PLOS ONE

RESEARCH ARTICLE

# Conserved motifs in nuclear genes encoding predicted mitochondrial proteins in *Trypanosoma cruzi*

**Lorena Becco**[1], **Pablo Smircich**[1,2], **Beatriz Garat**[1]*

**1** Laboratorio de Interacciones Moleculares, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay, **2** Departamento de Genómica, Instituto de Investigaciones Biológicas Clemente Estable, Ministerio de Educación y Cultura, Montevideo, Uruguay

* bgarat@fcien.edu.uy





**Citation:** Becco L, Smircich P, Garat B (2019) Conserved motifs in nuclear genes encoding predicted mitochondrial proteins in *Trypanosoma cruzi*. PLoS ONE 14(4): e0215160. https://doi.org/10.1371/journal.pone.0215160

**Editor:** M Carolina Elias, Instituto Butantan, BRAZIL

**Received:** January 18, 2019

**Accepted:** March 27, 2019

**Published:** April 9, 2019



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files.

**Funding:** This work received financial support from: Programa de Desarrollo de las Ciencias Básicas to LB, PS and BG (http://www.pedeciba.edu.uy/indice.php); Agencia Nacional de Investigación e Innovación, FCE_2_2011_1_6924, to BG (http://www.anii.org.uy/); Comisión Sectorial de Investigación Científica de la Universidad de la República, Uruguay, proyecto Grupos I+D 2014 108725 to BG (http://www.csic.edu.uy/). The funders had no role in study design, data collection

## Abstract

*Trypanosoma cruzi*, the protozoan parasite that causes Chagas' disease, exhibits peculiar biological features. Among them, the presence of a unique mitochondrion is remarkable. Even though the mitochondrial DNA constitutes up to 25% of total cellular DNA, the structure and functionality of the mitochondrion are dependent on the expression of the nuclear genome. As in other eukaryotes, specific peptide signals have been proposed to drive the mitochondrial localization of a subset of trypanosomatid proteins. However, there are mitochondrial proteins encoded in the nuclear genome that lack of a peptide signal. In other eukaryotes, alternative protein targeting to subcellular organelles via mRNA localization has also been recognized and specific mRNA localization towards the mitochondria has been described. With the aim of seeking for mitochondrial localization signals in *T. cruzi*, we developed a strategy to build a comprehensive database of nuclear genes encoding predicted mitochondrial proteins (MiNT) in the TriTryps (*T. cruzi*, *T. brucei* and *L. major)*. We found that approximately 15% of their nuclear genome encodes mitochondrial products. In *T. cruzi* the MiNT database reaches 1438 genes and a conserved peptide signal, M(L/F) R (R/S) SS, named TryM-TaPe is found in 60% of these genes, suggesting that the canonical mRNA guidance mechanism is present. In addition, the search for compositional signals in the transcripts of *T. cruzi* MiNT genes produce a list, being worth to note a conserved non-translated element represented by the consensus sequence DARRVSG. Taking into account its reported interaction with the *T. brucei* TRRM3 protein which is enriched in the mitochondrial membrane fraction, we here suggest a putative zip code role for this element. Globally, here we provide an inventory of the mitochondrial proteins in *T. cruzi* and give evidence for the existence of both peptide and mRNA signals specific to nuclear encoded mitochondrial proteins.

and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

*Trypanosoma cruzi* (Kinetoplastidae, Trypanosomatidae) is the protozoan parasite that causes Chagas' disease, also known as American trypanosomiasis [1]. This disease affects 6 to 7 millions of people, mostly from poor rural regions of 21 countries of Central and South America, where the vector responsible for the transmission to humans, diverse species of the Reduvidae, is found [2]. Nevertheless, since the parasite can also be transmitted by contaminated food, congenitally from mother to child and through contaminated blood or organ donations, disease has spread out world-wide.

Kinetoplastids are characterized by the presence of a dense structure of DNA and proteins at their unique mitochondrion, the kinetoplast. Therein two types of circular and concatenated DNA are assembled: the maxi and minicircles. The maxicircles (20–37 kb) are functionally equivalent to the mitochondrial DNA of other eukaryotes, and contain the genes encoding rRNAs (12S rRNA and 9S rRNA) and a reduced number of proteins (ND1; ND3;ND4; ND5; ND8; ND9; MURF1; MURF2; MURF5; COI; COII; COIII; Cyb; ATPase6; CR3; CR4; RPS12) [3]. Their transcripts need to be extensively edited (uridine addition/deletion) to solve features such as discontinuous open reading frames (ORFs), absence of essential elements for translation, i.e. initiation codons, or extensive modification to generate ORFs [4–10]. The minicircles (0.5–2.8 kb) contain the huge repository of sequences encoding the guide RNAs (gRNA) which drive the editing process [8,11,12], albeit a few gRNAs are encoded in the maxicircles [13]. All the other mitochondrial proteins are encoded in the nuclear parasite genome. So, in spite of the fact that the mitochondrial DNA constitutes up to 25% of total cellular DNA, the structure and functionality of the mitochondrion (oxidative ATP synthesis, redox balance [14], among others) are absolutely dependent on the expression of the nuclear genome. Proteomic analyses in *T. brucei* have enabled the identification of 1065 mitochondrial proteins from which only 18 are encoded in the mitochondrial genome [15].

In eukaryotes, the mitochondrial localization of proteins encoded in the nuclear genome is paradigmatically achieved through specific peptide signal [16]. This is also the case for trypanosomatids [17,18]. However, there are mitochondrial proteins encoded in the nuclear genome that do not present peptide signal. Afterwards, protein subcellular localization facilitated through mobilization of mRNA towards organelle surrounds has been recognized in eukaryotes [19] and later in *T. cruzi* [20]. Particularly, in eukaryotes but not yet in *T. cruzi*, specific mRNA localization towards the mitochondria has been described [21,22].

In order to study the existence of putative mitochondrial localization signals in *T. cruzi*, a database of nuclear genes encoding predicted mitochondrial proteins was built (MiNT). We could establish that *T. cruzi* MiNT database contains almost 14% of the nuclear genes. Following the same approach, similar results were obtained for the two other trypanosomatid models *T. brucei* and *L. major*, that together with *T. cruzi* conform the so called TriTryps. The presence of peptide signals was studied and a conserved peptide signal, M(L/F) R (R/S) SS, named TryM-TaPe, was found in 60% of the genes in the database. In addition, several nucleotide conserved elements were detected in both 3' and 5' untranslated regions. Amongst them, a compositional conserved element, DARRVSG was identified. Considering that in *T. brucei* this element is recognized by the TRRM3 protein which is mainly associated to the mitochondrial membrane [23], we here suggest a putative zip code role for this conserved element. Globally, the results here presented not only provide an inventory of the mitochondrial proteins in *T. cruzi*, but also give evidence for the existence of both peptide and mRNA signals specific to nuclear encoded mitochondrial proteins.

## Materials and methods

### Databases

This work was performed using data from the TriTrypDB 33 available at Tritrypdb.org [24].

### UTR sequences

To determine the boundaries and sequences of the UTRs of the transcripts in *T. cruzi* RNASeq data from Smircich *et al.* [25] and the UTRme software was used [26].

### Ortholog genes

TriTrypDB tools were used for massive ortholog gene finding when data for compared organisms were available. When this approach was not possible, Best Reciprocal BLAST hit was performed [27], applying in house bash scripts developed for this purpose.

### Gene ontology

Annotation of Gene ontology (GO) terms was performed using the online tool DAVID (Database for Annotation, Visualization and Integrated Discovery, v6.7) [28]. Background databases employed were different, as indicated in each case, depending on the analysis performed. This tool was also used for GO term enrichment analysis. As reported [29], an enrichment score higher than 1.3 was accepted as meaningful. A variant of the Exact Fisher Test (EASE score) was used for p-value calculation. For each ontological category the FDR (false discovery rate) was controlled with the Benjamini method [30], considering acceptable only those values lower than 0.05. Alternatively, TriTrypDB tools were used to analyze GO term enrichment of the gene lists using the whole genome as background. In this case the p-value cutoff was set on 0.01 and the search was limited to GO Slim terms.

### Compositional and structural analysis

Both CDS and UTR GC content was computed with Geecee tool (EMBOSS package). On the other hand, GC content at third codon position (GC3) for CDS was estimated using the INCA software (INteractive Codon usage Analysis) [31]. This tool also allowed the analysis of the codon usage bias using the MELP measure to quantify synonymous codon usage (MILC: Measure Independent of Length and Composition; MELP: MILC-based Expression Level Predictor) [32]. The RNAfold algorithm (Vienna RNA v2.0 package) was used to infer the minimum free energy (MFE) structure and the thermodynamic stability of the whole or partial region of the transcripts [33]. MFE was computed separately for the untranslated regions or the coding sequence.

### Conserved sequence signals

The search for sequence conserved signals was done using the MEME suite package tools [34]. The elicitation of discriminative regular expression motifs in specific data subsets was performed using DREME tool [35], applying the gene complement of the query subset as negative control. The motifs retrieved were compared against Ray motifs' database, using TOMTOM tool [34] from the same package. MEME tool was used to generate consensus regular expressions when comparing motifs from different searches.

When studying signal peptides, three different approximations were made. The first approximation was performed using MEME and FIMO tools [36] both from MEME suite

package with a threshold = 0.001. In addition, TargetP [16] and PredSL [37] signal peptide predictors were used with the default parameters selecting non-plant sequence.

For the motif searching analysis we chose to use a 5' and 3'UTR length corresponding to the 60$^{th}$ percentile of the distribution (S1 Table). For the 3'UTR, this gives 350 nt (3'UTR$_{350}$), while for the 5'UTR the length corresponds to 100 nt. With the purpose of comparing and controlling results obtained for both UTRs it is desirable to achieve equal length for both UTRs, so the 5'UTR was defined as the region starting at -100nt from the AUG to +250 nt downstream from it (5'UTR$_{350}$).

## Results and discussion

### Construction of MiNT: A database of nuclear genes encoding mitochondrial proteins in *T. cruzi*

In order to generate a database that comprehensively contains the nuclear transcripts that encode mitochondrial proteins (MiNT) in *T. cruzi*, an inclusive strategy was followed. First, all the genes whose products were annotated as "mitochondrial" in TriTrypDB 33 were used to conform a base-set which, surprisingly, consisted only on 145 genes. As this strategy revealed meager results, several other complementary approaches were performed (Fig 1). Using TriTrypDB tools, a search by ontology terms (Cellular component = "Mitochondrion"; GO:0005739) was done, obtaining 216 genes. As expected, most of these genes were included in the initial data-set, yielding only a little increment (total 283 genes). Therefore, to further expand the *T. cruzi* database, we use the TriTrypDB search tool based now on text, which allows the exploration of the enclosing author notes. We reasoned that this search could help to detect genes encoding mitochondrial products that have not yet been updated in regard with their function, localization or even for their complete sequence. Using this strategy and after a manual revision of the results, 474 genes including 193 new and 281 already present in the database, were identified. The inclusion of these new genes is supported by the enriched

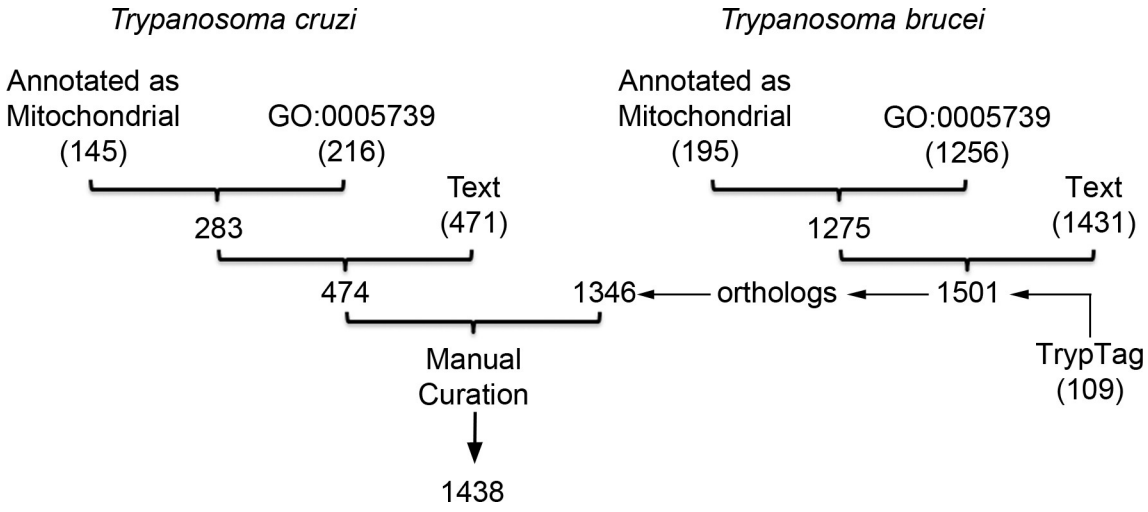


**Fig 1. Schematic representation of the *in-silico* strategy to build a database of the nuclear genes encoding mitochondrial proteins in *T. cruzi*.** The genomes of the mentioned trypanosomes were interrogated under selected categories (annotated as mitochondrial, GO and Text). For each, the output number of genes is shown underneath in brackets. The braces point to the number of genes obtained by the union of the involved outputs. When performed on the *T. cruzi* genome, the search yields only 474 genes. The same strategy in *T. brucei* genome, due to the state of the annotation, allowed to define a database containing 1501 nuclear genes encoding mitochondrial proteins. By comparing this later with the *T. cruzi* genome, a database of 1438 nuclear genes encoding mitochondrial proteins could be obtained.

https://doi.org/10.1371/journal.pone.0215160.g001

terms associated to cell respiration and Krebs cycle, found for this subset through Functional Annotation Clustering tool (DAVID). Since, in the case of *T. brucei*, at least 1065 mitochondrial proteins encoded by the nuclear genome were predicted [15], we considered that the current number of genes in the MiNT database was still too low.

Following the above described search approaches in *T. brucei*, we identified 195 genes annotated as mitochondrial products, 1256 genes under the GO term (GO:005739) and 1431 genes based on text search, summing up a total of 1501, out of the 11703 nuclear genes (13%), encoding mitochondrial products (S2 Table). This database includes not only all the genes that encode mitochondrial products as observed by fluorescent protein tag (109) [38] and most of the genes (1049/1061) proposed by Xiaobai Zhang *et al.* [15] but also adds 452 more genes. It was not surprising to found that our strategy in *T. brucei* yielded a higher number of nuclear genes encoding mitochondrial products than in *T. cruzi* since the annotation of the *T. brucei* genome is much more complete than the one of *T. cruzi*.

As a final approach, we decided to search for the *T. cruzi* genes orthologs to the ones we identified in *T. brucei* following our strategy. After manual revision, 1346 genes were found to have an ortholog gene in *T. cruzi*. With this strategy, 994 new genes were added to the *T. cruzi* database. Thus, we completed the database named as MiNT which, after a final manual curation, contains a total of 14% nuclear genes (1438 out of 10597) encoding mitochondrial proteins (S3 Table).

The same strategy was also performed to define the mitochondrial proteins encoded in the nuclear genome in *L. major*. As in *T. cruzi*, the results of the search by annotation, GO:0005739 or text (175, 313 and 467 genes respectively) gave a poor number of only 477 genes. As before, we added to this group *L. major* orthologs to *T. brucei* and *T. cruzi* MiNT genes. From *T. brucei* MiNT, 1433 ortholog genes in *L. major* were identified, increasing the number of nuclear genes encoding mitochondrial proteins to 1490. In addition, when using *T. cruzi* MiNT 1271 orthologs in *L. major* were found, yielding a MiNT database of 1558 gene (S1 Fig and S4 Table).

As expected, considering the number of genes encoding hypothetical proteins in *T. cruzi*, a great number of the MiNT genes (36%, 519/1438) are annotated as such. Nonetheless, 97% of them (501/519) have an ortholog gene either in *T. brucei* or in *L. major*, strongly suggesting a shared functional role in these organisms.

In brief, following a strict inclusion strategy to avoid false positives, we found that around 15% of the TriTryps' nuclear genome encodes mitochondrial proteins (14%, 13% and 17% for *T. cruzi*, *T. brucei* and *L. major* respectively).

## Compositional and structural analysis of *T. cruzi* MiNT

In order to study the compositional and structural features of the nuclear genes encoding mitochondrial products in *T. cruzi*, the transcripts belonging to MiNT were split into the CDS and UTRs and both were independently analyzed.

For the CDSs, no significant differences were found when comparing length (S2 Fig) or GC content (Fig 2A). However, both GC3 and the MELP (Fig 2B and 2C) were significantly higher in MiNT transcripts than in the rest of the transcriptome suggesting high protein production. Another difference between MiNT CDSs and the CDSs of the rest of the transcriptome (No-MiNT) is found at the structural level. Indeed, in accordance with the mitochondrial prokaryotic origin [39], lower structured CDS are predicted by the MFE per base analysis (Fig 2D) for the nuclear derived transcripts encoding mitochondrial proteins.

We also performed a comparison among MiNT and No-MiNT UTRs in epimastigotes (Fig 3). While no differences were observed when comparing the length of the 5'UTRs, MiNT

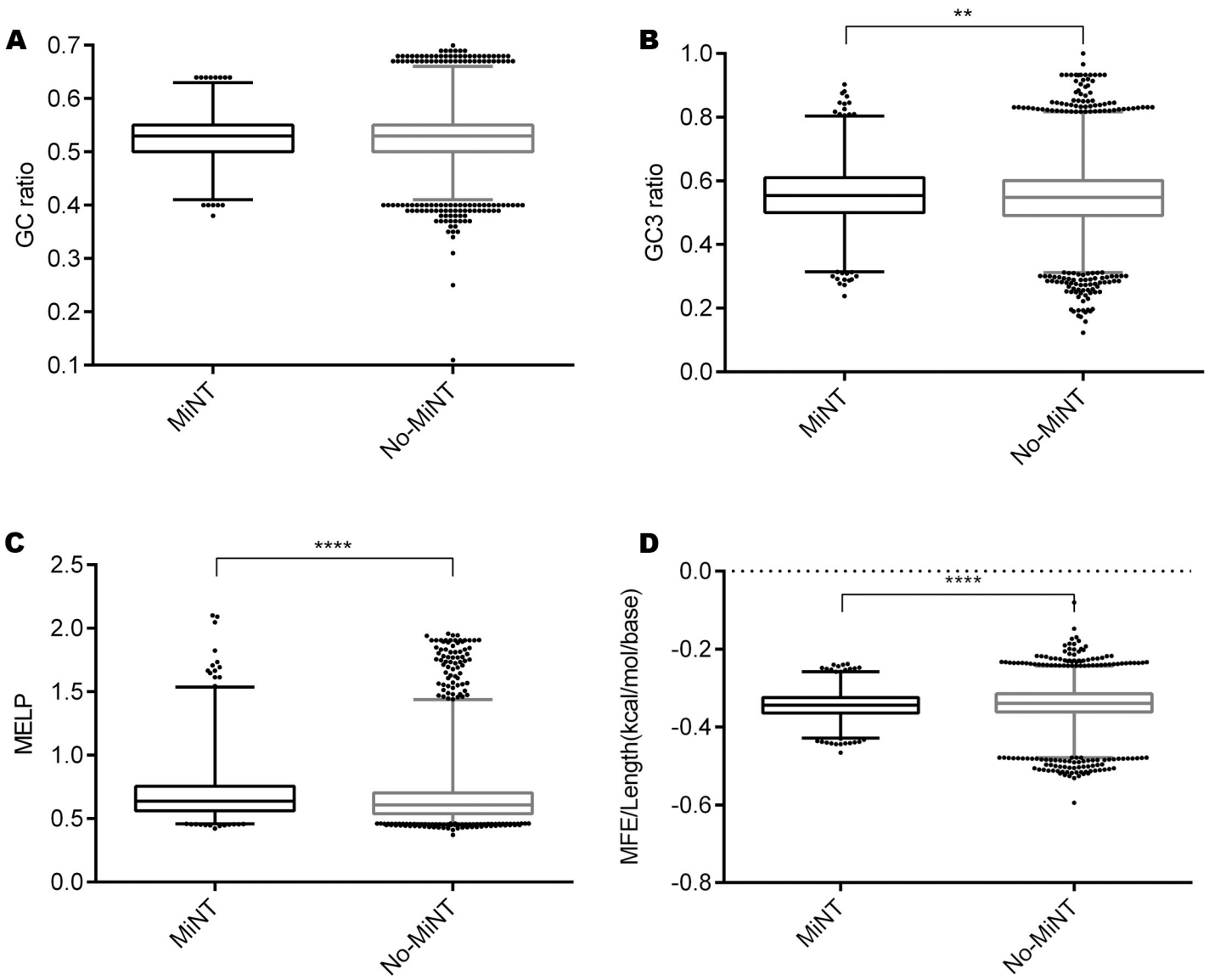


**Fig 2. Analysis of MiNT vs No-MiNT CDS parameters.** Box-plot with whiskers (percentile 1–99%) of (A) G+C Content ratio (*p-value* 0.4321), (B) G+C in the third codon position ratio (p-value < 0.002), (C) MELP, Measure Independent of Length and Composition -based Expression Level Predictor, (*p-value* < 0.001) and (D) MFE (minimum free energy)/Length ratio (*p-value* < 0.001).

https://doi.org/10.1371/journal.pone.0215160.g002

3'UTR are significantly shorter than those of No-MiNT genes (Fig 3A). Regarding the GC content, both 5' and 3' UTR of MiNT present a lower GC content than the ones of No-MiNT genes (Fig 3B). Finally, considering the minimum free energy level, the predicted secondary structure for both 5' and 3'UTR were found to be less stable in MiNT than in No-MiNT (Fig 3C). Similar results were obtained using the UTRs from the metacyclic trypomastigote stage (S3 Fig)

The distinctiveness observed for the coding and not for the regulatory regions of the nuclear genes encoding mitochondrial proteins may be explained by the more stringent requirements that govern their functionality. Indeed, the codon usage is highly non-random with respect to both GC3 and MELP.

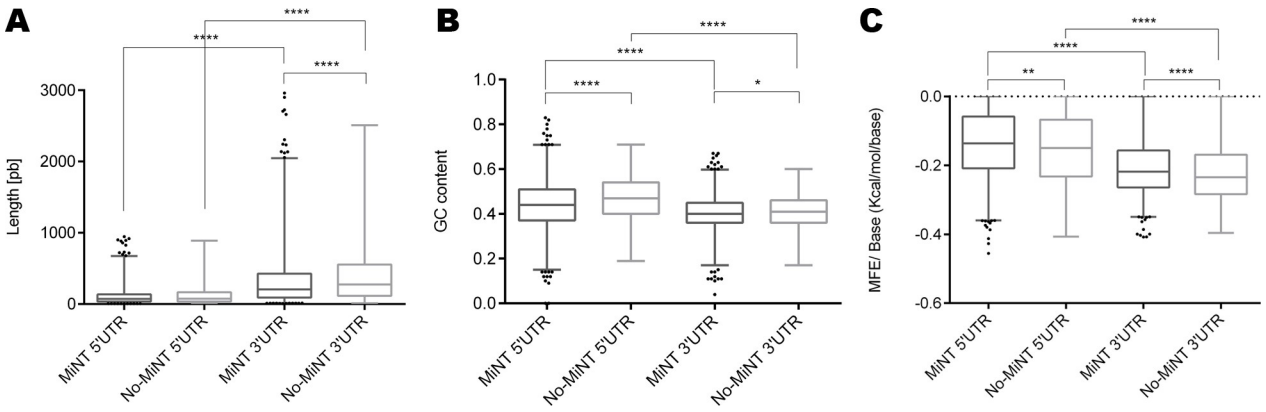


**Fig 3. Analysis of MiNT vs No-MiNT UTRs parameters in epimastigote stage.** Box-plot with whiskers (percentile 1–99%) of (A) Length; (B) G+C Content ratio (C) MFE (minimum free energy)/Length ratio. Multiple comparisons amongst groups were performed by Dunn's multiple comparison test and differences were seen by comparing mean ranks.

https://doi.org/10.1371/journal.pone.0215160.g003

Overall, the compositional and structural analysis of the nuclear encoded mitochondrial genes of *T. cruzi* revealed both high expression characteristic values (higher GC3 and MELP) and prokaryotic origin traces (less structure complexity at CDSs and UTRs) when compared with no-MiNT.

## Expression analysis of MiNT genes along the life cycle of *T. cruzi*

Expression evidence for the 99% of MiNT genes (1427/1438) whether in micro-arrays data [40], expressed sequence tags (EST) or RNA-seq data has been reported [25,41].

First, the micro-arrays data published by Mining *et al.* [40], was used to compare MiNT gene expression across the life cycle of the parasite (Fig 4A). This analysis revealed that MiNT genes have an exacerbated expression when compared with the rest of the genome, being higher in the replicative stages, and lower in both trypomastigotes forms. Similar results were obtained using the RNA-Seq data from Smircich *et al.* [25] (Fig 4B).

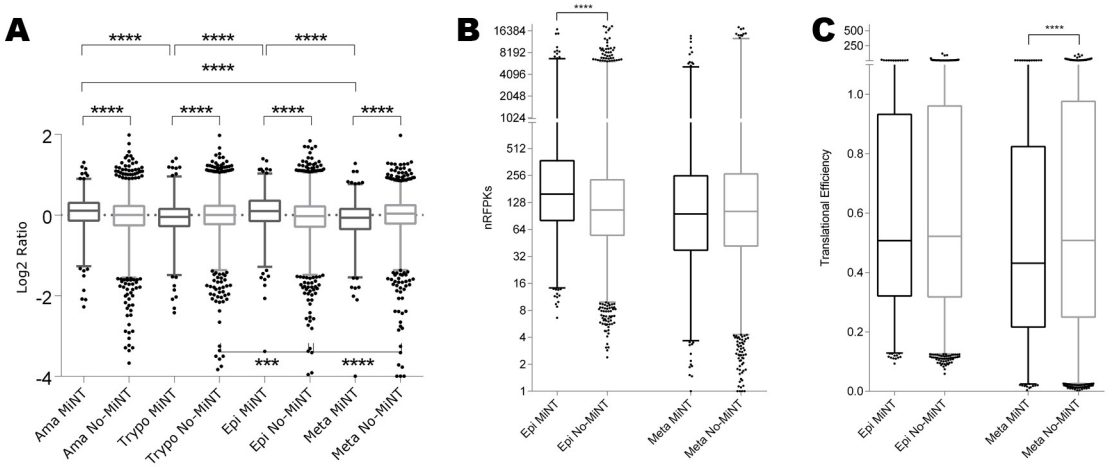


**Fig 4. Comparison of expression data for MiNT and No-MiNT genes in *T. cruzi* life cycle.** (A) Transcriptome data obtained from the microarray approach of Minning, et al. [40] as a ratio of the particular stage to an equal mixture of all four life cycle stages. (B) Transcriptome data obtained from RNA-Seq from Smircich, *et al.* [25] as number of reads per kilobase nRPK. (C) Ribosome footprinting data from RNA-Seq from Smircich, *et al.* [25] as number of footprints per transcript (translational efficiency). Parasite stages: Ama = amastigote; Trypo = trypomastigote; Epi = epimastigote; Meta = Metacyclic.

https://doi.org/10.1371/journal.pone.0215160.g004

As a complementary approach, the analysis of ribosome footprinting data available for the vector parasite stages (epimastigote / metacyclic trypomastigote) [25] was performed (Fig 4C). Higher ribosome occupancy for MiNT when compared to the rest of the genes at the replicative epimastigotes was revealed. No such effect was observed at the metacyclic trypomastigote stage. Indeed, the translation slowdown that is a characteristic of this infective stage is also clearly observed for MiNT.

## Search for an amino-terminal localization peptide signal in mitochondrial proteins encoded by nuclear genes in *T. cruzi*

While the presence of a signal peptide targeting proteins to its final localization is not mandatory, at least 70% of the mitochondrial proteins are demonstrated to carry a peptide that is responsible for their subcellular localization in yeast [42]. Certain loosely characteristics, such as an amphipathic character and the presence of at least two basic amino acids have been proposed for the mitochondrial signal peptide [43]. Nonetheless, there is not a conserved consensus sequence reported for this signal.

Aiming to define a consensus sequence to identify those proteins whose localization could be directed by a signal encoded in the aminoacidic sequence in *T. cruzi*, we firstly analyzed several experimentally tested signal peptides. Thirty-five previously reported signals [44] were used as an input to define a preliminary consensus sequence (Fig 5A). It was then submitted to FIMO analysis using the mitochondrial annotated proteins as the target database. MEME analysis found a consensus mitochondrial targeting sequence named as TryM-TaPe (Trypanosomal Mitochondrial Targeting Peptide) (Fig 5B).

We extended the search of TryM-TaPe to the complete MiNT database and found that 865 out of 1438 encoded proteins of MiNT database (60%) contained this conserved sequence, a number that is consistent with reports in other eukaryotes [42]. Though we cannot rule out the presence of other peptide signals, the existence and representation of TryM-TaPe validates the reliability of the *T. cruzi* MiNT database.

## Search for mRNA localization signals in nuclear genes encoding mitochondrial proteins in *T. cruzi*

The absence of a peptide signal in nuclear encoded mitochondrial proteins may be overcome by the presence of signals in the transcript mediating the approach of mRNAs to the mitochondria. To facilitate the search for these transcript localization signals, the identification of a reliable set of proteins without a mitochondrial localization sequence (MTS) would be advisable. For this purpose, we firstly identified all proteins in MiNT with an MTS (MiNT-MTS dataset) to then obtain those without MTS (MiNT-NoMTS dataset).

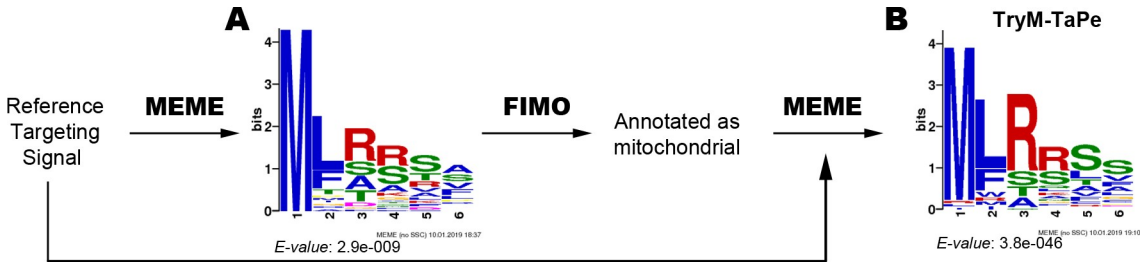


**Fig 5. TryM-TaPe definition strategy.** Mitochondrial targeting sequences available in [44] were used to obtain a preliminary signal (A) that was then used to search among the mitochondrial annotated proteins. This later search output was used as an input to obtain the final signal named as TryM-TaPe (B).

https://doi.org/10.1371/journal.pone.0215160.g005

Since TryM-TaPe may not be the only signal peptide that could be acting as a mediator for transcript or protein localization to the mitochondrial surrounding, two common MTS predictors (TargetP and PredSL) were also used. As shown above, FIMO search led to 865 proteins carrying TryM-TaPe, meanwhile TargetP and PredSL predicted 660 and 521 proteins encoded by MiNT transcripts carry an MTS, respectively (Fig 6). We decided to include in MiNT-MTS those genes predicted by at least two of the three methods (620), while the remaining 818 genes correspond to MiNT-NoMTS.

In order to investigate the presence of sequence elements enriched in MiNT-NoMTS with respect to MiNT-MTS, we analyzed the UTRs of those transcripts. Considering that nucleotide motifs may also complement the function of the signal peptide, the reciprocal search was also carried out. We used the DREME tool [35] on the $UTR_{350}$ (as described in Materials and Methods) to search for enriched motifs within each database and the TOMTOM tool [45] to search for putative interacting RBPs. As the database used in TOMTOM [46] includes *L. major* and *T. brucei gambiense* proteins, *T. cruzi* orthologs were searched.

The analysis of MiNT-NoMTS yielded 21 and 20 motifs in the 5'$UTR_{350}$ and 3'$UTR_{350}$ respectively and several interacting proteins were predicted (S5 Table and Table 1). Common trans-acting factors such as: the protein PABP (polyadenylate-binding protein), a general factor implied in different steps of mRNA metabolism and the protein DRBD12, known to destabilize the wide spread ARE-containing target genes[47], were found to interact to motifs in both the 3'$UTR_{350}$ and 5'$UTR_{350}$. In addition, motifs recognized by DRBD3 (one) and TRRM3 (three) were also found. In *T. brucei* DRBD3 was found to be associated with mRNAs encoding membrane proteins, playing a role in mRNA stabilization, splicing, translation and transport [48]. Interestingly, TRRM3 is found within the mitochondrion or associated to its membrane in *T. brucei* procyclic forms [23]. As usually found in mitochondrial proteins, TRRM3 is

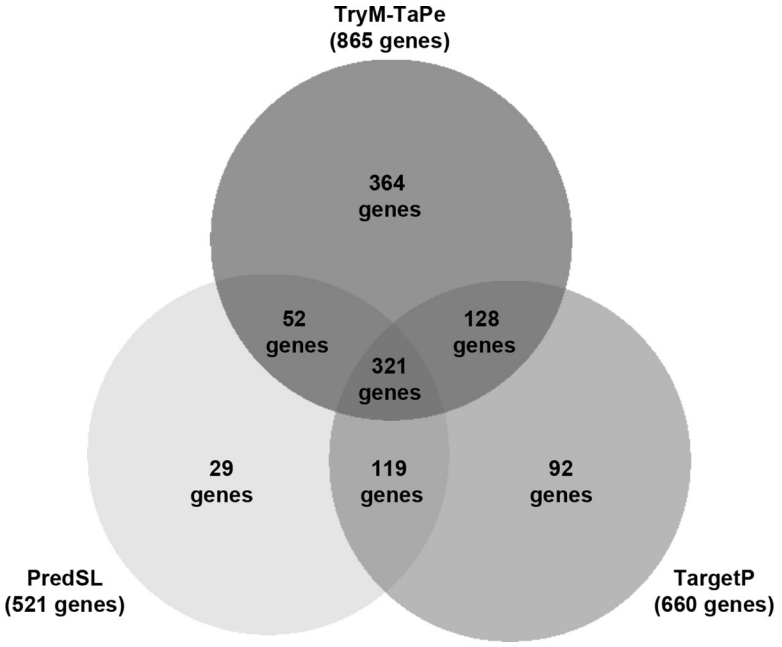


**Fig 6. Venn diagram of the different strategies used to define the subset of nuclear encoded mitochondrial protein carrying a signal peptide in *T. cruzi*.** The signal peptide predictors TargetP [16] and PredSL [37] predictors with the default parameters selecting non-plant sequence, and FIMO using TryM-TaPe were used on the *T. cruzi* MiNT database.

https://doi.org/10.1371/journal.pone.0215160.g006

**Table 1. Enriched motifs found in MiNT-NoMTS database.**

| MiNT-NoMTS vs No-MiNT | | | | | | |
|---|---|---|---|---|---|---|
| **5'UTR** | | | | | | |
| **Motif** | **Consensus Sequence** | **P-value** | **TomTom Result** | **Product Description** | **Ortholog gene** | **Product Description** |
| Motif 1 | MAAAR | 4.8e-008 | Tbg972.9.5210 | polyadenylate-binding protein 1 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| | | | LmjF.35.4130 | polyadenylate-binding protein 2 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| Motif 2 | TTGTKATT | 4.0e-009 | Tbg972.7.6230 | Double RNA binding domain protein 12 | TcCLB.506825.10 | Double RNA binding domain protein 12 |
| Motif 3 | TAGAATGG | 2.6e-008 | Tbg972.3.3970 | Triple RNA binding domain protein 3 | TcCLB.510149.140 | Triple RNA binding domain protein 3 |
| Motif 4 | AGAGAGGT | 9.3e-010 | Tbg972.6.2300 | RNA-binding protein, putative | TcCLB.509965.180 | RNA-binding protein, putative |
| **3'UTR** | | | | | | |
| **Motif** | **Consensus Sequence** | **P-value** | **TomTom Result** | **Product Description** | **Ortholog gene** | **Product Description** |
| Motif 5 | AAMARA | 2.5e-020 | Tbg972.9.5210 | polyadenylate-binding protein 1 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| | | | LmjF.35.4130 | polyadenylate-binding protein 2 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| Motif 6 | GDAAA | 5.8e-014 | Tbg972.9.5210 | polyadenylate-binding protein 1 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| Motif 7 | TTGTYGTT | 2.4e-008 | Tbg972.6.2300 | RNA-binding protein, putative | TcCLB.509965.180 | RNA-binding protein, putative |
| Motif 8 | HGTAG | 5.5e-008 | Tbg972.11.17950 | RNA-binding protein, putative | TcCLB.507037.20 | RNA-binding protein, putative |
| Motif 9 | AGAW | 1.1e-012 | Tbg972.11.17950 | RNA-binding protein, putative | TcCLB.507037.20 | RNA-binding protein, putative |
| Motif 10 | TTMTTW | 7.2e-011 | Tbg972.9.4840 | Double RNA binding domain protein 3 | TcCLB.506649.80 | Double RNA binding domain protein 3 |
| | | | Tbg972.7.6230 | Double RNA binding domain protein 12 | TcCLB.506825.10 | Double RNA binding domain protein 12 |
| Motif 11 | ARRGGG | 2.0e-019 | Tbg972.3.3970 | Triple RNA binding domain protein 3 | TcCLB.510149.140 | Triple RNA binding domain protein 3 |
| Motif 12 | WAGG | 3.8e-018 | Tbg972.3.3970 | Triple RNA binding domain protein 3 | TcCLB.510149.140 | Triple RNA binding domain protein 3 |
| Motif 13 | GAAGCCC | 9.8e-009 | Tbg972.3.3970 | Triple RNA binding domain protein 3 | TcCLB.510149.140 | Triple RNA binding domain protein 3 |

https://doi.org/10.1371/journal.pone.0215160.t001

regulated by arginine methylation [49]. Therefore, it is tempting to propose a zip code role for the TRRM3 target motifs.

For MiNT-MTS we found 12 overrepresented sequences in the 5´UTR$_{350}$ and 12 in the 3´UTR$_{350}$ (S6 Table). As expected, and validating our approach, the sequence encoding TryM--TaPe was found. TOMTOM analysis allowed to associate four motifs of the 3'UTR$_{350}$ to RNA binding proteins (Table 2). Two of them are recognized by the general factor PABP and two others by two isoforms of the double RNA binding domain, DRBD3 and DRBD9. Remarkably, one motif is recognized by TRRM3. As mentioned above, this is not surprising since subcellular localization signals may act in collaboration with peptide signals. Thus, this finding reinforces the proposed role. All the sequences, in the 3'UTR$_{350,}$ found to be associated to TRRM3 (Motifs 11, 12, 13 and 15) were used to obtain a consensus recognition motif (DARRVSG) which in turn, can also be recognized by TRRM3 according to the TOMTOM algorithm (*p*-value 8.10 e -03).

**Table 2. Enriched motif found in MiNT-MTS database.**

| MiNT-MTS vs No-MiNT | | | | | | |
|---|---|---|---|---|---|---|
| 3'UTR | | | | | | |
| **Motif** | **Consensus Sequence** | **P-value** | **TomTom Result** | **Product Description** | **Ortholog gene** | **Product Description** |
| Motif 14 | MAAA | 6.6e-008 | Tbg972.9.5210 | polyadenylate-binding protein 1 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| | | | LmjF.35.4130 | polyadenylate-binding protein 2 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| Motif 15 | AARRA | 1.4e-015 | Tbg972.9.5210 | polyadenylate-binding protein 1 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| | | | LmjF.35.4130 | polyadenylate-binding protein 2 | TcCLB.506885.70 | polyadenylate-binding protein 1, putative |
| | | | Tbg972.3.3970 | Triple RNA binding domain protein 3 | TcCLB.510149.140 | Triple RNA binding domain protein 3 |
| Motif 16 | CTTWTT | 6.7e-014 | Tbg972.9.4840 | Double RNA binding domain protein 3 | TcCLB.506649.80 | Double RNA binding domain protein 3 |
| Motif 17 | GAACGCCT | 1.2e-007 | LmjF.35.2550 | Double RNA binding domain protein 9 | TcCLB.510747.80 | Double RNA binding domain protein 9 |

https://doi.org/10.1371/journal.pone.0215160.t002

It is worth noting that the RNA binding domains for all the putative interactors here presented have a high identity to the respective *T. cruzi* ortholog, suggesting that they could recognize the same motifs (see alignments on S7 Table).

In spite of the fact that the relevance of the motifs found will require further study, it is tempting to propose that TRRM3 and its cognate recognition motif play a role as a zip code transporting specific mRNAs to the mitochondrial surroundings.

## Conclusions

Aiming to identify conserved signals among the nuclear genes encoding mitochondrial proteins in *T. cruzi*, we searched for the genes annotated as such in the TriTrypDB. Despite its availability since 2005 [50], and the many efforts to its improvement, completion and annotation from there on, we only obtained meager results. Thus, we undertook the task of obtaining a comprehensive list of nuclear genes encoding mitochondrial proteins which would not only serve as the dataset target for the aim of this work but also constitute by itself a contribution to the current state of the knowledge of *T. cruzi* genome. Following an in-silico strategy, a wide inventory of the nuclear genes encoding mitochondrial proteins, MiNT, in the TriTryps was obtained (1438, 1501 and 1558 for *T. cruzi*, *T. brucei* and *L. major* respectively). The search for enriched motifs in *T. cruzi* MiNT allowed the identification of a list of conserved signals. Signals involved in different metabolic steps were identified. For the well-known mitochondrial localization peptides, we could establish a consensus motifhere named TryM-TaPe, M(L/F) R (R/S) SS, present in 60% of *T. cruzi* MiNT database. In addition, a putative mitochondrial localization role is here proposed for the nucleic element DARRVSG that may be recognized by the conserved TRRM3 protein which is enriched in the mitochondrial membrane fraction in *T. brucei*. While work is in progress to analyze the role of this element, its actual interaction with TRRM3 and the function of this RBP, these findings suggest that in addition to the canonical peptide localization signal, mRNA localization could be guided to the mitochondria surroundings via zip code nucleic signals present in the UTRs in *T. cruzi*.

## Supporting information

**S1 Fig. Schematic representation of the *in-silico* strategy to build a database of the nuclear genes encoding mitochondrial proteins in *L. major*.** The genomes of the mentioned

trypanosomes were interrogated under selected categories (annotated as mitochondrial, GO and Text). For each, the output number of genes is shown underneath in brackets. The braces point to the number of genes obtained by the union of the involved outputs. When performed on the *L. major* genome, the search yields only 477 genes. This result was complemented with the orthologs of *T. brucei* and *T. cruzi*.
(TIF)

**S2 Fig. Analysis of MiNT vs No-MiNT CDS length.** Box-plot with whiskers (percentile 1–99%) of CDS length.
(TIF)

**S3 Fig. Analysis of MiNT vs No-MiNT UTRs parameters in trypomastigote stage.** Box-plot with whiskers (percentile 1–99%) of (A) Length; (B) G+C Content ratio (C) MFE/Length ratio. Multiple comparisons amongst groups were performed by Dunn's multiple comparison test and differences were seen by comparing mean ranks.
(TIF)

**S1 Table. *T. cruzi* UTR length and statistical analysis.** Length obtained for both 5' and 3'UTRs in epimastigotes and trypomastigotes stages data from Smircich *et al.* [25] using UTRme tool [26]. The descriptive statistics for each group are also included.
(XLSX)

**S2 Table. *T. brucei* nuclear genes encoding mitochondrial proteins.** The results of each step of the search strategy are presented.
(XLSX)

**S3 Table. *T. cruzi* nuclear genes encoding mitochondrial proteins.** The results of each step of the search strategy are presented.
(XLSX)

**S4 Table. L. major nuclear genes encoding mitochondrial proteins.** The results of each step of the search strategy are presented.
(XLSX)

**S5 Table. Enriched motifs in MiNT-NoMTS UTRs.** Complete set of motifs found by DREME search for both 5' and 3' $UTR_{350}$. TomTom results and the ortholog gene in *T. cruzi*, when found, are also described.
(XLSX)

**S6 Table. Enriched motifs in MiNT-MTS UTRs.** Complete set of motifs found in DREME search for both 5' and 3' $UTR_{350}$. TomTom results and the ortholog gene in *T. cruzi*, when found, are also described.
(XLSX)

**S7 Table. RRM domain of the interacting RBPs and *T. cruzi* ortholog alignments.** The RRM domain predicted by pfam in TriTrypDB for each RBP and its putative *T. cruzi* ortholog were aligned to confirm the similarity.
(XLSX)

## Acknowledgments

We acknowledge all members of the Laboratorio de Interacciones Moleculares at UDELAR and the Department of Genomics at IIBCE for constant discussion and technical support. Particularly, we kindly appreciate the technical support on UTR analyses provided by Santiago

Radio. We also thank several colleagues that have provided critical insight in this study during scientific meetings.

## Author Contributions

**Conceptualization:** Lorena Becco, Pablo Smircich, Beatriz Garat.

**Data curation:** Lorena Becco.

**Formal analysis:** Lorena Becco.

**Investigation:** Lorena Becco.

**Supervision:** Pablo Smircich, Beatriz Garat.

**Writing – original draft:** Lorena Becco, Pablo Smircich, Beatriz Garat.

**Writing – review & editing:** Lorena Becco, Pablo Smircich, Beatriz Garat.

## References

1. Chagas C. Nova tripanozomiaze humana. Estudos sobre a morfolojia e o ciclo evolutivo do Schizotrypanum cruzi n.gen. n.sp., ajente etiolojico de nova entidade morbida do homem. Mem Inst Oswaldo Cruz. 1909; 1: 159–218.
2. WHO. Chagas disease (American trypanosomiasis). In: Media centre [Internet]. 2015. Available: http://www.who.int/mediacentre/factsheets/fs340/en/
3. Genomics B, Westenberger SJ, Cerqueira GC, El-Sayed NM, Zingales B, Campbell DA, et al. Trypanosoma cruzi mitochondrial maxicircles display species-and strain-specific variation and a conserved element in the non-coding region. BMC Genomics. 2006;7. https://doi.org/10.1186/1471-2164-7-7
4. Westenberger SJ, Cerqueira GC, El-Sayed NM, Zingales B, Campbell D a, Sturm NR. Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. BMC Genomics. 2006; 7: 60. https://doi.org/10.1186/1471-2164-7-60 PMID: 16553959
5. Payne M, Rothwell V, Jasmer DP, Feagin JE, Stuart K. Identification of mitochondrial genes in Trypanosoma brucei and homology to cytochrome c oxidase II in two different reading frames. Mol Biochem Parasitol. 1985; 15: 159–70. PMID: 2989684
6. Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986; 46: 819–26. PMID: 3019552
7. Simpson L. The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution. Annu Rev Microbiol. 1987; 41: 363–382. https://doi.org/10.1146/annurev.mi.41.100187.002051 PMID: 2825587
8. Feagin JE, Shaw JM, Simpson L, Stuart K. Creation of AUG initiation codons by addition of uridines within cytochrome b transcripts of kinetoplastids. Proc Natl Acad Sci U S A. 1988; 85: 539–43. PMID: 2448777
9. Shaw JM, Feagin JE, Stuart K, Simpson L. Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell. 1988; 53: 401–11. PMID: 2452696
10. van der Spek H, van den Burg J, Croiset A, van den Broek M, Sloof P, Benne R. Transcripts from the frameshifted MURF3 gene from Crithidia fasciculata are edited by U insertion at multiple sites. EMBO J. 1988; 7: 2509–14. PMID: 2461295
11. Pollard VW, Rohrer SP, Michelotti EF, Hancock K, Hajduk SL. Organization of minicircle genes for guide RNAs in trypanosoma brucei. Cell. 1990; 63: 783–790. https://doi.org/10.1016/0092-8674(90)90144-4 PMID: 2171782
12. Sturm NR, Simpson L. Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. Cell. 1990; 61: 879–884. https://doi.org/10.1016/0092-8674(90)90198-N PMID: 1693097
13. Thomas S, Martinez LLIT, Westenberger SJ, Sturm NR. A population study of the minicircles in Trypanosoma cruzi: predicting guide RNAs in the absence of empirical RNA editing. BMC Genomics. 2007; 8: 133. https://doi.org/10.1186/1471-2164-8-133 PMID: 17524149

14. Tomás AM, Castro H. Redox metabolism in mitochondria of trypanosomatids. Antioxid Redox Signal. 2013; 19: 696–707. https://doi.org/10.1089/ars.2012.4948 PMID: 23025438

15. Zhang X, Cui J, Nilsson D, Gunasekera K, Chanfon A, Song X, et al. The Trypanosoma brucei MitoCarta and its regulation and splicing pattern during development. Nucleic Acids Res. 2010; 38: 7378–87. https://doi.org/10.1093/nar/gkq618 PMID: 20660476

16. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. Nature Publishing Group; 2007; 2: 953–71. https://doi.org/10.1038/nprot.2007.131 PMID: 17446895

17. Herrmann JM, Neupert W. Protein transport into mitochondria. Curr Opin Microbiol. 2000; 3: 210–4. PMID: 10744987

18. Herrmann JM, Neupert W. What fuels polypeptide translocation? An energetical view on mitochondrial protein sorting. Biochim Biophys Acta. 2000; 1459: 331–8. PMID: 11004448

19. Weis BL, Schleiff E, Zerges W. Protein targeting to subcellular organelles via mRNA localization. Biochim Biophys Acta—Mol Cell Res. 2013; 1833: 260–273. https://doi.org/10.1016/j.bbamcr.2012.04.004

20. Alves LR, Guerra-Slompo EP, de Oliveira A V, Malgarin JS, Goldenberg S, Dallagiovanna B. mRNA localization mechanisms in Trypanosoma cruzi. PLoS One. 2013; 8: e81375. https://doi.org/10.1371/journal.pone.0081375 PMID: 24324687

21. Gerber AP, Luschnig S, Krasnow M a, Brown PO, Herschlag D. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2006; 103: 4487–92. https://doi.org/10.1073/pnas.0509260103 PMID: 16537387

22. Quenault T, Lithgow T, Traven A. PUF proteins: repression, activation and mRNA localization. Trends Cell Biol. 2011; 21: 104–12. https://doi.org/10.1016/j.tcb.2010.09.013 PMID: 21115348

23. Niemann M, Wiese S, Mani J, Chanfon A, Jackson C, Meisinger C, et al. Mitochondrial outer membrane proteome of Trypanosoma brucei reveals novel factors required to maintain mitochondrial morphology. Mol Cell Proteomics. 2013; 12: 515–28. https://doi.org/10.1074/mcp.M112.023093 PMID: 23221899

24. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010; 38: D457–D462. https://doi.org/10.1093/nar/gkp851 PMID: 19843604

25. Smircich P, Eastman G, Bispo S, Duhagon MA, Guerra-Slompo EP, Garat B, et al. Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi. BMC Genomics. 2015; 16: 443. https://doi.org/10.1186/s12864-015-1563-8 PMID: 26054634

26. Radío S, Fort RS, Garat B, Sotelo-Silveira J, Smircich P. UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes. Front Genet. Frontiers; 2018; 9: 671. https://doi.org/10.3389/fgene.2018.00671 PMID: 30619487

27. Ward N, Moreno-Hagelsieb G, Altschul S, Madden T, Schaffer A, Zhang J, et al. Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss?. PLoS One.; 2014; 9: e101850. https://doi.org/10.1371/journal.pone.0101850 PMID: 25013894

28. Dennis Glynn. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. BioMed Central; 2003; 4: R60.

29. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4: 44–57. https://doi.org/10.1038/nprot.2008.211 PMID: 19131956

30. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001; 125: 279–284. PMID: 11682119

31. Supek F, Vlahovicek K. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics. 2004; 20: 2329–30. https://doi.org/10.1093/bioinformatics/bth238 PMID: 15059815

32. Supek F, Vlahovicek K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics. 2005; 6: 182. https://doi.org/10.1186/1471-2105-6-182 PMID: 16029499

33. Hofacker IL. Vienna RNA secondary structure server. Nucleic Acids Res. 2003; 31: 3429–31. PMID: 12824340

34. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37: W202–8. https://doi.org/10.1093/nar/gkp335 PMID: 19458158

35. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011; 27: 1653–9. https://doi.org/10.1093/bioinformatics/btr261 PMID: 21543442

36. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27: 1017–8. https://doi.org/10.1093/bioinformatics/btr064 PMID: 21330290

37. Petsalaki EI, Bagos PG, Litou ZI, Hamodrakas SJ. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics Proteomics Bioinformatics. 2006; 4: 48–55. https://doi.org/10.1016/S1672-0229(06)60016-8 PMID: 16689702

38. Dean S, Sunter JD, Wheeler RJ. TrypTag.org: A Trypanosome Genome-wide Protein Localisation Resource. Trends Parasitol. 2017; 33: 80–82. https://doi.org/10.1016/j.pt.2016.10.009 PMID: 27863903

39. Solaimuthu S. Natural selection on mRNA secondary structure and its correlation with protein functional groups. Faculty of New Jersey Institute of Technology. 2010.

40. Minning TA, Weatherly DB, Atwood J, Orlando R, Tarleton RL. The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. BMC Genomics. 2009; 10: 370. https://doi.org/10.1186/1471-2164-10-370 PMID: 19664227

41. Li Y, Shah-Simpson S, Okrah K, Belew AT, Choi J, Caradonna KL, et al. Transcriptome Remodeling in Trypanosoma cruzi and Human Cells during Intracellular Infection. PLoS Pathog. 2016; 12: e1005511. https://doi.org/10.1371/journal.ppat.1005511 PMID: 27046031

42. VöGtle F-N, Wortelkamp S, Zahedi RP, Becker D, Leidhold C, Gevaert K, et al. Global Analysis of the Mitochondrial N-Proteome Identifies a Processing Peptidase Critical for Protein Stability. Cell. 2009; 139: 428–439. https://doi.org/10.1016/j.cell.2009.07.045 PMID: 19837041

43. von Heijne G, Steppuhn J, Herrmann RG. Domain structure of mitochondrial and chloroplast targeting peptides. Eur J Biochem. 1989; 180: 535–45. PMID: 2653818

44. Häusler T, Stierhof YD, Blattner J, Clayton C. Conservation of mitochondrial targeting sequence function in mitochondrial and hydrogenosomal proteins from the early-branching eukaryotes Crithidia, Trypanosoma and Trichomonas. Eur J Cell Biol. 1997; 73: 240–51. PMID: 9243185

45. Gupta S, Stamatoyannopoulos JA, Bailey TTL, Noble WS, Maniatis T, Goodbourn S, et al. Quantifying similarity between motifs. Genome Biol. 2007; 8: R24. https://doi.org/10.1186/gb-2007-8-2-r24 PMID: 17324271

46. Ray D, Weirauch HKKBCMT, Najafabadi HS, Li X, Gueroussov S, Albu M, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499: 172–7. https://doi.org/10.1038/nature12311 PMID: 23846655

47. Kim D, Chiurillo MA, El-Sayed N, Jones K, Santos MRM, Porcile PE, et al. Telomere and subtelomere of Trypanosoma cruzi chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of T. cruzi telomeres. Gene. 2005; 346: 153–61. https://doi.org/10.1016/j.gene.2004.10.014 PMID: 15716016

48. Das A, Bellofatto V, Rosenfeld J, Carrington M, Romero-Zaliz R, Del Val C, et al. High throughput sequencing analysis of Trypanosoma brucei DRBD3/PTB1-bound mRNAs. Mol Biochem Parasitol. 2015; 199: 1–4. https://doi.org/10.1016/j.molbiopara.2015.02.003 PMID: 25725478

49. Lott K, Li J, Fisk JC, Wang H, Aletta JM, Qu J, et al. Global proteomic analysis in trypanosomes reveals unique proteins and conserved cellular processes impacted by arginine methylation. J Proteomics. 2013; 91: 210–25. https://doi.org/10.1016/j.jprot.2013.07.010 PMID: 23872088

50. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science. 2005; 309: 404–9. https://doi.org/10.1126/science.1112181 PMID: 16020724