Postprint of Low-Coverage Draft Genome Analysis of Haematococcus pluvialis

Authors: Chen Jun, Zheng Huajun, Liu Yaming, Zhao Guoping, Qin Song

Date: 2018-05-23T00:00:00+00:00

Abstract

Conducting genome sequencing research on *Haematococcus pluvialis* holds significant importance for elucidating the origin and evolution of green algae, the mechanisms of biological stress response, as well as for promoting the industrial development of *Haematococcus pluvialis*. High-throughput sequencing was performed on *Haematococcus pluvialis* using the Illumina Hiseq 2500 platform, yielding a low-coverage whole-genome draft. Through k-mer distribution analysis, the draft genome size was predicted to be approximately 547 Mbp, with a GC content of 59.2%, indicating a homozygous or haploid state. A total of 11,059 predicted genes were identified, with an average gene length of 1,711 bp and an average CDS length of 681 bp; each gene contained an average of 3.2 exons, with an average exon length of 353 bp. Metabolic pathway analysis revealed the presence of complete fundamental metabolic pathways, including glycolysis, the tricarboxylic acid cycle, the pentose phosphate pathway, and purine and pyrimidine synthesis.

Full Text

Preamble

Title: The Analysis of the Low-Coverage *Haematococcus pluvialis* Draft Genome

Authors: Chen Jun , Zheng Huajun , *Liu Yaming , Zhao Guoping , Qin Song*

Affiliations:

Key Laboratory of Coastal Biology and Bioresource Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China

University of Chinese Academy of Sciences, Beijing 100049, China

Chinese National Human Genome Center at Shanghai, Shanghai 201203, China

Abstract

Genomic sequencing of *Haematococcus pluvialis* holds significant importance for understanding the origin and evolution of green algae, elucidating mechanisms of stress response, and advancing the industrial development of this valuable microalga. In this study, we performed high-throughput sequencing of *H. pluvialis* using the Illumina HiSeq 2500 platform to obtain a low-coverage draft genome. K-mer distribution analysis predicted a genome size of approximately 547 Mbp with a GC content of 59.2%, suggesting a homozygous or haploid genome. A total of 11,059 protein-coding genes were predicted, with an average gene length of 1,711 bp and an average CDS length of 681 bp. Each gene contained on average 3.2 exons, with a mean exon length of 353 bp. Metabolic pathway analysis revealed the presence of complete essential metabolic pathways, including glycolysis, the tricarboxylic acid cycle, the pentose phosphate pathway, and purine and pyrimidine synthesis.

Keywords: *Haematococcus pluvialis*; genome sequencing; gene prediction; functional annotation

Introduction

Haematococcus pluvialis Flotow 1844 is a unicellular eukaryotic microalga belonging to the family Haematococcaceae, order Volvocales, class Chlorophyceae, and phylum Chlorophyta. It is widely distributed in various small water bodies and moist soils. This alga exhibits a complex life cycle comprising motile cells, non-motile cells, zoospores, and aplanospores. Under specific environmental conditions such as high light intensity, high salinity, and nitrogen deficiency, *H. pluvialis* transforms from motile cells into non-motile aplanospores and accumulates astaxanthin, which can account for 1–4% of its dry cell weight. Consequently, *H. pluvialis* is considered the best natural source of astaxanthin, a compound with exceptional antioxidant activity often referred to as "super vitamin E." The market price of natural astaxanthin reaches approximately $10,000/kg—about 200 times that of synthetic astaxanthin—yet production capacity remains far below market demand, with applications spanning food supplements, cosmetics, and aquaculture feed.

Recent advances in molecular biology have substantially improved our understanding of astaxanthin biosynthesis in *H. pluvialis*, with the metabolic pathway now largely elucidated. The emergence of high-throughput sequencing technologies has further enabled numerous transcriptomic, proteomic, and metabolomic studies. For instance, Chen et al. integrated transcriptomic and metabolomic approaches to demonstrate the coordinated regulation of astaxanthin and fatty acid synthesis, revealing that astaxanthin esterification promotes astaxanthin formation and accumulation. Similarly, Gwak et al. employed transcriptomics and lipidomics to investigate astaxanthin and fatty acid metabolism under high-light stress, highlighting the coordinated regulatory mechanisms during cyst formation. Despite these advances, fundamental questions remain: How

do environmental stresses precisely regulate astaxanthin synthesis? What cis-acting elements exist upstream of key rate-limiting enzyme genes, and how do they interact with transcription factors? Is astaxanthin accumulation merely a byproduct of reactive oxygen species scavenging during stress responses, or is it a primary defense strategy? Traditional PCR and transcriptomic methods provide incomplete information, necessitating whole-genome sequencing for a more comprehensive understanding of the genetic architecture. This study utilized second-generation Illumina HiSeq 2500 technology to generate a low-coverage draft genome of *H. pluvialis*, providing foundational data for future high-quality genome assembly.

Materials and Methods

1.1 Experimental Materials and Culture Conditions

The *Haematococcus pluvialis* strain used in this study was obtained from the Culture Collection of Algae and Protozoa (CCAP) in the United Kingdom and is maintained at the Key Laboratory of Coastal Biology and Bioresource Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences. Cells were inoculated into Bold' s Basal Medium (BBM) and cultured statically at 25°C under a light intensity of 40 mol/(m^2 · s) with a 12 h/12 h light/dark photoperiod.

1.2 Illumina Genome Library Construction and Sequencing

Genomic DNA was extracted from logarithmic-phase *H. pluvialis* cells using the Tiangen Plant Genomic DNA Extraction Kit (DP320-02, Beijing). DNA quality was assessed via 1% agarose gel electrophoresis, and concentration/purity (A /A) were measured using a micro-spectrophotometer. High-quality genomic DNA was used to construct paired-end libraries. DNA fragments of approximately 400 bp were generated using a Covaris S2 instrument (Covaris, USA), and libraries were prepared with the TruSeq™ DNA Sample Prep Kit –Set A (Illumina, USA). Fragments of 350–450 bp were gel-purified. Ten nanograms of the constructed library DNA underwent cluster generation using the TruSeq PE Cluster Kit (Illumina, USA) on a cBot system, followed by paired-end sequencing on the Illumina HiSeq™ 2500 platform.

1.3 Data Processing

Raw sequencing data were processed using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to obtain clean reads through the following steps: (1) adapter sequences were removed using fastx_clipper; (2) N bases were trimmed from the 3' to 5' end until the first non-N base; (3) consecutive low-quality bases (quality score q < 5) were removed from the 3' to 5' end, and reads shorter than 50 bp after trimming were discarded along with their paired reads; (4) paired-end reads were matched using a local script.

Genome size was estimated using Jellyfish v3.3.1 for k-mer distribution analysis. The actual sequencing depth (N) was calculated using the formula M = N × (L - K + 1) / L, where M represents the peak of the k-mer curve, L is the read length, and K is the k-mer length. The predicted genome size was then obtained by dividing the total sequence length by the actual sequencing depth N. Genome assembly was performed using Velvet, and the resulting contigs were aligned against 1,002 existing *H. pluvialis* EST sequences from NCBI to assess coverage. Gene prediction was conducted using Augustus, and predicted proteins were functionally annotated via blastp searches against NCBI' s non-redundant protein database (nr), KEGG, and UNIPROT (E-value < 1e-5). KEGG pathway maps were generated, and KOG classification was performed using the CDD database and RPS-BLAST.

Results

2.1 Data Preprocessing and Assembly

Sequencing yielded 38,189,673 paired-end reads (2 × 150 bp), totaling 11.45 Gb. After quality filtering, 37,185,329 high-quality paired reads (10.55 Gb) were retained. Analysis of base composition and quality distributions revealed initial fluctuations in A, C, G, and T frequencies that stabilized across read positions, with A T and C G as expected. The library exhibited uniform base distribution, extremely low N content, and high base quality, confirming suitability for downstream analysis.

2.2 Genome Size Prediction and Heterozygosity Analysis

A k-mer refers to a substring of k nucleotides derived from sequencing reads; a read of length L yields L - k + 1 k-mers. In this study, each 150 bp read was divided into 134 17-mers. The distribution of 17-mer frequencies was plotted with depth on the x-axis and total k-mer count on the y-axis, where the peak M approximates the actual sequencing depth N. Using Jellyfish v3.3.1, we analyzed 10.55 Gb of clean reads and obtained M = 17, corresponding to N = 19. K-mers appearing only once or a few times typically represent sequencing errors. As shown in [Figure 3: see original paper], the leftmost valley corresponds to a depth of 8; removing all 17-mers with frequency < 8 retained 98.5% of the data. The estimated genome size is therefore G = 10,556,751,790 × 0.985 / 19 = 547 Mb. The single peak observed in [Figure 3: see original paper] indicates that *H. pluvialis* is homozygous or haploid.

2.3 Sequence Assembly and Genome Coverage Assessment

Assembly of Illumina data using Velvet produced varying results across different k-mer values (). Since optimal assembly minimizes contig number while maximizing contig length, we selected the k-mer value yielding the highest average contig length. At k-mer = 75, we obtained 56,423 contigs with a total length of 104,818,003 bp, an average contig length of 1,857 bp, and a GC content of 59.2%.

Due to the effective coverage of only 17×, only partial genome assembly was achieved. Alignment against 1,002 existing *H. pluvialis* EST sequences from NCBI showed that 910 sequences (90.82%) matched assembled contigs, with 495 ESTs (49.40%) covered by a single contig at >90% length and 782 ESTs (78.04%) covered at >50% length (). Comparison of predicted genes with NCBI ESTs revealed matches for 412 sequences (41.12%). These results indicate that the draft genome covers approximately 90% of the protein-coding regions of *H. pluvialis*.

Comparison of assembled contigs with algal genome sequences in NCBI showed the highest matches to *H. pluvialis* genes (4,964 contigs, total matched length 9,534,797 bp, alignment length 696,150 bp), followed by *Chlamydomonas reinhardtii* (1,977 contigs, 4,068,934 bp total, 246,989 bp alignment) (). *C. reinhardtii*, a model green alga in the same phylum and order, showed high homology as expected.

2.4 Gene Prediction and Functional Annotation

Augustus gene prediction identified 11,059 genes with an average length of 1,711 bp, average CDS length of 681 bp, 3.2 exons per gene, and mean exon length of 353 bp. Blastp searches against nr, KEGG, and UNIPROT databases functionally annotated 6,890 proteins, with 3,117 proteins assigned KEGG orthologs (62.30% annotation rate). Homology analysis against existing algal protein sequences identified best matches to *Volvox carteri f. nagariensis* (2,148 proteins, 24.06% at E-value = 10^{-3}), *C. reinhardtii*, *Monoraphidium neglectum*, *Coccomyxa subellipsoidea* C-169, *Chlorella variabilis*, *Auxenochlorella protothecoides*, *H. pluvialis*, and *Dunaliella salina* ().

KEGG pathway analysis generated 230 metabolic maps, revealing complete essential pathways including glycolysis, TCA cycle, pentose phosphate pathway, and purine/pyrimidine synthesis. Highly represented pathways included carbohydrate metabolism, amino acid metabolism, energy metabolism, translation, cofactor and vitamin metabolism, protein folding/sorting/degradation, membrane transport, signal transduction, and lipid metabolism ([Figure 2: see original paper]). KOG classification using the CDD database annotated 5,233 proteins across 26 functional categories, predominantly general function prediction, posttranslational modification/protein turnover/chaperones, signal transduction mechanisms, amino acid transport/metabolism, translation/ribosomal structure/biogenesis, and carbohydrate transport/metabolism ([Figure 3: see original paper];).

Discussion

Eukaryotic algae originated through endosymbiotic events, exhibiting remarkable diversity and complex evolutionary histories. According to endosymbiotic theory, primary endosymbiosis of cyanobacteria gave rise to green, red, and glaucophyte algae, while secondary endosymbiosis of green and red algae produced

other microalgae such as cryptophytes. Green algae display morphological diversity and photosynthetic pigment systems similar to higher plants, containing both chlorophyll *a* and *b*. Their evolutionary position—intermediate between higher plants and prokaryotic cyanobacteria—makes them invaluable for understanding plant evolution. Genome sequencing of *H. pluvialis* is therefore crucial for deciphering green algal evolution and eukaryotic algal origins.

Haematococcus pluvialis is widely distributed and highly tolerant, adapting to nitrogen deficiency, high light, low oxygen, and high salinity while synthesizing astaxanthin. These complex stress responses have endowed it with sophisticated signal transduction and specialized secondary metabolic systems, making it an ideal model for studying biological stress responses. Whole-genome sequencing will establish an integrated systems biology framework encompassing genomics, transcriptomics, and metabolomics, enabling deeper investigation into energy storage, stress response mechanisms, and network regulation of complex traits. This will facilitate functional genomics modeling and provide essential guidance for metabolic engineering and synthetic biology approaches to construct superior production strains.

In summary, completing the whole-genome sequencing of *H. pluvialis* is urgently needed both for evolutionary studies and industrial development. This study provides a low-coverage draft genome and preliminary characterization of genomic features, laying groundwork for a high-quality reference genome. However, compared with recent high-quality green algal genomes, gaps remain. For example, Roth et al. (2017) combined third-generation PacBio and second-generation sequencing to achieve chromosome-level assembly of *Chromochloris zofingiensis*, identifying over 15,000 genes and elucidating astaxanthin accumulation mechanisms under varying light intensities through transcriptomics and carotenoid profiling. This comprehensive approach serves as a valuable reference for future *H. pluvialis* genome projects.

Several considerations are essential for future whole-genome sequencing efforts: (1) Establishing axenic culture systems. *H. pluvialis* often harbors symbiotic bacteria and is prone to contamination when cultured in organic media like EG:JM. Sterile cultivation is critical to avoid assembly artifacts. Zheng et al. (2017) demonstrated successful axenization of *H. pluvialis* strain FACHB-712 using sequential treatment with penicillin, gentamicin, and kanamycin. In this study, we ensured sample purity through repeated purification, microscopic examination, and 16S rRNA sequencing. (2) Integrating second- and third-generation sequencing technologies. Given the predicted genome size of ~500 Mb and likely multiple linear chromosomes, we recommend employing the PacBio SMRT system, which generates reads up to 10 kb for superior assembly, supplemented with Illumina HiSeq 2500 data to correct single-base indel errors and enhance accuracy.

References

[1] Hu H J, Wei Y X. The freshwater algae of China- Systematics, Taxonomy and Ecology[M]. Beijing: Science Press, 2006 (in Chinese).
[2] Liu J G, Yin M Y, Zhang J P, et al. Statues of cell cycle in *Haematococcus pluvialis*[J]. Oceanology et Limnologia Sinica, 2000, (02): 145-150 (in Chinese).
[3] Kobayashi M, Kurimura Y, Kakizono T, et al. Morphological changes in the life cycle of the green alga *Haematococcus pluvialis*[J]. Journal of Fermentation and Bioengineering, 1997, 84(1): 94-97.
[4] Lorenz R T, Cysewski G R. Commercial potential for *Haematococcus* microalgae as a natural source of staxanthin[J]. Trends in Biotechnology, 2000, 18(4):160-167.
[5] Borowitzka M A, High-value products from microalgae-their development and commercialisation[J]. Journal of Applied Phycology, 2013, 25(3): 743-756.
[6] Chen J, Wang Y, Benemann J R, et al. Microalgal industry in China: challenges and prospects[J]. Journal of Applied Phycology, 2016, 28(2): 715-725.
[7] Cui H L, Yu X N, Wang Y, et al. Evolutionary origins, molecular cloning and expression of carotenoid hydroxylases in eukaryotic photosynthetic algae[J]. BMC Genomics, 2013, 14: 457.
[8] Han D X, Li Y T, Hu Q. Astaxanthin in microalgae: pathways, functions and biotechnological implications[J]. Algae, 2013, 28(2): 131-147.
[9] Lu Y D, Jiang P, Liu S F, et al. Methyl jasmonate- or gibberellins A(3)-induced astaxanthin accumulation is associated with up-regulation of transcription of beta-carotene ketolase genes (bkts) in microalga *Haematococcus pluvialis*[J]. Bioresource Technology, 2010, 101(16): 6468-6474.
[10] Gao Z Q, Li Y, Wu G X, et al. Transcriptome analysis in *Haematococcus pluvialis*: Astaxanthin induction by salicylic acid (SA) and jasmonic acid (JA)[J]. Plos One, 2015, 10(10): e0140609.
[11] Li K, Cheng J, Lu H X, et al. Transcriptome-based analysis on carbon metabolism of *Haematococcus pluvialis* mutant under 15% CO2[J]. Bioresource Technology, 2017, 233: 313-321.
[12] Cheng J, Li K, Zhu Y X, et al. Transcriptome sequencing and metabolic pathways of astaxanthin accumulated in *Haematococcus pluvialis* mutant under 15% CO2[J]. Bioresource Technology, 2017, 228: 99-105.
[13] Chen G Q, Wang B B, Han D X, et al. Molecular mechanisms of the coordination between astaxanthin and fatty acid biosynthesis in *Haematococcus pluvialis* (Chlorophyceae)[J]. Plant Journal, 2015, 81(1): 95-107.
[14] Gwak Y, Hwang Y S, Wang B B, et al. Comparative analyses of lipidomes and transcriptomes reveal a concerted action of multiple defensive systems against photooxidative stress in *Haematococcus pluvialis*[J]. Journal of Experimental Botany, 2014, 65(15): 4317-4334.
[15] Wang S B, Chen F, Sommerfeld M, et al. Proteomic analysis of molecular response to oxidative stress by the green alga *Haematococcus pluvialis* (Chlorophyceae)[J]. Planta, 2004, 220(1): 17-29.
[16] Kim J D, Lee W S, Kim B, et al. Proteomic analysis of protein expression patterns associated with astaxanthin accumulation by green alga *Haematococ-*

cus pluvialis (Chlorophyceae) under high light stress[J]. Journal of Microbiology and Biotechnology, 2006, 16(8): 1222-1228.
[17] Gao Z Q, Miao X X, Zhang X W, et al. Comparative fatty acid transcriptomic test and iTRAQ-based proteomic analysis in *Haematococcus pluvialis* upon salicylic acid (SA) and jasmonic acid (JA) inductions[J]. Algal Research-Biomass Biofuels and Bioproducts, 2016, 17: 277-284.
[18] Su Y X, Wang J X, Shi M L, et al. Metabolomic and network analysis of astaxanthin-producing *Haematococcus pluvialis* under various stress conditions[J]. Bioresource Technology, 2014, 170: 522-529.
[19] Lv H X, Xia F, Liu M, et al. Metabolomic profiling of the astaxanthin accumulation process induced by high light in *Haematococcus pluvialis*[J]. Algal Research-Biomass Biofuels and Bioproducts, 2016, 20: 35-43.
[20] Boussiba, S. Carotenogenesis in the green alga *Haematococcus pluvialis*: Cellular physiology and stress response[J]. Physiologia Plantarum, 2000, 108(2): 111-117.
[21] Zerbino D R, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs[J]. Genome Research, 2008, 8(5): 821-829.
[22] Stanke M, Schoffmann O, Morgenstern B, et al. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources[J]. BMC Bioinformatics, 2006, 7: 62.
[23] Marchler-Bauer A, Bo Y, Han L, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures[J]. Nucleic Acids Research, 2017, 45(D1): D200-D203.
[24] Roth M S, Cokus S J, Gallaher S D, et al. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, E4296–E4305.
[25] Zheng L L, Zhang Q, Li T L, et al. Axenation of *Haematococcus pluvialis* and the effects of axenic cultivation on the growth and physiology of the strain[J]. Journal of Fujian Normal University (Natural Science Edition), 2017, 33(1): 44-50 (in Chinese).

Note: Figure translations are in progress. See original paper for figures.