Genetic and biological insights from whole-genome sequencing of the pathogenic microalga *Prototheca wickerhamii*

Tomasz Jagielski[1], Zofia Bakuła [1], Agnieszka Gromadka[3], Jan Gawor[2], Robert Gromadka[2], Paweł Siedlecki[3,4]

[1]Department of Applied Microbiology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Poland
[2]DNA Sequencing and Oligonucleotides Synthesis Unit at the Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
[3]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland;
[4]Department of Systems Biology, Institute of Experimental Plant Biology and Biotechnology, University of Warsaw, Poland

The study was financed by the National Science Centre «PRELUDIUM» (2013/09/N/NZ2/00248) and «SONATA» grants (2014/15/D/NZ7/01797).

Background: Among an estimated number of 72,500 algal species and 400,000 plant species worldwide, *Prototheca* are the only known organisms potentially pathogenic for both humans and animals. Of the eight currently recognized *Prototheca* species, four have been implicated in human infections, namely *P. wickerhamii*, *P. zopfii*, *P. cutis*, and *P. miyajii* with the former being responsible for the bulk of cases. The *Prototheca* algae and protothecosis in general have been much neglected areas of research. Importantly, sequencing of the entire chromosomal DNA of any *Prototheca* species has never been attempted before.
The aim of this study was to perform whole genome sequencing (WGS) of *P. wickerhamii* using Illumina (Illumina, USA) and PacBio (Pacific Biosciences, USA) technologies combined to provide a comprehensive picture of the genetic make-up, pathophysiology, and evolution of this species.

Methods: The type strain of *Prototheca wickerhamii* (ATCC 16529) was used in the study. Genomic DNA was isolated by using a newly designed DNA extraction method, essentially described elsewhere (Jagielski et al., Plant Methods. 2017;13:77)**.** WGS was performed with the Illumina MiSeq (Illumina, San Diego, USA) and PacBio (Pacific Biosciences, Menlo Park, USA) platforms. Genome completeness was assessed using QUAST, BUSCO, and CEGMA. Gene prediction and annotation was performed using the MAKER annotation pipeline. Repetitive elements were defined by RepeatMasker and RepBase. Sequence similarity was assessed using BLAST with E-value < 1e-10.
RNA for transcriptome analysis was isolated using the A&A Biotechnology (Poland) Total RNA Mini Plus kit, while sequencing libraries were generated with the KAPA Stranded RNA Seq kit (KAPA-Roche, Switzerland). Genetic content of *P. wickerhamii* was analyzed along with the genomes, publicly accessible via NCBI, of two chlorellacean species, closely related to *Prototheca*, namely *Auxenochlorella protothecoide*s (accession no. GCA_000733215.1) and *Helicosporidium* sp. (GCA_000690575.1).

Results: The contigs were assembled into 79 scaffolds spanning 32 Mbp, with a separate scaffold for mitochondrial and plastid DNA, resulting in a genome larger than that of *A. protothecoides* (22.9 Mbp) and *Helicosporidium* sp. (12.4 Mbp). The *P. wickerhamii* genome had a GC-rich genomic composition (64.4%), higher than in both *A. protothecoides* (63.5%) and *Helicosporidium* sp. (61.7%). Over 9,300 protein-encoding genes were predicted in *P. wickerhamii*, a number significantly higher than in *Helicosporidium* sp. (7,016) and *A. protothecoides* (6,033). Interspersed repeats and low complexity DNA sequences were shown to comprise 2.8% of the *P. wickerhamii* genome, with simple repeats and low complexity regions accounting for 2.1% and 0.2% of the genome, respectively. Small RNA sequences were detected in 0.3% of the genome and the most abundant transposable elements were the long terminal repeats (LTRs) Gypsy and Copia.
The *P. wickerhamii* genome was shown to encode more metalloproteases and protease inhibitors when compared with *Auxenochlorella* and *Helicosporidium*. A number of genes coding for proteins, originating from staphylococci, streptococci, *Candida albicans*, and *Cryptococcus neoformans*, plausibly involved in the pathogen–host interactions were disclosed.

Conclusions: This study provides a first insight into the genome of *P. wickerhamii.* A precise determination of the gene/protein functions and delineation of metabolic pathways is still in progress.

Keywords: *Prototheca wickerhamii*, genome, whole-genome sequencing, algae