DR MARÍA MUÑOZ-AMATRIAÍN (Orcid ID : 0000-0002-4476-1691)

Article type : Resource

The genome of cowpea (*Vigna unguiculata* [L.] Walp.)

Stefano Lonardi[1*†], María Muñoz-Amatriaín[2*†], Qihua Liang[1], Shengqiang Shu[3], Steve I. Wanamaker[2], Sassoum Lo[2], Jaakko Tanskanen[4,5,6], Alan H. Schulman[4,5,6], Tingting Zhu[7], Ming-Cheng Luo[7], Hind Alhakami[1], Rachid Ounit[1], Abid Md. Hasan[1], Jerome Verdier[8], Philip A. Roberts[9], Jansen R.P. Santos[9,10], Arsenio Ndeve[9], Jaroslav Doležel[11], Jan Vrána[11], Samuel A. Hokin[12], Andrew D. Farmer[12], Steven B. Cannon[13], Timothy J. Close[2]

[1]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA; [2]Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA; [3]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA; [4]Natural Resources Institute Finland (Luke), Helsinki, Finland; [5]Institute of Biotechnology, University of Helsinki, Helsinki, Finland; [6]Viikki Plant Science Centre, University of Helsinki, Helsinki, Finland; [7]Department of Plant Sciences, University of California, Davis, CA 95616, USA; [8]Institut de Recherche en Horticulture et Semences, INRA, Université d'Angers, 49071 Beaucouzé, France; [9]Department of Nematology, University of California, Riverside, CA 92521, USA; [10]Departamento de Fitopatologia, Instituto de Ciências Biológicas, Universidade de Brasília, DF, Brazil; [11]Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic; [12]National Center for Genome Resources, Santa Fe, NM 87505, USA; [13]US Department of Agriculture–Agricultural Research Service, Ames, Iowa, USA.

[†]Both authors contributed equally to this work.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.14349

*Corresponding authors: Stefano Lonardi (stelo@cs.ucr.edu), Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA; María Muñoz-Amatriaín (maria.munoz_amatriain@colostate.edu), Department of Botany & Plant Sciences, University of California, Riverside, CA 92521, USA. Current address: Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO 80523, USA.

Running Head: Cowpea reference genome

Keywords: Chromosomal Inversion, Cowpea, Domestication, Genome Annotation, Genome Evolution, Genome Size, Next-Generation Sequencing, Legumes, *Phaseolus vulgaris*, Repetitive Elements, *Vigna unguiculata*

ABSTRACT

Cowpea (*Vigna unguiculata* [L.] Walp.) is a major crop for worldwide food and nutritional security, especially in sub-Saharan Africa, that is resilient to hot and drought-prone environments. An assembly of the single-haplotype inbred genome of cowpea IT97K-499-35 was developed by exploiting the synergies between single molecule real-time sequencing, optical and genetic mapping, and an assembly reconciliation algorithm. A total of 519 Mb is included in the assembled sequences. Nearly half of the assembled sequence is composed of repetitive elements, which are enriched within recombination-poor pericentromeric regions. A comparative analysis of these elements suggests that genome size differences between *Vigna* species are mainly attributable to changes in the amount of *Gypsy* retrotransposons. Conversely, genes are more abundant in more distal, high-recombination regions of the chromosomes; there appears to be more duplication of genes within the NBS-LRR and the SAUR-like auxin superfamilies compared to other warm-season legumes that

have been sequenced. A surprising outcome is the identification of an inversion of 4.2 Mb among landraces and cultivars, which includes a gene that has been associated in other plants with interactions with the parasitic weed *Striga gesnerioides*. The genome sequence facilitated the identification of a putative syntelog for multiple organ gigantism in legumes. A revised numbering system has been adopted for cowpea chromosomes based on synteny with common bean (*Phaseolus vulgaris*). An estimate of nuclear genome size of 640.6 Mbp based on cytometry is presented.

INTRODUCTION

Cowpea (*Vigna unguiculata* [L.] Walp.) is one of the most important food and nutritional security crops, providing the main source of protein to millions of people in developing countries. In sub-Saharan Africa, smallholder farmers are the major producers and consumers of cowpea, which is grown for its grains, tender leaves and pods as food for human consumption, with the crop residues being used for fodder or added back to the soil to improve fertility (Singh, 2014). Cowpea was domesticated in Africa (D'Andrea et al., 2007, Faris, 1965), from where it spread into all continents and now is commonly grown in many parts of Asia, Europe, the United States, and Central and South America. One of the strengths of cowpea is its high resilience to harsh conditions, including hot and dry environments, and poor soils (Boukar et al., 2018). Still, as sub-Saharan Africa and other cowpea production regions encounter climate variability (Kotir, 2011, Serdeczny et al., 2016), breeding for more climate-resilient varieties remains a priority.

Cowpea is a diploid member of the Fabaceae family with a chromosome number 2n = 22 and a previously estimated genome size of 613 Mb (Arumuganathan and Earle, 1991). Its genome shares a high degree of collinearity with other warm season legumes (Phaseoleae tribe), including common bean (*Phaseolus vulgaris* L.) (Vasconcelos et al., 2015, Muñoz-Amatriaín et al., 2017). A highly

fragmented draft assembly and BAC sequence assemblies of IT97K-499-35 were previously generated (Muñoz-Amatriaín et al., 2017). Although these resources enabled progress on cowpea genetics (Yao et al., 2016, Carvalho et al., 2017, Misra et al., 2017, Huynh et al., 2018, Lo et al., 2018) they lacked the contiguity and completeness required for accurate genome annotation, detailed investigation of candidate genes or thorough genome comparisons. Here, we re-estimated the genome size of *Vigna unguiculata* and produced a genome assembly using single-molecule real-time sequencing combined with optical and genetic mapping. This reference sequence was used to identify repetitive elements, genes and gene families, and genetic variation, and for comparative analysis with three closely related legumes including common bean, which stimulated a change of chromosome numbering to facilitate comparative studies. The publicly available genome sequence lays the foundation for basic and applied research, enabling progress towards the improvement of this key crop plant for food and nutritional security.

RESULTS AND DISCUSSION

Estimation of *Vigna unguiculata* genome size

To assess the genome size of the sequenced accession IT97K-499-35, nuclear DNA content was estimated using flow cytometry (Dolezel, 2003), k-mer analysis and optical mapping (see Methods for more detail). In brief, cytometry indicated that the 2C nuclear DNA amount of *Vigna unguiculata* IT97K-499-35 is 1.310 ± 0.026 pg DNA (mean ± SD), which corresponds to 1C genome size of 640.6 Mbp (see Figure S1). This is slightly higher than the estimate of 613 Mbp by Arumuganathan and Earle (1991) but 841 Mbp smaller than the estimate of Parida et al. (1990). The higher estimate of DNA amount by Parida et al. (1990) could be due to incomplete removal of formaldehyde fixative prior to staining with Schiff's reagent, which binds to free aldehyde groups (Chieco and Derenzini, 1999). The estimate of Arumuganathan and Earle (1991) was obtained using Feulgen microdensitometry, which is considered a reliable method, and perfect agreement has been

observed between flow cytometric and microspectrophotometric estimates (Doležel et al., 1998). The small difference between the genome size estimates of Arumuganathan and Earle (1991) and the present work could be due to different values assigned to reference standards, instrument variation between laboratories (Doležel et al., 1998) or actual differences between accessions.

Also, a k-mer distribution analysis was carried out, providing a somewhat lower estimate of 560.3 Mbp (Figure S2). However, k-mer-based estimates suffer inaccuracies from overcounting low copy k-mers that result from errors introduced by PCR, undercounting k-mers that are repeated within gene families and conserved motifs, and vast undercounting of k-mers from highly repetitive sequences. As noted below, genome size estimates within this range also were obtained from optical mapping. As the cytometry analysis indicates, a genome size of 640.6 Mbp was used.

Sequencing and assembly using stitching

The elite breeding line IT97K-499-35, developed at the International Institute of Tropical Agriculture (IITA, Nigeria), was used previously for the development of genome resources (Timko et al., 2008, Muñoz-Amatriaín et al., 2017). Here, a fully homozygous (single haplotype; See Methods) stock was sequenced using PacBio (Pacific Biosciences of California, Inc., Menlo Park, CA, USA) single-molecule real-time (SMRT) sequencing. In total, 56.8 Gb of sequence data were generated (~91.7x genome equivalent), with a read N50 of 14,595 bp. Pre- and post-filter read length and quality distribution are reported in Figures S3-S6. Two Bionano Genomics (San Diego, California, USA) optical maps (Cao et al., 2014) were generated using nicking enzymes *BspQ*I and *BssS*I (Tables S1 and S2). The size of the *BsqQ*I optical map is 622.21 Mb, while the size of the *BssS*I optical map is 577.76 Mb.

With the PacBio data, eight draft assemblies were generated, six of which were produced with CANU (Berlin et al., 2015, Koren et al., 2017) using multiple parameter settings at the error correction stage, one with Falcon (Chin et al., 2016), and one with ABruijn (Lin et al., 2016). As Table S3 shows, CANU, Falcon, and ABruijn produced assemblies with significantly different assembly statistics, which made it difficult to designate one as "best". These tools are fundamentally different at the algorithmic level (e.g., CANU and Falcon are based on the overlap-layout-consensus paradigm, while ABruijn uses the de Bruijn graph), and their designers have made different choices in the tradeoff between maximizing assembly contiguity versus minimizing mis-joins. Here, we employed an alternative assembly methodology: instead of choosing one assembly, the optical maps were leveraged to merge multiple assemblies in what we call "stitching" (see Pan et al. (2018) and Methods). This method was applied to the eight assemblies in Table S3, after removing contaminated contigs and breaking chimeric contigs identified using the optical maps. The number of chimeric contigs ranged from 16 to 40 depending on the assembly. Each of the eight assemblies contributed a fraction of its contigs to the final assembly: 13% of the "minimal tiling path" (MTP) contigs were from the FALCON assembly, 8% from the ABruijn assembly and the rest (79%) from the six CANU assemblies, each ranging from 4% to 20%. Table 1 reports statistics of the stitched and polished (PacBio Quiver pipeline) assembly. PacBio Quiver enables consensus accuracies on genome assemblies approaching or exceeding Q60 (one error per million bases) when the sequencing depth is above 60X (Chin et al., 2013). All of the assembly statistics significantly improved compared to the eight individual assemblies (Table S3). For instance, the N50 for the stitched assembly (10.9 Mb) was almost double the highest N50 for any of the eight individual assemblies. Similarly, the longest contig for the stitched assembly increased by 4 Mb over the longest contig of any single assembly.

Scaffolds were obtained by mapping the stitched and polished assembly to both optical maps using the Kansas State University pipeline (Shelton et al., 2015). Briefly, a total of 519.4 Mb of sequence scaffold were generated with an N50 of 16.4 Mb (Table 1). Finally, a total of ten genetic maps containing 44,003 unique Illumina iSelect SNPs (Muñoz-Amatriaín et al., 2017) were used to

anchor and orient sequence scaffolds into eleven pseudochromosomes (i.e. pseudomolecules) via ALLMAPS (Tang et al., 2015). Details of the ten genetic maps can be found in Table S4. ALLMAPS was able to anchor 47 of the 74 scaffolds for a total of 473.4 Mb (91.1% of the assembled sequences), 30 of which were also oriented, resulting in 449 Mb of anchored and oriented sequence (Table 1). Only 46 Mb (8.9% of the total assembly) were unplaced. The average GC content of the assembly was 32.99%, similar to other sequenced legumes (Varshney et al., 2012, Schmutz et al., 2014, Yang et al., 2015). The quality of the chromosome-level assembly was evaluated using a variety of metrics. Several sequence datasets that were independently generated were mapped onto the assembly using BWA-mem with default settings, namely: (1) about 168M 149-bp paired-end Illumina reads (98.92% mapped of which 86.7% were properly paired and 75.53% had MAPQ of at least 30), (2) about 129 thousand contigs (500 bp or longer) of the WGS assembly generated previously (Muñoz-Amatriaín et al., 2017; 99.69% mapped of which 98.69% had MAPQ>30), (3) about 178 thousand BAC sequence assemblies generated previously (Muñoz-Amatriaín et al., 2017; 99.95% mapped of which 68.39% had MAPQ>30), and (4) about 157 thousand transcripts (Santos et al., 2018; 99.95% mapped of which 94.74% had MAPQ>30). All of these metrics indicate agreement with the pseudochromosomes. The original PacBio reads were also mapped onto the assembly using BLASR using default settings: 5.29 M long reads mapped for a total of about 46 x 10^9 bp. 88.68% of the bases of the long reads were present in the 519 Mbp assembly.

Revised chromosome numbering for cowpea

Several members of the Phaseoleae tribe are diploid with 2n = 22, but the numbering of chromosomes has been designated independently within and across species by each research group. The *P. vulgaris* genome sequence was the earliest among these species (Schmutz et al., 2014), thus establishing a precedent and rational basis for a more uniform chromosome numbering system. Extensive synteny has been previously observed between cowpea and common bean (Muñoz-Amatriaín et al., 2017), which facilitates a revised chromosome numbering system for

cowpea based on synteny with common bean. As summarized in Figure S7 and Table S5, six cowpea chromosomes are largely syntenic with six common bean chromosomes in one-to-one relationships, making the numbering conversion straightforward in those cases. Each of the remaining five cowpea chromosomes is related to parts of two *P. vulgaris* chromosomes. For each of those cases, the number of the common bean chromosome sharing the largest syntenic region with cowpea was adopted, with one exception: two cowpea chromosomes (previous linkage groups/chromosomes #1 and #5) both shared their largest block of synteny with *P. vulgaris* chromosome Pv08. However, there was only one optimum solution to the chromosome numbering of cowpea, assigning Vu08 to previous cowpea linkage group/chromosome #5 and assigning Vu05 to previous linkage group/chromosome #1 (Table S5). In addition, comparisons between cowpea genetic maps and chromosomal maps developed by fluorescence in situ hybridization (FISH) using cowpea BACs as probes (Iwata-Otsubo et al., 2016) revealed that the prior orientations of three linkage groups (now referred to as Vu06, Vu10, and Vu11) were inverted relative to their actual chromosome orientation. Hence, cowpea pseudochromosomes and all genetic maps were inverted for chromosomes Vu06, Vu10, and Vu11 to meet the convention of short arm on top and long arm on the bottom, corresponding to ascending cM values from the distal (telomeric) end of the short arm through the centromere and on to the distal end of the long arm. It is also of some interest that both Vu06 and Pv06 are acrocentric chromosomes, but although Pv09 is acrocentric the ratio of short to long arm in Vu09 (formerly cowpea linkage group 8) is 25.86 to 46.35 µm (Iwata-Otsubo et al., 2016). Clearly there are many structural similarities, but also some differences between common bean and cowpea chromosomes.

The revised numbering system is shown in Table S5 and used throughout the present manuscript. The Windows software HarvEST:Cowpea (harvest.ucr.edu), which includes a synteny display function, also has adopted updated numbering system.

Gene annotation and repetitive DNA

The assembled genome was annotated using *de novo* gene prediction and transcript evidence based on cowpea ESTs (Muchero et al., 2009) and RNA-seq data from leaf, stem, root, flower and seed tissue (Yao et al., 2016, Santos et al., 2018), and protein sequences of Arabidopsis, common bean, soybean, Medicago, poplar, rice and grape (see Methods). In total, 29,773 protein-coding loci were annotated, along with 12,514 alternatively spliced transcripts. Most (95.9%) of the 1,440 expected plant genes in BUSCO v3 (Simão et al., 2015) were identified in the cowpea gene set, indicating completeness of genome assembly and annotation. The average gene length was 3,881 bp, the average exon length was 313 bp, and there were 6.29 exons per gene on average. The GC content in coding exons was higher than in introns plus UTRs (40.82% vs. 24.27%, respectively). Intergenic regions had an average GC content of 31.84%.

Based on the results of an automated repeat annotation pipeline (Table S6), an estimated 49.5% of the cowpea genome is composed of the following repetitive elements: 39.2% transposable elements (TEs), 4% simple sequence repeats (SSRs), and 5.7% unidentified low-complexity sequences. The retrotransposons, or Class I TEs, comprise 84.6% of the transposable elements by sequence coverage and 82.3% by number. Of the long terminal repeat (LTR) retrotransposons, elements of the *Gypsy* superfamily (Wicker et al., 2007; code RLG) are 1.5 times more abundant than *Copia* (code RLC) elements, but non-autonomous TRIM elements appear to be very rare, with only 57 found. The LINEs (RIX) and SINEs (RSX), comprising the non-LTR retrotransposons, together amount to only 0.4% of the genome. The DNA, or Class II, transposons compose 6.1% of the genome, with the CACTA (DTC; 5.7% of the transposable element sequences), hAT (DTA; 3.5%), and MuDR (DTM; 2.4%) being the major groups of classical “cut-and-paste” transposons. The rolling-circle *Helitron* (DHH) superfamily is relatively abundant at 1.3% of the genome and 7013 individual elements. Only 6.4% of the TE sequences were unclassified.

Centromeric regions were defined based on a 455-bp tandem repeat that was previously identified by FISH as abundant in cowpea centromeres (Iwata-Otsubo et al., 2016). Regions containing this sequence span over 20.18 Mb (3.9% of the assembled genome; Table S7). Cowpea centromeric and pericentromeric regions are highly repetitive in sequence composition and exhibit low gene density and low recombination rates, while both gene density and recombination rate increase as the physical position becomes more distal from the centromeres (Figure 1; Figure S87; Data S1). Contrasting examples include Vu04, where the recombination rate near the telomeres of both arms of this metacentric chromosome are roughly ten times the rate across the pericentromeric region, versus Vu02 and Vu06, where the entire short arm in each of these acrocentric chromosomes has a low recombination rate (Figure S8). These patterns have been observed in other plant genomes including legumes (Schmutz et al., 2010, Schmutz et al., 2014) and have important implications for genetic studies and plant breeding. For example, a major gene for a trait that lies within a low recombination region can be expected to have high linkage drag when introgressed into a different background. Knowledge of the recombination rate can be integrated into decisions on marker density and provide weight factors in genomic selection models to favor rare recombination events within low recombination regions.

Cowpea genetic diversity

Single-nucleotide and insertion/deletion variation

Whole-genome shotgun (WGS) data from an additional 36 diverse accessions relevant to Africa, China and the USA were previously used to identify single nucleotide polymorphisms (Muñoz-Amatriaín et al., 2017). Almost all (99.83%) of the 957,710 discovered SNPs (hereinafter referred as the "1M list") were positioned in the reference sequence, including 49,697 SNPs that can be assayed using the Illumina iSelect Consortium Array (Muñoz-Amatriaín et al., 2017) (Data S2). About 35% of the SNPs in the 1M list were associated with genes (336,285 SNPs), while that

percentage increased to 62% in the iSelect array (31,708 SNPs; Data S2 and Table S8). This indicates that the intended bias towards genes in the iSelect array design (Muñoz-Amatriaín et al., 2017) was successful. The number of annotated cowpea gene models containing a SNP was 23,266 (78% of total), or 27,021 (91% of total) when considering genes within 10 kb of a SNP (Table S8). In general, SNP density was lowest near centromeric regions (Figure 1, Figure S9). This information enables formula-based selection of SNPs, including distance to gene and recombination rate. When these metrics are combined with minor allele frequency and nearness to a trait determinant, one can choose an optimal set of SNPs for a given constraint, for example cost minimization, on the number of markers.

The same WGS data described above were analyzed using BreakDancer v.1.4.5 (Chen et al., 2009) to identify structural variants. A total of 17,401 putative insertions and 117,403 putative deletions relative to the reference genome were identified (Data S3). The much smaller number of insertions than deletions may reflect limitations in the ability of the software to identify insertions when sequence reads are mapped to a reference genome. The presently available data from one reference-quality genome sequence and WGS short reads from 36 accessions are insufficient to create a comprehensive and reliable catalog of structural variants; additional high-quality *de novo* assemblies will be required to accomplish those goals.

Identification of a 4.2 Mb chromosomal inversion on Vu03

As explained above, ten genetic maps were used to anchor and orient scaffolds into pseudochromosomes. Plots of genetic against physical positions for SNPs on seven of those genetic maps showed a relatively large region in an inverted orientation (Figure 2A; Figure S10). The other three genetic maps showed no recombination in this same region, suggesting that the two parents in the cross had opposite orientations. The genotype data from all of the parental lines showed that one of the parents from each of those three populations, but not the other parent, had the same haplotype as IT97K-499-35, and hence presumably the same orientation (Data S4). To define the

inversion breakpoints, WGS data available from some of these accessions (Muñoz-Amatriaín et al., 2017) were used. In both breakpoint regions, contigs from accessions that presumably had the same orientation as the reference (type A) showed good alignments, while those from accessions with the opposite orientation (type B) aligned only until the breakpoints (Data S5). An additional *de novo* assembly of a 'type B' accession enabled a sequence comparison with the reference genome for the entire genomic region containing the inversion (Figure 2B). This provided a confirmation of the chromosomal inversion and the position of the two breakpoints in the reference sequence: 36,118,991 bp (breakpoint 1) and 40,333,678 bp (breakpoint 2) for a 4.21 Mb inversion containing 242 genes (Data S6). PCR amplifications of both breakpoint regions further validated this inversion (see Methods and Figure S11).

A set of 368 diverse cowpea accessions, including 243 landraces and 97 breeding accessions for which iSelect data existed, was used to estimate the frequency of the inversion among germplasm accessions. A total of 33 accessions (9%) had the same SNP haplotype as the reference genome across the entire region, which presumably indicates the same orientation. Among those 33 accessions, only three were landraces (1.2% of the landraces in the set), while the other 30 were breeding materials, including the reference genome. This suggests that the reference genome orientation of this region is rare among landraces and that its frequency has been increased among breeding lines. Also, a complete lack of recombination across this region is reflected in the genetic map derived from a cultivated x wild cross (Lo et al., 2018; IT99K-573-1-1 x TVNu-1158; Figure S10), which indicates that the wild parent has the opposite orientation of the cultivated accession. Since this cultivated parent has the same haplotype as the reference genome, and thus presumably also the same orientation, the lack of recombination across this region suggests that the opposite-to-reference orientation is the ancestral (wild) type while the reference orientation carries an inversion. A comparison between cowpea and adzuki bean (Figure S12) showed that IT97K-499-35 and adzuki

bean genome assemblies have opposite orientations in this region, consistent with the conjecture that the cowpea reference genome is inverted in this region with respect to an ancestral state that has been retained in the wild cowpea accession as well as in this representative congeneric species.

A direct effect of inversions is that they suppress recombination in heterozygotes, causing inverted regions to evolve independently. Selection can act to maintain an inversion when it carries one or more advantageous alleles or when an inversion breakpoint causes gene disruption or expression changes that are adaptive (Kirkpatrick, 2010, Puig et al., 2015). Two of the three landraces carrying the inversion (B-301 and B-171) originated from Botswana while the third (TVu-53) is a Nigerian landrace. B-301 was the donor of resistance to several races of *Striga gesnerioides*, a serious parasitic weed of cowpea, and is in the pedigree of many breeding lines that carry the inversion, most of which are also *Striga* resistant (including the reference genome IT97K-499-35). To explore whether the inversion is associated with *Striga* resistance, the map positions of previously identified QTLs for this trait (Ouédraogo et al., 2001, Ouédraogo et al., 2002, Boukar et al., 2004) were compared to the position of the inversion. QTLs for resistance to *Striga* Races 1 and 3 were located on a different chromosome/linkage group than the inversion on Vu03, ruling out the inversion as the basis of those resistances. However, it was noted that the sorghum gene *Sobic.005G213600* regulating *Striga* resistance via a presence/absence variation (Gobena et al., 2017) encodes a sulfotransferase that is homologous to the cowpea gene *Vigun03g220400*, which is located inside the inverted region on Vu03 (Data S6) and is highly expressed in root tissue (https://legumeinfo.org/feature/Vigna/unguiculata/gene/vigun.IT97K-499-35.gnm1.ann1.Vigun03g220400#pane=geneexpressionprofile). Therefore, it seems possible that the region containing *Vigun03g220400* may affect *Striga* interactions in a manner that has not yet been discovered; this hypothesis merits further testing. In addition to *Striga* considerations, a QTL for pod number (Xu et al., 2013; *Qpn.zaas-3*) is located inside the inverted region.

Although additional studies will be required to determine whether there is an adaptive consequence of the Vu03 inversion, awareness of it certainly is important for trait introgression and breeding, as this region represents nearly 1% of the cowpea genome and can be moderately active recombinationally during meiosis only when both chromatids carry the same orientation.

Synteny with other warm season legumes

Synteny analyses were performed between cowpea and its close relatives adzuki bean (*Vigna angularis*), mung bean (*Vigna radiata*) and common bean (*Phaseolus vulgaris*). Extensive synteny was observed between cowpea and the other three diploid warm-season legumes although, as expected, a higher conservation was observed with the two *Vigna* species (Figure 3A-C) than with common bean. Six cowpea chromosomes (Vu04, Vu06, Vu07, Vu09, Vu10 and Vu11) largely have synteny with single chromosomes in all three other species. Cowpea chromosomes Vu02, Vu03 and Vu08 also have one-to-one relationships with the other two *Vigna* species but one-to-two relationships with *P. vulgaris*, suggesting that these chromosome rearrangements are characteristic of the divergence of *Vigna* from *Phaseolus*. The remaining cowpea chromosomes Vu01 and Vu05 have variable synteny relationships, each with two chromosomes in each of the other three species, suggesting these chromosome rearrangements are more characteristics of speciation within the *Vigna* genus. It should be noted also that most chromosomes that have a one-to-two relationship across these species or genera are consistent with translocations involving the centromeric regions (Figure 3A-C). On the basis of these synteny relationships, adoption of the revised cowpea chromosome numbering for adzuki bean, mung bean and presumably other *Vigna* species would be straightforward. This would facilitate reciprocal exchange of genomic information on target traits from one *Vigna* species to another.

Repetitive elements and genome expansion

Using the same computational pipeline as for *V. unguiculata* (Vu), the repeats of the *V. angularis* (Yang et al., 2015; Va) and *V. radiata* (Kang et al., 2014; Vr) genomes also were annotated. Previous analyses placed cowpea phylogenetically closer to mung bean (Vr) than to adzuki bean (Va) (She et al., 2015), although the Va and Vr genomes are relatively similar in size, with cowpea respectively 11% and 12% larger. The annotated repeat spaces in the three genomes were examined to make inferences on their evolution. Comparing Vu with Vr, 94% of the 56 Mbp size difference can be explained by the differential abundance of TEs, and 57% by the differential abundance of superfamily *Gypsy* retrotransposons alone (Table S9). The differential abundance of *Gypsy* elements in cowpea amounts to 58% and 56% of the total contribution of TEs to its genome size difference with mung bean and adzuki bean, respectively. The non-LTR retrotransposons, composed of SINEs and LINEs, appear to have played only a minor role in genome size enlargement in cowpea. Helitrons contributed 10% (vs. Vr) or 11% (vs. Va) to the expansion of the cowpea genome, and increased in genome share by an order of magnitude. The DNA TEs together contributed 38% of the size difference between Vu and Vr and 40% between Vu and Va. CACTA contributed about the same amount (Va), or 35% more (vs. Vr) of DNA as hAT elements, to this growth. For both Vr and Va, far fewer unidentified LTR retrotransposons (RLX) were found than in the Vu genome, perhaps because the Vu genome appears to be less fragmented and more complete than the former two. Expansion of SSR content was very moderate in Vu vs Vr, and comprised a smaller genome share than in Va.

A similar comparison was made to the 473 Mb genome assembly of *Phaseolus vulgaris* (Schmutz et al., 2014; Pv) with a genome estimated to be only 9% smaller (587 Mbp; http://data.kew.org/cvalues). However, Pv has a higher TE content than cowpea, 45.2% vs. 39%, of which 39% vs. 33% are retrotransposons. In Pv the *Gypsy* elements comprise 25% of the genome vs. 18% in *V. unguiculata*, although the *Copia* elements are 2% less abundant than in cowpea. There are 23.5 Mb more *Gypsy* elements annotated in the *P. vulgaris* assembly than in Vu, although the total

TE coverage is only 10.8 Mb greater in Pv than in cowpea. While the assemblies represent similar shares of the estimated genomes (Vu, 81.1%; Pv, 80.5%), the contig N50 for *P. vulgaris* is 0.395 Mb vs. 10.9 Mb for Vu. These data may indicate that the true *P. vulgaris* genome is considerably larger than estimated by Feulgen densitometry, with the large fraction of TEs interfering with contig assembly.

Taken together, the cross-species comparisons suggest that differences in genome size in *Vigna* can be largely explained by TE abundance, especially by that of *Gypsy* retrotransposons. This can result from either differential amplification recently, or differential retention of ancient insertions. In the grasses, comparison, e.g., of the *Brachypodium distachyon* (Initiative, 2010) and *Hordeum vulgare* (Mascher et al., 2017) genomes suggests that differences in *Gypsy* content are largely due to differential retention. However, among the legumes examined here, annotated full-length retrotransposons appear to be of recent origin (less than 0.5 million years) in *P. vulgaris* (Schmutz et al., 2014).

Gene family changes in cowpea

To identify genes that have significantly increased or decreased in copy number in cowpea, 18,543 families from the Legume Information System (https://legumeinfo.org/search/phylotree and https://legumeinfo.org/data/public/Gene_families/) were analyzed. This set was constructed to capture genes originating at the legume taxonomic depth, based on orthology relationships and per-species synonymous-site rates for legume species and outgroup species. These families include 14 legume species, six of which are from the Phaseoleae tribe (soybean, common bean, adzuki bean, mung bean, pigeon pea, and cowpea). Among the 185 gene families in the top percentile in terms of cowpea gene membership in the family relative to average membership per legume species, the families include several in the following superfamily groups: NBS-LRR disease resistance genes, various receptor-like protein kinases, defensins, ribosomal proteins, NADH-quinone oxidoreductase components (Data S7). All of these families occur in large genomic arrays, which can expand or

contract, likely through slipped-strand mispairing of paralogous genes (Levinson and Gutman, 1987, Cannon et al., 2004, Li et al., 2016).

Gene families lacking cowpea membership are more difficult to interpret biologically, as these tend to be smaller gene families, likely showing stochastic effects of small families "falling out of" larger superfamilies, due to extinction of clusters of genes or to artifactual effects of family construction. Among 18,543 legume gene families, there were 2,520 families without cowpea gene membership, which is comparable to the average number of families without membership (3,057) for six other sequenced genomes in the Phaseoleae. The 2,520 "no-cowpea" families were enriched for the following superfamilies: UDP-glycosyltransferases, subtilisin-like serine proteases, several kinase superfamilies, several probable retrotransposon-related families, FAR1-related proteins, and NBS-LRR disease resistance families (Data S7). These superfamilies are generally organized in large genomic clusters that are subject to expansion and contraction (Cannon et al., 2004, Leister, 2004, Li et al., 2016). Several families in cowpea are notable for copy-number differences relative to other sequenced species in *Vigna* (adzuki bean and mung bean). The SAUR-like auxin superfamily contains 138 annotated genes in cowpea, vs. 90 and 52 in adzuki and mung bean, respectively. The NBS-LRR superfamily contains 402 annotated genes, vs. 272 and 86 in adzuki and mung bean, respectively (Data S7). In both superfamilies, adzuki and mung bean may have lost gene copies, rather than cowpea gaining genes, or their assemblies underrepresent them due to technological difficulties with short read assemblies capturing such clusters. The cowpea gene counts are more typical of the other annotated Phaseoleae species: 252 and 130 SAUR genes in *Phaseolus* and *Cajanus*, respectively, and 341 and 271 NBS-LRR genes in *Phaseolus* and *Cajanus*, respectively (Data S7). Of course, these comparisons are subject to revision as the respective genome sequences become more complete.

Identification of a candidate gene for multiple organ gigantism

Crop domestication typically involved size increases of specific organs harvested by humans (Doebley et al., 2006). Recently, a genomic region related to increased organ size in cowpea was identified on Vu08 using a RIL population derived from a domesticated x wild cross (Lo et al., 2018). This region contains a cluster of QTLs for pod length, seed size, leaf length, and leaf width (*CPodl8*, *CSw8*, *CLl8*, *CLw8*). The reference genome sequence described here was used to further investigate this domestication hotspot, which spans 2.21 Mb and includes 313 genes. Syntenic regions in the common bean genome were identified, the largest of which is located on common bean chromosome 8 (Pv08). That region contains a total of 289 common bean syntelogs, which were then compared with the list of common bean genes associated with domestication available from Schmutz et al. (Schmutz et al., 2014). The intersection of these two lists contained only a single gene, *Phvul.008G285800*, a *P. vulgaris* candidate gene for increased seed size that corresponds to cowpea *Vigun08g217000*. This gene codes for a histidine kinase 2 that is expressed in several cowpea tissues including root, seed, pod and leaf (https://legumeinfo.org). The Arabidopsis ortholog *AHK2* (*AT5G35750.1*) is a cytokinin receptor that has been shown to regulate, among other things, plant organ size (, Riefler et al., 2006; Bartrina et al., 2017). *Vigun08g217000* is thus a candidate gene for further investigation.

METHODS

Estimation of genome size

Flow cytometric estimation of genome size followed the protocol of Doležel et al. (2007). Briefly, suspensions of cell nuclei were prepared from 50 mg of young leaf tissue of cowpea IT97K-499-35, and of *S. lycopersicum* cv. Stupické polní rané as an internal standard. The tissues were chopped using a razor blade in 0.5 ml Otto I solution in a glass Petri dish. The homogenate was

filtered through a 50 µm nylon mesh to remove debris and kept on ice. Then, 1 ml Otto II solution containing 50 µg ml^{-1} propidium iodide and 50 µg ml^{-1} RNase was added and the sample was analyzed by a CyFlow Space flow cytometer (Sysmex Partec, Görlitz, Germany). The threshold on the PI detector was set to channel 40 and no other gating strategy was applied. Five thousand events were acquired in each measurement. The resulting histograms of relative DNA content (Figure S1) comprised two major peaks representing nuclei in the G1 phase of the cell cycle. The ratio of G1 peak positions was used to calculate the amount of DNA of cowpea. Five different plants of IT97K-499-35 were analyzed, each three times on three different days and the mean 2C DNA amount was calculated. Genome size was determined using the conversion factor 1 pg = 0.978 Mbp (Dolezel, 2003).

To estimate the cowpea IT97K-499-35 genome size using k-mer distribution, 168 M 149 bp paired-end Illumina reads were processed for a total of about 50 billion bp. Figure S2 shows the frequency distribution of 27-mers produced with KAT (https://github.com/TGAC/KAT). The x-axis represents the 27-mer multiplicity, the y-axis represents the number of 27-mers with that multiplicity. The peak of the distribution is 56, which represents the effective coverage. The total number of uniqe 27-mers in the range x=2-10000 is 31.381 x 10^9. As is usually done, 27-mers that appear only once are excluded because they are considered erronous, i.e., to contain sequencing errors. The estimated genome size based on the formula bp = (# of unique 27-mers–k+1)/peak depth of coverage the is thus 31.381 x 10^9/56 = 560,379,733 bp.

Bionano Genomics optical maps

High molecular weight DNA was isolated by Amplicon Express (Pullman, WA) from nuclei purified from young etiolated leaves (grown in the dark) of 100% homozygous, pure seeds of cowpea IT97K-499-35. The material was screened for homozygosity by genotyping with the Cowpea iSelect Consortium Array (Muñoz-Amatriaín et al., 2017; Data S8). The nicking endonucleases

Nt.*BspQ*I and Nb.*BssS*I (New England BioLabs, Ipswich, MA) were chosen to label DNA molecules at specific sequence motifs. The nicked DNA molecules were stained according to instructions of the IrysPrep Reagent Kit (Bionano Genomics, San Diego, CA) as per Luo et al. (2017). The DNA sample was loaded onto the nano-channel array of an IrysChip (Bionano Genomics, San Diego, CA) and then imaged using the Irys system (Bionano Genomics, San Diego, CA). For the *BspQ*I map, seven separate runs (132 unique scans) were generated, and a total of 108 Gb (~170x genome equivalent) of raw DNA molecules (>100 kb) were collected. Molecules of at least 180 kb in length were selected to generate a BNG map assembly. Table S1 shows the summary of raw molecule status and the BNG *BspQ*I map assembly. For the *BssS*I map, five separate runs (123 unique scans) were generated, and a total of 186 Gb (~310x genome equivalent) of DNA raw molecules (> 20 kb; 133 Gb molecules > 100 kb) were collected. Molecules of at least 180 kb in length were selected to generate a BNG map assembly. Table S2 shows the summary of raw molecules status and the BNG *BssS*I map assembly.

Whole-genome shotgun sequencing and assembly

High molecular weight gDNA and library preparation

Pure seeds of the fully inbred cowpea accession IT97K-499-35 were sterilized and germinated in the dark in crystallization dishes with filter paper and a solution containing antibacterial (cefotaxime, 50 µg/ml) and antifungal (nystatin, 100 units/ml) agents. About 70 g of seedling tissue was collected, frozen in liquid nitrogen, stored at -80C and shipped on dry ice. High molecular weight gDNA was prepared from nuclei isolated from the seedling tissue by Amplicon Express (Pullman, WA).

Pacific Biosciences sequencing

Pacific Biosciences reads were generated at Washington State University (Pullman, WA) following the "Procedure and Checklist-20 kb Template Preparation Using BluePippin Size Selection System" (P/N 100-286-000-5) protocol provided by Pacific Biosciences (Menlo Park, CA) and the Pacific Biosciences SMRTbell Template Prep kit 1.0 (P/N 100-259-100). Resulting SMRTbell libraries were size selected using the BluePippin (Sage Biosciences) according the Blue Pippin User Manual and Quick Guide. The cutoff limit was set to 15-50 kb to select SMRTbell library molecules with an average size of 20 kb or larger. The Pacific Biosciences Binding and Annealing calculator determined the appropriate concentrations for the annealing and binding of the SMRTbell libraries. SMRTbell libraries were annealed and bound to the P6 DNA polymerase for sequencing using the DNA/Polymerase Binding Kit P6 v2.0 (P/N100-372-700). The only deviation from standard protocol was to increase the binding time to 1-3 hours, compared to the suggested 30 minutes. Bound SMRTbell libraries were loaded onto the SMRT cells using the standard MagBead protocol, and the MagBead Buffer Kit v2.0 (P/N 100—642-800). The standard MagBead sequencing protocol followed the DNA Sequencing Kit 4.0 v2 (P/N 100-612-400) which is known as P6/C4 chemistry. PacBio RS II sequencing data were collected in six-hour movies and Stage Start was enabled to capture the longest sub-reads possible.

Sequence quality control

First, CLARK and CLARK-*S* (Ounit and Lonardi, 2016) were used to identify possible contamination from unknown organisms. CLARK and CLARK-*S* are classification tools that use discriminative (spaced, in the case CLARK-S) *k*-mers to quickly determine the most likely origin of each input sequence (*k=21* and *k=31*). The target database for CLARK/CLARK-*S* was comprised of: (i) a representative sample of ~5,000 bacterial/viral genomes from NCBI RefSeq, (ii) human genome,

Homo sapiens, assembly GRCh38, (iii) Illumina-based cowpea draft genome, *Vigna unguiculata* (Muñoz-Amatriaín et al., 2017), assembly v0.03), (iv) soybean, *Glycine max* (Schmutz et al., 2010), assembly Gmax_275_v2.0, (v) common bean, *Phaseolus vulgaris* (Schmutz et al., 2014), assembly Pvulgaris_218_v1.0, (vi) adzuki bean, *Vigna angularis* (Yang et al., 2015), assembly adzuki.ver3.ref.fa.cor, (vii) mung bean, *Vigna radiata* (Kang et al., 2014), assembly Vradi.ver6.cor, and (viii) a nematode that attacks the roots of cowpea, *Meloidogyne incognita* (Abad et al., 2008), assembly GCA_900182535.1_Meloidogyne_incognita_V3.

Whole genome assemblies

Eight draft assemblies were generated, six of which were produced with CANU v1.3 (Berlin et al., 2015, Koren et al., 2017), one with Falcon v0.7.3 (Chin et al., 2016), and one with Abruijn v0.4 (Lin et al., 2016). Hinge v0.41(Kamath et al., 2017) was also tested on this dataset, but at that time the tool required the entire alignment file (over 2Tb) to fit in primary memory and we did not have the computational resources to handle it. CANU v1.3 was run with different settings for the error correction stage on the entire dataset of ~6M reads (two CANU runs were optimized for highly repetitive genomes). Falcon and Abruijn were run on 3.54 M error-corrected reads produced by CANU (30.62Gbp, or 49.4x genome equivalent). Each assembly took about 4-15 days on a 512-core Torque/PBS server hosted at UC Riverside.

Removal of contaminants from the assemblies

To remove "contaminated" contigs, two sets of reference genomes were created, termed the *white* list and the *black* list. Black-list genomes included possible contaminants, whereas white-listed genomes included organisms evolutionarily close to cowpea. The black list included: (i) *Caulobacter segnis* (NCBI accession GCF 000092285.1), (ii) *Rhizobium vignae* (NCBI accession GCF

000732195.1), (iii) *Mesorhizobium sp. NBIMC P2-C3* (NCBI accession GCF 000568555.1), (iv) *Streptomyces purpurogeneiscleroticus* (NCBI accession GCF 001280155.1), (v) *Caulobacter vibrioides* (NCBI accession GCF 001449105.1), (vi) mitochondrion of *Vigna radiata* (Alverson et al., 2011; NCBI accession NC_015121.1), (vii) mitochondrion of *Vigna angularis* (NCBI accession NC_021092.1), (viii) chloroplast of *Vigna unguiculata* (NCBI accession NC_018051.1 and KJ468104.1), and (ix) human genome (assembly GRCh38). The white list included the genomes of: (i) soybean (*Glycine max*, Schmutz et al., 2010) assembly Gmax_275_v2.0), (ii) common bean (*Phaseolus vulgaris*, Schmutz et al., 2014) assembly Pvulgaris_218_v1.0), (iii) adzuki bean (*Vigna angularis*, Yang et al., 2015; assembly adzuki.ver3.ref.fa.cor), (iv) mung bean (*Vigna radiata*, Kang et al., 2014; assembly Vradi.ver6.cor), and (v) Illumina-based cowpea draft genome (*Vigna unguiculata*, Muñoz-Amatriaín et al., 2017, assembly v.0.03). Each assembled contig was BLASTed against the "white" genome and the "black" genomes, and all high-quality alignments (e-score less than $1e^{-47}$ corresponding to a bit score of at least 200, and covering at least 10% of the read length) were recorded. The percentage of each contig covered by white and black high-quality alignments was computed by marking each alignment with the corresponding identity score from the output of BLAST. When multiple alignments covered the same location in a contig, only the best identity alignment was considered. The sum of all these identity scores was computed for each contig, both for the black and the white list. These two scores can be interpreted as the weighted coverage of a contig by statistically significant alignments from the respective set of genomes. A contig was considered contaminated when the black score was at least twice as high as the white score. Chimeric contigs were identified by mapping them against the optical maps using RefAligner (Bionano Genomics), then determining at what loci to break chimeric contigs by visually inspecting the alignments using IrysView (Bionano Genomics).

Stitching of contaminant-free assemblies and polishing

This stitching method (i) uses optical map(s) to determine small subsets of assembled contigs from the individual assemblies that are mutually overlapping with high confidence, (ii) computes a minimum tiling path (MTP) of contigs using the coordinates of the contigs relative to the optical map, and (iii) attempts to stitch overlapping contigs in the MTP based on the coordinates of the contigs relative to the optical map. A series of checks are carried out before and after the stitching to minimize the possibility of creating mis-joins. Additional details about the stitching method can be found in Pan et al. (2018). The final stitched assembly was then polished via the PacBio Quiver pipeline (RS_resequencing.1 protocol) in SMRT Portal v2.3.0 (Patch 5) by mapping all the PacBio subreads against the assembly. The polishing step took about seven days on a 40-core server at UC Riverside.

Scaffolding via optical maps

Scaffolds were obtained from the polished assembly via the Kansas State University (KSU) stitching pipeline (Shelton et al., 2015) in multiple rounds. A tool called XMView (https://github.com/ucrbioinfo/XMView) developed in-house that enables the visual inspection of alignments of assembled contigs to two optical maps simultaneously, also displaying consensus genetic map coordinates for SNPs, was used to identify chimeric optical molecules that had to be excluded from the scaffolding step. The KSU stitching pipeline was iterated four times, alternating *BspQ*I and *BssS*I (twice each map) at which point no conflicts remained.

Pseudochromosome construction via anchoring to genetic maps

Pseudochromosomes were obtained by anchoring the scaffold sequences to SNP markers (BLAST of SNP design sequences, e^{-50} or less) in ten genetic maps (Table S4). Seven of these genetic maps were previously published, five of which are from Muñoz-Amatriaín et al. (2017), and one each from Santos et al. (2018) and Lo et al. (2018). The remaining three genetic maps were generated as part of this study after genotyping three additional RIL (Recombinant Inbred Line) populations with the Cowpea iSelect Consortium Array (Muñoz-Amatriaín et al., 2017). SNP calling and curation were done as described by Muñoz-Amatriaín et al. (2017) and linkage mapping was performed using MSTmap (Wu et al., 2008). Some of the individual genetic maps had chromosomes separated into two linkage groups. In those cases the cowpea consensus genetic map of Muñoz-Amatriaín et al. (2017) was used to join them by estimating the size of the gap (in cM). The final ordering and orientation of the scaffold was produced by ALLMAPS (Tang et al., 2015) from the SNP locations corresponding to the ten genetic maps. As noted elsewhere, 46 Mb of assembled sequences were not anchored. In addition, 24.5 Mb of the anchored sequences were oriented arbitrarily.

Annotation method and estimation of centromere positions

Transcript assemblies were made from ~1.5 B pairs of 2 x100 paired-end Illumina RNAseq reads (Yao et al., 2016, Santos et al., 2018) using PERTRAN (Shu, personal communication). 89,300 transcript assemblies were constructed using PASA (Haas et al., 2003) from EST-derived UNIGENE sequences (Muchero et al., 2009; P12_UNIGENES.fa; harvest.ucr.edu) and these RNAseq transcript assemblies. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from Arabidopsis, common bean, soybean, medicago, poplar, rice, grape and Swiss-Prot proteomes to repeat-soft-masked cowpea genome using RepeatMasker (Smit et al., 2017) with up to 2 kb extension on both ends unless extending into another locus on the same strand. The repeat

library consisted of de novo repeats identified by RepeatModeler (Smit et al., 2008) and Fabaceae repeats in RepBase. Gene models were predicted by homology-based predictors, FGENESH+, FGENESH_EST (similar to FGENESH+, EST as splice site and intron input instead of protein/translated ORF), GenomeScan (Yeh et al., 2001), PASA assembly ORFs (in-house homology constrained ORF finder) and from AUGUSTUS via BRAKER1 (Hoff et al., 2015). The best scored predictions for each locus were selected using positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subject to protein homology analysis to the proteomes mentioned above to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual best hit) BLASTP score and protein coverage is the highest percentage of protein aligned to the best homolog. PASA-improved transcripts were selected based on Cscore, protein coverage, EST coverage, and its CDS overlapping with repeats. A transcript was selected if the Cscore and protein coverage were at least 0.5, or if it had EST coverage while its CDS overlap with repeats was less than 20%. For gene models whose CDS overlap with repeats was more than 20%, its Cscore had to be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subjected to Pfam analysis, and gene models whose protein was more than 30% in Pfam TE domains were removed.

The centromere-abundant 455-bp repeat available from Iwata-Otsubo et al. (2016) was BLASTed against cowpea pseudochromosomes to identify approximate start and end positions of cowpea centromeres. Only alignments with an e-score ≤ $1e^{-50}$ were considered. The region extending from the beginning of the first hit to the end of the last hit was considered to define the centromeric region of each cowpea chromosome.

Recombination rate

A polynomial curve fit of cM position as a function of pseudochromosome coordinate was generated using R for each of the eleven linkage groups from maps of each of ten biparental RIL populations. The linear model R function *lm* was used to compute the linear regression. The R function *predict* was used to create the raster objects and the R function *polynomial* yielded the polynomial coefficients. For each curve, the best fit from polynomials ranging from 4th to 8th order was selected. The first derivative was then calculated for each of the 110 selected polynomials to represent the rate of recombination as cM/Mbp. The mean values of the recombination rates (first derivative) were then calculated along each of the eleven linkage groups after setting all negative values to zero and truncating values at the ends of each linkage group where the polynomial curve clearly was no longer a good fit. A polynomial was then derived for the mean values along each pseudochromosome to represent recombination rate as a function of nucleotide coordinate (cM/Mbp). Data S1 provides the polynomial formulae for each pseudochromosome.

Repeat Analysis

Repeats in the contigs and pseudochromosomes were analyzed using RepeatMasker. An initial library of elements was built by combining the output from Repet, RepeatModeler, LTRharvest/LTRdigest (genometools.org), elements in the Fabaceae section of the RepBase transposon library (Bao et al., 2015) and our own custom pipeline. Subsequent *Vigna*-specific libraries were built by iterative searches. The resulting *Vigna*-specific libraries were used again in iterative searches to build the set of elements in the genome. The set was supplemented with elements identified by similarities to expected domains, including LINE integrases for the LINEs and transposases for the DNA transposons. The set was supplemented by searches based on structural criteria typical of various groups of transposable elements. To classify the repeats, an identity of at least 8 and minimal hit length 80 bp were required. For the LTR retrotransposons, full-length

versions were identified with LTRharvest (Ellinghaus et al., 2008) using the following parameter settings: overlaps best -seed 30 -minlenltr 100 -maxlenltr 3000 -mindistltr 100 -maxdistltr 15000 -similar 80 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3. All candidates were annotated for PfamA domains with hmmer3 software (Eddy, 2011) and filtered for false positives by several criteria, the main ones being the presence of at least one typical retrotransposon domain (e.g., reverse transcriptase, RNaseH, integrase, Gag) and a tandem repeat content below 5%.

Identification of genetic variation

Nearly 1M SNPs with strong support were discovered previously by aligning WGS data from 36 diverse accessions to a draft assembly of IT97K-499-35 (Muñoz-Amatriaín et al., 2017). To position those SNPs on the cowpea reference genome, the 121-base sequences comprised of the SNP position and 60 bases on each side were BLASTed against the cowpea genome assembly with an e-score cutoff of e^{-50}. Only the top hit for each query was kept. The exact SNP position was then calculated. SNPs previously identified as organellar were excluded, together with those hitting multiple locations in the reference genome sequence.

For detection of insertions and deletions, WGS data from 36 diverse accessions (Muñoz-Amatriaín et al., 2017) were used. Reads from each cowpea accession were mapped to the genome assembly using BWA-MEM version 0.7.5a (Li, 2013). Variant calling was carried on each resulting alignment using BreakDancer version 1.4.5 (Chen et al., 2009), with a minimum mapping quality score of 30 and 10 as the minimum number of pair-end reads to establish a connection. The maximum structural variation size to be called by BreakDancer was set to 70 kb. A deletion was considered validated when at least 75% of the SNPs contained in the deletion region were “No Call”. Among the 5,095 putative deletions that spanned SNPs represented in the iSelect array, data were available to validate only 1,558 (30.6%) by this method, leaving the false positive rate uncertain.

To validate the inversion, the sequence assembly of the reference genome was compared to that of a cowpea accession typical of California breeding lines via MUMmer (Kurtz et al., 2004), using a minimum exact match of 100 bp and a minimum alignment length of 1 kb. PCR amplifications of the breakpoint regions were performed to further validate the Vu03 inversion. Four accessions were tested for each of the two orientations (type A and type B); these were parental lines of some of the ten genetic maps used for anchoring (Figure S10) and included one wild cowpea (TVNu-1158). Two primer pairs were designed for each breakpoint region: one to amplify the reference orientation and another to amplify the opposite orientation (Table S10). For the latter, the sequence assembly of the California accession was used to design primers. When primers were designed to amplify the reference orientation, they worked well in type A accessions, but they did not work for the type B accessions (Figure S11). When primers were designed to amplify the opposite orientation, there was PCR product only in the type B accessions (Figure S11). Only the wild cowpea accession did not yield an amplification product for either of the breakpoints, possibly due to sequence variation within the breakpoint regions.

Synteny between cowpea and *Phaseolus vulgaris, Vigna radiata* and *Vigna angularis*

The cowpea IT87K-499-35 genome sequence assembly was aligned to that of common bean v2.1 (Schmutz et al., 2014), adzuki bean (Sakai et al., 2015) and mung bean (Kang et al., 2014) using MUMmer v3.23 (Kurtz et al., 2004). Alignments were generated using pipeline 'nucmer', with a minimum length of an exact match set to 100 bp. Alignments with a length less than 1 kb were filtered out. The output alignments between genomes were visualized using Circos v0.69-3 (Krzywinski et al., 2009) (Figure 3).

Gene Families

The legume-focused gene families from the NSF Legume Federation project (NSF DBI#1444806) was used to compare annotated genes in cowpea with those from other legume proteomes. This is 18,543 gene families, monophyletic for the legume family, including proteomes from cowpea (this study), thirteen other crop and model legumes, and five non-legume species for phylogenetic rooting and evolutionary context (Table S11). Gene families were generated as follows (summarizing method details from https://github.com/LegumeFederation/legfed_gene_families). All-by-all comparisons of protein sequences were calculated using BLAST (Camacho et al., 2009), with post-processing filters of 50% query coverage and 60% identity. The top two matches were used to generate alignments of coding sequences, which were the used to calculate synonymous (Ks) counts per gene pair. For each species pair, histograms of Ks frequencies were the basis for choosing per-species Ks cutoffs for that species pair. A list of all-by-all matches, filtered to remove pairs with Ks values greater than the per-species-pair Ks cutoff, was used for Markov clustering implemented in the MCL program (Enright et al., 2002) with inflation parameter 1.2 and relative score values (transformed from Ks values) indicated with the -abc flag. Sequence alignments were then generated for all families using *muscle* (Edgar, 2004), and hidden Markov models (HMMs) were calculated using *hmmer* (Mistry et al., 2013). Family membership was evaluated relative to median HMM bitscores for each family, with sequences scoring less than 40% of the median HMM bitscore for the family being removed. The HMMs were then recalculated from families (without low-scoring outliers), and used as targets for HMM search of all sequences in the proteome sets, including those omitted during the initial Ks filtering. Again, sequences scoring less than 40% of the median HMM bitscore for the family were removed. Prior to calculating phylogenetic trees, the HMM alignments from the resulting family sets were trimmed of non-aligning characters (characters outside the HMM match states). Phylogenies were calculated using RAxML (Stamatakis et al., 2008), with model PROTGAMMAAUTO, and rooted using the closest available outgroup species.

Identification of a syntelog for increased organ size

The identification of QTLs on Vu08 for organ size (*CPodl8*, *CSw8*, *CLl8*, *CLw8*) is described in Lo, et al. (2018). The SNP markers associated with those QTLs span the genomic region Vu08:36035190-38248903, which contains 313 annotated genes. The corresponding syntenic segment in *Phaseolus vulgaris* (Chr08: 57594596-59622008) was determined using the legumeinfo.org instance of the Genome Context Viewer (GCV) (Cleary and Farmer, 2017). This region contained 289 Phaseolus genes, of which only one (*Phvul.008G285800*) was present in the intersection with a list of genes associated with domestication reported in Schmutz, et al. (2014) as determined using functions of cowpeamine and legumemine (https://mines.legumeinfo.org), which are instances of the InterMine data warehousing system (Kalderimis et al., 2014). The cowpea syntelog of that gene is *Vigun08g217000*, according to the genomic segment alignment provided by the GCV using the gene family assignments described above.

DATA AVAILABILITY

The genome assembly of cowpea IT97K-499-35 is available for browsing and downloadable through Phytozome (phytozome.jgi.doe.gov), the Legume Information System Data Store (https://legumeinfo.org/data/public/), and NCBI SRA BioSample accession SAMN06674009 (also ASM411807v1). Raw PacBio reads for cowpea accession IT97K-499-35 are available at NCBI SRA sample SRS3721827 (study SRP159026). As stated in Muñoz-Amatriaín et al. (2017), raw Illumina reads from 37 diverse cowpea accessions are available under SRA accession SRP077082. RNA-Seq raw reads are available as NCBI SRA biosample accessions SAMN071606186 through SAMN071606198, SAMN07194302 through SAMN07194309 and SAMN07194882 through SAMN07194909 and were described in Yao et al. (2016) and Santos et al. (2018). EST sequences and their GenBank accession numbers are available through the software HarvEST:Cowpea (harvest.ucr.edu) and were described in Muchero et al. (2009).

ACKNOWLEDGMENTS

Authors thank: Panruo Wu, Laxmi Buyhan, Zizhong Chen and Tamar Shinar (University of California Riverside, CA) for allowing access to their HPC server; Suresh Iyer, Amy Mraz, Jon Wittendorp, and Robert Bogden (Amplicon Express, Pullman, WA) for DNA extraction and library prep; Derek Pouchnik and Mark Wildung (Washington State University, Pullman, WA) for PacBio sequencing and library preparation; Thiru Ramaraj (National Center for Genome Resources), Matthew Seetin and Christopher Dunn (Pacific Biosciences of California, Inc., Menlo Park, CA), Brian Walenz (University of Maryland, College Park, MD), John Urban (Brown University, Providence, Rhode Island), and Xingtan "Tanger" Zhang for helpful discussion on assembly tools, trouble shooting, usage and parameter selection; Haibao Tang (JVCI) for help with ALLMAPS; Suk-Ha Lee (Seoul National University, South Korea) for allowing access to their most recent mung bean and adzuki bean genome assemblies; David Goodstein (Joint Genome Institute, Walnut Creek, CA) for assistance coordinating genome annotation; Yi-Ning Guo and Savannah St Clair (UC Riverside, CA) for technical assistance with DNA preparation; and Ira Herniter (UC Riverside, CA) for helpful discussion. This work was supported by the NSF IOS-1543963 ("Advancing the Cowpea Genome for Food Security"), NSF IIS-1526742 ("Algorithms for Genome Assembly of Ultra-Deep Sequencing Data") and NSF IIS-1814359 ("Improving *de novo* Genome Assembly using Optical Maps"). Estimation of the genome size was supported by the Czech Ministry of Education, Youth and Sports (award LO1204 from the National Program of Sustainability I). The analysis of gene families was provided through by in-kind contributions from the USDA Agricultural Research Service, project 5030-21000-069-00-D, while the repetitive elements analysis was supported by the Academy of Finland "Papugeno" (Decision 298314). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

TJC, StL and MMA conceived and supervised the study. StL coordinated the sequencing and executed the assembly with help from SIW. RO identified contamination sequences. TZ and MCL generated the optical maps. MMA, SaL and AN generated genetic maps. JV, PAR and JS contributed to the generation of transcriptome data. SS generated gene annotations. AHS and JT annotated and analyzed repeats. SIW, MMA and TJC contributed to SNP annotation and analysis. HA and AMH identified structural variants. MMA analyzed and validated the chromosomal inversion with help from SaL, SIW and ADF. TJC and SIW estimated recombination rates. QL performed synteny analyses, identified cowpea centromeres and generated a visualization of the distribution of genes, repeats and genetic variation across the genome. TJC, StL, QL and MMA developed the revised chromosome numbering for cowpea. SBC and ADF performed gene family analyses. SaL, SAH and ADF identified the syntelog for multiple organ gigantism. JD and JV estimated the genome size. StL and MMA wrote the manuscript with inputs from TJC, SBC, ADF, JD and AHS.

COMPETING INTERESTS

The authors declare no competing financial interests.

SUPPORTING INFORMATION

Table S10. Primer sequences used to validate the Vu03 inversion.

Table S11. Data sources and references for genome assemblies and annotations used in the gene family analysis.

Data S1. Polynomial formulae for each cowpea chromosome.

Data S2. SNP positions in the cowpea genome.

Data S3. Insertions and deletions identified in the cowpea genome.

Data S4. Haplotype information of parents in ten genetic maps for the Vu03 inversion region.

Data S5. Available WGS sequences for the inversion breakpoint regions.

Data S6. Cowpea gene models contained within the inversion region.

Data S7. Gene family analysis of different legume species.

Data S8. IT97K-499-35 genotypic information.

REFERENCES

ABAD, P., GOUZY, J., AURY, J.-M., CASTAGNONE-SERENO, P., DANCHIN, E. G. J., DELEURY, E., PERFUS-BARBEOCH, L., ANTHOUARD, V., ARTIGUENAVE, F., BLOK, V. C., CAILLAUD, M.-C., COUTINHO, P. M., DASILVA, C., DE LUCA, F., DEAU, F., ESQUIBET, M., FLUTRE, T., GOLDSTONE, J. V., HAMAMOUCH, N., HEWEZI, T., JAILLON, O., JUBIN, C., LEONETTI, P., MAGLIANO, M., MAIER, T. R., MARKOV, G. V., MCVEIGH, P., PESOLE, G., POULAIN, J., ROBINSON-RECHAVI, M., SALLET, E., SÉGURENS, B., STEINBACH, D., TYTGAT, T., UGARTE, E., VAN GHELDER, C., VERONICO, P., BAUM, T. J., BLAXTER, M., BLEVE-ZACHEO, T., DAVIS, E. L., EWBANK, J. J., FAVERY, B., GRENIER, E., HENRISSAT, B., JONES, J. T., LAUDET, V., MAULE, A. G., QUESNEVILLE, H., ROSSO, M.-N., SCHIEX, T., SMANT, G., WEISSENBACH, J. & WINCKER, P. 2008. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nature Biotechnology,* 26**,** 909.

ALVERSON, A. J., ZHUO, S., RICE, D. W., SLOAN, D. B. & PALMER, J. D. 2011. The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. *PLoS one,* 6**,** e16404.

ARUMUGANATHAN, K. & EARLE, E. 1991. Nuclear DNA content of some important plant species. *Plant molecular biology reporter,* 9**,** 208-218.

BAO, W., KOJIMA, K. K. & KOHANY, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA,* 6**,** 11.

BARTRINA, I., JENSEN, H., NOVAK, O., STRNAD, M., WERNER, T. & SCHMÜLLING, T. 2017. Gain-of-function mutants of the cytokinin receptors *AHK2* and *AHK3* regulate plant organ size, flowering time and plant longevity. *Plant physiology***,** pp. 01903.2016.

BERLIN, K., KOREN, S., CHIN, C.-S., DRAKE, J. P., LANDOLIN, J. M. & PHILLIPPY, A. M. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology,* 33**,** 623.

BOUKAR, O., KONG, L., SINGH, B., MURDOCK, L. & OHM, H. 2004. AFLP and AFLP-derived SCAR markers associated with *Striga gesnerioides* resistance in cowpea. *Crop Science,* 44**,** 1259-1264.

BOUKAR, O., BELKO, N., CHAMARTHI, S., TOGOLA, A., BATIENO, J., OWUSU, E., HARUNA, M., DIALLO, S., UMAR, M. L. & OLUFAJO, O. 2018. Cowpea (*Vigna unguiculata*): Genetics, genomics and breeding. *Plant Breeding*

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC bioinformatics,* 10**,** 421.

CANNON, S. B., MITRA, A., BAUMGARTEN, A., YOUNG, N. D. & MAY, G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC plant biology,* 4**,** 10.

CAO, H., HASTIE, A. R., CAO, D., LAM, E. T., SUN, Y., HUANG, H., LIU, X., LIN, L., ANDREWS, W. & CHAN, S. 2014. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience,* 3**,** 34.

CARVALHO, M., MUÑOZ-AMATRIAÍN, M., CASTRO, I., LINO-NETO, T., MATOS, M., EGEA-CORTINES, M., ROSA, E., CLOSE, T. & CARNIDE, V. 2017. Genetic diversity and structure of Iberian Peninsula cowpeas compared to world-wide cowpea accessions using high density SNP markers. *BMC genomics,* 18**,** 891.

CHEN, K., WALLIS, J. W., MCLELLAN, M. D., LARSON, D. E., KALICKI, J. M., POHL, C. S., MCGRATH, S. D., WENDL, M. C., ZHANG, Q. & LOCKE, D. P. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods,* 6**,** 677.

CHIECO, P. & DERENZINI, M. 1999. The Feulgen reaction 75 years on. *Histochemistry and cell biology,* 111**,** 345-358.

CHIN, C.-S., ALEXANDER, D. H., MARKS, P., KLAMMER, A. A., DRAKE, J., HEINER, C., CLUM, A., COPELAND, A., HUDDLESTON, J. & EICHLER, E. E. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10, 563.

CHIN, C.-S., PELUSO, P., SEDLAZECK, F. J., NATTESTAD, M., CONCEPCION, G. T., CLUM, A., DUNN, C., O'MALLEY, R., FIGUEROA-BALDERAS, R. & MORALES-CRUZ, A. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods,* 13**,** 1050.

CLEARY, A. & FARMER, A. 2017. Genome Context Viewer: visual exploration of multiple annotated genomes using microsynteny. *Bioinformatics,* 34**,** 1562-1564.

D'ANDREA, A. C., KAHLHEBER, S., LOGAN, A. L. & WATSON, D. J. 2007. Early domesticated cowpea (*Vigna unguiculata*) from Central Ghana. *Antiquity,* 81**,** 686-698.

DOEBLEY, J. F., GAUT, B. S. & SMITH, B. D. 2006. The molecular genetics of crop domestication. *Cell,* 127**,** 1309-1321.

DOLEZEL, J. 2003. Nuclear DNA content and genome size of trout and human. *Cytometry,* 51**,** 127-128.

DOLEŽEL, J., GREILHUBER, J., LUCRETTI, S., MEISTER, A., LYSÁK, M., NARDI, L. & OBERMAYER, R. 1998. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Annals of Botany,* 82**,** 17-26.

DOLEŽEL, J., GREILHUBER, J. & SUDA, J. 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature protocols,* 2**,** 2233.

EDDY, S. R. 2011. Accelerated profile HMM searches. *PLoS computational biology*, 7, e1002195.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32**,** 1792-1797.

ELLINGHAUS, D., KURTZ, S. & WILLHOEFT, U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics,* 9**,** 18.

ENRIGHT, A. J., VAN DONGEN, S. & OUZOUNIS, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research,* 30**,** 1575-1584.

FARIS, D. 1965. The origin and evolution of the cultivated forms of *Vigna sinensis*. *Canadian journal of genetics and cytology,* 7**,** 433-452.

GOBENA, D., SHIMELS, M., RICH, P. J., RUYTER-SPIRA, C., BOUWMEESTER, H., KANUGANTI, S., MENGISTE, T. & EJETA, G. 2017. Mutation in sorghum LOW GERMINATION STIMULANT 1 alters strigolactones and causes Striga resistance. *Proceedings of the National Academy of Sciences,* 114**,** 4471-4476.

HAAS, B. J., DELCHER, A. L., MOUNT, S. M., WORTMAN, J. R., SMITH JR, R. K., HANNICK, L. I., MAITI, R., RONNING, C. M., RUSCH, D. B. & TOWN, C. D. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research,* 31**,** 5654-5666.

HOFF, K. J., LANGE, S., LOMSADZE, A., BORODOVSKY, M. & STANKE, M. 2015. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics,* 32**,** 767-769.

HUYNH, B. L., EHLERS, J. D., HUANG, B. E., MUÑOZ□AMATRIAÍN, M., LONARDI, S., SANTOS, J. R., NDEVE, A., BATIENO, B. J., BOUKAR, O. & CISSE, N. 2018. A multi□parent advanced generation inter□cross (MAGIC) population for genetic analysis and improvement of cowpea (*Vigna unguiculata* L. Walp.). *The Plant Journal,* 93**,** 1129-1142.

INITIATIVE, I. B. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature,* 463**,** 763.

IWATA-OTSUBO, A., LIN, J.-Y., GILL, N. & JACKSON, S. A. 2016. Highly distinct chromosomal structures in cowpea (*Vigna unguiculata*), as revealed by molecular cytogenetic analysis. *Chromosome Research,* 24**,** 197-216.

KALDERIMIS, A., LYNE, R., BUTANO, D., CONTRINO, S., LYNE, M., HEIMBACH, J., HU, F., SMITH, R., ŠTĚPÁN, R. & SULLIVAN, J. 2014. InterMine: extensive web

services for modern biology. *Nucleic acids research,* 42**,** W468-W472.

KAMATH, G. M., SHOMORONY, I., XIA, F., COURTADE, T. A. & DAVID, N. T. 2017. HINGE: long-read assembly achieves optimal repeat resolution. *Genome research,* 27**,** 747-756.

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. 2014. Genome sequence of mungbean and insights into evolution within Vigna species. *Nature communications,* 5**,** ncomms6443.

KIRKPATRICK, M. 2010. How and why chromosome inversions evolve. *PLoS biology,* 8**,** e1000501.

KOREN, S., WALENZ, B. P., BERLIN, K., MILLER, J. R., BERGMAN, N. H. & PHILLIPPY, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research,* 27**,** 722-736.

KOTIR, J. H. 2011. Climate change and variability in Sub-Saharan Africa: a review of current and future trends and impacts on agriculture and food security. *Environment, Development and Sustainability,* 13**,** 587-605.

KRZYWINSKI, M., SCHEIN, J., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J. & MARRA, M. A. 2009. Circos: an information aesthetic for comparative genomics. *Genome research,* 19**,** 1639-1645.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome biology,* 5**,** R12.

LEISTER, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in genetics,* 20**,** 116-122.

LEVINSON, G. & GUTMAN, G. A. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution,* 4**,** 203-221.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

LI, Z., DEFOORT, J., TASDIGHIAN, S., MAERE, S., VAN DE PEER, Y. & DE SMET, R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *The Plant Cell***,** TPC2015-00877-LSB.

LIN, Y., YUAN, J., KOLMOGOROV, M., SHEN, M. W., CHAISSON, M. & PEVZNER, P. A. 2016. Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences,* 113**,** E8396-E8405.

LO, S., MUÑOZ-AMATRIAÍN, M., BOUKAR, O., HERNITER, I., CISSE, N., GUO, Y.-N.,

ROBERTS, P. A., XU, S., FATOKUN, C. & CLOSE, T. J. 2018. Identification of QTL controlling domestication-related traits in cowpea (*Vigna unguiculata* L. Walp). *Scientific reports,* 8**,** 6261.

LUO, M.-C., GU, Y. Q., PUIU, D., WANG, H., TWARDZIOK, S. O., DEAL, K. R., HUO, N., ZHU, T., WANG, L. & WANG, Y. 2017. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature,* 551.

MASCHER, M., GUNDLACH, H., HIMMELBACH, A., BEIER, S., TWARDZIOK, S. O., WICKER, T., RADCHUK, V., DOCKTER, C., HEDLEY, P. E. & RUSSELL, J. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature,* 544**,** 427.

MISRA, V. A., WANG, Y. & TIMKO, M. P. 2017. A compendium of transcription factor and Transcriptionally active protein coding gene families in cowpea (*Vigna unguiculata* L.). *BMC genomics,* 18**,** 898.

MISTRY, J., FINN, R. D., EDDY, S. R., BATEMAN, A. & PUNTA, M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research,* 41**,** e121-e121.

MUCHERO, W., DIOP, N. N., BHAT, P. R., FENTON, R. D., WANAMAKER, S., POTTORFF, M., HEARNE, S., CISSE, N., FATOKUN, C. & EHLERS, J. D. 2009. A consensus genetic map of cowpea [*Vigna unguiculata* (L) Walp.] and synteny based on EST-derived SNPs. *Proceedings of the national academy of sciences,* 106**,** 18159-18164.

MUÑOZ-AMATRIAÍN, M., MIREBRAHIM, H., XU, P., WANAMAKER, S. I., LUO, M., ALHAKAMI, H., ALPERT, M., ATOKPLE, I., BATIENO, 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., JACKSON, S. A., KUSI, F., LAWLEY, C. T., LUCAS, M. R., MA, Y., TIMKO, M. P., WU, J., YOU, F., BARKLEY, N. A., ROBERTS, P. A., LONARDI, S. & CLOSE, T. J. 2017. Genome resources for climate-resilient cowpea, an essential crop for food security. *The Plant Journal,* 89**,** 1042-1054.

OUÉDRAOGO, J., MAHESHWARI, V., BERNER, D., ST-PIERRE, C.-A., BELZILE, F. & TIMKO, M. 2001. Identification of AFLP markers linked to resistance of cowpea (*Vigna unguiculata* L.) to parasitism by *Striga gesnerioides*. *Theoretical and Applied Genetics,* 102**,** 1029-1036.

OUÉDRAOGO, J. T., TIGNEGRE, J.-B., TIMKO, M. P. & BELZILE, F. J. 2002. AFLP markers linked to resistance against *Striga gesnerioides* race 1 in cowpea (*Vigna unguiculata*). *Genome,* 45**,** 787-793.

OUNIT, R. & LONARDI, S. 2016. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics,* 32**,** 3823-3825.

PAN, W., WANAMAKER, S. I., AH-FONG, A. M., JUDELSON, H. S. & LONARDI, S. 2018. Novo&Stitch: accurate reconciliation of genome assemblies via optical maps. *Bioinformatics,* 34**,** i43-i51.

PARIDA, A., RAINA, S. & NARAYAN, R. 1990. Quantitative DNA variation between and within chromosome complements of Vigna species (Fabaceae). *Genetica,* 82**,** 125-133.

PUIG, M., CASILLAS, S., VILLATORO, S. & CÁCERES, M. 2015. Human inversions and their functional consequences. *Briefings in functional genomics,* 14**,** 369-379.

RIEFLER, M., NOVAK, O., STRNAD, M. & SCHMÜLLING, T. 2006. Arabidopsis cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism. *The Plant Cell,* 18**,** 40-54.

SAKAI, H., NAITO, K., OGISO-TANAKA, E., TAKAHASHI, Y., ISEKI, K., MUTO, C., SATOU, K., TERUYA, K., SHIROMA, A. & SHIMOJI, M. 2015. The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Scientific reports,* 5**,** 16780.

SANTOS, J. R. P., NDEVE, A. D., HUYNH, B.-L., MATTHEWS, W. C. & ROBERTS, P. A. 2018. QTL mapping and transcriptome analysis of cowpea reveals candidate genes for root-knot nematode resistance. *PloS one,* 13**,** e0189185.

SCHMUTZ, J., CANNON, S. B., SCHLUETER, J., MA, J., MITROS, T., NELSON, W., HYTEN, D. L., SONG, Q., THELEN, J. J. & CHENG, J. 2010. Genome sequence of the palaeopolyploid soybean. *nature,* 463**,** 178.

SCHMUTZ, J., MCCLEAN, P. E., MAMIDI, S., WU, G. A., CANNON, S. B., GRIMWOOD, J., JENKINS, J., SHU, S., SONG, Q. & CHAVARRO, C. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics,* 46**,** 707.

SERDECZNY, O., WATERS, E. & CHAN, S. 2016. Non-economic loss and damage in the context of climate change. *German Development Institute Discussion Paper,* 3**,** 2016.

SHE, C. W., JIANG, X. H., OU, L. J., LIU, J., LONG, K. L., ZHANG, L. H., DUAN, W. T., ZHAO, W. & HU, J. C. 2015. Molecular cytogenetic characterisation and phylogenetic analysis of the seven cultivated V igna species (Fabaceae). *Plant Biology,* 17**,** 268-280.

SHELTON, J. M., COLEMAN, M. C., HERNDON, N., LU, N., LAM, E. T., ANANTHARAMAN, T., SHETH, P. & BROWN, S. J. 2015. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC genomics,* 16**,** 734.

SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. &

ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31**,** 3210-3212.

SINGH, B. 2014. Cowpea: The Food Legume of the 21st Century, *Crop Science Society of America*.

SMIT, A., HUBLEY, R. & GREEN, P. 2017. 1996–2010. RepeatMasker Open-3.0.

SMIT, A., HUBLEY, R. R. & GREEN, P. R. 2008. Open-1.0. 2008–2015. *Institute for Systems Biology, Seattle, WA, USA*.

STAMATAKIS, A., HOOVER, P. & ROUGEMONT, J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic biology,* 57**,** 758-771.

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.

TIMKO, M. P., RUSHTON, P. J., LAUDEMAN, T. W., BOKOWIEC, M. T., CHIPUMURO, E., CHEUNG, F., TOWN, C. D. & CHEN, X. 2008. Sequencing and analysis of the gene-rich space of cowpea. *BMC genomics,* 9**,** 103.

VARSHNEY, R. K., CHEN, W., LI, Y., BHARTI, A. K., SAXENA, R. K., SCHLUETER, J. A., DONOGHUE, M. T., AZAM, S., FAN, G. & WHALEY, A. M. 2012. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature biotechnology,* 30**,** 83.

VASCONCELOS, E. V., DE ANDRADE FONSECA, A. F., PEDROSA-HARAND, A., DE ANDRADE BORTOLETI, K. C., BENKO-ISEPPON, A. M., DA COSTA, A. F. & BRASILEIRO-VIDAL, A. C. 2015. Intra- and interchromosomal rearrangements between cowpea [*Vigna unguiculata* (L.) Walp.] and common bean (*Phaseolus vulgaris* L.) revealed by BAC-FISH. *Chromosome Res,* 23**,** 253-66.

WICKER, T., SABOT, F., HUA-VAN, A., BENNETZEN, J. L., CAPY, P., CHALHOUB, B., FLAVELL, A., LEROY, P., MORGANTE, M., PANAUD, O., PAUX, E., SANMIGUEL, P. & SCHULMAN, A. H. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics,* 8**,** 973.

WU, Y., BHAT, P. R., CLOSE, T. J. & LONARDI, S. 2008. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet,* 4**,** e1000212.

XU, P., WU, X., WANG, B., HU, T., LU, Z., LIU, Y., QIN, D., WANG, S. & LI, G. 2013. QTL mapping and epistatic interaction analysis in asparagus bean for several characterized and novel horticulturally important traits. *BMC genetics,* 14**,** 4.

YANG, K., TIAN, Z., CHEN, C., LUO, L., ZHAO, B., WANG, Z., YU, L., LI, Y., SUN, Y. & LI, W. 2015. Genome sequencing of adzuki bean (*Vigna angularis*) provides

insight into high starch and low fat accumulation and domestication. *Proceedings of the National Academy of Sciences,* 112**,** 13213-13218.

YAO, S., JIANG, C., HUANG, Z., TORRES□JEREZ, I., CHANG, J., ZHANG, H., UDVARDI, M., LIU, R. & VERDIER, J. 2016. The *Vigna unguiculata* gene expression atlas (VuGEA) from de novo assembly and quantification of RNA□seq data provides insights into seed maturation mechanisms. *The Plant Journal,* 88**,** 318-327.

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.

TABLES

Table 1. Assembly statistics for stitched contigs, scaffolds, and pseudochromosomes.

	Stitched contigs	Scaffolds	Pseudochromosomes
N50 (bp)	10,911,736	16,417,655	41,684,185
L50	16	12	6
NG50 (bp)	9,203,620	15,388,583	41,327,797
LG50	21	15	7
total (bp)	518,799,885	519,432,264	519,435,864
contigs/scaffolds	765	722	686
contigs/scaffolds ≥ 100kbp	177	135	103
contigs/scaffolds ≥ 1Mbp	61	38	13
contigs/scaffolds ≥ 10Mp	18	21	11
longest contig/scaffold (bp)	22,343,392	30,539,429	65,292,630
% N	0.0%	0.523%	0.524%
mapped SNPs	49,888	49,888	49,888
GC	33.0%	32.994%	32.994%

FIGURE LEGENDS

Figure 1. **Landscape of the cowpea genome.** (a) Cowpea chromosomes in Mb, with red lines representing centromeric regions based on a 455-bp tandem repeat alignment (Iwata-Otsubo et al., 2016); (b) Recombination rate at each 1Mb; (c) Gene density in 1Mb windows; (d) Repeat coverage in 1Mb windows; (e) SNP density in 1Mb windows.

Figure 2. Large chromosomal inversion detected on Vu03. (a) The relationships between genetic and physical positions are shown for SNPs on four genetic maps (1 to 4). Maps (1) to (3) show a 4.2 Mb region in an inverted orientation (red arrow) while map (4) shows no recombination in that same region (area contained within red lines). (b) Sequence comparison between IT97K-499-35 (reference genome) and a "type B" accession for the region including the Vu03 chromosomal inversion. Red color indicates same orientation between both sequences, while in blue are shown those sequences having opposite orientations between accessions.

Figure 3. Synteny view between cowpea (Vu; *Vigna unguiculata*) and other closely related diploid species. These include: (a) adzuki bean (Va*; Vigna angularis*), (b) mung bean (Vr; *Vigna radiata*), and (c) common bean (Pv; *Phaseolus vulgaris*) using the revised cowpea chromosome numbering system.