PLOS PATHOGENS

RESEARCH ARTICLE

Strain-specific genome evolution in *Trypanosoma cruzi*, the agent of Chagas disease

Wei Wang[1], **Duo Peng**[1,2], **Rodrigo P. Baptista**[1,3], **Yiran Li**[3], **Jessica C. Kissinger**[1,3,4], **Rick L. Tarleton**[1,2]*

1 Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, Georgia, United States of America, **2** Department of Cellular Biology, University of Georgia, Athens, Georgia, United States of America, **3** Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States of America, **4** Department of Genetics, University of Georgia, Athens, Georgia, United States of America

* tarleton@uga.edu

Citation: Wang W, Peng D, Baptista RP, Li Y, Kissinger JC, Tarleton RL (2021) Strain-specific genome evolution in *Trypanosoma cruzi*, the agent of Chagas disease. PLoS Pathog 17(1): e1009254. https://doi.org/10.1371/journal.ppat.1009254

Editor: Nicolai Siegel, GERMANY

Received: August 17, 2020

Accepted: December 22, 2020

Published: January 28, 2021

Data Availability Statement: All raw and processed sequencing data generated in this study have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) under accession number SRX8355431-SRX8355434 (Brazil A4) and SRX8395372-SRX8395375(Y C6). The genome and annotation data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA512864 (Brazil A4) and PRJNA554625 (Y C6). The assembled and annotated genomes are also accessible in the TritrypDB database (https://tritrypdb.org/tritrypdb/).The strand-specific RNA-

Abstract

The protozoan *Trypanosoma cruzi* almost invariably establishes life-long infections in humans and other mammals, despite the development of potent host immune responses that constrain parasite numbers. The consistent, decades-long persistence of *T. cruzi* in human hosts arises at least in part from the remarkable level of genetic diversity in multiple families of genes encoding the primary target antigens of anti-parasite immune responses. However, the highly repetitive nature of the genome–largely a result of these same extensive families of genes–have prevented a full understanding of the extent of gene diversity and its maintenance in *T. cruzi*. In this study, we have combined long-read sequencing and proximity ligation mapping to generate very high-quality assemblies of two *T. cruzi* strains representing the apparent ancestral lineages of the species. These assemblies reveal not only the full repertoire of the members of large gene families in the two strains, demonstrating extreme diversity within and between isolates, but also provide evidence of the processes that generate and maintain that diversity, including extensive gene amplification, dispersion of copies throughout the genome and diversification via recombination and *in situ* mutations. Gene amplification events also yield significant copy number variations in a substantial number of genes presumably not required for or involved in immune evasion, thus forming a second level of strain-dependent variation in this species. The extreme genome flexibility evident in *T. cruzi* also appears to create unique challenges with respect to preserving core genome functions and gene expression that sets this species apart from related kinetoplastids.

Author summary

Many pathogens vary their surface antigenic profile in order to establish and maintain infections in the face of host immune responses. Although antigenic variation has been extensively documented in extracellular pathogens and is crucial in these cases to

seq reads are also available at SRA (SRX9261610 for Brazil A4 and SRX9264913 for Y C6). The scripts used for discovery of all members of the 6 largest multigene families is available at https://github.com/duopeng/Large-gene-family-search-pipeline. Scripts for the multidimensional display analysis pipeline are accessible at https://github.com/duopeng/Shiny-gene-families.

Funding: This work was supported by funding from the National Institutes of Health, (USA; nih.gov) grants R03 AI124228 and R01 AI124692 to RLT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

pathogen evasion of host antibody responses, there is scant understanding of the role that antigenic variation plays in immunity to intracellular pathogens, where cell-mediated immune responses are key to infection control and where a low frequency of switching from one predominant surface antigen to a new variant would be expected to have little impact on immune recognition. Herein we use comparative genome analysis to reveal the details of and mechanisms behind how the intracellular parasite *Trypanosoma cruzi*, agent of human Chagas disease, maintains a vast and varying array of antigens that are the targets of host immune responses. The process of diversification is so efficient that two isolates share not a single identical gene among the thousands of antigenic variants in their genomes, thus making the likelihood of generating protective vaccines, low. This genome flexibility also ensnares genes whose products are not targets of immune responses, thus further driving the isolate-specific biological diversity that characterizes this species.

Introduction

The protozoan parasite *Trypanosoma cruzi* is the causative agent of Chagas disease, the highest impact parasitic infection in the Americas, affecting 10 to 20 million humans and innumerable animals in many species. The study of *T*. *cruzi* and Chagas disease is particularly challenging for a number of reasons, including the complexity and unique characteristics of its genome. Over 50% of the *T*. *cruzi* genome is composed of repetitive sequences, which include numerous families of surface proteins (e.g. *trans*-sialidases, mucins and mucin-associated surface proteins) with hundreds to thousands of members each, as well as substantial numbers of transposable elements, microsatellites and simple tandem repeats [1–3]. This repetitive nature greatly hampered the assembly of the original CL Brener strain reference genome generated in 2005, resulting in a highly fragmented and draft assembly with extensively collapsed high repeat regions [2]. In addition, the CL Brener strain turned out to be a hybrid strain with divergent alleles at many loci. To scaffold the genome sequence, Weatherly *et al*. took advantage of the bacterial artificial chromosome (BAC) library sequencing data and combined with synteny analysis of two genomes from closely related species, *Trypanosoma brucei* and *Leishmania*, obtained the current reference genome with 41 chromosomes [4]. Nevertheless, a large number of gaps are still present in the chromosomes of the reference genome, and many unassigned contigs remain, making it impossible to determine the exact genome content and, in particular, the full repertoires of large gene families.

As in many pathogens, and best documented in the related trypanosomatid *Trypanosoma brucei*, families of variant surface proteins often serve as both the primary molecular interface with mammalian hosts and as the predominant target of host immune responses. Classical antigenic variation in these pathogens consists of the serial expression of a single (or a highly restricted number of) antigen variant(s) in the pathogen population at any one time, with switches to new variants becoming evident once the host immune response controls the dominant one. This largely "one-at-a-time" strategy appears particularly effective in pathogens exposed continuously to antibody-mediated immune control mechanisms. *T*. *cruzi*, however, appears to take a much different approach to antigenic variation, generating multiple very large families of genes encoding surface and secreted proteins, many of which are expressed simultaneously rather than serially. We believe that this strategy may reflect the primarily intracellular lifestyle of *T*. *cruzi* in mammalian hosts and the necessity of evading T cell recognition of infected host cells, although this has not yet been formally proven.

The advent of two advances in genome analysis has made it feasible to revisit and substantially improve upon the *T*. *cruzi* genome assembly and to advance our understanding of its composition. The long-read capability of PacBio Single-Molecule Real-Time (SMRT) sequencing provides read lengths capable of spanning long repetitive regions. The application of this technology [5–7] as well as nanopore sequencing [8], has resulted in much-improved contiguity and expansion of the members of large gene families in *T*. *cruzi*. Secondly, proximity ligation methods have allowed for the scaffolding of assemblies spanning highly repetitive regions. One of the methods, Hi-C, identifies extant inter-chromosomal interactions by capturing chromosome conformation, and has been used to create scaffolds at chromosomal scale [9–11]. A second approach termed Chicago, adapts this same methodology but reconstitutes the conformation of DNA *in vitro* by combining the DNA with purified histones and chromatin assembly factors [12]. These proximity ligation methods not only improve the contiguity of genomes by joining contigs, they also identify misjoins in the contigs and separate them to increase the accuracy of assemblies [12]. The combination of Chicago and Hi-C has now been applied to many genomes [13–17].

In this study, we have applied SMRT sequencing and proximity ligation methods to produce very high-quality assemblies from the Brazil (TcI) and Y (TcII) strains of *T*. *cruzi*. These two strains are representatives of the most ancestral lines that are hypothesized to have given rise to the 6 discrete typing units (DTUs, TcI-TcVI) lineages now composing this genetically diverse species [18–23]. Using these chromosomal-level assemblies with minimal gaps, we are now able to compare the full gene content of representatives of these founding lineages of the *T*. *cruzi* species, including the full repertoires of large gene families. The six largest of these gene families are of particular interest because each contains over 200 members and substantial numbers of pseudogenes, the latter likely a product of gene duplication and recombination events. The three largest of these families, the *trans*-sialidases, mucins and mucin-associated surface proteins (MASPs) also serve as the primary targets of host immune responses and are thus under immune pressure. Herein, we document a substantial diversity in individual chromosome content, including frequent allelic variants, but with an overall conserved gene content outside of the large gene families. Within these largest gene families, however, extreme diversification is evident, with no identical copies between the two strains, of genes in these families. These high-quality genomes also reveal the mechanisms behind the expansion and diversification of the large gene families, presumably in response to immunological pressure, and in the process, creating other challenges in terms of core genome stability and function.

Results

Genome sequencing and assembly

PacBio SMRT sequencing provided 1,264,527 (N50 = 9,560 bp) and 763,579 (N50 = 12,499 bp) filtered reads with ~9 Gb and ~6 Gb of sequence data for Brazil clone A4 (Brail A4) and Y clone C6 (Y C6), respectively, corresponding to ~200x and ~130x coverage based on the predicted genome size. Initial assembly resulted in sequences of 45.11 Mb and 46.98 Mb for Brazil A4 and Y C6 draft genomes, respectively, close to the estimated haploid genome size of *T*. *cruzi* [24] (Table 1). Furthermore, we applied the proximity-ligation tools, Hi-C and Chicago [12] to scaffold the draft assembly with the generation of joins and breaks (S1 Table). Density histograms mapped with Hi-C reads are provided in S1 Fig for scaffolds > 1Mb. The application of these two libraries decreased the L50 to half of that of the draft genomes, and the size of the largest scaffolds doubled (Table 1). Gap extension and base correction using Illumina reads ultimately resulted in 12 and 14 scaffolds in the Brazil A4 and Y C6 final assemblies, respectively, with a length greater than 1 Mb. Telomeric repeats [(TTAGGG)n] were identified

Table 1. Summary of assembly statistics.

Genome	Method used and coverage	Total size (Mbp)	Number of contigs or scaffolds	GC (%)	N50 (bp)	L50	Largest contig or scaffold length (bp)	# of gaps
Brazil A4								
Draft	PacBio RSII (200x)	44.97	677	51.56	227,072	48	1,236,815	0
Scaffolded	Chicago (125x) and Hi-C (46,451x)	45.00	402	51.56	907,746	18	2,710,165	295
Final	PBJelly, Pilon and iCorn	45.56	402	51.58	914,771	17	2,738,928	295
Y C6								
Draft	PacBio Sequel (130x)	46.98	351	51.57	410,475	33	1,547,313	0
Scaffolded	Chicago (2,096x) and Hi-C (21,551x)	47.00	266	51.57	890,993	18	2,951,407	106
Final	PBJelly, Pilon and iCorn	47.22	266	51.58	889,019	18	2,951,016	106

* Draft assemblies are sequences with zero gaps.

https://doi.org/10.1371/journal.ppat.1009254.t001

in 18 Brazil A4 and 15 Y C6 scaffolds, including on both ends of three scaffolds in Brazil A4, suggesting full chromosome assembly in these cases. The improvement in these new genomes is not only in integrity (S2 Table and S2 Fig), but also in filled gaps, recovered genes and deconvoluted repetitive regions (see examples in S3 Fig).

Genome features and content

Due to a lack of apparent chromosome condensation during replication [24,25], the karyotype of *T. cruzi* has not been completely elucidated. Moreover, chromosome size and content vary significantly between different *T. cruzi* strains and even among clones of the same strain based upon pulse-field gel electrophoresis (PFGE) analysis [25–29]. Based on criteria including size, repeat proportion, and gene number, 43 scaffolds of Brazil A4 and 40 scaffolds of Y C6 were designated as chromosomes (S4 Fig) and the remainder referred to as smaller scaffolds.

Repetitive sequences occupy 58.8% and 62.3% of the genome for Brazil A4 and Y C6 (S3 Table), substantially higher than the 50% that was estimated in the reference CL Brener genome, thus confirming the capability of long-read sequencing and assembly approaches to recover and place more repetitive DNA content. Approximately 50% of the sequence in chromosomes is repetitive sequences, compared to ~90% in smaller scaffolds (S4 Fig). Using conventional approaches with manual curation, gene models were identified in Brazil A4 and Y C6, respectively. We employed Benchmarking Universal Single-Copy Orthologs (BUSCO, v3.0.2) [30] to evaluate the completeness of Brazil A4 and Y C6 assembly in comparison to three other published long-read assembled *T. cruzi* genomes for which annotations are available [5,31]. Searching against single-copy orthologs that are expected to be present in either eukaryote lineage orthologous groups, or protist lineage orthologous groups, Brazil A4 and Y C6 contain the highest number of gene sets in one haplotype among all these genomes (S4 Table).

A major constituent of the repetitive regions in the *T. cruzi* genome is large gene families, including the *trans*-sialidases (TS), mucin associated surface proteins (MASP), mucins, and surface protease GP63 (all targets of immune responses), as well as retrotransposon hotspot (RHS) proteins and dispersed gene family 1 proteins (DGF-1) [2,32,33]. Our previous studies indicated the total copy number of TS genes was underestimated using conventional annotation approaches due in part to the failure to identify new variants and fragments of TS resulting from frequent recombination [34]. To complete the annotation of the members of large gene families, we developed a customized workflow (summarized in S5 Fig) and applied it to the six largest gene families. This allowed us to capture the full repertoire of the largest and

Fig 1. Distribution of large gene families and synteny between chromosomes in Brazil A4 (right, B) and Y C6, (left, Y). Tracks from outer to inner rings: chromosome number, TS, MASP, mucin, GP63, and synteny blocks.

https://doi.org/10.1371/journal.ppat.1009254.g001

most diverse gene families (S5 Table, copy numbers of which are summarized in S6 Table). Members of these gene families are unequally distributed among and along the chromosomes, with several of the largest chromosomes (e.g. TcBrA4_Chr2 and TcYC6_Chr1) composed nearly entirely of members of these large gene families (Figs 1 and S6). In contrast to previous reports suggesting the members of large gene families were mainly located in telomeric and subtelomeric regions [2,7], members of large gene families are not restricted to particular regions of chromosomes. Moreover, TS, MASP, mucin and GP63 have an overlapping distribution along the chromosomes, while RHS and DGF-1 genes are more dispersed.

After consolidating the predictions of large gene families with our conventional annotations, the Brazil A4 and Y C6 genomes contained 18,708 and 17,650 gene models, respectively (see annotation summary in S7 Table). The composition of gene content between two genomes is very similar, with ~25% as members of 6 largest gene families, ~40% as hypothetical proteins, and >90% of the remaining genes as orthologs of those in the related kinetoplastids *T. brucei* and *L. major*. That this gene model count in the two *T. cruzi* strains is substantially higher than that estimated for *T. brucei* and *L. major* is likely due to two factors: 1) the high number of members of large gene families in *T. cruzi*, and 2) a greater number of hypothetical genes in *T. cruzi*, a third of which are unique to *T. cruzi*, although the size distribution of the hypothetical proteins is similar in the 3 species (S7 Fig).

Allelic variation

The significant number of small scaffolds and the relatively high gene model numbers in some of them prompted us to consider whether these small scaffolds might represent regions of

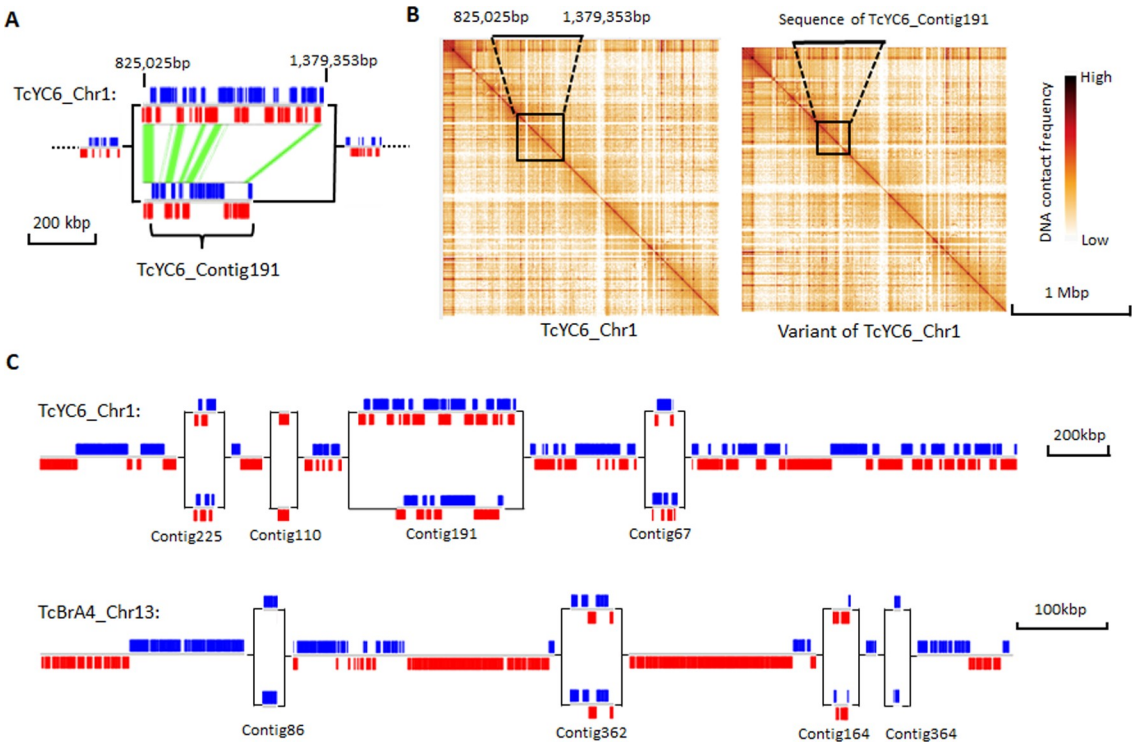

Fig 2. An example of homologous chromosomes with large allelic variations. (A) Synteny between two allelic variants in Chr1 of Y C6. Synteny blocks are marked with green. (B) Hi-C heat maps of TcYC6_Chr1 (left) and its homologous chromosome with TcYC6_Contig191 (boxed area) replacing the allelic region in TcYC6_Chr1 (boxed area). (C) Two chromosomes with multiple allelic variants. Blue blocks indicate genes on the forward strand, and red blocks indicate genes at the reverse strand.

https://doi.org/10.1371/journal.ppat.1009254.g002

allelic variation between sister chromosomes, as allelic variation is one of the factors that results in fragmentation during genome assembly for diploid genomes. Although TcI and TcII DTUs represented by the Brazil and Y strains, respectively, are considered homozygous lineages, we very conservatively detected 26 and 33 small scaffolds in each genome showing consistent synteny in multiple gene models to parts of the core chromosomes (S8 Table). An example is shown in Fig 2A in which scaffold TcYC6_Contig191 demonstrates regions of synteny within the 825,025–1,379,353 bp region in the first chromosome of Y C6 (TcYC6_Chr1). Confirmation of this chromosome variant was supplied by replacing the identified region in TcYC6_Chr1 with TcYC6_Contig191 and then mapping the chromosomal contacts in the Hi-C data for these 2 alternative versions for TcYC6_Chr1. As shown in Fig 2B, the Hi-C data are equally strong for both chromosome variants. Using Falcon-Phase, which phases diploid genome sequences by integrating long reads and Hi-C data [35], we identified an additional 18 and 7 allelic variations in Brazil A4 and Y C6, respectively. In combination, these analyses identified allelic variations in 24 chromosomes of Brazil A4 and 25 of Y C6, including chromosomes with multiple allelic variants, e.g. the largest chromosome in Y C6 (TcYC6_Chr1), and an intermediate-sized chromosome in Brazil A4 (TcBrA4_Chr13; Fig 2C). Thus, we suggest that many of the small scaffolds are variants of regions in the chromosome-size scaffolds. However, because the majority of these small scaffolds lack the conserved, non-gene family sequences required to prove synteny, and Falcon-Phase can only resolve haplotypes bearing divergence of $< 5\%$, identifying the position of all the small scaffolds on the chromosomes was not possible.

Structural comparison of the Brazil and Y sequences

The very high genome quality and contiguity provided by the combination of SMRT sequencing and Hi-C analysis enabled chromosome level comparison of the Brazil (TcI) and Y (TcII) clones (Fig 1). The synteny plots show that the majority of chromosomes from one genome are collinear with those in the other genome (synteny blocks with 1–1 orthologous pairs are summarized in S9 Table). For instance, Brazil A4 Chr4 showed continuous synteny to Y C6 Chr4 overall. However, as expected based upon previous gene mapping studies [26,36–38], some chromosomes corresponded to different regions in multiple chromosomes of the other genome, e.g. Brazil A4 Chr1 showed synteny to a combination of Chr20 (298,235–684,393bp), Chr9 (63,384–95,053bp) and Chr2 (20,327–1,438,658bp) in Y C6. Some inverted syntenies were also detected, e.g. between 388,900–968,190bp on Brazil A4 Chr8 and 11,711–556,982bp on Y C6 Chr16 (Fig 1). Notably, the diversity of sequences encoding members of the large gene families (see details below) prevented the detection of synteny in a substantial proportion of the two genomes, including in two of the largest chromosomes (e.g. TcBrA4_Chr2 and TcYC6_Chr1).

Variation in gene models within and between Brazil and Y strains is predominantly in the large gene families

A large number of genetic variations were identified in the non-repetitive regions, including heterozygous SNPs/Indels within respective strains (S10 Table), and homozygous SNPs/Indels between the two strains (S11 Table). We also detected aneuploidy in both genomes: 3 and 8 chromosomes in Brazil A4 and Y C6, respectively, exist in copy numbers greater than two, based on the results of both relative read depth and allele frequency (S8 Fig). Among these are the partially syntenic chromosomes (TcBrA4_Chr24 and TcYC6_Chr10), which also share synteny with Chr31 in CL Brener, reported to be supernumerary in many strains [39], thus suggesting a species-wide requirement for >2 copies of one or more genes in these regions. Additionally, variations exist in the copy number for a substantial number of individual genes characterized by OrthoFinder [40], with ~150 genes showing the greatest variation between the two strains (S12 Table). However, with respect to genes unique to either strain, we found 23 (Brazil A4) and 20 (Y C6) unique gene loci not present in the other strain and further validated this finding by examining the raw reads (S13 Table). All are annotated as hypothetical proteins, and most are small genes located in large gene family-rich regions of the genome and thus are likely the products of recombination events involved in gene family diversification (see below).

To fully assess the variation in the members of the 6 largest gene family between the two strains, we carried out a best match search for the protein sequence of putatively expressed genes in each family from Y C6 genome with those in Brazil A4 (S14 Table). As a control, the same analysis was performed for a subset of 291 mostly single-copy genes (BUSCO), as well as a small gene family of 35 members, beta galactofuranosyl glycosyltransferase (b-gal GT). As shown in Fig 3, high-identity matches could always be found for the BUSCO genes, and some of them (22 out of 291) have identical matches (100% identity) in the other strain. Similar to BUSCO genes, the identity between best matches for b-gal GT is also tightly distributed in the range of 90–97%. In contrast, all six large gene families exhibit a broad distribution of identity for their best matches relative to the BUSCO genes and b-gal GT genes, especially TS, MASP, mucin and RHS, with only a small proportion of best matches bearing 90% identity or more. Among the family members with the greatest similarity between the two strains are the small subset of TS genes containing the sialidase enzymatic domain as previously described [41], suggesting that this group of *trans*-sialidases has been selected for and conserved in both strains.

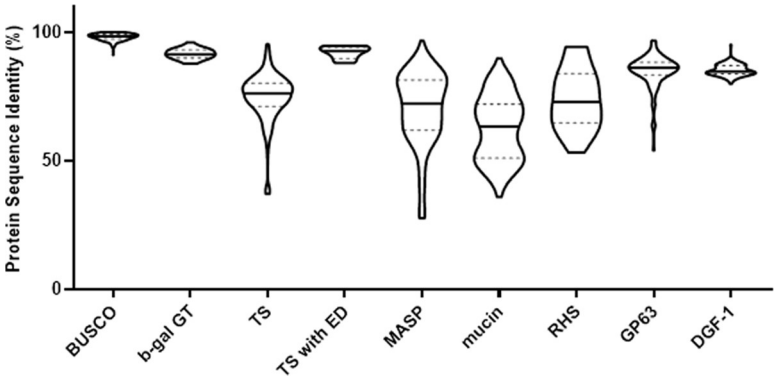

Fig 3. Protein best match analysis of gene families between Brazil A4 and Y C6. Solid lines indicate median and dashed lines indicate quantiles.

https://doi.org/10.1371/journal.ppat.1009254.g003

Evidence of large gene family expansion and diversification

The very high number and the impressive within- and between- strain variation in the genes composing the largest gene families in *T. cruzi* is indicative of a system under intense evolutionary pressure. We have taken advantage of the high contiguity of these two genome sequences, as well as the comprehensive prediction of all members of the 6 largest gene families, to attempt to understand better how this remarkable diversity is generated and maintained.

We first examined the genomes for evidence of gene duplication events that could increase the number of members in gene families. Multidimensional scaling (MDS) plots based on the pair-wise genetic distances of all members of each large gene family in each strain allowed us to identify tightly distributed gene clusters with high sequence identity (http://shiny.ctegd.uga.edu). In multiple cases, genes within these clusters were tandemly arrayed individually (TS; Fig 4A top) or as a set of genes (TS plus MASP; Fig 4A bottom). Such tandem amplifications are present in all large gene families (except DGF-1) and occur uniquely in each strain (S15 Table). A number of unusual amplification events were also noted, including inverted duplications creating a strand switch in between (Fig 4B), and an amplification involving several genes on both strands, replicated a total of 5 times (Fig 4C), thus creating a complex set of strand switches.

The majority of tandem amplification of genes in the 6 largest gene families in both *T. cruzi* genomes contained 10 or fewer replicates (S15 Table). However, one hypothetical protein (*HP) in the Y C6 occurs in a tandem array of 29 units with a TcMUCI gene (Fig 4D). Comparison to the syntenic region in Brazil A4 revealed a single TcMUCI ortholog (and no *HP sequence), indicating that at some point the *HP sequence was inserted next to the TcMUCI gene in Y C6, and the two genes were amplified together as a segment (Fig 4D). Although no particular protein domains were characterized in the *HP gene, 18% of its sequence shares similarity with several MASP sequences, implying that at least part of the gene might be derived from a MASP. The abnormally high number of replicates in this tandem array as well as the low diversity in the tandem copies suggested that this might be a recent amplification event. However, comparison to syntenic regions in other *T. cruzi* genomes sequenced with long-read technologies revealed the same TcMUCI+*HP tandem array in the TCC (TcIV) strain but not in the Dm28c (TcI), Sylvio (TcI), and Bug2148 (TcV) (Fig 4D). Additionally, phylogenetic analysis grouped all of the replicated copies from Y C6 and TCC together and distant from the single TcMUCI genes in the other 4 strains (Fig 4E). Using the model of DTU evolution in *T. cruzi* which postulates that the TcVI is derived from a hybridization event

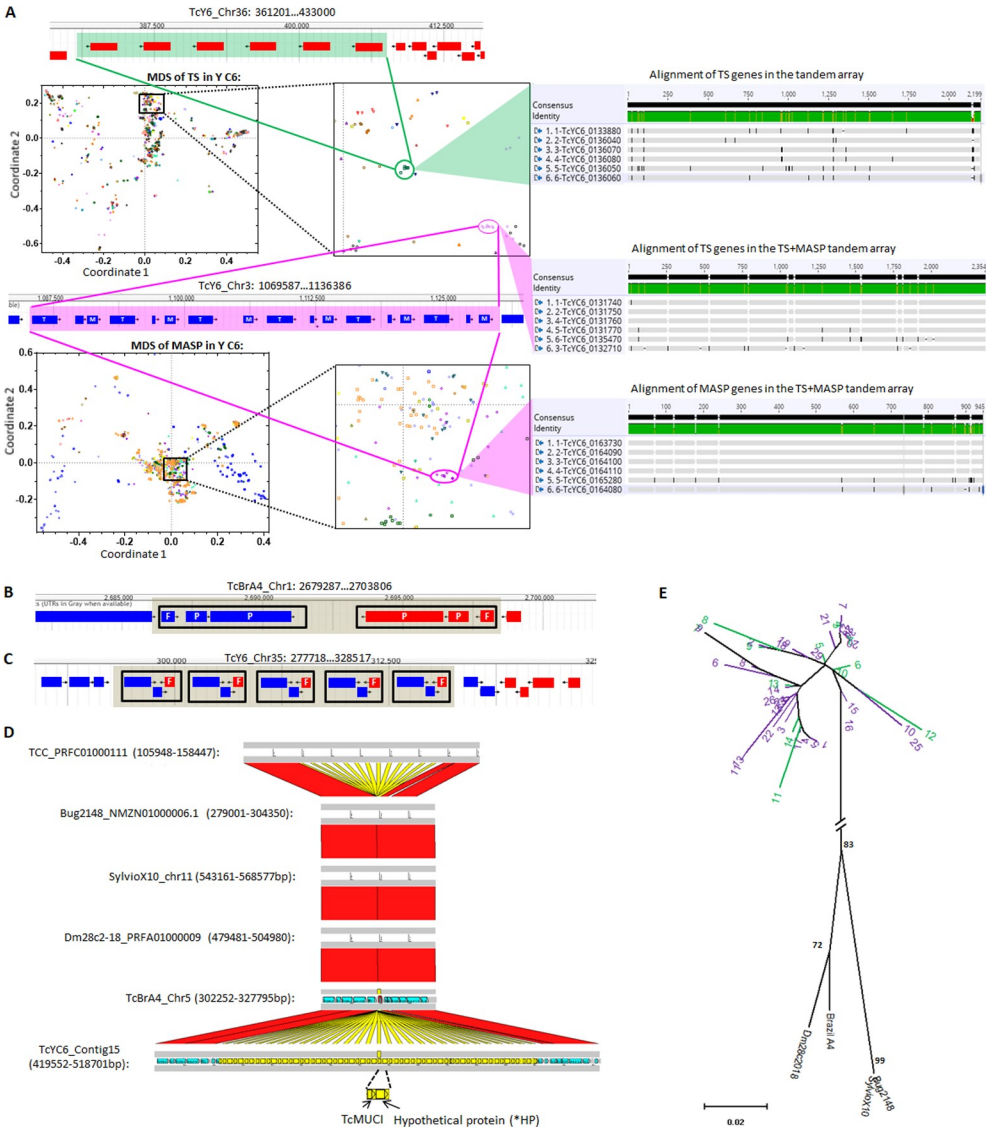

Fig 4. Gene amplification events in members of large gene families. (A) Tandem arrays of individual TS genes (top), and a TS+MASP pair (bottom) clustered based upon genetic distance in the MDS plots. Each chromosome is displayed as a separate pattern on the MDS plot. T: TS; M: MASP. Alignment of the genes in each MDS cluster (right) confirms high consensus (grey regions); black regions indicate SNPs and '–' indicate gaps. (B) Mirror-duplication of one fragmented RHS and two RHS pseudogenes. P: pseudogene; F: fragment. (C) One RHS (+), one hypothetical protein (+) and one fragmented glycosyltransferase (-) replicated 5 times, creating multiple strand switches. F: fragment. (D) Syntenic regions of the TcMUCI+*HP tandem array detected in 6 long-read sequenced *T. cruzi* strains. Synteny of TcMUCI orthologs are labeled in yellow. (E) Bayesian inference of phylogeny of all TcMUCI orthologs from the 6 strains. Note that TcMUCI genes from Y C6 (purple) and TCC (green) are intermingled in the top portion of the tree, indicating the retention of high similarity in these lineages, and are collectively distant from their next nearest mucins in 4 *T. cruzi* genomes lacking this array. Live MDS plots can be explored at http://shiny.ctegd.uga.edu. Alignments were performed using Geneious software (v11.0.4).

https://doi.org/10.1371/journal.ppat.1009254.g004

between TcII and TcIII [18,23], we propose that the TcMUCI+*HP amplification is an ancient event, occurring after the split of TcI and TcII but prior to the TcII/TcIII hybridization that yielded TcVI.

In addition to tandem clusters of genes in these large gene families, MDS analysis also revealed closely related family members located on multiple chromosomes (Fig 5A). An

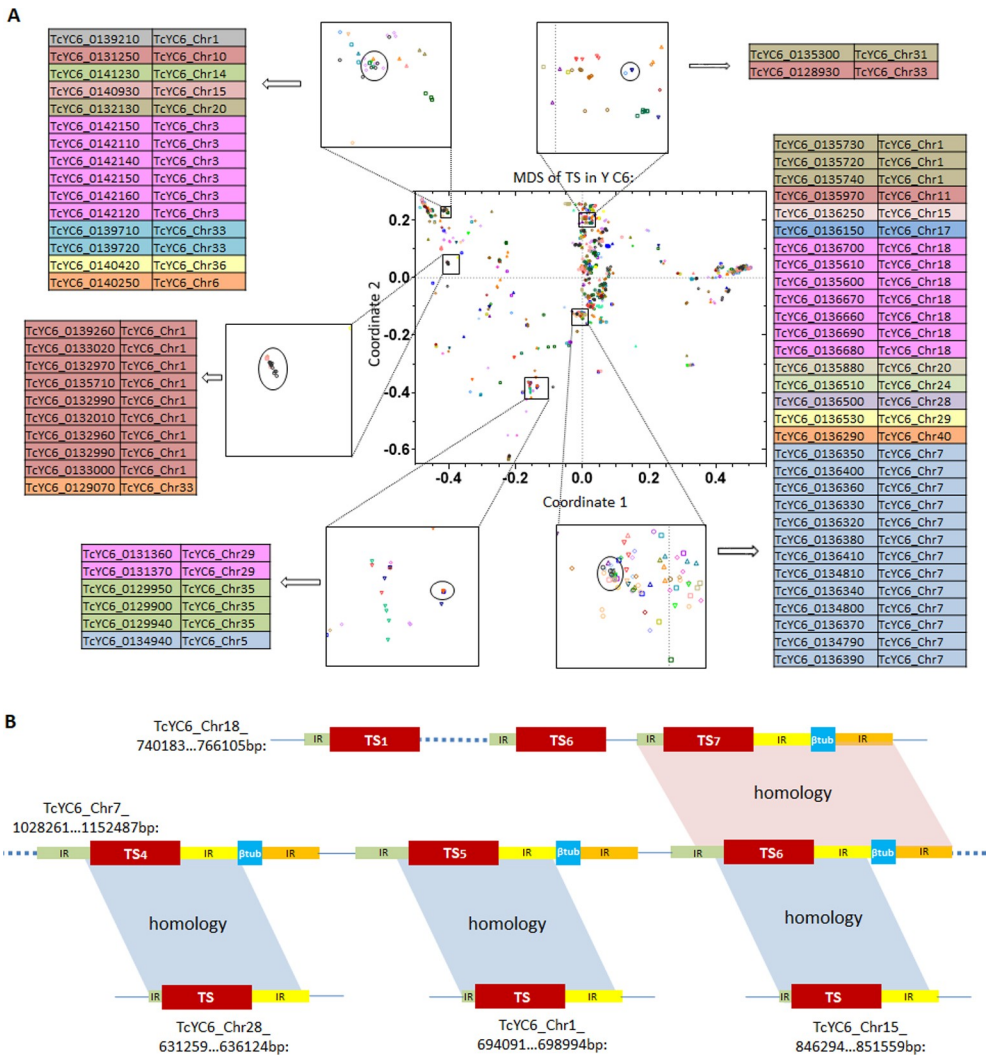

Fig 5. Examples of relocations of TS genes in Y C6. (A) Tight clusters of TS genes from MDS plot are distributed on different chromosomes. (B) Diagram of relocations in one of the TS clusters on the bottom right in (A). Blocks in the same color indicate genes or flanking sequences in high identity. Note that the segment size is not to scale. IR: intergenic region.

https://doi.org/10.1371/journal.ppat.1009254.g005

extreme case is the Y C6 gene cluster in the bottom right of Fig 5A which contained 31 TS genes with very high similarity distributed on 11 different chromosomes (S9A Fig). Interestingly, the 13 TS genes on Chr7 (Fig 5B, middle) are in tandem, interspersed with a beta tubulin gene, while the 7 TS genes on Chr18 (Fig 5B, top) are in tandem as TS genes alone (with one beta tubulin gene downstream of TS_7). The remaining 11 TS genes in this cluster are dispersed in the genome as singlets (3 of them are shown at the bottom of Fig 5B). Notably, the sequences upstream and downstream of the TS gene in the TS + beta tubulin array on Chr7 are homologous to those of the TS_7 gene in the Chr18 array, and the dispersed singlet TS also share a portion of the upstream and downstream sequences with the other TS in this cluster (S9B Fig). Together, these results suggest that all 31 TS genes in this cluster originated from one or more gene amplification/relocation events. Based on the phylogenetic analysis (S9C Fig), we propose that the TS + beta tubulin tandem copies have been generated in or relocated to Chr7 (13 copies) and Chr18 (1 copy), with another 4 TS copies as single genes beyond the TS + beta tubulin

Table 2. Recombination events detected within genes of large gene families in Brazil A4 and Y C6.

	Brazil A4				Y C6				CL Brener
	TS	MASP	Mucin	GP63	TS	MASP	Mucin	GP63	TS
# of genes	1644	1118	700	411	1465	1066	797	427	3209
Kb length total	3477.7	1011.9	352.1	460.6	2614.5	1115.2	458.9	619.7	4456.5
# of genes recombined	793	145	38	70	479	154	73	89	787
# of recombination events	2976	190	39	101	1334	221	85	153	2087
% of genes recombined	48.2	13.0	5.4	17.0	32.7	14.4	9.2	20.8	24.5
Average events per gene	1.8	0.2	0.1	0.2	0.9	0.2	0.1	0.4	0.7
Average events per kb	0.9	0.2	0.1	0.2	0.5	0.2	0.2	0.2	0.5
Number of genes with 'n' number of recombination events									
n = 1	137	111	37	51	162	110	61	58	324
n = 2	198	24	1	15	114	28	12	19	149
n = 3	98	9	0	1	66	11	0	3	110
n = 4	128	1	0	1	54	3	0	6	72
n = 5	68	0	0	1	33	2	0	1	52
n>5	164	0	0	1	50	0	0	2	80
Max of n	18	4	2	5	12	5	2	6	12

Max of n: the highest number of recombination events detected for one gene.

https://doi.org/10.1371/journal.ppat.1009254.t002

cassette on Chr18, while the single TS genes on other chromosomes may derive from the TS on Chr7.

We next used a pipeline previously designed to identify recombination events within TS genes in the CL Brener genome [34], to quantify recombination for 4 of the large gene families in the Brazil and Y strains (Table 2). This pipeline uses the RDP4 package that employs a variety of methods to detect signals of recombination and then approximates breakpoints and the recombinant sequence [42]. As expected, recombination events, including multiple events acting on the same gene, were detected in a large fraction of the genes but were particularly abundant (2-fold higher) in the TS family relative to the other three families examined. Interestingly, recombination events in the TS family were detected at a roughly 2-fold higher frequency in the Brazil strain as compared to the Y or the CL Brener strains.

As noted previously, our recombination pipeline is highly conservative in detecting relatively recent events that have not been obscured by subsequent accumulation of SNPs and Indels [34]. Such *in situ* diversification is evident in genes that are clustered in the MDS analysis but dispersed in the genome. An example is a cluster of GP63 genes in Brazil A4 which have low genetic distance based on MDS analysis (S10A Fig), but are located on different chromosomes and display a considerable degree of variation (SNPs and Indels; S10B Fig). However, because these genes also share similar upstream genes (a TS) and intergenic regions, all of these dispersed genes were likely derived via gene duplication. This hypothesis is further supported by the result that 9 out of 10 GP63 genes and their corresponding GP63 + flanking sequences (including upstream TS + intergenic region + GP63 + intergenic region) occupy identical positions in their respective phylogenetic trees (S10C Fig). Therefore, a TS/GP63 gene pair and associated intergenic regions underwent one or more duplication and relocation events with subsequent diversification through the accumulation of SNPs and Indels, yielding multiple, diverse genes spread through the genome.

The potential complexity generated by amplification, relocation, recombination and diversification make it challenging to track the specific set of events contributing to the evolution of

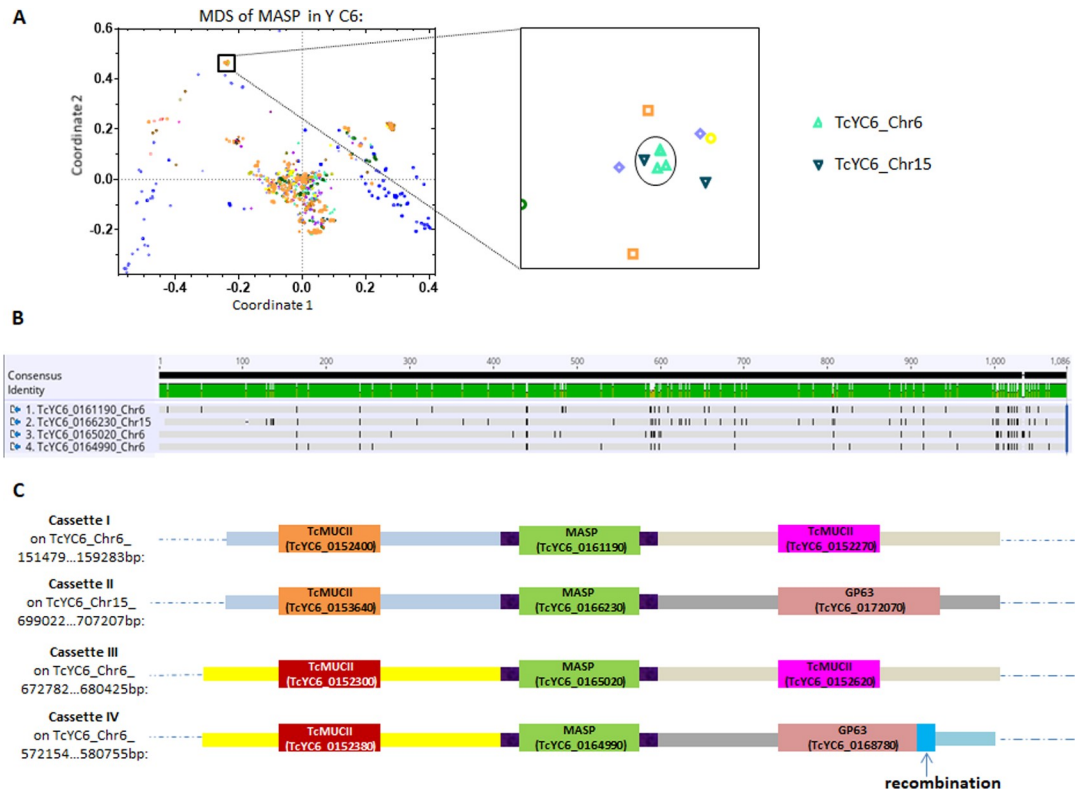

Fig 6. The combination of gene amplification, relocation, recombination and *in situ* diversification of members of large gene family. (A) A tight cluster of 4 MASP genes from the MDS plot are distributed on two chromosomes. (B) Alignment of the 4 MASP genes shows high identity with modest diversification of SNPs/Indels. Alignment was analyzed using the same method as in Fig 4. (C) MASP genes with flanking intergenic sequences and flanking genes. Blocks with the same color indicate sequences in high identity. Note that the segment sizes are not to scale.

https://doi.org/10.1371/journal.ppat.1009254.g006

individual members of these large gene families in *T. cruzi*. However, some gene sets reveal all of these processes at work. Fig 6C shows four cassettes located on different chromosomes or in distant sites on the same chromosome, each cassette with a central MASP and flanking region with high identity (Fig 6A and 6B), suggesting a common origin. SNPs/Indels indicate *in situ* diversification of the MASP genes, especially in the C terminus (Fig 6B). Cassette pairs I/II and III/IV share the same upstream gene and flanking sequence (mucin genes in both cases) while cassette pairs I/III and II/IV shared downstream mucin and GP63 genes, respectively. In addition, a recombination event was detected in the C terminus of the GP63 in cassette IV, creating divergence from the GP63 C terminus in cassette IV.

Potential impact of high genome flexibility on gene expression

Unlike in other classical models of antigenic variation in protozoa, the large gene families in *T. cruzi* are not restricted to particular regions of chromosomes (e.g. subtelomeric in the case of *T. brucei* [11,43–46]) but instead are spread throughout the genome (Fig 1). This presents the complication that the amplification and dispersion events common in the large gene families of *T. cruzi* might also impact non-gene family (core) genes as well. To investigate this possibility, we focused on core genes for which there were > 6 total paralogues for the two genomes and organized these paralogues on the basis of gene location (S12 Table). By doing this, we could identify tandemly distributed genes that likely resulted from gene amplification. For the

over 150 groups of genes in this analysis, many showed dramatic differences in gene copies in the two *T. cruzi* genomes with 26 instances of double-digit gene copies in one strain compared to only 1–3 copies in the other. This same high level of variation was evident for other *T. cruzi* genomes sequenced using long-read sequencing methods but not in similarly sequenced *T. brucei* and *Leishmania* isolates (S11 Fig). Additionally, dispersion patterns for these amplified genes differed widely between the Y C6 and Brazil A4 genomes (S12 Table). Thus, the mechanisms that provide for the generation and maintenance of diversity in the large gene families also appear to allow for substantial variation in copy number for selected core genes, representing a second major contributor to between-strain genetic variation in *T. cruzi* strains.

Most gene expression in trypanosomatids initiates in the absence of specific promoters and with the production of multi-gene mRNA transcripts that are then processed into single-gene mature mRNAs. These polycistronic transcriptional units (PTUs) of genes can be well over >100 kb in length and are marked by start and stop signals, including base modifications [47]. The apparent wide degree of freedom for amplification and dispersion both within and outside the *T. cruzi* large gene families, and particularly events that create tandem strand switches as shown in Fig 4C, would be expected to impact this normal multi-gene PTU structure. Indeed, the average PTU length was 116.5 and 126.8 kb in the core gene-rich regions of the Brazil A4 and Y C6, respectively, similar to that in *T. brucei* (148.3 kb). However, the average PTU length in the gene-family-enriched regions of both *T. cruzi* genomes was less than ¼ of that (29.3 kb in Brazil A4 and 33.8 kb in Y C6), indicating a disruption of the normal PTU structure. Interestingly, amplified but conserved tandem gene arrays like the 'mucin + *HP' array in Y C6 discussed above (Fig 4D) and the previously described TcSMUG family [48–50] are within large PTUs containing almost no members of the large gene families (S12A Fig) while many other tandem arrays or apparently diverging genes reside in gene-family-rich, short PTUs (S12B Fig). The disruption in PTU structure might also hamper the preservation of transcriptional control mechanisms, in particular the tight controls on transcriptional termination characterized in other kinetoplastids and mediated by base J and histone H3/4 variants [11,51–57]. To address this question, we mapped strand-specific RNA-seq reads to both sense and antisense strands to assess transcriptional termination relative to PTUs. Surprisingly, we found extensive antisense RNA levels throughout the genome (an average of sense:antisense = 114:1 in Brazil A4 and 84:1 in Y C6). Higher levels of antisense RNAs occurred at the strand switch regions of long PTUs (Fig 7A and 7B), but in some cases, matched or exceeded the sense strand transcripts in gene-family-enriched regions containing shorter PTUs (Fig 7C). Thus, unlike *T. brucei* and *Leishmania*, *T. cruzi* does not appear to tightly regulate antisense RNA production.

Discussion

T. cruzi is a highly heterogeneous species, with at least six DTUs and with extreme variation in phenotype and virulence among isolates even of the same DTU. Gaining new understanding of the genetic basis of this high strain-to-strain variation in the disease-causing potential in *T. cruzi* has been challenging due to the lack of a high-quality reference genome, and notwithstanding the excellent long-read sequencing-based assemblies recently provided [5–8], comparable quality genomes of diverse genetic types for in-depth comparison. The high content of repetitive sequences in the *T. cruzi* genome (>50%), including multiple families of surface protein-encoding genes each with >200 members, makes complete genome assembly from conventional short-read sequences impossible. This study reports very high-quality genomes for *T. cruzi* strains belonging to the presumed ancestral lineages of this species, TcI, represented here by the Brazil strain and TcII by the Y strain. This significantly improved resource was achieved by the application of long-read sequencing techniques and proximity ligation

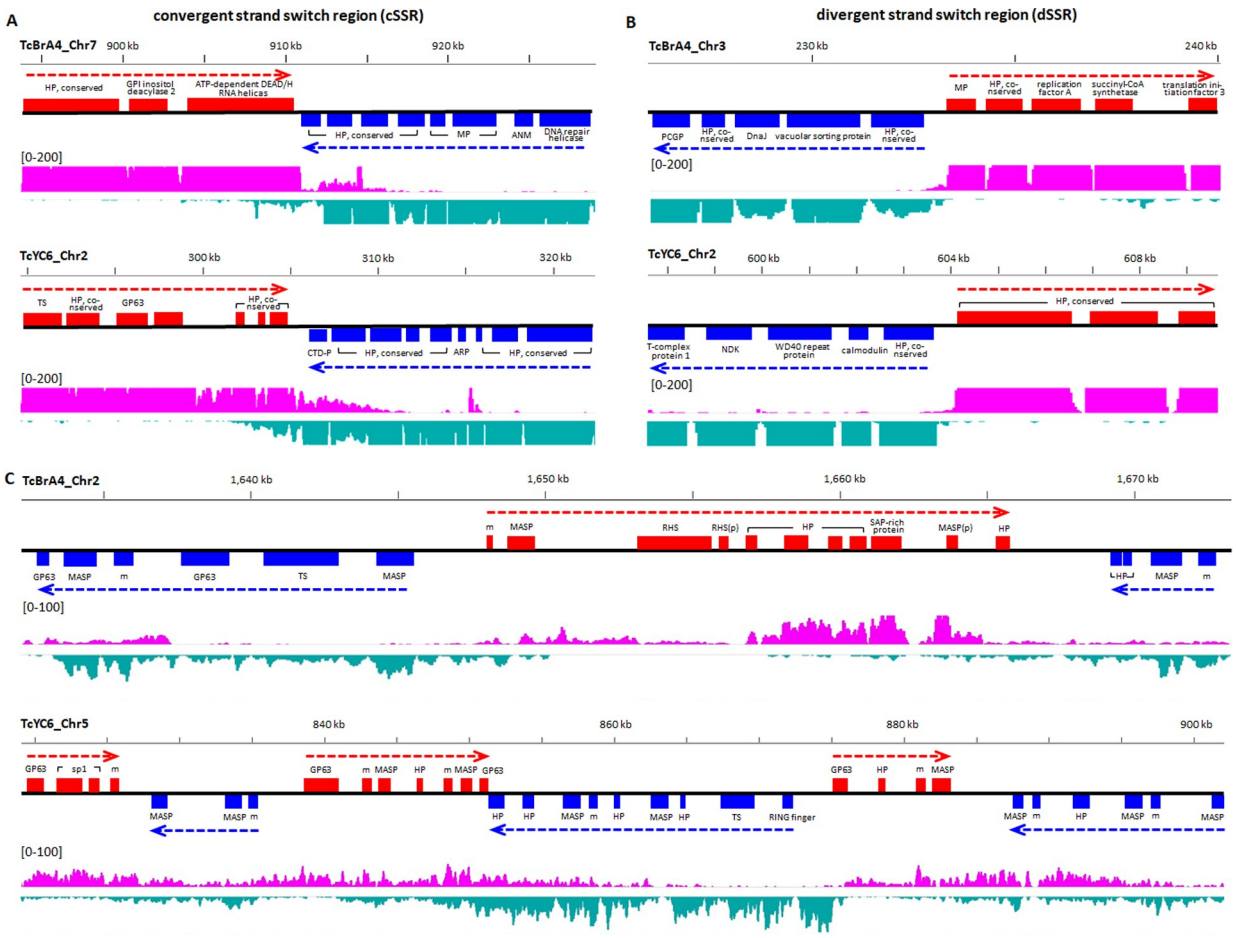

Fig 7. Antisense RNA levels in *T. cruzi* in relation to PTU structure, including both convergent strand switch regions (cSSR, A) and divergent strand switch regions (dSSR, B). (C) Gene-family-enriched regions with frequent strand switches where antisense RNA was detected in higher levels. HP, hypothetical protein; MP, mitochondrial protein; ANM, arginine N-methyltransferase; PCGP, parkin coregulated gene protein; CTD-P, TFIIF-stimulated CTD phosphatase; ARP, ankyrin repeat protein; NDK, nucleoside diphosphate kinase; SAP-rich protein, serine-alanine-and proline-rich protein; m, mucin; p, pseudogene.

https://doi.org/10.1371/journal.ppat.1009254.g007

libraries to better resolve the full repertoires of gene content, thus allowing a detailed comparison of genetic variation between these strains.

Although DTU-specific associations have been frequently proposed for characteristics such as virulence, disease presentation, geographic distribution, and host species restrictions, many of these linkages falter when more extensive sampling is done and none has been linked to DTU-specific genetic differences [58–60]. The current dataset provides the opportunity to begin examination of representative strains of *T. cruzi* lineages that diverged from each other an estimated 1–3 million years ago [23]. The most surprising revelations from this comparative analysis were not the variability in unique gene content between these isolates, but rather the extremes of the high similarity in core gene content and the comparative huge diversity in gene family-rich portions of the genomes. As anticipated based upon previous strain-based screens [39,61], a considerable degree of variation exists in the form of SNPs/Indels and additionally, a substantial number of strain-specific copy number variations were identified. However, the core (non-gene family) genome, contains only ~20 strain-unique gene models, and in all cases, these are hypothetical genes encoding proteins with no recognizable protein domain structures.

In very sharp contrast, the variation evident in the large gene families of *T*. *cruzi* is equally remarkable, demonstrating vast diversity within and between strains with no perfect matches and relatively few genes of the same family with even a 90% similarity. Structurally, these gene families comprise ~25% of the genome and their members are spread widely throughout the genome, with some on every chromosome and some of the largest chromosomes being almost entirely composed of members of large gene families. The use of synteny detection tools and Falcon-Phase validated by Hi-C methods allowed us to also conservatively document heterozygosity in more than half of the chromosomes in each genome, and we suspect that this heterozygosity extends to nearly all gene family-rich regions of the genome. Based upon the total base count of the repeat-rich small scaffolds not assigned to chromosomes, we estimate that up to 50% of all members of the large gene families have variants on the sister chromosome.

The quality of the genome assemblies also provided the opportunity to document the continuing diversification of these large gene families and to permit an understanding of how this process might work. Select members of the large gene families in *T*. *cruzi* have clear and critical functions in parasite biology, with the best-documented example being the enzymatically-active *trans*-sialidases required for acquisition of sialic acid by *T*. *cruzi* trypomastigotes [41,62–64]. However, the number, diversity, and potential for variation of genes in these large gene families, and the exposure of the gene products to and response by the host immune system, argue that these gene families evolve under intense immunological pressure. In all these respects, the three largest and most diverse gene families in *T*. *cruzi* (TS, MASP and mucin) are similar to other families of genes involved in antigenic variation in the protozoans *T*. *brucei* (variant surface glycoproteins, VSGs), *Plasmodium* (*var* genes) and Giardia (Variant-specific Surface Protein, VSPs) [65–69]. However in contrast to the "one-at-a-time" models of classical antigenic variation best characterized in the sister kinetoplastid *T*. *brucei* [65], *T*. *cruzi* expresses many gene family variants simultaneously. This difference in strategy likely relates to the fact that *T*. *cruzi* lives predominantly intracellularly in mammals and must effectively evade cell-mediated (rather than exclusively antibody-mediated) immunity. But expressing many antigen variants at one time may also require a larger antigen repertoire and/or an enhanced ability to generate new variants and a genetic system with the flexibility that such generation entails.

Classically, segmental duplication creates the source material on which mutational and recombinational events act to derive new genes and new gene functions [70]. The presence of segmental duplications (one gene or multiple genes as a unit) also encourages additional rounds of duplications that can rapidly change gene content [71–73]. These processes of gene duplication, recombination and mutation-driven diversification, functioning in concert to ensure high and constant antigenic diversity, is strongly evident in the large gene families of *T*. *cruzi*. Although we are able to track a significant number of these events, all occurring independently in these two *T*. *cruzi* strains, we are presumably only observing the most recent occurrences, as recombinations and mutations ultimately obscure the origins of new genes. Certainly the repeat-rich structure and dense representation of retrotransposons of the *T*. *cruzi* genome facilitates maintenance of these processes and the dispersion of members of large gene families throughout the genome, and interestingly not restricted to chromosome ends as is the case in *T*. *brucei* [11,43–46]. However, the specific structural elements that initially established and continue to allow for these apparently constant rearrangements throughout the genome, but without impacting overall genome integrity, remain unidentified. From our analysis, no consistent pattern of structures, such as the A/T tracks associated with gene application events in *Plasmodium* [74] were evident.

The apparent high frequency and continued evolution of gene families in *T*. *cruzi* also create structures and products apparently unique among closely related kinetoplastids, including

the lack of segregation of large gene families to chromosome ends, absence of partitioning of expression sites (as in *T*. *brucei* VSGs[45,75,76]), tolerance for the generation of short PTUs and frequent strand switching, and most surprisingly, an abundance of antisense RNAs. The latter may well explain why a high functioning RNAi system like that present in *T*. *brucei* is absent in *T*. *cruzi* [77,78]. The presence of abundant and nearly genome-wide antisense RNAs also suggests that *T*. *cruzi* does not adhere to the full set of rules for transcription termination as defined in *T*. *brucei* and *Leishmania* [57,79] and may explain the differential impact of base J knockdown in *T*. *cruzi* relative to other kinetoplastids [80].

Interestingly, there are several subsets of members of these large gene families that appear to be exceptions to these processes of recombination, diversification and distribution throughout the genome. The previously characterized SMUG families are the best examples. Two subgroups of TcSMUG genes, TcSMUG L and S, involved in development and infectivity of insect-dwelling stages [48–50], distribute as tandem arrays in the respective subgroups within the same PTU and exhibit minimal diversification. Here we also identify an ancient, lineage-specific duplication event that created a new hypothetical gene and a mucin gene in a tandem array and which, like the SMUGS, has remained with minimal changes. It will be of interest to determine if further diversification of this and other gene family subsets are restricted because of their location in the genome, or if, like the SMUGS, this hypothetical gene/mucin tandem is under selective pressure due to their unique function. One common feature of these tandem arrays is that they all locate in and are flanked by large PTUs (>220 kb) containing only core genes with no members from the large gene families (other than the mucins in the mucin +*HP array), suggesting that they are maintained in an environment largely devoid of large-gene-family-related diversification.

An additional strain-dependent difference documented here is the higher recombination frequency in Brazil A4 compared to Y and in CL Brener [34]. The ~2X greater number of recombination events in all large gene families in Brazil vs Y suggests that this is an inherent property of this strain and perhaps of DTUI strains in general. Alternatively, because we very conservatively call recombination events which then eventually become concealed by further mutations/recombinations over time, it is also possible that the Brazil A4 has been under stronger, or more recent, selective pressure.

The apparent high levels of gene amplification and dispersion readily documented in the large gene families in this species also extends to a fraction of core genes as well, and represents a second major source of between-strain diversity and perhaps the one primarily responsible for the broad between-strain phenotypic variation in *T*. *cruzi* (S12 Table). Retention of these core gene amplifications implies a fitness benefit, perhaps under certain environmental/host conditions; others may also occur regularly but engender a fitness cost and thus are lost.

In summary, the careful analysis of these two *T*. *cruzi* strains soundly confirms the vast genetic diversity of parasite lines within this species, and identifies the bulk of diversity to be represented in 3 compartments: 1) rapidly evolving families of genes involved in immune evasion, 2) a subset of "core" genes not linked to evasion but which vary greatly in copy number and perhaps expression, and 3) SNPs and Indels common to all genomes. We hypothesize that the gene family diversity is driven by immune selection and that the same processes that provide for this diversity also allow for copy number variation and diversification of select core genes. We believe it is this latter process, rather than DTU type, that accounts for much of the biological diversity of *T*. *cruzi* lines. With these high-quality genomes in hand for these strains, we can now test these hypotheses by further modifying these gene sets and exposing both wild-type and modified parasite lines to various levels of selection pressure and observing the genomes of the lineages that emerge.

Materials and methods

Parasite cultures, DNA/RNA extraction and sequencing

Epimastigotes of Brazil and Y strains were cultured at 26˚C in supplemented liver digested-neutralized tryptose (LDNT) medium as described previously [81]. Single-cell clones were made for each strain by depositing epimastigotes into a 96-well plate at a density of 0.5 cell/well by using a MoFlow cell sorter (Dako-Cytomation, Denmark). One healthy clone that has been confirmed to have cycled through all life stages was chosen for sequencing for each strain. High molecular weight DNA was isolated using MagAttract HMW DNA kit (Qiagen) before submitting to Duke Center for Genomic and Computational Biology (GCB) for SMRT sequencing. Brazil A4 was sequenced using PacBio RS II sequencer, while PacBio Sequel sequencer was used for Y C6.

Genomic DNA of the selected clone of both strains was isolated using QIAamp DNA blood mini kit (Qiagen) for whole genome sequencing using Illumina HiSeq 150 PE. An RNase treatment step was included to eliminate RNA in the samples. For RNA-seq sampling, extracellular amastigotes and trypomastigotes isolated from infected Vero cells were pooled with epimastigotes for total RNA-extraction. Following ribo-depleted RNA library construction, RNA sequencing using Illumina Nextseq 75PE was performed by Georgia Genomics and Bioinformatics Core (GGBC). Illumina reads from either DNA or RNA sequencing with mean quality lower than 30 (Phred Score based) were removed for analysis.

Genome assembly

The draft genome of Brazil A4 was assembled with SMRT Link v3.1, and Y C6 with SMRT Link v5.0. The parameters were set to default except for the expected genome size, which was set to 40 Mb for both strains. Chicago and Hi-C libraries were constructed and sequenced by Dovetail Genomics, and HiRise pipeline was run for scaffolding the draft assembly by incorporating data from both libraries.

The Chicago library was prepared as described previously [12]. Briefly, ~500ng of HMW gDNA (mean fragment length = 48Kbp) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library.

The Dovetail HiC library was prepared in a similar manner as described previously [82]. Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. Both Chicago and Hi-C libraries were sequenced on an HiSeqX to produce 2x151 bp paired end reads.

The draft assembly, Chicago library reads, and Dovetail HiC library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation

data to scaffold genome assemblies [12]. An iterative analysis was conducted. First, Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). Only reads with map quality > = 50 (uniquely mapped reads) are retained. The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After aligning and scaffolding Chicago data, Dovetail HiC library sequences were aligned and scaffolded following the same method. A gap was generated whenever two contigs were joined by HiRise, and since the distance between two contigs was unknown, all gaps were given 100 Ns.

Gap filling was performed by PBJelly [83] using the SMRT subreads with the minimum percent identity at 85% and minimum coverage at 5. There were 122 and 4 gaps extended for Brazil and Y, respectively. Correction of the genomes using Illumina short reads was run by Pilon [84] and iCORN2 [85] through multiple iterations to eliminate errors from SMRT sequencing.

Repeat annotation

RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler) was used to build a *de novo* repeats library using 'tcruzi' database, and then used RepeatMasker v 4.0.7 (http://www.repeatmasker.org) with search engine parameter as "ncbi" to annotate all the repetitive sequences.

Genome annotation

To develop open reading frame (ORF) in the new genome sequences, WebApollo 2.0 [86] was deployed with the genome sequence, and the following tracks of evidence were added:

1. Gene prediction from COMPANION [87] using *Trypanosoma brucei* as reference.
2. Gene prediction using AUGUSTUS [88,89] which was self-trained by CL Brener genome.
3. Annotation transfer from CL Brener by Exonerate [90].
4. ESTs from available EST sequencing libraries in *T*. *cruzi* (retrieved from https://tritrypdb.org/tritrypdb/app/record/dataset/DS_6889a51dab).
5. Proteins from available Mass spec data for *T*. *cruzi* [91].
6. Strand-specific RNA-seq alignment data, the pipeline of which was followed as previously described [79].

Each ORF along the genome was manually produced by the integration of all tracks.

InterProScan v5.31–70.0 [92] was used to detect protein families, domains and sites with all 11 default databases. Gene Ontology (GO) term was assigned by InterProScan based on the protein domains results. Besides, BLASTP was used to search protein homology against *T*. *cruzi* CL Brener, *T*. *brucei*, *L*. *major* databases from TriTrypDB release 39 (https://tritrypdb.org/tritrypdb/) and RefSeq non-redundant protein database, respectively, to determine the best hit for protein naming by in-house scripts. The parameter used for BLASTP was E value <1e-10, identity >70% and coverage (length of alignment/length of target protein) >70%. Predicted pseudogenes were named by homology in RefSeq non-redundant nucleotide database with E value <1e-30.

Annotation of large gene families

A customized computational pipeline automated using PERL and Python scripts were developed for identifying members of large gene families in the genome (https://github.com/

duopeng/Large-gene-family-search-pipeline). First, the annotated members of each gene family were searched against the Brazil A4 and Y C6 genome using BLASTN (version ncbi-blast-2.8.1+) with num_alignments and max_hsps arguments set to 100, the perc_identity argument set to 85. BLAST hits that have an overlap longer than 100 bp were merged if they match members from the same gene family. BLAST hits that were bracketed by longer hits from the same family were removed. A minimum length cutoff of 150 bp was applied to the BLAST hits. The remaining BLAST hits were considered new family member gene candidates. BLASTN was used to match the new candidate genes to all annotated transcripts from the genome of *T. cruzi* CL Brener strain (TriTrypDB release 34) (BLAST argument settings: num_alignments and max_hsps set to 50, perc_identity not set). Candidate genes were retained only if one of its top two best matches is a member of the candidate gene's corresponding gene family.

Next, the boundaries of the candidate genes were refined by using model genes of each family in two steps. (1) Extending the candidate gene boundaries to include possible segments missed by previous steps. Using model gene sequences of each family to search the new genomes, and compare the coordinates of the matches to that of candidate genes. If $> 50\%$ overlap was found, and the non-overlapping length was $< 1{,}000$ bp, then the boundary of the candidate was extended according to the genomic match of the model gene. (2) The boundary of candidate genes was next subjected to small-scale trimmings. The candidate genes were mapped against model genes of the corresponding gene family (num_alignments and max_hsps arguments set to 100, the perc_identity argument set to 85). If a match was found within 100bp distance to the boundary of candidate genes, the candidate gene boundary was trimmed to match that of the model gene.

The start of mucin candidate genes was further refined using a conserved signal peptide sequence (in an alignment format allowing for minor variations). BLASTN was used to match the signal peptide sequences to mucin candidate genes (BLAST argument settings: num_alignments and max_hsps set to 200, perc_identity set to 65, gapopen and gapextend set to 1). Sequence upstream of signal peptide matches in the candidate genes was removed.

A final trim was applied to the boundaries of all candidate genes, as many of our BLAST steps could lead to inaccurate boundary identification due to 25% chance of a random matching to an extra nucleotide base at the boundary and 6.25% chance for two extra bases and so forth, which could obscure start and stop codons. As an attempt to address this issue, we trimmed up to 10 bases which could reveal a start/stop codon that is in-frame with an existing stop/start codon.

Manual corrections of boundaries for members of large gene families were performed when necessary.

Comparative analysis and synteny detection

Protein sequences of the annotated genes in Brazil A4 were used for BLASTP search against all genes in Y C6, and only the top two hits with an e-value $< 10^{-5}$ were kept as homologous pairs. These homologous pairs were used for syntenic block detection through MCscanX [93] with a match score of 50, match size of 5, gap penalty of -1, overlap window of 5, e-value of 1e-5, and max gaps of 25. The output of MCScanX was further parsed using in-house script, and was submitted to Circa (OMGenomics, http://omgenomics.com/circa/) to draw comparative plot. The same workflow was applied to detect synteny blocks between smaller scaffolds and chromosomes for identifying allelic variations.

SNP calling

Illumina reads were first filtered by removing bases with quality score < 30, and then mapped to the genome with Bowtie2 using the parameters—*maxin 900—no-discordant—no-mixed*.

SNP calling was performed by the HaplotypeCaller module in the Genome Analysis Toolkit (GATK) version 3.4 [94] under default parameters.

Clustering of orthologous groups

The clustering of orthologous groups for *T. cruzi*, *T. brucei* and *Leishmania* strains was carried out using OrthoFinder [40] with default parameters. All sequences were retrieved from TriTrypDB database (https://tritrypdb.org/tritrypdb/) release-44.

Identification of recombination events

Discovery of gene recombination events was done as previously described [34]. In brief, the pipeline used the RDP4 program with default parameters (for details see the RDP4 handbook at http://web.cbio.uct.ac.za/~darren/martin%202015.pdf). RDP4 first detects recombination signals by employing a variety of methods, including RDP, BOOTSCAN, MAXCHI, CHIMAERA, 3SEQ, GENECONV, LARD, and SISCAN (see references for these in the RDP4 handbook above). Following the detection of a 'recombination signal' with these methods, RDP4 determines approximate breakpoint positions using a hidden Markov model, BURT, and then identifies the recombinant sequence using the PHYLPRO, VISRD, and EEEP methods.

Hi-C contact matrix

Hi-C contact matrix were analyzed by following the manual of https://github.com/hms-dbmi/hic-data-analysis-bootcamp/, and then visualized in HiGlass [95].

Multidimensional scaling

K-tuple distance between genes was calculated with Clustal-Omega 1.2.4 [96] using unaligned sequences with option parameters: "—full" and "—distmat-out". Full alignment distance between genes was calculated with Clustal-Omega 1.2.4 using aligned sequences (aligned with Clustal-Omega 1.2.4 using default parameters) with option parameters:—full—full-iter—distmat-out. MDS is performed with the "cmdscale" function built-in R 3.6.3 with the input of a matrix of either pairwise K-tuple distances or full alignment distances. The results of MDS are visualized using the Shiny package 1.4.0.2 in R 3.6.3. Scripts are available at https://github.com/duopeng/Shiny-gene-families.

Phylogenetic inference

Multiple sequence alignment was performed using MUSCLE [97]. The resulting alignment was manually edited. Bayesian inference of phylogeny was performed using MrBayes v.3.2.6 [98] with the following parameters: nst = 6, rates = invgamma, Ngammacat = 8, Ngen = 10,000,000, nruns = 2, nchains = 4, and burn-infraction = 0.5. Convergence was determined by 25,000 post burn-in samples from two independent runs. The resulting phylogenetic tree was rendered in Figtree v.1.4.4. Node support values are given in percent posterior probability.

Supporting information

S1 Fig. Link density histogram mapped with Hi-C reads. The x and y axes give the mapping positions of the first and second read in the read pair, respectively, grouped into bins. The color of each square gives the number of read pairs within that bin. White vertical and black horizontal lines have been added to show the borders between scaffolds. Only scaffolds > 1

Mb are shown.
(TIF)

S2 Fig. Overview of the Brazil A4 and Y C6 genomes. Tracks from outer to inner circles indicate: lengths, chromosome number, gaps, gene density (window size: 20kb, range: 6–23 in Brazil A4, 1–22 for Y C6), GC content (window size: 10kb, range: 0.36–0.70 in Brazil A4, 0.33–0.70 in Y C6), repetitive content (window size: 10kb, range: 0–10000), heterozygous SNPs (window size: 20kb, range: 0–120 in Brazil A4, 1–390 in Y C6) and heterozygous Indels (window size: 20kb, range: 65–1 in Brazil A4, 129–1 in Y C6).
(TIF)

S3 Fig. Assembly improvement compared to CL Brener. (A) An example of filled gaps. Syntenic regions between Chr1 in Brazil A4 and Chr8 in CL Brener were aligned with the Artemis Comparison Tool (ACT) [99]. All five gaps were filled in Brazil A4. (B) An example of recovered genes. Two pieces of an adenosine monophosphate (AMP) gene were identified flanking a gap, while the syntenic region in Brazil A4 shows the intact AMP gene. (C) An example of extended repeats. With 8 copies of histone H4 in Chr2 of CL Brener separated by a gap, the syntenic region of Brazil had the gap filled, extending the copy number of histone H4 to 41. Solid while boxes: gaps; green bars: genes.
(TIF)

S4 Fig. Repetitive composition of the scaffolds. Chromosomes are calculated individually, while small scaffolds are calculated by averaging a range of scaffolds as indicated on the x axis.
(TIF)

S5 Fig. Workflow used to predict full repertoire of large gene families (taking TS as an example).
(TIF)

S6 Fig. Distribution of large gene families and retrotransposons on the chromosomes. Rings from outer to inner: chromosome number, retrotransposons, TS, MASP, mucin, GP63, RHS and DGF-1 gene families.
(TIF)

S7 Fig. Size distribution of hypothetical proteins identified in kinetoplastids. Genomes of *T. brucei* TRE92 and *L. major* Friedlin were downloaded from TriTrypDB database (https://tritrypdb.org/tritrypdb/) release-44 [100].
(TIF)

S8 Fig. Estimation of chromosome copy number. (A) Relative read depth of each chromosome normalized to the mean read depth of all chromosomes at non-repetitive regions. Chromosomes with more than two copies are indicated in red. (B) Allele frequency calculated by the proportion of heterozygous SNPs/Indels at the non-repetitive regions of each chromosome. 'Counts (%)' on the Y axis indicate the percentage of SNPs/Indels called at certain frequency which was calculated as previously described [39]. A diploid chromosome shows the peak of allele frequency around 50% as shown in Chr8 and Chr31, whereas an aneuploid chromosome shows peak of allele frequency lower than 50% as shown in Chr24 and Chr28 in Brazil A4. Note that 5 chromosomes (Chr35, 36, 38, 39 and 42) in Brazil A4 were not included in this analysis due to their high proportion of repetitive features.
(TIF)

S9 Fig. Alignment of TS (A) **and their flanking regions** (B) **of the cluster in Fig 5, as well as the Bayesian inference of phylogeny of all the TS genes** (C). IR: intergenic region.

Alignments were analyzed using the same method as in Fig 4.
(TIF)

S10 Fig. An example of *in situ* diversification. (A) A tight cluster of GP63 genes from MDS plot are distributed in different chromosomes. (B) Alignment of these GP63 genes showed high identity as well as a number of diversifications including SNPs and Indels. (C) Bayesian inference of phylogeny of GP63 in the cluster (left), and GP63 plus flanking sequences on both sides (right). IR: intergenic region.
(TIF)

S11 Fig. Correlation analysis of copy number variation in kinetoplastids. Copy numbers of 152 orthologous gene sets from S12 Table are highly correlated (Spearman correlation > 0.7) in pairwise comparisons between *T*. *brucei* strains and subspecies and between *Leishmania* species, but poorly correlated between *T*. *cruzi* strains (Spearman correlation in a range between 0.006 and 0.6).
(TIF)

S12 Fig. Examples of long or short PTUs with tandem gene arrays. (A) Tandem arrays of conserved gene sets are contained within long PTUs devoid of large gene family members. Chromosomes containing TcSMUG S/L in Brazil A4 (top) and Y C6 (middle), and TcMUCI +*HP in Y C6 (bottom). In contrast, members of the large gene families, including some tandemly duplicated genes, show frequent strand switches and are in short PTUs (B). Blue bars indicate genes other than members of large gene families, while yellow bars indicate members of these gene families.
(TIF)

S1 Table. Number of joins and breaks generated by Chicago or Hi-C libraries.
(PDF)

S2 Table. Summary of assembly metrics among all available *T*. *cruzi* genomes assembled by long-read sequencing. All sequences were retrieved from TriTrypDB database (https://tritrypdb.org/tritrypdb/) release-44. *No scaffolding was applied to these genomes, so no gaps were generated. **47 are not *de novo* assembled contigs or scaffolds, but rather pseudomolecules produced by aligning the core regions of scaffolds to the core regions of CL Brener reference genome. Therefore, although the genome showed higher N50 and lower L50, it left an extensively high number of gaps behind. ***Genome sequence is not available.
(PDF)

S3 Table. Repetitive sequences characterized in Brazil A4 and Y C6.
(XLSX)

S4 Table. Assessment of genome assembly and annotation completeness using single-copy ortholog benchmarking. 5 PacBio-assembled genomes with annotation were compared using either 'eukaryota_odb9' or 'protists_ensembl' datasets using default parameters. The 'eukaryota_odb9' dataset contains 303 single-copy genes conserved in 100 eukaryote species, while the "protists_ensembl' dataset' contains 215 single-copy genes that are present among 33 protist species. *Note that TCC is a hybrid strain, so its genome is a mixture of two haplotypes, while all other genomes contain one haplotype.
(PDF)

S5 Table. Members of 6 largest gene families in the Brazil A4 and Y C6 genomes.
(XLSX)

S6 Table. Copy number of large gene families characterized in the new genomes.
(PDF)

S7 Table. Annotation summary.
(PDF)

S8 Table. Scaffolds that were detected to be allelic variants. Syntenies were examined between small scaffolds and chromosomes. Only those with multiple syntenic regions throughout the entire scaffold with part of the chromosome were considered as allelic variants.
(XLSX)

S9 Table. Synteny blocks identified between Brazil A4 and Y C6.
(XLSX)

S10 Table. Heterozygous SNPs/Indels identified in Brazil A4 and Y C6.
(XLSX)

S11 Table. Homozygous SNPs/Indels identified between Brazil A4 and Y C6.
(XLSX)

S12 Table. Orthologous groups in *T. cruzi*, *T. brucei* and *Leishmania* species with total gene count > 6. All sequences were retrieved from TriTrypDB database (https://tritrypdb.org/tritrypdb/) release-44.
(XLSX)

S13 Table. List of unique genes in the respective strains. Genes are unique under the condition that no orthologs or only orthologs with either < 50% identity or < 50% coverage were discovered in the other genome. Genes derived from amplification with high identity are considered as one unique gene.
(XLSX)

S14 Table. BLASTP result of the best match analysis in 6 large gene families between the two strains.
(XLSX)

S15 Table. Prominent tandem arrays of large gene families identified in Brazil A4 and Y C6 (> = 3 genes in the array, < = 5kb between two tandem genes).
(XLSX)

Acknowledgments

We thank Dr. Todd Minning for initial contributions to this project, Dr. Peng Qi from the University of Georgia for generous technical support, Dr. Robert Sabatini from the University of Georgia for insightful discussions, and Dr. Benedikt Brink from Ludwig-Maximilians-Universität for advice and for testing our data in his genome phasing pipeline.

Author Contributions

Conceptualization: Wei Wang, Rick L. Tarleton.

Data curation: Wei Wang, Duo Peng, Rodrigo P. Baptista, Yiran Li.

Formal analysis: Wei Wang, Duo Peng, Rodrigo P. Baptista, Yiran Li, Rick L. Tarleton.

Funding acquisition: Rick L. Tarleton.

Methodology: Wei Wang, Duo Peng, Rodrigo P. Baptista, Rick L. Tarleton.

Project administration: Rick L. Tarleton.

Software: Duo Peng, Rodrigo P. Baptista, Yiran Li.

Supervision: Jessica C. Kissinger.

Validation: Wei Wang.

Visualization: Wei Wang.

Writing – original draft: Wei Wang, Duo Peng, Rick L. Tarleton.

Writing – review & editing: Jessica C. Kissinger, Rick L. Tarleton.

References

1. De Pablos LM, Osuna A. Multigene families in *Trypanosoma cruzi* and their role in infectivity. Infect Immun. 2012; 80(7):2258–64. Epub 2012/03/21. https://doi.org/10.1128/IAI.06225-11 PMID: 22431647; PubMed Central PMCID: PMC3416482.
2. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science. 2005; 309(5733):409–15. Epub 2005/07/16. https://doi.org/10.1126/science.1112631 PMID: 16020725.
3. Weston D, Patel B, Van Voorhis WC. Virulence in *Trypanosoma cruzi* infection correlates with the expression of a distinct family of sialidase superfamily genes. Mol Biochem Parasitol. 1999; 98 (1):105–16. Epub 1999/02/24. https://doi.org/10.1016/s0166-6851(98)00152-2 PMID: 10029313.
4. Weatherly DB, Boehlke C, Tarleton RL. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. BMC Genomics. 2009; 10:255. Epub 2009/06/03. https://doi.org/10.1186/1471-2164-10-255 PMID: 19486522; PubMed Central PMCID: PMC2698008.
5. Berna L, Rodriguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G, et al. Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*. Microb Genom. 2018; 4(5). Epub 2018/05/01. https://doi.org/10.1099/mgen.0.000177 PMID: 29708484; PubMed Central PMCID: PMC5994713.
6. Callejas-Hernandez F, Rastrojo A, Poveda C, Girones N, Fresno M. Genomic assemblies of newly sequenced *Trypanosoma cruzi* strains reveal new genomic expansion and greater complexity. Sci Rep. 2018; 8(1):14631. Epub 2018/10/04. https://doi.org/10.1038/s41598-018-32877-2 PMID: 30279473; PubMed Central PMCID: PMC6168536.
7. Carlos Talavera-López JLR-C, Messenger Louisa A., Lewis Michael D., Yeo Matthew, et al. Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. bioRxiv. 2018. https://doi.org/10.1101/283531.
8. Diaz-Viraque F, Pita S, Greif G, de Souza RCM, Iraola G, Robello C. Nanopore sequencing significantly improves genome assembly of the protozoan parasite *Trypanosoma cruzi*. Genome Biol Evol. 2019; 11:1952–7. Epub 2019/06/21. https://doi.org/10.1093/gbe/evz129 PMID: 31218350.
9. Kaplan N, Dekker J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. Nat Biotechnol. 2013; 31(12):1143–7. Epub 2013/11/26. https://doi.org/10.1038/nbt.2768 PMID: 24270850; PubMed Central PMCID: PMC3880131.
10. Korbel JO, Lee C. Genome assembly and haplotyping with Hi-C. Nat Biotechnol. 2013; 31(12):1099–101. Epub 2013/12/10. https://doi.org/10.1038/nbt.2764 PMID: 24316648.
11. Muller LSM, Cosentino RO, Forstner KU, Guizetti J, Wedel C, Kaplan N, et al. Genome organization and DNA accessibility control antigenic variation in trypanosomes. Nature. 2018; 563(7729):121–5. Epub 2018/10/20. https://doi.org/10.1038/s41586-018-0619-8 PMID: 30333624; PubMed Central PMCID: PMC6784898.
12. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. Genome Res. 2016; 26(3):342–50. Epub 2016/02/06. https://doi.org/10.1101/gr.193474.115 PMID: 26848124; PubMed Central PMCID: PMC4772016.
13. Denton RSK Robert D., Malcom Jacob W., Du Preez Louis, and Malone John H. The African Bullfrog (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. bioRxiv. 2018. https://doi.org/10.1101/329847.
14. Kalbfleisch ESR Theodore S., DePriest Michael S. Jr, Walenz Brian P., Hestand Matthew S., Vermeesch Joris R., O'Connell Brendan L., Fiddes Ian T., Vershinina Alisa O., Petersen Jessica L., Finno Carrie J., Bellone Rebecca R., McCue Molly E., Brooks Samantha A., Bailey Ernest, Orlando Ludovic,

Green Richard E., Miller Donald C., Antczak Douglas F, MacLeod James N. EquCab3, an Updated Reference Genome for the Domestic Horse. bioRxiv. 2018. https://doi.org/10.1101/306928.

15. Elbers JP, Rogers MF, Perelman PL, Proskuryakova AA, Serdyukova NA, Johnson WE, et al. Improving Illumina assemblies with Hi-C and long reads: An example with the North African dromedary. Mol Ecol Resour. 2019; 19(4):1015–26. Epub 2019/04/12. https://doi.org/10.1111/1755-0998.13020 PMID: 30972949; PubMed Central PMCID: PMC6618069.

16. Salter JF, Johnson O, Stafford NJ 3rd, Herrin WF Jr., Schilling D, Cedotal C, et al. A Highly Contiguous Reference Genome for Northern Bobwhite (*Colinus virginianus*). G3 (Bethesda). 2019; 9(12):3929–32. Epub 2019/10/16. https://doi.org/10.1534/g3.119.400609 PMID: 31611345; PubMed Central PMCID: PMC6893191.

17. Schreiber M, Mascher M, Wright J, Padmarasu S, Himmelbach A, Heavens D, et al. A Genome Assembly of the Barley 'Transformation Reference' Cultivar Golden Promise. G3 (Bethesda). 2020; 10(6):1823–7. Epub 2020/04/04. https://doi.org/10.1534/g3.119.401010 PMID: 32241919; PubMed Central PMCID: PMC7263683.

18. Westenberger SJ, Barnabe C, Campbell DA, Sturm NR. Two hybridization events define the population structure of *Trypanosoma cruzi*. Genetics. 2005; 171(2):527–43. Epub 2005/07/07. https://doi.org/10.1534/genetics.104.038745 PMID: 15998728; PubMed Central PMCID: PMC1456769.

19. Zingales B, Andrade SG, Briones MR, Campbell DA, Chiari E, Fernandes O, et al. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem Inst Oswaldo Cruz. 2009; 104(7):1051–4. Epub 2009/12/23. https://doi.org/10.1590/s0074-02762009000700021 PMID: 20027478.

20. Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MM, et al. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. Infect Genet Evol. 2012; 12(2):240–53. Epub 2012/01/10. https://doi.org/10.1016/j.meegid.2011.12.009 PMID: 22226704.

21. de Freitas JM, Augusto-Pinto L, Pimenta JR, Bastos-Rodrigues L, Goncalves VF, Teixeira SM, et al. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. PLoS Pathog. 2006; 2(3):e24. Epub 2006/04/13. https://doi.org/10.1371/journal.ppat.0020024 PMID: 16609729; PubMed Central PMCID: PMC1434789.

22. Flores-Lopez CA, Machado CA. Analyses of 32 loci clarify phylogenetic relationships among *Trypanosoma cruzi* lineages and support a single hybridization prior to human contact. PLoS Negl Trop Dis. 2011; 5(8):e1272. Epub 2011/08/11. https://doi.org/10.1371/journal.pntd.0001272 PMID: 21829751; PubMed Central PMCID: PMC3149036.

23. Tomasini N, Diosque P. Evolution of *Trypanosoma cruzi*: clarifying hybridisations, mitochondrial introgressions and phylogenetic relationships between major lineages. Mem Inst Oswaldo Cruz. 2015; 110(3):403–13. Epub 2015/03/26. https://doi.org/10.1590/0074-02760140401 PMID: 25807469; PubMed Central PMCID: PMC4489478.

24. Souza RT, Lima FM, Barros RM, Cortez DR, Santos MF, Cordero EM, et al. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. PLoS One. 2011; 6(8):e23042. Epub 2011/08/23. https://doi.org/10.1371/journal.pone.0023042 PMID: 21857989; PubMed Central PMCID: PMC3155523.

25. Henriksson J, Dujardin JC, Barnabe C, Brisse S, Timperman G, Venegas J, et al. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is evolutionarily informative. Parasitology. 2002; 124(Pt 3):277–86. Epub 2002/04/02. https://doi.org/10.1017/s0031182001001093 PMID: 11922429.

26. Vargas N, Pedroso A, Zingales B. Chromosomal polymorphism, gene synteny and genome size in *T. cruzi* I and *T. cruzi* II groups. Mol Biochem Parasitol. 2004; 138(1):131–41. Epub 2004/10/27. https://doi.org/10.1016/j.molbiopara.2004.08.005 PMID: 15500924.

27. Pedroso A, Cupolillo E, Zingales B. Evaluation of *Trypanosoma cruzi* hybrid stocks based on chromosomal size variation. Mol Biochem Parasitol. 2003; 129(1):79–90. Epub 2003/06/12. https://doi.org/10.1016/s0166-6851(03)00096-3 PMID: 12798509.

28. Lima FM, Souza RT, Santori FR, Santos MF, Cortez DR, Barros RM, et al. Interclonal variations in the molecular karyotype of *Trypanosoma cruzi*: chromosome rearrangements in a single cell-derived clone of the G strain. PLoS One. 2013; 8(5):e63738. Epub 2013/05/15. https://doi.org/10.1371/journal.pone.0063738 PMID: 23667668; PubMed Central PMCID: PMC3646811.

29. Triana O, Ortiz S, Dujardin JC, Solari A. *Trypanosoma cruzi*: variability of stocks from Colombia determined by molecular karyotype and minicircle Southern blot analysis. Exp Parasitol. 2006; 113(1):62–6. Epub 2006/01/04. https://doi.org/10.1016/j.exppara.2005.11.016 PMID: 16388803.

30. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31 (19):3210–2. Epub 2015/06/11. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717.

31. Franzen O, Ochaya S, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, et al. Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener. PLoS Negl Trop Dis. 2011; 5(3):e984. Epub 2011/03/17. https://doi.org/10.1371/journal.pntd.0000984 PMID: 21408126; PubMed Central PMCID: PMC3050914.

32. Buscaglia CA, Campo VA, Frasch AC, Di Noia JM. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. Nat Rev Microbiol. 2006; 4(3):229–36. Epub 2006/02/21. https://doi.org/10.1038/nrmicro1351 PMID: 16489349.

33. Martin DL, Weatherly DB, Laucella SA, Cabinian MA, Crim MT, Sullivan S, et al. CD8+ T-Cell responses to Trypanosoma cruzi are highly focused on strain-variant trans-sialidase epitopes. PLoS Pathog. 2006; 2(8):e77. Epub 2006/08/02. https://doi.org/10.1371/journal.ppat.0020077 PMID: 16879036; PubMed Central PMCID: PMC1526708.

34. Weatherly DB, Peng D, Tarleton RL. Recombination-driven generation of the largest pathogen repository of antigen variants in the protozoan *Trypanosoma cruzi*. BMC Genomics. 2016; 17(1):729. Epub 2016/09/14. https://doi.org/10.1186/s12864-016-3037-z PMID: 27619017; PubMed Central PMCID: PMC5020489.

35. Kronenberg RJH Zev N., Hiendleder Stefan, Smith Timothy P. L., Sullivan Shawn T., Williams John L., Kingan Sarah B. FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. bioRxiv. 2018. https://doi.org/10.1101/327064.

36. Henriksson J AL, Macina RA, Franke de Cazzulo BM, Cazzulo JJ, Frasch AC, Pettersson U. Chromosomal localization of seven cloned antigen genes provides evidence of diploidy and further demonstration of karyotype variability in *Trypanosoma cruzi*. Mol Biochem Parasitol. 1990; 42:213–23. https://doi.org/10.1016/0166-6851(90)90164-h PMID: 2270104

37. Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, et al. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*. Mol Biochem Parasitol. 1995; 73(1–2):63–74. Epub 1995/07/01. https://doi.org/10.1016/0166-6851(95)00096-j PMID: 8577348.

38. CaroleBranchea S, LenaÅslundb, BjörnAnderssona. Comparative karyotyping as a tool for genome structure analysis of *Trypanosoma cruzi*. Mol Biochem Parasitol. 2006; 147:30–8. https://doi.org/10.1016/j.molbiopara.2006.01.005 PMID: 16481054

39. Reis-Cunha JL, Rodrigues-Luiz GF, Valdivia HO, Baptista RP, Mendes TA, de Morais GL, et al. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains. BMC Genomics. 2015; 16:499. Epub 2015/07/05. https://doi.org/10.1186/s12864-015-1680-4 PMID: 26141959; PubMed Central PMCID: PMC4491234.

40. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015; 16:157. Epub 2015/08/06. https://doi.org/10.1186/s13059-015-0721-2 PMID: 26243257; PubMed Central PMCID: PMC4531804.

41. Cremona ML, Sanchez DO, Frasch AC, Campetella O. A single tyrosine differentiates active and inactive *Trypanosoma cruzi trans*-sialidases. Gene. 1995; 160(1):123–8. Epub 1995/07/04. https://doi.org/10.1016/0378-1119(95)00175-6 PMID: 7628705.

42. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015; 1(1):vev003. Epub 2015/05/26. https://doi.org/10.1093/ve/vev003 PMID: 27774277; PubMed Central PMCID: PMC5014473.

43. El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, et al. The sequence and analysis of *Trypanosoma brucei* chromosome II. Nucleic Acids Res. 2003; 31(16):4856–63. Epub 2003/08/09. https://doi.org/10.1093/nar/gkg673 PMID: 12907728; PubMed Central PMCID: PMC169936.

44. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. Science. 2005; 309(5733):416–22. Epub 2005/07/16. https://doi.org/10.1126/science.1112642 PMID: 16020726.

45. Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, et al. Telomeric expression sites are highly conserved in *Trypanosoma brucei*. PLoS One. 2008; 3(10):e3527. Epub 2008/10/28. https://doi.org/10.1371/journal.pone.0003527 PMID: 18953401; PubMed Central PMCID: PMC2567434.

46. Mugnier MR, Stebbins CE, Papavasiliou FN. Masters of Disguise: Antigenic Variation and the VSG Coat in *Trypanosoma brucei*. PLoS Pathog. 2016; 12(9):e1005784. Epub 2016/09/02. https://doi.org/10.1371/journal.ppat.1005784 PMID: 27583379; PubMed Central PMCID: PMC5008768.

47. Clayton C. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. Open Biol. 2019; 9(6):190072. Epub 2019/06/06. https://doi.org/10.1098/rsob.190072 PMID: 31164043; PubMed Central PMCID: PMC6597758.

48. Yoshida N. Molecular basis of mammalian cell invasion by *Trypanosoma cruzi*. An Acad Bras Cienc. 2006; 78(1):87–111. Epub 2006/03/15. https://doi.org/10.1590/s0001-37652006000100010 PMID: 16532210.

49. Nakayasu ES, Yashunsky DV, Nohara LL, Torrecilhas AC, Nikolaev AV, Almeida IC. GPIomics: global analysis of glycosylphosphatidylinositol-anchored molecules of *Trypanosoma cruzi*. Mol Syst Biol. 2009; 5:261. Epub 2009/04/10. https://doi.org/10.1038/msb.2009.13 PMID: 19357640; PubMed Central PMCID: PMC2683718.

50. Gonzalez MS, Souza MS, Garcia ES, Nogueira NF, Mello CB, Canepa GE, et al. *Trypanosoma cruzi* TcSMUG L-surface mucins promote development and infectivity in the triatomine vector *Rhodnius prolixus*. PLoS Negl Trop Dis. 2013; 7(11):e2552. Epub 2013/11/19. https://doi.org/10.1371/journal.pntd.0002552 PMID: 24244781; PubMed Central PMCID: PMC3828161.

51. Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. Genes Dev. 2009; 23 (9):1063–76. Epub 2009/04/17. https://doi.org/10.1101/gad.1790409 PMID: 19369410; PubMed Central PMCID: PMC2682952.

52. Reynolds D, Hofmeister BT, Cliffe L, Alabady M, Siegel TN, Schmitz RJ, et al. Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in *Trypanosoma brucei*. PLoS Genet. 2016; 12(1):e1005758. Epub 2016/01/23. https://doi.org/10.1371/journal.pgen.1005758 PMID: 26796527; PubMed Central PMCID: PMC4721609.

53. Cliffe LJ, Siegel TN, Marshall M, Cross GA, Sabatini R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of *Trypanosoma brucei*. Nucleic Acids Res. 2010; 38(12):3923–35. Epub 2010/03/11. https://doi.org/10.1093/nar/gkq146 PMID: 20215442; PubMed Central PMCID: PMC2896530.

54. Schulz D, Zaringhalam M, Papavasiliou FN, Kim HS. Base J and H3.V Regulate Transcriptional Termination in *Trypanosoma brucei*. PLoS Genet. 2016; 12(1):e1005762. Epub 2016/01/23. https://doi.org/10.1371/journal.pgen.1005762 PMID: 26796638; PubMed Central PMCID: PMC4721952.

55. Kawasaki F, Beraldi D, Hardisty RE, McInroy GR, van Delft P, Balasubramanian S. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite *Leishmania*. Genome Biol. 2017; 18 (1):23. Epub 2017/02/01. https://doi.org/10.1186/s13059-017-1150-1 PMID: 28137275; PubMed Central PMCID: PMC5282726.

56. van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. Cell. 2012; 150(5):909–21. Epub 2012/09/04. https://doi.org/10.1016/j.cell.2012.07.030 PMID: 22939620; PubMed Central PMCID: PMC3684241.

57. Reynolds D, Cliffe L, Forstner KU, Hon CC, Siegel TN, Sabatini R. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei*. Nucleic Acids Res. 2014; 42(15):9717–29. Epub 2014/08/12. https://doi.org/10.1093/nar/gku714 PMID: 25104019; PubMed Central PMCID: PMC4150806.

58. Revollo S, Oury B, Laurent JP, Barnabe C, Quesney V, Carriere V, et al. *Trypanosoma cruzi*: impact of clonal evolution of the parasite on its biological and medical properties. Exp Parasitol. 1998; 89 (1):30–9. Epub 1998/05/29. https://doi.org/10.1006/expr.1998.4216 PMID: 9603486.

59. Rassi A Jr., Rassi A, Marcondes de Rezende J. American trypanosomiasis (Chagas disease). Infect Dis Clin North Am. 2012; 26(2):275–91. Epub 2012/05/29. https://doi.org/10.1016/j.idc.2012.03.002 PMID: 22632639.

60. Nguyen T, Waseem M. Chagas Disease (American Trypanosomiasis). StatPearls. Treasure Island (FL)2020.

61. Ackermann AA, Carmona SJ, Aguero F. TcSNP: a database of genetic variation in *Trypanosoma cruzi*. Nucleic Acids Res. 2009; 37(Database issue):D544–9. Epub 2008/11/01. https://doi.org/10.1093/nar/gkn874 PMID: 18974180; PubMed Central PMCID: PMC2686512.

62. Previato J, Andrade AFB, Pessolani MCV, and Mendonça-Previato L. Incorporation of sialic acid into *Trypanosoma cruzi* macromolecules. A proposal for a new metabolic route. Mol Biochem Parasitol. 1985; 16:85–96. https://doi.org/10.1016/0166-6851(85)90051-9 PMID: 2412116

63. Uemura H, Schenkman S., Nussenzweig V., and Eichinger D. Only some members of a gene family in *Trypanosoma cruzi* encode proteins that express both *trans*-sialidase and neuraminidase activities. EMBO J. 1992; 11:3837–44. PMID: 1396577

64. Frasch AC. Functional diversity in the *trans*-sialidase and mucin families in *Trypanosoma cruzi*. Parasitol Today. 2000; 16(7):282–6. Epub 2000/06/20. https://doi.org/10.1016/s0169-4758(00)01698-7 PMID: 10858646.

65. Cross GA. Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of *Trypanosoma brucei*. Parasitology. 1975; 71(3):393–417. Epub 1975/12/01. https://doi.org/10.1017/s003118200004717x PMID: 645

66. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, et al. Switches in expression of Plasmodium falciparum *var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell. 1995; 82(1):101–10. Epub 1995/07/14. https://doi.org/10.1016/0092-8674(95)90056-x PMID: 7606775; PubMed Central PMCID: PMC3730239.

67. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell. 1995; 82(1):89–100. Epub 1995/07/14. https://doi.org/10.1016/0092-8674(95)90055-1 PMID: 7606788.

68. Mowatt MR, Aggarwal A, Nash TE. Carboxy-terminal sequence conservation among variant-specific surface proteins of *Giardia lamblia*. Mol Biochem Parasitol. 1991; 49(2):215–27. Epub 1991/12/01. https://doi.org/10.1016/0166-6851(91)90065-e PMID: 1775165.

69. Pimenta PF, da Silva PP, Nash T. Variant surface antigens of *Giardia lamblia* are associated with the presence of a thick cell coat: thin section and label fracture immunocytochemistry survey. Infect Immun. 1991; 59(11):3989–96. Epub 1991/11/01. https://doi.org/10.1128/IAI.59.11.3989-3996.1991 PMID: 1937758; PubMed Central PMCID: PMC258987.

70. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000; 290 (5494):1151–5. Epub 2000/11/10. https://doi.org/10.1126/science.290.5494.1151 PMID: 11073452.

71. Lewis EB. Pseudoallelism and gene evolution. Cold Spring Harb Symp Quant Biol. 1951; 16:159–74. Epub 1951/01/01. https://doi.org/10.1101/sqb.1951.016.01.014 PMID: 14942737.

72. Sturtevant AH. The Effects of Unequal Crossing over at the Bar Locus in Drosophila. Genetics. 1925; 10:117–47. PMID: 17246266

73. Muller HJ. Bar Duplication. Science. 1936; 83(2161):528–30. Epub 1936/05/29. https://doi.org/10.1126/science.83.2161.528-a PMID: 17806465.

74. Huckaby AC, Granum CS, Carey MA, Szlachta K, Al-Barghouthi B, Wang YH, et al. Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome. Nucleic Acids Res. 2019; 47(4):1615–27. Epub 2018/12/24. https://doi.org/10.1093/nar/gky1268 PMID: 30576466; PubMed Central PMCID: PMC6393310.

75. Ersfeld K, Melville SE, Gull K. Nuclear and genome organization of *Trypanosoma brucei*. Parasitol Today. 1999; 15(2):58–63. Epub 1999/05/11. https://doi.org/10.1016/s0169-4758(98)01378-7 PMID: 10234187.

76. Navarro M, Gull K. A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*. Nature. 2001; 414(6865):759–63. Epub 2001/12/14. https://doi.org/10.1038/414759a PMID: 11742402.

77. Barnes RL, Shi H, Kolev NG, Tschudi C, Ullu E. Comparative genomics reveals two novel RNAi factors in *Trypanosoma brucei* and provides insight into the core machinery. PLoS Pathog. 2012; 8(5): e1002678. Epub 2012/06/02. https://doi.org/10.1371/journal.ppat.1002678 PMID: 22654659; PubMed Central PMCID: PMC3359990.

78. DaRocha WD, Otsu K, Teixeira SM, Donelson JE. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. Mol Biochem Parasitol. 2004; 133(2):175–86. Epub 2003/12/31. https://doi.org/10.1016/j.molbiopara.2003.10.005 PMID: 14698430.

79. Kieft R, Zhang Y, Marand AP, Moran JD, Bridger R, Wells L, et al. Identification of a novel base J binding protein complex involved in RNA polymerase II transcription termination in trypanosomes. PLoS Genet. 2020; 16(2):e1008390. Epub 2020/02/23. https://doi.org/10.1371/journal.pgen.1008390 PMID: 32084124; PubMed Central PMCID: PMC7055916.

80. Ekanayake D, Sabatini R. Epigenetic regulation of polymerase II transcription initiation in Trypanosoma cruzi: modulation of nucleosome abundance, histone modification, and polymerase occupancy by O-linked thymine DNA glucosylation. Eukaryot Cell. 2011; 10(11):1465–72. Epub 2011/09/20. https://doi.org/10.1128/EC.05185-11 PMID: 21926332; PubMed Central PMCID: PMC3209055.

81. Xu D, Brandan CP, Basombrio MA, Tarleton RL. Evaluation of high efficiency gene knockout strategies for *Trypanosoma cruzi*. BMC Microbiol. 2009; 9:90. Epub 2009/05/13. https://doi.org/10.1186/1471-2180-9-90 PMID: 19432966; PubMed Central PMCID: PMC2688506.

82. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science.

2009; 326(5950):289–93. Epub 2009/10/10. https://doi.org/10.1126/science.1181369 PMID: 19815776; PubMed Central PMCID: PMC2858594.

83. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012; 7(11):e47768. Epub 2012/11/28. https://doi.org/10.1371/journal.pone.0047768 PMID: 23185243; PubMed Central PMCID: PMC3504050.

84. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9 (11):e112963. Epub 2014/11/20. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509; PubMed Central PMCID: PMC4237348.

85. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010; 26(14):1704–7. Epub 2010/06/22. https://doi.org/10.1093/bioinformatics/btq269 PMID: 20562415; PubMed Central PMCID: PMC2894513.

86. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013; 14(8):R93. Epub 2013/09/05. https://doi.org/10.1186/gb-2013-14-8-r93 PMID: 24000942; PubMed Central PMCID: PMC4053811.

87. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, et al. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res. 2016; 44(W1):W29–34. Epub 2016/04/24. https://doi.org/10.1093/nar/gkw292 PMID: 27105845; PubMed Central PMCID: PMC4987884.

88. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004; 32(Web Server issue):W309–12. Epub 2004/06/25. https://doi.org/10.1093/nar/gkh379 PMID: 15215400; PubMed Central PMCID: PMC441517.

89. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005; 33(Web Server issue):W465–7. Epub 2005/06/28. https://doi.org/10.1093/nar/gki458 PMID: 15980513; PubMed Central PMCID: PMC1160219.

90. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005; 6:31. Epub 2005/02/17. https://doi.org/10.1186/1471-2105-6-31 PMID: 15713233; PubMed Central PMCID: PMC553969.

91. Queiroz RM, Charneau S, Motta FN, Santana JM, Roepstorff P, Ricart CA. Comprehensive proteomic analysis of *Trypanosoma cruzi* epimastigote cell surface proteins by two complementary methods. J Proteome Res. 2013; 12(7):3255–63. Epub 2013/05/21. https://doi.org/10.1021/pr400110h PMID: 23682730.

92. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30(9):1236–40. Epub 2014/01/24. https://doi.org/10.1093/bioinformatics/btu031 PMID: 24451626; PubMed Central PMCID: PMC3998142.

93. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012; 40(7):e49. Epub 2012/01/06. https://doi.org/10.1093/nar/gkr1293 PMID: 22217600; PubMed Central PMCID: PMC3326336.

94. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. Epub 2010/07/21. https://doi.org/10.1101/gr.107524.110 PMID: 20644199; PubMed Central PMCID: PMC2928508.

95. Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 2018; 19(1):125. Epub 2018/08/26. https://doi.org/10.1186/s13059-018-1486-1 PMID: 30143029; PubMed Central PMCID: PMC6109259.

96. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7:539. Epub 2011/10/13. https://doi.org/10.1038/msb.2011.75 PMID: 21988835; PubMed Central PMCID: PMC3261699.

97. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–7. Epub 2004/03/23. https://doi.org/10.1093/nar/gkh340 PMID: 15034147; PubMed Central PMCID: PMC390337.

98. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012; 61 (3):539–42. Epub 2012/02/24. https://doi.org/10.1093/sysbio/sys029 PMID: 22357727; PubMed Central PMCID: PMC3329765.

99. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 2005; 21(16):3422–3. Epub 2005/06/25. https://doi.org/10.1093/bioinformatics/bti553 PMID: 15976072.

100. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010; 38(Database issue): D457–62. Epub 2009/10/22. https://doi.org/10.1093/nar/gkp851 PMID: 19843604; PubMed Central PMCID: PMC2808979.