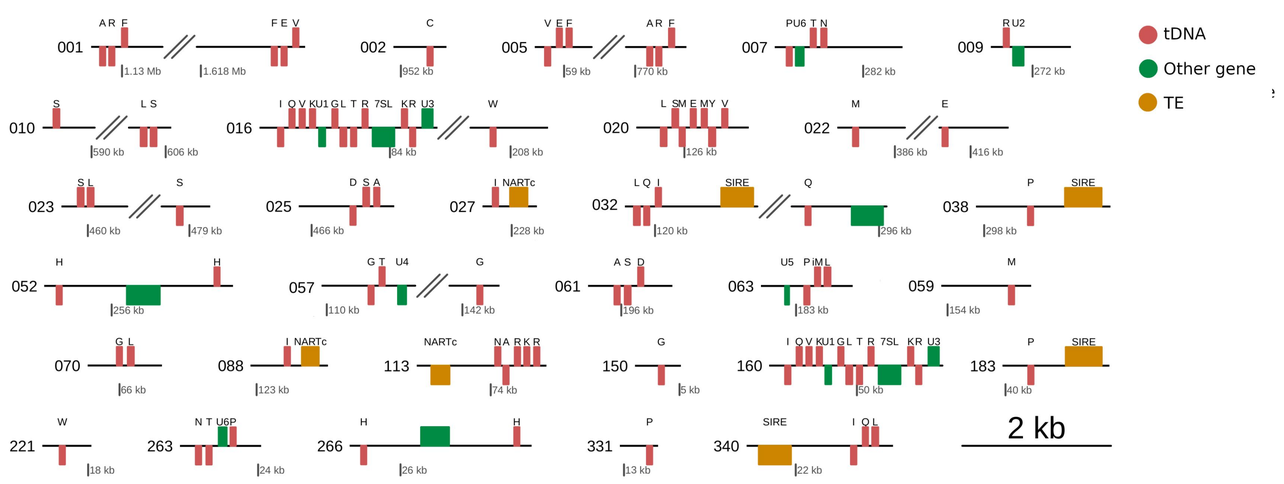
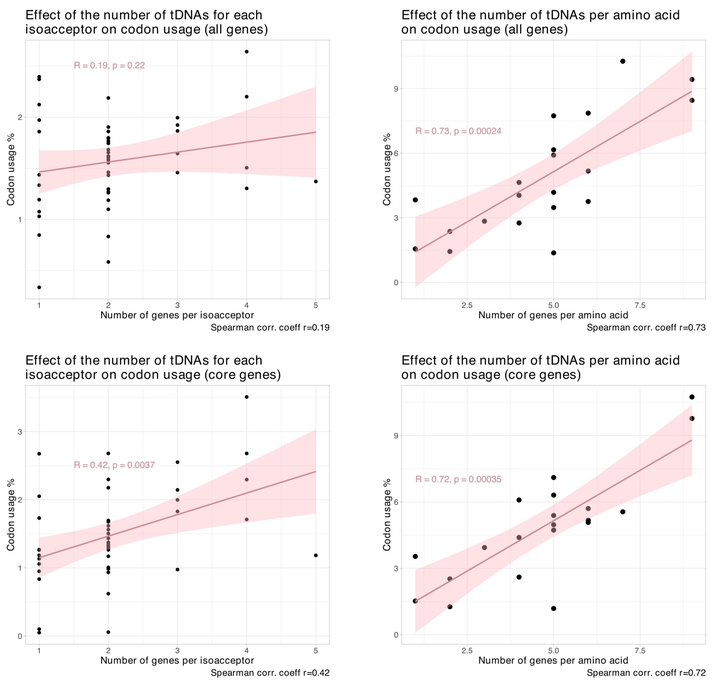
Revisiting the RNA world of trypanosomatids: Neglected genes from neglected parasites

Florencia Díaz- Viraqué[1], Carlos Robello[2]

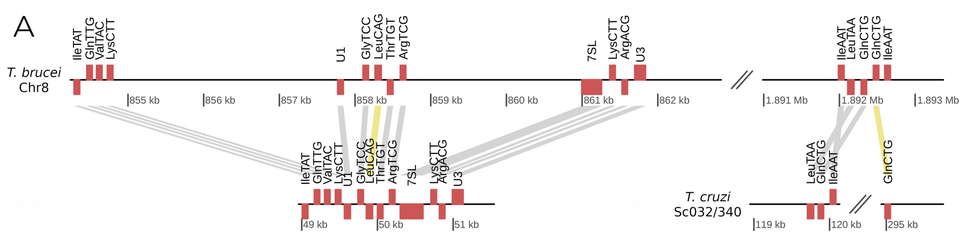
Trypanosomatids are unicellular eukaryotes that differ from the rest of the eukaryotes in several aspects regarding RNA metabolism, gene expression regulation and gene organization. Recently, with the advent of long-read technologies (PacBio and Oxford Nanopore Technologies), several trypanosomatid genome assemblies were published using third-generation sequencing technologies which improves genome sequencing contiguity. However, non-coding RNA annotation has not been thoroughly assessed. These RNA genes encode functional RNA products and they are often a neglected class of genes in large scale genome analysis probably due to their sequence and structure diversity that require more dedicated annotation. Since these genes do not present the features that define coding genes (e.g. long open reading frames) and instead present limited sequence conservation, classical strategies for gene annotation can not be used. We used several optimized algorithms depending on the RNA to re-annotate them providing a complete annotation, including the identification of previously undescribed non-coding RNAs as well as the correct annotation of genes that were previously incorrectly assigned. In sum, this work reports a highly curated genome annotation, and unveils the organization of non-coding RNAs in trypanosomatid genome assemblies.



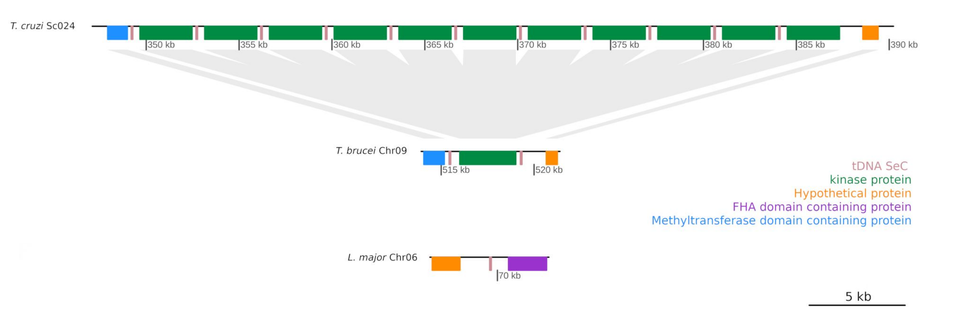
Genomic organization of tDNAs in T. cruzi. tRNA coding genes annotated using tRNA-scan SE were identified in 40 different locus of *T. cruzi* genome. 67% of the tDNAs are organized in clusters and in linear genome association with other Pol III transcribed genes. The association of tDNA with TE had not been previously described in trypanosomatid genomes.



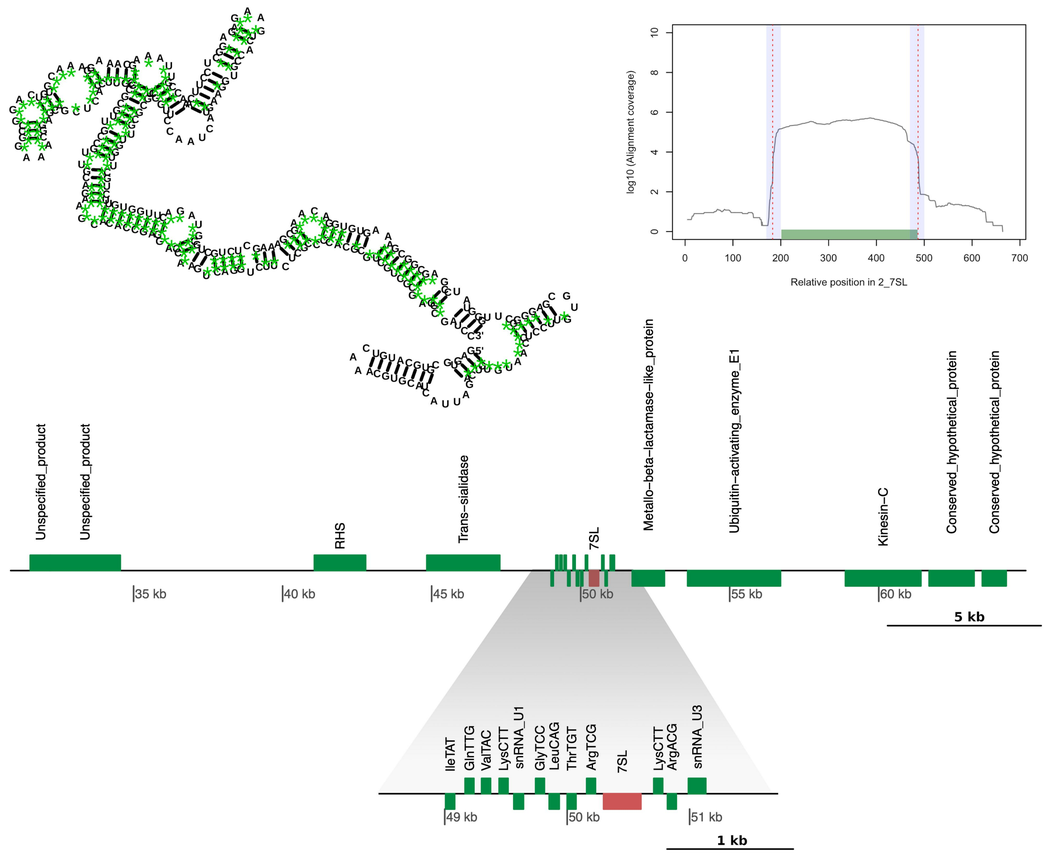
Correlation of codon usage with the number of tDNAs. The codon usage was calculated using all the annotated protein coding genes (upper panels) or eliminating genes from the disruptive compartment of the genome (lower panels). Spearman's correlation coefficient was used to measure the relationship of the variables. Relation between codon usage and the number of tRNA genes for each isoacceptor increase from ρ=0.19 to ρ=0.42 (left panel) meanwhile the corelation between codon usage and the number of tRNA genes per amino acid is mantained with the elimination of disruptives genes (right panel).



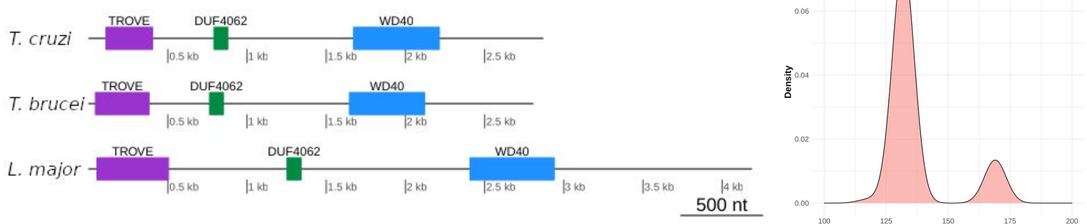
Syntenically conserved tDNAs between T. cruzi and T. brucei. Grey lines connect the orthologous genes in *T. brucei* chromosomes and *T. cruzi* scaffold. In yellow are linked genes that present an inverted relative position among the genes that compose the tDNA locus.

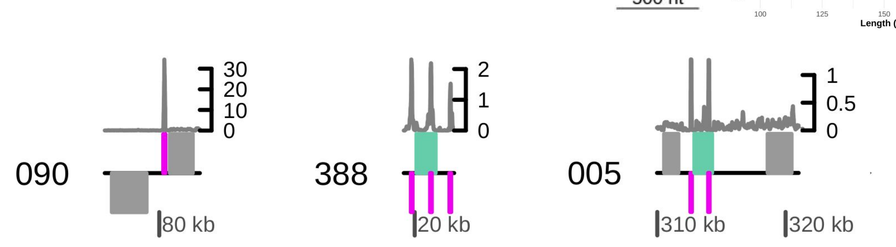


Comparison of tDNA^{Sec} locus in T. cruzi, T. brucei and L. major. The tRNA genes in *T. cruzi* and *T. brucei* are interspersed by kinases and this array is preceded by a hypothetical protein and followed by a Methyltransferase domain containing protein in both parasites. These genes are homologous and present 44%, 21% and 55% of identity, respectively. On the other hand, the two genes at both sides of tDNA^{Sec} in *L. major* are different.



Identification of 7SL RNA gene in T. cruzi. A covariance model constructed with 7SL RNA sequences from *T. brucei* and *T. congolense* was used to search for the gene in *T. cruzi*. It is shown 7SL RNA secondary structure of *T. cruzi*, the RNA expression and genome organization.





RNA and proteins from vault particle were annotated in T. cruzi genome. Interestingly, looking for the RNA, we identify several hits with two different lengths. The hits are clustered in three different groups that are expressed. One of these groups seems to be the classical vtRNA wile the others (shorter hits) are associated in the genome with a transposable element.

Concluding remarks

- 47% of tRNAs genes are located in strand-switch regions
- The multi copy tRNA^{Sec} gene in *T. cruzi* present sytheny with *T. brucei*
- Out of the 19 clusters of tDNAs, 18 present synteny with tDNAs clusters in *T. brucei*
- Codon usage patterns does not seem to be coadapter with the relative abundance of tRNAs isoacceptors in *T. cruzi*
- 7SL RNA gene in *T. cruzi* seems to be longer
- We identify vaultRNA gene in *T. cruzi* but also vaultRNA-like sequences

Non coding genes in trypanosomatids present gene expansions and sytheny. These features are not exclusive to protein coding genes as it was proposed.

[1] Laboratorio de Interacciones Hospedero-Patógeno - UBM, Institut Pasteur de Montevideo, Uruguay; [2] Departamento de Bioquímica, Facultad de Medicina, Universidad de la República, Uruguay