Indian Journal of Agricultural Sciences **88** (3): 442–6, March 2018/Article

Transcriptome sequencing of sesame (*Sesamum indicum*) using Illumina Platform

P SUPRIYA[1], A R RAO[2] and K V BHAT[3]

ICAR-Indian Agricultural Research Institute, New Delhi 110 012

Received: 01 November 2017 ; Accepted: 16 November 2017

ABSTRACT

Sesame is an important oil seed crop worldwide and has essential health and medicinal values. In the present study, a high-throughput transcriptome sequencing of sesame was performed using Illumina paired-end sequencing technology for gene and marker discovery. Approximately 6 Gb data was generated and assembled into 16548 unigenes with an N50 of 905 bp. In addition, a total of 1716 unigeneswere assigned to 22 KEGG pathways. The unigenes resulted from this study are involved in lipid metabolism and Glycan biosynthesis pathways etc. Furthermore, 1443 Simple Sequence Repeats (SSR) were detected and consequently primers were designed.Our study enhances the genomic resources of sesame and provides ample amount of information about the transcriptome and SSRs which could serve as a valuable basis for future studies.

Key words: Assembly, Simple sequence repeat, Transcriptome, Unigene

Sesame (*Sesamum indicum* L.) is an ancient oil yielding crop cultivated for its edible seed, high quality oil (Bhat *et al*. 1999, Suh *et al*. 2003) and it belongs to the family Pedaliaceae. Sesame seed contains high oil content (46% - 50%) with 20% proteins, 83% - 90% unsaturated fatty acids and minor nutrients such as vitamins and minerals. It also contains large amount of characteristic lignans such as sesamin, sesamol, sesamolin and tocopherols (Fukuda *et al*. 1985). The functional components give nutraceutical value to the crop by imparting resistance against oxidative deterioration. Hence, sesame seeds are consumed as a traditional health food with great amounts of nutritional components and for their anti-carcinogenic, anti-oxidative activity; anti-inflammatory and specific anti-hypertensive effect (Yokota *et al*. 2007, Hsu *et al*. 2005, Lee *et al*. 2004, Nakano 2008). Although the fact that majority of the wild species of the genus *Sesamum* are intuitive to sub-Saharan Africa, studies demonstrated that sesame was first domesticated in India (Bedigian 2004). Earlier studies on sesame have primarily focused on traditional genetic breeding (Were 2006), quantitative genetics (Wei *et al*. 2009), genetic relationships and diversity among sesame germplasm collections (Laurentin and Karlovsky 2006, Ercan *et al*. 2004). However, the advancement of molecular marker aided breeding and next generation sequencing techniques resulted in a quantum jump in its improvement. Transcriptome is the whole collection of transcripts at a specific developmental stage in a cell, which provides comprehensive and valuable information on gene regulation, gene expression, and amino acid content of proteins. Next-generation sequencing technologies such as Illumina paired-end sequencing technology have provided a novel method both for transcriptome analysis and gene mapping (Jiang *et al*. 2013).

In the present study, we performed the whole transcriptome sequencing of sesame using Illumina paired-end sequencing technology to obtain the comprehensive transcriptome profile of sesame and also detection of SSR markers for subsequent studies of sesame at levels of, molecular biology, genomics and genetics.

MATERIALS AND METHODS

Total RNA was isolated using XcelGen Plant RNA miniprep kit from leaf tissues. The quality of RNA was checked in 1% denatured agarose gel (loaded 3 µl) for the presence of 28S and 18S bands. The gel was run at 90 V for 40 min. The total RNA was quantified using Nanodrop-8000. The 1 µl of each sample was loaded in NanoDrop 8000 and Qubit fluorometer for determining concentration and A260/280 ratio.

Illumina *TruSeq RNA* sample preparation v2 kit(Illumina) was used for the preparation of paired end cDNA sequencing library. The mRNA fragmentation was followed by several steps, viz. reverse transcription, second-strand synthesis, end-repairing, adenylation of 3' ends, pair-end adapter ligation and finally ended with index PCR amplification

[1]Ph D Scholar (e mail: puramsupriya@gmail.com), [2]Professor and Principal scientist (e mail: ar.rao@icar.gov.in), Division of Bioinformatics, ICAR-IASRI. [3]Principal Scientist (e mail: kvbhat2001@yahoo.com), Division of Genomic Resources, ICAR-NBPGR, New Delhi.

of adaptor-ligated library. The libraries were in the size range of 400 bp to 2000bp. High Sensitivity (HS) DNA chip was used for analyzing the quality of amplified library on Bioanalyzer 2100.

The next generation sequencing run for whole transcriptome assembly was achieved through Illumina MiSeq. The raw data was filtered using Trimmomatic v0.30 (Bolger *et al*. 2014). The reads with adaptor contamination and of low quality (quality score <20) were filtered. The high quality reads obtained after adapter trimming and quality control were assembled using CLC genomics workbench on default parameters.

Unigenes were predicted from transcripts using ESTScan (Iseli *et al*. 1999) followed by BLASTx against the NCBI non-redundant (Nr) protein database, the Swiss-Prot protein database and the results were parsed by an in-house perl script. The longest ORF out of six frames was selected. Gene Ontology (GO) terms were assigned to unigenes using Blast2GO (Conesa *et al*. 2005) and functional classification was achieved through WEGO software (Ye *et al*. 2006).

GO assignments were used to classify the functions of the predicted unigenes. The GO mapping provides ontology of defined terms signifying gene product properties which are categorized into three main domains: Biological Process, Molecular Function and Cellular Component. GO mapping was carried out to retrieve GO terms for all the functionally annotated unigenes through BLASTX. For the retrieval of gene names or symbols BLASTX result accession IDs were used. The identified gene names or symbols searched against the species specific entries of the gene product tables of GO database. BLASTX result accession IDs are used in order to retrieve UniProt IDs with the use of PIR which includes PSD, UniProt, TrEMBL, SwissProt, GenPept, RefSeq and PDB databases. Further, accession IDs are searched in the dbxref table of GO database.

MIcroSAtellite (MISA, http://pgrc.ipk-gatersleben.de/misa/) tool was used for microsatellite mining. In the present study, the SSRs that contain motifs with two to six nucleotides were taken into consideration in size and a minimum of 6,4,3,3,3 contiguous repeat units for di, tri, tetra, penta, hexa repeats respectively. Compound SSRs were located by taking a condition of occurrence of two SSRs with a minimum of 75 nucleotide intervening sequences. In-house *perl* scripts were written to extract the flanking sequences of SSRs for designing primers. Batchprimer3 (http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cg) was used for designing primers with the criteria of GC content between 20-80%, primer length of 18-30 bp and annealing temperature between 57-63ºC.

RESULTS AND DISCUSSION

Read statistics

The total number of reads obtained by Illumina Miseq platform were 16452323 and 16452323 for paired end sequencing, while the total number of bases sequenced were 29,025,50,023(2.9 GB) and 28,144,82,671 (2.8 GB) respectively. The read length ranges from 40 to 250 with an average GC content of 54%.

Data analysis and assembly statistics

A total of 16612 transcripts were resulted from *de novo* assembly with N50 value of 905. The whole transcriptome length was 15,248,753. The average length of assembled transcript was 918. The minimum and maximum size of assembled transcript was 17830 and 408 respectively (Table 1).

Table 1 Assembly statistics

Description	Count
Number of transcripts	16612
Transcriptome length	15248753
Maximum transcript size	17830
Minimum transcript size	408
Mean transcript size	918
N50 value	905

GO mapping and unigene distribution

A total of 16548 unigenes were predicted and unigene length ranges from 100-1000. As many as 2739 unigenes were in the range of 500-600 length followed by 400-500 length. The predicted unigenes were annotated using BLASTx which resulted in the annotation of 15438 unigenes with an e-value less than 1e-5. These annotated unigenes were mapped on to GO database. Maximum percentage of unigenes significant similarity with *Ricinus communis* followed by *Cucumis sativus, Oryza sativa* and so forth. The GO mapping resulted in the retrieval of GO terms for unigenes using diverse databases. The GO terms related to BLASTX hits were found in UniProtKB in majority followed by TAIR database. Based on Nr annotation, 15438 unigenes were assigned a total of 58 gene ontology (GO) terms in three ontologies namely, biological process (23 GO terms), molecular function (16 GO terms) and cellular component (19 GO terms) (Fig 1).

The unigenes for biological process made up the majority (44.3%), followed by cellular components (34.5%), and molecular function (21.2%). The unigenes which were assigned functionally covered a broad range of GO categories. Under the biological process category, cellular process (32.8%) and metabolic process (32%) were represented in majority. Furthermore, 24.9% unigenes were involved in response to stimulus. Under the cellular component category, cell component (35.6%), cell part (35.6%) component and organelle extra cellular region represented the majorities. For the molecular function category, binding (30.5%) and catalytic activity (30.4%) represented the majorities.

Pathway mapping of transcripts by KEGG

Ortholog assignment and mapping of the transcripts to the biological pathways was performed using KEGG automatic annotation server (KAAS) (Moriya *et al*. 2007).

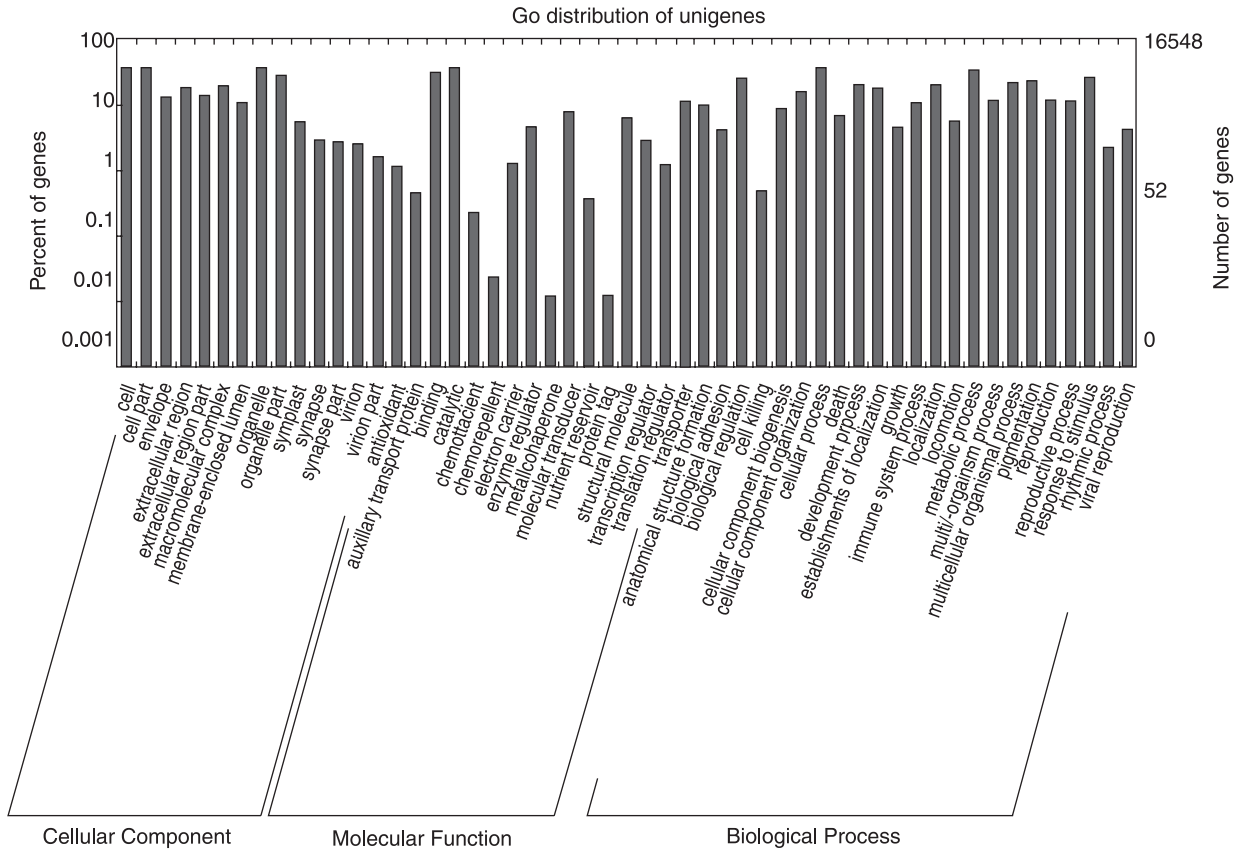

Fig 1 Distribution of GO terms of unigenes

The unigenes represented metabolic pathways of major biomolecules such as carbohydrates, lipids, nucleotides, amino acids, glycans, cofactors, vitamins, terpenoids, polyketides, etc. A total of 1716 unigenes had significant matches in the KEGG database and were assigned to 22 KEGG pathways. Among them, the translation pathway containing 233 unigenes is the largest one, followed by energy metabolism (195), carbohydrate metabolism (176) and amino acid metabolism. Furthermore, the unigenes also represented the genes involved in genetic information processing, environmental information processing and cellular processes, metabolism of cofactors and vitamins, signal transduction and lipid metabolism (Table 2).

Characterization of SSR markers

SSRs are a group of repetitive DNA sequences that denote a substantial portion of higher eukaryote genomes. SSRs are highly polymorphic, co-dominant and are a major source of marker systems for molecular breeding, genetic mapping, comparative genomics, gene mapping and population genetic analyses in a wide variety of species (Smith *et al*. 1997, Senior *et al*. 1998, Wang *et al*. 2002, Neeraja *et al*. 2007, Chapman *et al*. 2009, Bushman *et al*. 2011). Traditional methods for SSR development are laborious, expensive and time consuming. Transcriptome sequencing has become a powerful and cost-efficient tool with the advancement of high throughput sequencing technology (Verma *et al*. 2007). The transcriptome data was a tremendous source for SSR mining and had been

Table 2 List of KEGG Pathway Categories of unigenes

Pathway list	Count
Metabolism	
Carbohydrate metabolism	176
Energy metabolism	195
Lipid metabolism	85
Nucleotide metabolism	75
Amino acid metabolism	176
Metabolism of other amino acids	59
Glycan biosynthesis and metabolism	22
Metabolism of cofactors and vitamins	113
Metabolism of terpenoids and polyketides	51
Biosynthesis of other secondary metabolites	40
Xenobiotics biodegradation and metabolism	29
Genetic information processing	
Transcription	57
Translation	233
Folding, sorting and degradation	136
Replication and repair	27
Environmental information processing	
Membrane transport	13
Signal transduction	78
Cellular processes	
Transport and catabolism	57
Cell motility	6
Cell growth and death	35
Cell communication	9
Organismal systems	
Environmental adaptation	44

utilized in many species (Jiang *et al*. 2013, Guo *et al*. 2011, Kaur *et al*. 2011).

In the present study, a total of 16612 transcripts were examined from which we identified 1443 microsatellites. Tri-nucleotide repeats (610) were the most abundant repeats followed by tetra nucleotide repeats (410). Hexa nucleotide repeats were least among all (89). Among di nucleotide repeats, AG/CT was the most abundant repeat followed by AT/AT repeat where as AAG/CTT repeat was most abundant among tri nucleotide repeats (Table 3).

Table 3 Summary of SSR mining results

Search item	Number
Total number of sequences examined	166121
Total size of examined sequences (bp)	15248753
Total number of identified SSRs	1443
Number of SSR containing sequences	1112
Number of sequences containing more than 1 SSR	151
Number of SSRs present in compound formation	186

Conclusion

In the present study, the transcriptome of leaf tissue of sesame was sequenced by Illumina paired end sequencing technology. Approximately 30 million clean pair-end reads were obtained and further used for de novo assembly through CLC genomics workbench which resulted in 16612 transcripts with an average length of 918bp. A total of 16,548 unigenes were predicted and assigned 58 GO terms. Subsequently, 1443 SSRs were identified from our transcriptome data, more SSR primers can be designed for future research, involving genetic mapping, genetic diversity assessment, and marker-assisted breeding in sesame.

ACKNOWLEDGEMENTS

This study was funded by ICAR-National Agricultural Innovation Project. The first author acknowledges the fellowship from ICAR- IASRI for the Ph D programme. The facilities provided by IARI, ICAR-NBPGR are acknowledged.

REFERENCES

Bedigian D. 2004. History and lore of sesame in southwest Asia. *Economic Botany* **58**: 330–53.

Bhat K V, Babrekar P P and Lakhanpaul S. 1999. Study of genetic diversity in Indian and exotic sesame (*Sesamum indicum* L.) germplasm using random amplified polymorphic DNA (RAPD) markers. *Euphytica* **110**(1): 21-34.

Bolger A M, Lohse M and Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30(**15): 2114–20.

Bushman B S, Larson SR, Tuna M, West M S, Hernandez A G, Vullaganti D, Gong G, Robins J G, Jensen K B and Thimmapuram J. 2011. Orchardgrass (*Dactylis glomerata* L.) EST and SSR marker development, annotation, and transferability. *Theoretical and Applied Genetics* **123**(1): 119–29.

Chapman M A, Hvala J, Strever J, Matvienko M, Kozi, A, Michelmore R W, Tang S, Knapp S J and Burke J M. 2009. Development, polymorphism, and cross-taxon utility of EST–SSR markers from safflower (*Carthamus tinctorius* L.). *Theoretical and Applied Genetics* **120**(1): 85–91.

Conesa A, Gotz S, Garcia-Gomez J M, Terol J, Talon M and Robles M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674–6.

Ercan A G, Taskin M and Turgut K. 2004. Analysis of genetic diversity in Turkish sesame (*Sesamum indicum* L.) populations using RAPD markers. *Genetic Resources and Crop Evolution* **51**(6): 599–607.

Fukuda Y, Osawa T, Namiki M and Ozaki T. 1985. Studies on antioxidative substances in sesame seed. *Agricultural and Biological Chemistry* **49**(2): 301–6.

Guo S, Liu J, Zheng Y, Huang M, Zhang H, Gong G, He H, Ren Y, Zhong S, Fei Z and Xu Y. 2011. Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. *BMC genomics* **12**(1): 454.

Hsu D Z, Su S B, Chien S P, Chiang P J, Li Y H, Lo Y J and Liu M Y. 2005. Effect of sesame oil on oxidative-stress-associated renal injury in endotoxemic rats: involvement of nitric oxide and proinflammatory cytokines. *Shock* **24**(3): 276–80.

Iseli C, Jongeneel C V and Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, pp 138–48.

Jiang B, Xie D, Liu W, Peng Q and He X. 2013. *De novo* assembly and characterization of the transcriptome, and development of SSR markers in wax gourd (*Benicasa hispida*). *PloS one* **8 (**8): e71054.

Kaur S, Cogan N O, Pembleton L W, Shinozuka M, Savin K W, Materne M and Forster J W. 2011. Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* **12(**1): 265.

Laurenti H E and Karlovsky P. 2006. Genetic relationship and diversity in a sesame (*Sesamum indicum* L.) germplasm collection using amplified fragment length polymorphism (AFLP). *BMC Genetics* **7(**1): 10.

Lee C C, Chen P R, Lin S, Tsai S C, Wang B W, Chen W W, Tsai C E and Shyu K G. 2004. Sesamin induces nitric oxide and decreases endothelin-1 production in HUVECs: possible implications for its antihypertensive effect. *Journal of Hypertension* **22**(12): 2329–38.

Moriya Y, Itoh M, Okuda S, Yoshizawa A C and Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**(2): 182–5.

Nakano D, Kurumazuka D, Nagai Y, Nishiyama A, Kiso Y and Matsumura Y. 2008. Dietary sesamin suppresses aortic NADPH oxidase in DOCA salt hypertensive rats. *Clinical and Experimental Pharmacology* **35**(3): 324–6.

Neeraja C N, Maghirang-Rodriguez R, PamplonA, Heuer S, Collard B C Y, Septiningsih E M, Vergara G, Sanche D, Xu K, Ismail A M and Mackill D J. 2007. A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. *Theoretical and Applied Genetics* **115**(6): 767–76.

Senior M L, Murphy J P, Goodman M M and Stuber C W 1998. Utility of SSRs for determining genetic similarities and

relationships in maize using an agarose gel system. *Crop Science* **38**(4): 1088–98.

Smith J S C, Chin E C L, Sh H, Smit O S, Wall S J, Senior M L, Mitchell S E, Kresovich S and Ziegle 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. *Theoretical and Applied Genetics* **95**(1): 163–73.

Suh M C, Kim M J, Hur C G, Bae J M, Park Y I, Chung C H, Kang C W and Ohlrogge J B. 2003. Comparative analysis of expressed sequence tags from *Sesamum indicum* and *Arabidopsis thaliana* developing seeds. *Plant Molecular Biology* **52**(6): 1107.

Verma V K, Behera T K, Munshi A D, Parida S K and Mohapatra T. 2007. Genetic diversity of ash gourd [*Benincasa hispida* (Thunb.) Cogn.] inbred lines based on RAPD and ISSR markers and their hybrid performance. *Scientia Horticulturae* **113**(3): 231–7.

Wang Y, Georgi L L, Zhebentyayeva T N, Reighard G L, Scorza R and Abbott A G. 2002. High-throughput targeted SSR marker development in peach (*Prunus persica*). *Genome* **45**(2): 319–28.

Wei L B, Zhang H Y, Zheng Y Z, Miao H M, Zhang T Z and Guo W Z. 2009. A genetic linkage map construction for sesame (*Sesamum indicum* L.). *Genes and Genomics* **31**(2): 199–208.

Were B A I. 2006. Genetic improvement of oil quality in sesame *(Sesamum indicum* L.*)*. 12.

Ye J, Fang L, Zheng H K, Zhang Y and Chen J. 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research* **34**: W293–W297.

Yokota T, Matsuzaki Y, Koyama M, Hitomi T, Kawanaka M, Enoki-Konishi M, Okuyama Y, Takayasu J, Nishino H, Nishikawa A and Osawa T. 2007. Sesamin, a lignan of sesame, down-regulates cyclin D1 protein expression in human tumor cells. *Cancer Science* **98**(9): 1447–53.