Title: *Prototheca wickerhamii* genome sequencing project – a preliminary report

The study was financed by the National Science Centre grant «PRELUDIUM» (2013/09/N/NZ2/00248).

Authors: Bakuła Z. [1], Siedlecki P.[2], Gawor J.[3], Gromadka R.[3], Jagielski T.[1]

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

Background: *Prototheca* is a genus of aerobic, unicellular, colorless, yeast-like algae widely distributed in the environment. Although normally saprophytic, these organisms may, under certain conditions produce infections in humans and different animal species. *Prototheca* algae are thus the only known plants with pathogenic ability for humans and animals. Of the six currently recognized *Prototheca* species, three have been implicated in human disease: *P. wickerhamii, P. zopfii*, and P. *cutis*. The aim of the project is to perform a preliminary sequencing analysis of the whole genome of *P. wickerhamii,* a major etiological agent of human protothecosis.

Material/methods: The strain used in our study is *P. wickerhamii* PL1, originally isolated from the first case of human protothecosis in Poland [Żak I, Jagielski T, Kwiatkowski S, Bielecki J. *Prototheca wickerhamii* as a cause of neuroinfection in a child with congenital hydrocephalus. First case of human protothecosis in Poland. Diagn. Microbiol. Infect. Dis. 2012; 74:186-189]. The preparation of the strain for the purpose of genome sequencing involved its revitalization on culture media and large-scale nuclear DNA extraction using in-house isolation method specially modified for disruption of *P. wickerhamii* cell walls. The sequencing was performed on next generation sequencing instrument MiSeq (Illumina). The long paired end reads from the MiSeq were used to make contigs and the mate pair reads were used for ordering contigs into large scaffolds. The completeness and contiguity of the assembly was estimated. Two versions of *Prototheca wickerhamii* genome were screened for regions to be masked (v3 and v4). In order to ensure a reference free initial detection, inverted repeat finder (IRF) and RepeatModeler were used to generate transposon candidates *de novo*. Since IRF & RepeatModeler produce multiple overlapping hits, CD-HIT was utilized for sequence clustering with similarity threshold set at 100% and query coverage set at 99% of the shorter sequence. Pfam & CDD protein domain profiles were used to elucidate motifs typical for transposons. The above procedure resulted in a custom RepeatMasker library construction. With this tool coordinates of detected Transposable Elements and coordinates of detected simple repeats were estimated.

Results: Genome sequencing yielded 2,860 scaffolds with the total length of 29 Mbps. The size of the genome was estimated at about 35 Mbps. The study shows that *P. wickerhamii* is characterized by low abundance of transposable elements (TE). Using a custom, curated database of TE, developed in-house, we found 20 LTR_Gypsy elements (LTR, long terminal repeats) and 1 DIRS-Ngaro element in v3 genome results (24 LTR_Gypsy and 1 DIRS-Ngaro for v4 genome). Sequencing data showed sporadic occurrence of simple repeats (SR) in the *P. wickerhamii* genome (v3 has 52, v4 has 62). Most of them were of size ranging between 100 and 300 bp.

Conclusions: The analyses so far performed within the project provided preliminary information about the overall genome organization of *P. wickerhamii.* In the next step, gene content (with special focus on virulence genes) and metabolic capacities of the pathogen will be investigated.