Detection and correction mis-assemblies in genome of *Corynebacterium pseudotuberculosis*

Thiago de Jesus Sousa[1], Doglas Parise[1], Diego César Batista Mariano[2], Daniela Arruda Costa[1], Felipe Luiz Pereira[3], Henrique César Pereira Figueiredo[3], Artur Silva[4], Rommel Thiago Jucá Ramos[4], Vasco Azevedo[1].

[1]Laboratory of Cellular and Molecular Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais;
[2] Laboratory of Bioinformatics and Systems, Department of Computer Science, Federal University of Minas Gerais;
[3] National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais;
[4]Institute of Biological Sciences, Federal University of Pará.

Corynebacterium pseudotuberculosis is a bacterium that belongs to CMNR group, which includes the genus *Corynebacterium, Mycobacterium, Nocardia* and *Rhodococcus*. This specie is the etiologic agent of Causeous Lymphadenitis (CLA) in sheep and goats (*C. pseudotuberculosis* biovar *ovis*); and Ulcerative lymphangits in horses, cattles, buffaloes and camels (*C. pseudotuberculosis* biovar *equi*). The first genome sequenced of this organism was *C. pseudotuberculosis* strain 1002, in 2006, which was deposited in the National Center for Biotechnology Information (NCBI) in 2009. Currently, 43 strains has its genomes available in the NCBI database, in which previous studies have showed that these genomes may contain assembly errors. These findings have been allowed by the evolution of next generation sequencing platforms, which provides high precision and reduced cost data. In addition, through restriction enzymes, the use of genome optical mapping enabled improvements in data assembly. Moreover, high precision contigs sorting and high accuracy data have allowed the detection of large genomics rearrangements. This work aimed to update *C. pseudotuberculosis* strain I19 (CpI19) and *C. pseudotuberculosis* strain 162 (Cp162) genomes. For this purpose, a new sequencing of these strains were done using a 400 pb fragment library in Ion Torrent PGM™ platform. The new CpI19 and Cp162 sequencing generated respectively 376,308,624 bp with coverage 160.98-fold and 473,348,503 pb and coverage 200.09-fold. The scaffolding process was performed using MapSolver software to contigs sorting, which utilizes a strategy based on restriction maps constructed upon known restriction sites recognized by Kpnl enzyme *in vitro* by OPGEN,Inc. (Gaithersburg,USA). After the assembling process, a genomic inversion of 1.22MB in CpI19 and 0,85MB in Cp162 were identified. These results showed also a reduction of 146 bp in CpI19 and an addition of 72.215 pb in Cp162. A combined analysis utilizing optical mapping and sequencing data enabled the detection of assembly errors, and genomics inversions as well as genome size inconsistencies in the two previously deposited genomes, showing the optical mapping efficiency

data in revealing mis-assembled data and genomics reallocations.