Theor Appl Genet (2013) 126:1145–1149
DOI 10.1007/s00122-013-2042-8

ORIGINAL PAPER

De novo next-generation sequencing, assembling and annotation of *Arachis hypogaea* L. Spanish botanical type whole plant transcriptome

Ning Wu · Kanyand Matand · Huijuan Wu · Baoming Li · Yue Li · Xiaoli Zhang · Zheng He · Jialin Qian · Xu Liu · Stephan Conley · Marshall Bailey · George Acquaah

Received: 8 August 2012 / Accepted: 9 January 2013 / Published online: 22 January 2013

Abstract Peanut is a major agronomic crop within the legume family and an important source of plant oil, proteins, vitamins, and minerals for human consumption, as well as animal feed, bioenergy, and health products. Peanut genomic research effort lags that of other legumes of economic importance, mainly due to the shortage of essential genomic infrastructure, tools, resources, and the complexity of the peanut genome. This is a pioneering study that explored the peanut Spanish Group whole plant transcriptome and culminated in developing unigenes database. The study applied modern technologies, such as, normalization and next-generation sequencing. It overall sequenced 8,308,655,800 nucleotides and generated 26,048 unigenes amongst which 12,302 were annotated and 8,817 were characterized. The remainder, 13,746 (52.77 %) unigenes, had unknown functions. These results will be applied as the reference transcriptome sequences for expanded transcriptome sequencing of the remaining three peanut botanical types (Valencia, Runner, and Virginia), which is currently in progress, RNA-seq, exome identification, and genomic markers development. It will also provide important tools and resources for other legumes and plant species genomic research.

Introduction

Peanut is an important crop member of the legume family and a major source of plant oil, proteins, essential vitamins and minerals that can be used for human consumption, animal feed, bioenergy, and health products (Higgs 2002; Li et al. 2010, 2011). Besides its economic importance, peanut is a notorious food source of allergens that can cause human lethal allergic reactions to hypersensitive people (Cianferoni and Muraro 2012).

Generally, cultivated peanut is categorized into four botanical types, referred to as, Spanish, Valencia, Virginia, and Southeast Runner that represent related genetic variabilities (Krapovickas and Gregory 1994). However, genetic variations amongst cultivated peanut groups are reportedly very limited (He and Prakash 2001; Burow et al. 2001), which has significantly exacerbated the slow pace of developing the physical map and global genomic resources in cultivated peanut (Moretzsohn et al. 2009). Considering that there are overtly major morphological differences amongst cultivated peanut groups, high quality of total transcriptome sequencing with related intra and intergroup characterizations could lead toward enriching related genetic differences to currently available genomic resources and information. Thus, high-quality global sequencing and characterization of the whole plant transcriptome of individual cultivated group and gene profiling, as currently pursued at Langston University, are strongly encouraged in peanut; and this study is the first major step into globally quantifying genetic variations amongst those botanical groups. Further, it is critical to significantly develop

Communicated by H. T. Nguyen.

N. Wu (✉) · K. Matand (✉) · S. Conley · M. Bailey · G. Acquaah
Center for Biotechnology Research and Education,
Langston University, Langston, OK 73050, USA
e-mail: nwu@langston.edu

K. Matand
e-mail: kmatand@langston.edu

H. Wu · B. Li · Y. Li · X. Zhang · Z. He · J. Qian · X. Liu
Beijing Center for Physical and Chemical Analysis,
100089 Beijing, China

specific genomic resources and information in peanut to fully understand its uniqueness that cannot be explained by merely using model crops' genomic resources.

While significantly intensive genomic research has been conducted and considerable progress achieved in model legumes such as soybean (Wilson and Grant 2010; Severin et al. 2010; Woody et al. 2011) and medicago (Cannon et al. 2005; Bell et al. 2001), relatively limited progress has been achieved in peanut (Zhang et al. 2012), considering that milestone genomic resources, such as, physical genome map and whole genome characterization and functional analysis studies have not been systemically developed in cultivated peanut. In addition to advanced technological deficiencies, intrinsic genetic factors such as little genetic variations amongst cultivated groups, differential genomic make-up of the cultivated allotetraploid species, and large genome size similar to humans justify the slow pace for developing robust-global peanut genomic resources. The National Center for Biotechnology Information (NCBI) Expressed Sequence Tag (EST) database search results of July 6, 2012 showed that only 246,733 general records of peanut ESTs, comparing to, 1,529,920 and 10,442,386 similar records for soybean and human, respectively, were available. Besides the smaller number, publicly available peanut ESTs are significantly disparate that their global impact in deciphering a significant portion of genome's genetic information of cultivated peanut is very limited. This is underpinned by the fact that most, if not all, cultivated peanut ESTs are generated based upon individually narrow genetic research frameworks, which are generally built around individual organ, such as seed (Zhang et al. 2012; Bi et al. 2010), or individual agronomic trait, such as, biotic or abiotic afflictions (Feng et al. 2012), trait-based molecular markers (Feng et al. 2012), and oil content (Zhang et al. 2012), etc. Thus, there is no systematic and comprehensive analysis of peanut gene expression profile for different cultivated peanut botanical groups. EST sequencing can provide a robust sequence resource that can be exploited for gene discovery, genome annotation, and comparative genomic studies (Adams et al. 1991). As an alternative and supplement to the whole genome sequencing, continuous development of strategic peanut ESTs remains a current priority in peanut genomic research.

This investigation reports, for the first time, the whole plant's globally expressed genomic information, which is not trait-based, in cultivated peanut botanical groups, in general, and Spanish group, in particular. It probed for all active genes in young and mature whole peanut plants, and has resulted in developing the first most comprehensive unigene database that represents the whole Spanish peanut plant's genetic content.

Materials and methods

Plant material preparation and sampling

Seeds of the peanut Spanish group variety, Spanco, used in this study were supplied by the Agricultural Experimental Station of the Oklahoma State University (Stillwater, Oklahoma). Because of the lack of synchronized plant organ formation and to ensure that most organ genes are broadly represented, organ tissues were collected twice at ten-day-old and at plant maturation. At ten-day-old, only leaf, stem, and root tissues were collected; whereas at the maturation stage all organs were represented, including, leaf, stem, root, flower, peg, and pod (variable growth stages ranging from early formation to seed-filled stage) (Boote 1982). Tissues from the two sample sets were used to cover all the genes expressed during plant growth and development.

Peanut plant total RNAs isolation and mRNAs purification

Total RNA of individual peanut organ was isolated, respectively, using QIAGEN Total RNA Isolation Kit (QIAGEN, Valencia, CA) according to the manufacturer's instruction. All total RNAs were pooled together for mRNA purification. The latter was purified with QIAGEN Oligotex beads (QIAGEN, Valencia, CA) also following the manufacturer's manual.

Normalized cDNA library construction

The normalization process was carried out on mRNA level. Briefly, first strand cDNA was synthesized by using SuperScript® III First-Strand Synthesis SuperMix (Life Technologies, Carlsbad, CA) in situ on QIAGEN's dT-Oligotex beads. After block extra dT in the beads with oligo-dA, mRNAs were hybridized with first strand cDNA on the dT-Oligotex beads according to the re-association kinetics-based approach (Bonaldo et al. 1996). Normalized mRNAs were, first, separated from beads by centrifugation followed by ethanol precipitation, then, characterized by RT-PCR with primers of Actin depolymerizing factor-like cDNA, prior to library construction. Then, normalized cDNA library was constructed by applying SuperScript® Plasmid System (Life Technologies, Carlsbad, CA) and Oligo-dT-Not I primer following the manufacturer's instruction. Further, synthesized cDNAs were cloned into pCMV Sport vector and transformed into DH10B electrocompetent cells. Platting assay and colony PCR were performed for normalized cDNA library quality control purpose.

De novo next-generation DNA sequencing

Normalized peanut whole plant cDNA library was PCR-amplified with PrimeSTAR GXL DNA Polymerase (Takara Bio Inc., Shiga, Japan) and T7 and SP6 standard primers in 50 µl reaction volume at 95 °C for 2 min 30 s; 35 cycles at 94 °C for 15 s, 55 °C for 15 s, and 72 °C for 2 min 30 s; and then 72 °C for 10 min. Double strand cDNAs were then purified with the QIAquick PCR Purification Kit (QIAGEN, Valencia, CA) followed by fragmentation with the Biorupter UCD-300 sonication device (Diagenode, Liège, Belgium). Sequencing library was prepared from sheared cDNAs with the average size of 150 bp by applying the NEBNext® DNA Library Preparation Kit (NEB, Ipswich, MA). The library was sequenced from both directions on HiSeq 2000 System (illumina, San Diego, CA) with 100 bp of data collected per run by applying TruSeq PE Cluster and TruSeq SBS Kits (illumina, San Diego, CA). Data analysis and base calling were achieved by applying the illumina instrument software.

Bioinformatics analysis

Transcriptome data processing and assembly were performed with the application of modified Velvet to construct unique consensus sequences (Zerbino and Birney 2008). The gene expression profile was developed by mapping trimmed transcriptome reads onto the unique consensus sequences using SOAP2 (Li et al. 2009). The unigenes were identified by sequence similarity comparison against both the SWISS-PROT database of the European Bioinformatics Institute (http://www.ebi.ac.uk/uniprot) (Altschul and Gish 1996) by applying BLAST at the cutoff E value $\leq$1e$-$10 and the NCBI non-redundant nucleotide and non-redundant protein databases (http://blast.ncbi.nlm.nih.gov/Blast.cgi) where BLASTN and BLASTX were applied, respectively, with the same cutoff E value $\leq$1e$-$5 (Altschul et al. 1997). Functional annotation and classification of resulted unigenes were performed by sequence similarity comparison against NCBI's clusters of orthologous groups (COG) database (http://www.ncbi.nlm.nih.gov/COG) (Tatusov et al. 2003) and SWISS-PROT database with BLAST at the same cutoff E value $\leq$1e$-$10, respectively. For assembled genes, whose sequence comparison results matched the records of other non-plant organisms, additional manual analysis based on score, coverage, identity, and E value were performed.

Results

This study depicts succinctly, for the first time, the fully expressed genome information of the Spanish peanut whole plant. The results showed that the normalized peanut whole plant cDNA library consisted of 2.01 × 10^5 clones with the recombinant rate of 87.5 % and average insert size of about 1,000 bp. This library was PCR-amplified, and the related product was sheared mechanically to generate the mixture of 150 bp fragments for HiSeq 2000 system sequencing.

Overall, the next-generation sequencing generated 41,543,279 reads from 8,308,655,800 bp. The sequencing quality scores of Q30 (99.9 % base call accuracy) were achieved across all bases in forward direction sequencing and from 1 to 84 bases in reverse direction sequencing. The de novo assembly of all sequencing data using the modified Velvet program was followed by the application of the SOAP2 program for unigene identification that culminated in generating 26,048 unigenes of the average sequence length of 550 bp and sequence length range of 201–2,431 bp (Fig. 1), that represent the total sequence length of 14,305,500 bp of the peanut genome. All assembled unigene sequences were annotated by sequence similarity comparisons against both SWISS-PROT and NCBI NR nucleotide and protein databases.

The functional annotation of all unigenes showed that 12,302 and 8,817 unigenes matched the records in SWISS-PROT database (version 20120304) and COG database (Table 1), respectively. Further proceeding showed that the analysis of COG's classification of all known function unigenes resulted in categorizing them into 23 clusters based on their common functions, while the functions of

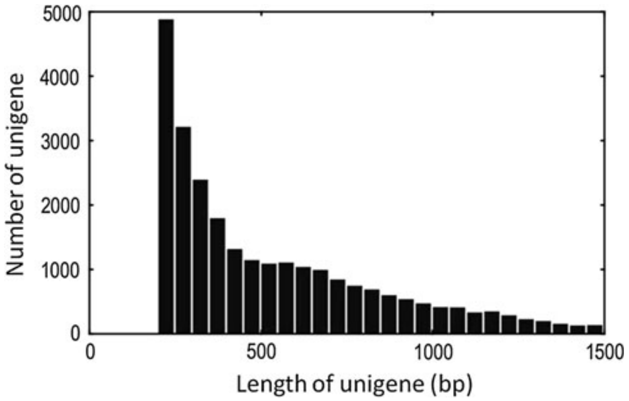

Fig. 1 Sequence length of 26,048 assembled peanut unigenes

Table 1 Unigene annotation

	Total		E value cut off	Database version
	Annotated unigenes	%		
Total unigene	26,048	100.00	–	–
SWISS-PROT	12,302	47.23	1e$-$10	20120304
COG	8,817	33.85	1e$-$10	–

Table 2 COG classification

Classification	Number of genes	Percentage (%)
Translation, ribosomal structure and biogenesis	1,378	15.63
Carbohydrate transport and metabolism	1,129	12.8
Energy production and conversion	1,127	12.78
General function prediction only	1,091	12.37
Post-translational modification, protein turnover, chaperones	795	9.02
Amino acid transport and metabolism	769	8.72
Lipid transport and metabolism	495	5.61
Inorganic ion transport and metabolism	318	3.61
Coenzyme transport and metabolism	217	2.46
Function unknown	209	2.37
Cell wall/membrane/envelope biogenesis	209	2.37
Secondary metabolites biosynthesis, transport and catabolism	174	1.97
Transcription	153	1.74
Nucleotide transport and metabolism	141	1.6
Signal transduction mechanisms	140	1.59
Intracellular trafficking, secretion, and vesicular transport	108	1.22
Cytoskeleton	94	1.07
Replication, recombination and repair	89	1.01
Chromatin structure and dynamics	66	0.75
Cell cycle control, cell division, chromosome partitioning	49	0.56
Defense mechanisms	42	0.48
RNA processing and modification	14	0.16
Cell motility	7	0.08
Nuclear structure	3	0.03
Total	8,817	100.00

209 genes remained unknown (Table 2). Overall, there still were 13,746 (52.77 %) unigenes that had no significant match with current records in either SWISS-PROT or NCBI databases.

Discussion

The peanut tissues used for this study were collected in two growing stages to cover all peanut organs and potential gene differential expressions in young and mature stages. A normalization strategy was applied to the peanut whole plant mRNA for library construction to enhance gene discovery by reducing redundant genes number and relatively increasing the rare gene copies in the cDNA library (Soares et al. 1994). Based upon previous peanut EST studies, normalizing of expressed genes prior to high-throughput sequencing process, has generated greater reads variety than using regular cDNA library (Bi et al. 2010; Huang et al. 2012; Koilkonda et al. 2012). The next-generation sequencing technology was applied in this study to maximize generating mRNA sequencing data of about 8.3G bp, from the peanut whole plant transcriptome. It represents about 250× sequencing depth of the peanut transcriptome, based on the final assembling results. Such sequencing depth provides a greater number of repeat sequence reads for individual gene and greater accuracy and reliability of sequencing data, especially during the assembling process. Although HiSeq2000 generates only 100 bp for each read, based upon the instrumental limitation, the sheared sequencing library with the size of about 150 bp was constructed to facilitate overlaps sequencing from both directions. By combining both greater bidirectional sequencing depth and overlaps sequencing strategy, we were able to generate unigenes sequencing data of great quality and reliability, even without available reference genome sequences. Although the assembled gene sizes varied from 201 to 2,431 bp, which might not represent the full-length of all genes in the database, the current minimum gene size is satisfactory for future EST probe applications (Koilkonda et al. 2012).

The resulting 41,543,279 reads were assembled into 26,048 contigs, creating a foundational reference transcriptome for the peanut Spanish group whole plant. About 47 % of this reference transcriptome was annotated and about 34 % was functionally characterized (Table 2). This study resulted in the construction of the first comprehensive peanut Spanish botanical type whole plant unigene database with the total of 26,048 individual records. Irrespective of the genotypic differences of experimental plant materials used, a recently reported peanut study (Zhang et al. 2012) and this investigation's annotation results were similar; the number of annotated unigenes was 8,252 and 8,817, respectively. This study is being followed up by similar research on the remaining three peanut botanical types (Valencia, Runner, and Virginia), which has also been funded by the United States Department of Agriculture (USDA).

Finally, by applying normalization and next-generation sequencing technologies, we successfully constructed the first peanut Spanish botanical type whole plant unigenes database. The resulted peanut reference transcriptomic sequences will be used for our ongoing expanded transcriptome sequencing project for the remaining three peanut botanical types. Upon the completion of all transcriptome sequencing, assembling, and annotating for all the four peanut botanical groups with related intergroup comparisons, the most comprehensive peanut whole plant unigene database will be finalized and applied more accurately for cross-species comparisons with other publicly available legumes genome databases and, then, released for public access, in time for its application to benefit the International Peanut Genomic Research Initiative (Wilson et al. 2011). The results of this study will also be applied to peanut RNA-seq, exome identification, and genomic markers development.

Acknowledgments Authors are grateful to Ms. Kayla Love for her laboratory assistance during the conduct of this experiment. The manuscript has been reviewed by all authors and all listed authors have agreed to this submission without conflict of interest. This work was supported by USDA-NIFA Evans-Allen Formula Grant (Accession No: 0209894).

References

Adams MD, Kelley JM, Gocayne JD et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Altschul SF, Gish W (1996) Local alignment statistics. Methods Enzymol 266:460–480

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bell CJ, Dixon RA, Farmer AD, Flores R, Inman J, Gonzales RA, Harrison MJ, Paiva NL, Scott AD, Weller JW, May GD (2001) The Medicago genome initiative: a model legume database. Nucleic Acids Res 29:114–117

Bi YP, Liu W, Xia H, Su L, Zhao CZ, Wan SB, Wang XJ (2010) EST sequencing and gene expression profiling of cultivated peanut (*Arachis hypogaea* L.). Genome 53:832–839

Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6:791–806

Boote KJ (1982) Growth stages of peanut (*Arachis hypogaea* L.). Peanut Sci 9:35–40

Burow MD, Simpson CE, Starr JL, Paterson AH (2001) Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.): broadening the gene pool of a monophyletic polyploidy species. Genetics 159:823–837

Cannon SB, Crow JA, Heuer ML, Wang X, Cannon EKS, Dwan C et al (2005) Databases and information integration for the *Medicago truncatula* genome and transcriptome. Plant Physiol 138:38–46

Cianferoni A, Muraro A (2012) Food-induced anaphylaxis. Immunol Allergy Clin North Am 32:165–195

Feng S, Wang X, Zhang X, Dang PM, Holbrook CC, Culbreath AK, Wu Y, Guo B (2012) Peanut (*Arachis hypogaea*) expressed sequence tag project: progress and application. Comp Funct Genomics 2012:373768. doi:10.1155/2012/373768

He G, Prakash C (2001) Evaluation of genetic relationships among botanical varieties of cultivated peanut (*Arachis hypogaea* L.) using AFLP markers. Genet Resour Crop Evol 48:347–352

Higgs J (2002) The beneficial role of peanuts in the diet—an update and rethink! Peanuts and their role in CHD. Nutr Food Sci 32:214–218

Huang J, Yan L, Lei Y, Jiang H, Ren X, Liao B (2012) Expressed sequence tags in cultivated peanut (*Arachis hypogaea*): discovery of genes in seed development and response to *Ralstonia solanacearum* challenge. J Plant Res 25:755–769. doi:10.1007/s10265-012-0491-9

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, Suzuki S, Hasegawa M, Kiyoshima H, Isobe S (2012) Large-scale development of expressed sequence tag-derived simple sequence repeat markers and diversity analysis in Arachis spp. Mol Breed 30:125–138

Krapovickas A, Gregory WC (1994) Taxonomı′a del ge′nero Arachis (Leguminosae). Bonplandia 8:1–186

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967

Li X, Hou S, Su M, Yang M, Shen S, Jiang G, Qi D, Chen S, Liu G (2010) Major energy plants and their potential for bioenergy development in China. Environ Manage 46:579–589

Li X, Rezaei R, Li P, Wu G (2011) Composition of amino acids in feed ingredients for animal diets. Amino Acids 40:1159–1168

Moretzsohn MC, Barbosa AVG, Alves-Freitas DMT, Teixeira C, Leal-Bertioli SCM, Guimarães PM, Pereira RW, Lopes CR, Cavallari MM, Valls JFM, Bertioli DJ, Gimenes MA (2009) A linkage map for the B-genome of *Arachis* (Fabaceae) and its synteny to the A-genome. BMC Plant Biol 9:40. doi:10.1186/1471-2229-9-40

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE et al (2010) RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. BMC Plant Biol 10:160

Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A (1994) Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci USA 91:9228–9232

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Wilson RF et al (2011) International peanut genomic research initiative strategic plan for 2012–2016 characterization of the peanut genome. http://www.peanutbioscience.com/images/IPGRI_StratPlan_DRAFT_v4_1_Aug11a.pdf

Wilson RF, Grant D (2010) Soybean genomics research program accomplishments report. http://soybase.org/SoyGenStrat2007/SoyGenStratPlan2008-2012-Accomplishments%20v1.6.pdf

Woody JL, Severin AJ, Bolon YT, Joseph B, Diers BW, Farmer AD, Weeks N, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC (2011) Gene expression patterns are correlated with genomic and genic structure in soybean. Genome 54:10–18

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Brujin graphs. Genome Res 18:821–829

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 Genomics 13:90. doi:10.1186/1471-2164-13-90