DATA NOTE
Draft genome assembly and transcriptome sequencing of the golden algae *Hydrurus foetidus* (Chrysophyceae) [version 1; peer review: 1 approved with reservations]

Jon Bråte¹, Janina Fuss², Kjetill S. Jakobsen³, Dag Klaveness⁴

¹Section for Genetics and Evolutionary Biology (EVOGENE), Department of Biosciences, University of Oslo, Oslo, 0316, Norway
²Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, 24118, Germany
³Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, 0316, Norway
⁴Section for Aquatic Biology and Toxicology (AQUA), Department of Biosciences, University of Oslo, Oslo, 0316, Norway

Abstract
*Hydrurus foetidus* is a freshwater alga belonging to the phylum Heterokonta. It thrives in cold rivers in polar and high alpine regions. It has several morphological traits reminiscent of single-celled eukaryotes, but can also form macroscopic thalli. Despite its ability to produce polyunsaturated fatty acids, its life under cold conditions and its variable morphology, very little is known about its genome and transcriptome. Here, we present an extensive set of next-generation sequencing data, including genomic short reads from Illumina sequencing and long reads from Nanopore sequencing, as well as full length cDNAs from PacBio IsoSeq sequencing and a small RNA dataset (smaller than 200 bp) sequenced with Illumina. We combined this data with, to our knowledge, the first draft genome assembly of a chrysophyte algae. The assembly consists of 5069 contigs to a total assembly size of 171 Mb and a 77% BUSCO completeness. The new data generated here may contribute to a better understanding of the evolution and ecological roles of chrysophyte algae, as well as to resolve the branching patterns within the Heterokonta.

Keywords
*Hydrurus foetidus*, Chrysophyceae, golden algae, genome, transcriptome, Nanopore, PacBio

This article is included in the Draft Genomes collection.
Corresponding author: Jon Bråte (jon.brate@ibv.uio.no)

Author roles: Bråte J: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; Fuss J: Data Curation, Formal Analysis, Investigation, Methodology, Writing – Review & Editing; Jakobsen KS: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; Klaveness D: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: JB was funded by the Research Council of Norway (grant numbers: 213703 and 240284).

Copyright: © 2019 Bråte J et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Bråte J, Fuss J, Jakobsen KS and Klaveness D. Draft genome assembly and transcriptome sequencing of the golden algae Hydrurus foetidus (Chrysophyceae) [version 1; peer review: 1 approved with reservations] F1000Research 2019, 8:401 (https://doi.org/10.12688/f1000research.16734.1)

First published: 08 Apr 2019, 8:401 (https://doi.org/10.12688/f1000research.16734.1)
Introduction

Here, we present extensive genome sequencing data, including a multiapproach assembly, as well as transcriptome data of mRNAs and small RNAs of the golden algae *Hydrurus foetidus* (Villars) Trevisan.

There has been considerable interest in the golden algae for many reasons: they are ecologically diverse, important as primary producers (phototrophs) in oligotrophic to dystrophic lakes (Kristiansen, 2005; Nicholls & Wujek, 2015), some are also mixotrophs, phagotrophs or osmotrophs (Kristiansen & Preisig, 2001; Pringsheim, 1963). The chrysophytes span a large range of feeding and nutrient uptake modes (Kristiansen, 2005) and therefore play a significant role in aquatic food webs. Chrysophytes also make up a significant fraction of sequence reads and novel operational taxonomic units in clone libraries from freshwater environmental samples (del Campo & Massana, 2011).

However, chrysophytes have also attracted considerable interest from an evolutionary point of view. They belong to the division (phylum) Heterokonta (Cavalier-Smith, 1986), an immensely diverse group of eukaryotes with many basal branches in the phylogeny still not resolved, despite numerous molecular phylogenetic studies, including multigene phylogenomics (e.g. Burki, 2014; Grossmann et al., 2016; Risberg et al., 2009; Scoble & Cavalier-Smith, 2014). One reason for this is the presence of cryptic species and many groups with extremely similar morphology (Grossmann et al., 2016). Another reason is the complex evolutionary history of the Heterokonta, including an elaborate plastid evolution (e.g. Kim et al., 2019) and heterotrophic lineages which have lost the plastids altogether (Graupner et al., 2018; Pringsheim, 1963). The lack of genomic or transcriptomic data from many taxa, and even whole orders, which limits the power of multigene phylogenies (Beisser et al., 2017), is yet another motivation for genomic and transcriptomic studies. However, recently, there has been a significant addition of transcriptomic data for chrysophyte taxa (e.g. Beisser et al., 2017; Graupner et al., 2018; Keeling et al., 2014; Kraus et al., 2019; Lie et al., 2017).

*Hydrurus foetidus* is not a typical representative of the golden algae. It is macroscopic and benthic (e.g. Klaveness et al., 2011; Rostafinski, 1882: Tab II, Szklarczyk, 1953), whereas most chrysophytes are microscopic single cells or colonial plankton (Sandgren, 1988; Kristiansen, 2005). Furthermore, *Hydrurus* is native to polar, peri-glacial and alpine rivers in Norway and similar regions around the world (e.g. Klaveness, 2019; Rott et al., 2006; Rott & Schneider, 2014) and can only live in cold waters (2–10 °C) (Bursa, 1934; Kann, 1978). Members of the *Hydrurus* clade may cause colored snow and ice, and may be found on permanent ice sheets (Klaveness et al., 2011; Lutz et al., 2018; Remias et al., 2013).

*Hydrurus* has a number of peculiar morphological characteristics relevant for understanding chrysophyte and heterokont evolution. Although it is multicellular, the cells in the thalli are not physically connected, and under some growth conditions the cells may slide away from each other in their wall-less polysaccharide tubes, or be released as single-celled swarmers (Klaveness et al., 2011). Other characteristic features, which may be considered primitive for a thallose alga, are contractive vacuoles, often more than one in each cell (Fott, 1959; Klaveness, 2019).

We have assembled a draft genome of *Hydrurus foetidus* using a combination of short-read Illumina sequencing and long-read Nanopore sequencing. The assembly consists of 5069 contigs yielding a total size of 171 Mb and a 77% BUSCO completeness. In addition to the deep genomic sequencing, we have also sequenced full-length poly(A) transcripts using PacBio IsoSeq, as well as sequencing the expressed small RNAs. This extensive dataset will be important, not only for studies of heterokont and chrysophyte evolution but also for elucidating the genetic mechanisms behind cold water adaptation, like the production of polyunsaturated fatty acids (Klaveness, 2017) and the regulation of a complex multicellular lifestyle.

Materials and methods

Culturing of *H. foetidus*

The specimen of *Hydrurus foetidus* (Villars) Trevisan (strain G070301) used in this study was sampled from the river at the Finse Alpine Research Center (60°36’ N, 07°30’ E) in March 2007 and is currently kept in culture at University of Oslo. *H. foetidus* was isolated in an adapted Guillard & Lorenzen’s WC (Wright’s Chu) medium (Guillard & Lorenzen, 1972) as described by Klaveness & Lindstrøm (2011). To prepare for DNA isolation, the growth of large thalli was promoted by repeated transfer of individual thalli into fresh culture media. Large thalli (0.5–1.0 g wet frozen weight) were collected by removal from the culture medium and immediate transfer to -80°C and storage until further processing. The culture will be deposited in a special culture collection, at the Fraunhofer Culture Collection of Cryophilic Algae (CCCryo).

Illumina sequencing of genomic DNA

Six individual thalli were used for the DNA isolation. DNA isolation was performed using the DNeasy Plant Mini Kit from Qiagen (Qiagen Inc., Valencia, CA, US). To ensure efficient lysis and homogenization of the external polysaccharide sheath, a few titanium beads were added to the frozen samples and the tubes were shaken using a thermomixer at 65°C and 800 rpm with 20 second intervals for 30 mins. After adding Buffer AP2 to the lysate, the incubation was done on ice for 15 minutes to allow for better precipitation of the polysaccharides. Further, the extraction kit protocol was followed as is, until the second elution step. Here we reused the flow through elution buffer and the RNase, tubes were placed in a thermomixer at 65°C and 800 rpm with 20 second intervals for 30 mins. After adding Buffer AP2 to the lysate, the incubation was done on ice for 15 minutes to allow for better precipitation of the polysaccharides. Further, the extraction kit protocol was followed as is, until the second elution step. Here we reused the flow through from the previous step elution to avoid excessive dilution of the samples. Afterwards, the samples were de-salted and concentrated by ethanol precipitation and resuspension in 100 μl of MilliQ water. Finally, the samples were concentrated even further by pooling all the samples and freeze drying with a Leybold Heraeus Lyovac GT2 (Leybold-Heraeus, Köln, Germany).
The isolated and freeze-dried genomic DNA was sent to the Norwegian Sequencing Center (NSC) at the University of Oslo for library preparation and sequencing. The library was made with 600-700 bp fragment size and sequenced on two lanes of Illumina HiSeq 2500 with 250 bp paired-end reads (Table 1).

### Nanopore sequencing of genomic DNA

Genomic DNA was isolated from two thalli as described above, except that tissue lysis was done using MagNA Lyser Green Beads (Roche, Penzberg, Germany) and shaken for 15 sec at 4 m/sec and the incubation at 65°C was done for 10 min. In addition, the supernatant (after adding buffer AP2) was run through QiaShredder columns to further homogenize the lysate. The DNA was eluted (twice, but re-using the elution buffer) in 50 μl AE elution buffer. To further clean and concentrate the samples, the samples were pooled and cleaned using the Zymo DNA Clean & Concentrate kit (Zymo Research, CA, US). The sample was double-eluted (as before) in 50 μl kit provided elution buffer.

DNA sequencing was done using the MinION (MIN-101B) sequencer, the R9.5 Flow Cell and following the SQK-LSK108 protocol (version GDE_9002_v108_revT_18Oct2016) (Oxford Nanopore, Oxford, UK). Approximately 1 μg of starting DNA was used and inspection of the DNA on a 0.7% agarose gel run at 30 volts from 18 hours showed that the majority of the DNA was between 20-30 kbp, but with a long tail of shorter fragments. The sequencing was run using the MinKNOW software (Oxford Nanopore, Oxford, UK; downloaded October 2017) on an iMac and stopped after 36 hours. Base-calling of the raw Nanopore sequence data was done using Albacore v.2.1.10 (Linux, Python 3.5 version) with default settings. The process was run on the Abel computing cluster at the University of Oslo (Table 1).

### PacBio transcriptome sequencing

Total RNA was isolated from one frozen thallus using Qiagen RNaseq Plant kit, including a QiaShredder column and lysis using MagNA lyser beads as described above, otherwise following the kit protocol. Isolated RNA was sent to NSC for library preparation and PacBio sequencing. Three size fractions (1-2 kbp, 2-3 kbp and 3-5 kbp) were prepared using the IsoSeq library preparation protocol and sequenced on RSII SMRT cells (Pacific Biosciences, CA, US) (Table 1 and Table 2).

### Illumina small RNA sequencing

Small RNAs (below 200 bp) were isolated from a frozen thallus using the Sigma mirPremier kit (Sigma-Aldrich, MO, US) following the manufacturer’s instructions, but including lysis with MagNA beads as described above. The sample was sent to the NSC for library preparation and sequencing. Sequencing library (up to approx. 40 nt fragment size) was prepared and sequenced with Illumina NextSeq 500 as single-end 75 bp reads (Table 1).

### Draft genome assembly

The basecalled Nanopore reads were processed with Porechop v0.2.2 using default parameters to remove sequencing adapters. Next, the reads were filtered with Nanofilt v2.0.0 (De Coster et al., 2018) to remove reads shorter than 500 bp and average quality below 9. The filtered reads were further error-corrected with LoRDEC v0.7 (Salmela & Rivals, 2014) using the Illumina reads. First the Illumina reads were quality assessed by removing sequencing adapters and bases with an average quality below 20 (average score across 4 bases), in addition to leading and trailing bases with a quality below 20. This was done using Trimmomatic v0.36 (Bolger et al., 2014). Then lordec-correct (options -k 21 -s 3) was run with the trimmed Illumina reads.

---

**Table 1. Overview of datasets produced in this study.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hfoetidus_ACAGTG_L001_R1_001.fastq.gz</td>
<td>Genomic DNA sequenced with Illumina HiSeq 2500.</td>
<td>ERR2882522</td>
</tr>
<tr>
<td>Hfoetidus_ACAGTG_L001_R1_001.fastq.gz</td>
<td>Genomic DNA sequenced with Illumina HiSeq 2500.</td>
<td>ERR3188711</td>
</tr>
<tr>
<td>Hfoetidus_ACAGTG_L002_R1_001.fastq.gz</td>
<td>Genomic DNA sequenced with Illumina HiSeq 2500.</td>
<td>ERR2887871</td>
</tr>
<tr>
<td>Hydrurus_nanopore_fastaq_files.tar.gz</td>
<td>Basecalled Oxford Nanopore reads</td>
<td>ERR2887871</td>
</tr>
<tr>
<td>IsoSeq_1-2kb_polished_low_qv_consensus_isoforms.fastq.gz</td>
<td>mRNA sequenced with PacBio SMRT RSII.</td>
<td>ERR2882521</td>
</tr>
<tr>
<td>IsoSeq_1-2kb_polished_high_qv_consensus_isoforms.fastq.gz</td>
<td>mRNA sequenced with PacBio SMRT RSII.</td>
<td>ERR2869477</td>
</tr>
<tr>
<td>IsoSeq_2-3kb_polished_low_qv_consensus_isoforms.fastq.gz</td>
<td>mRNA sequenced with PacBio SMRT RSII.</td>
<td>ERR2869481</td>
</tr>
<tr>
<td>IsoSeq_2-3kb_polished_high_qv_consensus_isoforms.fastq.gz</td>
<td>mRNA sequenced with PacBio SMRT RSII.</td>
<td>ERR2869478</td>
</tr>
<tr>
<td>IsoSeq_3-6kb_polished_low_qv_consensus_isoforms.fastq.gz</td>
<td>mRNA sequenced with PacBio SMRT RSII.</td>
<td>ERR2869484</td>
</tr>
<tr>
<td>IsoSeq_3-6kb_polished_high_qv_consensus_isoforms.fastq.gz</td>
<td>mRNA sequenced with PacBio SMRT RSII.</td>
<td>ERR2869483</td>
</tr>
<tr>
<td>1-Hfo-miRNA_S6_R1_001.fastq.gz</td>
<td>Small RNA sequenced with Illumina NextSeq 500.</td>
<td>ERR2869485</td>
</tr>
<tr>
<td>pilon_round3.fasta.gz</td>
<td>Draft genome assembly</td>
<td>ERZ780628</td>
</tr>
</tbody>
</table>
to correct the filtered Nanopore reads. Then the corrected reads were run through Canu v1.6 (Koren et al., 2017) for further correction (canu -correct with genome Size set to 300 m) and trimming (canu -trim) before assembly (canu -assemble). The assembly was done with two different corrected error rates, 0.144 and 0.146. The two assemblies were almost identical, but the results from using the corrected error rate of 0.144 were used further because the total size was slightly larger and also had the largest contig. The Canu assembly was then polished using the trimmed Illumina reads (described above) by running three rounds of Pilon v1.22 (Walker et al., 2014). The final genome assembly consisted of 5069 contigs with a total length of 171 183 409 nt. The N50 was 43,856 nt and the longest contig of 5,118,963 nt (Table 3).

**Table 2. Summary of the read numbers in the different file types of the IsoSeq data set.**

<table>
<thead>
<tr>
<th>Library</th>
<th>Size fraction</th>
<th>&lt; 1kb</th>
<th>1-2kb</th>
<th>2-3kb</th>
<th>3-4kb</th>
<th>4-5kb</th>
<th>5-6kb</th>
<th>&gt; 6kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2kb_high</td>
<td>7310</td>
<td>31953</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1-2kb_low</td>
<td>903</td>
<td>5908</td>
<td>170</td>
<td>110</td>
<td>89</td>
<td>39</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>2-3kb_high</td>
<td>596</td>
<td>2703</td>
<td>37399</td>
<td>217</td>
<td>443</td>
<td>147</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2-3kb_low</td>
<td>78</td>
<td>586</td>
<td>7749</td>
<td>116</td>
<td>301</td>
<td>134</td>
<td>215</td>
<td></td>
</tr>
<tr>
<td>3-6kb_high</td>
<td>13</td>
<td>552</td>
<td>28621</td>
<td>4603</td>
<td>20</td>
<td>388</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-6kb_low</td>
<td>0</td>
<td>6</td>
<td>268</td>
<td>8830</td>
<td>1535</td>
<td>74</td>
<td>418</td>
<td></td>
</tr>
</tbody>
</table>

1-2kb_high – accession ERR2869477; 1-2kb_low - accession ERR2882521; 2-3kb_high - accession ERR2869478; 2-3kb_low - accession ERR2869481; 3-6kb_high - accession ERR2869483; 3-6kb_low - accession ERR2869484.

**Table 3. Statistics of the Hydrurus foetidus draft genome assembly.**

<table>
<thead>
<tr>
<th>Number of contigs &gt; 1000 bp</th>
<th>Largest contig</th>
<th>Contig N50</th>
<th>Assembly size</th>
<th>Estimated genome size</th>
<th>Complete and fragmented BUSCO orthologs</th>
</tr>
</thead>
<tbody>
<tr>
<td>5069</td>
<td>5118963 bp</td>
<td>43856 bp</td>
<td>171 Mb</td>
<td>299.9 Mb</td>
<td>77.2%</td>
</tr>
</tbody>
</table>

*The genome size estimation was based on k-mer frequencies on the Illumina data.

**Data availability**

All Hydrurus foetidus datasets produced in this study are available, study accession number PRJEB29405: https://identifiers.org/ena.embl/PRJEB29405.

**Author contributions**

JB performed DNA and RNA isolation, Nanopore sequencing, genome assembly, prepared and submitted data and wrote the manuscript. JF monitored and performed culturing, isolated DNA and RNA and prepared for Illumina and PacBio sequencing, prepared and submitted data and commented on the manuscript. KSJ planned and monitored the entire project and commented on the manuscript. DK maintained and developed the cultures and isolation protocol and wrote the manuscript.

**Grant information**

JB was funded by the Research Council of Norway (grant numbers: 213703 and 240284).

_The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript._

**Acknowledgements**

We thank the Norwegian Sequencing Centre for all sequence library preparations and sequencing services.

**References**


del Campo J, Massara R: Emerging diversity within chrysophytes, choano flagellates and bicosoecids based on molecular surveys. Protoplast. 2011; 162(3): 435–448. Published Abstract | Publisher Full Text

Fott B: Algenkunde. VEB Gustav Fischer Verlag Jena.1959; 482 S. Reference Source


Ojcowie. 1953; 482 S. Reference Source


Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 09 May 2019

https://doi.org/10.5256/f1000research.18291.r47028

© 2019 Beisser D et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniela Beisser  
Department of Biodiversity, University of Duisburg-Essen, Essen, Germany

Stephan Majda  
Department of Biodiversity, University of Duisburg-Essen, Essen, Germany

Please find below the comments to the article “Draft genome assembly and transcriptome sequencing of the golden algae *Hydrurus foetidus* (Chrysophyceae)” by Jon Bråte, Janina Fuss, Kjetill S. Jakobsen and Dag Klaveness.

The authors present in the article draft genome data for the chrysophyte *Hydrurus foetidus*. *H. foetidus* was chosen because it is a special chrysophyte which is, in contrast to most other chrysophytes, macroscopic, but cells can slide apart under some growth conditions, benthic and thrives under cold conditions in polar and high alpine regions.

The genome was sequenced using Illumina and Nanopore sequencing, in addition mRNA reads sequenced with PacBio and small RNA reads sequenced with Illumina are provided. The genome was assembled in a hybrid approach, resulting in a length of 171 Mb and 5,069 contigs with a BUSCO completeness of 77%.

DNA extraction, sequencing and assembly generation are described in detail, but some information is missing, which is listed below.

Overall, the work is technically sound and should be indexed after minor corrections.

Minor comments:

1. Abstract: “We combined this data with, to our knowledge, the first draft genome assembly of a chrysophyte algae”. The data is not combined with the draft genome, maybe it should be “We combined this data to create, to our knowledge, the first draft genome assembly of a chrysophyte algae”. Further, there are two draft genomes listed at JGI for *Ochromonas* and *Paraphysomonas* species (https://genome.jgi.doe.gov/portal/OchCCMStandDraft_FD/OchCCMStandDraft_FD.info.html, https://genome.jgi.doe.gov/portal/ParimpEvaluation_FD/ParimpEvaluation_FD.info.html). They cannot be accessed without registration, but perhaps they should be cited.

3. Abstract: An assembly of 171 Mb was obtained, is this size expected? Is there other data available which suggests this genome size? On the other hand, the k-mer based genome size estimation suggests a much larger genome.

4. Introduction: “multiapproach assembly”, should better be described as hybrid assembly.

5. Culturing of H. foetidus: The introduction mentions plastid reduction and heterotrophy in chrysophyte species. H. foetidus in called alga, so I assume it is phototrophic, but it would be good to state that somewhere explicitly. Otherwise one would wonder if the cultures were axenic.

6. Concerning the sequencing: Was there quality control performed by the sequencing center (RIN values etc.) before sequencing? Which library preparation protocols were used for the different sequencing data? Was the RNA rRNA-depleted before sequencing? What was the average sequence quality before and after filtering? An overview table of all the sequencing data including quality and number of reads/yield should be added.

7. Nanopore sequencing of genomic DNA: “Albacore v2.1.10 (Linux, Python 3.5 version)” Please provide a reference for Albacore. The information that it was run under Linux with Python 3.5 is probably not necessary.

8. Next line: Table 1 is not the correct reference for the sentence.

9. Table 2: It is not described in the text or caption what “_high” and “_low” means.

10. Table 3: The genome size estimation based on k-mers and BUSCO are not described in the methods section. Maybe the GC content could also be added to the table.

References


Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes
Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** High-throughput sequencing analysis

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com