

# Molecular marker development from transcript sequences and germplasm evaluation for cultivated peanut (*Arachis hypogaea* L.)

Ze Peng<sup>1</sup> · Maria Gallo<sup>2</sup> · Barry L. Tillman<sup>1</sup> · Diane Rowland<sup>1</sup> · Jianping Wang<sup>1,3</sup>

Received: 28 January 2015 / Accepted: 4 September 2015 / Published online: 11 September 2015  
© Springer-Verlag Berlin Heidelberg 2015

**Abstract** Molecular markers are important tools for genotyping in genetic studies and molecular breeding. The SSR and SNP are two commonly used marker systems developed from genomic or transcript sequences. The objectives of this study were to: (1) assemble and annotate the publicly available ESTs in *Arachis* and the in-house short reads, (2) develop and validate SSR and SNP markers, and (3) investigate the genetic diversity and population structure of the peanut breeding lines and the U.S. peanut mini core collection using developed SSR markers. An NCBI EST dataset with 252,951 sequences and an in-house 454 RNAseq dataset with 288,701 sequences were assembled separately after trimming. Transcript sequence comparison and phylogenetic analysis suggested that peanut is closer to cowpea and scarlet bean than to soybean, common bean and *Medicago*. From these two datasets, 6455 novel SSRs and 11,902 SNPs were identified. Of the discovered SSRs, 380 representing various SSR types were selected for PCR validation. The amplification rate was 89.2 %. Twenty-two (6.5 %) SSRs were polymorphic between at least one pair

of four genotypes. Sanger sequencing of PCR products targeting 110 SNPs revealed 13 true SNPs between tetraploid genotypes and 193 homoeologous SNPs within genotypes. Eight out of the 22 polymorphic SSR markers were selected to evaluate the genetic diversity of Florida peanut breeding lines and the U.S. peanut mini core collection. This marker set demonstrated high discrimination power by displaying an average polymorphism information content value of 0.783, a combined probability of identity of  $10^{-11}$ , and a combined power of exclusion of 0.99991. The structure analysis revealed four sub-populations among the peanut accessions and lines evaluated. The results of this study enriched the peanut genomic resources, provided over 6000 novel SSR markers and the credentials for true peanut SNP marker development, and demonstrated the power of newly developed SSR markers in genotyping peanut germplasm and breeding materials.

**Keywords** Peanut · Diversity · Marker · Polymorphism · Transcript · SSR · SNP

Communicated by S. Hohmann.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00438-015-1115-6) contains supplementary material, which is available to authorized users.

✉ Jianping Wang  
wangjp@ufl.edu

<sup>1</sup> Agronomy Department, University of Florida, Gainesville, FL 32610, USA

<sup>2</sup> Molecular Biosciences and Bioengineering Department, University of Hawai'i-Mānoa, Honolulu, HI 96822, USA

<sup>3</sup> Genetics Institute, Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL 32610, USA

## Introduction

Cultivated peanut or groundnut (*Arachis hypogaea* L.) is an important oilseed legume crop. It is an annual herbaceous plant grown throughout the world, primarily in Asia, Africa, South and North America (Tillman and Stalker 2009). The cultivated peanut is an allotetraploid species ( $2n = 4x = 40$ , AABB) in *Arachis* with two different genomes, A and B. The genome size was estimated to be 2.8 Gb (Feng et al. 2012). Molecular markers are valuable tools to use in linkage map construction, gene mapping, marker-assisted selection (MAS), genomic selection, and gene discovery (Hyten et al. 2010). They are also a useful

tool for estimation of genetic diversity and discrimination between individuals (Kirst et al. 2005). The molecular marker development in peanut is slow compared to other legume species such as *Glycine max* and *Medicago truncatula*, primarily due to limited genomic resources and genome complexity. Still, progress has been made in peanut genomic resource development (Pandey et al. 2012a, b; Varshney et al. 2013). Recently, the public has been given access to the reference genomes of two wild species, *A. duranensis* for the A genome and *A. ipaensis* for the B genome, the purported progenitors of cultivated peanut. The reference genome of cultivated peanut is not yet available.

Transcript sequences, such as expressed sequence tags (ESTs) and short reads from next generation sequencing (NGS) of RNA transcripts (RNAseq), are valuable genomic resources that provide direct access not only to gene content, but also gene sequence variance and the dynamics of gene expression (Pandey et al. 2012a, b; Varshney et al. 2013). The first report on EST development in peanut was published by Luo et al. (2005), cataloging 1345 ESTs into Genbank. Proite et al. (2007) contributed 8785 ESTs derived from cDNA libraries from leaves and roots of *A. stenosperma*, a wild relative of peanut. Later, more ESTs were added from different tissues such as seeds (Guo et al. 2008; Bi et al. 2010; Song et al. 2010) and leaves (Guo et al. 2009). Koilkonda et al. (2012) and Nagy et al. (2010) contributed 10,102 and 84,229 ESTs utilizing mixed tissues. Although it was estimated that the peanut genome contains 50,000 genes, the majority of the genome consists of repetitive sequences and non-coding regions (Schmidt 2007). By April of 2014, a total of 254,541 *Arachis* ESTs became available to the peanut research community and it is likely that this number will continue to grow. These ESTs can be a great genomic resource for sequence-based simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) marker development.

The SSR and SNP markers are mostly co-dominant and highly abundant in eukaryotic genomes. SSRs in *Arachis* were first reported in 1999, and out of a total of 26 SSRs identified at that time, six proved to be polymorphic (Hopkins et al. 1999). Since then, more than 15,000 SSRs have been made available to the public (Guo et al. 2013), while the majority are EST-SSRs. However, little has been reported on SNP marker development for tetraploid peanuts. Although ample amounts of SNP markers are available throughout the genome, the peanut's tetraploid genomes (AABB) complicate SNP development due to polymorphism between the homoeologous A and B genomes. Two research groups reported the utilization of Illumina GoldenGate arrays containing 1536 and 768

SNPs, respectively (Nagy et al. 2012a, b; Mallikariuna and Varshney 2014), which were only useful for diploid peanut genotyping. Recently, Zhou et al. (2014) developed a large number of SNP markers through NGS and 1621 of them were mapped on linkage groups. The availability of large numbers of SNP markers will greatly increase the marker density and contribute to the genome assembly of cultivated peanut. With a large amount of EST resources in *Arachis*, it is advisable to fully exploit the cost-efficient EST-SSRs within these sequences. In addition, the investigation of the complex nature of SNPs in cultivated peanut will provide better understanding and guidance for large-scale SNP marker development.

Germplasm collections are an effective way to store and maintain genetic diversity and resources. Peanut germplasm collections are an important genetic resource for developing peanut cultivars with improved pod yield, seed grade, and biotic/abiotic stress resistance (Tillman and Stalker 2009). However, with almost 10,000 accessions in cultivated peanut germplasm collections, trait evaluation and gene identification become impractical and inefficient. The core and mini core collections have fewer accessions but are representative of the entire collection. They make thorough evaluation and further utilization of the germplasm approachable (Upadhyaya 2005; Anderson et al. 1996). The U.S. peanut mini core collection containing 112 accessions had been genotypically characterized using SSR markers in different studies (Kottapalli et al. 2007; Barkley et al. 2007; Wang et al. 2011). The average number of alleles per primer pair ranged from 8 to 10, indicating high levels of genetic variation between the mini core accessions. In addition, these studies all reported the existence of population structure within the mini core accessions, which was associated with subspecies and botanical types of peanut. In China, 196 peanut cultivars were genotyped using 46 SSR markers and the genotyping generated three alleles per locus on average, indicating relatively low genetic variability among these cultivars (Ren et al. 2014) and the necessity to broaden the genetic basis through genome introgression from other germplasm collections or wild species. Investigating the genetic diversity and population structure of the breeding lines and germplasm using molecular markers will facilitate parental line selection and provide guidance for peanut breeding programs. The objectives of this study were to: (1) assemble and annotate the publicly available ESTs in *Arachis* and the in-house short reads, (2) develop and validate SSR and SNP markers, and (3) investigate the genetic diversity and population structure of the peanut breeding lines and U.S. peanut mini core collection using developed SSR markers. The results from this study will provide valuable tools for peanut breeding programs.

## Materials and methods

### Sequence resources for marker development

Two different transcript sequence datasets were used in this study. One dataset was composed of 288,701 sequences generated by the 454 Genome Sequencer FLX+ system from two peanut cultivars, Georgia Green (Branch 1996), and C7616 (Corley Holbrook, personal communication, 2013), which were available in our laboratory. The other database contained 252,951 *Arachis* ESTs downloaded from the National Center for Biotechnology Information (NCBI) on February 2013.

### Sequence *de novo* assembly and annotation

The 454 sequences were first assembled using the GS *de novo* Assembler, Newbler v2.6 (454 Life Sciences, Roche) with a trimming file containing vector and primer sequences. The ‘sff’ file was input into Newbler using ‘-cdna’ and ‘-vt’ option followed by the trimmer file. Then the output contigs and singletons from Newbler were further assembled using CAP3 (Huang and Madan 1999), with ‘-o’ set to 40, ‘-p’ set to 95. The use of a second assembler was mainly to produce longer and more complete contigs, thus making it possible to design primers in flanking regions of the markers. Our efforts in minimizing mis-assemblies from CAP3 were to use highly stringent parameters (‘-o’ 40, ‘-p’ 95). All the other settings were default for both assemblers. The resulting combination of contigs and singletons were considered as the first transcript set.

The *Arachis* NCBI EST database, was preprocessed using SeqClean (<http://www.tigr.org/tdb/tgi/software/>) to screen for vector contaminants, low quality sequences, as well as adaptor and linker sequences. During this process, a generic vector database UniVec (downloaded from NCBI) was used as required by SeqClean. Default settings were used for SeqClean. The sequences generated from SeqClean were assembled using CAP3 with ‘-o’ set to 40 and ‘-p’ set to 95. The other options were default. This parameter set was consistent with previous studies on EST analysis (Quackenbush et al. 2000). These resulting contigs and singletons were considered as the second transcript set.

Each transcript set was annotated by blast against GenBank non-redundant protein (nr) and nucleotide (nt) database using BLASTX (‘-b’ 20, ‘-p’ blastx, ‘-e’ 0.000001, ‘-m’ 7) and BLASTN (‘-v’ 1, ‘-b’ 1, ‘-p’ blastn, ‘-e’ 0.000001, ‘-m’ 8) programs, respectively. Gene Ontology (GO) terms were assigned to the transcripts using Blast2Go (Conesa et al. 2005). The transcripts were analyzed by TargetIdentifier (<http://proteomics.yzu.edu/tools/TargetIdentifier.html>) to predict full-length cDNAs. All the settings

followed authors’ recommendations. Each transcript set was compared to five available legume species’ gene indices including *Glycine max* (75.7 Mb), *Medicago truncatula* (53.3 Mb), *Phaseolus vulgaris* (20.6 Mb), *Vigna unguiculata* (25.6 Mb) and *Phaseolus coccineus* (9.6 Mb) downloaded from DFCI at <http://compbio.dfci.harvard.edu/tgi/plant.html>. A set of gene markers (Zimmer and Wen 2012) commonly utilized to resolve phylogenetic relationships of related plant species were searched in peanut transcripts and the five legume species’ gene indices by BLAST. The gene sequences of 26S RNA, chalcone synthase (CHS), and RPB2 found hits for all six species. For each species, the gene sequences corresponding to these three genes were concatenated together. Based on the sequences of this set of common genes across the six species, MEGA6 software (Tamura et al. 2013) was used to construct a phylogenetic tree following the protocol described by Hall (2013). The two transcript sets were also compared to each other for detection of novel ESTs. The *e* value cutoff for BLAST was set to  $10^{-6}$ .

### OrthoMCL clustering analysis

To identify orthologous gene groups, the protein sequences from *A. hypogaea*, the plant model species *Arabidopsis thaliana*, and three legume species *G. max*, *M. truncatula*, *P. vulgaris* were analyzed using OrthoMCL software (Fischer et al. 2011). The protein sequences of *A. thaliana* TAIR10 (Lamesch et al. 2012), *G. max* Wm82.a2.v1 (Schmutz et al. 2010), *M. truncatula* Mt4.0v1 (Young et al. 2011), and *P. vulgaris* v1.0 (Schmutz et al. 2014) were obtained from Phytozome website (<http://phytozome.jgi.doe.gov/pz/portal.html>). The protein sequences from ‘PrimaryTranscriptOnly’ were used. To obtain protein sequences of *A. hypogaea*, the novel ESTs from the 454 database together with assembled sequences from the NCBI EST database were input into OrfPredictor (Min et al. 2005). The resulting protein sequences were used as representatives of peanut proteome. Based on the recommendations, a pre-prepared BLASTX output file (to nr) was also input into OrfPredictor. All the options were following author’s suggestion (‘-v’ 1, ‘-b’ 1, ‘-e’  $1e^{-5}$  for BLASTX; default for OrfPredictor). The OrthoMCL analysis on these protein sequences followed the ‘Basic Protocol 2’ from Fischer et al. (2011). An *e* value of ‘ $1e^{-5}$ ’ was used for all-versus-all BLASTp comparisons. The inflation value (-I) was set to ‘1.5’ (OrthoMCL default).

### SSR and SNP discovery

The two assembled transcript sets were used for SSR and SNP discovery. Both SSR detection and primer design were performed using a combined script of MISA and Primer3

**Table 1** Summary statistics of SSR detection and validation

	454 database			NCBI EST database			Total
	Di-	Tri-	>3	Di-	Tri-	>3	
SSRs in the database	809	735	79	3421	2244	229	7517
SSRs designed with primers	371	437	30	3240	2163	214	6455
Ordered SSR primer pairs	52	93	12	105	100	18	380
Amplified SSRs	45	88	11	89	89	17	339
Polymorphic SSRs	5	2	0	11	3	2	22
Polymorphic rate	11.1 %	2.2 %	0 %	12.4 %	3.4 %	11.8 %	6.5 %

(You et al. 2008). As the NCBI EST database was a public resource, a comprehensive literature search for publicly available *Arachis* SSR primer sequences was carried out to avoid developing duplicated SSRs. The collected *Arachis* SSR primer sequences were aligned to the NCBI EST transcript set using Bowtie (Langmead et al. 2009) with paired-end alignment. The allowed mismatch was set to two and the insert size for valid paired-end alignments ranged from 100 to 1000 bp. Only transcripts without hits were utilized in the SSR development.

For SNP discovery, cleaned sequence reads resulting from Newbler and SeqClean were aligned to their respective contigs using default settings in BWA-SW (Li and Durbin 2010). The SNP calling was carried out with Samtools (Li et al. 2009). The ‘SAM’ files from BWA-SW were converted to the binary format ‘BAM’. Options ‘-u -D -g -f’ were used in the ‘mpileup’ step. The information on SNP calling was stored in the ‘VCF’ files. The ‘BAM’ and ‘VCF’ files were viewed using Integrative Genomics Viewer (Thorvaldsdottir et al. 2013). SNPs with low quality or a read depth of less than two were filtered out using ‘vcfutils.pl varFilter -d 2’.

### SSR marker validation

For SSR marker validation, DNA samples isolated from four peanut cultivars: Georgia Green, C7616, Tifrunner (Holbrook and Culbreath 2007), and NC3033 (Beute et al. 1976) were used as templates for SSR amplification and to evaluate the amplification rate, amplicon size, and polymorphism rate. Georgia Green and C7616 were used because the 454 sequences were derived from them. Thus these two genotypes served as template positive controls. Tifrunner was chosen due to its future role as a reference genome for tetraploid peanuts. Moreover, Tifrunner and NC3033 are two parental lines for a recombinant inbred line population segregating on white mold disease resistance (Holbrook, personal communication, 2013). Genomic DNA was isolated based on the method described by Dellaporta et al. (1983). The extraction buffer from the protocol was modified with the addition of Polyvinylpyrrolidone (PVP) in order to remove phenolic compounds. In

addition, an extra precaution was taken to remove lipids and proteins with chloroform prior to the usage of isopropanol. All SSR primers (Table 1) were synthesized by Invitrogen™, Life Technologies. PCR reactions were performed in 10 µl volumes containing 1 µl of 10 × PCR buffer; 1.5 µl of magnesium chloride (25 mM); 0.8 µl of dNTP (2 mM); 0.5 µl of Taq enzyme; 1.4 µl of forward and reverse primers (2 mM); 1 µl of DNA template (10 ng/µl); 3.8 µl of double distilled water. The PCR program was as follows: initial denaturation at 94 °C for 3 min; 35 cycles of 94 °C for 30 s, 60 °C for 20 s, 72 °C for 40 s; and final extension at 72 °C for 7 min. PCR products were separated through 6 % non-denatured polyacrylamide gel electrophoresis (PAGE) under 320 volts for 3 h and imaged using silver staining.

### SNP validation

Single nucleotide polymorphism validation was conducted by sequencing amplicons of selected SNP regions using the Sanger method. The primers for SNP region amplification were designed from flanking sequences of the targeted SNPs using Primer 3 (Table S1). The amplicon sizes were set to be within 120 and 600 bp to allow complete sequencing of the amplicon in both directions using Sanger sequencing technology. All SNP primers were synthesized by Invitrogen™, Life Technologies. The PCR templates were DNA samples isolated from the same four genotypes as used for SSR validation: Georgia Green, C7616, Tifrunner, and NC3033. A PCR protocol for Phusion® High-Fidelity DNA polymerase (M0530) from New England Biolabs, Inc. was used in accordance with manufacturer recommendations. The PCR program was as follows: initial denaturation at 94 °C for 3 min; 5 cycles of 94 °C for 30 s, 68 °C for 20 s, 72 °C for 40 s; 5 cycles of 94 °C for 30 s, 65 °C for 20 s, 72 °C for 40 s; 5 cycles of 94 °C for 30 s, 60 °C for 20 s, 72 °C for 40 s; 25 cycles of 94 °C for 30 s, 55 °C for 20 s, 72 °C for 40 s; and final extension at 72 °C for 7 min. The resulting PCR products were purified using the GenCatch Advanced PCR Extraction kit from Epoch Life Science (Catalog No. 23-60250). Purified PCR products were then sequenced using the Sanger method at the

**Table 2** The number of alleles, genotypes, polymorphism information content, probability of identity, power of exclusion, labeled dye, genomic location, and potential function of the eight SSR markers

SSR	No. of alleles	No. of genotypes	PIC	I	Q	Dye	Location	SSR-containing EST's function
AHW064	4	6	0.619	0.161	0.424	FAM	A09, B09	High similarity to <i>Lotus japonicas</i> chloroplast DNA
AHW0803	6	15	0.831	0.040	0.699	NED	B04	NA
AHW0813	6	15	0.852	0.032	0.732	PET	A09, B08	NA
AHW0818	8	13	0.745	0.076	0.590	VIC	A04	NA
AHW0833	13	38	0.943	0.006	0.891	FAM	NA	NA
AHW0839	6	8	0.642	0.140	0.459	NED	B01	NA
AHW0843	5	8	0.723	0.095	0.543	PET	A07, B07	Protease inhibitor, seed storage lipid transfer protein family
AHW0845	6	22	0.912	0.012	0.835	VIC	A03	NA
Overall	6.75 <sup>a</sup>	15.6 <sup>a</sup>	0.783 <sup>a</sup>	1 × 10 <sup>-11b</sup>	0.99991 <sup>b</sup>			

'NA' means unknown

<sup>a</sup> Indicates the average of the above values

<sup>b</sup> Indicates the combined probability of *I* and *Q*, considering all SSR markers

Interdisciplinary Center for Biotechnology Research, University of Florida.

To distinguish the A and B genomes for SNP allele origin, DNA samples isolated from two wild diploid species: *A. duranensis*, PI 468321 and *A. ipaensis*, PI 468322 were used as templates for amplicon generation and sequencing from selected primer pairs. Sequences from the two genotypes were compared to those generated from the four cultivated peanut genotypes and visualized using DNA Baser Sequence Assembler v3.x (2012), Heracle BioSoft SRL.

### SSR genotyping of peanut lines

Eight highly polymorphic SSR primer pairs were used to genotype the 142 peanut lines (Table S2). The 142 lines included 107 U.S. peanut mini core accessions (USDA, ARS, 2011), 31 lines (Barry L Tillman, personal communication, 2013) from the Florida Peanut Breeding Program, and the four genotypes used in SSR validation. DNA samples of these lines were extracted following the same method as SSR validation. For PCR amplification of the SSRs, an oligo "tail", 'ACG ACG TTG TAA AAC GAC', was added to the 5' end of forward primers. Primer pairs were labeled with NED, PET, VIC, and FAM dyes (Table 2). The resulting PCR product was multiplexed for genotyping. PCR reactions were performed in 10 µl volumes containing 1 µl of 10 × PCR buffer; 1.5 µl of magnesium chloride (25 mM); 0.8 µl of dNTP (2 mM); 0.5 µl of Taq enzyme; 1.4 µl of forward and reverse primer (2 mM); 1 µl of DNA template (10 ng/µl); 1 µl of 10 × ammonium sulfate; 0.14 µl dye as mentioned above; 2.66 µl of double distilled water. The PCR reaction followed a touchdown program: initial denaturation at 95 °C for 5 min; followed

by 5 cycles of 96 °C for 1 min, 68 °C for 5 min (decreasing 2 °C per cycle), 72 °C for 1 min; 5 cycles of 96 °C for 1 min, 58 °C for 2 min (decreasing 2 °C per cycle), 72 °C for 1 min; 25 cycles of 96 °C for 1 min, 50 °C for 1 min, 72 °C for 1 min; and final extension at 72 °C for 5 min. The PCR amplicons were checked on 1 % agarose gel. The four different dye-labeled PCR products from the same genotype were pooled together following the ratio: 3.0 µl FAM: 3.0 µl VIC: 4.0 µl NED: 6.0 µl PET. The GeneScan™ 600 LIZ™ size standard (Applied Biosystems, USA) and Hi-Di formamide were added to the combined PCR products for the ABI3730 sequencer system. The bands were separately scored using GeneMarker software (Holland and Parson 2011). The presence of a band was scored as "1"; absence of a band was scored as "0"; missing data were scored as "9".

### SSR data analysis

The number of alleles for SSR markers across 142 genotypes was recorded and polymorphism information content (PIC) was calculated using the formula from Botstein et al. (1980). In this study, the frequency of a unique genotype was considered the allele frequency (Milbourne et al. 1997; Pan 2006). The PIC value was calculated using an online tool PICcalc (Nagy et al. 2012a, b). The probability of identity (*I*) and the power of exclusion (*Q*) were calculated using the formulas according to Paetkau et al. (1995) and Jamieson (1994) respectively.

To investigate the population structure among the 142 individuals, each SSR marker band was treated as a dominant marker and the genotyping data was input into STRUCTURE software Ver 2.3.4 (Pritchard et al. 2000).

The program ran 20 iterations with the ‘K’ value ranging from one to 13 (50,000 Burnin Period, 50,000 of MCMC Reps after Burnin). The admixture model was used as suggested by Falush et al. (2007). Two methods were used to determine the best ‘K’: the use of  $L(K)$  (Rosenberg et al. 2001) and the ad hoc quantity method (Evanno et al. 2005). The same genotyping data were input into DARwin V5.0.137 (Perrier et al. 2003) to obtain a phylogenetic tree. The Neighbor-Joining method was used with 500 bootstrap replicates.

## Results

### Sequence assembly

The 454 database was composed of 288,701 reads, with an average length of 215 bp. The first assembly by Newbler grouped 233,685 reads (84.6 %) into 2290 contigs with 42,700 reads remaining as singletons. The average contig length was 425 bp with an ‘N50’ of 679 bp (Table S3). After excluding 409 sequences with lengths less than 20 bp and 35 sequences containing only N’s, all sequences from Newbler in conjunction with quality score information were input into CAP3 for a secondary assembly. As a result, more contigs were generated (Fig. 1a). In total, 8942 contigs were generated with an average length of 353 bp and an ‘N50’ of 362 bp. In addition, 35,065 singletons remained with an ‘N50’ of 228 bp. In summation, 44,007 transcripts were obtained with an average length of 198 bp and an ‘N50’ of 240 bp (Table S3). The final assembly rate was 87.3 % (241,320 reads) with 12.7 % (35,065 reads) remaining as singletons.

Out of the 252,951 ESTs input from the NCBI database, 54,977 were trimmed and 1420 sequences were discarded. As a default option of SeqClean, all sequences less than 100 bp were removed. Therefore, a total of 251,531 ESTs passed SeqClean and were assembled using CAP3. Consequently, 21,392 contigs were obtained with an average length of 852 bp and an ‘N50’ of 870 bp (Table S4). In total 50,905 singletons remained with an ‘N50’ of 579 bp. In summation, 72,297 transcripts were generated with an average length of 606 bp and an ‘N50’ of 661 bp (Table S4).

### Sequence annotation

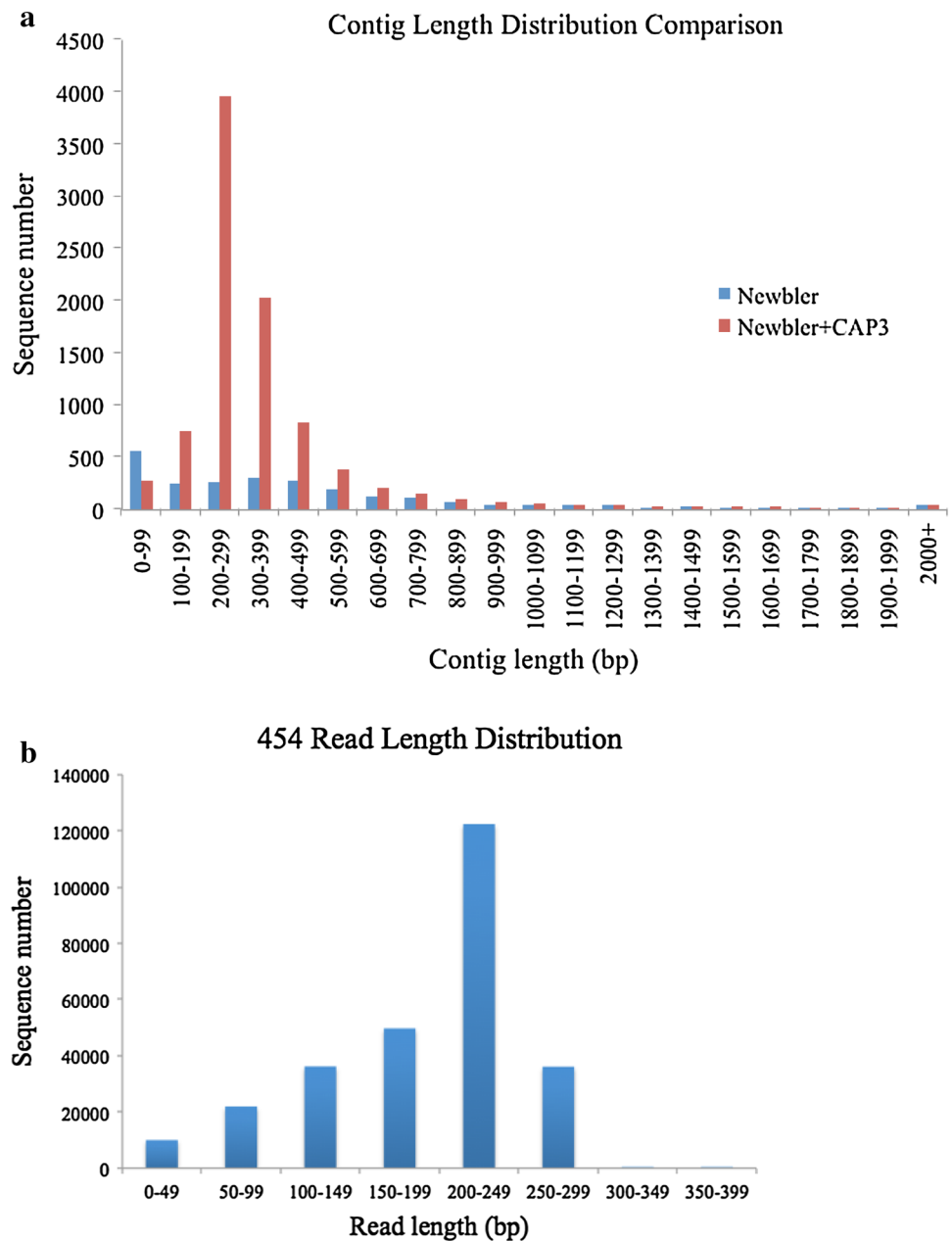
Transcripts assembled from the 454 database were blasted against the NCBI nr and nt databases separately. In total 19,479 (44.3 %) and 24,234 (55.1 %) transcripts assembled from the 454 sequences had hits to nr and nt databases, respectively. There were 4707 (10.7 %) transcripts identified as full-length cDNAs. Blast2Go was used to assign the

functional categories for each transcript with a hit in the nr database. A total of 14,803 annotation counts were generated for the whole transcript set. As sorted in the biological process, the majority of the transcripts are involved in metabolic (20 %) and cellular (20 %) processes. The remainder of the transcripts are involved in single-organism processes (12 %), response to stimulus (9 %), biological regulation (8 %), and other activities. For the molecular function classes, the two major categories of these transcripts were catalytic (47 %) and binding (41 %) activities. For cellular components, most of the sequences were in the cell (37 %), followed by organelle (31 %), and membrane (14 %) categories (Fig. 2).

Transcripts assembled from the NCBI EST database were also aligned to the NCBI nr and nt databases via BLAST. The alignment revealed that 38,161 (52.8 %) of the transcripts had hits in the nr database and 51,122 (70.7 %) in the nt database. There were 13,801 (19.1 %) transcripts predicted as full-length cDNAs. In total 34,136 annotation counts were obtained from Blast2Go. Most sequences fell into either metabolic (20 %) or cellular (20 %) processes. For molecular function, catalytic (45 %) and binding (43 %) activities constituted most of the categories, followed by transporter (6 %) and structural molecule (4 %) activities. As for the cellular component class, the three major categories were cell (37 %), organelle (28 %) and membrane (16 %) (Fig. 2).

The two transcript sets were compared to the five available legume species’ gene indices. For the 454 database, more sequences (21,867, 49.7 %) successfully matched the *G. max* gene indices, than with *M. truncatula* (19,258, 43.8 %), *V. unguiculata* (17,683, 40.2 %), *P. vulgaris* (17,425, 39.6 %), and *P. coccineus* (11,109, 25.2 %) (Table 3). Similarly, for the NCBI EST database, 46,340 transcripts (64 %) had hits to *G. max* gene indices, 41,135 transcripts (57 %) matched *M. truncatula*, 38,347 transcripts (53 %) matched *V. unguiculata*, 34,202 transcripts (47 %) matched *P. vulgaris*, and 24,183 transcripts (33 %) matched *P. coccineus* (Table 3). However, the sequence identity of the hits in each database showed the highest percentage values (92.3 %, 454 database; 88 %, NCBI EST database) relating to scarlet bean (Table 3). Gene sequences of 26S RNA, CHS, and RPB2 were obtained from all six species and were used in the phylogenetic analysis, which revealed that peanut has a closer relationship with cowpea and scarlet bean (Fig. 3) than with the other three species. To further evaluate the relationship between *A. hypogaea* and the other four model species, OrthoMCL was used to cluster the orthologous groups (gene families) among the five species. In total, 25,936 orthologous groups were clustered from 159,236 protein sequences (Fig. 4). The five species shared 8880 gene families, which may represent conservative genes among plant species. Among grouped

**Fig. 1** The 454 contig length and read length distribution. **a** Comparison of contig length distribution between assembly by Newbler and the sequential assembly with two assemblers. The *blue color* indicates contigs from Newbler and the *red color* indicates contigs from the two assemblers. **b** The 454 read length distribution



gene families, the *A. hypogaea* shared 11,240 (69.33 %) gene families with *G. max*, followed by *P. vulgaris* (10,981; 67.73 %), *M. truncatula* (10,841; 66.87 %), and *A. thaliana* (9372; 57.81 %). There were 4526 (27.92 %) gene families and 46,507 protein sequences unique to *A. hypogaea* (Fig. 4).

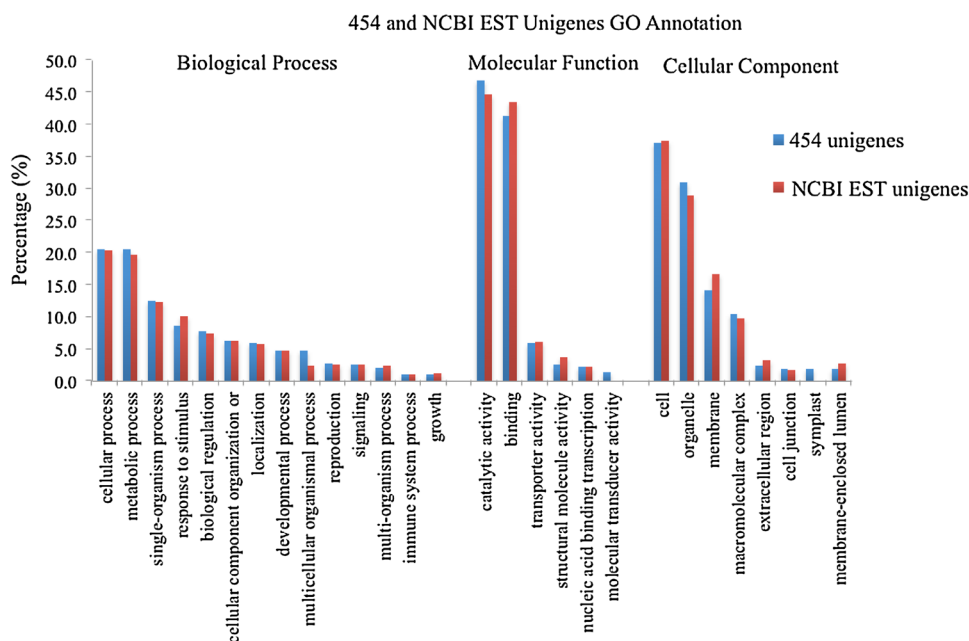
The transcripts assembled from the 454 database were compared to the transcripts assembled from the NCBI EST database via BLAST. In total, 31,399 (71.4 %) transcripts from the 454 database had hits with 15,550 (21.5 %) transcripts in the NCBI EST database (Table S5). The remaining 4913 transcripts from the 454 database with a length of 200 bp or more were considered as novel sequences and

deposited into the DDBJ/EMBL/GenBank (TSA accession ID: GDDN00000000, first version).

### SSR detection

Using the 454 transcript set, 1623 SSRs including 809 di-nucleotide SSRs, 735 tri-nucleotide SSRs, 54 tetra-nucleotide SSRs, 13 penta-nucleotide SSRs, and 12 hexa-nucleotide SSRs were discovered (Table S6). In general, the SSR presence decreased with an increase in the nucleotide number of repeat units. Among the 45 SSR motifs, the most frequent motif type was AG/CT (30.4 %), followed by AAG/CTT (16.6 %), AT/AT (11.5 %) and

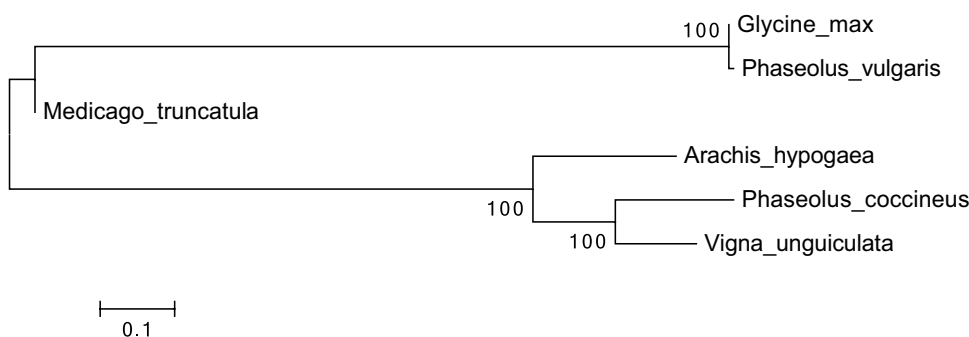
**Fig. 2** Gene ontology term distribution of 454 and NCBI EST transcript sets. The *blue color* indicates the 454 transcript set and the *red color* indicates the NCBI EST transcript set. There are three major Gene Ontology categories, i.e. biological process, molecular function, and cellular component



**Table 3** Comparison of 454 transcripts and NCBI EST transcripts to the gene indexes of five legumes

	Soybean		Medicago		Scarlet bean		Bean		Cowpea	
	454	NCBI EST	454	NCBI EST	454	NCBI EST	454	NCBI EST	454	NCBI EST
No. of unique hits	21,867	46,340	19,258	41,135	11,109	24,183	17,425	34,202	17,683	38,347
Percentage (%)	49.7	64.1	43.8	56.9	25.2	33.4	39.6	47.3	40.2	53.0
Average sequence identity (%)	90.5	87.0	90.0	86.1	92.3	88.0	90.8	86.5	90.0	86.4

**Fig. 3** Phylogenetic tree of six legume species constructed using MEGA6 software



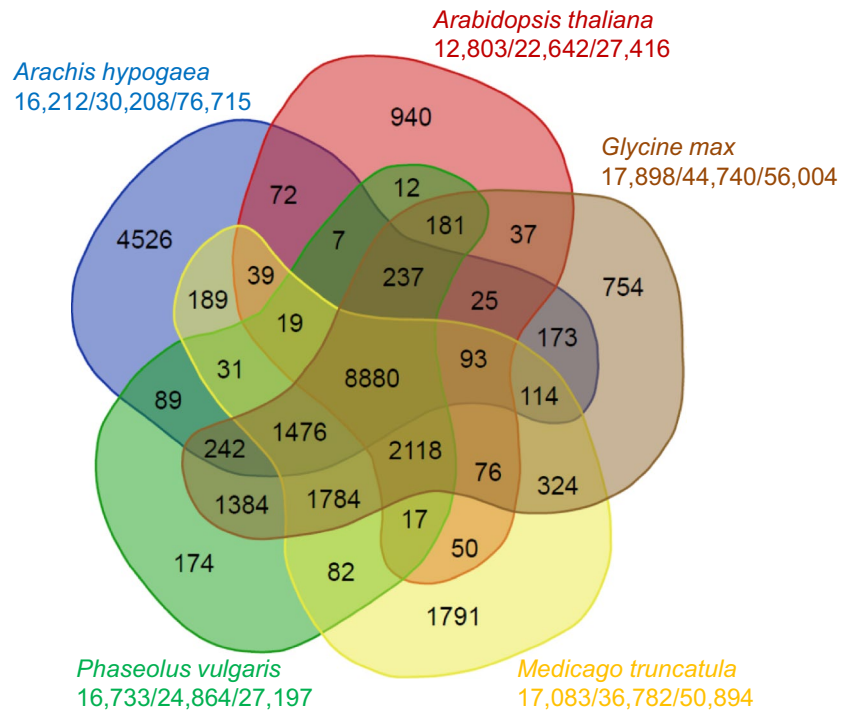
AC/GT (7.6 %) (Table S6). Primers were successfully designed for 838 SSRs consisting of 371 di-nucleotide SSRs, 437 tri-nucleotide SSRs, 20 tetra-nucleotide SSRs, two penta-nucleotide SSRs, and 8 hexa-nucleotide SSRs (Table 1).

Out of the 72,297 transcripts acquired from the NCBI EST transcript set, 7831 (10.8 %) sequences contained SSRs. To eliminate the publicly available SSRs, a comprehensive literature search for *Arachis* SSR markers was performed. This search yielded 14,780 SSR primer pairs (Table S7). The 14,780 primer pair sequences were aligned

to the 7831 transcripts as paired-end sequences. The 2485 transcripts with hits were excluded from further SSR marker development. The remaining 5346 transcripts contained 5894 novel SSRs. Among the 74 SSR motif types, the most frequent was AG/CT (39.6 %), followed by AAG/CTT (15.1 %), AT/AT (11.6 %), and AC/GT (6.5 %) (Table S8). Primers were successfully designed for 5617 SSRs, out of which 3240 were di-nucleotide SSRs, 2163 were tri-nucleotide SSRs, 136 were tetra-nucleotide SSRs, 37 were penta-nucleotide SSRs, and 41 were hexa-nucleotide SSRs (Table 1).



**Fig. 4** A venn diagram describing the distribution of shared gene families among five species. The five species are *Arachis hypogaea*, *Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, and *Phaseolus vulgaris* (OrthoMCL 2.0.7,  $E$  value  $<1e-5$ ). Under species names shows the number of gene families, the gene number in the families, and total number of genes (protein sequences) input into OrthoMCL



In total 6455 novel SSRs were designed with primers. A total of 2775 (42.99 %) SSR-containing transcripts were assigned with GO terms. In terms of biological process, the majority SSR containing transcripts involved in regulation of transcription (246 transcripts), oxidation–reduction process (146 transcripts), and serine family amino acid metabolic process (87 transcripts). For molecular function, the most frequent SSR containing transcripts was ATP binding (224 transcripts), followed by DNA binding (161 transcripts), and zinc ion binding (151 transcripts). Regarding cellular component, most SSR containing transcripts were assigned to nucleus (393 transcripts), plasma membrane (379 transcripts), and mitochondrion (223 transcripts).

#### SSR marker validation

A total of 380 SSR markers were selected for validation. These included 157 di-nucleotide SSRs, 193 tri-nucleotide SSRs, and 30 tetra-, penta- and hexa-nucleotide SSRs (Table 1). In total, 339 (89.2 %) of the selected 380 SSRs were successfully amplified with expected PCR product size. For di-nucleotide SSRs, a polymorphic rate of 11.1 % for 454 database and 12.4 % for NCBI EST database was observed (Table 1, Table S9). From both databases, only seven out of the tri-, tetra-, penta-, and hexa-nucleotide SSRs were polymorphic (Table 1). Sixteen SSRs were polymorphic between Tifrunner and NC3033. Twenty-two (6.5 %) SSRs were polymorphic between at least two of the four genotypes (Table S9).

#### SNP calling

A total of 559 SNPs were detected from the 454 dataset with an average SNP density of one in every 271 bp. Out of these 559 SNPs (three multi-allelic), 383 were transition SNPs and 199 were transversion SNPs. In addition, 45 were shown to have polymorphism between Georgia Green and C7616. From the NCBI EST dataset, 11,343 SNPs were detected. Among these SNPs (21 multi-allelic), 6788 were transitions and 4902 were transversions. The average SNP density was one in every 286 bp. Out of the 11,902 predicted SNPs, 8797 (73.9 %) were located within annotated transcripts.

#### SNP validation

Forty-five SNPs from the 454 database were chosen for validation through Sanger sequencing based on their supposed polymorphism between Georgia Green and C7616 (Table S1). In addition, 65 SNPs (Table S1) from the NCBI EST database were selected for validation. Out of the 110 chosen SNPs (Table 4), 97 were successfully designed with primers (Table S10). Forty-six SNP primer pairs amplified specific bands (Table S10). By comparing the amplicon sequences, 13 SNPs between at least two of the four tetraploid genotypes were confirmed (which are derived from four SNP primer pairs) and these SNP regions all have read depth of more than 30. Forty-two SNP primers yielded PCR products with either no polymorphism between the sequences or only homoeologous SNPs (SNPs within

**Table 4** SNPs selected for PCR amplification and Sanger sequencing validation

Primer name	Ref allele	Alt allele	Dataset	Proposed SNP no.	Observed SNP no.	Comments
SNPW0078	C	T	NCBI EST	4	6	Homoeologous SNPs
SNPW0095	C	T	NCBI EST	3	0	
SNPW0051	G	T	NCBI EST	2	0	
SNPW0053	A	G	NCBI EST	8	22	Homoeologous SNPs
SNPW0057	A	C	NCBI EST	1	13	Homoeologous SNPs
SNPW0058	C	T	NCBI EST	9	24	Homoeologous SNPs
SNPW0059	T	C	NCBI EST	5	5	Homoeologous SNPs
SNPW0060	C	A	NCBI EST	5	4	Homoeologous SNPs
SNPW0062	C	T	NCBI EST	2	13	Homoeologous SNPs
SNPW0066	C	T	NCBI EST	10	0	
SNPW0072	C	T	NCBI EST	5	0	
SNPW0073	A	G	NCBI EST	4	0	
SNPW0079	C	T	NCBI EST	6	11	Homoeologous SNPs
SNPW0083	C	T	NCBI EST	2	13	Homoeologous SNPs
SNPW0084	G	T	NCBI EST	3	0	
SNPW0089	A	G	NCBI EST	1	1	Homoeologous SNPs
SNPW0090	A	G	NCBI EST	14	5	Homoeologous SNPs
SNPW0091	T	C	NCBI EST	27	19	Homoeologous SNPs
SNPW101	T	G	454	1	8	Homoeologous SNPs
SNPW102	A	G	454	1	0	
SNPW106	A	G	454	1	0	
SNPW109	G	A	454	1	4	Homoeologous SNPs
SNPW110	T	C	454	12	noise	
SNPW111	A	G	454	1	1	Homoeologous SNPs
SNPW113	T	A	454	1	noise	
SNPW114	C	G	454	1	noise	
SNPW116	A	G	454	1	0	
SNPW117	C	T	454	1	noise	
SNPW118	G	A	454	1	1	Homoeologous SNPs
SNPW122	G	A	454	1	2	Homoeologous SNPs
SNPW123	T	A	454	1	13	Homoeologous SNPs
SNPW124	C	T	454	1	noise	
SNPW125	A	T	454	1	8	Homoeologous SNPs
SNPW126	T	C	454	1	2	Homoeologous SNPs
SNPW128	G	A	454	4	noise	
SNPW131	G	T	454	3	noise	
SNPW132	G	A	454	5	0	
SNPW0134	T	C	NCBI EST	4	3	Homoeologous SNPs
SNPW0135	C	T	NCBI EST	7	7	Homoeologous SNPs
SNPW0136	C	T	NCBI EST	2	0	
SNPW0137	C	T	NCBI EST	13	3	3 SNPs
SNPW0139	T	A	NCBI EST	11	7	1 SNP + 6 homoeologous SNPs
SNPW0140	A	C	NCBI EST	8	4	3 SNPs + 1 homoeologous SNPs
SNPW0141	A	T	NCBI EST	4	7	6 SNPs + 1 homoeologous SNPs
SNPW0143	C	T	NCBI EST	2	noise	
SNPW0146	G	A	NCBI EST	6	0	

**Table 5** SNP validation comparison between cultivated peanut and the two progenitors *Arachis duranensis* and *Arachis ipaensis*

Allele types from Sanger sequencing			No.	Percentage (%)
<i>Arachis hypogaea</i>	<i>Arachis duranensis</i> (A genome)	<i>Arachis ipaensis</i> (B genome)		
M/N	M	N	116	66.3
M/N	M	M/N	35	20.0
M/N	M/N	M	3	1.7
M/N	M/N	M/N	7	4.0
M/N	M	M	6	3.4
M/N	noise	M	7	4.0
M/N/V	M	N/V	1	0.6
Total			175	100

'M', 'N', and 'V' indicate three different bases among 'A', 'T', 'G', and 'C'. 'M/N' indicates two overlapping peaks at the investigated position. 'M' or 'N' indicates a single peak at the position. 'M/N/V' indicates three overlapping peaks at the position

the genotypes). The amplicon sequences of 20 out of the 42 SNP primer pairs contain 175 homoeologous SNPs. After validation with diploid progenitors, the majority of them (162, 93 %) were confirmed as homoeologous SNPs (Table 5). For 116 (72 %) of these homoeologous SNPs, the wild diploid genotypes were homozygous at the SNP positions. In addition, there were some positions where the A genome progenitor was heterozygous (2 %), B genome progenitor was heterozygous (22 %), or both of them were heterozygous (4 %).

#### SSR allele frequency and genotype frequency of tested peanut breeding lines and germplasm accessions

Eight selected polymorphic markers were used to genotype 142 peanut lines including 32 breeding lines from Florida peanut breeding program and 110 U.S. minicore accessions. The SSR marker AHW0833 yielded 13 unique alleles and 38 unique genotypes from the 142 peanut individuals and generated the highest PIC value (0.943) among the eight SSR markers (Table 2). In contrast, there were only four unique alleles and six unique genotypes for marker AHW064, with the lowest PIC value of 0.619 (Table 2). Five of the SSRs were located at different chromosomes. The location of AHW0833 was unknown. AHW064 and AHW0813 both had hits to A09 chromosome (2.8 Mb apart) (Table 2). A total of 54 alleles were amplified from the eight primer pairs across the 142 genotypes. The average number of unique alleles and genotypes per primer pair of the eight SSRs were 6.75 and 15.60, respectively. Low probability of identity was obtained for most of the markers, ranging from 0.006 (AHW0833) to 0.161 (AHW064). Under the assumption that these eight loci segregate independently and are not in linkage disequilibrium, the combined probability of identity was close to zero ( $10^{-11}$ ). The power of exclusion value was highest for marker AHW0833 (0.891) and lowest for marker AHW064

(0.424) (Table 2). The combined power of exclusion for these eight markers was approximately one (0.99991).

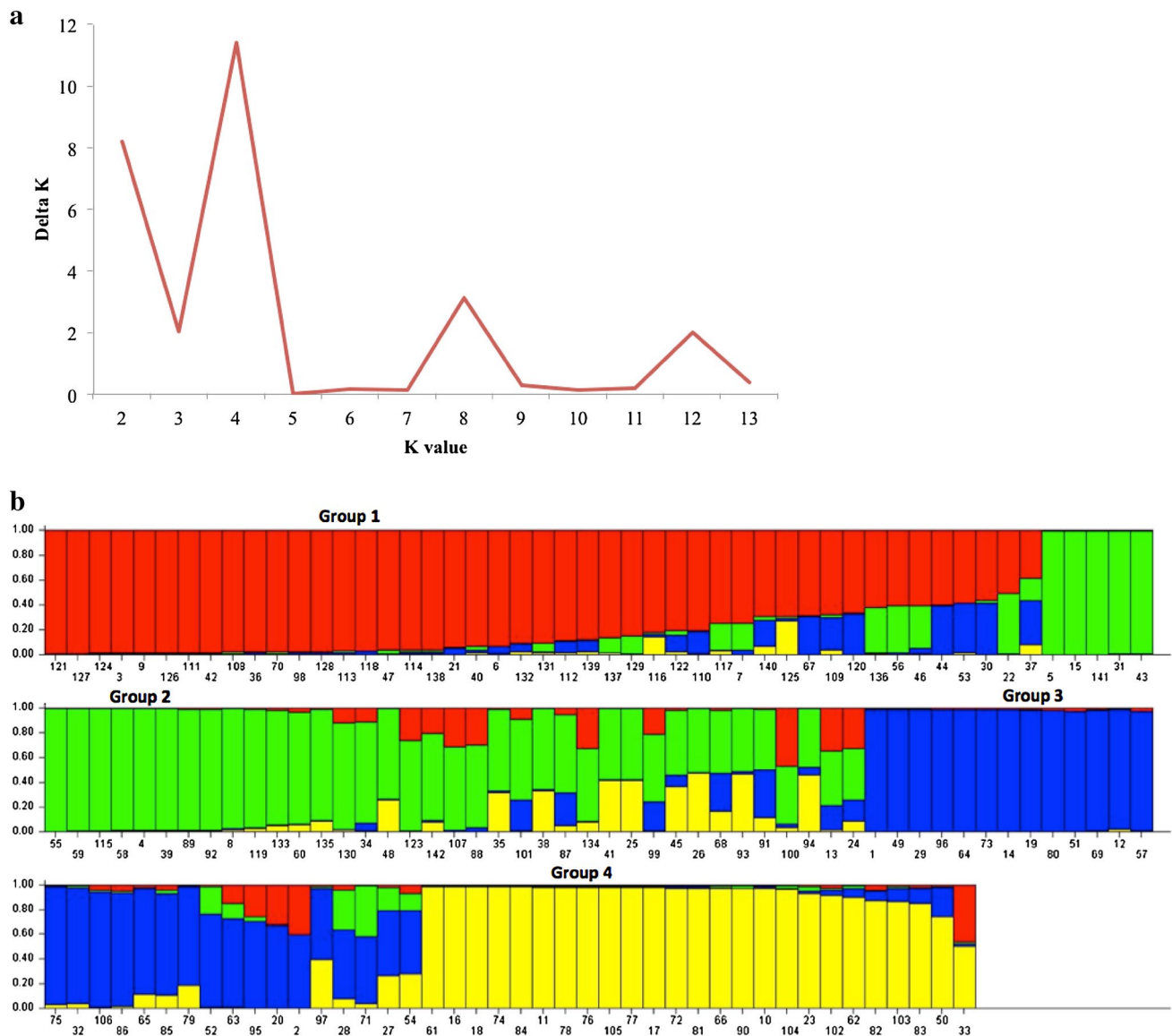
#### Population structure of the peanut mini core and breeding lines

Population structure analysis of the peanut mini core and breeding lines was conducted based on the genotyping results of the 54 SSR alleles. The number of populations ( $K$ ) was determined by the  $L(K)$  and Delta  $K$  values (Table S11). No obvious cutoff was observed when using the  $L(K)$  method; however, an obvious peak was observed at  $K = 4$  when analyzing with Delta  $K$  (Fig. 5). This indicated that four sub-populations can represent the major structure of the U.S. minicore and Florida breeding lines. The individuals assigned to sub-population Groups 2 and 4 are mainly comprised of subspecies *hypogaea* and *fastigiata*, respectively (Table S2). Most of the breeding lines were allocated to sub-population Groups 1 and 2 while the remaining lines were assigned with admixed ancestries (Fig. 5). The phylogenetic tree (Fig. 6) generated from DARwin software also revealed four major sub-groups among the 142 lines.

#### Discussion

##### Sequence assembly and annotation

The first assembly from the 454 database using Newbler generated 2290 contigs with an average length of 425 bp. The CAP3 was used to obtain an integrative assembly. Since these two assemblers utilize different algorithms, the contig number increased to 8942 with an average sequence length of 353 bp (Fig. 1a). The final assembly rate of 87.3 % was comparable to previous studies utilizing *de novo* assembly strategies in *Jatropha* (81.6 %) (Natarajan and Parani 2011) and in blueberry (90.8 %) (Rowland et al.



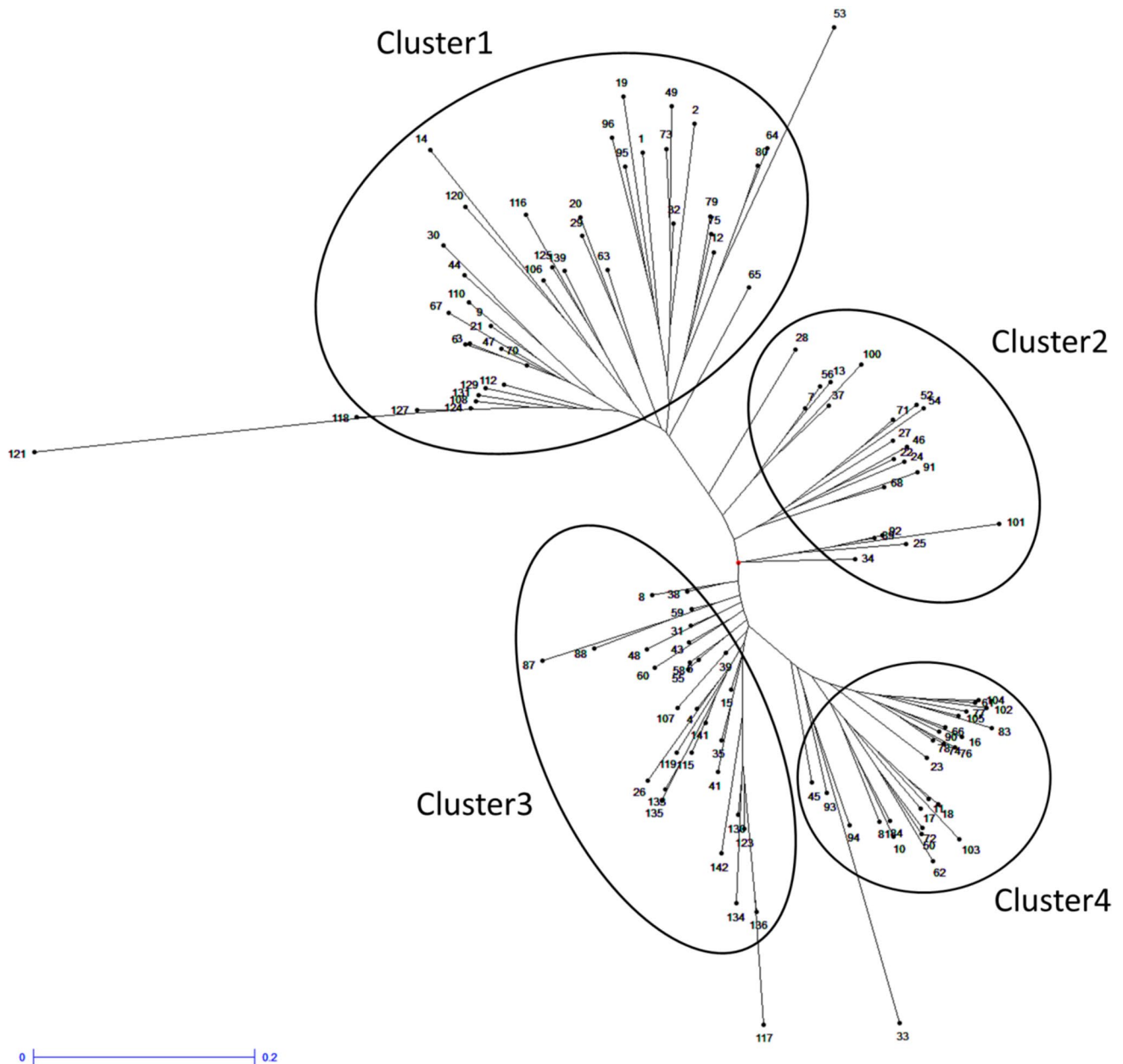
**Fig. 5** **a** Delta  $K$  versus different  $K$  values for determining the best  $K$ .  $X$  axis shows different  $K$  values.  $Y$  axis shows the value of Delta  $K$  corresponding to each  $K$  value; a clear peak occurred when  $K = 4$ . **(b)** Population structure of 142 peanut genotypes generated from STRUCTURE software at  $K = 4$ . The values on the  $y$  axis represent

the proportion of the individual's genetic ancestries that are derived from a certain sub-population. The  $x$  axis indicates the 142 individuals. The accession names and breeding material ID were provided in the Table S2

2012). However, the rate was lower than that observed in a recent study on cultivated peanut (96.7 %, Yin et al. 2013), which may due to the relative shallow sequence depth of this 454 sequence set. The large number (35,065) of singletons in this data set could be attributed to the low coverage, low expression levels, artifacts, or sequencing errors. In contrast to the 454 reads (215 bp), the original NCBI EST entries provided a longer average length (541 bp). Consequently, a better assembly was obtained, which resulted in 21,392 contigs with an average length of 852 bp. The average length of transcripts from the NCBI EST database

was 606 bp which is comparable to several previous studies (619 bp, Zhang et al. 2012; 751 bp, Yin et al. 2013) on cultivated peanut.

The GO annotation results of these two data sets were highly consistent with each other. Similar results on GO annotation of peanut transcript sequences from seeds were reported by Zhang et al. (2012) and Yin et al. (2013). But the presence of a single-organism process, which falls under biological processes, has not been mentioned in previous studies. With no hit in GenBank, a large proportion of transcripts (66.4 % in 454 database; 52.8 % in NCBI EST



**Fig. 6** The phylogenetic tree showing the relatedness of the 142 peanut genotypes

database) was left without annotation, which could be due to peanut-specificity (Yin et al. 2013), or the limitations in the annotation approaches. The 18,508 full-length cDNAs predicted in this study can be further applied in studying gene functions.

The transcript comparison between peanut and five legume species could only reveal the portion of the gene sequences that shared similarity. A proper normalization method is needed to take into account the variance among the database sizes for the five legume species (*G. max*, 75.7 Mb; *M. truncatula*, 53.3 Mb; *V. unguiculata*, 25.6 Mb; *P. vulgaris*, 20.6 Mb; *P. coccineus*, 9.6 Mb). Both of the

databases in this study had more hits with soybean than any of the other species (Table 3), which may due to the large number of sequences in the soybean gene index database instead of a close phylogenetic relationship. This was consistent with the OrthoMCL clustering analysis. The *A. hypogaea* shared more orthologous groups with *G. max* than the other three species (Fig. 4). Based on the sequence similarity of the hits, scarlet bean showed the largest percentage of sequence identity to peanut among the five legume species (Table 3) indicating a relative low divergence between peanut and scarlet bean. This was supported by the phylogenetic analysis revealing that the peanut ESTs were

more similar to scarlet bean and cowpea than to the other three species (Fig. 3). The addition of novel ESTs such as the 4913 presented in this study may aid in the discovery of new genes and their functions and also enable the identification of additional SSR and SNP markers. The 4526 gene families unique to *A. hypogaea* as well as the 46,507 protein sequences that were not grouped should represent species-specific genes. Since the proteomes from several model plant species were used in OrthoMCL clustering analysis, these well-annotated proteomes will provide reliable function information corresponding to the discovered orthologous groups.

### SSR development

The development of polymorphic genetic markers is important in learning about the genetic basis of agronomic traits, population genetic structures, MAS, and gene mapping. Microsatellites or SSR markers are one of the most useful molecular markers and have been used in constructing several peanut genetic maps and in evaluating peanut genetic diversity (Guo et al. 2013). ESTs have become an important resource for developing markers that are associated with biological functions (Gadaleta et al. 2007). This study has contributed 6455 novel EST-SSRs (838 SSRs from the 454 database; 5617 SSRs from the NCBI EST database) (Table S12) with corresponding primers. With the previously available 14,780 SSRs, a total of 21,235 SSRs will now be accessible to the public (a 43.7 % increase).

The di-nucleotide SSR type had a higher frequency in both datasets than any other type. This was consistent with a previous study on dicotyledonous species (Kumpatla and Mukhopadhyay 2005). The most abundant type of di-nucleotide SSR was AG/CT (30.4 %), followed by AT/TA (11.5 %). This pattern was consistent with previous studies on cultivated peanut (Liang et al. 2009; Luo et al. 2005; Proite et al. 2007). AAG/CTT (16.7 %) was the most frequent tri-nucleotide SSR type, which was in agreement with other studies on legumes (Liang et al. 2009; Peakall et al. 1998) and may indicate a common SSR motif feature among legume species. It was suggested by Cardle et al. (2000) and Gao et al. (2003) that the motif AAG/CTT is common in dicots. In monocot species, the most common motif types are AAC/GTT (wheat; Kantety et al. 2002), AGG/CCT (rice; Varshney et al. 2002), and CCG/CGG (barley; Thiel et al. 2003).

### SSR validation

The total SSR amplification rate was 89.2 %. This result was slightly higher than that in some previous studies (83.4 %, Gadaleta et al. 2007; 86.5 %, Liang et al. 2009; 79.9 %, Pandey et al. 2012a, b), and comparable to the amplification

rate of 92 % found in a study by Saha et al. (2004). The amplification rate in this study has demonstrated the success of primer design. The lack of amplification of some primer pairs may be due to a number of reasons: a primer sequence may have overlapped the exon-exon junction; a large intron may be between the primer pair; or there may be sequencing errors in the primer sequences (Nicot et al. 2004). This study's EST-SSR polymorphic rate of 6.5 % was relatively low when compared to the rate from other studies in cultivated peanut such as 7.5 % (Moretzsohn et al. 2005), 8.9 % (Guo et al. 2009), 10.3 % (Liang et al. 2009), 11.5 % (Koilkonda et al. 2012), and 20 % (Luo et al. 2005). The low polymorphic rate may be a result of using only four genotypes in SSR validation. More diverse genotypes can be used to identify polymorphic SSRs from the amplifiable ones. The relatively low rate of polymorphism was observed in tri-nucleotide SSRs. Only five out of 193 tri-nucleotide SSRs were discovered to be polymorphic. In contrast, the polymorphic rate of di-nucleotide SSRs (11.9 %) was relatively high. The other reason for the low polymorphic rate might be related to the genetic backgrounds of the chosen genotypes for SSR validation. Typically cultivated peanut plants have low genetic variability (Kottapalli et al. 2007) even though a high level of phenotypic variation is observed in the field (Barry L Tillman, personal communication, 2014), which may be due to epigenetic modification, gene-gene interactions, or environmental effects.

The lower polymorphic rate of tri-nucleotide SSRs than di-nucleotide SSRs was consistent with results in other studies on *Arachis* (Liang et al. 2009; Proite et al. 2007) and other crops such as wheat (Nicot et al. 2004; Gadaleta et al. 2007), switchgrass (Liu and Wu 2012), barley (Struss and Plieske 1998), soybean (Rongwen et al. 1995), and maize (Smith et al. 1997). One possible explanation is that the location of the di-nucleotide SSRs might be in UTRs (Morgante et al. 2002), which have higher sequence divergence between individuals than the coding regions. However, a lower polymorphic rate in di-nucleotide SSRs than in tri-nucleotide SSRs was also found in some studies on cultivated peanut (Zhao et al. 2012; Koilkonda et al. 2012). In these studies, the lower polymorphic rate for di-nucleotide SSRs may be explained by their deleterious nature. To clarify, a repeat number change of di-nucleotide SSRs is more likely to change the reading frame of the genes (Zhao et al. 2012).

The progenitor reference genomes for cultivated peanut have been released and may eventually lead to significant amounts of genomic SSRs. This study, however, specifically targeted transcribed regions and has produced 6455 (838 + 5617) SSRs. When associating SSR markers to traits of interest, markers from transcribed regions may be preferred for their proximity to genes. The further investigation into the functional annotation of these SSR regions in this study amplifies their usefulness.

## SNP validation

Single nucleotide polymorphisms are widely considered the marker of choice for applications in genetics and genomics studies due to their potent performance in automation, genome-wide coverage, analytical simplicity, and data quality. To date, few studies on peanut SNP marker development have been reported. However, utilization of EST sequences for SNP development have been reported in *Arabidopsis* (Schmid et al. 2003), barley (Kota et al. 2003), sugarcane (Grivet et al. 2003), maize (Batley et al. 2003), and cassava (Lopez et al. 2005). The availability of diverse genotypes in EST databases and the presence of redundant gene sequences have aided EST SNP marker development (Batley et al. 2003). One of the major challenges in peanut SNP marker development is the complex genome composition with two sets of closely related genomes, A and B. In a homozygous peanut, one primer pair could amplify two different fragments of the same size, one from the A genome and the other from the B genome. As expected, most of the SNP positions being investigated presented two different nucleotides. To prove this, an investigation of 175 SNP positions derived from 20 amplified fragments of the two progenitors was conducted. Predominantly, the progenitors were homozygous at the same position and each shared one polymorphic nucleotide with the tetraploid peanut (Table 5). The heterozygosity rate in wild peanut lines is very low, which may be associated with their outcrossing rate. Thus, *A. ipaensis* (PI 468322) should possess a higher outcrossing rate than *A. duranensis* (PI 468321). Occasionally (3.4 %), the progenitors were homozygous (without SNP) at the position of interest, while the tetraploid contained a SNP. This phenomenon may be the result of natural mutations that occurred after cultivated peanut evolved from the hybridization of its progenitors. The low frequency of these cases indicates a relatively short evolutionary course period for cultivated peanut. Only one case was observed in which the tetraploid was tri-allelic (M/N/V) where the A genome progenitor possessed one allele (M), and the B genome progenitor possessed two alleles (N/V) (Table 5).

In peanut, this is a pioneer report on SNP marker development for tetraploid genotypes. The abundance of homoeologous SNPs discovered in this study may benefit the comparative and evolutionary studies between A and B genomes. Due to the highly similar A and B genome sequences in the cultivated peanut, the allelic SNP discovery is very challenging. For allelic SNP discovery, a single sequence source based or NGS-based method will not be enough. It will be necessary to combine with a secondary approach such as amplicon sequencing or calling SNPs based on a segregating population. So far, 13 SNPs have been validated in cultivated peanut. The primer pairs

corresponding to them were all designed for SNPs with a read depth of more than 30. Therefore, a cutoff of 30 is recommended for cultivated peanut SNP calling. In our study, 506 SNPs were called with a depth over 30, which will be further investigated for high-throughput genotyping method development. For SNP marker validation, high-throughput 454 sequencing of amplicons may be more cost- and time effective than Sanger sequencing.

## Allele and genotype frequency

The U.S. peanut mini core accessions coupled with the breeding lines from the Florida Peanut Breeding Program are comprised of individuals with diverse genetic backgrounds. This panel of lines is particularly suited for the evaluation of the discrimination power and genetic information content of newly developed SSR markers. This study's average number of alleles per locus (6.75) was much higher than the 2.99 detected using 46 SSR markers and over 196 cultivars by Ren et al. (2014). The low average number of alleles reported by that study may indicate a low genetic variability in their cultivars from China. He et al. (2003) and Li et al. (2011) also detected a lower number of average alleles per locus, 4.25 and 2.5, respectively. This was at least partially because they used only 24 and 16 genotypes, respectively. In contrast, higher numbers of average alleles per locus have been reported with the utilization of U.S. peanut mini core accessions (7.88, Kottapalli et al. 2007; 10.1, Barkley et al. 2007; 8.1, Wang et al. 2011). This study used only SSR markers derived from genic regions, which may explain the relatively lower average number of alleles per locus.

## Genetic information content parameters

Polymorphism information content was introduced by Botstein et al. (1980) mainly to measure the usefulness of a marker in a linkage analysis. A positive correlation between the number of alleles and the PIC value has been revealed by numerous studies (Siwach et al. 2004; Rabbani et al. 2010; Islam et al. 2012), which was consistent with the results from this study. Marker AHW0833 had the most (13) alleles and consequently displayed the highest PIC value (0.94). The lowest PIC value (0.62) was detected for marker AHW064 with only four alleles. The average PIC (0.783) for the eight markers in this study was higher than that from several similar studies (Kottapalli et al. 2007; Barkley et al. 2007; Wang et al. 2011), indicating a high discrimination power in the eight markers. The probability of identity was low for most cases, ranging from 0.161 (AHW064) to 0.006 (AHW0833). Assuming they are segregating independently, the combined probability of identity will be equal to the product of the values of these eight

loci. A value close to zero indicates that the probability of finding two randomly chosen individuals displaying the same genotype is almost zero. These results illustrated the discrimination power of these markers. The overall power of exclusion for these eight markers was approximately one, indicating they are able to exclude most of the falsely assigned parentages.

### Population structure

Based on the genotyping data from the 54 alleles, four major sub-groups were identified. This result was consistent with a previous study using 96 U.S. peanut mini core accessions (Belamkar et al. 2011). In this study, most of the individuals assigned to Groups 2 and 4 are mainly comprised of subspecies *hypogaea* and *fastigiata*, respectively (Table S2). The remaining individuals used in this study had admixed ancestries. The study by Belamkar et al. (2011) found a similar correlation when characterizing the U.S. peanut mini core collection using SSR markers. Among the four sub-groups discovered, two predominantly consisted of accessions of subspecies *hypogaea* and *fastigiata*, respectively. The other two were classified as “mixed”, containing accessions of different subspecies. Another study specifically detecting structures in Chinese peanut cultivars found that the subpopulation clustering was closely related to germplasm origin (Ren et al. 2014). However, the sub population clustering of the 142 genotypes in this study did not reflect their different origins, indicating the germplasm or breeding lines had similar patterns of genetic composition to cultivated peanut. The structure analysis revealed that about half of the breeding lines from the Florida Peanut Breeding Program had admixed ancestries, also indicating relatively diverse genetic backgrounds in these lines. However, the genetic components of the remaining half of the breeding lines used in this study were mostly assigned to sub-population Groups 1 and 2. To further enrich the gene pool of the breeding program at the University of Florida, it may be worthwhile to select some germplasm collections from sub-population Groups 3 and 4 into the parent pool for breeding.

In summary, this study has enriched the peanut genomic resources, provided more than 6000 novel SSR markers across the whole peanut genomes, investigated the SNP marker development, and explored the challenges of SNP discovery in cultivated peanut. The eight chosen SSR markers demonstrated high discrimination power in germplasm and breeding line evaluation, which will assist association studies and breeding activities. The newly developed SSR markers can also be used for identifying markers linked to agronomically important traits to facilitate the peanut breeding programs. The credentials of SNP discovery in

peanut provided reference for future high-throughput and automated SNP marker development.

**Acknowledgments** The research presented in this article was sponsored by the Florida Peanut Producers Association.

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

### References

- Anderson WF, Holbrook CC, Culbreath AK (1996) Screening the core collection for resistance to tomato spotted wilt virus. *Peanut Sci* 23:57–61
- Barkley NA, Dean RE, Pittman RN, Wang ML, Holbrook CC, Pederson GA (2007) Genetic diversity of cultivated and wild-type peanuts evaluated with M13-tailed SSR markers and sequencing. *Genet Res* 89:93–106
- Batley J, Barker G, O’Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data1. *Plant Physiol* 132:84–91
- Belamkar V, Selvaraj MG, Ayers JL, Payton PR, Puppala N, Burrow MD (2011) A first insight into population structure and linkage disequilibrium in the US peanut minicore collection. *Genetica* 139:411–429
- Beute MK, Wynne JC, Emery DA (1976) Registration of NC 3033 peanut germplasm. *Crop Sci* 16:887
- Bi Y, Liu W, Xia H, Su L, Zhao C, Wan S, Wang X (2010) EST sequencing and gene expression profiling of cultivated peanut (*Arachis hypogaea* L.). *Genome* 53:832–839
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331
- Branch WD (1996) Registration of ‘Georgia Green’ peanut. *Crop Sci* 36:806
- Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156:847–854
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676
- Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA miniprep: version II. *Plant Mol Biol Rep* 1:19–21
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* 14:2611–2620
- Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 7:574–578
- Feng S, Wang X, Zhang X, Dang PM, Holbrook C, Culbreath AK, Wu Y, Guo B (2012) Peanut (*Arachis hypogaea*) expressed sequence tag project: progress and application. *Comp Funct Genomics*. doi:10.1155/2012/373768
- Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ (2011) Using OrthoMCL to



- assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinform* Chapter 6: Unit 6.12.1-19
- Gadaleta A, Mangini G, Mulè G, Blanco A (2007) Characterization of dinucleotide and trinucleotide EST-derived microsatellites in the wheat genome. *Euphytica* 153:73–85
- Gao L, Tang J, Li H, Jia J (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol Breed* 12:245–261
- Grivet LL, Glaszmann J-C, Vincentz MM, da Silva FF, Arruda PP (2003) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. *Theor Appl Genet* 106:190–197
- Guo B, Chen X, Dang P, Scully BT, Liang X, Holbrook CC, Yu J, Culbreath AK (2008) Peanut gene expression profiling in developing seeds at different reproduction stages during *Aspergillus* parasiticus infection. *BMC Dev Biol* 8:12
- Guo B, Chen X, Hong Y, Liang X, Dang P, Brennen T, Holbrook CC, Culbreath A (2009) Analysis of gene expression profiles in leaf tissues of cultivated peanuts and development of EST-SSR markers and gene discovery. *Int J Plant Genomics*. doi:10.1155/2009/715605
- Guo BZ, Pandey M, He GH, Zhang XY, Liao BS, Culbreath A, Varshney R, Nwosu V, Wilson R, Stalker T (2013) Recent advances in molecular genetic linkage maps of cultivated peanut (*Arachis hypogaea* L.). *Peanut Sci* 40:95–106
- Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. *Mol Biol Evol* 30:1229–1235
- He G, Meng R, Newman M, Gao G, Pittman RN, Prakash CS (2003) Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). *BMC Plant Biol* 3:3
- Holbrook CC, Culbreath AK (2007) Registration of ‘Tifrunner’ Peanut. *J Plant Registr* 1:124
- Holland MM, Parson W (2011) GeneMarker® HID: a reliable software tool for the analysis of forensic STR data. *J Forensic Sci* 56:29–35
- Hopkins MS, Casa AM, Wang T, Mitchell SE, Dean RE, Kochert GD, Kresovich S (1999) Discovery and characterization of polymorphic simple sequence repeats (SSRs) in peanut. *Crop Sci* 39:1243–1247
- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* 9:868–877
- Hytien DL, Song Q, Fickus EW, Quigley CV, Lim J, Choi I, Hwang E, Pastor-Corrales M, Cregan PB (2010) High-throughput SNP discovery and assay development in common bean. *BMC Genom* 11:475
- Islam S, Haque MS, Emon RM, Islam MM, Begum SN (2012) Molecular characterization of wheat (*Triticum aestivum* L.) genotypes through SSR markers. *Bangladesh. J Agric Res* 37(3):389–398
- Jamieson A (1994) The effectiveness of using co-dominant polymorphic allelic series for (1) checking pedigrees and (2) distinguishing full-sib pair members. *Anim Genet* 1:37–44
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48:501–510
- Kirst M, Cordeiro CM, Rezende G, Grattapaglia D (2005) Power of microsatellite markers for fingerprinting and parentage analysis in *Eucalyptus grandis* breeding populations. *J Hered* 96:161–166
- Koilkonda P, Sato S, Tabata S, Shirasawa K, Hirakawa H, Sakai H, Sasamoto S, Watanabe A, Wada T, Kishida Y, Tsuruoka H, Fujishiro T, Yamada M, Kohara M, Isobe S, Suzuki S, Hasegawa M, Kiyoshima H (2012) Large-scale development of expressed sequence tag-derived simple sequence repeat markers and diversity analysis in *Arachis* spp. *Mol Breed* 30:125–138
- Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Mol Genet Genomics* 270:24–33
- Kottapalli KR, Burow MD, Burow G, Burke J, Puppala N (2007) Molecular characterization of the US peanut mini core collection using microsatellite markers. *Crop Sci* 47:1718–1727
- Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48:985–998
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucl Acids Res* 40:D1202–D1210
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
- Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079
- Li Y, Chen CY, Knapp SJ, Culbreath AK, Holbrook CC, Guo BZ (2011) Characterization of simple sequence repeat (SSR) markers and genetic relationships within cultivated peanut (*Arachis hypogaea* L.). *Peanut Sci* 38:1–10
- Liang X, Chen X, Hong Y, Liu H, Zhou G, Li S, Guo B (2009) Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. *BMC Plant Biol* 9:35
- Liu L, Wu Y (2012) Development of a genome-wide multiple duplex-SSR protocol and its applications for the identification of selfed progeny in switchgrass. *BMC Genom* 13:522
- Lopez C, Piégue B, Cooke R, Delseny M, Tohme J, Verdier V (2005) Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). *Theor Appl Genet* 110:425–431
- Luo M, Dang P, Guo BZ, He G, Holbrook CC, Bausher MG, Lee RD (2005) Generation of expressed sequence tags (ESTs) for gene discovery and marker development in cultivated peanut. *Crop Sci* 45:346–353
- Mallikariuna N, Varshney RK (2014) Molecular markers, genetic maps and QTLs for molecular breeding in peanut. In: Mallikariuna N, Varshney RK (eds) *Genetics, genomics and breeding of peanuts*. CRC Press, Boca Raton, pp 79–113
- Milbourne D, Meyer R, Bradshaw J, Baird E, Bonar N, Provan J, Powell W, Waugh R (1997) Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. *Mol Breed* 3:127–136
- Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucl Acids Res* 33:W677–W680
- Moretzsohn MC, Leoi L, Proite K, Guimarães PM, Leal-bertioli SCM, Gimenes MA, Martins WS, Valls JFM, Grattapaglia D, Bertioli DJ (2005) A microsatellite-based, gene-rich linkage map for the AA genome of *Arachis* (Fabaceae). *Theor Appl Genet* 111:1060–1071
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30:194–200
- Nagy E, Guo S, Khanal S, Taylor C, Ozias-Akins P, Stalker HT, Nielsen N (2010) Developing a high density molecular map of the A-genome species *A. duranensis*. In: proceedings of American Peanut Research and Education Society (APRES), 12–15th July, Florida, USA
- Nagy S, Poczai P, Cernák I, Gorji AM, Hegeds G, Tallér J (2012a) PICcalc: an online program to calculate polymorphic

- information content for molecular genetic studies. *Biochem Genet* 50:670–672
- Nagy ED, Guo Y, Tang S, Bowers JE, Okashah RA, Taylor CA, Zhang D, Khanal S, Heesacker AF, Khalilian N, Farmer AD, Carrasquilla-Garcia N, Penmetsa RV, Cook D, Stalker HT, Nielsen N, Ozias-Akins P, Knapp SJ (2012b) A high-density genetic map of *Arachis duranensis*, a diploid ancestor of cultivated peanut. *BMC Genom* 13:469
- Natarajan P, Parani M (2011) De novo assembly and transcriptome analysis of five major tissues of *Jatropha curcas* L. using GS FLX titanium platform of 454 pyrosequencing. *BMC Genom* 12:191
- Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy P, Bernard M, Sourdille P (2004) Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). *Theor Appl Genet* 109:800–805
- Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* 4:347–354
- Pan YB (2006) Highly polymorphic microsatellite DNA markers for sugarcane germplasm evaluation and variety identity testing. *Sugar Tech* 8(4):246–256
- Pandey MK, Gautami B, Jayakumar T, Sriswathi M, Upadhyaya HD, Varshney RK, Gowda MVC, Radhakrishnan T, Bertioli DJ, Knapp SJ, Cook DR (2012a) Highly informative genic and genomic SSR markers to facilitate molecular breeding in cultivated groundnut (*Arachis hypogaea*). *Plant Breeding* 131:139–147
- Pandey MK, Nigam SN, Upadhyaya HD, Janila P, Varshney RK, Monyo E, Ozias-Akins P, Liang X, Guimarães P, Zhang X, Guo B, Cook DR, Michelmore R, Bertioli DJ (2012b) Advances in *Arachis* genomics for peanut improvement. *Biotechnol Adv* 30:639–651
- Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998) Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. *Mol Biol Evol* 15:1275–1287
- Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (eds) Genetic diversity of cultivated tropical plants. Science Publishers, Montpellier, pp 43–76
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Proite K, Leal-Bertioli SC, Bertioli DJ, Moretzsohn MC, Da Silva FR, Martins NF, Guimarães PM (2007) ESTs from a wild *Arachis* species for gene discovery and marker development. *BMC Plant Biol* 7:7
- Quackenbush J, Liang F, Holt I, Perteu G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucl Acids Res* 28:141
- Rabbani MA, Masood MS, Shinwari ZK, Yamaguchi-Shinozaki K (2010) Genetic analysis of basmati and non-basmati Pakistani rice (*Oryza sativa* L.) cultivars using microsatellite markers. *Pak J Bot* 42(4):2551–2564
- Ren X, Jiang H, Yan Z, Chen Y, Zhou X, Huang L, Lei Y, Huang J, Yan L, Qi Y, Wei W, Liao B (2014) Genetic diversity and population structure of the major peanut (*Arachis hypogaea* L.) cultivars grown in China by SSR markers. *Plos One* 9(2):e88091
- Rongwen J, Akkaya Bhagwat AA, Lavi U, Cregan PB (1995) The use of microsatellite DNA markers for soybean genotype identification. *Theor Appl Genet* 90:43–48
- Rosenberg NA, Burke T, Eloe K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Mäki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K, Weigend S (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159:699–713
- Rowland LJ, Alkharouf N, Darwish O, Ogden EL, Polashock JJ, Bassil NV, Main D (2012) Generation and analysis of blueberry transcriptome sequences from leaves, developing fruit, and flower buds from cold acclimation through deacclimation. *BMC Plant Biol* 12:46
- Saha MC, Mian MA, Rouf Eujayl I, Zwonitzer JC, Wang L, May GD (2004) Tall fescue EST-SSR markers with transferability across several grass species. *Theor Appl Genet* 109:783–791
- Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, Weissshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res* 13(6):1250–1257
- Schmidt M (2007) Transformation and functional genomics in legumes. International workshop on advances in arachis through genomics and biotechnology. Atlanta, Georgia
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183
- Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Gefroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS, Jackson SA (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet* 46:707–713
- Siwach P, Jain S, Saini N, Chowdhury VK, Jain RK (2004) Allelic diversity among basmati and non-basmati long-grain Indica rice varieties using microsatellite markers. *J Plant Biochem Biotechnol* 13:25–32
- Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegler J (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree. *Theor Appl Genet* 95:163–173
- Song GQ, Li MJ, Xiao H, Wang XJ, Tang RH, Xia H, Zhao CZ, Bi YP (2010) EST sequencing and SSR marker development from cultivated peanut (*Arachis hypogaea* L.). *Electron J Biotechnol*. doi:10.2225/vol13
- Struss D, Plieske J (1998) The use of microsatellite markers for detection of genetic diversity in barley populations. *Theor Appl Genet* 97:308–315
- Tamura K, Stecher G, Peterson D, Filipowski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30:2725–2729
- Thiel TT, Michalek WW, Varshney RR, Graner AA (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 106:411–422
- Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178–192
- Tillman BL, Stalker HT (2009) Peanut. In: Vollmann J, Rajcan I (eds) Oil Crops, handbook of plant breeding 4. Springer, New York, pp 287–315
- Upadhyaya HD (2005) Variability for drought resistance related traits in the mini core collection of peanut. *Crop Sci* 45:1432–1440

- Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett* 7:537–546
- Varshney RK, Mohan SM, Gaur PM, Gangarao NVPR, Pandey MK, Bohra A, Sawargaonkar SL, Chitikineni A, Kimurto PK, Janila P, Saxena KB, Fikre A, Sharma M, Rathore A, Pratap A, Tripathi S, Datta S, Chaturvedi SK, Mallikarjuna N, Anuradha G, Babbar A, Choudhary AK, Mhase MB, Bharadwaj C, Mannur DM, Harer PN, Guo B, Liang X, Nadarajan N, Gowda CLL (2013) Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnol Adv* 31:1120–1134
- Wang ML, Sukumaran S, Barkley NA, Chen Z, Chen CY, Guo B, Pittman RN, Stalker HT, Holbrook CC, Pederson GA, Yu J (2011) Population structure and marker-trait association analysis of the US peanut (*Arachis hypogaea* L.) mini-core collection. *Theor Appl Genet* 123:1307–1317
- Yin D, Wang Y, Zhang X, Li H, Lu X, Zhang J, Zhang W, Chen S (2013) De novo assembly of the peanut (*Arachis hypogaea* L.) seed transcriptome revealed candidate transcripts for oil accumulation pathways. *Plos One* 8(9):e73767
- You FM, Huo N, Gu YQ, Luo M, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinform* 9:253
- Young ND, Debellé F, Oldroyd GED et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480:520–524
- Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, Zhang Q, Liang X, Li Y (2012) De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in peanut (*Arachis hypogaea* L.). *BMC Genom* 13:90
- Zhao Y, Prakash CS, He G (2012) Characterization and compilation of polymorphic simple sequence repeat (SSR) markers of peanut from public database. *BMC Res Notes* 5:362
- Zhou X, Xia Y, Ren X, Chen Y, Huang L, Huang S, Liao B, Lei Y, Yan L, Jiang H (2014) Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genom* 15:351
- Zimmer EA, Wen J (2012) Using nuclear gene data for plant phylogenetics: progress and prospects. *Mol Phylogenet Evol* 65(2):774–785