Peer Community In Genomics

Reference genome for the lichen-forming green alga *Coccomyxa viridis* SAG 216–4

Iker Irisarri *based on peer reviews by* ***Fabian Haas***, ***Elisa Goldbecker*** *and 2 anonymous reviewers*

Anton Kraege, Edgar Chavarro-Carrero, Nadège Guiglielmoni, Eva Schnell, Joseph Kirangwa, Stefanie Heilmann-Heimbach, Kerstin Becker, Karl Köhrer, Philipp Schiffer, Bart P. H. J. Thomma, Hanna Rovenich (2024) High quality genome assembly and annotation (v1) of the eukaryotic terrestrial microalga *Coccomyxa viridis* SAG 216-4. bioRxiv, ver. 2, peer-reviewed and recommended by Peer Community in Genomics.
https://doi.org/10.1101/2023.07.11.548521

Submitted: 08 April 2024, Recommended: 13 July 2024

Cite this recommendation as:
Irisarri, I. (2024) Reference genome for the lichen-forming green alga *Coccomyxa viridis* SAG 216–4. *Peer Community in Genomics*, 100300. 10.24072/pci.genomics.100300

Published: 13 July 2024

Green algae of the genus *Coccomyxa* (family Trebouxiophyceae) are extremely diverse in their morphology, habitat (i.e., in marine, freshwater, and terrestrial environments) and lifestyle, including free-living and mutualistic forms. *Coccomyxa viridis* (strain SAG 216–4) is a photobiont in the lichen *Peltigera aphthosa*, which was isolated in Switzerland more than 70 years ago (cf. SAG, the Culture Collection of Algae at the University of Göttingen, Germany). Despite the high diversity and plasticity in *Coccomyxa*, integrative taxonomic analyses led Darienko et al. (2015) to propose clear species boundaries. These authors also showed that symbiotic strains that form lichens evolved multiple times independently in *Coccomyxa*.

Using state-of-the-art sequencing data and bioinformatic methods, including Pac-Bio HiFi and ONT long reads, as well as Hi-C chromatin conformation information, Kraege et al. (2024) generated a high-quality genome assembly for the *Coccomyxa viridis* strain SAG 216–4. They reconstructed 19 complete nuclear chromosomes, flanked by telomeric regions, totaling 50.9 Mb, plus the plastid and mitochondrial genomes. The performed quality controls leave no doubt of the high quality of the genome assemblies and structural annotations. An interesting observation is the lack of conserved synteny with the close relative *Coccomyxa subellipsoidea*, but further comparative studies with additional *Coccomyxa* strains will be required to grasp the genomic evolution in this genus of green algae. This project is framed within the ERGA pilot project, which aims to establish a pan-European genomics infrastructure and contribute to cataloging genomic biodiversity and producing resources that can inform conservation strategies (Formenti et al. 2022). This complete reference genome represents an important step towards this goal, in addition to contributing to future genomic analyses of *Coccomyxa* more generally.

References:

Darienko T, Gustavs L, Eggert A, Wolf W, Pröschold T (2015) Evaluating the species boundaries of green microalgae (*Coccomyxa*, Trebouxiophyceae, Chlorophyta) using integrative taxonomy and DNA barcoding with further implications for the species identification in environmental samples. PLOS ONE, 10, e0127838. https://doi.org/10.1371/journal.pone.0127838

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, Malukiewicz J, Mouton A, Oomen RA, Paez S, Palsbøll PJ, Pampoulie C, Ruiz-López MJ, Svardal H, Theofanopoulou C, de Vries J, Waldvogel A-M, Zhang G, Mazzoni CJ, Jarvis ED, Bálint M, European Reference Genome Atlas Consortium (2022) The era of reference genomes in conservation genomics. Trends in Ecology & Evolution, 37, 197–202. https://doi.org/10.1016/j.tree.2021.11.008

Kraege A, Chavarro-Carrero EA, Guiglielmoni N, Schnell E, Kirangwa J, Heilmann-Heimbach S, Becker K, Köhrer K, WGGC Team, DeRGA Community, Schiffer P, Thomma BPHJ, Rovenich H (2024) High quality genome assembly and annotation (v1) of the eukaryotic terrestrial microalga *Coccomyxa viridis* SAG 216-4. bioRxiv, ver. 2 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2023.07.11.548521

Reviews

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/2023.07.11.548521
Version of the preprint: 1

Authors' reply, 01 July 2024

We have also uploaded an updated version of the manuscript to bioRxiv.
Download author's reply
Download tracked changes file

Decision by Iker Irisarri, posted 22 May 2024, validated 22 May 2024

Minor revision

Dear authors,

Thank you very much for submitting your study to PCI Genomics. Your study has been seen by four Reviewers, who provided thorough comments that I believe could help further improve the manuscript. Given that this is a genome note, I do not think it needs to evolve into a comparative genomics paper, but I would appreciate a bit more context regarding the interest of sequencing this genome and the availability of further Coccomyxa genomes in NCBI/ENA.

In addition to the Reviewers' comments, I could add the following minor points:

L41 Prasinodermophyta has been proposed as a third major lineage of Chloroplatida besides chlorophytes & streptophytes: https://www.nature.com/articles/s41559-020-1221-7

Fig. 2 I assume the two dots with lower GC% correspond to the mitochondrial and plastid genomes, as suggested in the caption. But could you indicate which is which?

L83 abbreviation for hour is h

L94 quantity and quality?

Reviewed by anonymous reviewer 1, 30 April 2024

The authors presented a high-quality assembly genome of microalga Coccomyxa viridis, and did the annotation. This manuscript provides useful resources of microalgae. I have some questions on the manuscript.

1. To evaluate the completeness of genome assembly, do the authors perform the genome size estimation of the microalgal based on experimental and computational method?

2. The authors showed that the assembly is chromosome-scale level, I wonder if the authors have any data on the chromosome number of this agal.

3. Line 202-205, the authors gave conclusion that scaffold 20 and 21 are chloroplast and mitochondrial genomes, these just only based on the length and GC content, I think it may be not correct, same as the conclusion in Figure 1a legend. Did the author map the scaffolds to reference plastome and mitogenome?

4. I want to ask if the authors have examined the scaffold 1-19 containing any plastome or mitogenome fragments?

Title and abstract

Does the title clearly reflect the content of the article? [ ] Yes,

Does the abstract present the main findings of the study? [ ] Yes,

Introduction

Are the research questions/hypotheses/predictions clearly presented? [ ] Yes,

Does the introduction build on relevant research in the field? [ ] Yes,

Materials and methods

Are the methods and analyses sufficiently detailed to allow replication by other researchers? [ ] Yes,

Are the methods and statistical analyses appropriate and well described? [ ] Yes,

Results

In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [ ] Yes,

Are the results described and interpreted correctly? [ ] Yes

Discussion

Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [ ] Yes,

Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [ ] Yes,

Reviewed by Elisa Goldbecker, 02 May 2024

Kraege et al. provide the first genome of the chlorophyte and lichen photobiont Coccomyxa viridis (SAG 216-4). They generated a high-quality assembly using long-reads by PacBio-HiFi and Oxford Nanopore, that were scaffolded using Hi-C. The assembly was further annotated using RepeatMasker and Braker software. The paper outline is very clear and concise. I will not comment on assembly methods, as this falls outside of my expertise. However, I have some small remarks regarding general things and the annotation:

Introduction:

Terms such as “early diverging” (line 43) should be avoided as they can lead to false tree thinking. (McDaniel, 2021), https://doi.org/10.1111/nph.17241

Methods:

RNAseq

It is not mentioned how many RNAseq samples were generated.

Annotation

It is stated that BRAKER was run using transcriptome evidence only, however BRAKER2 is cited, which describes the implementation of BRAKER using protein data. The citation should be changed to BRAKER1 e.g. Hoff et al. 2016 https://doi.org/10.1093/bioinformatics/btv661

Results:

The claim that the average level of alternative splicing is predicted to be very low is in my opinion too speculative, as apparently only RNAseq data from one condition was used and also the number of RNAseq samples is unknown.

Data availability:

Data should be made available upon publishing.

Reviewed by Fabian Haas, 09 May 2024

Title and abstract

Does the title clearly reflect the content of the article? **Yes**

Does the abstract present the main findings of the study? **Yes**

Introduction

Are the research questions/hypotheses/predictions clearly presented? **No** – The history and differentiation of Coccomyxa was shown. And the question of the molecular mechanisms that determine the various symbiotic lifestyles was asked. I'm missing a clear statement how this new genome assembly will help answering this question.

Does the introduction build on relevant research in the field? **Yes**

Materials and methods

Are the methods and analyses sufficiently detailed to allow replication by other researchers? **Yes**

Are the methods and statistical analyses appropriate and well described? **Yes**

Results

In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? No negative results

Are the results described and interpreted correctly? **Yes**

Discussion

Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? **No** – The results are representing not everything the data could show. Some analyses are missing.

Are the conclusions adequately supported by the results (without overstating the implications of the findings)? **Yes**

Review Kraege et al.

In this manuscript, 'High quality genome assembly and annotation (v1) of the eukaryotic terrestrial microalga Coccomyxa viridis SAG 216-4' posted July 12, 2023 at bioRxiv, the authors present the first fully assembled genome of the eukaryotic terrestrial microalga Coccomyxa viridis SAG 216-4. Besides the genome assembly the authors performed repeat masking, gene annotation, contamination analysis, synteny detection, and a ploidy test.

The manuscript presents the resource of the genome and is kept technical. I'm missing the biological meaning and some more analyses. At the introduction the authors are asking the question of the molecular mechanisms that determine the various symbiotic lifestyles. The manuscript does not show the approach to answer

this question. E.g. the article published by Tagirdzhanova et al., 2023 (Sci Rep), uses, among other things, the genome assembly by Kraege et al. and shows some more biological context. Is there any gene loss or gene transfer at Coccomyxa viridis compared to free living Coccomyxa species?

Suggestions of additional analyses for this paper with the existing dataset:

Hi-C: The telomere boundaries were mentioned. What about centromeres? Are there TADs or other structural elements or A/B compartments? Is the Hi-C resolution high enough to say anything about the 3D structure?

Nanopore (ONT): The ONT data can be used to detect methylation (e.g. 6mA or 5mC). https://github.com/nanoporetech/dorado

RNA-seq: Are there alternative splicing sides, start codons, rDNA arrays?

Assembly: Does the assembly contain endogenous viral element(s)? Are there any interesting TE structures like the Chlorella zepp retro TE at the centromere? Are there sub-telomere structures or TEs at the telomeres?

A few minor points:

Line 28: nineteen => 19

Line 81: 3x vitamins => which?

Line 93/94: DNA quality and quality => quantity

Line 109: Why was the Rapid Sequencing Kit used?

Line 111: Flow Cell 9.4.1 => which device?

Line 146: Why manually at the first place? Who many gaps were left after Hi-C? Usually, ARCS (doi:10.1093/bioinformatics/btx6 or TGS-gapcloser (doi:10.1093/gigascience/giaa094) are performing well.

Line 175: Were protein files of other green algae included at the braker run or only the RNA-seq bam files?

Download the review