Modular variability of multigene families encoding surface proteins uncovers differential composition of motifs among Trypanosoma cruzi strains

João Luís Reis-Cunha[1], Gabriela F. Rodrigues-Luiz[1], Rodrigo P. Baptista[2], Laila Viana de Almeida[1], Mariana Santos Cardoso[1], Gustavo Coutinho Cerqueira[3], Daniella C. Bartholomeu[1]

1-Universidade Federal de Minas Gerais, 2-The University of Georgia, 3-Broad Institute.

Among the Tritryps, T. cruzi owns the largest expansion of multigene families encoding surface proteins. Despite playing crucial role in host-parasite interactions, one third of these gene families were not incorporated into the 41 putative chromosomes in the T.cruzi reference strain CL Brener. The large number of members of these families also hinders the assignment of reads to a specific gene, as they can map/align with the same reliability to several loci. Although these families are highly polymorphic, they also present motifs shared among distinct members, resulting in a mosaic structure that may favor the generation of sequence variability by rearrangement of defined blocks through recombination. The relative abundance of these conserved motifs can be used to estimate the variability of these regions among T.cruzi strains. To this end, we developed a methodology to evaluate the copy number variation of motifs derived from mucin-associated surface protein (MASP), TcMUC mucins and trans-sialidases multigene families. This methodology is assembly independent and only requires next generation sequence reads for a given isolate and a reference genome. The first step of this methodology consists in retrieving all reads that map with all the genes of each family, generating all possible kmers of 30 nucleotides present in these reads. The kmers are then clustered by sequence similarity to generate conserved motifs. Finally, the deep of coverage of each motif is computed and compared among T. cruzi strains. Our methodology was used to estimate the relative abundance of all motifs identified in MASP, mucin and trans-sialidase families in different T. cruzi DTUs, revealing several differences in their abundance within and among DTUs. Dendrograms based on the abundance of these motifs presented discordances with the phylogeny based on single copy genes, reinforcing the hypothesis that different selective pressures shape the evolution of these two T. cruzi genomic regions.