





Long-Read–Based Hybrid Genome Assembly and Annotation of Snow Algal Strain CCCryo 101-99 (cf. *Sphaerocystis* sp., Chlamydomonadales)

Ozan Çiftçi ^{1,*}, Athanasios Zervas ², Stefanie Lutz^{1,3}, Helen Feord¹, Christoph Keusching¹, Thomas Leya ⁴, Martyn Tranter², Alexandre M. Anesio², Liane G. Benning ^{1,5}

¹Interface Geochemistry, GFZ German Research Centre for Geosciences, Potsdam, Germany

²Department of Environmental Science, Aarhus University, Roskilde, Denmark

³Plant-Soil Interactions, Department of Agroecology and Environment, Agroscope, Zurich, Switzerland

⁴Fraunhofer Institute for Cell Therapy and Immunology, Branch Bioanalytics and Bioprocesses IZI-BB, Potsdam, Germany

⁵Department of Earth Sciences, Freie Universität Berlin, Berlin, Germany

*Corresponding author: E-mail: ozan.ciftci@gfz-potsdam.de.

Accepted: June 20, 2024

Abstract

Polar regions harbor a diversity of cold-adapted (cryophilic) algae, which can be categorized into psychrophilic (obligate cryophilic) and cryotrophic (nonobligate cryophilic) snow algae. Both can accumulate significant biomasses on glacier and snow habitats and play major roles in global climate dynamics. Despite their significance, genomic studies on these organisms remain scarce, hindering our understanding of their evolutionary history and adaptive mechanisms in the face of climate change. Here, we present the draft genome assembly and annotation of the psychrophilic snow algal strain CCCryo 101-99 (cf. *Sphaerocystis* sp.). The draft haploid genome assembly is 122.5 Mb in length and is represented by 664 contigs with an N50 of 0.86 Mb, a Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness of 92.9% ($n = 1,519$), a maximum contig length of 5.3 Mb, and a guanine-cytosine (GC) content of 53.1%. In total, 28.98% of the genome (35.5 Mb) contains repetitive elements. We identified 417 noncoding RNAs and annotated the chloroplast genome. The predicted proteome comprises 14,805 genes with a BUSCO completeness of 97.8%. Our preliminary analyses reveal a genome with a higher repeat content compared with mesophilic chlorophyte relatives, alongside enrichment in gene families associated with photosynthesis and flagella functions. Our current data will facilitate future comparative studies, improving our understanding of the likely response of polar algae to a warming climate as well as their evolutionary trajectories in permanently cold environments.

Key words: genomes, snow, cryophilic, algae, climate change, chlorophyta.

Significance

In polar regions, cold-adapted algae thrive and bloom, yet genomic studies on these organisms are scarce. This is hindering our understanding of their adaptive mechanisms in the face of climate change. Here, we provide the first genome assembly of the psychrophilic snow algal strain CCCryo 101-99, revealing a higher repeat content compared with mesophilic relatives, and enrichments in gene families crucial for photosynthesis and flagellar functions. This genome data will allow further comparative analyses and help us obtain critical insights into the adaptive strategies of polar algae and their responses to a changing climate.

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Polar and alpine regions are characterized by extreme environmental conditions with snow and ice covering ~12% of the Earth's land surface. Significant yet not well-studied components of the polar biota are cold-adapted eukaryotic snow and glacier ice algae, which are the main primary producers in these ecosystems. They shape local biogeochemical cycles and affect albedo through their pigmentation, thus enhancing melting (Lutz et al. 2016; Chevrollier et al. 2023). On semipermanent and permanent snow fields, different snow algae phenotypes bloom during the summer melt seasons, despite the harsh conditions characterized by low temperatures, high irradiation, and oligotrophic nutrient levels. Snow algae produce green, yellow, orange, and red algal blooms on snow fields worldwide (Lutz et al. 2016; Segawa et al. 2018). The ever-lengthening melt seasons, in particular, in future warming climate scenarios, will lead to an expansion of the habitat extent of snow algae, further exacerbating glacier melt (Lutz et al. 2014; Segawa et al. 2018).

Various chlorophyte algae species belonging to the genera *Chlainomonas*, *Chlamydomonas*, *Chlorominima*, *Chloromonas*, *Limnomonas*, *Microglena*, *Raphidonema*, *Sanguina*, and *Scotiella* have been described from snow habitats, and their ecology, physiology, biogeographical distribution, and taxonomic diversity have been characterized (e.g. Remias et al. 2018; Procházková et al. 2019; Tesson and Pröschold 2022; Novis et al. 2023 and references therein). Molecular studies, particularly metabarcoding and metagenomic approaches, have only recently provided the first critical insights into polar biodiversity and evolution and adaptation to permanently cold environments (Clark et al. 2023). However, most work has focused on marine settings, and there are only a few available psychrophilic algae genomes (Blanc et al. 2012; Mock et al. 2017; Zhang et al. 2020, 2021; Hulatt et al. 2024). This hampers our ability to understand their evolutionary history, metabolic potential, and cellular responses and to better address their potential roles in future global warming climate scenarios. Here, we present extensive DNA and RNA sequencing (RNA-seq) data, a long-read-based draft assembly, annotations, and preliminary functional enrichment results of the genome of the psychrophilic snow algal strain CCCryo 101-99 (cf. *Sphaerocystis* sp.).

Results and Discussion

Genome Assembly

DNA sequencing using Illumina, PacBio, and Nanopore generated 31,984,466 (7.55 Gb), 1,986,964 (14.1 Gb, read N50: 12.8 kb), and 1,591,524 (1.9 Gb, read N50: 2 kb, max: 400 kb) reads, respectively. For the estimated genome size of 160 Mb, PacBio and Nanopore reads cumulatively

provided 98x sequencing depth (86x and 12x, respectively). For error correction with Illumina short reads, 21,366,832 (3.7 Gb) trimmed and quality-filtered reads were mapped onto the long-read assembly with a mapping rate of 98.8%. We compared the metrics and qualities of the assemblies produced by different error correction strategies and assemblers (supplementary table S1, Supplementary Material online). Canu hybrid long-read assembly corrected with Illumina short reads was the longest and most contiguous and had the highest mapping rate and BUSCO coverage and smallest number of assembly errors. An initial error correction step with long reads decreased the BUSCO coverage and contig N50 metrics and introduced additional assembly errors, so the correction was performed using short reads only.

In contaminant screening, 107 contigs comprising 73% of the total genome length (~90 Mb) were assigned to Chlorophyta and 8 contigs (1.8 Mb) to Streptophyta, and 557 contigs (~30 Mb) had no hits. In total, 66 contigs of 380 kb length that were assigned to Actinomycetota and Pseudomonadota were removed. The corrected and filtered haploid draft genome is 122.5 Mb in length represented by 664 contigs (68.3x average read coverage) with an N50 of 0.86 Mb, a BUSCO completeness of 92.9% (complete: 92.9% [single-copy: 92.4%, duplicated: 0.5%], fragmented: 2.5%, missing: 4.6%, chlorophyta_odb10, $n = 1,519$, mode = genome), a maximum contig length of 5.3 Mb, and a guanine-cytosine (GC) content of 53.1% (Table 1). The total genome size was estimated to be around 160 Mb based on k -mer counts of Illumina sequencing reads. One contig, 458 kb in length and assigned to Chlorophyta, was identified as a candidate chloroplast genome based on GC content (36.5%) and read coverage (2,586x).

Genome Annotation

The repeat content of the genome is 28.98% (35.5 Mb), and the most abundant category was unidentified repeats (13.24%; supplementary table S2, Supplementary Material online). Long terminal repeats (LTRs) comprised 8.85% of the genome, and the most abundant LTR family was Gypsy/DIRS1 (8.41%) which has been found in diverse eukaryotic species including green algae (Piednoël et al. 2011). The second most abundant repeat category was long interspersed retrotransposable elements (LINEs, 4.38%), and the most abundant LINE family was L1/CIN4 (2.67%). CCCryo 101-99 genome has a slightly higher repeat content compared with its mesophilic chlorophyte relatives, 22% in *Chlamydomonas reinhardtii* (Payne et al. 2023) and 24.76% in *Volvox carteri* (Prochnik et al. 2010), but not as high as other cold-adapted chlorophytes, 64% and 49% in *Chlamydomonas* sp. ICE-L and UWO, respectively (Zhang et al. 2020, 2021). Increased content of

Table 1 Genome assembly and annotation statistics of CCCryo 101-99 in comparison with other chlorophyte algae genomes

		Assembly size (kb)	Number of contigs/scaffolds	Longest contig (kb)	BUSCO (%)	N50	L50 (kb)	GC (%)	Repeat content (%)
Assembly	CCCryo 101-99	122,425	664	5,321	92.9	39	855	53.1	29.98
	<i>Chlamydomonas</i> sp. UWO241 ^a	211,600	2,458	nd	85.0	165	370	60.6	49
	<i>Limnomonas spitsbergensis</i> ^b	260,248	124	10,736	nd	21	3,933	54.1	33.16
	<i>C. reinhardtii</i> (v5.5) ^c	111,098	17 ^d	9,731	96.5	7	7,780	64.1	21.64
		Number of protein-coding genes	Mean gene length (kb)	Mean protein length (aa)	BUSCO (%)	Mean number of exons per gene	Mean exon length (bases)	Mean number of introns per gene	Mean intron length (bases)
Annotation	CCCryo 101-99	14,805	5.98	632	97.8	7.8	314	7.3	516
	<i>Chlamydomonas</i> sp. UWO241 ^a	16,325	8.06	nd	nd	nd	1,563	10.1	934
	<i>Limnomonas spitsbergensis</i> ^b	18,277	nd	nd	95.2	6.7	nd	nd	788
	<i>C. reinhardtii</i> (v5.6) ^c	16,656	5.48	nd	96.1	nd	nd	7.8	229

nd, not determined.

^aZhang et al. 2021. ^bHulatt et al. 2024. ^cCraig et al. 2021. ^d17 chromosomes + 37 unassembled scaffolds.

tandem repeats and transposable elements in the genome can potentially be helpful for organisms that survive in extreme habitats by creating genetic diversity that can drive the emergence of new adaptive traits (Schrader and Schmitz 2019).

Altogether, 197 transfer RNA (tRNA) genes were predicted, including one pseudogene and one selenocysteine tRNA gene, as well as 100 ribosomal RNA (rRNA) genes including 6 tandem repeats of the ribosomal DNA (rDNA) cistron on the same contig (18S, ITS1, 5.8S, ITS2, and 28S). These were 5.8 kb each and were separated by approximately 3 kb long intergenic spacers. Five of these six rDNA copies have identical sequences for their whole length, and one additional copy was identified in another contig. Other non-coding RNA (ncRNA) classifications include 84 histone 3' untranslated region (UTR) stem-loop, 17 spliceosomal RNA (U1, U2, U4, U5, and U6), 15 small nucleolar RNA (R38, R160, SNORD14, SNORD24, U3, U36a, Z157, and Z159), 2 mRNA (mir-191), 1 RNase MRP, and 1 antisense RNA (isrR).

The assembled chloroplast genome is circular, has one large (254 kb) and one small (202 kb) single-copy region, and has an inverted repeat comprising five rRNA and one tRNA genes (Fig. 1). In total, 112 genes were annotated in the chloroplast genome including 10 rRNA and 24 tRNA genes. In addition, four group I and three group II catalytic introns were identified. We performed additional analyses to assemble the mitochondrial genome using 65 *Chlamydomonadales* mitogenomes as a reference; however, no genes could be annotated on assembled contigs. The mitochondrial genomes of chlorophytes harbor significant diversity in terms of gene content and genome structure (Popescu and Lee 2007). *Chlamydomonas*-like mitochondrial genomes have reduced organizational pattern

characterized by small genome size, limited gene content, and the presence of fragmented and scrambled rRNA coding regions (Nedelcu et al. 2000). Our results suggest that CCCryo 101-99 follows this *Chlamydomonas*-like mitochondrial genome evolution.

The Illumina RNA-seq generated 8,748,326 (1.32 Gb) reads. After quality filtering, 4,772,938 (0.72 Gb) reads were mapped onto the genome assembly with a 92.4% mapping rate. Using the RNA-seq mapping as extrinsic evidence, 17,206 transcripts and 14,805 genes were predicted, where 5' and 3' UTR regions were identified for 94% of the transcripts, and 99.5% of the genes had both start and stop codons. The BUSCO completeness of the predicted proteome is 97.8% (complete: 97.8% [single-copy: 96.4%, duplicated: 1.4%], fragmented: 0.5%, missing: 1.7%, chlorophyta_odb10, $n = 1,519$, mode = proteome; Table 1).

Genome annotation assigned a protein or domain to 13,707 genes in total (92.5%) (supplementary table S3, Supplementary Material online). The most enriched category (enrichment score = 37.81) in the functional enrichment analysis against the *C. reinhardtii* genome included the Gene Ontology (GO) term "chloroplast thylakoid membrane" and photosynthetic genes (i.e. *psbW*, *petH*, and *rubA*). Duplicates and increased protein accumulation of the photosynthetic genes have been reported (Zhang et al. 2020) in the cold-adapted Antarctic green alga *Chlamydomonas* sp. UWO241 (i.e. compared with the mesophilic relative *C. reinhardtii*). These authors hypothesized that this is an indication of an adaptation to the cold. The second most enriched category in our annotation (enrichment score = 37.29) contained flagellum-associated genes (CFAP58/91/157/221). During its asexual life cycle, CCCryo 101-99 produces motile spores (zoospores) representing

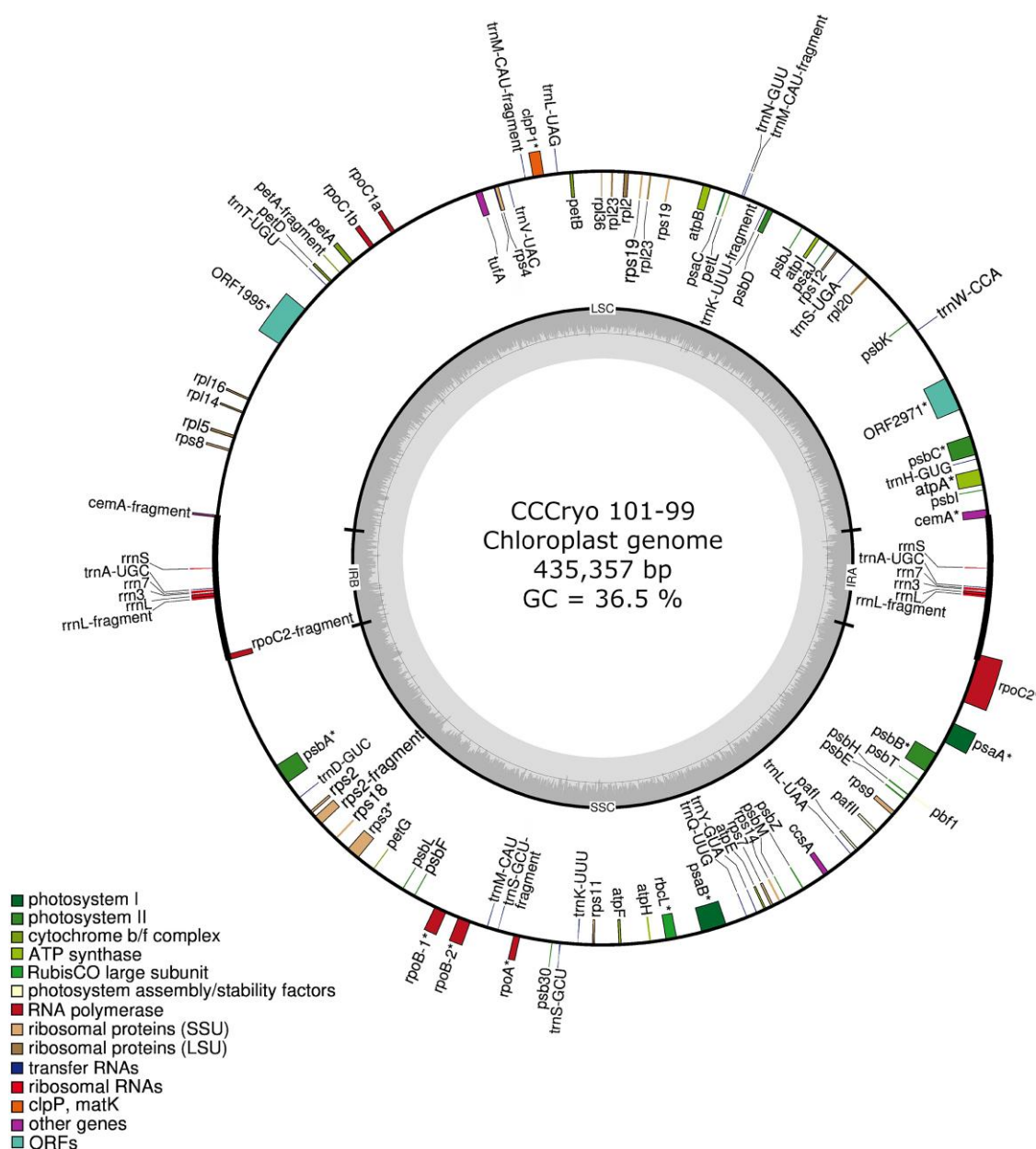


Fig. 1. The chloroplast genome of CCCryo 101-99. LSC, long single-copy region; SSC, short single-copy region; IRA, inverted repeat A; IRB, inverted repeat B.

young stages of haploid spores. These are not frequently observed in laboratory-grown cultures as they shed their flagella shortly after attaching to surfaces and develop into nonmotile adult spores that divide several times by multiple mitoses and develop into sporangia. Later, young motile zoospores are released from these sporangia to fulfill the asexual life cycle. Nothing is known about their sexual reproduction.

In summary, our analyses on the genome of the snow algal strain CCCryo 101-99 (cf. *Sphaerocystis* sp.) revealed a slightly higher (29.98%) repeat content, compared with the

mesophilic chlorophyte relative *C. reinhardtii*. Furthermore, the gene families associated with photosynthesis and flagellar functions were enriched. The former may contribute to the adaptation to cold climates by maintaining a high level of CO₂ fixation and, thus, optimizing energy production, while the latter may be important for an effective dispersal strategy in snow habitats by maintaining motility in young cell stages. Our draft-assembled snow algal genome and ongoing work on comparative snow algal genomics will allow in-depth comparisons with the increasing number of available genomes. This will improve our understanding of the evolution and

diversification of photosynthetic eukaryotes adapted to living in permanently cold climates and their responses to future global warming climate scenarios.

Materials and Methods

Strain Description

The snow algal strain CCCryo 101-99 was isolated in 1998 from moss patches along melt streams flowing down from snow fields above on Brøggerhalvøya northwest of Ny-Ålesund, Svalbard, Norway. The strain has been deposited and since then been maintained at the CCCryo Culture Collection of Cryophilic Algae at the Fraunhofer IZI-BB, Potsdam, Germany (Leya 2022). The taxonomic identity of CCCryo 101-99 still is under debate; therefore, for the time being, it is called *cf. Sphaerocystis* sp. According to its phylogenetic marker (*rbcl*), it shows relatedness (data not shown) to the genus *Achoma* (Novis and Visnovsky 2012). Detailed studies are currently in progress. It can be classified as a psychrophilic snow alga as its optimum temperature for growth lies at around 14 °C; its maximum temperature lies above 18 °C (Leya et al. 2009), but above that it starts ceasing growth and eventually dies. A progeny of CCCryo 101-99 had spent 531 d outside the International Space Station in its desiccated cyst stage as part of the BIOMEX experiments and survived (de Vera et al. 2019). CCCryo 101-99 extracts are also used in cosmetic products to guard against aging of the skin caused by extrinsic factors, such as ultraviolet radiation or air pollution, or intrinsic factors such as aging-specific gene expression levels (Stutz et al. 2010). The strain is known to synthesize the antioxidants and pigments: astaxanthin, adonixanthin, canthaxanthin, echinenone, hydroxyechinenone, and saccharose (Leya 2022).

Cultivation and DNA/RNA Extractions

The cultures for nucleic acid extractions were grown in a 3N Bold's Basal Medium at pH 5.5 and 2 °C (green biomass) or 12 °C (orange biomass induced from green biomass) under axenic conditions and continuous illumination. Genomic DNA (gDNA) was extracted from green biomass using the PowerSoil DNA Isolation kit (QIAGEN, Germany) for Illumina sequencing and the QIAGEN Genomic-tips extraction kit for PacBio sequencing, following the manufacturer's instructions. gDNA extraction for Nanopore sequencing was performed using a cetyltrimethylammonium bromide (CTAB)/phenol extraction of green biomass following Chaux-Jukic et al. (2024). We assessed the size, quantity, and integrity of the extracts on a TapeStation 4150 (Agilent Technologies) and a Qubit 4 Fluorometer (Thermo Scientific) and via gel electrophoresis with a low concentration agar (0.4%). For RNA extractions, freeze-dried green and orange cells were processed following Sanz-Luque and Montaigu

(2018) with the only modification being the replacement of the sodium dodecyl sulfate lysis buffer with CTAB.

Library Preparations and Sequencing

Illumina library preparations and sequencing were done at the Genome Analysis Centre (Earlham Institute, UK) on an Illumina MiSeq with the 500-cycle v2 chemistry kit in 250 bp pair-end mode. PacBio library preparation was performed using SMRTbell Template Prep Kit v1.0, and the sequencing was done on a PacBio Sequel SMRT Cell (2.1 chemistry) at the NERC Biomolecular Analysis Facility, Liverpool. Nanopore library preparation followed the SQK-LSK110 protocol (Oxford Nanopore, Oxford, UK), and sequencing was performed on a MinION with a FLO-MIN106 flow cell, controlled using MinKNOW (19.10.1) at Aarhus University. Raw Nanopore fast5 reads were basecalled with GPU-Guppy (3.2.6+afc8e14) under default settings.

Eukaryotic mRNA was selected from 1 µg of the RNA pool using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, #E7490) following the manufacturer's instructions. Sequencing libraries were prepared using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs, #E7765) following the manufacturer's instructions, with 5× dilution of the NEBNext Adaptor and 14 library amplifications cycles. The resulting libraries were measured on a Qubit 4 (Thermo Scientific) using the dsDNA High Sensitivity assay and visualized on a TapeStation 4150 (Agilent Technologies) using a D1000 ScreenTape and reagents. Sequencing was performed on an Illumina NextSeq 500 with the 300-cycle v2.5 chemistry kit in 151 bp pair-end mode.

Genome Assembly, Error Correction, and Contaminant Screening

The qualities of the raw Illumina, Nanopore, and PacBio sequencing reads were inspected using FastQC (Andrews 2010) and RabbitQC (Yin et al. 2021). Adapters and low-quality ends of Illumina reads were trimmed using `bbduk.sh` script with parameters "`k = 23 mink = 11 hdist = 1 tbo trimq = 30 qtrim = r1 minlength = 50 maq = 30`" (<https://sourceforge.net/projects/bbmap/>). The basecalled Nanopore reads were processed with Porechop v.0.2.4 (<https://github.com/rwick/porechop>) using default parameters to remove sequencing adapters. PacBio and Nanopore long reads were assembled using a hybrid approach with Canu v2.2 (Koren et al. 2017). Two rounds of error correction were performed with Pilon v1.24 (Walker et al. 2014) on the hybrid long-read assembly using the filtered Illumina reads and minimap v2.26 (Li 2018). We assessed the completeness of the assembly based on the Chlorophyta (odb10, 2020-08-05) data set of BUSCO v5.4.6 (Manni et al. 2021), computed summary statistics, and evaluated

assembly errors using Inspector v1.0.2 (<https://github.com/ChongLab/Inspector>).

Assembled contigs were screened for contaminants by assigning taxonomy using National Center for Biotechnology Information (NCBI) Foreign Contaminant Screening tool (FCS-GX; <https://github.com/ncbi/fcs-gx>). Genome size was estimated using the *k*-mer analysis tool jellyfish v.2.3 (<https://github.com/gmarcais/Jellyfish/>) using Illumina reads and following an online tutorial (<https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/>).

Repeat Masking and Structural and Functional Annotations

A de novo repeat library for strain CCCryo 101-99 was constructed with RepeatModeler v2.0.4 (Smit and Hubley 2008–2015) and RepeatScout v1.0.6 (Price et al. 2005) using the default parameters. First, simple and complex repeats were identified based on the Chlorophyta repeat library of Repbase (version 29.04) using RepeatMasker v4.1.5 (Smit et al. 2013–2015; rmbblastn version v2.13.0), and de novo repeats were then identified. Finally, all repeats were combined, and the genome was soft-masked using bedtools v2.26.0 (<https://github.com/arq5x/bedtools2/>).

Adapter removal and end trimming of Illumina RNA-seq reads were performed using Trimmomatic v.0.39 (Bolger et al. 2014) with parameters “leading: 5 trailing: 5 sliding-window: 5:10 minlen: 36.” Filtered RNA-seq reads were mapped onto the soft-masked genome using STAR v2.5.2b (Dobin et al. 2013). Coding regions were predicted by incorporating RNA-seq mapping data as extrinsic evidence with BRAKER v3.06 (Stanke et al. 2006, 2008), which is an automated pipeline that utilizes ab initio gene prediction tools GeneMark and AUGUSTUS. The 5′ and 3′ UTR regions were predicted with GUSHR (<https://github.com/Gaius-Augustus/GUSHR>) based on RNA-seq coverage information, and the predictions were added as gene features to structural annotations. The models and splice patterns were manually checked for a subset of genes with IGV v2.14.1 (Robinson et al. 2011). The longest isoform for each gene was selected for the downstream analyses.

Functional annotations of the BRAKER predicted translations were performed by obtaining the top BLAST hits using BLASTP (*E*-value < $1e^{-5}$) against SwissProt (UniProt 2023) and DIAMOND BLASTP (<https://github.com/bbuchfink/diamond>; *E*-value < $1e^{-5}$, sensitive mode) against UniRef90 (release version: 2023-01) databases. InterPro protein families and domains, Pfam domains, and GO terms were annotated using InterProScan v5.65.97 (Jones et al. 2014). The predictions of the transmembrane domains by TMHMM v2.0c (Krogh et al. 2001) and signal sequences by SignalP v4.1 (Nielsen 2017) were incorporated into InterProScan results. Functional gene enrichment analysis against the *C. reinhardtii* genome was performed using the DAVID web

server (<https://david.ncicrf.gov/>). The tRNA genes were annotated by tRNAscan-SE v2.0.9 (Chan and Lowe 2019), rRNA genes were annotated using RNAMMER v1.2 (Lagesen et al. 2007), and ncRNA families were annotated using Infernal v1.1.4 (Nawrocki and Eddy 2013) and Rfam database (v14.10). One contig assigned to the chloroplast genome (i.e. based on coverage and GC content) was circularized using a python script (<https://github.com/Kzra/Simple-Circularise>), annotated using GeSeq (Tillich et al. 2017) and the *C. reinhardtii* chloroplast genome as the reference, and visualized using OGDRAW (Greiner et al. 2019). Default parameters for eukaryotes were used for all tools. All bioinformatic analyses were performed at the GFZ servers in Potsdam.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Thanh Hang Pham, Marc Christian, and Alex Barthelemos (GFZ) for the laboratory and bioinformatics server support, Felix Jorde (IZI-BB) for providing algal biomass, Tina Thane (Aarhus University) for the mRNA library preparations, and NBAF Liverpool staff for the Illumina sequencing.

Author Contributions

Conceptualization: L.G.B., T.L., S.L., and O.C. Data curation: L.G.B., S.L., H.F., and O.C. Formal analysis: O.C. Investigation and methodology: O.C. and A.Z. Funding acquisition and project administration: A.M.A., M.T., and L.G.B. Writing—original draft: O.C. Writing—review & editing: A.Z., S.L., H.F., C.K., T.L., A.M.A., M.T., and L.G.B.

Funding

O.Ç. was funded by the Helmholtz Recruiting Initiative, awarded to L.G.B. (award no. I-044-16-01). This project also received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 856416. PacBio WGS experiments were conducted with NERC Biomolecular Analysis Facility grant Snow Algae Genomics (NBAF1083), and Illumina WGS experiments were conducted with a personal research grant to L.G.B. by the University of Leeds.

Data Availability

The raw sequences of CCCryo 101-99 are available on NCBI under BioProject accession PRJNA1036577. The Whole Genome Shotgun project has been deposited at DDBJ/ENA/

GenBank under the accession JBEFCQ000000000. The version described in this paper is version JBEFCQ010000000. The annotation files are deposited to Zenodo (doi:10.5281/zenodo.10799857). Illumina whole genome sequencing data for CCCryo snow algae strains 002b-99 (*Microglena* sp.), 011a-99 (*Raphidonema sempervirens*), and 005-99 and 047-99 (*Chloromonas remiasii*) are also uploaded under the same BioProject accession. The cultivation and DNA extraction of these strains followed the same protocols as 101-99 as described in the Materials and Methods section. The sequencing of these strains was performed by the Earlham Institute, UK. This additional genome data is currently used by our group for further studies on snow algae.

Literature Cited

- Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010 [accessed 2023 Mar 12]. www.bioinformatics.babraham.ac.uk/projects/fastqc.
- Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* 2012;13(5):R39. <https://doi.org/10.1186/gb-2012-13-5-r39>.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
- Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol.* 2019;1962:1–14. https://doi.org/10.1007/978-1-4939-9173-0_1.
- Chaux-Jukic F, Agier N, Eberhard S, Xu Z. Extraction and selection of high-molecular-weight DNA for long-read sequencing from *Chlamydomonas reinhardtii*. *PLoS One.* 2024;19(2):e0297014. <https://doi.org/10.1371/journal.pone.0297014>.
- Chevrollier L-A, Cook JM, Halbach L, Jakobsen H, Benning LG, Anesio AM, Tranter M. Light absorption and albedo reduction by pigmented microalgae on snow and ice. *J. Glaciol.* 2023;69(274):333–341. <https://doi.org/10.1017/jog.2022.64>.
- Clark MS, Hoffman JI, Peck LS, Bargelloni L, Gande D, Havermans C, Meyer B, Patarnello T, Phillips T, Stoof-Leichsenring KR, et al. Multi-omics for studying and understanding polar life. *Nat Commun.* 2023;14(1):7451. <https://doi.org/10.1038/s41467-023-43209-y>.
- Craig RJ, Hasan AR, Ness RW, Keightley PD. Comparative genomics of *Chlamydomonas*. *Plant Cell.* 2021;33(4):1016–1041. <https://doi.org/10.1093/plcell/koab026>.
- de Vera J-P, Alawi M, Backhaus T, Baqué M, Billi D, Böttger U, Berger T, Bohmeier M, Cockell C, Demets R et al. Limits of life and the habitability of Mars: the ESA space experiment BIOMEX on the ISS. *Astrobiology.* 2019;19(2):145–157. <https://doi.org/10.1089/ast.2018.1897>.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
- Greiner S, Lehwarck P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 2019;47(W1):W59–W64. <https://doi.org/10.1093/nar/gkz238>.
- Hulatt C, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. The genome of the Arctic snow-alga *Limnomonas spitsbergensis* (Chlamydomonadales). G3 (Bethesda). 2024;jkae086. <https://doi.org/10.1093/g3journal/jkae086>, preprint: not peer reviewed.
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England).* 2014;30(9):1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27(5):722–736. <https://doi.org/10.1101/gr.215087.116>.
- Krogh A, Larsson B, Heijne Gv, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001;305(3):567–580. <https://doi.org/10.1006/jmbi.2000.4315>.
- Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 2007;35(9):3100–3108. <https://doi.org/10.1093/nar/gkm160>.
- Leya T. The CCCryo Culture Collection of Cryophilic Algae as a valuable bioresource for algal biodiversity and for novel, industrially marketable metabolites. *Appl Phycol.* 2022;3(1):167–188. <https://doi.org/10.1080/26388081.2020.1753572>.
- Leya T, Rahn A, Lütz C, Remias D. Response of Arctic snow and permafrost algae to high light and nitrogen stress by changes in pigment composition and applied aspects for biotechnology. *FEMS Microbiol Ecol.* 2009;67(3):432–443. <https://doi.org/10.1111/j.1574-6941.2008.00641.x>.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- Lutz S, Anesio AM, Jorge Villar SE, Benning LG. Variations of algal communities cause darkening of a Greenland glacier. *FEMS Microbiol Ecol.* 2014;89(2):402–414. <https://doi.org/10.1111/1574-6941.12351>.
- Lutz S, Anesio AM, Raiswell R, Edwards A, Newton RJ, Gill F, Benning LG. The biogeography of red snow microbiomes and their role in melting Arctic glaciers. *Nat Commun.* 2016;7(1):11968. <https://doi.org/10.1038/ncomms11968>.
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38(10):4647–4654. <https://doi.org/10.1093/molbev/msab199>.
- Mock T, Otilar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ, et al. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature.* 2017;541(7638):536–540. <https://doi.org/10.1038/nature20803>.
- Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* 2013;29(22):2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>.
- Nedelcu AM, Lee RW, Lemieux C, Gray MW, Burger G. The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. *Genome Res.* 2000;10(6):819–831. <https://doi.org/10.1101/gr.10.6.819>.
- Nielsen H. Predicting secretory proteins with SignalP. *Methods Mol Biol.* 2017;1611:59–73. https://doi.org/10.1007/978-1-4939-7015-5_6.
- Novis PM, Dhami M, Podolyan A, Matsumoto M, Kodner R. The austral biflagellate *Chloromonas rubroleosa* (Chlorophyceae) is the closest relative of the unusual quadriflagellate genus *Chlainomonas*, both

- found in snow. *J Phycol.* 2023;59(2):342–355. <https://doi.org/10.1111/jpy.13318>.
- Novis PM, Visnovsky G. Novel alpine algae from New Zealand: chlorophyta. *Phytotaxa.* 2012;39(1):1–30. <https://doi.org/10.11646/phytotaxa.39.1.1>.
- Payne ZL, Penny GM, Turner TN, Dutcher SK. A gap-free genome assembly of *Chlamydomonas reinhardtii* and detection of translocations induced by CRISPR-mediated mutagenesis. *Plant Commun.* 2023;4(2):100493. <https://doi.org/10.1016/j.xplc.2022.100493>.
- Piednoël M, Gonçalves IR, Higuët D, Bonnavard E. Eukaryote DIRS1-like retrotransposons: an overview. *BMC Genomics.* 2011;12(1):621. <https://doi.org/10.1186/1471-2164-12-621>.
- Popescu CE, Lee RW. Mitochondrial genome sequence evolution in *Chlamydomonas*. *Genetics.* 2007;175(2):819–826. <https://doi.org/10.1534/genetics.106.063156>.
- Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics.* 2005;21(Suppl 1):i351–i358. <https://doi.org/10.1093/bioinformatics/bti1018>.
- Procházková L, Leya T, Křížková H, Nedbalová L. *Sanguina nivaloides* and *Sanguina aurantia* gen. et spp. nov. (Chlorophyta): the taxonomy, phylogeny, biogeography and ecology of two newly recognised algae causing red and orange snow. *FEMS Microbiol Ecol.* 2019;95(6):fiz064. <https://doi.org/10.1093/femsec/fiz064>.
- Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carterii*. *Science.* 2010;329(5988):223–226. <https://doi.org/10.1126/science.1188800>.
- Remias D, Procházková L, Holzinger A, Nedbalová L. Ecology, cytology and phylogeny of the snow alga *Scotiella cryophila* K-1 (Chlamydomonadales, Chlorophyta) from the Austrian Alps. *Phycologia.* 2018;57(5):581–592. <https://doi.org/10.2216/18-45.1>.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24–26. <https://doi.org/10.1038/nbt.1754>.
- Sanz-Luque E, Montaigu AD. Phenol-based extraction of RNA from *Chlamydomonas reinhardtii*. *Bio Protoc.* 2018;Bio-101:e2910. <https://doi.org/10.21769/BioProtoc.2910>.
- Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. *Mol Ecol.* 2019;28(6):1537–1549. <https://doi.org/10.1111/mec.14794>.
- Segawa T, Matsuzaki R, Takeuchi N, Akiyoshi A, Navarro F, Sugiyama S, Yonezawa T, Mori H. Bipolar dispersal of red-snow algae. *Nat Commun.* 2018;9(1):3094. <https://doi.org/10.1038/s41467-018-05521-w>.
- Smit A, Hubley R. RepeatModeler open-1.0; 2008–2015 [accessed 2023 Apr 11]. <http://www.repeatmasker.org>.
- Smit A, Hubley R, Green P. RepeatMasker open-4.0; 2013–2015 [accessed 2023 Apr 05]. <http://www.repeatmasker.org>.
- Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntetically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* 2008;24(5):637–644. <https://doi.org/10.1093/bioinformatics/btn013>.
- Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* 2006;7(1):62. <https://doi.org/10.1186/1471-2105-7-62>.
- Stutz CS, Schmid D, Züllig F, inventors: Use of an extract from snow algae in cosmetic or dermatological formulations. U.S. Patent No. US20100316720A1. 2010.
- Tesson SVM, Pröschold T. Description of *Limnomonas* gen. nov., *L. gaiensis* sp. nov. and *L. spitsbergensis* sp. nov. (Chlamydomonadales, Chlorophyta). *Diversity (Basel).* 2022;14(6):481. <https://doi.org/10.3390/d14060481>.
- Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. Geseq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 2017;45(W1):W6–W11. <https://doi.org/10.1093/nar/gkx391>.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963. <https://doi.org/10.1371/journal.pone.0112963>.
- Yin Z, Zhang H, Liu M, Zhang W, Song H, Lan H, Wei Y, Niu B, Schmidt B, Liu W. RabbitQC: high-speed scalable quality control for sequencing data. *Bioinformatics.* 2021;37(4):573–574. <https://doi.org/10.1093/bioinformatics/btaa719>.
- Zhang X, Cvetkovska M, Morgan-Kiss R, Hüner NPA, Smith DR. Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. *iScience.* 2021;24(2):102084. <https://doi.org/10.1016/j.isci.2021.102084>.
- Zhang Z, Qu C, Zhang K, He Y, Zhao X, Yang L, Zheng Z, Ma X, Wang X, Wang W, et al. Adaptation to extreme antarctic environments revealed by the genome of a sea ice green alga. *Curr Biol.* 2020;30(17):3330–3341.e7. <https://doi.org/10.1016/j.cub.2020.06.029>.

Associate editor: Sujal Phadke