Transcriptome Sequencing and Marker Development for Four Underutilized Legumes

Mark A. Chapman

1Manuscript received 21 November 2014; revision accepted 24 January 2015.

This work was funded by a start-up fund to M.A.C. at the University of Southampton. The author would like to thank the Wellcome Trust Centre for Human Genetics for carrying out the sequencing, Nazmul Haq for insightful discussions pertaining to underutilized crops, and Nikol Voutsina for critical review of the manuscript.

1Centre for Biological Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom; and 2Centre for Underutilised Crops, University of Southampton, Southampton SO17 1BJ, United Kingdom

Keywords: beans; drought; legumes; microsatellites; simple sequence repeats (SSRs); transcriptome.

Threats to food security, resulting from climate change and a growing population, are the focus of major international organizations such as the Food and Agriculture Organization of the United Nations (von Braun et al., 2003; FAO et al., 2012). It is clear that major changes will be needed to the way we produce, process, and transport food. Only a handful of crops provide the basic nutritional intake for the majority of the human population; however, dozens of species are cultivated at a smaller scale, and several thousand more have a variety of food and nonfood uses (Mangelsdorf, 1966). Many of the less-studied crops have advantages in terms of environmental tolerances and unique nutritional profiles when compared to the widely cultivated crops. As such, investigations into less-studied crops (hereafter referred to as "underutilized crops") have the potential to reveal crops, crop varieties, and unique alleles that could be applied in future agriculture and crop breeding (Varshney et al., 2010).

Legumes are a vitally important source of protein in developing countries, where meat protein is consumed infrequently. They are also often grown in poor soils, because of their associated atmospheric N2 fixation, and where water availability is low or unpredictable. With this motivation, this study reports transcriptome sequences for four underutilized legumes: hyacinth bean (Lablab purpureus (L.) Sweet), grasspea (Lathyrus sativus L.), winged bean (Psophocarpus tetragonolobus (L.) Sweet), and Bambara groundnut (Vigna subterranea (L.) Verdc.). Hyacinth bean is a very versatile crop, likely originating in Africa, and is cultivated throughout the tropics, often as forage for livestock. The leaves and flowers can be eaten by humans raw or cooked, and the roots can be eaten cooked; notably, this species is one of the most drought-tolerant legumes (Ewansiha and Singh, 2006; Maruthi et al., 2006). Grasspea is widely grown in East Africa and Asia and is often considered an "insurance crop" because of its ability to grow under drought conditions even when other crops fail (Polignano et al., 2005). The seed, while being a good source of protein (Abd El-Moneim et al., 2000), contains the neurotoxin ODAP (β-N-Oxalyl-γ-αβ-diaminopropionic acid), which can cause health problems if it is the sole or major protein source in the diet. Winged bean (P. tetragonolobus) is very versatile, with all parts of the plant (leaves, flowers, tubers, and seeds) edible (National Academy of Sciences, 1975). Finally, Bambara groundnut (V. subterranea) grows similarly to the peanut in that its flowers develop into pods underground. It is thought to have originated in West Africa and is widely cultivated in sub-Saharan Africa, as well as in some regions of Asia, and can be grown on marginal soils where other legumes fail to grow (Jørgensen et al., 2010).

For future investigations, including population genetic and linkage or quantitative trait locus (QTL) mapping, transcriptome sequences can be mined for simple sequence repeats (SSRs, or microsatellites). These markers are typically highly polymorphic and are often transferable to closely related species; for example, 45–91% of 127 Vigna radiata (L.) R. Wilczek SSRs successfully amplified in other Vigna species (Tangpaphornmuang et al., 2009). Furthermore, a comparative analysis of transcriptome sequences can be carried out to isolate markers with the potential to amplify in multiple species. These are typically PCR-based markers (i.e., a conserved orthologous set [COS], sensu Fulton et al., 2002) and are designed to amplify across a range of species, typically within a botanical family, for comparative

1Applications in Plant Sciences 2015 3(2): 1400111; http://www.bioone.org/loi/apps © 2015 Chapman. Published by the Botanical Society of America. This work is licensed under a Creative Commons Attribution License (CC-BY-NC-SA).
QTL analyses and/or phylogenetics (e.g., in the Asteraceae [Chapman et al., 2007], Brassicaceae [Jeong et al., 2014], Pinaceae [Liewlaksaneeyanawin et al., 2009], Rosaceae [Cabrera et al., 2009], and Solanaceae and allies [Wu et al., 2006]). So far, however, there has been no attempt to develop COS markers for the Fabaceae, despite the large size of the family and the variety of economically important species within (Stevens, 2006; Meyer and Purugganan, 2013).

In this study, seedling transcriptomes from these four underutilized crops were sequenced using Illumina technology, assembled, annotated, and mined for SSRs and COS markers. The resultant data have been made publicly available as a valuable resource for investigations into gene expression and genetic variation in these and other leguminous crops.

**METHODS AND RESULTS**

Seeds were obtained from the U.S. Department of Agriculture (Ames, Iowa, USA; Table 1). The seeds were soaked overnight in 0.5 mg·mL\(^{-1}\) gibberellic acid, then rinsed and placed on damp filter paper at 22°C until germination. Germinated seeds were transferred to a 2:1 mixture of Levington’s F2+S and vermiculite in a greenhouse with 16-h days, supplemented with artificial light. RNA extraction was carried out from true leaves of seedlings using a QIAGEN RNeasy Kit (QIAGEN, Crawley, United Kingdom) following the manufacturer’s protocol, with on-column DNase digestion (RNase-free DNase, QIAGEN). Three micrograms of RNA was sent to the Wellcome Trust Centre for Human Genetics (Oxford, United Kingdom) for library preparation using the TruSeq Stranded mRNA Sample Prep Kit (Illumina, San Diego, California, USA) and sequencing on a partial lane of an Illumina HiSeq 2500 with 100-bp paired-end (PE) reads.

The following analyses were carried out for each of the four crops separately. Sequence data were trimmed of adapters and low-quality sequences (q < 10) using cutadapt (http://code.google.com/p/cutadapt/; Martin, 2011), and the Perl script resyncMates.pl (https://github.com/percyfal/ratatosk.ext.scilife/blob/master/scripts/resyncMates.pl) was used to maintain proper read pairing where one of a pair was removed by cutadapt. The resulting clean reads were used in transcriptome assembly with Trinity ver. 2011.04.17 (Grabherr et al., 2011). Libraries were normalized to a kmer coverage of 30 so as to reduce computation time, and then assembled using Trinity with the settings "--min_kmer_cov 2" to increase the stringency for reads being assembled together, and "--max_diffs_same_path 4 --max_internal_gap_same_path 15" which allowed for more divergent reads (up to four nucleotide differences and up to a 15-bp gap) to

| **Table 1.** Summary statistics for the four de novo assembled transcriptomes. |
| **Species (accession)** | **Lablab purpureus** (AustTRCF309497) | **Lathyrus sativus** (REWA2) | **Psophocarpus tetragonolobus** (Ibadan Local-1) | **Vigna subterranea** (KAHEMBA) |
| **Descriptor** | **No. of BLAST hits to** | **No. of BLAST hits to** | **No. of BLAST hits to** | **No. of BLAST hits to** |
| | **Pvul** | **Pvul** | **Pvul** | **Pvul** |
| | | | | |
| **No. of transcripts** | 52,019 | 49,280 | 52,083 | 47,759 |
| **No. of genes** | 32,446 | 33,042 | 33,595 | 34,401 |
| **No. of normalized reads** | 41,878,949 | 34,478,291 | 37,006,438 | 38,961,627 |
| **No. of reads** | 16,190,774 | 15,779,854 | 16,625,155 | 7,887,745 |
| **Country of origin** | Malawi | India | Nigeria | Africa |
| | USA; | | | |

**Note:** PI no. = Plant Introduction number; **Pvul** = Phaseolus vulgaris; SSR = simple sequence repeat; USDA = U.S. Department of Agriculture.

**Table 2.** Primer sequences for 12 COS markers tested in 12 members of the Fabaceae.

| F primer name | F primer sequence | R primer name | R primer sequence |
|---------------|-------------------|---------------|-------------------|
| COS1_F245     | ARCAAAAGCATGGAAGAGATGA | COS1_R728     | CYACATCTCCATGGTACTC |
| COS2_F267     | CAAAGGAGGGAGGAGAGGAGG | COS2_R754     | GATTTTCACACAGTCCTCAT |
| COS3_F1596    | GCTGAGGGGAGGAGGAGGAGG | COS3_R2073    | AGGCTCTTGGGCTGCTGCT |
| COS4_F645     | AGAGACGTGGAGGACCGGTT | COS4_R844     | AAATCTGACCATCTATCTAG |
| COS5_F319     | TCCCTCAACGGAGGAGCTTGA | COS5_R406     | AATTGATCCACAGATATCCT |
| COS6_F104     | AAGGTTTACAGGAGGAGGAGG | COS6_R347     | CKRCGAGAGAGAGATGGT |
| COS7_F633     | AAACAAACTTCTTCATTTCAAGG | COS7_R787     | CCGATATCGAACACAT |
| COS8_F291     | GCCATCGGAAACATGTGTT | COS8_R610     | ATATCATGTCGGCAGGAAATG |
| COS9_F448     | GCGCCCATCCTCGTCTCCTAG | COS9_R530     | AGGCGATCGATCCAGAAGT |
| COS10_F295    | GCGGACCGGAGGAGGAGGAGG | COS10_R537    | TGGCTARGAGSRYGGAGA |
| COS11_F419    | TCCCTTCGTTCCTTGCTTG | COS11_R674    | GCCAATTATCACATTGAG |
| COS12_F446    | TTKCAGAGAAGAGATSVYNTCAT | COS12_R550    | TCMAGATAWCCCTC |

http://www.bioone.org/loi/apps
be assembled into the same contig, i.e., taking into account the potential het-
rozygosity of the species. Heterozygosity was assessed by mapping the reads
back to the respective transcriptomes, parsing the data through SAMtools (Li
et al., 2009; mapq settings -q 3 -Q 20), vcffilter (settings -d 3), seqtk, and
outputting .fa files. Putative functions of transcripts were assigned based on
comparisons to transcripts derived from the common bean (Phaseolus vulgaris
L.) genome sequencing project (Schmutz et al., 2014; http://www.phytozome
.net/commonbean.php) with BLASTn and an E-value cut-off of e-30. SSRs were
identified in the transcriptomes by running misa.pl (http://gpcr.ipk-gatersleben
.de/) with the minimum number of repeats set at 8, 6, and 4 for dinucleo-
tides, trinucleotides, and tetranucleotides, respectively.

To identify COS loci, the four transcriptomes were BLASTed against each other in a “round-robin” fashion (Lablab vs. Lathyris, Lathyris vs. Psophocarpus, Psophocarpus vs. Vigna, and Vigna vs. Lablab) and reciproc-
cal best BLAST hits retained. Where all four comparisons successfully pro-
duced a reciprocal best BLAST hit this was considered a potential COS
locus. For 12 such loci (picked at random, except for avoiding short loci,
<300 bp), the sequences were aligned using Clustal Omega (Sievers and Hig-
gins, 2014) and checked for presences of introns by comparison to the bean
gnome (Schmutz et al., 2014; http://www.phytozome.net/commonbean.php);
degenerate primers were then designed to flank the introns using Primaclade
(Gadberry et al., 2005; Table 2). These were then used to PCR amplify DNA
extracted from the four species used in the transcriptome assembly as well as
eight other legume taxa (Crotalaria brevidens Bentham., Lupinus mutabilis
Sweet, Mimosa pudica L., Phaseolus cocineus L. [runner bean], Pisum sati-
tum L. [field pea], Vigna aconitifolia (Jacq.) Marechal [moth bean], V. un-
guiculata subsp. unguiculata (L.) Walp. [black-eyed pea], and V. unguiculata
subsp. sesquipedalis (L.) Verdc. [yardlong bean]). DNA extraction was car-
rried out with a cetyltrimethylammonium bromide (CTAB)-based method,
and PCR amplification took place using a touchdown program with 55°C final
annealing temperature (see Chapman and Burke, 2012 for details). PCR prod-
ucts were resolved on agarose gels stained with GelRed (Biotium, Hayward,
California, USA).

Between 7.9 and 16.6 million paired-end 100-bp reads were generated for
each of the bean samples (reads have been deposited in the National Center
for Biotechnology Information [NCBI] Sequence Read Archive [http://www.ncbi.nlm.nih.gov/sra] under BioProject ID PRJNA273585). After
normalization, the number of paired reads to assemble was 2.8 to 3.8
million (Table 1). Assembly of the transcriptomes results in between 32,446
and 34,401 genes, corresponding to 47,759 to 52,083 transcripts (Table 1;
Fig. 1A). The number of genes and number of transcripts was similar across
species. In these four species, there are therefore approximately 1.4 to 1.6
isoforms per gene, a value comparable to the proportion of Arabidopsis
intron-containing genes that have more than one alternative transcript
(ca. 42%; Filichkin et al., 2010). The FASTA-formatted transcriptome se-
quences are available from the Dryad Digital Repository (http://dx.doi
.org/10.5061/dryad.k8h76; Chapman, 2015). N50, mean, and median tran-
script lengths were largely comparable across species; however, they were lower for V. subterranea, for which the smallest number of reads was ob-
tained (Fig. 1B). In line with the above statistics, the distribution of tran-
script lengths was qualitatively similar across species; however, more of the
longer transcripts were resolved for Lablab and more of the shorter tran-
scripts for Vigna (Fig. 1C). BLASTn searches against the common bean (P.
vulgaris) coding DNA sequences (one sequence per common bean locus)
resulted in between 71% and 92% of transcripts having a hit (Appendix S1).

Heterozygosity was low, as predicted for an inbred accession from a gene
bank, varying from 0.014% in Lablab and Vigna to 0.028% in Psophocar-
pus and 0.149% in Lathyris.

Mining for SSRs identified between 1106 (Lathyris) and 2427 (Lablab)
SSR-containing loci (with one to four SSRs per locus) per transcriptome
(Table 1; Appendix S2). This corresponds to 4.7%, 2.2%, 3.5%, and 2.6% of
the transcripts containing at least one SSR for Lablab, Lathyris, Psophocar-
pus, and Vigna, respectively. A large proportion of the SSRs (36–43% per
transcriptome) were found in the first or last 50 bp of the sequence, which
may hamper primer design for this subset of loci. The proportions of di-, tri-,
and tetranucleotide repeat SSRs were similar across Lablab, Lathyris, and
Psophocarpus, with di- and trinucleotide repeat SSRs each making up a simi-
lar proportion of SSRs (36–47% each; see above). To test the efficacy of this strategy in pro-
ducing reliably amplifiable cross-family loci, primers were designed for the
first 12 loci (avoiding three very short loci; Table 2) and PCR amplified from
DNA from 12 legume species (see above). Of the 12 primer pairs, one failed
to amplify and four amplified in at least 10 of the 12 species. The remainder suc-
cessfully amplified between two and nine samples (Fig. 3).
**DISCUSSION**

To identify, understand, and benefit from crops with adaptive stress tolerances, it is important to characterize the genetic variation within the crop, with one goal being to relate genetic to phenotypic variation (Burke et al., 2007; Meyer and Purugganan, 2013). Legumes are vitally important crops due to their high protein content, especially in countries where meat protein is rarely consumed (they are sometimes called the “meat of the poor”; de Jager, 2013). The aim of this study was not simply to provide a comprehensive set of transcripts from these species but to pave the way, using SSR and COS markers, for further investigations including population genetics, QTL mapping, and marker-assisted selection.

The percentage of transcripts with a putative orthologue from the fully sequenced *P. vulgaris* was quite high (60–82%; Table 1; Appendix S1), and was highest for the two species most closely related to *Phaseolus* (Lablab and Vigna) and lowest for *Lathyrus*, which is found in a different clade within the Papilionoideae legumes. The percentage of 27,197 *P. vulgaris* transcripts (taking just the longest transcript from each locus) with a BLAST hit in these four legumes was lower (49–70%; Table 1); however, this is not unexpected given the relatively shallow sequencing effort undertaken. Again, the percentage of hits was highest for Lablab and Vigna and lowest for Lathyrus (Table 1).

The development of cross-family markers to use the legume genomic resources in other species with limited resources would be of great benefit. Between 1139 and 2567 SSRs were identified in 1106 to 2427 transcripts (Table 1; Appendix S2) corresponding to approximately 2.2–4.7% of transcripts containing at least one SSR, a value similar to other studies (2.5–21.1% in a review by Ellis and Burke, 2007). In this paper, approximately
1800 potential COS markers were identified (Fig. 2; Appendix S3), and, although only a small subset was tested for cross-family amplification, these markers show promise for fulfilling the aims of COS loci, i.e., they cross-amplify in a diverse assemblage of species and are thus useful in anchoring QTL maps or for phylogenetic reconstructions of the members of the family. This study demonstrates that a modest sequencing effort in selected taxa can lead to very large resources for future crop-specific and comparative analyses.

LITERATURE CITED

Abo El-Moneim, A. M., B. van Dorresteen, M. Baum, and W. Mullega. 2000. Improving the nutritional quality and yield potential of grasspea (Lathyrus sativus L.). Food and Nutrition Bulletin 21: 493–496.

Burke, J. M., J. C. Burger, and M. A. Chapman. 2007. Crop evolution: From genetics to genomics. Current Opinion in Genetics & Development 17: 525–532.

Cabrera, A., A. Kozik, W. Howad, P. Arus, A. F. Iezzoni, and E. van der Kaar. 2009. Development and bin mapping of a Rosaceae Conserved Ortholog Set (COS) of markers. BMC Genomics 10: 562.

Chapman, M. A. 2015. Data from: Transcriptome sequencing and marker amplification, these markers show promise for fulfilling the aims.

Delgado-Salinas, A., M. Thuillet, R. Pasquet, N. Weissen, and M. Lavín. 2011. Vigna (Leguminosae) sensu lato: The names and identities of the American segregate genera. American Journal of Botany 98: 1694–1715.

Ellis, J. R., and J. M. Burke. 2007. EST-SSRs as a resource for population genetic analyses. Heredity 99: 125–132.

Ewansha, S. U., and B. B. Singh. 2006. Relative drought tolerance of important herbaceous legumes and cereals in the moist and semi-arid regions of West Africa. Journal of Food Agriculture and Environment 4: 188–190.

FAO, WFP, and IFAD. 2012. The state of food insecurity in the world 2012. FAO, Rome, Italy.

Filichkin, S., H. Priest, S. Givan, R. Shen, D. Bryant, S. Fox, W.-K. Wong, et al. 2010. Genome-wide approach of alternative splicing in Arabidopsis thaliana. Genome Research 20: 45–58.

Fulton, T. M., R. van der Heuven, N. T. Eenetta, and S. D. Tankersley. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14: 1457–1467.

Gaddbery, M. D., S. T. Malcolm, A. N. Dout, and E. A. Kellogg. 2005. Primaclade: a flexible tool to find conserved PCR primers across multiple species. Bioinformatics (Oxford, England) 21: 1263–1264.

Graberre, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644–652.

Jeong, Y.-M., W.-H. Chung, H. Chung, N. Kim, B.-S. Park, K.-B. Lim, H.-J. Yu, et al. 2014. Comparative analysis of the radish genome based on a conserved ortholog set (COS) of Brassica. Theoretical and Applied Genetics 127: 1975–1989.

Jørgensen, S. T., F. Liu, M. Overdagoo, W. H. Ntundu, J. Sarkazin, and J. L. Christiansen. 2010. Drought responses of two Bambara groundnut (Vigna subterranea L. Verd.) landraces collected from a dry and a humid area of Africa. Journal of Agronomy and Crop Science 196: 412–422.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Martin, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25: 2078–2079.

Liewak easeeyanawin, C., J. Zhuang, M. Tang, N. Farzaneh, G. Lueng, C. Cullis, S. Findlay, et al. 2009. Identification of COS markers in the Pinaceae. Tree Genetics & Genomes 5: 247–255.

Mangelsdorf, P. C. 1966. Genetic potentials for increasing yields of food crops and animals. Proceedings of the National Academy of Sciences, USA 56: 370–375.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBNet journal 17: 10–12.

Marieth, M. N., B. Manjunatha, A. R. Rekha, M. R. Govindappa, J. Cohen, and V. Munivappa. 2006. A tropical forage solution to poor quality ruminant diets: A review of Lablab purpureus. Annals of Applied Biology 149: 187–195.

Meyer, R. S., and M. D. Purugganan. 2013. Evolution of crop species: Genetics of domestication and diversification. Nature Reviews. Genetics 14: 840–852.

National Academy of Sciences. 1975. The winged bean. A high protein crop for the tropics. National Academy of Sciences, Washington, D.C., USA.

Policiano, G. B., P. Uggenti, V. Alba, V. Bischgani, and C. Della Gatta. 2005. Morpho-agronomic diversity in grasspea (Lathyrus sativus L.). Plant Genetic Resources: Characterization and Utilization 3: 29–34.

Schmutz, J., P. E. McLean, S. Mamidi, G. A. Wu, S. B. Cannon, J. Grimson, J. Jenkins, et al. 2014. A reference genome for common bean and genome-wide analysis of multiple domestications. Nature Genetics 46: 707–713.

Severs, F., and D. G. Higgins. 2014. Clustal Omega, accurate alignment of very large numbers of sequences. In D. J. Russell [ed.], Multiple sequence alignment methods, vol. 1079, Methods in molecular biology, 105–116. Humana Press, Totowa, New Jersey, USA.

Stevens, P. F. 2006. Angiosperm Phylogeny Website. Version 7, May 2006. Website: http://www.mobot.org/MOBOT/research/APweb/ [accessed 26 June 2015].

Tangpsatthornruang, S., P. Somta, P. Uthaiapansong, J. Chanprasert, D. Sangsaruk, W. Serehalak, W. Sommanas, et al. 2009. Characterization of microsatellites and gene contents from genome shot gun sequences of mungbean (Vigna radiata (L.) Wilczek). BMC Plant Biology 9: 137.

Varshney, R. K., J.-C. Glazmann, H. Lueng, and J.-M. Ribaout. 2010. More genomic resources for less-studied crops. Trends in Biotechnology 28: 452–460.

Von Braun, J., M. S. Swaminathan, and M. W. Rosegrant. 2003. Agriculture, food security, nutrition, and the millennium development goals: Annual report essay. Institutional Food Policy Research Institute, Washington, D.C., USA.

Worochowski, M. F., M. Lavín, and M. J. Sanderson. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. American Journal of Botany 91: 1846–1862.

Wu, F.-N., L. A. Mueller, D. Crouch, P. Etienne, and S. D. Tanksley. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: A test case in the euasterid plant clade. Genetics 174: 1407–1420.