RESOURCES

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

Maria Munoz-Amatriain1*,1, Hamid Mirebrahim1,2, Pei Xu3, Steve I. Wanamaker1, MingCheng Luo4, Hind Alhakami2, Matthew Alpert2, Ibrahim Atokple5, Benoit J. Batien6, Ousmane Boukar7, Serdar Bozdag2,8, Ndiaga Cisse9, Issa Drabo6, Jeffrey D. Ehlers1,10, Andrew Farmer11, Christian Fatokun12, Yong Q. Gu13, Yi-Ning Guo1, Bao-Lam Huynh14, Scott A. Jackson15, Francis Kusi5, Cynthia T. Lawley16, Mitchell R. Lucas1, Yaqin Ma1,14, Michael P. Timko17, Jiajie Wu4, Frank You4,8, Noelle A. Barkley19, Philip A. Roberts14, Stefano Lonardi2 and Timothy J. Close1*

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

© 2016 The Authors. The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
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 available from 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 GAIL, 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 HindIII and MboI (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 HindIII and 63.2% from MboI) 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 ESTs, which are available at http://phymap.ucdavis.edu/cowpea. Additional 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 ESTs, which are available at http://phymap.ucdavis.edu/cowpea.

**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 LGs between maps. 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. The resulting consensus map contains 37,372 SNP loci mapped to 3,280 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

---

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

<table>
<thead>
<tr>
<th>Genetic map</th>
<th>Characteristic LG1</th>
<th>LG2</th>
<th>LG3</th>
<th>LG4</th>
<th>LG5</th>
<th>LG6</th>
<th>LG7</th>
<th>LG8</th>
<th>LG9</th>
<th>LG10</th>
<th>LG11</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tvu-1476 × IT84S-2246-4</td>
<td>1156</td>
<td>1447</td>
<td>1266</td>
<td>1307</td>
<td>1308</td>
<td>1246</td>
<td>909</td>
<td>1152</td>
<td>1508</td>
<td>1703</td>
<td>1360</td>
<td>17500</td>
</tr>
<tr>
<td>Sanzí × Vita7</td>
<td>1156</td>
<td>1447</td>
<td>1266</td>
<td>1307</td>
<td>1308</td>
<td>1246</td>
<td>909</td>
<td>1152</td>
<td>1508</td>
<td>1703</td>
<td>1360</td>
<td>17500</td>
</tr>
<tr>
<td>ZN016 × Zhijiang822</td>
<td>1156</td>
<td>1447</td>
<td>1266</td>
<td>1307</td>
<td>1308</td>
<td>1246</td>
<td>909</td>
<td>1152</td>
<td>1508</td>
<td>1703</td>
<td>1360</td>
<td>17500</td>
</tr>
<tr>
<td>CB46 × IT95S-903-1</td>
<td>1156</td>
<td>1447</td>
<td>1266</td>
<td>1307</td>
<td>1308</td>
<td>1246</td>
<td>909</td>
<td>1152</td>
<td>1508</td>
<td>1703</td>
<td>1360</td>
<td>17500</td>
</tr>
<tr>
<td>CR27 × IT83E-18</td>
<td>1156</td>
<td>1447</td>
<td>1266</td>
<td>1307</td>
<td>1308</td>
<td>1246</td>
<td>909</td>
<td>1152</td>
<td>1508</td>
<td>1703</td>
<td>1360</td>
<td>17500</td>
</tr>
</tbody>
</table>

© 2016 The Authors. 
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, and...
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 \( \Delta K \) 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 \( \geq 0.8 \); 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 \( \left( F_{ST} \right) \) 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 \( \left( F_{ST} = 0.02 \right) \), in accordance with STRUCTURE.
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 ($H_e$) and nucleotide diversity ($\pi$) 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 $H_e$ and $\pi$ 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, $H_e$ and $\pi$ 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, $H_e$ and $\pi$ 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, $H_e$, and $\pi$ are highly correlated, only expected heterozygosity ($H_e$) values for each subpopulation, and for landraces and breeding lines are shown at each SNP in Data S7.

$H_e$ 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-st-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.

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

**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 (*FST*) between the two subpopulations. *He* and *FST* 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 *FST* 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 (*FST*).
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 than 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 π = 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-S) 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 HindIII and Mbol. After partial digestion with restriction enzymes, high MW cowpea DNA fragments were ligated with HindIII or BamHI 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 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 HindIII library and 130 kb for the Mbol library.

BAC clones from the two libraries (36 996 from HindIII and 23 312 from Mbol) were fingerprinted using the SNaPshot-based fingerprinting procedure (Luo et al., 2003). BAC DNAs were simultaneously digested with five restriction enzymes (BamHI, EcoRI, XbaI, Xhol, and HaelII), 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 Sultanon score cutoff values (down to 1 × 10^-30). 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'-TACGGCAAGCTTTAGGTGAGA-3') and BigDye terminator chemistry (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 (Lorandi 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 detail by Lorandi 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 1 × 10^-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 1 × 10^-40 and 1 × 10^-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 1 × 10^-40 and 1 × 10^-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 1 038 981 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), using e-value cutoffs of 1 × 10^-40 and 1 × 10^-25 respectively.
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 bead types. 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., 2008; 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/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 conflicting 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. Ln(PD) 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 v5.6 (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; 45 620 polymorphic SNPs). PIC was calculated using the method of Botstein et al. (1980). He (for two alleles) was calculated as 

\[
He = 1 - \frac{\sum_{i=1}^{k} P_i^2}{P^2},
\]

where P is the frequency of the _P_ allele among a total of _k_ alleles. π was evaluated as in Xu et al. (2016). FST values (Nei, 1977) were calculated per locus for accessions of subpopulations 1 and 2, and also for landraces and breeding lines. He and FST 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 MATU0000000. WGS sequence raw reads from 37 diverse cowpea accessions are available under SRA accession SRP077862.

ACKNOWLEDGEMENTS

Authors acknowledge John Weger (Genome Core Facility, UC Riverside), Greg D. May (NGCRI) and BGI Americas for sequencing services, David Van Den Berg (University of Southern California) for iSelect genotyping services, and Sassoum Lo and Savannah 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:Algorithm Challenge Program ‘Tropical Legumes 1’ project. MTP BAC assembly characteristics and anchoring to the genetic map.

Table S6. Information on West African accessions used in the SNP discovery panel.

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.

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 SNP discovery.
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).

REFERENCES
