Nature Communications

https://doi.org/10.1038/s41467-026-73494-2

Article in Press

Pan-genome analysis uncovers extensive structural variations in Africa's indigenous legume crop cowpea

Received: 15 October 2025

Accepted: 13 May 2026

Cite this article as: Sun, S., Wei, C., Pearson, S. *et al.* Pan-genome analysis uncovers extensive structural variations in Africa's indigenous legume crop cowpea. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-73494-2

Shichao Sun, Chuanzheng Wei, Sofie Pearson, Alan Cruickshank, Yinzi Wang, Yanbo Wang, Tashi Dorjee, Yi Zhou, David Jordan, Yan Yang, Hongru Wang, Xingtan Zhang, Emma Mace & Yongfu Tao

Pan-genome analysis uncovers extensive structural variations in Africa's indigenous legume crop cowpea

Shichao Sun[1,§], Chuanzheng Wei[1,§], Sofie Pearson[2,§], Alan Cruickshank[3†], Yinzi Wang[1,4,5], Yanbo Wang[1], Tashi Dorjee[1], Yi Zhou[1], David Jordan[2], Yan Yang[6], Hongru Wang[7], Xingtan Zhang[1], Emma Mace[2,3,*], Yongfu Tao[1,*]

[1]State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China

[2]Queensland Alliance for Agriculture and Food Innovation (QAAFI), The University of Queensland, Hermitage Research Facility, Warwick, QLD 4370, Australia

[3]Queensland Department of Primary Industries (DPI), Hermitage Research Facility, Warwick, QLD, 4370, Australia

[4]School of Life Sciences, Henan University, Kaifeng 475004, China

[5]Shenzhen Research Institute of Henan university, Shenzhen 518000, China

[6]Tropical Crops Genetic Resources Institute, National Key Laboratory for Tropical Crop Breeding, Chinese Academy of Tropical Agricultural Sciences, Haikou/Sanya, Hainan 571101/572024, China

[7]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China

[§]These authors contributed equally.

[†]Deceased: Alan Cruickshank.

[*]These authors jointly supervised this work.

*Corresponding authors: Yongfu Tao (taoyongfu@caas.cn); Emma Mace (emma.mace@uq.edu.au)

Abstract

Cowpea (*Vigna unguiculata* L. Walp.) is an Africa-originated legume providing a vital source of protein for millions across resource constrained areas of Africa, Asia and Latin America. Understanding genetic diversity in cowpea is pivotal to its genetic improvement. Here, we assemble 20 high-quality genomes representing cowpea global diversity and construct a graph-based pan-genome to characterize genetic variation. The pan-genome comprises 29,557 orthogroups, with approximately 73% conserved across all accessions. Extensive structural variations (SVs) are uncovered, affecting thousands of coding sequences and gene expression. Thousands of SVs are found under selection between two cowpea subspecies, highlighting their roles in subspecies divergence. We pinpoint key SVs underlying variation in pod length and seed number per pod. Two SV clusters appear to modulate pod length through altering expression and function of *VuWAK* and *VuGA2ox2* genes. These findings provide insights into cowpea genetic diversity and establish a fundamental resource for cowpea genetic improvement.

Introduction

Africa is facing an unprecedented hunger and malnutrition crisis due to food shortages (FAO)[1]. Indigenous crops, with their adaptation to local environments, play a pivotal role in ensuring regional food security. Among these, cowpea (*Vigna unguiculata* L. Walp., 2n = 22) is the second most widely cultivated legume in Africa, owing to its resilience in hot and drought-prone environments[2]. As a legume crop with efficient nitrogen fixation, cowpea provides a critical source of protein for hundreds of millions of people across sub-Saharan Africa, East Asia and Latin America[3,4].

Substantial variation in morphological traits has been observed in cowpea[5]. As a result, two major cowpea subspecies have been identified, including the grain-type *V. unguiculata* ssp. *unguiculata* (Vu), predominantly used for dry seed production in Africa, and the vegetable-type *V. unguiculata* ssp. *sesquipedalis* (Vs), favored for its uniquely tender and elongated pods in East Asia[6]. The two subspecies show marked differences in a range of agronomic traits, including plant architecture, growth habit and pod morphology. Improving cowpea productivity through breeding relies on the effective exploration and utilization of cowpea genetic diversity. However, the underlying genetic diversity in cowpea remains largely untapped[7].

Genetic variation is the raw material upon which human selection acts to improve crop productivity[8]. The assembly of reference genomes and subsequent re-sequencing studies have reported millions of sequence variations in cowpea by mapping short-read sequences to the cowpea reference genome[6,9,10]. Evidence suggests a single reference genome is inadequate to fully capture genetic variation within a species[11-13]. Reference-based genome approaches often overlook gene content variation and structural variations (SVs), which are increasingly recognized as key drivers of trait variation and crop evolution[14,15]. Characterizing these types of variation and their impact is critical for cowpea improvement and breeding. Large structural variants have been reported in cowpea through a recent comparison of six cowpea genomes assembled using short read

sequencing[16]. While informative, the genotypes selected represent only a subset of the global diversity of cowpea germplasm (Supplementary Fig. 1).

Here, we assemble 20 high-quality cowpea genomes representative of global cowpea diversity and construct a pan-genome to characterize gene content variation and SVs. Leveraging the graph-based cowpea pan-genome, we explore the potential roles of SVs in subspecies divergence and the genetic regulation of key agronomic traits**.** This study deepens our understanding of genetic diversity in cowpea and offers a valuable resource to accelerate its improvement and breeding.

Results

High-quality *de novo* genome assemblies and annotation of cowpea accessions

To capture the diversity in cowpea, SNP variation within 9,609 accessions were assessed from a global collection sourced from five major global genebanks (see Methods). A total of 20 accessions representing the genetic diversity of this collection were selected for genome sequencing (Fig.1a, b, and Supplementary Figs. 1-3). Six previously published high-quality cowpea genomes (NJ, A147, G98, G323, IT97K-499-35, and FC6)[6,10,17-19] were also utilized to develop a globally diverse cowpea pan-genome consisting of 26 genomes in total. These 26 accessions included six accessions from the Vs subspecies and 20 accessions from the Vu subspecies (Fig.1c and Supplementary Data 1).

Cowpea is a predominantly self-pollinating, cleistogamous crop with low heterozygosity (Supplementary Table 1). A total of 408.23 gigabases (Gb) of PacBio circular consensus sequencing reads were generated for the 20 cowpea accessions with an average depth of 36× for each accession (Supplementary Data 2). Each genome was assembled individually using Hifiasm. The resulting genomes ranged from 557.03 to 606.59 megabases (Mb) in size, with contig N50 values > 24.24 Mb (Table 1). Genome completeness assessments revealed that the 20 newly generated genomes captured an average of 98.55% of benchmarking universal single-copy

orthologues (BUSCO)[20] with assembly consensus quality values (QV)[21] ≥ 59.65 (Table 1). Genome error rates were assessed using the Clipping Reveals Assembly Quality[22] (CRAQ) analysis, generating regional assembly quality indicators (R-AQI) and overall assembly quality indicators (S-AQI) averaged at 98.83 and 98.40, respectively, for the 20 genomes (Supplementary Data 1). These quality metrics indicated the 20 assembled genomes are high-quality. A telomere-to-telomere (T2T) genome of the cowpea elite cultivar FC6 was used as a reference in the subsequent analyses[6,23]. Reference-guided scaffolding for the 20 selected accessions achieved an average anchoring rate of 95.75% (Table 1). To evaluate chromosome anchoring and orientation, three represent accessions (Vung05, Vung14, and Vung17) were selected and subjected to Hi-C sequencing. Collinearity analysis between the Ragtag scaffold assemblies and the Hi-C anchored assemblies showed high concordance and no obvious structural inconsistencies (Supplementary Fig. 4), supporting the reliability of the assemblies. Compared to the previously published cowpea pan-genome[16], our assembled genomes exhibit a longer contig N50 and a higher BUSCO value (Supplementary Table 2). A high degree of genome consistency was observed among the 26 cowpea accessions (Fig. 1d) with an average synteny relationship index[24] value as high as 0.94.

To minimize the annotation discrepancies caused by different pipelines, the 26 cowpea genomes were annotated using a standardized approach integrating *ab initio* gene prediction, homology-based searches and RNA-seq evidence. An average of 29,297 protein-coding genes were annotated for these assemblies, with a mean BUSCO completeness of 98.70% (Table 1). These annotated genes were well supported by expression and homology evidence. Over 90% of the annotated genes had RNA-seq matches (Supplementary Table 3). An average of 87.09% of annotated genes showed matches with the known functional databases, including Pfam[25], SwissProt[26], GO[27], and KEGG[27] (Supplementary Table 4). Transposable elements (TEs) accounted for an average of 42.26% of the genome (range: 39.61–44.72%; Fig. 1e and Supplementary Fig. 5), and their abundance increased with genome size (R = 0.61, p = 0.04; Supplementary Fig. 6). Taken together, these data suggested consistent and robust gene annotations.

Pan-genome construction and diversity analysis in cowpea

To investigate gene content variation in cowpea, we conducted a pan-genome analysis of the 26 genomes. A gene-based pan-genome analysis demonstrated that cowpea has a closed pan-genome, as the number of orthogroups plateaued around 19 genomes (Fig. 2a). A total of 29,557 orthogroups were identified in the developed cowpea pan-genome. The cowpea pan-genome was found to be relatively conserved, comprising 21,620 core orthogroups (73.2%) shared among all 26 genomes. By contrast, 7,867 dispensable orthogroups (26.6%) were found between 2 to 25 genomes, while 70 private orthogroups (0.3%) appeared in only one genome (Fig. 2b and Supplementary Fig. 7). On average, core genes accounted for 78.7% of genes in the cowpea genome (Fig. 2c). The high proportion of core genes in the pan-genome and individual genomes of cowpea mirrors a pattern observed in cultivated soybean and mung bean[28,29].

Core genes exhibited longer coding DNA sequences with lower nucleotide diversity (π) and non-synonymous to synonymous substitution ratios (K_a/K_s) compared to dispensable and private genes (Fig. 2d-f), suggesting their greater functional conservation. Expression analysis revealed that core genes had the highest mean expression levels, with dispensable genes showing moderate expression and private genes the lowest (Fig. 2g). Further functional enrichment analysis showed that core genes were enriched in fundamental biological processes, including nucleobase-containing compound metabolism, cellular processes, and cellular component organization (Fig. 2h). By contrast, dispensable and private genes were enriched with stress-related functions, such as response to biotic stimuli, response to external stimuli, and secondary metabolic processes (Fig. 2i). This indicates the fast-evolving dispensable genes could play a pivotal role in environmental adaptation in cowpea.

Genomic landscape of structural variations in cowpea

Given the critical roles that SVs play in regulating agronomic traits and adaptation[30], we investigated the genomic landscape of SVs in cowpea using our high-quality genomes. By comparing the 25 genomes with the FC6 reference, we identified a total of 62,591 non-redundant SVs, including 29,625 insertions (INS), 25,489 deletions (DEL), 368 inversions (INV), 4,096 duplications (DUP), and 3,013 translocations (TRANS) (Fig. 3a). These SVs were unevenly distributed across 11 chromosomes with an enrichment in heterochromatin regions ($p = 1.56\times10^{-8}$, Wilcoxon test, Supplementary Fig. 8 and Supplementary Table 5). An average of 9,358 SVs were detected in each genome with Vu accessions containing more SVs than Vs accessions, probably due to the closer relationship between Vs and the FC6 reference (Fig. 3b and Supplementary Fig. 9). The number of pan-SVs has not reached a plateau with the 26 genomes, indicating that more SVs could be detected when extra genomes were included (Fig. 3c). We randomly selected 20 SVs located within genic region for PCR validation, all of which were confirmed (Supplementary Fig. 10 and Supplementary Table 6), demonstrating the high accuracy of the SV detection.

These SVs covered 329.2 Mb of genomic sequence, equivalent to > 50% of the reference genome. Significant differences in variation length were observed among the five types of SVs, with DUP, INV, and TRANS being longer than DEL and INS (Fig. 3d and Supplementary Fig. 11a). INV, which accounted for only 0.6% of the total SV number, contributed to 44.13% of the total SV length (Supplementary Table 7). There were 26 SVs larger than 1 Mb, each encompassing multiple genes. They included three TRANSs, seven DUPs, and 16 INVs (Supplementary Table 8). For example, a 4.8 Mb inversion containing 85 genes was detected on chromosome 10 (Fig. 3e, Supplementary Fig. 12 and Supplementary Data 3). These 85 genes included a cluster of homologs of the powdery mildew resistance gene *RUN1*[31] (*Vu10G0620*, *Vu10G0623*, *Vu10G0625-Vu10G0627*, *Vu10G0630*, *Vu10G0641* and *Vu10G0642*), a seed maturation-related gene *VuMYB118* (*Vu10G0599*)[32], and a petal morphogenesis gene *VuAS1* (*Vu10G0648*)[33]. Three haplotypes of this SV were identified, which showed a clear separation alongside the phylogenetic

tree (Fig. 3f and Supplementary Fig. 13). Most of the cowpea SVs were low-frequency variation, with over 60% of them present in only one or two accessions (Supplementary Fig. 11b).

To understand the potential functional impact of SVs, SVs were annotated based on their location relative to the protein-coding genes. Although the majority (62%) of SVs were located in the intergenic region, a total of 18,946 genes (including gene bodies and their 5-kb flanking regions) overlapped with 24,220 SVs (Fig. 3g and Supplementary Fig. 14). SVs caused marked alteration of gene coding sequence with 900 high-impact gene-disrupting events identified, including transcript ablations, frameshift mutations, and gain or loss of stop codons (Supplementary Table 9). Further tests demonstrated that SVs also affected expression of their overlapping genes ($p \leq 2.2\times10^{-16}$, Wilcoxon test) (Supplementary Fig. 15). Gene ontology enrichment showed that the genes which overlapped with SVs were enriched in biological processes related to environmental adaptation, such as response to stress, cell communication, signal transduction, and secondary metabolic process (Supplementary Data 4).

The observed enrichment of SVs in heterochromatic regions prompted an investigation into the factors shaping their distribution. We found TEs were enriched in SV surrounding regions in cowpea ($p \leq 2.2\times10^{-16}$, Wilcoxon test; Fig. 3h). A sliding window test showed that the number of SVs in each window was positively correlated with the number of TEs, particularly long terminal repeat retrotransposon (LTR) and terminal inverted repeat transposons (TIR) (SV versus TE: R = 0.26, $p = 3.78\times10^{-10}$; SV versus LTR: R = 0.44, $p \leq 2.0\times10^{-16}$; SV versus TIR: R = 0.12, p = 0.01) (Fig. 3i and Supplementary Fig. 16). Further investigation determined that LTR was the dominant type of TE surrounding all four types of SVs (Supplementary Fig. 17). The evidence indicates that TE movement is a major driver of SV formation. SVs were classified into TE-derived and non-TE-derived groups according to their overlap with TEs to evaluate their differences. Although TE-derived SVs were significantly longer than non-TE-derived SVs, they were less likely to co-locate with genes (Fig. 3j). In addition, a significant negative correlation was observed

between SV number and recombination rate in cowpea (R = 0.56, $p \leq 2.2\times10^{-16}$) (Supplementary Fig. 18a), indicating that high recombination rates may facilitate SV purging given the generally deleterious effects of SVs on gene function. SV number were significantly influenced by recombination rate, especially INS and DEL (Supplementary Fig. 18b). Overall, our evidence indicated that movement of TEs and recombination could influence SV distribution.

The cowpea pan-genome uncovers structural variations associated with subspecies divergence

The cowpea pan-genome presents an excellent opportunity to investigate the roles of SVs in the divergence of two cowpea subspecies. To enable accurate genotype calling in the resequencing samples, a graph-based pan-genome of 26 genomes was constructed. The 775.55-Mb cowpea pan-genome incorporated extensive genetic variation, including 5,412,029 single-nucleotide polymorphisms (SNPs, 1 bp), 545,048 insertions and deletions (indels, 2-49 bp), and 69,245 SVs (≥50 bp), with a strong overlap between the SVs identified by the pan-genome method and those detected by SyRI (Supplementary Fig. 19). Short-read sequences of 619 diverse cowpea accessions from previous studies[6,10] were mapped to the graph-based cowpea pan-genome, identifying 61,612 SVs and 4,248,171 SNPs. Population structure analysis identified 89 Vu accessions and 235 Vs accessions with group membership >70%as group representatives for subsequent analyses (Supplementary Fig. 20).

These racial representatives were subsequently employed to identify genomic regions underlying subspecies divergence through an SV-based genome-wide association study (SV-GWAS) with racial identity as the phenotype, resulting in the identification of 126 genetic loci (Fig. 4a and Supplementary Data 5). A complementary method integrating a Pi-ratio, fixation index (F_{ST}), and a cross-population composite likelihood ratio (XP-CLR) was also employed to screen genomic regions under selection between the Vs and Vu subspecies (Fig. 4b, and Supplementary Fig. 21). A total of 299 genomic regions covering 30.06 Mb sequences and 1,851 protein-coding genes were

identified (Fig. 4b and Supplementary Data 6). Combining the two approaches resulted in the identification of 2,374 non-redundant candidate genes associated with cowpea subspecies divergence (Supplementary Data 7). These candidate genes were enriched in various biological processes, including responses to various stimuli (external, biotic, stress, light), cellular processes (programmed cell death, protein modification, secondary metabolism, catabolic processes), and regulation of molecular function (Supplementary Fig. 22 and Supplementary Data 8), indicating that the divergence between the two subspecies involves multiple complex biological processes. Comparing these results with previous SNP-based selective sweep analyses reveals that 21.5% of the identified candidate genes overlap (expected overlap ≈ 76 genes; hypergeometric test, $p < 1 \times 10^{-100}$; Supplementary Fig. 23a,b)[10], supporting a strong concordance between the two studies. Functional annotation of the non-overlapping genes shows an enrichment of biological functions related to environmental adaptation, supporting their role in racial divergence (Supplementary Fig. 23c).

Multiple candidate genes in the differentiated regions were associated with the distinct agronomic traits between the two cowpea subspecies (Fig. 4a,b). *VuDAO* (*Vu02G1064*), encoding a 2-oxoglutarate-dependent dioxygenase, is a candidate for regulating pod development and length through auxin biosynthesis[34]. *VuBEE1*(*Vu04G0211*) is a basic helix-loop-helix type transcription factor, controlling photoperiod flowering in *Arabidopsis thaliana*[35]. *VuTAC1* (*Vu10G1353*), encoding a tiller angle control protein, regulates lateral branch angles through influencing polar auxin transport in peach[36]. Likewise, *Vu4CL1* (*Vu06G0638*), encoding a *4-coumarate CoA ligase*, is a crucial gene in lignin biosynthesis in *Arabidopsis* and may contribute to plant architecture and erect growth habit[37]. Interestingly, the cowpea homolog of the *Arabidopsis thaliana* circadian rhythm and flowering regulator gene *EARLY FLOWERING 4* (*AtELF4*)[38], *VuELF4* (*Vu01G2216*), lies within a differentiated region at 47.97 Mb on chromosome 1 (Fig. 4c,d). A 78-bp deletion in the *VuELF4* promoter, encompassing the core promoter element TATA-box and the light-responsive GA-motif, was identified, implying an alteration in gene expression (Fig. 4c).

Expression analysis showed that *VuELF4* is predominantly expressed in cowpea leaves, and this deletion reduces its expression level (Fig. 4e), potentially leading to delayed flowering[39]. This deletion was present in the 50% Vu accessions, but was fixed in the Vs accessions (Fig. 4f), suggesting a potential role in environmental adaptation.

Allele frequency analyses of SVs identified 578 and 109 SVs that were fixed in the Vu and Vs subspecies group, respectively (Supplementary Fig. 24 and Supplementary Data 9). A total of 89 genes were identified that overlapped with these fixed SVs. Functional enrichment of these genes showed they were related to flower development, responsiveness to endogenous stimulus, lipid metabolic processes, and transport (Supplementary Data 10), suggesting they may be critical in the adaptation of the two subspecies to different environmental conditions and human selection pressure. These fixed SVs provide critical resources for functional studies to elucidate molecular mechanisms underlying cowpea subspecies divergence.

Structural variations underlying key cowpea agronomic traits

To investigate the role of SVs in genetic regulation of agronomic traits in cowpea, we conducted SV-based GWAS of pod length (PL) and seed number per pod. Eight QTL associated with pod length were identified, located on chromosome 3 (*PL 3.1*), chromosome 5 (*PL 5.1*, *PL 5.2* and *PL 5.3*), chromosome 8 (*PL 8.1* and *PL 8.2*), chromosome 9 (*PL 9.1*), and chromosome 10 (*PL 10.1*) (Fig. 5a, Supplementary Fig. 25a and Supplementary Table 10). *PL3.1* and *PL9.1* overlapped with pod length QTL reported in a previous study[6], while the other QTLs are located outside previously reported pod-length regions. Six of these QTL that were significant related to pod length were further validated using a haplotype-based analysis (Fig. 5b and Supplementary Fig. 26). Accessions carrying more favorable alleles of the six loci exhibited longer pods (Fig. 5c and Supplementary Fig. 27), indicating potential pyramiding effects on regulating pod length.

Candidate genes and SVs in these pod length QTL were examined to identify possible causal variations. A tandem duplication consisting of five *Wall-Associated Receptor Kinase* (*VuWAK*) genes, involved in the regulation of cell shape and size and affect silique length in *Arabidopsis thaliana*[40], was found in *PL9.1* with the highest significance signal in GWAS analysis (Fig. 5d). Two deletions (SV9:35552732, 1,067 bp; SV9:35554053, 68 bp) in the promoter of *VuWAK-1* and two insertions (SV9:35563200, 4,080 bp; SV9:35563251, 205 bp) in the promoter of *VuWAK-2* were identified, each having significant effects on pod length (Fig. 5d,e). Two *VuWAK* genes were predominantly expressed in pods, and genotypes carrying different alleles of these SVs exhibited variable expression levels of the two *WAK* genes during pod development (Supplementary Fig. 28). Luciferase reporter assays showed that the SV9:35552732 deletion in the promoter region of *VuWAK-1* markedly enhanced promoter activity, whereas the SV9:35563251 insertion in the *VuWAK-2* promoter reduced promoter activity (Supplementary Fig. 29).

Four haplotypes (Hap1-4) were defined by the four SVs due to high LD among these physically close SVs (Fig. 5d). Significant differences in pod length were observed among these four types of haplotypes with Hap4 accessions displaying the longest pods (Fig. 5f). Subspecies Vs exhibits significantly longer pods than subspecies Vu, reflecting human-mediated selection for longer pods for fresh consumption (Supplementary Fig. 30). Hap3 and Hap4 were mainly found in the Vs population, while more than half of Hap1 and Hap2 were present in Vu accessions (Fig. 5g). These results suggest that the four SVs may contribute to pod length difference between Vu and Vs through altering expression of the two *WAK* genes.

Additionally, a tandem duplication comprising two *VuGA2ox2* genes (*Vu05G1623* and *Vu05G1624*) was found at the second highest significance loci (*PL5.2*) (Fig. 5h). GA2ox2 encodes a key enzyme that catalyzes the deactivation of bioactive gibberellins, which negatively regulates pod length in *Arabidopsis thaliana*[41,42]. Two SVs including a 250-bp insertion in the downstream of *Vu05G1623* and a 51-bp insertion in the third exon of *Vu05G1624* were significantly correlated

with pod length (Fig. 5i). Alignment of the *Vu05G1624* sequence in the pan-genome and its ortholog in related *Vigna* species demonstrated that this 51-bp DNA fragment is absent in nearly all Vs accession, but present in all Vu accessions and related *Vigna* species (Supplementary Fig. 31). Protein structure modelling showed the absence of this 51-bp sequence in *VuGA2ox2-2* led to the deletion of a *β*-sheet in 3D structure of the encoded protein, which reduced its binding affinity to bioactive gibberellins (Fig. 5j and Supplementary Table 11). Consistently, overexpression of *VuGA2ox2-2* containing the 51-bp insertion (VuGA2ox2-2-ALT) in *Arabidopsis thaliana* significantly reduced pod length, whereas overexpression of the reference type (VuGA2ox2-2-REF) showed no significant effect, suggesting that the insertion may reduce overall gibberellin levels and consequently inhibit pod elongation (Fig. 5k,l). Three haplotypes combining the two SVs showed significant effects on pod length regulation (Fig. 5m). The distinct subspecies composition in the three haplotypes supports the combination of two SVs underlying the variation of pod length between the two subspecies (Fig. 5n).

Seed number per pod is one of the key yield components of grain cowpea[43]. Our SV-based GWAS of seed number per pod identified three QTLs, including *seed number per pod 4*, which overlaps with a previous reported QTL[6]. While the previous study could not pinpoint a causal gene, our graph-based pan-genome provided the resolution to identify a high-confidence candidate. The strong association signal on chromosome 4, which harbored *Vu04G0313* encoding a pentatricopeptide repeat (PPR) protein (Fig. 6a,b and Supplementary Fig. 25b). *VuPPR* is a homolog of the *EMBRYO-DEFECTIVE* gene *AtEMB1586*[44], that controls embryo development and fertility rate in *Arabidopsis thaliana* (Supplementary Fig. 32). A 83-bp deletion was found in the exon of *VuPPR* causing a premature stop codon at the 134th amino acid (Fig. 6c). Accessions with the deletion exhibited significantly lower seed number per pod (Fig. 6d). The two cowpea subspecies had pronounced differences in allelic composition of this SV (Fig. 6e). The deletion was predominately found in Vu accessions, consistent with their lower seed number per pod (Fig.

6f). These results suggest that the 83-bp deletion in *VuPPR* may influence variation of seed number per pod in cowpea.

Discussion

Genetic improvement of cowpea holds the key to ensuring food security in the sub-Saharan region, which in turns relies on effective utilization of genetic diversity within global germplasm[1,4,11]. In this study, our construction of the graph-based cowpea pan-genome using 20 new *de novo* assemblies and six existing genomes represents a paradigm shift from single-reference genomics to a population scale understanding of cowpea. Our study systematically characterizing gene content variation and structural variations. We found that approximately 73% of orthogroups in cowpea pan-genome are core genes, indicating a relatively conserved gene composition in cultivated cowpea. This high core gene proportion in cowpea is comparable to other self-pollinating legumes like soybean and mungbean, yet stands in stark contrast to the highly fluid pan-genomes of outcrossing cereals such as maize[13,28,29,45], possibly reflecting unique features of legumes in terms of its cleistogamous mating system and specific TE composition[30]. However, the dispensable genome, which comprises 26.6% of orthogroups, is significantly enriched in stress-responsive and secondary metabolic pathways. This suggests that while the core genes remains rigid to maintain fundamental biological processes, the accessory genes act as a dynamic repository of genetic variation, possibly facilitating cowpea's renowned resilience in the semi-arid tropics.

While SNPs have long been the focus of cowpea genetic studies[5], our identification of 62,591 non-redundant SVs reveals a hidden layer of diversity, affecting over 50% of the reference genome. We observed that SV distribution is positively correlated with TE density (specifically LTRs) and negatively correlated with recombination rates. Consistent with the patterns observed in soybean[11] and *Brassica oleracea*[46], these SVs in cowpea are predominantly localized in intergenic regions. This distribution is primarily attributed to strong purifying selection against deleterious mutations in genic regions, which limits SV accumulation in functional sequences[47]. Conversely, the high density of TEs in intergenic spaces facilitates SV formation via unequal crossing over[48]. This genomic landscape aligns with the sheltering hypothesis, where low-recombination

heterochromatic regions serve as refugia for large SVs and TE insertions that might otherwise be purged in high-recombination regions[49,50,51]. The discovery of a 4.8 Mb inversion on chromosome 10, which sequesters a cluster of powdery mildew resistance (*RUN1*) and seed maturation (*VuMYB118*) genes, suggests that SVs play a critical role in maintaining adaptive gene complexes that are protected from crossover-mediated disruption[52,53].

The divergence between Vu and Vs cowpea provides a compelling model for human-mediated selection for distinct agricultural niches[54]. Our SV-based GWAS and selection scans identified over 2,300 candidate genes underlying subspecies divergence, highlighting pronounced genetic differentiation. The fixation of a 78-bp deletion in the *VuELF4* promoter in Vs accessions highlights the role of circadian rhythm regulation in subspecies divergence. By reducing the expression of this light-responsive regulator, this SV may have facilitated the delayed flowering and environmental adaptation required for vegetable-type cultivation in East Asia[38].

Structural variations have been shown to be major sources of gene functional variation and play a critical role in the genetic regulation of agronomic traits in a range of species[55-58]. We identified two distinct structural mechanisms driving the elongated pod phenotype in Vs. First, SVs in the promoters of *VuWAK* genes likely modulate expression to alter cell shape and size[59]. Second, a 51-bp deletion in the exon of *VuGA2ox2-2*, which fixed in the Vs subspecies, results in a structural alteration of the protein. This modification potentially reduces the enzyme's affinity for bioactive gibberellins, thereby increasing hormonal levels to promote pod elongation[41,42]. Compared to previous studies[6,60], our SV-based GWAS analysis identified novel genetic loci and possible causal mutations controlling important traits, emphasizing that SVs could complement SNPs in genetic analysis. The identified SVs in the cowpea pangenome and their associations with key agronomic traits offer a valuable foundation for genetic improvement through the pyramiding of favorable alleles.

Our study confirms a closed cowpea pan-genome with a high proportion of core genes, consistent with previous report[16]. More importantly, we constructed a graph-based pan-genome that captures complex structural variations underrepresented in earlier short-read-based studies[16], and further explored their functional relevance through SV-based genome-wide association analyses (SV-GWAS) and candidate gene analysis, representing a substantial advance. However, limitations remain. The absence of wild *Vigna* limits our ability to capture genetic variation lost during domestication. In addition, although the statistical associations identified are robust, high-throughput functional validation, such as CRISPR-based approaches, will be essential to establish causality for candidate genes.

In conclusion, our study establishes a comprehensive, graph-based pan-genomic resource for cowpea, providing a fundamental framework for understanding its global genetic diversity. We demonstrate that SVs are key drivers of subspecies divergence and agronomic trait variation, influencing key biological processes and gene functions. These resources constitute an essential molecular toolkit to accelerate the development of climate-smart varieties, ensuring food security in resource-constrained regions.

Methods

Sample selection and sequencing

The 20 cowpea accessions were selected from a large diversity analysis of 9,609 accessions from multiple genebanks and collections globally[61], including the United States Department of Agriculture (USDA), International Institute of Tropical Agriculture (IITA), Australian Grains Genebank (AGG), the Japanese National Agriculture and Food Research Organization (NARO), the Mozambique Genebank and other previously established collections including the UCR mini-core[62]. The accessions were sequenced using the DArTseq™ reduced-representation sequencing platform, resulting in the identification of 2,302 high quality SNPs. A maximum likelihood phylogeny was subsequently constructed with IQ-TREE (v2.2.2.3)[63].

Based on phylogenetic relationships, geographic origin, and subgroup clustering information, 20 representative cowpea accessions were selected for genome sequencing. These accessions originated from 11 countries, encompassing Venezuela, Mexico, Thailand, India, Philippines, Afghanistan, Japan, South Africa, Nigeria, Madagascar, and Morocco, representing a broad geographical distribution across the Americas, Asia, and Africa. Genomic DNA isolated from young leaves of the 20 cowpea accessions were used to construct PacBio SMRT libraries for each sample, following the recommended standard protocols. In briefly, high-quality genomic DNA was first mechanically sheared into fragments of approximately 15 - 20 kb using g-TUBE devices. The resulting DNA fragments underwent a series of enzymatic reactions for DNA damage repair (nick repairing) and end repair, followed by the ligation of SMRTbell adapters to both ends of the double-stranded DNA. To ensure high library purity, the ligation products were treated with exonucleases to selectively degrade failed or incompletely ligated fragments. Subsequently, the libraries were subjected to size selection using the BluePippin system with a target threshold of >10 kb to enrich for long-insert fragments. The final SMRTbell libraries were assessed for size distribution and concentration using the Agilent 4200 TapeStation and Qubit Fluorometer, respectively. Sequencing was performed on the PacBio Sequel II platform in Circular Consensus Sequencing mode using 8M SMRT Cells. A 30-hour movie time was utilized to ensure sufficient passes for each molecule, generating a total of 15.01 - 24.45 Gb of high-fidelity (HiFi) reads ($Q \geq 20$) per accession.

Young leaf tissues were collected from Vung05, Vung14, and Vung17 for Hi-C sequencing. Samples were cross-linked with 1% for maldehyde and digested using Dpn II. DNA ends were biotin-labeled and proximity-ligated to capture spatial interactions. Following streptavidin-mediated enrichment and library preparation, sequencing was performed on the Illumina NovaSeq 6000 platform using a PE150 strategy. Valid interaction pairs were identified for chromosome-level scaffolding analysis.

Genome assembly, annotation and quality assessment

We assembled the genomes of 20 cowpea accessions sequenced with HiFi reads. Briefly, HiFi reads were firstly assembled using Hifiasm (v0.20.0-r639) with default parameters[64]. The resulting assemblies were anchored and oriented onto the chromosomes of the FC6 reference genome using the reference-guided scaffolding tool RagTag[65] (v2.1.0) with default settings. Structural accuracy of these assemblies was assessed using HiFi reads and Hi-C data of Vung05, Vung14 and Vung17 , which detected no evidence of structural inconsistencies or misassemblies. The quality of genome assemblies was evaluated using BUSCO[20] (v5.7.1) with the *fabales_odb10* lineage dataset, Clipping Reveals Assembly Quality[22] (CRAQ, v1.0.9-alpha), and Merqury[21] (v1.4.1).

TEs were annotated using the Extensive *de novo* TE Annotator (EDTA, v2.2.0), which integrates multiple TE prediction tools[66]. Briefly, long terminal repeat retrotransposons (LTR-RTs) were initially identified using LTR_FINDER_parallel (v1.0.7) and LTR_HARVEST (v1.1), followed by quality filtering and refinement with LTR_retriever (v2.9.4)[67-69]. Terminal inverted repeat (TIR) elements were detected using TIR-Learner (v3.0), while Helitron elements were annotated with HelitronScanner[70] (v1.1). The remaining TE components were identified using RepeatModeler (v2.0.5) and further annotated through homology-based searches with RepeatMasker (v4.1.5), incorporating the TE library generated by EDTA[71,72]. The combined results from all tools were integrated to generate the final comprehensive TE annotation. Gene structures were predicted for each genome assembly using the BRAKER3 pipeline[73] (v3.0.8), which integrates *ab initio* gene predictions, transcript evidence, and homologous protein evidence. To improve gene model prediction accuracy, GeneMark-ETP (v1) was first trained using both transcript and protein evidence, followed by training of AUGUSTUS (v3.5.0) based on the GeneMark-ETP predictions[74]. Protein homology evidence was derived from the UniProt90 database[26]. Publicly available cowpea RNA-seq data from Wu *et al*.[6] (NCBI accessions: PRJNA970477 and PRJNA954189) were used to support gene annotation. These datasets include roots, seeds, leaves, flower buds, and pods at 0,

4, 8, and 12 days after anthesis for the two cowpea subspecies. The reads were aligned to the each soft-masked genomes using HISAT2[75] (v2.2.1).

The quality of gene annotation was assessed using BUSCO[20] (v5.7.1) with the *fabales_odb10* dataset. Functional annotation of predicted protein-coding sequences was performed using BLASTP searches against the Swiss-Prot database. Conserved protein domains were identified using InterProScan[76] (v5.45-80.0). Gene Ontology (GO) terms (http://geneontology.org/) and KEGG pathway annotations (https://www.genome.jp/kegg/) were assigned using eggNOG-mapper[77].

Phylogenetic tree, and collineariry analysis

Six previously published cowpea genomes[6,10,17-19] (NJ, A147, G98, G323, IT97K-499-35, and FC6) with comparable quality were combined with the 20 assembled genomes for following genome analysis. To investigate the phylogenetic relationships among the 26 cowpea accessions, *Vigna mungo* was used as an outgroup[78]. OrthoFinder[79] was used to group protein-coding genes into orthogroups and a total of 14,174 sigle-copy orthologous genes were identified. A maximum likelihood (ML) phylogenetic tree was constructed based on the concatenated sequences of single-copy orthologous genes using the RAxML[80] software package (v8.2.12) with 1,000 bootstrap replicates. Pairwise genome synteny analysis was performed using the JCVI[81] package (v1.4.22). The syntenic relationship index (SRI) was calculated to quantify the level of synteny between two genomes (https://github.com/Yujiaxin419/SRI-Pipeline).

Gene-based pan-genome analysis

To construct the gene-based pan-genome, protein sequences from 26 cowpea accessions were analyzed using OrthoFinder[79] (v2.5.5), which applies the Markov Cluster Algorithm (MCL) to group protein-coding genes into orthogroups. The resulting orthogroups were categorized as core (present in all 26 genomes), dispensable (present in 2–25 genomes), or private (unique to a single

genome). The nonsynonymous to synonymous substitution ratio (*Ka*/*Ks*) for each orthogroup category was estimated using KaKs_Calculator[82] (v2.0) with the Yang–Nielsen (YN) method. Nucleotide diversity (π) was calculated with vcftools[83] (v0.1.16). Wilcoxon rank-sum tests were performed to assess statistically significant differences among the gene categories.

SV detection

To identify SVs across the 26 assembled cowpea genomes with high resolution, a whole-genome alignment-based approach was employed. Specifically, 25 genome assemblies were aligned to the FC6 reference genome using Minimap2[84] (v2.28) with default parameters. The resulting whole-genome alignments were analyzed using SyRI[85] (v1.7.0) to detect SVs larger than 49 bp. To merge SVs across all genomes, we applied SURVIVOR[86] (v1.0.7) with the following parameters: 1000 1 1 0 0 50, which allows a maximum distance of 1,000 bp between breakpoints for merging SVs across samples.

Graph-based pan-genome construction and population SV genotyping

To eliminate bias introduced by the reference genome, when performing SV genotyping on resequenced samples[6,10], we constructed a graph-based pan-genome using Minigraph-Cactus[87] (v2.9.1) with default parameters, incorporating all 26 assembled cowpea genomes. Briefly, minigraph was first used to generate a draft variation graph from all 26 assemblies, and the draft graph was subsequently refined and fully aligned using Cactus to produce a whole-genome multiple alignment graph. This graph represents all shared and accession-specific sequences, including structural variants, insertions, deletions, and presence–absence regions across the 26 genomes. Resequencing reads were mapped to the graph genome using vg giraffe[88] (v1.63.1). Mapping coverage was analyzed using vg pack[89] (v1.63.1), and SV genotyping was performed with vg call[90] (v1.63.1). Genotype data for each sample were subsequently indexed using tabix (v1.21, https://github.com/tabixio/tabix). Finally, all individual genotype VCF files were merged

using bcftools[91] (v1.21) for downstream analysis. ODGI[92] (v0.9.4) is used for graph-based pangenome visualization.

Selective sweep analysis

To minimize potential misclassification caused by subspecies assignment, we retained only individuals with an ancestral component greater than 70% under the ADMIXTURE population structure analysis[93,94] at $K = 2$. This commonly used threshold effectively excludes highly admixed individuals while preserving sufficient sample size for group comparision[5,95]. As a result, 72 accessions were retained for the Vu group and 283 accessions for the Vs group. To eliminate bias due to unequal sample sizes, we randomly selected 72 accessions from the Vs group to match the Vu group for downstream analyses.

Selective sweeps potentially associated with artificial selection were identified by integrating three approaches: XP-CLR (v1.1.2), nucleotide diversity ratio (π_{Vu}/π_{Vs}), and F_{ST}. Due to the limited number of SVs, XP-CLR analysis was conducted using only SNP data. XP-CLR[96] was run with a 50-kb window size, a 10-kb step size, and a maximum of 200 SNPs per window. Nucleotide diversity (π) and fixation index (F_{ST}) were calculated using vcftools[83] (v0.1.16) with a sliding window of 50 kb and a step size of 10 kb. For each method, the top 5% of scores was used as the threshold to define candidate selective sweep regions. Regions identified by at least two of the three methods were considered as putative selective sweeps. Overlapping regions were determined using bedtools[97] (v2.31.1).

Genome-wide association studies

For the SV-based genome-wide association study (SV-GWAS), only biallelic structural variants were retained for analysis. To account for population structure and relatedness among samples, a kinship matrix was included as a covariate. GWAS was performed using the R package rMVP[98]. Phenotypic data for pod length and grain number per pod were obtained from a previously

published study[6]. The best linear unbiased predictions (BLUPs) across multiple years were calculated using the lme4[99] package and used as phenotypic inputs for the SV-GWAS. Both general linear models (GLM), mixed linear models (MLM), and the Fixed and random model Circulating Probability Unification (FarmCPU) model were evaluated for association testing. The FarmCPU model was ultimately selected as the optimal model due to its better control of false positives and higher statistical power. The genome-wide significance threshold was determined using the Bonferroni correction based on the effective number of independent SVs , as calculated by the genetic type I error calculator. A uniform threshold of 0.05/N was applied, resulting in a final significance cutoff of 2.14×10^{-6}. It is worth noting that FarmCPU iteratively uses significant markers as covariates, which frequently results in fragmented or interrupted peaks in Manhattan plots around true causal loci. PopLDdecay[100] (v3.43) was used to generate linkage disequilibrium heatmaps.

To identify genomic regions associated with divergence between Vu and Vs , we performed a binary-trait GWAS. Accessions with group membership > 70% were selected as racial representatives (72 Vu and 283 Vs), and their racial identify was assigned as a numeric trait (Vu = 1, Vs = 0). Association analysis was conducted in rMVP using a general linear mixed model, with the centered kinship matrix and the first five principal components as covariates.

RNA-seq analysis

The RNA-seq raw data were retrieved from the study by Wu *et al.*[6]. The dataset includes transcriptomes of seeds, leaves, flower buds, and pods collected at 0, 4, 8, and 12 days after anthesis from three cowpea accessions representing two subspecies. To quantify the expression of genes in this study, RNA-seq reads from different tissues were trimmed using the fastp[101] (v0.24.0) program. The clean reads were then mapped against the reference gene models using HISAT2[75] (v.2.2.1). The featureCounts[102] (v2.0.6) package was used for estimating FPKM values.

Protein structure and binding site prediction

The three-dimensional structure of the protein was predicted using AlphaFold3[103]. To identify potential ligand-binding sites, DeepSite (available at https://playmolecule.com/deepsite/) was employed with default parameters.

Dual-luciferase assay

The synthetic 2000-bp upstream fragments as the promoter of the two *VuWAK1* alleles and three *VuWAK2* alleles were ligased into luciferase (LUC) expression vector pGreenII 0800-LUC, and the empty vector and CaMV 35S promoter were used as the negative and positive control, respectively. These constructed were introduced into *Agrobacterium tumefaciens* GV3101, which was then used to transiently transform *Nicotiana benthamiana* leaves via infiltration. After two days of dark incubation and one day of light exposure, total protein was extracted. Luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega), and signals from Firefly LUC and Renilla luciferase (REN) were detected with a Synergy H1 multimode reader (BioTek).

Vector construction, plant transformation and phenotypic characterization

The coding sequences of *GA2OX2-2*-REF and *GA2OX2-2*-ALT alleles were cloned into the binary vector pCAMBIA1301 under the control of the CaMV 35S promoter. The constructs were introduced into *Agrobacterium tumefaciens* strain GV3101 and used to transform wild-type *Arabidopsis thaliana* (Col-0) via the floral dip method[104]. Transgenic lines were selected on hygromycin-containing medium, and positive transformants were confirmed by PCR with gene-specific primers (Supplementary Table 12). To measure silique length, mature siliques from the middle portion of the main inflorescence stem were collected.

Data availability

The raw sequence and the pan-genome assembly and annotation data have been deposited in the

National Genomics Data Center (NGDC; https://ngdc.cncb.ac.cn/gsa/) under accession PRJCA051570 [https://ngdc.cncb.ac.cn/gsa/search?searchTerm=PRJCA051570] and PRJCA044967 [https://ngdc.cncb.ac.cn/gwh/search/advanced/result?search_category=&search_term=&source=0&query_box=PRJCA044967]. The genotype data can be accessed at Figshare [https://doi.org/10.6084/m9.figshare.30931271]. The previously released transcriptome data used in this study are available at NCBI under accession PRJNA970477 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA970477] and PRJNA954189 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA954189][6]. The previous released resequencing data used in this study are available at NCBI under accession PRJNA889224 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA889224][6] and PRJNA890023 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA890023][10]. Source data are provided with this paper.

References

1. FAO, AUC, ECA & WFP. Africa - Regional Overview of Food Security and Nutrition 2023: Statistics and Trends. FAO, Accra (2023).
2. 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).
3. Carvalho, M. et al. Cowpea: a legume crop for a challenging environment. *J. Sci. Food Agric.* **97**, 4273–4284 (2017).
4. Abebe, B. K. & Alemayehu, M. T. A review of the nutritional use of cowpea (*Vigna unguiculata* L. Walp) for human and animal diets. *J. Agric. Food Res.* **10**, 100383 (2022).
5. Fiscus, C. J. et al. The pattern of genetic variability in a core collection of 2,021 cowpea accessions. *G3-Genes Genom. Genet.* **14**, jkae071 (2024).
6. Wu, X. et al. Differential selection of yield and quality traits has shaped genomic signatures of

cowpea domestication and improvement. *Nat. Genet.* **56**, 992–1005 (2024).

7. Mahalakshmi, V. et al. Cowpea [*Vigna unguiculata* (L.) Walp.] core collection defined by geographical, agronomical and botanical descriptors. *Plant Genet. Resour.* **5**, 113–119 (2007).
8. Swarup, S. et al. Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* **61**, 839–852 (2021).
9. Muñoz-Amatriaín, M. et al. Genome resources for climate-resilient cowpea, an essential crop for food security. *Plant J.* **89**, 1042–1054 (2017).
10. 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).
11. Liu, Y. et al. Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162–176 (2020).
12. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
13. Tao, Y. et al. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat. Plants* **7**, 766–773 (2021).
14. Yuan, Y. et al. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **19**, 2153–2163 (2021).
15. Li, W. et al. Plant pan-genomics: recent advances, new challenges, and roads ahead. *J. Genet. Genomics* **49**, 833–846 (2022).
16. Liang, Q. et al. A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.] Walp.). *Plant Genome* **17**, e20319 (2024).
17. Lonardi, S. et al. The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant J.* **98**, 767–782 (2019).
18. Liang, L. et al. Genome and pan-genome assembly of asparagus bean (*Vigna unguiculata* ssp. *sesquipedialis*) reveal the genetic basis of cold adaptation. *Front. Plant Sci.* **13**, 1059804 (2022).
19. Yang, Y. et al. A near-complete assembly of asparagus bean provides insights into anthocyanin

accumulation in pods. *Plant Biotechnol. J.* **21**, 2473–2489 (2023).

20. Simão, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
21. Rhie, A. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
22. Li, K. et al. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat. Commun.* **14**, 6556 (2023).
23. Wei, C. et al. Telomere-to-telomere cowpea genomes reveal insights into centromere evolution in Phaseoleae. *Hotic. Res.* uhaf359 (2025).
24. Chen, S. et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nat. Plants* **9**, 1986–1999 (2023).
25. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
26. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
27. Cantalapiedra, C. P. et al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
28. Torkamaneh, D., Lemay, M. A. & Belzile, F. The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. *Plant Biotechnol. J.* **19**, 1852–1862 (2021).
29. Liu, C. et al. High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Commun.* **3**, 100352 (2022).
30. Tao, Y. et al. Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant* **12**, 156–169 (2019).
31. Barker, C. L. et al. Genetic and physical mapping of the grapevine powdery mildew resistance gene, *Run1*, using a bacterial artificial chromosome library. *Theor. Appl. Genet.* **111**, 370–377 (2005).

32. Barthole, G. et al. *MYB118* represses endosperm maturation in seeds of *Arabidopsis*. *Plant Cell* **26**, 3519–3537 (2014).

33. Xu, B. et al. Arabidopsis genes *AS1*, *AS2*, and *JAG* negatively regulate boundary-specifying genes to promote sepal and petal development. *Plant Physiol.* **146**, 566–575 (2008).

34. Porco, S. et al. Dioxygenase-encoding *AtDAO1* gene controls IAA oxidation and homeostasis in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11016–11021 (2016).

35. Wang, F. et al. BES1-regulated *BEE1* controls photoperiodic flowering downstream of blue light signaling pathway in *Arabidopsis*. *New Phytol.* **223**, 1407–1419 (2019).

36. Dardick, C. et al. *PpeTAC1* promotes the horizontal growth of branches in peach trees and is a member of a functionally conserved gene family found in diverse plants species. *Plant J.* **75**, 618–630 (2013).

37. Hamberger, B. & Hahlbrock, K. The *4-coumarate:CoA ligase* gene family in *Arabidopsis thaliana* comprises one rare, sinapate-activating and three commonly occurring isoenzymes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2209–2214 (2004).

38. Doyle, M. R. et al. The *ELF4* gene controls circadian rhythms and flowering time in *Arabidopsis thaliana*. *Nature* **419**, 74–77 (2002).

39. Liu, Z. et al. Dual roles of pear *EARLY FLOWERING 4 -like* genes in regulating flowering and leaf senescence. *BMC Plant Biol.* **24**, 1117 (2024).

40. Lally, D. et al. Antisense expression of a cell wall-associated protein kinase, WAK4, inhibits cell elongation and alters morphology. *Plant cell* **13**, 1317–1331 (2001).

41. Kubalová, M. et al. Gibberellin-deactivating GA2ox enzymes act as a hub for auxin-gibberellin cross talk in *Arabidopsis thalian*a root growth regulation. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2425574122 (2025).

42. Schomburg, F. M. et al. Overexpression of a novel class of gibberellin 2-oxidases decreases gibberellin levels and creates dwarf plants. *Plant Cell* **15**, 151–163 (2003).

43. Bianchi, J. S. et al. Changes in leaflet shape and seeds per pod modify crop growth parameters, canopy light environment, and yield components in soybean. *Crop J.* **8**, 351–364 (2020).

44. Meinke, D. W. Genome-wide identification of *EMBRYO-DEFECTIVE* (*EMB*) genes required for growth and development in Arabidopsis. *New Phytol.* **226**, 306–325 (2020).

45. Hufford, M. B. et al. *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).

46. Li, X. et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in *Brassica oleracea*. *Nat. Genet.* **56**, 517–529 (2024).

47. Flagel, L. E. et al. The standing pool of genomic structural variation in a natural population of *Mimulus guttatus. Genome Biol. Evol.* **6**, 53–64 (2014).

48. Munasinghe, M. et al. Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. *PLoS Genet.* **19**, e1011086 (2023).

49. Yuan, Y. et al., Current status of structural variation studies in plants. *Plant Biotechnol. J.* **19**, 2153–2163 (2021).

50. Akakpo, R. et al. The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.* **226**, 44–49 (2020).

51. Muller, H. J. The relation of recombination to mutational advance. *Mutat Res.* **106**, 2–9 (1964).

52. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).

53. Schwander, T. et al. Supergenes and complex phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).

54. Wu, X. et al. 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).

55. Feng, C. et al. Genomic and genetic insights into Mendel's pea genes. *Nature* **642**, 980–989 (2025).

56. Guo, L. et al. Super pangenome of *Vitis* empowers identification of downy mildew resistance genes for grapevine improvement. *Nat. Genet.* **57**, 741–753 (2025).

57. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression

differences in humans and often affect multiple nearby genes. *Genome Res.* **31**, 2249–2257 (2021).

58. Igolkina, A. A. et al. A comparison of 27 *Arabidopsis thaliana* genomes and the path toward an unbiased characterization of genetic polymorphism. *Nat. Genet.* **57**, 2289–2301 (2025).
59. Wagner, T. A. & Kohorn, B. D. Wall-associated kinases are expressed throughout plant development and are required for cell expansion. *Plant Cell* **13**, 303–318 (2001).
60. Han, L. et al. Identification of novel genomic regions associated with yield-related traits in cowpea (*Vigna unguiculata* [L.] Walp) landraces. *Mol Breed.* **45**, 65 (2025).
61. Pearson, S. M. et al. Scaling up orphan crop research: A global genetic perspective of cowpea (*Vigna unguiculata*) diversity from 10,617 accessions. *Plant J.* **125**, e70777 (2026).
62. Muñoz-Amatriaín, M. et al. The UCR Minicore: A resource for cowpea research and breeding. *Legume Sci*. **3**, e95 (2021).
63. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
64. Cheng, H. et al. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
65. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
66. Ou, S. et al. Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome Res.* **34**, 1140–1153 (2024).
67. Ou,S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
68. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* **9**, 18 (2008).
69. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
70. Lu, T. & Ou, S. TIR-Learner v3: New generation TE annotation program for identifying TIRs.

figshare Presentation (2024).

71. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
72. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Ch. 4**, 4.10.1–4.10.14 (2009).
73. Gabriel, L. et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
74. Stanke, M. et al. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644 (2008).
75. Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
76. Jones, P. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
77. Cantalapiedra, C. P. et al. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
78. Pootakham, W. et al. A chromosome-scale assembly of the black gram (*Vigna mungo*) genome. *Mol. Ecol. Resour.* **21**, 238–250 (2021).
79. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
80. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
81. Tang, H. et al. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
82. Wang, D. et al. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics Bioinf.* **8**, 77–80 (2010).
83. Petr, D. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
84. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–

4574 (2021).

85. Goel, M. et al. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
86. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
87. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* 42, 663–673 (2024).
88. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
89. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
90. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
91. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
92. Guarracino, A. et al. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
93. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
94. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **12**, 246 (2011).
95. Leamy, L. J. et al. Environmental versus geographical effects on genomic variation in wild soybean (*Glycine soja*) across its native range in northeast Asia. *Ecol. Evol.* **6**, 6332–6344 (2016).
96. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
97. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

98. Yin, L. et al. rMVP: A Memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics Bioinf.* **19**, 619–628 (2021).
99. Luke, S. G. Evaluating significance in linear mixed-effects models in R. *Behav. Res. Methods* **49**, 1494–1502 (2017).
100. Zhang, C. et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
101. Chen, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
102. Liao, Y. et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
103. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
104. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J* **16**, 735–743 (1998).

Acknowledgements

The conclusions and opinions expressed in this work are those of the authors alone and shall not be attributed to the Gates Foundation. Under the grant conditions of the foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. We thank Anna Koltunow for helpful comments on the manuscript.

Funding statement

This work was supported, in part, by the Gates Foundation [Hy-Gain INV-002955] and the National Natural Science Foundation of China (Overseas) awarded to Y.T.

Author contributions

Y.T., E.M. and D.J. designed this project and coordinated the research activities. A.C., S.P., E.M., D.J. and Y.Y. collected and provided plant materials. S.P. prepared the sequenced samples. X.Z. and H.W. provided constructive feedback throughout the research process. S.S., C.W., Yanbo Wang, Yinzi Wang, T.D., Y.Z. and S.P. performed data analyses. S.S. wrote the manuscript. Y.T., E.M., H.W., S.P. and D.J. revised the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Table 1. Statistics of assemblies and annotation of cowpea accession genomes.

Sample ID	Assembly size (Mb)	Anchored rate (%)	Contig N50 (Mb)	GC content (%)	TE content (%)	Gene number	Annotation completeness (BUSCO, %)	QV
Vung01	580.03	95.73	29.26	39.33	40.82	29,683	98.7	63.63
Vung02	577.16	96.08	26.39	36.70	44.72	29,521	98.8	63.79
Vung03	582.98	95.81	43.20	40.90	43.39	28,839	98.8	63.85
Vung04	577.57	95.68	34.14	42.13	40.82	28,726	98.7	64.36
Vung05	574.07	97.40	35.00	38.55	42.51	28,736	98.7	64.72
Vung06	606.59	93.84	25.92	37.37	44.27	30,459	98.8	61.19
Vung07	569.46	96.02	40.46	36.00	41.60	29,244	98.8	64.15
Vung08	586.79	92.71	30.27	34.38	43.01	33,180	98.6	59.65
Vung09	580.20	95.34	24.80	38.18	41.16	30,779	98.8	63.71
Vung10	565.12	95.50	24.24	40.64	42.89	28,328	98.9	63.74
Vung11	574.56	96.63	32.79	35.98	41.92	28,826	98.8	63.79
Vung12	569.10	95.27	26.88	37.15	40.46	30,605	98.6	63.36
Vung13	558.21	96.98	30.05	38.30	43.28	28,603	98.8	63.99
Vung14	569.13	96.04	33.02	41.22	43.05	29,115	98.9	63.96
Vung15	558.53	96.47	40.31	36.28	39.96	30,008	98.6	63.55
Vung16	561.33	96.17	26.85	38.26	43.89	28,911	98.6	63.17
Vung17	569.87	94.87	31.84	35.83	39.99	31,599	98.8	62.82
Vung18	557.03	96.60	25.60	37.27	39.61	29,792	98.7	63.71
Vung19	572.19	95.54	27.50	39.78	43.39	29,386	98.9	63.43
Vung20	557.49	96.27	32.89	36.25	42.58	29,485	98.9	63.81
FC6*	567.48	95.60	46.12	33.61	43.14	29,940	98.8	63.42
IT97K**	519.44	91.15	-	32.82	39.87	27,781	97.5	-
G323**	592.27	93.31	-	33.81	43.08	27,679	98.7	-
G98**	632.31	89.87	-	34.28	42.62	28,372	98.8	-
NJ**	550.31	99.37	-	33.41	42.78	27,610	98.7	-
A147**	593.98	90.80	-	33.83	43.93	31,708	98.6	-

*The full PacBio and ONT raw reads of FC6 were downloaded from the National Genomics Data Center under accession CRA007957.

**The genome assemblies of these five cultivars were accessed from previous publications[6,10,17-19] and re-analyzed in this study.

Figure legends

Fig. 1. Geographic distribution and genomic feature analysis of cowpea accessions. a, Principal component analysis (PCA) based on single nucleotide polymorphisms (SNPs) of cowpea accessions. Each dot represents an accession, with colored dots indicating cultivars selected for pan-genome analysis. **b,** Geographic distribution of cultivated cowpea accessions, with dot size proportional to sample size. **c,** Phylogenetic tree of 26 representative cowpea accessions, using *Vigna mungo* as the outgroup. Red solid dots represent *Vigna unguiculata* ssp. *unguiculata*, and blue solid dots represent *Vigna unguiculata* ssp. *sesquipedalis*. **d,** Genomic collinearity among pan-genome accessions. **e,** Genome size (Mb) and proportions of non-repetitive (light grey) and repetitive (dark grey) sequences for each cowpea accession.

Fig. 2. Pan-genome analysis of 26 cowpea accessions. a, Counts of pan and core gene families, based on 1,000 random samplings per count. **b,** Composition of the pangenome. The histogram shows the number of orthogroups, and the pie chart illustrates the proportions of core, dispensable, and private orthogroups. **c,** Composition of individual genomes, with each row representing an accession. **d–f,** Comparison of nucleotide diversity (π), gene length, and K_a/K_s ratio among core, dispensable, and private genes. **g,** Expression levels of core, dispensable, and private genes, measured in fragments per kilobase of transcript per million mapped reads (FPKM). **h,** Gene Ontology term enrichment for core genes. **i,** Gene Ontology term enrichment for variable (dispensable and private) genes. The dot size indicating gene count and color representing the *p*-value. In box plots, boxes span the 25^{th} to 75^{th} percentiles, with central lines at the median and whiskers extending to 1.5× the interquartile range. Statistical significance was determined using two-sided Wilcoxon tests, with sample sizes (n) indicated for each group. Source data are provided as a Source Data file.

Fig. 3. Structural variant landscape and pan-SV analysis in cowpea. a, Distribution of SVs in the FC6-T2T reference genome. (A) Chromosomes, (B) Gene density, (C) Deletions, (D)

Insertions, (E) Inversions, (F) Duplications, (G) Translocations. **b,** Composition of SVs across 25 cowpea accessions. **c,** Numbers of pan and core SVs identified in different number of genomes, based on 1,000 random sampling per genome count. **d,** Length distribution of the five types of SVs. **e,** Example of a complex inversion event on chromosome 10. Left: Dot plot showing genome alignment to FC6-T2T. Right: Genomic collinearity of the inverted region. **f,** Illustration of inversion types across 26 accessions. **g–h,** SVs distribution relative to genes (**g**) and transposable elements (TEs) (**h**). **i,** Violin plots showing numbers of TE-derived SVs and non-TE-derived SVs within 1-Mb windows (n = 549 windows). **j,** Comparison of SVs length between TE-derived (n =38,345) and non-TE-derived (n =23,794) SVs. In box plots, boxes span the 25th to 75th percentiles, with central lines at the median and whiskers extending to 1.5× the interquartile range. Statistical significance was determined using a two-sided Wilcoxon test. Source data are provided as a Source Data file.

Fig. 4. Identification of structural variants underlying divergence between *V. unguiculata* subsp. a, Manhattan plot showing SV-based GWAS of subspecies identity in cowpea. Colored dots represent significant loci, and the horizontal line indicates the genome-wide significance threshold. Statistical significance was assessed using 0.05/SV number. **b,** Manhattan plot displaying XP-CLR scores between Vs and Vu, with horizontal dotted lines indicating the top 5% threshold. Regions identified by at least two of XP-CLR (top 5%), Pi-ratio (bottom 5%), and F_{ST} (top 5%) are highlighted in orange. **c-d,** Example of an identified candidate gene underlying subspecies divergence: gene structure of *VuELF4* and its synteny with the homologous *AtELF4* in *Arabidopsis thaliana*. **e,** Expression levels of *VuELF4* in different tissues and their comparison between REF and ALT genotype accession. **f,** Frequency of the 78-bp deletion in *VuELF4* across Vu and Vs populations.

Fig. 5. Structural variants associated with pod length in cowpea. a, Manhattan plot displaying SV-based GWAS results for pod length, with linkage disequilibrium heatmaps of the top two

significant loci. The horizontal line denotes the genome-wide significance threshold. **b,** Box plots validating effects of identified significant loci on pod length. **c,** Box plots illustrating the effects of different combinations of favorable alleles on pod length. **d,** Gene structure of *VuWAK* genes and locations of four SVs. *VuWAKs* include *VuWAK-1* (*Vu09G1544*), *VuWAK-2* (*Vu09G1545*), *VuWAK-3* (*Vu09G1546*), *VuWAK-4* (*Vu09G1547*) and *VuWAK-5* (*Vu09G1548*). **e,** Effects of the four SVs in *VuWAK-1* and *VuWAK-2* on pod length. **f,** Box plots showing effects of combined haplotypes of the four SVs in *VuWAK-1* and *VuWAK-2* on pod length. Each dot represents an accession. **g,** Pie charts depicting the subspecies composition of the combined haplotypes of the four SVs in *VuWAK-1* and *VuWAK-2*. **h,** Gene structure of *VuGA2ox2s* and locations of two SVs. *VuGA2ox2s* include *VuGA2ox2-1* (*Vu05G1623*) and *VuGA2ox2-2* (*Vu05G1624*). **i,** Effects of the two SVs in *VuGA2ox2* genes on pod length. **j,** Changes of protein structure led by the 51-bp insertion in *VuGA2ox2-2*. Three-dimensional protein structure was modeled by AlphaFold3. Regions colored in red correspond to the 51-bp insertion and 5 amino acid flanking regions. **k,** Representative siliques from wild type (WT), VuGA2ox2-2-REF, and VuGA2ox2-2-ALT overexpression lines in *Arabidopsis thaliana*. **l,** Pod length measurements of WT (control) and transgenic lines. Each dot represents an individual transgenic line. Pod length was calculated as the mean of five siliques sampled from the middle portion of each plant. Each line represents a biological replicate, with at least 10 independent transgenic lines included. **m,** Box plots showing the effects of combined haplotypes of the two SVs in *VuGA2ox2* genes. Each dot represents an accession. **n,** The subspecies composition of combined haplotypes of the two SVs in *VuGA2ox2* genes. In box plots, boxes represent the 25th to 75th percentiles, with central lines at the median and whiskers extending to 1.5× the interquartile range. Significance was determined by two-sided Wilcoxon tests, with sample sizes (n) indicated for each group. Source data are provided as a Source Data file.

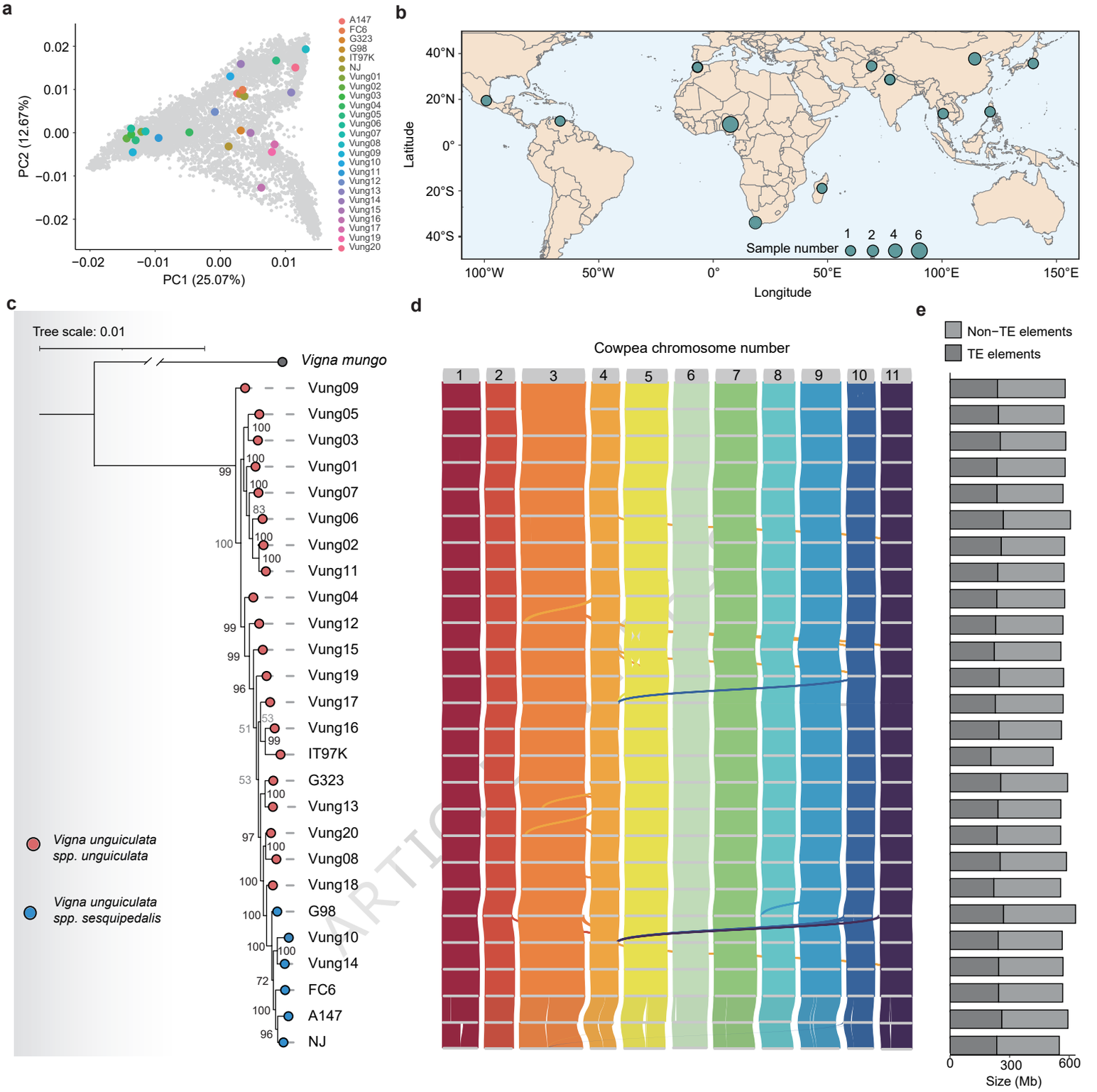
Fig. 6. GWAS of structural variants linked to seed number per pod in cowpea. a, Manhattan plot displaying SV-based GWAS results for the seed number per pod. The horizontal line denotes the genome-wide significance threshold. **b,** Linkage disequilibrium heatmap surrounding the

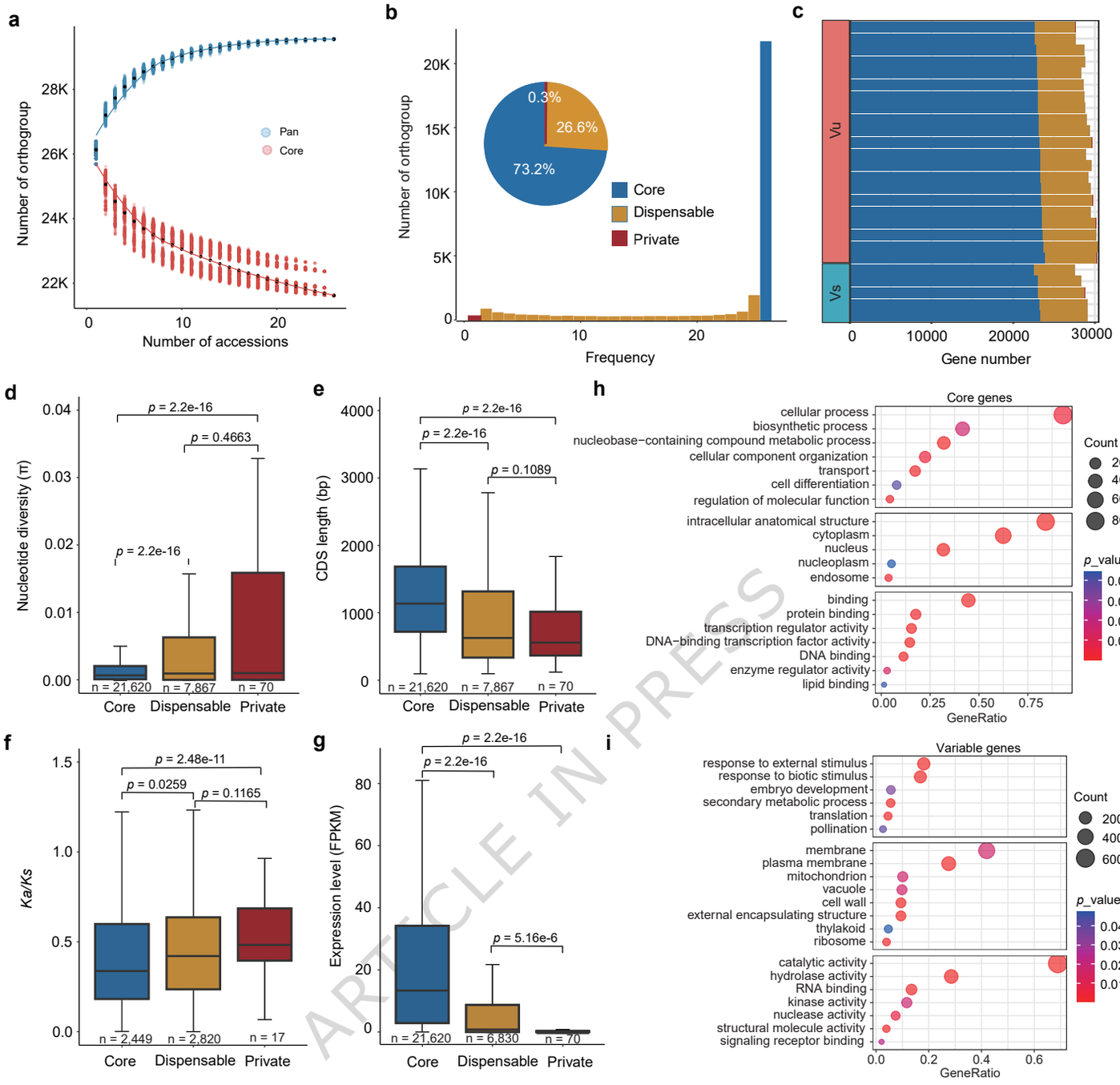
candidate gene *VuPPR*. **c,** Gene structure of *VuPPR* and the position of the 83-bp deletion. **d,** Effects of the 83-bp deletion in *VuPPR* on seed number per pod. **e,** Comparison of seed number per pod between Vu and Vs. **f,** Allele frequency differences for the 83-bp deletion between Vu and Vs. In boxplots, boxes represent the 25th to 75th percentiles, with central lines at the median and whiskers extending to 1.5× the interquartile range. Significance was assessed using two-sided Wilcoxon tests, with sample sizes (n) indicated for each group. Source data are provided as a Source Data file.

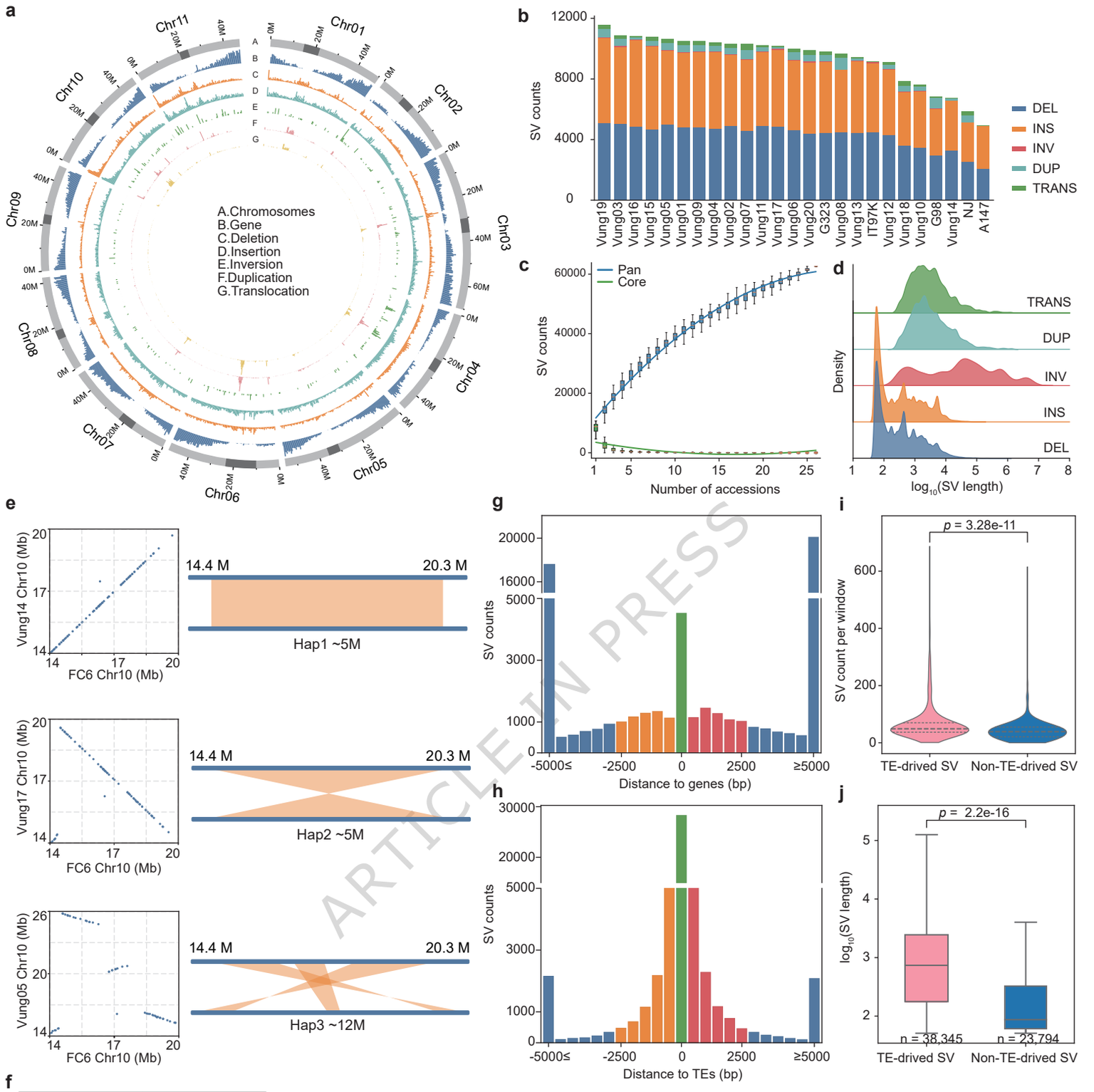
Editor's Summary

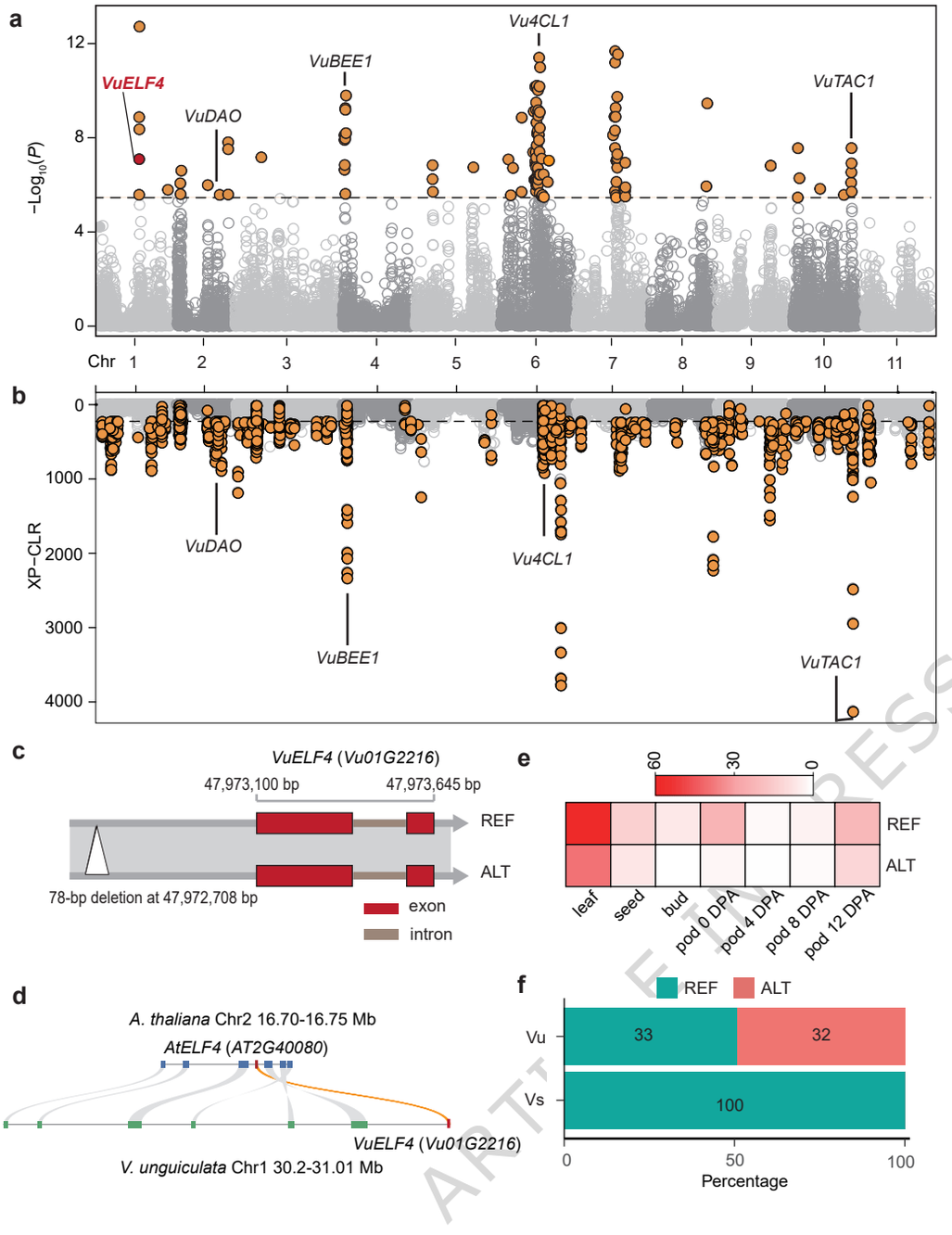
Cowpea is a nutritious legume that provides an important source of plant-based protein worldwide. Here, the authors assemble 20 genomes representing global diversity and construct a pan-genome to characterize genetic variations and their associations with pod length and seed number per pod.

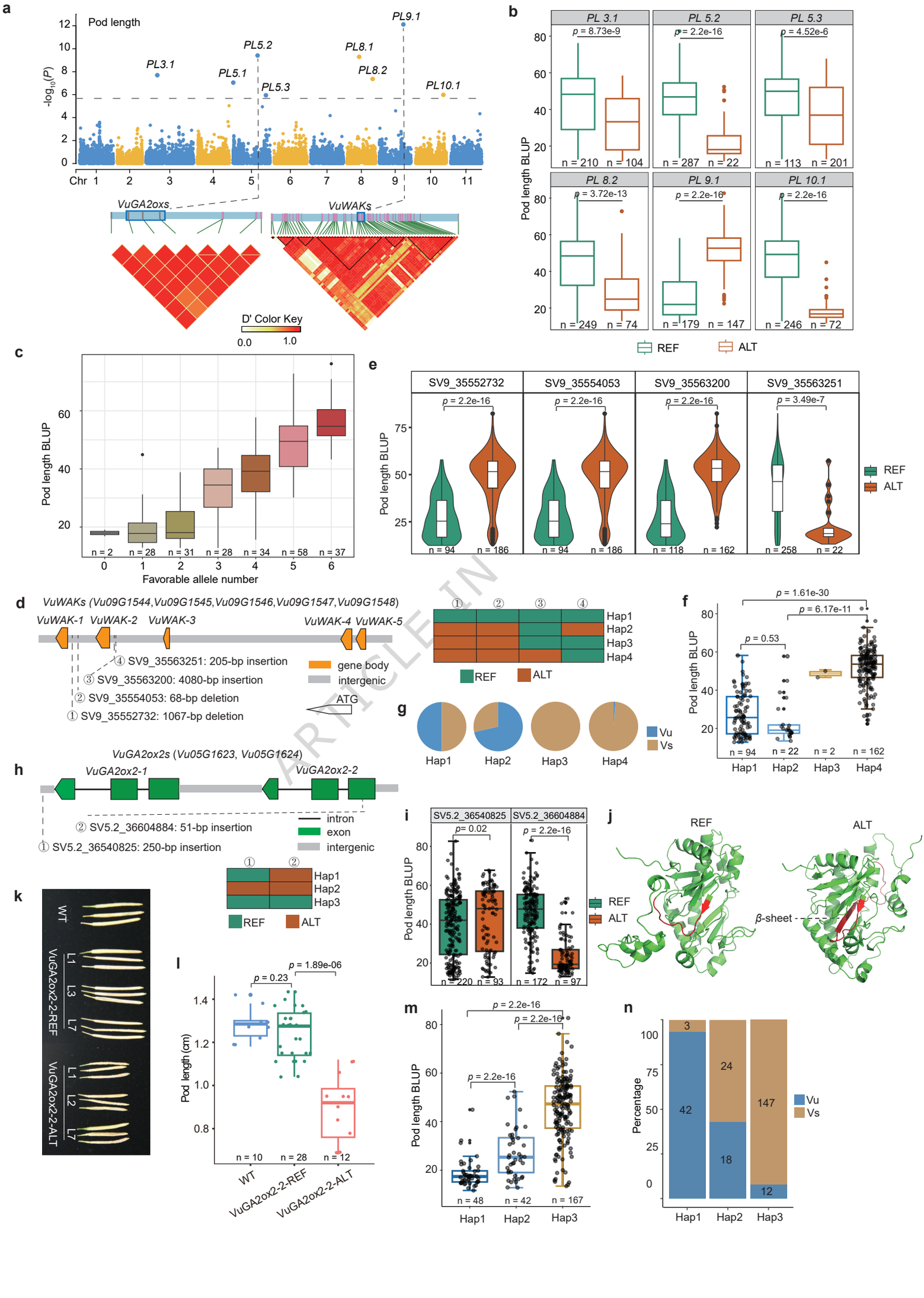
Peer review information: *Nature Communications* thanks Li Guo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.


a
PC2 (12.67%)
PC1 (25.07%)
A147
FC6
G323
G98
IT97K
NJ
Vung01
Vung02
Vung03
Vung04
Vung05
Vung06
Vung07
Vung08
Vung09
Vung10
Vung11
Vung12
Vung13
Vung14
Vung15
Vung16
Vung17
Vung19
Vung20
b
Latitude
Longitude
Sample number
1
2
4
6
c
Tree scale: 0.01
Vigna mungo
Vung09
Vung05
Vung03
Vung01
Vung07
Vung06
Vung02
Vung11
Vung04
Vung12
Vung15
Vung19
Vung17
Vung16
IT97K
G323
Vung13
Vung20
Vung08
Vung18
G98
Vung10
Vung14
FC6
A147
NJ
Vigna unguiculata spp. unguiculata
Vigna unguiculata spp. sesquipedalis
d
Cowpea chromosome number
e
Non−TE elements
TE elements
Size (Mb)








a
Pod length
PL3.1
PL5.1
PL5.2
PL5.3
PL8.1
PL8.2
PL9.1
PL10.1
-log10(P)
Chr 1 2 3 4 5 6 7 8 9 10 11
VuGA2oxs
VuWAKs
D' Color Key
0.0 1.0
b
PL 3.1
PL 5.2
PL 5.3
PL 8.2
PL 9.1
PL 10.1
p = 8.73e-9
p = 2.2e-16
p = 4.52e-6
p = 3.72e-13
p = 2.2e-16
p = 2.2e-16
Pod length BLUP
n = 210 n = 104
n = 287 n = 22
n = 113 n = 201
n = 249 n = 74
n = 179 n = 147
n = 246 n = 72
REF
ALT
c
Pod length BLUP
n = 2 n = 28 n = 31 n = 28 n = 34 n = 58 n = 37
Favorable allele number
e
SV9_35552732
SV9_35554053
SV9_35563200
SV9_35563251
p = 2.2e-16
p = 2.2e-16
p = 2.2e-16
p = 3.49e-7
Pod length BLUP
n = 94 n = 186
n = 94 n = 186
n = 118 n = 162
n = 258 n = 22
REF
ALT
d
VuWAKs (Vu09G1544,Vu09G1545,Vu09G1546,Vu09G1547,Vu09G1548)
VuWAK-1 VuWAK-2 VuWAK-3 VuWAK-4 VuWAK-5
④ SV9_35563251: 205-bp insertion
③ SV9_35563200: 4080-bp insertion
② SV9_35554053: 68-bp deletion
① SV9_35552732: 1067-bp deletion
gene body
intergenic
ATG
① ② ③ ④
Hap1
Hap2
Hap3
Hap4
REF
ALT
f
p = 1.61e-30
p = 6.17e-11
p = 0.53
Pod length BLUP
n = 94 n = 22 n = 2 n = 162
Hap1 Hap2 Hap3 Hap4
g
Hap1 Hap2 Hap3 Hap4
Vu
Vs
h
VuGA2ox2s (Vu05G1623, Vu05G1624)
VuGA2ox2-1
VuGA2ox2-2
② SV5.2_36604884: 51-bp insertion
① SV5.2_36540825: 250-bp insertion
intron
exon
intergenic
① ②
Hap1
Hap2
Hap3
REF
ALT
i
SV5.2_36540825
SV5.2_36604884
p= 0.02
p = 2.2e-16
Pod length BLUP
n = 220 n = 93
n = 172 n = 97
REF
ALT
j
REF
ALT
β-sheet
k
WT
VuGA2ox2-2-REF
VuGA2ox2-2-ALT
L1
L2
L3
L7
l
p = 1.89e-06
p = 0.23
Pod length (cm)
n = 10 n = 28 n = 12
WT
VuGA2ox2-2-REF
VuGA2ox2-2-ALT
m
p = 2.2e-16
p = 2.2e-16
p = 2.2e-16
Pod length BLUP
n = 48 n = 42 n = 167
Hap1 Hap2 Hap3
n
Percentage
3
42
24
18
147
12
Vu
Vs
Hap1 Hap2 Hap3