Physiol Mol Biol Plants (January–March 2014) 20(1):139–142
DOI 10.1007/s12298-013-0204-4

SHORT COMMUNICATION

Analysis of *Jatropha curcas* transcriptome for oil enhancement and genic markers

Atul Grover · Maya Kumari · Sadhana Singh · Shivender Singh Rathode · Sanjay Mohan Gupta · Pankaj Pandey · Sween Gilotra · Devender Kumar · Mohommad Arif · Zakwan Ahmed

Received: 15 May 2013 / Revised: 29 August 2013 / Accepted: 2 September 2013 / Published online: 26 September 2013

Abstract Oil-rich seeds of *Jatropha curcas* are being focussed as a source of bio-diesel. However, prior to its industrial use, a lot of crop improvement efforts are required in Jatropha. Availability of a large number of EST sequences of Jatropha in public domain allow identification of candidate genes for several agronomic characters including oil content in seeds. Here, we have analysed 42,477 ESTs of Jatropha spanning 22.9 Mbp for microsatellites and fatty acid metabolism related sequences. Unigene sequences were built using CAP 3 programme resulted in 12,358 contigs and 5,730 singlets. Nearly, 8 % unigenes showed presence of microsatellites, slightly over-represented compared to their occurrence in ESTs. Most of the microsatellites were either di- or tri-nucleotide repeats, while other categories of tetra-, penta- and hexa-nucleotide repeats together constituted ~4 % of total microsatellites. Assessment of functional relevance of unigenes was carried out using Blast2GO using its default settings. The overall sequence similarity level against sequences in 'nr' database was >80 %. A total of 931 sequences that participated in any of the pathways related to fatty acid or lipid metabolism were found at GO level 6. Among these, GO terms "Fatty acid metabolic process" and "Fatty acid biosynthetic process" were most over-represented. Overall, our work has due relevance in identifying molecular markers for the candidate genes for oil content in Jatropha seeds, and will prove to be an important reference for further studies for identification of trait specific markers in Jatropha.

Keywords *Jatropha curcas* · ESTs · Unigenes · Microsatellites · Gene ontology · Fatty acids

Electronic supplementary material The online version of this article (doi:10.1007/s12298-013-0204-4) contains supplementary material, which is available to authorized users.

A. Grover (✉) · M. Kumari · S. Singh · S. S. Rathode ·
S. M. Gupta · P. Pandey · D. Kumar · M. Arif · Z. Ahmed
Biotechnology Division, Defence Institute of Bio-Energy Research,
Goraparao, P.O. Arjunpur, Haldwani 263 139, India
e-mail: iatulgrover@gmail.com

S. Gilotra
Department of Biotechnology, Indian Institute of Technology,
Guwahati, India

Z. Ahmed
Directorate of Management Services, Defence Research
and Development Organization, DRDO Bhawan,
New Delhi 110011, India

Present Address:
P. Pandey
National Agri-Food Biotechnology Institute, C-127 Industrial Area,
SAS Nagar Phase 8, Mohali 160071, Punjab, India

Present Address:
D. Kumar
RVC Center and College, c/o 56 APO, Meerut Cantt 900468, India

Introduction

Jatropha curcas L. is one of the major plants with non edible oil rich (27–40 %) seeds (Acthen et al. 2007), currently being focussed and developed for biodiesel production. Its oil contain high amount of unsaturated fatty acids, and can thus serve as a biodiesel feedstock meeting international standards (Tiwari et al. 2007). Unfortunately, potential seed production of Jatropha is not agronomically as standardized, as it is in food crops. Crop improvement in Jatropha through conventional breeding is difficult owing to its perennial, heterozygous and outbreeding nature. Thus, further genetic improvement of the plant by biotechnological intervention is currently required (Natarajan et al. 2010). For better knowledge of ecological and agronomic properties like growth conditions, biomass production, seed yield and genetics, a large scale analysis of its genome and transcriptome is required.

In recent years, there have been increasing number of molecular level studies towards understanding of biosynthetic pathways, metabolic flux control, genes encoding economically important proteins and enzymes, and also towards development of molecular markers and genetic maps (Natarajan et al. 2010; Sato et al. 2011; Wang et al. 2011). Thus, in a short span, the numbers of Jatropha ESTs in dbEST of NCBI has grown to more than 42,000. Gene discovery using high throughput transcriptome sequencing using classical or Next Generation Sequencing and subsequently through sequence similarity and comparative genomics has become an attractive tool in recent years, especially for non model species (Annadurai et al. 2012; Roorkiwal and Sharma 2012; Gupta et al. 2013; Pathak et al. 2013; Xiao et al. 2013). Coding regions, in particular can be exploited for developing transferable DNA markers like microsatellites for comparative studies. Microsatellites occur preferentially in and around genic regions in plants (Grover and Sharma 2007). They are important tools for comparative mapping because of their high polymorphism and transferability (Varshney et al. 2005). Here, we describe comparative genomics based analysis of Jatropha ESTs for occurrence of microsatellites as well as for limited functional prediction.

Methods

ESTs of *Jatropha curcas* were downloaded from dbEST of NCBI. To nullify redundancy prevalent in ESTs, we used the sequence assembly program CAP3 (Huang and Madan 1999) to cluster ESTs into contigs/singletons and generated non-redundant unigenes. A perl script, MISA (MIcroSAtelitte; http://pgrc.ipk-gatersleben.de/misa/) (Thiel et al. 2003) was used to identify microsatellites in all the clustered as well as non clustered set of sequences. A simple sequence repeat with motif length varying from 2–6 bp was identified as a microsatellite. Mononucleotide repeats were excluded from the analysis because of the abundance of poly A/T repeats mostly resulting from sequencing artefacts and poly A tails. Repeat-motifs like AG, GA, TC and CT were considered in the same class considering complementary sequences and/or different reading frames. The analysis of mined microsatellites was done on the basis of their motif length (di- to hexa-nucleotide), number and type of repeats, relative frequency of occurrence and length as class I (≥20 nucleotides) and class II (12–20 nucleotides) types (Temnykh et al. 2001). Assessment of functional relevance of unigenes of Jatropha was carried out using Blast2GO (Conesa et al. 2005). Unigene sequences not showing any match were considered unique to Jatropha.

Results and discussion

Availability of large collections of ESTs of Jatropha has allowed us to explore these resources for the presence of different microsatellite repeats, as well as different genes of agronomic interest. However, EST collections are often redundant and full of relatively shorter sequences. Unigenes, on the other hand are longer and non redundant. EST-SSRs representing the coding regions of the genome, are expected to be conserved with a high rate of cross species transferability in comparison to genome derived SSRs in addition to being polymorphic enough to be exploited for population studies (Grover et al. 2009; Jain et al. 2010).

A total of 42,477 ESTs (~22.9 Mbp) were downloaded from dbEST of NCBI, which were annotated into 12,358 contigs (4.76 Mbp) and 5,730 singlets (6.50 Mbp). These 18,088 transcriptomic sequences were searched for the presence of microsatellites, which were found in 7.91 % of sequences yielding 3,557 microsatellite motifs. Thus, the number of microsatellite containing transcriptomic sequences in Jatropha was found comparable to previously reported data from other plants (Cavagnaro et al. 2010; Grover and Sharma 2012). Significantly, while the number of unigenes was only 42.6 % of sequences in dbEST, non redundant set of microsatellites sequences are 55.4 % of the originally searched in Jatropha sequences in dbEST, thus over-represented in unigene sequences. It may be noted that the variable abundance of microsatellites is known to be dependent on the search criteria, the size of the dataset, the database-mining tools and the species concerned (Grover et al. 2012). Thus, 18.6 % of oak ESTs have been reported to contain microsatellites, with shorter threshold length using the tool SSRIT (Durand et al. 2010). The criteria and the mining tool used in this study (MISA) is a popular one, and a number of microsatellite abundance studies (Thiel et al. 2003; Grover and Sharma 2007; Grover et al. 2007; Cavagnaro et al. 2010) have been based on its default settings.

In earlier reports, trinucleotide repeats generally formed the most common motif in various plant species, regardless of the EST-SSR search criteria. However, abundance of dinucleotide repeats has also been reported in many of the dicot species (Grover and Sharma 2012). Interestingly in Jatropha, their numbers were comparable, together constituting little more than 95 % of all the microsatellites, while tetranucleotide (~2.7 %), pentanucleotide (<0.8 %) and hexanucleotide (~0.7 %) repeats were rare.

In terms of single SSR motif, the dinucleotide motif AG/CT was most frequent, as reported earlier as well in other plants (Kantety et al. 2002; Kumpatla and Mukhopadhyay 2005; Grover and Sharma 2012). The two most dominant motif types recorded in our search were AG and AAG (Table 1), which are known to be abundant in genic sequences (Grover and Sharma 2007). Low abundance of "CG" repeats may be attributed to their tendency of forming secondary structures (hairpins), leading to a selective pressure against 'CG' accumulation in genomes. Microsatellites were also classified into two classes on length basis. Firstly, Class I microsatellites, which included microsatellites more than or equal to 20 nucleotides in length, and secondly, Class II microsatellites of less than 20 nucleotides. Class II microsatellites were more abundant (73 %) both in unclusterred as well as clustered set of sequences. In silico identification of SSRs from various sequence resources like genomic sequences, ESTs or unigenes is a low cost and easy method for development of microsatellite markers. The present work itself has been a source of development of twenty four EST SSR markers for Jatropha (Kumari et al. 2013), with PIC values 0.024–0.495, demonstrated on a panel of 41 accessions, collected from all over India. Importantly, the designated primers targeting EST SSRs amplified at more than one locus, leading to multiple bands in the 4 % agarose gel analysis. The polymorphism could thus not be strictly attributed to EST SSR alone that had been targeted. Various other factors like duplications, or occurrences of pseudogenes could have been the reasons for such amplifications. Nevertheless, these markers can still be used for understanding the nature and possible biological functions alongwith genomic evolutionary events, and on-going evolutionary activities of the Jatropha genome.

We also performed sequence similarity search against non redundant 'nr' database of NCBI. More than 70 % of unigenes showed homology to genes having known function, and nearly 4,000 unigenes showed significant similarity with castorbean sequences (Alignment score > 200). As the sequence similarity level was high (ranging 80–90 % for most sequences), default parameters of Blast2GO were used for annotation of these sequences (Gotz et al. 2008). Most of the unigene sequences represented enzymes of general metabolism as reported earlier (Costa et al. 2010). On the basis of GO annotation, unigenes were assigned GO terms associated with biological process, cellular component and molecular function. Unigenes related to biological process such as in various metabolic pathways were further studied for their possible roles in fatty acid biosynthesis pathways. Maximum sequences (931) with their roles in fatty acid metabolism annotating to 60 different GO terms were identified at GO level 6. GO terms "Fatty acid metabolic process" and "Fatty acid biosynthetic process" were most over-represented at GO level 6, with 143 and 95 sequences annotated to this category respectively (Supplementary Table 1). A GO term obtained at higher GO level is more precise, and accordingly the number of genes assigned per GO term decreases as the GO level is raised (Al-Shahrour et al. 2004). KEGG maps have been generated wherever possible for identification of metabolic reactions being catalyzed by the enzymes identified using Blast2GO annotation. This has also helped to obtain the enzyme codes (EC) for the sequences of interest. These gene sequences can further be targeted for development of trait-specific markers and utilize them in assigning marker trait associations and allele variance across diverse collections of Jatropha.

Table 1 Distribution of the selected microsatellite motifs in Jatropha transcriptomic sequences. Unclustered sequences refer to the raw EST files, while upon clustering, sequences were distributed in 'Singlets' and 'Contigs', as refered in the text. Microsatellite motifs shown in the table below were selected on the basis of the published reports on their preferential occurrences. Motifs AG/CT are known to be abundant in UTRs, AT/TA, AAT/TTA and AAAT/ATTT are generally most abundant motif types of their categories (di-, tri- and tetra-nucleotide) in the genomes, in non-coding DNA, AAG/CTT are generally most abundant repeats in exons, CCG/CGG are most common repeats in some species (mostly monocots) and motifs GC/CG are extremely rare

Motif	Unclustered sequences	Singlets	Contigs
AG/CT	1118	321	265
AT/TA	302	143	64
GC/CG	1	1	0
AAG/CTT	414	156	110
AAT/TTA	274	80	35
CCG/CGG	62	10	11
AAAT/ATTT	19	11	6

Future lines

Mining microsatellites from coding sequences have diverse applications, and foremost among them is to answer basic questions on the functional and structural evolution of genes and genomes. The scientific community is still in a process to understand fundamental principles which allow accommodation of these repetitive and so-called hypervariable sequences in coding regions. Though the molecular mechanisms that lead to persistence of these repeats in transcriptome are not yet known, a number of biological roles as in adaptive advantages have been suggested in the past both by in silico and wet experiments (reviewed by Grover and Sharma 2011). As more and more positional information on microsatellite occurrence gathers, comparative analysis becomes easier, and more concrete evidences can be collected to comment upon their significance in a genome or a transcriptome.

Another equally important area is to use them as functional molecular markers. A representative set of microsatellite sequences, preferably with length > 30 bp have already been developed into molecular markers for their abilities to discriminate different accessions of *Jatropha curcas* collected from all over India (Kumari et al. 2013). We further aim to use this information alongwith the gene ontology information towards development of trait-specific markers for rapid identification of high yielding genotypes in near future.

References

Acthen WMJ, Mathijs E, Verchot L, Singh VP, Aerts R, Muys B (2007) *Jatropha* biodiesel fueling sustainability? Biofuels Bioproduct Bioresour 1:283–291

Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with group of genes. Bioinformatics 20:578–580

Annadurai RS, Jayakumar V, Mugasimangalam RC, Katta MAVSK, Anand S, Gopinathan S, Sarma SP, Fernandes SJ, Mullapudi N, Murugesan S, Rao SN (2012) Next generation sequencing and de novo transcriptome analysis of *Costus pictus* D. Don, a non-model plant with potent anti-diabetic properties. BMC Genomics 13:363

Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, Huang S, Weng Y (2010) Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genomics 11:569

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Costa GGL, Cardoso KC, Del Bem LEV, Lima AC, Cunha MAS, de Campos-Leite L, Vicentini R, Papes F, Moreira RC, Yunes JA, Campos FAP, Da Silva MF (2010) Transcriptome analysis of the oil-rich seed of the bioenergy crop *Jatropha curcas* L. BMC Genomics 11:462

Durand J, Bodenes C, Chancerel E, Frigerio J-M, Vendramin G, Sebastiani F, Buonamici A, Gailing O, Koelewijn H-P, Villani F, Mattioni C, Cherubini M, Goikoetxea P, Herran A, Ikaran Z, Cabane C, Ueno S, Alberto F, Dumoulin P-Y, Guichoux E, Daruvar A, Kremer A, Plomion C (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. BMC Genomics 11:570

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SJ, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435

Grover A, Sharma PC (2007) Microsatellite motifs with moderate GC content are clustered around genes on *Arabidopsis thaliana* chromosome 2. In Silico Biol 7:201–213

Grover A, Sharma PC (2011) Is spatial occurrence of microsatellites in genome a determinant of their function and dynamics contributing to genome evolution? Curr Sci 100:859–869

Grover A, Sharma PC (2012) Tandem repetitions in transcriptomes of some Solanaceae species. Am J Mol Biol 2:140–152

Grover A, Aishwarya V, Sharma PC (2007) Biased distribution of microsatellite motifs in the rice genome. Mol Genet Genomics 277:469–480

Grover A, Ramesh B, Sharma PC (2009) Development of microsatellite markers in potato and their transferability in some members of solanaceae. Physiol Mol Biol Plants 15:343–358

Grover A, Aishwarya V, Sharma PC (2012) Searching microsatellites in DNA sequences: approaches used and tools developed. Physiol Mol Biol Plants 18:11–19

Gupta P, Goel R, Pathak S, Srivastava A, Singh SP, Sangwan RS, Asif MH, Trivedi PK (2013) De novo assembly, functional annotation and comparative analysis of *Withania somnifera* leaf and root transcriptomes to identify putative genes involved in the Withanolides biosynthesis. PLoS One 8:e62714

Huang X, Madan A (1999) A DNA sequence assembly program. Genome Res 9:868–877

Jain A, Ghangal R, Grover A, Raghuvanshi S, Sharma PC (2010) Development of new EST based SSR markers in seabuckthorn. Physiol Mol Biol Plants 16:375–378

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

Kumari M, Grover A, Patade VY, Arif M and Ahmed Z (2013) Development of EST-SSR markers through data mining and their use for genetic diversity study in Indian accessions of *Jatropha curcas* L.: a potential energy crop. Genes Genomics (In press)

Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985–998

Natarajan P, Kanagasabapathy D, Gunadayalan G, Panchalingam J, Shree N, Sugantham PA, Singh KK, Madasamy P (2010) Gene discovery from *Jatropha curcas* by sequencing of ESTs from normalized and full-length enriched cDNA library from developing seeds. BMC Genomics 11:606

Pathak S, Lakhwani D, Gupta P, Mishra BK, Shukla S, Asif MH, Trivedi PK (2013) Comparative transcriptome analysis using high papaverine mutant of *Papaver somniferum* reveals pathway and uncharacterized steps of papverine biosynthesis. PLoS One 8:e65622

Roorkiwal M, Sharma PC (2012) Sequence similarity based identification of abiotic stress responsive genes in chickpea. Bioinformation 8:92–97

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima U, Minami C, Muraki A, Nakazaki N, Takahashi C, Nakayama S, Kishida Y, Kohara M, Yamada M, Tsuruoka H, Sasamoto S, Tabata S, Aizu T, Toyoda A, Shin-i T, Minakuchi Y, Kohara Y, Fujiyama A, Tsuchimoto S, Kajiyama S, Makigano E, Ohmido N, Shibagaki N, Cartagena JA, Wada N, Kohinata T, Atefeh A, Yuasa S, Matsunaga S, FukuI K (2011) Sequence analysis of the genome of an oil bearing tree, *Jatropha curcas* L. DNA Res 18:65–76

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

Thiel T, Michalek W, Varshney RK, Graner A (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

Tiwari AK, Kumar A, Raheman H (2007) Biodiesel production from Jatropha (*Jatropha curcas*) with high free fatty acids: an optimized process. Biomass Bioenerg 31:569–575

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23: 48–55

Wang CM, Liu P, Yi C, Gu K, Sun F, Li L, Lo LC, Liu X, Feng F, Lin G, Cao S, Hong Y, Yin Z, Yue GH (2011) A first generation microsatellite and SNP-based linkage map of *Jatropha*. PLoS One 6:e23632

Xiao M, Zhang Y, Chen X, Lee E-J, Barber CJS, Chakrabarty R, Desgagne-Penix I, Haslam TM, Kim Y-B, Liu E, MacNevin G, Masada-Atsumi S, Reed DW, Stout JM, Zerbe P, Zhang Y, Bohlmann J, Covello PS, De Luca V, Page JE, Ro D-K, Martin VJJ, Facchini PJ, Sensen CW (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J Biotechnol 166:122–134