

# UC Riverside

## UC Riverside Previously Published Works

### Title

A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.] Walp.)

### Permalink

<https://escholarship.org/uc/item/7wn084zt>

### Journal

The Plant Genome, 17(1)

### ISSN

1940-3372

### Authors

Liang, Qihua

Munoz-Amatriaín, María

Shu, Shengqiang

et al.

### Publication Date

2024-03-01

### DOI

10.1002/tpg2.20319

### Copyright Information


This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

SPECIAL SECTION: GENOMICS OF ABIOTIC STRESS TOLERANCE  
AND CROP RESILIENCE TO CLIMATE CHANGE

# A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.] Walp.)

Qihua Liang<sup>1</sup> | María Muñoz-Amatriáin<sup>2,3</sup> | Shengqiang Shu<sup>4</sup> | Sassoum Lo<sup>2,5</sup> |  
 Xinyi Wu<sup>6</sup> | Joseph W. Carlson<sup>4</sup> | Patrick Davidson<sup>4</sup> | David M. Goodstein<sup>4</sup> |  
 Jeremy Phillips<sup>4</sup> | Nadia M. Janis<sup>7</sup> | Elaine J. Lee<sup>7</sup> | Chenxi Liang<sup>7</sup> |  
 Peter L. Morrell<sup>7</sup> | Andrew D. Farmer<sup>8</sup> | Pei Xu<sup>9</sup> | Timothy J. Close<sup>2</sup>  |  
 Stefano Lonardi<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, University of California Riverside, Riverside, CA, USA<sup>2</sup>Department of Botany and Plant Sciences, University of California Riverside, Riverside, CA, USA<sup>3</sup>Departamento de Biología Molecular, Universidad de León, León, Spain<sup>4</sup>US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA<sup>5</sup>Department of Plant Sciences, University of California Davis, Davis, CA, USA<sup>6</sup>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, China<sup>7</sup>Department of Agronomy and Plant Genetics, University of Minnesota Twin Cities, Saint Paul, MN, USA<sup>8</sup>National Center for Genome Resources, Santa Fe, NM, USA<sup>9</sup>Key Lab of Specialty Agri-Product Quality and Hazard Controlling Technology of Zhejiang Province, China Jiliang University, Hangzhou, China**Correspondence**

Timothy J. Close, Department of Botany and Plant Sciences, University of California Riverside, Riverside, CA, USA.

Email: [timothy.close@ucr.edu](mailto:timothy.close@ucr.edu)

Stefano Lonardi, Department of Computer Science and Engineering, University of California Riverside, Riverside, CA, USA.

Email: [stelo@cs.ucr.edu](mailto:stelo@cs.ucr.edu)

Assigned to Associate Editor Rajeev Varshney.

**Abstract**

Cowpea, *Vigna unguiculata* L. Walp., is a diploid warm-season legume of critical importance as both food and fodder in sub-Saharan Africa. This species is also grown in Northern Africa, Europe, Latin America, North America, and East to Southeast Asia. To capture the genomic diversity of domesticates of this important legume, de novo genome assemblies were produced for representatives of six subpopulations of cultivated cowpea identified previously from genotyping of several hundred diverse accessions. In the most complete assembly (IT97K-499-35), 26,026 core and 4963 noncore genes were identified, with 35,436 pan genes when considering all seven accessions. GO terms associated with response to stress and defense response were

**Abbreviations:** BED, browser extensible data; BUSCO, benchmarking universal single-copy orthologue; BWA, Burrows-Wheeler Aligner; CDS, coding sequence; CNV, copy number variation; DNA, deoxyribonucleic acid; GCV, Genome Context Viewer; GFF, general feature format; GMI, gene model improvement; GO, gene ontology; GVCF, genomic variant call format; IGC, integrated gene call; indel, insertion–deletion; JGI, Joint Genome Institute; LIS, legume information system; PacBio, Pacific Biosciences; PAV, presence–absence variation; RNA, ribonucleic acid; SNP, single nucleotide polymorphism; UCR, University of California Riverside; UTR, untranslated region; VCF, variant call format; VeP, variant effect predictor; WGS, whole genome shotgun.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *The Plant Genome* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

**Funding information**

United States Agency for International Development, Grant/Award Number: Cooperative Agreement AID-OAA-A-13-00070; National Natural Science Foundation of China, Grant/Award Number: 32172568; U.S. Department of Agriculture, Grant/Award Number: Hatch Project CA-R-BPS-5306-H; National Science Foundation, Grant/Award Numbers: IIS 1814359, IOS 1543963; Agricultural Research Service, Grant/Award Number: Non-Assistance Cooperative Agreement 58-5030-7-069; National Ten-Thousand Talents Program of China; U.S. Department of Energy, Grant/Award Number: Contract No. DE-AC02-05CH11231; Major Science and Technology Project of Plant Breeding in Zhejiang Province, Grant/Award Number: 2021C02065-6-3

highly enriched among the noncore genes, while core genes were enriched in terms related to transcription factor activity, and transport and metabolic processes. Over 5 million single nucleotide polymorphisms (SNPs) relative to each assembly and over 40 structural variants >1 Mb in size were identified by comparing genomes. Vu10 was the chromosome with the highest frequency of SNPs, and Vu04 had the most structural variants. Noncore genes harbor a larger proportion of potentially disruptive variants than core genes, including missense, stop gain, and frameshift mutations; this suggests that noncore genes substantially contribute to diversity within domesticated cowpea.

**1 | INTRODUCTION**

Individuals within a species vary in their genomic composition. The genome of any individual does not include the full complement of genes contained within the species. A pan-genome includes genes core to the species (shared among all individuals) and those absent from one or more individuals (noncore, dispensable, or variable genes). This pan-genome concept started to be applied to plants by Morgante et al. (2007) but began in bacterial species (reviewed by Golicz et al., 2020). Due to the complexity of plant genomes, the first studies exploring gene presence–absence variation (PAV) in plants used reduced-representation approaches, including array comparative genomic hybridization and sequencing of transcriptomes (Hirsch et al., 2014; Muñoz-Amatriaín et al., 2013; Springer et al., 2009). Once sequencing of multiple plant genomes became feasible, several pan-genomes of variable degrees of completeness were generated, and it was soon understood that PAV is prevalent in plants and that the pan-genome of any plant species is larger than the genome of any individual accession (reviewed by Lei et al., 2021). Moreover, many of the genes absent in reference accessions have functions of potential adaptive or agronomic importance, such as time to flowering, and response to abiotic and biotic stresses (Bayer et al., 2020; Gordon et al., 2017; Montenegro et al., 2017), making the construction of a pan-genome a crucial task for crops of global importance.

Cowpea is a diploid ( $2n = 22$ ) member of the family Fabaceae tribe Phaseoleae, closely related to mung bean, common bean, soybean, and several other warm-season legumes. Cowpea was domesticated in Africa, but its cultivation has spread throughout most of the globe (Herniter et al., 2020). The inherent resilience of the species to drought and high temperatures (Hall, 2004), together with its nutritional value as a

reliable source of plant-based protein and folic acid, position cowpea favorably as a component of sustainable agriculture in the context of global climate change. Most cowpea production and consumption presently occur in sub-Saharan Africa, especially in the Sudano-Sahelian Zone, with production mainly by smallholder farmers, often as an intercrop with maize, sorghum, or millet (Boukar et al., 2019). Tender green seeds are often consumed during the growing season, and immature pods are eaten as a vegetable, especially in East and Southeast Asia. In addition, fresh leaves are sometimes consumed, and dry haulms are harvested and sold as fodder for livestock. Spreading varieties are also utilized as cover crops to prevent soil erosion and weed control.

A reference genome sequence of cowpea cv. IT97K-499-35 was previously generated (Lonardi et al., 2019). Preliminary sequence comparisons using whole genome shotgun (WGS) data of 36 accessions suggested that extensive single nucleotide polymorphism (SNP) and structural variation exists within domesticated cowpea (Lonardi et al., 2019). Cowpea also displays a wide range of phenotypic variation, and genetic assignment approaches have identified six subpopulations within cultivated cowpea germplasm (Muñoz-Amatriaín et al., 2021). These observations support the need to develop cowpea pan-genome resources based on diverse cowpea accessions.

This study reports *de novo* assemblies of six cultivated cowpea accessions. Each accession was annotated using transcriptome sequences from the accession along with *ab initio* methods. These genome sequences, together with the previously reported sequence of IT97K-499-45 (Lonardi et al., 2019), constitute a pan-genome resource for domesticated cowpea. Using annotations for the seven genomes, including genes, along with variant calls for SNPs and short insertion–deletion (indels), and larger structural variants, the following

questions were addressed: (i) What proportion of genes are core and noncore, and do core and noncore genes differ in size or functional class? (ii) What proportion of large-effect variants are created by single nucleotide variants versus structural variants (including indels), and do the proportions of large-effect variants differ among core and noncore genes? (iii) To what extent are gene content and gene order consistent across accessions within the species *Vigna unguiculata* and across species within the genus *Vigna* and the tribe Phaseoleae? The results suggest that both extensive structure differences among individual accessions and the nature of variation in noncore genes are important considerations in efforts to identify genetic variation with adaptive potential.

## 2 | MATERIALS AND METHODS

Cowpea accessions were used in this work (Table S01).

Accessions chosen for sequencing and de novo assembly represented the six subpopulations of domesticated cowpea described in Muñoz-Amatriáin et al. (2021), as indicated in Figure 1. The intention of choosing accessions that cover each subpopulation was to maximize the discovery of genetic variations relevant to cultivated cowpea using a small number of samples. As shown by Gordon et al. (2017) in *Brachypodium distachyon*, the addition of individuals from subpopulations not previously sampled contributes much more to increasing the pan-genome size than adding closely related individuals.

IT97K-499-35 is a blackeye variety with resistance to the parasitic plants *Striga* and *Alectra*, developed at the International Institute of Tropical Agriculture in Ibadan, Nigeria (Singh et al., 2006) and provided by Michael Timko (University of Virginia, Charlottesville, Virginia, USA) to the University of California Riverside (UCR) in 2006. The sequence assembly and annotation of IT97K-499-35 were described in Lonardi et al. (2019). CB5-2 is a fully inbred isolate closely related to CB5, the predominant Blackeye of the US Southwest for several decades. CB5 (Blackeye 8415) was bred by W. W. Mackie at the University of California (Mackie, 1946) to add resistances to *Fusarium* wilt and nematodes to a California Blackeye landrace, and provided to UCR by K. Foster, University of California, Davis, in 1981. Suvita-2, also known as Gorom Local (IITA accession TVu-15553, US NPGR PI 583259), is somewhat resistant to bruchids and certain races of *Striga* and is relatively drought tolerant. This landrace was collected from a local market by V. D. Aggarwal at the Institut de l'Environnement et de Recherches Agricoles (INERA) in Burkina Faso (Aggarwal et al., 1984) and provided to UCR by V. D. Aggarwal in 1983. Sanzi is an early flowering, small-seeded landrace from Ghana with resistance to flower bud thrips (Boukar et al., 2013), provided by K. O. Marfo, Nyankpala Agricultural Experiment Station, Tamale, Ghana to UCR in 1988. UCR779 (PI 583014) is a landrace from Botswana (de Mooy, 1985; Ehlers et al., 2002) that

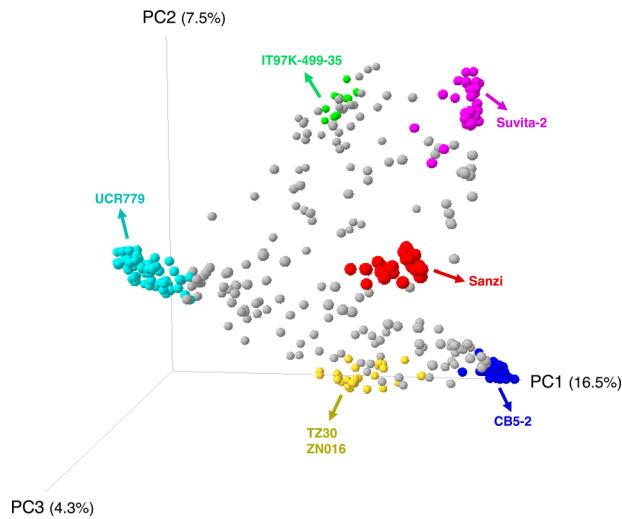
### Core Ideas

- The genetic attributes of cultivated cowpea germplasm are best represented by a pan-genome.
- Seven diverse accessions of cultivated cowpea converge on a core gene set, while not exhausting noncore content.
- Noncore genes are enriched for stress response and contribute a relatively large portion of variants.
- Several large, high-frequency inversions within chromosomes exist in cultivated cowpea germplasm.
- Breeding for improved climate adaptation must consider variants related to inversions and noncore genes.

was provided to UCR as B019-A in 1987 by C. J. de Mooy of Colorado State University. Yardlong bean or asparagus bean (cv.-gr. *Sesquipedalis*), the vegetable type of cowpea, is widely grown in Asian countries for the consumption of tender long pods. TZ30 is an elite Chinese variety with a pod length of around 60 cm. ZN016 is a landrace originating from Southeastern China with a pod length of about 35 cm and showing resistance to multiple major diseases of cowpea. TZ30 and ZN016 were used previously as parents of a mapping population to study the inheritance of pod length (Xu et al., 2017).

### 2.1 | DNA sequencing and de novo assembly of seven cowpea accessions

The annotated genome (v1.0) of African variety IT97K-499-35 was assembled from Pacific Biosciences (Menlo Park, California, USA) long reads, two Bionano Genomics (San Diego, California, USA) optical maps and 10 genetic linkage maps as described previously (Lonardi et al., 2019). The six additional de novo assemblies were produced by Dovetail Genomics (Scotts Valley, California, USA) using Illumina (San Diego, California, USA) short reads (150 × 2). DNA was extracted by Dovetail Genomics from seedling tissue of CB5-2, TZ30, and ZN016, and seeds of CB5-2, Suvita-2, Sanzi, and UCR779. Meraculous (Chapman et al., 2011) was used to assemble the reads, then sequences from Dovetail Chicago® and Dovetail Hi-C® libraries were added (using their proprietary pipeline) to resolve misassemblies and increase contiguity. These assemblies were further refined using ALLMAPS (Tang et al., 2015). This analysis used 10 previously reported genetic linkage maps to relate assemblies to the standard orientations and numbering of the 11 cowpea chromosomes, as described in Lonardi et al. (2019) for



**FIGURE 1** Principal component analysis of the UCR Minicore, indicating the accessions selected for sequencing and the subpopulation they belong to. Accessions in the plot are colored by the result of STRUCTURE for  $K = 6$ , as shown in Muñoz-Amatriaín et al. (2021).

IT97K-499-35. See “Data Availability Statement” for access to raw data and assemblies.

## 2.2 | Calling of SNPs, indels, and structural variants

SNPs and indels were called using each reference genome versus the reads from the six other accessions. Reads of each accession described above for genome assemblies, plus short-read sequences produced by 10× Genomics from IT97K-49-35, were mapped to all assemblies using Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009). SNPs and indels were called using the GATK 4.2.0 pipeline in GVCF mode for each accession. All the per-sample GVCFs were gathered in joint genotyping to produce a set of joint-called SNPs and indels. Both per-sample SNPs and joint-called SNPs were filtered with the same parameters of “ $QD < 2.0 \parallel FS > 60.0 \parallel MQ < 40.0 \parallel MQRankSum < -12.5 \parallel ReadPosRankSum < -8.0 \parallel SOR > 4.0$ ”. Indels were filtered with “ $QD < 2.0 \parallel FS > 200.0 \parallel ReadPosRankSum < -20.0 \parallel SOR > 10.0$ ”.

Each pair of individual genomes was aligned using minimap2 (Li, 2018), producing  $\binom{7}{2} = 21$  alignment files. Structural variants, including inversions and translocations, were identified from the alignment files using SyRI (Goel et al., 2019). Figures were produced using PlotSR (Goel & Schneeberger, 2022). Depth analyses were carried out using Mosdepth (Pedersen & Quinlan, 2018). The average nucleotide diversity within and between populations was calculated from a VCF file using Pixy (Korunes & Samuk, 2021).

## 2.3 | Annotation of genes and repeats

All genomes were annotated using the JGI plant genome annotation pipelines (Shu et al., 2014), integrated gene call (IGC), and gene model improvement (GMI). Both IGC and GMI are evidence-based gene call pipelines. In IGC, a gene locus was defined by peptide alignments of related organism homologous peptides and with alignments of within-organism transcriptome assemblies. Genes were predicted by homology-based gene prediction programs FGENESH+ (Salamov & Solovyev, 2000), FGENESH\_EST, and GenomeScan (Yeh et al., 2001), and a JGI in-house homology-constrained transcriptome assembly ORF finder. Homologous proteomes included *Arabidopsis thaliana* and those from common bean (*Phaseolus vulgaris*), soybean (*Glycine max*), barrel medic (*Medicago truncatula*), poplar (*Populus trichocarpa*), rice (*Oryza sativa*), grape (*Vitis vinifera*), and Swiss-Prot. For transcript-based annotations of the six new assemblies, RNA for RNA-seq was extracted using Qiagen RNeasy Plant (Hilden, Germany) from each accession from well-hydrated and drought-stressed young seedling root and leaves, immature flower buds, and pods 5 days after pollination, and from developing seeds of Suvita-2, TZ30, and ZN016 (not CB5-2, Sanzi, or UCR779) 13 days after pollination. RNA quality was assessed, and concentrations were determined using an Agilent 2100 BioAnalyzer (Santa Clara, California, USA) and the Agilent RNA 6000 Nano Kit. The RNA-seq short reads from each accession were assembled using a JGI in-house genome-guided assembler, PERTRAN (Shu et al., 2013), using each genome assembly. Each short-read-based assembly and UNIGENE sequences (P12\_UNIGENES.fa from harvest.ucr.edu) were fed into PASA (Haas et al., 2003) to produce transcriptome assemblies. The best gene per locus (based on evidence) was defined using PASA from alignment of transcriptome assemblies for splicing correctness, alternative transcripts, and UTR addition. The PASA genes were filtered to obtain the final gene set, including an automated repeat coding sequence (CDS) overlap filter, a manual low-quality gene filter, and an automatic filter from transposable element (TE) protein domain assignments. This process was repeated once with one additional homology seeding of non-self, high-confidence gene models.

## 2.4 | Determination of core and noncore genes among seven accessions

Core and noncore genes were determined by running the GET\_HOMOLOGUES-EST tool ([https://github.com/ead-csic-compbio/get\\_homologues](https://github.com/ead-csic-compbio/get_homologues)) on the primary transcripts of the seven cowpea accessions provided in nucleotide and protein formats. GET\_HOMOLOGUES-EST was run

in orthoMCL mode, as suggested by the authors for pan-genome analyses (Contreras-Moreira et al., 2017). The other GET\_HOMOLOGUES-EST options “-M-c -z -t 0 -A -L” were used to obtain orthoMCL gene clusters, which had genes in 1–7 accessions. The term “core” means that a matching gene was identified in all seven accessions and “noncore” means that a matching copy gene was identified in less than all seven accessions.

GO-term enrichment analyses were performed in agriGO v2.0 (Tian et al., 2017) for core and noncore genes using GO terms available from the Legume Information System (LIS; <https://www.legumeinfo.org/>). Given the large number of GO terms in both the core and noncore gene sets, GO slims (Onsongo et al., 2008) were extracted. The full list of core and noncore genes, with GO and other annotations, is available from the Google Drive noted in the Data Availability Statement.

## 2.5 | Annotation of variants in core and noncore genes

To test if variants in noncore genes have been subject to reduced selective constraint, variant effect predictor (VeP) (McLaren et al., 2016) was used to annotate variants identified in the primary transcripts of core and noncore genes. Gene annotations for IT97K-499-35 were used to identify intervals that overlap core and noncore genes, and filtering of the VCF file used BEDtools intersect (Quinlan & Hall, 2010) with variants called relative to the IT97K-499-35 assembly using the six other assemblies. Scripts used for these analyses are at [https://github.com/MorrellLAB/Cowpea\\_Pangenome](https://github.com/MorrellLAB/Cowpea_Pangenome). VeP was run separately for SNPs and indels, reporting classes of variants with potentially large effects, including missense, stop gains, start or stop changes, and frameshifts. The numbers of synonymous changes and in-frame indels are also reported.

## 2.6 | Relative size of core and noncore genes

The physical sizes of core and noncore genes were compared in the total annotated length and the length of the coding portion of the primary transcript of each gene. The length of each gene was extracted from the general feature format (GFF) annotations. The CDS length was calculated based on the primary transcript identified in Phytozome annotations ([https://phytozome-next.jgi.doe.gov/cowpeapan/info/Vunguiculata\\_v1\\_2](https://phytozome-next.jgi.doe.gov/cowpeapan/info/Vunguiculata_v1_2)). The full list of core and noncore genes, with gene and CDS sizes indicated, is available from the Google Drive noted in the Data Availability Statement.

## 2.7 | Nucleotide sequence diversity in cowpea

Tajima's (1983) estimate of  $\theta = 4N_e\mu$  was used to determine the level of sequence diversity in the pan-genome accessions. “Callable” regions were identified based on coverage estimates in mosdepth (Pederson & Quinlan, 2018), with “callable” regions defined as those with coverage between 5× and 400×. This estimate was derived from a sample with ~200× average coverage. The callable regions were used to create a BED file used for filtering genomic regions. This approach is intended to avoid variant calls in regions with inadequate sequence depth or regions where very high coverage may indicate non-unique mapping of sequence reads. The callable regions and the VCF file of filtered variants mapped to the IT97K-499-35 reference were used with pixy (Korunes & Samuk, 2021), a tool designed to deal with missing data in genome-level resequencing datasets.

## 2.8 | Physical locations of SNPs from genotyping platforms

The physical positions of SNPs in the Illumina iSelect Cowpea Consortium Array (Muñoz-Amatriáin et al., 2017), whose positions in the IT97K-499-35 genome were provided in Lonardi et al. (2019), were mapped using BWA MEM (Li & Durbin, 2009) within each of the seven assemblies using the contextual sequence that flanked each variant. The resulting alignment file was processed with SAMtools (Li et al., 2009) and SNP\_Utills ([https://github.com/MorrellLAB/SNP\\_Utills](https://github.com/MorrellLAB/SNP_Utills)) to report positions in a VCF file. The positions of iSelect SNPs relative to all seven genome assemblies are provided in Table S02, and an updated summary map for the 51,128 iSelect SNPs is in Table S03. The positions identified for iSelect SNPs relative to the IT97K-499-35 assembly were used to annotate the variants. The annotation used VeP (McLaren et al., 2016) with the GFF file provided by Phytozome (<https://genome.jgi.doe.gov/portal/>) and SNP positions in VCF files ([https://github.com/MorrellLAB/cowpea\\_annotation/blob/main/Results/IT97K-499-35\\_v1.0/iSelect\\_cowpea.vcf](https://github.com/MorrellLAB/cowpea_annotation/blob/main/Results/IT97K-499-35_v1.0/iSelect_cowpea.vcf); see Data Availability Statement).

## 2.9 | Synteny analysis among genome assemblies

To assess the conservation of gene content and ordering between genome assemblies from diverse species, MCScanX (Wang et al., 2012) was run for every genome pair, using default settings and homologous gene pairings derived from

gene family assignments defined as the best match of the longest protein product with an E-value of  $1e-10$  or better from hmmsearch (Eddy, 2011) applied to the legfed\_v1\_0 families (Stai et al., 2019).

### 3 | RESULTS AND DISCUSSION

#### 3.1 | Development of six de novo assemblies and pan-genome construction

Summary statistics for the seven assemblies (assembly characteristics, repetitive content, genes, BUSCO completeness) are reported in Table 1. More detailed statistics of the intermediate assembly steps are reported in Table S04. The contiguity of the new six assemblies, as indicated by their N50s, is comparable to the PacBio assembly for IT97K-499-35 despite being based on short-read sequences. In all six new assemblies, each of the 11 chromosomes of cowpea is represented by a single scaffold. These six assembled genomes are similar to each other in size, ranging from 447.58 to 453.97 Mb, with a mean of 449.91 Mb. IT97K-499-35 had a ~15% larger (more complete) assembled size (519.44 Mb) than these six accessions, with the difference attributable to long-read sequencing and optical mapping providing a more complete assembly. Assemblies of the six additional accessions share the same percentage of repetitive content of about 45%–46% (Table 1 and Figure S1). The IT97K-499-35 assembly has a somewhat higher repetitive content than the assemblies of these six accessions. This may be attributable to more complete resolution of unique positions of repetitive sequences within long sequence reads than is possible from only short reads. A difference between the sequencing methods in the resolution of repetitive sequences is evident in centromeric regions, which are typically abundant in repetitive sequences, where some chromosomes of the six newly sequenced accessions appear to be missing from the assemblies. Centromeric regions were defined based on a 455-bp tandem repeat previously identified by fluorescence in situ hybridization (Iwata-Otsubo et al., 2016). Table S05 shows the coordinates of the putative centromeric regions in IT97K-499-35 for all 11 chromosomes for a total span of 20.18 Mb, in CB5-2 on five chromosomes for a total span of 5.6 Mb, in Sanzi on one chromosome for a total span of 0.59 Mb, in ZN016 on four chromosomes for a total of 7.13 Mb, and TZ30 on one chromosome for 1.32 Mb. The tandem repeat was not found in any assembled chromosome of Suvita-2 or UCR779, nor in the other chromosome assemblies where coordinates are not listed.

RNA was prepared from each accession to support gene annotation, and the same annotation protocol was applied to each accession (see Materials and Methods). This is important when comparing genomes at the gene level, as it reduces the technical variability that can otherwise obfuscate the

TABLE 1 Summary of assembly statistics, repetitive content, gene content, and BUSCO4 completeness for the seven genomes

	IT97K-499-35	CB5-2	Suvita-2	Sanzi	UCR779	ZN016	TZ30
Assembly size (bp)	519,435,864	448,043,751	447,585,192	447,277,261	453,970,486	451,130,807	451,468,680
N50 (bp)	41,684,185	36,897,245	36,142,647	34,759,918	35,700,653	37,764,243	36,906,789
#Contigs/scaffolds	686	6534	9123	11,268	12,939	7032	6771
#Contigs/scaffolds $\geq$ 100 kbp	103	28	28	17	13	28	48
#Contigs/scaffolds $\geq$ 1Mbp	13	11	11	11	11	11	11
#Contigs/scaffolds $\geq$ 10Mp	11	11	11	11	11	11	11
Longest contig (bp)	65,292,630	60,086,998	58,539,223	58,655,738	58,369,212	60,653,587	59,481,915
Repetitive content	47.25%	45.52%	45.43%	45.50%	45.89%	45.68%	45.76%
Annotated genes (#)	31,948	28,297	28,545	28,461	28,562	27,723	27,742
<b>BUSCO completeness</b>							
Genome	1595	1574	1580	1581	1574	1589	1583
Transcripts	1594	1570	1582	1585	1581	1584	1580
Proteins	1595	1569	1584	1587	1585	1584	1582
	98.8%	97.5%	97.8%	97.9%	97.9%	98.2%	98.1%
	98.8%	97.2%	98.0%	98.2%	98.2%	98.1%	97.8%
	98.8%	97.3%	98.2%	98.3%	98.2%	98.1%	98.0%

interpretation of results (Lei et al., 2021). The number of genes annotated in the six new assemblies ranged from 27,723 to 28,562, with a mean of 28,222 (Table 1). IT97K-499-35 had ~13% more annotated genes, with a total of 31,948, reflecting deeper transcriptome sequencing and, to some extent, the more complete assembly of its genome. Table S06 summarizes the number of alternative transcripts, exon statistics, gene model support, and ontology annotations (Panther, PFam, KOG, KEGG, and E.C.). The number of alternative transcripts in the six new assemblies ranged from 15,088 to 17,115. Again, IT97K-499-35 had a higher number of alternative transcripts, a total of 22,536, than the other six accessions. The average number of exons was 5.4 in each of the six new assemblies and 5.2 in IT97K-4899-35, with a median length ranging from 162 to 169 bp. Gene and repeat density were computed in 1 Mb non-overlapping sliding windows along each chromosome and each accession (Figure S1). All chromosomes have a higher gene density in their more recombinationally active regions, while repeat density peaks in the low-recombination centromeric and pericentromeric regions (see also Figure S8 in Lonardi et al., 2019). All seven accessions have similar gene and repeat density, and high BUSCO v4 completeness at the genome, transcript, and protein levels (Table S07), with somewhat higher numbers for IT97K-499-35 than the six new assemblies.

As stated above (Materials and Methods), genes annotated in the seven genomes were classified as core if a matching gene was present in all accessions and noncore if absent in one or more of the seven accessions. In IT97K-499-35, a total of 26,026 core genes (in 24,476 core clusters) and 4963 noncore genes (in 4285 noncore clusters) were identified (Table S08). When considering all seven accessions itemized in Table S08, a total of 26,494 core genes and 9042 noncore genes (in 8157 noncore clusters) were identified, resulting in a total of 35,536 pan genes in 32,633 pan gene clusters.

To determine if adding accessions significantly changed the numbers and proportions of core and noncore genes, we took advantage of the analysis results produced by GET\_HOMOLOGUES-EST. GET\_HOMOLOGUES-EST produces pan or core genome growth simulations by adding accessions in random order, using 20 permutations. Figure 2 shows the growth of core and pan genomes for an increasing number of accessions. A fitted Tettelin function (Tettelin et al., 2005) is plotted in green. As expected, the number of pan genes increases as additional accessions are “added” to the pan-genome, while the number of core genes decreases. However, the fact that the core gene plot is flattening considerably (approaching an asymptotic limit) for six and seven accessions indicates that most core genes have been identified with these seven diverse accessions. In contrast, the pan-genome plot has not flattened, indicating that there may be many more noncore genes not included among these seven accessions. Figure 2 provides an estimated 29,659 pan

gene clusters and an estimated 24,439 core gene clusters as the output of GET\_HOMOLOGUES-EST from 20 random samplings. Roughly, it appears that the pan-genome defined by the seven cultivated cowpea accessions is composed of about 80% core genes, constituting nearly the entire set of core genes in cultivated cowpea, and 20% noncore genes. Clearly, more noncore genes would be revealed with a larger number of accessions.

A GO term enrichment analysis was performed for genes within the two components of the pan-genome (core and non-core) using agriGO v2 (Tian et al., 2017). Many GO terms for all three ontology aspects (biological process, cellular component, and molecular function) were significantly enriched in both core and noncore genes (Table S09). Given the high number of significant GO terms, GO Slim terms (Onsongo et al., 2008) were extracted and used for Figure 3. Terms enriched in the core genes were related to transport and some metabolic processes and molecular functions involving DNA-binding transcription factor activity (Figure 3; Table S09). This supports the idea that the core genome contains genes that perform essential cellular functions that are highly conserved at the species level. The output was quite different for the noncore genes, with very high enrichment of the GO term “response to stress” (Figure 3), in particular “defense response” ( $-\log_{10}q = 123.7$ ; Table S09). This is consistent with previous research showing that the “dispensable” genome encodes genes involved in defense response and other beneficial functions for some individuals (Golicz et al., 2016; Gordon et al., 2017; Montenegro et al., 2017).

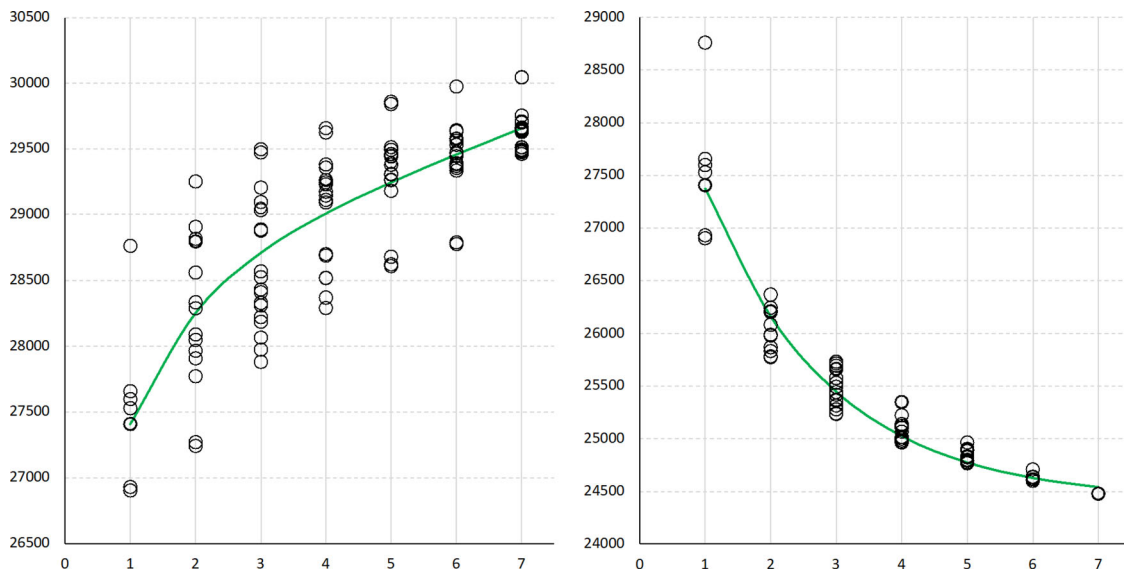
### 3.2 | Genetic variation analysis

In addition to identifying gene PAVs, the seven assemblies were used to identify other types of variation. Variants were detected using two different software pipelines, depending on their size. SNPs and indels of length up to 300 nucleotides, both considered small variants, were detected using GATK (see Materials and Methods). Larger structural variations, including deletions, duplications, inversions, and translocations, were detected using SyRI (Goel et al., 2019).

Across all “callable” regions of the genome, average  $\theta_{\pi} = 0.0111 (\pm 0.0549)$ . At the pseudomolecule level, average diversity was highest on Vu05, with  $\theta_{\pi} = 0.0155 (\pm 0.0723)$ , and lowest on Vu10, with  $\theta_{\pi} = 0.0095 (\pm 0.0447)$  (Table S10). A mean diversity of ~1% is higher than many grain crops, such as barley (Morrell et al., 2014; Schmid et al., 2018) and roughly comparable to maize (Tittes et al., 2021). The observed diversity in the cowpea pan-genome sample is above average for herbaceous plants (Corbett-Detig et al., 2015; Leffler et al., 2012; Miller & Gross, 2011).

For SNPs and indels, the genome of each accession was used in turn as the “reference,” mapping the reads for each





**FIGURE 2** The number of genes identified in the pan-genome (pan genes) and core genome (core genes) as new accessions are added. Green curves are fitted Tettelin functions.

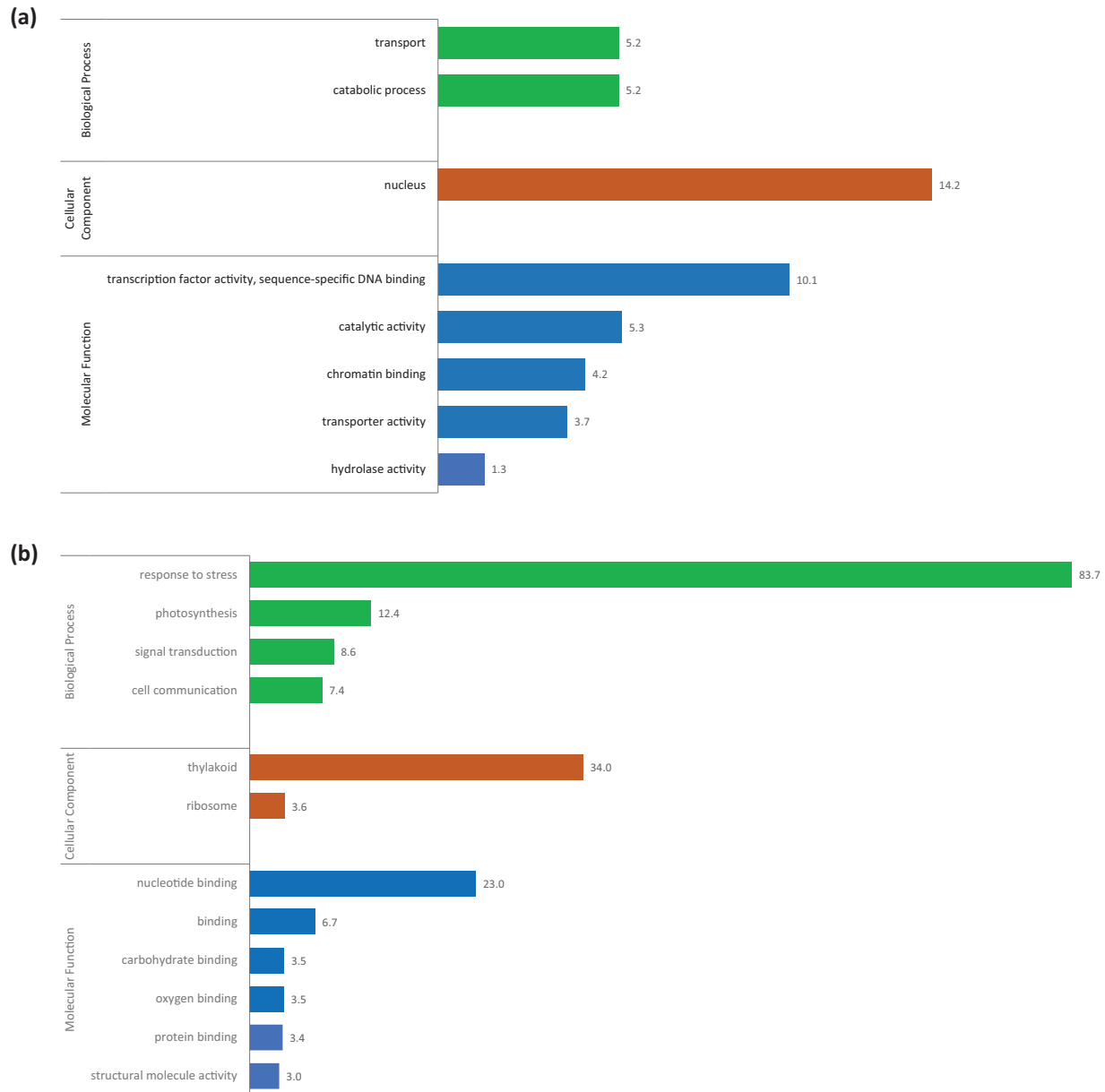
of the six other accessions against that genome. For each, the six SNP sets produced by GATK were merged by taking the union of the SNPs based on their location (i.e., an SNP in two accessions was counted only once if it appeared in the same genomic position). Table S11 summarizes the number of SNPs detected, where the reference genome is listed on each row. For instance, using Suvita-2 as the reference, 1,489,850 SNPs were detected using mapped reads from CB5-2, compared to 2,625,678 SNPs using the reads from UCR779. Combining the SNPs by counting all distinct SNPs in the union of the six sets of SNPs, the number of SNPs for Suvita-2 was 5,292,933.

When UCR779 was used as the reference, a much higher number of SNPs was detected in every pairwise comparison, indicating that UCR779 is the most divergent among these seven accessions. Conversely, CB5-2 (a California cultivar) has fewer SNPs in pairwise comparisons to TZ30 or ZN016 (both from China) than in pairwise comparisons to other accessions. This suggests that CB5-2 is more similar to these two accessions than to the other four accessions. This is consistent with genetic assignment analyses reported by Muñoz-Amatriain et al. (2021) and historical considerations discussed in Herniter et al. (2020). Table S12 provides a similar analysis for indels, where again, UCR779 stands out as the most different among the seven accessions. Summary statistics for SNPs and indels for each chromosome and each accession can be found in the file “SNPs\_indels\_stats.xlsx,” available from the Google Drive indicated in the Data Availability Statement.

GATK requires a minimum coverage of 5 $\times$  to call SNPs. Coverage analysis with Mosdepth indicated that the average read coverage of IT97K-499-35 is very high (e.g., about

$\sim 190\times$  when mapping CB5-2 reads to IT97K-499-35), thus a very high fraction of IT97K-499-35 chromosomes was covered by at least five reads. The lowest was Vu10 with 85.1%, the highest was Vu07 with 98.6%, and the overall percentage of SNPs in IT97K-499-35 that were in a “callable” region (i.e., with coverage 5 $\times$ –400 $\times$ ) was 88.96%. The frequency of SNPs, as the number of unique SNPs identified (Table S11) divided by the size of the assembled genome (Table 1), ranges from one in 139 to one in 309 bp, and the indel frequency (Table S12) ranges from one in 486 to one in 529 bp. Circos plots for SNP density (SNPs per Mb) on each chromosome using each accession as the reference are in Figure S2A–G, where it is evident, for example, that Vu04 and Vu10 have the highest SNP frequency. In contrast, Vu05 and Vu09 have the lowest. This was observed previously when mapping nearly one million SNPs on the IT97K-499-35 reference genome (Lonardi et al., 2019). Also, when using UCR779 as the reference (Figure S2E), the number of SNPs on Vu04 and Vu10 is significantly higher than when any other accession is used as the reference, again consistent with UCR779 being the most different among the seven accessions.

Structural variations were identified using SyRI (Goel et al., 2019) from the alignment of each pair of individual genomes and visualized using PlotSR (Goel & Schneeberger, 2022) (Figure 4). The visualization shows a relatively large number of apparent structural rearrangements between the seven cowpea genomes, which are more abundant in the centromeric and pericentromeric regions of all chromosomes. Vu04 is the chromosome with the highest abundance of structural variants (Figure 4). A summary of all the structural variants identified in all pairs of accessions is reported in Table S13. The table shows that Suvita-2 versus UCR779 had



**FIGURE 3** Gene ontology (GO) term enrichment analysis. Significantly enriched GO terms for core (a) and noncore genes (b) are shown for GO-Slim categories belonging to Biological Process, Cellular Component, and Molecular Function aspects (in different colors).  $-\log_{10}$  of FDR-adjusted  $p$ -values ( $q$ -values) are shown on the right of each bar.

the largest number of inversions (2008) and translocations (1822). This intuitively makes sense since these two accessions belong to two different genetic subpopulations separated by the first principal component (Figure 1).

Inversions are a common type of rearrangement with important consequences for cross-over frequency and distribution, as they suppress recombination in heterozygotes (Kirkpatrick, 2010). While inversion can be important to maintain locally adaptive variants (Kirkpatrick & Barton, 2006), crossover inhibition can impede plant breeding efforts. Table 2 summarizes the genomic coordinates of all inversions larger than 1 Mbp. For example, the first column of Table 2,

corresponding to IT97K-499-35, shows 27 inversions that were identified by comparing the reference genome against the other six accessions. The same inversion can appear in multiple sub-tables. For instance, the  $\sim 4.2$  Mb inversion on chromosome 3 previously described in Lonardi et al. (2019) occurs in the same orientation in six accessions and the opposite orientation only in IT97K-499-35, so it is listed six times in the column for IT97K-499-35.

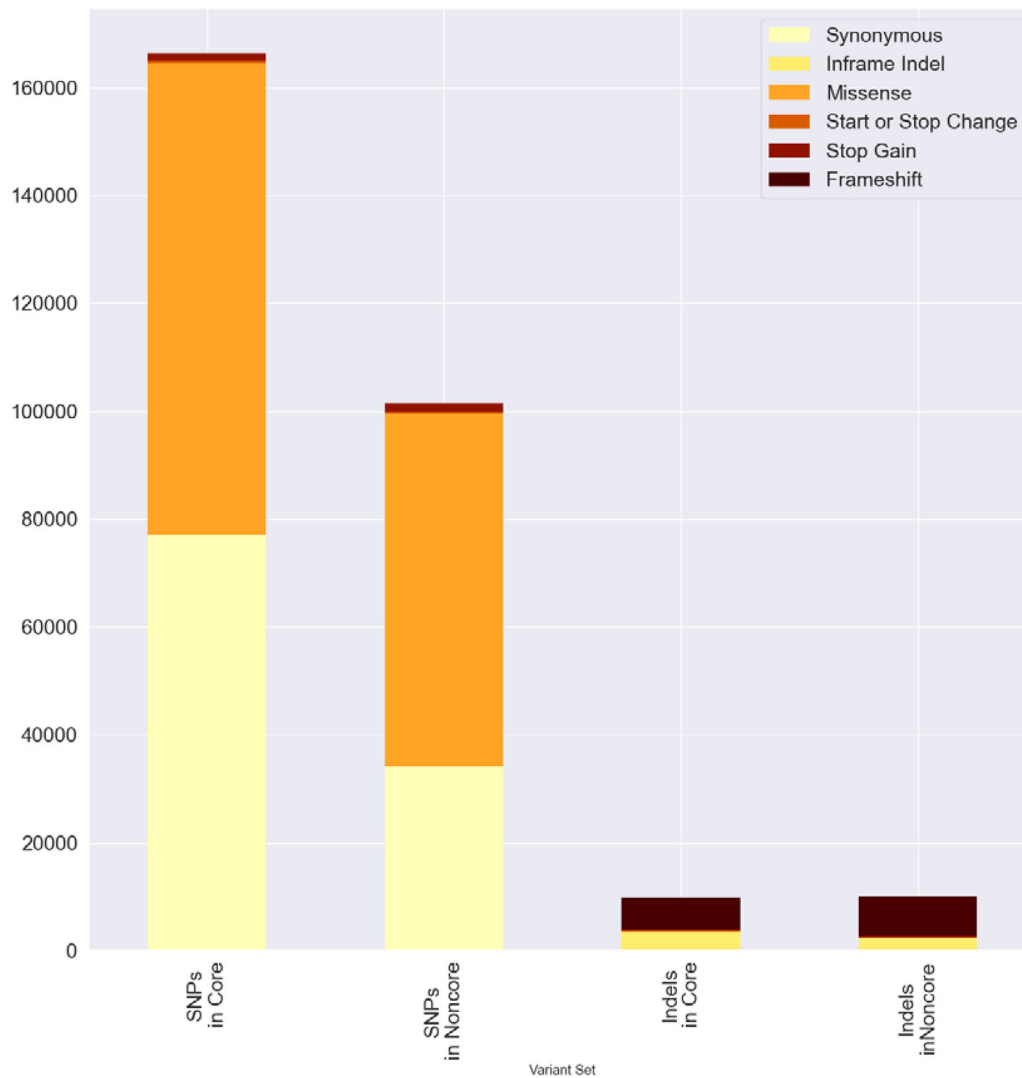
Similarly, the inversions on Vu04 and Vu05 are detected against five accessions. The  $\sim 9.0$  Mb inversion on Vu06 is the largest inversion found by SyRI, and its orientation is unique to Suvita-2. However, this inversion appears to be



**FIGURE 4** Representation of structural variations (of any size) detected by SyRI from the output of whole-genome pairwise alignments between the seven cowpea accessions. The black track indicates gene density in the reference genome IT97K-499-35, while the blue track indicates SNP density in the reference genome IT97K-499-35.

due to an assembly imperfection. It is reported as unoriented in the ALLMAPS output (Table S14), and comparisons between optical maps derived from Suvita-2 and another cowpea accession not included here indicate a non-inverted orientation in Suvita-2 (unpublished). Also, as shown in Lonardi et al. (2019) and Figure S3, this entire region has a very low recombination rate and comprises nearly the entire short arm of acrocentric chromosome 6 (Iwata-Otsubo et al., 2016). These factors can account for a spurious orientation assignment for this region in the Suvita-2 Vu06 assembly.

The positions of the largest inversions shown in Figure 4 are provided in Table 2, for example, the inversions on Vu03 in IT97K-499-35 reported by Lonardi et al. (2019), and the inversion on Vu06 in Suvita-2 likely due to a mis-assembly, as discussed above. It should be noted that regions with apparently low synteny within several chromosomes are low-recombination centromeric and pericentromeric regions (Lonardi et al., 2019), which are notoriously hard to assemble due to their high repetitive content and hard to orient due to a paucity of mapped and recombinationally ordered SNPs.



**FIGURE 5** Variant effect predictor (VeP) annotations for SNPs and indels found in the core and noncore genes present in IT97K-499-35. Values on the y-axis are the absolute number of variants in each variant class.

In these regions, it is expected to find compressed contigs, gaps, and misassemblies, any of which might be flagged as apparent structural variations. The number of false-positive structural variations can likely be reduced by increasing the completeness of the assemblies within these regions using long-read sequencing and optical mapping. Figure S4A–U shows all 21 SyRi+PlotSR alignments between all pairs of cowpea accessions.

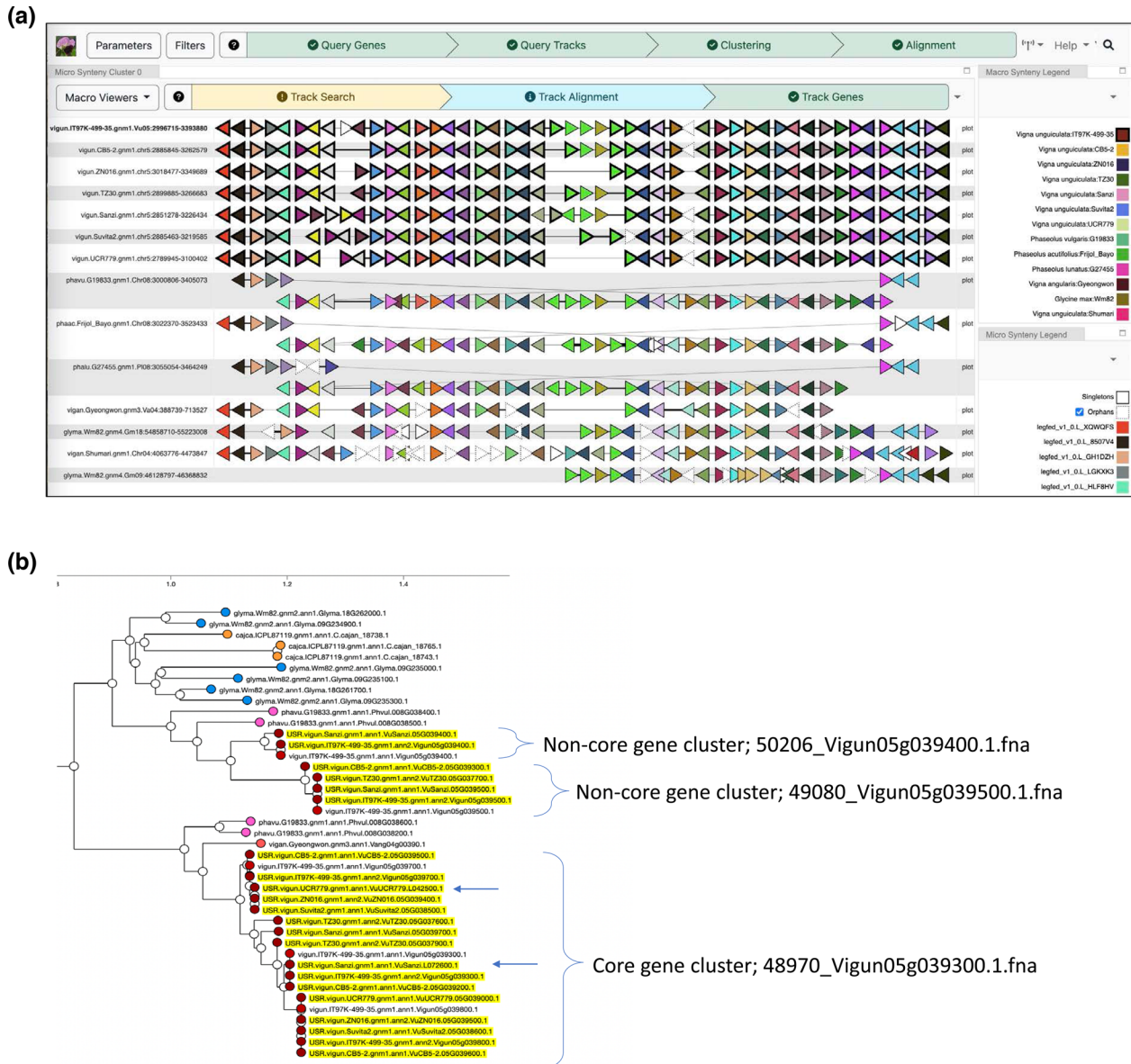
### 3.3 | Further characterization of core and noncore genes

Partitioning SNPs into those found in core versus noncore genes in IT97K-499-35 resulted in 702,073 SNPs in core genes and 239,100 SNPs in noncore genes. The indel comparison involves 161,900 indels in core genes and 39,845 in noncore genes. The numbers of variants with potential conse-

quences are summarized in Figure 5 and Table S15. Counting both SNPs and indels, there are 80,693 potentially benign variants among core genes (3.10 per gene) and 36,519 in noncore genes (7.36 per gene), which is a 2.37-fold higher frequency in noncore versus core genes. Likewise, potentially harmful variants, including missense, stop gained, start or stop change, and frameshift total 95,465 among core genes (3.67 per gene) and 75,048 in noncore genes (15.12 per gene), which is a 4.12-fold higher incidence in noncore versus core genes. Among these, noncore genes have a much higher incidence of frameshift variants (1.48 per gene) than do core genes (0.23 per gene), this being a 6.43-fold difference. In each of these comparisons, noncore genes contribute proportionally a larger number of variants than do core genes, whether benign or potentially harmful.

Based on the gene annotations, core gene primary transcripts are longer than noncore gene primary transcripts, with a mean length of 4226.08 ( $\pm$  4047.234) for IT97K-499-35





**FIGURE 6** Conservation of gene content within and across species. (a) A region depicting gene content conservation and variability among cowpea genomes and other representative Phaseoleae species. Triangular glyphs represent order and orientation of genes, with color representing gene family memberships (<https://vigna.legumeinfo.org/tools/gcv>). (b) All cowpea proteins assigned to the family whose members exhibit copy number variation in (a) are shown augmenting a dynamically recomputed gene tree at the Legume Information System, with genes from unanchored contigs not present in the chromosomes aligned in (a) indicated with arrows (<https://mines.legumeinfo.org/cowpeamine>).

whole genome duplication in soybean. The region serves as a breakpoint for the syntenic block in Gm09, which, taken together with the other structural variation, suggests that the expansion of gene copy number here has had consequences for the stability of the chromosome in these regions over evolutionary time (Hastings et al., 2009).

Although the GCV view shows good evidence for CNV, there are some limitations to what may be inferred from that alone. First, since the viewer only has access to gene family assignment information, it cannot determine which elements among those in tandem arrays have the highest sequence simi-

larity and provide insight into which copies have been deleted. Second, because it relies on the surrounding genomic context of each gene to place it into correspondence, it will have limited capability for finding genes that are present in the assembly but are largely isolated on small scaffolds that were not incorporated into the main pseudomolecules. Another tool at LIS that provides a complementary view based on the underlying sequence identity of the different copies of the expanded gene family is shown in Figure 6b. Here, the InterMine (Kalderamis et al., 2014) instance for cowpea (<https://mines.legumeinfo.org/cowpeamine/begin.do>) was used to collect all

protein sequences for cowpea genes assigned to the given family. A dynamic tree construction procedure invoked based on hmalign-derived (<http://www.csb.yale.edu/userguides/seq/hmmer/docs/node18.html>; Eddy, 2011) additions of these genes to the multiple sequence alignment for the founding members of the family. The resulting tree (a subtree of which is shown) allows the user to determine the best correspondences of the copies in each genome and pulls in two additional genes on unanchored contigs that likely belong to the region.

### 3.5 | Pan-genome core genes and cross-species synteny

To explore the question of how within-species gene content conservation compares with gene content shared between species in other species and genera, we used the LIS gene family assignments to define homology pairings between all members of each gene family, then used the resulting data to determine collinearity blocks among all pairwise comparisons of the cowpea genomes, as well as to soybean and representative genomes from *Vigna* and *Phaseolus* spp. The counts of genes participating in at least one collinear block were tallied for each genome in each pairwise comparison. As expected, intra-specific comparisons between cowpea accessions yield higher numbers of conserved collinear genes than inter-specific comparisons. On the other hand, there is no appreciable difference in the extent of conserved collinearity when comparing cowpea genomes to other species within the *Vigna* genus versus species from *Phaseolus* or *Glycine* genera (Figure S5). Because soybean has an additional whole genome duplication relative to all other species in the comparison, the total number of soybean genes found in collinear blocks is higher than in other comparisons. Comparisons between all species and the *Vigna radiata* version 6 genome (Kang et al., 2014) show fewer conserved collinear genes, but this is presumably due to missing data in that assembly, given that all other inter-specific comparisons are similar.

#### AUTHOR CONTRIBUTIONS

**Qihua Liang:** Data curation; formal analysis; investigation; methodology; writing—original draft; writing—review and editing. **María Muñoz-Amatriáin:** Conceptualization; formal analysis; funding acquisition; investigation; methodology; writing—original draft; writing—review and editing. **Shengqiang Shu:** Data curation; formal analysis; methodology; software; writing—original draft. **Sassoum Lo:** Conceptualization; resources; writing—review and editing. **Xinyi Wu:** Data curation; funding acquisition; resources; writing—review and editing. **Joseph W. Carlson:** Data curation; visualization; writing—review and editing. **Patrick Davidson:** Data curation; visualization. **David M. Goodstein:** Data

curation; visualization. **Jeremy Phillips:** Data curation; visualization. **Nadia M. Janis:** Formal analysis; visualization. **Elaine J. Lee:** Data curation; formal analysis; visualization; writing—review and editing. **Chenxi Liang:** Formal analysis; writing—review and editing. **Peter L. Morrell:** Data curation; formal analysis; investigation; methodology; software; supervision; visualization; writing—original draft; writing—review and editing. **Andrew D. Farmer:** Data curation; formal analysis; investigation; methodology; resources; software; supervision; visualization; writing—original draft; writing—review and editing. **Pei Xu:** Conceptualization; funding acquisition; investigation; resources; supervision; writing—review and editing. **Timothy J. Close:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; writing—original draft; writing—review and editing. **Stefano Lonardi:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; software; supervision; writing—original draft; writing—review and editing.

#### ACKNOWLEDGMENTS

Development of an initial pan-genome of domesticated cowpea (*Vigna unguiculata* ssp. *unguiculata* [L.] Walp.) was an objective of the National Science Foundation BREAD project “Advancing the Cowpea Genome for Food Security” (IOS 1543963). Funding was also provided by NSF IIS-1814359 (“Improving de novo Genome Assembly using Optical Maps”). The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, was supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Support for the work of ADF provided by the U.S. Department of Agriculture, Agricultural Research Service, Non-Assistance Cooperative Agreement 58-5030-7-069. This work also benefited from and addressed breeding and germplasm management objectives of the Feed the Future Innovation Lab for Climate Resilient Cowpea (USAID Cooperative Agreement AID-OAA-A-13-00070). Partial support was also provided by the National Natural Science Foundation of China (32172568), the Major Science and Technology Project of Plant Breeding in Zhejiang Province (2021C02065-6-3), and the National Ten-Thousand Talents Program of China (to Pei Xu). Hatch Project CA-R-BPS-5306-H also provided partial support. The authors thank Staff Research Associate Yi-Ning Guo (UCR) for RNA preparation; Programmer Steve Wanamaker (UCR) for informatics assistance; Jeffrey Ehlers (Bill & Melinda Gates Foundation), Phillip Roberts (UCR), and Anthony Hall (UCR) for discussions on the accessions for sequencing; Bao Lam Huynh (UCR) and Mitchell Lucas (UCR) for assistance with greenhouse operations for pure seed production; Richard Hayes (Department

of Energy Joint Genome Institute, Berkeley, California, USA) for assistance with Phytozome; Samuel Hokin (National Center for Genome Resources, Santa Fe, New Mexico, USA) for data curation for loading data into CowpeaMine; and Bruno Contreras-Moreira (Spanish National Research Council) for helpful discussions on the use of GET\_HOMOLOGUES-EST. The University of Minnesota Supercomputing Institute also provided hardware and software support.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The genome assemblies and annotations described in this manuscript are available from CowpeaPan (<https://phytozome-next.jgi.doe.gov/cowpeapan>). Raw DNA and RNA sequence data from IT97K-499-35 and whole genome shotgun DNA sequences for 36 diverse cowpea accessions used for SNP discovery in Muñoz-Amatriáin et al. (2017) are available from the National Center for Biotechnology Information (NCBI) as SRA accessions SRS3721827, SRP077082, SAMN071606186 through SAMN071606198, SAMN07194302 through SAMN07194309, and SAMN07194882 through SAMN07194909, as stated in Lonardi et al. (2019). Raw DNA and RNA sequence data from the six additional accessions providing de novo assemblies in this report, and sequences produced by 10× Genomics from IT97K-49-35, are available as BioProject PRJNA836573 from the National Center for Biotechnology Information (NCBI). More complete annotation files, assemblies and SNPs are also available via the <https://drive.google.com/drive/folders/1iQaLW4SLmN2IP7q4k3uovHK3SvsxGbVi?usp=sharing> Google shared drive link.

## ORCID

Timothy J. Close  <https://orcid.org/0000-0002-9759-3775>

## REFERENCES

- Aggarwal, V. D., Muleba, N., Drabo, I., Souma, J., & Mbewe, M. (1984). *Inheritance of Striga gesnerioides resistance in cowpea*. In C. Parker, L. J. Musselman, R. M. Polhill, & A. K. Wilson (Eds.), *Proceedings of 3rd International Symposium on Parasitic Weeds* (pp. 143–147). ICARDA.
- Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6, 914–920. <https://doi.org/10.1038/s41477-020-0733-0>
- Boukar, O., Belko, N., Chamarthi, S., Togola, A., Batiemo, J., Owusu, E., Haruna, M., Diallo, S., Umar, M. L., Olufajo, O., & Fatokun, C. (2019). Cowpea (*Vigna unguiculata*): Genetics, genomics and breeding. *Plant Breeding*, 138, 415–424. <https://doi.org/10.1111/pbr.12589>
- Boukar, O., Bhattacharjee, R., Fatokun, C., Kumar, P. L., & Gueye, B. (2013). Cowpea. In M. Singh, H. D. Upadhyaya, & I. S. Bisht (Eds.), *Genetic and genomic resources of grain legume improvement* (Chapter 6, pp. 137–156). Elsevier.
- Chapman, J. A., Ho, I., Sunkara, S., Luo, S., Schroth, G. P., & Rokhsar, D. S. (2011). Meraculous: *De novo* genome assembly with short paired-end reads. *PLoS ONE*, 6, e23501. <https://doi.org/10.1371/journal.pone.0023501>
- Cleary, A., & Farmer, A. (2018). Genome Context Viewer: Visual exploration of multiple annotated genomes using microsynteny. *Bioinformatics*, 34, 1562–1564. <https://doi.org/10.1093/bioinformatics/btx757>
- Contreras-Moreira, B., Cantalapiedra, C. P., García-Pereira, M. J., Gordon, S. P., Vogel, J. P., Igartua, E., Casas, A. M., & Vinuesa, P. (2017). Analysis of plant pan-genomes and transcriptomes with GET\_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Frontiers in Plant Science*, 8, 184. <https://doi.org/10.3389/fpls.2017.00184>
- Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *PLOS Biology*, 13, e1002112. <https://doi.org/10.1371/journal.pbio.1002112>
- Dash, S., Campbell, J. D., Cannon, E. K. S., Cleary, A. M., Huang, W., Kalberer, S. R., Karingula, V., Rice, A. G., Singh, J., Umale, P. E., Weeks, N. T., Wilkey, A. P., Farmer, A. D., & Cannon, S. B. (2016). Legume information system (LegumeInfo.org): A key component of a set of federated data resources for the legume family. *Nucleic Acids Research*, 44, D1181–1188. <https://doi.org/10.1093/nar/gkv1159>
- de Mooy, B. E. (1985). *Germplasm evaluation of Botswana cowpea (Vigna unguiculata [L.] Walp.) landraces* (M.S. Thesis). Michigan State University.
- Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7, e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- Ehlers, J. D., Fery, R. L., & Hall, A. E. (2002). Cowpea breeding in the USA: New varieties and improved germplasm. In C. S. Fatokun, S. A. Tarawali, B. B. Singh, P. M. Kormawa, & M. Tamò (Eds.), *Challenges and Opportunities for Enhancing Sustainable Cowpea Production. Proceedings of the World Cowpea Conference III held at the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria, 4–8 September 2000* (Chapter 1.6, pp. 62–77). IITA.
- Goel, M., & Schneeberger, K. (2022). plots: Visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, 38, 2922–2926. <https://doi.org/10.1093/bioinformatics/btac196>
- Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 20, 277. <https://doi.org/10.1186/s13059-019-1911-0>
- Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pan-genome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, 7, 13390. <https://doi.org/10.1038/ncomms13390>
- Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics comes of age: From bacteria to plant and animal applications. *Trends in Genetics*, 36, 132–145. <https://doi.org/10.1016/j.tig.2019.11.006>
- Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., Stritt, C., Roulin, A. C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T. E., Amasino, R., ... Vogel, J. P. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, 8, 2184. <https://doi.org/10.1038/s41467-017-02292-8>



- Haas, B. J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*, 5654–5666. <https://doi.org/10.1093/nar/gkg770>
- Hall, A. E. (2004). Breeding for adaptation to drought and heat in cowpea. *European Journal of Agronomy*, *21*, 447–454. <https://doi.org/10.1016/j.eja.2004.07.005>
- Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Review Genetics*, *10*, 551–564. <https://doi.org/10.1038/nrg2593>
- Herniter, I. A., Muñoz-Amatriaín, M., & Close, T. J. (2020). Genetic, textual, and archeological evidence of the historical global spread of cowpea (*Vigna unguiculata* [L.] Walp.). *Legume Science*, *3*, 57. <https://doi.org/10.1002/leg3.57>
- Herniter, I. A., Muñoz-Amatriaín, M., Lo, S., Guo, Y.-N., & Close, T. J. (2018). Identification of candidate genes controlling black seed coat and pod tip color in cowpea (*Vigna unguiculata* [L.] Walp.). *G3*, *8*, 3347–3355. <https://doi.org/10.1534/g3.118.200521>
- Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., De Leon, N., Kaeppeler, S. M., & Buell, C. R. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, *26*, 121–135. <https://doi.org/10.1105/tpc.113.119982>
- Iwata-Otsubo, A., Lin, J.-Y., Gill, N., & Jackson, S. A. (2016). Highly distinct chromosome structures in cowpea (*Vigna unguiculata*), as revealed by molecular cytogenetic analysis. *Chromosome Research*, *24*, 197–216. <https://doi.org/10.1007/s10577-015-9515-3>
- Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., Hu, F., Smith, R., Štěpán, R., Sullivan, J., & Micklem, G. (2014). InterMine: Extensive web services for modern biology. *Nucleic Acids Research*, *42*, W468–472. <https://doi.org/10.1093/nar/gku301>
- Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., Jun, T. H., Hwang, W. J., Lee, T., Lee, J., Shim, S., Yoon, M. Y., Jang, Y. E., Han, K. S., Taepayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K. S., ... Lee, S.-H. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature Communications*, *5*, 5443. <https://doi.org/10.1038/ncomms6443>
- Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLOS Biology*, *8*, e1000501. <https://doi.org/10.1371/journal.pbio.1000501>
- Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*, 419–434. <https://doi.org/10.1534/genetics.105.047985>
- Korunes, K. L., & Samuk, K. (2021). PIXY: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*, 1359–1368. <https://doi.org/10.1111/1755-0998.13326>
- Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P., & Przeworski, M. (2012). Revisiting an old riddle: What determines genetic diversity levels within species. *PLOS Biology*, *10*, e1001388. <https://doi.org/10.1371/journal.pbio.1001388>
- Lei, L., Goltsman, E., Goodstein, D., Wu, G. A., Rokhsar, D. S., & Vogel, J. P. (2021). Plant pan-genomics comes of age. *Annual Reviews of Plant Biology*, *72*, 411–435. <https://doi.org/10.1146/annurev-arplant-080720-105454>
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 genome project data processing subgroup. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lonardi, S., Muñoz-Amatriaín, M., Liang, Q., Shu, S., Wanamaker, S. I., Lo, S., Tanskanen, J., Schulman, A. H., Zhu, T., Luo, M.-C., Alhakami, H., Ounit, R., Hasan, A. M., Verdier, J., Roberts, P. A., Santos, J. R. P., Ndeve, A., Doležel, J., Vrána, J., ... Close, T. J. (2019). The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *The Plant Journal*, *98*, 767–782. <https://doi.org/10.1111/tpj.14349>
- Mackie, W. W. (1946). Blackeye beans in California. *Bulletin*, *696*, 1–56. University of California Agricultural Experiment Station.
- Mclaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl variant effect predictor. *Genome Biology*, *17*, 122. <https://doi.org/10.1186/s13059-016-0974-4>
- Miller, A. J., & Gross, B. L. (2011). From forest to field: Perennial fruit crop domestication. *American Journal of Botany*, *98*, 1389–1414. <https://doi.org/10.3732/ajb.1000522>
- Moghaddam, S. M., Oladzad, A., Koh, C., Ramsay, L., Hart, J. P., Mamidi, S., Hoopes, G., Sreedasyam, A., Wiersma, A., Zhao, D., Grimwood, J., Hamilton, J. P., Jenkins, J., Vaillancourt, B., Wood, J. C., Schmutz, J., Kagale, S., Porch, T., Bett, K. E., ... Mcclean, P. E. (2021). The tepary bean genome provides insight into evolution and domestication under heat stress. *Nature Communications*, *12*, 2638. <https://doi.org/10.1038/s41467-021-22858-x>
- Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*, 1007–1013. <https://doi.org/10.1111/tpj.13515>
- Morgante, M., Depaoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinions in Plant Biology*, *10*, 149–155. <https://doi.org/10.1016/j.pbi.2007.02.001>
- Morrell, P. L., Gonzales, A. M., Meyer, K. K. T., & Clegg, M. T. (2014). Resequencing data indicate domestication's modest effect on barley diversity: A cultigen with multiple origins. *Journal of Heredity*, *105*, 253–264. <https://doi.org/10.1093/jhered/est083>
- Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., Mayer, K. F., Taudien, S., Platzer, M., Jeddloh, J. A., Springer, N. M., Muehlbauer, G. J., & Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, *14*, R58. <https://doi.org/10.1186/gb-2013-14-6-r58>
- Muñoz-Amatriaín, M., Lo, S., Herniter, I. A., Boukar, O., Fatokun, C., Carvalho, M., Castro, I., Guo, Y.-N., Huynh, B.-L., Roberts, P. A., Carnide, V., & Close, T. J. (2021). The UCR Minicore: A resource for cowpea research and breeding. *Legume Science*, *3*, e95. <https://doi.org/10.1002/leg3.95>
- Muñoz-Amatriaín, M., Mirebrahim, H., Xu, P., Wanamaker, S. I., Luo, M., Alhakami, H., Alpert, M., Atokple, I., Batienco, B. J., Boukar, O., Bozdag, S., Cisse, N., Drabo, I., Ehlers, J. D., Farmer, A., Fatokun, C., Gu, Y. Q., Guo, Y.-N., Huynh, B.-L., ... Close, T. J. (2017). Genome resources for climate-resilient cowpea, an essential crop for food

- security. *The Plant Journal*, 89, 1042–1054. <https://doi.org/10.1111/tbj.13404>
- Onsongo, G., Xie, H., Griffin, T. J., & Carlis, J. (2008). Generating GO slim using relational database management systems to support proteomics analysis. *21st IEEE International Symposium on Computer-Based Medical Systems*. <https://doi.org/10.1109/CBMS.2008.77>
- Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, 34, 867–868. <https://doi.org/10.1093/bioinformatics/btx699>
- Quinlan, A. R., & Hall, I. M. (2010). BEDtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–841. <https://doi.org/10.1093/bioinformatics/btq033>
- Sakai, H., Naito, K., Ogiso-Tanaka, E., Takahashi, Y., Iseki, K., Muto, C., Satou, K., Teruya, K., Shiroma, A., Shimoji, M., Hirano, T., Itoh, T., Kaga, A., & Tomooka, N. (2015). The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Scientific Reports*, 5, 16780. <https://doi.org/10.1038/srep16780>
- Salamov, A. A., & Solovyev, V. V. (2000). *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Research*, 10, 516–522. <https://doi.org/10.1101/gr.10.4.516>
- Schmid, K., Kilian, B., & Russell, J. (2018). Barley domestication, adaptation and population genomics. In N. Stein & G. Muehlbauer (Eds.), *The barley genome. Compendium of plant genomes* (pp. 317–336). Springer.
- Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., ... Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, 46, 707–713. <https://doi.org/10.1038/ng.3008>
- Shu, S., Goodstein, D., & Rokhsar, D. (2013). PERTRAN: Genome-guided RNA-seq read assembler. <https://www.osti.gov/biblio/1241180>
- Shu, S., Rokhsar, D., Goodstein, D., Hayes, D., & Mitros, T. (2014). *JGI plant genomics gene annotation pipeline*. Lawrence Berkeley National Laboratory.
- Singh, B. B., Olufajo, O. O., Ishiyaku, M. F., Adeleke, R. A., Ajeigbe, H. A., & Mohammed, S. G. (2006). Registration of six improved germplasm lines of cowpea with combined resistance to *Striga gesnerioides* and *Alectra vogelii*. *Crop Science*, 46, 2332–2333. <https://doi.org/10.2135/cropsci2006.03.0148>
- Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddleloh, J. A., Nettleton, D., & Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, 5, e1000734. <https://doi.org/10.1371/journal.pgen.1000734>
- Stai, J. S., Yadav, A., Sinou, C., Bruneau, A., Doyle, J. J., Fernández-Baca, D., & Cannon, S. B. (2019). Cercis: A non-polyploid genomic relic within the generally polyploid legume family. *Frontiers in Plant Science*, 10, 345. <https://doi.org/10.3389/fpls.2019.00345>
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105, 437–460. <https://doi.org/10.1093/genetics/105.2.437>
- Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015). ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biology*, 16, 3. <https://doi.org/10.1186/s13059-014-0573-1>
- Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit, Y., Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., ... Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13950–13955. <https://doi.org/10.1073/pnas.0506758102>
- Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. (2017). agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, 45, W122–W129. <https://doi.org/10.1093/nar/gkx382>
- Tittes, S., Lorant, A., McGinty, S., Doebley, J. F., Holland, J. B., de Jesus Sánchez-González, Seetharam, A., Tenaillon, M., & Ross-Ibarra, J. (2021). Not so local: The population genetics of convergent adaptation in maize and teosinte. *BioRxiv*. <https://doi.org/10.1101/2021.09.09.459637>
- Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S., Brown, A. V., Ren, L., Jenkins, J., Chung, C. Y.-L., Chan, T.-F., Daum, C. G., Plott, C., Hastie, A., Baruch, K., Barry, K. W., Huang, W., Patil, G., Varshney, R. K., Hu, H., Batley, J., ... Nguyen, H. T. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant Journal*, 100, 1066–1082. <https://doi.org/10.1111/tbj.14500>
- Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40, e49. <https://doi.org/10.1093/nar/gkr1293>
- Xu, P., Wu, X., Muñoz-Amatriaín, M., Wang, B., Wu, X., Hu, Y., Huynh, B.-L., Close, T. J., Roberts, P. A., Zhou, W., Lu, Z., & Li, G. (2017). Genomic regions, cellular components and gene regulatory basis underlying pod length variations in cowpea (*V. unguiculata* L. Walp). *Plant Biotechnology Journal*, 15, 547–557. <https://doi.org/10.1111/pbi.12639>
- Yeh, R.-F., Lim, L. P., & Burge, C. B. (2001). Computational inference of homologous gene structures in the human genome. *Genome Research*, 11, 803–816. <https://doi.org/10.1101/gr.175701>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Liang, Q., Muñoz-Amatriaín, M., Shu, S., Lo, S., Wu, X., Carlson, J. W., Davidson, P., Goodstein, D. M., Phillips, J., Janis, N. M., Lee, E. J., Liang, C., Morrell, P. L., Farmer, A. D., Xu, P., Close, T. J., & Lonardi, S. (2023). A view of the pan-genome of domesticated cowpea (*Vigna unguiculata* [L.] Walp.). *The Plant Genome*, e20319. <https://doi.org/10.1002/tpg2.20319>