Lawrence Berkeley National Laboratory
LBL Publications

Title
The pan-genome of Emiliania huxleyi

Permalink
https://escholarship.org/uc/item/52c5x6qr

Authors
Kuo, Alan
Grigoriev, Igor
Read, Betsy

Publication Date
2010-03-24

The pan-genome of Emiliania huxleyi

Alan Kuo[1]*, Betsy Read[2], and Igor Grigoriev[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA, USA. [2]California State University, San Marcos, CA, USA
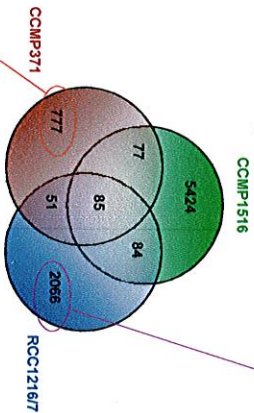*Correspondence: akuo@lbl.gov

I. Abstract

Coccolithophores are a major group of phytoplankton and play a significant role in the carbon cycle, as their calcareous exoskeletons comprise half the carbonate in deep-sea sediments. The bloom species Emiliania huxleyi is globally distributed with multiple strains isolated from around the world. A model for coccolithophore biology, it has been the object of numerous morphological, physiological, and transcriptomic studies, and it is now the first coccolithophore to have its genome sequenced. The reference genome provides an opportunity for us to address an outstanding issue of E. huxleyi biology: its remarkable morphological, physiological, and genomic variation between isolates. Variability suggests that E. huxleyi may have a "pan-genome", where different strains of a single species possess a heterogeneous gene complement. We use the JGI Annotation Pipeline and the JGI EST Pipeline to exploit the reference genome in combination with transcriptomic data from two other strains and discover potentially strain-specific genes. Such genes provide evidence of inter-strain heterogeneity, though they do not fully characterize the E. huxleyi pan-genome.

III. Stategy to use the genomic assembly and ESTs from multiple strains to discover strain-specific genes

Reference strain
CCMP1516

Annotation Pipeline → 7809 genomic scaffolds
Blat
103689 aligned ESTs (78%)
39126 genes (33341 in Reduced Model set)

132554 ESTs
80162 CCMP1516
14006 CCMP371
38386 RCC1216/7

2 other strains
CCMP371 and RCC1216/7

EST Pipeline
28865 nonaligned ESTs (22%)
Annotation Pipeline → 17126 EST consensi
8564 ORFs

Classify consensi (and thus ORFs) based on strain origin of component ESTs

II. Genome heterogeneity across strains, as measured by flow cytometry, and by Solexa resequencing of genomic DNA

E. huxleyi Genome Size Variation
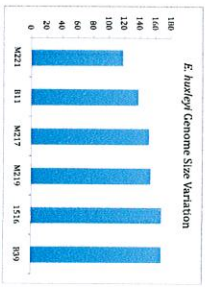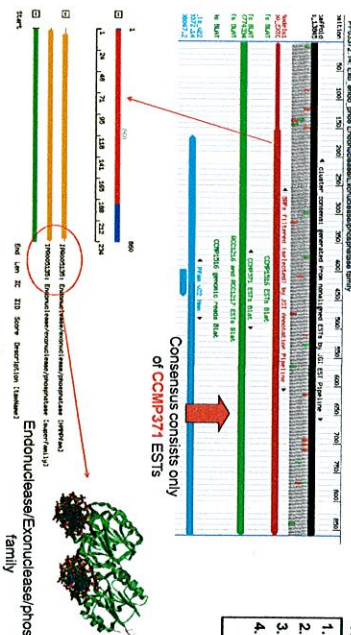
Strain	Raw Reads	Filtered Reads	Alignments	% Coverage	Genes
Strain 1516	12,174,869	12,004,762	9,265,323	77.18	
Strain M219	7,384,973	3,654,804	2,036,000	55.70	28,684
Strain M217	12,011,216	7,321,212	5,435,165	74.23	32,911
Strain M221	12,622,275	6,881,847	1,333,138	19.43	25,784
Strain B39	8,154,341	5,381,208	4,001,719	74.36	32,244
Strain B11	9,819,349	9,477,524	6,020,321	63.52	32,388
Totals				83.60	

IV. ORFs binned by strain

CCMP371 — 777
CCMP1516 — 5424
RCC1216/7 — 2066
77
51
85
84

V. A candidate CMP371-specific hydrolase

Consensus consists only of CCMP371 ESTs

Endonuclease/Exonuclease/phosphatase family

VI. A candidate RCC1216/7-specific transcription factor

Consensus consists only of RCC1216/7 ESTs
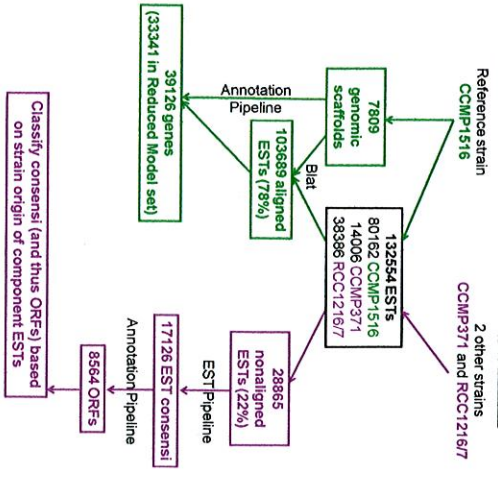
MYND finger
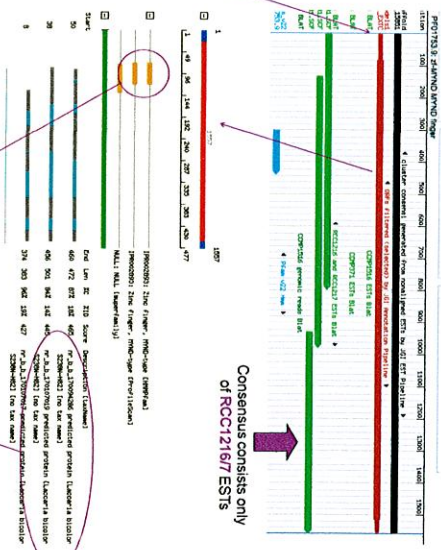
No similar CCMP1516 genes. Most similar proteins are fungal.

VII. Conclusions

1. 22% ESTs do not align with scaffolds.
2. EST Pipeline assembles 17126 consensi from the nonaligned ESTs.
3. Annotation Pipeline predicts 8564 ORFs on the consensi.
4. Tracking ESTs provides evidence for a pan-genome by revealing 2894 potentially strain-specific ORFs.

VIII. Credits

CSUSM
Xiaoyu Zhang
Betsy Read

SB-Roscoff
Peter von Dassow
Hiroyuki Ogata
Colomban de Vargas

JGI EST Pipeline
Jasmyn Pangilinan
Jeff Froula
Erika Lindquist

JGI Annotation Pipeline
Bobby Otillar
Frank Korzeniewski
Asaf Salamov
Andrea Aerts