DATA NOTE

REVISED Assembled genomic and tissue-specific transcriptomic data resources for two genetically distinct lines of Cowpea (*Vigna unguiculata* (L.) Walp) [version 2; peer review: 3 approved]

Andrew Spriggs[1], Steven T. Henderson[2], Melanie L. Hand[2,3], Susan D. Johnson[2], Jennifer M. Taylor[1], Anna Koltunow[2]

[1]CSIRO Agriculture and Food, Acton, ACT, 2601, Australia
[2]CSIRO Agriculture and Food, Urrbrae, SA, 5064, Australia
[3]University of Adelaide, Adelaide, SA, 5000, Australia

v2 **First published:** 09 Feb 2018, **2**:7
https://doi.org/10.12688/gatesopenres.12777.1
Latest published: 18 Jun 2018, **2**:7
https://doi.org/10.12688/gatesopenres.12777.2

Abstract

Cowpea (*Vigna unguiculata* (L.) Walp) is an important legume crop for food security in areas of low-input and smallholder farming throughout Africa and Asia. Genetic improvements are required to increase yield and resilience to biotic and abiotic stress and to enhance cowpea crop performance. An integrated cowpea genomic and gene expression data resource has the potential to greatly accelerate breeding and the delivery of novel genetic traits for cowpea. Extensive genomic resources for cowpea have been absent from the public domain; however, a recent early release reference genome for IT97K-499-35 (*Vigna unguiculata* v1.0, NSF, UCR, USAID, DOE-JGI, http://phytozome.jgi.doe.gov/) has now been established in a collaboration between the Joint Genome Institute (JGI) and University California (UC) Riverside. Here we release supporting genomic and transcriptomic data for two transformable cowpea varieties, IT97K-499-35 and IT86D-1010. The transcriptome resource includes six tissue-specific datasets for each variety, with particular emphasis on reproductive tissues that extend and support the *V. unguiculata* v1.0 reference. Annotations have been included in our resource to allow direct mapping to the v1.0 cowpea reference. The resource described here is supported by downloadable raw and assembled sequence data.

Keywords

cowpea, genome, transcriptome, male and female gametogenesis, seed

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 18 Jun 2018			
version 1 09 Feb 2018	✓ view	✓ view	✓ view

1. **Valerie Hecht**, University of Tasmania, Hobart, Australia
2. **Ian D. Godwin**, University of Queensland, Brisbane, Australia
3. **Timothy J. Close**, University of California, Riverside, Riverside, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Anna Koltunow (anna.koltunow@csiro.au)

Author roles: Spriggs A: Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Henderson ST**: Investigation, Methodology, Resources; **Hand ML**: Investigation, Methodology; **Johnson SD**: Investigation, Methodology, Resources; **Taylor JM**: Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Koltunow A**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Bill and Melinda Gates Foundation [OPP1076280].

How to cite this article: Spriggs A, Henderson ST, Hand ML *et al.* **Assembled genomic and tissue-specific transcriptomic data resources for two genetically distinct lines of Cowpea (*Vigna unguiculata* (L.) Walp) [version 2; peer review: 3 approved]** Gates Open Research 2018, **2**:7 https://doi.org/10.12688/gatesopenres.12777.2

First published: 09 Feb 2018, **2**:7 https://doi.org/10.12688/gatesopenres.12777.1

REVISED **Amendments from Version 1**

All of the reviewer comments were very helpful towards improving the clarity, accuracy and utility of the data presented and all were addressed in the text as detailed below. In addition the data repository supporting the paper has been extended to provide the quality metrics across all the raw DNA and RNA sequence files used in this data collection.

- Changes in Abstract:
  - We more clearly stated here that both cowpea varieties sequenced are transformable.
- Changes in the Introduction:
  - We clarified the contributions to the generation of the *Vigna unguiculata v1.0* reference assembly to be "generated by UC Riverside and subsequently annotated by the Joint Genome Institute".
- Changes in the Methods:
  - We clarified that the CSIRO lines were not maintained through single seed descent.
  - We added further detail of the sourcing of the material for the IT97K-499-35 reference assembly (*Vigna unguiculata v1.0*).
  - We added further detail in the description of sequencing library construction.
  - We deposited an additional data file in the online repository describing a range of quality metrics associated with all raw sequence files to assist in assessment of the quality of the raw sequence used in the paper. We added text in the methods to direct the reader to this file.
- Changes in Data Set Validation:
  - We improved the text to clarify the trend observed for 70–90% overlap thresholds of assembled contigs to the reference genome.
  - We improved the text to clarify when *denovo* assembled transcript contigs were being counted, in contrast to predicted gene models in the genome reference.

See referee reports

Introduction

Cowpea (*Vigna unguiculata* (L.) Walp) is a versatile grain legume crop, also cultivated for vegetative consumption and animal fodder. The grain provides a rich source of protein (25% by weight) for human consumption. Cowpea was domesticated in sub-Saharan Africa and is relatively resilient to heat and drought stress. It has the ability to fix atmospheric nitrogen, and cowpea is often intercropped with cereals or used in crop rotations. Cowpea is grown frequently on subsistence and smallholder farms in mixed crop-livestock systems, particularly in low-input farming systems in the semi-arid regions of West and Central Africa, South America, and Asia (Singh, 2014). Cowpea is a vital component for nutrient security in global agricultural communities.

Cowpea crop improvement has been led by the International Institute of Tropical Agriculture (IITA) through the generation of multiple varieties with improved yield and stress tolerance. However, further improvement is required as many varieties in use exhibit low yield, disease susceptibility, and are prone to abiotic stress (Hall, 2012). Reproductive characteristics have been revisited in cowpea recently and developmental calendars developed for two cowpea varieties developed by IITA, IT86D-1010 and IT97K-499-35 together with supporting developmental experimental tools to support seed yield improvements (Salinas-Gamboa *et al.*, 2016). One approach to increase yield aims to alter sexual reproductive development in high yielding hybrids to an asexual mode in order to assess if it is feasible to save hybrid cowpea seed each growing season (Salinas-Gamboa *et al.*, 2016; Capturing Heterosis OPP1076280). Technological advances in genetic profiling and DNA sequencing approaches over the last decade have facilitated the recent establishment of genomic resources for cowpea (Muñoz-Amatriaín *et al.*, 2017). These data resources have the potential to rapidly accelerate cowpea crop improvement through molecular assisted breeding, characterisation of population diversity and various genomic editing technologies.

The cowpea genome (2*n*=22) has an estimated size of 620 megabases (Mb) (Chen *et al.*, 2007). Analyses of cDNA libraries from 17 different cowpea accessions were used to identify 183,118 expressed sequence tags (ESTs) and 29,728 'unigene' sequences (Muchero *et al.*, 2009). Subsequently, high-throughput sequencing and EST-derived single nucleotide polymorphisms (SNPs) have formed the basis for rapid improvement in consensus genetic maps for cowpea (Lucas *et al.*, 2011; Muchero *et al.*, 2009; Muñoz-Amatriaín *et al.*, 2017). The current consensus map contains 37,372 SNP loci mapped to 3,280 bins and spans 837.11 cM with sub-centimorgan average density (0.26 cM) (Muñoz-Amatriaín *et al.*, 2017).

Most genomic characterisation to date has focussed on the cowpea variety IT97K-499-35, adapted for West Africa. A substantial new genomic resource for IT97K-499-35 containing 97,777 assembled DNA contigs of greater than 1 kb in length, representing 323 Mb of the cowpea genome, has been recently released (Muñoz-Amatriaín *et al.*, 2017). This assembly was combined with sequencing data from two genomic bacterial artificial chromosome (BAC) libraries to generate a BAC physical map (Muñoz-Amatriaín *et al.*, 2017). Despite the substantial contribution and utility of these resources, they did not represent a complete contiguous sequence or 'reference' assembly of the cowpea genome.

University California Riverside (UCR) in collaboration with the Joint Genome Institute have since generated an early release of an annotated genome reference for cowpea (IT97K-499-35) (***Vigna unguiculata*** NSF, UCR, USAID, DOE-JGI. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vunguiculata_er). This resource incorporates long-read sequence technology enabling the assembly of 519.4 Mb into 11 pseudo-molecules and 722 scaffolds generated by UC Riverside and subsequently annotated by the Joint Genome Institute. When finalised, this resource will be foundational to future advances in cowpea crop improvement and will serve as an important unified resource for cowpea crop research.

In this publication, we describe and release survey genome assemblies and tissue-specific transcriptome assemblies derived from IT86D-1010 and IT97K-499-35 to supplement and extend the existing cowpea sequence resources. These cowpea varieties, of different pedigrees, are transformable using *Agrobacterium*-mediated gene insertion (Popelka *et al.,* 2006). They therefore represent important genetic resources for investigating and substantiating gene function. In addition, their genomic and transcriptomic characterisation will enable identification and testing of cell-type specific promoters and genic tools that should facilitate the examination and synthesis of reproductive pathways to improve seed yield in cowpea. We have therefore developed transcriptomic resources to characterise expressed genes in leaf and importantly floral tissues undergoing male and female gametogenic development, and early seed initiation.

The survey genome assembly of IT97K-499-35 supports the reference genome assembly, *Vigna unguiculata* v1.0, of IT97K-499-35; however, the IT86D-1010 data resource is the first public genome-scale resource for this variety. Additional cowpea transcriptome resources are provided for leaf and reproductive tissues for both IT86D-1010 and IT97K-499-35. In accordance with the policies of early release genomes, an extensive comparative analysis of data provided here with the reference assembly (*Vigna unguiculata* v1.0) is not provided. However, we have annotated our transcriptomic and genomic contig data with coordinates of the v1.0 reference, based on IT97K-499-35, to further facilitate integration of publicly available cowpea genome and transcriptome resources.

Transcriptomes of multiple tissues derived from IT97K-499-35 have been generated and previously published (Yao *et al.,* 2016; http://vugea.noble.org). Tissues previously profiled were predominately vegetative and included leaf, stem, root and flower from 5-week-old plants, empty seed pods at 6, 10 and 16 days after pollination and seeds at 8, 10, 14 and 18 days after pollination (DAP) (Yao *et al.,* 2016). In this publication, we provide the first transcriptomic characterisation in both IT97K-499-35 and IT86D-1010 for floral tissues undergoing male and female gametogenic development, and early seed initiation.

The work described in this publication provides a unique and valuable extension to emerging genomic and transcriptomic resources in cowpea. These foundational resources will enable identification and testing of cell-type specific promoters and genic tools that should facilitate the examination and synthesis of reproductive pathways to improve seed yield in cowpea. All transcriptomic and genomic resources are provided with coordinate-based annotation to the IT97K-499-35 reference genome (*V. unguiculata* v1.0) providing integration of these resources to assist coordinated scientific progression of the cowpea research community.

Methods

Plant materials and tissue collection

Cowpea lines IT86D-1010 and IT97K-499-35 were originally sourced from the International Institute of Tropical Agriculture (IITA) and their pedigrees are provided in Supplementary Figure 1. Lines have been maintained in CSIRO for more than 10 generations (not through single seed descent). Material (IT97K-499-35) used in the generation of the reference assembly (*Vigna unguiculata* v1.0) was sourced from Mike Timko at the School of Medicine, University of Virginia, who had previously received the material from IITA. UC Riverside took the IT97K-499-35 line through single seed descent and confirmed 100% homozygosity before bulking. Analysis is underway to compare the CSIRO lines and UC Riverside lines to quantitatively assess genetic similarity of the independently sourced seed stocks. The plants were grown as described by Salinas-Gamboa *et al.* (2016). Young unexpanded leaves were collected for DNA and total RNA extraction for both lines. The reproductive calendars developed for these varieties by Salinas-Gamboa *et al.* (2016) were used to harvest a set of five reproductive tissue types from both IT86D-1010 and IT97K-499-35. Anther tissues containing developing male gametophytes at pollen mother cell, tetrad and mature bicellular pollen stages were pooled to form a pooled male gametophyte (PMG) sample for both lines. In addition, ovules were extracted from both lines from floral buds to provide individual tissue samples containing differentiated megaspore mother cells (MMCs), female meiotic tetrads (FMT), and mature female gametophytes (MFG) at anthesis. Finally, early developing seeds (ES) were collected post-fertilization containing a mixture of zygotes and early globular embryos with proliferating endosperm.

Nucleic acid extraction and sequencing

DNA and RNA extractions were carried out using a Qiagen maxi DNA kit and Qiagen RNeasy plant mini kit, respectively, as per the manufacturer's instructions. Illumina library preparation and sequencing of DNA and RNA was undertaken by the Australian Genome Research Facility (AGRF) with 2 × 100 bp standard insert paired-end sequencing using a Hiseq 2500 system. Shotgun sequencing libraries from single IT86D-1010 and IT97K-499-35 genomic DNA samples were prepared using the Illumina TruSeq Nano DNA Library Prep Kit as per the manufacturer's instructions. Whereas the Illumina TruSeq Stranded mRNA Library Prep Kit was used to prepare poly(A) mRNA sequencing libraries from total RNA samples as per the manufacturer's instructions. Three replicate libraries were prepared and sequenced for each of the RNA samples except for the IT97K-499-35 MMC and FMT samples, where two replicates were sequenced.

Sequence analysis

Raw genomic DNA sequencing reads from IT86D-1010 and IT97K-499-35 were separately assembled into contigs using Biokanga (version 4.3.6) in a multi-step process. First, raw reads were run through 'biokanga filter', where common adapter, primer, and vector contaminants were identified and trimmed. Redundant copies of identical paired-end read pairs were removed, and pairs with no sequence overlap to other raw sequence were also removed as they provided no value to the assembly. Filtered paired-end reads were then assembled into contigs, using 'biokanga assemb', with default parameterisation that allows 1 base substitution per 100bp of sequence overlap. Resulting

contigs were run through a second 'reassembly' step with 'biokanga assemb', allowing up to 5 base substitutions per 100bp of sequence overlaps to provide reduction in redundant sequence representations. Finally, 'biokanga scaffold' added ordering to some contigs, by identifying raw paired-end pairs that match to ends of different contigs under assumptions of sequencing insert fragment size of 110–1500bp. Raw tissue-specific RNASeq reads were separately assembled into transcriptome contigs using Biokanga, with the same multi-step process as used for the genomic DNA reads (above), without the reassembly step to retain putative transcript isoforms. Summary quality metrics for all DNA and RNA sequence reads are provided in the data repository associated with this paper as Supplementary Data 1. These metrics include read length, average GC content and average quality score across the length of the read, the read midpoint and read end.

The assembled genomic DNA sequences of IT86D-1010 and IT97K-499-35 were annotated for predicted gene regions using Augustus v3.1.0 (Stanke & Waack, 2003). From the available Augustus training sets, tomato (*Solanum lycopersicum,* ITAG2.4) gene sequences were selected in Augustus on the basis of the greater percentage of cowpea RNA reads covered by the resulting gene predictions. Predictions from the Augustus approach also encompassed gene predictions from both DNA strands, partial gene predictions and predictions of untranslated regions (UTRs). The resulting protein sequence predictions, with a minimum length of 100 amino acids, were annotated through matches to the NCBI's 'nr' protein sequence database (downloaded 8th August 2017) using 'blastp' with an e-value threshold of 1e-50.

To complete sequence alignment analysis, the genomic DNA sequencing reads and the tissue-specific RNASeq reads from IT86D-1010 and IT97K-499-35 were pre-processed by 'biokanga filter' as described above, prior to alignment with the genomic sequence assemblies of IT86D-1010 and IT97K-499-35 and to the *Vigna unguiculata* v1.0 reference genome sequences. The software 'biokanga align' was used for these alignments and unique-best alignments for each paired-end sequence with an insert fragment size of 100–1000bp to a genomic sequence were reported, with at most 3 base substitutions per 100bp. Auto-end-trimming (read chimera detection) was permitted to 50bp where required. Detection and reporting of SNPs between DNA or RNA sequencing reads and assembled genomes was enabled where there was coverage of at least 5 reads.

Dataset validation

Genomic sequence data for IT86D-1010 and IT97K-499-35

A total of 527 and 303 million pair-end DNA sequence reads from IT86D-1010 and IT97K-499-35, were generated, respectively. These were assembled into 39,123 contigs for IT86D-1010 and 57,690 contigs for IT97K-499-35 with average lengths of 15.6 and 9.8 kilobases (kb), respectively (Table 1). The contig assemblies generated were able to incorporate 68–73% of the raw DNA reads generated (Table 2). The majority (>87%) of the assembled genomic contigs from IT86D-1010 and IT97K-499-35 could be mapped to the *V. unguiculata* v1.0 reference genome (Table 3) with a minimum of 70% contig overlap. When the required contig overlap increased to 90%, contig mapping to the reference decreased to an average of 63% across assembled datasets. Possible causes for lack of mapping at higher stringencies are loss of contiguous alignment or loss of fidelity of assembly towards the end of the survey assembly contigs. *In-silico* gene prediction identified approximately 60,000 putative coding sequences in both IT86D-1010 and IT97K-499-35 and nearly 70% of these could be annotated to published protein sequences within the NCBI nr public database (Table 4).

Leaf and reproductive cell-type and seed transcriptomes and genomic comparisons

RNA sequencing of the six tissue transcriptomes for each variety generated read counts varying from 125 to 265 million pair-end sequences. These could be assembled into transcript

Table 1. Details of IT97K-499-35 and IT86D-1010 genomic DNA contigs generated and assembled in this study. Contigs of less than 1000 base pairs were excluded in this summary. Comparison to the *V. unguiculata* genome v1.0 of IT97K-499-35 is provided.

	IT97K-499-35 gDNA[1]	IT86D-1010 gDNA[1]	V.Ung v1.0[2]
Number of sequences	57,690	39,123	686
Combined length[3]	568,059,011	609,523,031	519,435,864
Minimum length[3]	1,000	1,000	2,922
Average length[3]	9,847	15,580	757,195
N50 length[3]	17,952	36,693	41,684,185
Maximum length[3]	150,032	347,074	65,292,630

1. Genomic DNA assembled contigs (gDNA)
2. *Vigna unguiculata* v1.0, NSF, UCR, USAID, DOE-JGI, http://phytozome.jgi.doe.gov/
3. Sequence lengths are in basepairs (bp)

Table 2. Proportion of filtered DNA paired-end reads that uniquely align to the assembled genomic DNA sequence sets from IT97K-499-35 and IT86D-1010, and to the *Vigna unguiculata* genome v1.0 assembly of IT97K-499-35. IT86D-1010 and IT97K-499-35 are the genome contig assemblies generated in this resource. Alignments were accepted if they were unique pair-end alignments within 1 kilobase of each other, with auto end-trimming of reads where required, and up to 3 mismatches per 100 base pairs.

Raw read set	IT86D-1010	IT97K-499-35	V.Ung v1.0[1]
IT86D-1010 gDNA[2]	72.6%	64.8%	64.8%
IT97K-499-35 gDNA	62.5%	68.1%	65.9%

1. *Vigna unguiculata* v1.0, NSF, UCR, USAID, DOE-JGI, http://phytozome.jgi.doe.gov/

2. Genomic DNA assembled contigs (gDNA)

Table 3. Proportion of assembled DNA and tissue-specific transcript contigs that align to the *Vigna unguiculata* v1.0 reference genome at three thresholds of overlap.

	Query sequences	50%[1]	70%[1]	90%[1]
IT86D-1010 gDNA[2]	131,241	98.0%	91.1%	65.4%
IT97K-499-35 gDNA	57,690	98.2%	87.8%	60.1%
IT86D-1010 Leaf-tr	73,278	87.7%	80.3%	56.8%
IT86D-1010 PMG[3]-tr[4]	36,179	90.7%	83.9%	59.8%
IT86D-1010 MMC[5]-tr	36,058	91.8%	84.5%	60.1%
IT86D-1010 FMT[6]-tr	40,158	92.2%	87.0%	66.2%
IT86D-1010 MFG[7]-tr	37,710	91.4%	86.6%	65.5%
IT86D-1010 ES[8]-tr	38,623	91.5%	86.6%	65.8%
IT97K-499-35 Leaf-tr	73,967	88.8%	81.9%	59.9%
IT97K-499-35 PMG-tr	35,503	91.8%	85.3%	61.9%
IT97K-499-35 MMC-tr	41,783	92.5%	86.0%	64.7%
IT97K-499-35 FMT-tr	41,580	92.0%	85.7%	64.1%
IT97K-499-35 MFG-tr	36,592	92.4%	87.8%	68.0%
IT97K-499-35 ES-tr	37,470	92.9%	88.0%	67.8%

1. Minimum overlap of query contig required within the target reference genome *Vigna unguiculata* v1.0.

2. Genomic DNA contigs (gDNA)

3. Pooled male gametophyte (PMG)

4. Transcript contigs (tr)

5. Megaspore mother cell stage (MMC)

6. Female meiotic tetrads (FMT)

7. Mature female gametophyte (MFG)

8. Early seeds (ES)

sets varying in size between 35,000 to 74,000 transcript contigs averaging 1 kilobase in length (Table 5 and Table 6). In both cowpea varieties, leaf transcriptomes were the largest in terms of *de novo* assembled contig numbers and the anther transcriptomes were the smallest. In subsequent analyses RNA sequence read alignment to predicted gene models within the assembled genome resources were used to compare expression counts across tissues. The assembled genome resources for both cowpea varieties provided good coverage for the analysis of RNA sequence reads as approximately 70% of reads across all tissues

Table 4. Details of predicted coding gene sequences with 300bp minimum length predicted by Augustus within the assembled genomic DNA contig sets from IT97K-499-35 and IT86D-1010. Matches to NCBI's 'nr' protein sequence database found through 'blastp' of translated predicted genes, with an e-value threshold of 1e-50.

Augustus[1] predicted genes	IT97K-499-35 gDNA[2]	IT86D-1010 gDNA[2]
Number of predicted CDS[3]	61,195	62,963
Combined length[4]	81,479,968	87,223,042
Minimum length[4]	300	300
Average length[4]	1,331	1,385
N50 length[4]	1,791	1,887
Maximum length[4]	14,583	15,909
Number with 'nr'[5] match	41,874	43,253
Percentage with 'nr'[5] match	68%	69%

1. Augustus *in-silico* gene prediction (bioinf.uni-greifswald.de/augustus/; Stanke & Waack, 2003)
2. Genomic DNA assembled contigs (gDNA)
3. Coding DNA Sequence (CDS)
4. Sequence lengths are in basepairs (bp)
5. NCBI 'nr' database downloaded 8th August 2017

Table 5. Details of assembled tissue-specific polyA RNA sequence sets from IT86D-1010. Assembled contigs of less than 300 base pairs were excluded in this analysis.

IT86D-1010	Leaf-tr[1]	PMG[2]-tr	MMC[3]-tr	FMT[4]-tr	MFG[5]-tr	ES[6]-tr
Number of sequences	73,278	36,179	36,058	40,158	37,710	38,623
Combined length[7]	68,247,480	40,853,458	42,326,934	43,555,218	41,562,341	41,760,972
Minimum length[7]	300	300	300	300	300	300
Average length[7]	931	1,129	1,174	1,085	1,102	1,081
N50 length[7]	1,208	1,602	1,660	1,494	1,538	1,501
Maximum length[7]	14,930	12,310	12,441	12,276	11,392	12,272

1. Transcript contigs (tr)
2. Pooled male gametophyte (PMG)
3. Megaspore mother cell stage (MMC)
4. Female meiotic tetrads (FMT)
5. Mature female gametophyte (MFG)
6. Early seeds (ES)
7. Sequence lengths are in base pairs

could be aligned uniquely to all three genomic resources. Transcriptomes derived from IT86D-1010 displayed slightly greater alignment to the IT86D-1010 genomic resource, than the corresponding comparisons for IT97K-499-35 (Table 7). The majority of transcript contigs (80 to 88%) across all tissues in both cultivars could be mapped to the *V. unguiculata* v1.0 reference genome with a minimum of 70% contig coverage (Table 3). The remaining unmapped percentage could represent a range of scenarios including IT86D-1010 specific contigs, missing regions in the *V. unguiculata* v1.0 reference genome, tissue-specific extensions to the IT97K-499-35 resource or misassembled transcript contigs. Predicted gene models were considered expressed if they accrued at least 20 uniquely aligning RNASeq reads. In all tissues, approximately 30% of predicted gene models

Table 6. Details of assembled tissue-specific polyA RNA sequence sets from IT97K-499-35. Assembled contigs of less than 300 base pairs were excluded in this analysis.

IT97K-499-35	Leaf-tr[1]	PMG[2]>-tr	MMC[3]-tr	FMT[4]-tr	MFG[5]-tr	Seed[6]-tr
Number of sequences	73,967	35,503	41,783	41,580	36,592	37,470
Combined length[7]	69,053,233	40,224,171	46,244,665	45,725,500	39,970,331	41,525,557
Minimum length[7]	300	300	300	300	300	300
Average length[7]	934	1,133	1,107	1,100	1,092	1,108
N50 length[7]	1,223	1,619	1,565	1,557	1,528	1,547
Maximum length[7]	13,965	12,238	12,960	13,799	12,605	16,435

1. Transcript contigs (tr)
2. Pooled male gametophyte (PMG)
3. Megaspore mother cell stage (MMC)
4. Female meiotic tetrads (FMT)
5. Mature female gametophyte (MFG)
6. Early seeds (ES)
7. Sequence lengths are in base pairs

Table 7. Proportion of filtered raw RNASeq paired-end reads that uniquely align to the assembled genomic DNA sequence sets from IT97K-499-35 and IT86D-1010, and to the *Vigna unguiculata* genome v1.0 assembly of IT97K-499-35. Alignments by 'biokanga align', with up to 3 substitutions per 100 base pairs, paired-ends retained within 1 kilobase of each other and auto end-trimming of reads where required.

Raw read set	IT86D-1010	IT97K-499-35	V.Ung v1.0[1]
IT86D-1010 Leaf-tr[2]	69.2%	68.6%	68.2%
IT86D-1010 PMG[3]-tr	72.6%	72.1%	70.7%
IT86D-1010 MMC[4]-tr	73.5%	72.7%	72.3%
IT86D-1010 FMT[5]-tr	71.4%	70.7%	70.1%
IT86D-1010 MFG[6]-tr	73.3%	72.7%	72.3%
IT86D-1010 ES[7]-tr	71.3%	70.7%	70.3%
IT97K499-35 Leaf-tr	66.6%	67.2%	66.7%
IT97K-499-35 PMG-tr	69.2%	70.0%	68.5%
IT97K-499-35 MMC-tr	69.9%	70.4%	69.9%
IT97K-499-35 FMT-tr	69.3%	69.8%	69.3%
IT97K-499-35 MFG-tr	69.7%	70.1%	69.6%
IT97K-499-35 ES-tr	68.4%	68.8%	68.1%

1. *Vigna unguiculata* v1.0, NSF, UCR, USAID, DOE-JGI, http://phytozome.jgi.doe.gov/
2. Transcript contigs (tr)
3. Pooled male gametophyte (PMG)
4. Megaspore mother cell stage (MMC)
5. Female meiotic tetrads (FMT)
6. Mature female gametophyte (MFG)
7. Early seeds (ES)

(Table 8) showed expression and 6% of predicted gene models displayed strong tissue-specific expression. We found that on average 90% of IT86D-1010 transcript contigs could be mapped within a IT86D-1010 genomic contig and that the median genomic contig size was 67 kb relative to median transcript contig size of 1.3 kb. This indicates that this resource contains substantial amounts of genomic sequence context around expressed genes in these tissues. This will be important for future

Table 8. Proportion of predicted gene models that accrue RNA sequencing reads. Counts shown for gene models with more than 20 uniquely aligning RNASeq reads.

Transcriptome	August Gene Models[1] Expressed	Proportion of total gene models
IT86D-1010 Leaf-tr[2]	21,024	31%
IT86D-1010 PMG[3]-tr	21,315	31%
IT86D-1010 MMC[4]-tr	20,672	31%
IT86D-1010 FMT[5]-tr	21,356	32%
IT86D-1010 MFG[6]-tr	20,290	30%
IT86D-1010 ES[7]-tr	20,486	30%
IT97K-499-35 Leaf-tr	21,088	31%
IT97K-499-35 PMG-tr	20,953	31%
IT97K-499-35 MMC-tr	20,905	31%
IT97K-499-35 FMT-tr	20,274	30%
IT97K-499-35 MFG-tr	20,005	30%
IT97K-499-35 ES-tr	20,871	31%

1. Augustus *in-silico* gene prediction on IT86D genomic contigs (bioinf.uni-greifswald.de/augustus/; Stanke & Waacke, 2003)
2. Transcript contigs (tr)
3. Pooled male gametophyte (PMG)
4. Megaspore mother cell stage (MMC)
5. Female meiotic tetrads (FMT)
6. Mature female gametophyte (MFG)
7. Early seeds (ES)

explorations of *cis*-regulatory regions associated tissue-specific gene expression.

Data availability

All data associated with this publication are provided on the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Data Access Portal (http://data.csiro.au). Data are available at the direct link: https://doi.org/10.4225/08/5b1723666d6a5 (Spriggs *et al.*, 2017).

Competing interests
No competing interests were disclosed.

Grant information
Bill and Melinda Gates Foundation [OPP1076280].

Acknowledgements
Dr TJ Higgins of CSIRO Agriculture and Food; Dr T Close, Dr Maria Muñoz-Amatriaín, Dr Stefano Lonardi of UC Riverside; Dr BB Singh, Dr O Boukar and IITA for providing IT86D-1010 and IT97K-499-35 cowpea lines for use in the research and the associated pedigree information.

Supplementary material

Supplementary Figure 1: Pedigree maps of IT97K-499-35 and IT86D-1010.

Click here to access the data.

References

Chen X, Laudeman TW, Rushton PJ, *et al.*: **CGKB: an annotation knowledge base for cowpea (*Vigna unguiculata L.*) methylation filtered genomic genespace sequences.** *BMC Bioinformatics.* 2007; **8**: 129.
PubMed Abstract | Publisher Full Text | Free Full Text

Hall AE: **Phenotyping cowpeas for adaptation to drought.** *Front Physiol.* 2012; **3**: 155.
PubMed Abstract | Publisher Full Text | Free Full Text

Lucas MR, Diop NN, Wanamaker S, *et al.*: **Cowpea-soybean synteny clarified through an improved genetic map.** *Plant Genome.* 2011; **4**(3): 218–225.
Publisher Full Text

Muchero W, Diop NN, Bhat PR, *et al.*: **A consensus genetic map of cowpea [*Vigna unguiculata* (L) Walp.] and synteny based on EST-derived SNPs.** *Proc Natl Acad Sci U S A.* 2009; **106**(43): 18159–64.
PubMed Abstract | Publisher Full Text | Free Full Text

Muñoz-Amatriaín M, Mirebrahim H, Xu P, *et al.*: **Genome resources for climate-resilient cowpea, an essential crop for food security.** *Plant J.* 2017; **89**(5): 1042–1054.
PubMed Abstract | Publisher Full Text

Popelka JC, Gollasch S, Moore A, *et al.*: **Genetic transformation of cowpea (*Vigna unguiculata* L.) and stable transmission of the transgenes to progeny.** *Plant Cell Rep.* 2006; **25**(4): 304–12.
PubMed Abstract | Publisher Full Text

Salinas-Gamboa R, Johnson SD, Sánchez-León N, *et al.*: **New observations on gametogenic development and reproductive experimental tools to support seed yield improvement in cowpea [*Vigna unguiculata* (L.) Walp.].** *Plant Reprod.* 2016; **29**(1–2): 165–177.
PubMed Abstract | Publisher Full Text | Free Full Text

Singh BB: **Cowpea: the food Legume of the 21st Century.** Madison, WI: Crop Science Society of America, Inc. 2014.
Publisher Full Text

Spriggs A, Henderson S, Taylor J, *et al.*: **Cowpea genome and transcriptome data resource. v2.** CSIRO. Data Collection. 2017.
Publisher Full Text

Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics.* 2003; **19**(Suppl 2): ii215–25.
PubMed Abstract | Publisher Full Text

***Vigna unguiculata* v1.0.** NSF, UCR, USAID, DOE-JGI.
Reference Source

Yao S, Jiang C, Huang Z, *et al.*: **The *Vigna unguiculata* Gene Expression Atlas (VuGEA) from *de novo* assembly and quantification of RNA-seq data provides insights into seed maturation mechanisms.** *Plant J.* 2016; **88**(2): 318–327.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 1

Reviewer Report 07 March 2018

https://doi.org/10.21956/gatesopenres.13837.r26246

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

The authors have produced genomic and transcriptomic sequences and assemblies from two cowpea accessions that have been possible to transform, IT97K-499-35 and IT86D-1010. They have used this information to develop annotated gene sequences. This is a useful contribution of information on the cowpea genome, adding significantly to the body of knowledge that has been developed from other cowpea genome sequencing and transcriptome efforts, of which there are only a few. They made good use of the IT97K-499-35 reference genome that became available in 2017. The biological and bioinformatic methods are sufficiently explained and seem appropriate. The various tables indicate a consistent level of quality across samples. I have just a few questions or comments.

1. I am somewhat confused by the consistency of the proportion of total gene models that accrue RNA sequencing reads (Table 8) versus the statement on page 5 that leaf transcriptomes had the largest number of contigs and anthers the smallest. I must not be grasping the counting methods. Please clarify.

2. Plant materials and tissue collection. Were these two lines/accessions propagated by single seed descent, such that one would expect every plant within each line/accession to be genetically identical or nearly so? Or was more than on plant taken forward at each generation? Single seed descent for many generations accomplishes homozygosity, providing a single haplotype, which simplifies sequence assembly and alignments. If records are available to address this, then please add that information.

3. Table 3. The values in the 90% column all are considerably lower than the 50% and 70% columns. Is there a simple explanation for this?

4. We received IT97K-499-35 from Mike Timko, who had previously received it from IITA. We took it through three rounds of SSD before bulking and, fortunately, found that the plant used to establish the bulk seed was 100% homozygous. Please adjust the wording on p.4 to indicate that Mike Timko was our source of this accession.

5. JGI annotated the pseudomolecules that were developed at UC Riverside.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 01 March 2018

https://doi.org/10.21956/gatesopenres.13837.r26245

Ian D. Godwin
School of Agriculture and Food Sciences, University of Queensland, Brisbane, Qld, Australia

This is a useful note on the production and assembly of genomic and transcriptomic data for two lines of cowpea. Like most "resource" papers, the manuscript is not particularly exciting or engaging reading, but that is not really the point. The point is to provide the data and allow interested users to access the data and share results. In my opinion, it can help readability to at least include something of biological interest in such a paper, such as a small case study interpreting the data, however, this is not necessary for the paper to stand on its own scientific merits.

Some minor comments.

1. The way I read the abstract, it appears that IT86D-1010 is the second variety and by way of disctinction, that it is transformable. However, in the Introduction, it is made clear that both lines are transformable. The Abstract should be modified to improve the clarity of this.

2. Page 5 para 2: varyingin ... varying in

3. Final paragraph under Data Availability: Data is .... Data are

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genetics, genomics and biotechnology, crop improvement

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 28 February 2018

https://doi.org/10.21956/gatesopenres.13837.r26247

Valerie Hecht
School of Plant Science, University of Tasmania, Hobart, Tas, Australia

The Data Note “Assembled genomic and tissue-specific transcriptomic data ressources for two genetically disctinct lines of Cowpea (*Vigna unguiculata* (L.) Walp)” by A. Spriggs and co-workers describes the obtention of genomic and transcriptomic data from two Cowpea varieties, IT97K-499-35 and IT86D-1010.

While an unclustered genome assembly is already publicly available for Cowpea (*Vigna unguiculata* v1.1; https://phytozome.jgi.doe.gov/pz/portal.html), the genomic dataset presented in this Data Note for two distinct varieties of Cowpea increases the sequence availability for this species.

The transcriptomics data obtained mainly focusses on reproductive tissues, including anthers and ovules from dissected floral buds at different stages. The five reproductive stages used for RNAseq are pooled male gametophyte (PMG), megaspore mother cell (MMC), female meiotic tetrad (FMT), mature female gametophyte (MFG) and early developing seeds (ES) containing young developing

embryos. In addition to these, young unexpanded leaves were also used. All those samples were sequenced in triplicates for both varieties of cowpea, except MMC and FMT from IT97K-499-35 where only two replicates were sequenced. These six tissue specific datasets per variety will be a very useful ressource for differencial expression analysis in reproductive studies of Cowpea.

The rationale to create the datasets is clearly and well explained in the introduction section. The methods and protocols used could be expanded as detailed below. The datasets are accessible though CSIRO Data Access portal, and are presented appropriately in the article as data analysis summaries in tables.

Comments:

- Table 5 and 6 show “details of the assembled tissue-specific polyA RNA sequences” from both varieties, but the nucleic acid extraction section of the Methods does not mention polyA RNA isolation. If polyA RNA was obtained from each tissue, the protocol used for it should be detailed in the methods.

- There are no details about the genomic and cDNA library constructions. Even if those were done commercially by AGRF, details on how the libraries were made should be provided.

- The analysis of the raw sequences was done using Biokanga, a CSIRO developed suite of Next Generation Sequencing analysis tools. Some information about the quality control analysis of the reads obtained for each library should be presented (Basic statistics, Per base and per sequence quality scores, GC content, Miscalled bases...) in order to evaluate the quality of the raw data.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant developmental biology, molecular biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.