Plant Biotechnology 21(5), 349–353 (2004)

Original Paper

Expressed sequence tags from callus of *Euphorbia tirucalli*: A resource for genes involved in triterpenoid and sterol biosynthesis

Masataka Kajikawa[1], Katsuyuki T. Yamato[1], Yoshito Kohzu[1], Ryoko Sakata[1], Hideya Fukuzawa[1], Hidenobu Uchida[2], Kanji Ohyama[2]*

[1] Division of Integrated Life Science, Graduate School of Biostudies, Kyoto University, Kyoto 606-8502, Japan
[2] Research Institute of Agricultural Resources, Ishikawa Agricultural College, Ishikawa 921-8836, Japan
* E-mail: kohyama@ishikawa-c.ac.jp Tel: +81-76-248-3137 (ext 5139) Fax: +81-76-248-4191

Received August 31, 2004; accepted September 29, 2004 (Edited by K. Yazaki)

Abstract We report generation of 9,301 expressed sequence tags (ESTs) derived from callus cells of *Euphorbia tirucalli* in search of candidate genes involved in the triterpenoid and sterol biosyntheses. After assembling 4,342 redundant ESTs into 1,252 clusters, a total of 6,211 non-redundant sequences were obtained. Database search revealed that 4,449 out of the 6,211 sequences shared significant similarities to known nucleotide or amino acid sequences, while the remaining 1,762 showed no significant matches and appear to represent novel genes in *E. tirucalli*. The annotations assigned to the hit database entries suggest that 48 of the unique sequences are involved in triterpenoid and sterol biosyntheses. Although functions of genes tagged by the 48 sequences are yet to be determined, the EST resource described here should contribute to identification of genes participating in the triterpenoid and sterol biosyntheses in *E. tirucalli*.

Key words: *Euphorbia tirucalli*, expressed sequence tags (ESTs), triterpenoid, sterol metabolisms.

The genus *Euphorbia* contains a few thousand of species, the most popular of which is probably poinsettia *E. pulcherrima*. However, some other species of *Euphorbia* are considered as a future source of biomass for energy conversion. Nielsen et al. (1977) suggested that several plants including *Euphorbia* species might be cultivated as renewable sources of hydrocarbon-like photosynthetic products. Furthermore, investigations on triterpenoid and sterol contents in cultured cells of *Euphorbia* species also demonstrated that this group of plant can be a rich source of triterpenoids and sterols (Biesboer and Mahlberg 1979; Yamamoto et al. 1981). *E. tirucalli* produces high concentrations of euphol (triterpene) and β-sitosterol (sterol), and thus presumably possesses a set of genes encoding enzymes with high activity in the triterpenoid and sterol biosyntheses (Ohyama et al. 1984b). An array of expressed sequence tags (ESTs) is a powerful tool that has emerged from genomics researches. Collections of EST can provide gene expression patterns, transcriptional regulations and sequence diversity. Availability of cDNA libraries of various sources and efficient high-throughput sequencing has made ESTs an effective means of gene discovery in focused metabolic situations (Sterky et al. 1998; Ohlrogge and Benning 2000). This approach was first applied to the isolation of oleate hydroxylase from castor (van de Loo et al. 1995). In order to better understand and isolate relevant genes in the *Euphorbia* triterpenoid and sterol biosyntheses, here we report generation and analysis of ESTs from an *E. tirucalli* callus cDNA library and search for candidate genes that are involved in the triterpenoid and sterol biosyntheses.

Materials and methods

Callus induction and chemical analysis

Explants of stems of *E. tirucalli* were sterilized by immersion in 70% ethanol for 5 min, rinsed with sterile water, and placed in a 25% sodium hypochlorite solution for 15 min. After extensive rinsing in sterile water, the stems were cut into 0.5 mm long segments and placed horizontally on B5 agar (0.8%) medium (Gamborg 1970; Ohyama et al. 1984a) containing 2,4-D (1 mg l^{-1}), NAA (2 mg l^{-1}) and BA (1 mg l^{-1}). Induced calli were subcultured on a B5 agar medium supplemented with 2,4-D (1 mgl^{-1}), NAA (2 mgl^{-1}), BA (1 mgl^{-1}) and

Abbreviations: BA, benzyladenine; EST, expressed sequence tag; NAA, naphthaleneacetic acid; 2,4-D, 2,4-dichlorophenoxyacetic acid.
The nucleotide sequence data reported in this paper appear in the GenBank/EMBL/DDBJ nucleotide sequence database with the accession numbers, BP953477 to BP962777.

casamino acid ($1\,g\,l^{-1}$). Fast growing calli were selectively subcultured and used in this study.

cDNA library construction

Poly(A)$^+$RNA was isolated from cultured cells using the Concert Plant RNA Reagent (Invitrogen, Carlsbad, CA), and the PolyATract System 1000 (Promega, Madison, WI) according to the manufacturers' instructions. A cDNA library was constructed using the SUPERSCRIPT^{TM} Plasmid System with the GATEWAY^{TM} Technology for cDNA Synthesis and Cloning (Invitrogen) according to the manufacturer's instructions. Insert size of each cDNA clone was estimated by agarose gel electrophoresis after PCR amplification with T7 primer: 5′-TAATACGACTCACTATAGGG-3′ and M13 forward primer: 5′-TGTAAAACGACGGCCAGT-3′ under the following condition: 35 cycles of 94°C for 20 s, 54°C for 20s and 72°C for 3 min.

DNA sequencing

For preparation of template DNAs for sequencing reaction, cDNA clones were first cultured in 96-well deep-well blocks containing 1.0 ml LB supplemented with 50μg/ml ampicillin. Deep-well blocks were incubated at 37°C with agitation (600 rpm) for 12 h and were centrifuged for 1 min at 2,000 rpm. Each 1 μl supernatant of the medium was used as a PCR template. ESTs were obtained by sequencing the 5′ ends of cDNA clones using the T7 promoter primer and the DYEnamic ET Dye Terminator Sequencing Kit (Amersham Biosciences, Piscataway, NJ) with a MegaBACE DNA Analysis System (Amersham Biosciences).

Sequence analysis and annotation

Sequence reads were first processed with a base-caller program, phred (Ewing et al. 1998; Ewing and Green 1998), to obtain nucleotide sequences with quality values assigned to each nucleotide residue. For sequence clean-up, a sequence processing software Paracel Transcript-Assembler^{TM} (PTA; Paracel, Pasadena, CA) was used. Low-quality sequences, vector sequence and polyA tail at the end of each sequence were trimmed. Trimmed sequences of $<$90 bp in length were rejected. Subsequently, sequences originated from *E. coli*, organelles and structural RNAs were removed. ESTs were clustered using PTA, and each cluster was visually inspected and edited for consistency using consed (Gordon et al. 1998). The complete set of ESTs were submitted to DNA Data Bank of Japan (DDBJ), and assigned accession numbers are BP953477 to BP962777. Functional classification of the ESTs was performed as follows. First, each EST was searched against the protein database UniProt (Apweiler et al. 2004) using BLASTX algorithm with E-value threshold of 1×10^{-5} (Altschul et al. 1997). The first UniProt entry in the hit list for each EST was automatically selected, and its Gene Ontology (GO) assignment was retrieved from a GO Annotation (GOA) file for UniProt (Camon et al. 2004), which had been slimmed down using a GO slim for plant provided by the Gene Ontology Consortium. ESTs are then automatically classified into categories according to their assigned GO terms. The ESTs were also searched against the EST database at the DNA Data Bank of Japan using BLASTN algorithm with E-value threshold of 1×10^{-50} (Altschul et al. 1997).

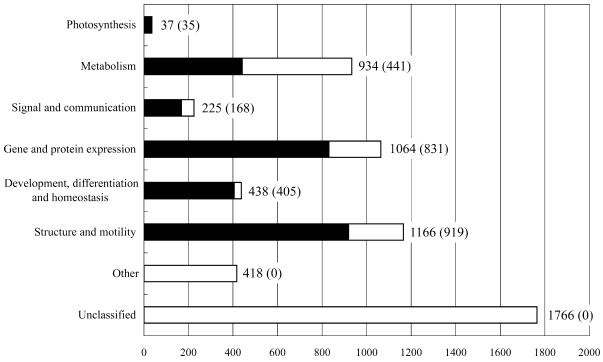

Figure 1. Classification of the 4,449 ESTs identified according to putative gene functions. White bars indicate ESTs classified into a single category and black bars indicate those classified into more than one category. The number at the end of each bar refers to the number of ESTs classified into the category, as well as the number of those classified into other categories in parentheses. The category 'unclassified' includes ESTs homologous to those from other species and to functionally unknown proteins such as hypothetical proteins.

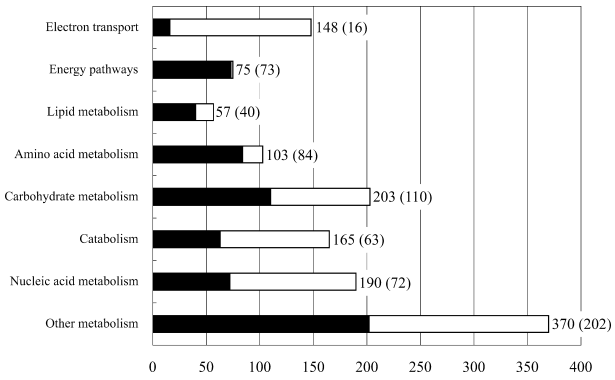

Figure 2. Further classification of the ESTs in the category 'metabolism'. White bars indicate ESTs classified into a single category and black bars indicate those classified into more than one category. The number at the end of each bar refers to the number of ESTs classified into the category, as well as the number of ESTs classified into other categories in parentheses.

Results and discussion

Qualification of cDNA library and evaluation of ESTs

The average insert size of the *E. tirucalli* callus cDNA library was estimated as 1.2 kb by agarose gel electrophoresis after PCR amplification with T7 and M13 forward primers. A total of 9,301 sequences were obtained, and the average read length was 550 bp. The average GC content of the ESTs was 41.6%, similar to that of *Arabidopsis thaliana* ESTs (Asamizu et al. 2000). Sequence comparison among 9,301 ESTs clones revealed that the 4,342 ESTs had overlaps with one or more EST clones and the remaining 4,959 ESTs were independent. Those redundant sequences were assembled into 1,252 independent clusters. Therefore, the redundancy of the EST set was 46.7%, and the number of unique EST sequences was 6,211.

Database search

Database search was performed for the non-redundant 6,211 ESTs to identify ESTs that are homologous to already characterized genes. The 1,762 ESTs showed no significant matches and might represent novel genes in

Table 1. Sterol, triterpene and other isoprenoid biosynthesis-related ESTs.

EST[a]	UniProt Id[b]	Putative Identification	E-value[c]
ESTs appear to code for enzymes required for the 2,3-epoxysqualene synthesis.			
ETC020E09	Q944G3	Acetyl CoA acetyltransferase	6.0E-35
ETC000A12	Q944F9	Truncated acetyl CoA acetyltransferase	4.0E-75
ETC090E03	Q944F9	Truncated acetyl CoA acetyltransferase-like protein	7.0E-95
ETC013G06	Q6QLW8	HMG-CoA synthase 2	2.0E-49
PTA.877.C1 (2)	Q944G2	Mevalonate kinase	1.0E-117
ETC060F07	Q944G1	Phosphomevalonate kinase	1.0E-28
ETC017F04	Q8S3L8	Isopentenyl pyrophosphate isomerase IDI2	3.0E-36
ETC104A10	Q6YZ75	Putative methylcrotonyl-CoA carboxylase beta chain, mitochondrial	2.0E-14
ETC015B08	Q8GTR1	Geranylgeranyl diphosphate synthase	2.0E-84
ETC041B05	Q94ID7	Geranylgeranyl diphosphate synthase	1.0E-91
PTA.906.C1 (2)	Q8L7F4	Farnesyl diphosphate synthase	1.0E-108
ETC033F09	Q94IE8	Putative FPP synthase 2	2.0E-82
ESTs code for enzymes required for the triterpene synthesis from 2,3-epoxysqualene.			
ETC010D09	Q8W3Z1	β-amyrin synthase	2.0E-66
ESTs code for proteins required for the sterol biosynthesis from 2,3-epoxysqualene.			
ETC031B10	O82139	Cycloartenol synthase	1.0E-95
ETC094C04	Q8LCB0	Sterol-C-methyltransferase	5.0E-63
PTA.331.C1 (2)	Q8L7L3	24-sterol C-methyltransferase	3.0E-99
ETC092C01	Q9SAA9	Obtusifoliol 14-demethylase	1.0E-101
ETC045H03	Q8GZV0	Obtusifoliol-14-demethylase	3.0E-82
ETC032C09	Q71V05	Sterol-4-methyl-oxidase	1.0E-37
ETC034C09	Q9FR36	Putative steroid reductase	2.0E-17
ETC059D01	Q9FWZ2	Putative sterol desaturase	2.0E-89
ETC116H04	Q38944	Probable steroid reductase DET2	1.0E-40
ETC020F02	Q940Y1	Oxydosterol binding protein	8.0E-93
ETC095B10	Q940Y1	Oxydosterol binding protein	8.0E-51
ETC058F03	Q84Z91	Oxysterol-binding protein-like	3.0E-43
PTA.675.C1 (4)	Q9XFM5	Putative progesterone-binding protein homolog	2.0E-68
ETC032A05	Q9ATR0	Brassinosteroid biosynthetic protein LKB	2.0E-86
ETC065H09	Q6PQJ9	Progesterone 5-beta-reductase	5.0E-46
ETC088H08	Q84TL0	Progesterone 5-beta-reductase	3.0E-89
ETC053H05	Q9FMN0	SCP-2 sterol transfer family	2.0E-39
ETC020D04	Q6XSH3	Putative lecithine cholesterol acyltransferase	2.0E-92
ETC054G05	Q9XIG1	Putative UDP-glucose:sterol glucosyltransferase (At1g43620)	1.0E-07
ETC083B09	Q8H9B4	UDP-glucose:sterol 3-O-glucosyltransferase	3.0E-73
ETC099G03	Q9FT43	Sterol glucosyltransferase-like protein	1.0E-24
ETC006F04	Q9M1V2	Sulfotransferase-like protein (At3g45070)	3.0E-17
ETC011B01	Q9M1V1	Sulfotransferase-like protein	6.0E-38
ETC089B08	O82410	Steroid sulfotransferase 3	3.0E-61
ETC095F05	Q8L5A7	Steroid sulfotransferase-like protein (At5g07010)	1.0E-55
PTA.750.C1 (2)	Q8L5A7	Steroid sulfotransferase-like protein (At5g07010)	1.0E-40
ETC034B09	Q7XIL3	Putative 5-alpha-taxadienol-10-beta-hydroxylase	1.0E-31
ETC104C04	Q7XIL3	Putative 5-alpha-taxadienol-10-beta-hydroxylase	9.0E-44

[a] EST names beginning with "PTA" indicate clusters, and the number of clones in each cluster is given in the parentheses.
[b] The accession numbers of the UniProt database at the European Bioinformatics Institute.
[c] The E (expect)-value is the probability that the associated match is due to randomness. The lower the E value, the more specific/significant is the match.

E. tirucalli (data not shown). The remaining 4,449 ESTs were found to be similar to genes coding for protein or ESTs registered in the public databases and classified by putative functions (Figure 1). It should be noted that the ESTs were classified to more than one category. The 1,166 ESTs belonged to the largest functional category of 'structure and motility'. The next largest category contains 1,064 ESTs that are presumably involved in 'gene and protein expression'. The category 'unclassified' includes 1,766 ESTs homologous to those from other species and to functionally unknown proteins such as hypothetical proteins.

The category 'metabolism' was further classified into eight categories (Figure 2). Two-hundred and three ESTs belong to the largest functional category 'carbohydrate metabolism'. The next largest category contains 190 ESTs that are presumably involved in 'nucleic acid metabolism'.

Putative genes for triterpenoid and sterol biosyntheses

Based on the annotations assigned to their hit entries in the UniProt database, 48 ESTs were found to presumably participate in the biosynthetic pathway starting from acetyl CoA to triterpenoids or sterols (Table 1).

Squalene is a key precursor of both triterpenoids and sterols, and 14 out of the 48 ESTs appear to code for enzymes required for the squalene biosynthesis, such as hydroxymethylglutaryl-CoA (HMG-CoA) synthase, mevalonate kinase and farnesyl diphosphate synthase (Table 1). After epoxydized to an intermediate (2,3-epoxysqualene), squalene is cyclized to multiple triterpenes, such as β-amyrin, and to cycloartenol, which is a precursor of sterols, respectively. Thirty-three ESTs appear to code for proteins required for biosynthesis of diverse sterols, such as 24-sterol C-methyltransferase, obtusifoliol 14-demethylase and sterol-4-methyl-oxidase (Table 1).

Cytochrome P450 is known to participate in a variety of biochemical reactions, including triterpenoid and sterol metabolisms. We found 42 ESTs which putatively code for cytochrome P450-like proteins (Table 2). Although further functional identification is indispensable, it is likely that some of the cytochrome P450-like proteins indeed mediate reactions required for triterpene and sterol biosyntheses in *E. tirucalli*.

We have identified a number of ESTs which represent candidate genes for the triterpenoid and sterol metabolisms in *Euphorbia*. Isolation of full-length clones and their functional characterization in the future should lead to development of a biological system for triterpenoid and sterol production.

Table 2. ESTs putatively coding for cytochrome P450.

EST[a]	UniProt Id[b]	Putative Identification	E-value[c]
ETC051A01	Q6WNQ8	CYP81E8 P450	4.0E-61
ETC070C03	Q6WNQ8	CYP81E8 P450	6.0E-38
ETC030A06	O48925	CYP82C1p P450	1.0E-45
ETC010C10	Q9XFX1	Cytochrome P450	2.0E-16
ETC021E08	Q9LUC5	Cytochrome P450	4.0E-40
ETC046E06	Q9AVQ2	Cytochrome P450	2.0E-50
ETC056A03	Q9LUD3	Cytochrome P450	2.0E-37
ETC087H10	Q8H0I8	Cytochrome P450	2.0E-52
ETC092A12	Q6QNI1	Cytochrome P450	2.0E-44
ETC107F07	Q8H0I7	Cytochrome P450	4.0E-43
PTA.587.C1 (3)	Q9LSF8	Cytochrome p450	2.0E-44
PTA.590.C1 (4)	Q9FVM9	Cytochrome P450	1.0E-134
PTA.646.C1 (5)	Q6QNI1	Cytochrome P450	5.0E-63
ETC092E06	Q944N6	Cytochrome P450 (Fragment)	4.0E-10
PTA.400.C1 (3)	Q944N6	Cytochrome P450 (Fragment)	1.0E-63
PTA.942.C1 (2)	O81970	Cytochrome P450 71A9 (P450 CP1)	2.0E-49
ETC111D04	P93531	Cytochrome P450 71D7	3.0E-49
PTA.544.C1 (2)	O81971	Cytochrome P450 71D9 (P450 CP3)	6.0E-46
ETC091G07	P37122	Cytochrome P450 76A2 (CYPLXXVIA2) (P-450EG7)	7.0E-30
PTA.113.C1 (3)	Q9FG65	Cytochrome P450 81D1	7.0E-93
ETC080D08	Q8LF37	Cytochrome P450, putative	2.0E-55
ETC051H09	O49394	Cytochrome P450-like protein	2.0E-32
ETC064H12	Q8S7S6	Cytochrome P450-like protein	6.0E-44
PTA.98.C1 (2)	Q9ASR3	Putative cytochrome P450	2.0E-18
ETC033H05	O64697	Putative cytochrome P450 (At2g34500)	5.0E-68
ETC100E12	Q94FM4	Elicitor-inducible cytochrome P450	3.0E-67

[a] EST names beginning with "PTA" indicate clusters, and the number of clones in each cluster is given in the parentheses.
[b] The accession numbers of the UniProt database at the European Bioinformatics Institute.
[c] The E (expect)-value is the probability that the associated match is due to randomness. The lower the E value, the more specific/significant is the match.

Acknowledgements

This work was performed as one of the technology development projects of the "Green Biotechnology Program" supported by NEDO (New Energy and Industrial Technology Development Organization). We thank Y. Watanabe and F. Sato for the use of their facility and invaluable discussions. M.K., Y.K. and R.S. were supported by the 21st Century COE Program of the Ministry of Education, Culture, Sports, Science and Technology.

References

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

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein Knowledgebase. *Nucl Acids Res* 32: D115–D119

Asamizu E, Nakamura Y, Sato S, Tabata S (2000) A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. *DNA Res* 7: 175–180

Biesboer DD, Mahlberg PG (1979) The effect of medium modification and selected precursors on sterol production by short-term callus cultures of *Enphorbia tirucalli*. *J Natur Products* 42: 648–657

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucl Acids Res* 32: D262–D266

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175–185

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194

Gamborg OL (1970) The effects of amino acids and ammonium on the growth of plant cells in suspension culture. *Plant Physiol* 45: 372–375

Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. *Genome Res* 8: 195–202

Nielsen PE, Nishimura H, Otvos JW, Calvin M (1977) Plant crops as a source of fuel and hydrocarbon-like materials. *Science* 198: 942–944

Ohlrogge J, Benning C (2000) Unraveling plant metabolism by EST analysis. *Curr Opin Plant Biol* 3: 224–228

Ohyama K, Misawa N, Yamano Y, Komano T (1984a) Protoplast isolation from *Euphorbia tirucalli* L. cell suspension cultures and sustained cell division. *Z Pflanzenphysiol* 113: 367–370

Ohyama K, Uchida Y, Misawa N, Komano T, Fujita M, Ueno T (1984b) Oil body formation in *Euphorbia tirucalli* L. cell suspension cultures. *Plant Cell Rep* 3: 21–22

Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarroel R, Van Montagu M, Sandberg G, Olsson O, Teeri TT, Boerjan W, Gustafsson P, Uhlen M, Sundberg B, Lundeberg J (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. *Proc Natl Acad Sci USA* 95: 13330–13335

van de Loo FJ, Broun P, Turner S, Somerville C (1995) An oleate 12-hydroxylase from *Ricinus communis* L. is a fatty acyl desaturase homolog. *Proc Natl Acad Sci USA* 92: 6743–6747

Yamamoto Y, Mizuguchi R, Yamada Y (1981) Chemical Constituents of cultured cells of *Euphorbia tirucalli* and *E. milii*. *Plant Cell Rep* 1: 29–30