De novo assembly of *Trypanosoma cruzi* strain CL Brener transcriptome

Eddie Luidy Imada[1], Mainá Bitar[2], Maíra Ribeiro Rodrigues[1], Daniela Ferreira Chame[1], Helaine Grazielle dos Santos Vieira[3], Michele Araújo Pereira[1], André Martins Reis[1], Dominik Kaczorowski[3], Willian Santos Prado[1], Andréa M. Macedo[1], Carlos R. Machado[1], Martin A. Smith[3], Glória R. Franco[1]

[1]Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte – Brazil, [2]QIMR Berghofer Medical Research Institute, Brisbane – Australia,[3] Garvan Institute of Medical Research, Sydney – Australia

Chagas disease is a neglected tropical disease caused by *Trypanosoma cruzi* that is estimated to affect at least 6 million people worldwide. The first genome draft of the hybrid clone CL Brener was published in 2005 as several scaffolds and contigs, which were latter partially assembled into 82 chromosomes in mid 2009. Despite the great improvement of the *T. cruzi* genome assembly it still presented many gaps and unplaced contigs. Another caveat of the current genome is that its current annotation are almost exclusively based on automatic ORF detection, which might skew transcriptome analysis due to the lack of unannotated genomic elements. To address these problems we have assembled a *de novo* transcriptome of *T. cruzi* CL Brener using ~70 million paired end Illumina Hiseq 2500 reads with Trinity. The assembly resulted in 57,084 transcripts that were clustered using CD-HIT-EST into 24,844 non-redundant clusters. The clusters representatives (herein refered simply as transcripts) were annotated using a 2-steps methodology by first searching for homologues with BLASTN at nucleotide level against current predicted CL Brener transcripts and then searching at protein level with BLASTX against the TriTrypDB for those failing to align at nucleotide level. As a result, 93% of the transcripts were assigned to 14883 annotations, with 6038 genes being completely covered and 8,665 genes assembled by at least half of its predict length. Transcripts expression evaluation with Kallisto and GO annotation showed that the most expressed genes were related to basic metabolism as expected since steady epimastigote cultures were used in this work. Furthermore, using the assembled transcriptome we were able to close over 400 gaps in the current genome, improve/create UTR annotations for 9206 genes and correct some misplaced contigs in the assembled chromosomes. Our approach shows how de novo transcriptomes can be leveraged not only to functional studies, but to improve genome assemblies and annotations as well.

Financial Support: CAPES, FAPEMIG