The first complete genome sequence of *Streptococcus dysgalatiae* subsp. *dysgalactiae* an emerging fish pathogen

Alexandra A. Urrutia Zegarra, Felipe L. Pereira, Fernanda A. Dorella, Alex F. Carvalho, Gustavo Morais Barony, Carlos A. G. Leal, and Henrique C. P. Figueiredo

Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

Streptococcus dysgalatiae subsp. *dysgalactiae* (SDD) is a Gram-positive cocci, it autoaggregates in saline, forms long chains in growth medium, it is catalase negative and α-hemolytic on blood agar. In 2002, it caused the first outbreak in southern Japanese farms. During the subsequent years fish farms in the country suffered huge losses. In Brazil, outbreaks of streptococcosis are common in the freshwater fish species Nile tilapia, *Oreochromis niloticus* (L.). In 2007, the first disease outbreak caused by SDD was spotted in Ceará state. The disease has spread worldwide and despite its increasing clinical and economic significance up until the moment, none SDD genome was fully sequenced. Therefore, considering the importance of a complete genome to characterize this fish pathogen strategy, a next-generation sequence genome initiative was managed. To obtain the SDD genome the sample was isolated from an overnight culture with the Maxwell 16 tissue DNA purification kit using the Maxwell 16 system (both from Promega, USA). A first run was conducted on the Ion Torrent PGM™ sequencing system (Life Technologies, USA) using a 200bp (~ 300- fold coverage) fragment library kit. However, as it resulted in an overly fragmented assembly, another runs were performed using a 400bp (~870-fold coverage) fragment library kit and a 400bp (~ 107 fold coverage) mate-pair kit with an insert of 6kbp. Additional runs were conducted on the Illumina® MiSEQ sequencing system using paired-end 2x150bp (~638-fold coverage) and mate-pair (~658-fold coverage), with an insert of 6kbp. Yet, as no improvements were reached in the assembly fragmentation matter an optical map was acquired. The sequences were assembled with SPAdes 3.8.0, and Newbler 2.9 software, the assembly with higher N50 was selected and aligned with the Optical Map (OpGen Inc, USA) in order to verify the orientation and start scaffolding. Additionally, CONTIGuator software and the assembly_graph text file from the assembly output were used for further scaffold construction. Initially 167 contigs were obtained with an N50 value of 26,993bp and the largest contig with a 141,256bp length size and a ~44% of whole genome map (WGM) coverage. The first scaffolds constructed were used as input in a new assembly, this strategy lead to a better N50 (28,066bp) and fewer contigs (148). The procedure was repeated and ~52% of WGM coverage was reached. Currently, 84% coverage of the WGM was reached and gap filling with CLC Genomics Workbench 7 (Qiagen, USA) still in process. The present study empowers the use of optical mapping as a tool in the assembly of highly repetitive genomes. Further results as the first SDD complete genome announcement are expected.

This study was supported by the MAPA; FAPEMIG; CNPq and the INCT.