Chapter

Biology of the *Trypanosoma cruzi* Genome

Luisa Berná, Sebastián Pita, María Laura Chiribao, Adriana Parodi-Talice, Fernando Alvarez-Valin and Carlos Robello

Abstract

The genome of *Trypanosoma cruzi* was first made available in 2005, and the intrinsic genome complexity of this parasite has hindered high-quality genome assembly and annotation. Recent technological developments in long read sequencing allowed to circumvent this problem, showing very interesting features in the genome architecture of *T. cruzi*, allowing to accurately estimate gene copy numbers, abundance and distribution of repetitive sequences (including satellites and retroelements), and the complexity of multigene families implied in host-parasite interactions. The genome of *T. cruzi* is composed of a "core compartment" and a "disruptive compartment" which exhibit opposite GC content and gene composition, with high differences on their regulatory regions. The novel tandem and dispersed repetitive sequences identified, in addition to recombination events, allows to conclude that genome plasticity is a key survival strategy during its complex life cycle.

Keywords: genome, *Trypanosoma cruzi*, compartmentalization, core and disruptive compartments

1. Introduction

The complex genome of *Trypanosoma cruzi* reflects its complex life. These parasites are able to invade almost any kind of cell to freely circulate in blood or extracellular matrix, to pass through the digestive tract of its insect vector and survive after being eliminated in feces. This stressful lifestyle strongly requires a fine regulation of gene expression, which in turn is reflected on its genome organization. Although the focus of this chapter is the nuclear genome (hereinafter called generically "genome"), it is worth mentioning that *Trypanosoma*tids have another genome that contained in their single mitochondria called kinetoplast DNA. This exhibits unique architectural and functional features: it consists of a dense network of two types of circular DNA molecules called maxicircles and minicircles. Maxicircles, of several kb in length, are equivalent to regular mtDNA of other eukaryotes, whereas minicircles are much shorter (seldom longer than 2 kb) and encode gRNAs. These are short RNA molecules responsible for guiding RNA editing, a process of posttranscriptional modifications that consists in the addition and deletion of uridines. Although editing is not exclusive of trypanosomatids, only in this group it involves massive changes in several (mitochondrialy encoded) genes.

Trypanosomes have peculiarities in transcription and genome organization that differentiate them from the majority of eukaryotes. Protein-coding genes are organized in clusters separated by relatively short intergenic regions, located on the same DNA strand [1] and—with a few exceptions—do not contain introns. Clusters are transcribed as long nuclear polycistronic units, and maturation implies 3′ polyadenylation—characteristic of eukaryotes—and trans-splicing, a peculiar mechanism of mRNA maturation. Trans-splicing is the process by which two RNAs encoded in different genome locations (trans) react to form a unique transcript, where the 5′ moiety contains the spliced leader sequence (~40 nt), and the rest contains the transcribed gene [2, 3]. The spliced leader (SL) is transcribed from a tandem array as a precursor of ~140 nt whose 3′ end is removed and SL inserted to an AG splice-acceptor site on a pre-mRNA molecule, through a molecular mechanism that resembles cis-splicing [4–6]. Usually polypyrimidine-rich motifs precede AG splice acceptor. Since SL-RNA is the target of capping, trans-splicing is responsible for the addition of the 7-methylguanosine cap-like (cap4) on RNAs [7]. It has been described decades ago that this process is coupled to the polyadenylation of the 3′ end of the upstream gene, co-transcriptionally. As a consequence, a molecule of mature mRNA (capped, trans-spliced and polyadenylated) is released from the polycistron and exported to the cytoplasm, where it can be translated. Unlike other organisms, where trans-splicing also occurs, in trypanosomatids it affects almost all genes. Therefore, in trypanosomes the 5′ UTR is the sequence segment located between the SL and the start codon, whereas the 3′ UTR is defined in the same way as in other eukaryotes. With the exception of genes tandemly repeated, polycistronic units do not contain functionally related genes, and usually individual genes from the same transcription unit can show markedly different expression patterns along life cycle [1, 3]. Gene expression in trypanosomes is regulated mainly at the posttranscriptional level, and numerous studies have shown the relevance of 3′ UTR regions in regulation, affecting mRNA stability or translation, and hence differential expression [3, 8]. Different elements in the 3′ UTRs together with the presence of a high number of RNA binding proteins could explain, at least in part, differential expression [9–11], although the exact mechanisms allowing gene specificity are still unknown.

An important issue that still is not clear is whether *T. cruzi* constitutes a single species or a complex of species. Initially two groups of *T. cruzi* were described (I and II) based on biological and biochemical criteria as well as molecular techniques [12]. The first study using molecular phylogeny (sequences of coding genes) clearly showed that at least three major lineages (A, B and C) were present in this parasite [13], and that the distances between these groups are equivalent to the distance between different species of *Leishmania*. Currently six groups or discrete typing units (DTUs) named TcI-TcVI were proposed [14], and *T. cruzi* isolates from bats were included as a seventh DTU [15, 16]; where TcV and TcVI are hybrid lineages derived from haplotypes TcII and TcIII [16]. However, the high biological and genetic diversity of the *T. cruzi* strains, even at the intra-DTU level, indicates that DTUs constitute a useful working definition, but not a definitive classification. The new era of genomic studies through next generation sequencing (NGS) is providing new insights on the above-mentioned unsolved questions.

2. Genome organization

2.1 Chromosomes

In *T. cruzi* mitosis occurs without a complete disruption of nuclear envelope. In addition, although nucleosomes are present, chromatin does not condense up to

chromosomes, so they cannot be visualized by microscopy. This feature has made classic cytogenetic studies unsuitable for these parasites. Instead, *T. cruzi* karyotype has been determined by molecular biology techniques, mainly pulsed field gel electrophoresis (PFGE) in combination with Southern blot [17–19]. Early studies revealed complex chromosomal patterns, evidenced by different PFGE profiles among strains, and allowed to infer that *T. cruzi* was at minimum diploid [20]. Size of chromosomal bands ranges from 0.45 to 4 Mb, without minichromosomes, and the number of chromosomes was estimated mainly through probes used as genetic markers. Depending on the probes and PFGE conditions, chromosomes ranged between 19 to 40 per haploid genome, showing that *T. cruzi* is mainly diploid, although the sizes of homologous chromosomes can differ significantly [17–19, 21–23].

A milestone was achieved in 2005 when the draft genomes of *L. major*, *T. brucei*, and *T. cruzi* were simultaneously published and referred as to the "TriTryps" [24–26]. This opened a new era in biology research on these parasites. A distinctive feature in *T. cruzi* was the already known highly repetitive nature of its genome (50%): in fact 5–10% of the genome is composed by the 195 bp satellite, and the rest of the repetitive DNA is composed of multigene families, tandem repeats and retrotransposable elements [27]. This feature gave rise to a highly fragmented assembly, resulting in that chromosome number and structure or, at least large contigs, could not be obtained. Attempts to recover full length chromosome sequencing, used a combined strategy based on synteny maps with *T. brucei* chromosomes and BAC ends sequencing. By this means 41 virtual chromosomes were obtained for the hybrid CL-Brener strain. Although this strategy represented a substantial improvement in comparison to previous versions of the genome, the issue of assembly fragmentation remained as a limitation for diverse types of analyses that require high precision. A recent milestone in the area was the first publication of long read sequencing of two *T. cruzi* genomes (Dm28c and TCC strains), which allowed to circumvent the limitation of high fragmentation imposed by the Sanger method [28], as well as by short reads NGS methods. Using this approach, also described for Bug strain [29], contigs of more than 1 Mb were obtained, probably covering whole chromosomes, but fragmentation still persists in some regions of the genomes. The exact number of chromosomes and their organization will be finally achieved through the combination of long read sequencing methods, optical maps techniques and polymer-based modeling, a field that has undergone a dramatic acceleration in the last decade [30].

2.2 Ploidy

Although PFGE and fluorescence cytophotometry were useful methods to depict the complex variability of *T. cruzi* karyotypes, it was not until the advent of next generation sequencing techonologies (NGS) that ploidy—or chromosomal copy number variation (CCNV)—analyses could be studied more in detail. Aneuploidy, the gain or loss of chromosomal copies, is of particular importance since it gives clues about the relevance of genome plasticity in the context of parasite fitness. This phenomenon has been detailed studied in *Leishmania spp.*, whose "mosaic" aneuploidies—ploidy variations within isolates from a strain and even between individual cells from a population – were related to drug resistance, regulation on gene expression, or host adaptation [31–33]. On the contrary, PFGE, fluorescence cytophotometry and high-throughput sequencing data analyses agreed on the ploidy stability of *T. brucei* and its subspecies: *T. b. brucei*, *T. b. gambiense* and *T. b. rhodesiense* [34–36]. Remarkably, a field isolated *T. congolense* triploid was reported, suggesting that Salivarian evolutionary lineage species, such as *T. brucei* and *T. congolense*, can sustain euploidies but not massive aneuploidies [37].

Strains	DTU	Size (Mb)	Contigs	N50	L50	GC%	Genes	Proteins	Sequencing plataform	References
Trypanosoma cruzi										
Dm28c	TcI	53,3	636	317.638	47	51,6	18759	15319	PacBio + illumina	[28]
TCC	TcVI (hybrid)	87,1	1.236	264.196	92	51,7	29109	24191	PacBio + illumina	[28]
Bug2148	TcV (hybrid)	55,2	929	200.364	64	51,3	-	-	PacBio	[29]
CL Brener	TcVI (hybrid)	89,9	29.495	88.624	212	51,7	23696	19607	Sanger	[25]
Esmeraldo-like*		32,5	41	--	--	--	11106	10338	--	[49]
Non-Esmeraldo-like*		32,5	41	--	--	--	11398	10831	--	[49]
Dm28c	TcI	27,3	1.210	78.389	86	50,6	11398	11348	454	[81]
G	TcI	25,2	1.450	74.655	91	47,4	13488	12708	454	[82]
CL	TcVI (hybrid)	65,0	7.764	73.547	95	39,8	34248	32278	454	[82]
S23b	TcII	28,1	7.145	20.992	332	45,2	-	-	Illumina	[38]
S92a	TcII	27,1	7.134	20.493	310	46,4	-	-	Illumina	[38]
S11	TcII	28,5	7.855	18.630	346	45,1	-	-	Illumina	[38]
S44a	TcII	17,2	4.971	17.818	232	45,0	-	-	Illumina	[38]
S15	TcII	27,5	9.197	17.779	370	46,2	-	-	Illumina	[38]
231	TcIII	35,4	8.469	14.202	586	48,6	-	-	Illumina	[83]
S162a	TcII	27,3	8.588	12.390	448	45,3	-	-	Illumina	[38]
Y	TcII	39,0	9.821	11.962	561	49,8	-	-	Illumina	[29]
Ycl4	TcII	26,1	6.664	10.716	560	46,6	-	-	Illumina	[38]
Ycl2	TcII	25,9	6.884	10.600	563	46,6	-	-	Illumina	[38]
Ycl6	TcII	25,8	6.967	10.394	549	46,6	-	-	Illumina	[38]
S154a	TcII	19,3	6.946	5.877	859	49,6	-	-	Illumina	[38]
Y	TcII	30,0	8.952	5.474	1305	50,6	-	-	454	[39]
Colombiana	TcI	30,9	9.338	5.189	1394	50,8	-	-	454	[39]
Sylvio X10/1	TcI	38,6	27.019	2.307	2599	51,2	10861	10847	454	[53, 84]
Arequipa	TcI	19,1	10.224	1.932	3156	50,9	-	-	454	[39]
Trypanosoma cruzi marinkellei										
B7	--	34,2	23154	2846	2511	50,9	10117	10104	454 + Illumina	[84]

Table 1.
Genomes of Trypanosoma cruzi

In *T. cruzi*, since CCNV analysis deeply depends on high quality, chromosome-level assembled reference genomes, it was extremely difficult to implement. However, in spite of this limitation, some approaches were done using CLBrener genome as reference [38]. Taking into account the poorly assemble reference

genome at that moment, and the repetitive nature of *T. cruzi* genome, only reads with high mapping quality were used in CCNV estimations. The single-copy genes ploidy estimation (SCoPE) was the methodology utilized by the authors. In this methodology, estimation of chromosomal somy is based on the ratio between the mean coverage of all single-copy genes (unique genomic sequences) in a given chromosome and the genome coverage. After including several *T. cruzi* strains from different DTUs, authors proposed that—as was observed in *Leishmania*—the aneuploidy pattern varies among and within *T. cruzi* lineages. In addition, as observed with PFGE, CCNV is considerably frequent between *T. cruzi* strains, including those within a same DTU. Authors propose that TcI appears to be more stable, and TcII had large differences between strains, suggesting that this mechanism is widely used by the parasite to expand groups of genes [39]. Nevertheless, unlike *L. donovani*, CCNV on *T. cruzi* seems to be stable on parasite population, at least for TcII analysis on Y strain and derived clones [38].

2.3 Genome size

The genome size of *T. cruzi* has been estimated by different methodologies such as flow cytometry, renaturation kinetic analysis, microfluorometry, chemical analysis, molecular karyotyping and genome sequencing. Every approach agreed on that *T. cruzi* genome size is variable. Polymorphism has been shown between DTUs, between strains within the same DTU, and even between isolates from the same strain [40–47]. From a wide genome size quantification and analysis including more than fifty strains from DTUs TcI to TcVI [46] it was found that: (i) maximum difference observed between strains was 47.5%; (ii) TcI was the smallest genome, (iii) TcV and TcVI were the least variable, (iv) parental genomes mean gene content (TcI: 88.4 Mb, TcII: 106.5 Mb, TcIII: 119.2 Mb), and similar results on the reduced size of TcI, with few exceptions was further observed [47].

Genome size estimation by bioinformatic analysis of NGS data, as was mentioned before, is hampered due to the massive presence of repetitive sequence regions, which reach up to 50% of the genome [25, 48]. This generates assembly fragmentation and collapse—gene and repetitive sequences, leading to copy number underestimation—which represents a challenge to the correct genome size estimation. In fact, as reflected on **Table 1**, the assembly size is far below the estimations made by DNA measurements methods. Only third generation sequenced genomes appear to represent more accurate figures [28, 29].

3. Genome architecture and composition

The publication the of first *T. cruzi* genome in 2005 [25] was a cornerstone of the study of its genome complexity. Although the CL-Brener sequenced strain turned out to be a hybrid that made the analyzes more arduous, at that time it was corroborated that more than 50% of the genome of *T. cruzi* corresponds to repetitive sequences—mainly retrotrasposons, multigenic families and tandem repeats—including the discovery of the new gene family of a new family of mucin associated surface proteins (MASP). Around 12,500 genes could be identified, but the assembly was fragmented into more than 5400 scaffolds (ordered contigs usually joined with unknown sequences filled as "N"), and the complete sequence of the genome was not obtained, being the total genome size about 67 Mb (half of it corresponding to each haplotype). Later on, based on the scaffolds already defined [25], BAC ends sequencing and synteny maps with *T. brucei*, it was possible to recover full length pseudo-chromosomes [49], although

still maintaining thousands of sequences as "unassigned contigs." Since these initial publications, several *T. cruzi* genomes have been reported to be sequenced by NGS, although massive sequencing could not improve the low resolution in complex and highly fragmented regions (**Table 1** and cites therein).

The advent of long read sequencing technologies helped to tackle part of the assembly fragmentation issue, and to better understand *T. cruzi* genome complexity. In 2018 the genomes of two *T. cruzi* strains (Dm28c and TCC, belonging to TcI and TcVI respectively) were sequenced by using Pacbio technology, showing substantial improvements: assemblies of Dm28c and TCC were of 53.2 and 86.7 Mb distributed in 599 and 1142 contigs, respectively, which implied a high reduction of fragmentation [28] (see N50 stats, **Table 1**). Completeness of these genomes was achieved, obtaining for the case of Dm28c all its haploid genome, totaling 53.3 Mb. This size is consistent with the most precise estimations made by fluorescent nucleic acid dye [47]. For the hybrid strain TCC, composed of two relatively divergent parental lineages, it is assumed that the diploid size that includes both parental haplotypes should be recovered, i.e., 106–122 Mb for TCC [46, 47], which compared with the 86.7 Mb indicates that segregation cannot be achieved in those regions with high identity. The ability to separate haplotypes opens new possibilities for the study of the evolutionary processes that occurred in *T. cruzi* and can be useful to provide insights on how hybrids were generated and evolved. Moreover, recombination events can be identified and studied [28]. The hybrid strain Bug 2148 (TcV) was recently long-read sequenced and assembled in 934 contigs, also resolving the fragmentation in a large degree; although the expected genome size is 106–135 Mb [46, 47], the total assembly size is 55.2 Mb and it is striking that there is no evidence of haplotype separation [29] as would be expected for a hybrid strain.

Even using this new technology, these assemblies still have some fragmentation mainly due to the size of the tandem repeats. In particular, the well-characterized 195 bp satellite that can reach clusters of 50 kb, contributes as a major factor to assembly fragmentation avoiding its complete resolution [50–52]. In fact, these genomes contain several contigs entirely composed of this repeat, which together encompasses more than 5% of the genome (see below).

3.1 Genome compartments and gene composition

Since genomic annotation, especially in *T. cruzi*, is arduous and often the goal of genomic sequencing escapes the annotation, it has not been performed in all genomes. For those genomic projects of *T. cruzi* that have the annotation (see **Table 1**) quite similar number of coding genes per haplotype was determined, a minimum of ~10,800 for Sylvio [53] and a maximum of 15,300 for Dm28c [28]. These genes can be divided into two large groups, those of well conserved core genes, and those coding for the multigenic surface families, several of which are unique for *T. cruzi* (see below). In fact, the improvements in the assemblies allowed us to determine that the genome of *T. cruzi* is composed of two compartments. These compartments, called "core" and "disruptive" [28] vary in gene content and nucleotide composition. The "core compartment" is composed of conserved and hypothetical conserved genes, it has a lower GC content (48%) and exhibits synteny conservation with *T. brucei* and *L. major*, whereas the "disruptive compartment" is mainly composed by the surface multigene families trans-sialidase, MASP, and mucins, and exhibits a higher GC content (53%).

3.2 Gene organization

As mentioned, genes in trypanosomatids are organized into non-overlapping clusters on the same DNA strand with unrelated predicted functions. Genes are

transcribed as polycistrons and subsequently trans-spliced and polyadenylated. In *T. cruzi* gene clusters can range from ~30 to 500 kb separated by divergent or convergent strand-switch regions (SSR) [54]. Although no evidence of shared consensus motive or patterns has been found among them, the SSR are functionally active. For instance, transcription initiation and termination take place [2, 55, 56], but it is also observed that they are involved in the origin of DNA replication [57], and centromeric function [58, 59]. The SSRs exhibit some properties such as a different composition in comparison to the rest of the genome and higher intrinsic curvature [60, 61], associated in turn with transcriptional regulation. Indeed SSRs from the disruptive compartment are longer than those from the core compartment (mean length ~4.5kb and ~1.5kb respectively).

4. *Trypanosoma cruzi* repetitive genome

One of the outstanding features of the *T. cruzi* genome is its repetitive nature. Three types of sequences contribute to this characteristic: multigenic families, retrotransposons and satellite DNA (tandem repeat sequences).

4.1 Multigene Families

A main characteristic of *T. cruzi* genome is the large number of multigene families, many of them having hundreds of members. The largest families in *T. cruzi* genome are shown in **Table 2**. TS, Mucins and MASP are located in the disruptive compartment of the genome, whereas GP63, DGF-1 and RHS are distributed in both compartments [28]. We will focus on families from the disruptive compartment (MASP, Mucins and TS), and GP63 as an example of a very expanded family in *T. cruzi*. It is noteworthy that these families code for proteins directly involved in interaction with the host, both at the cellular level (adhesion, invasion, infection) and in immune modulation responses, mainly because most of TS, Mucins, MASP and GP63 proteins are GPI anchored, i.e., they are constitutive part of the functionally relevant cell surface of *T. cruzi*.

Family*	Dm28c	TCC
trans-sialidase (TS)	1491	1734
MASP	1045	1332
RHS	784	1222
Mucins	574	1018
GP63	378	710
DGF-1	215	491
UDP-Gal or UDP-GlcNAc-dependent glycosyltransferase	115	118
Elongation factor (1-alpha, 1-gamma and 2)	81	167
Glutamamyl carboxypeptidase	71	86
Protein Associated with Differentiation	61	60
Kinesin	58	81
TASV	45	92
Syntaxin binding protein	45	85
Heat shock protein 70	35	43

*including pseudogenes

Table 2.
Gene families groups in T. cruzi.

4.1.1 Trans-sialidases

Trans-sialidases and trans-sialidase-like proteins (TS) constitute a large and polymorphic superfamily [25, 28, 29] whose name comes from the ability to transfer sialic acid from host glycoconjugates to parasite's mucins [70, 62]. This activity is highly relevant since *T. cruzi* is unable to synthesizes sialic acid *de novo*, and sialic acid containing glycoproteins are demonstrated to be relevant for infection [70, 62]. However, only a very few members of TS family are predicted to be enzymatically active [29], whereas the rest of them have other relevant roles such as binding to host molecules, immunomodulation, apoptosis or invasion [64]. It should be very important to rename this family since its current denomination leads to confusion. The hallmark of the family is the presence of the canonical amino acid motif VTVXNVXLYNR, although some members have a degenerated version of it [64]. TS proteins can be secreted or membrane anchored, in which case they exhibit an *N*-terminal signal peptide and GPI signal sequence at the *C*-terminal region of the protein. Genomic analysis of TS gene family in CL-Brener revealed that TS family was clustered in eight groups, which were classified by the presence or absence of additional motifs like FRIP, Asp box and the SAPA [65, 66]. In this classification the Group I is defined as those sequences with a predicted enzymatic activity, which corresponds to 4% of the total TS genes [67]. By long read sequencing, a more precise gene copy number could be determined on TCC and Dm28c strains: 1734 and 1491 TS genes respectively; with these new protein sequences the classification should be updated. Draws the attention that both strains exhibit a substantially high percentage of pseudogenes: 41.6% in TCC and 38% in Dm28c [28], which suggest that they could not constitute "inert material." This point deserves further studies to determine if pseudogenes are expressed, and/or can constitute a source of variability, among their possible functions. Most of TS genes are overexpressed in trypomastigotes, but a small percentage are upregulated in amastigotes or epimastigotes at the transcriptional level [68].

4.1.2 Mucins

Mucins and mucin like glycoproteins are the main acceptors of sialic acid through the trans-sialidase TS activity [69], and participate in adhesion, protection against lysis, invasion and immune evasion [70]. The first mucin-like gene cloned and the predicted protein exhibited an internal tandem repeat with the canonical sequence T_8LP_2, flanked by an *N*-terminal signal peptide and a *C*-terminal GPI anchor signal sequence. Further studies revealed the presence of a complex family with genes coding for proteins with similar N and C termini but with non-repetitive, variable and serine and threonine rich domains, also classified as mucins. Those groups with repetitive domains and without repetitive domains were designated TcMUCI and TcMUCII [10], and the presence of a mosaic sequences between both groups led to the proposal of a common ancestor and further diversification [70]. Another group of smaller mucin genes, TcSMUG [71], are expressed in the insect stages, and were subclassified in large and small TcSMUG (L and S) [70]. Due to the complexity of this family manual curation is needed for annotation of these genes. Our group used the following criteria: genes exhibiting an *N*-terminal signal peptide, a C-terminal GPI anchor signaling, and T rich sequences such as T_8KP_2, $T_{6\text{-}8}$KAP or $T_{6\text{-}8}$QAP, finding 1018 and 574 mucin genes in TCC and Dm28c respectively [28], and around 20% were classified as pseudogenes in both strains. Regarding the expression of TcMUC and TcSMUG in life stages of *Trypanosoma cruzi*, trypomastigotes presented higher expression levels of both TcMUC groups,

and in contrast with biochemical reports [70], in amastigotes the highest expressed mucins belong to TcMUCII instead of TcMUCI [68].

4.1.3 MASPs

One of the most surprising result after assembly and annotation of the first *T. cruzi* genome [25], was the discovery of a new gene family composed of approximately 1300 genes, and named as mucin associated surface protein (MASP), because of their location in proximity or clustered with mucin genes. MASP family is characterized by conserved *N*-terminal signal peptide, a conserved *C*-terminal domain containing a GPI anchor addition site, and a variable central region [25]. One of the proposed roles of this gene family is the immune system evasion during the acute phase of Chagas disease [72]. CL-Brener clone contains 1377 *masp* genes, among which 771 appear to be intact genes and 433 (31%) are pseudogenes [25], and analysis in Dm28c and TCC yield similar results: 1045 and 1332 genes where 36 and 33% respectively are pseudogenes [28]. Regarding the expression of this gene family, 97% of *masp* genes are upregulated in trypomastigotes, and a discrete number of genes are expressed specifically in amastigotes or epimastigotes [68].

4.1.4 GP63

GP63 are GPI anchored metalloproteases present in the Tritryps. However, whereas *L. major* contains six *gp63* genes and *T. brucei* has thirteen copies, in *T. cruzi* this family is widely expanded: 400 genes or pseudogenes were identified in CL-Brener [25] and Dm28c [28], and more than 700 in TCC [28]. Strikingly, more than 60% of these genes on the three strains are annotated as pseudogenes. Although the role of this family in innate immune evasion and invasion, has been extensively studied in *Leishmania* [73, 74], little is known about its role in *T. cruzi*. The reason of the expansion of this gene family in the *T. cruzi* genome remains to be elucidated as well as its role on this parasite. Transcriptomic analysis revealed that most of the members are highly expressed in trypomastigotes, whereas a few genes are expressed almost exclusively in amastigotes. Interestingly, phylogenetic analysis using 3′ UTR sequences of *gp63* genes showed three groups of sequences clearly distinguished; one group associated with genes highly expressed in trypomastigotes, another one with genes highly expressed in amastigotes, and a third group of genes with almost no expression in any stage of the parasite [68]. This result strongly supports a major role of the 3′ UTR in posttranscriptional regulation of this family that deserves further studies.

4.2 Transposable elements

Transposable elements (TEs) are repeated DNA sequences, which have the ability to move from one to another *locus* in the genome. This was why they have been referred to as "junk" DNA, selfish sequences or genomic parasites. However, growing evidence is indicating the great importance that TEs play in the evolution of genes and genomes in a wide range of organisms, including trypanosomatids [75, 76]. *T. cruzi* genome lacks class II elements (DNA transposons), bearing only class I retroelements. Within them—according to Wicker [77] TEs classification—*T. cruzi* presents three autonomous families: VIPER, a tyrosine recombinase (YR) element which belongs to the DIRS order; L1Tc, a non-LTR element of the *ingi* clade; and CZAR, also a non-LTR element from the CRE clade which is site-specific, inserting only on the SL gene [25, 76, 78]. On the other hand, non-autonomous elements have been also

identified. SIRE, have similarity with the VIPER 5′ and 3′ ends, resembling what nowadays are called solo-LTR. NARTc is the non-autonomous couple of L1Tc elements, as has been classically described for LINE/SINE-like couples. Finally, TcTREZO has been described as another site-specific retroelement, inserted within *masp* genes [79]. Although it has been characterized as a non-LTR retroelement due to the presence of a poly-A tail and a secondary structure which will be promoting its retrotranscription, no conserved domains have been detected on this element. Hence, TcTREZO could be an ancient non-autonomous retroelement. All of the VIPER, CZAR and TcTREZO copies are defective (no complete domains where found), whereas L1Tc was the only one which showed putative active copies [28].

4.3 Tandem repeats

Although NGS platforms implied an enormous progress for our knowledge about genomes composition and evolution, tandem repeats were not that benefited. Tandem repeats are commonly classified in micro, mini and macro-satellite, depending on their monomer or cluster length. Microsatellites are those whose monomers are from 2 to 5 bp, minisatellites from 15 to 100 bp, and finally macrosatellites or just called satellites involves repeats greater than 100 bp [80]. Even with very deep genome coverage, short read lengths cause problems for *de novo* assemblies, especially in tandem repeat rich regions. Due to this trouble, tandem repeats can be considered as neglected sequences in the majority of genome analyses. Although great efforts were done, fragmentation of the genome assembly occurs frequently where repeated sequences are located. In fact, the massive major 195 bp satellite (TcSAT1 named in repbase) described for the first time by Sloof et al. [50], represents a huge challenge for contig assembly. Although PacBio reads enable to develop an improved assembly and characterization of tandem repeats characterization and assembly, the size of some clusters exceeds that of the reads. In fact, some small-size contigs (50 kb) are composed entirely by the 195 bp satellite sequence.

In summary, genomic studies are essential for understanding *Trypanosoma cruzi* biology, and the new technologies will give responses to still unanswered questions: Which molecular mechanisms allow to regulate specific genes, without consensus sequences? Is *Trypanosoma cruzi* a unique species? How many chromosomes do they have? How are chromosomes organized? Which role plays the highly repetitive genome on its plasticity? And we can continue, to reinforce the idea that much remains to be done.

Financial support

This work was supported by Institut Pasteur de Montevideo (S.P. postdoctoral fellowship) from UK Research and Innovation via the Global Challenges Research Fund under grant agreement 'A Global Network for Neglected Tropical Diseases' grant number MR/P027989/1. LB, APT, FAV and CR are members of the Sistema Nacional de Investigadores (SNI-ANII, Uruguay).

Author details

Luisa Berná[1], Sebastián Pita[1,2], María Laura Chiribao[1,3], Adriana Parodi-Talice[1,2], Fernando Alvarez-Valin[4] and Carlos Robello[1,3]*

1 Laboratory of Host Pathogen Interactions-UBM, Institut Pasteur de Montevideo, Uruguay

2 Sección Genética, Facultad de Ciencias, Universidad de la República, Uruguay

3 Departamento de Bioquímica, Facultad de Medicina, Universidad de la República, Uruguay

4 Sección Biomatemática, Facultad de Ciencias-Universidad de la República, Uruguay

*Address all correspondence to: robello@pasteur.edu.uy

References

[1] Vanhamme L, Pays E. Control of gene expression in trypanosomes. Microbiology and Molecular Biology Reviews. 1995;**59**(2):223-240

[2] Martínez-Calvillo S et al. Gene expression in trypanosomatid parasites. BioMed Research International. 2010;**2010**

[3] Kramer S. Developmental regulation of gene expression in the absence of transcriptional control: The case of kinetoplastids. Molecular and Biochemical Parasitology. 2012;**181**(2):61-72

[4] Sutton RE, Boothroyd JC. Evidence for trans splicing in trypanosomes. Cell. 1986;**47**(4):527-535

[5] Michaeli S. Trans-splicing in trypanosomes: Machinery and its impact on the parasite transcriptome. Future Microbiology. 2011;**6**(4):459-474

[6] Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa: Why be normal? Microbes and Infection. 2003;**5**(13):1231-1240

[7] Freistadt MS et al. Direct analysis of the mini-exon donor RNA of *Trypanosoma brucei*: Detection of a novel cap structure also present in messenger RNA. Nucleic Acids Research. 1987;**15**(23):9861-9879

[8] Clayton CE. Gene expression in kinetoplastids. Current Opinion in Microbiology. 2016;**32**:46-51

[9] da Silva RA, Bartholomeu DC, Teixeira SM. Control mechanisms of tubulin gene expression in *Trypanosoma cruzi*. International Journal for Parasitology. 2006;**36**(1):87-96

[10] Di Noia JM et al. AU-rich elements in the 3′-untranslated region of a new mucin-type gene family of *Trypanosoma cruzi* confers mRNA instability and modulates translation efficiency. The Journal of Biological Chemistry. 2000;**275**(14):10218-10227

[11] Jager AV, Muia RP, Campetella O. Stage-specific expression of *Trypanosoma cruzi* trans-sialidase involves highly conserved 3′ untranslated regions. FEMS Microbiology Letters. 2008;**283**(2):182-188

[12] Tibayrenc M et al. Genetic characterization of six parasitic protozoa: Parity between random-primer DNA typing and multilocus enzyme electrophoresis. Proceedings of the National Academy of Sciences. 1993;**90**(4):1335-1339

[13] Robello C et al. Evolutionary relationships in *Trypanosoma cruzi*: Molecular phylogenetics supports the existence of a new major lineage of strains. Gene. 2000;**246**(1-2):331-338

[14] Zingales B et al. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: Second revision meeting recommends TcI to TcVI. Memórias do Instituto Oswaldo Cruz. 2009;**104**(7):1051-1054

[15] Marcili A et al. A new genotype of *Trypanosoma cruzi* associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA. Parasitology. 2009;**136**(6):641-655

[16] Zingales B et al. The revised *Trypanosoma cruzi* subspecific nomenclature: Rationale, epidemiological relevance and research applications. Infection, Genetics and Evolution. 2012;**12**(2):240-253

[17] Henriksson J, Åslund L, Pettersson U. Karyotype variability in

Trypanosoma cruzi. Parasitology Today. 1996;**12**(3):108-114

[18] Santos MR et al. The *Trypanosoma cruzi* genome project: Nuclear karyotype and gene mapping of clone CL Brener. Memórias do Instituto Oswaldo Cruz. 1997;**92**(6):821-828

[19] Henriksson J et al. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 1995;**73**(1-2):63-74

[20] Gibson WC, Miles MA. The karyotype and ploidy of *Trypanosoma cruzi*. The EMBO Journal. 1986;**5**(6):1299-1305

[21] Henriksson J et al. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is evolutionarily informative. Parasitology. 2002;**124**(3):277-286

[22] Vargas N, Pedroso A, Zingales B. Chromosomal polymorphism, gene synteny and genome size in *T. cruzi* I and *T. cruzi* II groups. Molecular and Biochemical Parasitology. 2004;**138**(1):131-141

[23] Cano MI et al. Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* genome project. Molecular and Biochemical Parasitology. 1995;**71**(2):273-278

[24] Berriman M et al. The genome of the African trypanosome *Trypanosoma brucei*. Science. 2005;**309**(5733):416-422

[25] El-Sayed NM et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science. 2005;**309**(5733):409-415

[26] Ivens AC et al. The genome of the kinetoplastid parasite, *Leishmania* major. Science. 2005;**309**(5733):436-442

[27] El-Sayed NM et al. Comparative genomics of trypanosomatid parasitic protozoa. Science. 2005;**309**(5733):404-409

[28] Berná L et al. Expanding an expanded genome: Long-read sequencing of *Trypanosoma cruzi*. Microbial Genomics. 2018;**4**(5)

[29] Callejas-Hernández F, Gironès N, Fresno M. Genome Sequence of *Trypanosoma cruzi* Strain Bug2148. Genome Announcements. 2018;**6**(3):e01497-e01417

[30] Parmar JJ, Woringer M, Zimmer C. How the genome folds: The biophysics of four-dimensional chromatin organization. Annual Review of Biophysics. 2019;**48**

[31] Downing T et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Research. 2011;**21**(12):2143-2156

[32] Dujardin J-C et al. Mosaic aneuploidy in *Leishmania*: The perspective of whole genome sequencing. Trends in Parasitology. 2014;**30**(12):554-555

[33] Mannaert A et al. Adaptive mechanisms in pathogens: Universal aneuploidy in *Leishmania*. Trends in Parasitology. 2012;**28**(9):370-376

[34] Almeida LV et al. Chromosomal copy number variation analysis by next generation sequencing confirms ploidy stability in *Trypanosoma brucei* subspecies. Microbial Genomics. 2018;**4**(10)

[35] Borst P et al. On the DNA content and ploidy of trypanosomes. Molecular and Biochemical Parasitology. 1982;**6**(1):13-23

[36] Hope M et al. Analysis of ploidy (in megabase chromosomes) in

Trypanosoma brucei after genetic exchange. Molecular and Biochemical Parasitology. 1999;**104**(1):1-9

[37] Tihon E et al. Evidence for viable and stable triploid *Trypanosoma* congolense parasites. Parasites & Vectors. 2017;**10**(1):468

[38] Reis-Cunha JL et al. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains. BMC Genomics. 2015;**16**(1):499

[39] Reis-Cunha JL, Valdivia HO, Bartholomeu DC. Gene and chromosomal copy number variations as an adaptive mechanism towards a parasitic lifestyle in trypanosomatids. Current Genomics. 2018; **19**(2):87-97

[40] Castro C, Craig SP, Castañeda M. Genome organization and ploidy number in *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 1981;**4**(5-6):273-282

[41] Dvorak JA et al. *Trypanosoma cruzi*: Flow cytometric analysis. I. Analysis of total DNA/organism by means of mithramycin-induced fluorescence 1, 2. The Journal of Protozoology. 1982;**29**(3):430-437

[42] Kooy RF et al. On the DNA content of *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 1989;**36**(1):73-76

[43] Lanar DE, Levy LS, Manning JE. Complexity and content of the DNA and RNA in *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 1981;**3**(5):327-341

[44] McDaniel JP, Dvorak JA. Identification, isolation, and characterization of naturally-occurring *Trypanosoma cruzi* variants. Molecular and Biochemical Parasitology. 1993;**57**(2):213-222

[45] Thompson CT, Dvorak JA. Quantitation of total DNA per cell in an exponentially growing population using the diphenylamine reaction and flow cytometry. Analytical Biochemistry. 1989;**177**(2):353-357

[46] Lewis MD et al. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. International Journal for Parasitology. 2009;**39**(12):1305-1317

[47] Souza RT et al. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. PLoS One. 2011;**6**(8):e23042

[48] Pita S et al. The Tritryps comparative repeatome: Insights on repetitive element evolution in Trypanosomatid pathogens. Genome Biology and Evolution. 2019;**11**(2):546-551

[49] Weatherly DB, Boehlke C, Tarleton RL. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. BMC Genomics. 2009;**10**(1):255

[50] Sloof P et al. Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*. Journal of Molecular Biology. 1983;**167**(1):1-21

[51] Gonzalez A et al. Minichromosomal repetitive DNA in *Trypanosoma cruzi*: Its use in a high-sensitivity parasite detection assay. Proceedings of the National Academy of Sciences. 1984;**81**(11):3356-3360

[52] Elias MCQ et al. Organization of satellite DNA in the genome of *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 2003;**129**(1):1-9

[53] Franzén O et al. Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio

X10/1 and comparison with T. cruzi VI CL Brener. PLoS Neglected Tropical Diseases. 2011;**5**(3):e984

[54] Myler PJ et al. *Leishmania* major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proceedings of the National Academy of Sciences. 1999;**96**(6):2902-2906

[55] Tosato V et al. Secondary DNA structure analysis of the coding strand switch regions of five *Leishmania* major Friedlin chromosomes. Current Genetics. 2001;**40**(3):186-194

[56] Worthey E et al. *Leishmania* major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene. Nucleic Acids Research. 2003;**31**(14):4201-4210

[57] McDonagh PD, Myler PJ, Stuart K. The unusual gene organization of *Leishmania* major chromosome 1 may reflect novel transcription processes. Nucleic Acids Research. 2000;**28**(14):2800-2803

[58] Obado SO et al. Functional mapping of a trypanosome centromere by chromosome fragmentation identifies a 16-kb GC-rich transcriptional "strand-switch" domain as a major feature. Genome Research. 2005;**15**(1):36-43

[59] Obado SO et al. Repetitive DNA is associated with centromeric domains in Trypanosoma brucei but not *Trypanosoma cruzi*. Genome Biology. 2007;**8**(3):R37

[60] El-Sayed NM et al. The sequence and analysis of *Trypanosoma brucei* chromosome II. Nucleic Acids Research. 2003;**31**(16):4856-4863

[61] Smircich P, El-Sayed NM, Garat B. Intrinsic DNA curvature in trypanosomes. BMC Research Notes. 2017;**10**(1):585

[62] Schenkman S et al. A novel cell surface trans-sialidase of *Trypanosoma cruzi* generates a stage-specific epitope required for invasion of mammalian cells. Cell. 1991;**65**(7):1117-1125

[63] Schenkman S et al. Mucin-like glycoproteins linked to the membrane by glycosylphosphatidylinositol anchor are the major acceptors of sialic acid in a reaction catalyzed by trans-sialidase in metacyclic forms of *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 1993;**59**(2):293-303

[64] Freire-de-Lima L et al. The trans-sialidase, the major *Trypanosoma cruzi* virulence factor: Three decades of studies. Glycobiology. 2015;**25**(11):1142-1149

[65] Buscaglia CA et al. Tandem amino acid repeats from *Trypanosoma cruzi* shed antigens increase the half-life of proteins in blood. Blood. 1999;**93**(6):2025-2032

[66] Cazzulo J, Frasch A. SAPA/trans-sialidase and cruzipain: Two antigens from *Trypanosoma cruzi* contain immunodominant but enzymatically inactive domains. The FASEB Journal. 1992;**6**(14):3259-3264

[67] Freitas LM et al. Genomic analyses, gene expression and antigenic profile of the trans-sialidase Superfamily of *Trypanosoma cruzi* reveal an undetected level of complexity. PLoS One. 2011;**6**(10):e25914

[68] Berna L et al. Transcriptomic analysis reveals metabolic switches and surface remodeling as key processes for stage transition in *Trypanosoma cruzi*. PeerJ. 2017;**5**:e3017

[69] Acosta A, Schenkman RP, Schenkman S. Sialic acid acceptors of different stages of *Trypanosoma cruzi* are mucin-like glycoproteins linked to the parasite membrane by GPI anchors. Brazilian Journal of Medical and Biological Research. 1994;**27**(2):439-442

[70] Buscaglia CA et al. *Trypanosoma cruzi* surface mucins: Host-dependent coat diversity. Nature Reviews Microbiology. 2006;**4**(3):229-236

[71] Urban I et al. Molecular diversity of the *Trypanosoma cruzi* TcSMUG family of mucin genes and proteins. The Biochemical Journal. 2011;**438**(2):303-313

[72] dos Santos SL et al. The MASP Family of *Trypanosoma cruzi*: Changes in gene expression and antigenic profile during the acute phase of experimental infection. PLoS Neglected Tropical Diseases. 2012;**6**(8)

[73] Yao C, Donelson JE, Wilson ME. The major surface protease (MSP or GP63) of *Leishmania* sp. biosynthesis, regulation of expression, and function. Molecular and Biochemical Parasitology. 2003;**132**(1):1-16

[74] Brittingham A et al. Role of the *Leishmania* surface protease gp63 in complement fixation, cell adhesion, and resistance to complement-mediated lysis. The Journal of Immunology. 1995;**155**(6):3102-3111

[75] Bringaud F et al. Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. PLoS Pathogens. 2007;**3**(9):e136

[76] Thomas MC et al. The biology and evolution of transposable elements in parasites. Trends in Parasitology. 2010;**26**(7):350-362

[77] Wicker T et al. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;**8**(12):973

[78] Bringaud F et al. Role of transposable elements in trypanosomatids. Microbes and Infection. 2008;**10**(6):575-581

[79] Souza RT et al. New *Trypanosoma cruzi* repeated element that shows site specificity for insertion. Eukaryotic Cell. 2007;**6**(7):1228-1238

[80] Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;**371**(6494):215

[81] Grisard EC et al. *Trypanosoma cruzi* clone Dm28c draft genome sequence. Genome Announcements. 2014;**2**(1):pii:e01114-13

[82] Bradwell KR et al. Genomic comparison of *Trypanosoma conorhini* and *Trypanosoma rangeli* to *Trypanosoma cruzi* strains of high and low virulence. BMC Genomics. 2018;**19**(1):770

[83] Baptista RP et al. Assembly of highly repetitive genomes using short reads: The genome of discrete typing unit III *Trypanosoma cruzi* strain 231. Microbial Genomics. 2018;**4**(4)

[84] Franzén O et al. Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi* marinkellei. BMC Genomics. 2012;**13**:531