Algal Research 35 (2018) 449–461

Contents lists available at ScienceDirect

Algal Research

journal homepage: www.elsevier.com/locate/algal

Genomic characterization reveals significant divergence within *Chlorella sorokiniana* (Chlorellales, Trebouxiophyceae)

Blake T. Hovde[a,1], Erik R. Hanschen[a,1], Christina R. Steadman Tyler[a], Chien-Chi Lo[a], Yuliya Kunde[a], Karen Davenport[a], Hajnalka Daligault[a], Joseph Msanne[b,c], Stephanie Canny[d], Seong-il Eyun[g], Jean-Jack M. Riethoven[d,e], Juergen Polle[f], Shawn R. Starkenburg[a,*]

[a] *Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87544, United States of America*
[b] *Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[c] *School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[d] *Center for Biotechnology, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[e] *School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[f] *Brooklyn College, City University of New York, New York, NY, 11210, United States of America*
[g] *Department of Life Science, Chung-Ang University, Seoul 06974, Korea*

ARTICLE INFO

Keywords:
Chlorella sorokiniana
Genome analysis
Meiosis
DNA methylation
Polyketide synthetase (PKS)

ABSTRACT

Selection of highly productive algal strains is crucial for establishing economically viable biomass and bioproduct cultivation systems. Characterization of algal genomes, including understanding strain-specific differences in genome content and architecture is a critical step in this process. Using genomic analyses, we demonstrate significant differences between three strains of *Chlorella sorokiniana* (strain 1228, UTEX 1230, and DOE1412). We found that unique, strain-specific genes comprise a substantial proportion of each genome, and genomic regions with > 80% local nucleotide identity constitute < 15% of each genome among the strains, indicating substantial strain specific evolution. Furthermore, cataloging of meiosis and other sex-related genes in *C. sorokiniana* strains suggests strategic breeding could be utilized to improve biomass and bioproduct yields if a sexual cycle can be characterized. Finally, preliminary investigation of epigenetic machinery suggests the presence of potentially unique transcriptional regulation in each strain. Our data demonstrate that these three *C. sorokiniana* strains represent significantly different genomic content. Based on these findings, we propose individualized assessment of each strain for potential performance in cultivation systems.

1. Introduction

Development and deployment of a productive, stable, and economically viable algal cultivation system requires detailed genetic and phenotypic knowledge of the platform strain(s) [1,2]. This knowledge is gained initially through sequencing and characterization of the genomic content, enabling the formation of testable hypotheses to accelerate algal strain improvement. Identification of both conserved and strain specific pathways will facilitate strain improvement through targeted genetic modification or selective breeding. However, high quality genome assemblies from microalgae production strain candidates are not widely available and many algal genomes are sequenced with short read sequence data, resulting in highly fragmented assemblies, thus impeding accurate gene annotation, transcriptomic analyses, and *in silico* metabolic modeling.

Nearly finished genomes of microalgae production strains inform many other biological functions relevant to bioproduct and biofuel production. Five functions of particular interest include (1) conservation and divergence in energy capture, (2) metabolism and carbon storage, (3) the capacity for sexual reproduction (thereby facilitating artificial selection of desirable traits), (4) the capacity for epigenetic modifications, enabling more nuanced regulation of metabolic and energy storage pathways, and (5) the production of antibiotic compounds, which may assist in crop defense or design of synthetic antibiotics.

First, given the importance of energy capture through photosynthesis in algal growth and production, the genes underlying photosynthesis are predicted to be highly conserved. Protein functionality in core

* Corresponding author at: PO Box 1663, Los Alamos, NM 87544, United States of America.
E-mail address: shawns@lanl.gov (S.R. Starkenburg).
[1] Equal contributors.

https://doi.org/10.1016/j.algal.2018.09.012
Received 3 May 2018; Received in revised form 10 August 2018; Accepted 16 September 2018
Available online 01 October 2018

photosynthetic processes includes carbohydrate metabolism; capturing electron excitation energy; and the anabolism of pigments, lipids, and amino acids. Previous phylogenomic analyses have characterized the set of genes restricted to photosynthetic organisms [3,4]. Loss of these genes may indicate a relative specialist photosynthetic strategy [4] as a result of adaptation to environmental niches. Such specialization may have subsequent consequences on production capabilities.

Second, once energy has been conserved, it may be utilized during metabolism or stored in different ways, which may have profound effects on product value. Understanding, and subsequently targeting, specific biochemical pathways *via* genome editing techniques can also lead to accumulation of additional lipids or other high value products [5].

Third, the ability to reproduce sexually has numerous consequences on both long-term evolution and laboratory selection. Recombination during sexual reproduction results in the purging of deleterious alleles and creation of novel gene combinations [6,7]. Accordingly, sexual reproduction has been used to accelerate artificial or experimental selection throughout eukaryotes [8–15]. The capacity for sexual reproduction in candidate production strains may be exploited through strategic breeding to improve biomass and bioproduct yields as used in traditional food cultivation.

Fourth, while genomic information helps inform functionality of an organism, understanding the factors that regulate the accessibility of the genome is necessary to analyze the phenotypic productivity of a given algae species. These factors collectively constitute chromatin remodeling mechanisms that, when inherited after mitotic activity, are deemed epigenetic in nature. Epigenetic machinery is responsible for posttranslational modification of amino acids in histone proteins or nucleic acids in RNA and DNA. DNA methylation, particularly on cytosine residues, is important for genome protection from opportunistic genetic elements, gene expression, and genomic stability. Two seminal reports of silencing mechanisms employed by microalgae species suggest the existence of DNA modification machinery, including RNA-mediated DNA methylation, and DNA modifications that are present in plants, but not in mammals [16,17]. However, for many algal species, these modifications are uncharacterized and only a handful of genetic signatures of epigenetic machinery have been identified in select species [16,18–22].

Finally, algae produce an array of defensive compounds, including some products of polyketide synthase (PKS) enzymes. PKS enzymes are large, multi-domain genes that encode a variety of naturally-occurring biotoxins and defensive compounds. Polyketide secondary metabolites include antimicrobial, antifungal, insecticide, and immunosuppressive chemicals [23]. Thus, we are interested in understanding polyketide diversity, evolution, and synthesis to determine the value of PKS products as a potential algal bioproduct. Synthesized from acetyl- or malonyl-CoA, polyketides are produced by polyketide synthase genes that possess a constrained set of canonical functional domains, including ketoacyl synthase (KS), acyl transferase (AT), ketoacyl reductase (KR), dehydratase (DH), enoyl reductase (ER), acyl carrier protein (ACP; also known as a phosphopantetheine attachment site), and a thioreductase (TE). PKS genes are categorized into three major structural groups. Type I PKSs are large, modular proteins which are found throughout bacteria, fungi, and algae [24,25]. Each module elongates and modifies the polyketide. Type II PKSs are smaller, aggregative proteins present in green algae and bacteria that iteratively act on polyketide chains [24,26]. Type III PKSs are restricted to streptophyte green algae and bacteria [27,28] and operate as homodimers. Phylogenomic investigations of polyketide synthases in green algae have found three PKS genes in green algae *Ostreococcus lucimarinus* and *Ostreococcus tauri* [25,29]. This relatively high abundance of PKS genes (1.5% of the genome length) suggests important, but unknown function in green algae [25,30].

Chlorella sorokiniana, a freshwater chlorophyte, is being evaluated for utilization as a feedstock for biofuels and bioproducts given its high degree of productivity during short periods of cultivation [1,31,32]. Although a few phenotypic comparisons between multiple *C. sorokiniana* strains have been performed [33–36], the genetic basis for the varied phenotypes remains unknown. Here, we present the genome sequences and gene annotations of three strains of *C. sorokiniana* and results of a comparative analysis of gene content between these strains (DOE1412, 1228, and UTEX 1230). The use of long read technologies and optical mapping generated high-quality, chromosome-level, genome assemblies of *C. sorokiniana*. We report a significant disparity of gene content, with each strain containing a large complement of unique genes and high genomic divergence. Defining the genomic variation among *C. sorokiniana* strains is a necessary step for realizing the potential of *C. sorokiniana* as a commodity feedstock and will inform differences in growth patterns and growth conditions between strains. While genomic differences may underlie differences in growth patterns, the basis for sexual reproduction, PKS defense, and photosynthesis are conserved among *C. sorokiniana* strains. These results highlight the potential to develop and improve *C. sorokiniana* for use in industrial applications through epigenetic modification, sexual reproduction, and bioengineering of different genomic elements.

2. Methods

2.1. Strain information

C. sorokiniana UTEX 1230 (hereafter 1230) is one of the most productive strains identified and is being evaluated for utilization as a biofuel feedstock [31]. *C. sorokiniana* has an optimal growth temperature of 37 °C and is able to grow heterotrophically on a variety of sugars that enhance oil accumulation [37]. Growth in an optimized mixotrophic and heterotrophic bioreactor supplemented with glucose enabled *C. sorokiniana* 1230 to accumulate 30–40% of its cell mass as lipids [38]. Furthermore, while growing in the absence of nitrogen (following pre-growth with ammonia at dry weight production rates equivalent to growth in the presence of ammonia), the energy content of the algae increased by nearly 50% on a dry weight basis (Dr. Sangeeta Negi, unpublished results).

C. sorokiniana strain 1228 (hereafter 1228) was first studied by Phycal, Inc., and was described to be a clonal isolate from a UTEX 1230 cultivation sample. *C. sorokiniana* 1228 has genomic content distinct from the other *C. sorokiniana* strains [37] and serves as a possible reservoir of genomic material for genetic engineering applications.

C. sorokiniana DOE1412 (equivalent to UTEX B 3016; hereafter 1412) was isolated through prospecting efforts by Dr. Juergen Polle's laboratory and demonstrates excellent growth characteristics and lipid accumulation potential [33]. Initial phylogenetic analysis of the rDNA 18S gene identified this strain as *C. sorokiniana*. However, another molecular marker for phylogenetic analysis, the rDNA internal transcribed spacer region 2 (ITS2), revealed that strain 1412 falls into the Chlorellales order, although the family, genus, and species was not resolved. Productivity data on the indoor and outdoor performance of this strain can be found in the final NAABB report [31,32].

2.2. Genome sequencing and assembly

2.2.1. DNA preparation and sequencing

For the short-read assembly of *C. sorokiniana* 1230, genomic DNA was purified following a standard protocol. Briefly, cells were resuspended in SDS-EB buffer (2% SDS, 400 mM NaCl, 50 mM EDTA, 100 mM Tris-HCl [pH 8]) by vortexing and extracted twice with phenol/chloroform/isoamyl alcohol (25:24:1 by volume). Nucleic acids were precipitated by adding two volumes of 100% ethanol. Resuspended nucleic acids were then treated with RNase A for 1 h and genomic DNA precipitated with CTAB (10% *w/v*) in 0.7 M NaCl, to remove polysaccharides. The DNA pellet was finally resuspended in TE buffer.