Identifying gene clusters in the genome of *Trypanosoma cruzi*

Prado, W. S.; Reis, A. L. M.; Castro, T. B. R.; Franco, G. R.;

Universidade Federal de Minas Gerais

The World Health Organization still considers American Trypanosomiasis, Chagas Disease, as a neglected tropical disease. The causative agent, *Trypanosoma cruzi*, has a complex digenetic life cycle and diverges significantly from other eukaryotes regarding transcriptional regulation and genomic organization. Similar to prokaryotes, transcription is polycistronic in this parasite; however, translation depends on monocistronic mRNAs, which are processed individually from the polycistronic transcript by coupling two reactions: the spliced leader trans-splicing at the 5'-end and polyadenylation at the 3'-end. Only a few promoters have been described for RNA Polymerase II and the boundaries of polycistrons are yet to be defined in this parasite. Some studies in other organisms use the Pearson's correlation coefficient to verify expression patterns and identify networks of genes spatially close in the genome. Nevertheless, such approach has not yet been applied to *T. cruzi*. Based on the assumption that most mRNAs from a commom polycistronic transcript should have a related expression, the main goal of this study is to identify gene clusters, as an attempt to rebuild polycistrons, even partially. Thus, we developed a Python script that integrates the genome annotation to RNA-seq data originated from epimastigotes of the CL Brener strain exposed and not exposed to 500 Gy of gamma radiation. In the experimental group, total RNA was extracted 4, 24 and 96 hours after exposure to ionizing radiation and all samples had two biological replicates. The program uses a sliding window without a fixed size to compare the expression of adjacent genes in the chromosome. If the genes are correlated, then they are assigned to the same cluster. Using this approach, we found 1,859 gene clusters in the Esmeraldo-like haplotype of the CL Brener strain of *T. cruzi*, which covers 5,448 out of the 10,597 genes annotated. Most clusters are dicistrons (58, 7%) or tricistrons (21%). The maximum size of a cluster was 19 genes. Several clusters in the range of 5 to 10 genes consisted of surface proteins (mucins, trans-sialidases and MASPs), which are involved in the evasion of the host immune system by the parasite. If virtually all protein-coding genes are polycistronically transcribed, finding 51.4% of them in clusters seems to be an underrepresentation. Therefore, these findings support the hypothesis that regulation of gene expression in *T. cruzi* occurs post-transcriptionally.

Financial support: CNPq, CAPES, FAPEMIG, UFMG