10-1-2016

Genome Resources for Climate-Resilient Cowpea, An Essential Crop for Food Security

María Muñoz-Amatriaín
University of California - Riverside

Hamid Mirebrahim
University of California - Riverside

Pei Xu
Zhejiang Academy of Agricultural Sciences

Steve Wanamaker
University of California - Riverside

MingCheng Luo
University of California - Davis

See next page for additional authors

Published version. *Plant Journal,* Vol. 89, No. 5 (October 2016): 1042-1054. DOI.

Authors
María Muñoz-Amatriaín, Hamid Mirebrahim, Pei Xu, Steve Wanamaker, MingCheng Luo, Hind Alhakami, Matthew Alpert, Ibrahim Atokple, Benoit J. Batineno, Ousmane Boukar, Serdar Bozdag, Ndiaga Cisse, Issa Drabo, Jeffrey D. Ehlers, Andrew Farmer, Christian Fatokun, Yong Q. Gu, Yi-Ning Guo, Bao-Lam Huynh, Scott A. Jackson, Francis Kusi, Cynthia T. Lawley, Mitchell R. Lucas, Yaqin Ma, Michael P. Timko, Jiajie Wu, Frank You, Noelle A. Barkley, Philip A. Roberts, Stefano Lonardi, and Timothy J. Close

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2016) doi: 10.1111/tpj.13404

RESOURCE

Genome resources for climate-resilient cowpea, an essential crop for food security

María Muñoz-Amatriaín[1,*,†], Hamid Mirebrahim[2,†], Pei Xu[3], Steve I. Wanamaker[1], MingCheng Luo[4], Hind Alhakami[2], Matthew Alpert[2], Ibrahim Atokple[5], Benoit J. Batieno[6], Ousmane Boukar[7], Serdar Bozdag[2,8], Ndiaga Cisse[9], Issa Drabo[6], Jeffrey D. Ehlers[1,10], Andrew Farmer[11], Christian Fatokun[12], Yong Q. Gu[13], Yi-Ning Guo[1], Bao-Lam Huynh[14], Scott A. Jackson[15], Francis Kusi[5], Cynthia T. Lawley[16], Mitchell R. Lucas[1], Yaqin Ma[1,4], Michael P. Timko[17], Jiajie Wu[4], Frank You[4,18], Noelle A. Barkley[19], Philip A. Roberts[14], Stefano Lonardi[2] and Timothy J. Close[1,*]

[1]*Department of Botany and Plant Sciences, University of California, Riverside, CA, USA,*
[2]*Department of Computer Science and Engineering, University of California, Riverside, CA, USA,*
[3]*Institute of Vegetables, Zhejiang Academy of Agricultural Sciences (ZAAS), Hangzhou 310021, China,*
[4]*Department of Plant Sciences, University of California, Davis, CA, USA,*
[5]*Council for Scientific and Industrial Research, Savanna Agricultural Research Institute, Tamale, Ghana,*
[6]*Institut de l'Environnement et de Recherches Agricoles, Saria, Burkina Faso,*
[7]*International Institute of Tropical Agriculture, Kano, Nigeria,*
[8]*Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI, USA,*
[9]*Institut Sénégalais de Recherches Agricoles, Thiès, Senegal,*
[10]*The Bill & Melinda Gates Foundation, Seattle, WA, USA,*
[11]*National Center for Genome Resources, Santa Fe, NM, USA,*
[12]*International Institute of Tropical Agriculture, Ibadan, Nigeria,*
[13]*USDA-ARS Western Regional Research Center, Albany, CA, USA,*
[14]*Department of Nematology, University of California, Riverside, CA, USA,*
[15]*Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA,*
[16]*Illumina, Inc., San Francisco, CA, USA,*
[17]*Department of Biology, University of Virginia, Charlottesville, VA, USA,*
[18]*Agriculture and Agri-Food Canada, Morden, MB, Canada, and*
[19]*USDA-ARS Plant Genetic Resources Conservation Unit, Griffin, GA, USA*

Received 9 June 2016; revised 16 October 2016; accepted 18 October 2016.
*For correspondence (e-mails maria.munoz-amatriain@ucr.edu; timothy.close@ucr.edu).
†These authors contributed equally.

SUMMARY

Cowpea (*Vigna unguiculata* L. Walp.) is a legume crop that is resilient to hot and drought-prone climates, and a primary source of protein in sub-Saharan Africa and other parts of the developing world. However, genome resources for cowpea have lagged behind most other major crops. Here we describe foundational genome resources and their application to the analysis of germplasm currently in use in West African breeding programs. Resources developed from the African cultivar IT97K-499-35 include a whole-genome shotgun (WGS) assembly, a bacterial artificial chromosome (BAC) physical map, and assembled sequences from 4355 BACs. These resources and WGS sequences of an additional 36 diverse cowpea accessions supported the development of a genotyping assay for 51 128 SNPs, which was then applied to five bi-parental RIL populations to produce a consensus genetic map containing 37 372 SNPs. This genetic map enabled the anchoring of 100 Mb of WGS and 420 Mb of BAC sequences, an exploration of genetic diversity along each linkage group, and clarification of macrosynteny between cowpea and common bean. The SNP assay enabled a diversity analysis of materials from West African breeding programs. Two major subpopulations exist within those materials, one of which has significant parentage from South and East Africa and more diversity. There are genomic regions of high differentiation between subpopulations, one of which coincides with a cluster of nodulin genes. The new resources and knowledge help to define goals and accelerate the breeding of improved varieties to address food security issues related to limited-input small-holder farming and climate stress.

Keywords: BAC sequencing, consensus genetic map, cowpea, genetic anchoring, iSelect genotyping array, *Phaseolus vulgaris* L., synteny, *Vigna unguiculata* L. Walp., West Africa, WGS sequencing.

INTRODUCTION

Cowpea (*Vigna unguiculata* (L.) Walp.), native to Africa and a member of the Fabaceae family, is a primary source of protein in sub-Saharan Africa, where it is grown for fresh and dry grains, foliage, and forage. Cowpea is also an important crop in parts of Asia, South America, and the USA (Singh, 2014). Because of its adaptability to harsh conditions, cowpea is a successful crop in arid and semi-arid regions where few other crops perform well. Cowpea is important to the nutrition and income of smallholder farmers in Africa, while also contributing to sustainability of the cropping system through fixation of atmospheric nitrogen and prevention of soil erosion. Despite its relevance to agriculture in the developing world and its stress resilience, actual yields of cowpea are much lower than the known yield potential, and cowpea genome resources have lagged behind those developed for other major crop plants.

Cowpea is a diploid with a chromosome number 2n = 22 and an estimated genome size of 620 Mb (Chen *et al.*, 2007). Its genome shares a high degree of collinearity with other warm season legumes, especially common bean (*Phaseolus vulgaris* L.) (Vasconcelos *et al.*, 2015). Diverse cowpea germplasm is available from collections in Africa (International Institute of Tropical Agriculture [IITA], Nigeria), the USDA repository in Griffin, GA (USA), the University of California, Riverside, CA (USA), and India (National Bureau of Plant Genetic Resources [NBPGR] in New Delhi). These collections contain diversity relevant to pests, pathogens, plant architecture, seed characteristics and adaptation to marginal environments. Resources that were developed previously to support adoption of markers for breeding include a 1536-SNP GoldenGate assay (Muchero *et al.*, 2009), which has enabled linkage mapping and QTL analysis (e.g. Lucas *et al.*, 2011; Muchero *et al.*, 2013; Pottorff *et al.*, 2014) as well as an assessment of the diversity of landraces throughout Africa (Huynh *et al.*, 2013).

IT97K-499-35, developed at IITA, was released in Nigeria in 2008 as a line that is resistant to most races of the parasitic weed *Striga gesnerioides* that are prevalent in West Africa. This black-eyed variety has also been released as a cultivar in Mali and Ghana under the names 'Djiguiya' and 'Songotra', respectively. Gene-space sequences accounting for approximately 160 Mb of the IT97K-499-35 genome were previously published (Timko *et al.*, 2008). In addition, 29 728 'unigene' consensus sequences, derived from 183 118 ESTs from cDNA libraries of 17 different cowpea accessions are available in the software HarvEST:Cowpea (harvest.ucr.edu) (Muchero *et al.*, 2009).

Here we present additional resources from IT97K-499-35 including sequence assemblies from 65× coverage whole-genome shotgun (WGS) short reads and minimal tiling path (MTP) BACs, a BAC physical map, more than 1 million SNPs discovered from sequences of 36 diverse accessions, and an Illumina Cowpea iSelect Consortium Array which represents a publicly accessible resource for screening 51 128 SNPs. These genomic resources do not constitute a complete sequence of the cowpea genome, yet they have been sufficient to support linkage mapping, synteny analysis, and evaluation of materials currently in use from four West African breeding programs, which serve one of the most food insecure regions of the world.

RESULTS

Whole-genome shotgun sequencing and assembly

A WGS approach using short-read sequencing was followed to assemble sequences of the cowpea genome. WGS data from cowpea accession IT97K-499-35 included 394 million paired-end short reads for a total of 40.6 Gb of sequence data (approximately 65× coverage) from Illumina GAII, and Illumina HiSeq sequences from one 5 kb long-insert paired-end (LIPE) library. These two datasets were assembled using SOAPdenovo (Luo *et al.*, 2012) together with the Sanger BAC-end sequences (BES) described below and the 'gene-space' sequences available from Timko *et al.* (2008). The resulting assembly has over 600 000 scaffolds (97 777 of 1 kb or longer), accounting for 323 Mb of the cowpea genome (724 Mb of total scaffold length including Ns; Table S1). This highly-fragmented assembly reflects the short length of the reads and the expected highly-repetitive genome; its close relatives common bean (Schmutz *et al.*, 2014) and adzuki bean (Yang *et al.*, 2015) are approximately 45% repetitive. Despite the fragmentation, the assembly yielded high BLAST hits to 97.2% of the available EST-derived 'unigene' consensus sequences available from HarvEST:Cowpea (http://harvest.ucr.edu). This may be an underestimate of the representation of genes in IT97K-499-35 because the 17 cowpea accessions used for the EST libraries may contain genes not present in IT97K-499-35. The WGS assembly also produced BLAST hits to 24 712 common bean gene models, which is 90.9% of the total number of predicted protein-coding loci (Schmutz *et al.*, 2014). The average GC content

of the WGS assembly was 35.96%, similar to other sequenced legumes (Varshney *et al.*, 2012; Schmutz *et al.*, 2014; Yang *et al.*, 2015).

Physical map and BAC sequencing

Two BAC libraries were constructed from IT97K-499-35 using restriction enzymes *Hind*III and *Mbo*I (36 864 clones each with 150 and 130 kb average clone insert size, respectively). High-quality BES were generated from 30 343 BACs using the Sanger method. BES had an average read length of 674 bp, a GC content of 37.2%, and accounted for 20.5 Mb. They were included in the WGS assembly described above. For physical mapping, 59 408 BACs (97.9% from *Hind*III and 63.2% from *Mbo*I) were fingerprinted using the method of Luo *et al.* (2003). After quality filtering, 43 717 clones were assembled into 829 contigs (40 952 BACs) and 2765 singletons using FPC (Soderlund *et al.*, 2000). The total number of fingerprints in the physical map represents an equivalent of 11-fold haploid genome coverage. The resulting cowpea physical map is available at http://phymap.ucdavis.edu/cowpea.

In total, 4355 MTP clones were sequenced in combinatorial pools (Lonardi *et al.*, 2013) using Illumina HiSeq2000. Reads were assigned to individual BACs and then assembled using SPAdes (Bankevich *et al.*, 2012). BAC assemblies had an average N50 of 18.5 kb, an average L50 of 5.7 contigs, and a total length of 496.9 Mb (Table S2). The GC content was 34.05%. Analysis of overlap between sequenced BACs provided an estimate of non-redundant genome coverage at 372.8 Mb (approximately 60.1% of the cowpea genome; see Experimental Procedures for more details). Sequence comparison revealed that the BAC assemblies contain 17 216 (57.9%) of 29 728 cowpea EST-derived cowpea unigenes (http://harvest.ucr.edu). In addition, the BAC sequences had high homology with 15 617 (57.4%) of the 27 197 protein-coding gene models in common bean (Schmutz *et al.*, 2014). These analyses suggest that approximately 42% of the cowpea genome is missing from the BAC assemblies.

Development of the Cowpea iSelect Consortium Array

In total, 36 additional cowpea accessions relevant to Africa, China and the USA were shotgun sequenced using Illumina HiSeq 2500 (12.5× average coverage) and aligned to the WGS assembly of IT97K-499-35 to discover SNPs. These accessions were chosen to represent the geographic, phenotypic and genetic diversity of cultivated cowpea (Figure S1 and Table S3). An additional set of 12.5× HiSeq data was also produced from IT97K-499-35 and included as a control against spurious SNP calls. The reads were mapped to the reference sequence (Data S1) using BWA (Li and Durbin, 2009) to generate a .bam file. Then, SAMtools (Li *et al.*, 2009), SGSautoSNP (Lorenc *et al.*, 2012) and FreeBayes (Garrison and Marth, 2012) were used to generate three overlapping sets of candidate SNPs, from which the intersection yielded about 1 million SNPs. No accession contributed substantially more SNPs than any other, highlighting the broad coverage of diversity within the set of accessions used for SNP discovery. The most distant accession from IT97K-499-35 was UCR 779 (differing at 25% of SNP loci) followed closely by the four Chinese accessions (22–24%). The accessions most closely related to IT97K-499-35 were the IITA breeding lines IT89KD-288, IT93K-503-1 and IT84S-2246 (12–13%). The set of approximately 1 million SNPs was filtered to 55 496 SNPs for the design of an Illumina iSelect Consortium Array (see Experimental Procedures for details on filtering criteria). The design also included 1163 SNPs from the prior GoldenGate assay (Muchero *et al.*, 2009) and 60 presumed organelle SNPs, for a total of 56 719 intended SNPs (60 000 assays). From those, 51 128 SNPs (90.1%) were represented in the final product manifest (Data S2). The Cowpea iSelect Consortium Array is available from Illumina (Illumina Inc., San Diego, CA, USA; http://www.illumina.com/areas-of-interest/agrigenomics/consortia.html).

Construction of a consensus genetic map for cowpea

Five bi-parental RIL populations were used to develop a consensus genetic map (Table S4). Monomorphic SNPs and those with an excessive number of missing and/or heterozygous calls were eliminated, as well as individuals that were duplicated or highly heterozygous. The number of lines per population used for mapping ranged from 94 to 135 (Table S4) for a total of 575 RILs. A genetic map was constructed using MSTmap (Wu *et al.*, 2008; http://mstmap.org/) at LOD 10 for each RIL population. Linkage groups (LGs) were numbered and oriented based on a previous cowpea consensus map (Lucas *et al.*, 2011). Individual maps and the genotype data used for their construction can be found in Data S3. Two maps (Sanzi × Vita7 and CB27 × IT82E-18) each had two chromosomes separated into two LGs (Table S4 and Data S3) due to regions where parents lack polymorphisms. One region of identity between CB27 and IT82E-18 on LG4 impacted the number of polymorphisms and marker bins, and the total size of that LG in the specific genetic map (Table 1). Genetic map sizes varied among the five populations, from 803.4 cM in ZN016× Zhijiang282 to 917.1 cM in Sanzi × Vita7 (Table 1).

Individual maps were merged into a consensus map using MergeMap (Wu *et al.*, 2011; http://mergemap.org/). Equal weight was given to each individual map. MergeMap identified a few conflicts in marker order, which were resolved by deleting a few conflicted markers with priority given to the map with the highest resolution in the particular LG (i.e. more bins). No SNP was placed on different LGs between maps. As MergeMap's coordinate calculations for a consensus map are inflated relative to cM

Table 1 Distribution of SNPs in the individual component maps and the consensus map

Genetic map	Characteristic	LG1	LG2	LG3	LG4	LG5	LG6	LG7	LG8	LG9	LG10	LG11	All
Tvu-14676 × IT84S-2246-4	Markers	1563	1447	2765	1207	1618	1326	953	1250	691	774	1066	14 660
	Bins	126	105	186	111	131	119	82	126	82	67	81	1216
	cM	84.50	56.85	121.29	66.96	67.14	70.46	68.74	94.18	70.41	56.57	55.80	812.90
Sanzi × Vita7	Markers	1156	1920	2912	1256	1050	973	1376	1343	1389	851	1393	15 619
	Bins	92	157	215	115	112	85	77	116	110	83	103	1265
	cM	59.44	101.01	156.40	77.59	85.03	54.36	50.97	100.46	80.51	66.37	84.95	917.10
ZN016 × Zhijiang282	Markers	800	426	1551	857	562	690	580	791	393	456	858	7964
	Bins	70	49	123	55	56	60	57	83	52	43	49	697
	cM	91.70	54.48	124.01	71.49	75.47	62.88	65.27	94.07	62.28	56.64	45.09	803.38
CB46 × IT93K-503-1	Markers	1342	2050	2745	1374	1336	1151	1138	1759	1227	1601	855	16 578
	Bins	109	109	179	88	91	74	92	116	76	83	66	1083
	cM	94.54	84.46	132.85	73.09	67.26	57.64	67.87	85.50	59.77	62.87	56.00	841.84
CB27 × IT82E-18	Markers	1520	1295	2640	737	1255	1539	1534	1195	1230	1778	1843	16 566
	Bins	100	100	159	37	105	62	68	109	75	69	93	977
	cM	81.57	78.47	132.64	24.45	90.07	60.01	46.54	86.69	68.30	59.76	82.06	810.56
Consensus	Markers	3317	3877	6183	3050	2790	3125	2597	3350	2993	3063	3027	37 372
	Bins	311	350	544	264	295	248	241	346	241	203	237	3280
	cM	82.40	75.04	133.27	62.82	77.00	60.96	59.74	92.29	68.21	60.49	64.89	837.11

distances in individual maps, consensus LG lengths were normalized to the mean cM length from the individual maps. The resulting consensus map contains 37 372 SNP loci mapped to 3280 bins (Table 1 and Data S4). This is a 34-fold increase in marker density and a four-fold increase in resolution (number of bins) over the consensus map of Lucas *et al.* (2011). The new consensus map includes 757 SNPs that were included in the prior GoldenGate assay (Muchero *et al.*, 2009). The map spans 837.11 cM at an average density of one bin per 0.26 cM and 11.4 SNPs per bin. The new consensus map has dense coverage of all 11 cowpea LGs, with 1.85 cM on LG1 being the largest gap (Figure S2 and Data S4).

Syntenic relationships between cowpea and common bean

Similar to cowpea, common bean is a diploid member of the Phaseoleae tribe with 2n = 22 chromosomes. The iSelect SNP design sequences were compared to *P. vulgaris* gene models (Schmutz *et al.*, 2014) to clarify the syntenic relationships of cowpea with this closely related species. The 26 550 SNPs that were mapped in *V. unguiculata* and matched a *P. vulgaris* gene model provided a view of synteny (Figure 1). Six cowpea LGs (VuLG2, VuLG6, VuLG8, VuLG9, VuLG10 and VuLG11) are largely collinear with six common bean pseudomolecules (Pv7, Pv6, Pv9, Pv11, Pv10 and Pv4, respectively), while the rest have synteny mainly with two common bean pseudomolecules (Figure 1 and Table S5). From these five cowpea LGs with one-to-two relationships, three (VuLG3, VuLG4, and VuLG7) have a higher number of links, and over a longer genome interval, with one *P. vulgaris* chromosome (Pv3, Pv1 and Pv2, respectively; Figure 1 and Table S5). The other two cowpea LGs, VuLG1 and VuLG5, both have their largest block of homologous synteny with Pv8, followed by Pv5 and Pv1, respectively (Figure 1 and Table S5).

The same numbering scheme for common bean and cowpea chromosomes would facilitate comparative studies between the two species. Adoption of the chromosome numbers of *P. vulgaris* according to synteny relationships with LGs of cowpea seems sensible, but additional cowpea sequence information will be needed to clarify the relationships between VuLG1 and VuLG5 with Pv1, Pv5 and Pv8. The BAC-FISH analysis by Iwata-Otsubo *et al.* (2016) that correlates the genetic and chromosome maps in cowpea can be used to orient the cowpea genetic map so that it meets the convention of displaying the short arm on top.

Genetic anchoring of WGS scaffolds and BACs

The 37 372-SNP consensus map was used to anchor WGS and BAC assemblies to genetic map positions. The iSelect SNP design sequences were used as BLAST queries to search against WGS and BAC sequences, and matches with an *e*-value = $1e^{-50}$ or better were tallied. Assembled

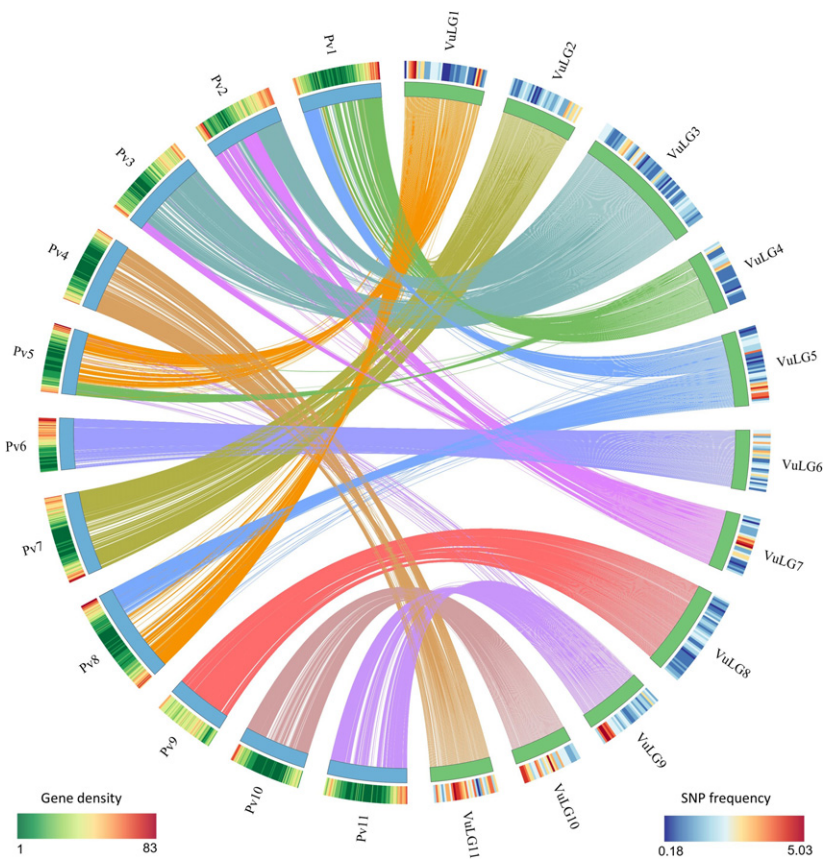

Figure 1. Circos illustration of synteny between cowpea linkage groups (VuLG) and common bean pseudomolecules (Pv).
SNP frequencies calculated for 2 cM windows and normalized to the total anchored scaffold size (in kb) are shown for the 11 cowpea LGs. *Phaseolus vulgaris* gene densities are also shown for 500 kb windows.

sequences were considered anchored to the genetic map if 100% of the matching SNPs mapped to the same LG, and were at most 5 cM apart (Data S5 and S6). The anchored sequences contain 100 Mb of the WGS assembly (237 Mb scaffold size including Ns; Table S1 and Data S5) and 420 Mb of BAC assemblies (Table S2 and Data S6). For BACs, this is an overestimate of the actual genome coverage because BAC sequences have approximately 23% overlap (see Physical map and BAC sequencing), resulting in a reduced estimate of 323 Mb of unique sequences within anchored BACs. Also, observe in Table S1 that 95.3% of the anchored WGS scaffolds are larger than 1 kb and they comprise 99.1% of the anchored non-N sequence. Thus, the anchored portion of the WGS assembly, which is comprised mainly of 24 342 scaffolds larger than 1 kb among 25 537 anchored scaffolds, contains many fewer fragments than the entire WGS assembly (644 126 scaffolds). This is the outcome of having selected SNPs in the largest WGS contigs as a final criterion in SNP selection (see Experimental Procedures).

Distribution of genetic variation

The anchoring of WGS scaffolds to the genetic map enabled investigation of the frequency and positional distribution of genetic diversity in the cowpea genetic map. Nearly half of the 1 036 981 SNPs discovered from the 37 diverse cowpea accessions were anchored to the genetic map based on the anchoring of 25 537 WGS scaffolds using mapped iSelect SNPs. This information was used to examine the SNP frequency and distribution across the 11 cowpea LGs. Frequencies were calculated for 2 cM intervals and normalized to the total anchored scaffold size. SNP frequencies were not uniformly distributed across the genetic map (Figure 1). LG11 and LG10 had significantly higher SNP frequencies than all other cowpea linkage groups. Relatively higher SNP frequencies were also observed in the distal ends of LG5 and LG9, in the centromeric region of LG7, and toward the ends of LG1. In contrast, LG8 had relatively low SNP diversity (Figure 1). There was no clear relationship between the most diverse cowpea genomic regions and the gene-dense syntenic regions of common bean (Figure 1).

Genetic diversity and structure in West African breeding programs

Evaluation of genetic diversity has important implications for breeding programs and the conservation of genetic resources. A total of 146 West African cultivated cowpea accessions were evaluated using the Cowpea iSelect Consortium Array. This included 105 cultivars and breeding lines from the breeding programs of IITA (Nigeria), INERA (Institut de l'Environnement et de Recherches Agricoles,

Burkina Faso), ISRA (Institut Senegalais de Recherches Agricoles, Senegal), and CSIR-SARI (Council for Scientific and Industrial Research, Savanna Agricultural Research Institute, Ghana), and 41 landraces collected from these same countries (Table S6). It should be noted that these landraces were chosen by the different breeding programs and may not represent the full range of genetic diversity available in the West African landrace germplasm.

STRUCTURE analysis and principal component analysis (PCA) were performed to evaluate population structure and to clarify the genetic relationships between accessions. STRUCTURE (Pritchard *et al.*, 2000) was run for $K = 1$–6 and, although the estimated log probabilities of the data reached a plateau at $K = 5$ (Figure 2a), at that level of population subdivision there were individuals not strongly assigned to one subpopulation or another (Figure S3). When applying the Evanno *et al.* (2005) method, the maximum ΔK value was reached at $K = 2$ (Figure 2a), which would be consistent with two major subpopulations. PCA showed a clear separation of the two subpopulations on the first component (PC1; Figure 2b), which were not differentiated by breeding program or by improvement status (Figure 2c,d and Table S6). The 45 accessions belonging to subpopulation 1 (i.e. ancestry ≥0.8; Table S6) included 23 landraces from the four countries and 22 breeding accessions. From the 44 accessions belonging to subpopulation 2, 14 were landraces (mostly from Senegal; Table S6) while the remaining 30 were either IITA breeding lines or lines from other programs derived from IITA lines. Pedigree history that was available from IITA revealed that members of subpopulation 2 contain South and East Africa parentage whereas subpopulation 1 parentages are restricted to West Africa. All admixed accessions but one (59–30) are cultivars and breeding lines. PCA also shows that the four West African breeding programs are working with very similar materials, except for somewhat narrower diversity within the Ghana program (Figure 2c). Landraces were less dispersed than cultivars and breeding lines, mostly distributed along the first component (Figure 2d). Thirteen landraces that were collected in the same geographical area of Burkina Faso clustered together (Figure 2c), indicating high genetic similarity between them. Fixation index (F_{ST}) values were calculated between the two major subpopulations and between landraces and cultivars/breeding lines. The F_{ST} value for subpopulations 1 and 2 was 0.18, indicating moderate population differentiation. Little genetic differentiation was found for landraces vs. breeding materials ($F_{ST} = 0.02$), in accordance with STRUCTURE

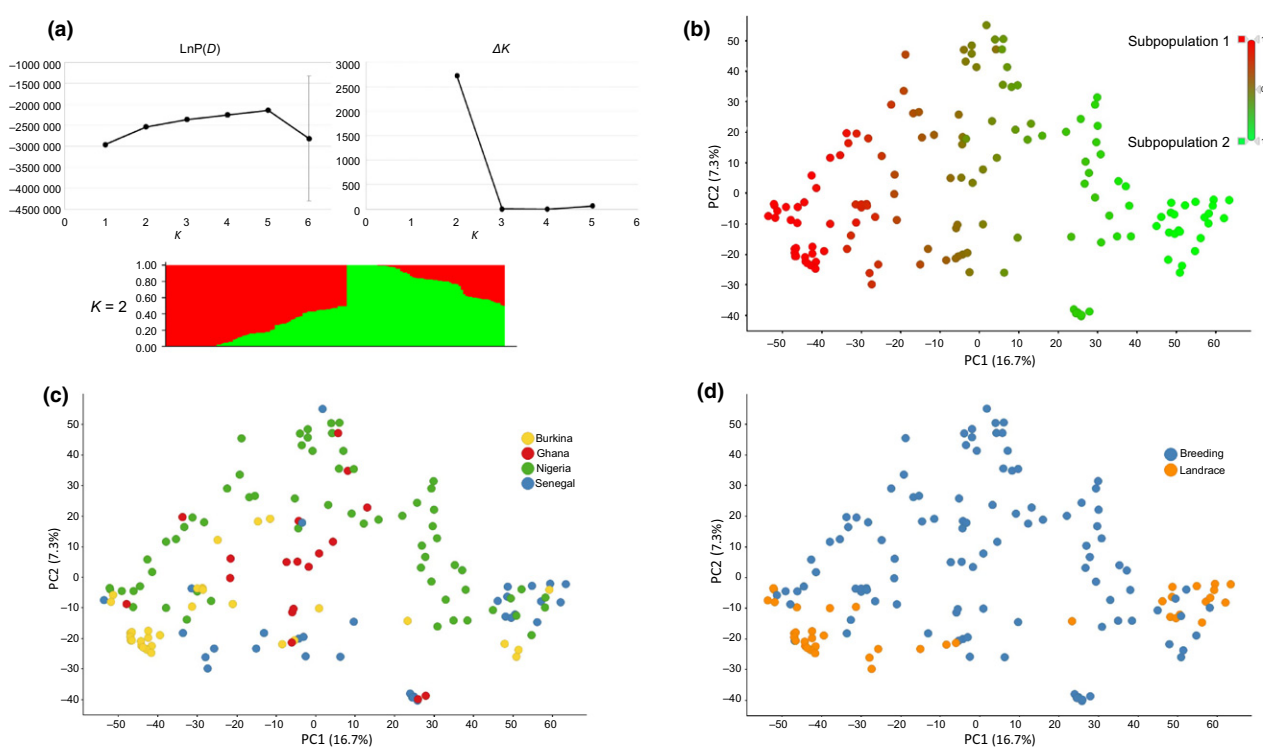

Figure 2. Population structure analysis of 146 cultivated cowpea accessions from West Africa.
(a) The plot on the left displays the log probability of the data for each K between 1 and 6, while the plot on the right shows ΔK values calculated as proposed by Evanno *et al.* (2005). The plot of ancestry estimates for $K = 2$ is shown in the bottom, where each individual is shown as a vertical bar.
(b–d) Principal component analysis of all cowpea accessions colored by the result of STRUCTURE (b), by breeding program (c), and by their improvement status (breeding vs. landrace; plot (d)).

and PCA results. F_{ST} values between subpopulations, and between landraces and breeding materials at SNPs across the genome are provided in Data S7.

Polymorphism information content (PIC), expected heterozygosity (*He*) and nucleotide diversity (π) were calculated for the entire set of West African accessions, for each of the two major subpopulations, and for landraces and breeding materials. Average PIC was 0.247, while *He* and π averaged 0.307 and 0.308, respectively, when considering the whole dataset. Average values for all three diversity measures were higher in subpopulation 2 than in subpopulation 1: average PIC, *He* and π were 0.158, 0.193 and 0.195, respectively, in subpopulation 1, while they were 0.229, 0.284 and 0.288 in subpopulation 2. Breeding lines had slightly higher PIC, *He* and π values than landraces, being 0.242, 0.301 and 0.303, respectively, in breeding materials, while they were 0.234, 0.290 and 0.293 in landraces. However, given the small sample size of local landraces and the fact that they were biased toward the interests of breeders, this may be an inaccurate estimate of the diversity in West African landrace germplasm. Since PIC, *He*, and π are highly correlated, only expected heterozygosity (*He*) values for each subpopulation, and for landraces and breeding lines are shown at each SNP in Data S7.

He values for subpopulation 1 and 2 were plotted to explore the spatial patterns of diversity across the 11 LGs (Figures 3 and S4, upper plots). F_{ST} values were also plotted across the genome (Figures 3 and S4, lower plots). The greater diversity within subpopulation 2 is apparent throughout most of the genome (Figures 3 and S4), with some exceptions. An extreme example of an exception is a region extending from 30 to 35 cM on LG7, where diversity is very low in subpopulation 2 (Figure 3). In addition, in regions where diversity is low in one subpopulation, it tends to be moderate to high in the other subpopulation. One exception to this latter trend is near 63 cM on LG1, where both subpopulations have very low diversity and contain the same alleles (low F_{ST}; Figure 3). This region coincides with a QTL for pod length (Xu *et al.*, 2016). Another exception is on LG3 (approximately 82–85 cM; Figure S4), in a region coinciding with a QTL for heat tolerance (Lucas *et al.*, 2013). These plots also revealed regions of very high population differentiation (F_{ST}) on LG 4, 7, and 8 (Figures 3 and S4). The smallest of these regions (LG8 at 53 cM) contains seven SNPs. The design sequences for six of them are contained within two sequenced BACs (H084G18 and M006L23) which contain many sequences related to *P. vulgaris* nodulin gene models (Phvul.009G135300.1 and Phvul.009G135400.1). This suggests the presence of a cluster of nodulin genes in this region. The number of genes that could be relevant in the larger regions of LG 4 and LG 7 is too large to be considered in detail.

HarvEST:Web and HarvEST:Cowpea allow easy access to available cowpea genome resources

The new cowpea genome resources must be easily accessed if they are to be widely utilized for basic research and agricultural development. In addition to all sequence data being deposited in permanent, public repositories at the National Center for Biotechnology Information (NCBI; see Accession numbers), information presented in this manuscript is available through HarvEST:Web (http://harvest-web.org/) or in the Windows software HarvEST:Cowpea (download from http://harvest.ucr.edu). WGS and BAC sequences, and their annotations can be retrieved in HarvEST:Web by specifying 'scaffold name' or 'BAC address,' respectively. These sequences can be searched by BLAST via http://www.harvest-blast.org. SNP names can also be used as inputs for sequence and annotation retrieval. In addition, a synteny viewer has been implemented in HarvEST:Cowpea, enabling facile comparisons between cowpea and either common bean, soybean or Arabidopsis. Macrosynteny and microsynteny are clearly evident between cowpea and the two closely related warm season legumes.

DISCUSSION

Increase in climate variability is projected to have the greatest negative consequences on agricultural and human systems in the tropical and subtropical developing world, aggravating food insecurity in already vulnerable populations (Thornton *et al.*, 2014). Cowpea is a relatively drought and heat-tolerant crop that provides protein to nearly 200 million Africans and cash income to smallholder farmers (Thomson, 2008). The limited availability of genome resources for cowpea has contributed to the relatively slow development of higher yielding varieties adapted to tolerate abiotic and biotic stresses. This report presents 323 Mb of WGS and 497 Mb of BAC sequence information, a tool to simultaneously test 51 128 single nucleotide variants, and a high-density genetic map providing coordinates for most of those sequences and variants. Application of these resources can be made for genome-wide association studies (GWAS) of cowpea germplasm to discover favorable alleles for simple and complex traits, as is already being conducted in other legume crops (e.g. Kujur *et al.*, 2015; Ray *et al.*, 2015). Useful variation can then be connected to assembled genome sequences–including BACs–annotated for *P. vulgaris* syntenic gene models, thereby increasing the precision and speed of cowpea improvement.

One of the biggest obstacles in comparing and using results obtained by different research groups is the lack of a common nomenclature for cowpea linkage groups. With a high SNP coverage of the genome and connections to cowpea genome sequences, this study provides the basis for a unified chromosome nomenclature for the cowpea

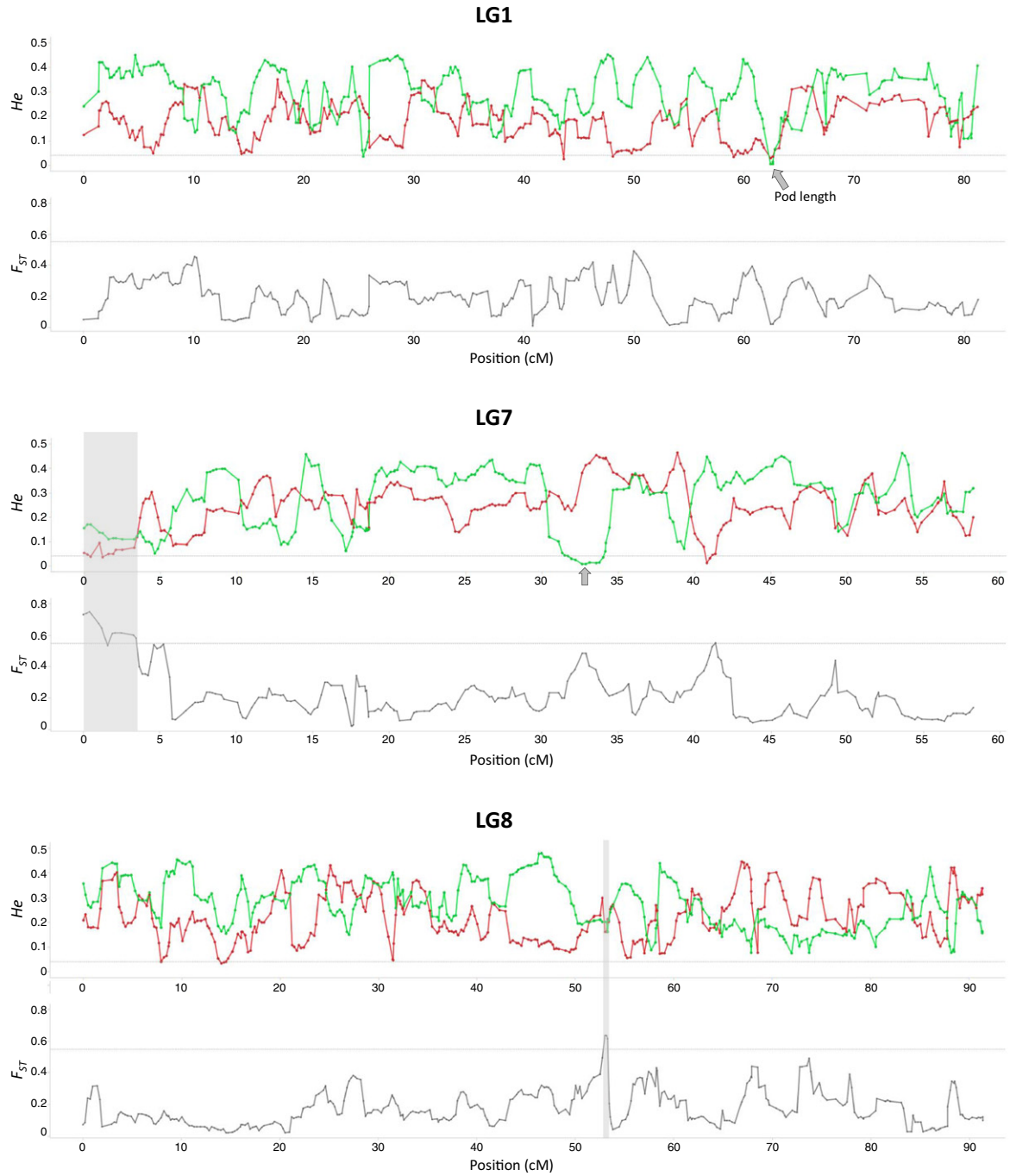

Figure 3. Genetic diversity and population differentiation across linkage groups 1, 7, and 8.
Upper plots show expected heterozygosities (*He*) for subpopulations 1 (red line) and 2 (green line), while lower plots show genetic differentiation (F_{ST}) between the two subpopulations. *He* and F_{ST} values were averaged across a sliding window of 5 genetic bins with a step of one bin. The dashed lines indicate the bottom and top 1% of *He* and F_{ST} values, respectively. Arrows indicate regions with a markedly depletion of genetic diversity in one or both subpopulations, while shadowed areas indicate genomic regions of very high genetic differentiation (F_{ST}).

research community. Such common nomenclature could adopt the *P. vulgaris* chromosome numbering on the basis of synteny comparisons between both species as well as cytogenetic studies in cowpea (Iwata-Otsubo *et al.*, 2016) and between cowpea and common bean (Vasconcelos *et al.*, 2015). While several cowpea LGs are largely syntenic with one *P. vulgaris* chromosome, further resolution is needed to satisfy a single nomenclature for those LGs whose syntenic relationships with common bean are less clear. The goal would be to extend a standard

chromosome numbering to other diploid *Vigna* species whose genomes have been sequenced and are integrated into a genome database (Sakai *et al.*, 2016). This would facilitate the transfer of genomic information on target traits from one Fabaceae species to another.

West Africa is the region with the largest production and consumption of cowpea in the world (FAOSTAT, 2012; Singh, 2014). Evaluating the genetic diversity present in the West African breeding germplasm is important to manage breeding programs and assure future genetic gains. By applying the Cowpea iSelect Consortium Array to 146 breeding lines and landraces, we have provided a useful overview of genetic variability in West African cultivated germplasm. Two subpopulations were found in the evaluated materials, which seem to coincide with the two major African gene pools (GP1–West, North and Central Africa; GP 2–East, South and Southeast Africa; Huynh *et al.*, 2013). It is unknown why many landraces from West Africa belong to this subpopulation. One can speculate that different subsets of the broader germplasm were carried by humans during different waves of migration. Subpopulation 2 is more diverse that subpopulation 1, which may be expected since it contains germplasm from outside West Africa. Since all of these accessions have been adapted to West Africa, the existence of two major subpopulations at the present time means that relatively wide crosses can be made without compromising adaptation. The new genetic knowledge helps guide crossing strategies. The common agro-ecological zones which extend across cowpea production areas of the four included countries of Burkina Faso, Ghana, Nigeria and Senegal facilitates coordination of breeding activities and exchange of germplasm. The IITA breeding program in Nigeria has been a regional distributor of new breeding materials during the last few decades, setting an excellent precedent which can now be revitalized and expanded using genome knowledge.

Average diversity values for entire genomes should be interpreted cautiously because patterns of diversity vary across LGs. In fact, although the overall genetic diversity within the West African breeding population is relatively high (*He* and $\pi = 0.31$), we identified genomic regions of diversity depletion. Those regions may contain favorable alleles for important traits that became fixed during domestication and breeding selection. The lowest *He* values in LG1 coincide with the position of SNPs associated with pod length in Chinese germplasm of *V. unguiculata* subspecies *sesquipedalis* (Xu *et al.*, 2016). One interpretation could be that there has been selection for a preferred pod length in these materials. Also, a previously reported QTL for heat tolerance (*Cht-5*) coincides with a low-diversity region of LG3 (Lucas *et al.*, 2013). Favorable alleles at this QTL were donated by the line IT82E-18, the African parent of the RIL population (Lucas *et al.*, 2013; Table 1). The low diversity in this region of LG3 may reflect selection for better yield performance of West African cowpeas under higher growing season temperatures. There are several genome regions where F_{ST} is much higher than the genome-wide average, indicating high genetic differentiation between subpopulations. Interestingly, a cluster of nodulins was annotated in BACs located in one of these regions. As nodulins play a key role in the establishment of symbiosis with *Rhizobium* bacteria (Legocki and Verma, 1980), perhaps different nodulin alleles are correlated with different rhizobial symbionts for the two subpopulations. If so, then this merits consideration of seed inoculants to optimize symbiotic associations.

The *He* and F_{ST} values for each SNP (Data S7) comprise another valuable resource stemming from this work. They are shown for the two subpopulations, and for landraces and breeding materials, providing breeders with a useful resource to increase the genetic diversity in their breeding programs or to incorporate unique alleles into their breeding populations. Also, *He* values can be used as criteria for selecting efficient subsets of markers for conversion to other platforms. Customized, maximally informative subsets of markers have numerous applications including routine tests of seed purity, validation of germplasm fidelity, verification of successful crosses and guidance of progeny selection in later generations during trait introgression into preferred backgrounds via backcrossing.

EXPERIMENTAL PROCEDURES

Physical mapping and BAC-end sequencing

Cowpea accession IT97K-499-35 was grown for three generations by single seed descent and then increased to provide a supply of seed for DNA isolation. The material was screened with the Illumina GoldenGate assay (Muchero *et al.*, 2009), establishing that homozygosity was attained. Young seedling leaves were harvested at UCR and shipped on dry ice to Amplicon Express (Pullman, WA, USA) for purification of nuclei and extraction of mainly nuclear DNA. Two BAC libraries were then constructed by Amplicon Express from high molecular weight DNA using restriction enzymes *Hin*dIII and *Mbo*I. After partial digestion with restriction enzymes, high MW cowpea DNA fragments were ligated with *Hin*dIII or *Bam*HI linearized BAC vector pCC1. Ligated DNA molecules were introduced into *Escherichia coli* DH10B cells by electroporation and plated on LB agar containing 12.5 μg/ml chloramphenicol, 0.5 mM IPTG and 40 μg ml^{-1} X-Gal and cultured overnight. White colonies were picked and inoculated into 384-well plates containing LB freezing buffer. Cultures were incubated at 37°C for 24 h with aeration, and then stored at −80°C. The libraries contained 36 864 clones each, with average insert sizes of 150 kb for the *Hin*dIII library and 130 kb for the *Mbo*I library.

BAC clones from the two libraries (36 096 from *Hin*dIII and 23 312 from *Mbo*I) were fingerprinted using the SNaPshot-based fingerprinting procedure (Luo *et al.*, 2003). BAC DNAs were simultaneously digested with five restriction enzymes (*Bam*H1, *Eco*RI, *Xba*I, *Xho*I, and *Hae*III), and then labeled with the SNaPshot labeling kit (Luo *et al.*, 2003). The fragments were sized on an ABI3730XL instrument with the GS1200Liz size-standard (Gu *et al.*, 2009). Fragment sizes in the range of 100–1000 bp were compiled

for computational assembly. After removing substandard fingerprints, potential cross contamination and clones with less than 40 total fragments, fingerprints from 43 717 clones (73.6%) were used for an initial contig assembly using the FPC software (Soderlund *et al.*, 2000). This initial assembly was performed with a relatively high stringency (1×10^{-45}) to minimize co-assembly of clones from unrelated regions of the genome. The 'DQer' function of the FPC software was used for second stage assembly by disassembling contigs containing more than 15% questionable clones. The 'Single-to-End' and 'End-to-End' merging function of FPC was used for a final, third stage assembly by stepwise decreases of assembly stringency based on Sulston score cutoff values (down to 1×10^{-35}). Finally, the 10% largest contigs were subjected to manual editing, examining with CB map analysis and disjoining contigs with CB analysis results at 1×10^{-30}.

The same BAC DNA used for fingerprinting was also used for BES. BAC clones were sequenced using pIndigoBAC5 Reverse End Sequencing primer (5′-TACGCCAAGCTATTTAGGTGAGA-3′) and BigDye terminator chemistry (Applied Biosystems, Foster City, CA, USA) on an ABI3730XL automated sequencer (Applied Biosystems, Foster City, CA, USA). Raw sequence reads were trimmed with the Phred program using a quality score of 20 (Ewing and Green, 1998). BES from vector sequences, *E. coli*, mitochondria and chloroplasts were identified using BLASTN. The chloroplast sequences of common bean (DQ886273.1), soybean (DQ317523), *Medicago truncatula* (AC093544), *Lotus japonicus* (AP002983), and mitochondrial DNA sequences of Arabidopsis (Y08501.2) and rice (DQ167399.1) were used to identify organelle contaminations. The resulting high-quality BES were then processed with the RepeatMasker program (www.repeatmasker.org) to identify characterized repeats. Cowpea BES with more than 80% of the sequence length showing homology to known repeats were removed, otherwise the BES were kept but the repetitive region was marked using letter N. Self-comparisons were conducted with the RepeatMasker processed sequences to further filter the cowpea-specific repeat elements.

MTP sequencing and BAC assembly

A set of MTP BACs was chosen using the FMTP method of Bozdag *et al.* (2013). MTP BACs were paired-end sequenced (2×100 bases) using Illumina HiSeq2000 (Illumina, Inc, San Diego, CA, USA). Sequencing was done in two sets of 2197 BACs (Vu1 and Vu2) applying a combinatorial pooling design (Lonardi *et al.*, 2013). After quality-trimming, reads in each pool were 'sliced' into smaller samples of optimal size, deconvoluted, and then assembled BAC-by-BAC using SPAdes (Bankevich *et al.*, 2012), as explained in detailed by Lonardi *et al.* (2015). From the 4394 intended BACs, 4355 produced sufficient reads to generate an assembly.

To estimate the percentage of overlapping BAC sequences, 19-mers occurring at least four times were identified and used for repeat-masking of sequences. Repeat-masked sequences were then BLASTed against themselves using an *e*-value cutoff of e^{-40}. Only overlapping sequences >300 bp were considered to be overlaps. To estimate the gene content of the BAC assemblies, BAC sequences were compared to cowpea EST-derived 'unigenes' (http://harvest.ucr.edu) and *P. vulgaris* gene models (Schmutz *et al.*, 2014) using BLAST (*e*-value cutoffs of e^{-40} and e^{-25}, respectively).

Whole-genome shotgun sequencing and assembly

The same batch of IT97K-499-35 nuclear DNA that was used for BAC library construction was used for WGS sequencing. About 394 M paired-end reads (equivalent to approximately 65× coverage) with an average read length of approximately 100 bases after quality-trimming were produced at the National Center for Genome Resources (NCGR; Santa Fe, NM, USA) on an Illumina GAII sequencing instrument. An additional approximately 90 M Illumina reads were produced using an Illumina HiSeq sequencing instrument at NCGR from one 5 kb long-insert paired-end (LIPE) library made from the same batch of nuclear DNA.

For the assembly, two additional sets of Sanger sequences were included. One set of Sanger sequences was the basis of a prior publication on 'gene-space sequences' (GSS; Timko *et al.*, 2008), comprised of approximately 250 000 reads from methyl filtered fragments of IT97K-499-35. The other set of Sanger sequences included the BES described above. The assembly combined the paired-end short reads, LIPE, GSS, and BES data using SOAPdenovo with $k = 31$ (Luo *et al.*, 2012). To estimate the gene content of the WGS assembly, sequences were BLASTed against cowpea EST-derived 'unigenes' (http://harvest.ucr.edu) and *P. vulgaris* gene models (Schmutz *et al.*, 2014), using *e*-value cutoffs of e^{-40} and e^{-25}, respectively.

SNP discovery and design of the Cowpea iSelect Consortium Array

A total of 32 accessions were sequenced to 12.5× coverage by the Beijing Genomics Institute (BGI) using Illumina HiSeq 2500 (Illumina, Inc.). Four additional accessions from China (see Table S3) were sequenced at the Majorbio Pharm Technology Co. Ltd (Shanghai, China). Additional sequences of IT97K-499-35 were produced in the Genomics Core Facility at the University of California, Riverside.

The WGS assembly from IT97K-499-35 described above was used as the reference to map each of these 36 sets of reads, and the new set of HiSeq sequences from the reference genotype sequenced at the University of California Riverside. This 37th set was used as a control (i.e. SNPs call in this accession were considered false positives). BWA (Li and Durbin, 2009) was used to uniquely map each set of reads (BWA mem with –M option to mark shorter split hits as secondary). Reads which mapped to multiple locations were excluded from further analysis. Alignment files were merged with the software tool Picard to a single 'sam' file. Reads that 'hanged off' the end of the contigs in the reference sequence were clipped with Picard. Also, to avoid skewed variant calling, duplicated reads were marked with Picard.

To filter putative SNPs to a shorter list of highest confidence variants, three software packages were used, namely SAMtools (Li *et al.*, 2009), SGSautoSNP (Lorenc *et al.*, 2012), and FreeBayes (Garrison and Marth, 2012). It was not possible to utilize GATK (McKenna *et al.*, 2010) because GATK requires a relatively large set of confirmed training SNPs for the base quality score recalibration phase, and no such set of SNPs was available for cowpea. In total, SAMtools discovered 4 629 826 SNPs using mpileup with default parameters, SGSautoSNP detected 2 488 797 SNPs and FreeBayes called a total of 8 269 140 SNPs. An intersection set of SNPs was then identified, leading to 1 036 981 SNPs that were identified by all three methods. Additional filtering was required to reduce the number of SNPs to the target density of 60 000 SNP assays designed as a community resource for future germplasm characterization. These filtering steps included: (i) designability score based upon Illumina's Assay Design Tool; (ii) avoidance of a SNP whose adjacent sequences occurred frequently in the genome assembly; (iii) consideration of allele frequency, generally avoiding SNPs with only one accession carrying the minor allele; (iv) selection of two SNPs in or near each inferred cowpea gene based on MUMmer sequence alignment with *P. vulgaris* gene

models (Schmutz *et al.*, 2014); (v) requirement for a minimum distance from a SNP that had already been selected; (vi) preference against an A/T or C/G SNP since these require two beadtypes (assay space); and (vii) location within a relatively larger WGS contig to maximize the amount of WGS contigs that could subsequently be anchored to a SNP-based genetic map.

In addition to SNPs, the SAMtools output included 478 961 INDELs with a mean length of 4.48 bp. These were not included in the iSelect design.

In addition to SNPs discovered by WGS sequencing of diverse accessions, 1163 SNPs previously validated on the GoldenGate platform (Muchero *et al.*, 2009) were included in the design to facilitate comparisons with prior genotyping research. In total, 56 719 SNPs were submitted for assay design using 60 000 beadtypes. This yielded 51 128 assays in the final manifest for the publicly available Cowpea iSelect Consortium Array.

Consensus genetic map construction

Five bi-parental RIL populations developed previously (Muchero *et al.*, 2009; Lucas *et al.*, 2011) were genotyped with the Cowpea iSelect Consortium Array at the University of Southern California. SNPs were called using the GenomeStudio software (Illumina, Inc.). To meet assumptions of the clustering algorithm, 'synthetic heterozygotes' were constructed and included in the initial set of 96 genotyped samples by creating 1:1 mixtures of DNA samples from individuals known from prior work to be most genetically distant from each other. The data from these individuals provided the signal needed for the algorithm to position clusters for heterozygotes. SNPs with low GenTrain scores were visually inspected based upon manufacturer's published best practice for optimizing accuracy in genotyping projects (http://www.illumina.com/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf). The resulting cluster file is available upon request.

SNP data from each population were exported from Genome Studio and curated to eliminate: (i) monomorphic SNP loci; (ii) SNPs with >20% missing or heterozygous calls; and (iii) segregation-distorted markers (MAF < 0.25). RILs were also curated to remove individuals with >10% heterozygous loci or those carrying many non-parental alleles. Identical individuals were also thinned to one such individual prior to mapping. Genetic maps for each RIL population were constructed at LOD 10 using MSTmap (Wu *et al.*, 2008; http://mstmap.org/). Because the level of residual heterozygosity varied among populations, different population type options were chosen for map construction in MSTmap (RIL 7 for Tvu-14676 × IT84S-2246-4; RIL 6 for Sanzi × Vita7 and ZN016 × Zhijiang282; and RIL5 for CB46 × IT93K-503-1 and CB27 × IT82E-18). Other parameters for MSTmap included: grouping LOD criteria = 10; no mapping size threshold = 2; no mapping distance threshold = 10 cM; try to detect genotyping errors = no; and genetic mapping function = kosambi. Output maps were inspected to identify and remove data that would result in presumably spurious double recombination events, unless supported by several markers or moderate to large genetic distances.

Linkage groups from each population were numbered and oriented based on the previous cowpea consensus map (Lucas *et al.*, 2011) and then merged into a consensus map using MergeMap (Wu *et al.*, 2011; http://mergemap.org/). Equal weight was given to each individual map (weight = 1.0). MergeMap identified a few conflicts in marker order, which were resolved by deleting a few conflicted markers with priority given to the map with the highest resolution in the particular LG (i.e. more bins). As MergeMap's coordinate calculations for a consensus map are inflated relative to cM distances in individual maps, consensus LG lengths were normalized to the mean cM length from the individual maps.

Synteny with *P. vulgaris*

The cowpea genome assembly described above was compared to *P. vulgaris* pseudomolecules and unanchored scaffolds (from https://phytozome.jgi.doe.gov/pz/portal.html) using MUMmer (Kurtz *et al.*, 2004). Alignments that were further used had a minimum identity of 55.11% and a mean identity of 89.24%. The positions of *P. vulgaris* gene models within the aligned regions was used to position each cowpea SNP relative to *P. vulgaris* gene models. A synteny plot was constructed based on SNPs that had a cM position in the cowpea consensus map and fell within the region of the cowpea sequence that was aligned with a common bean gene model. Circos v.67-7 (Krzywinski *et al.*, 2009) was used to illustrate the synteny between each cowpea linkage group and common bean chromosome that shared 50 or more SNPs. Cowpea LGs were plotted according to cM lengths, while common bean chromosomes were plotted as physical length.

Cowpea SNP frequencies were based on the number of discovered SNPs per genetic bin and the total size of the WGS scaffolds allocated into the corresponding bin. For every 2 cM window the number of SNPs allocated within that window was divided by the sum of the corresponding WGS scaffold sizes in kb. Two outlying values were replacing by a maximum value so that all of the other calculated values could be easily visualized. *Phaseolus vulgaris* gene densities were calculated as the number of genes available from Schmutz *et al.* (2014) per 500 kb windows.

Genetic analyses of West African accessions

In total, 146 accessions were genotyped with the Cowpea iSelect Consortium Array. Monomorphic loci were eliminated, as were SNPs with missing or heterozygous calls in more than 20% of the samples. A total of 46 620 polymorphic SNPs passed this filtering. The software STRUCTURE v.2.3.4 (Pritchard *et al.*, 2000) was used to infer population structure. SNPs with minor allele frequencies (MAF) <0.05 were excluded. STRUCTURE was run four times for each hypothetical number of subpopulations (K) between 1 and 6, with a burn-in period of 10 000 and 50 000 Monte Carlo Markov Chain iterations. LnP(D) values were plotted and ΔK values were calculated according to Evanno *et al.* (2005) to estimate the optimum number of subpopulations. A final run at the inferred K ($K = 2$) was performed to assign individuals to subpopulations based on a membership probability ≥0.80. Those accessions with probabilities lower than 0.80 were considered 'admixed.' A total of 45 accessions were assigned to subpopulation 1, 44 were assigned to subpopulation 2, and 57 were considered 'admixed' (Table S6). PCA was conducted in TASSEL v5.0 (Bradbury *et al.*, 2007) using SNPs with MAF > 0.05, and results were displayed using TIBCO Spotfire® 6.5.0 (TIBCO Software Inc., Palo Alto, CA, USA).

PIC, *He*, and π values were calculated for all 46 620 SNPs in the entire set of samples, and then separately for subpopulation 1 and subpopulation 2 (45 and 44 samples, respectively; 45 820 polymorphic SNPs). PIC was calculated using the method of Botstein *et al.* (1980), *He* (for two alleles) was calculated as $\mathrm{He} = 1 - \sum_{i=1}^{k} Pi^2$, where P_i is the frequency for the i^{th} allele among a total of k alleles. π was evaluated as in Xu *et al.* (2016). F_{ST} values (Nei, 1977) were calculated per locus for accessions of subpopulations 1 and 2, and also for landraces and breeding lines. *He* and F_{ST} were plotted along the consensus genetic map by averaging values across a sliding window of 5 bins in 1 bin steps. Figures were made using TIBCO Spotfire® 6.5.0.

Accession numbers

Sequence data are available through the National Center for Biotechnology Information, as follows. Raw BAC sequence reads from IT97K-499-35 are available under SRA accessions SRA052227 and SRA052228. BAC assemblies are HTGS accessions AC270865 to AC275219. The WGS assembly of IT97K-499-35 is genome accession MATU00000000. WGS sequence raw reads from 37 diverse cowpea accessions are available under SRA accession SRP077082.

ACKNOWLEDGEMENTS

Authors acknowledge John Weger (Genome Core Facility, UC Riverside), Greg D. May (NCGR) and BGI Americas for sequencing services, David Van Den Berg (University of Southern California) for iSelect genotyping services, and Sassoum Lo and Savanah St. Clair (UC Riverside) for DNA isolation. Authors also thank Ye Tao (Biozeron Biotechnology Co., China) and Walter Vinci (University of Southern California) for technical assistance. Development of the cowpea iSelect was supported by the Feed the Future Innovation Laboratory for Climate Resilient Cowpea (USAID Cooperative Agreement AID-OAA-A-13-00070) and the 2014 Illumina Agricultural Greater Good Initiative. BAC library construction, physical mapping and BAC-end sequencing was supported by the Generation Challenge Program 'Tropical Legumes 1' project. MTP BAC sequencing was supported by NSF IIS-1526742 ('III:Small:Algorithms for Genome Assembly'). Partial support was also provided by Hatch Project CA-R-BPS-5306-H. 'GSS' sequencing (Timko *et al.*, 2008) was supported by Kirkhouse Trust. The sequencing and genotyping of Chinese germplasm was partially supported by the National Key Technology Research & Development Program of China (2013BAD01B04-12) and the National Ten-Thousand Talents Program of China (to P. Xu).

CONFLICT OF INTEREST

Cynthia T. Lawley recognizes a competing interest as an employee of Illumina, Inc.

SUPPORTING INFORMATION

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

Figure S1. Principal component analysis (PCA) of 729 samples representing the diversity of cultivated cowpea (blue) and distribution of 34 of the 36 accessions included in the SNP discovery panel (red).

Figure S2. Graphical representation of the iSelect SNP consensus genetic map for cowpea. Each horizontal line is a bin.

Figure S3. STRUCTURE cluster plot for $K = 5$. Each bar is an accession.

Figure S4. Genetic diversity and population differentiation across linkage groups 2, 3, 4, 5, 6, 9, 10, and 11.

Table S1. WGS assembly characteristics and anchoring to the genetic map.

Table S2. BAC assembly characteristics and anchoring to the genetic map.

Table S3. Information on cowpea accessions used for SNP discovery.

Table S4. Information on the individual mapping population data used for consensus map construction.

Table S5. Pairwise counts of the number of links between cowpea linkage groups (VuLG) and common bean pseudomolecules (Pv).

Table S6. Information on West African accessions used in the study.

Data S1. Mapping statistics for 37 cowpea accessions.

Data S2. Information of SNPs included in the final Cowpea iSelect Consortium Assay.

Data S3. Five individual genetic maps (each sheet) and the genotype dataset used for their construction.

Data S4. iSelect SNP consensus genetic map for cowpea.

Data S5. List of WGS scaffolds and their genetic anchoring information.

Data S6. List of sequenced BACs and their genetic anchoring information.

Data S7. Expected heterozygosity (*He*) values at SNPs across linkage groups for the entire sample set (146 West African accessions; *He*_All), for subpopulation 1 (*He*_Subp.1) and subpopulation 2 (*He*_Subp.2) accessions, and for landraces (*He*_Landraces) and breeding materials (*He*_Breeding).

REFERENCES

Bankevich, A., Nurk, S., Antipov, D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331.

Bozdag, S., Close, T. and Lonardi, S. (2013) A graph-theoretical approach to the selection of the minimum tiling path from a physical map. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10**, 352–360.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Chen, X., Laudeman, T.W., Rushton, P.J., Spraggins, T.A. and Timko, M.P. (2007) CGKB: an annotation knowledge base for cowpea (*Vigna unguiculata* L.) methylation filtered genomic genespace sequences. *BMC Bioinformatics*, **8**, 129.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620.

Ewing, B. and Green, F. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

FAOSTAT (2012) Statistical database of the Food and Agriculture Organization of the United Nations. http://faostat.fao.org/site/339/default.aspx.

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*, **1207**, 3907.

Gu, Y.Q., Ma, Y., Huo, N. *et al.* (2009) A BAC-based physical map of *Brachypodium distachyon* and its comparative analysis with rice and wheat. *BMC Genom.* **10**, 496.

Huynh, B.L., Close, T.J., Roberts, P.A. *et al.* (2013) Gene pools and the genetic architecture of domesticated cowpea. *Plant Genome*, **6**, doi: 10.3835/plantgenome2013.03.0005.

Iwata-Otsubo, A., Lin, J.Y., Gill, N. and Jackson, S.A. (2016) Highly distinct chromosomal structures in cowpea (*Vigna unguiculata*), as revealed by molecular cytogenetic analysis. *Chromosome Res.* **24**, 197–216.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

Kujur, A., Bajaj, D., Upadhyaya, H.D. *et al.* (2015) A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. *Sci. Rep.* **5**, 11166.

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.

Legocki, R.P. and Verma, D.P. (1980) Identification of 'nodule-specific' host proteins (nodulins) involved in the development of Rhizobium–legume symbiosis. *Cell*, **20**, 153–163.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. Marth, G., Abecasis, G. and Durbin, R. and **1000 Genome Project Data Processing Subgroup**. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lonardi, S., Duma, D., Alpert, M. *et al.* (2013) Combinatorial pooling enables selective sequencing of the barley gene space. *PLoS Comput. Biol.* **9**, e1003010.

Lonardi, S., Mirebrahim, H., Wanamaker, S., Alpert, M., Ciardo, G., Duma, D. and Close, T.J. (2015) When less is more: 'slicing' sequencing data improves read decoding accuracy and *de novo* assembly quality. *Bioinformatics*, **31**, 2972–2980.

Lorenc, M.T., Hayashi, S., Stiller, J. *et al.* (2012) Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology (Basel)*, **1**, 370–382.

Lucas, M.R., Diop, N.N., Wanamaker, S., Ehlers, J.E., Roberts, P.A. and Close, T.J. (2011) Cowpea–soybean synteny clarified through an improved genetic map. *Plant Genome*, **4**, 218–225.

Lucas, M.R., Ehlers, J.D., Huynh, B.L., Diop, N.N., Roberts, P.A. and Close, T.J. (2013) Markers for breeding heat-tolerant cowpea. *Mol. Breeding*, **31**, 529–536.

Luo, M.C., Thomas, C., You, F.M., Hsiao, J., Ouyang, S., Buell, C.R., Malandro, M., McGuire, P.E., Anderson, O.D. and Dvorak, J. (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, **82**, 378–389.

Luo, R., Liu, B., Xie, Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, **1**, 18.

McKenna, A., Hanna, M., Banks, E. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Muchero, W., Diop, N.N., Bhat, P.R. *et al.* (2009) A consensus genetic map of cowpea *Vigna unguiculata* (L) Walp. and synteny based on EST-derived SNPs. *Proc. Natl Acad. Sci. USA*, **106**, 18159–18164.

Muchero, W., Roberts, P.A., Diop, N.N., Drabo, I., Cisse, N., Close, T.J., Muranaka, S., Boukar, O. and Ehlers, J.D. (2013) Genetic architecture of delayed senescence, biomass, and grain yield under drought stress in cowpea. *PLoS ONE*, **8**, e70041.

Nei, M. (1977) F-statistics and analysis of gene diversity in sub-divided populations. *Ann. Hum. Genet.* **41**, 225–233.

Pottorff, M.O., Li, G., Ehlers, J.D., Close, T.J. and Roberts, P.A. (2014) Genetic mapping, synteny, and physical location of two loci *for Fusarium oxysporum* f. sp. *tracheiphilum* race 4 resistance in cowpea *Vigna unguiculata* (L.) Walp. *Mol. Breeding*, **33**, 779–791.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Ray, J.D., Dhanapal, A.P., Singh, S.K. *et al.* (2015) Genome-wide association study of ureide concentration in diverse maturity group IV soybean [*Glycine max* (L.) Merr.] accessions. *G3*, **5**, 2391–2403.

Sakai, H., Naito, K., Takahashi, Y., Sato, T., Yamamoto, T., Muto, I., Itoh, T. and Tomooka, N. (2016) The Vigna Genome Server, 'VigGS': a genomic knowledge base of the genus Vigna based on high-quality, annotated genome sequence of the Azuki Bean, *Vigna angularis* (Willd.) Ohwi & Ohashi. *Plant Cell Physiol.* **57**, e2.

Schmutz, J., McClean, P.E., Mamidi, S., *et al.* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713.

Singh, B.B. (2014) *Cowpea: The Food Legume of the 21st Century*. Madison, WI: Crop Science Society of America, Inc. DOI: 10.2135/2014.cowpea.

Soderlund, C., Longden, I. and Mott, R. (2000) FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535.

Thomson, J.A. (2008) The role of biotechnology for agricultural sustainability in Africa. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 905–913.

Thornton, P.K., Ericksen, P.J., Herrero, M. and Challinor, A.J. (2014) Climate variability and vulnerability to climate change: a review. *Glob. Chang. Biol.* **20**, 3313–3328.

Timko, M.P., Rushton, P.J., Laudeman, T.W., Bokowiec, M.T., Chipumuro, E., Cheung, F., Town, C.D. and Chen, X. (2008) Sequencing and analysis of the gene-rich space of cowpea. *BMC Genom.* **9**, 103.

Varshney, R.K., Chen, W., Li, Y. *et al.* (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Vasconcelos, E.V., de Andrade Fonsêca, A.F., Pedrosa-Harand, A., de Andrade Bortoleti, K.C., Benko-Iseppon, A.M., da Costa, A.F. and 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**, 253–266.

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.

Wu, Y., Close, T.J. and Lonardi, S. (2011) Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 381–394.

Xu, P., Wu, X., Muñoz-Amatriaín, M. *et al.* (2016) Genomic regions, cellular components and gene regulatory basis underlying pod length variations in cowpea (*V. unguiculata* L. Walp). *Plant Biotechnol. J.* doi:10.1111/pbi.12639.

Yang, K., Tian, Z., Chen, C. *et al.* (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.