Gene mining a marama bean expressed sequence tags (ESTs) database: Embryonic seed development genes and microsatellite marker identification

Emilia N. Sheehama and Percy M. Chimwamurombe*

Department of Biological Sciences, University of Namibia, P. Bag 13301, Windhoek, Namibia.

Received 14 October, 2014; Accepted 30 September, 2015

Tylosema esculentum (marama bean) is one of the underutilized legumes that have potential to provide protein and fatty acids to ensure food security in dry parts of Southern Africa. In order to establish rapid domestication programs for the plant, it is important to explore the plant's genome and identify functional genes molecular markers like microsatellites in order to develop molecular tools. With the advent of high-throughput sequencing technologies and associated bioinformatics methods, expressed sequence tags (ESTs) have been developed for many plant species. These are being developed as an economic means of obtaining large numbers of gene sequences. The aim of this study was to identify genes with important roles for valuable agronomic traits and microsatellite sequences for marama bean. The authors reported the identification of genes associated with embryonic development and microsatellite sequences. The future direction will entail characterization of these genes using gene over-expression and mutant assays.

Key words: Namibia, simple sequence repeats (SSR), data mining, homology searches, bioinformatics, Tylosema esculentum.

INTRODUCTION

In order to meet the future food and nutrition demands of an increasing population in southern Africa, and to make optimal use of marginal land, there is need to start research on little known edible plant species that offer great potential. Tylosema esculentum (Marama bean) is one of those research neglected plants. Marama bean is found in Namibia and Botswana in large populations and small populations in Gauteng, South Africa (Chingwaru et al., 2011). Marama bean is a species in the legume family that produces pods and bean-like seeds perennially. It is native to dry areas of Kalahari agro-ecological zones with little seasonal rainfall. It is particularly important in subsistence agriculture (Müsele and Schönfeldt, 2006). These neglected crops are usually accepted by the local population and better adapted to existing environmental conditions. The potential to provide a more stable food supply for a drought stricken Africa has been reported (Müsele and Schönfeldt, 2006). The plant is a nutritional and valuable food source and can be successfully used in programs
specifically aimed at improving household and food security and in programs aimed to improve protein deficiency in southern Africa.

*T. esculentum* is a non-nodulating, undomesticated tuber-producing legume, abundant in protein, oil and starch (Takundwa et al., 2010). The bean and tuberous root extracts of the plant have also been used as medicine (Chingwaru et al., 2011). Despite abundance of protein, oil and starch, the plant has low yields, producing one or two seeds per pod. With the advent of bioinformatics, researchers have sequenced some legume genomes. The prominent ones are soybean (*Glycine max*), barrel medic (*Medicago truncatula*) and birdfoot trefoil (*Lotus japonicus*), common bean (*Phaseolus vulgaris*), mungbean (*Vigna radiata*), red bean (*Vigna acutifolia*), narrow-leaved lupin (*Lupinus angustifolius*), wild peanut (*Arachis duranensis* and *Arachis ipaensis*), pigeon pea (*Cajanus cajan*) and chickpea (*Cicer arietinum*). The impact of these assembled, annotated genomes has been enormous. These genome sequences are useful for genome comparisons and to transfer information from these biological models to other crop species and vice versa (Cannon et al., 2009). Besides the genome sequencing of some legumes, researchers have also analyzed and exploited ESTs of some plant species in order to understand them better. These powerful tools are used to gain further insight in the molecular manifestations of growth, development, ripening and survival of the organism studied. ESTs have proven to be an economically feasible alternative for gene discovery in species lacking a draft genome sequence (Matukumalli et al., 2004), such as the *T. esculentum*.

An expressed sequence tag (EST) is a short sub-sequence of cDNA derived from cellular mRNA and thus represents part of a protein-coding gene (expressed genes). ESTs are short (200-800 nucleotide bases in length); unedited, randomly selected single-pass sequence reads derived from cDNA libraries (Nagaraj et al., 2006). EST libraries have been developed for plant species such as tomato, apple, rice, grape and citrus (Gonzalez-Ibeas et al., 2007). However, amongst the comprehensive ones are *Arabidopsis thaliana* and *Oryzae sativa* which are the common models for analysis (Gonzalez-Ibeas et al., 2007). Bioinformatics tools can be used to identify and dissect biological processes that are of great technological importance such as flavor development and fruit ripening through the analysis of ESTs (Gonzalez-Ibeas et al., 2007). Gene mining can be used to select candidate genes that are associated with traits of interest (Frank et al., 2004; Higgs and Attwood, 2005). The EST collections can also be used to develop microarrays to identify genes expressed during plant developmental stages and/or responding to environmental stimuli as well as to gain deeper understanding of the common regulatory mechanisms amongst diverse fruit species and ripening physiological patterns (Gonzalez-Ibeas et al., 2007; Fei et al., 2004). Some previous studies have used this analysis to identify genes involved in fruit ripening and pathogen defense (Gonzalez-Ibeas et al., 2007).

*T. esculentum* has no genome draft. Nonetheless, due to its economic and agricultural potential, it is imperative to explore what genes and microsatellites can be efficiently and rapidly mined and identified. Delayed or inefficient analysis due to tool constraints or lack thereof may impede development of potential products such as molecular markers, beneficial genes and useful biochemical pathways. The objectives of this study were to identify genes and microsatellites represented in the ESTs library developed for marama bean.

**MATERIALS AND METHODS**

**ESTs generation and bioinformatic analyses**

RNA was extracted from the embryogenic axis of germinating marama bean seeds using a Qiagen RNA extraction kit (Qiagen, Germany) and this RNA was used to construct the ESTs library using an oligo-dT primer based cDNA synthesis kit (Roche, Germany). Pyro-sequencing with 454 Sequencing technology was used to directly sequence the resultant derived cDNAs without using vectors. For the analysis of datasets, a Window 7 professional, 32-bit operating system and Intel (R) Celeron (R) CPU at 1.80 GHz computer was used together with an internet connection. *T. esculentum* ESTs datasets were analyzed using on-line detached programs. There were two EST datasets that were analyzed: the marama bean single reads and the marama contigs datasets. On average, the ESTs were between 50 and 276 bp for the single reads and 100 and 718 bp for the contigs.

The single reads dataset contained 13,582 sequences which were multiple aligned using ClustalW (www.clustalw.com). This was the preliminary processing to ensure minimum redundancy of sequences. Sequences (20) were aligned at a time. After multiple sequence alignment, 10,660 sequences remained. The sequences clustered as similar scored 90% or higher. The longest sequence of each batch was selected for downstream processing.

A BLASTn search was run against the non-redundant nucleotide database of NCBI's Genebank (www.ncbi.nlm.nih.gov/BLAST/). Default search parameters were used. After the BLASTn, a tBLASTx search was done on the sequences that produced significant alignment hits. Non-plant genes and similarity alignments with E-value >0.01 were disregarded.

The marama contigs were also processed similarly, multiple aligned using ClustalW and then searched against the *Arabidopsis* database, using the default TAIR BLASTn search parameters (www.arabidopsis.org/BLAST/). The sequences before and after multiple alignment were 924 contigs. The alignments with E-value < 0.5 were considered significant. Contig sequences (50) were analyzed. The single reads that gave significant similarities were scanned for SSRs using an SSR search tool (SSRIT) (www.gramene.org/db/markers/ssrtool).

**RESULTS**

After the analysis of 3247 out of 10660 sequences in the single reads dataset, 227 genes and proteins were identified to be of plant origin. The genes identified were...
found to be involved in essential cellular and metabolic processes in other various plants (Table 1). These were classified as housing keeping genes (79% of the total predicted proteins) and those that did not exhibit high frequencies are classified as specialized (29% of the total predicted proteins) (Figure 1). It was also observed that some of the important putative marama bean genes that were identified and are worth investigating were similar to ribosomal protein S2; disease resistance; retrotransposons B38a yara_autonomous TY1-type; glycosyltransferase CAZY family GT47; tRNA-Lys (trmK) gene intron and mature K (matK) gene; centromeric retrotransposon Psat1-6 mutant gag-pol polyprotein gene; inverted repeat B; RING/FYVE/PHD zinc finger superfamily and transposable element gene. Tables 2 and 3 show the genes that were identified with BLASTx from the single reads data base and TAIR BLASTn from the large contigs, respectively. Table 4 shows the microsatellite repeats that could be mined in the GRAMENE database using SSRIT microsatellite search tool.

For the large contigs dataset, 50 out of 924 sequences were searched against the Arabidopsis database and 34 genes with high similarities were found. In this study, microsatellite sequences were identified and genes associated with these SSR markers were identified to be close to CBL interacting protein kinase (MTR_2g049790) with (CT) repeats; mitochondrion like with (GA) repeats; NA Damage-repair/tolerance protein DRT111 and chloroplastic gene with (TC) repeats and lastly galactosyl transferase 11-like gene with a (TTG) repeats.

**DISCUSSION**

The objectives of this study were to identify genes and microsatellites from the EST single reads and contigs libraries as the first approach of identifying functional genes in marama bean at the embryonic seed stage. The plant lacks a genome draft and therefore has an unknown genome size. Due to the potential of the plant and the endeavors to domesticate it, functional genomic information is necessary to identify and map biochemical pathways and also to design primers for microsatellites. Genes (180) and proteins were identified in the single reads dataset that are involved in photosynthetic and energy processes. Genes (47) from the single reads dataset and the 34 genes identified in the contigs dataset are involved in processes such as transcription, transport, cellular communication, disease resistance and DNA repair.

Within all the genes identified in both the single reads and contig datasets, 7 genes identified have important uses in plant disease resistance as well as in plant biotechnology. For instance, rps2 gene is involved in disease resistance, while retrotransposons and transposons can be used in mutagenesis and plant evolutionary studies (Kumar and Bennetzen, 1999). In this study, the longest marker identified contains three base repeats and the rest contain two bases (dinucleotide repeats). Some genes associated with markers are involved in cellular transport and DNA repair such as DNA repair protein RAD51 homolog 2-like. It still remains to be evaluated how useful will these markers be in the selection and breeding of marama bean with desired superior traits. Similar studies have been done on plant to develop and use microsatellite markers for genetic variation analysis in the Namibian germplasm within and between populations using ESTs. The markers are now available for use in efforts of domestication and conservation. Takundwa et al. (2010) stated that it is desirable to isolate and characterize more DNA markers in the plant for more productive genetic studies such as genetic mapping, marker associated selection and gene discovery. In a study by Bombarely et al. (2010), ESTs were generated and analyzed in the evaluation of Fragaria xananassa at a genetic and molecular level. The analysis of the transcription analysis generated knowledge and molecular tools that would be essential in ongoing breeding programs and had also allowed the development of molecular markers that have been applied to germplasm characterization. ESTs have also been used in studies of plants such as tomato to understand tissue specific genes and biological responses in fruit ripening (Fei et al., 2004), and the fruit traits were studied using ESTs for melon (Cucumis melo). The genes of interest were the genes in the essential traits such as fruit development, fruit maturation and disease resistance, and to speed up the process of breeding new and better adapted melon varieties, such genes are yet to be studied in marama bean.

**Conclusion**

This study has demonstrated the first significant progress in the identification of genes using EST database gene mining for advancing molecular breeding and biotechnological crop improvement for this species, T. esculentum. If a sequence is known, microsatellites and markers can be identified, and then marama bean-specific primers can be developed. Genes that have been identified in marama bean are involved in energy generation, disease resistance, transcription, maturation and DNA repair.

There are a lot more genes to be discovered and studied beyond what this study has discovered for marama bean. In marama bean, traits of interest are, but not limited to increasing number of seeds per pod produced by the plant, selecting for early flowering and early germination (Takundwa et al., 2010). In breeding programs, traits of interest can be linked to markers, which can be used for marker associated selection which is time-saving than traditional breeding. The legumes are remarkably well positioned in the genomic era.
Table 1. NCBI BLASTn search outputs against a NR nucleotide database: marama bean single reads dataset.

| Protein/gene | Number of hits | Identity (%) | EST length (bp) | E-value    | Species                                           | Accession number |
|--------------|----------------|--------------|-----------------|------------|--------------------------------------------------|------------------|
| Chloroplast  | 55             | 92           | 268             | 1.00E-17   | *Eleutherococcus senticosus*                     | JN637765.1       |
| Plastid      | 15             | 98           | 100             | 1.00E-11   | *Quercus rubra*                                  | JX970937.1       |
| Mitochondrion (Mitochondrial DNA) | 53         | 100          | 84              | 5.00E-06   | *Carica papaya*                                  | EU431224.1       |
| ycf2         | 7              | 100          | 192             | 2.00E-72   | *Lacistema robustum*                             | JX6643392.1      |
| ATP synthase subunit α (atpA) | 7          | 98           | 115             | 3.00E-09   | *Medicago truncatula*                            | XM003638699.1    |
| Uncharacterized | 5          | 78           | 269             | 3.00E-17   | *Glycine max*                                    | XM003545696.1    |
| Putative β-1.3 galactosyltransferase 11-like | 1          | 92           | 240             | 4.00E-28   | *Glycine max*                                    | XM003526636.1    |
| S-adenosyl-L-homocysteine hydrolase | 1          | 88           | 149             | 2.00E-20   | *Beta vulgaris*                                  | AB221012.1       |
| Glutamic acid rich-protein-like | 1          | 91           | 84              | 6.00E-17   | *Cicer arietinum*                                | XM004498226.1    |
| GC-rich-sequence DNA-binding factor 1-like | 1          | 96           | 97              | 2.00E-16   | *Glycine max*                                    | XM003528521.1    |
| Chloroplast partial PsA gene for photosystem I P700 chlorophyll a apoprotein A1 α-tubulin 7 | 1          | 94           | 126             | 1.00E-24   | *Vitis riparia*                                  | HF585117.1       |
| Mitochondrial, ATP 1, NAD 4 genes for hypothetical protein, ATP synthase subunit 1, NADH dehydrogenase subunit polygalacturonase-like | 3          | 98           | 241             | 8.00E-61   | *Solanum melongena*                              | AB762698.1       |
| Psa B        | 1              | 97           | 116             | 2.00E-17   | *Glycine max*                                    | XM003551901.1    |
| wbABI 3 mRNA for ABI-3 homolog | 2          | 85           | 231             | 7.00E-16   | *Erythroxylum areolatum*                         | JX662950.1       |
| Serine hydroxymethyl transferase 3 | 3          | 91           | 163             | 4.00E-22   | *Psophacarpus tetragonalobus*                    | AB164427.1       |
| ATP synthase subunit β (atp β) | 6          | 98           | 126             | 9.00E-30   | *Averrhoa carambola*                             | JX663789.1       |
| Photosystem II D2 protein & photosystem II CP43 protein genes (psb D & psb C) | 2          | 98           | 76              | 1.00E-09   | *Petermannia cirrosa*                            | AY465689.1       |
| NADH dehydrogenase subunit 5 gene (nad 5) | 1          | 100          | 119             | 8.00E-24   | *Anthericum ramosum*                             | JX182968.1       |
| Ndh B (Ndh B) | 1              | 94           | 116             | 5.00E-26   | *Drypetes roxburghii*                            | JX664317.1       |
| Ribulose biphosphate carboxylase large chain (rbcL) (1,5 bisphosphate) | 3          | 99           | 203             | 2.00E-100  | *Tylosem esculentum*                             | AJ584710.1       |
| Ribosomal protein S4 mitochondrial-like (rps4) | 1          | 99           | 256             | 7.00E-106  | *Cicer arietinum*                                | XM004488640.1    |
| RNA polymerase β chain (rpo C2) | 6          | 96           | 153             | 1.00E-41   | *Quillaja saponaria*                             | EU002536.1       |
| rpl 14       | 1              | 96           | 132             | 1.00E-27   | *Pera bumellifolia*                              | JX664267.1       |
| 18S ribosomal RNA | 2          | 99           | 178             | 2.00E-60   | *Metanarthecium luteoviride*                     | AB679366.1       |
| rPOB subunit (RNA polymerase B) | 2          | 95           | 246             | 7.00E-41   | *Podocalyx loranthoides*                         | JX663494.1       |
Table 1. Contd.

| Genotype | Assembly | Protein | Length | Amino Acid Length | Accession | Description |
|----------|----------|---------|--------|-------------------|-----------|-------------|
| trn-trnG intergenic spacer and tRNA-Gly (trnG gene) | 1 | 95 | 221 | 1.00E-18 | JN942525.1 | Cercis racemosa |
| tRNA-Lys (trnK) gene intron and matrK (matK) gene | 1 | 99 | 255 | 2.00E-36 | JN881458.1 | Tephrosia rhodesica |
| Photosystem I assembly protein ycf4 (ycf4) gene | 1 | 96 | 252 | 5.00E-87 | HM048910.1 | |
| Photosystem II CP43 chlorophyll apoprotein (psb C) gene | 2 | 96 | 192 | 4.00E-56 | GQ998106.1 | Cornus florida |
| Glucan endo-1,3-beta-glucosidase 4-like | 1 | 83 | 214 | 2.00E-32 | XM002283512.1 | Vitis vinifera |
| SRG-1-like protein | 3 | 77 | 201 | 2.00E-14 | XM004303154.1 | |
| Cytochrome C heme attachment protein (ccsA gene) | 1 | 94 | 214 | 1.00E-32 | GQ997938.1 | Berberidopsis corallina |
| Endoglucanase 11-like | 1 | 85 | 206 | 1.00E-33 | XM003518482.1 | |
| Manganese-dependent ADP-ribose/CDP-alcohol diphosphate like | 1 | 86 | 154 | 4.00E-16 | XM004299102.1 | |
| Centromeric retrotransposon Pisat1-6 mutant gag-pol polyprotein gene | 1 | 84 | 185 | 4.00E-27 | GU136552.1 | |
| Mitogen-activated protein kinase 19-like | 1 | 88 | 256 | 6.00E-47 | XM00448631.1 | |
| Glycosyltransferase, CAZy family GT47 | 1 | 89 | 142 | 1.00E-08 | XM002313394.1 | |
| Mitochondrial voltage-dependent anion-selective channel | 1 | 93 | 159 | 1.00E-17 | DQ072165.1 | |
| Retrotransposon B39_yara_autonomous-Ty1-type | 1 | 89 | 182 | 5.00E-08 | KC608799.1 | |
| PsA putative Pentatricopeptide repeat-containing protein Atg68930-like | 1 | 89 | 138 | 8.00E-21 | XM004298286.1 | |
| SS rRNA gene | 1 | 100 | 185 | 2.00E-16 | Z25803.1 | |
| Psb D gene | 1 | 97 | 268 | 2.00E-60 | JX662334.1 | |
| CBL-interacting protein kinase (MTR_2g049790) | 1 | 94 | 191 | 3.00E-33 | XM003595548.1 | |
| Heterogeneous nuclear ribonucleoprotein D-like | 1 | 91 | 202 | 6.00E-22 | NM001252787.2 | |
| Phosphoenolpyruvate-carboxylase pepc1 isoform | 2 | 88 | 239 | 1.00E-12 | AJ011302.1 | |
| DNA repair protein RAD51 homolog 2-like | 1 | 96 | 201 | 1.00E-09 | XM003547460.1 | |
| psb E-Pet L intergenic spacer | 1 | 80 | 246 | 3.00E-29 | DQ2826228.1 | |
| Histone-lysine N-methyltransferase EZAlike | 1 | 90 | 165 | 8.00E-11 | XM004164933.1 | |
| Uridine nucleosidase 1-like | 1 | 86 | 215 | 1.00E-18 | NM001255381.2 | |
| Histone H3 1-like variant 1 | 1 | 87 | 239 | 7.00E-31 | XM002746095.1 | |
| DNA damage-repair/toleration protein DRT111, chloroplast-like | 1 | 85 | 241 | 8.00E-46 | XM002281707.1 | |

Afr. J. Biotechnol.
Table 1. Contd.

| Putative | Sequence | E-value | Genus          | Accession   |
|-----------|----------|---------|----------------|-------------|
| 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase 3 | 1 | 84 | 266 | 6.00E-42 | Faqs sylvatica | DQ166521.1 |
| U-box domain-containing-protein 4-like | 1 | 82 | 254 | 5.00E-41 | Glycine max | XM003551125.1 |
| Thylakoid structural protein (Psb B gene) | 1 | 96 | 256 | 4.00E-91 | Ceratophyllum sp SM-2010 | GU902269.1 |
| Rps 2 (rps 2 gene) | 1 | 98 | 255 | 1.00E-08 | Passiflora ciliata | JX663163.1 |
| Magnesium transporter MRS2-1-like | 1 | 91 | 191 | 2.00E-50 | Glycine max | XM003543660.1 |
| PetB (petB gene) | 1 | 92 | 285 | 2.00E-87 | Caloncoba echinata | JX663902.1 |
| Nucleobase-ascorbate transporter 1-like | 1 | 90 | 215 | 1.00E-32 | Cicer arietinum | XM004501987.1 |
| Inverted repeat B (transposon boundary in chloroplast) | 1 | 91 | 258 | 6.00E-86 | Rhodeleia championii | EF207455.1 |
| UDP-arabinose 4-epimerase 1-like | 1 | 89 | 142 | 6.00E-22 | Glycine max | XM003546247.1 |
| Photosystem Q (B) protein-like | 1 | 96 | 247 | 1.00E-92 | Cicer arietinum | XM004515165.1 |
| Ndh1 (ndhI) | 1 | 93 | 237 | 2.00E-16 | Vismia ferruginea | JX662090.1 |
| Acetyl-CoA carboxylase carboxyltransferase β | 1 | 88 | 260 | 2.00E-62 | Camellia oleifera | FJ965289.1 |

Figure 1. The overall classification of the genes identified (in single reads dataset) as housekeeping genes or specialized.

**Future perspectives**

In the future, it will be important to identify and characterize more genes and traits, and to extend new genomic tools to orphan species like marama bean. Some of the most critical work does not only rely on new high-throughput sequencing or genomic technologies. This includes characterizing and managing germplasm collections and breeding lines in many species; developing mapping populations for various traits of interest in less-studied species. Working with indigenous farmers ensures that the by-product of centuries of conservation and indigenous knowledge are not lost. Investigating protocols for hybrid seed production in...
Table 2. NCBI tBLASTx search results against NR nucleotide marama bean database from single reads.

| Protein/gene | Identity (%) | Number of positive hits | EST length (bp) | E-Value    | Species                        | Accession       |
|--------------|--------------|-------------------------|-----------------|------------|--------------------------------|-----------------|
| Chloroplast  | 81           | 80                      | 260             | 1.00E-29   | Camellia cuspidata              | KF156833.1      |
| Plastid      | 100          | 100                     | 100             | 0.026      | Quercus rubra                   | JX970937.1      |
| PsbE-PetL Intergenic spacer | 64           | 83                      | 246             | 2.00E-08   | Prunus virginiana               | DQ826228.1      |
| Plastid Genes | 98           | 100                     | 192             | 2.00E-27   | Acrotrema costatum              | HQ664618.1      |
| Chloroplast  | 60           | 63                      | 115             | 4.80E-02   | Berberis bealei                 | KF176554.1      |
| Centromeric retrotransposon PsSat 1-6 mutant gag-pol polyprotein gene | 68 | 84 | 269 | 2.00E-25 | Pisum sativum | GU136552.1 |
| Mitochondrial sequence | 100 | 100 | 268 | 2.00E-24 | Cucumis sativum subsp. melo | JF412793.1 |
| Mitochondrial orf227, atp1, nad4 genes for hypothetical protein, ATP synthase subunit 1, NADH dehydrogenase subunit 4 | 86 | 94 | 241 | 2.00E-21 | Solanum melongena | AB762698.1 |
| Tubulin alpha- 4 chain-like | 84 | 85 | 121 | 3.50E-02 | Glycine max Turnera ulmifolia | XM003555953.1 |
| psaA (psaA gene) | 100          | 100                     | 126             | 5.00E-07   | Glycine max                     | JX664233.1      |
| GC -rich sequence DNA-Binding Factor 1-like | 89 | 89 | 97 | 6.00E-04 | Cicer arietinum Trifolium pratense | XM004498226.1 |
| Glutamic acid-rich protein-like | 100 | 100 | 84 | 4.90E-02 | Rhus communis Trachium caeruleum | AB236805.1      |
| RNA for putative adenosylhomocysteinase | 97 | 96 | 149 | 3.00E-10 | Drypetes roxburghii             | EU909187.1      |
| Putative beta-1,3-galactosyltransferase sqv-2 | 89 | 94 | 240 | 1.00E-14 | Cicer arietinum                 | XM004516889.1   |
| nad 5        | 92           | 95                      | 119             | 2.00E-06   | Lygodium flexuosum              | AJ131135.1      |
| Chloroplast  | 89           | 91                      | 268             | 4.00E-47   | Rhus communis                   | EU909187.1      |
| Ndh B (ndh B) gene | 86           | 86                      | 116             | 1.00E-07   | Trevesia baviensis              | JX664317.1      |
| NAD(P)H-quinone oxidoreductase chain 4 chloroplastic-like | 97 | 96 | 214 | 6.00E-13 | Cicer arietinum                 | XM004516889.1   |
| Chloroplast  | 94           | 94                      | 258             | 4.00E-26   | Lotus japonicus                 | AP002983.1      |
| Unknown      | 80           | 88                      | 237             | 2.00E-21   | Lotus japonicus                 | BT146355.1      |
| 18S ribosomal RNA gene | 94           | 95                      | 178             | 1.00E-23   | Marine streptophyte             | EU143544.1      |
| 5S ribosomal RNA and nontranscribed spacer | 55 | 70 | 185 | 9.00E-06 | Trevesia baviensis              | AY304751.1      |
| Ribosomal protein S4 mitochondrial-like | 99 | 98 | 256 | 3.00E-42 | Cicer arietinum                 | XM004488640.1   |
| Rpl 14 (rpl) gene | 92           | 92                      | 132             | 7.00E-10   | Averrhoa carambola              | JX664237.1      |
| tRNA-Lys (trnK) gene, intron; and maturase K (matK) gene | 91 | 93 | 255 | 2.00E-11 | Bauhinia scandens | JN881423.1 |
| Chromosome POP064-N07 | 89 | 92 | 191 | 7.00E-16 | Populus trichocarpa             | AC209224.1      |
| Gene Name                                                                 | Identity | E-Value | Accession     |
|--------------------------------------------------------------------------|----------|---------|---------------|
| Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit            | 100      | 239     | AB586306.1    |
| Putative Pentatricopeptide repeat-containing protein Atg68930-like       | 93       | 139     | XM002267577.1 |
| Mitogen-activated protein kinase 19-like                                  | 95       | 256     | XM004488631.1 |
| Manganese-dependent ADP-ribose/CDP-alcohol diphosphatase-like           | 80       | 154     | XM00501808.1  |
| Endoglucanase 11-like                                                   | 80       | 206     | XM003518482.1 |
| SRG1-like protein                                                        | 68       | 201     | XM003547143.1 |
| Protein                                                                 | 76       | 214     | XM002297602.1 |
| Chloroplast                                                             | 94       | 285     | KF242504.1    |
| Chloroplast                                                             | 92       | 192     | KF242502.1    |
| Serine hydroxymethyltransferase (SHMT2)                                 | 96       | 163     | EF148390.1    |
| ATP synthase beta chain (atpB) gene                                     | 88       | 126     | JN168037.1    |
| Photosystem II protein D1 (psbA) gene                                   | 100      | 247     | FJ458214.1    |
| Photosystem I assembly protein ycf4 (ycf4) gene                         | 97       | 252     | GQ998510.1    |
| Photosystem II CP47 protein (psbB) gene, Photosystem II subunit (psbT) and photosystem II subunit (psbN) and photosystem II subunit (psbH) genes | 97 | 256 | HQ901410.1 |
| RNA polymerase beta subunit (rpoB) gene                                 | 84       | 246     | JQ274454.1    |
| RpoC2 (rpoC2) gene                                                      | 100      | 153     | JX662688.1    |
| U-box domain-containing protein 4-like                                  | 75       | 254     | XM003538281.1 |
| Phospho-2-dehydro-3-deoxyheptonate aldolase 2, chloroplastic like      | 90       | 266     | XM003545637.1 |
| Uridine nucleosidase 1-like                                             | 74       | 215     | XM002283117.2 |
| Putative enhancer of zeste, ezh                                          | 68       | 165     | XM002515233.1 |
| Histone H3 type 2                                                       | 100      | 239     | XM001862639.1 |
| DNA-damage-repair/tolerance protein DRT111, chloroplastic like          | 81       | 241     | XM0022881707.1|
| DNA repair protein RAD51 homolog 2-like                                 | 54       | 201     | XM004251232.1 |
| Phosphoenolpyruvate carboxylase (PepC-large) gene                       | 92       | 239     | AF333864.1    |
Table 2. Contd.

| Heterogeneous nuclear ribonucleoprotein D-like | 93 | 92 | 202 | 3.00E-11 | Glycine max | NM001252787.2 |
| UDP-arabinose 4-epimerase 1-like | 93 | 92 | 142 | 5.00E-07 | Glycine max | XM003546247.1 |
| Nucleobase-ascorbate transporter 1 | 80 | 92 | 215 | 3.00E-22 | Arabidopsis thaliana | NM126592.2 |
| Magnesium transporter MRS2-1-like | 85 | 89 | 191 | 4.00E-20 | Glycine max | XM003543660.1 |
| probable Polygalacturonase-like | 95 | 100 | 116 | 2.00E-03 | Setaria italica | XM004951228.1 |
| B3 domain-containing transcription factor ABI3-like | 81 | 93 | 231 | 4.00E-10 | Vitis vinifera | XM003632349.1 |
| Uncharacterized | 82 | 89 | 181 | 5.00E-08 | Cicer arietinum | XM004496867.1 |

Table 3. TAIR BLASTn search outputs.

| Contig number | Gene/protein | E- Value | Identity (%) | Marama bean Contig size (bp) | Accession Number |
|---------------|--------------|----------|--------------|-----------------------------|-----------------|
| contig00001   | Unknown Protein | 0.064 | 100 | 132 | AT5G28910.2 |
| contig00003   | Homeo domain glabrous 2 | 0.026 | 95 | 203 | AT1G05230.4 |
| contig00005   | RING/FYVE/PHD Zinc finger superfamily protein | 0.33 | 100 | 170 | AT3G47550.6 |
| contig00008   | GDP-D-mannose 3',5' -epimerase | 0.24 | 100 | 445 | AT5G28840.2 |
| contig00009   | thalianol synthase 1 (THAS 1) | 0.45 | 100 | 225 | AT5G48010 |
| contig00013   | Nucleoporin, Nup133/Nup155 - like | 0.095 | 95 | 188 | AT2G05120.2 |
| contig00014   | phosphotidyl serine synthase family protein | 9.00E-05 | 96 | 174 | AT1G15110.2 |
| contig00015   | Putatative lysine decarboxylase family protein (LOG 1, ATLOG 1) | 0.002 | 100 | 225 | AT2G28035.1 |
| contig00016   | Laccase | 0.097 | 92 | 193 | AT5G01190.1 |
| contig00018   | putative methyl transferase family protein | 0.095 | 100 | 188 | AT5G06050.1 |
| contig00020   | prenylated RAB acceptor 1.B5 (PRA1.B5) | 0.45 | 100 | 221 | AT5G16260.1 |
| contig00021   | Cytochrome P450 superfamily protein (CYP81D1) | 0.17 | 100 | 221 | AT5G28740.1 |
| contig00024   | Tudor/PWWP/MBT domain containing protein | 0.33 | 100 | 412 | AT2G48160.1 |
| contig00026   | photosystem II reaction center protein B (PSBB) | 5.00E-51 | 87 | 203 | ATCG00680.1 |
| contig00027   | high affinity K+ transporter 5 (HAK5, ATHAK5) | 0.011 | 100 | 229 | AT4G34240.1 |
| contig00030   | Transcription factor Jumonsi (jmj) family protein/zinc finger (CSHCZ type) family protein | 0.38 | 100 | 132 | AT2G38950.1 |
| contig00031   | phosphatidic acid phosphohydrolase 2 (PAH 2) | 0.18 | 100 | 230 | AT5G42870.2 |
| contig00032   | plastid - encoded CLP p (CLPP 1, PCPLPP) | 0.11 | 100 | 155 | ATCG00670.1 |
| contig00033   | phytoene desaturation (POS1, HPD) | 0.4 | 100 | 140 | AT1G06570.2 |
| contig00034   | ATP synthase subunit 1 (ATP1) | 2.00E-39 | 95 | 119 | ATMG01190.1 |
| contig00035   | NRAMP metal ion transporter family protein (NRAMP5, ATNRAMP5) | 0.092 | 100 | 128 | AT4G18790.1 |
| contig00037   | F - Box and associated interaction domains- containing protein | 0.41 | 100 | 143 | AT5G62660.1 |
| contig00038   | Transposable element gene | 0.16 | 100 | 208 | AT3G44000.1 |
| contig00040   | Galactose Oxidase/ kelch repeat superfamily protein | 0.037 | 100 | 193 | AT1G55270.1 |
| contig00041   | lysm domain GP1-anchored protein 2 precursor (LYM2) | 0.34 | 100 | 120 | AT2G17120.1 |
| contig00043   | photosystem II reaction center protein N (PSBN) | 0.01 | 87 | 209 | ATCG00700.1 |
various legumes; and working to maintain and develop under-studied legumes for use in diverse, challenging growing environments around the globe is a responsibility to help diversity crops for a changing world climate (Cannon et al., 2009).

The rapid increment in the information and data generation in plant science, demands for tools and methods in data management, visualization integration, analysis, modeling and prediction has also increased (Useche et al., 2001, Rhee et al., 2006; Frank et al., 2004). In this regard, bioinformatic analysis is a utility. This specific knowledge can then be used to produce stronger, more drought resistant crops and improve the quality of livestock, making them healthier, more disease resistant and more productive (Singh et al., 2011).

Conflict of interests

The authors did not declare any conflict of interest.

REFERENCES

Bombarely A, Merchante C, Csukasi F, Cruz-Rus E, Caballero JL, Medina-Escobara N, Blanco-Portales R, Botella MA, Munoz-Blanco J, Sanchez-Sevilla JF, Valpuesta V (2010). Generation and analysis of ESTs from strawberry (Fragaria xananassa) fruits and evaluation of their utility in genetic and molecular studies. BMC Genomics 11(503):1-17.

Cannon SB, May GD, Jackson SA (2009). Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. Plant Physiol. 151: 970-977.

Chingwaru W, Majinda TR, Yeboah SO, Jackson CJ, Kapewangolo PT, Kandawa-Schulz M, Cencic A (2011). Tylosoma esculentum (Marama) tuber and bean extracts are strong antiviral agents against rotavirus infection. Evid. Based Complement. Alternat Med. Article ID 284795, 11 pages, 2011. doi:10.1155/2011/284795.

Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ (2004). Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. Plant J. 40: 47-59.

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004). Data mining in bioinformatics using Weka. Bioinformatics 20:2479-2481.

Gonzalez-Ibeas D, Blanca J, Roig C, Gonzalez-To M, Pico B, Truniger V, Gomez P, Deleu W, Cano-Delgado A, Arus P, Nuez F, Garcia-Mas, J, Puigdomenech P, Aranda MA (2007). MELOGEN: an EST database for Melon functional genomics. BMC Genomics 8(306):1-17.

Higgs PC, Atwood TK (2005). Bioinformatics and Molecular Evolution. UK & USA: Blackwell Science Ltd.

Kumar A, Bennetzen JL (1999). Plant retrotransposons. Annu. Rev. Genet. 33:479-532.

Matukumalli LK, Grefenstette JJ, Sonstergard TS, Van Tassell CP (2004). EST-PAGE: Managing and analyzing EST data. Bioinformatics 20(2):286-288.

Mügeler DL, Schönfeldt HC (2006). The Nutrient content of the Marama Bean (Tylosoma esculentum), an underutilized legume from Southern Africa. Agricul.16:7-13.

Nagaraj SH, Gasser RB, Ranganathan S (2006). A Hitchhiker's guide to expressed sequence tag (EST) analysis. Brief. Bioinform. 8(1):6-21.

Rhee SY, Dickerson J, Xu D (2006). Bioinformatics and its applications in Plant Biology. Annu. Rev. Plant Biol. 57:335-360.

Singh VK, Singh AK, Chand R, Kushwaha C (2011). Role of bioinformatics in agriculture and sustainable development. Int. J. Bioinform. Res. 3(2):221-226.

Takundwa M, Chimwamurombe PM, Kunert K, Cullis CA (2010). Isolation and characterisation of microsatellite repeats in marama bean (Tylosoma esculentum). Afr. J. Agric. Res. 5(7):561-566.

Useche FJ, Gao G, Hanafey M, Rafalski A (2001). High-Throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform. 12:194-203.

Table 3. Contd.

| contig00044 | 2 - oxoglutarate (2G) and Fe (II) - dependent oxygenase superfamily protein | 0.17 | 100 | 219 | AT3G18210.2 |
| contig00045 | Reticulon family protein | 0.17 | 100 | 224 | AT4G28430.1 |
| contig00046 | pseudogene, similar to NADH dehydrogenase | 2.00E-59 | 90 | 219 | AT2G07709.1 |
| contig00047 | F-Box/ RN1- like domains- containing protein | 0.18 | 90 | 229 | AT1G16930.1 |
| contig00048 | s-locus lectin protein kinase family protein | 0.37 | 95 | 131 | AT5G35370.1 |
| contig00049 | chloroplast ribosomal protein S14 (RPS14) | 7.00E-21 | 93 | 122 | ATCG00330.1 |

Table 4. SSRs identified: di- to tri-nucleotide (2-3) repeat motifs search outputs on GRAMENE database for 66 sequences in single reads.

| Sequence | Motif | Number of repeats | SSR start | SSR end | Sequence length |
|----------|-------|-------------------|-----------|---------|-----------------|
| 024042_2232_1498 | CT | 5 | 181 | 190 | 191 |
| 031209_2673_1063 | GA | 4 | 51 | 58 | 84 |
| 026256_2398_2536 | TC | 4 | 166 | 173 | 241 |
| 003796_2321_0642 | TTG | 4 | 207 | 218 | 240 |

Chingwaru W, Majinda TR, Yeboah SO, Jackson CJ, Kapewangolo PT, Kandawa-Schulz M, Cencic A (2011). Tylosoma esculentum (Marama) tuber and bean extracts are strong antiviral agents against rotavirus infection. Evid. Based Complement. Alternat Med. Article ID 284795, 11 pages, 2011. doi:10.1155/2011/284795.