nature genetics

Article

https://doi.org/10.1038/s41588-024-01722-w

Differential selection of yield and quality traits has shaped genomic signatures of cowpea domestication and improvement

Received: 30 November 2022

Accepted: 19 March 2024

Published online: 22 April 2024

Check for updates

Xinyi Wu[1,7], Zhongyuan Hu[2,7], Yan Zhang[3], Mao Li[1,4], Nanqiao Liao[2], Junyang Dong[1,4], Baogen Wang[1,4], Jian Wu[1,4], Xiaohua Wu[1,4], Ying Wang[1,4], Jian Wang[1,4], Zhongfu Lu[1,4], Yi Yang[3], Yuyan Sun[1,4], Wenqi Dong[1,4], Mingfang Zhang[2,5,6] ✉ & Guojing Li[1,4] ✉

Cowpeas (tropical legumes) are important in ensuring food and nutritional security in developing countries, especially in sub-Saharan Africa. Herein, we report two high-quality genome assemblies of grain and vegetable cowpeas and we re-sequenced 344 accessions to characterize the genomic variations landscape. We identified 39 loci for ten important agronomic traits and more than 541 potential loci that underwent selection during cowpea domestication and improvement. In particular, the synchronous selections of the pod-shattering loci and their neighboring stress-relevant loci probably led to the enhancement of pod-shattering resistance and the compromise of stress resistance during the domestication from grain to vegetable cowpeas. Moreover, differential selections on multiple loci associated with pod length, grain number per pod, seed weight, pod and seed soluble sugars, and seed crude proteins shaped the yield and quality diversity in cowpeas. Our findings provide genomic insights into cowpea domestication and improvement footprints, enabling further genome-informed cultivar improvement of cowpeas.

Legumes are crops with a high potential to provide balanced nutrition to the human diet and sustain food and nutritional security in developing regions, particularly in African countries. Cowpea (*Vigna unguiculata* L. Walp., $2n = 2x = 22$), which is endemic to sub-Saharan Africa, is cultivated as a grain, vegetable or livestock feed worldwide[1], and the wild subspecies *V. unguiculata* ssp. *dekindtiana* var. *spontanea* is considered its progenitor[2–4]. The domesticated cowpea has formed two major subspecies: the grain cowpea (*V. unguiculata* L. Walp. ssp. *unguiculata*) in Africa and the vegetable or garden cowpea (*V. unguiculata* L. Walp. ssp. *sesquipedalis*) in Asia. Global annual production of grain cowpeas is ~8.9 million tonnes, 85% of which is produced in West Africa (FAOSTAT, 2020; https://www.fao.org/faostat/en/#data/QCL). This subspecies provides an excellent source of starch, dietary protein, fiber and micronutrients in developing countries as an important

[1]State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products, Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou, P. R. China. [2]Laboratory of Vegetable Germplasm Innovation and Molecular Breeding, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, P. R. China. [3]Guangdong Key Laboratory for New Technology Research of Vegetables, Vegetable Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, P. R. China. [4]Key Laboratory of Vegetable Legumes Germplasm Enhancement and Molecular Breeding in Southern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Zhejiang Academy of Agricultural Sciences, Hangzhou, P. R. China. [5]Hainan Institute of Zhejiang University, Yazhou Bay Science and Technology City, Sanya, P. R. China. [6]Key Laboratory of Horticultural Plant Growth and Development, Ministry of Agriculture and Rural Affairs, Hangzhou, P. R. China. [7]These authors contributed equally: Xinyi Wu, Zhongyuan Hu. ✉e-mail: mfzhang@zju.edu.cn; ligj@zaas.ac.cn

cereal substitute for the human population or as feed for livestock. Vegetable cowpea, also known as asparagus bean or yardlong bean, is cultivated for its long immature pods (40–100 cm in length) and is often consumed as a vegetable[2,5]. The vegetable cowpea is predominant in East and Southeast Asia and is ranked among the top ten Asian vegetables owing to its high tolerance to heat and drought as well as diverse nutrition enrichment[6].

Grain and vegetable cowpeas vary greatly in many important agronomic traits such as pod length (PL), grain number per pod (GNP) and nutrition. Trait differentiation in cowpea subspecies is postulated to be caused by human selection for favorable usage[7,8]. Previous studies have reported some quantitative trait loci (QTLs) controlling cowpea domestication and improvement traits, such as pod shattering (PS), PL, pod quality and seed size[9–16]. However, the genome-wide genetic variations associated with subspecies divergence remain largely unknown. Although a high-quality grain cowpea genome has been released[17], the paucity of information regarding the vegetable cowpea genome is still hindering the elucidation of the genomic basis and selection signatures of key traits for shaping the subspecies' differentiation and improvement.

In this study, we report two chromosome-scale genome assemblies of grain and vegetable cowpeas by combining PacBio, chromatin conformation capture (Hi-C) and Illumina sequencing technology toolkits. A genetic diversity panel encompassing 344 accessions of landraces, wild species and breeding lines was re-sequenced to clarify the phylogenomic evolution of cowpeas. Furthermore, genome-wide association studies (GWAS) were performed to identify the genes that are responsible for crucial agronomic traits. Our study reveals a global landscape of genome structural variations (SVs) between two subspecies and provides insights into cowpea domestication and improvement under selection.

Results

Genome assemblies and gene annotations

One vegetable-landrace cowpea (G98) with a super-long-pod and one grain cowpea (G323) with strong disease resistance were selected for de novo genome sequencing (Fig. 1a,b). Through *k*-mer analysis, the genome sizes of G98 and G323 were estimated to be 623.16 Mb and 597.42 Mb, respectively (Supplementary Table 1). Next, draft assemblies of G98 (632.54 Mb) and G323 (593.26 Mb) were constructed using PacBio sequences (Supplementary Tables 2–5). Finally, the G98 and G323 genomes were adjusted to 568.24 Mb (scaffold N50 = 49.41 Mb) and 552.66 Mb (scaffold N50 = 49.35 Mb), respectively, using the Hi-C approach (Fig. 1c,d and Supplementary Tables 5–7).

Multiple genome assessments validated the high quality of the two genome assemblies. Firstly, 99.19% and 99.69% of Illumina reads were mapped to the G98 and G323 assemblies, respectively (Supplementary Table 8). Secondly, over 98% of the Core Eukaryotic Gene Mapping Approach (CEGMA) core eukaryotic genes and 95% of the Benchmarking Universal Single-Copy Orthologue (BUSCO) genes could be properly mapped to the two assemblies (Supplementary Tables 9 and 10). In addition, Merqury analysis showed the high quality of G98 (44.48) and G323 (47.17), which is similar to the improved pea genome (ZW6; 44.5)[18]. Using the IT97K-499-35 and A147 genomes as references[17,19], the nucleotide accuracy rates of our two assemblies reached 99.6% and 97.74%, respectively.

We annotated 33,159 and 33,222 genes in the G98 and G323 genomes, respectively (Supplementary Table 11); 89.63% and 90.17% of the transcriptome data could be mapped onto the predicted genes in the G98 and G323 genomes, indicating the high fidelity of the gene predictions. Furthermore, we also annotated 8,087 transfer RNA, 15,077 ribosomal RNA, 83 microRNA and 325 pseudogenes throughout the G98 genome assembly, as well as 5,119 transfer RNA, 8,737 ribosomal RNA, 91 microRNA and 286 pseudogenes in the G323 genome.

In total, 56.97% and 55.25% of the G98 and G323 assemblies, respectively, were annotated as repetitive sequences (Supplementary Table 12), of which the transposable elements content was 44.56% in G98 and 44.57% in G323. Class I retrotransposons were the most abundant transposable elements in both genomes, of which the *Gypsy* long-terminal repeat retrotransposons (LTRs) were the leading type in the G98 (17.19%) and G323 (18.72%) genomes. Class II DNA transposons comprised 9.49% and 8.82% of the G98 and G323 genomes, respectively. Among them, hAT was the major transposon (Supplementary Table 12).

Phylogenomic relationships and SVs

A maximum-likelihood phylogenetic tree of 25 plant genomes revealed that the cowpea is closely related to the adzuki bean (*V. angularis*) and mung bean (*V. radiata*) and that they apparently diverged about 6–27 million years ago (Fig. 2a and Supplementary Table 13), which is consistent with previous reports[19,20]. Moreover, 512 and 396 gene families displayed significant expansions in the G98 and G323 genomes. The expanded genes in G98 were significantly enriched in the glycosphingolipid metabolism pathway, which is involved in membrane organization[21–23]. As pod elongation is largely caused by cell division[24,25], these expanded genes possibly contribute to the longer pods in the vegetable cowpea. Conversely, the expanded genes in G323 were significantly enriched in energy production and conversion pathways such as amino sugar and nucleotide sugar metabolism, glucosinolate gamma-glutamyl hydrolase and galactose metabolism (Fig. 2b,c and Supplementary Table 14), which might relate to the higher carbohydrate accumulation and defense response in the grain cowpea.

We also performed a comparative genome analysis on the two genomes using G323 as a reference. A total of 2,219,947 single nucleotide polymorphisms (SNPs) were identified in the G98 genome, of which 38,420 SNPs may cause changes in gene functions (Supplementary Table 15). Meanwhile, a total of 407,119 insertions and deletions (InDels; 2–49 bp) were identified in G98, 62.50% of which could cause protein encoding alterations (Supplementary Table 16). In addition, 13,541 SVs (≥50 bp) were identified in G98, including 963 translocation variations (TRANS), 74 inversion variations (INVs) and 3,701 duplications (DUPs), 7,112 presence–absence variations (PAVs; ≥50 bp) and 1,691 gene copy number variations (CNVs) (Supplementary Table 17).

Notably, we found five large SV regions (>1 Mb) (Fig. 2d–f). Chromosome 1 contained a 7.5 Mb INV (Chr01: 20,118,943 ~ 27,655,522) and multiple adjacent TRANS (Chr01: 39,164,314 ~ 40,968,319), which harbors 61 and 31 genes, respectively (Fig. 2d). Chromosome 6 contained a 4.73 Mb INV region (Chr06: 42,392 ~ 4,775,193) and a 5.14 Mb region (Chr06:13,882,828 ~ 19,025,451) containing two INVs, two DUPs and two TRANS. These two regions involve 42 and 52 gene variations, respectively (Fig. 2e). Chromosome 10 contained the largest number of INVs (Chr10: 13,491,13 ~ 31,515,899), comprising 224 genes (Fig. 2f). A total of 13 other SV regions were detected on the rest of the chromosomes (Supplementary Fig. 1).

Population structure and divergence of cowpea subspecies

In total, 344 cowpea accessions collected from various geographic regions were selected for whole-genome re-sequencing (Supplementary Table 18). We identified 7,982,974 SNPs and 1,874,358 InDels, with an average of 12.63 SNPs and 2.97 InDels per kb. After filtration, 1,262,497 high-confidence SNPs were selected for population structure analysis. Principal component analysis (PCA) divided the 344 accessions into two clusters (Fig. 3a), which were mainly formed by grain (cluster I) and vegetable cowpeas (cluster II). Using the common bean as an outgroup, the phylogenetic tree of the 344 accessions revealed three groups centered on grain cowpea (G), vegetable cowpea landraces (VL) and vegetable cowpea cultivars (VC) (Fig. 3b). Group I corresponded to cluster I in the PCA, including the 2 wild cowpeas, 69 grain cowpeas, 5 vegetable cowpeas and 1 uncertain usage. Group II constituted 147 vegetable cowpeas, most of which (127) belong to landraces. Group III included 2 grain cowpeas and 92 vegetable cowpeas, 61.96% (57) of which were cultivars or breeding lines. Population structural analysis

https://doi.org/10.1038/s41588-024-01722-w

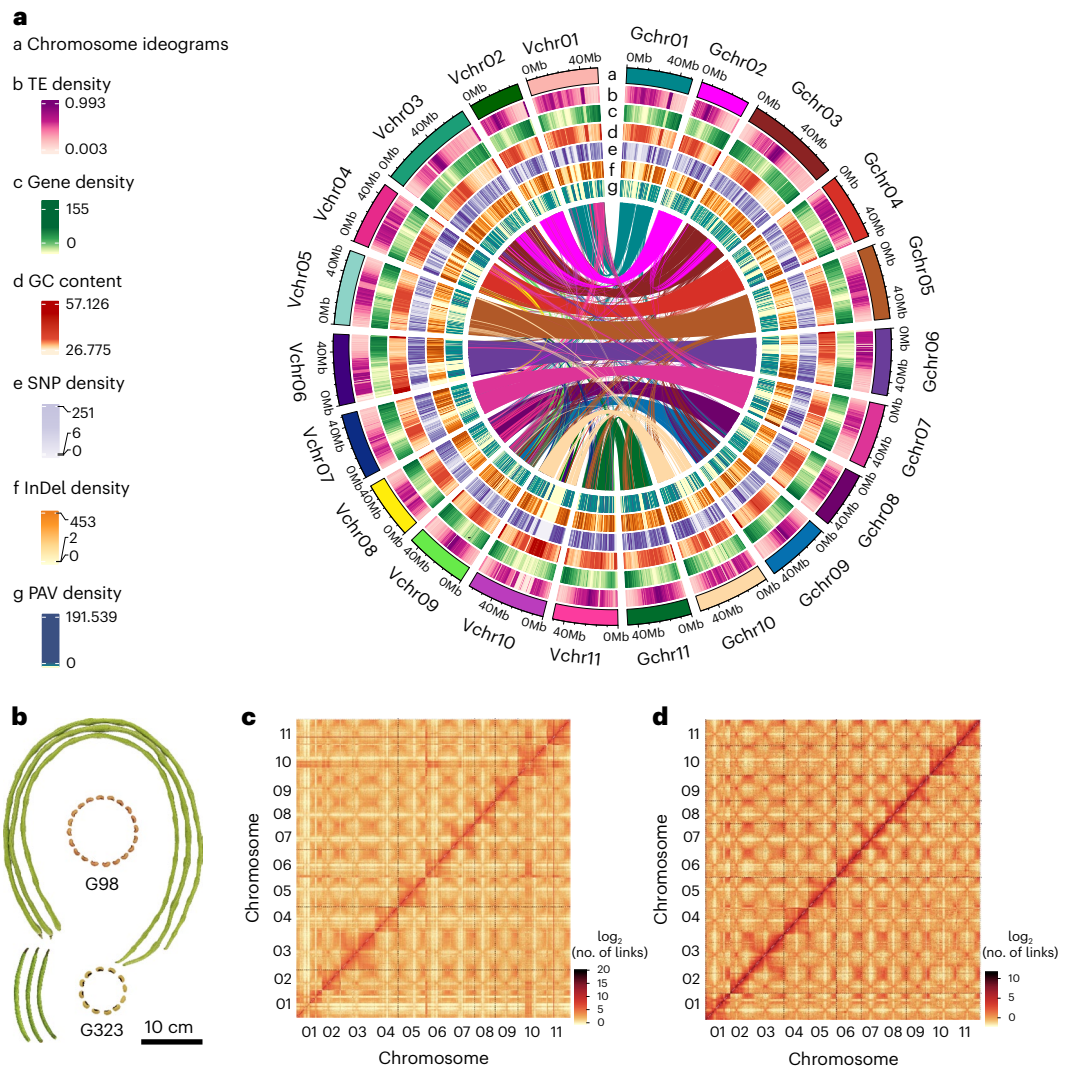


Fig. 1 | High-quality genome assembly of vegetable cowpea G98 and grain cowpea G323. a, Genome features of the two genomes. TE, transposable element **b**, The immature pod and number of grains per pod for G98 and G323. The scale bar provides a size comparison. **c**,**d**, Whole-genome Hi-C heat map for G98 (**c**) and G323 (**d**).

also supported the clades classification of the PCA and phylogenetic tree (Fig. 3b). When $k = 2$, two clusters were formed, corresponding to the grain and vegetable cowpeas in the PCA. When $k = 3$, cluster II was classified into two subclades divided among the landraces and cultivars or breeding lines (Fig. 3b).

Nucleotide diversity (π) and population divergence (fixation index, F_{ST}) in the three subpopulations or groups with multiple accessions were estimated (Fig. 3c). The G group ($\pi = 0.0007$) had much higher nucleotide diversity than the VL ($\pi = 0.00047$) and VC ($\pi = 0.00024$) groups. The F_{ST} values for G–VC (0.1903) and G–VL (0.0924) were higher than VC–VL (0.0498). Moreover, linkage disequilibrium decayed faster in the G group than in the VL and VC groups, indicating a higher degree of genetic recombination in grain cowpeas (Fig. 3d).

Genomic signatures of domestication and improvement

In many crops, human selection for specific traits reflects plant domestication and leads to different edible types such as grain and vegetable usage in legume crops as well as oilseed and vegetable *Brassica juncea*[26–28]. To investigate how natural or artificial selection has affected cowpea differentiation, we searched for selection signatures in the cowpea genome by comparing the selective sweeps among three subgroups. Using cross-population composite likelihood ratio test (XP-CLR) analyses, we identified a total of 189, 156 and 196 potential selective loci in G versus VL, VL versus VC and G versus VC, respectively (Supplementary Table 19). A total of 3,212 and 2,972 genes located in the selective regions are associated with differentiation in G versus VL and G versus VC, respectively, while 2,650 genes are in the regions associated with improvement in VL versus VC (Fig. 4a and Supplementary Table 19). Among them, 239 genes were identified in all three pairwise groups, implying that these genes might have been exposed to long-term and continuous selection during cowpea domestication and improvement or were unintentionally selected owing to the hitchhiking effect of their neighboring loci. Meanwhile, we also identified numerous putative selective sweeps by F_{ST} and π values, and multiple sweeps overlapped with the selected regions in XP-CLR (Supplementary Fig. 2 and Tables 20 and 21).

In total, 18 signals associated with PS, pod total starch content (PTS), seed total starch content (STS), GNP, pod soluble sugar content (PSS) and PL were co-identified by both selective sweep detection and GWAS analyses. We also found that over 70 known loci for drought tolerance, disease resistance and agronomic traits overlapped with the selective regions (Fig. 4a, Supplementary Fig. 2 and Supplementary Table 22). Seventeen signals for PS, one of the most conspicuous domestication syndrome-related traits, were identified in SNP-GWAS, and eight of them were also detected in InDel-GWAS (Fig. 5a and Supplementary Tables 23–25). Among them, PS-3.2, which overlapped

https://doi.org/10.1038/s41588-024-01722-w

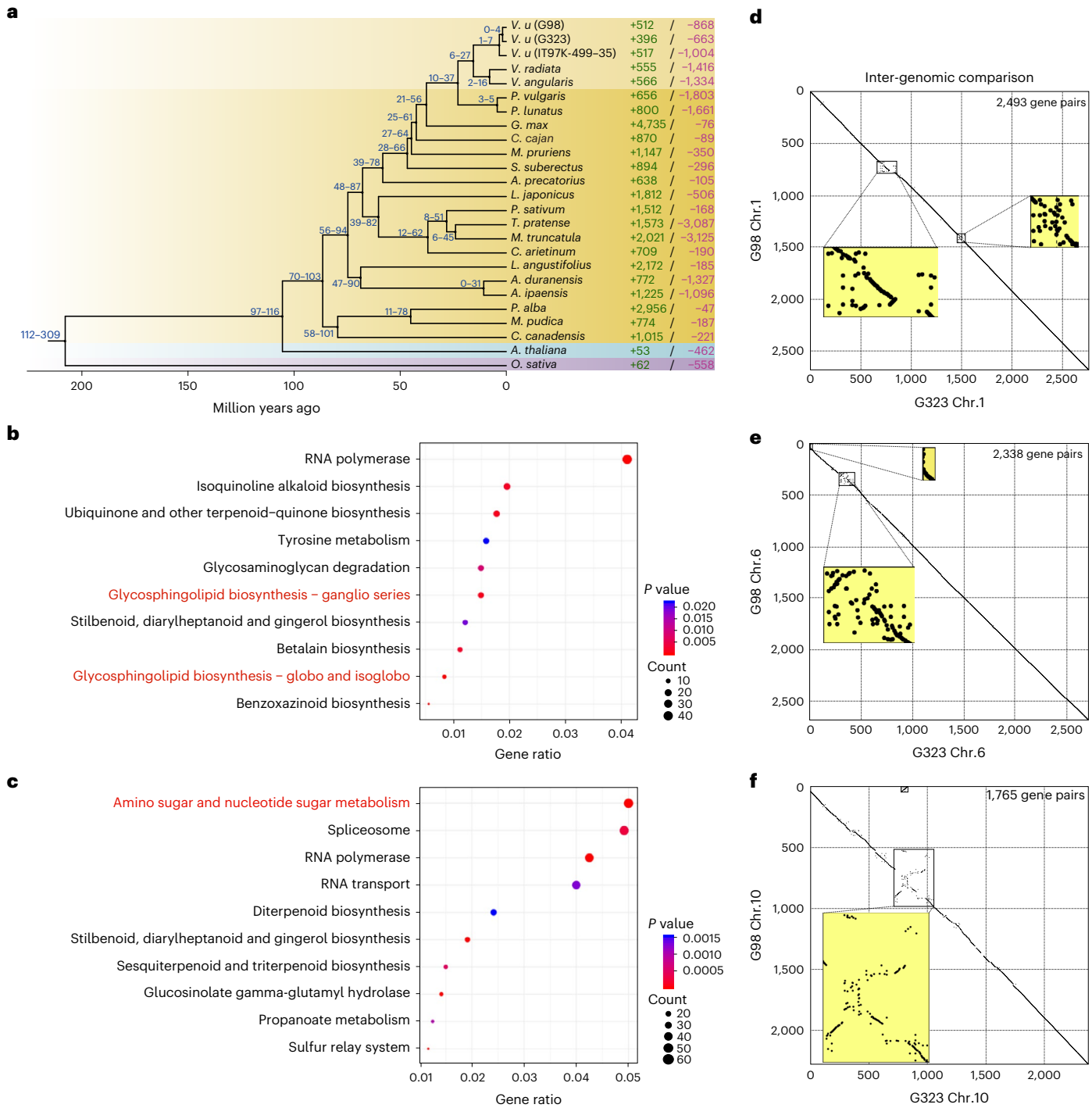


Fig. 2 | Phylogenetic analysis and genome structure variations of cowpeas. a, The phylogenetic tree of 25 plant species and the evolution of the gene families. Numerical values beside each node show the estimated divergence time of each node. The numbers in the right panel indicated the expanded (green) and contracted (purple) gene families. **b**,**c**, The top ten enriched KEGG terms of the expanded genes in G98 (**b**) and G323 (**c**). The enrichment test was performed based on a hypergeometric distribution. *P* values were false discovery rate (FDR)-adjusted. **d**–**f**, The large SVs (>1 Mb) between the two genomes on chromosome 1 (**d**), 6 (**e**) and 10 (**f**).

with the known *CPshat3* (refs. 9,29), contains an MYB transcription factor (VuG9803G034000). This gene is homologous to the cell wall biosynthesis-related *AtMYB50* that affects shattering habits in different species[30]. Two haplotypes of PS-3.2 were observed in the 344 accessions, with HapII accessions have a higher shattering ratio (48%) (Fig. 4b). Similarly, E3 ubiquitin-protein ligase gene (*VuPS2*, VuG9801G016510; Fig. 5b), zinc finger CCCH domain protein (*VuPS6*, VuG9804G001420; Fig. 5c)[31], MYB transcription factor (*VuPS8*, VuG9806G012680; Fig. 5d) and three uncharacterized proteins (*VuPS4*, VuG9803G014760; *VuPS7*, VuG9805G024700; *VuPS10*, VuG9808G017090; Fig. 5e–g) were considered as the related genes for other PS signals. Polymorphisms of these six PS-related loci were identified in the 344 accessions, which resulted in different haplotypes with significant variations in PS resistance (Fig. 5 and Supplementary Tables 23–25). Three other PS loci (PS-3.3, PS-4.2, PS-10.3) were detected in InDel-GWAS. However, only slight differences in the shattering ratio were observed between different haplotypes of

https://doi.org/10.1038/s41588-024-01722-w

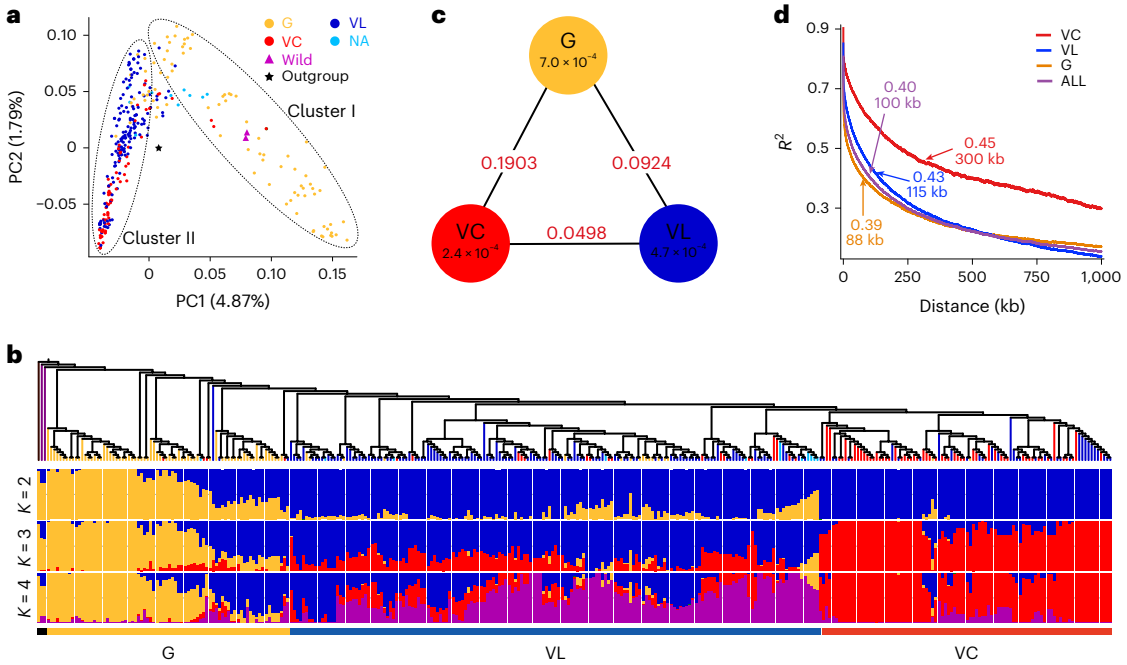


Fig. 3 | Population structure and genomic diversity of 344 cowpea accessions. **a**, PCA of the 344 re-sequenced cowpea accessions. NA, uncertain accessions. **b**, Phylogenetic tree and model-based clustering ($K = 2$–4) of the 344 sequenced accessions. **c**, Nucleotide diversity (π; numbers in the circles) and population divergence (F_{ST}; numbers between the circles) across the three subpopulations. **d**, Linkage disequilibrium decay analysis in the three subpopulations.

the three loci (Supplementary Table 25). Interestingly, 6 of 14 PS-related loci were found in the domestication sweeps of G versus VL (Fig. 4a and Supplementary Fig. 2), which partially explains the differences in PS resistance between these 2 subspecies. Surprisingly, five PS loci were also observed in improvement sweeps in VL versus VC (Fig. 4a and Supplementary Fig. 2), although all accessions of both subpopulations exhibit strong PS resistance. Most PS-related candidate genes showed different expression patterns between the pods of G-type and V-type cowpeas. For instance, PS-1.1-related, PS-2.1-related and PS-6.1-related genes (VuG9801G008190, VuG9802G010740 and VuG9806G012680) had higher expression in pods of non-shattering V-type accessions at anthesis, but with fairly low expression in G-type cowpeas with higher shattering rates, indicating that these genes might be negative regulators of PS. PS-1.2-related and PS-3.2-related genes (VuG9801G016510 and VuG9803G034000) showed the highest expression level in the later stages of pod development in G-type accessions, suggesting a possible positive correlation between gene transcriptions and PS (Supplementary Fig. 3a,b and Supplementary Table 26).

One PTS-1.1 locus was co-identified by selective sweep and GWAS analyses that contains two ribokinase-like genes (VuG9801G016420 and VuG9801G016430) acting in native starch granule degradation[32]. VuG9801G016420 is likely the candidate gene (*VuPTS1*) contributing to the pod starch content variation among accessions, as two haplotypes of this gene led to significantly different PTS phenotypes (Fig. 4c). VuG9801G016430 is less likely to be the candidate for PTS-1.1, as no sequence polymorphism could be observed. Interestingly, the sustained higher expression of both genes in V-type cowpeas might also contribute to their final low PTS phenotype (Supplementary Fig. 3b and Supplementary Table 26). STS-1.1 contained a soluble metal binding protein encoding gene VuG9801G016790 (*VuSTS1*), whose ortholog is specifically expressed in companion cells of the phloem and involved in starch accumulation in *Arabidopsis*[33]. Three haplotypes of *VuSTS1* were identified in the different accessions, and in one test, *VuSTS1*-HapIII was found to result in lower seed starch content than the others (Fig. 4d). In addition, one locus significantly associated with GNP (*GNP-4.1*), an important factor affecting grain yield in cowpea, was identified by all analyses. The putative gene VuG9804G003460 (*VuGNP1*) in this region encodes a KNOX2 protein, which was found to mediate panicle length and spikelet number in rice[34]. *VuGNP1* displayed two haplotypes; the average GNP in the *VuGNP1*-HapI accessions was higher than that in *VuGNP1*-HapII accessions (Fig. 4e).

Yield and quality variations

Large variations in PL were observed in the 344 accessions, with the G subpopulation generally showing significantly shorter PLs than those of the VL and VC subpopulations (Fig. 6a). In total, four PL-related signals (PL-3.1, PL-3.2, PL-5.1 and PL-9.1) were detected in SNP-GWAS; PL-3.1, PL-3.2 and PL-9.1 were also detected in InDel-GWAS (Fig. 6b and Supplementary Figs. 4a and 5a). Two nitrate transporter 1 and peptide transporter family genes (VuG9803G015800 and VuG9803G015810) were identified in the PL-3.1 locus. Nitrate and peptide transporter family proteins transport numerous substrates[35,36] and are essential for plant development. VuG9803G015800 is less likely to be the candidate because no amino acid substitution was observed in any accessions. InDel-GWAS revealed an A/AG mutation in an exon–intron junction site of VuG9803G015810, which led to alternatively spliced transcripts between G98 and G323 (the third exon of 219 bp in G323 became a part of the second intron in G98) (Supplementary Fig. 6). Two haplotypes of this gene in different accessions caused significantly different PL phenotypes (Fig. 6b), supporting that VuG9803G015810 is the putative *VuPL1* in the PL-3.1 locus. The PL-9.1 locus contains six tandem-duplicating *Wall-Associated Receptor Kinase* (*WAK*) genes. Among them, VuG9809G017980, which is homologous to *Arabidopsis WAK2* (AT1G21270, a cell wall-associated kinase required for invertase activity and cell growth)[37,38], is closest to the peak SNP (Chr09: 35607906) and was considered as the putative *VuPL4*. Three haplotypes of *VuPL4* were observed, and *VuPL4*-HapI likely contributes to long PL (Fig. 6b). Transcriptions of five *WAK*s (VuG9809G017960, VuG9809G017970, VuG9809G017980, VuG9809G017990 and VuG9809G018010) were detected in developing pods but were rarely observed in seeds at the same stage (Supplementary Fig. 3a). Their higher expressional levels were usually observed in VC or VL cowpea

https://doi.org/10.1038/s41588-024-01722-w

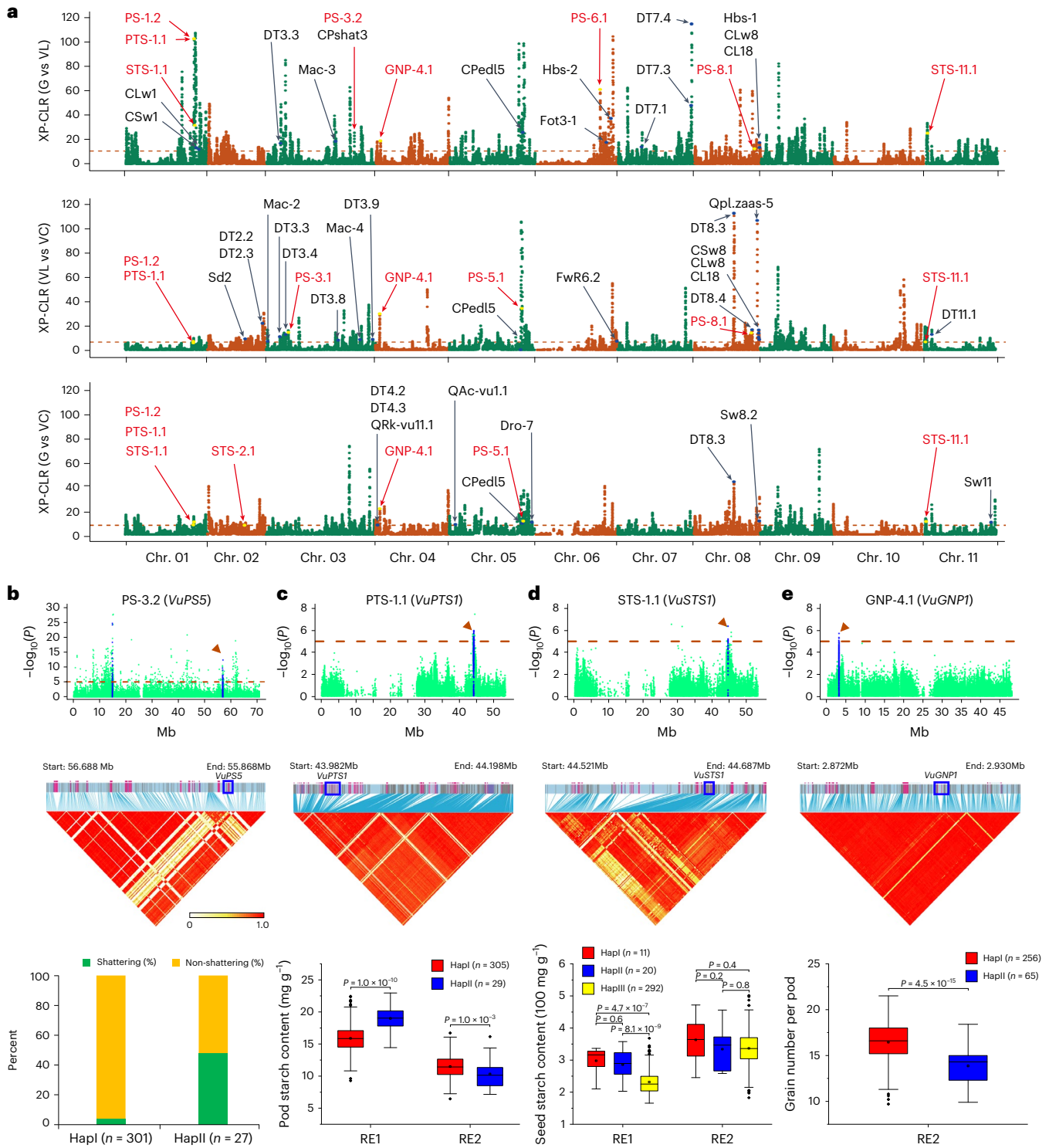


Fig. 4 | Genome-wide distribution of selective sweeps in cowpea and GWAS for four different traits. a, Selective sweeps through comparisons of G versus VL, VL versus VC and G versus VC. The GWAS signals identified in this study are labeled in red and the known QTLs are labeled in black. **b**–**e**, Manhattan plots (top), local linkage disequilibrium (LD) block analysis (middle) and haplotype analysis (bottom) for PS-3.2 (**b**), PTS-1.1 (**c**), STS-1.1 (**d**) and GNP-4.1 (**e**). The putative genes for each signal are shown in the blue boxes in the LD heat maps. For the haplotype analysis, the *n* values in the histogram and boxplots indicate the accession number with the corresponding haplotypes. In the boxplots, the 25% and 75% quartiles are shown as the lower and upper edges of boxes, respectively; the central lines denote the median and the small hollow squares indicate the mean. The whiskers extend to 1.5× the interquartile range and the small solid diamonds indicate outliers. The *P* values for two-sided Student's *t*-test are shown above the boxplots.

pods rather than those of G cowpeas (Supplementary Fig. 3b), implying another possibility that *WAK*s might contribute to the PL difference in dose/transcripts-dependent manner. Furthermore, both putative genes in the PL-3.2 (VuG9803G016720, *VuPL2*) and PL-5.1 (VuG9805G030040, *VuPL3*) loci encode uncharacterized proteins and both have two haplotypes that are related to PL variations (Supplementary Fig. 4a). A new

https://doi.org/10.1038/s41588-024-01722-w

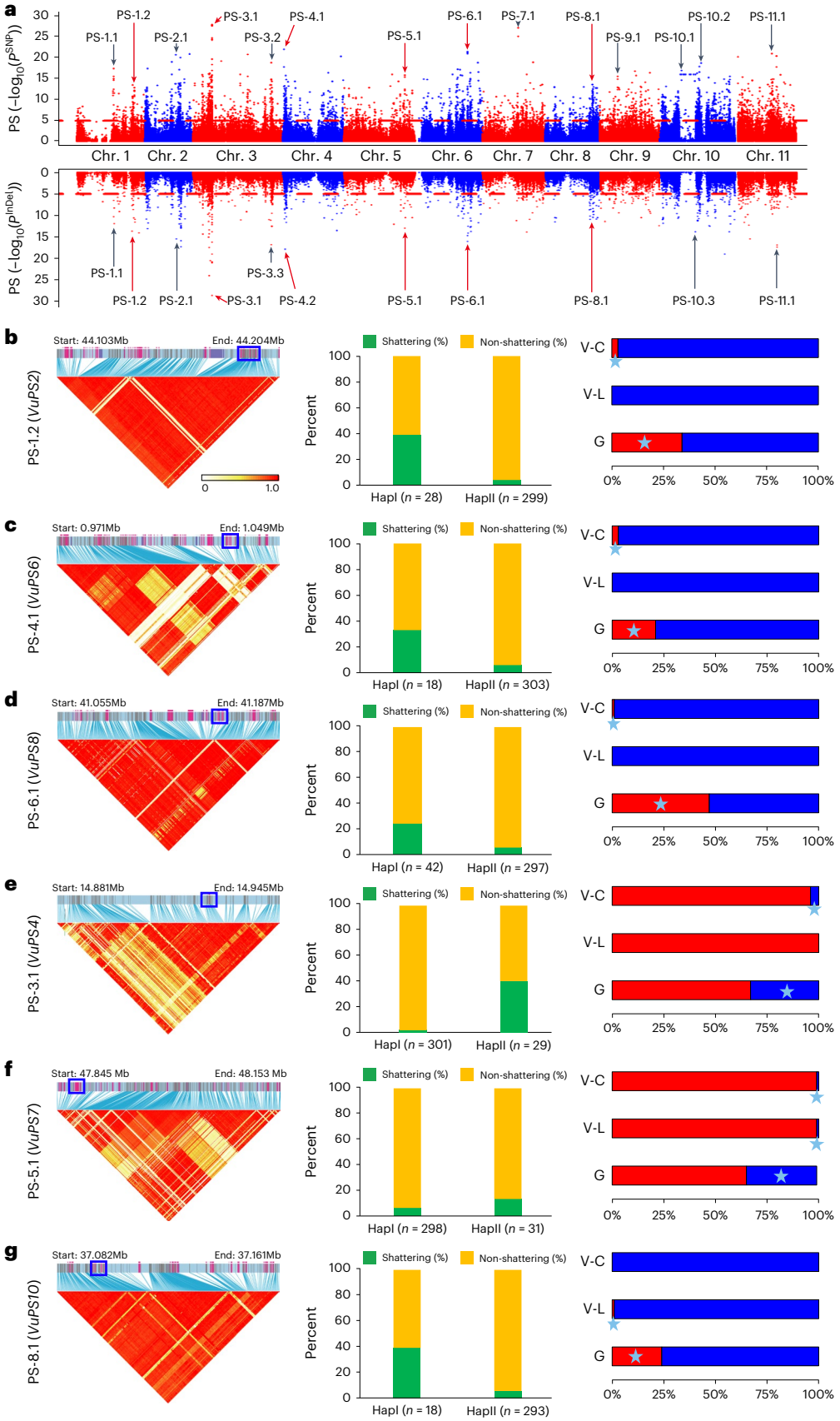


Fig. 5 | Identification of PS-related genes in cowpeas. a, GWAS signals of PS phenotypes based on SNP genotypes (top) and InDel genotypes (bottom); **b**–**g**, local LD block analysis (left), haplotype analysis in the 344 accessions (middle) and the haplotype distribution in different cowpea subpopulations (right) for the PS genes *VuPS2* (**b**); *VuPS6* (**c**); *VuPS8* (**d**); *VuPS4* (**e**); *VuPS7* (**f**) and *VuPS10* (**g**). The putative genes for each signal are shown in the blue box in the LD heat map. The *n* values in the histograms indicate the accession number with the corresponding haplotypes. The light-blue star indicates the haplotypes related to a higher shattering rate.

PL-related QTL on VuA147Chr03 (61899120-62027427)[18] likely contained our PL-3.2 locus, even though their genomic locations were different, possibly because of a large INV between the two genome assembles (Supplementary Fig. 7 and Supplementary Table 27). In addition, transcriptional profiles of both genes indicated their potential function at an earlier pod development stage (Supplementary Fig. 3b).

The favorable alleles of the putative genes in PL-9.1 and PL-3.1 seem to have a stronger function on the PL phenotype than those in other loci (Fig. 6c), and PL increased along with the number of favored alleles of PL loci in single accession (Fig. 6d). Moreover, *VuPL4*-HapI and *VuPL1*-HapII were strongly selected during cowpea domestication and improvement, respectively (Fig. 6a). The function of both *VuPL1* and *VuPL4* were further validated in a recombinant inbred line (RIL) population (Supplementary Fig. 8a,b and Supplementary Table 28), and *VuPL1* seems to have a stronger role than *VuPL4* in this population (Supplementary Fig. 8c). The PL-9.1 signal was also identified by several selective sweep detections (Supplementary Fig. 2a–c). Unsurprisingly, abundant InDel diversity was found in six tandem-duplicating *WAK* genes in G-type cowpeas, which was rarely observed in the VC type (Supplementary Table 29), indicating that selections on this locus probably facilitated cowpea domestication and improvement.

Two GNP-associated signals were detected (Fig. 6e), of which GNP-4.1 was stably identified by domestication sweeps and GNP-10.1 was identified in improvement sweeps (Fig. 4a,e and Supplementary Fig. 2b). Putative *VuGNP2* (VuG9810G016140) in GNP-10.1 encoded 1-aminocyclopropane-1-carboxylate synthase (ACS), which is a rate-limiting enzyme of ethylene biosynthesis[39]. Among the three haplotypes of *VuGNP2*, *VuGNP2*-HapIII is identified as the favored type (Fig. 6e). Seed weight is another important yield-related trait that has undergone strong selection in crop domestication and improvement processes[20,26,40]. Two signals associated with thousand seeds weight (TSW) were detected by both SNP-GWAS and InDel-GWAS (Fig. 6f and Supplementary Fig. 5b). A transcription termination factor (MTERF8) coding gene VuG9809G001620 (*VuTSW1*) was anchored near the peak SNP of the TSW-9.1 locus, whose product (MTERF proteins) has multiple roles in plant development[41]. Among the three haplotypes of *VuTSW1*, *VuTSW1*-HapII is probably the favored type (Fig. 6f). The peak SNP of TSW-9.2 is located inside VuG9809G007710 (*VuTSW2*), which encodes an endoplasmic reticulum membrane protein that serves many roles in the cell, including calcium storage, protein synthesis and lipid metabolism in plants[42,43]. Three haplotypes of *VuTSW2* were observed, and *VuTSW2*-HapIII showed the largest TSW (Fig. 6f). Moreover, the functions of *VuTSW1* and *VuTSW2* were further validated in an RIL population and an F_2 population, respectively (Supplementary Fig. 8d–f).

Soluble sugar, total starch and crude protein content are three basic quality traits of legume crops[44,45]. We detected three signals (PSS-9.1, PSS-9.2, PSS-11.1) that were significantly associated with PSS (Fig. 7a). PSS-9.1 contained two bidirectional sugar transporter *SWEET10*-like genes (VuG9809G011400 and VuG9809G011410), which have important roles in transporting sucrose and hexose[46–50]. VuG9809G011400 showed higher expression patterns in developing pods while VuG9809G011410 displayed nearly no expression in all tested tissues (Supplementary Fig. 3a,b), indicating that VuG9809G011400 is probably the function-relevant gene for PSS-9.1 (*VuPSS1*). A xylulose kinase protein VuG9809G020860, located 662 bp downstream of the peak SNP of PSS-9.2, was proposed as the putative *VuPSS2*. However, no SNP mutations were identified in any accessions, indicating that *VuPSS2* may affect PSS content by its *cis*-regulatory elements. The expression pattern of this gene also supports this hypothesis (Supplementary Fig. 3b). The PSS-11.1 locus contained three *beta-galactosidase* genes (VuG9811G017350, VuG9811G017390 and VuG9811G017400), a negative regulator of cell galactose levels[51], and VuG9811G017400 (*VuPSS3*) is the closest gene to the peak SNP. Four haplotypes were investigated in *VuPSS3*, and *VuPSS3*-HapIV exhibited the highest PSS content (Fig. 7b). The low transcription levels of VuG9811G017390 and VuG9811G017400 in VC cowpea might relate to its high PSS content (Supplementary Fig. 4b). In addition, we discovered two signals (PCP-7.1 and PCP-8.1) associated with pod crude protein (PCP) content by both SNP-GWAS and InDel-GWAS (Fig. 7c and Supplementary Fig. 5d). PCP-7.1 contains an uncharacterized protein gene (*VuPCP1*), and *VuPCP1*-HapII probably leads to higher PCP content (Supplementary Table 25). *VuPCP2* (VuG9808G012140) in PCP-8.1 encodes a *bHLH* transcription factor that involves the E3 ubiquitin pathway, and *VuPCP2*-HapI was the predominant type in PCP (Fig. 7c,d).

Furthermore, GWAS also revealed one signal associated with seed soluble sugar (SSS) content (Fig. 7e), which contained a FAR1-RELATED SEQUENCE (FAR1) family protein (*VuSSS1*, VuG9810G003320). FAR1 has roles in starch synthesis as well as sugar transport and degradation[52]. *VuSSS1* generated two haplotypes and *VuSSS1*-HapI accessions displayed higher SSS content (Fig. 7f). In addition, we detected two signals associated with seed crude protein (SCP) content (Fig. 7g,h and Supplementary Fig. 4b). SCP-3.1 contains a phosphoinositide phosphatase SAC9 protein gene (*VuSCP1*, VuG9803G018410) that belongs to the SAC domain-containing family involved in protein regulation[53,54]. Three main haplotypes were found, and *VuSCP1*-HapIII showed the highest SCP values (Fig. 7h). The SCP-4.1 region contained a RING-type E3 ubiquitin transferase protein VuG9804G016420 (*VuSCP2*), *VuSCP2*-HapIII and *VuSCP2*-HapIV usually led to higher SCP (Supplementary Fig. 4b). Similar expression patterns of *VuSCP1* and *VuSCP2* suggest that they may affect the SCP phenotype at the mid-stage of seed development (Supplementary Fig. 3c).

In total, five signals for STS were detected by GWAS (Fig. 7i and Supplementary Figs. 4c and 5h). *VuSTS2* (VuG9802G007660) in STS-2.1 contained an MYB transcription factor and *VuSTS5* (VuG9811G001010) in the STS-11.1 locus encodes a phosphatidylserine decarboxylase that possibly affects a key plant development regulator, phosphatidylserine[55]. Different haplotypes of both genes showed varied STSs in the two environments, suggesting that they were largely influenced by environmental conditions (Fig. 7j and Supplementary Fig. 4c). VuG9803G004130 (*VuSTS3*) in STS-3.1 encoded an uncharacterized protein, and *VuSTS3*-HapIII exhibited the largest effect on STS. STS-6.1 contained a NAD-dependent protein deacetylase VuG9806G009220 (*VuSTS4*), which possibly affects starch biosynthesis and regulation[56], and *VuSTS4*-HapI showed a stronger effect on STS (Supplementary Fig. 4c).

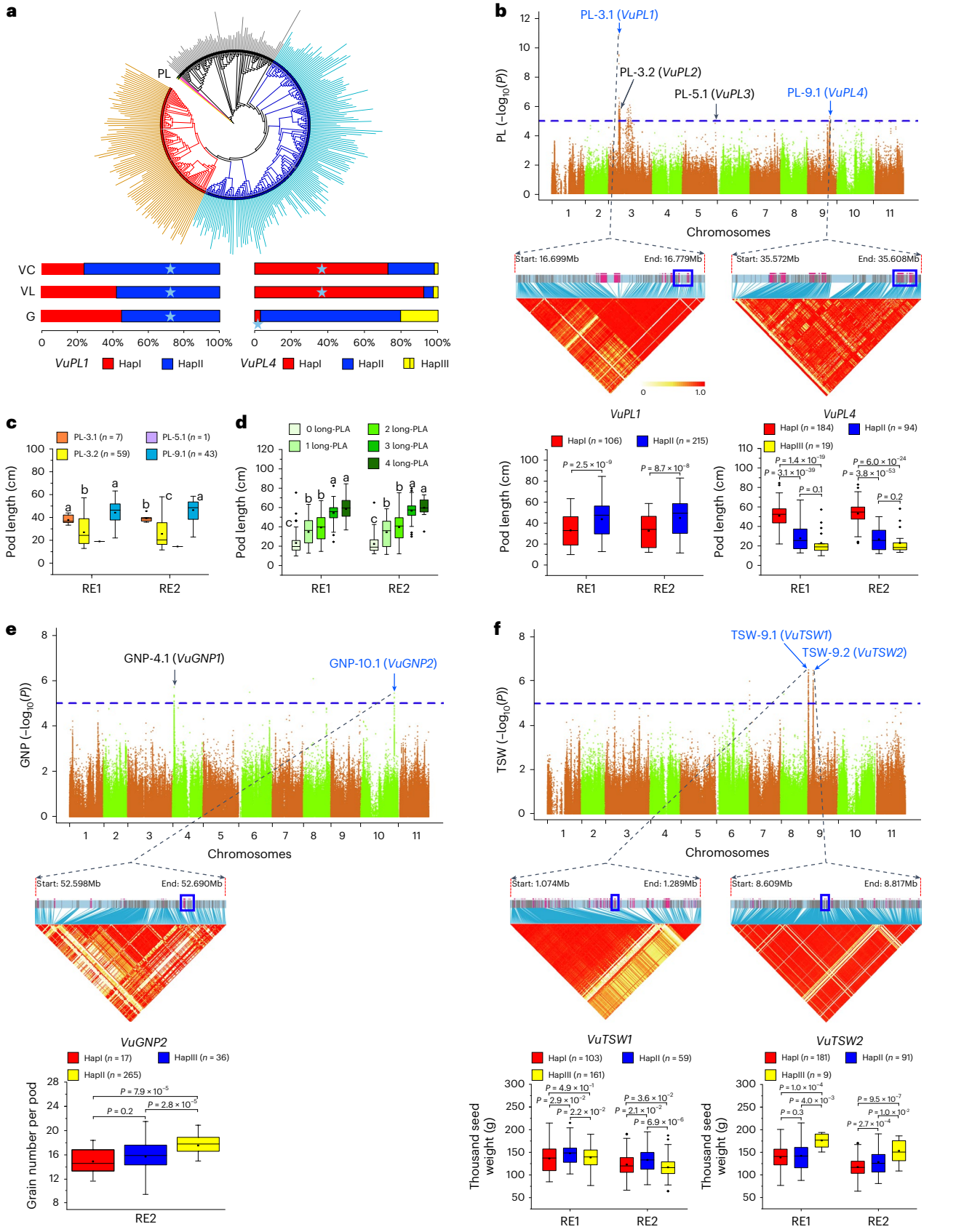
Discussion

Cowpeas, which are abundant in balanced starch and protein, are suitable substitutes for cereals and animal proteins in developing countries. Here, we assembled two chromosome-scale genomes. The differentially expanded genes in vegetable G98 and grain G323 are probably

Fig. 6 | Yield traits-related gene mining in cowpea. a, The PL and haplotype distribution of *VuPL1* and *VuPL4* in different subpopulations. The columns in the outer circle of the evolutionary tree represent PL values of the cowpea accessions; the light-blue stars indicate the haplotypes related to longer PL. **b**, GWAS for PL trait, pairwise LD heat map and haplotype analysis for the candidate genes in the PL-3.1 and PL-9.1 loci. **c**, The effect of different signals on PL. **d**, The additive effects of different allele pyramids on PL. **e**, GWAS for the trait of GNP and candidate gene analysis of *VuGNP2*. **f**, GWAS for the TSW trait and candidate gene analysis of *VuTSW1* and *VuTSW2*. The putative genes for each signal are shown in the blue box in the LD heat map. In the boxplots, the *n* values indicate the accession number with the corresponding haplotypes, the 25% and 75% quartiles are shown as lower and upper edges of boxes, respectively, central lines denote the median and the small hollow square indicates the mean. The whiskers extend to 1.5× the interquartile range and the small solid diamonds indicate outliers. In **b**, **c**, **e** and **f**, the *P* values for two-sided Student's *t*-test are shown above the boxplot. In **d**, significant levels were determined using a least significant difference test and the different lowercase letters above the boxplots represent significant differences ($P \leq 0.05$).

https://doi.org/10.1038/s41588-024-01722-w

responsible for their long PL and high STS, respectively (Fig. 2). Moreover, numerous SVs containing genes enriched in Gene Ontology (GO) terms such as hormone transport and cell wall modification might have been strongly selected during cowpea domestication. SVs possessing genes related to GO terms like protein catabolic process and L-amino acid transmembrane transporter activities have probably been subjected to artificial selection during vegetable cowpea improvement (Supplementary Fig. 9).



https://doi.org/10.1038/s41588-024-01722-w

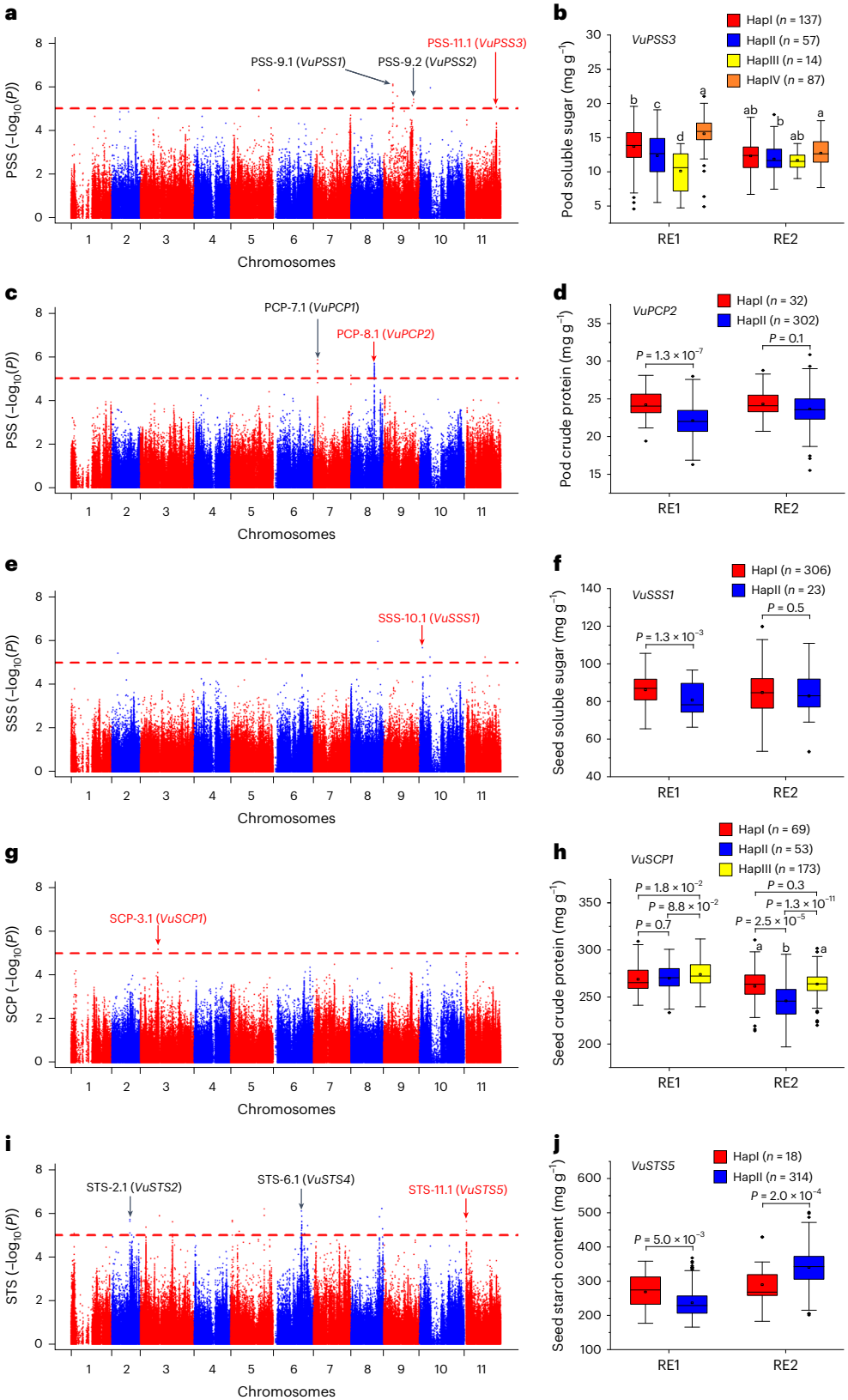


Fig. 7 | Quality traits-related gene mining in cowpea. a, GWAS for the trait of PSS content. **b**, Haplotype analysis of *VuPSS3*. **c**, GWAS for the trait of PCP content. **d**, Haplotype analysis of *VuPCP2*. **e**, GWAS for the trait of SSS content. **f**, Haplotype analysis of *VuSSS1*. **g**, GWAS for the trait of SCP content. **h**, Haplotype analysis of *VuSCP1*. **i**, GWAS for the trait or STS content. **j**, Haplotype analysis of *VuSTS5*. In boxplots, the *n* values indicate the accession number with the corresponding haplotypes, the 25% and 75% quartiles are shown as lower and upper edges of boxes, respectively, central lines denote the median and the small hollow square indicates the mean. The whiskers extend to 1.5× the interquartile range and the small solid diamonds indicate outliers. In **b**, significant levels were determined using a least significant difference test; different lowercase letters above the boxplots represent significant differences ($P \leq 0.05$). In **d**, **f**, **h** and **j**, the *P* values for two-sided Student's *t*-test are shown above the boxplots.

https://doi.org/10.1038/s41588-024-01722-w

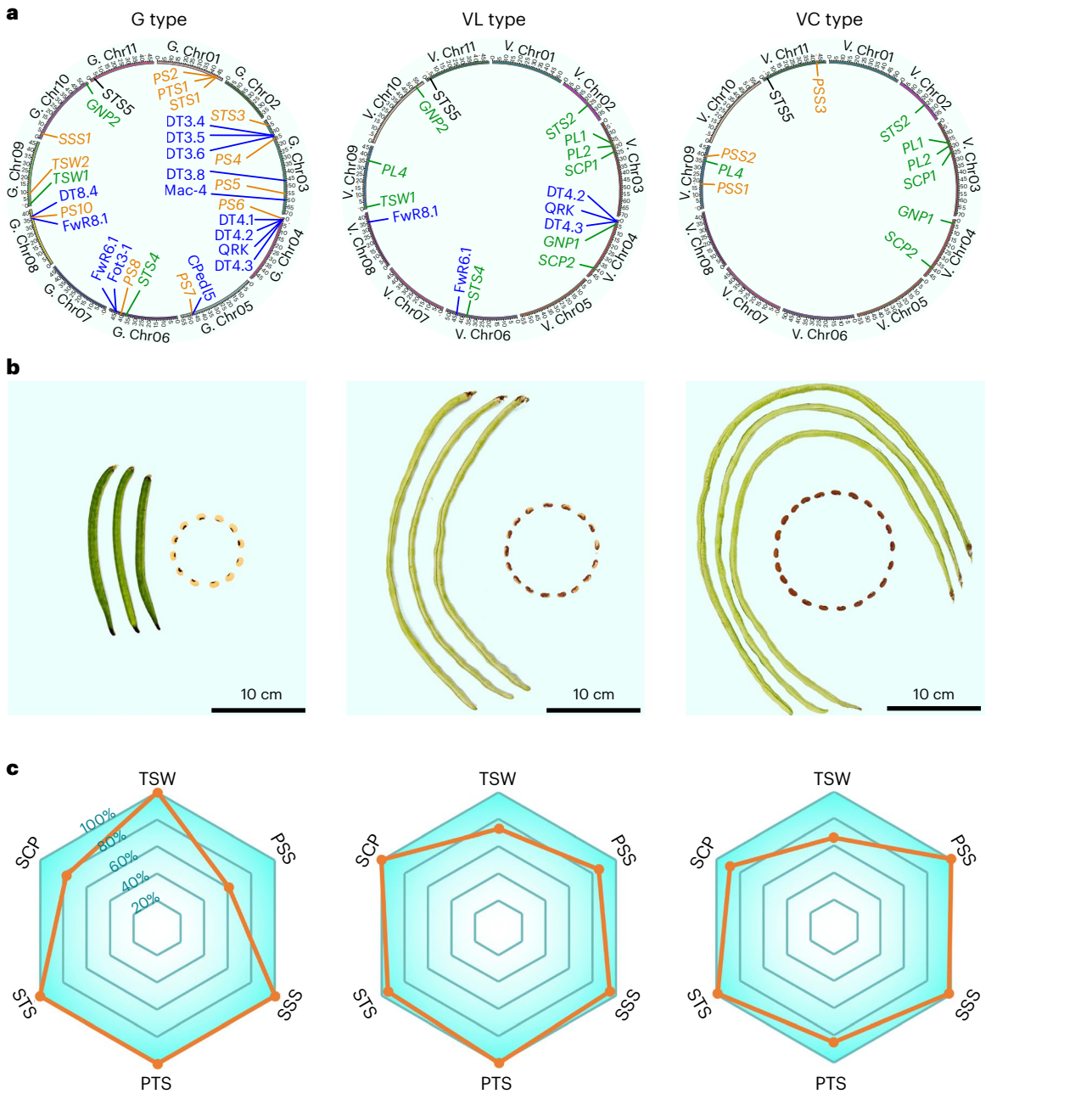


Fig. 8 | Proposed model of cowpea domestication and improvement. **a**, Favorable alleles of selected genes distributed in the genomes of the three subpopulations. The data were derived from three representative accessions: G323, ZN016 and TZ30. Gene names in orange font indicate that their favorable alleles were found only in G-type or VC-type cowpeas. Gene names in green indicate that their favorable alleles were found in two neighboring subpopulations of cowpeas (G to VL; VL to VC). Gene names in black indicate that their favorable alleles were found in all subpopulations. Loci names in blue indicate the favorable haplotypes of stress-resistance-related QTLs located adjacent to the PS genes. **b**, The PL and GNP phenotypes of the three selected accessions. Scale bar is provided for size comparison. **c**, Relative values of TSW, PSS, SSS, PTS, STS and SCP traits among three accessions.

Loss of PS is recognized as a domestication syndrome event[26,57]. Thus, strong and continuous selection on PS loci is unsurprisingly observed during G to VL cowpea domestication (Fig. 4 and Supplementary Fig. 2). However, during the vegetable cowpea improvement (VL to VC), PS-1.2, PS-3.1, PS-4.1 and PS-6.1 were sporadically re-introduced into four vegetable cultivars (Fig. 5b–e and Supplementary Table 30). A possible reason for this phenomenon is that these loci might have been selected alongside their adjacent genes through a 'hitchhiking' effect, as several drought tolerance, disease resistance and agronomic trait-related genes or loci (for example, PTS-1.1, DT3.4, DT4.1, QRk-vu11.1, CPdel5 and DT8.4)[9,15,58,59] were found in the vicinity of these PS loci (Fig. 4a, Supplementary Fig. 2 and Supplementary Tables 22 and 30). The findings of multiple PS loci will raise the potential of pyramiding beneficial non-shattering alleles into the shattering-prone grain cowpea, enabling the breeding of PS-resilient cultivars. On the other hand, a diversity of loci related to drought resistance or disease resistance (DTs, Mac-4, QRk-vu1.1, Fot3-1 and FwRs)[10,58–69] is rarely observed in VC cowpeas (Fig. 8 and Supplementary Tables 22 and 30), which is largely consistent with the lower stress-resistance of VC than those of G cowpeas[19,68,70]. The coexistence of beneficial stress-relevant alleles and unfavorable PS loci in four VC accessions (Supplementary Table 30) suggests potential linkage drags between them. This finding provides

a genomic roadmap toward avoiding the re-introduction of PS genes during stress-resistance improvement in vegetable cowpeas.

PL, GNP and TSW are important yield factors, whereas starch is a critical factor of crop quality and nutrition in cowpeas[32]. VL and VC groups normally show higher PL and GNP values than G-group cowpeas, which might be attributed to their favorable haplotypes of the *VuPL1*, *VuPL4* and *VuGNP1* genes (Fig. 6a, Supplementary Fig. 10a and Supplementary Table 30). The dry seed yield of grain cowpeas from China is about 1,022 kg ha^{-1}, much higher than that in Africa (581.4 kg ha^{-1}; FAOSTAT, 2020; https://www.fao.org/faostat/en/#data/QCL), whereas the seed yield of vegetable cowpeas ranges from 1,500 to 2,250 kg ha^{-1} (unpublished data), largely owing to their longer PL and higher GNP. Thus, our findings point to the promising applicability of these beneficial haplotypes in yield enhancements for grain cowpeas, which would help to mitigate hunger and malnutrition in developing regions.

The polymorphism and distribution of yield-related and quality-related genes among the three subpopulations (G, VL and VC) dissected the genetic basis of phenotypic differentiation in cowpeas (Fig. 8 and Supplementary Fig. 10). For instance, grain cowpeas usually exhibit higher TSW, PTS, STS and SSS values than vegetable cowpeas. Unsurprisingly, nine favorable alleles for these four traits were commonly identified in grain cowpeas. Conversely, favorable alleles of *VuPTS1*, *VuSTS1* and *VuTSW2* genes were rarely observed in vegetable cowpeas (Fig. 8a), and the favorable *VuTSW1*-HapII and *VuSSS1*-HapI were more strongly selected in grain cowpeas than in vegetable cowpeas (Supplementary Fig. 10b,f). Surprisingly, grain cowpeas possess lower SCP, an important nutritional ingredient, than vegetable cowpeas, which might relate to the over-accumulation of SSS and STS as well as their higher TSW (Fig. 8c and Supplementary Fig. 10), as both total sugar and seed size were negatively correlated with protein content in soybeans[49,50]. Moreover, artificial selection for taste quality of grain may be another reason for low SCP in grain cowpeas, as higher protein content may increase seed hardness[71–73]. Interestingly, the favorable alleles of two *VuSCP* genes as well as favorable *VuSTS2*-HapII and *VuSTS5*-HapII were more frequently observed in vegetable cowpeas, displaying their potential as genetic resources for balanced starch and protein improvement in grain cowpeas. By contrast, the increase in PSS during cowpea domestication is likely driven by stacking the favorable *VuPSS* alleles (Fig. 8 and Supplementary Fig. 10c). This finding provides useful guidance for the precision breeding of high-quality vegetable cowpeas.

In conclusion, this study provides a global landscape of genome-wide genetic variations associated with important agronomic traits in cowpeas and offers genomic insights into the domestication and improvement of cowpea subspecies. The differential genomic selections of yield and quality traits will facilitate the establishment of genetic resource toolkits for the bidirectionally reciprocal improvement of grain and vegetable cowpeas.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01722-w.

References

1. Singh, B. B. *Cowpea: The Food Legume of the 21st Century* (Crop Science Society of America, 2014).
2. Timko, M. P & Singh, B. in *Plant Genetics and Genomics: Crops and Models* (eds Moore, P. H. & Ming, R.) 227–258 (Springer, 2008).
3. Pasquet, R. S. & Padulosi, S. Genus *Vigna* and cowpea (*Vigna unguiculata* (L.) Walp.) taxonomy: current status and prospects. In *Proc. 5th World Cowpea Conference* (eds Boukara, O. et al.) 66–87 (International Institute of Tropical Agriculture, 2012).
4. Herniter, I. A., Muñoz-Amatriaín, M. & Close, T. J. Genetic, textual, and archeological evidence of the historical global spread of cowpea (*Vigna unguiculata* (L.) Walp.). *Legume Sci.* **2**, e57 (2020).
5. Timko, M. P., Ehlers, J. D. & Roberts, P. A. in *Genome Mapping and Molecular Breeding in Plants*, Vol. 3 (ed. Kole, C.) 49–67 (Springer, 2007).
6. National Research Council. *Lost Crops of Africa*, Vol. 2 (The National Academies Press, 2006).
7. Som, M. G. & Hazra, P. in *Genetic Improvement of Vegetable Crops* (eds Kalloo G. & Bergh, B. O.) 339–354 (Elsevier, 1993).
8. Kongjaimun, A. et al. The genetics of domestication of yardlong bean, *Vigna unguiculata* (L.) Walp. ssp. *unguiculata* cv.-gr. *sesquipedalis*. *Ann. Bot.* **109**, 1185–1200 (2012).
9. Lo, S. et al. Identification of QTL controlling domestication-related traits in cowpea (*Vigna unguiculata* L. Walp). *Sci. Rep.* **8**, 6261 (2018).
10. Lo, S. et al. A genome-wide association and meta-analysis reveal regions associated with seed size in cowpea [*Vigna unguiculata* (L.) Walp]. *Theor. Appl. Genet.* **132**, 3079–3087 (2019).
11. Suanum, W. et al. Co-localization of QTLs for pod fiber content and pod shattering in F_2 and backcross populations between yardlong bean and wild cowpea. *Mol. Breed.* **36**, 80 (2016).
12. Andargie, M., Pasquet, R. S., Gowda, B. S., Muluvi, G. M. & Timko, M. P. Molecular mapping of QTLs for domestication-related traits in cowpea (*V. unguiculata* (L.) Walp.). *Euphytica* **200**, 401–412 (2014).
13. Herniter, I. A., Muñoz-Amatriaín, M., Lo, S., Guo, Y.-N. & Close, T. J. Identification of candidate genes controlling black seed coat and pod tip color in cowpea (*Vigna unguiculata* [L.] Walp). *G3* **8**, 3347–3355 (2018).
14. Kongjaimun, A. et al. QTL mapping of pod tenderness and total soluble solid in yardlong bean [*Vigna unguiculata* (L.) Walp. subsp. *unguiculata* cv.-gr. *sesquipedalis*]. *Euphytica* **189**, 217–223 (2013).
15. Xu, P. et al. Genomic regions, cellular components and gene regulatory basis underlying pod length variations in cowpea (*V. unguiculata* L. Walp). *Plant Biotechnol. J.* **15**, 547–557 (2017).
16. Wu, X. B., Cortés, A. J. & Blair, M. W. Genetic differentiation of grain, fodder and pod vegetable type cowpeas (*Vigna unguiculata* L.) identified through single nucleotide polymorphisms from genotyping-by-sequencing. *Mol. Hortic.* **2**, 8 (2022).
17. Lonardi, S. et al. The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant J.* **98**, 767–782 (2019).
18. Yang, T. et al. Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat. Genet.* **54**, 1553–1563 (2022).
19. Pan, L. et al. Comprehensive genomic analyses of *Vigna unguiculata* provide insights into population differentiation and the genetic basis of key agricultural traits. *Plant Biotechnol. J.* **21**, 1426–1439 (2023).
20. Guan, J. T. et al. Genomic analyses of rice bean landraces reveal adaptation and yield related loci to accelerate breeding. *Nat. Commun.* **13**, 5707 (2022).
21. Rennie, E. A. et al. Identification of a sphingolipid α-glucuronosyltransferase that is essential for pollen function in *Arabidopsis*. *Plant Cell* **26**, 3314–3325 (2014).
22. Chen, L. Y. et al. The *Arabidopsis* alkaline ceramidase TOD1 is a key turgor pressure regulator in plant cells. *Nat. Commun.* **6**, 6030 (2015).
23. Haslam, T. M. & Feussner, I. Diversity in sphingolipid metabolism across land plants. *J. Exp. Bot.* **73**, 2785–2798 (2022).
24. Liu, J. et al. Natural variation in ARF18 gene simultaneously affects seed weight and silique length in polyploid rapeseed. *Proc. Natl Acad. Sci. USA* **112**, E5123–E5132 (2015).

25. Shi, L. L. et al. A CACTA-like transposable element in the upstream region of *BnaA9.CYP78A9* acts as an enhancer to increase silique length and seed weight in rapeseed. *Plant J.* **98**, 524–539 (2019).
26. Abbo, S. et al. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends Plant Sci.* **19**, 351–360 (2014).
27. Yang, J. H. et al. Genomic signatures of vegetable and oilseed allopolyploid *Brassica juncea* and genetic loci controlling the accumulation of glucosinolates. *Plant Biotechnol. J.* **19**, 2619–2628 (2021).
28. Kang, L. et al. Genomic insights into the origin, domestication and diversification of *Brassica juncea*. *Nat. Genet.* **53**, 1392–1402 (2021).
29. Lo, S. et al. Genetic, anatomical, and environmental patterns related to pod shattering resistance in domesticated cowpea [*Vigna unguiculata* (L.) Walp]. *J. Exp. Bot.* **72**, 6219–6229 (2021).
30. Taylor-Teeples, M. et al. An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* **517**, 571–575 (2015).
31. Zhang, D. M. et al. An uncanonical CCCH-tandem zinc-finger protein represses secondary wall synthesis and controls mechanical strength in rice. *Mol. Plant.* **11**, 163–174 (2018).
32. Huang, L. C., Tan, H. Y., Zhang, C. Q., Li, Q. F. & Liu, Q. Q. Starch biosynthesis in cereal endosperms: an updated review over the last decade. *Plant Commun.* **2**, 100237 (2021).
33. Tian, H. et al. *Arabidopsis* NPCC6/NaKR1 is a phloem mobile metal binding protein necessary for phloem function and root meristem maintenance. *Plant Cell* **22**, 3963–3979 (2010).
34. Su, S. et al. Gibberellins orchestrate panicle architecture mediated by DELLA-KNOX signalling in rice. *Plant Biotechnol. J.* **19**, 2304–2318 (2021).
35. Léran, S. et al. A unified nomenclature of NITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER family members in plants. *Trends Plant Sci.* **19**, 5–9 (2014).
36. Corratgé-Faillie, C. & Lacombe, B. Substrate (un)specificity of *Arabidopsis* NRT1/PTR FAMILY (NPF) proteins. *J. Exp. Bot.* **68**, 3107–3113 (2017).
37. Kanneganti, V. & Gupta, A. K. Wall associated kinases from plants—an overview. *Physiol. Mol. Biol. Plants* **14**, 109–118 (2008).
38. Kohorn, B. D. et al. An *Arabidopsis* cell wall-associated kinase required for invertase activity and cell growth. *Plant J.* **46**, 307–316 (2006).
39. Barry, C. S., Llop-Tous, M. I. & Grierson, D. The regulation of 1-aminocyclopropane-1-carboxylic acid synthase gene expression during the transition from system-1 to system-2 ethylene synthesis in tomato. *Plant Physiol.* **123**, 979–986 (2000).
40. Kaga, A., Isemura, T., Tomooka, N. & Vaughan, D. A. The genetics of domestication of the azuki bean (*Vigna angularis*). *Genetics* **178**, 1013–1036 (2008).
41. Robles, P. & Quesada, V. Research progress in the molecular functions of plant mTERF proteins. *Cells* **10**, 205 (2021).
42. Schwarz, D. S. & Blower, M. D. The endoplasmic reticulum: structure, function and response to cellular signaling. *Cell. Mol. Life Sci.* **73**, 79–94 (2016).
43. Zang, J. Z., Kriechbaumer, V. & Wang, P. W. Plant cytoskeletons and the endoplasmic reticulum network organization. *J. Plant Physiol.* **264**, 153473 (2021).
44. Bouchenak, M. & Lamri-Senhadji, M. Nutritional quality of legumes, and their role in cardiometabolic risk prevention: a review. *J. Med. Food* **16**, 185–198 (2013).
45. Zhang, M. et al. Progress in soybean functional genomics over the past decade. *Plant Biotechnol. J.* **20**, 256–282 (2022).
46. Chen, L. Q. Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* **468**, 527–532 (2010).
47. Yang, J. L. et al. *SWEET11* and 15 as key players in seed filling in rice. *N. Phytol.* **218**, 604–615 (2018).
48. Sosso, D. et al. Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. *Nat. Genet.* **47**, 1489–1493 (2015).
49. Wang, S. D. et al. Simultaneous changes in seed size, oil content and protein content driven by selection of *SWEET* homologues during soybean domestication. *Natl Sci. Rev.* **7**, 1776–1786 (2020).
50. Duan, Z. B. et al. Natural allelic variation of *GmST05* controlling seed size and quality in soybean. *Plant Biotechnol. J.* **20**, 1807–1818 (2022).
51. Paniagua, C. et al. Antisense down-regulation of the strawberry β-galactosidase gene *FaβGal4* increases cell wall galactose levels and reduces fruit softening. *J. Exp. Bot.* **67**, 619–631 (2016).
52. Ma, L., Xue, N., Fu, X. Y., Zhang, H. S. & Li, G. *Arabidopsis thaliana* FAR-RED ELONGATED HYPOCOTYLS3 (FHY3) and FAR-RED-IMPAIRED RESPONSE1 (FAR1) modulate starch synthesis in response to light and sugar. *N. Phytol.* **213**, 1682–1696 (2017).
53. Zhong, R. Q. & Ye, Z. H. The SAC domain-containing protein gene family in *Arabidopsis*. *Plant Physiol.* **132**, 544–555 (2003).
54. Ilsley, J. L., Sudol, M. & Winder, S. J. The WW domain: linking cell signalling to the membrane cytoskeleton. *Cell. Signal.* **14**, 183–189 (2002).
55. Yao, H. Y. & Xue, H. W. Phosphatidic acid plays key roles regulating plant development and stress responses. *J. Integr. Plant Biol.* **60**, 851–863 (2018).
56. Zhang, H., Lu, Y., Zhao, Y. & Zhou, D. X. OsSRT1 is involved in rice seed development through regulation of starch metabolism gene expression. *Plant Sci.* **248**, 28–36 (2016).
57. Parker, T. A., Lo, S. & Gepts, P. Pod shattering in grain legumes: emerging genetic and environment-related patterns. *Plant Cell* **33**, 179–199 (2021).
58. Xu, P. et al. Natural variation and gene regulatory basis for the responses of asparagus beans to soil drought. *Front. Plant Sci.* **6**, 891 (2015).
59. Santos, J. R. P., Ndeve, A. D., Huynh, B. L., Matthews, W. C. & Roberts, P. A. QTL mapping and transcriptome analysis of cowpea reveals candidate genes for root-knot nematode resistance. *PLoS ONE* **13**, e0189185 (2018).
60. Huynh, B. L. et al. Genetic mapping and legume synteny of aphid resistance in African cowpea (*Vigna unguiculata* L. Walp.) grown in California. *Mol. Breed.* **35**, 36 (2015).
61. Huynh, B. L. et al. A major QTL corresponding to the *Rk* locus for resistance to root-knot nematodes in cowpea (*Vigna unguiculata* L. Walp.). *Theor. Appl. Genet.* **129**, 87–95 (2016).
62. Muchero, W., Ehlers, J. D., Close, T. J. & Roberts, P. A. Mapping QTL for drought stress-induced premature senescence and maturity in cowpea [*Vigna unguiculata* (L.) Walp.]. *Theor. Appl. Genet.* **118**, 849–863 (2009).
63. Muchero, W., Ehlers, J. D. & Roberts, P. A. Restriction site polymorphism-based candidate gene mapping for seedling drought tolerance in cowpea [*Vigna unguiculata* (L.) Walp.]. *Theor. Appl. Genet.* **120**, 509–518 (2010).
64. Muchero, W., Ehlers, J. D., Close, T. J. & Roberts, P. A. Genic SNP markers and legume synteny reveal candidate genes underlying QTL for *Macrophomina phaseolina resistance* and maturity in cowpea [*Vigna unguiculata* (L) Walp.]. BMC Genom. 12, 8 (2011).
65. Pottorff, M. et al. Genetic and physical mapping of candidate genes for resistance to *Fusarium oxysporum* f. sp. *tracheiphilum* race 3 in cowpea [*Vigna unguiculata* (L.) Walp.]. PLoS ONE 7, e41600 (2012).
66. Pottorff, M. et al. Identification of candidate genes and molecular markers for heat-induced brown discoloration of seed coats in cowpea [*Vigna unguiculata* (L.) Walp.]. *BMC Genom.* **15**, 328 (2014).

67. Ravelombola, W., Shi, A. & Huynh, B. L. Loci discovery, network-guided approach, and genomic prediction for drought tolerance index in a multi-parent advanced generation intercross (MAGIC) cowpea population. *Hortic. Res.* **8**, 24 (2021).
68. Wu, X. Y. et al. Association mapping for fusarium wilt resistance in Chinese asparagus bean germplasm. *Plant Genome* **8**, 1–6 (2015).
69. Wu, X. Y. et al. SNP marker-based genetic mapping of rust resistance gene in the vegetable cowpea landrace ZN016. *Legum. Res.* **41**, 222–225 (2018).
70. Heng, T., Kaga, A., Chen, X. & Somta, P. Two tightly linked genes coding for NAD-dependent malic enzyme and dynamin-related protein are associated with resistance to Cercospora leaf spot disease in cowpea (*Vigna unguiculata* (L.) Walp.). *Theor. Appl. Genet.* **133**, 395–407 (2020).
71. Geater, C. W. & Fehr, W. R. Association of total sugar content with other seed traits of diverse soybean cultivars. *Crop Sci.* **40**, 1552–1555 (2000).
72. Yoshikawa, Y., Chen, P. Y., Zhang, B., Scaboo, A. & Orazaly, M. Evaluation of seed chemical quality traits and sensory properties of natto soybean. *Food Chem.* **153**, 186–192 (2014).
73. Bu, Y. P. et al. Conditional and unconditional QTL analyses of seed hardness in vegetable soybean (*Glycine max* L. Merr.). *Euphytica* **214**, 237 (2018).

Methods

Plant materials

A total of 344 cowpea accessions were selected for re-sequencing, of which 307 accessions were collected in the Institute of Vegetables, Zhejiang Academy of Agricultural Sciences (ZAAS), 35 accessions were introduced from the National Crop Genebank of China, Institute of Crop Sciences, Chinese Academy of Agriculture Sciences (CAAS) and 2 wild cowpea accessions (*V. unguiculata* subsp. *baoulensis*, *V. unguiculata* subsp. *letouzeyi*) were exchanged from Dr Remy S. Pasquet at Université Montpellier in France. Among these accessions, 83 (24.13%) were collected from genebanks outside of China, and the remaining 261 (75.87%) were from a new core collection in China[15]. Among the 342 cultivated cowpeas, 87 belonged to the grain cowpea, 244 belonged to the vegetable cowpea and 11 were uncertain usage. Detailed information for each accession is provided in Supplementary Table 17.

Genome library construction and sequencing of G98 and G323

To extract high molecular weight DNA for constructing different sequencing libraries, the seeds of G98 and G323 were sowed in plastic pots and grown in a growth chamber for 2 weeks. The seedlings were transferred to a dark room for 24 h before sample collection. Genomic DNA was isolated from young leaves using the cetyltrimethylammonium bromide method[74].

The Illumina sequencing libraries for G98 and G323 were prepared using the TruSeq Nano DNA High Throughput Library Prep Kit following the manufacturer's instructions (Illumina). In brief, at least 1 µg DNA from each genotype was sheared using Covaris M220. The fragments were subjected to end repairing and adaptor ligation, and then separated on 2% agarose gel. The final paired-end libraries with insertion sizes of around 350 bp were sequenced on an Illumina HiSeq 2000 platform.

PacBio sequencing libraries of G98 and G323 were constructed following the standard single molecular real-time bell construction protocol (PacBio). The DNA was randomly sheared using g-TUBE (Covaris), the fragments were treated with end repairing, adaptor ligation and exonuclease digestion, and the desired fragments of 10–50 kb were selected using BluePippin electrophoresis. Finally, the libraries were constructed and sequenced on the PacBio CCS system with P6-C4 chemistry.

Hi-C libraries were created from tender leaves of G98 and G323 following the proximo Hi-C plant protocol. The samples were first fixed in formaldehyde to keep the cross-linking between DNA and protein and maintain their 3D structure in cells. Then the DNA was digested with the restriction endonuclease *HindIII* to generate different fragments with sticky ends. The fragments were treated with end repairing and adaptor ligation, resulting in the formation of chimeric circles. Finally, the cyclized fragments were disconnected and purified, and fragments with a size of 300–700 bp were selected to construct libraries and sequenced on the Illumina X Ten platform.

Genome assembly of two genotypes

The identical bioinformatic process was used to assemble the genomes of G98 and G323. Firstly, clean, short Illumina reads were used to estimate genome size by *k*-mer distribution[75] with KAT (v.2.4.1; https://github.com/TGAC/KAT). Subsequently, the PacBio CCS data was used to assemble a draft genome using hifiasm (v.0.12) software[76]. Next, the Hi-C data was aligned into contigs and anchored into chromosomes using LACHESIS (v.2.0) software[77].

To assess the quality of the genome assembly for G98 and G323, the 458 conserved core genes in the CEGMA (v.2.5) database and the 1,614 core eukaryotic genes in the BUSCO (v.4) database were used to evaluate genome completeness, and then the Illumina clean reads were mapped to the assembled genome using BWA (v.0.7.8) to assess coverage rate and average depth. Finally, Merqury (v.1.3) was used to assess the consensus quality value and completeness of the genome assembly[78].

Transcriptome sequencing

For genome annotation, G98 and G323 were grown in a greenhouse under normal watering and drought-stress conditions. The drought treatment was performed from the third week until the fifth week after sowing. Then the well-hydrated and drought-stressed young seedling roots and leaves were collected. Young flower buds, pods 5 days after pollination and developing seeds 13 days after pollination under normal watering conditions were also collected. Total RNA for these samples was extracted using the QIAGEN RNeasy Plant Mini Kit (Hilden, Germany). RNA-sequencing (RNA-seq) libraries were prepared using NEBNext UltraTM RNA Library Prep Kit for Illumina (NEB, USA) following the manufacturer's recommendations and sequenced on an Illumina HiSeq 2000 platform.

Genome annotation in G98 and G323 genomes

For the annotation of repeats, de novo repeat libraries for the G98 genome and the G323 genome were created first using RepeatModeler (v.1.05)[79]. The predicted repeats were classified using RepeatClassifier[79] based on the known Repbase (v.19.06)[80], REXdb (v.3.0)[81] and Dfam (v.3.2)[82] databases. The LTRs were identified using LTRharvest (v.1.5.9) (-minlenltr 100 -maxlenltr 40,000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes)[83] and LTR_finder (v.1.1) (-D 40,000 -d 100 -L 9,000 -l 50 -p 20 -C -M 0.9)[84]. Then the de novo predicted results and repeats identified from the known databases were combined to form a species-specific transposable elements library for G98 and G323, respectively. Finally, the transposable element sequences were identified and classified by a homology search against the library using RepeatMasker (v.4.10)[85]. In addition, tandem repeats were annotated by Tandem Repeats Finder[86] and the MIcroSAtellite identification tool (MISA v.2.1)[87].

For the annotation of protein-coding genes, three strategies, including de novo prediction, homology-based prediction and RNA-seq-based prediction, were used to identify and annotate candidate genes for G98 and G323, respectively. The de novo prediction was performed using two ab initio gene-prediction software tools: Augustus (v.2.4)[88] and SNAP (v.2013-11-29)[89]. In addition, the reference gene models from *Arabidopsis thaliana*, *Phaseolus vulgaris*, *Vigna unguiculata*, *V. radiata* and *V. angularis* were used to conduct homology-based predictions using the software GeMoMa (v.1.7)[90]. The RNA-seq data from different tissues were mapped to the cowpea reference genome using Hisat (v.2.0.4)[91] and assembled by Stringtie (v.1.2.3)[92], and candidate genes were predicted based on the assembled transcripts using GeneMarkS-T (v.5.1)[93]. Finally, the predicted genes from different approaches were combined using the EVM software (v.1.1.1)[94] and updated by PASA (v.2.2.0)[95]. All the predicted genes were annotated by searching the GenBank non-redundant (v.20200921), TrEMBL (v.202005), Pfam (v.33.1), SwissProt (v.202005), eukaryotic orthologous groups (KOG, v.20110125), GO (v.20200615) and Kyoto Encyclopedia of Genes and Genomes (KEGG, v.20191220) databases.

Meanwhile, non-coding RNAs were identified using different software. tRNA was identified using the tRNAscan-SE (v.1.3.1)[96], rRNA was identified by barrnap (v.0.9)[97], miRNA was identified by the miRBase (v.21) database and snoRNA and snRNA were predicted using the INFERNAL against the Rfam (v.12.0) database[98,99].

To identify the pseudogenes, the GenBlastA (v.1.0.4)[100] program was used to scan the whole genomes to search the candidates after masking predicted functional genes, then the non-mature mutations and frame-shift mutations in the candidates were further identified using GeneWise (v.2.4.1)[101] to confirm the pseudogenes.

Comparative genomics analysis in legume crops

A total of 25 plant genomes including G98, G323, IT97K-499-35, a monocot (*Oryza sativa*)[102], an eudicot (*Arabidopsis thaliana*)[103] and 20 legume crops (including *Vigna radiata*[104], *V. angularis*[105], *Phaseolus vulgaris*[106], *P. lunatus*[107], *Glycine max*[108], *Cajanus cajan*[109], *Mucuna*

pruriens (https://www.ncbi.nlm.nih.gov/genome/71552), *Spatholobus suberectus* Dunn[110], *Abrus precatorius* (https://www.ncbi.nlm.nih.gov/genome/74709), *Lotus japonicus*[111], *Pisum sativum* L.[112], *Trifolium pratense*[113], *Medicago truncatula*[114], *Cicer arietinum*[115], *Lupinus angustifolius*[116], *Arachis duranensis*[117], *A. ipaensis*[117], *Prosopis alba* (https://www.ncbi.nlm.nih.gov/genome/79095), *Mimosa pudica*[118] and *Cercis canadensis*[118]) were used to construct a phylogenetic tree. The protein sequences of ortholog genes were identified among these genomes first using OrthoFinder (v.2.3.9)[119], and 1,030 single-copy ortholog genes were used to estimate their phylogenetic relationships by constructing a phylogenetic tree using IQ-TREE (v.1.6.11)[120]. A Markov chain Monte Carlo tree program embedded in PAML (v.4.9)[121] was used to calculate the divergence time. The expansion and contraction of orthologous genes were searched using CAFE (v.4.2) (https://github.com/hahnlab/CAFE). GO and KEGG enrichment analyses were performed using the R package ClusterProfiler (v.3.18.0)[122] and *P* values of <0.05 indicated significantly enriched genes. For G98 and G323, the positive selection genes were also identified using CodeML in PAML (v.4.9). Gene collinearity analysis was performed using diamond (v.0.9.29.130) and MCScanX (jcvi) to determine the pairwise similarity.

SV analysis between G98 and G323

To compare two genomes, we used MUMmer (v.4.0)[123] to perform whole-genome alignment using the G323 genome as a reference and then used SyRI (v.1.4)[124] to detect the SNPs, small InDels (2–49 bp) and larger-scale SVs (≥50 bp). The detected SVs included TRANS, INVs, DUPs, PAVs (≥50 bp) and CNVs. The locations of the SNPs and InDels in the genome were determined using the ANNOVAR package (v.2020-10-7)[124]. GO and KEGG analyses for the genes with functional alterations were conducted using clusterProfiler (v.3.18.0)[122].

Genome re-sequencing of 344 accessions and SNP calling

The genomic DNA extraction, construction of the Illumina libraries and re-sequencing of 344 accessions were done using the same protocol as for G98 and G323 above. The sequencing depth for each accession is about 10×. The raw reads were filtered by removing the adaptor sequences, low-quality reads with >10% N and Q10 > 50% to generate clean reads. All clean reads for each accession were aligned to the G98 genome using the 'MEM' algorithm in the Burrows–Wheeler Aligner (bwa-mem2 v.2.2)[125]. After filtering the redundant reads using Samtools (v.1.9)[126], the HaplotypeCaller module in GATK (v.4.1.5.0)[127] was used to generate gvcf files for each accession and then to identify SNPs and InDels in the panel. SNP and InDel annotations were conducted based on the G98 genome using SnpEff (v.5.1)[128]. The original SNPs were further filtered following the criterion that only SNPs or InDels with a minor allele frequency greater than 5% and less than 20% missing data were considered high-quality SNPs or InDels. Finally, a total of 1,262,497 high-quality SNPs and 298,495 high-quality InDels were obtained and used for further analysis.

Population structure analysis

A common bean accession was used as an outgroup for population structure analysis after combining into the 344 accessions. A rooted neighbor-joining phylogenetic tree was conducted using MEGA7 (ref. 129) with 500 bootstraps. PCA was performed using EIGENSOFT (v.7.2.1)[130]. Population structure analysis was performed using admixture (v.1.23)[131]. Admixture analyses were run 20 times for each *K* value ranging from 2 to 12.

According to the population structure analysis results, values of π and F_{ST} of each subgroup were calculated using VCFtools (v.0.1.15; https://vcftools.github.io/index.html). For each subgroup, 50 accessions were randomly sampled each time and repeated 100 times to calculate the average value of π and F_{ST}. In addition, linkage disequilibrium decay was calculated for all pairs of SNPs within 500 kb using PopLDdecay (v.3.27)[132] with parameters '-MaxDist 500 -Het 0.1 -Miss 0.1'.

Field experiments and phenotyping

All accessions were planted for two repeated tests in the spring of 2021. Baiyun research base of GAAS in Guangzhou (RE1; 23° 07′ N, 113° 34′ E) and Hangzhou (RE2; 29° 50′ N, 120° 04′ E) were selected as two main habitats representing the different environmental conditions of South and East China. Two experimental replications were conducted at each site. PS was recorded as 0, 'no shattering' or 1, 'pods opened and twisted'. PL was determined by measuring the length of ten representative pods for each accession. The GNPs were only calculated based on five representative pods in Hangzhou. Seed weight was measured with SC-G software (Hangzhou Wanshen Detection Technology). The fresh pods at the commodity period and the mature seeds were collected for pod quality and seed quality assessment, respectively. The soluble sugar and total starch content were assessed using the quality analysis kit following the manufacturer's instructions (Suzhou Keming). Crude protein was measured using the Kjeldahl method[133].

GWAS analysis

GWAS for the ten traits was performed using the SNP and InDel data under the efficient mixed-model association expedited (EMMAX) program in GEMMA (v.0.94.1)[134]. A kinship (*K*) matrix in the emmax-kin-intel package of EMMAX was used to correct the population structure. The significance threshold of SNP-trait associations was established with a false-detection-rate-adjusted $P < 0.05$ using the Benjamini–Hochberg procedure[135], which corresponds to an uncorrected *P* value of approximately 1.0×10^{-5}. Given that the genome-wide average distance of linkage disequilibrium decay ($r^2 = 0.40$) is 100 kb, adjacent GWAS loci within 100 kb were considered as a GWAS interval or signal. To compare our PL signals with the recently reported PL loci[19], we used MUMmer (v.4.0)[123] to perform whole-genome alignment among four genomes (G98, G323, IT97K-499-35 and A147) to determine their collinearity.

All genes located directly in or within 100 kb of the GWAS signal were selected as the putative genes for the GWAS loci. LDblockShow (v.1.32) was used to examine the local linkage disequilibrium of candidate regions. To determine the possible candidate gene for each signal, the SNPs and InDels inside the linkage disequilibrium block were sorted in ascending order of *P* value ($<10^{-5}$). Those genes close to or covering those SNPs or InDels with the lowest *P* values are the possible candidate genes. Information on gene annotation and function evidence of homologous genes in *Arabidopsis thaliana* or other plants were used to assist with the selection of putative target genes. Meanwhile, a haplotype analysis of the candidate gene was conducted to investigate the correlation between haplotypes and phenotypes at the population level to confirm its genetic effect on the target trait. Then the expression profiles of these genes during pod and seed development in different cowpeas were also considered as references for their function in the target trait.

CDS sequences of the possible candidate genes were used to blast the transcripts of *Arabidopsis thaliana* to search the homologous genes. In addition, the genomic syntenic analysis between G98 and *Arabidopsis thaliana* was also performed using diamond (v.0.9.29.130) and MCScanX (jcvi) to determine the collinearity blocks.

Sweep selection analysis

A cross-population composite likelihood ratio was calculated using the XP-CLR package (v.1.0)[136] with sliding windows of 100 kb and a step size of 10 kb. The selective sweeps were identified by comparing the G group versus the VL group for differentiation sweeps and the VL group versus the VC group for improvement sweeps. To further confirm the selective sweeps, we also investigated the F_{ST} value and π ratio in different subgroup comparisons by a slide window approach with a window size of 100 kb and a step of 10 kb using Vcftools (v.0.1.15). The top 5% of regions were assigned to candidate selective regions, and genes in these regions were considered candidate genes.

Expression of the candidate genes

Digital RNA-seq[137] was conducted for the developing pods and seeds from D413 (VC type), D445 (VL type) and D722 (G type) at 0, 5, 10 and 15 days after anthesis, respectively, with three biological replicates. In brief, total RNA was extracted using the QIAGEN RNeasy Plant Mini Kit (QIAGEN). The RNA-seq libraries were prepared using KC-DigitalTM Stranded mRNA Library Prep Kit for Illumina (Wuhan Seqhealth) following the manufacturer's instructions and each cDNA molecule was labeled using a unique molecular identifier of eight random bases[138]. Sequencing was performed on a DNBSEQ-T7 sequencer (MGI Tech). After filtering the raw data, clean reads were clustered according to the unique molecular identifier sequences, and consensus sequences were generated based on the sequences identified through pairwise alignment and multiple sequence alignment. The consensus sequences were aligned to the G98 genome using STAR software (v.2.5.3a) with default parameters to calculate the reads per kb per million reads. The expression patterns of candidate genes were displayed using edgeR package (v.3.12.1)[139]. Enrichment significance (*P* value) was calculated using the hypergeometric test (one-sided).

PL and TSW gene verification in bi-parental populations

An RIL population (183 lines)[15], created by single-seed descent from the cross of a VL 'ZN016' and a VC 'Zhijiang282', and an F_2 population (165 individuals), constructed by the cross of G98 and G323, were used for validation of the GWAS signals. Different alleles of *VuPL1* (HapI and HapII), *VuPL4* (HapI and HapII) and *VuTSW1* (HapI and HapIII) were observed in the parents of the RIL population. Different alleles of *VuPL1* (HapI and HapII), *VuPL2* (HapI and HapII), *VuPL3* (HapI and HapII), *VuPL4* (HapI and HapIII), *VuTSW1* (HapI and HapIII) and *VuTSW2* (HapI and HapIII) were observed in the parents of the F_2 population. The phenotypes of TSW and PL were investigated in Haining County (30° 32′ N, 120° 41′ E) in 2013 (RE3) and 2019 (RE4). The haplotypes of candidate genes were examined by KASP marker SNPs from the candidate genes of PL, and TSW signals were selected to convert into KASP markers and amplify in these two populations. KASP primer design and genotyping followed a previous publication[140].

Statistical tests used

Details of the statistics applied are provided in the figure legends. Pairwise comparisons were conducted using a two-tailed Student's *t*-test. Multiple comparisons were analyzed using the least significant difference method with Bonferroni correction. Statistical analyses and plotting were performed using Origin (v.9.0).

Reporting summary

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

Data availability

The genome assemblies of G98 (accession number JBALLC000000000) and G323 (accession number JAZDUG000000000) and the re-sequencing data of 344 accessions have been deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA889224; the RNA-seq data for gene annotation have been deposited in the NCBI Sequence Read Archive under BioProject accession number PRJNA954189; the transcriptome data of three cowpea accessions have been deposited in the NCBI Sequence Read Archive under BioProject accession number PRJNA970477. The genotype and phenotype data can be accessed in figshare (https://doi.org/10.6084/m9.figshare.21646556)[141].

Code availability

All codes and tools used in this study are described in Methods and Reporting Summary.

References

74. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
75. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
76. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
77. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
78. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
79. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
80. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
81. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, 1 (2019).
82. Wheeler, T. J. et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2013).
83. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
84. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
85. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **4**, 4.10.1–4.10.14 (2009).
86. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
87. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
88. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
89. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
90. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, E89 (2016).
91. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
92. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
93. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, E78 (2015).
94. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
95. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

96. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
97. Loman, T. *A Novel Method for Predicting Ribosomal RNA Genes in Prokaryotic Genomes*. MSc thesis, Lund Univ. (2017).
98. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
99. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006).
100. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
101. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
102. Song, J. M. et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* **14**, 1757–1767 (2021).
103. Swarbreck, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2008).
104. Kang, Y. J. et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* **5**, 5443 (2014).
105. Yang, K. et al. Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proc. Natl Acad. Sci. USA* **112**, 13213–13218 (2015).
106. Schmutz, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).
107. Garcia, T. et al. Comprehensive genomic resources related to domestication and crop improvement traits in Lima bean. *Nat. Commun.* **12**, 702 (2021).
108. Valliyodan, B. et al. Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J.* **100**, 1066–1082 (2019).
109. Varshney, R. K. et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89 (2012).
110. Qin, S. et al. A draft genome for *Spatholobus suberectus*. *Sci. Data* **6**, 113 (2019).
111. Kamal, N. et al. Insights into the evolution of symbiosis gene copy number and distribution from a chromosome-scale *Lotus japonicus* Gifu genome sequence. *DNA Res.* **27**, dass015 (2020).
112. Kreplak, J. et al. A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* **51**, 1411–1422 (2019).
113. De Vega, J. J. et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Sci. Rep.* **5**, 17394 (2015).
114. Young, N. D. et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011).
115. Varshney, R. K. et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246 (2013).
116. Hane, J. K. et al. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant–microbe interactions and legume evolution. *Plant Biotechnol. J.* **15**, 318–330 (2017).
117. 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).
118. Griesmann, M. et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, eaat1743 (2018).
119. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
120. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
121. Yang, Z. H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).
122. Yu, G. C., Wang, L. G., Han, Y. Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
123. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
124. Wang, K., Li, M. Y. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
125. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
126. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
127. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
128. 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).
129. Kumar, S. et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
130. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
131. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
132. Zhang, C., Dong, S. S., Xu, J. Y., He, W. M. & Yang, T. L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
133. Varelis, P. *Food Chemistry and Analysis* (Elsevier, 2016).
134. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
135. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
136. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
137. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl Acad. Sci. USA* **109**, 1347–1352 (2012).
138. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
139. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
140. Wu, X. Y. et al. Development of a core set of single nucleotide polymorphism markers for genetic diversity analysis and cultivar fingerprinting in cowpea. *Legume Sci.* **3**, e93 (2021).
141. Wu, Xi. & Li, G. Differential selection of yield and quality traits has shaped genomic signatures of cowpea domestication and improvement. *figshare* https://doi.org/10.6084/m9.figshare.21646556 (2024).

Acknowledgements
We thank J. Yang from Zhejiang University for valuable comments on the paper. This work was supported by the Key R&D Program of Zhejiang Province (2022C02016 to Xinyi Wu), the Major Science and Technology Project of Plant Breeding in Zhejiang Province (2021C02065 to B.W.), Key R&D Program of Guangdong Province (2020B020220002 to Xinyi Wu), the Crop Germplasm Identification Project of General Seed Management Station of Agriculture and Rural Affairs Department in Zhejiang Province (2022R23T60D01 to Xinyi Wu), the Fundamental Research Funds for the Central Universities (+226-2022-00100 to M.Z.) and Biological Breeding Project of ZAAS Program for Transdisciplinary Research (to Xinyi Wu).

Author contributions
Xinyi Wu, M.Z. and G.L. conceived and designed the project. Xinyi Wu, Z.H. and N.L. performed genome assembly and assessment, comparative genome analysis and other bioinformatic analyses. Xinyi Wu, B.W., Xiaohua Wu, Y.W., J.W., Z.L., Y.S. and W.D. performed field cultivation and phenotype investigation in Hangzhou, and Y.Z. and Y.Y. performed field cultivation and phenotype investigation in Guangzhou. M.L., J.D. and J.W. worked on quality testing and GWAS data analysis. J.D. performed gene expression analysis and gene function validation on the bi-parental population. Xinyi Wu and Z.H. wrote the paper, M.Z. revised the paper and all authors read and approved the 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/s41588-024-01722-w.

Correspondence and requests for materials should be addressed to Mingfang Zhang or Guojing Li.

Peer review information *Nature Genetics* thanks Caroline Belser and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Mingfang Zhang, Guojing Li.

Last updated by author(s): Oct 14, 2023

Reporting Summary

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

Statistics

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

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection	Thousand seed weight was measured with SC-G software (Hangzhou Wanshen Detection Technology).
Data analysis	Genome size estamation was performed using KAT (https://github.com/TGAC/KAT, v2.4.1). Genome assembly and assessment were performed using: hifiasm (v0.12), LACHESIS (v2.0), BWA (v0.7.8), Merquery (v1.3). Genome annotation was performed using: RepeatModeler (v1.05), Repbase (v19.06), REXdb (v3.0), Dfam (v3.2), LTRharvest (v1.5.9), LTR_finder (v1.1), RepeatMasker (v4.10), MIcroSAtellite (MISA v2.1), Augustus (v2.4), SNAP (v2013-11-29), GeMoMa (v1.7), Hisat (v2.0.4), Stringtie (v1.2.3), GeneMarkS-T (v5.1), EVM (v1.1.1), PASA (v2.2.0), tRNAscan-SE (v1.3.1), barrnap (v0.9), GenBlastA (v1.0.4), GeneWise (v2.4.1). Comparative genomics analysis was performed using: OrthoFinder (v2.3.9), IQ-TREE (v1.6.11), PAML (v4.9), CAFE (v4.2), ClusterProfiler (v3.18.0), diamond (v0.9.29.130), MCScanX (jcvi). Genome structural variation analysis was performed using MUMmer (v4.0), SyRI (v1.4), ANNOVAR package (v-2020-10-7) and ClusterProfiler (v3.18.0). SNP calling and population structure analysis were performed using: Burrows–Wheeler Aligner (bwa-mem2 v2.2), samtools (v1.9), GATK (v4.1.5.0), SnpEff (v5.1), VCFtools (V0.1.15), MEGA (V7.0), Eigensoft (V7.2.1), Admixture (v1.23), PopLDdecay (v3.27). GWAS was conducted using GEMMA (v0.94.1). LD block analysis was performed using LDblockShow (v1.32). Sweep selection analysis was performed using XP-CLR package (v1.0). The expression patterns of candidate genes were displayed using STAR (v2.5.3a) and edgeR package (v3.12.1). Pairwise comparisons were conducted using the two-tailed Student's t-test. Multiple comparisons were analyzed using the least significant difference (LSD) method with Bonferroni correction. Statistical analyses and plotting were performed using Original (v 9.0). All codes and tools used in this study are described in Methods.

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 raw sequencing data and the final assemblies of G98 and G323, and the resequencing data of 344 accessions are deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA889224, the RNA-seq data for gene annotation are deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA954189, the transcriptome data of three cowpea accessions are deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA970477. The genotype data is deposited in figshare (https://doi.org/10.6084/m9.figshare.21646556).

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	No statistical methods were used to determine sample size. The sample size of 344 cowpea accessions for resequencing were chosen based on accessions available in our lab to cover geographic distribution of cowpea. A RIL population containing 183 lines and a F2 population with 165 individuals were used for validation of the GWAS signals.
Data exclusions	No data were excluded from analysis in this study.
Replication	The phenotype evaluation of 344 accessions was conducted in two environments with two biological replicates. Pod length and thousand seeds weight evaluation of the RIL and F2 population were performed with two biological replicates in two years at one site. All replications were successful and reported in the manuscript.
Randomization	Plants were randomly allocated in the field.
Blinding	Blinding was not relevant for this study and phenotypic data were collected without known genetic data.

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	All the 344 cowpea accessions were saved in Zhejiang Academy of Agricultural Sciences.
Novel plant genotypes	No novel plant genotypes were produced.
Authentication	No relevant information.