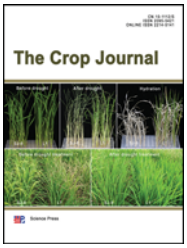
THE CROP JOURNAL XX (2017) XXX–XXX

Available online at www.sciencedirect.com

ScienceDirect



Development and validation of simple sequence repeat markers from *Arachis hypogaea* transcript sequences

Houmiao Wang[a,b], *Yong Lei*[a], *Liying Yan*[a], *Liyun Wan*[a], *Yan Cai*[a], *Zefeng Yang*[b], *Jianwei Lv*[a], *Xiaojie Zhang*[a], *Chenwu Xu*[b,*], *Boshou Liao*[a,*]

[a]*Key Laboratory of Oil Crop Biology of the Ministry of Agriculture, CAAS-ICRISAT Joint Laboratory for Groundnut Aflatoxin Management, Oil Crops Research Institute of Chinese Academy of Agricultural Sciences, Wuhan 430062, Hubei, China*

[b]*Jiangsu Key Laboratory of Crop Genetics and Physiology/Co-Innovation Center for Modern Production Technology of Grain Crops, Key Laboratory of Plant Functional Genomics of the Ministry of Education, Yangzhou University, Yangzhou 225009, Zhejiang, China*

ARTICLE INFO

Article history:
Received 2 May 2017
Received in revised form 11 September 2017
Accepted 11 October 2017
Available online xxxx

Keywords:
Arachis hypogaea L.
Molecular marker
Polymorphism
SSR

ABSTRACT

Simple sequence repeats (SSRs) are important molecular markers for assessing genetic diversity in *Arachis hypogaea* L. and many other crops and constructing genetic linkage maps for important agricultural traits. In this study, 29,357 potential SSRs were identified in 22,806 unigenes assembled from *A. hypogaea* transcript sequences. Of these unigenes, 1883 and 4103 were annotated and assigned in Kyoto Encyclopedia of Genes and Genomes Orthology and Eukaryotic Orthologous Groups databases, respectively. Among the SSR motifs, mono- (19,065; 64.94%) and trinucleotide (5033; 17.14%) repeats were the most common, and the three most dominant motifs were A/T (18,358; 62.54%), AG/CT (2804; 9.55%), and AAG/CTT (1396; 4.76%). Polymerase chain reaction (PCR) primer pairs were designed for 4340 novel SSR markers and 210 new SSRs were validated using 24 *A. hypogaea* varieties. Of the 210, 191 (91%) yielded PCR products, with 37 (18%) identifying polymorphisms. The 37 polymorphic SSR markers detected 146 alleles (2–10 alleles per locus), and the average polymorphic information content was 0.403 (with a range of 0.077 to 0.819). The new SSRs enrich the current marker resources for *A. hypogaea* and may also be useful for genetic diversity analysis, functional genomics research, and molecular breeding.

1. Introduction

The cultivated peanut (*Arachis hypogaea* L.) is an important oilseed and cash crop in most tropical and subtropical areas of the world, and one of the primary sources of vegetable oil and protein for human consumption. The species is a self-pollinating allotetraploid (AABB) with two different genomes (A and B), and the genome size is estimated to be 2.8 Gb [1]. The most likely wild diploid progenitors of *A. hypogaea* are *A. duranensis* (AA, 2*n* = 2x = 20) and *A. ipaensis* (BB, 2*n* = 2x = 20) [2]. Reference genomes of *A. duranensis* (A genome) and *A. ipaensis* (B genome) have been released recently in a public database (http://peanutbase.org/). However, a reference genome of *A. hypogaea* is not yet available.

Molecular markers are valuable tools for linkage map construction, quantitative trait locus (QTL) analyses, genomic

* *Corresponding authors.*
E-mail addresses: qtls@yzu.edu.cn (C. Xu), liaoboshou@caas.cn (B. Liao).
Peer review under responsibility of Crop Science Society of China and Institute of Crop Science, CAAS.

https://doi.org/10.1016/j.cj.2017.09.007

Please cite this article as: H. Wang, et al., Development and validation of simple sequence repeat markers from *Arachis hypogaea* transcript sequences, The Crop Journal (2017), https://doi.org/10.1016/j.cj.2017.09.007

selection, gene discovery, and marker-assisted selection for crop improvement [3]. They are also useful for estimating diversity and discriminating among genotypes [4]. Progress has been made in the development of molecular markers and genetic resources in *A. hypogaea* [3,5–7]. However, the application of molecular markers is more advanced in the legume species *Glycine* max and *Medicago truncatula* than in *A. hypogaea*, primarily because of the genome complexity and the narrow genetic base of *A. hypogaea*. Simple sequence repeats (SSRs) and single-nucleotide polymorphisms (SNPs) are currently the standard DNA markers used for gene mapping and marker-assisted selection in many crops [8]. SSR and SNP markers share similar advantages, as both are codominant, abundant throughout genomes, and highly polymorphic. However, SSRs are often multi-allelic, whereas most SNPs are biallelic. SSRs can be easily detected by polymerase chain reaction (PCR) and gel electrophoresis [9]. SSRs have been widely applied in *A. hypogaea* for verification of cultivar identity, diversity studies [7,10–15], linkage map construction [16–18], and QTL analysis [19–23]. SSRs are classified into genomic SSRs and expressed sequence tag-SSRs (EST-SSRs) depending on the origin of the sequences used for the initial identification of these markers. Genomic SSRs are not necessarily expected either to have genetic functions or to be closely linked to transcribed regions of the genome, whereas EST-SSRs are tightly linked with functional genes that may influence important agronomic characters. Because of these advantages, EST-SSRs have been developed and used in many plant species [3,8,24–31]. Although a major disadvantage of EST-SSRs is sequence redundancy, resulting in multiple sets of markers at the same locus, the problem can be circumvented by assembling the ESTs and short reads of RNA transcripts into unigenes. With a large number of EST resources of *A. hypogaea* in public databases, it is advisable to fully exploit the EST-SSRs within these sequences. Since *Arachis* species SSRs were first reported in 1999, a total of 14,390 *A. hypogaea* SSRs have been deposited to date in the public database (Peanut Marker Database, http://marker.kazusa.or.jp/Peanut/). However, the number of SSR markers reported for *A. hypogaea* is still far fewer than that reported for *Glycine* max [32].

The application of next-generation sequencing technologies has efficiently and cost-effectively generated a massive amount of genetics sequence data. Additionally, new techniques have enabled whole-transcriptome sequencing [i.e., RNA sequencing (RNA-seq)] and analysis of crops [33]. RNA-seq is an effective approach for detecting functional genes and characterizing gene expression patterns and associated regulatory networks. This technique has been used successfully to analyze the transcriptome of *A. hypogaea* under different conditions [34–39]. RNA-seq has also allowed the rapid identification of SSR loci derived from ESTs in many crops [3,8,24–31].

We previously reported the first study of the post-harvest *A. hypogaea* transcriptome using RNA-seq and de novo assembly via Illumina paired-end sequencing [40]. The raw sequencing data from that study were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive database (SRP061959), and 128,725 unigenes of *A. hypogaea* were obtained [40]. In this study, these 128,725 unigene sequences were used to detect SSRs for the large-scale development and characterization of SSR markers.

2. Materials and methods

2.1. Plant materials and DNA extraction

Twenty-four *A. hypogaea* varieties from 14 provinces in China were used for analyzing the polymorphism of SSR markers (Table S1). All 24 varieties were planted in the experimental greenhouses of the Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences (CAAS-OCRI), Wuhan, China. Genomic DNA was extracted from fresh leaves of each variety following the hexadecyltrimethyl ammonium bromide (CTAB) method [41]. The quality and integrity of the extracted DNA were evaluated by 1.0% agarose gel electrophoresis and the concentrations were determined with a Beckman DU-650 spectrophotometer (Beckman Coulter, Inc., Brea, CA, USA).

2.2. Expressed sequence tag simple sequence repeat detection and primer design

SSRs present in the 128,725 unigenes were detected using the MIcroSAtellite program (MISA, http://pgrc.ipk-gatersleben.de/misa/misa.html) [31]. The default criteria were based on the minimum number of repeats, which were set as follows: 10 repeating units for mononucleotides, six repeating units for dinucleotides, and five repeating units for tri-, tetra-, penta- and hexanucleotides. The maximum distance between two SSRs was specified as 100 bases. Primer pairs specific for the flanking regions of potential SSRs were designed for each SSR locus using Primer3 (http://primer3.sourceforge.net/releases.php) [25]. Primers were designed based on the following criteria: 1) GC content between 40% and 60%, 2) primer length between 18 and 27 bp, 3) melting temperature between 57 °C and 63 °C, and 4) expected PCR product sizes from 100 to 280 bp.

2.3. Functional classification of simple sequence repeat-containing unigenes

All unigenes containing an SSR motif were classified into Clusters of Orthologous Groups (COG) categories according to the results of National Center for Biotechnology Information (NCBI) BLAST (version 2.2.28+) searches against amino acid sequences in the Eukaryotic Orthologous Groups (KOG) database with an E-value threshold of 10^{-3} (http://www.ncbi.nlm.nih.gov/COG/) [42]. To comprehensively characterize the biological functions and interactions of these SSR-containing unigenes, pathways were assigned based on the KEGG database [43] using BLASTX with an E-value threshold of 10^{-5}.

2.4. Validation of simple sequence repeats

Two hundred and ten SSR markers (Table S2) were validated using 24 *A. hypogaea* varieties (Table S1). PCR reactions were performed as previously described [41]. The PCR-amplified products were separated by nondenaturing

Please cite this article as: H. Wang, et al., Development and validation of simple sequence repeat markers from *Arachis hypogaea* transcript sequences, The Crop Journal (2017), https://doi.org/10.1016/j.cj.2017.09.007