Accessing the variability of multicopy genes in complex genomes using unassembled short reads: the case of *Trypanosoma cruzi* multigene families

João Luís Reis-Cunha[1,2], Anderson Coqueiro-dos-Santos[1], Samuel Alexandre Pimenta Carvalho[1], Larissa Pinheiro Marques[1], Gabriela F. Rodrigues-Luiz[3], Rodrigo P. Baptista[4], Laila Viana de Almeida[1], Nathan Ravi Medeiros Honorato[1], Francisco Pereira Lobo[5], Vanessa Gomes Fraga[1], Lucia Maria da Cunha Galvão[1,6], Lilian Lacerda Bueno[1], Ricardo Toshio Fujiwara[1], Mariana Santos Cardoso[1], Gustavo Coutinho Cerqueira[7], Daniella C. Bartholomeu[1]

Abstract

Multicopy genes and other repetitive elements cause assembly fragmentation in complex eukaryotic genomes, limiting the study of their variability. The genome of *Trypanosoma cruzi*, the protozoan parasite that causes Chagas disease, has a high repetitive content, which consist of multigene families, transposable elements, tandem repeats, and satellite sequences. Although many *T. cruzi* multigene families encode surface proteins that play pivotal roles in host-parasite interactions, their variability is currently underestimated, as their high repetitive content results in collapsed gene variants, even in current long-read assemblies. Also, there are few studies comparing multigene family's variability among Discrete Typing Units (DTUs), which are usually performed at the level of assembled genomes, using a limited number of strains. To estimate sequence variability and copy number variation of multigene-repetitive families, we have developed a whole-genome-sequencing read-based approach that is independent of gene-specific mapping and *de novo* assembly. Reads from each parasite isolate are mapped in a reference containing genomic sequences from representative strains, and reads that map to any given gene of a family of interest are

recovered and fragmented in 30 nucleotide long k-mers. These k-mers are clustered based on sequence similarity to reduce redundancy. Finally, sums of counts of all k-mers in each cluster are assumed as the cluster copy number. This methodology was used to estimate the copy number and variability of MASP, TcMUC and Trans-Sialidase (TS), the three largest *T. cruzi* multigene families, in 36 *T. cruzi* strains, including members of all six parasite DTUs. This analysis has shown that *T. cruzi* multigene families present a specific pattern of variability and copy number among the distinct parasite DTUs. TcI isolates had the lowest, while hybrid strains present the highest sequence variability, suggesting that maintaining a larger content of their members after hybridization could be advantageous. There were differences observed between the hybrid strains CL Brener and Tulahuen, which suggests that they could have resolved the hybridization differently. The three evaluated multigene families vary in antigenicity in murine model, where the antibody response to MASP and TS had respectively the highest and lowest diversification with chronification. The reactivity of sera from chronic Chagasic human patients was focused on TS antigens, suggesting that targeting TS conserved sequences could be a potential avenue to improve diagnosis and vaccine design against Chagas disease. Finally, the proposed approach can be applied to study multicopy genes in any organism, providing new possibilities to access sequence variability in complex genomes.

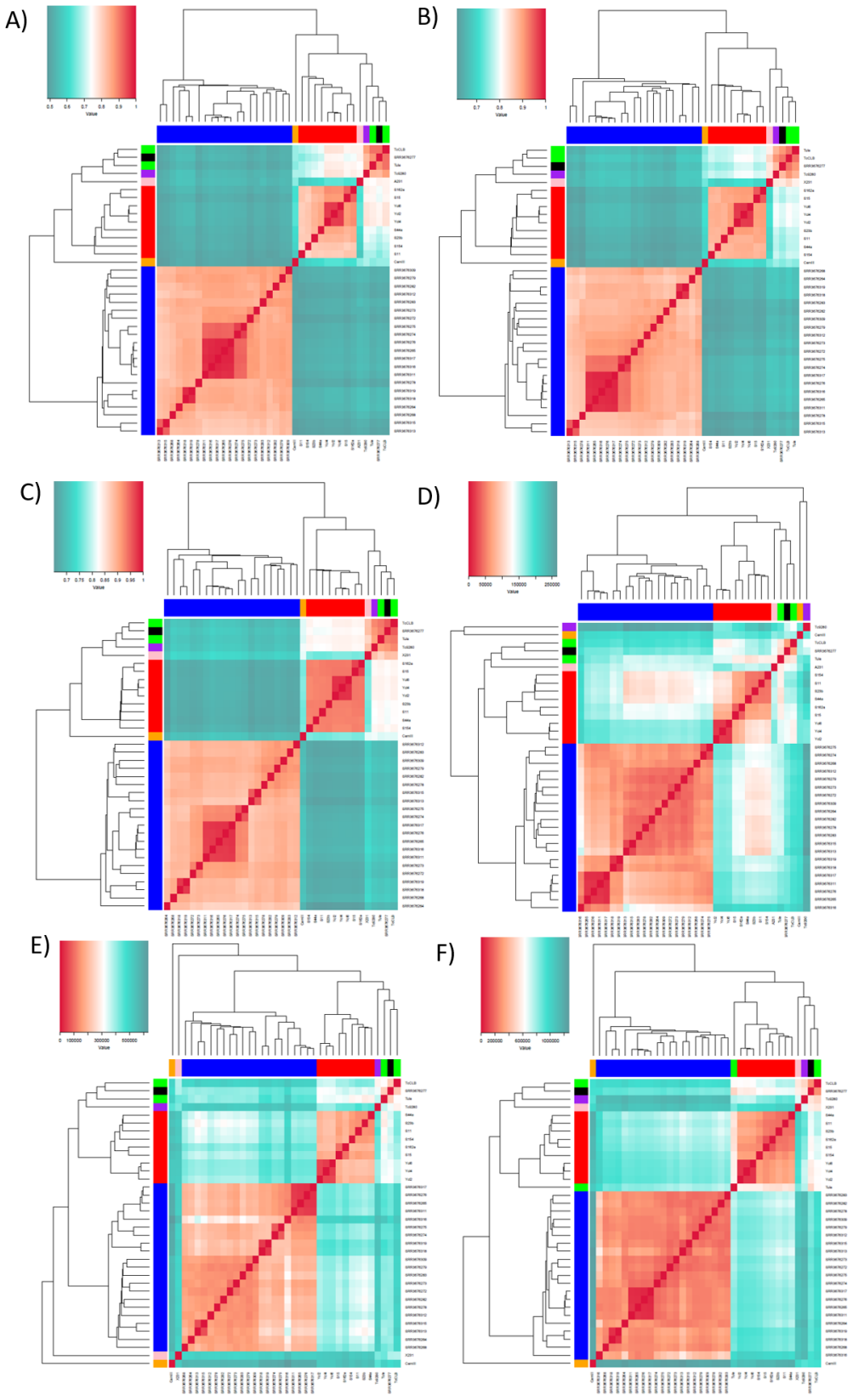

Fig 1: Heatmap of the cluster variability and copy number among *T. cruzi* isolates.

Cluster variability estimated by Jaccard Coefficient (JC) based on the presence/absence of

clusters for each multigene family: **A)** TcMUC, **B)** MASP and **C)** TS. JC values are represented in a scale from green (low), white (medium) to red (high) similarity. Cluster copy number variability estimated by Manhattan Distance for each multigene family: **D)** TcMUC, **E)** MASP and **F)** TS. Manhattan distance values are represented in a scale from green (high), white (medium) to red (low) distances. In this image, each line and column correspond to a *T. cruzi* isolate. The DTU of each isolate is represented by colored lateral strips, where blue, red, pink, orange, purple, green and black correspond to, respectively, TcI, TcII, TcIII, TcIV, TcV, TcVI and 6277 (unknown DTU). Lateral dendrograms were generated by UPGMA clustering.

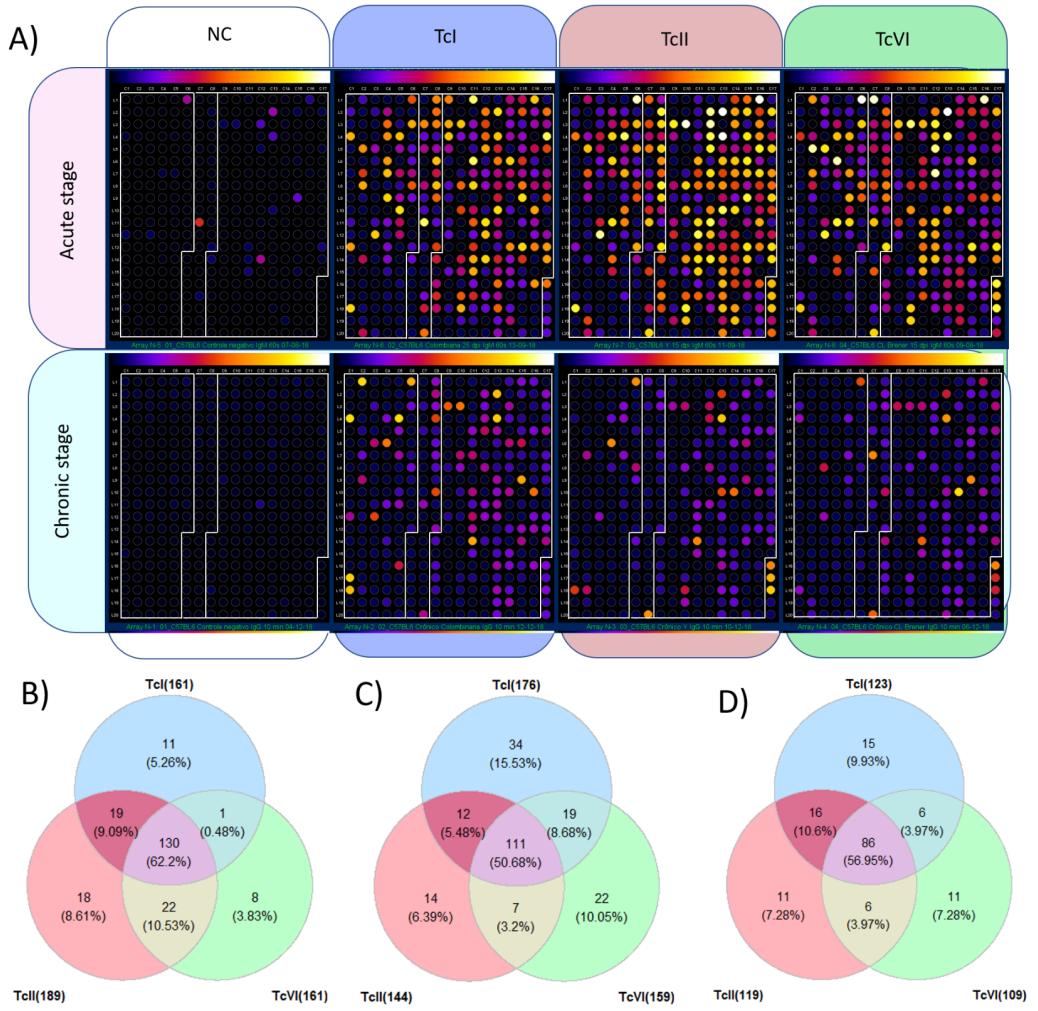

Fig 2: Antigenicity of peptides derived from the multigene families using sera of mice infected with different *T. cruzi* DTUs. A) In this image, each dot corresponds to a peptide,

and the white boxes in each panel separate the peptides from the MASP (left) TcMUC (middle) and TS (right) multigene families. The reactivity of each peptide is represented in a scale from black (low reactivity), orange (median reactivity) to white (high reactivity). The panels representing the reactivity of the sera from mice in the acute phase are circumvented horizontally by a pink box, while the ones representing the sera from mice in the chronic phase are by a cyan box. The panels vertically circumvented by white, blue, salmon and green boxes represent, respectively, the reactivity from the peptides to the sera of: non-infected mice (NC), or mice infected with TcI, TcII or TcVI strains. Venn diagrams representing the number of peptides with above cutoff reactivity for the pool of sera collected during the acute **B)**, chronic **C)** or both acute and chronic **D)** phase of infection. Percentage values correspond to the fraction of the reactive peptides that were observed in each quadrant.