Title: 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] and Rick L. Tarleton[1,3,#]

[1] Center for Tropical and Emerging Global Diseases; [2] Department of Cellular Biology; [3] Institute of Bioinformatics; [4] Department of Genetics, University of Georgia, Athens, GA, USA.

#Corresponding author: tarleton@uga.edu

Running title: Genome evolution in *T. cruzi*

Keywords: Chagas disease, *Trypanosoma cruzi*, immune evasion, gene amplification and diversification, antigenic variation, strain-specific evolution.

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 gene family members 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. These processes also impact genes not required for or involved in immune evasion, creating unique challenges with respect to preserving core genome function while maximizing genetic diversity.

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 (Weston et al. 1999; El-Sayed et al. 2005; De Pablos and Osuna 2012). 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 (El-Sayed et al. 2005). 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 (Weatherly et al. 2009). 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 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 yet not 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 (Berna et al. 2018; Callejas-Hernandez et al. 2018) as well as nanopore sequencing (Diaz-Viraque et al. 2019) has resulted in much-improved contiguity and expansion of gene family members 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 (Kaplan and Dekker 2013; Korbel and Lee 2013). A second approach termed Chicago, adapts this same methodology but reconstitutes the confirmation of DNA *in vitro* by combining the DNA with purified histones and

chromatin assembly factors (Putnam et al. 2016). 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 (Putnam et al. 2016). The combination of Chicago and Hi-C has now been applied to many genomes (Robert D. Denton 2018; Theodore S. Kalbfleisch 2018; Elbers et al. 2019; Salter et al. 2019; Schreiber et al. 2020).

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 (Westenberger et al. 2005; de Freitas et al. 2006; Zingales et al. 2009; Flores-Lopez and Machado 2011; Zingales et al. 2012; Tomasini and Diosque 2015). 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. 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 gene families, however, extreme diversification is evident with no genes of the identical sequence within strains or shared between these strains. 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* (Souza et al. 2011) (Table 1). Application of the *in vitro* proximity-ligation tools Hi-C and Chicago (Putnam et al. 2016), decreased the L50 to half of that of the draft genomes, and the size of the largest scaffolds doubled (Table 1). Filling gaps 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 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 (Supplemental Table S1 and Supplemental Fig. S1), but also in filled gaps, recovered genes and extended repetitive regions (see examples in Supplemental Fig. S2).

Genome assembly	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)	45.11	677	51.50	227,072	48	1,236,815	0
Scaffolded	Chicago (125x) and Hi-C (46,451x)	45.16	402	51.50	907,746	18	2,710,165	295
Final	PBJelly, Pilon and iCorn	45.56	402	51.53	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	231
Final	PBJelly, Pilon and iCorn	47.22	266	51.58	889,019	18	2,951,016	231

Table 1. Summary of assembly statistics.

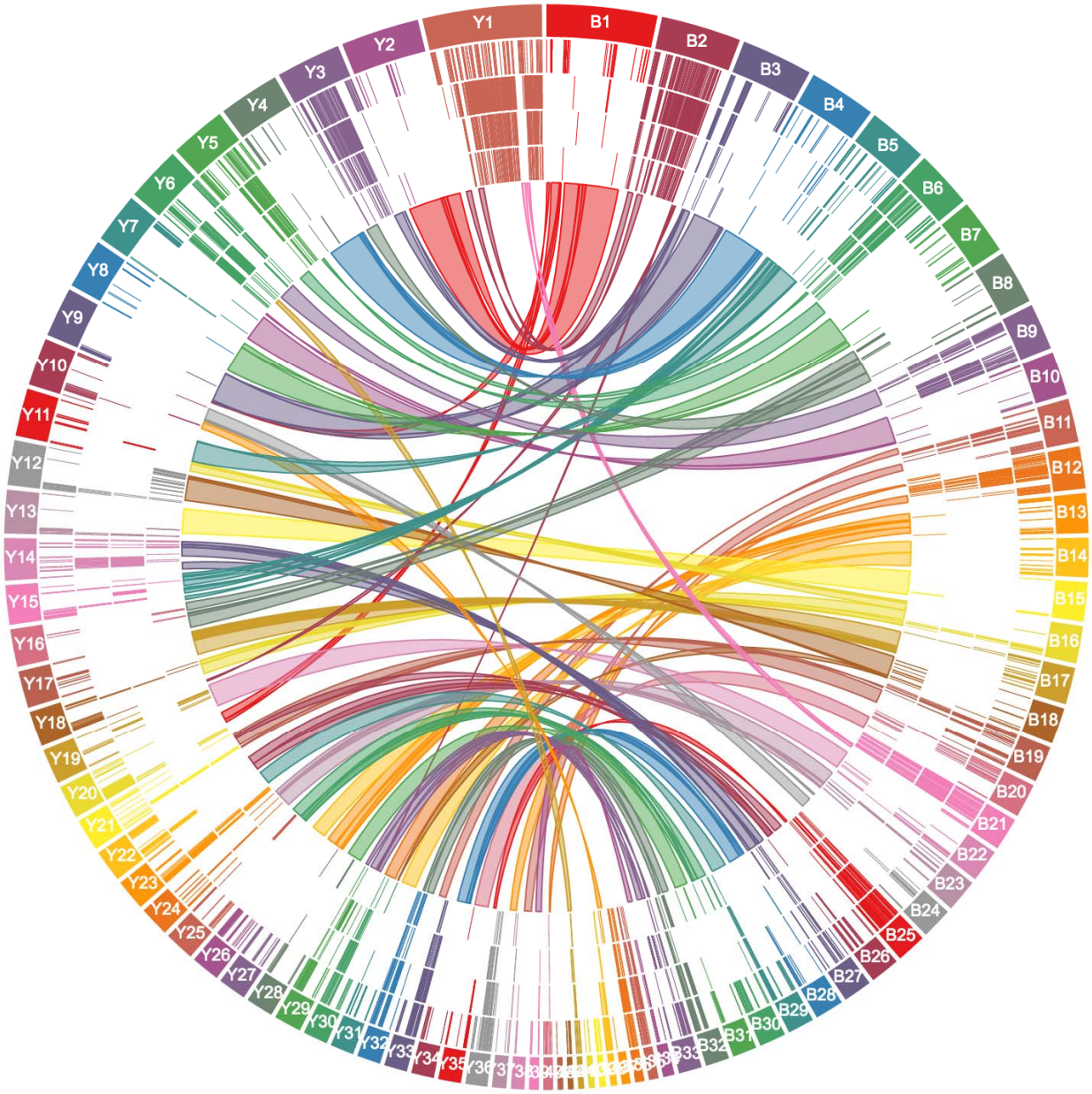
Genome features and content

Due to a lack of apparent chromosome condensation during replication (Henriksson et al. 2002; Souza et al. 2011), 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-filed gel electrophoresis (PFGE) analysis (Henriksson et al. 2002; Pedroso et al. 2003; Vargas et al. 2004; Triana et al. 2006; Lima et al. 2013). 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 (Supplemental Fig. S3) 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 (Supplemental Tables S2 and S3), 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 (Supplemental Fig. S3). Using conventional approaches with manual curation, gene models were identified in Brazil A4 and Y C6, respectively. Based on BUSCO assessment, the Brazil A4 and Y C6 contain the highest number of single-copy gene sets among assembled *T. cruzi* genomes (Supplemental Table S4).

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) (El-Sayed et al. 2005; Buscaglia et al. 2006; Martin et al. 2006). Our previous studies indicated the total copy number of TS genes was underestimated using conventional annotation

137 approaches due in part to the failure to identify new variants and fragments of TS resulting from
138 frequent recombination (Weatherly et al. 2016). To complete the annotation of the members of large
139 gene families, we developed a customized workflow (summarized in Supplemental Fig. S4) and applied it
140 to the six largest gene families. This allowed us to capture the full repertoire of gene family members
141 (copy numbers of which are summarized in Supplemental Table S5), the distribution of which were
142 plotted in Fig. 1 and Supplemental Fig. S5. Gene family members are unequally distributed among and
143 along the chromosomes with several of the largest chromosomes (e.g. TcBrA4_Chr2 and TcYC6_Chr1)
144 composed nearly entirely of gene family members. In contrast to previous reports suggesting the
145 members of large gene families were mainly located in telomeric and subtelomeric regions (Carlos
146 Talavera-López 2018) (El-Sayed et al. 2005), gene family members are not restricted to particular regions
147 of chromosomes. Moreover, TS, MASP, mucin and GP63 have an overlapping distribution along the
148 chromosomes, while RHS and DGF-1 genes are more dispersed.
149



152

153 **Figure 1.** Distribution of large gene families and synteny between chromosomes in Brazil A4 (right, B) and Y C6,

154 (left, Y). Tracks from outer to inner rings: chromosomes, TS, MASP, mucin, GP63, and synteny blocks.

155

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 Supplemental Table S6). The composition of gene content between two genomes is very similar, with ~25% as members of large gene families, ~40% as hypothetical proteins, and >90% of the remaining genes are orthologs of those in the related kinetoplastids *T. brucei* and *Leishmania 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 large gene family members 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 (Supplemental Fig. S6).

Allelic variation

The significant number of small scaffolds and the relatively high gene model numbers in some of the small scaffolds prompted us to consider whether these small scaffolds might represent regions of 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 (Supplemental Table S7). 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 (Zev N. Kronenberg 2018), 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.

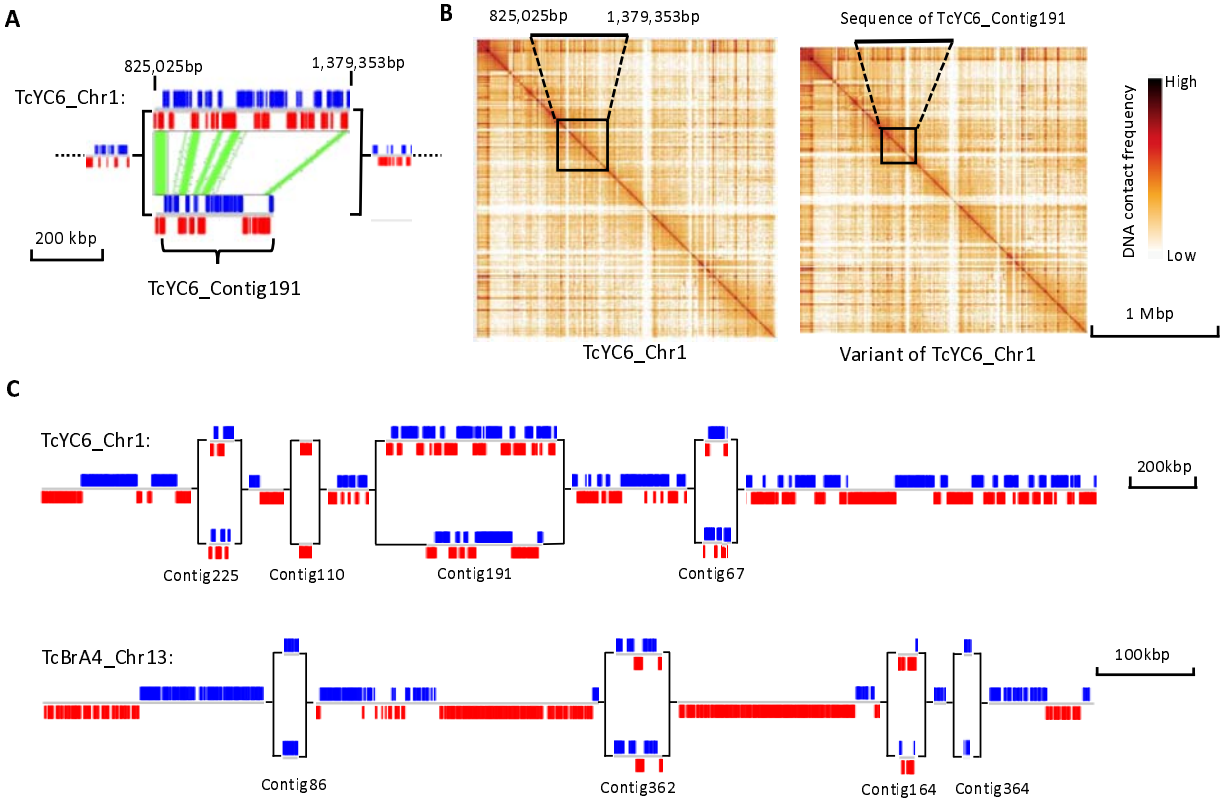


Figure 2. An example of homologous chromosomes with large allelic variations. (A) Synteny between two allelic variants in Chr1 of Y C6. Synteny blocked are marked with green. (B) Hi-C heat maps of TcYC6_Chr1 (left) and its homologous chromosome with TcYC6_Contig191 (boxed area) replacing the allelic region in TcYC6_Chr1 (boxed

193 area). (C) Two chromosomes with multiple allelic variants. Blue blocks indicate genes on the forward strand, and
194 red blocks indicate genes at the reverse strand.
195
196 **Structural comparison of the Brazil and Y sequences**
197 The very high genome quality and contiguity provided by the combination of SMRT sequencing and Hi-C
198 analysis enabled chromosome level comparison of the Brazil (TcI) and Y (TcII) clones (Fig. 1). The
199 synteny plots show that the majority of chromosomes from one genome collinear with those in the
200 other genome. For instance, Brazil A4 Chr4 showed continuous synteny to Y C6 Chr4 overall. However,
201 as expected based upon previous gene mapping studies (Henriksson J 1990) (Henriksson et al. 1995)
202 (Vargas et al. 2004) (CaroleBranchea 2006), some chromosomes corresponded to different regions in
203 multiple chromosomes of the other genome, e.g. Brazil A4 Chr1 showed synteny to a combination of
204 Chr20 (298,235 - 684,393bp), Chr9 (63,384 - 95,053bp) and Chr2 (20,327 - 1,438,658bp) in Y C6. Some
205 inverted syntenies were also detected, e.g. between 388,900 - 968,190bp on Brazil A4 Chr8 and 11,711 -
206 556,982bp on Y C6 Chr16 (Fig. 1). Notably, the diversity of sequences encoding members of the large
207 gene families (see details below) prevented the detection of synteny in a substantial proportion of the
208 two genomes, including in two of the largest chromosomes (e.g. TcBrA4_Chr2 and TcYC6_Chr1).
209
210 **Variation in gene models within and between Brazil and Y strains is predominantly in the large gene**
211 **families**
212 A large number of genetic variations were identified in the non-repetitive region, including heterozygous
213 SNPs/Indels within respective strains, and homozygous SNPs/Indels between the two strains
214 (Supplemental Tables S8, S9 and S10). We also detected aneuploidy in both genomes: 3 and 8
215 chromosomes in Brazil A4 and Y C6, respectively, exist in copy numbers greater than two, based on the
216 results of both relative read depth and allele frequency (Supplemental Fig. S7). Among these are the

partially syntenic chromosomes (TcBrA4_Chr24 and TcYC6_Chr10), which also share synteny with chromosome 31 in CL Brener, reported to be supernumerary in many strains (Reis-Cunha et al. 2015), thus suggesting a species-wide requirement for > 2 copies of one or more genes in these regions. Additionally, variation exist in the copy number for a substantial number of individual genes characterized by OrthoFinder, with ~150 genes showing the greatest variation between the two strains (Supplemental Table S11). However, with respect to genes unique to either strain, we found 23 (Brazil A4) and 20 (Y C6) gene loci not present in the other strain and further validated this finding by examining the raw reads (Supplemental Table S12). All are annotated as hypothetical proteins and most are small genes located in 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 large gene family members between the two strains, we carried out a best match search for the protein sequence of putatively expressed genes in each large gene family from Y C6 genome with those in Brazil A4. As a control, the same analysis was performed for a subset of 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 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 (Cremona et al. 1995), suggesting that this group of *trans*-sialidases has been selected for and conserved in both strains.

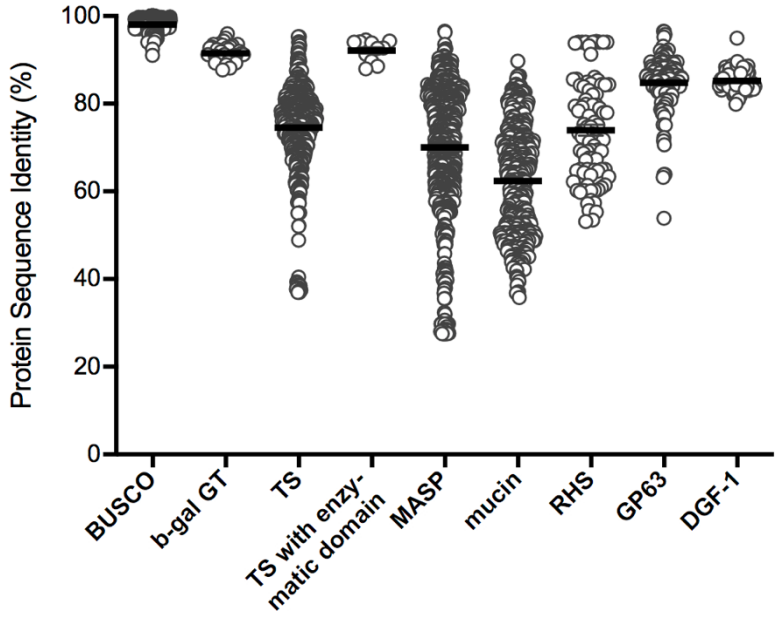


Figure 3. Protein best match analysis of gene families between Brazil A4 and Y C6.

Evidence of gene family expansion and diversification

The very high number and the impressive within- and between- strain variation in the genes composing the large 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 large 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 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 gene families (except DGF-1) and occur uniquely in each strain (Supplemental Table S14). 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 gene families in both *T. cruzi* genomes contained 10 or fewer replicates (Supplemental Table S14). However, one hypothetical protein (*HP) in the Y C6 occurr 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 share 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 suggest that this might be a recent amplification event. However, comparison to syntenic regions in other long-read sequenced *T. cruzi* genomes 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 three strains (Fig. 4E). Using the model of DTU evolution in *T. cruzi* which postulates that the TcVI is derived from a hybridization event between TcII and TcIII (Westenberger et al. 2005; Tomasini and Diosque 2015), 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.

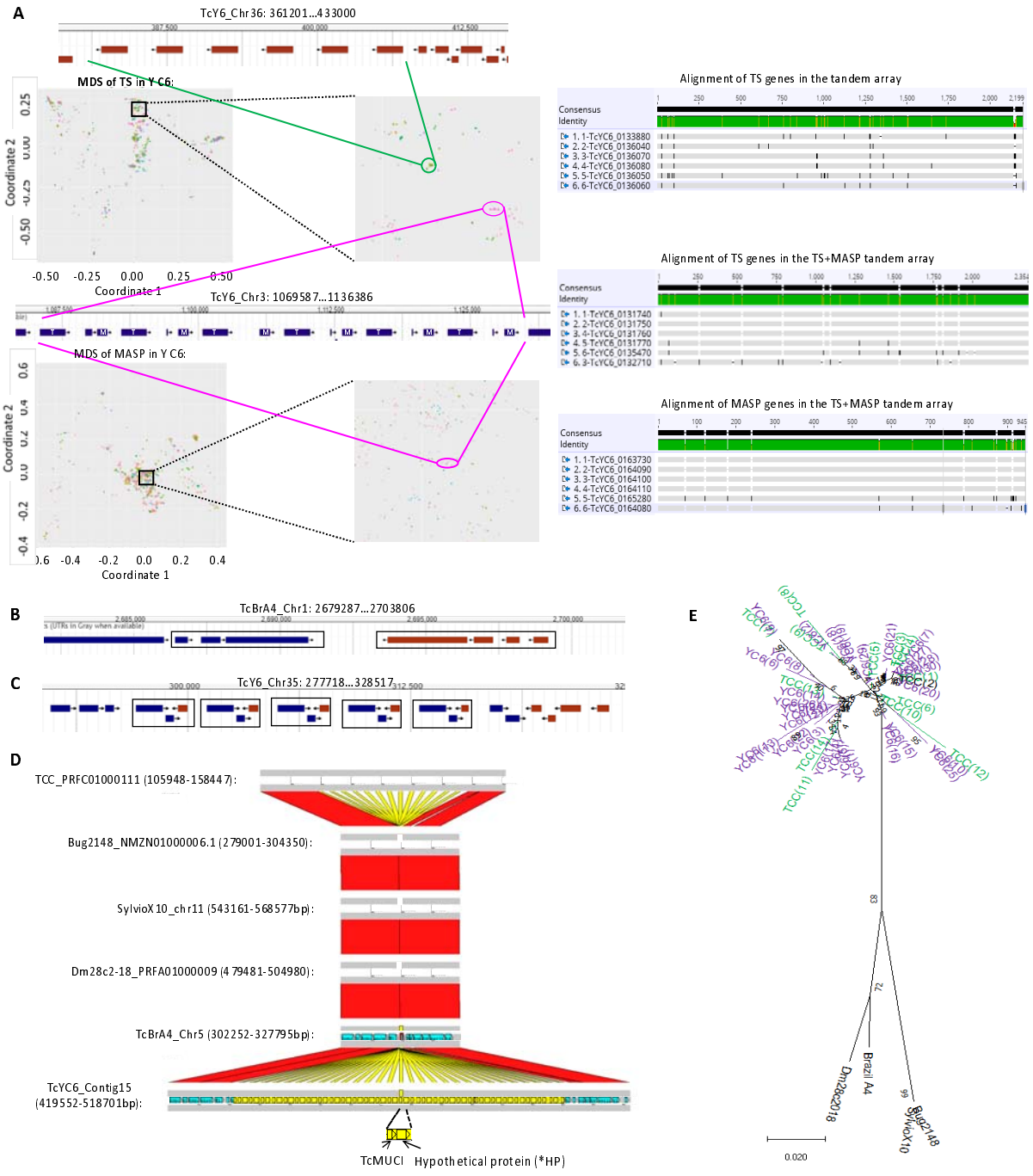


Figure 4. Gene amplification events in members of large gene families. (A) Tandem arrays of individual TS genes (top), and a TS+MASP pair (bottom) clustered based upon genetic distance in the MDS plots. Each chromosome is displayed as a separate pattern on the MDS plot. T: TS; M: MASP. Alignment of the 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 pseudo RHS genes. (C) One RHS (+), one hypothetical protein (+) and

one fragmented glycosyltransferase (-) replicated 5 times, creating multiple strand switches. (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) Phylogenetic tree 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. Live MDS plots can be explored at http://shiny.ctegd.uga.edu.

In addition to tandem clusters of gene family members, MDS analysis also revealed closely related gene family members located on multiple chromosomes (Fig. 5A). An 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 (Supplemental Fig. S8A). Interestingly, the 13 TS on Chr7 (Fig. 5B, middle) are in tandem, interspersed with a beta tubulin gene, while the 7 TS 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 (Supplemental Fig. S8B). 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 (Supplemental Fig. S8C), 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 cassette on Chr18, while the single TS genes on other chromosomes may derive from TS on Chr7.

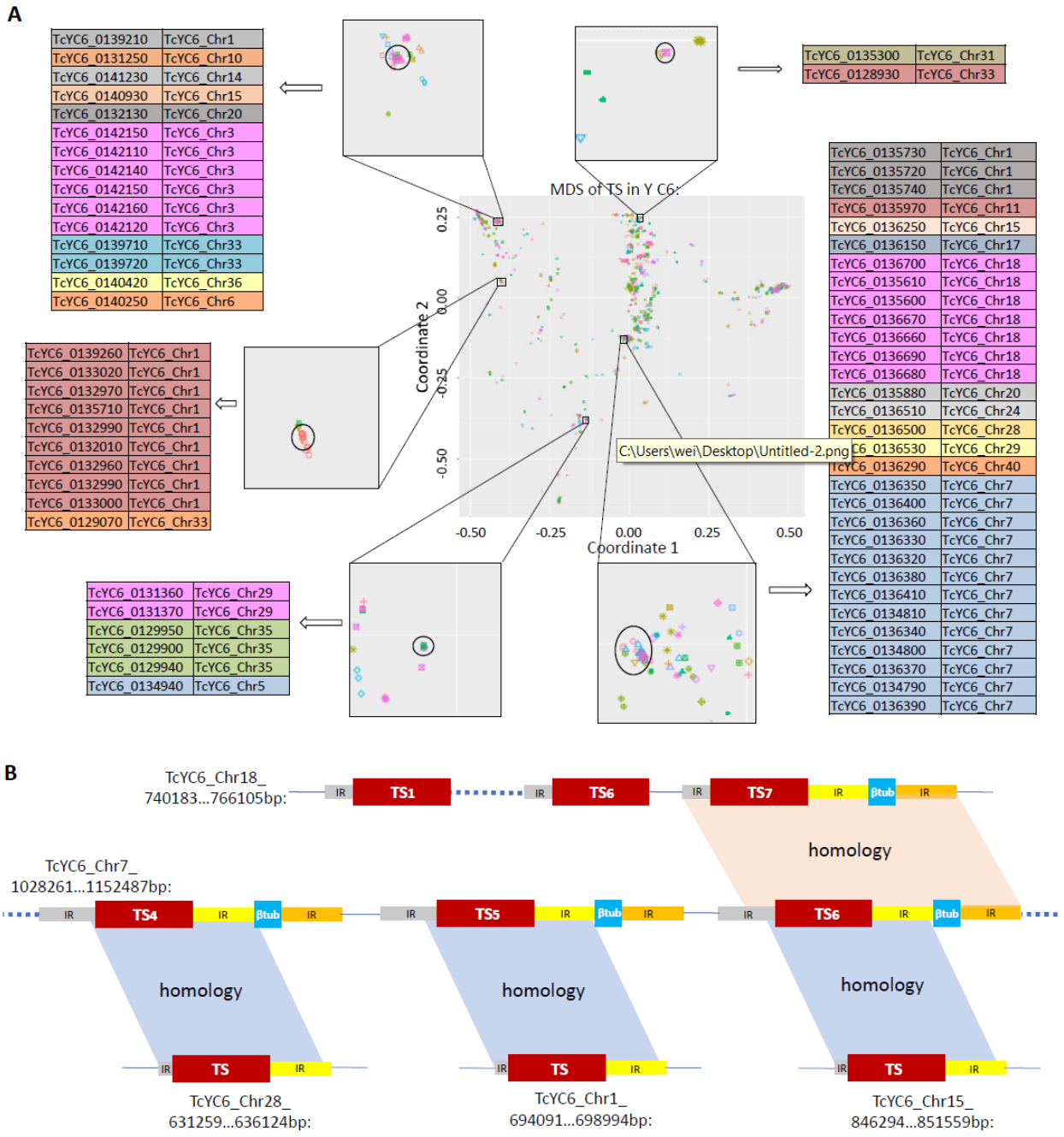


Figure 5. Examples of relocations of TS genes in Y C6. (A) Tight clusters of TS genes from MDS plot are distributed on different chromosomes. (B) Diagram of relocations in one of the TS clusters on the bottom right in (A). Blocks in the same color indicate genes or flanking sequences in high identity.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

We next used a pipeline previously designed to identify recombination events within TS genes in the CL Brener genome (Weatherly et al. 2016), to quantify recombination for 4 of the large gene families in the Brazil and Y strains (Table 2). 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.

	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 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

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

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 (Weatherly et al.

2016). 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 (Supplemental Fig. S9A), but are located on different chromosomes and display a considerable degree of variation (SNPs and Indels; Supplemental Fig. S9B). 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 (Supplemental Fig. S9C). 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 individual gene family members 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 B), 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.

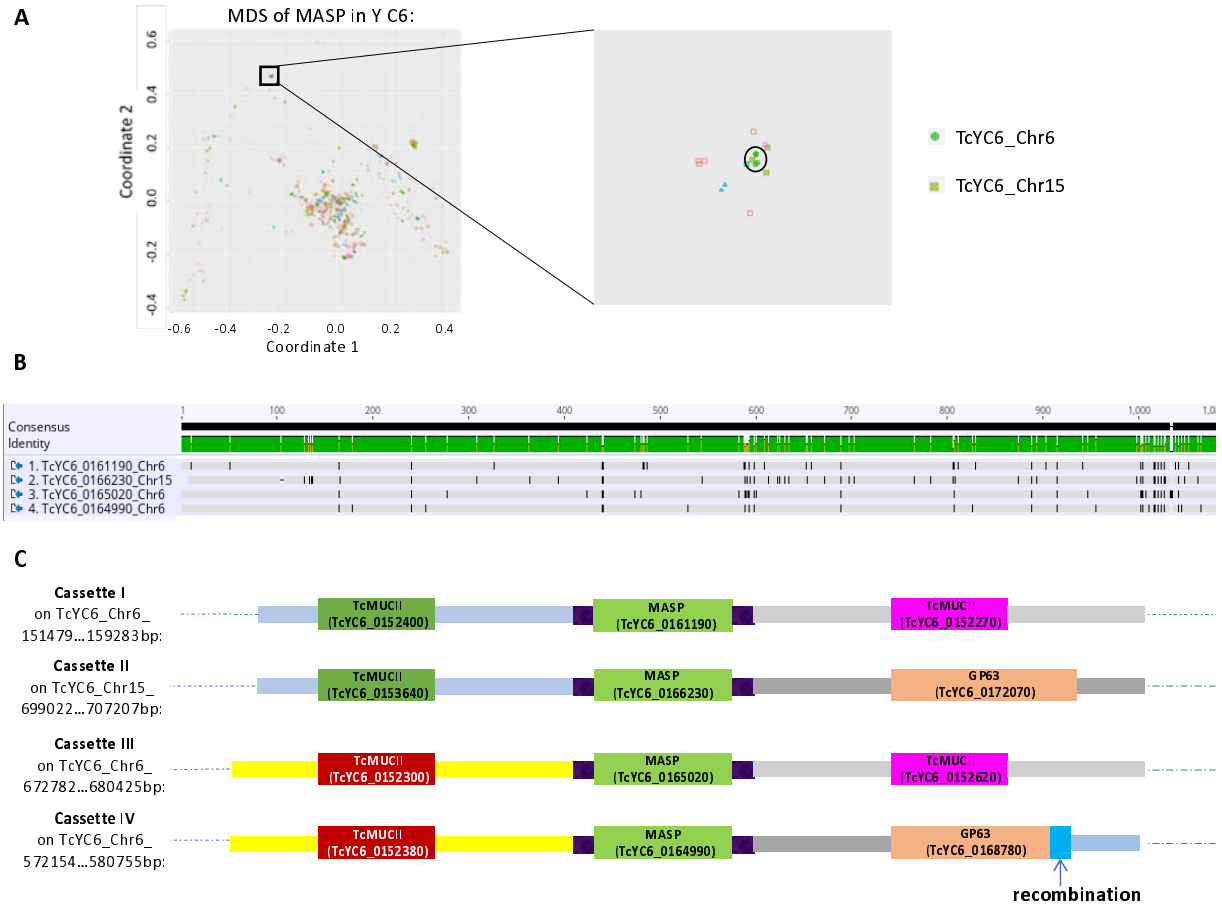


Figure 6. The combination of gene amlification, relocation, recombination and *in situ* diversification of large gene family members. (A) A tight cluster of 4 MASP genes from MDS plot are distributed on two chromosomes. (B) Alignment of the 4 MASP genes shows high identity with modest diversification of SNPs/Indels. (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.

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* (El-Sayed et al. 2003; Berriman et al. 2005; Hertz-Fowler et al. 2008; Mugnier et al. 2016; Muller et al. 2018)] but

362 instead are spread throughout the genome (Fig. 1). This presents the complication that the
363 amplification and dispersion events common in the large gene families of *T. cruzi* might also impact non-
364 gene family (core) genes as well. To investigate this possibility, we focused on core genes for which
365 there were > 6 total paralogues for the two genomes, and organized these paralogues on the basis of
366 gene location (Supplemental Table S11). By doing this, we could identify tandemly distributed genes that
367 likely resulted from gene amplification. For the over 150 groups of genes in this analysis, many showed
368 dramatic differences in gene copies in the two *T. cruzi* genomes with 26 instances of double-digit gene
369 copies in one strain compared to only 1-3 copies in the other. This same high level of variation was also
370 evident for other *T. cruzi* genomes sequenced using long-read sequencing methods but not in similarly
371 sequenced *T. brucei* and *Leishmania* isolates (Supplemental Fig. S10). Additionally, dispersion patterns
372 for these amplified genes differed widely between the Y C6 and Brazil A4 genomes (Supplemental Table
373 S11). Thus, the mechanisms that provide for the generation and maintenance of diversity in the large
374 gene families also appear to allow for substantial variation in copy number for selected core genes, and
375 representing a second major contributor to between-strain genetic variation in *T. cruzi* strains.
376
377 Most gene expression in trypanosomatids initiates in the absence of specific promoters and with the
378 production of multi-gene mRNA transcripts that are then processed into single-gene mature mRNAs.
379 These polycistronic transcriptional units (PTUs) of genes can be well over >100 kb in length and are
380 marked by start and stop signals, including base modifications (Clayton 2019). The apparent wide degree
381 of freedom for amplification and dispersion both within and outside the *T. cruzi* large gene families, and
382 particularly events that create tandem strand switches as shown in Fig 4C, would be expected to impact
383 this normal multi-gene PTU structure. Indeed, the average PTU length was ~116.5 and 126.8 kb in the
384 core gene-rich regions of the Brazil A4 and Y C6, respectively, similar to that in *T. brucei* (148.3 kb).
385 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 (Yoshida 2006; Nakayasu et al. 2009; Gonzalez et al. 2013) are within large PTUs containing almost no gene family members (Supplemental Fig. S11A) while many other tandem arrays or apparently diverging genes reside in gene-family-rich, short PTUs (Supplemental Fig. S11B). 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 (Siegel et al. 2009; Cliffe et al. 2010; van Luenen et al. 2012; Reynolds et al. 2014; Reynolds et al. 2016; Schulz et al. 2016; Kawasaki et al. 2017; Muller et al. 2018). 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 B), 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 regulate antisense RNA production so tightly.

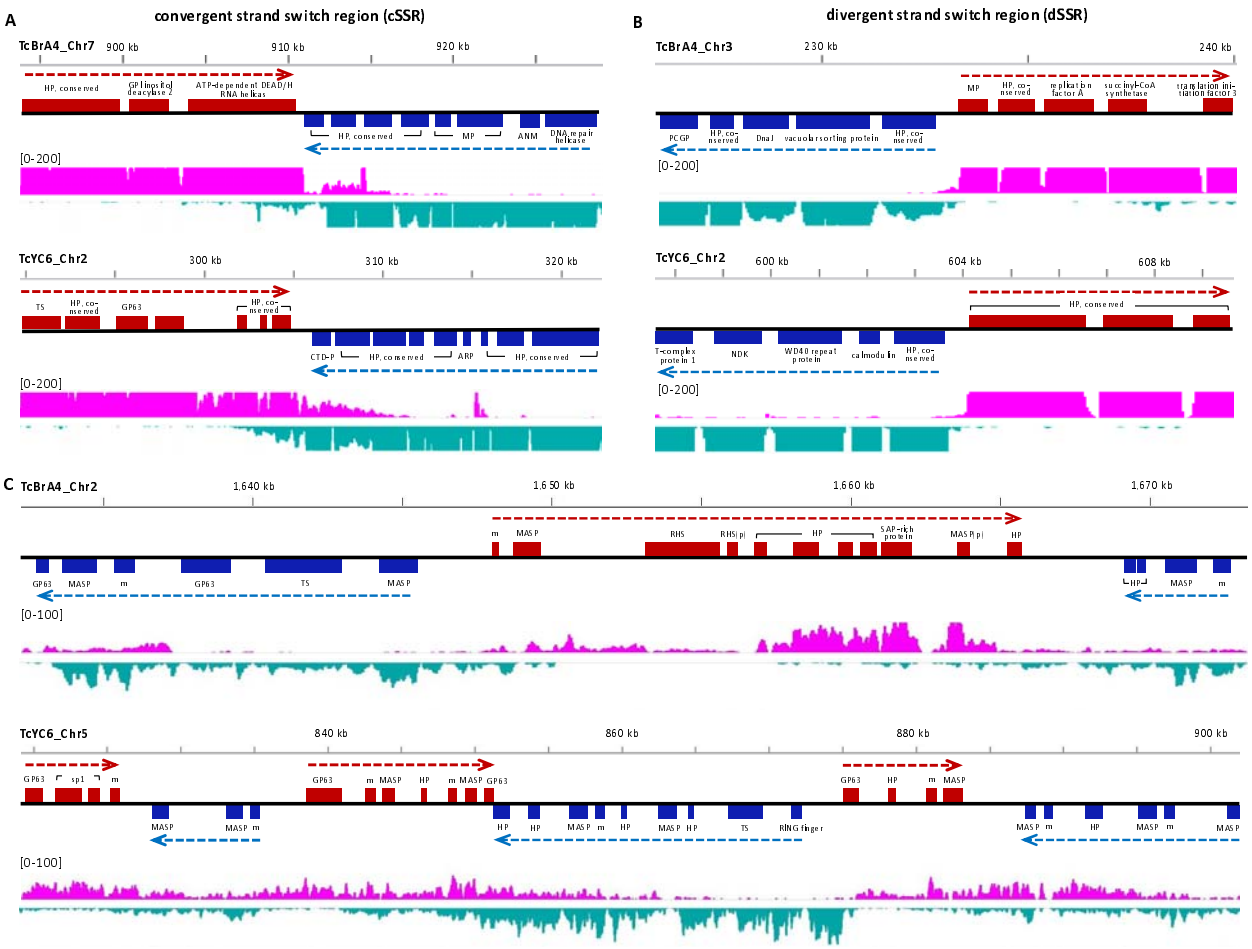


Figure 7. Antisense RNA levels in *T. cruzi* in relation to PTU structure, including both convergent strand switch regions (cSSR, A) and divergent strand switch regions (dSSR, B). (C) Gene-family-enriched regions with frequent strand switches where antisense RNA were detected in higher levels. HP, hypothetical 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.

Discussion

T. cruzi is a highly heterogeneous specie, 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 disease-causing potential in *T. cruzi* has been challenging due to the lack of high-quality reference genome. 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 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 (Revollo et al. 1998; Rassi et al. 2012; Nguyen and Waseem 2020). 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 (Tomasini and Diosque 2015). 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 (Ackermann et al. 2009) (Reis-Cunha et al. 2015), a considerable degree of variation exists in the form of SNPs/Indels and additionally, a substantial number of strain-specific copy number differences 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 family members make up ~25% of the genome and are spread widely throughout the genome, with some members on every chromosome and some of the largest chromosomes being almost entirely composed of gene family members. 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 gene family members 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 the beginning of 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 enzyme-active *trans*-sialidases required for acquisition of sialic acid by *T. cruzi* trypomastigotes (Previato 1985) (Uemura 1992; Cremona et al. 1995) (Frasch 2000). 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 this respect, 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

463 glycoproteins, VSGs), *Plasmodium* (*var* genes) and Giardia (Variant-specific Surface Protein, VSPs) (Cross
464 1975; Mowatt et al. 1991; Pimenta et al. 1991; Smith et al. 1995; Su et al. 1995). However in contrast to
465 the "one-at-a-time" models of classical antigenic variation best characterized in the sister kinetoplastid
466 *Trypanosoma brucei* (Cross 1975), *T. cruzi* expresses many gene family variants simultaneously. This
467 difference in strategy may relate to the fact that *T. cruzi* lives predominantly intracellularly in mammals
468 and must effectively evade cell-mediated (rather than exclusively antibody-mediated) immunity. But
469 expressing many antigen variants at one time also likely requires a larger antigen repertoir and/or an
470 enhanced ability to generate new variants. In African trypanosomes, a comparison of the genome
471 sequences of 2 different **subspecies** (*T. brucei brucei* vs *T. brucei gambienese*) revealed that >86% of the
472 genes – including most VSGs – varied by <1% between the subspecies and only 69 ortholog pairs
473 (including 35 VSG gene pairs) had less than 95% nucleotide identify (Jackson et al. 2010). The diversity
474 of antigen variants in *T. cruzi* among the 2 **strains** examined in this study is vastly greater, with the
475 average similarity of orthologues pairs ranging from 62.4% to 84.8% in the 5 largest gene families (Fig.
476 2). This finding suggests high pressure to generate variants and a genetic system that accommodate the
477 genomic flexibility that such generation would require.
478
479 Classically, segmental duplication creates the source material on which mutational and recombinational
480 events act to derive new genes and new gene functions (Lynch and Conery 2000). The presence of
481 segmental duplications (one gene or multiple genes as a unit) also encourages additional rounds of
482 duplications that can rapidly change gene content (Sturtevant 1925; Muller 1936; Lewis 1951). These
483 processes of gene duplication, recombination and mutation-driven diversification, functioning in concert
484 to ensure high and constant antigenic diversity, is strongly evident in the large gene families of *T. cruzi*.
485 Although we are able to track a significant number of these events, all occurring independently in these
486 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 retrotranposons of the *T. cruzi* genome facilitates maintenance of these processes and the dispersion of gene family members throughout the genome, and interestingly not restricted to chromosomes ends as is the case in *T. brucei* (El-Sayed et al. 2003) (Berriman et al. 2005) (Hertz-Fowler et al. 2008) (Mugnier et al. 2016) (Muller et al. 2018). 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* (Huckaby et al. 2019) were evident.

The apparent high frequency and continued evolution of gene families in *T. cruzi* also create structures and products unique among the kinetoplastids, including the lack of segregation of gene families to chromosome ends, absence of partitioning of expression sites [as in *T. brucei* VSGs (Ersfeld et al. 1999) Navarro and Gull 2001; (Navarro and Gull 2001; Hertz-Fowler et al. 2008)], tolerance for the generation of short PTUs and frequent strand switching, and most surprisingly, the tolerance of antisense RNA production. The latter may well explain the absence of the machinery for RNAi in *T. cruzi* (DaRocha et al. 2004) (Barnes et al. 2012). 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* (Reynolds et al. 2014) (Kieft et al. 2020).

Interestingly, there are several subsets of gene family members 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

(Yoshida 2006; Nakayasu et al. 2009; Gonzalez et al. 2013), 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 (Weatherly et al. 2016). The ~2X greater number of recombination events in all 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, strong selective pressure.

The apparent high levels of gene amplification/diversification 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*. Retention of these core gene amplifications imply 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, and this later process, rather than DTU type, 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.

Supplementary files

Supplemental Figure S1. Overview of the Brazil A4 and Y C6 genomes. Tracks from outer to inner circles indicate: sizes, chromosomes, 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).

Supplemental Figure S2. 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) (Carver et al. 2005). 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. Blocks: gaps; green bars: genes.

Supplemental Figure S3. Repetitive composition in the scaffolds. Chromosomes are calculated individually, while small scaffolds are calculated by averaging a range of scaffolds as indicated on the x axis.

Supplemental Figure S4. Workflow of predicting full repertoire of large gene families (taking TS as an example).

Supplemental Figure S5. Distribution of large gene families and retrotransposons on the chromosomes. Rings from outer to inner: chromosomes, retrotransposons, TS, MASP, mucin, GP63, RHS and DGF-1 gene families.

Supplemental Figure S6. 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 (Aslett et al. 2010).

Supplemental Figure S7. Analysis 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 was indicated in red. (B) Allele frequency calculated by the proportion of heterozygous SNPs/Indels at the non-repetitive regions of each chromosome. A diploid chromosome showed the peak of allele frequency around 50% as shown in chr8 and chr31, whereas a multi-ploid chromosome showed peak of allele frequency lower than 50% as shown in chr24 and chr28

580 in Brazil A4. Note that 5 chromosomes (Chr35, 36, 38, 39 and 42) in Brazil A4 were not included in this
581 analysis due to their high proportion of repetitive features.

582 **Supplemental Figure S8.** Alignment of TS (A) and their flanking regions (B) of the cluster in Figure 5, as
583 well as the phylogenetic tree of all the TS genes (C).

584 **Supplemental Figure S9.** An example of *in situ* diversification. (A) A tight cluster of GP63 genes from
585 MDS plot are distributed in different chromosomes. (B) Alignment of these GP63 genes showed high
586 identity as well as a number of diversifications including SNPs and Indels. (C) Phylogenetic trees of GP63
587 in the cluster (left), and GP63 plus flanking sequences on both sides (right).

588 **Supplemental Figure S10.** Copy numbers of 152 orthologue genes sets from Supplemental Table S11 are
589 highly correlated (Spearman correlation > 0.7) in pairwise comparisons between *T. brucei* strains and
590 subspecies and between *Leishmania* species, but poorly correlated between *T. cruzi* strains (range 0.006
591 - 0.6).

592 **Supplemental Figure S11**. Examples of long or short PTUs with tandem gene arrays. (A) Tandem arrays
593 of conserved gene sets are contained within long PTUs devoid of gene family members. Chromosomes
594 containing TcSMUG S/L in Brazil A4 (top) and Y C6 (middle), and TcMUCl+*HP in Y C6 (bottom). In
595 contrast, large gene family members, including some tandemly duplicated genes, show frequent strand
596 switches and are in short PTUs (B). Blue bars indicate genes other than large gene family members,
597 while yellow bars indicate large gene families.

598

599 **Supplemental Table S1.** Evaluation of assembly metrics among all available *T. cruzi* genomes assembled
600 by long-read sequencing. *No scaffolding was applied to these genomes, so no gaps were generated.
601 **47 are not *de novo* assembled contigs or scaffolds, but rather pseudomolecules produced by aligning
602 the core regions of scaffolds to the core regions of CL Brener reference genome. Therefore, although the

603 genome showed higher N50 and lower L50, it left an extensively high number of gaps behind.

604 ***Genome sequence is not available.

605 **Supplemental Table S2.** Repetitive sequences characterized in Brazil A4.

606 **Supplemental Table S3.** Repetitive sequences characterized in Y C6.

607 **Supplemental Table S4.** BUSCO assessment of gene completeness for *T. cruzi* genomes with annotation

608 available. Not that TCC is a hybrid strain, so its genome is a mixture of two haplotypes, while all other

609 genomes contain one haplotype.

610 **Supplemental Table S5.** Copy number of large gene families characterized in the new genomes.

611 **Supplemental Table S6.** Annotation summary.

612 **Supplemental Table S7.** Scaffolds that were detected to be allelic variants. Syntenies were examined

613 between small scaffolds and chromosomes. Only those with multiple syntenic regions throughout the

614 entire scaffold with part of the chromosome were considered as allelic variants.

615 **Supplemental Table S8.** Heterozygous SNPs/Indels identified in Brazil A4.

616 **Supplemental Table S9.** Heterozygous SNPs/Indels identified in Y C6.

617 **Supplemental Table S10.** Homozygous SNPs/Indels identified between Brazil A4 and Y C6.

618 **Supplemental Table S11.** Orthologue groups in *T. cruzi*, *T. brucei* and *Leishmania* species with total gene

619 count > 6. All sequences were retrieved from TritrypDB database (https://tritrypdb.org/tritrypdb/)

620 release-44.

621 **Supplemental Table S12.** List of unique genes in the respective strains.

622 **Supplemental Table S13.** BLAST result of the best match analysis in 6 large gene families between the

623 two strains.

624 **Supplemental Table S14.** Prominent tandem arrays of large gene families identified in Brazil A4 and Y C6.

625

626

Methods

Parasite cultures, DNA/RNA extraction and sequencing

Epimastigotes of Brazil and Y were cultured at 26°C in supplemented liver digested- neutralized tryptose (LDNT) medium as described previously (Xu et al. 2009). 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 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 and 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 at default except the expected genome size, which was set at 40 Mb for both

650 strains. Chicago and Hi-C libraries were constructed and sequenced by Dovetail Genomics, and HiRise
651 pipeline was run for scaffolding the draft assembly by incorporating data from both libraries. A gap was
652 generated whenever two contigs were joined or one contig was broken by HiRise and since the distance
653 between two contigs was unknown, all gaps were given 100 Ns.
654
655 Gap filling was performed by PBJelly (English et al. 2012) using the SMRT subreads with the minimum
656 percent identity at 85%. 122 and 4 gaps were extended for Brazil and Y, respectively. Correction of the
657 genomes using Illumina short reads was run by Pilon (Walker et al. 2014) and iCORN2 (Otto et al. 2010)
658 through multiple iterations to eliminate errors from SMRT sequencing.
659
660 **Repeat annotation**
661 RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler) was used to build a *de novo*
662 repeats library, and then used RepeatMasker v 4.0.7 (http://www.repeatmasker.org) with search engine
663 parameter as "ncbi".
664
665 **Genome annotation**
666 To develop open reading frame (ORF) in the new genome sequences, WebApollo 2.0 (Lee et al. 2013)
667 was deployed with the genome sequence and the following tracks of evidence were added:
668 1. Gene prediction from COMPANION (Steinbiss et al. 2016) using *Trypanosoma brucei* as
669 reference.
670 2. Gene prediction using AUGUSTUS (Stanke et al. 2004; Stanke and Morgenstern 2005) which was
671 self-trained by CL Brener genome.
672 3. Annotation transfer from CL Brener by Exonerate (Slater and Birney 2005).

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* (Queiroz et al. 2013).
6. Strand-specific RNA-seq alignment data, the pipeline of which was followed as previously described (Kieft et al. 2020).

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

InterProScan v5.31-70.0 (Jones et al. 2014) was used to detect protein families, domains and sites with all 11 default databases. Gene Ontology (GO) term was assigned also by InterProScan based on the protein domains results. Besides, BLASTP was used to search protein homology against *T. cruzi* CL Brener, *T. brucei*, *Leishmania 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 evalue <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 evalue <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. 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. The new candidate genes were BLASTNed against 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 BLASTed 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). The signal peptide sequences were BLASTNed against 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 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 gene family members were performed when necessary.

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 (Kerpedjiev et al. 2018).

Multidimensional scaling

K-tuple distance between genes are calculated with Clustal-Omega 1.2.4 (Sievers et al. 2011) using unaligned sequences with option parameters: “--full" and “--distmat-out". Full alignment distance between genes are calculated with Clustal-Omega 1.2.4 using aligned sequences (aligned with Clustal-Omega 1.2.4 using default parameters) with options 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.

Phylogenetic inference

744 Multiple sequence alignment was performed using MUSCLE (Edgar 2004). The resulting alignment was
745 manually edited. Bayesian inference of phylogeny was performed using MrBayes v.3.2.6 (Ronquist et al.
746 2012) with the following parameters: nst = 6, rates = invgamma, Ngammacat = 8, Ngen = 10,000,000,
747 nruns = 2, nchains = 4, and burn-infraction = 0.5. Convergence was determined by 25,000 post burn-in
748 samples from two independent runs. The resulting phylogenetic tree was rendered in Figtree v.1.4.4.
749 Node support values are given in percent posterior probability.
750

751 Data access

752

753 All raw and processed sequencing data generated in this study have been submitted to the NCBI
754 Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) under accession number
755 SRR118039885- SRR118039888 (Brazil A4) and SRR11845028- SRR11845031 (Y C6). The genome and
756 annotation data generated in this study have been submitted to the NCBI BioProject database
757 (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA512864 (Brazil A4) and
758 PRJNA554625 (Y C6). The assembled and annotated genomes are also accessible in the TritrypDB
759 database (https://tritrypdb.org/tritrypdb/).
760

761 Acknowledgements

762

763 We thank Dr. Todd Minning for initial contributions this project, Dr. Robert Sabatini from the University
764 of Georgia for insightful discussions, and Dr. Benedikt Brink from Ludwig-Maximilians-Universität for
765 advice and for testing our data in his genome phasing pipeline. This work was supported by funding
766 from the National Institutes of Health, (USA) grants R03 AI124228 and R01 AI124692 to RLT.

Disclosure Declaration

The authors declare no conflicts of interest.

References

Ackermann AA, Carmona SJ, Aguero F. 2009. TcSNP: a database of genetic variation in *Trypanosoma cruzi*. *Nucleic Acids Res* **37**: D544-549.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38**: D457-462.

Barnes RL, Shi H, Kolev NG, Tschudi C, Ullu E. 2012. Comparative genomics reveals two novel RNAi factors in *Trypanosoma brucei* and provides insight into the core machinery. *PLoS Pathog* **8**: e1002678.

Berna L, Rodriguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G, Alvarez-Valin F, Robello C. 2018. Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*. *Microb Genom* **4**.

Berriman M Ghedin E Hertz-Fowler C Blandin G Renauld H Bartholomeu DC Lennard NJ Caler E Hamlin NE Haas B et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**: 416-422.

Buscaglia CA, Campo VA, Frasch AC, Di Noia JM. 2006. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. *Nat Rev Microbiol* **4**: 229-236.

Callejas-Hernandez F, Rastrojo A, Poveda C, Girones N, Fresno M. 2018. Genomic assemblies of newly sequenced *Trypanosoma cruzi* strains reveal new genomic expansion and greater complexity. *Sci Rep* **8**: 14631.

Carlos Talavera-López JLR-C, Louisa A. Messenger, Michael D. Lewis, Matthew Yeo, Daniella C. Bartholomeu, José E. Calzada, Azael Saldaña, Juan David Ramírez, Felipe Guhl, Sofía Ocaña-Mayorga, Jaime A. Costales, Rodion Gorchakov, Kathryn Jones, Melissa Nolan Garcia, Edmundo C. Grisard, Santuza M. R. Teixeira, Hernán Carrasco, Maria Elena Bottazzi, Peter J. Hotez, Kristy O. Murray, Mario J. Grijalva, Barbara Burleigh, Michael A. Miles, Björn Andersson. 2018. Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. *bioRxiv* doi: https://doi.org/10.1101/283531.

CaroleBranchea S, LenaÅslundb, BjörnAnderssona. 2006. Comparative karyotyping as a tool for genome structure analysis of *Trypanosoma cruzi*. *Mol Biochem Parasitol* **147**: 30–38.

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics* **21**: 3422-3423.

Clayton C. 2019. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol* **9**: 190072.

Cliffe LJ, Siegel TN, Marshall M, Cross GA, Sabatini R. 2010. 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* **38**: 3923-3935.

Cremona ML, Sanchez DO, Frasch AC, Campetella O. 1995. A single tyrosine differentiates active and inactive *Trypanosoma cruzi trans*-sialidases. *Gene* **160**: 123-128.

Cross GA. 1975. Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of *Trypanosoma brucei*. *Parasitology* **71**: 393-417.

DaRocha WD, Otsu K, Teixeira SM, Donelson JE. 2004. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. *Mol Biochem Parasitol* **133**: 175-186.

de Freitas JM, Augusto-Pinto L, Pimenta JR, Bastos-Rodrigues L, Goncalves VF, Teixeira SM, Chiari E, Junqueira AC, Fernandes O, Macedo AM et al. 2006. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. *PLoS Pathog* **2**: e24.

De Pablos LM, Osuna A. 2012. Multigene families in *Trypanosoma cruzi* and their role in infectivity. *Infect Immun* **80**: 2258-2264.

Diaz-Viraque F, Pita S, Greif G, de Souza RCM, Iraola G, Robello C. 2019. Nanopore sequencing significantly improves genome assembly of the protozoan parasite *Trypanosoma cruzi*. *Genome Biol Evol* **11**: 1952–1957.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, Wanless D, Peterson J, Hou L, Taylor S et al. 2003. The sequence and analysis of *Trypanosoma brucei* chromosome II. *Nucleic Acids Res* **31**: 4856-4863.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G et al. 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**: 409-415.

Elbers JP, Rogers MF, Perelman PL, Proskuryakova AA, Serdyukova NA, Johnson WE, Horin P, Corander J, Murphy D, Burger PA. 2019. Improving Illumina assemblies with Hi-C and long reads: An example with the North African dromedary. *Mol Ecol Resour* **19**: 1015-1026.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**: e47768.

Ersfeld K, Melville SE, Gull K. 1999. Nuclear and genome organization of *Trypanosoma brucei*. *Parasitol Today* **15**: 58-63.

Flores-Lopez CA, Machado CA. 2011. Analyses of 32 loci clarify phylogenetic relationships among *Trypanosoma cruzi* lineages and support a single hybridization prior to human contact. *PLoS Negl Trop Dis* **5**: e1272.

Frasch AC. 2000. Functional diversity in the *trans*-sialidase and mucin families in *Trypanosoma cruzi*. *Parasitol Today* **16**: 282-286.

Gonzalez MS, Souza MS, Garcia ES, Nogueira NF, Mello CB, Canepa GE, Bertotti S, Durante IM, Azambuja P, Buscaglia CA. 2013. *Trypanosoma cruzi* TcSMUG L-surface mucins promote development and infectivity in the triatomine vector *Rhodnius prolixus*. *PLoS Negl Trop Dis* **7**: e2552.

Henriksson J AL, Macina RA, Franke de Cazzulo BM, Cazzulo JJ, Frasch AC, Pettersson U. 1990. Chromosomal localization of seven cloned antigen genes provides evidence of diploidy and further demonstration of karyotype variability in *Trypanosoma cruzi*. *Mol Biochem Parasitol* **42**: 213–223.

852 Henriksson J, Dujardin JC, Barnabe C, Brisse S, Timperman G, Venegas J, Pettersson U, Tibayrenc M,
853 Solari A. 2002. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is
854 evolutionarily informative. *Parasitology* **124**: 277-286.
855 Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, Cazzulo JJ, Frasch AC, Pettersson U. 1995.
856 Chromosome specific markers reveal conserved linkage groups in spite of extensive
857 chromosomal size variation in *Trypanosoma cruzi*. *Mol Biochem Parasitol* **73**: 63-74.
858 Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, Brooks K, Churcher C, Fahkro S,
859 Goodhead I et al. 2008. Telomeric expression sites are highly conserved in *Trypanosoma brucei*.
860 *PLoS One* **3**: e3527.
861 Huckaby AC, Granum CS, Carey MA, Szlachta K, Al-Barghouthi B, Wang YH, Guler JL. 2019. Complex DNA
862 structures trigger copy number variation across the *Plasmodium falciparum* genome. *Nucleic*
863 *Acids Res* **47**: 1615-1627.
864 Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A,
865 Melville SE et al. 2010. The genome sequence of *Trypanosoma brucei* gambiense, causative
866 agent of chronic human african trypanosomiasis. *PLoS Negl Trop Dis* **4**: e658.
867 Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al.
868 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-
869 1240.
870 Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from *in vivo* DNA interaction frequency.
871 *Nat Biotechnol* **31**: 1143-1147.
872 Kawasaki F, Beraldi D, Hardisty RE, McInroy GR, van Delft P, Balasubramanian S. 2017. Genome-wide
873 mapping of 5-hydroxymethyluracil in the eukaryote parasite *Leishmania*. *Genome Biol* **18**: 23.
874 Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A,
875 Kumar N et al. 2018. HiGlass: web-based visual exploration and analysis of genome interaction
876 maps. *Genome Biol* **19**: 125.
877 Kieft R, Zhang Y, Marand AP, Moran JD, Bridger R, Wells L, Schmitz RJ, Sabatini R. 2020. Identification of
878 a novel base J binding protein complex involved in RNA polymerase II transcription termination
879 in trypanosomes. *PLoS Genet* **16**: e1008390.
880 Korbel JO, Lee C. 2013. Genome assembly and haplotyping with Hi-C. *Nat Biotechnol* **31**: 1099-1101.
881 Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE.
882 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14**: R93.
883 Lewis EB. 1951. Pseudoallelism and gene evolution. *Cold Spring Harb Symp Quant Biol* **16**: 159-174.
884 Lima FM, Souza RT, Santori FR, Santos MF, Cortez DR, Barros RM, Cano MI, Valadares HM, Macedo AM,
885 Mortara RA et al. 2013. Interclonal variations in the molecular karyotype of *Trypanosoma cruzi*:
886 chromosome rearrangements in a single cell-derived clone of the G strain. *PLoS One* **8**: e63738.
887 Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:
888 1151-1155.
889 Martin DL, Weatherly DB, Laucella SA, Cabinian MA, Crim MT, Sullivan S, Heiges M, Craven SH,
890 Rosenberg CS, Collins MH et al. 2006. CD8+ T-Cell responses to *Trypanosoma cruzi* are highly
891 focused on strain-variant *trans*-sialidase epitopes. *PLoS Pathog* **2**: e77.
892 Mowatt MR, Aggarwal A, Nash TE. 1991. Carboxy-terminal sequence conservation among variant-
893 specific surface proteins of *Giardia lamblia*. *Mol Biochem Parasitol* **49**: 215-227.
894 Mugnier MR, Stebbins CE, Papavasiliou FN. 2016. Masters of Disguise: Antigenic Variation and the VSG
895 Coat in *Trypanosoma brucei*. *PLoS Pathog* **12**: e1005784.
896 Muller HJ. 1936. Bar Duplication. *Science* **83**: 528-530.
897 Muller LSM, Cosentino RO, Forstner KU, Guizetti J, Wedel C, Kaplan N, Janzen CJ, Arampatzi P, Vogel J,
898 Steinbiss S et al. 2018. Genome organization and DNA accessibility control antigenic variation in
899 trypanosomes. *Nature* **563**: 121-125.

Nakayasu ES, Yashunsky DV, Nohara LL, Torrecilhas AC, Nikolaev AV, Almeida IC. 2009. GPIomics: global analysis of glycosylphosphatidylinositol-anchored molecules of *Trypanosoma cruzi*. *Mol Syst Biol* **5**: 261.

Navarro M, Gull K. 2001. A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*. *Nature* **414**: 759-763.

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

Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**: 1704-1707.

Pedroso A, Cupolillo E, Zingales B. 2003. Evaluation of *Trypanosoma cruzi* hybrid stocks based on chromosomal size variation. *Mol Biochem Parasitol* **129**: 79-90.

Pimenta PF, da Silva PP, Nash T. 1991. 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* **59**: 3989-3996.

Previato J, Andrade AFB, Pessolani MCV, and Mendonça-Previato L. 1985. Incorporation of sialic acid into *Trypanosoma cruzi* macromolecules. A proposal for a new metabolic route. *Mol Biochem Parasitol* **16**: 85–96.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW et al. 2016. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res* **26**: 342-350.

Queiroz RM, Charneau S, Motta FN, Santana JM, Roepstorff P, Ricart CA. 2013. Comprehensive proteomic analysis of *Trypanosoma cruzi* epimastigote cell surface proteins by two complementary methods. *J Proteome Res* **12**: 3255-3263.

Rassi A, Jr., Rassi A, Marcondes de Rezende J. 2012. American trypanosomiasis (Chagas disease). *Infect Dis Clin North Am* **26**: 275-291.

Reis-Cunha JL, Rodrigues-Luiz GF, Valdivia HO, Baptista RP, Mendes TA, de Morais GL, Guedes R, Macedo AM, Bern C, Gilman RH et al. 2015. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains. *BMC Genomics* **16**: 499.

Revollo S, Oury B, Laurent JP, Barnabe C, Quesney V, Carriere V, Noel S, Tibayrenc M. 1998. *Trypanosoma cruzi*: impact of clonal evolution of the parasite on its biological and medical properties. *Exp Parasitol* **89**: 30-39.

Reynolds D, Cliffe L, Forstner KU, Hon CC, Siegel TN, Sabatini R. 2014. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei*. *Nucleic Acids Res* **42**: 9717-9729.

Reynolds D, Hofmeister BT, Cliffe L, Alabady M, Siegel TN, Schmitz RJ, Sabatini R. 2016. Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in *Trypanosoma brucei*. *PLoS Genet* **12**: e1005758.

Robert D. Denton RSK, Jacob W. Malcom, Louis Du Preez, and John H. Malone. 2018. The African Bullfrog (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. *bioRxiv* doi: https://doi.org/10.1101/329847.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**: 539-542.

Salter JF, Johnson O, Stafford NJ, 3rd, Herrin WF, Jr., Schilling D, Cedotal C, Brumfield RT, Faircloth BC. 2019. A Highly Contiguous Reference Genome for Northern Bobwhite (*Colinus virginianus*). *G3 (Bethesda)* **9**: 3929-3932.

946 Schreiber M, Mascher M, Wright J, Padmarasu S, Himmelbach A, Heavens D, Milne L, Clavijo BJ, Stein N,
947 Waugh R. 2020. A Genome Assembly of the Barley 'Transformation Reference' Cultivar Golden
948 Promise. *G3 (Bethesda)* **10**: 1823-1827.
949 Schulz D, Zaringhalam M, Papavasiliou FN, Kim HS. 2016. Base J and H3.V Regulate Transcriptional
950 Termination in *Trypanosoma brucei*. *PLoS Genet* **12**: e1005762.
951 Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, Wang X, Dewell S, Cross GA. 2009.
952 Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma*
953 *brucei*. *Genes Dev* **23**: 1063-1076.
954 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et
955 al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using
956 Clustal Omega. *Mol Syst Biol* **7**: 539.
957 Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC*
958 *Bioinformatics* **6**: 31.
959 Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, Pinches R, Newbold CI, Miller
960 LH. 1995. Switches in expression of Plasmodium falciparum *var* genes correlate with changes in
961 antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**: 101-110.
962 Souza RT, Lima FM, Barros RM, Cortez DR, Santos MF, Cordero EM, Ruiz JC, Goldenberg S, Teixeira MM,
963 da Silveira JF. 2011. Genome size, karyotype polymorphism and chromosomal evolution in
964 *Trypanosoma cruzi*. *PLoS One* **6**: e23042.
965 Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows
966 user-defined constraints. *Nucleic Acids Res* **33**: W465-467.
967 Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in
968 eukaryotes. *Nucleic Acids Res* **32**: W309-312.
969 Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD. 2016. Companion: a
970 web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* **44**: W29-34.
971 Sturtevant AH. 1925. The Effects of Unequal Crossing over at the Bar Locus in Drosophila. *Genetics* **10**:
972 117–147.
973 Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE. 1995.
974 The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic
975 variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82**: 89-100.
976 Theodore S. Kalbfleisch ESR, Michael S. DePriest Jr., Brian P. Walenz, Matthew S. Hestand, Joris R.
977 Vermeesch, Brendan L. O'Connell, Ian T. Fiddes, Alisa O. Vershinina, Jessica L. Petersen, Carrie J.
978 Finno, Rebecca R. Bellone, Molly E. McCue, Samantha A. Brooks, Ernest Bailey, Ludovic Orlando,
979 Richard E. Green, Donald C. Miller, Douglas F. Antczak, James N. MacLeod. 2018. EquCab3, an
980 Updated Reference Genome for the Domestic Horse. *bioRxiv* doi:
981 https://doi.org/10.1101/306928.
982 Tomasini N, Diosque P. 2015. Evolution of *Trypanosoma cruzi*: clarifying hybridisations, mitochondrial
983 introgressions and phylogenetic relationships between major lineages. *Mem Inst Oswaldo Cruz*
984 **110**: 403-413.
985 Triana O, Ortiz S, Dujardin JC, Solari A. 2006. *Trypanosoma cruzi*: variability of stocks from Colombia
986 determined by molecular karyotype and minicircle Southern blot analysis. *Exp Parasitol* **113**: 62-
987 66.
988 Uemura H, Schenkman, S., Nussenzweig, V., and Eichinger, D. 1992. Only some members of a gene
989 family in *Trypanosoma cruzi* encode proteins that express both *trans*-sialidase and
990 neuraminidase activities. *EMBO J* **11**: 3837–3844.
991 van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A,
992 Ramasamy G et al. 2012. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional
993 readthrough in *Leishmania*. *Cell* **150**: 909-921.

Vargas N, Pedroso A, Zingales B. 2004. Chromosomal polymorphism, gene synteny and genome size in *T. cruzi* I and *T. cruzi* II groups. *Mol Biochem Parasitol* **138**: 131-141.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Weatherly DB, Boehlke C, Tarleton RL. 2009. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. *BMC Genomics* **10**: 255.

Weatherly DB, Peng D, Tarleton RL. 2016. Recombination-driven generation of the largest pathogen repository of antigen variants in the protozoan *Trypanosoma cruzi*. *BMC Genomics* **17**: 729.

Westenberger SJ, Barnabe C, Campbell DA, Sturm NR. 2005. Two hybridization events define the population structure of *Trypanosoma cruzi*. *Genetics* **171**: 527-543.

Weston D, Patel B, Van Voorhis WC. 1999. Virulence in *Trypanosoma cruzi* infection correlates with the expression of a distinct family of sialidase superfamily genes. *Mol Biochem Parasitol* **98**: 105-116.

Xu D, Brandan CP, Basombrio MA, Tarleton RL. 2009. Evaluation of high efficiency gene knockout strategies for *Trypanosoma cruzi*. *BMC Microbiol* **9**: 90.

Yoshida N. 2006. Molecular basis of mammalian cell invasion by *Trypanosoma cruzi*. *An Acad Bras Cienc* **78**: 87-111.

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

Zingales B, Andrade SG, Briones MR, Campbell DA, Chiari E, Fernandes O, Guhl F, Lages-Silva E, Macedo AM, Machado CR et al. 2009. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. *Mem Inst Oswaldo Cruz* **104**: 1051-1054.

Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MM, Schijman AG, Llewellyn MS, Lages-Silva E, Machado CR et al. 2012. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect Genet Evol* **12**: 240-253.