The Complete Genome Sequence of the Glaucophyte Alga *Cyanophora paradoxa*

Debashish Bhattacharya (PI)[1], Jeffrey L. Boore (Co-PI)[2]

[1]Department of Biology and Roy J. Carver Center for Comparative Genomics, University of Iowa, Iowa City, IA 52242

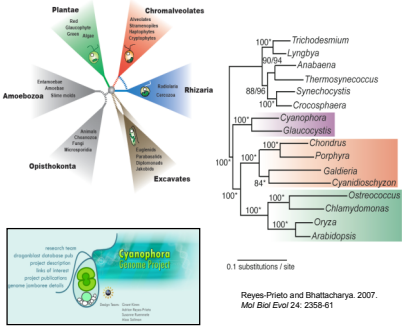
[2]SymBio Corporation, 1455 Adams Drive Menlo Park, CA 94025

Lab of Molecular Evolution

Grant Abstract

How was photosynthesis established in eukaryotes? To gain insights into this fundamental step in the evolution of our planet, we are determining to high coverage the 140 million bp nuclear genome sequence of the unicellular alga *Cyanophora paradoxa*. *Cyanophora* is a member of the remaining group of photosynthetic eukaryotes (Glaucophyta) that still lacks a complete genome sequence. The single cyanobacterial primary endosymbiosis that gave rise to all plastids (e.g., chloroplasts) occurred in the common ancestor of *Cyanophora*, other algae, and plants **(the Plantae, Figure 1).** This was a pivotal and ancient (~1.5 billion years ago) event in the Earth's history that laid the foundation for modern terrestrial ecosystems. A critical step in plastid establishment was the transfer of endosymbiont genes to the "host" nucleus **(Figure 2).** It is unclear however whether this massive transfer was limited to genes strictly involved in plastid metabolism or whether the host profited from the captured genome to explore other novel functions via recruitment of genes from the cyanobacterium. The *Cyanophora* genome sequence will enable us to rigorously test this idea in a relatively "simple" algal model. Beyond its contribution to understanding endosymbiosis, the *Cyanophora* genome sequence will allow countless other insights which include identifying a set of core genes shared by algae and plants that can be studied in detail to understand the origin of plant-specific characters. In addition the *Cyanophora* genome will be invaluable for guiding the annotation of the genomes of plants and other protists.

The earliest diverging photosynthetic eukaryote



Reyes-Prieto and Bhattacharya. 2007. *Mol Biol Evol* 24: 2358-61

***Cyanophora* genome web site**
http://cyanophora.biology.uiowa.edu/

Figure 1

From Free Living Cell to Obligate Endosymbiont

Critical Events in Plastid Origin

Metabolite Exchange Systems
Membrane Transporters

Plastid Protein Import Machinery
Translocation Complexes
Transit Peptides

Endosymbiotic Gene Transfer
Massive Gene Transfer into Host Genome
Acquisition of Nuclear Regulatory Elements
Successful Transcription and Translation
Acquisition of Signal Peptides
Integration into Host Metabolism

Endosymbiont Genome Reduction

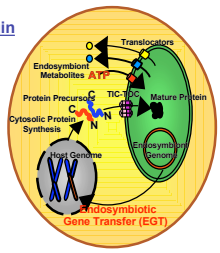
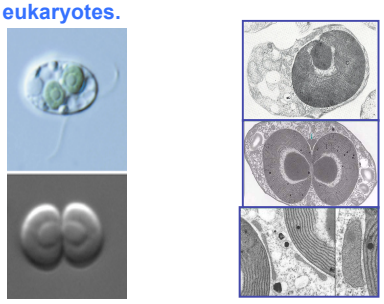
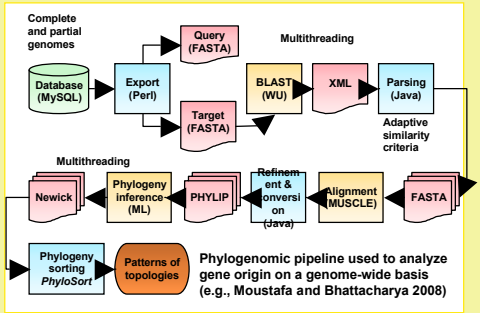


Figure 2

Our Model:

***Cyanophora paradoxa*: a living fossil of the algal world and member of an independent lineage of photosynthetic eukaryotes.**





Phylogenomic pipeline used to analyze gene origin on a genome-wide basis (e.g., Moustafa and Bhattacharya 2008)

Annotation Jamboree
This event will take place at Rutgers University in New Brunswick, New Jersey in Fall 2009. A team of scientists has been assembled to help us annotate and analyze the *Cyanophora* nuclear genome.

Training
Grant funds were used thus far to train graduate students Heather Tyra and Ahmed Moustafa, two undergraduate interns, post-doctoral fellow Adrian Reyes-Prieto, and two web-designers Susanne Ruemmele and Alaa Soliman. All of these individuals have learned about *Cyanophora* and genomics from this project and contributed to project success. In the future, remaining grant funds will be used to generate small RNA sequences from cells grown in
4 different culture conditions using the ABI platform and to do whole genome tiling with the NimbleGen platform. This work will be done by post-docs Jeferson gross and Tovah Salcedo in collaboration with John Manak (University of Iowa).

***Cyanophora* Web Site**
Project web designers have designed a web site for the dissemination of project data and results:
http://cyanophora.biology.uiowa.edu/

Overall Progress
To date, we have produced 701,334 sequencing reads. This has produced about 316 million usable nucleotides of sequence after trimming vector sequence and all but the most high quality nucleotides. We have generated an assembly and BLAST analysis as part of quality control to ensure we are on track to finish with a complete high quality sequence assembly of the *Cyanophora* genome.

Data Analysis
SymBio Corporation has performed a series of assemblies on the sequences using an industry-standard assembly software package (Arachne). A genome size estimate was calculated based on these results as well as by using a cDNA matching strategy; this is in agreement with the predetermined value of 140 million nucleotides. BLAST analyses verify that there was little or no contamination from bacterial or organelle DNA.

Assembly
SymBio has performed multiple assemblies on SymBio's Titan Assembly Server, including the most current assembly (7/1/08), containing 701,334 reads.

Results
The genome coverage is 2.26X based on an estimated genome size of 140,000,000 bases. The largest contig is 50,006 bases in length, whereas the largest supercontig is 89,842 bases in length. The assembly yielded 2,318 contigs (Table 1), of which 448 can be further joined by gap-spanning reads.

Genome Size Estimate
Methods
SymBio Corporation mathematically estimated the genome size as the mean of three methods. Reads were aligned in an all-by-all BLAST search and read length plotted versus number of matches, where it can be shown that: (1) the number of alignments at the mean read length divided by 2 equals the depth of coverage, (2) the Y-intercept of the best fit line is equal to the depth of coverage, and (3) the slope of this line is the depth of coverage divided by the average read length, which can be easily converted into the genome size.

Results
SymBio Corporation estimates the genome size to be 139,287,455 bases. This number is nearly identical to the original estimate of 140,000,000 bases.

BLAST Analysis
The 2,318 contigs (from assembly performed 7/1/08, Figure 3) were queried against the NCBI nucleotide (nt) database using BLASTN. There were a total of 279 strong (e-value < 0.0001) hits to various organisms in the database, including 42 hits to *Chlamydomonas reinhardtii*, 5 hits to *Monosiga brevicollis*, and 28 hits to *Cyanophora paradoxa*. These *Cyanophora* results included hits to 26S rRNA, mRNAs for proteins of the 60S and 40S ribosomal subunits, several hits to *Cyanophora*'s cyanelle, and to several other genes, including ATP citrate lyase and RNA polymerase II.

Table 1: Assembly Statistics Summary

Number of Reads Into Assembly	701,334	Prior to screening
Number of Reads Retained for Assembly	493,752	After vector and quality screening
Percent Reads Retained for Assembly	70	
Mean Trimmed Read Length	641	After vector and quality trimming
Total Usable Nucleotides	316,495,032	
Number of Contigs	2,318	Minimum of two reads per contig
Number nucleotides in contigs	90,201,084	
Percentage of nucleotides in contigs	28.5	
Assembly size	12,244,107	Sum of the length of all contigs
Mean Contig Length	5,282	
Mean Reads/Contig	61	
Largest Contig Size	50,006	
Largest Supercontig Size	89,842	Contigs joined by gap-spanning reads

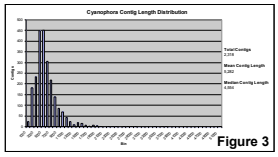
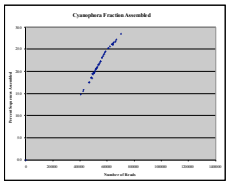


Figure 3

Total Sequencing Reads vs. Reads Assembled



Revised Plan for Project Completion Using Next-Generation Sequencing

	Original Plan	Revised Plan
Capillary Lanes	1,294,452	750,000
Nucleotides from capillary lanes	910,000,000	336,525,000
Nucleotides from 454-type sequencing	0	1,120,000,000
Nucleotides from Solexa-type sequencing	0	1,500,000,000
Total nucleotides determined	910,000,000	2,956,525,000
Expected total coverage	6.5X	ca. 21X

Bhattacharya lab: