Transcriptome analysis of an oil-rich race A strain of *Botryococcus braunii* (BOT-88-2) by *de novo* assembly of pyrosequencing cDNA reads

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

[a] Graduate School of Life and Environment Sciences, University of Tsukuba, Tennoudai 1-1-1, Tsukuba, Ibaraki, 305-8572 Japan

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

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

[1] These authors contributed equally to this work.

Abstract

To gain genetic information of oil-producing algae *Botryococcus braunii*, a novel dataset of 185,936 complementary DNA (cDNA) reads was obtained via pyrosequencing for the representative race A strain (strain BOT-88-2) exhibiting high oil productivity. The cDNA reads were assembled to retrieve 29,038 non-redundant sequences and 964 of them were successfully annotated based on similarity to database sequences. The transcriptome data embraced candidate genes for majority of enzymes involved in the biosynthesis of unsaturated very long-chain fatty acids. The transcriptome dataset has been deposited in the GenBank/EMBL/DDBJ database.

Keywords: Expressed sequence tag (EST); Hydrocarbon; 454 pyrosequencing;

Transcriptome; Very long-chain fatty acid

1. Introduction

The oil-producing green alga *Botryococcus braunii* is one of the relatively promising renewable sources of petroleum substitutes. *B. braunii* have been found in fresh and brackish water worldwide. *B. braunii* strains are grouped into different races, namely, race A, B, and L, depending on the types of liquid hydrocarbons they produce. The race A strains produce alkadienes and alkatrienes, while the race B strains produce triterpenoid hydrocarbons called botryococcenes and methylated squalenes. The race L strains synthesize tetraterpenes referred to as lycopadienes (reviewed in Banerjee et al., 2002; Metzger and Largeau, 2005).

The race A hydrocarbon products, alkadienes and alkatrienes, are thought to be synthesized from very long-chain fatty acids (VLCFAs). Unsaturated fatty acids are the direct precursors of VLCFA elongation (reviewd in Banerjee et al. 2002; Metzger and Largeau 2005). Conversion of fatty aldehyde, one of the VLCFA derivates, into alkadiene accompanying the terminal double bond insertion seems to be the final step of the alkadiene synthesis in *B. braunii* (Dennis and Kolattukudy, 1991).

Here expression sequence tags (ESTs) of a representative race A strain of *B. braunii* (BOT-88-2) were examined. The BOT-88-2 strain was used because of its excellence in both growth and oil productivity. This strain produces alkadienes ((1*E*,18*E*)-1,18-heptacosadiene [55%]) and alkatrienes ((1*E*,3*E*,18*E*)-1, 3, 18-heptacosatriene [45%]) derived from unsaturated VLCFAs (personal communication with Takako Tanoi, Masanobu Kawachi, and Kunimitsu Kaya). The EST dataset has

been deposited in the GenBank/EMBL/DDBJ database.

2. Materials and Methods

The original axenic culture of the BOT-88-2 strain of *B. braunii* was provided from Dr. Masanobu Kawachi (National Institute for Environmental Studies). The cells were grown under commonly used optimal growth conditions as described in Ioki et al. (2011, this issue) at 25 °C under continuous white light of 150 μmol$m^{-2}s^{-1}$ with bubbling of filtered air. Because of relatively slow growth of *B. braunii* with doubling time of approximately 1 week, the culture was maintained as an active culture (i.e. no lagging or stationary phases). Under these conditions, the BOT-88-2 strain exhibits oil contents around 50 % oil per dry weight (personal communication with Takako Tanoi, Masanobu Kawachi, and Kunimitsu Kaya). The cells for the present study were harvested at a random time from the actively growing culture using a 5-μm filter and immediately frozen in liquid nitrogen.

The Genome Sequencer-FLX System (Roche diagnostics, Basel, Switzerland) was used for the EST acquisition. Total RNA was extracted from cells harvested from a 250-ml culture using RNeasy Plant Mini Kit (Qiagen Inc., Chatsworth, CA, USA). Contaminating DNA was eliminated using RNase-Free DNase Set (Qiagen Inc.). Amplification of antisense RNA (aRNA) was performed as described by Vega-Arreguín et al. (2009) using 2 μg of total RNA. Reverse transcription was performed using 10 μg of aRNA and random 6 mers. Then double-stranded cDNA was synthesized using biotinylated oligo(dT) primers with the GS-FLX adaptor. The aRNA-derived cDNA was fragmented by sonication and separated by electrophoresis on an agarose gel.

Fragments of 300 to 800 bp were excised out of the gel. Then, 3'-terminal fragments were collected using streptavidin-coated magnetic beads (Dynal, Invitrogen, Carlsbad, CA). On the 5'-ends of these fragments, the other GS-FLX adaptors were conjugated. sstDNA were collected using magnetic beads. The resulting cDNA library was quantified by fluorometry (Quant-iT Ribogreen, Invitrogen), clonally amplified by emulsion PCR and sequenced according to the standard procedures for 454 sequencing.

The cDNA reads were filtered, clustered, and assembled into non-redundant sequences using the Paracel TranscriptAssembler^{TM} Version 2.6 software (Paracel, Pasadena, CA) under the auspices of the Dragon Genomics Center, Takara Bio Co. Ltd, Japan. In the filtering process, (i) poly A/T sequences longer than 7 bases within 30-base regions of 5'- and 3'-ends of the cDNA reads detected by the HASTE algorithm and (ii) repetitive sequences of more than 21 bases detected by the DUST algorithm were marked not to be used in the clustering process. Terminal sequences with low quality values, the adaptor sequences, and short sequences of less than 50 bases were removed. In the clustering process, the filtered sequences were grouped into clusters locally sharing common sequences detected by one-to-one comparisons. The sequences within each cluster were subjected to the assembling procedure to retrieve non-redundant sequences based on global similarity detected by the CAP4 algorithm.

The non-redundant sequences were annotated and classified into different functional categories using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) program (Moriya et al., 2007) based on the BLASTX algorithm. Additional hydrocarbon biosynthesis-related genes were identified using the EST Viewer software (Dragon Genomics Center, Takara Bio Co. Ltd, Japan) according to the enzyme nomenclature (accepted and alternative names) of the

BLASTX search hits from the non-redundant (nr) database (ver. 2009.01.09) compiled by National Center for Biotechnology Information (NCBI). All the BLASTX searches were performed with cutoff bit score of 50, which corresponds to the *e*-value of approximately 1 x 10^{-5}.

3. Results and Discussion

The pyrosequencing generated 185,936 cDNA reads. Majority of the cDNA reads were of 200 to 300 bases in length. The average read length was 217 bases (Fig. 1a). The cDNA reads are available in the DDBJ Sequence Read Archive (http://trace.ddbj.nig.ac.jp/dra/index_e.shtml) as DRR000585.

By assembling the cDNA reads, 29,038 non-redundant sequences of up to over 2 kb in length were retrieved. The average length of non-redundant sequences was 296 bases (Fig. 1b). Majority of the non-redundant sequences consisted of a single cDNA read, while some non-redundant sequences consisted of more than 2,000 cDNA reads (Table 1, Fig. 1c). The average GC content estimated using all non-redundant sequences was 49.5%. For 964 non-redundant sequences, the gene functions were successfully predicted based on homology to previously characterized genes of other organisms (Table S1). Metabolism-related genes comprised a significant portion of the annotated non-redundant sequences as in other algal transcriptomes (Rismani-Yazdi et al., 2011; Wahlund et al., 2004; Weber et al., 2004). More than 50% of the 964 non-redundant sequences were predicted to be associated with metabolism, and many of the metabolic genes were predicted to be associated with fatty acid biosynthesis, reflecting the race A hydrocarbon products of the BOT-88-2 strain. The non-redundant sequences can be

found in the DDBJ (http://www.ddbj.nig.ac.jp/index-e.html) under accession numbers FX056085 through FX085122.

Top 30 list of non-redundant sequences with largest EST counts is shown in Table 1. Photosynthesis-related genes were expressed at high levels. The non-redundant sequences with accession numbers FX056503 (677 reads), FX056901 (561 reads) and FX056504 (442 reads) showed similarity to photosystem-related genes, psbO, LHCB4 and psbP, respectively. The non-redundant sequence with accession number FX056922 (442 reads) showed similarity to an aquaporin, PIP. The non-redundant sequences with accession number FX056339 (419 reads) showed similarity to a carbonic anhydrase, cynT, which is involved in the CO_2 concentrating mechanism in many microalgal species (Giordano et al. 2005). Only 5 of the 30 non-redundant sequences showed significant similarity to known genes. Many of the highly expressed genes did not exhibit homology to database sequences (Table 1). Further analysis of these sequences may lead to discovery of novel genes.

Search for genes encoding enzymes potentially associated with hydrocarbon oil biosynthesis from acetyl-CoA retrieved 30 non-redundant sequences (accession numbers shown in Table S2). While the transcriptome data embraced putative genes associated with unsaturated VLCFA biosynthesis, the genes for the final conversion of unsaturated VLCFA to the hydrocarbon end products still remained unidentified (Table 2).

Fatty acid biosynthesis begins with conversion of acetyl-CoA into malonyl-CoA by acetyl-CoA carboxylase [EC 6.4.1.2] (Chan and Vogel, 2010). The BOT-88-2 transcriptome contained a total of 423 reads of ESTs for acetyl-CoA carboxylase [EC 6.4.1.2] (FX056429, 306 reads; FX056394, 112 reads; FX056430, 3

reads; FX056966, 1 reads; FX056123, 1 reads) (Table 2, Table S2). Such prevalence of acetyl-CoA carboxylase transcripts has been unprecedented in other algal transcriptomes and seemed to reflect the vigorous fatty acid synthesis in BOT-88-2 (Rismani-Yazdi et al., 2011; Wahlund et al., 2004; Weber et al., 2004).

In plants, fatty acid elongation mainly takes place in chloroplasts in the acyl carrier protein (ACP)-bound form (Ohlrogge et al., 1979; Chan and Vogel, 2010). Two classes of acyl-ACP elongation systems are known. The type I system involves a large multifunctional enzyme and is present primarily in eukaryotes, while the type II system involves discretely expressed mono-functional enzymes and is present primarily in bacteria and plants (Chan and Vogel, 2010). As expected, the BOT-88-2 transcriptome contained many ESTs associated with the type II system, namely, beta-ketoacyl-ACP synthase I [EC 2.3.1.41] (FX056905, 112 reads; FX056121, 1 read) and 3-oxoacyl-ACP reductase [EC 1.1.1.100] (FX056134, 1 read). Interestingly, an EST for the type I fatty-acid synthase [EC 2.3.1.85] (FX056119, 1 read) was also retrieved (Table 2, Table S2). This transcript suggests that fatty acid elongation may take place not only in chloroplasts but also in cytosol in BOT-88-2.

The fatty acid desaturation process, in general, can take place either the ACP-bound or the CoA-bound forms. The BOT-88-2 EST dataset contained both stearoyl-CoA 9-desaturase (FX056093, 91 reads; FX056090, 15 reads; FX056094, 2 reads) and acyl-ACP desaturase (FX056729, 2 reads), suggesting the desaturation leading to oleic acid production is catalyzed not only in ACP-bound form but also in CoA-bound form (Table 2, Table S2). This may be a unique feature of *B. braunii*. To our knowledge, stearoyl-CoA 9-desaturase is not expressed in other algae or plants. In animal, yeast, and fungal cells, acyl-CoA desaturases are bound to the endoplasmic

reticulum (Reviewed in Los and Murata, 1998). Mechanisms underlying the formation of the third desaturation of the alkatrienes remained completely unknown.

Interestingly, the BOT-88-2 transcriptome data embraced many genes associated with glycerolipid metabolism, namely 1-acylglycerol-3-phosphate O-acyltransferase [EC 2.3.1.51] (FX057047, 1 read, $2e^{-22}$), phospholipid: diacylglycerol acyltransferase [EC 2.3.1.158] (FX056197, 9 reads), glycerol kinase [EC 2.7.1.30] (FX056103, 126 reads; FX056104, 87 reads; FX056102, 53 reads; FX056105, 1 read), diacylglycerol kinase [EC 2.7.1.107] (FX056091, 3 reads; FX056234, 1 read), triacylglycerol lipase [EC 3.1.1.3] (FX056122, 10 reads; FX056099, 2 reads; FX056253, 1 read; FX056113, 1 read), lipoprotein lipase [EC 3.1.1.34] (FX056107, 6 reads) and phosphatidate phosphatase [EC 3.1.3.4] (FX078476, 1 read) were identified (Table 2, Table S2). Glycerolipids are important storage molecules of carbon in various organisms (Yen et al., 2008). The BOT-88-2 transcripts suggested active carbon exchanges between fatty acids and glycerolipids prior to entry into the VLCFA elongation pathway.

Unsaturated fatty acids are further elongated via the VLCFA synthesis pathway in plants (Plant Metabolic Network [http://plantcyc.org/]). Candidate genes associated with VLCFA elongation encoding enoyl-CoA hydratase [EC 4.2.1.17] (FX056771, 2 reads) and trans-2-enoyl-CoA reductase (NADPH) [EC 1.3.1.38] (FX056096, 46 reads; FX056097, 2 reads) were found in the BOT-88-2 transcriptome (Table 2, Table S2). For the final conversion of unsaturated VLCFA to the hydrocarbon end products, candidate BOT-88-2 genes still remained unidentified. As characterized in bacteria, head-to-head condensation of fatty acids catalyzed by a condensing enzyme referred to as OleA may take place to form hydrocarbons (Sukovich et al., 2010). However, no putative

homologs were identified in the *B. braunii* transcriptome with reliable homology (threshold *e*-value of 1x 10^{-5}) for the *oleA* gene.

4. Conclusion

In this study, the transcriptome sequencing data were obtained for the oil-rich BOT-88-2 strain of *B. braunii*. The sequenced transcriptome embraced genes for majority of enzymes associated with the race A oil biosynthesis of *B. braunii*. In the light of the world's urgent need for petroleum substitutes, the dataset could be a rich source of genes useful in biofuel development.

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.

Chan, D.I., Vogel, H.J., 2010. Current understanding of fatty acid biosynthesis and the acyl carrier protein. Biochem. J. 430, 1-19.

Dennis, M.W., Kolattukudy, P.E., 1991. Alkane biosynthesis by decarbonylation of aldehyde catalyzed by a microsomal preparation from *Botryococcus braunii*. Arch. Biochem. Biophys. 287(2), 268-275.

Giordano, M., Beardall, J., Raven, J.A., 2005. CO_2 Concentrating Mechanisms in Algae: Mechanisms, Environmental Modulation, and Evolution. Annu. Rev. Plant Biol. 56, 99-131.

Ioki, M., Ohkoshi, M., Nakajima, N., Nakahira-Yanaka, Y., Watanabe, M.M., 2011. Isolation of herbicide-resistant mutants of *Botryococcus braunii*. Bioresour. Technol., this issue, in press, http://dx.doi.org/10.1016/j.biortech.2011.07.101.

Los, D.A., Murata, N., 1998. Structure and expression of fatty acid desaturases. Biochim. Biophys. Acta. 1394(1), 3-15.

Metzger, P., Largeau, C., 2005. Botryococcus braunii: a rich source for hydrocarbons and related ether lipids. Appl. Microbiol. Biotechnol. 66(5), 486-496.

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.

Ohlrogge, J.B., Kuhn, D.N., Stumpf, P.K., 1979. Subcellular localization of acyl carrier protein in leaf protoplasts of *Spinacia oleracea*. Proc. Natl. Acad. Sci. U S A 76(3), 1194-1198.

Rismani-Yazdi, H., Haznedaroglu, B.Z., Bibby, K., Peccia, J., 2011. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: Pathway description and gene discovery for production of next-generation biofuels. BMC Genomics 12, 148.

Sukovich, D.J., Seffernick, J.L., Richman, J.E., Gralnick, J.A., Wackett, L.P., 2010. Widespread head-to-head hydrocarbon biosynthesis in bacteria and role of OleA. Appl. Environ. Microbiol. 76(12), 3850-3862.

Vega-Arreguín, J.C., Ibarra-Laclette, E., Jiménez-Moraila, B., Martínez, O., Vielle-Calzada, J.P., Herrera-Estrella, L., Herrera-Estrella, A., 2009. Deep sampling of the *Palomero* maize transcriptome by a high throughput strategy of pyrosequencing. BMC Genomics 10, 299.

Wahlund, T.M., Hadaegh, A.R., Clark, R., Nguyen, B., Fanelli, M., Read, B.A., 2004. Analysis of expressed sequence tags from calcifying cells of marine coccolithophorid (*Emiliania huxleyi*). Mar. Biotechnol. 6, 278-290.

Weber, A.P., Oesterhelt, C., Gross, W., Bräutigam, A., Imboden, L.A., Krassovskaya, I.,

Linka, N., Truchina, J., Schneidereit, J., Voll, H., Voll, L.M., Zimmermann, M., Jamai, A., Riekhof, W.R., Yu, B., Garavito, R.M., Benning, C., 2004. EST-analysis of the thermo-acidophilic red microalga Galdieria sulphuraria reveals potential for lipid A biosynthesis and unveils the pathway of carbon export from rhodoplasts. Plant Mol. Biol. 55(1), 17-32.

Yen, C.L., Stone, S.J., Koliwad, S., Harris, C., Farese, R.V. Jr., 2008. Thematic review series: glycerolipids. DGAT enzymes and triacylglycerol biosynthesis. J. Lipid Res. 49(11), 2283-2301.

Figure Captions

Fig. 1. Features of the BOT-88-2 EST dataset generated by pyrosequencing. a, Distribution of cDNA read lengths. **b**, Length distribution of non-redundant sequences obtained after assembling of cDNA reads. **c**, Distribution of EST counts.

P.S. Please check fig. and tables at

http://dx.doi.org/10.1016/j.biortech.2011.10.033