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

Permalink
https://escholarship.org/uc/item/28f458sf

Journal
The Plant journal: for cell and molecular biology, 98(5)

ISSN
0960-7412

Authors
Lonardi, Stefano
Muñoz-Amatriaín, María
Liang, Qihua
et al.

Publication Date
2019-06-01

DOI
10.1111/tpj.14349

Peer reviewed
RESERVE

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

Stefano Lonardi1,*,†, Maria Muñoz-Amatriain2,*,†,‡, Qihua Liang1, Shengqiang Shu3, Steve I. Wanamaker2, Sassoum Lo2, Jaakko Tanskanen4,†,‡, Alan H. Schulman4,†,‡, Tingting Zhu5, Ming-Cheng Luo6, Hind Alhakami7, Rachid Ounit1, Abid Md. Hasan1, Jerome Verdier8, Philip A. Roberts9, Jansen R.P. Santos8,†,‡, Arsenio Ndeve9, Jaroslav Dolezel11, Jan Vrana11, Samuel A. Hokin12, Andrew D. Farmer12, Steven B. Cannon13 and Timothy J. Close2

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

Received 28 January 2019; revised 25 March 2019; accepted 28 March 2019; published online 24 April 2019.

*For correspondence (e-mails stelo@cs.ucr.edu and maria.munoz.amatriain@colostate.edu).
†Present address: Department of Soil and Crop Sciences Colorado State University Fort Collins CO 80523 USA.
‡These authors contributed equally to this work.

SUMMARY

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 with 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.

Keywords: chromosomal inversion, cowpea, domestication, genome annotation, genome evolution, genome size, next-generation sequencing, legumes, *Phaseolus vulgaris*, repetitive elements, *Vigna unguiculata*.

Linked article: This paper is the subject of a Research Highlight article. To view this Research Highlight article visit https://doi.org/10.1111/tpj.14382.
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 (Faris, 1965; D’Andrea et al., 2007), from where it spread into all continents and now is commonly grown in many parts of Asia, Europe, USA, 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; Munoz-Amatriain et al., 2017). A highly fragmented draft assembly and BAC sequence assemblies of IT97K-499-35 were previously generated (Muñoz-Amatriain 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 V. 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 in chromosome numbering to facilitate comparative studies. The publicly available genome sequence lays the foundation for basic and applied research, enabling progress towards the improvement in 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 (Dolezél, 2003), k-mer analysis and optical mapping (see Experimental procedures for more detail). In brief, cytometry indicated that the 2C nuclear DNA amount of V. unguiculata IT97K-499-35 is 1.310 ± 0.026 pg DNA (mean ± SD), which corresponds to 1C genome size of 640.6 Mbp (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 (Dolezél 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 (Dolezél 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 polymerase chain reaction (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; Munoz-Amatriain et al., 2017). Here, a fully homozygous (single haplotype; see Experimental procedures) stock was sequenced using Pacific Biosciences of California, Menlo Park, CA, USA) single-molecule real-time (SMRT) sequencing. In total, 56.8 Gb of sequence data were generated (~91.7 x 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, CA, USA) optical maps (Cao et al., 2014) were generated using nicking enzymes BspQI and BssSI (Tables S1 and S2). The size of the BsaQI optical map is 622.21 Mb, while the size of the BssSI 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.,
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 10 genetic maps containing 44 003 unique Illumina iSelect SNPs (Muñoz-Amatriain et al., 2017) were used to anchor and orient sequence scaffolds into 11 pseudochromosomes (i.e., pseudomolecules) via ALLMAPS (Tang et al., 2015). Details of the 10 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: (i) 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); (ii) about 129 000 contigs (500 bp or longer) of the whole-genome shotgun (WGS) assembly generated previously (Muñoz-Amatriain et al., 2017; 99.69% mapped of which 98.69% had MAPQ > 30); (iii) about 178 000 BAC sequence assemblies generated previously (Muñoz-Amatriain et al., 2017; 99.95% mapped of which 98.39% had MAPQ > 30); and (iv) about 157 000 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 × 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-Amatriain 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 (telomiceric) 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 25.86:46.35 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 an 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 Experimental procedures). In total, 29 773 protein-coding loci were annotated, along with 12 514 alternatively spliced transcripts. Most (95.9%) of the 1440 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 3881 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% versus 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 TEs 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 TE 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 (Figures 1 and S8; Data S1). Contrasting examples include Vu04, where the recombination rate near the telomeres of both arms of this metacentric chromosome are roughly 10 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, 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 data from an additional 36 diverse accessions relevant to Africa, China and USA were previously used to identify single-nucleotide polymorphisms (Muñoz-Amatriain et al., 2017). Almost all (99.83%) of the 957 710 discovered single nucleotide polymorphisms (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-Amatriain 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; Table S8). This indicates that the intended bias towards genes in the iSelect array design (Muñoz-Amatriain 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 (Figures 1 and S8). 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, 10 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 (Figures 2a and 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-Amatriain 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

© 2019 The Authors.
The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology., The Plant Journal, (2019), 98, 767–782

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 1 Mb. (c) Gene density in 1 Mb windows. (d) Repeat coverage in 1 Mb windows. (e) Single nucleotide polymorphism (SNP) density in 1 Mb windows.
The 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 Experimental procedures; 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 × wild cross (Lo et al., 2018; IT99K-573-1-1 × TVNu-1158; Figure S10), which indicates that the wild parent has the opposite orientation of the cultivated accession. Because 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 (Ouedraogo et al., 2001, 2002; Boukar et al., 2004) were compared with 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 *Vigun03 g220400*, which is located inside the inverted region on Vu03 (Data S6) and is highly expressed in root tissue (https://legumeinfo.org/...
Synteny analyses were performed between cowpea and its close relatives adzuki bean (Vigna angularis), mung bean (P. vulgaris) as well as V. radiata and common bean (P. 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 superfamilies 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% (versus Vr) or 11% (versus 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 (versus 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 versus Vr, and comprised a smaller genome share than in Va.

A similar comparison was made to the 473 Mb genome assembly of P. vulgaris (Schmutz et al., 2014; Pv) with a genome estimated to be only 9% smaller (587 Mb; http://data.kew.org/cvalues). However, Pv has a higher TE content than cowpea, 45.2% versus 39%, of which 39% versus 33% are retrotransposons. In Pv, the Gypsy elements comprise 25% of the genome versus 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 versus 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, for example, 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 (<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

**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.
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 2520 families without cowpea gene membership, which is comparable to the average number of families without membership (3057) for six other sequenced genomes in the Phaseoleae. The 2520 ‘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, versus 90 and 52 in adzuki and mung bean, respectively. The NBS-LRR superfamily contains 402 annotated genes, versus 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 (Doebely et al., 2006). Recently, a genomic region related to increased organ size in cowpea was identified on Vu08 using a recombinant inbred line (RIL) population derived from a domesticated × 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. (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 Vigun08 g217000. 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). Vigun08 g217000 is thus a candidate gene for further investigation.

EXPERIMENTAL PROCEDURES

Estimation of genome size

Flow cytometric estimation of genome size followed the protocol of Dolezel et al. (2007). Briefly, suspensions of cell nuclei were prepared from 50 mg of young leaf tissue of cowpea IT97K-499-35, and of Solanum 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 (PI) 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 unique 27-mers in the range x = 2–10 000 is 31.381 × 10⁹. As is usually done, 27-mers that appear only once are excluded because they are considered erroneous, that is to contain sequencing errors. The estimated genome size based on the formula bp = (# of unique 27-mers–k+1)/peak depth of coverage is thus 31.381 × 10⁹/56 = 560 379 733 bp.
Bionano Genomics optical maps

High-molecular-weight DNA was isolated by Amplicon Express (Pullman, WA, USA) 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-Amatriain et al., 2017; Data S8). The nicking endonucleases Nt.BspQI and Nb.BssSI (New England Biolabs, Ipswich, MA, USA) 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) as per Luo et al. (2017). The DNA sample was loaded onto the nanochannel array of an IrysChip (Bionano Genomics) and then imaged using the Irys system (Bionano Genomics). For the BspQI map, seven separate runs (132 unique scans) were generated, and a total of 108 Gb (>90 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 BssSI map assembly. For the BssSI map, five separate runs (123 unique scans) were generated, and a total of 186 Gb (>310 kb 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 molecule status and the BNG BssSI 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⁻¹) and antifungal (nystatin, 100 units per ml) agents. About 70 g of seedling tissue was collected, frozen in liquid nitrogen, stored at –80°C and shipped on dry ice. High-molecular-weight gDNA was prepared from nuclei isolated from the tissue by Amplicon Express (Pullman, WA, USA).

Pacific Biosciences sequencing. Pacific Biosciences reads were generated at Washington State University (Pullman, WA, USA) following the ‘Procedure and Checklist-20 kb Template Preparation Using BluePippin Size Selection System’ (PN 100-286-000-5) protocol provided by Pacific Biosciences (Menlo Park, CA, USA) and the Pacific Biosciences SMRTbell Template Prep kit 1.0 (PN 100-100-100-100). Resulting SMRTbell libraries were size selected using the BluePippin (Sage Biosciences) according to 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 (PN N100-372-700). The only deviation from standard protocol was to increase the binding time to 1–3 h, compared with the suggested 30 min. Bound SMRTbell libraries were loaded onto the SMRT cells using the standard MagBead protocol, and the MagBead Buffer Kit v2.0 (PN 100-642-800). The standard MagBead sequencing protocol followed the DNA Sequencing Kit 4.0 v2 (PN 100-612-400), which is known as P6/C4 chemistry. PacBio RS II sequencing data were collected in 6-h movies and Stage Start was enabled to capture the longest subreads possible.

Sequence quality control. First, CLARK and CLARK-S (Unlut 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 ~5000 bacterial/viral genomes from NCBI RefSeq; (ii) human genome, Homo sapiens, assembly GRCh38; (iii) Illumina-based cowpea draft genome, V. unguiculata (Muñoz-Amatriain et al., 2017), assembly v0.03; (iv) soybean, Glycine max (Schmutz et al., 2010), assembly Gmax.275_v2.0; (v) common bean, P. vulgaris (Schmutz et al., 2014), assembly Pvulgaris.218_v1.0; (vi) adzuki bean, V. angularis (Yang et al., 2015), assembly adzuki.vers3.ref.fa.cor; (vii) mung bean, V. radiata (Kang et al., 2014), assembly Vradi.vers6.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 ABRUJNI 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 2 Tb) 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 ~6 M reads (two CAnU runs were optimized for highly repetitive genomes). Falcon and ABRUJNI were run on 3.54 M error-corrected reads produced by CAnU (30.62 Gbp, or 49.4 x genome equivalent). Each assembly took about 4–5 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. White-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 00009285.1); (ii) Rhizobium vigneae (NCBI accession GCF 000732195.1); (iii) Mesorhizobium sp. NBMC P2-C3 (NCBI accession GCF 000568555.1); (iv) Streptomyces purpurogenesclerotici (NCBI accession GCF 001280155.1); (v) Caulobacter vibrioides (NCBI accession GCF 001449105.1); (vi) mitochondrion of V. radiata (Alversen et al., 2011; NCBI accession NC_015121.1); (vii) mitochondrion of V. angularis (NCBI accession NC_021092.1); (viii) chloroplast of V. unguiculata (NCBI accession NC_018051.1 and KJ488104.1); and (ix) human genome (assembly GRCh38). The white list included the genomes of: (i) soybean (G. max; Schmutz et al., 2010; assembly Gmax.275_v2.0); (ii) common bean (P. vulgaris; Schmutz et al., 2014; assembly Pvulgaris.218_v1.0); (iii) adzuki bean (V. angularis; Yang et al., 2015; assembly adzuki.vers3.ref.fa.cor); (iv) mung bean (V. radiata; Kang et al., 2014; assembly Vradi.vers6.cor); and (v) Illumina-based cowpea draft genome (V. unguiculata; Muñoz-Amatriain et al., 2017; assembly v0.03). Each assembled contig was BLASTed against the ‘white’ genome and the ‘black’ genomes, and all high-quality alignments (e-score > 1e⁻⁷, 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 sub-sets of assembled contigs from the individual assemblies that are mutually overlapping with high confidence; (ii) computes a 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 polished assembly via the Kansas State University (KSU) assembly alignments and/or EXONERATE alignments of proteins to the end of the last hit was considered to define the centromeric region of each cowpea chromosome. Only alignments with an e-score ≥30% in Pfam TE domains were removed. 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 corresponding 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 pseudochromosomal coordinates was generated using R for each of the 11 linkage groups after setting all negative values to zero. The mean values of the recombination rates (first derivative) were then calculated along each of the 11 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.
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 (genomertools.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 TEs. To classify the repeats, an identity of at least 80% or minimal hit length 80 bp were required. For the LTR retrotransposons, full-length versions were identified withLTRharvest/ELLinghaus et al., 2008) using the following parameter settings: overlapping best seed -30 minlenltr 100 -maxlenltr 3000 -mindistltr 100 -maxdistltr 15000 -similar 80 -mintsd 4 -maxtsd 7 -del 3. All candidates were annotated for PfamA domains with mindistltr 100 -maxdistltr 15000 -similar 80 -mintsd 4 -maxtsd 7 -del 3. All candidates were annotated for PfamA domains with

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 IT87K-499-35 (Muñoz-Amatriain 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-02. 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-Amatriain 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 6595 putative insertions that spanned SNPs represented in the iSelect array, data were available to validate only 1558 (30.6%) by this method, leaving the false-positive rate uncertain.

To validate the inversion, the sequence assembly with paired reads of the 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 10 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 <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 Federa
tion project (NSF DBI#1444806) were 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), 13 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 80% identity. The top two matches were used to generate alignments of coding sequences, which were then 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 bit-scores 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 PROTGRAMMAUTO, and rooted using the closest available outgroup species.
Identification of a syntelog for increased organ size

The identification of QTLs on Vv08 for organ size (CPod18, CSw8, CL18, CLw8) is described in Lo et al. (2018). The SNP markers associated with those QTLs span the genomic region Vv08: 36035190-38248903, which contains 313 annotated genes. The corresponding syntenic segment in P. vulgaris (Chr08: 57594596-59622008) was present in the intersection with a list of genes associated with domestication reported in Schmutz et al. (2014) as determined using functions of cowpea and legumemine (https://mines.legumeinfo.org), which are instances of the InterMine data warehousing system (Kalderimis legumemine (https://mines.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 cowpea 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 Vign08 g217000, 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 is downloadable through Phytozome (phytozome.jgi.doe.gov), the Legume Information System Data Store (https://legumeinfo.org/data/public/), and NCBI SRA BioSample accession SAMN0674009 (also ASM 411807v1). Raw PacBio reads for cowpea accession IT97K-499-35 are available at NCBI SRA sample SRS3712827 (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 accesses 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).

ACKNOWLEDGEMENTS

The authors thank Panruo Wu, Laxmi Buyhan, Zichong Chen and Tamar Shinar (University of California Riverside, CA, USA) for allowing access to their HPC server; Suresh Iyer, Amy Mraz, Jon Wittendorp and Robert Bogden (Amplicon Express, Pullman, WA, USA) for DNA extraction and library prep; Derek Pouchnik and Mark Wildung (Washington State University, Pullman, WA, USA) 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, USA), Brian Walenz (University of Maryland, College Park, MD, USA), John Urban (Brown University, Providence, Rhode Island, USA) and Xingtan ‘Tanger’ Zhang for helpful discussion on assembly tools, trouble shooting, usage and parameter selection; Haibo 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, USA) for assistance coordinating genome annotation; Yi-Ning Guo and Savannah St Clair (UC Riverside, CA, USA) for technical assistance with DNA preparation; and Ira Herniter (UC Riverside, CA, USA) 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 in-kind contributions from the USDA Agricultural Research Service, project 5030-21000-068-00-D, while the repetitive elements analysis was supported by the Academy of Finland ‘Papugeno’ (Decision 298314). The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US 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

Additional Supporting Information may be found in the online version of this article.

Figure S1. Estimation of cowpea genome size using flow cytometry.

Figure S2. 27-mer distribution of occurrences.

Figure S3. Pre-filter PacBio read length distribution.

Figure S4. Post-filter PacBio read length distribution.

Figure S5. Subread Filtering PacBio read length distribution.

Figure S6. Post-filter PacBio read quality distribution.

Figure S7. Synteny view between cowpea and common bean using the previous chromosome nomenclature.

Figure S8. Gene and repeat densities, and recombination rate in the cowpea genome.

Figure S9. SNP distribution in the cowpea genome.

© 2019 The Authors.
The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology., The Plant Journal, (2019), 98, 767–782
Figure S10. Cowpea pseudochromosome Vu03 reconstructed from 10 genetic maps using ALLMAPS.

Figure S11. PCR amplification of the regions surrounding the two breakpoints of the inversion.

Figure S12. Comparison between cowpea and adzuki bean for the cowpea inversion region.

Table S1. Statistics for BspQI optical map.

Table S2. Statistics for the BssSI optical map.

Table S3. Assembly statistics for the eight individual draft assemblies.

Table S4. Characteristics of the 10 genetic maps used for pseudochromosome construction.

Table S5. Cross-reference between old and revised chromosome numbers for cowpea (Vu).

Table S6. Annotated repeat abundances in cowpea.

Table S7. Centrotermere position prediction.

Table S8. Number and location of SNPs relative to annotated cowpea genes.

Table S9. Comparative repeat abundance in Vigna species.

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 10 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. Table S11, 499-35 genotypic information.

REFERENCES

Abad, P., Gouzy, J., Aury, J.-M. et al. (2008) Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat. Biotechnol. 26, 909.

Alverson, A.J., Zhuo, S., Rice, D.W., Sloan, D.B. and Palmer, J.D. (2011) The molecular genetics of the leafy vine Vigna radiata and the analysis of recombination across short mitochondrial repeats. PLoS One 6, e16404.

Arumuganathan, K. and Earle, E. (1991) Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9, 208-218.

Bao, W., Kojima, K.K. and 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. and Schmulling, T. (2017) Gain-of-function mutants of the cytokinin receptors AHK2 and AHK3 regulate plant organ size, flowering time and plant longevity. Plant Physiol. 173, 1783-1797.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M. and Phillippy, A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623.

Boukar, O., Kong, L., Singh, B., Murdock, L. and Ohm, H. (2004) AFLP and AFLP-derived SCAR markers associated with Striga gesnerioides resistance in cowpea. Crop Sci. 44, 1259-1264.

Boukar, O., Belko, N., Chamarti, S., Togola, A., Batiens, O., Ozwus, E., Haruna, M., Diallo, S., Umar, M.L. and Olufajo, O. (2018) Cowpea (Vigna unguiculata) genetics, genomics and breeding. Plant Breed. 1-10, https://doi.org/10.1111/pbr.12969.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2008) BLAST+: architecture and applications. BMC Bioinform. 10, 421.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol. 4, 10.

Cao, H., Hastie, A.R., Cao, D., Lam, E.T., Sun, Y., Huang, H., Liu, X., Lin, L., Andrews, W. and Chan, S. (2014) Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience 3, 34.

Carvalho, M., Munoz-Amatriain, M., Castro, L., U-no.Neto, T., Matos, M., Egea-Cortines, M., Rosa, E., Close, T. and Camide, V. (2017) Genetic diversity and structure of Iberian Peninsula cowpeas compared to world-wide cowpea accessions using high density SNP markers. BMC Genom. 18, 891.

Chen, K., Wallis, J.W., McClellan, M.D., Larson, D.E., Kallick, J.M., Pohl, C.S., McGrath, S.D., Wendel, M.C., Zhang, Q. and Locke, D.P. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods, 6, 677.

Chico, P. and Derenzeni, M. (1999) The Feulgen reaction 75 years on. Histochim. Cell Biol. 111, 345-358.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Hudson, J. and Eichler, E.E. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods, 10, 563.

Chin, C.-S., Peluso, P., Sedlacek, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O’Malley, R., Figueroa-Balderas, R. and Morales-Cruz, A. (2018) Phased diploid genome assembly with single-molecule realtime sequencing. Nat. Methods, 13, 1050.

Cleary, A. and 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. and Watson, D.J. (2007) Early domesticated cowpea (Vigna unguiculata) from Central Ghana. Antiquity, 81, 689-698.

Doebley, J.F., Gaut, B.S. and 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, 57, 127-128.

Dolezel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysák, M., Nardi, L. and Obermayer, R. (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. Ann. Bot. 82, 17-26.

Dolezel, J., Greilhuber, J. and Suda, J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat. Protoc. 2, 2233.

Eddy, S.R. (2011) Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792-1797.

Ellinghaus, D., Kurtz, S. and Willhoft, U. (2008)LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinform. 9, 18.

Enright, A.J., van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575-1584.

Farias, D. (1965) The origin and evolution of the cultivated forms of Vigna sinensis. Can. J. Genet. Cytol. 7, 433-452.

Gobena, D., Shimels, M., Rich, P.J., Ruyter-Spira, C., Bouwmeester, H., Kanuganti, S., Mengiste, T. and Ejeta, G. (2017) Mutation in sorghum LOW GERMINATION STIMULANT 1 alters strigolactones and causes Striga resistance. Proc. Natl Acad. Sci. USA, 114, 4471-4476.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K. Jr., Hancock, L.I., Maiti, R., Ronning, C.M., Rusch, D.B. and Town, C.D. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment schemes. Nucleic Acids Res. 31, 5654-5666.

Hoff, K.J., Lange, S., Lompsadze, A., Borodovsky, M. and Stanke, M. (2015) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics, 32, 767-769.

Huyhn, B.L., Ehlers, J.D., Huang, B.E., Munoz-Amatriain, M., Lonardi, S., Santos, J.R., Ndeve, A., Batiens, B.J., Boukar, O. and Cisse, N. (2016) A multi-parent advanced generation inter-cross (MAGIC) population for genetic analysis and improvement of cowpea (Vigna unguiculata L. Walp.). Plant J. 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. and Jackson, S.A. (2016) Highly distinct chromosomal structures in cowpea (Vigna unguiculata), as

The Plant Journal published by John Wiley & Sons Ltd and Experimental Biology, The Plant Journal, (2019), 98, 767–782.
revealed by molecular cytogenetic analysis. Chromosoma Res. 24, 197–216.
Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., Hu, F., Smith, R., Stepán, R. and Sullivan, J. (2014) InterMine: extensive web services for modern biology. Nucleic Acids Res. 42, W487-W487.
Kamath, G.M., Shehata, L., Xia, F., Courtade, T.A. and David, N.T. (2017) HINGE: long-read assembly achieves 704 consistent repeat resolution. Genome Res. 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. and Lee, J. (2014) Genome sequence of mungbean and insights into evolution within Vigna species. Nat. Commun. 5, ncomms7443.
Kirkpatrick, M. (2010) How and why chromosome inversions evolve. PLoS Biol. 8, e1000501.
Koren, S., Wain, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2010) How and why chromosome inversions evolve. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.
Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. Genome Biol. 5, R12.
Leister, D. (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends Genet. 20, 116–122.
Lewinsohn, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221.
Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
Li, Z., Deloort, J., Tasdighian, S., Maere, S., de, Van Peer, Y. and De Smet, P.A. (2018) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A. (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA. 113, E8396-E8405.
Lo, S., Munoz-Amatriain, M., Boukar, O., Herniter, I., Cisse, N., Guo, Y.-N., Roberts, P.A., Xu, S., Fatokun, C. and Close, T.J. (2018) Identification of QTL controlling domestication-related traits in cowpea [Vigna unguiculata (L.) Walp]. Ecol. Sci. 8, 6261.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., De, Van Peer, Y. and De Smet, P.A. (2016) Genome sequence of mungbean large genomes. Genome Res. 26, 326–344.
Wu, Y., Bhat, P.R., Close, T.J. and 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. and Li, G. (2013) QTL mapping and epistatic interaction analysis in asparagus bean for several characterized and novel horticulturally important traits. *BMC Genet.* 14, 4.

Yang, K., Tian, Z., Chen, C., Luo, L., Zhao, B., Wang, Z., Yu, L., Li, Y., Sun, Y. and Li, W. (2015) Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proc. Natl Acad. Sci. USA,* 112, 13213-13218.

Yao, S., Jiang, C., Huang, Z., Torres-Jerez, I., Chang, J., Zhang, H., Udvardi, M., Liu, R. and 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. *Plant J.* 88, 318-327.

Yeh, R.-F., Lim, L.P. and Burge, C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.* 11, 803-816.