Assembly, annotation and comparison of *Corynebacterium pseudotuberculosis* lineages

Doglas Parise[1], Thiago de Jesus Sousa[1], Mariana Teixeira Dornelles Parise[1], Adrian Valentín Muñoz Bucio[2], Felipe Luiz Pereira[1], Fernanda Alves Dorella[1], Efrén Díaz Aparicio[2], Henrique Figueiredo[1], Daniela Arruda Costa[1], Vasco Ariston de Carvalho Azevedo[1]

Federal University of Minas Gerais[1], National Autonomous University of Mexico[2]

Corynebacterium pseudotuberculosis (*Cp*) is a pathogenic bacterium that belongs to CMNR group (*Corynebacterium, Mycobacterium, Nocardia e Rhodococcus*). This group presents high CG content (46 – 74%) and cell wall composed of peptidoglycan, arabinogalactan and mycolic acids. Such bacterium is the etiological agent of caseous lymphadenitis in small ruminants and can affect other mammals as horses, buffaloes, camels and even humans. This work aims to characterize six *Cp* linages of both biovars (*ovis* and *equi)*, isolated from Mexico. A key feature concerning those strains is that it is the first time *Cp* biovar *equi* is isolated from Mexico and any *Cp* isolated from this country is sequenced. The lineages *Cp* MEX1 and *Cp* MEX9 were isolated from goats, *Cp* MEX25 and *Cp* MEX29 were isolated from sheep and *Cp* MEX30 and *Cp* MEX31 were isolated from horses. The sequencing was performed in AQUACEN (UFMG) laboratory in Ion Torrent platform with a 400 base pairs (bp) fragment library kit. It generated an amount of data varying from 210,064,890 to 316,695,111 bp with coverage varying from 88.75-fold to 135.48-fold and medium phred quality varying from either 23 or 27 to each sequencing. The assembly methodology was based on hybrid strategy, performing *ab initio* assembly and contigs alignment through reference. Newbler 2.9, Mira 3.9 and SPAdes 3.6.0 assemblers were used and the best assembly was chosen considering the following criteria: number of contigs, genome size, maximum and minimum contig size, and N50. The selected assemblies varied from six to 33 contigs and were aligned against a reference genome utilizing CONTIGuator software. To obtain complete genomes the remaining gaps were filled using CLC software. Structural and functional annotation was performed in the next step, through automatic annotation and manual curation. The first was performed using RAST pipeline and an *in house* script to transfer the annotation of a previously curated genome, after this process both outputs were merged and manually curated. Until this moment, all strains except *Cp* MEX1 have already been annotated and curated. Such genomes presented ~2.3 mega bases of size, GC content of ~52%, ~2000 CDs, 1-77 pseudogenes, 12 rRNAs and 48-49 tRNAs. The linages *Cp* MEX9 and *Cp* MEX25 are available in NCBI, with the respective accession numbers: NZ_CP014543.1 and NZ_CP013697.1. As perspectives of this work, it is intended to finish the annotation and manual curation of *Cp* MEX1; deposit all strains and perform a comparative genomics study with them.