Genome sequencing of *Prototheca zopfii* genotypes 1 and 2 provides evidence of a severe reduction in organellar genomes

Marco Severgnini, Barbara Lazzari, Emanuele Capra, Stefania Chezza, Mario Luini, Roberta Bordoni, Bianca Castiglioni, Matteo Ricchi & Paola Cremonesi

*Prototheca zopfii* (*P. zopfii*, class Trebouxiophyceae, order Chlorellales, family Chlorellaceae), a non-photosynthetic predominantly free-living unicellular alga, is one of the few pathogens belonging to the plant kingdom. This alga can affect many vertebrate hosts, sustaining systemic infections and diseases such as mastitis in cows. The aim of our work was to sequence and assemble the *P. zopfii* genotype 1 and genotype 2 mitochondrial and plastid genomes. Remarkably, the *P. zopfii* mitochondrial (38 Kb) and plastid (28 Kb) genomes are models of compaction and the smallest known in the Trebouxiophyceae. As expected, the *P. zopfii* genotype 1 and 2 plastid genomes lack all the genes involved in photosynthesis, but, surprisingly, they also lack those coding for RNA polymerases. Our results showed that plastid genes are actively transcribed in *P. zopfii*, which suggests that the missing RNA polymerases are substituted by nuclear-encoded paralogs. The simplified architecture and highly-reduced gene complement of the *P. zopfii* mitochondrial and plastid genomes are closer to those of *P. stagnora* and the achlorophyllous obligate parasite *Helicosporidium* than to those of *P. wickerhamii* or *P. cutis*. This similarity is also supported by maximum likelihood phylogenetic analyses inferences. Overall, the *P. zopfii* sequences reported here, which include nuclear genome drafts for both genotypes, will help provide both a deeper understanding of the evolution of *Prototheca* spp. and insights into the corresponding host/pathogen interactions.

Organisms belonging to the genus *Prototheca* are chlorophyllous algae widespread in the environment. The genus is classified within the class of Trebouxiophyceae, order Chlorellales and family Chlorellaceae, and historically encompasses six species: *P. stagnora*, *P. ulnea*, *P. wickerhamii*, *P. blaschkeae*, *P. zopfii* and *P. cutis*. A seventh species, *P. miyajii*, has very recently been isolated in a patient with systemic protothecosis and classified as a separate species due to some genetic and phenotypical differences from *P. wickerhamii*. Finally, an eighth species, *P. moriformis*, is not currently considered a species per se because of its biochemical/genetic resemblance to *P. zopfii* and because of its high intraspecific heterogeneity. All *Prototheca* species have forfeited their photosynthetic capabilities, and, consequently, their ability to harvest energy from light and fix carbon, having undergone an evolutionary transition from autotrophy to heterotrophy, favoured also by the ability of some species to sustain infectious diseases in both humans and animals. *P. wickerhamii*, *P. cutis* and *P. blaschkeae*, in particular, have been associated with human diseases, especially in the presence of impaired immunological-cellular systems. Nevertheless, *P. wickerhamii*, *P. blaschkeae* and *P. zopfii* can also infect animals, especially dogs and cows. Among them, *P. blaschkeae* and *P. zopfii* are the most important species in the veterinary field because of their ability to sustain bovine mastitis, whereas genotype 2 is the most isolated *Prototheca* in bovine mastitis outbreaks worldwide.

1 Institute of Biomedical Technologies, National Research Council (ITB-CNR), Segrate, Milan, Italy. 2 PTP-Science Park, Lodi, Italy. 3 Institute of Agricultural Biology and Biotechnology, National Research Council (IBBA-CNR), Lodi, Italy. 4 Lombardy and Emilia Romagna Experimental Zootechnic Institute (IZSLER), Lodi, Italy. 5 Lombardy and Emilia Romagna Experimental Zootechnic Institute (IZSLER), Piacenza, Italy. Marco Severgnini, Barbara Lazzari and Emanuele Capra contributed equally. Correspondence and requests for materials should be addressed to M.R. (email: matteo.ricchi@izsler.it)
The sequences of *Prototheca* species currently available in public databases are those of the 18S rDNA (small subunit of rDNA, SSU) and 28S rDNA (large subunit of rDNA, LSU)\(^1\), and those of the Internal Transcribed Spacer regions (ITS), as well as some mitochondrial and plastid genomes. Notably, this information is not available for all species and, full-length sequences are very often missing\(^1,2^{15},2^{21}\); complete sequences of the organelar DNA are currently only available for *P. wickerhammitii* (both mitochondrion and plastid)\(^1,2^{15}\), *P. cutis* and *P. stagnora* (plastid only)\(^24\).

In this paper, we present the complete and manually annotated genomes of mitochondria and plastids of *P. zopfii* genotypes 1 and 2, and the first draft assembly of the whole nuclear genomes of both *P. zopfii* subspecies. Our work gives, for the first time, a representative overview of the extreme reduction which occurs within the mitochondrial and plastid genomes of these algae and provides basic information for further investigations.

Materials and Methods

**Strains and culture conditions.** *P. zopfii* genotype 1 (SAG 2063) and *P. zopfii* genotype 2 (SAG 2021) were obtained from the Culture Collection of Algae at Göttingen University (“Sammlung von Algenkulturen der Universität Göttingen”, SAG, Göttingen, Germany). The strains were aerobically sub-cultured on Sabouraud Dextrose Agar plates for 48–72 h at 30 °C until DNA isolation was carried out.

**Isolation of DNA and RNA.** Genomic DNA and RNA were extracted starting from approximately one g of *P. zopfii* genotype 1 and genotype 2 resuspended in 20 ml wash solution (0.6 M Sucrose, 20 mM Tris, 20 mM MgCl₂ and 1 mM DTT, pH 7.5) and centrifuged at 2200 g for 2 min at 4 °C. For DNA, washed pellets were resuspended in 500 µl of TRIS EDTA lysis buffer (10 mM Tris-HCl, 10 mM EDTA, 250 mM NaCl, pH 8), supplemented with 25 µl of proteinase K (20 mg/ml) (Sigma-Aldrich, St. Louis, MO, USA), 25 µl of SDS 10% and incubated at 56 °C for 2 h. Next, 25 µl of RNaseA (20 mg/ml) (Sigma-Aldrich) were added and the suspensions were incubated at 56 °C for 30 min. DNA was extracted using an equal volume of 1:1 (v/v) phenol:chloroform and precipitated with one volume of cold isopropanol. DNA was rinsed with 70% (v/v) cold ethanol, air dried, resuspended in 30 µl of ultrapure water and stored at −20 °C until use. DNA concentration and quality were estimated by PicoGreen (Thermo Fisher, Waltham, MA, USA) and by agarose gel electrophoresis.

Total RNA was extracted from washed pellets with TRIzol (Invitrogen, Carlsbad, CA, USA) and purified by NucleoSpin miRNA kit (Macherey-Nagel, Duren, Germany), following the manufacturer’s protocol, in combination with TRIzol lysis. RNA concentration (ng/µl) and quality RNA Integrity Number (RIN) were determined by Agilent 2100 Bioanalyzer (Santa Clara, CA, USA). RNA extracts were stored at −80 °C until use.

**DNA and RNA library preparation and sequencing.** *P. zopfii* DNA libraries for Illumina sequencing (paired-end and mate-pair sequencing) (Illumina, San Diego, CA, USA) were prepared and sequenced on an Illumina MiSeq instrument, following the manufacturer’s instructions as detailed in Supplementary Table 1. DNA library preparation for GS-FLX sequencing was performed using the GS FLX Titanium Rapid Library Preparation Kit (Roche, Basel, Switzerland) as follows: 1 µg DNA for each *P. zopfii* strain was used in the preparation of shotgun libraries by genomic DNA fragmentation by nebulization and ligation to specific adapters. According to the manufacturer’s instructions, libraries were subjected to clonal amplification by emulsion PCR reaction, recovered by isopropanol breaking and enriched for positive reaction beads. Each library was separately loaded onto one region of the GS-FLX PicoTiter Plate and sequenced according to the 454 GS-FLX Titanium XL protocol. One µg of RNA was used for libraries construction using TruSeq® RNA Sample Preparation v2 Kit (Illumina), according to the manufacturer’s instructions using poly(A) enrichment and sequenced on a 2 × 101-cycles Hiseq 2000 run (Illumina).

**De novo genome assembly.** The whole genomic sequence collections - composed of Illumina paired-end and mate-pair sequences, plus GS-FLX reads - from *P. zopfii* genotype 1 and 2 were assembled with SPAdes\(^25\) in read error correction and assembling mode. Six K values, ranging from K21 to K127, were automatically selected by the algorithm based on the read length and dataset type. CAP3\(^26\) (-p 96 -o 500) was run on the resulting scaffolds to further assemble contiguous regions. Illumina reads were mapped back to scaffolds by means of bowmem (v. 0.7.1027), keeping only those having ≤2 mismatches with the reference and Pilon (v 1.22)\(^28\) in assembly improvement mode was employed in order to correct substitutions and short indels.

**Organelles assembly.** Among the assembled scaffolds, sequences of mitochondrion and plastid were detected by homology with a set of 322 mitochondrial and 605 plastid/chloroplast sequences from related species (i.e.: green algae) and circularized, giving rise to full circular sequences for both organelles in both *P. zopfii* genotypes. Assembly of the mitochondrial and plastid genomes of *P. zopfii* was independently confirmed by a custom assembly strategy that, starting from “seeds”, implements an iterative procedure aimed at finding reads partially overlapping with the seed and assembling them with the original contig. Then, overlaps among the new contigs were used to generate a “supercontig”. Further alignment-assembly-extension steps were performed on each side of the supercontig until an overlap between the 5′ and 3′ ends of the sequence was found, meaning that the whole circular genome had been covered. A full description of this custom assembly procedure is available in Supplementary Methods.

**Nuclear genome annotation.** Gene prediction in nuclear genomes was performed with Augustus\(^29\) using *Chlorella variabilis* as reference species. A double annotation procedure was carried out on the predicted protein models, employing both BLAST\(^30\) (-evalue 1e-10) versus the UniProtKB database (http://www.uniprot.org/uniprot/), and InterProScan 5\(^31\) to provide functional analysis. BLAST comparison (-evalue 1e-10) was also
| Organellum | Accession number | GenBank Accession |
|------------|-----------------|------------------|
| C. variabilis | 124,793 | NC_015359.1 |
| A. protothecoides | 84,576 | NC_023775.1 |
| Helicosporidium sp. | 37,454 | NC_001613.1 |
| P. wickerhamii | 55,636 | NC_026099.1 |
| P. zopfi gen. 1 | 38,164 | NC_025413.1 |
| P. zopfi gen. 2 | 39,222 | NC_017841.1 |
| C. variabilis | 124,793 | NC_015359.1 |
| A. protothecoides | 84,576 | NC_023775.1 |
| Helicosporidium sp. | 37,454 | NC_001613.1 |
| P. wickerhamii | 55,636 | NC_026099.1 |
| P. zopfi gen. 1 | 38,164 | NC_025413.1 |
| P. zopfi gen. 2 | 39,222 | NC_017841.1 |

Table 1. Organelles and nuclear genome annotation statistics. For each organism, the nuclear or organelar DNA size, the number of scaffolds in the assembly, the percentage of GC content (%GC), the number of introns and size, the number of genes (Total features), subdivided into: coding sequences (CDS), the number of transfer RNAs (tRNA), and of ribosomal RNAs (rRNA) are reported; last column reports NCBI GenBank Accession number for the assemblies. 1Original WGS sequencing (BioProject PRJNA182710, year: 2014) covers 21,856,191 bp, whereas a more recent sequencing (PRJNA362700, year: 2017) estimates a total length of 32,730,026 bp. 2As reported in51. cPlastid also includes 3 ORFs of unknown function. dP. stagnora plastid also includes 3 ORFs of unknown function.
Results

The genomes of *P. zopfii* genotype 1 and 2 were sequenced using a combination of different approaches (Supplementary Table 1) resulting in 45,166,626 and 66,488,185 total reads, respectively.

Sequence assembly. Sequence assembly yielded a nuclear genome of 26,448,891 and 24,744,895 bp for *P. zopfii* genotype 1 and genotype 2, respectively; organelles were very small: mitochondria being 38,164 and 39,222 bp for genotypes 1 and 2, respectively, whereas plastids being 28,698 and 28,686 bp, for genotype 1 and 2 respectively (Table 1).

Mitochondrial structure and annotation. The mitochondrial genomes of both *P. zopfii* genotype 1 and 2 are sized at about 38–39 Kb (Fig. 1A, Supplementary Fig. 1A). They are extremely compact, with only about 32% of non-coding DNA and are characterized by a substantial loss of any intron-exon structure in their genes. Only *P. zopfii* genotype 2 shows a single intron (length: 777 bp) in the long ribosomal subunit (*lrn*) gene, whereas the other species belonging to the Chlorellales (i.e.: *C. variabilis*, *A. protothecoides*, *Helicosporidium* sp., and *P. wickerhamii*) display a more complex structure, with intron length reaching 4,000–8,000 bp. A putative LAGLIDADG homing endonuclease, a class of restriction enzymes directly involved in the DNA cutting process⁴², is encoded within the intron of *lrn* gene.

The annotation revealed a number of coding genes for both *P. zopfii* species similar to those observed in other Chlorellales species. All the genes encoding for cytochrome units, NADH dehydrogenase, ATP synthases and ribosomal proteins, as well as all the tRNAs and the ribosomal units, have a conserved structure (Table 1, Fig. 2A). As expected, mitochondrial protein similarity between *P. zopfii* genotype 1 and 2 is above 90%, while similarity with *Helicosporidium* sp., *C. variabilis*, *A. protothecoides* and *P. wickerhamii* is around 60%.

Plastid structure and annotation. The structure of the plastid genomes of *P. zopfii* genotype 1 and 2 (Fig. 1B, Supplementary Fig. 1B) is very similar to that of other non-photosynthetic algae belonging to the Trebouxiophyceae class, being extremely compact. The genomes are sized only about 28.7 Kb for both genotypes and are the shortest plastid genomes within their class. Both *P. zopfii* genotypes possess only 19 CDS, and as a result are simpler than those of *Helicosporidium* sp. (26 CDS), *P. wickerhamii*, *P. cutis* (both possessing 40 CDS) and *P. stagnora* (28 CDS). On the other hand, the tRNAs and rRNAs are conserved among all these species and other photosynthetic algae (i.e.: *C. variabilis* and *A. protothecoides*) (Table 1, Fig. 2B). Plastid protein similarity between genotype 1 and 2 is about 85% while similarity with other organisms is around 41–43%, with the best match being with *P. stagnora* (i.e.: 49.9% and 49.5% similarity to *P. zopfii* 1 and 2, respectively) (Fig. 3A). Comparison with other related organisms shows that the plastid genomes of *P. zopfii* genotype 1 and 2, like those of *Helicosporidium* sp., *P. wickerhamii*, *P. cutis* and *P. stagnora*, lack the genes associated with photosynthesis (photosystem I and II, chlorophyll biosynthesis and cytochrome components), whereas ATP synthase genes were

---

**Figure 1.** *P. zopfii* genotype 2 mitochondrion and plastid circular plot. Circular plots depicting the annotation of *P. zopfii* genotype 2 mitochondrion (A) and plastid (B). Gene annotation is reported on the outermost circle of the plot; CDS are in blue, tRNA are in green and rRNA are in red. Innermost circles represent gene orientation, GC content and skew. Other rings report the extent and the % identity of the plastid features with those of proximal organisms (*C. variabilis*, *A. protothecoides*, *Helicosporidium* sp., *P. wickerhamii* plus *P. cutis* and *P. stagnora* for plastid only) and with *P. zopfii* genotype 1. Transparency is proportional to the degree of identity between *P. zopfii* genotype 2 and each reference genome; no transparency indicates 100% identity. % identity was calculated on the basis of BLASTn (for tRNA and rRNA) and BLASTp (for CDS) matches with the corresponding features on the reference plastid genome.
maintained only in *P. wickerhamii* and *P. cutis*. Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (*rbcL*, or RuBiCO) is absent in *P. zopfii*, as well as in all other *Prototheca* spp. and in *Helicosporidium* sp. Moreover, differently to the other Trebouxiophyceae, *P. zopfii* also lacks all plastid-encoded RNA polymerases (i.e.: *rpoA*, *rpoB*, *rpoC1* and *rpoC2*) (Table 2, Supplementary Table 2). Our transcriptome analysis demonstrated that most plastid genes encoding rRNA and proteins were expressed in *P. zopfii* (Supplementary Fig. 2A), although three of them (i.e.: *rps4*, *rps7*, *rpl20*) had low RPKM values, suggesting that nuclear-encoded counterparts could compensate for the loss of plastid-encoded RNA polymerases.

Both *P. zopfii* genotypes plastids contain the genes of most of the ribosomal proteins, with the exception of RPL12, RPL19, RPL23, RPL32, RPS2, RPS18 and RPS9. The same proteins, with the exception of RPL19 and RPL32, are also absent in *P. stagnora*, which, apparently, is the most similar organism. The phylogenetic tree inferred from the super-alignment of 9 shared ribosomal proteins confirms that *Helicosporidium* sp., *P. stagnora* and *P. zopfii* are closely related genera, whereas *P. wickerhamii* is more distant, closer to *A. protothecoides* than to other *Prototheca* species (Fig. 3B).

**Nuclear genome annotation.** From the genome assembly procedure, 6,956 and 4,555 scaffolds with lengths >1 Kb representing the nuclear genome were obtained for *P. zopfii* genotype 1 and 2, respectively, indicating a genome size of about 26.5 Mbp and 24.7 Mbp for genotypes 1 and 2, respectively, a size between those of *Helicosporidium* sp. (12.4 Mbp) and *C. variabilis* (46.2 Mbp) and comparable to the genome of *A. protothecoides* (~23 Mbp) and *P. wickerhamii* (~29 Mbp) (Table 1).

The maximum scaffold length was 97,625 and 57,068 bp, with 1,289 and 1,708 contigs exceeding 5 Kb in length and a N50 value of 6,686 and 7,940 for *P. zopfii* genotypes 1 and 2, respectively.
Augustus gene prediction led to the individuation of 6,884 and 6,381 gene models for the two genotypes. 56.5% and 62.0% of *P. zopfii* genotype 1 gene models were annotated versus UniProtKB and InterPro databases, respectively, while the corresponding percentages for genotype 2 were 59.4% and 67.2%. The main features of the nuclear genome, such as gene number and coding density, were consistent for both genotypes, and similar to their counterparts in *A. protothecoides* genome (i.e.: gene density: 0.26 genes/Kb and 0.32 genes/Kb, for both *P. zopfii* genotypes and *A. protothecoides*, respectively). On the contrary, the average exon and intron sizes were higher compared to those previously observed in related species (Supplementary Table 3). BLAST comparison of the predicted proteins of the two *P. zopfii* genotypes resulted in a set of 6,134 common entities, representing a core set of homologous proteins conserved in the two genotypes. Among the predicted genes and transcripts, we found evidence of the presence and expression of nuclear-encoded polymerases (NEPs) (Supplementary Data 1 and 2). *P. zopfii* genotype 1 and 2 possess 21 and 19 genes annotated as NEPs, respectively, and both showed a RNA-Seq signal indicating active transcription of many of them (Supplementary Fig. 2B,C). Prediction of target peptides highlighted at least one NEP per genotype as a high-confidence candidate for containing a chloroplast transit peptide (cTP) (genes g5108 and g2780 for genotypes 1 and 2, respectively, for which all the four prediction software employed were concordant); moreover, PredAlgo also suggested two more genes (g3914 and g4216 for genotypes 1 and 2, respectively) to be plastid-directed NEPs. In addition to that, we found evidence of some mitochondrial targeting peptides (mTPs) in more gene models (Supplementary Table 4).

Table 2. Comparison of genes of the *C. variabilis* chloroplast with genes of the *P. zopfii* and *P. stagnora* plastids. Genes are grouped according to function. ‘a’Function and names of chloroplast genes follow the categorization and nomenclature reported in52. bGenes present in *P. zopfii*. Genes present in *P. zopfii* genotype 1 and 2 but absent in *P. stagnora* plastid are in boldface. cGenes absent in both *P. zopfii* and *P. stagnora*. Genes present in *P. stagnora* but absent in *P. zopfii* genotype 1 and 2 plastid are in boldface.
Discussion
In this paper, we describe the complete, manually annotated, circular sequence of both mitochondrial and plastid organellar DNA of *P. zopfii* genotype 1 and genotype 2, as well as a first draft of the complete genome, by whole genome shotgun sequencing.

Structure of the mitochondrial genomes of both *P. zopfii* genotypes was revealed to be smaller in size and extremely condensed when compared to that of some related organisms, i.e.: *P. wickerhamii* and *Helicosporidium* sp., but similarly functional, with the size reduction mostly due to the lack of intron-exon structures.

An extremely compact and simplified structure was also observed in *P. zopfii* plastid genomes, which showed a substantially reduced size (about 28.7 Kb), smaller than those of all other algae belonging to the class of Trebouxiophyceae. As previously observed in the genomes of non-photosynthetic algae belonging to this class, *P. zopfii* plastsids lack all the genes for the synthesis of the proteins involved in the photosynthesis process, and for the RubisCO large and small subunits. Fundamental plastid-related functions, however, seem to have been preserved, as indicated by the presence of a RNA-Seq signal on the genes of plastid-encoded ribosomal proteins. The low expression of some of them, however, cannot preclude their presence in the genome as pseudogenes in *P. zopfii*. However, further experiments should be carried on in order to confirm this observation. More interestingly, the entire set of rpo genes (i.e. *rpoA, rpoB, rpoC1* and *rpoC2*), which codify for the plastid-encoded RNA polymerases (PEPs), was lost in *P. zopfii*, an unprecedented observation within this class of algae. Loss of PEPs was previously reported for other non-photosynthetic parasitic plants, such as *Cuscuta obtusiflora* and *Rhizanthella gardneri*, but not in apicomplexans and algae: plastid genomes of *Plasmodium falciparum*, *A. protothecoides*, *P. wickerhamii*, *P. cutis*, *P. stagnora* and *Helicosporidium* sp. have all retained the complete set of rpo genes. We found no evidence of PEP sequences either in plastid assemblies or in the nuclear draft genomes, whereas nuclear genome contigs contained evidence of 21 and 19 RNA-driven, nuclear-encoded polymerases (NEPs) for *P. zopfii* genotype 1 and 2, respectively, and at least one of them, per genotype, was predicted to contain plastid-targeting signal peptides. It is therefore possible that *P. zopfii* codes for other NEPs able to target the plastid, making it possible to transcribe its genetic information.

As previously suggested, considering the degree of similarity between the few structural genes preserved in its plastid genome and the evidence from the phylogenetic analysis, *P. zopfii* seems to be more closely related to *P. stagnora* and to *Helicosporidium* sp., rather than to *P. wickerhamii* or *P. cutis*. Moreover, it is noteworthy that *P. wickerhamii* appears not to be closely related to *P. zopfii*, instead to *A. protothecoides*, strengthening the evidence that *P. wickerhamii* is only loosely related to other *Prototheca* spp, as previously revealed by plastid genome comparison and supporting the proposal of either moving *P. wickerhamii* into Auxenochlorella genus or creating a new genus.

Nuclear genome assemblies of *P. zopfii* genotype 1 and 2 had a size estimated at about 25–26 Mb for both, consistent with that reported in a previous work. Although further studies are certainly needed to elucidate the structure of nuclear genomes of *P. zopfii*, this work adds information to the growing body of genome resources for the plant kingdom, being, although a preliminary draft, the first report of the assembly of nuclear DNA of *P. zopfii*.

In conclusion, we believe that the information reported herein will be important for the understanding of the evolution and genomic organization of *Prototheca* spp., with a particular focus on the progressive loss of functions of plastids in the shift from autotrophic photosynthetic, to obligate, heterotrophic, parasitic algae.

Data Availability
The complete genome sequencing project has been registered in the NCBI BioProject portal (https://www.ncbi.nlm.nih.gov/bioproject/) under the accession number PRJNA388740. Raw DNA sequence reads for *P. zopfii* genotype 1 and 2 have been deposited into the NCBI Short-Read Archive (SRA) under the accession numbers SRR7091517-SRR7091518. Full sequences of mitochondria and plastids are available in GenBank, under accessions MF197533, MF197534, MF197535, and MF197536. The Whole Genome Shotgun project has been deposited (as non-annotated contigs) at DDBJ/ENA/GenBank under the accessions PEIA00000000 and PGFX00000000. The versions described in this paper are PEIA01000000 and PGFX01000000. Sequences annotated as nuclear-encoded polymerases (NEPs) are available as amino acid FASTA files in Supplementary Data 1 and 2.

References
1. Satoh, K., Ooe, K., Nagayama, H. & Makimura, K. *Prototheca cutis* sp. nov., a newly discovered pathogen of protothecosis isolated from inflamed human skin. *Int J Syst Evol Microbiol.* **60**, 1236–40, https://doi.org/10.1099/ijs.0.016402-0 (2010).
2. Rosler, U., Möller, A., Hensel, A., Baumann, D. & Truyen, U. Diversity within the current algal species *Prototheca zopfii*: a proposal for two *Prototheca zopfii* genotypes and description of a novel species, *Prototheca blaschkeae* sp. nov. *Int J Syst Evol Microbiol.* **56**, 1419–25 (2006).
3. Pore, R. S., Barnett, E. A., Barnes, W. C. Jr & Walker, J. D. *Prototheca* ecology. *Mycopathologia*. **81**, 49–62 (1983).
4. Masuda, M., Hirose, N., Ishikawa, T., Ikawa, Y. & Nishimura, K. *Prototheca miyajii* sp. nov., isolated from a patient with systemic protothecosis. *Int J Syst Evol Microbiol.* **66**, 1510–1520, https://doi.org/10.1099/ijsem.0.009911 (2016).
5. Pore, R. S. Selective medium for the isolation of *Prototheca*. *Appl Microbiol.* **26**, 648–9 (1973).
6. Jänschi, S. et al. Review of the microbiological, pathological, and clinical aspects of bovine mastitis caused by the alga *Prototheca zopfii*. *Vet Q.* **23**, 58–61 (2001).
7. Lass-Förl, C. & Mayr, A. Human protothecosis. *Clin Microbiol Rev.* **20**, 230–40 (2007).
8. Nelson, A. M., Neafie, R. C. & Connor, D. H. Cutaneous protothecosis and chlorelliosis, extraordinary “aquatic-borne” algal infections. *Clin Dermatol.* **5**, 76–87 (1987).
9. Shank, A. M., Dubziel, R. D. & Teixeira, L. B. Canine ocular protothecosis: A review of 14 cases. *Vet Ophthalmol.* **18**, 437–42, https://doi.org/10.1111/vop.12239 (2015).
10. Sternen, V. J. et al. *Protothecosis* in 17 Australian dogs and a review of the canine literature. *Med Mycol.* **45**, 249–66 (2007).
11. Ricchi, M. et al. Molecular characterization of *Prototheca* strains isolated from Italian dairy herds. *J Dairy Sci.* **93**, 4625–31, https://doi.org/10.3168/jds.2010-3178 (2010).
 evidence supporting the reclassification of Helicosporidia as green algae (Chlorophyta) and related taxa inferred from SSU and LSU ribosomal DNA partial sequence data. 

 characteristic structure useful for identification and genotyping. 

 identification of 

gene content and genome organization. 

 Complete sequence of the mitochondrial DNA of the chlorophyte alga 

 and 

 plastid genome sequences give insight into the origins of non-photosynthetic algae. 

 Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae 

 has 

 with 

 associated with outbreak of bovine protothecal mastitis. 

 Genome Res. 9. 868–77 (1999). 

 Aligning sequence reads with 

 Basic local alignment search tool. 

 Nucleic Acids Research. 33, W665–W673, https://doi.org/10.1093/nar/gkj459 (2005). 

 InterProScan: protein domains identifier. 

 PloS ONE. 8(1), 940 (2018). 

 A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. 

 and 

 PredAlgo: a new subcellular localization prediction tool dedicated to green algae. 

 Mol Biol Evol. 29(12), 3625–39, https://doi.org/10.1093/molbev/msx178. (2012). 

 Predicting subcellular localization using sequence-biased recurrent networks. 

 Mol. Syst. Biol. 26, 3125–6, https://doi.org/10.1038/mmb.2010.151 (2010). 

 First outbreak of bovine mastitis caused by 

 Short communication: Molecular typing of 

 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. 

 STAR: ultrafast universal RNA-seq aligner. 

 InterProScan: protein domains identifier. 

 An optimized method for high quality DNA extraction from microalga 

 Plasmodium falciparum 

 and 

 Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. 

 59(3), 307–21 (2010). 

 ultrafast universal RNA-seq aligner. 

 29(1), 15–21, https://doi.org/10.1093/bioinformatics/btaa035 (2013). 

 Brefoula et al. 

 26(6), 841–842, https://doi.org/10.1093/bioinformatics/bto333 (2010). 

 DNA polymerase I genes of 

 Helicosporidium. PLoS One. 5(1), e9894 (2010). 

 Helicosporidium sp. is highly 

 BMC Biol. 21, 4:12 (2006). 

 Mycopathologia. 173, 275–81, https://doi.org/10.1007/s10461-011-9510-y (2012). 

 bacterial type II restriction enzyme. 

 PLoS ONE. 10(9.034) (2011). 

 2011-5101 (2012). 

 59(3), 1005–1016 (2000). 

 Bioinformatics. 20, 3525–5 (2004). 

 7119, 14465, https://doi.org/10.1038/srep14465 (2015). 

 Nucleic Acids Research. 33, W116–W120, https://doi.org/10.1093/nar/gkt442 (2005). 

 Bioinformatics. 2279–2286 (2005). 

 Bioinformatics. 26, 455–477, https://doi.org/10.1093/cmb/crb090 (2012). 

 Bioinformatics. 20, 719–725, https://doi.org/10.1093/molbev/msv028 (2011). 

 few months earlier. 

 Bioinformatics. 237, 75–86 (1994). 

 Phylogeny of the non-photosynthetic green micro-algal genus 

 Systematic Biology 59, 15–21, https://doi.org/10.1093/sysbio/syv008 (2010). 

 PLoS ONE. 8(11), e112963, https://doi.org/10.1371/journal.pone.0112963 (2014). 

 Bioinformatics. 220(1), e8954 (2010). 

 F. i. has 

 J Dairy Sci. 95, 4442–6, https://doi.org/10.3168/jds.2010-4192 (2015). 

 149

 B. R. Chloroplast genomes of photosynthetic eukaryotes. 

 135

 Bioinformatics. 220(11), e112963, https://doi.org/10.1371/journal.pone.0112963 (2014). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 Bioinformatics. 220(1), 15–21, https://doi.org/10.1093/sysbio/syv008 (2010). 

 868–77 (1999). 

 Bioinformatics. 4, 981–9 (2005). 

 Bioinformatics. 237, 2238–2266 (2007). 

 10.1093/molbev/msx178. (2012). 

 10.1093/molbev/msv028 (2011). 

 10.1111/j.1365-2461.2008.02345.x (2008). 

 Bioinformatics. 220(1), e8954 (2010). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1093/molbev/msv028 (2011). 

 10.1111/j.1365-2461.2008.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008). 

 10.1111/j.1365-2461.2006.02345.x (2008).
Acknowledgements
We thank Dr. Tomasz Jagielski from the University of Warsaw (Poland) for providing P. zopfii genotype 2 strain. We also thank Mr. Bryan Manchi for his valuable revision of the English text.

Author Contributions
M.S. and B.L. assembled and annotated the genome, conducted analyses and wrote the paper. E.C. prepared P. zopfii genome samples, performed genomic analyses and wrote the paper. R.B. prepared and sequenced P. zopfii samples on GS-FLX platform. M.R. helped during the conception of the study, provided P. zopfii and wrote the paper. M.L. provided P. zopfii and helped draft the manuscript. S.C. sequenced P. zopfii samples on Illumina MiSeq instrument and helped draft the manuscript. B.C. conceived the study and helped draft the manuscript. P.C. conceived the study, helped in preparing P. zopfii genome samples, sequenced P. zopfii samples on Illumina MiSeq instrument and wrote the paper. All authors read and approved the final manuscript.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-32992-0.

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018, corrected publication 2022