

Variability of the Pr77 sequence of L1Tc retrotransposon among six *T. cruzi* strains belonging to different discrete typing units (DTUs)

Inmaculada Gómez^a, Manuel Carlos López^a, Alberto Rastrojo^b, Fabián Lorenzo-Díaz^c, José María Requena^b, Begoña Aguado^b, Basilio Valladares^c, M. Carmen Thomas^{a,*}

^a Instituto de Parasitología y Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas; PTS-Granada, Spain

^b Centro de Biología Molecular Severo-Ochoa (CBMSO) (CSIC-UAM), Consejo Superior de Investigaciones Científicas, Universidad Autónoma de Madrid, Madrid, Spain

^c Instituto Universitario de Enfermedades Tropicales y Salud Pública de Canarias. Universidad de La Laguna. La Laguna, Spain

ARTICLE INFO

Keywords:

Trypanosoma cruzi
Chagas disease
Pr77-hallmark
genomic variability
retroelements
Discrete Type Units

ABSTRACT

All trypanosomatid genomes are colonized by non-LTR retrotransposons which exhibit a highly conserved 77-nt sequence at their 5' ends, known as the Pr77-hallmark (Pr77). The wide distribution of Pr77 is expected to be related to the gene regulation processes in these organisms as it has promoter and HDV-like ribozyme activities at the DNA and RNA levels, respectively. The identification of Pr77 hallmark-bearing retrotransposons and the study of the associations of mobile elements with relevant genes have been analyzed in the genomes of six strains of *Trypanosoma cruzi* belonging to different discrete typing units (DTUs) and with different geographical origins and host/vectors. The genomes have been sequenced, assembled and annotated. BUSCO analyses indicated a good quality for the assemblies that were used in comparative analyses. The results show differences among the six genomes in the copy number of genes related to virulence processes, the abundance of retrotransposons bearing the Pr77 sequence and the presence of the Pr77 hallmarks not associated with retroelements. The analyses also show frequent associations of Pr77-bearing retrotransposons and single Pr77 hallmarks with genes coding for trans-sialidases, RHS, MASP or hypothetical proteins, showing variable proportion depending on the type of retroelement, gene class and parasite strain. These differences in the genomic distribution of active retroelements and other Pr77-containing elements have shaped the genome architecture of these six strains and might be contributing to the phenotypic variability existing among them.

1. Introduction

Trypanosoma cruzi is a protozoan parasite that causes Chagas disease, also known as American trypanosomiasis. The disease represents a major public health challenge in Latin America, where 6 to 7 million people are infected (World Health Organization, 2020), with a growing number of cases in nonendemic countries due to the increase in population migration (Gascon et al., 2010). The parasite is mainly transmitted to humans through the feces or urine of hematophagous triatomine bugs as well as nonvectorial routes such as mother to fetus, blood transfusion, organ transplantation or consumption of contaminated food. Chagas disease shows variable clinical features ranging from asymptomatic to chronic disease, including cardiac, digestive, neurological, and mixed presentations (Rassi Jr. et al., 2012). Nevertheless, to date, it has not been possible to determine an association between the *T. cruzi* genotype and the manifestation of Chagas disease in the chronic

phase due to the scarcity of parasites in blood and the difficulty of sampling tissue parasites (Messenger et al., 2015; Zingales, 2018).

T. cruzi strains have been subdivided into discrete typing units (DTUs) designated as TcI to TcVI and Tcbat due to their high genotypic and phenotypic heterogeneity (Lima et al., 2015; Zingales et al., 2012). Parasites in all DTUs participate in sylvatic and domestic cycles and show differences in their distribution between insect and mammalian host species and their geographic location (Brenière et al., 2016; Zingales et al., 2012). The different DTUs have been reported to show striking cellular and molecular peculiarities regarding tissue and organ tropisms, antigenicity, degree of virulence, drug susceptibility, ability to infect vectors, number of chromosomes, and gene content (Macedo et al., 2004; Macedo and Pena, 1998; Zingales et al., 2012). TcI presents a high genetic heterogeneity, in which five intra-DTU TcI genotypes (Ia, Ib, Ic, Id, Ie) have been described based on sequence polymorphisms in the mini-exon intergenic region (Cura et al., 2010; Falla et al., 2009). It

* Corresponding author at: Instituto de Parasitología y Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas; PTS-Granada, Spain.

E-mail address: mcthomas@ipb.csic.es (M.C. Thomas).

<https://doi.org/10.1016/j.actatropica.2021.106053>

Received 5 December 2020; Received in revised form 15 June 2021; Accepted 11 July 2021

Available online 15 July 2021

0001-706X/© 2021 The Authors.

Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

is the most abundant and widely distributed *T. cruzi* lineage, and it is associated with chagasic cardiomyopathy (Zingales, 2018; Zingales et al., 2012). TcII has a more limited geographical distribution and is associated with cardiac symptomatology and alterations of the meg-esophagus and megacolon (Zingales et al., 2012). TcIII is mostly associated with the sylvatic cycle and terrestrial niche in Brazil and adjacent countries and is infrequently found in human infections. TcIV is also predominantly associated with the sylvatic cycle, and TcV and TcVI are associated with domestic cycles (Zingales et al., 2012).

The first draft genome of a *T. cruzi* strain (CL Brener, DTU VI) was reported in 2005 (El-Sayed et al., 2005a). More recently, in parallel to the advances in sequencing methodologies, genomic drafts for new strains have been reported. A salient feature of the *T. cruzi* genome is the tremendous number of repetitive sequences, which are mainly derived from transposable elements (TEs) (Thomas et al., 2010). TEs, also known as LINES (Long Interspersed Nuclear Elements), constitute an important fraction of the genome in most eukaryote organisms. TEs are dynamic elements that reshape host genomes by generating rearrangements with the potential to create or disrupt genes, shuffle existing genes, and modulate their patterns of expression (Thomas et al., 2010). The sequencing of trypanosomatid genomes has revealed that they all contain a large number of retrotransposons (active or degenerate) that together make up to 5% of the nuclear genome (El-Sayed et al., 2005b). Among trypanosomatids, the best characterized LINE is L1Tc of *T. cruzi* (Martin et al., 1995; Thomas et al., 2010) which, together with the ingi element of *T. brucei* (Tbingi) conforms the ingi/L1Tc clade (Bringaud et al., 2006). L1Tc encodes its own retrotransposition machinery: AP endonuclease (Olivares et al., 1997), reverse transcriptase (García-Pérez et al., 2003), RNase H (Olivares et al., 2002) and nucleic acid chaperone protein (Heras et al., 2005). L1Tc and ingi homologous elements were also identified in the *T. vivax* (Tvingi) and *T. congolense* (L1Tco and Tcoingi) genomes (Bringaud et al., 2009). Truncated versions of these LINES were found in the *T. cruzi* (NARTc, Non-Autonomous Retrotransposon in *T. cruzi*) and in *T. brucei*, *T. vivax* and *T. congolense* genomes (TbRIME, TvRIME and TcoRIME, respectively, Ribosomal Mobile Elements), (Bringaud et al., 2009; Bringaud et al., 2002)). In addition, long and short degenerated versions (DIREs and SIDERs) have been identified in most genomes of trypanosomatids (Bringaud et al., 2006). A common feature of retrotransposons and degenerated elements is the presence of a 77-long nucleotide sequence at their 5' end; this sequence is known as Pr77, the Pr77 signature, or the Pr77 hallmark. This Pr77 hallmark has been shown to possess DNA promoter activity at the DNA level (Heras et al., 2007) and RNA HDV-like ribozyme activity at the RNA level (Sánchez-Luque et al., 2011), acting as an active dual system (Sánchez-Luque et al., 2012). L1Tc and homologous elements have been reported to be responsible for spreading the Pr77 active sequence throughout the trypanosomatid genomes (Sánchez-Luque et al., 2014).

To improve our knowledge regarding the *T. cruzi* genome and to better understand the similarities and differences of the Pr77-hallmark bearing retroelements among the different DTUs and strains, herewith we have analyzed the genomic features and the presence and distribution of the L1Tc and NARTc retroelements in the genome of six *T. cruzi* strains. Comparative studies among the different strains have also been performed after genome sequencing, assemblies and gene annotations for the genomes of the B. M. López (TcIa), Dm28 (TcId), Y (TcII), Ikiakarora (TcIII), SOL (TcV) and CL Brener (TcVI) *T. cruzi* strains.

2. Material and methods

2.1. Strains, parasite cultures, DNA isolation and genotyping

Epimastigotes from B. M. López, Dm28, Y, Ikiakarora, SOL, and CL Brener *T. cruzi* strains (see Table 1 for strain details) were grown at 28°C in liver infusion tryptose (LIT) medium supplemented with 10% (v/v) heat-inactivated fetal bovine serum (Flow Laboratory, Irvine, UK). The parasite cultures were collected by centrifugation at the logarithmic

Table 1

DTU, geographical origin, and host/vector of the *T. cruzi* strains used in this study (Hamuy et al., 2013; Higuera et al., 2013; Murcia et al., 2013; Rodríguez et al., 1998).

Strain	DTU	Geographic location	Host/Vector
B. M. López	Ia	Paratebueno, Cundinamarca (Colombia)	<i>Homo sapiens</i>
Dm28	Id	Carabobo (Venezuela)	<i>Didelphis marsupialis</i>
Y	II	Sao Paulo (Brazil)	<i>Homo sapiens</i>
Ikiakarora	III	Catatumbo, Norte Santander (Colombia)	<i>Rhodnius prolixus</i>
SOL	V	Murcia (Spain)	<i>Homo sapiens</i>
CL Brener	VI	Encruzilhada, Rio Grande do Sul (Brazil)	<i>Triatoma infestans</i>

phase of growth ($10\text{--}20 \times 10^6$ parasites/ml), washed twice with cold PBS and lysed by resuspension in PBS-1% NP40. Nuclei were collected by centrifugation at 13000 rpm for 5 minutes and suspended in PBS containing 0.5% SDS. The genomic DNA was purified by phenol-chloroform extraction. The genotyping of the parasites was carried out by PCR amplification of the mini-exon, A10-e fragment, 18S and 24S genes. The PCR amplification of the mini-exon and 24Sα rRNA was performed as described by Brisse and colleagues (Brisse et al., 2001). PCR for amplification of the size-variable domain of the 18S rRNA sequence was carried out as described by Clark and colleagues (Graham Clark and Pung, 1994). Amplification of the A10-e fragment was performed by PCR as described by Brisse and colleagues (Brisse et al., 2000). TcI haplotype characterization was based on nucleotides polymorphisms of the minixon intergenic region as described by (Falla et al., 2009).

2.2. Library preparation and genome sequencing

Whole genomes of *Trypanosoma cruzi* B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener strains were sequenced using Ion Torrent next-generation technology (Thermo Fischer Scientific). Genomic DNA samples (500 ng) in 10 mM Tris pH 7.6 buffer were used for automatic library construction in the AB Library Builder system using the Ion Xpress Plus Library Kit (Thermo Fischer Scientific) following the manufacturer's instructions. Libraries were size-selected to an optimal length range of 450 bp by the E-Gel system and SizeSelect 2% Agarose Gels (Invitrogen). For quality control of DNA samples, concentration, determination and estimation of fragment length distribution, the Agilent 2100 Bioanalyzer system was employed using a high-sensitivity DNA Kit (Agilent Technologies). DNA content was also determined by the Quant-IT dsDNA assay kit and the Qubit Fluorometer system (Invitrogen). DNA samples were then serially diluted to a final concentration of 23 pM and subjected to emulsion PCR and enrichment using the One Touch 2 system (Ion OneTouch 400 Template kit; Life Technologies). The enriched samples were loaded on Ion 316(v2) chips and sequenced on the Ion Torrent PGM™ system with Hi-Q™ Sequencing Chemistry (Thermo Fischer Scientific).

2.3. Quality control of raw reads

Raw reads were analyzed with FastQC v0.10.1 (default settings) (Andrews, 2010). Prinseq v0.20.4 (Schmieder and Edwards, 2011) was used iteratively for quality filtering using the following parameters: -derep 14, -ns_max_p 1 -ns_max_n 3 -trim_ns_left 1 -trim_ns_right 1, -trim_qual_right 20 -trim_qual_type mean -trim_qual_window 5 -trim_qual_step 1, -trim_qual_right 20 -trim_qual_type mean -trim_qual_window 1 -trim_qual_step 1, -trim_qual_left 20 -trim_qual_type mean -trim_qual_window 5 -trim_qual_step 1, -trim_qual_left 20 -trim_qual_type mean -trim_qual_window 1 -trim_qual_step 1, -lc_method entropy -lc_threshold 50, -min_qual_mean 25, -min_len 50.

2.4. Genome assembly

Genome assemblies were performed using CLC Genomic Workbench v8.0 (Qiagen) run with default parameters modifying only the length fraction and similarity fraction, which were set to 0.90 and 0.97, respectively. Additionally, a minimum contig length of 500 bp was established.

2.5. Assessment of genome assembly completeness

BUSCO v4.0.5 (Benchmarking Universal Single-Copy Orthologs) analysis (Seppey et al., 2019) was performed to evaluate genome completeness. The assembled genomes were compared with a pre-defined orthologue set of euglenozoan v10, including 130 BUSCOs (parameters: -m genome, -l euglenozoa_odb10). The program was also executed with automatic detection of the closest lineage, in which euglenozoa were detected (-m genome -autolineage).

2.6. Genome annotation

Gene annotation of each genome was performed using the AUGUSTUS gene-prediction tool (<http://bioinf.uni-greifswald.de/webaugustus/>) as described elsewhere (Hoff and Stanke, 2013). Parameter files, required for 'Augustus Submit Prediction', obtained from available *T. cruzi* reference genomes were generated for each strain in 'Augustus Submit Training'. Selection of the reference genome to be used for gene annotation of each genome was carried out using Best Hit analysis performed by identification of the open reading frames in contigs of each genome followed by BLASTP searching (Altschul et al., 1990) against a protein database created with the proteins downloaded from the *T. cruzi* genomes available at TriTrypDB (<http://tritrypdb.org/>) using a custom Python script (Python.org., 2020). The reference genome (downloaded from TriTrypDB, version 46) for each of the six sequenced genomes was selected based on the highest identity score obtained in each case (Table 3). Thus, Dm28c was selected for annotation of the B. M. López and Dm28 genomes, CL Brener Esmeraldo-like for the Ikiakarora and Y genomes and CL Brener non-Esmeraldo-like for the SOL and CL Brener genomes. Gene prediction was executed using 'Augustus Submit Prediction' by providing the parameter files obtained for each strain. Subsequently, the OrthoMCL workflow (Li et al., 2003) in the VEuPathDB Galaxy Site (<https://veupathdb.globusgenomics.org/>) was used to assign the set of predicted proteins to OrthoMCL groups using OrthoMCL 6r1 proteins blastDB (default settings). The description of the functions of the assigned OrthoMCL groups of proteins was downloaded (https://orthomcl.org/common/downloads/release-6/defines_OrthoMCL-6.txt.gz) and added together with OrthoMCL groups to annotation files using an in-house Python script.

2.7. Analysis of the characteristics of the assembled genomes

For identification of the characteristics and properties of assembled genomes, commands in Linux terminal in Unix OS and customized scripts in Python (Python.org., 2020) and Perl ("The Perl Programming Language - www.perl.org," (The Perl Programming Language, 2020) were used.

2.8. Identification and validation of transposable elements

Probes corresponding to 5' and 3' ends of the L1Tc or NARTc elements were employed for the identification of retrotransposons into reads of the six sequenced genomes. Two independent searches were performed executing a customized Python script. The 5' probe corresponded to the Pr77-hallmark. The Pr77 probe was fragmented into 7-nt overlapping 8-nt-long seeds. The sequences containing the Pr77-hallmark were extracted from the reads and aligned with the Pr77 probe, setting 70% as the identity threshold. The identity threshold was adjusted depending

on the length of the identified sequence according to the following

formula: $\text{mic} = \text{mi} + (1 - \text{mi}) * \left(1 - \left(\frac{\text{hit_length}}{\text{probe_length}}\right)\right)$, where mic is the

minimum corrected identity, mi is the minimum identity (0.7), hit_length is the length of the Pr77-hallmark fragment found in the reading, and probe_length is the length of the searched probe. Only sequences higher than 15 nt were analyzed. Following identification of the Pr77-hallmark, Pr77 downstream sequences were analyzed to classify them as L1Tc or NARTc based on their identity to the 40-nt-long sequences employed as the identifier probes (L1Tc probe, 5'CCCATCCGCTGCCCGCGGAGAGGCAGGAGGCGCCGCACAA3' and NARTc probe 5'TTCATGCTTCAAACCCGATGAGTAGTTGTTACTTAGTTT3'). Only sequences longer than 15 nt and having identity values higher than 70% were retrieved for further classification as L1Tc or NARTc. For identification of the 3' ends, the last 55 nt of L1Tc (5'CCACCTCTTCGGCACTCAGATGGCACTGTATAGCTAGACGCGCTGGTAAGAGTAG3) and NARTc (5'TACGCTTACCAGAGCTGCACAATTGAGGTGTAGTTACTACGACTGGTAAGAGAAG3), located upstream of the poly(A) tails in each case, were used as probes to classify the elements as L1Tc or NARTc. Only sequences with an identity higher than 75% and longer than 15 nucleotides were considered. Finally, the retrieved sequences were validated by BLASTN search (Altschul et al., 1990) against a database composed of the different sequences corresponding to L1Tc (Supplementary file 1) or NARTc (Supplementary file 2) elements using a Python script. A minimum sequence identity of 70% and an alignment coverage $\geq 70\%$ were required to ascribe the elements as *bona fide* L1Tc or NARTc.

2.9. Genomic context of transposable elements

A custom Python script was used for the genomic context analysis. The sequences located upstream of the Pr77-hallmark and downstream of the TEs were extracted. A minimum length of 15 nt outside the TE sequence was required to further analyze a given read. To assign the gene descriptions, the obtained sequences were aligned using bowtie2 (Langmead et al., 2009) v2.2.0 (default parameters) with the fasta files of the assembled genomes and, then, the gtf files of genome annotations.

2.10. Estimation of gene family content

Estimation of the copy number of genes encoding 13 proteins and the percentage occupied by the selected genes in each genome were analyzed in assembled and annotated genomes using a custom Perl script (<https://bitbucket.org/ipbln/AnnotationStats/>) in the .gtf annotation file of each genome. The selected genes were trans-sialidases (TS), mucins, mucin-associated surface proteins (MASP), retrotransposon hot spot proteins (RHS), TASV (Trypomastigote, Alanine, Serine, Valine proteins), kinesins, RNA helicases, protein kinases, glycosyltransferases, cysteine peptidases, heat shock protein 70 (HSP70), Paraflagellar Rod (PFR) Proteins and RNA-binding proteins. The content of genes coding for hypothetical proteins was also analyzed in all the genomes.

2.11. Contig visualization

Integrative Genomics Viewer (IGV) software (Thorvaldsdottir et al., 2013) was used to visualize the alignments of reads on the genomic assemblies.

2.12. Data availability

Trypanosoma cruzi assembled genomes and raw reads have been deposited at GenBank and in the Sequence Read Archive (SRA), respectively, under the following accession numbers: B. M. López (WWP Y000000000; BioProject: PRJNA595079), Dm28 (JACVEY000000000; BioProject: PRJNA661283), Y (JACVEZ000000000; BioProject: PRJNA661288), Ikiakarora (WWPZ000000000; BioProject: PR

JNA595095), SOL (BioProject: PRJNA661295) and CL Brener (BioProject: PRJNA661279).

3. Results

3.1. Genome sequencing, assemblies and annotations in the different *T. cruzi* strains and their comparisons

The aim of this study was to identify genomic changes associated to the Pr77-hallmark bearing retroelements among the different DTUs and strains. For this purpose, genomic DNA samples from several strains of the parasite belonging to different DTUs were sequenced, and the sequences were assembled into contigs to perform comparative genomic studies. The following strains were analyzed: B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener (see Table 1 for strain details).

Genomic DNA was sequenced using the Ion Torrent technology, assembled into contigs and annotated following the workflow depicted in Figure 1 (see also Material and methods for further details). Table 2 summarizes the sequencing and assembling metrics. A total of 5.4×10^6 , 3.4×10^6 , 2.3×10^6 , 3.9×10^6 , 3.3×10^6 and 3.5×10^6 raw reads were obtained from the libraries derived from strains B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener, respectively. After filtering by sequence quality, the remaining reads (4.6×10^6 , 2.9×10^6 , 2×10^6 , 3.3×10^6 , 2.9×10^6 and 3×10^6 for B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener strains, respectively) were assembled into contigs using the CLC Genomic Workbench software. A total of 5,923 (B. M. López), 6,541 (Dm28), 6,942 (Y), 11,096 (Ikiakarora), 11,946 (SOL) and 11,101 (CL Brener) contigs were obtained, with the longest contig having a size of 45,876 (B. M. López), 33,325 (Dm28), 17,707 (Y), 33,607 (Ikiakarora), 26,881 (SOL) and 23,056 (CL Brener) base pairs (bp). The average contig sizes (in bp) were 3,124 for B. M. López, 2,633 for Dm28, 2,239 for Y, 1,666 for Ikiakarora, 1,679 for SOL and 1,759 for CL Brener. The assembled genomes were 18.5 Mb, 17.2 Mb, 15.5 Mb, 18.5 Mb, 20.1 Mb and 19.5 Mb for the B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener strains, respectively. Finally, the N50 values calculated (in bp) for each strain were 5,125 (B. M. López), 3,659 (Dm28), 2,891 (Y), 2,193 (Ikiakarora), 2,171 (SOL) and 2,298 (CL Brener).

The completeness of the six different assembled genomes was assessed through comparison with the Benchmarking Universal Single-

Copy Orthologs (BUSCO) for the euglenozoa dataset as the default, which contains 130 target BUSCO proteins that are expected to be present as single-copy genes. The gene sets were classified according to BUSCO analysis parameters (Figure 2) as complete (C), single (S), duplicated (D), fragmented (F) or missing (M). The obtained results graphically represented in Figure 2 indicate that the percentage of the identified complete and single copy BUSCOs genes were 96.9% (B. M. López), 96.9% (Dm28), 87.7% (Y), 78.5% (Ikiakarora), 76.2% (SOL), and 86.2% (CL Brener) from the 130 BUSCO searched groups (Figure 2). These percentages would reflect the completeness of the assembled genomes.

To estimate the genome size of the strains under study, the number of total bases identified in each assembled genome (Table 2) and the percentage of complete and single-copy BUSCO genes (Figure 2) were used taking into account the genome size reported of CL Brener strain as reference genome (55 Mb (El-Sayed et al., 2005b)). These calculations, described in Supplementary Table 1, led to an estimation of the genome size for each strain of 49.7 Mb for B.M. López, 46.3 Mb for Dm28, 43.5 Mb for Y, 54.2 Mb for Ikiakarora, and 59.5 Mb for SOL (Supplementary Table 1). Gene annotations for each genome were performed using the AUGUSTUS gene-prediction tool and OrthoMCL workflow to assign the set of predicted proteins to OrthoMCL groups as indicated in Figure 1. To select the reference genome to be used for gene annotation of each genome in Augustus, a Best Hit Analysis was performed against a protein database created with the proteins downloaded from *T. cruzi* genomes available at TriTrypDB. The reference genome for each of the six sequenced genomes was selected based on the highest identity score obtained in each case (Table 3). Thus, Dm28c was selected for annotation of the B. M. López and Dm28 genomes with 3,402 and 4,329 best hits, CL Brener Esmeraldo-like was used for the Y and Ikiakarora genome annotation as CL Brener Esmeraldo-like was the strain with which Y and Ikiakarora maintained the highest identity with 6,892 and 5,485 best hits, respectively. On this base, CL Brener Non-Esmeraldo-like was used for the annotation of SOL and CL Brener genomes with 6,120 and 5,717 best hits, respectively.

The G+C content of the genomes was 48.3%, 48.6%, 49.8%, 48.7%, 49.5% and 49.5% for the B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener strains, respectively (Table 4). The number of annotated genes in each genome was 7,661 (B. M. López), 7,720 (Dm28), 8,297 (Y), 11,358

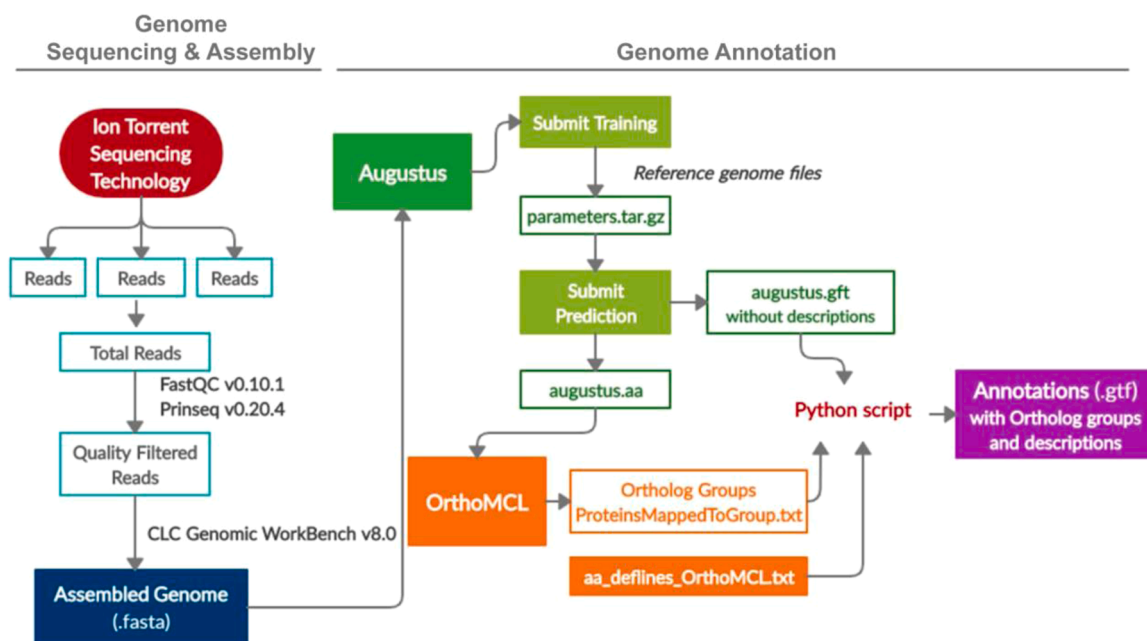


Figure 1. Schematic diagram describing the sequencing, assembly and annotation processes carried out for generating the genomic assemblies and annotations of the different *T. cruzi* strains described in this study.

Table 2Summary of sequenced and assembled genomes from the B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener *Trypanosoma cruzi* strains.

Strain	Sequencing Total reads	Quality filtered reads ^a	Assembly Total contigs	N50 (bp)	Shortest ^b contig (bp)	Mean	Longest	Total bases
B. M. López	5,415,819	4,591,877	5,923	5,125	500	3,124	45,876	18,508,455
Dm28	3,354,685	2,947,182	6,541	3,659	500	2,633	33,325	17,227,559
Y	2,345,376	1,972,462	6,942	2,891	500	2,239	17,707	15,548,466
Ikiakarora	3,928,712	3,338,764	11,096	2,193	500	1,666	33,607	18,492,845
SOL	3,332,449	2,888,011	11,946	2,171	500	1,679	26,881	20,061,745
CL Brener	3,468,705	2,955,230	11,101	2,298	500	1,759	23,056	19,533,022

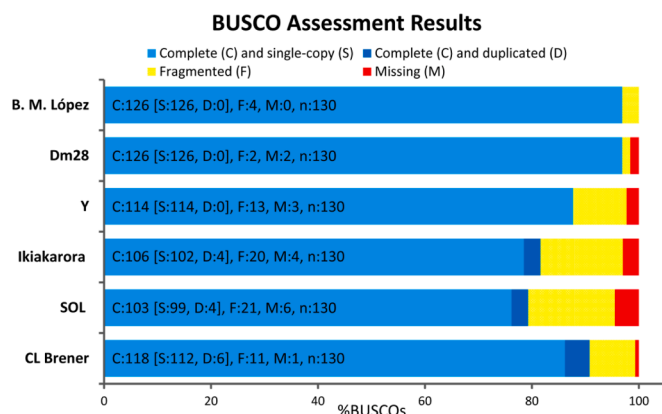
^a : Number of reads after FastQC analysis and Prinseq quality filtering.^b : The minimum contig size set in the assembly process is 500 bp.

Figure 2. BUSCO Analysis of Genome Assembly Completeness. The BUSCO (Benchmarking set of Universal Single-Copy Orthologues) dataset of the euglenozoa odb 10 including 130 BUSCOs was used for genome assembly evaluation. The bar chart shows the following percentages: complete and single copy genes (light blue), complete and duplicated genes (dark blue), fragmented genes (yellow) and missing genes (red) in the assemblies.

(Ikiakarora), 12,829 (SOL), and 11,909 (CL Brener), representing 62.9% (B. M. López), 64.1% (Dm28), 77.7% (Y), 73.2% (Ikiakarora), 72.6% (SOL), and 73.3% (CL Brener) of the total base pairs of each assembled genome. Of the total number of annotated genes, functional information was added to 5,036 (B. M. López), 5,145 (Dm28), 5,590 (Y), 7,739 (Ikiakarora), 8,600 (SOL) and 8,019 (CL Brener) genes based on their sequence identity with homologous proteins in other organisms. The remaining genes were annotated as coding for hypothetical proteins (Table 4).

3.2. Analysis of retroelement content

To identify the Pr77-hallmark bearing L1Tc and NARTc retrotransposons in the reads of the six sequenced genomes, two searches were conducted to identify the 5' and 3' ends of the L1Tc and NARTc elements. First, the 5' end probe was used, which corresponded to the Pr77 hallmark as it was always present in both elements. The 40-nt-long sequences located downstream of Pr77 downstream were employed to identify the class of element (L1Tc or NARTc), and at least 70% sequence identity was requested for the assignment. To define the 3' end, a probe corresponding to the last 55 nt of the L1Tc and NARTc elements was employed for the L1Tc and NARTc classification. Finally, the candidates were validated by BLASTN search against a database composed of 12 and 7 different sequences corresponding to L1Tc (Supplementary file 1) or NARTc (Supplementary file 2) elements, respectively. Only sequences

Table 3

Summary of Best Hit Analysis. The table shows the number of proteins of six sequenced genomes with the best hits for each reference genome. The genomes used were CL Brener, CL Brener non-Esmeraldo-like, CL Brener Esmeraldo-like, Dm28c, SylvioX10-1 and Marinkellei B7 subspecies.

Strain	Reference Strains					
	CL Brener	CL Brener Non-Esmeraldo-like	CL Brener Esmeraldo-like	Marinkellei B7	Dm28c	Sylvio X10-1
B. M. López	62	922	474	62	3,402	4,266
Dm28	52	795	391	65	4,329	3,753
Y	108	1,859	6,892	148	572	704
Ikiakarora	139	4,427	5,485	111	685	972
SOL	191	6,120	5,078	158	852	1,212
CL Brener	214	5,717	4,983	122	699	968

Table 4Characteristics of the assembled genomes and annotated genes in the *T. cruzi* strains.

Strain	GC % ^a	Total genes	Genes average length ^b (bp)	genome % identified genes ^c	Homology with previous identified proteins ^d	Hypothetical proteins ^e
B. M. López	48.26	7,661	1,518	62.86	5,036	2,625
Dm28	48.59	7,720	1,429	64.05	5,145	2,575
Y	49.79	8,297	1,456	77.74	5,590	2,707
Ikiakarora	48.72	11,358	1,191	73.20	7,739	3,619
SOL	49.48	12,829	1,135	72.63	8,600	4,229
CL Brener	49.53	11,909	1,201	73.26	8,019	3,890

^a : Guanine–cytosine content.^b : Sum of the length of all identified genes divided by the total number of identified genes.^c : Percentage of the genome occupied by the annotated genes.^d : Number of genes having functional information.^e : Number of genes of unknown function (hypothetical proteins).

maintaining an identity higher than 70% and with an aligned length of at least 70% of the sequence element for validation were retrieved for analysis. A number of Pr77-containing sequences that did not correspond either to L1Tc or NARTc were found in all the studied strains; consequently, these sequences were classified as single Pr77 hallmarks (Table 5).

To compare the frequency of the identified Pr77 hallmarks among the strains, a Pr77 index was calculated by dividing the number of identified Pr77 sequences found in each genome by the total filtered reads obtained in each sequenced genome. As shown in Table 6 (Pr77 Index), B. M. López was identified as the strain with a lower number of Pr77-hallmarks (9.5 Index) and the Ikiakarora strain as that with the largest number of Pr77 signatures (14.4 Index). To identify the relative number of Pr77 hallmarks in each strain relative to that detected in the B. M. López strain, the presence of Pr77 hallmarks relative to the abundance of Pr77 hallmarks identified in the B. M. López strain was analyzed using the following formula:

Relative abundance of Pr77 hallmarks

$$= \frac{(\text{Strain Index} - \text{B. M. López Index})}{\text{B. M. López index}} + 1$$

The results (Table 5) showed a 1.4, 1.2, 1.5, 1.3 and 1.5-fold enrichment of Pr77-hallmarks in the Dm28, Y, Ikiakarora, SOL and CL Brener genomes relative to the B. M. López genome, respectively.

Analysis of the number of full-length NARTc relative to the total reads in each genome showed an index of 0.88, 1.11, 1.18, 1.11, 1.06 and 1.21 for B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener (Table 5, Index^c). The number of Pr77 hallmarks from the total reads of each genome that were not associated with either L1Tc or NARTc were 0.33, 0.17, 0.42, 0.93, 0.87 and 1.17 in the B. M. López, Dm28, Y, Ikiakarora, SOL and CL Brener genomes, respectively (Table 5, Index^d).

3.2.1. Proportion of L1Tc and NARTc Retroelements

Differences in the number and proportion of L1Tc and NARTc retroelements were noted after comparing their abundances in the genomes of six *T. cruzi* strains. For comparison, data retrieved from the two searches conducted with the Pr77 hallmark and the L1Tc or NARTc 3' probes were considered. The results indicated that all the genomes contained a higher number of L1Tc copies than NARTc. Moreover, in all the analyzed strains, the L1Tc/NARTc ratio of identified elements using the element 3' probe was higher than that detected using the Pr77 identifier. In conclusion, all the genomes would have at least four times more L1Tc elements than NARTc copies (Table 6).

Table 5
Comparison of transposable element content.

	B. M. López	Dm28	Y	Ikiakarora	SOL	CL Brener
Pr77 Index ^a	9.5	13.3	11.7	14.4	12.5	13.8
Pr77 presence respect strain with lower Index ^b	1	1.4	1.2	1.5	1.3	1.5
Full-length NARTc (present Pr77 and 3' end) Index ^c	0.88	1.11	1.18	1.11	1.06	1.21
Pr77 not associated with L1Tc or NARTc Index ^d	0.33	0.17	0.42	0.93	0.87	1.17

^a : Index obtained by dividing the number of identified Pr77 sequences from the total filtered reads of each genome and multiplying by 10⁴.

^b : Abundance of Pr77 hallmarks in the different strains relative to the Pr77 abundance in the B. M. López genome (strain with the lowest Pr77Index^a).

^c : Index calculated by dividing the number of full-length NARTc by the total filtered reads of each genome and multiplying by 10⁴.

^d : Index calculated by dividing the number of identified Pr77 not associated with L1Tc or NARTc elements by the total filtered reads of each genome and multiplying by 10⁴.

Table 6

Relative abundance of retroelements identified by (a) Pr77 hallmark plus 40 nt located downstream of Pr77, (b) 3' ends (55 last nt of L1Tc or NARTc) or (c) the average of elements retrieved from both 5' and 3' ends.

Proportion	B. M. López	Dm28	Y	Ikiakarora	SOL	CL Brener
L1Tc / NARTc (Identified from Pr77) ^a	4	5.02	3.39	4	4.37	4.19
L1Tc / NARTc (Identified from 3' end) ^b	5.64	7.08	5.59	6.73	4.80	6.38
L1Tc / NARTc (Total) ^c	4.88	6.09	4.44	5.27	4.62	5.26

^a Pr77 hallmark plus 40 nt located downstream of Pr77

^b 3' ends (55 last nt of L1Tc or NARTc)

^c the average of elements retrieved from both 5' and 3' ends.

3.3. Genomic context of retroelements

To determine the location and genomic context in which the identified Pr77 hallmark bearing retroelements were inserted, the sequences located upstream of the Pr77-hallmarks and of the 3' retrotransposon ends were retrieved from the reads and aligned to the contigs of the assembled genomes and to the annotated genomes for sequence identification.

The results, summarized in Table 7, indicated that L1Tc and NARTc were in all the strains, frequently positioned close to trans-sialidase (TS), mucin-associated protein (MASP), retrotransposon hot spots (RHS) genes and to genes encoding hypothetical proteins (HPs). Moreover, the results indicated that these genes could be found both upstream and downstream of L1Tc, NARTc and single Pr77 hallmarks (Sh), although at different proportions depending on the type of retroelement, gene class and parasite strain. Analysis of the percentage represented by each associated sequence (retroelement and particular sequence) regarding the total number of identified Pr77-bearing retroelements into each genome (Table 7) showed that these were not random associations because they were frequently found in all cases.

The data showed that the percentage of elements associated with TS-coding genes varied between strains, demonstrating a greater frequency in the SOL strain (40.28%) and mainly located upstream of the Pr77 hallmarks. However, the Pr77 hallmarks were most frequently associated with genes encoding hypothetical proteins that were commonly located both upstream and downstream of Pr77 signatures in all strains, ranging from 407 and 368 in B. M. López and from 76 and 113 in Y (see Table 7 for details).

The number of retroelements flanked by MASP was lower than those flanked by TS. Differences were also observed between strains, with CL Brener presenting an elevated percentage of elements flanked by these proteins (7.58% DW and 0.25% UP) and the Y strain the lowest percentage (0.76% UP), in which no MASP was located downstream of the identified retroelements. The number of retroelements which were upstream and downstream flanked by RHS-coding genes was higher in the SOL and Ikiakarora strains, represented by a total of 22.34% and 16.53%, respectively, as this percentage was 5.7%, 2.5% and 0.24% in CL Brener, Y and B. M. López. No retroelements flanked by RHS were identified in Dm28. In the B. M. López strain, RHS was only identified upstream of the retroelements (Table 7). The percentages corresponding to hypothetical proteins ranged from 28.9% to 60.26% UP and from 34.59% to 67.66% DW, depending on the strain. This percentage was higher in Dm28 (60.26% UP and 53.06% DW), and the lowest values were detected in the SOL strain (39.23% UP and 34.59% DW) (Table 7).

Table 7

Genomic context where retroelements are inserted in the different genomes.

		B. M. López	Dm28	Y	Ikiakarora	SOL	CL Brener
(b)		L1Tc/NARTc/Sh/NC	L1Tc/NARTc/Sh/NC	L1Tc/NARTc/Sh/NC	L1Tc/NARTc/Sh/NC	L1Tc/NARTc/Sh/NC	L1Tc/NARTc/Sh/NC
TS	UP	2 / 1 / 22 / 7 ^(a) 32 (3.87%)	11 / 0 / 13 / 4 28 (4.64%)	13 / 3 / 28 / 16 60 (22.81%)	2 / 36 / 0 / 3 41 (6.43%)	13 / 41 / 95 / 39 188 (34.31%)	31 / 13 / 0 / 10 54 (13.33%)
	DW	20 / 68 / nd / nd 88 (12.79%)	14 / 9 / nd / nd 23 (6.71%)	-	15 / 6 / nd / nd 21 (7.37%)	17 / 2 / nd / nd 19 (5.97%)	10 / 1 / nd / nd 11 (5.56%)
MASP	UP	13 / 0 / 0 / 1 14 (1.69%)	5 / 0 / 0 / 6 11 (1.82%)	1 / 0 / 0 / 1 2 (0.76%)	6 / 3 / 7 / 0 16 (2.51%)	1 / 0 / 0 / 0 1 (0.18%)	1 / 0 / 0 / 0 1 (0.25%)
	DW	24 / 3 / nd / nd 27 (3.92%)	11 / 0 / nd / nd 11 (3.21%)	-	1 / 1 / nd / nd 2 (0.70%)	14 / 1 / nd / nd 15 (4.72%)	8 / 7 / nd / nd 15 (7.58%)
RHS	UP	0 / 1 / 0 / 1 2 (0.24%)	-	4 / 1 / 0 / 0 5 (1.90%)	52 / 30 / 0 / 10 92 (14.42%)	13 / 4 / 0 / 2 19 (3.47%)	7 / 8 / 0 / 4 19 (4.69%)
	DW	-	-	0 / 1 / nd / nd 1 (0.60%)	3 / 3 / nd / nd 6 (2.11%)	60 / 0 / nd / nd 60 (18.87%)	2 / 0 / nd / nd 2 (1.01%)
HP	UP	101 / 157 / 42 / 107 407 (49.27%)	210 / 68 / 20 / 66 364 (60.26%)	8 / 33 / 25 / 10 76 (28.9%)	98 / 80 / 24 / 34 236 (36.99%)	99 / 40 / 28 / 48 215 (39.23%)	49 / 72 / 29 / 33 183 (45.19%)
	DW	202 / 166 / nd / nd 368 (53.49%)	104 / 78 / nd / nd 182 (53.06%)	97 / 16 / nd / nd 113 (67.66%)	107 / 70 / nd / nd 177 (62.11%)	51 / 59 / nd / nd 110 (34.59%)	69 / 59 / nd / nd 128 (64.65%)

(a) Number of L1Tc and NARTc retrotransposons bearing Pr77 hallmark (L1Tc and NARTc), single Pr77 hallmarks not associated with retrotransposons (Sh) and non-classified Pr77 hallmarks due to read length (NC).

(b) Trans-sialidases (TS), Mucin- associated surface proteins (MASP), Retrotransposon hot spot protein (RHS), and hypothetical protein (HP) identified upstream (UP) or downstream (DW) of Pr77-bearing retrotransposons or single Pr77 hallmarks. The number of times that each association was found in each genome is represented in bold face, and the percentage representing the total identified associations is shown as %. nd: Not determined due to the read length.

3.4. Gene number variations of representative gene families among the different *T. cruzi* strains

To estimate the gene composition content and the similarity degree and divergences among the genomes under study, the copy number of selected genes coding for 13 proteins families of interest and the percentage that they occupied among the total identified genes were determined. The amount of hypothetical proteins was also determined as an additional group in all of them. The gene families analyzed were those coding for trans-sialidases (TS), mucins, mucin-associated surface proteins (MASPs), retrotransposon hot spot proteins (RHSs), and TASV (Trypomastigote, Alanine, Serine, Valine proteins). These proteins are associated with parasite virulence processes and retroelements. In addition constitutive genes such as kinesins, RNA helicases, protein kinases, glycosyltransferases, cysteine peptidases, heat shock protein 70 (HSP70), Paraflagellar Rod (PFR) proteins and RNA-binding proteins were also analyzed.

The results, summarized in Table 8, show a similar gene content of the analyzed constitutive genes as well as of the identified hypothetical proteins (ranging from 31.86% to 34.26% of the total amount of identified genes) among the six *T. cruzi* genomes under study. However, differences in the copy number of particular genes and, therefore, the percentage that they represented in the genomes were identified among the six analyzed *T. cruzi* strains. Thus, the trans-sialidase gene family had

a greater representation in the genomes of strains Y (0.87%), CL Brener (0.85%) and SOL (0.83%) containing more than double the number of TS genes than the other strains analyzed, corresponding to 0.48%, 0.34% and 0.19% of the Ikiakarora, B. M. López and Dm28 genomes, respectively. The copy number of genes coding for MASP and TASV proteins was also superior reaching one order of magnitude higher in the Y, CL Brener and SOL genomes. Mucin genes were found in the largest proportion in the CL Brener genome (0.34%), followed by the Y (0.23%), Ikiakarora (0.20%) and SOL (0.19%) genomes, with percentages that were much higher than those found in the genomes of Dm28 (0.04%) and B. M. López (0.02%) strains. The percentage of identified RHS genes was higher in four strains (0.29% in Y, 0.26% in CL Brener, and 0.17% in Ikiakarora and SOL) than in the genomes of Dm28 (0.08%) and B. M. López (0.13%) strains.

4. Discussion

Numerous studies have reported that the different *T. cruzi* strains causing Chagas disease present phenotypic variations as well as different behaviors in terms of tissue tropism, virulence, and physiopathology, consequently inducing heterogeneous immunological responses in the hosts (Franzen et al., 2011; Rassi Jr. et al., 2012; Revollo et al., 1998; Rodriguez et al., 2014). It has been reported that *T. cruzi* strains from different DTU, as well as strains from the same DTU, show differences in

Table 8

Comparison of gene family content. Copy number of each gene in the assembled genome and representation of the total identified genes (in percentage) in each case.

	B.M. López	Dm28	Y	Ikiakarora	SOL	CL Brener
Trans-sialidase	26 (0.34)	15 (0.19)	72 (0.87)	54 (0.48)	107 (0.83)	101 (0.85)
Kinesin	47 (0.61)	48 (0.62)	61 (0.74)	81 (0.71)	99 (0.77)	99 (0.83)
Mucin	2 (0.02)	3 (0.04)	19 (0.23)	23 (0.20)	24 (0.19)	40 (0.34)
RNA helicase	41 (0.54)	43 (0.56)	49 (0.59)	69 (0.61)	71 (0.55)	72 (0.60)
MASP	14 (0.18)	12 (0.16)	108 (1.30)	99 (0.87)	201 (1.57)	195 (1.64)
RHS	10 (0.13)	6 (0.08)	24 (0.29)	19 (0.17)	22 (0.17)	31 (0.26)
Kinase protein	276 (3.60)	296 (3.83)	317 (3.82)	412 (3.63)	456 (3.55)	423 (3.55)
Glycosyltransferase	15 (0.2)	14 (0.18)	15 (0.18)	23 (0.28)	27 (0.21)	26 (0.22)
Cysteine peptidase	21 (0.27)	21 (0.27)	32 (0.39)	47 (0.41)	52 (0.41)	43 (0.36)
TASV	3 (0.04)	1 (0.01)	7 (0.08)	3 (0.03)	7 (0.05)	7 (0.06)
Heat Shock - 70	7 (0.09)	6 (0.08)	8 (0.10)	10 (0.09)	15 (0.12)	9 (0.07)
PFR-proteins	21 (0.27)	20 (0.26)	23 (0.28)	29 (0.26)	41 (0.32)	35 (0.29)
RNA-binding protein	79 (1.03)	69 (0.89)	73 (0.88)	96 (0.85)	115 (0.90)	104 (0.87)
Hypothetical protein	2,625 (34.3)	2,575 (33.3)	2,707 (32.6)	3,619 (31.9)	4,229 (32.9)	3,890 (32.7)

Abbreviations used: **MASP**: Mucin- associated surface proteins; **RHS**: Retrotransposon hot spot protein; **PFR-proteins**: Paraflagellar Rod Proteins.

chromosome number, genome size and DNA content (Lewis et al., 2009; Vargas et al., 2004). Differences in the gene content and copy number of target genes employed for parasite detection and diagnosis (Duffy et al., 2009; Ramírez et al., 2015), together with differences in susceptibility to drugs (Teston et al., 2013), greatly hinders the follow up of patients and development of new drugs and vaccines. Thus, genetic characterization of the different strains will facilitate the management of patients and disease control.

To improve our knowledge of the *T. cruzi* genetic variability, strains of different DTUs were sequenced and analyzed at the genomic level. Although retrotransposons constitute a huge amount of eukaryote genomes, they are the bottleneck for genome assembly due to their long length and high copy number. Thus, transposable elements are mis-annotated sequences and parts of the genomes that are difficult to study. With the aim of improving our knowledge of these mobile elements, the content of the L1Tc and NARTc retroelements which bear the Pr77 hallmark at their 5' ends was analyzed in six *T. cruzi* sequenced genomes. The strains selected were B. M. López (TcIa), Dm28 (TcId), Y (TcII), Ikiakarora (TcIII), SOL (TcV) and CL Brener (TcVI) as they were isolated in different endemic and nonendemic geographical areas from Venezuela, Spain, and different locations from Colombia and Brazil, from different mammal hosts (*Homo sapiens*, *Didelphis marsupialis*) and insect vectors (*Triatoma infestans* or *Rhodnius prolixus*) (Hamuy et al., 2013; Higuera et al., 2013; Rodríguez et al., 1998). The sequencing and assembly of the B. M. López and Ikiakarora strains have recently been reported (Gomez et al., 2020; Gómez et al., 2020). The SOL strain was isolated from a baby who had been vertically infected and born in Spain, a nonendemic region (Murcia et al., 2013).

Following sequencing and assembly, analysis of the sequencing data revealed similar characteristics regarding raw reads which enabled comparative studies among the six strains. Thus, the quality-filtered reads that were assembled into a contig provided from 5,923 (B. M. López strain) to 11,946 (SOL strain) contigs, reaching a total size ranging from 15,548,466 (Y strain) to 20,061,745 (SOL strain) bp. The contigs ranged from a minimum size of 500 bp for all strains, which was the threshold established in the assembly process, to the largest of 45,876 bp obtained for the B. M. López strain. In spite of the limitation of the high number of the generated contigs derived from sequencing using the Ion Torrent PGM™ system, analysis of the completeness of the genome assemblies carried out by BUSCO revealed a higher percentage of identified complete and single copy BUSCO genes in all strain, ranging from 76.2% (SOL) to 96.9% (B. M. López and Dm28 strains). This fact, together with the small number of duplicated BUSCOs, suggested that haplotypes were correctly collapsed and indicated a good quality of the assembled genomes. For genome annotation, Submit Training was performed prior to gene prediction to generate a specific parameter file for each genome annotation. The reference strains selected were in all cases those with the largest number of best hits, except for B. M. López, in which Dm28c with 3,402 was chosen as the reference genome instead of Sylvio X10 (4,266) because the latter had a poor annotation with a large number of proteins with unknown function.

The G+C content of each genome was close to 50% in all cases, ranging from 48.26% (B. M. López strain) to 49.79% (Y strain). The number of predicted genes in each genome ranged from 7,661 (B. M. López strain) to 12,829 (SOL strain), representing from 62.86% to 77.74% of the total base pairs of each assembled genome. It should be noted that the highest number of predicted genes were identified in the CL Brener and SOL strains (11,909 and 12,829, respectively), the two hybrid strains in which the highest number of total bases were also identified (19.5×10^6 and 20.1×10^6) and also those with the largest estimated genome sizes. These differences could be related to the genome size of TcI strains (B.M. López and Dm28) which are smaller than the hybrid CL Brener and SOL strains (TcVI and TcV). The average length of the annotated genes ranges from 1,191 to 1,518 bp, with the latter identified in the B. M. López strain. According to the annotation data assigned to the analyzed genomes, a large proportion of identified

genes in the present six genomes had an identity with previously identified proteins in other genomes. Thus, from 5,036 protein coding genes from B. M. López strain to 8,600 proteins identified in the SOL strain were found to have counterparts in other organisms. In addition, approximately one-third of the identified predicted genes did not match with genes of known function, ranging from 2,575 (Dm28 strain) to 4,229 hypothetical proteins found in the SOL strain.

The genome of *Trypanosoma cruzi* consists of more than 50% tandemly repeated sequences, which is much higher than those represented in other trypanosomatid genomes such as *Leishmania* (20%) and *T. brucei* (25%) (Berriman et al., 2005; El-Sayed et al., 2005a, 2005b; Ivens et al., 2005). This repetitive multigene content, together with the existence of hybrid strains and aneuploidies, makes *T. cruzi* an organism with a complex genomic architecture that hinders the completeness of genome assemblies (Reis-Cunha et al., 2015; Reis-Cunha and Bartholomeu, 2019; Vargas et al., 2004). These repeats are mainly due to large multigene families of surface proteins, retrotransposons, and subtelomeric repeats.

Analysis of the Pr77-bearing retrotransposon content in the sequenced genomes showed that these retrotransposons were highly represented in all the strains and particularly abundant in the Ikiakarora and CL Brener strains. In addition, the proportion of L1Tc was at least 4-fold higher than the NARTc elements in all the strains. This proportion could be related to the potentially autonomous character of the L1Tc retrotransposon which could be used by non-autonomous NARTc for *trans* mobilization. Moreover, the high proportion of L1Tc elements may well be related to the need for the repair properties associated with NL1Tc, a protein encoded by ORF1 of L1Tc, which has been shown to be endowed with AP endonuclease as well as 3'-phosphatase and 3'-phosphodiesterase activities (Olivares et al., 1999). In addition to the implications that these activities have in the first step of the integration process of these retrotransposons, these enzymatic functions *in vivo* have been shown to lead to a reduction of approximately 60% of DNA damage caused by daunorubicin treatment of parasites, providing protection from the negative effects of daunorubicin and gamma-radiation (6 or 9 Gy) on parasite survival and growth rate (Olivares et al., 2003).

Although the Pr77-hallmark was one of the two probes used to search the L1Tc and NARTc elements in the six genomes, the subsequent analyses carried out by BLAST searches indicated that retrotransposons lacking the Pr77 hallmark were not found. By contrast, Pr77 hallmarks not associated with retrotransposons were identified in the six *T. cruzi* genomes; these new elements were called single Pr77 hallmarks (Sh). The Pr77 sequence was originally described in the L1Tc 5' end and was shown to be able to activate the transcription of downstream genes and generation of abundant transcripts via RNA polymerase II (Heras et al., 2007). The Pr77 sequence was shown to contain a downstream core promoter element (DPE) that is conserved in terms of nucleotide composition and location to the transcription start site (TSS) in the consensus sequence of L1Tc from *T. congolense*, the *ingi* elements of *T. brucei*, *T. congolense* and *T. vivax*, the elements corresponding to the truncated versions of L1Tc and the *ingi*-retrotransposons (NARTc and RIME) of *T. cruzi*, *T. brucei* and *T. vivax*, in SIDER 1 of *T. congolense* and SIDER1a of *T. vivax*, SIDER 2 from *T. brucei*, and SIDER 2A of *L. infantum*, *L. mexicana*, *L. braziliensis*, *L. panamensis* and *L. major* (Macías et al., 2016). In addition, Pr77 was the first dual promoter/ribozyme system to be discovered that works at the DNA/RNA levels as a promoter and HDV-like ribozyme, respectively (Sanchez-Luque et al., 2012). Thus, it was suggested that Pr77 signature-bearing retrotransposons may be responsible for the expansion of these functions across trypanosomatid genomes (Sanchez-Luque et al., 2014). Recent studies carried out in our laboratory have shown that some of the above mentioned single Pr77 hallmarks (Sh) have HDV-like ribozyme activity *in vitro* (Gomez I, Afonso-Lehmann R, et al, manuscript in preparation).

In the present paper, we also described that an important proportion of L1Tc and NARTc are frequently associated with RHSS, MASPs, and trans-sialidases. Furthermore, this study showed differences between

the analyzed strains; thus, the SOL strain was the one in which most of the TS genes were associated with retroelements. Strikingly, most of them were associated with the single Pr77-hallmarks not associated with L1Tc and NARTc elements (Sh). The presence of active sequences suggested that the association of Pr77-hallmarks (as part of retrotransposons or as single Pr77 hallmarks) and these particular sequences may have a functional sense probably related to a regulatory role in the gene expression patterns that has allowed to these associations to coevolve together and be maintained during evolution. The relatively high degree of dispersion exhibited by L1Tc and noncoding but transcriptionally active NARTc elements reinforces the idea that these non-LTR retrotransposons play a regulatory role in gene expression and are responsible for the transcription of adjacent genes and polycistrons as it was previously suggested (Heras et al., 2007). Previous studies based on genomic distribution of L1Tc and NARTc retrotransposons in the *T. cruzi* genome revealed the existence of concatemers of these elements inserted into noncoding sequences as well as associated with gene coding sequences, particularly the one encoding trans-sialidase in the CL strain (Olivares et al., 2000). This association was taken as an additional indication of the domestication of retrotransposons by the host genomes, which due to their high content of mobile genetic elements are far from being static information systems. Once acquired, these retrotransposons became sufficiently necessary to persist and become a part of the genomes.

The most abundant multigene families encode surface proteins, such as mucins, members of the trans-sialidase (TS) superfamily, mucin-associated surface proteins (MASPs) and retrotransposon hot spot proteins (RHSS) (El-Sayed et al., 2005a). The copy number of genes encoding these proteins together with others of interest in trypanosomatids with a relevant biological role and genes with constitutive expression was analyzed in the six genomes under study. The results showed differences in the copy number of genes associated with virulence processes, but not in constitutive gene content, among the six *T. cruzi* genomes. The copy numbers of MASP, TS and TASV-coding genes varied greatly among the strains being the abundance of these genes higher in the Y (TcII), SOL (TcV), and CL Brener (TcVI) strains with respect to the other analyzed strains. Studies of these three groups of proteins have suggested that they are related to host-parasite interactions, as all of them have been identified in the trypomastigote stage, in which the parasite interacts with the mammalian host (Bartholomeu et al., 2009; Bernabó et al., 2013; Caeiro et al., 2018; De Pablos et al., 2011; El-Sayed et al., 2005a; Frasc, 2000; Kulkarni et al., 2009). Interestingly, TS constitutes one of the main antigens present in the parasite (Frasc, 2000), and the MASP multigene family was first described in the *T. cruzi* CL Brener genome (El-Sayed et al., 2005a). Several studies have suggested that MASP may play an important role in immune evasion and host-parasite interaction processes, presenting a high variability among the *T. cruzi* strains (Bartholomeu et al., 2009; De Pablos et al., 2011; dos Santos et al., 2012; Seco-Hidalgo et al., 2015). CL Brener and Y were also the strains with the highest content of RHS, a multigene family associated with mobile element insertions (Bringaud et al., 2002). Remarkably, the Dm28 and B. M. López genomes, both of which belong to DTU I, contained the lowest frequency of mucins, with large differences compared with the other analyzed strains. These strains were also the ones with the lowest contents of the TS, MASP, TASV, RHS, cysteine peptidases and kinesin proteins coding genes. Strikingly, a study of *T. rangeli*, which is closely related to *T. cruzi*, revealed that its genome has fewer copies of the multigene family MASP, TS and mucins compared with *T. cruzi*. *T. rangeli* is nonpathogenic to the mammalian host, which may be the result of the reduction of multigene families, highlighting the important role of these proteins in cellular invasion processes (Reis-Cunha and Bartholomeu, 2019; Stoco et al., 2014).

Since the sequencing of these genomes contributes to the current *T. cruzi* repository, the finding reported herein highlight the variability among the different strains of this complex parasite in which the Pr77

hallmark seems to have a relevant functional role.

Author Statement

Conceptualization: MCL and MCT; Formal analysis: IG, AR, BA, FLD, MCL and MCT; Funding acquisition: BV, JMR, BA, MCL and MCT; Investigation: IG, AR, JMR, FLD and MCT; Methodology: IG, AR, MCL and MCT; Discussion of Results: IG, JMR, MCL and MCT; Writing first draft: IG and MCT; Writing – review & editing final manuscript: IG, MCL and MCT

Supporting information captions

Supplementary file 1. Database composed of the different sequences corresponding to L1Tc.

Supplementary file 2. Database composed of the different sequences corresponding to NARTc.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dr. Eduardo A. León (IPBLN-CSIC) for his help with the estimation of gene family content and BUSCO bioinformatics analyses and to Almudena López-Barajas (IPBLN) for her technical collaboration in the identification of the lineage of *T. cruzi* strains. This work is part of the Ph.D. thesis of the student Inmaculada Gómez at the University of Granada in the Biochemistry and Molecular Biology Program.

This research was funded by grants PID2019-109090RB-I00 from the Programa Estatal I+D+i, Ministry of Science and Innovation of Spain (MICINN) and the Network of Tropical Diseases Research RICET, Instituto de Salud Carlos III (RD16/0027/0001, RD16/0027/0005 and RD16/0027/0008) and FEDER. The CBMSO receives institutional grants from the Fundación Ramón Areces and from the Fundación Banco de Santander.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.actatropica.2021.106053.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Andrews, S., 2010 FastQC: a quality control tool for high throughput sequence data. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Bartholomeu, D.C., Cerqueira, G.C., Leão, A.C.A., daRocha, W.D., Pais, F.S., Macedo, C., Djikeng, A., Teixeira, S.M.R., El-Sayed, N.M., 2009. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen *Trypanosoma cruzi*. *Nucleic Acids Res.* 37, 3407–3417. <https://doi.org/10.1093/nar/gkp172>.
- Bernabó, G., Levy, G., Ziliani, M., Caeiro, L.D., Sánchez, D.O., Tekiel, V., 2013. TcTASV-C, a protein family in *trypanosoma cruzi* that is predominantly trypomastigote-stage specific and secreted to the medium. *PLoS One* 8. <https://doi.org/10.1371/journal.pone.0071192>.
- Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., Bohme, U., Hannick, L., Aslett, M.A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U.C., Arrowsmith, C., Atkin, R.J., Barron, A.J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.J., Churcher, C., Clark, L.N., Corton, C.H., Cronin, A., Davies, R.M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M.C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B.R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A.X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P.J., Moule, S., Martin, D.M., Morgan, G.W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C.S., Peterson, J., Quail, M.A., Rabinowitz, E., Rajandream, M.A.,

- Reitter, C., Salzberg, S.L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A.J., Tallon, L., Turner, C.M., Tait, A., Tivey, A.R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M.D., Embley, T.M., Gull, K., Ullu, E., Barry, J. D., Fairlamb, A.H., Opperdoes, F., Barrell, B.G., Donelson, J.E., Hall, N., Fraser, C.M., Melville, S.E., El-Sayed, N.M., 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309 (80-), 416–422. <https://doi.org/10.1126/science.1112642>.
- Brenière, S.F., Waleckx, E., Barnabé, C., 2016. Over Six Thousand *Trypanosoma cruzi* Strains Classified into Discrete Typing Units (DTUs): attempt at an inventory. *PLoS Negl. Trop. Dis.* 10, 1–19. <https://doi.org/10.1371/journal.pntd.0004792>.
- Bringaud, F., Berriman, M., Hertz-Fowler, C., 2009. Trypanosomatid genomes contain several subfamilies of ingi-related retrotransposons. *Eukaryot. Cell* 8, 1532–1542. <https://doi.org/10.1128/EC.00183-09>.
- Bringaud, Frédéric, Biteau, N., Melville, S.E., Hez, S., El-Sayed, N.M., Leech, V., Berriman, M., Hall, N., Donelson, J.E., Baltz, T., 2002. A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*. *Eukaryot. Cell* 1, 137–151. <https://doi.org/10.1128/EC.1.1.137-151.2002>.
- Bringaud, F., Garcia-Perez, J.L., Heras, S.R., Ghedin, E., El-Sayed, N.M., Andersson, B., Baltz, T., Lopez, M.C., 2002. Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. *Mol. Biochem. Parasitol.* 124, 73–78.
- Bringaud, F., Ghedin, E., Blandin, G., Bartholomeu, D.C., Caler, E., Levin, M.J., Baltz, T., El-Sayed, N.M., 2006. Evolution of non-LTR retrotransposons in the trypanosomatid genomes: *Leishmania major* has lost the active elements. *Mol. Biochem. Parasitol.* 145, 158–170. <https://doi.org/10.1016/j.molbiopara.2005.09.017>.
- Brise, S., Dujardin, J.C., Tibayrenc, M., 2000. Identification of six *Trypanosoma cruzi* lineages by sequence-characterised amplified region markers. *Mol. Biochem. Parasitol.* 111, 95–105. [https://doi.org/10.1016/S0166-6851\(00\)00302-9](https://doi.org/10.1016/S0166-6851(00)00302-9).
- Brise, S., Verhoef, J., Tibayrenc, M., 2001. Characterisation of large and small subunit rRNA and mini-exon genes further supports the distinction of six *Trypanosoma cruzi* lineages. *Int. J. Parasitol.* 31, 1218–1226. [https://doi.org/10.1016/S0020-7519\(01\)00238-7](https://doi.org/10.1016/S0020-7519(01)00238-7).
- Caeiro, L.D., Alba-Soto, C.D., Rizzi, M., Solana, M.E., Rodriguez, G., Chidichimo, A.M., Rodriguez, M.E., Sánchez, D.O., Levy, G.V., Tekiel, V., 2018. The protein family TcTASV-C is a novel *Trypanosoma cruzi* virulence factor secreted in extracellular vesicles by trypomastigotes and highly expressed in bloodstream forms. *PLoS Negl. Trop. Dis.* 12. <https://doi.org/10.1371/journal.pntd.0006475>.
- Cura, C.I., Mejía-Jaramillo, A.M., Duffy, T., Burgos, J.M., Rodríguez, M., Cardinal, M.V., Kjos, S., Gurgel-Gonçalves, R., Blanchet, D., De Pablos, L.M., Tomasini, N., da Silva, A., Russomando, G., Cuba, C.A.C., Aznar, C., Abate, T., Levin, M.J., Osuna, A., Gürtler, R.E., Diosque, P., Solari, A., Triana-Chávez, O., Schijman, A.G., 2010. *Trypanosoma cruzi* I genotypes in different geographical regions and transmission cycles based on a microsatellite motif of the intergenic spacer of spliced-leader genes. *Int. J. Parasitol.* 40, 1599–1607. <https://doi.org/10.1016/j.ijpara.2010.06.006>.
- De Pablos, L.M., González, G.G., Parada, J.S., Hidalgo, V.S., Lozano, I.M.D., Samblás, M. M.G., Bustos, T.C., Osuna, A., 2011. Differential expression and characterization of a member of the mucin-associated surface protein family secreted by *Trypanosoma cruzi*. *Infect. Immun.* 79, 3993–4001. <https://doi.org/10.1128/IAI.05329-11>.
- dos Santos, S.L., Freitas, L.M., Lobo, F.P., Rodrigues-Luiz, G.F., Mendes, T.A., de, O., Oliveira, A.C.S., Andrade, L.O., Chiari, E., Gazzinelli, R.T., Teixeira, S.M.R., Fujiwara, R.T., Bartholomeu, D.C., 2012. The MASP family of trypanosoma *cruzi*: changes in gene expression and antigenic profile during the acute phase of experimental infection. *PLoS Negl. Trop. Dis.* 6, e1779. <https://doi.org/10.1371/journal.pntd.0001779>.
- Duffy, T., Bisio, M., Altcheh, J., Burgos, J.M., Diez, M., Levin, M.J., Favaloro, R.R., Freilij, H., Schijman, A.G., 2009. Accurate real-time PCR strategy for monitoring bloodstream parasitic loads in chagas disease patients. *PLoS Negl. Trop. Dis.* 3, e419. <https://doi.org/10.1371/journal.pntd.0000419>.
- El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G., Westenberger, S.J., Caler, E., Cerqueira, G.C., Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D.A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J.F., de Jong, P., Edwards, K., Englund, P.T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A.C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M.J., Lorenzi, H., Louie, T., Machado, C.R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J.C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J.L., Rinta, J., Robertson, L., Salzberg, S.L., Sanchez, D.O., Seyler, A., Sharma, R., Shetty, J., Simpson, A.J., Sisk, E., Tammi, M.T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P.N., Wickstead, B., Wortman, J., White, O., Fraser, C.M., Stuart, K. D., Andersson, B., 2005a. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309 (80-), 409–415. <https://doi.org/10.1126/science.1112631>.
- El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renaud, H., Worthey, E.A., Hertz-Fowler, C., Ghedin, E., Peacock, C., Bartholomeu, D.C., Haas, B.J., Tran, A.N., Wortman, J.R., Alsmark, U.C., Angiuoli, S., Anupama, A., Badger, J., Bringaud, F., Cadag, E., Carlton, J.M., Cerqueira, G.C., Creasy, T., Delcher, A.L., Djikeng, A., Embley, T.M., Hauser, C., Ivens, A.C., Kummerfeld, S.K., Pereira-Leal, J.B., Nilsson, D., Peterson, J., Salzberg, S.L., Shallom, J., Silva, J.C., Sundaram, J., Westenberger, S., White, O., Melville, S.E., Donelson, J.E., Andersson, B., Stuart, K.D., Hall, N., 2005b. Comparative genomics of trypanosomatid parasitic protozoa. *Science* (80-). 309, 404–409. <https://doi.org/10.1126/science.1112181>.
- Falla, A., Herrera, C., Fajardo, A., Montilla, M., Vallejo, G.A., Guhl, F., 2009. Haplotype identification within *Trypanosoma cruzi* I in Colombian isolates from several reservoirs, vectors and humans. *Acta Trop.* 110, 15–21. <https://doi.org/10.1016/j.actatropica.2008.12.003>.
- Franzen, O., Ochaya, S., Sherwood, E., Lewis, M.D., Llewellyn, M.S., Miles, M.A., Andersson, B., 2011. Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio X10/1 and comparison with T. *cruzi* VI CL Brenner. *PLoS Negl. Trop. Dis.* 5, e984. <https://doi.org/10.1371/journal.pntd.0000984>.
- Frasch, A.C.C., 2000. Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*. *Parasitol. Today.* [https://doi.org/10.1016/S0169-4758\(00\)01698-7](https://doi.org/10.1016/S0169-4758(00)01698-7).
- García-Pérez, J.L., González, C.I., Thomas, M.C., Olivares, M., López, M.C., 2003. Characterization of reverse transcriptase activity of the LITc retroelement from *Trypanosoma cruzi*. *Cell. Mol. Life Sci.* 60, 2692–2701. <https://doi.org/10.1007/s00018-003-3342-y>.
- Gascon, J., Bern, C., Pinazo, M.J., 2010. Chagas disease in Spain, the United States and other non-endemic countries. *Acta Trop.* 115, 22–27. <https://doi.org/10.1016/j.actatropica.2009.07.019>.
- Gómez, I., Rastrojo, A., Lorenzo-Díaz, F., Sánchez-Luque, F.J., Macías, F., Aguado, B., Valladares, B., Requena, J.M., López, M.C., Thomas, M.C., 2020. *Trypanosoma cruzi* Ikiakara (TcII) Draft Genome Sequence. *Microbiol. Resour. Announc.* 9. <https://doi.org/10.1128/MRA.00453-20>.
- Gomez, I., Rastrojo, A., Sanchez-Luque, F.J., Lorenzo-Diaz, F., Macias, F., Valladares, B., Aguado, B., Requena, J.M., Lopez, M.C., Thomas, M.C., 2020. Draft genome sequence of the trypanosoma *cruzi* B. M. Lopez Strain (TcIa), isolated from a colombian patient. *Microbiol. Resour. Announc.* 9. <https://doi.org/10.1128/MRA.00031-20>.
- Graham Clark, C., Pung, O.J., 1994. Host specificity of ribosomal DNA variation in sylvatic *Trypanosoma cruzi* from North America. *Mol. Biochem. Parasitol.* 66, 175–179. [https://doi.org/10.1016/0166-6851\(94\)90052-3](https://doi.org/10.1016/0166-6851(94)90052-3).
- Hamuy, R., Vera, B.N., Ferreira, M.E., Acosta, N., Lopez, E., 2013. Determination of the in vitro sensitivity of different *Trypanosoma cruzi* strains to benzimidazole and the leaf extract of the plant *Zanthoxylum chiloperone*. *Mem. Inst. Invest. Cienc. Salud* 9, 16–25.
- Heras, S.R., López, M.C., García-Pérez, J.L., Martin, S.L., Thomas, M.C., 2005. The LITc C-Terminal domain from trypanosoma *cruzi* Non-Long terminal repeat retrotransposon codes for a protein that bears two C2H2 Zinc finger motifs and is endowed with nucleic acid chaperone activity. *Mol. Cell. Biol.* 25, 9209–9220. <https://doi.org/10.1128/mcb.25.21.9209-9220.2005>.
- Heras, S.R., Lopez, M.C., Olivares, M., Thomas, M.C., 2007. The LITc non-LTR retrotransposon of *Trypanosoma cruzi* contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. *Nucleic Acids Res.* 35, 2199–2214. <https://doi.org/10.1093/nar/gkl1137>.
- Higuera, S.L., Guhl, F., Ramírez, J.D., 2013. Identification of *Trypanosoma cruzi* Discrete Typing Units (DTUs) through the implementation of a High-Resolution Melting (HRM) genotyping assay. *Parasite. Vectors* 6, 1–6. <https://doi.org/10.1186/1756-3305-6-112>.
- Hoff, K.J., Stanke, M., 2013. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucl. Acids Res.* 41, W123–W128. <https://doi.org/10.1093/nar/gkt418>.
- Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S.M., Bianchinetti, G., Borzym, K., Bothe, G., Bruschi, C.V., Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R.M., Cronin, A., Cruz, A.K., Davies, R.M., De Gaudenzi, J., Dobson, D.E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A.C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J.C., Muller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O’Neil, S., Pentony, M., Pohl, T. M., Price, C., Purnelle, B., Quail, M.A., Rabinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J.C., Rutter, S., Saunders, D., Schafer, M., Schein, J., Schwartz, D.C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D.F., Blackwell, J.M., Stuart, K.D., Barrell, B., Myler, P.J., 2005. The genome of the kinetoplast parasite, *Leishmania major*. *Science* 309 (80-), 436–442. <https://doi.org/10.1126/science.1112680>.
- Kulkarni, M.M., Olson, C.L., Engman, D.M., McGwire, B.S., 2009. *Trypanosoma cruzi* GP63 proteins undergo stage-specific differential posttranslational modification and are important for host cell infection. *Infect. Immun.* 77, 2193–2200. <https://doi.org/10.1128/IAI.01542-08>.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10. <https://doi.org/10.1186/gb-2009-10-3-r25>.
- Lewis, M.D., Llewellyn, M.S., Gaunt, M.W., Yeo, M., Carrasco, H.J., Miles, M.A., 2009. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. *Int. J. Parasitol.* 39, 1305–1317. <https://doi.org/10.1016/j.ijpara.2009.04.001>.
- Li, L., Stoeckert Jr., C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. <https://doi.org/10.1101/gr.1224503>.

- Lima, L., Espinosa-Álvarez, O., Ortiz, P.A., Trejo-Varón, J.A., Carranza, J.C., Pinto, C.M., Serrano, M.G., Buck, G.A., Camargo, E.P., Teixeira, M.M.G., 2015. Genetic diversity of *Trypanosoma cruzi* in bats, and multilocus phylogenetic and phylogeographical analyses supporting Tcbat as an independent DTU (discrete typing unit). *Acta Trop* 151, 166–177. <https://doi.org/10.1016/j.actatropica.2015.07.015>.
- Macedo, A.M., Machado, C.R., Oliveira, R.P., Pena, S.D.J., 2004. *Trypanosoma cruzi*: Genetic structure of populations and relevance of genetic variability to the pathogenesis of chagas disease. *Mem. Inst. Oswaldo Cruz*. <https://doi.org/10.1590/S0074-02762004000100001>.
- Macedo, A.M., Pena, S.D.J., 1998. Genetic variability of *Trypanosoma cruzi*: Implications for the pathogenesis of Chagas disease. *Parasitol. Today*. [https://doi.org/10.1016/S0169-4758\(97\)01179-4](https://doi.org/10.1016/S0169-4758(97)01179-4).
- Macías, F., López, M.C., Thomas, M.C., 2016. The *Trypanosomatid* Pr77-hallmark contains a downstream core promoter element essential for transcription activity of the *Trypanosoma cruzi* L1Tc retrotransposon. *BMC Genomics* 17, 1–14. <https://doi.org/10.1186/s12864-016-2427-6>.
- Martin, F., Maranon, C., Olivares, M., Alonso, C., Lopez, M.C., 1995. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes. *J Mol Biol* 247, 49–59. <https://doi.org/10.1006/jmbi.1994.0121>.
- Messenger, L.A., Miles, M.A., Bern, C., 2015. Between a bug and a hard place: *Trypanosoma cruzi* genetic diversity and the clinical outcomes of Chagas disease. *Expert Rev. Anti. Infect. Ther.* 13, 995–1029. <https://doi.org/10.1586/14787210.2015.1056158>.
- Murcia, L., Carrilero, B., Muñoz-Davila, M.J., Thomas, M.C., López, M.C., Segovia, M., 2013. Risk factors and primary prevention of congenital chagas disease in a nonendemic country. *Clin. Infect. Dis.* 56, 496–502. <https://doi.org/10.1093/cid/cis910>.
- Olivares, M., Alonso, C., López, M.C., 1997. The open reading frame 1 of the L1Tc retrotransposon of *Trypanosoma cruzi* codes for a protein with apurinic-apyrimidinic nuclease activity. *J. Biol. Chem.* 272, 25224–25228. <https://doi.org/10.1074/jbc.272.40.25224>.
- Olivares, M., del Carmen Thomas, M., Lopez-Barajas, A., Requena, J.M., Garcia-Perez, J. L., Angel, S., Alonso, C., Lopez, M.C., 2000. Genomic clustering of the *Trypanosoma cruzi* nonlong terminal L1Tc retrotransposon with defined interspersed repeated DNA elements. *Electrophoresis* 21, 2973–2982. [https://doi.org/10.1002/1522-2683\(20000801\)21:14<2973::AID-ELPS2973>3.0.CO;2-4](https://doi.org/10.1002/1522-2683(20000801)21:14<2973::AID-ELPS2973>3.0.CO;2-4).
- Olivares, M., García-Pérez, J.L., Thomas, M.C., Heras, S.R., López, M.C., 2002. The Non-LTR (Long Terminal Repeat) Retrotransposon L1Tc from *Trypanosoma cruzi* Codes for a Protein with RNase H Activity. *J. Biol. Chem.* 277, 28025–28030. <https://doi.org/10.1074/jbc.M202896200>.
- Olivares, M., López, M.C., García-Pérez, J.L., Briones, P., Pulgar, M., Thomas, M.C., 2003. The endonuclease NL1Tc encoded by the LINE L1Tc from *Trypanosoma cruzi* protects parasites from daunorubicin DNA damage. *Biochim. Biophys. Acta - Gene Struct. Expr.* 1626, 25–32. [https://doi.org/10.1016/S0167-4781\(03\)00022-8](https://doi.org/10.1016/S0167-4781(03)00022-8).
- Olivares, M., Thomas, M.C., Alonso, C., López, M.C., 1999. The L1Tc, long interspersed nucleotide element from *Trypanosoma cruzi*, encodes a protein with 3'-phosphatase and 3'-phosphodiesterase enzymatic activities. *J. Biol. Chem.* 274, 23883–23886. <https://doi.org/10.1074/jbc.274.34.23883>.
- Python.org., 2020 Welcome to Python.org [Internet]. Available from: www.python.org/.
- Ramírez, J.C., Cura, C.I., Da Cruz Moreira, O., Lages-Silva, E., Juiz, N., Velázquez, E., Ramírez, J.D., Alberti, A., Pavia, P., Flores-Chávez, M.D., Muñoz-Calderón, A., Pérez-Morales, D., Santalla, J., Marcos Da Matta Guedes, P., Peneau, J., Marcet, P., Padilla, C., Cruz-Robles, D., Valencia, E., Crisante, G.E., Greif, G., Zulantay, I., Costales, J.A., Alvarez-Martínez, M., Martínez, N.E., Villarroel, R., Villarroel, S., Sánchez, Z., Bisio, M., Parrado, R., Maria Da Cunha Galvão, L., Da Câmara, A.C.J., Espinoza, B., De Noya, B.A., Puerta, C., Riarte, A., Diosque, P., Sosa-Estani, S., Guhl, F., Ribeiro, I., Aznar, C., Britto, C., Yadón, Z.E., Schijman, A.G., 2015. Analytical validation of quantitative real-time PCR methods for quantification of *trypanosoma cruzi* DNA in blood samples from chagas disease patients. *J. Mol. Diagnostics* 17, 605–615. <https://doi.org/10.1016/j.jmoldx.2015.04.010>.
- Rassi Jr., A., Rassi, A., Marcondes de Rezende, J., 2012. American trypanosomiasis (Chagas disease). *Infect Dis Clin North Am* 26, 275–291. <https://doi.org/10.1016/j.idc.2012.03.002>.
- Reis-Cunha, J.L., Bartholomeu, D.C., 2019. *Trypanosoma cruzi* genome assemblies: Challenges and milestones of assembling a highly repetitive and complex genome. *Methods Mol. Biol.* 1955, 1–22. https://doi.org/10.1007/978-1-4939-9148-8_1.
- Reis-Cunha, J.L., Rodrigues-Luiz, G.F., Valdivia, H.O., Baptista, R.P., Mendes, T.A.O., de Moraes, G.L., Guedes, R., Macedo, A.M., Bern, C., Gilman, R.H., Lopez, C.T., Andersson, B., Vasconcelos, A.T., Bartholomeu, D.C., 2015. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains. *BMC Genomics* 16. <https://doi.org/10.1186/s12864-015-1680-4>.
- Revollo, S., Oury, B., Laurent, J.P., Barnabé, C., Quesney, V., Carrière, V., Noël, S., Tibayrenc, M., 1998. *Trypanosoma cruzi*: Impact of clonal evolution of the parasite on its biological and medical properties. *Exp. Parasitol.* 89, 30–39. <https://doi.org/10.1006/expr.1998.4216>.
- Rodríguez, H.O., Guerrero, N.A., Fortes, A., Santi-Rocca, J., Gironès, N., Fresno, M., 2014. *Trypanosoma cruzi* strains cause different myocarditis patterns in infected mice. *Acta Trop* 139, 57–66. <https://doi.org/10.1016/j.actatropica.2014.07.005>.
- Rodríguez, P., Montilla, M., Nicholls, S., Zarante, I., Puerta, C., 1998. Isoenzymatic characterization of Colombian strains of *Trypanosoma cruzi*. *Mem. Inst. Oswaldo Cruz* 93, 739–740. <https://doi.org/10.1590/S0074-02761998000600008>.
- Sanchez-Luque, F., Lopez, M.C., Macías, F., Alonso, C., Thomas, M.C., 2012. Pr77 and L1TcRz: A dual system within the 5'-end of L1Tc retrotransposon, internal promoter and HDV-like ribozyme. *Mob Genet Elem* 2, 1–7. <https://doi.org/10.4161/mge.19233>.
- Sanchez-Luque, F.J., Lopez, M.C., Carreira, P.E., Alonso, C., Thomas, M.C., 2014. The wide expansion of hepatitis delta virus-like ribozymes throughout trypanosomatid genomes is linked to the spreading of L1Tc/ingi clade mobile elements. *BMC Genomics* 15, 340. <https://doi.org/10.1186/1471-2164-15-340>.
- Sánchez-Luque, F.J., López, M.C., Macías, F., Alonso, C., Thomas, M.C., 2011. Identification of an hepatitis delta virus-like ribozyme at the mRNA 5'-end of the L1Tc retrotransposon from *Trypanosoma cruzi*. *Nucleic Acids Res* 39, 8065. <https://doi.org/10.1093/nar/gkr478>.
- Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. <https://doi.org/10.1093/bioinformatics/btr026>.
- Seco-Hidalgo, V., De Pablos, L.M., Osuna, A., 2015. Transcriptional and phenotypical heterogeneity of *Trypanosoma cruzi* cell populations. *Open Biol* 5. <https://doi.org/10.1098/rsob.150190>.
- Seppely, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* 1962, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14.
- Stoco, P.H., Wagner, G., Talavera-Lopez, C., Gerber, A., Zaha, A., Thompson, C.E., Bartholomeu, D.C., Lückemeyer, D.D., Bahia, D., Loreto, E., Prestes, E.B., Lima, F.M., Rodrigues-Luiz, G., Vallejo, G.A., Filho, J.F.da S., Schenkman, S., Monteiro, K.M., Tyler, K.M., Almeida, L.G.P.de, Ortiz, M.F., Chiurillo, M.A., Moraes, M.H.de, Cunha, O.de L., Mendonça-Neto, R., Silva, R., Teixeira, S.M.R., Murta, S.M.F., Sincero, T.C.M., Mendes, T.A., de O., Urmenyi, T.P., Silva, V.G., DaRocha, W.D., Andersson, B., Romanha, A.J., Steindel, M., Vasconcelos, A.T.R.de, Grisard, E.C., 2014. Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. *PLoS Negl. Trop. Dis.* 8. <https://doi.org/10.1371/journal.pntd.0003176>.
- Teston, A.P.M., Monteiro, W.M., Reis, D., Bossolani, G.D.P., Gomes, M.L., de Araújo, S. M., Bahia, M.T., Barbosa, M.G.V., Toledo, M.J.O., 2013. *In vivo* susceptibility to benznidazole of *Trypanosoma cruzi* strains from the western Brazilian Amazon. *Trop. Med. Int. Heal.* 18, 85–95. <https://doi.org/10.1111/tmi.12014>.
- The Perl Programming Language - www.perl.org [WWW Document], 2020 URL <http://www.perl.org/>.
- Thomas, M.C., Macías, F., Alonso, C., López, M.C., 2010. The biology and evolution of transposable elements in parasites. *Trends Parasitol.* <https://doi.org/10.1016/j.pt.2010.04.001>.
- Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Br. Bioinform* 14, 178–192. <https://doi.org/10.1093/bib/bbs017>.
- Vargas, N., Pedroso, A., Zingales, B., 2004. Chromosomal polymorphism, gene synteny and genome size in *T. cruzi* I and *T. cruzi* II groups. *Mol. Biochem. Parasitol.* 138, 131–141. <https://doi.org/10.1016/j.molbiopara.2004.08.005>.
- Organization, World Health, 2020. Chagas disease (also known as American trypanosomiasis) [WWW Document]. WHO Media Cent. URL: [https://www.who.int/en/news-room/fact-sheets/detail/chagas-disease-\(american-trypanosomiasis\)](https://www.who.int/en/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)).
- Zingales, B., 2018. *Trypanosoma cruzi* genetic diversity: Something new for something known about Chagas disease manifestations, serodiagnosis and drug sensitivity. *Acta Trop.* <https://doi.org/10.1016/j.actatropica.2017.09.017>.
- Zingales, B., Miles, M.A., Campbell, D.A., Tibayrenc, M., Macedo, A.M., Teixeira, M.M., Schijman, A.G., Llewellyn, M.S., Lages-Silva, E., Machado, C.R., Andrade, S.G., Sturm, N.R., 2012. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect Genet Evol* 12, 240–253. <https://doi.org/10.1016/j.meegid.2011.12.009>.