Transcriptome analysis of an oil-rich race B strain of *Botryococcus braunii* (BOT-70) by *de novo* assembly of 5'-end sequences of full-length cDNA clones

Motohide Ioki [a, 1], Masato Baba [b, 1], Nobuyoshi Nakajima [a, b, *], Yoshihiro Shiraiwa [b], Makoto M. Watanabe [b]

[a] Center for Environmental Biology and Ecosystem Studies, National Institute for Environmental Studies, Onogawa 16-2, Tsukuba, Ibaraki, 305-8506 Japan

[b] Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan

Footnotes

* Corresponding author: E-mail, naka-320@nies.go.jp; FAX, 81-29-850-2490

[1] These authors contributed equally to this work.

Abstract

Here the transcriptome of an oil-rich race B strain of *Botryococcus braunii* (BOT-70) was analyzed to mine genetic information useful in biofuel development. A full-length-enriched cDNA library was constructed via the oligo-capping method and the 5' ends of 11,904 randomly chosen cDNA clones were sequenced. Homology search using BLASTX identified candidate BOT-70 genes for majority of the reactions required for biosynthesis of botryococcenes through the mevalonate-independent pathway. The sequence retrieval from the transcriptome dataset implicated that an alternative entry route into the mevalonate-independent pathway via

xylulose-5-phosphate, rather than the conventional entry route via 1-deoxy-D-xylulose-5-phosphate, is predominantly active. Analysis of N-terminal sequences of the retrieved genes indicated that the final reactions of botryococcene biosynthesis are likely to take place outside of chloroplasts. The transcriptome dataset has been deposited in the GenBank/EMBL/DDBJ database.

Keywords: Expressed sequence tag (EST); Hydrocarbon; Triterpenoid; Transcriptome; Non-mevalonate pathway

1. Introduction

The hydrocarbon oil produced by *Botryococcus braunii* resembles petroleum and this organism is anticipated for practical industrial use. To improve the cost effectiveness of oil production using *B. braunii* for potential industrial use, understanding the oil biosythesis pathways and their regulation is critical. There are two major types of hydrocarbon oils produced by *B. braunii* as represented by the A and B chemical races: hydrocarbons derived from unsaturated very long-chain fatty acids and triterpenes, respectively. Race L strains, similarly to race B strains, synthesize terpenoid-derived hydrocarbons referred to as lycopadienes. Race A and B strains are more abundant in nature and generally exhibit higher oil contents than race L strains (reviewed in Banerjee et al., 2002).

Terpenoid-derived hydrocarbons produced by race B strains are called botryococcenes. Botryococcenes are synthesized from an intermediate compound called isopentenyl-pyrophosphate (IPP). IPP is synthesized via either the mevalonate pathway

or the mevalonate-independent pathway. While the mevalonate pathway is widely accepted for the isoprenoid biosynthesis, the mevalonate-independent pathway is often predominantly active in phototrophic eukaryotes (reviewed in Rohmer, 1999). Previously, feeding experiments using radioactive isotopes have demonstrated that botryococcenes are synthesized through the mevalonate-independent pathway rather than the mevalonate pathway (Sato et al., 2003). Typically, pyruvate and glyceraldehyde-3-phosphate are the starting materials for the IPP synthesis via the mevalonate-independent pathway. In addition, an uncharacterized pathway analogous to such a mevalonate-independent pathway has been reported to be the major pathway for terpenoid biosynthesis in the cyanobacterium *Synechocystis* sp. strain PCC6803 under photosynthetic conditions (Ershov et al., 2002; Poliquin et al., 2004). However, it is not clear which entry route is active in *B. braunii*. The genes encoding enzymes associated with hydrocarbon biosynthesis in *B. braunii* are largely unknown except for the squalene synthase gene and its paralogs (Okada et al., 2000; Niehaus et al., 2011). Here the transcriptome sequencing dataset of a representative race B strain of *B. braunii* (BOT-70) was analyzed to mine genetic information useful in biofuel development.

2. Methods

Total RNA used for the complementary DNA (cDNA) library construction was extracted using RNeasy Plant Mini Kit (Qiagen Inc., Chatsworth, CA, USA) according to the manufacturer's instruction. The BOT-70 strain was cultured axenically in the modified Chu medium (0.2 g/L KNO_3, 0.04 g/L K_2HPO_4, 0.1 g/L $MgSO_4 \cdot 7H_2O$, 0.054 g/L $CaCl_2 \cdot 2H_2O$, 0.98 mg/L $FeCl_3 \cdot 6H_2O$, 0.18 mg/L $MnCl_2 \cdot 4H_2O$, 0.11 mg/L $ZnSO_4 \cdot$

$7H_2O$, 12.5 ng/L $Na_2MoO_4 \cdot 2H_2O$, 0.02 mg/L $CoCl_2 \cdot 6H_2O$, 5 mg/L $Na_2EDTA \cdot 2H_2O$, pH 7) at 25°C under white fluorescence light. The photosynthetic photon flux density (PPFD) was 50 µmole photons m^{-2} s^{-1}, as measured with a quantum meter (LI-1000, Li-Cor, Lincoln, NE, USA). The active cultures were maintained by inoculating 5 L of a fresh medium with 2 L of the preceding culture twice a month. Cells were harvested at 1, 2, 3, and 4 weeks after the inoculation using a 5-µm filter and immediately frozen in liquid nitrogen. Contaminating DNA was eliminated using RNase-Free DNase Set (Qiagen Inc.). The total RNA from the 1, 2, 3, and 4-week samples were combined with equal proportions for the cDNA construction.

A full-length-enriched cDNA library of the BOT-70 strain of *B. braunii* was constructed via the oligo-capping method as described by Tang et al. (2006). To estimate the average length of the cDNA insertions, the plasmid DNAs isolated from randomly chosen 16 independent bacterial clones were subjected to restriction digestion at the *Xho*I restriction sites, and subsequently separated according the molecular weight on a agarose gel by electrophoresis.

A novel dataset of ESTs was acquired by determining the 5'-end nucleotide sequences of approximately 10,000 randomly chosen cDNA clones. The 5' ends of the cDNA insertions were sequenced using a primer designed upstream of the cloning site (5'- TACGGAAGTGTTACTTCTGC -3'). The sequencing was performed via the dye terminator method using TaKaRa PCR Thermal Cycler GP (Takara Bio Co. Ltd, Shiga, Japan) and MegaBACE4000 (GE Healthcare UK Ltd, Little Chalfont, Buckinghamshire, England) according to the manufacturer's instructions. The cDNA reads were filtered, clustered, and assembled into non-redundant sequences using the Paracel TranscriptAssembler^{TM} Version 2.7 software (Paracel, Pasadena, CA) under the

auspices of the Dragon Genomics Center, Takara Bio Co. Ltd, Japan.

The non-redundant sequences were annotated and classified into different functional categories using KEGG Automatic Annotation Server (KAAS) based on amino acid sequence homology determined by the BLASTX algorithm (Moriya et al., 2007). Additional hydrocarbon biosynthesis-related genes were identified based on the enzyme nomenclature (accepted and alternative names) using the EST Viewer platform (Dragon Genomics Center, Takara Bio Co. Ltd, Japan) fed with the BLASTX search hits from the non-redundant (nr) database (ver. 2009.05.21) compiled by National Center for Biotechnology Information (NCBI). The cutoff value for BLASTX was set for the bit score of 50, which corresponds to the *e*-value of approximately 1×10^{-5}.

To predict the subcellular localization of the enzymes encoded by these genes, the TargetP software was used to examine the presence of N-terminal sequence motifs directing proteins to the secretory pathway, mitochondria, and chloroplasts as described by Emanuelsson et al. (2007).

3. Results and Discussion

In this study, a full-length-enriched cDNA library was constructed for the BOT-70 strain. This strain was selected because of its excellence in both growth and oil productivity (Tanoi et al., 2011). This strain produces liquid hydrocarbons ($C_{34}H_{58}$ cyclohexene) derived from triterpenes (personal communication with Takako Tanoi, Masanobu Kawachi, and Kunimitsu Kaya). The average insertion length of the cDNA library was estimated to be approximately 1.8kb (data not shown).

Expressed sequence tags (ESTs) were obtained by sequencing the 5’ ends of randomly chosen 11,866 cDNA clones via the conventional Sanger method. By filtering the cDNA reads, 9,345 high-quality sequences were obtained. The high-quality cDNA sequences can be found in the GenBank/EMBL/DDBJ data libraries under accession numbers FY358876 through FY368220. Majority of the cDNA reads were of 500 to 700 base pairs in length (Fig. 1a). The high-quality sequences were assembled into 1,868 non-redundant sequences (Table S1). The average length of the non-redundant sequences was 650 base pairs and there were sequences of up to 2,424 base pairs in length (Fig. 1b). The GC content calculated from all 1868 non-redundant sequences was 51.7%. Some of the non-redundant sequence consisted of more than 400 reads, while singleton sequences were the most abundant (Fig. 1c). Sequence data of the non-redundant sequences can be found in Table S1.

Successful enrichment of full-length clones was confirmed by analyzing all 201 non-redundant sequences exhibiting homology to publicly available database sequences with BLASTX bit scores greater than 200. Among the 201 non-redundant sequences, 162 sequences (80.6%) contained 5’-untranslated regions. The mean length and GC content of these 5’-untranslated regions were 94.8 base pairs and 46.4%, respectively.

To grasp the overall picture of the EST dataset, the non-redundant sequences were annotated and classified into different functional categories. Relatively reliable prediction of function was feasible for 509 non-redundant sequences with the cutoff bit score of 50 determined by the BLASTX algorithm. The annotations of the non-redundant sequences can be found in Table S1. A relatively large number of genes (46.8 % of lipid metabolisms-related genes) were predicted to be associated with

terpenoid biosynthesis, reflecting the triterpene-producing nature of the BOT-70 strain (Table S2). The top 30 genes with largest EST counts are shown in Table 1. Genes associated with photosynthesis were the most abundantly expressed. Interestingly, many of the genes with large EST counts did not exhibit any significant homology to database sequences. This may indicate that nucleotide and amino acid sequences tend to be distinctive in *B. braunii*.

Ten non-redundant sequences, consisting of 19 cDNA reads, associated with botryococcene biosynthesis were discovered from the transcriptome data. No genes were found for enzymes in the mevalonate pathway for IPP biosynthesis (Table 2). Contrarily, candidate genes were retrieved for enzymes catalyzing 4 out of the 8 reactions of the mevalonate-independent pathway. For the triterpene biosynthesis from IPP, ESTs were retrieved for 4 out of 5 reactions (Table 2, Table S3). The transcriptome profile of the BOT-70 strain, therefore, indicated that the botryococcene biosynthesis is unlikely to take place via the mevalonate pathway. To date, there is no document of active mevalonate pathway in any green algae. According to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, many of the mevalonate pathway-related genes are missing in the genomes of green algal species with sequenced genomes such as *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, *Ostreococcus lucimarinus*. The BOT-70 transcriptome was also consistent with preceding biochemical studies demonstrating that the mevalonate pathway is not active in *B. braunii* (Sato et al., 2003). The BOT-70 transcriptome provided a new piece of evidence supporting the idea that the mevalonate-independent pathway is the major pathway for terpenoid biosynthesis in green algae.

The sequence retrieval from the BOT-70 EST dataset indicated that a unique

shunt into the mevalonate-independent pathway, which so far has been reported only for cyanobacteria, is active in *B. braunii*. Besides the conventional entry route into the mevalonate-independent pathway via the formation of 1-deoxy-D-xylulose-5-phosphate by 1-deoxy-D-xylulose-5-phosphate synthase (EC 2.2.1.7), a shortcut entry route via the xylulose-5-phosphate formation has been suggested in cyanobacteria. This alternative entry path allows direct flow of photosynthetic products of the reductive pentose phosphate cycle into IPP biosynthesis and is the major pathway for isoprenoid synthesis in *Synechocystis* sp. PCC6803 under photosynthetic conditions. It is not clear whether the same or different enzymes are responsible for these two analogous mevalonate-independent pathways (Ershov et al., 2002; Poliquin et al., 2004). Failure in EST retrieval for enzymes associated with the conventional entry route and abundant ESTs for enzymes involved in the synthesis of xylulose-5-phosphate suggested the dominance of metabolic flow via the alternative path in BOT-70 (Table 2, Table S3).

In order to gain insight into the cytological site of botryococcene oil biosynthesis, the N-terminal amino acid sequences encoded by the botryococcene biosynthesis-related genes were inspected for the presence of localization signal sequences using the TargetP software (Emanuelsson et al., 2007). As a result, a chloroplast transit peptide was found in the N-terminus of one of the mevalonate-independent pathway enzymes, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (EC 4.6.1.12). On the other hand, absence of a chloroplast transit peptide was predicted with high reliability for the enzyme catalyzing the final reaction of the pathway, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase

(EC 1.17.1.2). A signal peptide directing the protein to the secretory pathway was found in the N-terminus of the putative isopentenyl-diphosphate Delta-isomerase (EC 5.3.3.2), which catalyzes the final reaction in the IPP biosynthesis. This enzyme, therefore, is likely to be directed into the secretory pathway. The final cytological destination of this enzyme, however, remains unclear. It is possible that that the reaction takes place outside the cell. Alternatively, the reaction may take place within the cell but outside the chloroplasts. For the other enzymes associated with the mevalonate-independent pathway and the triterpene biosynthesis pathway, no cytological localization signals were detected with adequate reliability. Taken together, the early reactions of botryococcene biosynthesis seem to take place in the chloroplasts and the final reactions outside the chloroplast. This is not surprising given that the isopentenyl-diphosphate Delta-isomerase does not localize in chloroplasts in *Arabidopsis* (Okada et al., 2008).

4. Conclusions

In this study, the transcriptome dataset for the oil-rich BOT-70 strain of *B. braunii* was acquired by determining the 5'-end nucleotide sequences of 9,345 full-length cDNA clones. The large set of full-length cDNA clones with determined 5' sequences would be a powerful and versatile genetic resource for biofuel development. The sequenced transcriptome embraced many genes associated with botryococcene biosynthesis from photosynthetic products through the mevalonate-independent pathway. N-terminal amino acid sequences of the botryococcene biosynthesis-related genes suggested that botryococcenes are synthesized via isoprenoid synthesis in chloroplasts followed by triterpene synthesis elsewhere.

Acknowledgements

Takako Tanoi, Yurie Akutsu, and Haniyeh Bidadi (National Institute for Environmental Studies) provided technical assistance. This research was supported by the Core Research for Evolutionary Science and Technology program of Japan Science and Technology Agency.

References

Banerjee, A., Sharma, R., Chisti, Y., Banerjee, U.C., 2002. *Botryococcus braunii*: a renewable source of hydrocarbons and other chemicals. Crit Rev Biotechnol. 22(3), 245-279.

Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2(4), 953-971.

Ershov, Y.V., Gantt, R.R., Cunningham, F.X. Jr., Gantt, E., 2002. Isoprenoid Biosynthesis in *Synechocystis* sp. Strain PCC6803 Is Stimulated by Compounds of the Pentose Phosphate Cycle but Not by Pyruvate or Deoxyxylulose-5-Phosphate. J. Bacteriol. 184(18), 5045-5051.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., Kanehisa, M., 2007. KAAS, an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res.

35, W182-185.

Niehaus, T.D., Okada, S., Devarenne, T.P., Watt, D.S., Sviripa, V., Chappell, J. 2011. Identification of unique mechanisms for triterpene biosynthesis in *Botryococcus braunii*. Proc Natl Acad Sci U S A. 108(30), 12260-12265.

Okada, K., Kasahara, H., Yamaguchi, S., Kawaide, H., Kamiya, Y., Nojiri, H., Yamane, H., 2008. Genetic evidence for the role of isopentenyl diphosphate isomerases in the mevalonate pathway and plant development in Arabidopsis. Plant Cell Physiol. 49(4), 604-616.

Okada, S., Devarenne, T.P., Chappell, J., 2000. Molecular characterization of squalene synthase from the green microalga *Botryococcus braunii*, race B. Arch. Biochem. Biophys. 373(2), 307-317.

Poliquin, K., Ershov, Y.V., Cunningham, F.X. Jr., Woreta, T.T., Gantt, R.R., Gantt, E., 2004. Inactivation of sll1556 in Synechocystis strain PCC 6803 impairs isoprenoid biosynthesis from pentose phosphate cycle substrates in vitro. J. Bacteriol. 186(14), 4685-4693.

Rohmer, M., 1999. The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. Nat. Prod. Rep. 16, 565–574.

Sato, Y., Ito, Y., Okada, S., Murakami, M., Abe, H., 2003. Biosynthesis of the

triterpenoids, botryococcenes and tetramethylsqualene in the B race of *Botryococcus braunii* via the non-mevalonate pathway. Tetrahedron Lett. 44, 7035-7037.

Tang, S.W., Chang, W.H., Chao, Y.W., Lin, C.Y., Chen, H.F., Lai, Y.H., Zhan, B.W., Su, Y.C., Jane, S.W., Chen, Y.C., Hsu, C.I., Lin, W.C., Wang, K.C., Lai, M.K., Lin, J.Y., 2006. Identification of differentially expressed genes in clear cell renal cell carcinoma by analysis of full-length enriched cDNA library. J. Biomed. Sci. 13(2), 233-240.

Tanoi, T., Kawachi, M., Watanabe, M.M., 2011. Effects of carbon source on growth and morphology of *Botryococcus braunii*. J. Appl. Phycol. 23(1), 25-33.

Figure Captions

Fig. 1 Features of the BOT-70 EST dataset generated by 5'-end sequencing of clones in full-length-enriched cDNA library. a, Distribution of EST lengths (after filtering). **b**, Length distribution of non-redundant sequences obtained by assembling of ESTs. **c**, Distribution of EST counts for individual non-redundant sequences.

P.S. please check fig and tables at

http://dx.doi.org/10.1016/j.biortech.2011.11.047