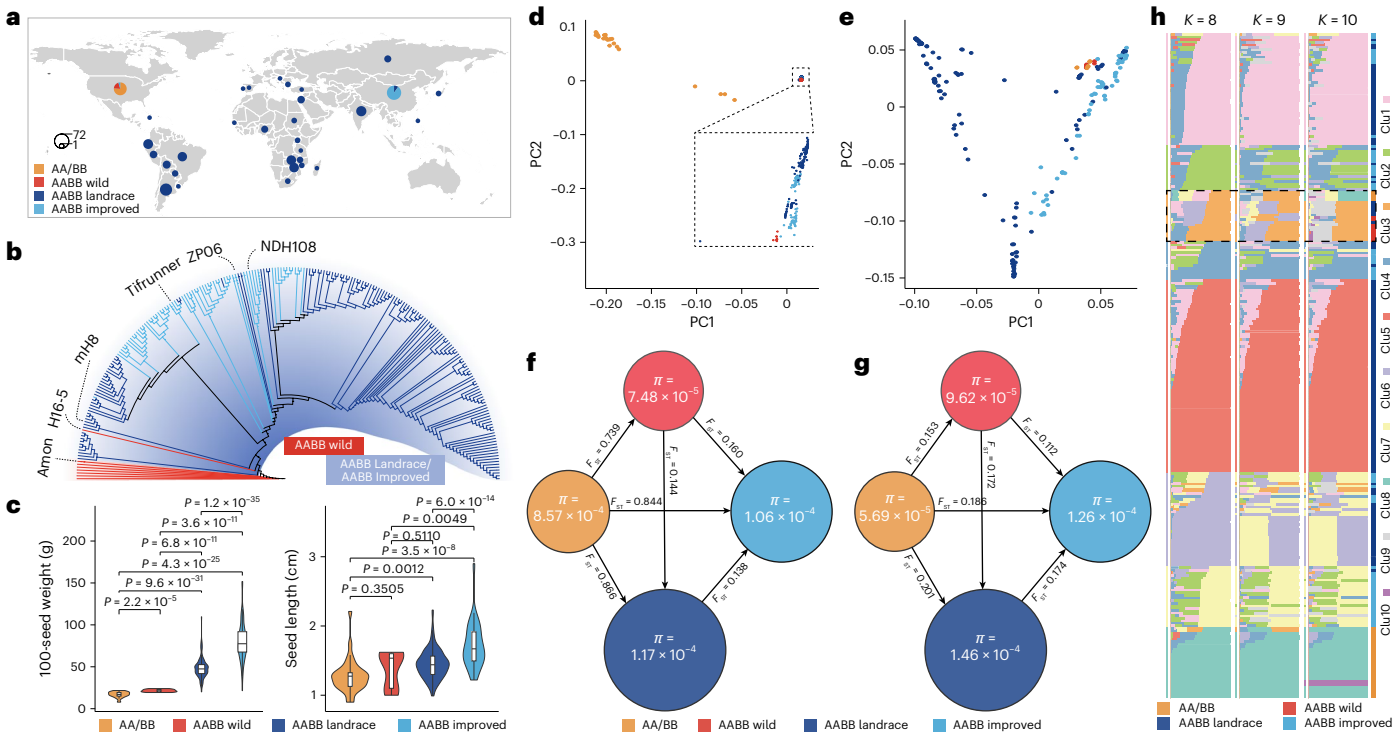

Fig. 1 | Genetic diversity of 269 wild and cultivated peanut accessions. **a**, Different species (*A. hypogaea*, their tetraploid *A. monticola* ancestors, and two wild diploid ancestors *A. duranensis* and *A. ipaensis*) and populations (AA/BB, AABB wild, AABB landrace and AABB improved) of peanut, including 269 peanut accessions with different seed sizes, are widely distributed geographically. **b**, Relationships of peanut accessions are illustrated using a phylogenetic tree based on subsampling 100,000 SNPs. The labeled accessions were used for downstream high-quality genome assembly or pangenome construction. **c**, Seed size (seed weight and length) differs significantly across four peanut populations. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively. *P* values were calculated by two-tailed Student's *t*-test. **d**,**e**, PCA based on SNPs from the A subgenome (**d**) and the B subgenome (**e**). **f**,**g**, Summary of nucleotide diversity (π) and population divergence (F_{ST}) in the A subgenome (**f**) and B subgenome (**g**) between peanut populations. **h**, Results of population structure analysis using ADMIXTURE clustering when $K = 8$, 9 and 10. The cluster with combination of AABB wild, landrace and improved accessions is highlighted in a dashed box. Clu, cluster; PC1, first principal component; PC2, second principal component.

the Notch signaling pathway (Fig. 2i). Core genes found in cultivated peanuts but absent from wild peanuts were enriched in photosynthesis and lipid metabolism, indicating that they underwent convergent evolution during domestication (Supplementary Fig. 6a,b).

We also investigated members from the distributed and private genes families of accession ZP06, which had the largest pod size (Methods). These genes were enriched in Pfam terms such as NB-ARC (nucleotide-binding adaptor shared by Apaf-1, R proteins, and Ced-4) domain, Rx (Potato Virus X Resistance Protein) N-terminal domain, F-box domains and hydroxymethylglutaryl-coenzyme A synthase N-terminal domain (Fig. 2j). For example, most members of the distributed and private gene families with domain PF00931 (NB-ARC domain) were present in the A subgenome (Fig. 2k). However, a member of the gene family OG0025534 in the A subgenome of ZP06 (*SL03G07910*) was unique among the gene families of improved peanuts (Fig. 2l). We noted that gain of a core gene family OG0003529 in NDH108 and ZP06 occurred at chr. 3 of the A subgenome during hybridization and polyploidization, which mainly includes the plant growth regulator indole-3-acetic acid (IAA) amino acid hydrolase ILR1-like 4 or 6 genes (Supplementary Fig. 7a,b).

SVs potentially influencing gene expression

A graph-based peanut pangenome was constructed utilizing a combination of assembly-based and read-based SV detection methods, exemplified by the NB-ARC gene that contains both SVs and small variants (Fig. 3a). Initially, we conducted a synteny analysis across multiple genomes, which revealed significant collinearity in the chromosomes of both the A and B subgenomes (Supplementary Fig. 8). From eight peanut genome assemblies, we identified a total of 86,308 high-confidence SVs, with insertions and deletions being the predominant types (Fig. 3b and Supplementary Fig. 9a,b). We performed genotyping on tetraploid accessions using resequencing data, and validated ten SVs in genes by polymerase chain reaction analysis (PCR; Supplementary Fig. 10a–k). Insertions were positively correlated with deletions, duplications, inversions and SNPs or insertion–deletions (indels) (Supplementary Fig. 11a). There were more SVs in subgenome A than in subgenome B, and these were not evenly distributed across each chromosome, mainly spreading in chr. 8 of A subgenome (Fig. 3c and Supplementary Fig. 11b,c). The size of SVs in the A and B subgenomes varied from 50 to 100,000 bp. The size of insertions was lower in the A subgenome than in the B subgenome, while the sizes of deletions and duplications were higher in A subgenome (Fig. 3d). Of the different SVs, 10–16% overlapped with genes and 15–30% overlapped with the upstream (promoter) or downstream 3,000 bp of genes (Supplementary Fig. 12a); 40–80% of SVs overlapped with repeat elements, particularly LTRs (Supplementary Fig. 12b). The sizes of insertions in promoter and exon regions were extremely variable and were in the three ranges, 50–500 bp, 500–5,000 bp and 5,000+ bp, while the sizes of deletions in promoter and exon regions were not similar (Fig. 3e and Supplementary Fig. 13a). Up to 40% of insertions and deletions in the promoter and exon regions were LTR and DNA transposons (Fig. 3f and Supplementary Fig. 13b,c). Genes with insertions in exons were enriched in glyoxylate and dicarboxylate metabolism and the PPRP (Peroxisome proliferator-activated receptors) signaling pathway, while the promoter insertions were enriched in beta-alanine metabolism, and valine, leucine and isoleucine degradation

https://doi.org/10.1038/s41588-025-02170-w

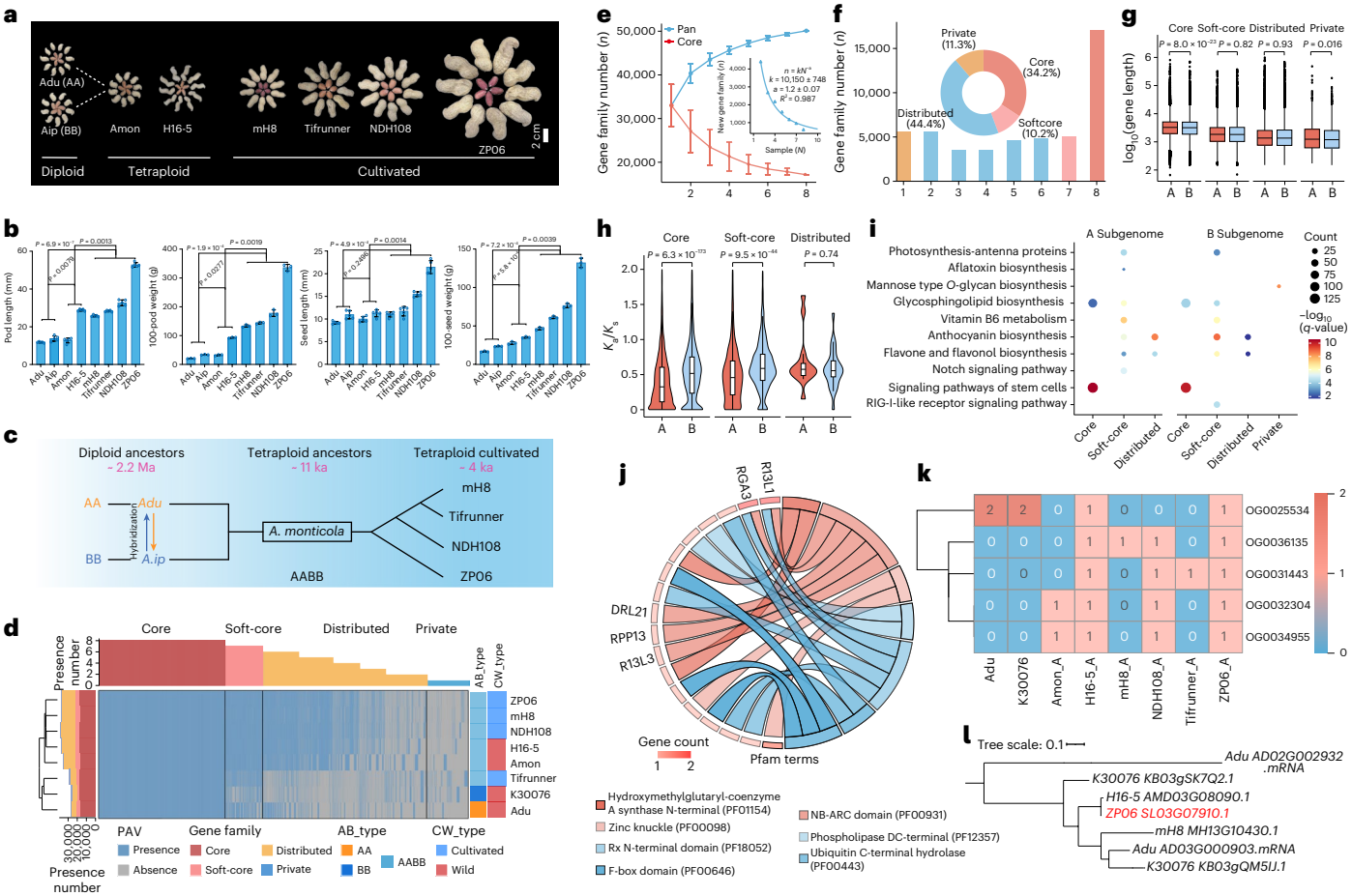

Fig. 2 | Gene-level pangenome in peanuts. a, Comparison of pods from representative diploid, tetraploid and cultivated peanuts. **b**, Pod size measurements, namely pod length, pod weight, seed length and seed weight. Data are given as the mean ± s.d. of five biological replicates. **c**, Possible ancestry and evolution of cultivated peanuts. **d**, Gene family PAVs in representative accessions. **e**, Pan-core gene family curve in representatives. $n = 100$ (iterations of random sampling without replacement). Fitting of power law model of new gene family (n) and sample (N) with $n = kN^{-a}$, where k is a constant, $-a$ is the law's exponent. R^2 is the Pseudo R-squared. **f**, Number and percentage of core (eight of eight), soft-core (seven of eight), distributed (two of eight to six of eight) and private (one of eight) gene members in representative accessions. **g**–**i**, Gene length (**g**), K_a/K_s values (**h**) of core, soft-core, distributed and enriched KEGG pathways, and private gene family members in the A and B subgenomes (**i**). **g**, Core: A $n = 145,544$, B $n = 158,044$; soft-core: A $n = 34,027$, B $n = 41,937$; distributed: A $n = 60,610$, B $n = 72,138$; private: A $n = 3,144$, B $n = 4,884$. **h**, Core: A $n = 15,556$, B $n = 11,793$; soft-core: A $n = 2,839$, B $n = 2,185$; distributed: A $n = 14$, B $n = 25$. **j**, Pfam enrichment analysis of distributed and private gene members in the super-large pod accession ZP06. **k**, PAV of genes with the domain PF00931 (NB-ARC domain) present in the A subgenome. The white number in the boxes is the hit count of protein–genome alignment, which is unlike the strict standard in gene annotation. **l**, Phylogenetic tree of genes in the family OG0025534 based on multiple sequence alignment. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively. P values were calculated by two-tailed Student's t-test (**b**, **g** and **h**). ka, thousand years ago.

(Supplementary Fig. 14). SVs in promoter and exon rather than downstream regions could regulate gene expression (Fisher's exact test) (Fig. 3g and Supplementary Fig. 15a–f).

Our analysis particularly concentrated on accession ZP06 with the largest pod size. In this accession, we observed 2,130 insertions and 2,160 deletions in the A subgenome, and 1,401 insertions and 1,352 deletions in the B subgenome (Fig. 4a). In addition, fewer SVs were found to overlap with CDS, promoter regions and downstream regions in both subgenomes from Amon to ZP06 than from diploids to Amon (Amon to ZP06, 36–338; K30076 to ZP06, 40–650; Adu to ZP06, 269–6,185) (Supplementary Fig. 16a–d). There were more deletions in the B subgenome than in the A subgenome from Amon to ZP06 (SVs overlapping CDS, exon, promoter, gene, downstream: 55, 59, 140, 167 and 231 (A); 77, 86, 223, 186 and 338 (B)) (Supplementary Fig. 16c,d). These SV-related genes in the A subgenome were enriched in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways such as arachidonic acid metabolism and metabolism of xenobiotics by cytochrome P450, whereas those in the B subgenome were enriched in KEGG pathways such as stilbenoid, diarylheptanoid and gingerol biosynthesis, and inositol phosphate metabolism (Fig. 4b). SVs had the potential to affect gene structure or gene expression directly, including some important genes related to yield and disease resistance. For example, *NDH16G00520* (Snakin-2) expression was higher in the flowers and fruits of ZP06, in which there was a 214-bp DNA/DTM (DNA transposon, Mutator) insertion in the promoter, than in other accessions without the insertion (Fig. 4c). In the exon of *NDH14G21970* (glutamine-dependent NAD(+) synthetase), the gene with a 7.9-kb deletion had strong evidence of supported whole-genome sequencing (WGS) of ZP06 (Fig. 4d). A 1.2-kb insertion of a DNA/DTA (DNA transposon, hAT) transposon in an exon of *NDH05G28200* (glycosyl transferase) divided it into two genes (Fig. 4e). A 1.7-kb deletion in ZP06 (encompassing exons of *NDH18G35100* and *NDH18G35110*) resulted in two truncated genes (Fig. 4f). These results indicate that SVs have varied effects on gene functions.

Selective sweeps of trait-related genes in domestication

To better understand how genomic variations affect the function of genes during domestication, we further investigated SVs under

https://doi.org/10.1038/s41588-025-02170-w

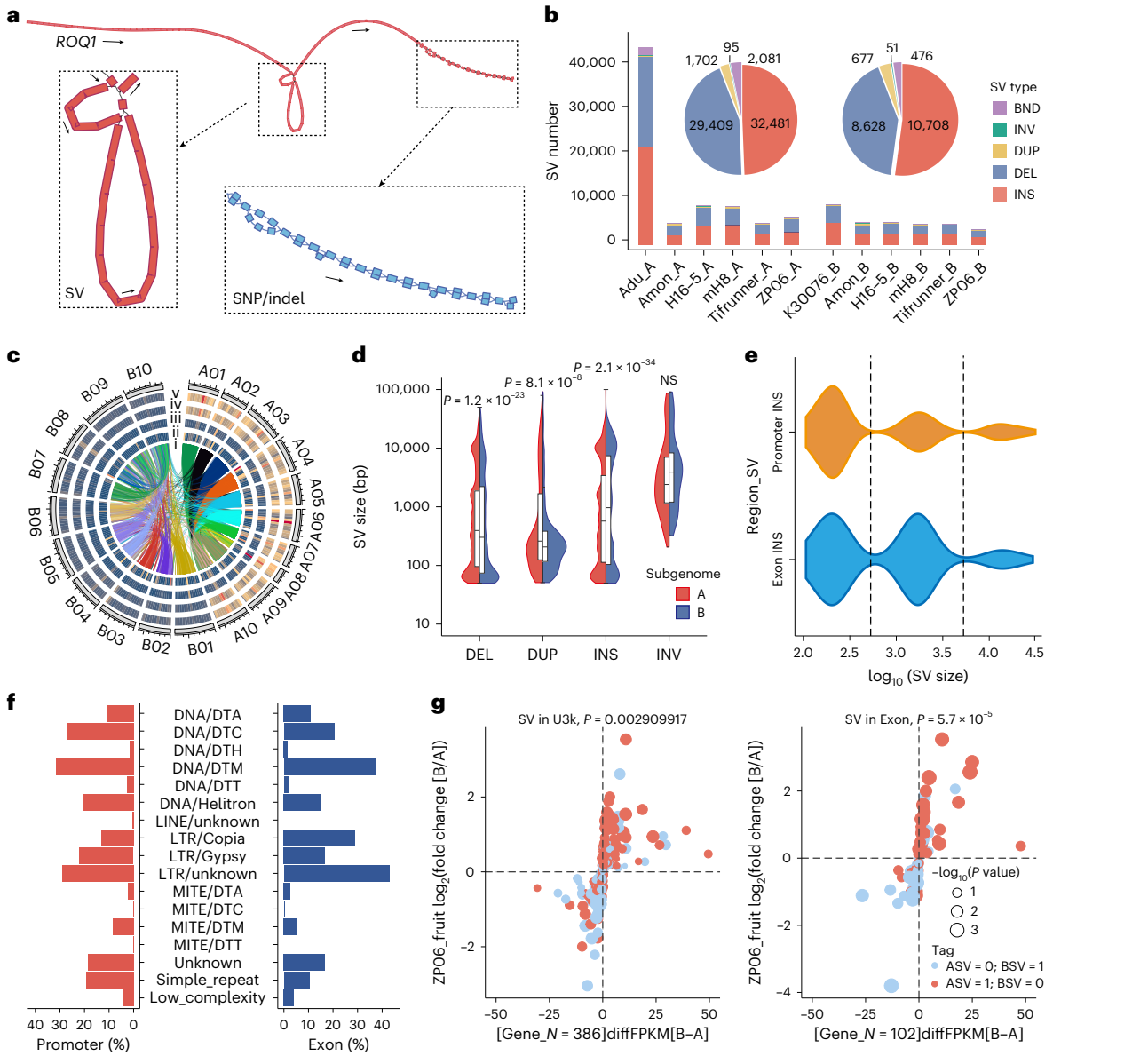

Fig. 3 | Characteristics of SVs in pangenome. a, Plot of a nucleotide-binding oligomerization domain-like receptor (NLR) gene, *ROQ1*, with both large SVs (red) and small variants (SNPs and indels) (blue). **b**, SVs in the A and B subgenomes of representative accessions. **c**, Distribution of SVs (the circles from outer to inner are nonredundant insertions (INS), deletions (DEL), duplications (DUP), inversions (INV) and translocations) in chromosomes. **d**, Sizes of deletions, duplications, insertions and inversions in the A and B subgenomes. DEL: A n = 29,279, B n = 8,495; DUP: A n = 1,661, B n = 666; INS: A n = 32,243, B n = 10,578; INV: A n = 85, B n = 43. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively. **e**, Sizes of insertions in promoter and exon regions. **f**, Repeat content of insertions in promoter and exon regions. **g**, Expression of single-copy genes with SVs in promoter and exon regions. *P* values were calculated by two-tailed Student's *t*-test. ASV, SV in subgenome A; BND, break ends; BSV, SV in subgenome B; DTA, hAT. DTC, CACTA; DTC, CACTA; DTH, PIF-Harbinger; DTM, Mutator; DTT, Tc1-Mariner; FKPM, Fragments per kilobase of transcript per million mapped fragments; MITE, Miniature inverted repeat transposable elements; NS, no significant difference; U3k, Upstream 3,000 bp promoter.

selection between wild and landrace accessions. Genome-wide analysis of selective sweeps using F_{ST}, the nucleotide diversity ratio (π_W/π_L or π_L/π_I) and XP-CLR (the cross population composite likelihood ratio test) showed the unbalanced selection of chromosomes during the evolution of landrace accessions from wild accessions (AmToAhL) and improved accessions from landraces (AhLToAhI) (Fig. 5a,b and Supplementary Fig. 17a,b). Chr. 1 (14.6 Mb), chr. 9 (7.8 Mb), chr. 13 (5.3 Mb) and chr. 20 (5.7 Mb) had more selection during the evolution of landrace accessions from wild accessions, whereas chr. 3 (10.3 Mb), chr. 10 (14.5 Mb), chr. 16 (27.4 Mb) and chr. 19 (31.4 Mb) had more regions under selection during the development of improved accessions from landraces (Supplementary Fig. 17c). Regions more than twice the size (AhLToAhI: 29 Mb (A) and 67 Mb (B)) were under selection in the B subgenome than in the A subgenome during the development of improved accessions from landraces (Fig. 5c).

We identified 1,335 domestication-related SVs overlapping 329 genes (Supplementary Tables 7–9). Significantly, 108 SV-related genes in the A subgenome were subject to selection during the evolution of landraces from wild accessions, which were enriched in the salt stress response, antifungal and cullin family, whereas 114 genes in the B subgenome were enriched in bromodomain, leucine-rich repeat and sucrose synthase. During the evolution of improved accessions from landraces, 59 SV-related genes under selection in the A subgenome exhibited enrichment in the fasciclin domain and aminotransferase class-V,

https://doi.org/10.1038/s41588-025-02170-w

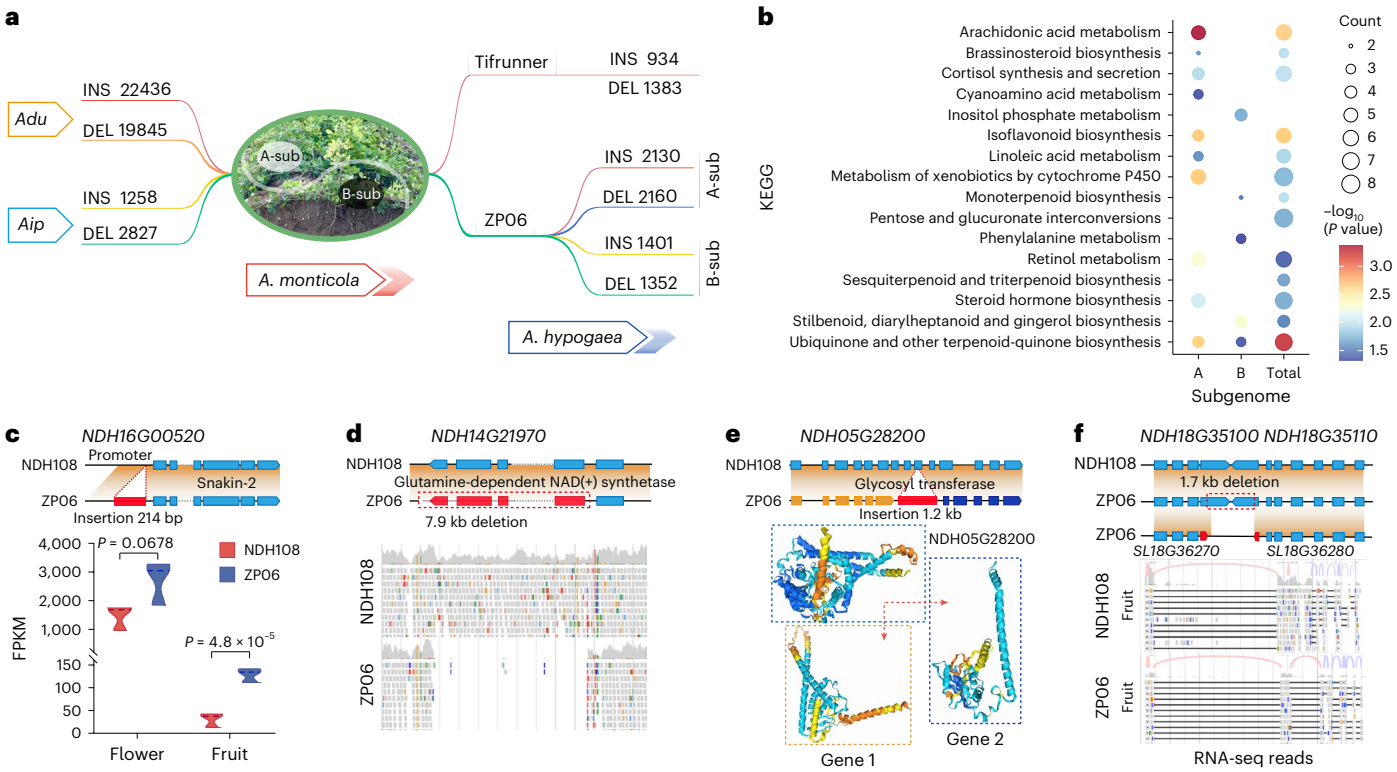

Fig. 4 | Structural variants and gene expression. a, Number of insertions and deletions from diploids to tetraploid in the A and B subgenomes. **b**, Enriched KEGG pathways for genes with SVs in the A and B subgenomes of ZP06. **c**, An allele of *NDH16G00520* (Snakin-2) with a 214-bp insertion has higher expression levels in flowers and fruits than the allele without the insertion. **d**, *NDH14G21970* (glutamine-dependent NAD(+) synthetase) with a 7.9-kb deletion has strong evidence of supported WGS. **e**, A 1.2-kb insertion in *NDH05G28200* (glycosyl transferase) divides it into two genes. **f**, A 1.7-kb deletion *in NDH18G35100* (enhanced disease susceptibility 1) and *NDH18G35110* (γ-aminobutyric acid, GABA permease) generated two truncated genes with different protein sequences.

while 50 genes in the B subgenome were enriched in the legume lectin domain and pathogenesis-related protein bet v1 family (Fig. 5d). Some 41–55% of domestication-related SVs overlap with LTR or Gypsy elements and 35–55% domestication-related SVs overlapped with simple repeats. Of domestication-related SVs from the A and B subgenomes, the percentage of SVs with simple repeats (35% ≪ 52.8–54.7%), Helitron (8% ≪ 15–20%) and low complexity (3% ≪ 7–8%) in the B subgenome were the lowest during the evolution of improved accessions from landrace accessions (Fig. 5e).

Among the genes with domestication-related SVs, 19 were associated with traits such as fruit size or known disease resistance (Supplementary Tables 8 and 9). For example, a higher proportion of diploids and wild allotetraploids had a 629-bp deletion in *CRK26* (a calcium-dependent protein kinase gene), which is found in a cluster of up to 19 *CRK* genes on chr. 9: 114.3–114.9 Mb (Fig. 6a). A 27.7-kb deletion in *IRK* (a leucine-rich repeat gene) was difficult to detect because of alignment deviation near tandem gene duplications, but assemblies can accurately report its position and predict its length (Fig. 6b). A tandem unit consisting of the pod size gene *NTF6* (mitogen-activated protein kinase, brassinosteroid signaling and homeostasis and regulation of grain size and plant height) and the disease-resistance gene *FBRL2* (negative transcriptional regulator of immune responsive genes, including *PR1*), had high copy number variants (Fig. 6c, left), which was unique in peanut and did not exist in rice, soybean and cotton genomes. In addition, higher expression levels of both *NTF6* and *FBRL2* were observed in accessions with larger pod sizes (Fig. 6c, middle), whereas lower expression levels of both were observed in roots after bacterial wilt infection (Fig. 6c, right), indicating strong consistency in the expression of these two genes. These results indicate that the pangenome provides a valuable resource for gene exploration and expansion associated with fruit size and disease resistance during peanut domestication.

SVs in the *CKX6* gene reduces its expression and promotes seed expansion

Cytokinins play crucial roles in plant development by regulating cell division. The mechanisms of cytokinin metabolism and signal transduction are well understood[40,41]. Zeatin and isopentenyladenine-type are considered the predominant active forms of cytokinins[42,43], and cytokinin oxidase/dehydrogenase (CKX) catalyzes the irreversible degradation of cytokinins[44,45]. In this study, we identified 117 SVs or indels associated with seed weight ($P < 0.001$), and one significant SV for seed weight was identified on chr. 3 by SV-GWAS ($P = 1.84 \times 10^{-4}$) (Fig. 7a and Supplementary Table 10). Linkage disequilibrium (LD) block analyses revealed a block (chr. 3: 9.71–9.73 Mb) next to the significantly associated SV in *NDH03G08990* (Fig. 7b). Gene annotation showed that *NDH03G08990* (*AhCKX6*) encodes a CKX containing an FAD-binding domain and a CK-binding domain (Fig. 7c). This gene might be involved in regulating cell division, thereby affecting yield, as previously reported in wheat[46] and chickpea[47]. Gene structure analysis revealed that *AhCKX6* has two insertions (TTACAAATATTATTTATTACTGT and CAAATTTGTC) in the 3′-UTR of accessions with larger seeds (Fig. 7d). Genotype-based association analysis revealed that the 3′-UTR insertions were absent in all 61 wild species. By contrast, three-quarters of the landrace and improved varieties exhibited insertions in the 3′-UTR (Fig. 7e). Furthermore, association analysis revealed that these insertions were significantly correlated with seed length and 100-seed weight (Fig. 7f and Supplementary Table 11).

https://doi.org/10.1038/s41588-025-02170-w

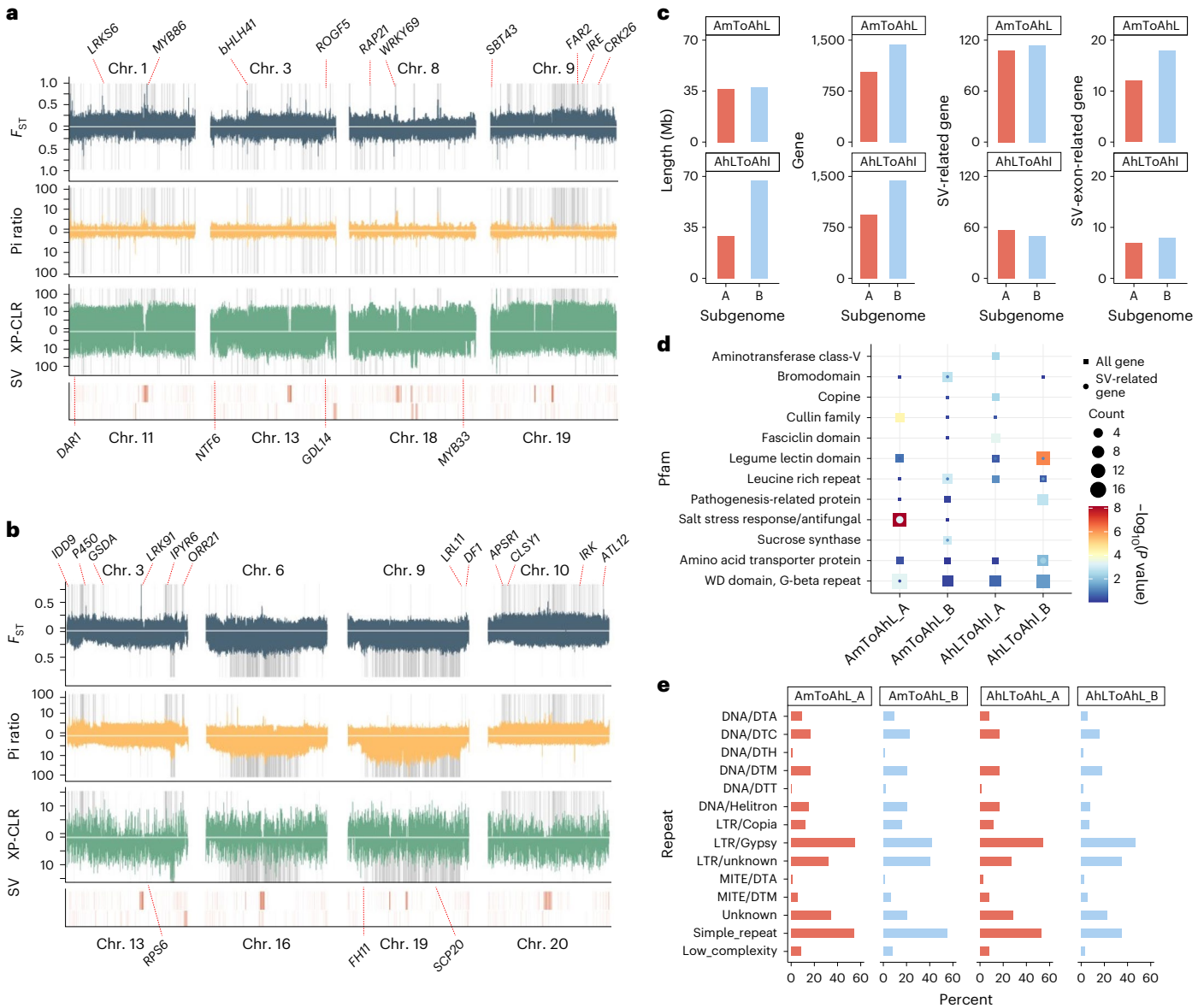

Fig. 5 | Genome-wide selective sweeps during the process of domestication. **a**,**b**, Major genomic regions with evidence of selective sweeps during the evolution of landrace accessions from wild accessions (**a**) and during the evolution of improved accessions from landrace accessions (**b**). F_{ST}, nucleotide diversity ratio (π_W/π_L or π_L/π_I) and XP-CLR tests were used for selection analysis. Regions with evidence of selection (top 5% for at least two metrics) are highlighted. Different SVs from two populations were used for searches of SV-related genes. **c**, Genomic length, genes, SV-related genes and SV exon-related genes under selection in the A (sub-A) and B subgenomes (B-sub). **d**, Pfam enrichment analysis of genes and SV-associated genes with evidence of selection. **e**, Repeat overlap with SVs of selection in the A and B subgenomes.

CKX evolution was traced back only to the first primitive land plant, *Physcomitrium patens*[48]. The ancestral CKX gene evolved in the kingdom Plantae along with the diversification into monocot and eudicot plant species >150 million years ago (Ma)[49]. Following a whole-genome duplication event approximately 58 Ma, *CKX* genes appeared in legumes. They then emerged in the *Arachis* genus around 2.2 Ma (ref. 37). A phylogenetic association study revealed that *AhCKX* genes exhibit a high degree of similarity to soybean *CKX* genes (Fig. 7g and Supplementary Table 12). The expression levels of *AhCKX6* were observed to be higher in the roots, leaves and seeds of HapI compared with HapII (Fig. 7h). In addition, *AhCKX6* expression levels were elevated in small-seed accessions relative to large-seed accessions at various stages of seed development (Fig. 7i). Conversely, large-seed accessions exhibited higher cytokinin content levels (Fig. 7j and Supplementary Table 13). In summary, the insertion in the 3′-UTR of the *CKX6* gene results in reduced expression levels. This increase in cytokinin promotes early cell division, ultimately leading to the development of larger seeds (Fig. 7k).

AhARF2-2 negatively regulates seed size

Seed size is a critical agronomic trait that is closely associated with grain yield. Here, we identified 73 SVs associated with seed size ($P < 0.001$) and a gene (*NDH08G29450*) exhibiting an exon SV was found encode auxin response factor 2 (*ARF2*). This was the most significant SV associated with seed length in SV-GWAS ($P = 3.63 \times 10^{-6}$) (Fig. 8a, Supplementary Fig. 18 and Supplementary Table 14), which acts as a negative regulator of seed size and weight in *Arabidopsis*[50]. In the ZP06 variety, the *AhARF2-2* gene contains a 275-bp deletion and a 7-bp insertion in the 12th exon, as well as premature stop codons in the 13th exon (HapII) (Fig. 8b and Supplementary Fig. 19), resulting in loss of the conserved

https://doi.org/10.1038/s41588-025-02170-w

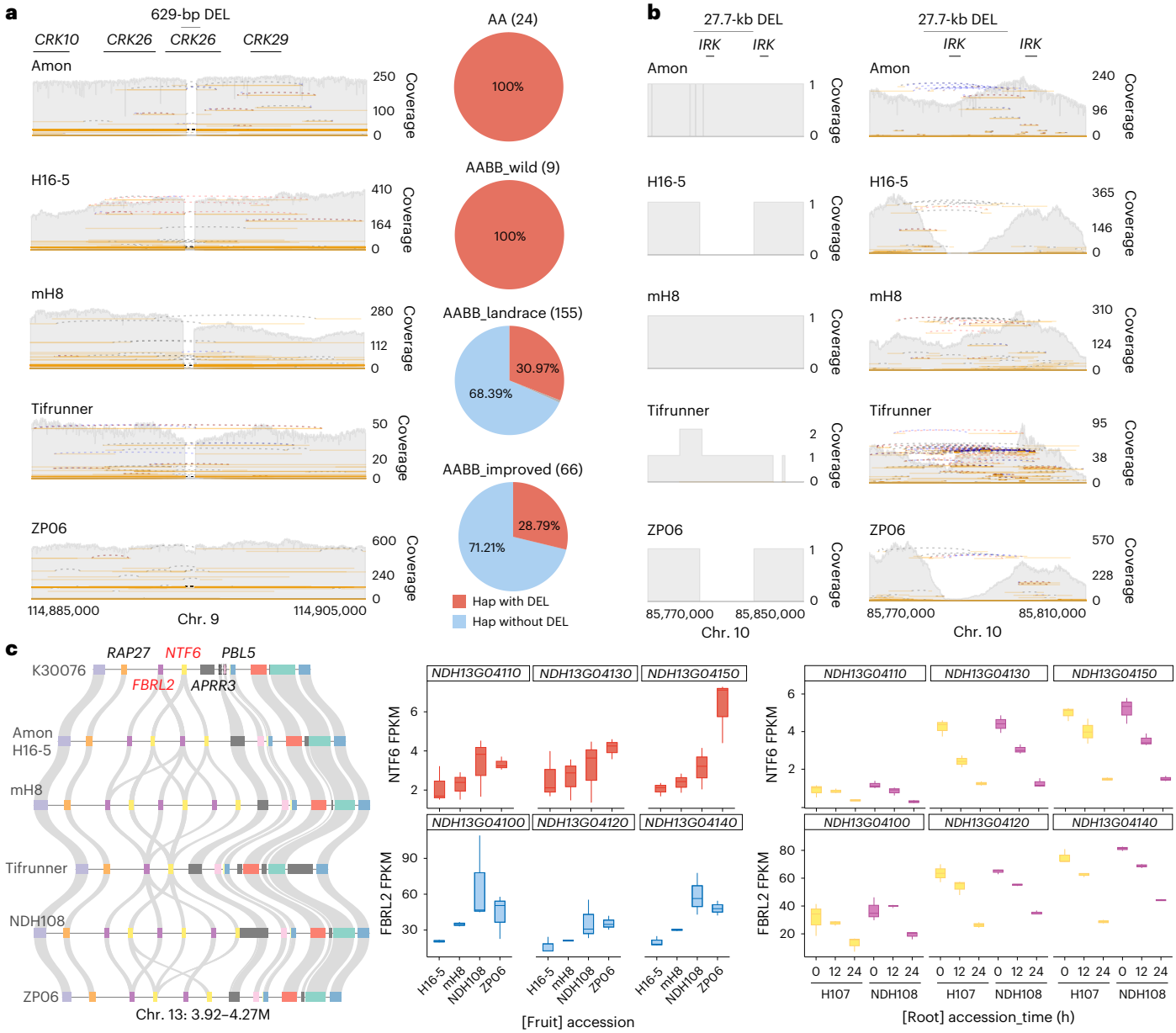

Fig. 6 | Trait-related genes with structural variants in selective sweep regions. **a**, A 629-bp deletion in *CRK26* (*NDH09G29320*) from a *CRK* gene cluster at chr. 9, and genotyping of the population for *CRK26* genes with (red) or without a deletion (steel blue). **b**, A 27.7-kb deletion in an *IRK* gene (*NDH10G19410*) from an *IRK* gene cluster at chr. 10 from haplotype assembly and long-read alignment. **c**, Genome collinearity (left), expression level of seed (middle) from various accessions and expression level of root from accession NDH108 with bacterial wilt infection (right) in the tandem unit of pod size genes *NTF6* (*NDH13G04110/NDH13G04130/NDH13G04150*, mediator of RNA polymerase II transcription subunit 36a) and disease-resistant related genes *FBRL2* (*NDH13G04100/NDH13G04120/NDH13G04140*, protein kinase superfamily protein) at chr. 13. Three biological replicates. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively.

Auxin/Indole-3-Acetic Acid domain (AUX/IAA, PF02309) (Fig. 8c and Supplementary Fig. 20a). This loss is significantly correlated with variations in seed length and 100-seed weight (Fig. 8d and Supplementary Table 15). Molecular markers have been developed based on this SV to differentiate between the HapI and HapII haplotypes (Supplementary Fig. 20b). The *AhARF2-2* gene was highly expressed in the seed and its expression was substantially higher in HapII accessions than in HapI accessions (Fig. 8e). Furthermore, after exogenous application of 1-naphthalene acetic acid (NAA) at the seedling stage, HapI accessions showed a significant upregulated response compared with HapII accessions (Fig. 8f). Further analysis indicated that the mutation did not alter the subcellular localization of its corresponding protein (Supplementary Fig. 20c).

Bimolecular fluorescence complementation and yeast two-hybrid (Y2H) assays confirmed that the *AhARF2* protein from HapI accessions can interact with *AhIAA13* and inhibit *AhTPL*, whereas the protein from HapII accessions has lost this interaction capability (Fig. 8g–i). In peanut, transient overexpression of the *AhARF2-2*^{HapI} and *AhARF2-2*^{HapII} genes resulted in a significant decrease in *AhGRF5* expression, with HapI causing more pronounced downregulation than HapII (Supplementary Fig. 21). These findings suggest that *AhARF2-2* may negatively regulate *AhGRF5* expression in peanuts. Yeast one-hybrid (Y1H) assays and firefly luciferase (LUC) complementation assays demonstrated that both the HapI and HapII *AhARF2-2* proteins can bind to the promoter of the *growth-regulating factor 5* (*AhGRF5*) gene (Fig. 8j). Reverse transcription quantitative PCR (RT–qPCR) analysis revealed that *AhGRF5*

https://doi.org/10.1038/s41588-025-02170-w

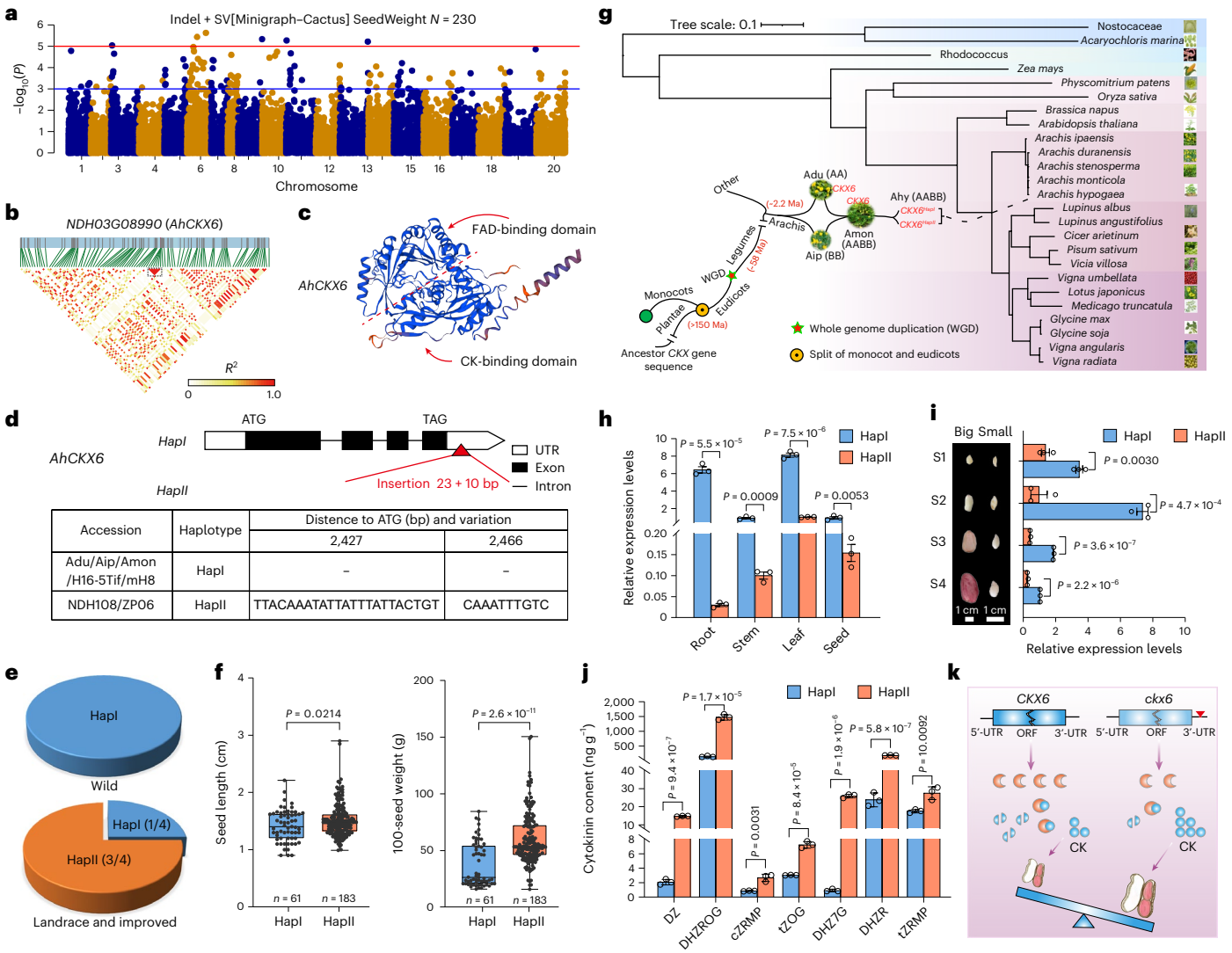

Fig. 7 | SV-GWAS and *AhCKX6* with SVs reduces its expression and promotes seed expansion. **a**, Graph-based SV-GWAS using genotype (SVs and indels from Minigraph-Cactus) and phenotype (seed weight) data from resequencing accessions. Horizontal lines represent the significant threshold ($P = 1 \times 10^{-3}$, 1×10^{-5}). **b**, LD heatmap and candidate gene *AhCKX6*. **c**, Conserved domains of AhCKX6. **d**, Gene structure of *AhCKX6* and location of the insertion. **e**, Proportions of wild, landrace and improved peanut accessions with the two haplotypes of *AhCKX6*. **f**, Box plots for seed length and 100-seed weight according to the genotype of the SV in *AhCKX6*. The numbers of accessions with the HapI and HapII genotypes are 61 and 183, respectively. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively. **g**, Evolution of CKX6 proteins in different plant lineages. Phylogenetic analysis of CKX proteins from various plant lineages using the alignment of 25 full-length CKX6 protein sequences. The ancestral *CKX* gene sequence evolved in the kingdom Plantae along with the diversification into monocot and eudicot plant species >150 Ma. **h**, Expression levels of *AhCKX6* in different tissues of HapI and HapII. **i**, The expression level of *AhCKX6* was higher in HapI than in HapII in developing seeds. **j**, Seeds of HapII accessions contain more cytokinins than those of HapI accessions. $n = 3$ biological replicates. **k**, Model showing the mechanism by which *AhCKX6* regulates peanut seed size. Data in **h**, **i** and **j** are given as mean ± s.e.m. $n = 3$ biological replicates. *P* values were calculated by two-tailed Student's *t*-test (**f**, **h**, **i** and **j**). CK, cytokinin; cZRMP, *cis*-zeatin riboside monophosphate; DHZ7G, dihydrozeatin-7-glucoside; DHZR, dihydrozeatin ribonucleoside; DHZROG, dihydrozeatin-*O*-glucoside riboside; DZ, dihydrozeatin; tZOG, *trans*-zeatin-*O*-glucoside; tZRMP, 9-ribosyl-*trans*-zeatin 5′-monophosphate.

exhibited the highest expression levels in seeds, with significantly higher expression in HapII accessions than in HapI accessions (Supplementary Fig. 22). Moreover, *AhARF2-2* could inhibit *AhGRF5* expression, with HapI exhibiting a stronger inhibitory effect than HapII (Fig. 8k). TOPLESS (TPL) acts as an inhibitory factor in multiple biological processes[51]. AhARF2 has the capability to recruit TPL, thereby further suppressing the expression of *AhGRF5*; thus, HapI exhibits stronger inhibitory capability than HapII (Supplementary Fig. 23). Results from transgenic *Arabidopsis* showed that the seed area and seed length of HapII lines were significantly larger than Hap I transgenic lines (Fig. 8l–p and Supplementary Fig. 24a–e). In summary, we propose that *AhARF2-2* interacts with *AhIAA13* via the C-terminal AUX/IAA domain. In the presence of auxin, *AhARF2-2* is released and subsequently recruits TPL to suppress the expression of downstream *AhGRF5*. However, in large-seed accessions with SV deletion, *AhARF2-2* is unable to interact with *AhIAA13* and TPL, which leads to the reduction of the suppression of *AhGRF5*, thereby promoting seed expansion (Fig. 8q).

Discussion

Peanut (*Arachis hypogaea* L.), recognized as one of the most important oilseed and food legume crops, exhibits the unique characteristic of ripening its seeds underground. Throughout the processes of domestication and cultivation, there has been a progressive increase in the size and weight of peanut pods and seeds[1,12,49]. However, our comprehension

https://doi.org/10.1038/s41588-025-02170-w

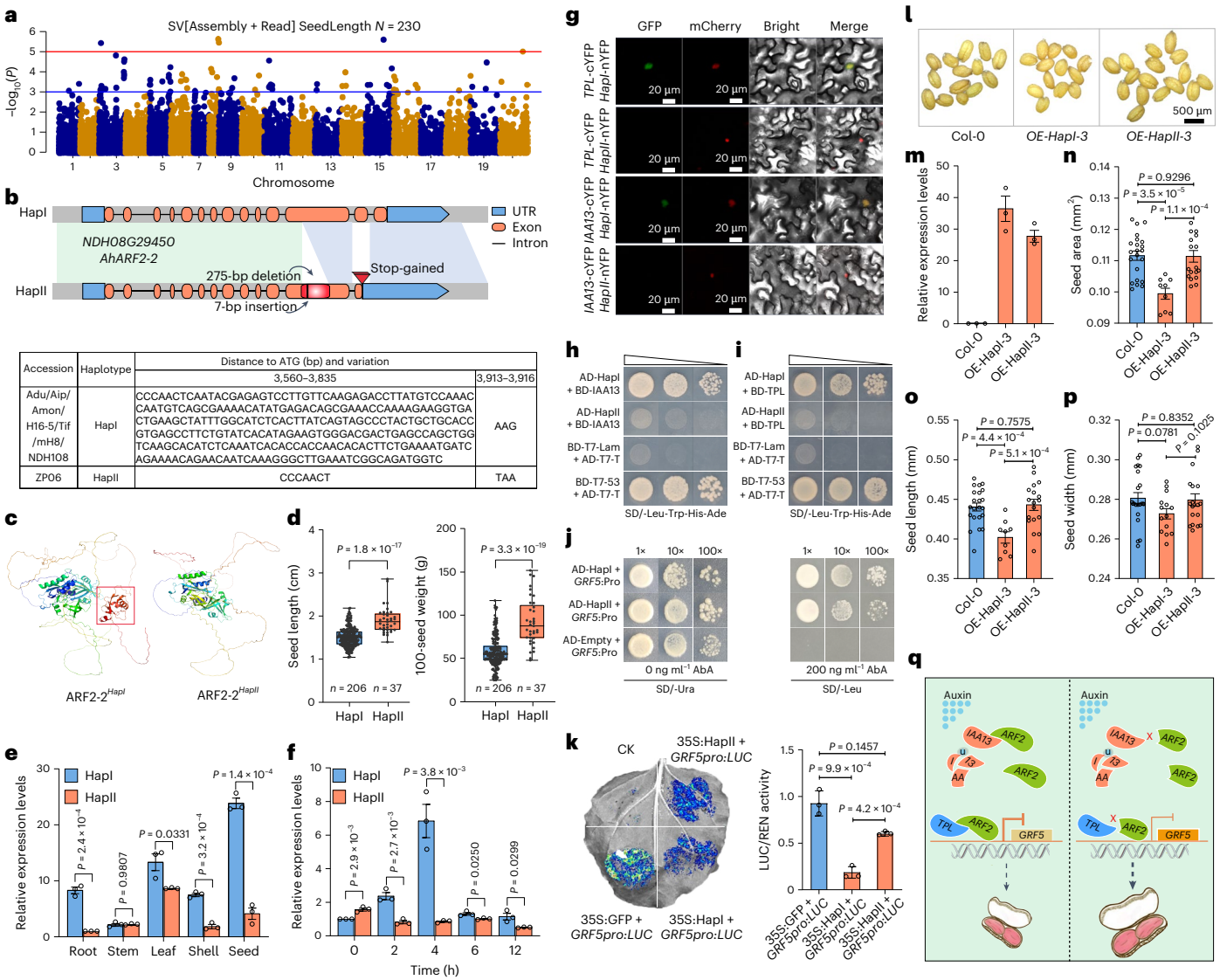

Fig. 8 | *AhARF2-2* negatively regulates seed size. a, Graph-based SV-GWAS using genotype (SVs from Assembly and Read) and phenotype (seed length) data from resequencing accessions. Horizontal lines represent the significance threshold ($P = 1 \times 10^{-3}, 1 \times 10^{-5}$). **b**, Gene structure of *AhARF2-2*. **c**, Protein structure of AhARF2-2. Red boxes represent the missing domain. **d**, Box plots for seed length and 100-seed weight according to the genotype of the SV in *AhARF2-2*. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively. **e**, Relative expression levels of *AhARF2-2* in different tissues. **f**, Expression pattern of *ARF2-2* after NAA treatment. **g**, Bimolecular fluorescence complementation (BiFC) assay of the interaction of AhARF2-2, AhIAA13 and AhTPL. **h**,**i**, Y2H of the interaction of AhARF2-2 with AhIAA13 (**h**) or AhTPL (**i**). BD-T7-Lam (pGBKT7-Lam)/AD-T7-T (pGADT7-T) was used as the negative control, and BD-T7-53 (pGBKT7-53)/AD-T7-T was used as the positive control. AD-HapI/II, pGADT7-*AhARF2-2*$^{HapI/II}$, BD-IAA13/TPL, pGBKT7-IAA13/TPL. **j**, Y1H assay of *AhARF2-2* and *AhGRF5* promoter. *GRF5*:Pro, *pAbAi-AhGRF5* promoter. **k**, Dual-luciferase assay in tobacco leaves expressing LUC. Renilla luciferase (REN) was used as the internal control to normalize values in each assay. 35S:HapI/HapII, pCAMBIA1300-*HapI/HapII*-GFP. **l**, Overexpression of *AhARF2-2*$^{HapI/HapII}$ in *Arabidopsis thaliana*. **m**, Relative expression levels of overexpression (OE) and wild-type (Col) *Arabidopsis*. **n**–**p**, Comparison of seed area (**n**), seed length (**o**) and seed width (**p**) between the OE and Col plants. Col $n = 22$, HapI-3 $n = 9$, HapII-3 $n = 19$. **q**, Model of *AhARF2-2* determining peanut seed size: *AhIAA13* binds *AhARF2-2* via the AUX/IAA domain. Auxin releases *AhARF2-2*, which recruits TPL to suppress *AhGRF5*. Data are given as mean ± s.e.m. $n = 3$ biologically independent samples (**e**, **f**, **k** and **m**). *P* values were calculated by two-tailed Student's *t*-test (**d**, **e**, **f**, **k**, **m**, **n**, **o** and **p**).

of the molecular mechanisms and evolutionary factors influencing peanut pod size and weight remains limited. In this study, we developed comprehensive peanut pangenomes by integrating data from eight high-quality peanut genomes and 269 resequenced accessions. This resulted in an extensive resource of genomic variations that contribute to key agronomic traits in peanuts. The newly constructed pangenomes and identified variations enhance our understanding of the genetic basis underlying traits such as seed size and weight, and will facilitate advancements in crop science and peanut breeding, thereby potentially improving global food security.

Along with SNPs and epigenetic differences, SVs are newly emerging as important variation features contributing to the genetic and phenotypic diversity observed in and between species. Understanding the impact of SVs on plant phenotypic variation is crucial for breeders aiming to develop superior cultivars[51]. SV identification has long been challenging in terms of both a lack of accuracy and comprehensiveness when using short-read resequencing data. Nevertheless, the important role of SVs has been highlighted in recent crop studies[51,52]. Our analysis revealed that combining assembly-based and read-based methods using long-read sequencing data improves the identification of SVs in

plant genomes, especially large SVs. Notably, a significant observation is that most insertion SVs are associated with LTR retrotransposons and DNA transposons. We identified 1,335 domestication-related SVs and 190 SVs associated with seed size or weight. Our study revealed that SVs could influence gene expression, functional dynamics and uneven domestication between the two subgenomes, ultimately affecting seed size and weight.

Throughout the extensive process of domestication and evolution, differential selection pressures have been exerted on subgenomes A and B, resulting in distinct functional specializations, similar to those observed in cotton (*Gossypium* spp.)[53] and rapeseed (*Brassica napus*)[54]. During the transition from landraces to improved cultivars, chr. 3, 16 and 19 experienced intense artificial selection. The genes in the B subgenome were enriched in legume lectin domain and pathogenesis-related protein, such as a tandem unit of two peanut-specific genes associated with pod size and disease resistance separately (Fig. 6c). In crop breeding, enhancing both disease resistance and yield presents major challenges, with trade-offs between these traits being a common phenomenon in crop production[55]. Our research suggested that resistance was compromised as yields increased during peanut domestication. SV-GWAS identified two significant regions potentially related to seed size and weight. Subsequent experiments confirmed that SVs in these genes influenced seed size and weight. Notably, an SV in the 3′-UTR of the *AhCKX6* gene on chr. 3 was associated with seed weight. Another significant candidate gene, *AhARF2-2*, encodes an auxin response factor and negatively regulates seed size. However, our research still has the limitation on the sample size of tetraploid wild peanuts in population selection, owing to the specificity of *Arachis monticola* acting as an intermediate between cultured and wild peanut.

In summary, we have developed an extensive dataset comprising high-quality reference genomes, pangenomes and significant genomic variations, such as SVs, elucidating the influence of SVs on critical traits, including seed size and weight in peanuts. This study furnishes a genetic resource for the identification of functional genes associated with yield and disease resistance, offering valuable tools for breeding and crop enhancement for peanut as well as other polyploid species.

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/s41588-025-02170-w.

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. 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).
3. 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).
4. 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).
5. Liu, Y. et al. Multi-omics profiling identifies candidate genes controlling seed size in peanut. *Plants (Basel)* **11**, 3276 (2022).
6. Guo, F. et al. Transcriptome analysis and gene expression profiling of the peanut small seed mutant identified genes involved in seed size control. *Int. J. Mol. Sci.* **23**, 9726 (2022).
7. Wu, Y. et al. Comparative transcriptomics analysis of developing peanut (*Arachis hypogaea* L.) pods reveals candidate genes affecting peanut seed size. *Front. Plant Sci.* **13**, 958808 (2022).
8. Yang, L. et al. Global transcriptome and co-expression network analyses revealed hub genes controlling seed size/weight and/or oil content in peanut. *Plants (Basel)* **12**, 3144 (2023).
9. Yang, H. et al. Fine mapping of *qAHPS07* and functional studies of *AhRUVBL2* controlling pod size in peanut (*Arachis hypogaea* L.). *Plant Biotechnol. J.* **21**, 1785–1798 (2023).
10. 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).
11. Zhao, K. et al. *PSW1*, an LRR receptor kinase, regulates pod size in peanut. *Plant Biotechnol. J.* **21**, 2113–2124 (2023).
12. Bertioli, D. J. et al. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* **51**, 877–884 (2019).
13. Chen, X. et al. Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Mol. Plant.* **12**, 920–934 (2019).
14. Yin, D. et al. Genome of an allotetraploid wild peanut *Arachis monticola*: a de novo assembly. *GigaScience* **7**, giy066 (2018).
15. 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).
16. Xue, H. et al. A near complete genome of *Arachis monticola*, an allotetraploid wild peanut. *Plant Biotechnol. J.* **22**, 2110–2112 (2024).
17. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
18. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
19. Zhang, F. et al. Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* **32**, 853–863 (2022).
20. Zekic, T., Holley, G. & Stoye, J. Pan-genome storage and analysis techniques. *Methods Mol. Biol.* **1704**, 29–53 (2018).
21. Shi, J., Tian, Z., Lai, J. & Huang, X. Plant pan-genomics and its applications. *Mol. Plant* **16**, 168–186 (2023).
22. Pandey, M. K. et al. Advances in *Arachis* genomics for peanut improvement. *Biotechnol. Adv.* **30**, 639–651 (2012).
23. Lu, Q. et al. A genomic variation map provides insights into peanut diversity in China and associations with 28 agronomic traits. *Nat. Genet.* **56**, 530–540 (2024).
24. Zheng, Z. et al. Chloroplast and whole-genome sequencing shed light on the evolutionary history and phenotypic diversification of peanuts. *Nat. Genet.* **56**, 1975–1984 (2024).
25. Tian, G. et al. Allelic variation of *TaWD40-4B.1* contributes to drought tolerance by modulating catalase activity in wheat. *Nat. Commun.* **14**, 1200 (2023).
26. Tian, F. et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
27. Li, H. et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* **45**, 43–50 (2013).
28. Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
29. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
30. Zhou, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).

31. 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).
32. Kang, L. et al. Genomic insights into the origin, domestication and diversification of *Brassica juncea*. *Nat. Genet.* **53**, 1392–1402 (2021).
33. 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).
34. Varshney, R. K. et al. A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* **599**, 622–627 (2021).
35. Zhang, X. et al. Genome-wide association study of major agronomic traits related to domestication in peanut. *Front. Plant Sci.* **8**, 1611 (2017).
36. 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).
37. Yin, D. et al. Comparison of *Arachis monticola* with diploid and cultivated tetraploid genomes reveals asymmetric subgenome evolution and improvement of peanut. *Adv. Sci.* **7**, 1901672 (2019).
38. Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
39. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
40. Sakakibara, H. CYTOKININS: activity, biosynthesis, and translocation. *Annu. Rev. Plant Biol.* **57**, 431–449 (2006).
41. Werner, T. & Schmülling, T. Cytokinin action in plant development. *Curr. Opin. Plant Biol.* **12**, 527–538 (2009).
42. Schmitz, R. Y. et al. Cytokinins: synthesis and biological activity of geometric and position isomers of Zeatin1. *Plant Physiol.* **50**, 702–705 (1972).
43. Einset, J. W. Conversion of N^6-isopentenyladenine to zeatin by *Actinidia* tissues. *Biochem. Biophys. Res. Commun.* **124**, 470–474 (1984).
44. Frébort, I. et al. Evolution of cytokinin biosynthesis and degradation. *J. Exp. Bot.* **62**, 2431–2452 (2011).
45. Zhang, L. et al. *TaCKX6-D1*, the ortholog of rice *OsCKX2*, is associated with grain weight in hexaploid wheat. *New Phytol.* **195**, 574–584 (2012).
46. Khandal, H. et al. Root-specific expression of chickpea cytokinin oxidase/dehydrogenase 6 leads to enhanced root growth, drought tolerance and yield without compromising nodulation. *Plant Biotechnol. J.* **18**, 2225–2240 (2020).
47. Paces, V., Werstiuk, E. & Hall, R. H. Conversion of N-(delta-isopentenyl) adenosine to adenosine by enzyme activity in tobacco tissue. *Plant Physiol.* **48**, 775–778 (1971).
48. von Schwartzenberg, K. et al. Cytokinins in the bryophyte *Physcomitrella patens*: analyses of activity, distribution, and cytokinin oxidase/dehydrogenase overexpression reveal the role of extracellular cytokinins. *Plant Physiol.* **145**, 786–800 (2007).
49. Khuman, A., Kumar, V. & Chaudhary, B. Evolutionary expansion and expression dynamics of cytokinin-catabolizing CKX gene family in the modern amphidiploid mustard (*Brassica* sp.). *3 Biotech.* **12**, 233 (2022).
50. 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).
51. Yuan, Y., Bayer, P. E., Batley, J. & Edwards, D. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **19**, 2153–2163 (2021).
52. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
53. Wang, M. et al. Asymmetric subgenome selection and *cis*-regulatory divergence during cotton domestication. *Nat. Genet.* **49**, 579–587 (2017).
54. Hu, J. et al. Genomic selection and genetic architecture of agronomic traits during modern rapeseed breeding. *Nat. Genet.* **54**, 694–704 (2022).
55. Derbyshire, M. C. et al. The complex relationship between disease resistance and yield in crops. *Plant Biotechnol. J.* **22**, 2612–2623 (2024).

Methods

Plant materials and growth conditions

All 269 analyzed accessions (two accessions, Tifrunner and Shitouqi, were previously published and others were newly sequenced) were collected from major peanut-growing regions, including India, the United States and China. These accessions, which are currently one of the most complete and representative collections worldwide, were used to analyze the major haplotype of the candidate gene related to the seed traits. Plant DNA was extracted using the Plant DNA Extraction Kit (catalogue number SM262; Seven). Peanut seedlings were grown in pots in a growth chamber set to 28 °C, with a 16 h light and 8 h dark photoperiod and ~60% relative humidity. Experiments were conducted when the seedlings reached the four-leaf stage. All *Arabidopsis thaliana* wild-type and transgenic plants were from the Columbia (Col-0) ecotype background. *Arabidopsis thaliana* Col-0 plants and tobacco (*Nicotiana benthamiana*) plants were grown in pots in a growth chamber at 23 °C with a 16 h light and 8 h dark photoperiod and ~60% relative humidity.

SNP detection from resequencing of 269 peanut accessions

In this study, previously collected WGS and phenotypic data from 269 peanut accessions with various seed sizes were reanalyzed using an updated high-quality reference genome NDH108. Raw sequencing data was preprocessed using FASTP (v.0.23.2)[56]. Trimmed reads were aligned to the reference genome (Near T2T genome: NDH108) using BWA (v.0.7.17)[57]. Moreover, for wild diploid accessions, homologous chromosomes of subgenomes were retained as the reference genomes (NDH108-A or NDH108-B subgenome). SNPs and indels were called using bcftools (v.1.10.2)[58] (parameter: bcftools mpileup --min-ireads 1 -Ou -f $ref ${sample}.bam | bcftools call --ploidy 1 -mv -Oz). The SNPs and indels were then filtered using bcftools (parameter: bcftools filter -g3 -G10 -e 'QUAL<30 || DP<5 || MQ<20'). Variants from all accessions were merged into one file using bcftools (parameter: merge -0). To conduct the analysis for both diploid and tetraploid peanut genomes, SNPs with the A or B subgenome were filtered with a minor allele frequency of >0.05 and a missing rate of <0.5 using VCFtools (v.0.1.16)[59] (parameter: --max-missing 0.5 --maf 0.05), respectively. These SNPs and indels were annotated using SnpEff (v.5.1)[60].

Population-genetic analyses using resequencing data

SNPs in variant call format (VCF) were converted to PHYLIP format using Vcf2phylip (v.2.8)[61] for phylogenetic analysis. The phylogenetic relationship was inferred using the neighbor-joining method with PHYLIP (v.3.697)[62]. PCA was performed using PLINK (v.1.90b6.24)[63] (parameter: --pca). To estimate the genetic similarity in and among populations, we calculated genetic diversity (π or pi) in four population using VCFtools (v.0.1.16)[59] (parameter: --window-pi 100000 --window-pi-step 20000), as well as the population fixation statistics F_{ST} using VCFtools (v.0.1.16)[59] (parameter: --fst-window-size 100000 --fst-window-step 20000). LD pruning was performed using PLINK (v.1.90b6.24)[63] (parameter: --indep-pairwise 50 10 0.2). Population structure was estimated using ADMIXTURE (v.1.3.0)[64], which identified different numbers of clusters (K) according to cross-validation error. Considering both agronomic traits and genetic variability, we newly selected several representative accessions with smaller to larger pods for high-quality genome assembly. These include an accession of Ad (*Arachis duranensis*; Adu), two accessions of Am (*Arachis monticola*; Amon and H16-5) and three accessions of Ahy (*Arachis hypogaea*; mH8, NDH108 and ZP06).

High-quality genome assembly combining several sequencing technologies

Raw Oxford Nanopore Technologies (ONT) long reads with a mean quality score >7 were retained at the pre-processing step. De novo genome assembly was performed with ONT Ultra-long reads or PacBio HiFi reads using NextDenovo (v.2.4.0)[38]. To improve the accuracy of the assembly, the contigs were refined with Racon (v.1.3.1)[65] using long reads and Nextpolish (v.1.3.1)[66] using Illumina short reads with default parameters.

Quality controlling of Hi-C raw data was performed using HiC-Pro (v.2.8.1)[39] as former research. First, low-quality sequences (quality scores <20), adapter sequences and sequences shorter than 30 bp were filtered out using FASTP (v.0.23.2)[56], and the clean paired-end reads were then mapped to the draft assembled sequence using Bowtie 2 (v.2.3.2)[67] (parameter: -end-to-end --very-sensitive -L 30) to get the unique mapped paired-end reads. Valid interaction paired reads were identified and retained by HiC-Pro from unique mapped paired-end reads for further analysis. Invalid read pairs, including dangling-end, self-cycle, re-ligation and dumped products were filtered by HiC-Pro. The scaffolds were further clustered, ordered and oriented scaffolds onto chromosomes by LACHESIS[68] (parameters: CLUSTER_MIN_RE_SITES=100, CLUSTER_MAX_LINK_DENSITY=2.5, CLUSTER NONINFORMATIVE RATIO=1.4, ORDER MIN N RES IN TRUNK=60, ORDER MIN N RES IN SHREDS=60).

Repeat annotation

Transpose elements (TEs) were annotated using EDTA (v.2.0.0)[69] () pipeline (parameter: --sensitive 1 --anno 1), which incorporates several well-performed structure-based and homology-based programs, including LTRharvest (v.1.5.10)[70] (in GenomeTools), LTR_FINDER (v.1.07)[71], LTR_retriever (v.2.9.0)[72], Generic Repeat Finder (v.1.0)[73], TIR-learner (v.2.5)[74], HelitronScanner (v.1.0)[75], TEsorter (v.1.3.0)[76], RepeatModeler (v.2.0.2)[77] and RepeatMasker (v.4.1.1)[78]. The interspersed repeats and tandem repeats of genomes were annotated using RepeatMasker (v.4.1.1)[78] with a TE library generated from EDTA (v.2.0.0)[69] and TRF (v.4.09)[79].

Gene annotation

Three independent approaches, ab initio prediction, homology search and reference guided transcriptome assembly, were used for gene prediction in a repeat-masked genome. In detail, GeMoMa (v.1.6.1)[80] was used to align the homologous peptides from related species to the assembly and then obtain the gene structure information, which was homolog prediction. For transcript-based gene prediction, filtered RNA-seq reads were aligned to the reference genome using STAR (v.2.7.3a)[81]. The transcripts were then assembled using Stringtie (v.1.3.5)[82] d and ORFs were predicted using PASA (v.2.3.3)[83]. For the de novo prediction, RNA-seq reads were de novo assembled using Stringtie and analyzed with PASA to produce a training set. Augustus (v.3.3.1)[84] with default parameters was then utilized for ab initio gene prediction with the training set. Finally, EVidenceModeler (v.1.1.1)[85] was used to produce an integrated gene set from which genes with TE were removed using the TransposonPSI package (http://transposonpsi.sourceforge.net/)[86] and the miscoded genes were further filtered. UTRs and alternative splicing regions were determined using PASA (v.2.3.3)[83] based on RNA-seq assemblies.

Protein assessment and functional annotation

Proteins of each accession were assessed using BUSCO (v.5.3.2)[87] with database embryophyta_odb10 v2020-09-10. Domains of proteins were annotated using InterProScan (v.5.55-88.0)[88]. The Gene Ontology of proteins was predicted using PANNZER2 (v.15.12.2020)[89]. Enrichment analysis was preformed using ClusterProfiler (v.3.16.1)[90].

Gene family PAVs

The longest protein sequences of each gene were selected as representatives. Short proteins with <50 amino acids were ignored. The orthogroups (or gene families) were found using Orthofinder (v.2.5.4)[91] with the parameter '-S blast'. For each accession, orthogroups with at least one member gene were defined as being present. The orthogroups

PAVs matrix was used for downstream analysis. The power model[92] was chosen to fit the new gene family count (n) and sample count (N) as reported previously. Every new gene family count was the median of values from 100 iterations of random sampling without replacement. The curve was fitted using the 'nls' function in R v.4.0.2. Pseudo-R^2 was computed using the R package aomisc v.0.648. The orthogroups in all (eight of eight) accessions were considered as core families. Orthogroups present in more than 90% of samples but not all (seven of eight) accessions were considered as soft-core families, the same threshold used for 'soft-core' in previous studies[18,93]. Orthogroups present in between two and six of eight accessions were considered as distributed families. The orthogroups present in only one (of eight) accession were considered as private families. Single-copy gene orthogroups with fewer than one member per accession were aligned in parallel using ParaAT (v.2.0)[94]. The gene pairs were aligned using MUSCLE (v.5.1)[95] and nonsynonymous (K_a) and synonymous (K_s) substitution rates were calculated using KaKs_Calculator (v.3.0)[96] with the model averaging method[97].

Synteny and collinearity

Synteny and collinearity analyses were performed at both the genome and gene levels. We used D-Geneies (v.1.4)[98] with whole genomes' alignment from Minimap2 (v.2.22)[99] and MCscan (in JCVI) v.1.2.7 (ref. 100) with the CDS of longest protein sequences of each gene alignment from Last (v.1418)[101].

SV detection and pangenome construction

SVs of each accession were called using assembly-based and read-based SV detection methods, generating three graph-based pangenome sets. (1) Assembly-based SVs, graph pangenome construction from variants (SVAss): SVs of each accession were detected with assembly-based methods Svim-asm (v.1.0.2)[102] (parameter: haploid) and CuteSV (v.1.0.13)[103] (parameter: -s 1 --genotype --report_readid -p -1 -mi 500 -md 500 –max_cluster_bias_INS 1000 --diff_ratio_merging_INS 0.9 --max_cluster_bias_DEL 1000 --diff_ratio_merging_DEL 0.5). Insertions (INS) and deletions (DEL) from the intersection of two tools were compared and selected using Truvari (v.3.4.0)[104] (parameter: bench). Duplications (DUP), inversions (INV) and breakends (BND) were selected from the results of CuteSV (v.1.0.13)[103]. SVs with supporting sequence names (RNAMES in the INFO column) from the same chromosomes or unplaced contigs were retained. Moreover, for two wild diploid accessions (Adu and K30076), the SVs from homologous chromosomes (Adu, A subgenome; K30076, B subgenome) or unplaced contigs were retained. The nonredundant SVs set was merged using SURVIVOR (v.1.0.7)[105] (parameter: merge 1000 1 1 -1 -1 50). We constructed graph-based pangenome from variants and reference sequences using vg v.1.53.0. (2) Assembly-based SVs and small variants, graph pangenome construction from assemblies (FASTA) (Minigraph-Cactus): base-level pangenome graphs were constructed of each chromosome using Minigraph-Cactus (v.2.7)[106]. The DNA sequences (nodes) in the sequence graph (GFA) of each chromosome were renamed (chromosome prefixes + raw node identification) and merged into a graph of all chromosomes. (3) Assembly-based + read-based SVs, graph pangenome construction from variants (SV Assembly and Read): we added a read-based SV detection method to avoid heterozygous and unassembled-region SVs omitted in the single haplotype of genome assembly. SVs of each accession were detected using the read-based method CuteSV (v.2.0.3)[103] from ONT reads and PacBio HiFi reads (parameter: --max_cluster_bias_INS 100 --diff_ratio_merging_INS 0.3 --max_cluster_bias_DEL 100 --diff_ratio_merging_DEL 0.3 (ONT); --max_cluster_bias_INS 1000 --diff_ratio_merging_INS 0.9 --max_cluster_bias_DEL 1000 --diff_ratio_merging_DEL 0.5 (PacBio HiFi)). The SVs with the 'PRECISE' and 'PASS' tags were filtered. Next, these SVs were divided into two subsets according to the regions of assembly alignment: SVs in assembly-covered regions and SVs in assembly-uncovered regions. For subset 1, heterozygous insertions and deletions were compared and merged with assembly-based SVs using Truvari (v.3.4.0)[104] (parameter: bench). For subset 2, insertions and deletions were added with assembly-based SVs. The final SVs dataset included both assembly-based SVs and read-based SVs. We constructed graph-based pangenome from variants and reference sequences using vg (v.1.53.0)[107].

SV repeat element and gene expression analysis

SVs were annotated with gene regions (gene, exon, CDS and gene upstream and downstream 3,000 bp) and repeat regions (TE, low complexity and simple repeat) using Vcfanno (v.0.3.3)[108]. The alternative allele sequences of insertions were extracted and annotated for repeats using RepeatMasker (v.4.1.1)[78] with a TE library generated from EDTA (v.2.0.0)[69] and TRF (v.4.09)[79].

Five tissues (flower, fruit, stem, leaf and root) of accessions Adu, Amon, H16-5, mH8, NDH108 and ZP06 were selected for RNA-seq experiments. The raw paired-end reads were filtered and trimmed using FASTP (v.0.23.2)[56]. The trimmed reads were mapped to the reference genome (NDH108) using HISAT2 (v.2.1.0)[109] (parameter: --dta) and sorted using Samtools (v.1.10)[110]. The coverage of each transcript of genes was calculated using Bedtools (v.2.29.2)[111] (parameter: genomecov -bga -split -ibam). The expression of each transcript of genes and each gene was calculated using Stringtie (v.1.3.5)[82]. The FPKM was used to estimate the gene expressions.

Identification of regions of domestication sweeps, selected genes and SV-related genes

We used three metrics to identify selective region of domestication sweeps, including genetic diversity ratio, F_{ST} and XP-CLR[112]. We compared the genetic diversity ratio (π_{AA}/π_{AABBw}, π_{BB}/π_{AABBw}, π_{AABBw}/π_{AABBc}) in four groups in 100-kb windows sliding 20 kb using VCFtools (v.0.1.16)[59] (parameter: --window-pi 100000 --window-pi-step 20000). Windows with the top 5% of values were identified as candidate domestication-sweep regions. To identify additional domestication effects, we calculated the population fixation statistics (F_{ST}) in 100-kb windows sliding 20 kb using VCFtools (v.0.1.16)[59] (parameter: --fst-window-size100000 --fst-window-step 20000). Windows with the top 5% F_{ST} value were regarded as highly differentiated regions. We also calculated XP-CLR[112] to scan for domestication-sweep regions in 100-kb windows sliding 20 kb and identified windows with the top 5% XP-CLR values (parameter: --ld 0.95 --size 100000 --step 20000). Windows filtered with at least two metrics were considered as domestication-sweep regions.

Experiment design and RNA-seq of accessions with bacterial wilt infection

NDH108 was cultivated using the hydroponic method (14 h light and 10 h dark photoperiod). When peanut roots reached 5–6 cm, seedlings were inoculated with *Ralstonia solanacearum* (strain number 180731-1) provided by the Institute of Plant Protection, Henan Provincial Academy of Agricultural Sciences, China. We performed RNA-seq of roots in NDH108 with bacterial wilt infection, at three time points (0, 12 and 24 h) of three independent experiments in replicates. The analysis process of raw reads was the same as described in the section 'Gene expression analysis'.

Genotyping and SV-GWAS

For genotyping of resequencing accessions, we used an ensemble genotyper EVG (v.1.2.0)[113] combining GraphTyper2 (v.2.7.7)[114], GraphAligner (v.1.0.13)[115], vg (v.1.53.0)[107] and Pangenie (v.3.0.1)[116] with the graph pangenome constructed from both genome assemblies and long reads to genotype variations of SVs and indels (≥10 bp) from short-read sequencing data in tetraploid accessions. For SV-GWAS, two phenotypes (seed weight and seed length) were surveyed and the insertions

and deletions from the graph pangenome were selected. Genotype files were prepared using PLINK (v.1.90b6.24)[63]. GWAS was performed using EMMAX (v.beta-07Mar2010)[117].

Data visualization

Plots were generated using R (v.4.0.2) with the R packages ggplot2 (v.3.3.6), ggmap (v.4.0.0) (Plotting spatial data), GOplot (v.1.0.2), ggbreak (v.0.1.1) (axis breaks plots), gghalves (v.0.1.4) (half plots), qqman (v.0.1.8) (GWAS Manhattan plots), ggpubr (v.0.4.0) (significant level in box plots) Pheatmap (v.1.0.12) (heatmap; https://CRAN.R-project.org/package=pheatmap), iTOL (v.6.6)[118] (phylogenetic tree), Bandage (v.0.8.1)[119] (graph pangenome visualization) and Rcirco (v.1.2.2)[120] (genomic circos plots). Plots of the genome browser including tracks of gene annotations, repeat annotations, SVs and alignments were generated using Samplot (v.1.3.0)[121], JBrowse2 (v.2.3.2)[122] and IGV (v.2.5.3)[123]. The LD heatmap was generated using LDBlockShow (v.1.40)[124].

Phylogenetic analysis

The amino acid sequences were retrieved from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) and UniProt (https://www.uniprot.org/). The phylogenetic tree was constructed using MEGA 5.2 (ref. 125) (maximum-likelihood method), FastTree (v.2.1.11)[126] (approximately maximum-likelihood method) and modified by iTOL (v.6.6)[118]. The multiple sequence alignment was performed using MUSCLE (v.5.1)[95]. The protein structure was predicted using AlphaFold (v.2.3)[127] and displayed using PyMol (v.2.3.4) (http://www.pymol.org).

Subcellular localization analysis and transient overexpression

The CDSs of *AhARF2-2*^{HapI} and *AhARF2-2*^{HapII} (*NDH08G29450*) were cloned into a pCAMBIA1300-GFP vector using the Hieff Clone Universal One Step Cloning Kit (catalogue number 10911ES20, Yeasen). The recombinant plasmids pCAMBIA1300-*HapI/HapII*-GFP (35S:HapI/HapII-GFP), marker vector and empty control vector were then transformed into *Agrobacterium tumefaciens* strain GV3101 (AC1001, WEIDI). *Agrobacterium* mediated transient expression assays in tobacco (*Nicotiana tabacum*) and peanut according to our previous work[11]. Fluorescence images of transiently infected tobacco leaf subepidermal cells were captured utilizing laser scanning confocal microscopy (catalogue number LSM710, Carl Zeiss).

Hormone treatments and RT–qPCR analysis

Two-week-old seedlings (mH8 and ZP06), cultivated on plates, were subjected to treatment through spraying with 1 µM NAA. Samples were collected at 0, 2, 4, 6 and 12 h following the treatment with NAA. Total RNA was extracted using FastPure Universal Plant Total RNA Isolation Kit (catalogue number RC411, Vazyme) and *TransScript* One-Step gDNA Removal and cDNA Synthesis SuperMix Kit (catalogue number AT311-02, TransGen Biotech). Gene expression levels were assayed using the PerfectStart Green qPCR SuperMix Kit (catalogue number AQ601, TransGen Biotech) in conjunction with the Bio-Rad CFX96 real-time PCR system. The $2^{-\Delta\Delta CT}$ method was used to calculate the relative expression levels of each gene[128]. *Ahactin*7 (XM_025826875) and *Atactin*2 (U37281.2) were used as the internal control genes. Primers were designed using Primer 3 (v.4.1.0)[129]. Cytokinin measurements were performed by Wuhan Metware Biotechnology Co., Ltd, utilizing seeds from the Amon and ZP06 varieties, approximately 15 days following the peg's penetration into the soil.

Yeast one-hybrid assays

The *AhGRF5* (*Arahy.V829EQ*) promoter was integrated into the pAbAi vector. The recombinant plasmid *pAbAi-AhGRF5* was digested with the restriction enzyme BstBI (catalogue number R0519V, New England Biolabs) and transferred into the Y1H yeast strain and tested on SD/-Ura medium (catalogue number PM2271, Coolaber) with different concentrations of aureobasidin A. The CDS of *AhARF2-2*^{HapI} and *AhARF2-2*^{HapII} were cloned into the pGADT7 and transferred into the Y1H yeast strain, containing pAbAi–*AhGRF5* plasmid, tested on SD/-Leu medium (catalogue number PM2201, Coolaber) with 200 ng ml^{-1} of aureobasidin A.

Bimolecular fluorescence complementation

The CDSs of *AhARF2-2*^{HapI} and *AhARF2-2*^{HapII} were cloned into the N-terminal half of yellow fluorescent protein (nYFP). The CDSs of *AhIAA13* and *AhTPL* were cloned into the C-terminal half of YFP (cYFP) to generate *HapI-nYFP*, *AhIAA13-cYFP*, *HapII-nYFP* and *AhTPL-cYFP* vectors. The appropriate pairs of constructs (*AhTPL-cYFP* and *HapI-nYFP*; *AhTPL-cYFP* and *HapII-nYFP*; *AhIAA13-cYFP* and *HapI-nYFP*; *AhIAA13-cYFP* and *HapII-nYFP*) were infiltrated into *N. benthamiana* leaves via *Agrobacterium*-mediated transient infiltration and detected by laser scanning confocal microscopy (catalogue number LSM710, Carl Zeiss).

Yeast two-hybrid assays

AhIAA13 (*Arahy.MDB4JZ*) and *AhTPL* (*Arahy.UFRA39*) CDSs were cloned into pGBKT7 vector (bait vector). These recombinant vectors were transformed into the Y2H yeast strain. The bait vectors pGBKT7-53 (BD-T7-53) and pGBKT7-Lam (BD-T7-Lam), serving as positive and negative controls respectively, were cotransformed with the prey vector pGADT7-T. Yeast cells and control cells, carrying recombinant plasmids, were grown on deficient medium: lacking tryptophan and leucine (SD/-Trp/-Leu; catalogue number PM2221, Coolaber) and leucine, tryptophan histidine and adenine (SD/-Leu/-Trp/-His/-Ade; catalogue number PM2221, Coolaber), respectively.

Dual-luciferase reporter assay

The CDSs of *AhTPL*, *AhARF2-2*^{HapI} and *AhARF2-2*^{HapII} were cloned into pCAMBIA1300-GFP vector to construct the pCAMBIA1300-*TPL*-GFP (35S:TPL), 35S:HapI and 35S:HapII as effector plasmids. The *AhGRF5* promoter (850 bp) was inserted into pGreenII 0800-LUC to construct the *GRF5pro:LUC* reporter plasmid. These vectors were individually transformed into GV3101 (harboring the pSoup plasmid). The pairs of constructs (35S:GFP and *GRF5pro:LUC*; 35S:HapI and *GRF5pro:LUC*; 35S:HapII and *GRF5pro:LUC*; 35S:HapI, 35S:TPL and *GRF5pro:LUC*; 35S:HapII, 35S:TPL and *GRF5pro:LUC*) were infiltrated into *N. benthamiana* leaves using the method reported previously[130]. LUC activity was detected using a low-light cooled CCD imaging apparatus (Tanon 5200). LUC or Renilla luciferase was assayed using the Double-Luciferase Reporter Assay Kit (catalogue number FR201, TransGen Biotech) with dual LUC assay reagents (Promega). Primers used for the LUC assay are listed in Supplementary Table 16.

Arabidopsis transformation

Agrobacterium tumefaciens GV3101 containing the recombinant vectors 35S:HapI-GFP or 35S:HapII-GFP was used to transform *Arabidopsis* (Col-0) using the floral dipping method[131] Transgenic *Arabidopsis* plants were selected on MS medium supplemented with 30 mg l^{-1} hygromycin and their identity was confirmed through PCR. Subsequently, the T_3 generation of transgenic plants was utilized for phenotypic analysis of the candidate gene. Seed sizes were photographed with an anatomical microscope and were tested using the Ween SC-G automated test analysis system (WSeen). Primers used for gene cloning are listed in Supplementary Table 16.

Statistical analyses

The statistical analyses were mostly performed in R (v.4.0.2). For biochemical and molecular biology analysis, at least three individuals were mixed in each sample with three biological replicates. Standard deviations and *P* values were calculated using two-tailed Student's *t*-test and analysis of variance.

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

Data availability
The eight genomic sequences for pangenome and all the 269 genomic sequence data for GWAS analysis have been deposited in the National Genomics Data Center (NGDC) database under BioProject numbers PRJCA029798 and PRJCA029800 (GSA numbers CRA019399 and CRA019260). The published transcriptomic datasets for candidate gene expression analysis can be downloaded from the NGDC under accession number PRJCA029802 (GSA number CRA019264). The transcriptomic datasets for gene expression analysis on accessions with bacterial wilt infection can be downloaded from the NGDC under accession number PRJCA030060 (GSA numbers CRA019399 and CRA020567). The result data has been mirrored deposited at https://cgm.sjtu.edu.cn/PeanutPan/index.html. Source data are provided with this paper.

Code availability
The codes for this study are available via Zenodo at https://doi.org/10.5281/zenodo.15003999 (ref. 132) and via GitHub at https://github.com/SJTUCGM/PeanutPan.

References

56. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
59. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
60. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
61. Edgardo M. O. Vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. *Zenodo* https://doi.org/10.5281/zenodo.2540861 (2019).
62. Felsenstein, J. PHYLIP—phylogeny inference package (Version 3.2). *Q. Rev. Biol.* **64**, 539–541 (1989).
63. 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).
64. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
65. Vaser, R. et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2013).
66. Hu, J. et al. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
68. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
69. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
70. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. In *Proc. IEEE/ACM Transactions on Computational Biology and Bioinformatics* Vol. 10 645–656 (IEEE, 2013).
71. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
72. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
73. Shi, J. & Liang, C. Generic Repeat Finder: a high-sensitivity tool for genome-wide de novo repeat detection. *Plant Physiol.* **180**, 1803–1815 (2019).
74. Su, W., Gu, X. & Peterson, T. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol. Plant* **12**, 447–460 (2019).
75. Xiong, W. et al. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111**, 10263–10268 (2014).
76. Zhang, R. G. et al. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).
77. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
78. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **4**, 4.10.1–4.10.14 (2009).
79. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
80. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016).
81. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
82. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
83. Campbell, M. A. et al. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics.* **7**, 327 (2006).
84. Stanke, M. et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
85. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
86. Urasaki, N. et al. Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res.* **24**, 51–58 (2017).
87. Manni, M. et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
88. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
89. Törönen, P., Medlar, A. & Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).
90. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
91. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
92. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).
93. Li, X. et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in *Brassica oleracea*. *Nat. Genet.* **56**, 517–529 (2024).

94. Zhang, Z. et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012).
95. Edgar, R. C. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
96. Zhang, Z. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics Proteomics Bioinformatics* **20**, 536–540 (2022).
97. Posada, D. Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Curr. Protoc. Bioinform.* **6**, 6.5.1–6.5.14 (2003).
98. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
99. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
100. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
101. Kiełbasa, S. M. et al. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
102. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
103. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
104. English, A. C. et al. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
105. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
106. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
107. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
108. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.* **17**, 118 (2016).
109. Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
110. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
111. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
112. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
113. Du, Z. Z., He, J. B. & Jiao, W. B. A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome Biol.* **25**, 91 (2024).
114. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
115. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
116. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
117. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
118. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
119. Wick, R. R. et al. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
120. Zhang, H., Meltzer, P. & Davis, S. RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics* **14**, 244 (2013).
121. Belyeu, J. R. et al. Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol.* **22**, 161 (2021).
122. Diesh, C. et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol.* **24**, 74 (2023).
123. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
124. Dong, S. S. et al. LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief Bioinform.* **22**, bbaa227 (2021).
125. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
126. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
127. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
128. Pfafl, M. W. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res.* **29**, e45 (2001).
129. Untergasser, A. et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
130. Zhang, Y. et al. The circadian-controlled PIF8-BBX28 module regulates petal senescence in rose flowers by governing mitochondrial ROS homeostasis at night. *Plant Cell* **33**, 2716–2735 (2021).
131. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
132. Xue, H. PeanutPAN: pan-genome for peanut genomics study (1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.15003999 (2025).

Acknowledgements

The computations in this paper were run on the π 2.0 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University. This work was supported by grants from the Key Program of National Natural Science Foundation of China (NSFC)-Henan United Fund (grant number U22A20475), Key Scientific and Technological Project of Henan Province (grant numbers 221111110500, 222301420026 and HARS-22-05-G1), National Natural Science Foundation of China (grant number 32170643), Natural Science Foundation of Shanghai (grant numbers 20ZR1428200, and 22ZR1433600), Shanghai Key Program of Computational Biology (grant number 23JS1400800) and National key R&D program (grant number 2023YFF1001600). R.K.V. acknowledges the start-up grant from the Food Futures Institute of Murdoch University, Australia. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.Y., C.W. and R.K.V. conceived and designed the study. H.X., G.L., A.C., X.D. and H.L. performed data analysis. Kunkun Zhao, Y.F., Z.C., D.Q., L.Z., A.C. and Kai Zhao performed the experiments. L.Z., D.Q. and Kai Zhao prepared the samples and reagents. Y.F., Z.C., G.L., Kai Zhao, R.R., F.G., Z.L. and X.M. measured the agronomic traits. D.Y., C.W., R.K.V., H.X. and Kunkun Zhao wrote the paper. D.Y., C.W., R.K.V. and S.W. revised the paper. All authors read and approved the final paper.

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/s41588-025-02170-w.

Correspondence and requests for materials should be addressed to Rajeev K. Varshney, Chaochun Wei or Dongmei Yin.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Rajeev K. Varshney , Chaochun Wei and Dongmei Yin

Last updated by author(s): Feb 15, 2025

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	Data were sequenced from PacBio, Nanopore and Illumina platform.
Data analysis	All the softwares used in this study are cited in the manuscript and listed as follows: FASTP (v0.23.2), BWA (v0.7.17), Bcftools (v1.10.2), Vcftools (v0.1.16), SnpEff (v5.1), Vcf2phylip (v2.8), PHYLIP (v3.697), PLINK (v1.90b6.24), Admixture (v1.3.0), Racon (v1.3.1), Nextpolish (v1.3.1), Bowtie2 (v2.3.2), LACHESIS, EDTA (v2.0.0), LTR_FINDER (v1.07), LTR_retriever (v2.9.0), Generic Repeat Finder (v1.0), TIR-learner (v2.5), HelitronScanner (v1.0), TEsorter (v1.3.0), RepeatModeler (v2.0.2), RepeatMasker (v 4.1.1), TRF (v4.09), GeMoMa (v1.6.1), STAR (v2.7.3a), Stringtie (v1.3.5), PASA (v2.3.3), Augustus (v3.3.1), EVidenceModeler (EVM) v1.1.1, TransposonPSI, BUSCO (v5.3.2), InterProScan (v5.55-88.0), PANNZER2 (v15.12.2020), ClusterProfiler (v3.16.1), Orthofinder (v2.5.4), Power model, ParaAT (v2.0), MUSCLE (v5.1), KaKs_Calculator, D-Geneies (v1.4), Minimap2 (v2.22), MCscan (in JCVI) v1.2.7, Last (v1418), Svim-asm (v1.0.2), CuteSV (v1.0.13), CuteSV (v2.0.3), Truvari (v3.4.0), SURVIVOR (v1.0.7), Minigraph-cactus (v2.7), vg (v1.53.0), Vcfanno (v0.3.3), HISAT2 (v2.1.0), Samtools (v1.10), Bedtools (v2.29.2), Pangenie (v3.0.1), EMMAX (vbeta-07Mar2010), iTOL (v6.6), Bandage (v0.8.1), Rcirco (v1.2.2), Samplot (v1.3.0), JBrowse2 (v2.3.2), IGV (v2.5.3), LDBlockShow (v1.40), MEGA 5.2, FastTree (v2.1.11), Primer 3 (v4.1.0), EVG (v1.2.0), GraphTyper2 (v2.7.7), GraphAligner (v1.0.13), AlphaFold (v2.3), PyMol (v2.3.4)

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 eight genomic sequences for pan-genome and all the 269 genomic sequence data for GWAS analysis have been deposited in the National Genomics Data Center (NGDC) database under BioProject number PRJCA029798 and PRJCA029800. The published transcriptomic datasets for candidate gene expression analysis can be downloaded from the NGDC under accession numbers PRJCA029802. The transcriptomic datasets for gene expression analysis on accessions with bacterial wilt infection can be downloaded from the NGDC under accession numbers PRJCA030060.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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

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	The germplasm used in this study included 32 two diploid wild species, 8 tetraploid wild species, 155 tetraploid cultivated landrace, 67 tetraploid improved accessions, and 7 unclassified accessions. There are a total of 296 germplasm lines, originating from over 30 countries worldwide and encompassing major peanut-growing regions. This represents one of the most comprehensive research collections to date.
Data exclusions	No data exclusions. Sequencing data was quality filtered, as described in manuscript.
Replication	All the experiments were performed using independent biological replicates as indicated in the manuscript, figure and table legends, and supplementary information data.
Randomization	A randomized complete block design was used in planting for phenotype data collection.
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

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*