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

25 and *in situ* mutations. These processes also impact genes not required for or involved in immune
26 evasion, creating unique challenges with respect to preserving core genome function while maximizing
27 genetic diversity.
28

29 ## Introduction

30

31 The protozoan parasite *Trypanosoma cruzi* is the causative agent of Chagas disease, the highest impact
32 parasitic infection in the Americas, affecting 10 to 20 million humans and innumerable animals in many
33 species. The study of *T. cruzi* and Chagas disease is particularly challenging for a number of reasons,
34 including the complexity and unique characteristics of its genome. Over 50% of the *T. cruzi* genome is
35 composed of repetitive sequences, which include numerous families of surface proteins (e.g. *trans*-
36 sialidases, mucins and mucin-associated surface proteins) with hundreds to thousands of members
37 each, as well as substantial numbers of transposable elements, microsatellites and simple tandem
38 repeats (Weston et al. 1999; El-Sayed et al. 2005; De Pablos and Osuna 2012). This repetitive nature
39 greatly hampered the assembly of the original CL Brener strain reference genome generated in 2005,
40 resulting in a highly fragmented and draft assembly with extensively collapsed high repeat regions (El-
41 Sayed et al. 2005). In addition, the CL Brener strain turned out to be a hybrid strain with divergent alleles
42 at many loci. To scaffold the genome sequence, Weatherly *et al.* took advantage of the bacterial artificial
43 chromosome (BAC) library sequencing data and combined with synteny analysis of two genomes from
44 closely related species, *Trypanosoma brucei* and *Leishmania*, obtained the current reference genome
45 with 41 chromosomes (Weatherly et al. 2009). Nevertheless, a large number of gaps are still present in
46 the chromosomes of the reference genome, and many unassigned contigs remain, making it impossible
47 to determine the exact genome content and, in particular, the full repertoires of large gene families.
48

49 As in many pathogens, and best documented in the related trypanosomatid *Trypanosoma brucei,*
50 families of variant surface proteins often serve as both the primary molecular interface with mammalian
51 hosts and as the predominant target of host immune responses. Classical antigenic variation in these
52 pathogens consists of the serial expression of a single (or highly restricted number of) antigen variant(s)
53 in the pathogen population at any one time, with switches to new variants becoming evident once the
54 host immune response controls the dominant one. This largely "one-at-a-time" strategy appears
55 particularly effective in pathogens exposed continuously to antibody-mediated immune control
56 mechanisms. *T. cruzi*, however, appears to take a much different approach to antigenic variation,
57 generating multiple very large families of genes encoding surface and secreted proteins, many of which
58 are expressed simultaneously rather than serially. We believe that this strategy may reflect the primarily
59 intracellular lifestyle of *T. cruzi* in mammalian hosts and the necessity of evading T cell recognition of
60 infected host cells, although this has yet not been formally proven.
61
62 The advent of two advances in genome analysis has made it feasible to revisit and substantially improve
63 upon the *T. cruzi* genome assembly and to advance our understanding of its composition. The long-read
64 capability of PacBio Single-Molecule Real-Time (SMRT) sequencing provides read lengths capable of
65 spanning long repetitive regions. The application of this technology (Berna et al. 2018; Callejas-
66 Hernandez et al. 2018) as well as nanopore sequencing (Diaz-Viraque et al. 2019) has resulted in much-
67 improved contiguity and expansion of gene family members in *T. cruzi*. Secondly, proximity ligation
68 methods have allowed for the scaffolding of assemblies spanning highly repetitive regions. One of the
69 methods, Hi-C, identifies extant inter-chromosomal interactions by capturing chromosome
70 conformation, and has been used to create scaffolds at chromosomal scale (Kaplan and Dekker 2013;
71 Korbel and Lee 2013). A second approach termed Chicago, adapts this same methodology but
72 reconstitutes the confirmation of DNA *in vitro* by combining the DNA with purified histones and

73 chromatin assembly factors (Putnam et al. 2016). These proximity ligation methods not only improve
74 the contiguity of genomes by joining contigs, they also identify misjoins in the contigs and separate
75 them to increase the accuracy of assemblies (Putnam et al. 2016). The combination of Chicago and Hi-C
76 has now been applied to many genomes (Robert D. Denton 2018; Theodore S. Kalbfleisch 2018; Elbers
77 et al. 2019; Salter et al. 2019; Schreiber et al. 2020).
78
79 In this study, we have applied SMRT sequencing and proximity ligation methods to produce very high-
80 quality assemblies from the Brazil (TcI) and Y (TcII) strains of *T. cruzi*. These two strains are
81 representatives of the most ancestral lines that are hypothesized to have given rise to the 6 discrete
82 typing units (DTUs, TcI-TcVI) lineages now composing this genetically diverse species (Westenberger et
83 al. 2005; de Freitas et al. 2006; Zingales et al. 2009; Flores-Lopez and Machado 2011; Zingales et al.
84 2012; Tomasini and Diosque 2015). Using these chromosomal-level assemblies with minimal gaps, we
85 are now able to compare the full gene content of representatives of these founding lineages of the *T.*
86 *cruzi* species, including the full repertoires of large gene families. Herein, we document a substantial
87 diversity in individual chromosome content, including frequent allelic variants, but with an overall
88 conserved gene content outside of the large gene families. Within these gene families, however,
89 extreme diversification is evident with no genes of the identical sequence within strains or shared
90 between these strains. These high-quality genomes also reveal the mechanisms behind the expansion
91 and diversification of the large gene families, presumably in response to immunological pressure, and in
92 the process, creating other challenges in terms of core genome stability and function.
93

94 ## Results

95

96 **Genome Sequencing and Assembly**

97 PacBio SMRT sequencing provided 1,264,527 (N50=9,560 bp) and 763,579 (N50=12,499 bp) filtered
98 reads with ~9 Gb and ~6 Gb of sequence data for Brazil clone A4 (Brail A4) and Y clone C6 (Y C6),
99 respectively, corresponding to ~200x and ~130x coverage based on the predicted genome size. Initial
100 assembly resulted in sequences of 45.11 Mb and 46.98 Mb for Brazil A4 and Y C6 draft genomes,
101 respectively, close to the estimated haploid genome size of *T. cruzi* (Souza et al. 2011) (Table 1).
102 Application of the *in vitro* proximity-ligation tools Hi-C and Chicago (Putnam et al. 2016), decreased the
103 L50 to half of that of the draft genomes, and the size of the largest scaffolds doubled (Table 1). Filling
104 gaps and base correction using Illumina reads ultimately resulted in 12 and 14 scaffolds in the Brazil A4
105 and Y C6 final assemblies, respectively, with a length greater than 1 Mb. Telomeric repeats [(TTAGGG)n]
106 were identified in 18 Brazil A4 and 15 Y C6 scaffolds, including on both ends of three scaffolds in Brazil
107 A4, suggesting full chromosome assembly in these cases. The improvement in these new genomes is not
108 only in integrity (Supplemental Table S1 and Supplemental Fig. S1), but also in filled gaps, recovered
109 genes and extended repetitive regions (see examples in Supplemental Fig. S2).
110

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 ICorr	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 ICorr	47.22	266	51.58	889,019	18	2,951,016	231

111

112 **Table 1.** Summary of assembly statistics.

113

114 **Genome features and content**

115 Due to a lack of apparent chromosome condensation during replication (Henriksson et al. 2002; Souza et
116 al. 2011), the karyotype of *T. cruzi* has not been completely elucidated. Moreover, chromosome size and
117 content vary significantly between different *T. cruzi* strains and even among clones of the same strain
118 based upon pulse-filed gel electrophoresis (PFGE) analysis (Henriksson et al. 2002; Pedroso et al. 2003;
119 Vargas et al. 2004; Triana et al. 2006; Lima et al. 2013). Based on criteria including size, repeat
120 proportion, and gene number, 43 scaffolds of Brazil A4 and 40 scaffolds of Y C6 were designated as
121 chromosomes (Supplemental Fig. S3) and the remainder referred to as smaller scaffolds.

122

123 Repetitive sequences occupy 58.8% and 62.3% of the genome for Brazil A4 and Y C6 (Supplemental
124 Tables S2 and S3), substantially higher than the 50% that was estimated in the reference CL Brener
125 genome, thus confirming the capability of long-read sequencing and assembly approaches to recover
126 and place more repetitive DNA content. Approximately 50% of the sequence in chromosomes is
127 repetitive sequences, compared to ~90% in smaller scaffolds (Supplemental Fig. S3). Using conventional
128 approaches with manual curation, gene models were identified in Brazil A4 and Y C6, respectively. Based
129 on BUSCO assessment, the Brazil A4 and Y C6 contain the highest number of single-copy gene sets
130 among assembled *T. cruzi* genomes (Supplemental Table S4).

131

132 A major constituent of the repetitive regions in the *T. cruzi* genome is large gene families, including the
133 *trans*-sialidases (TS), mucin associated surface proteins (MASP), mucins, and surface protease GP63 (all
134 targets of immune responses), as well as retrotransposon hotspot (RHS) proteins and dispersed gene
135 family 1 proteins (DGF-1) (El-Sayed et al. 2005; Buscaglia et al. 2006; Martin et al. 2006). Our previous
136 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

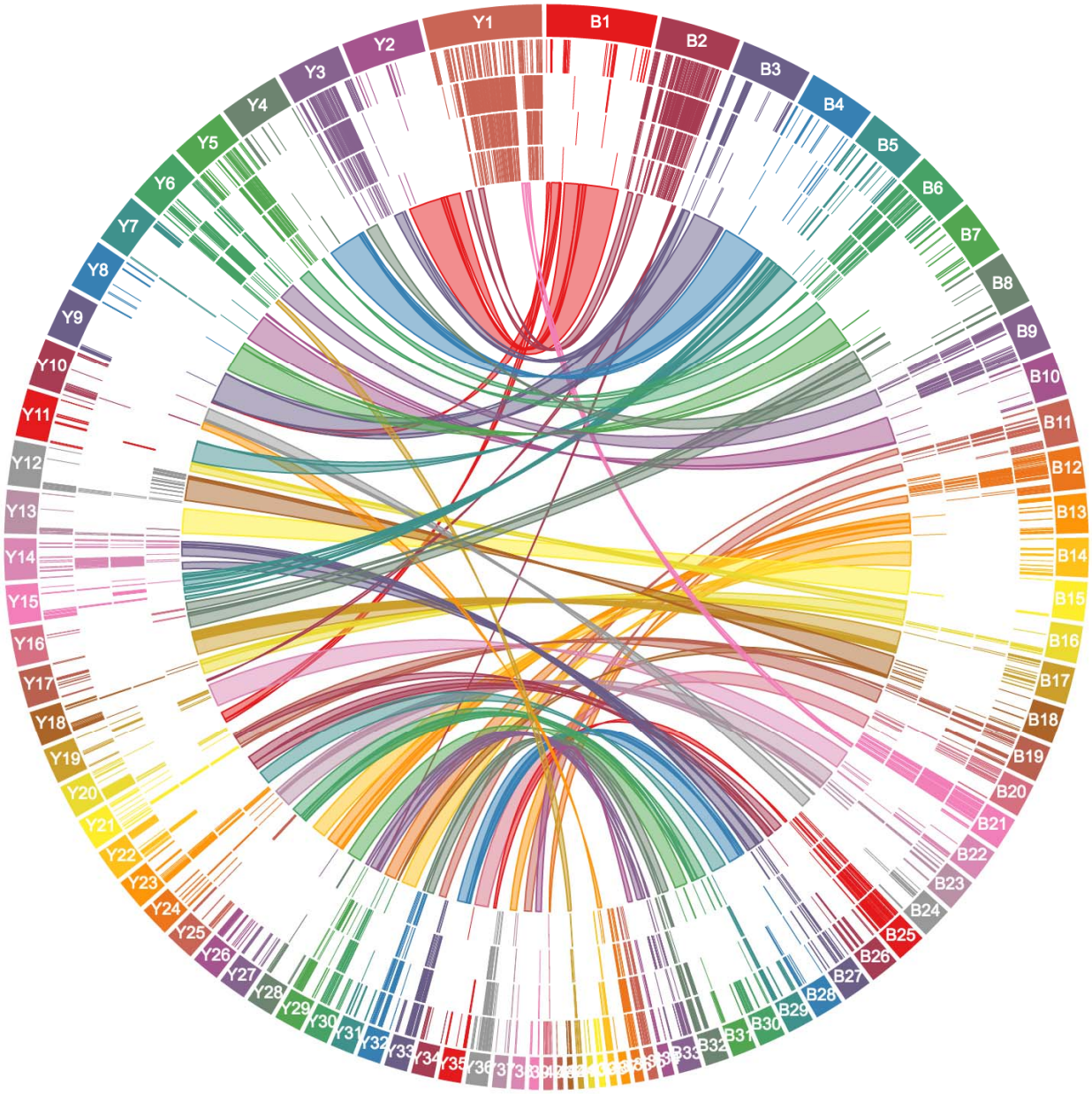
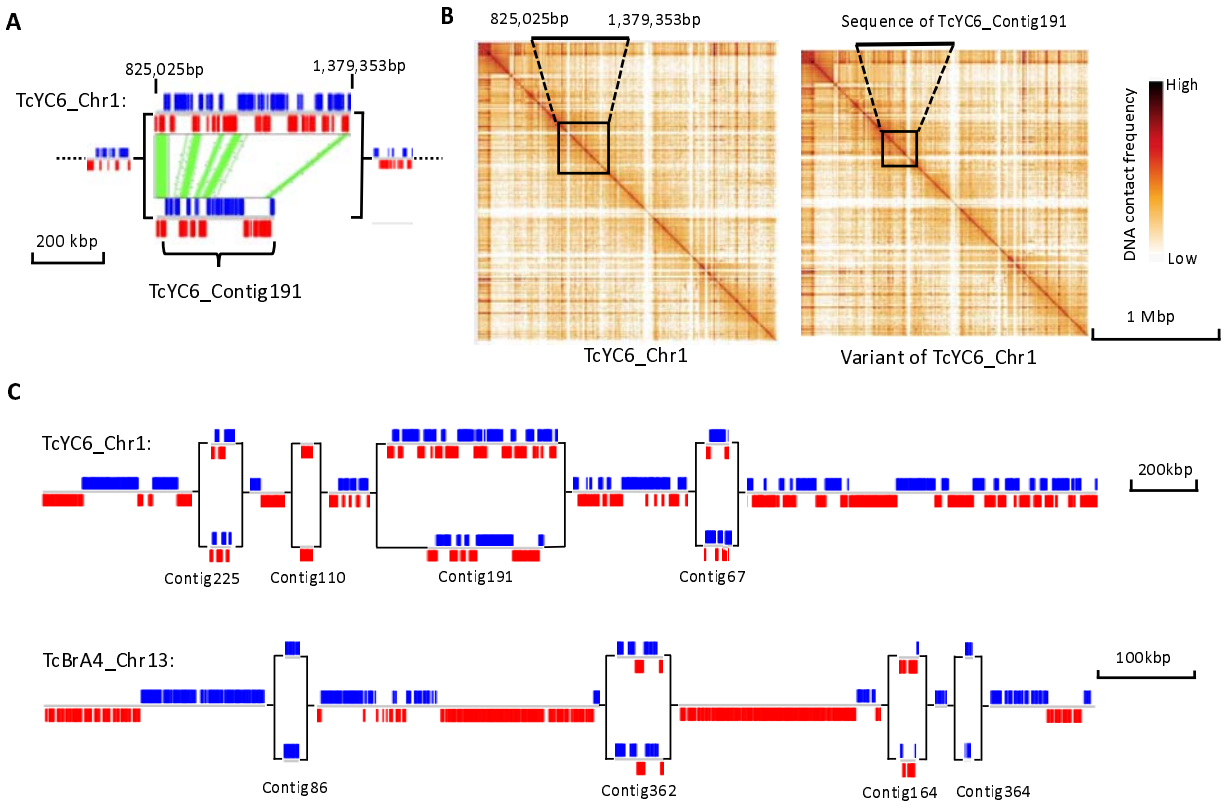


Figure 1. Distribution of large gene families and synteny between chromosomes in Brazil A4 (right, B) and Y C6, (left, Y). Tracks from outer to inner rings: chromosomes, TS, MASP, mucin, GP63, and synteny blocks.

156 After consolidating the predictions of large gene families with our conventional annotations, the Brazil
157 A4 and Y C6 genomes contained 18,708 and 17,650 gene models, respectively (see annotation summary
158 in Supplemental Table S6). The composition of gene content between two genomes is very similar, with
159 ~25% as members of large gene families, ~40% as hypothetical proteins, and >90% of the remaining
160 genes are orthologs of those in the related kinetoplastids *T. brucei* and *Leishmania major*. That this gene
161 model count in the two *T. cruzi* strains is substantially higher than that estimated for *T. brucei* and *L.*
162 *major* is likely due to two factors: 1) the high number of large gene family members in *T. cruzi*, and 2) a
163 greater number of hypothetical genes in *T. cruzi*, a third of which are unique to *T. cruzi*, although the size
164 distribution of the hypothetical proteins is similar in the 3 species (Supplemental Fig. S6).
165
166 **Allelic variation**
167 The significant number of small scaffolds and the relatively high gene model numbers in some of the
168 small scaffolds prompted us to consider whether these small scaffolds might represent regions of allelic
169 variation between sister chromosomes, as allelic variation is one of the factors that results in
170 fragmentation during genome assembly for diploid genomes. Although TcI and TcII DTUs represented by
171 the Brazil and Y strains, respectively, are considered homozygous lineages, we very conservatively
172 detected 26 and 33 small scaffolds in each genome showing consistent synteny in multiple gene models
173 to parts of the core chromosomes (Supplemental Table S7). An example is shown in Fig. 2A in which
174 scaffold TcYC6_Contig191 demonstrates regions of synteny within the 825,025 – 1,379,353 bp region in
175 the first chromosome of Y C6 (TcYC6_Chr1). Confirmation of this chromosome variant was supplied by
176 replacing the identified region in TcYC6_Chr1 with TcYC6_Contig191 and then mapping the
177 chromosomal contacts in the Hi-C data for these 2 alternative versions for TcYC6-Chr1. As shown in Fig.
178 2B, the Hi-C data are equally strong for both chromosome variants. Using Falcon-Phase, which phases
179 diploid genome sequences by integrating long reads and Hi-C data (Zev N. Kronenberg 2018), we

180 identified an additional 18 and 7 allelic variations in Brazil A4 and Y C6, respectively. In combination,
181 these analyses identified allelic variations in 24 chromosomes of Brazil A4 and 25 of Y C6, including
182 chromosomes with multiple allelic variants, e.g. the largest chromosome in Y C6 (TcYC6_Chr1), and an
183 intermediate-sized chromosome in Brazil A4 (TcBrA4_Chr13; Fig. 2C). Thus, we suggest that many of the
184 small scaffolds are variants of regions in the chromosome-size scaffolds. However, because the majority
185 of these small scaffolds lack the conserved, non-gene family sequences required to prove synteny, and
186 Falcon-Phase can only resolve haplotypes bearing divergence of < 5%, identifying the position of all the
187 small scaffolds on the chromosomes was not possible.
188



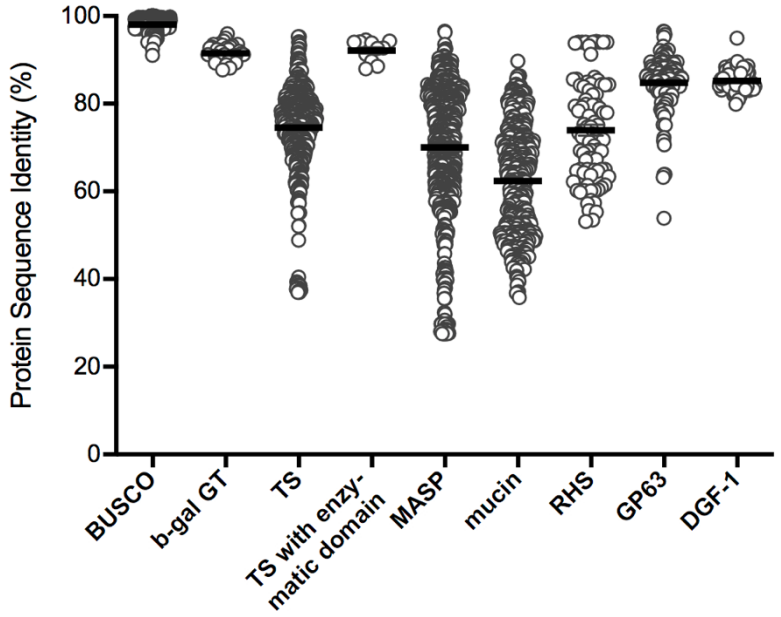
189

190 **Figure 2.** An example of homologous chromosomes with large allelic variations. (A) Synteny between two allelic
191 variants in Chr1 of Y C6. Synteny blocked are marked with green. (B) Hi-C heat maps of TcYC6_Chr1 (left) and its
192 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

217 partially syntenic chromosomes (TcBrA4_Chr24 and TcYC6_Chr10), which also share synteny with
218 chromosome 31 in CL Brener, reported to be supernumerary in many strains (Reis-Cunha et al. 2015),
219 thus suggesting a species-wide requirement for > 2 copies of one or more genes in these regions.
220 Additionally, variation exist in the copy number for a substantial number of individual genes
221 characterized by OrthoFinder, with ~150 genes showing the greatest variation between the two strains
222 (Supplemental Table S11). However, with respect to genes unique to either strain, we found 23 (Brazil
223 A4) and 20 (Y C6) gene loci not present in the other strain and further validated this finding by
224 examining the raw reads (Supplemental Table S12). All are annotated as hypothetical proteins and most
225 are small genes located in gene family-rich regions of the genome and thus are likely the products of
226 recombination events involved in gene family diversification (see below).
227
228 To fully assess the variation in the large gene family members between the two strains, we carried out a
229 best match search for the protein sequence of putatively expressed genes in each large gene family
230 from Y C6 genome with those in Brazil A4. As a control, the same analysis was performed for a subset of
231 mostly single-copy genes (BUSCO), as well as a small gene family of 35 members, beta galactofuranosyl
232 glycosyltransferase (b-gal GT). As shown in Fig. 3, high-identity matches could always be found for the
233 BUSCO genes, and some of them (22 out of 291) have identical matches (100% identity) in the other
234 strain. Similar to BUSCO genes, the identity between best matches for b-gal GT is also tightly distributed
235 in the range of 90-97%. In contrast, all large gene families exhibit a broad distribution of identity for
236 their best matches relative to the BUSCO genes and b-gal GT genes, especially TS, MASP, mucin and RHS,
237 with only a small proportion of best matches bearing 90% identity or more. Among the family members
238 with the greatest similarity between the two strains are the small subset of TS genes containing the
239 sialidase enzymatic domain as previously described (Cremona et al. 1995), suggesting that this group of
240 *trans*-sialidases has been selected for and conserved in both strains.

241



242

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

244

245 **<u>Evidence of gene family expansion and diversification</u>**

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

251

252 We first examined the genomes for evidence of gene duplication events that could increase the number
253 of members in gene families. Multidimensional scaling (MDS) plots based on the pair-wise genetic
254 distances of all members of each gene family in each strain allowed us to identify tightly distributed
255 gene clusters with high sequence identity (http://shiny.ctegd.uga.edu). In multiple cases, genes within
256 these clusters were tandemly arrayed individually (TS; Fig. 4A top) or as a set of genes (TS plus MASP;

257 Fig. 4A bottom). Such tandem amplifications are present in all gene families (except DGF-1) and occur
258 uniquely in each strain (Supplemental Table S14). A number of unusual amplification events were also
259 noted, including inverted duplications creating a strand switch in between (Fig. 4B), and an amplification
260 involving several genes on both strands, replicated a total of 5 times (Fig. 4C), thus creating a complex
261 set of strand switches.
262
263 The majority of tandem amplification of gene families in both *T. cruzi* genomes contained 10 or fewer
264 replicates (Supplemental Table S14). However, one hypothetical protein (*HP) in the Y C6 occurr in a
265 tandem array of 29 units with a TcMUCI gene (Fig. 4D). Comparison to the syntenic region in Brazil A4
266 revealed a single TcMUCI ortholog (and no *HP sequence), indicating that at some point the *HP
267 sequence was inserted next to the TcMUCI gene in Y C6, and the two genes were amplified together as a
268 segment (Fig. 4D). Although no particular protein domains were characterized in the *HP gene, 18% of
269 its sequence share similarity with several MASP sequences, implying that at least part of the gene might
270 be derived from a MASP. The abnormally high number of replicates in this tandem array as well as the
271 low diversity in the tandem copies suggest that this might be a recent amplification event. However,
272 comparison to syntenic regions in other long-read sequenced *T. cruzi* genomes revealed the same
273 TcMUCI+*HP tandem array in the TCC (TcIV) strain but not in the Dm28c (TcI), Sylvio (TcI), and Bug2148
274 (TcV) (Fig. 4D). Additionally, phylogenetic analysis grouped all of the replicated copies from Y C6 and TCC
275 together and distant from the single TcMUCI genes in the other three strains (Fig. 4E). Using the model
276 of DTU evolution in *T. cruzi* which postulates that the TcVI is derived from a hybridization event between
277 TcII and TcIII (Westenberger et al. 2005; Tomasini and Diosque 2015), we propose that the TcMUCI+*HP
278 amplification is an ancient event, occurring after the split of TcI and TcII but prior to the TcII/TcIII
279 hybridization that yielded TcVI.

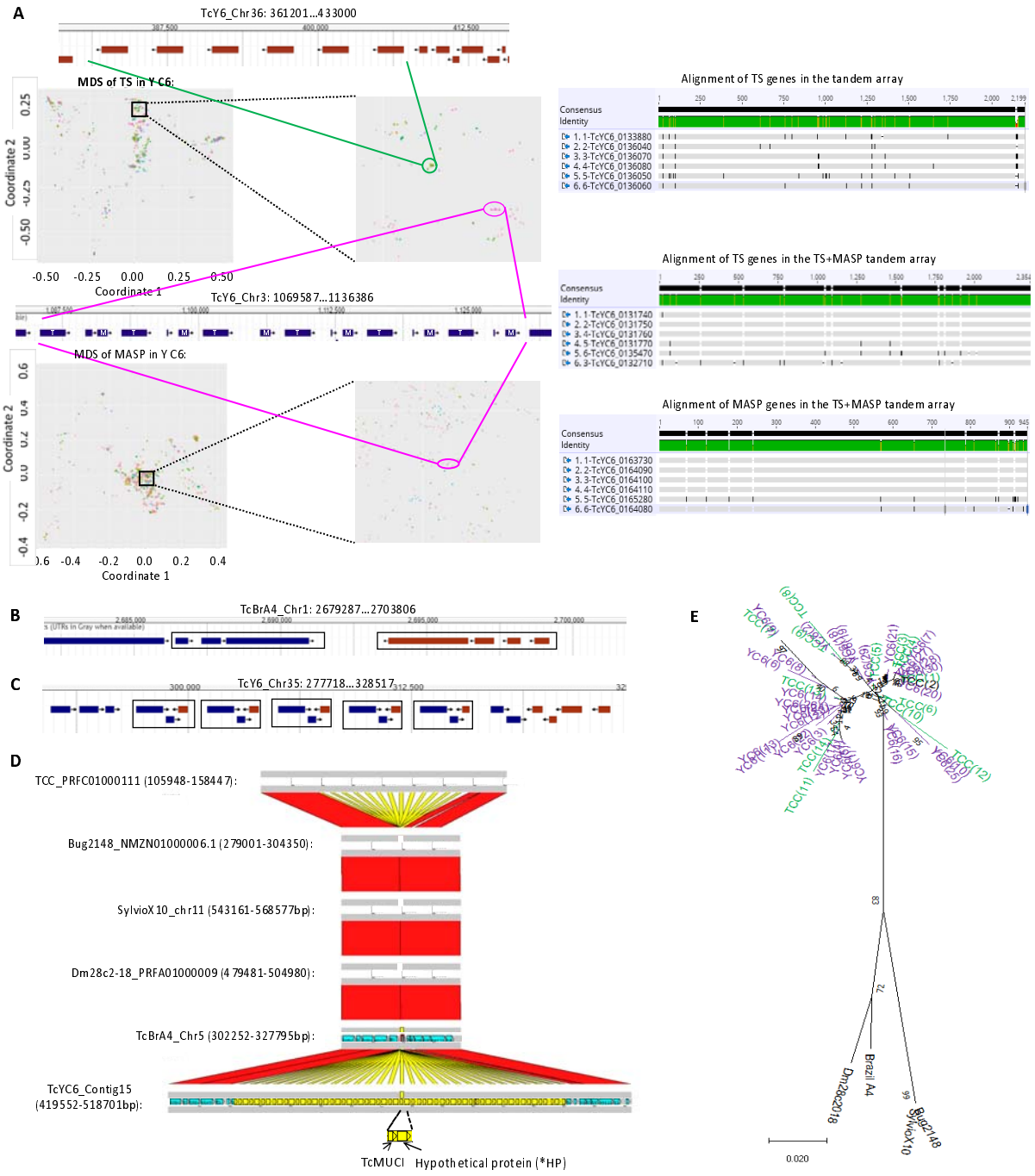


Figure 4. Gene amplification events in members of large gene families. (A) Tandem arrays of individual TS genes (top), and a TS+MASP pair (bottom) clustered based upon genetic distance in the MDS plots. Each chromosome is displayed as a separate pattern on the MDS plot. T: TS; M: MASP. Alignment of the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

286 one fragmented glycosyltransferase (-) replicated 5 times, creating multiple strand switches. (D) Syntenic regions
287 of the TcMUCI+*HP tandem array detected in 6 long-read sequenced *T. cruzi* strains. Synteny of TcMUCI orthologs
288 are labeled in yellow. (E) Phylogenetic tree of all TcMUCI orthologs from the 6 strains. Note that TcMUCI genes
289 from Y C6 (purple) and TCC (green) are intermingled in the top portion of the tree. Live MDS plots can be explored
290 at http://shiny.ctegd.uga.edu.
291
292 In addition to tandem clusters of gene family members, MDS analysis also revealed closely related gene
293 family members located on multiple chromosomes (Fig. 5A). An extreme case is the Y C6 gene cluster in
294 the bottom right of Fig. 5A which contained 31 TS genes with very high similarity distributed on 11
295 different chromosomes (Supplemental Fig. S8A). Interestingly, the 13 TS on Chr7 (Fig. 5B, middle) are in
296 tandem, interspersed with a beta tubulin gene, while the 7 TS on Chr18 (Fig. 5B, top) are in tandem as
297 TS genes alone (with one beta tubulin gene downstream of TS_7). The remaining 11 TS genes in this
298 cluster are dispersed in the genome as singlets (3 of them are shown at the bottom of Fig. 5B). Notably,
299 the sequences upstream and downstream of the TS gene in the TS + beta tubulin array on Chr7 are
300 homologous to those of the TS_7 gene in the Chr18 array, and the dispersed singlet TS also share a
301 portion of the upstream and downstream sequences with the other TS in this cluster (Supplemental Fig.
302 S8B). Together, these results suggest that all 31 TS genes in this cluster originated from one or more
303 gene amplification/relocation events. Based on the phylogenetic analysis (Supplemental Fig. S8C), we
304 propose that the TS + beta tubulin tandem copies have been generated in or relocated to Chr7 (13
305 copies) and Chr18 (1 copy), with another 4 TS copies as single genes beyond the TS + beta tubulin
306 cassette on Chr18, while the single TS genes on other chromosomes may derive from TS on Chr7.
307

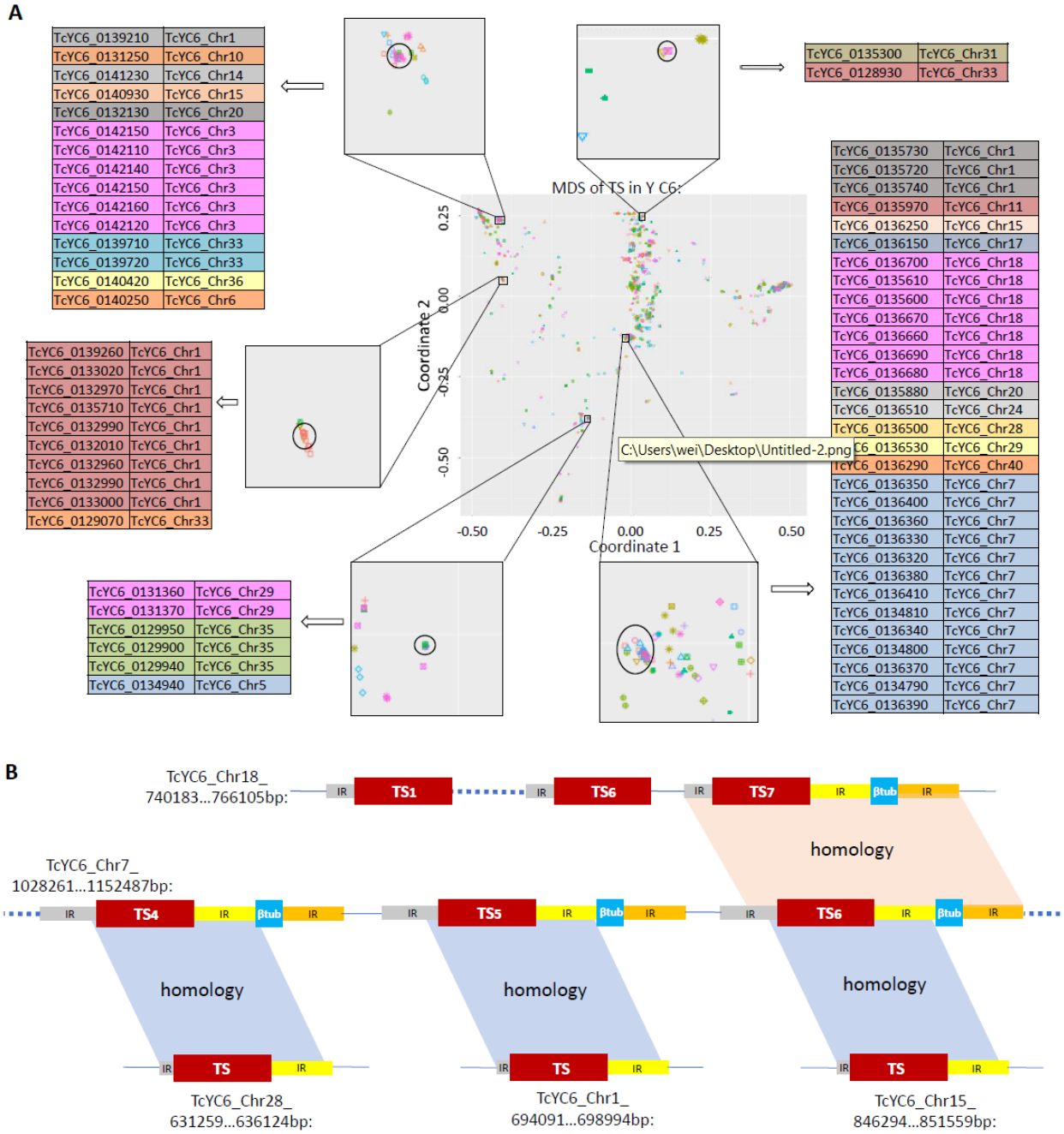


Figure 5. Examples of relocations of TS genes in Y C6. (A) Tight clusters of TS genes from MDS plot are distributed on different chromosomes. (B) Diagram of relocations in one of the TS clusters on the bottom right in (A). Blocks in the same color indicate genes or flanking sequences in high identity.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

314

315 We next used a pipeline previously designed to identify recombination events within TS genes in the CL
316 Brener genome (Weatherly et al. 2016), to quantify recombination for 4 of the large gene families in the
317 Brazil and Y strains (Table 2). As expected, recombination events, including multiple events acting on
318 the same gene, were detected in a large fraction of the genes but were particularly abundant (2-fold
319 higher) in the TS family relative to the other three families examined. Interestingly, recombination
320 events in the TS family were detected at a roughly 2-fold higher frequency in the Brazil strain as
321 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

322

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

324

325 As noted previously, our recombination pipeline is highly conservative in detecting relatively recent
326 events that have not been obscured by subsequent accumulation of SNPs and Indels (Weatherly et al.

327 2016). Such *in situ* diversification is evident in genes that are clustered in the MDS analysis but
328 dispersed in the genome. An example is a cluster of GP63 genes in Brazil A4 which have low genetic
329 distance based on MDS analysis (Supplemental Fig. S9A), but are located on different chromosomes and
330 display a considerable degree of variation (SNPs and Indels; Supplemental Fig. S9B). However, because
331 these genes also share similar upstream genes (a TS) and intergenic regions, all of these dispersed genes
332 were likely derived via gene duplication. This hypothesis is further supported by the result that 9 out of
333 10 GP63 genes and their corresponding GP63 + flanking sequences (including upstream TS + intergenic
334 region + GP63 + intergenic region) occupy identical positions in their respective phylogenetic trees
335 (Supplemental Fig. S9C). Therefore, a TS/GP63 gene pair and associated intergenic regions underwent
336 one or more duplication and relocation events with subsequent diversification through the
337 accumulation of SNPs and Indels, yielding multiple, diverse genes spread through the genome.
338
339 The potential complexity generated by amplification, relocation, recombination and diversification make
340 it challenging to track the specific set of events contributing to the evolution of individual gene family
341 members in *T. cruzi.* However, some gene sets reveal all of these processes at work. Fig. 6C shows four
342 cassettes located on different chromosomes or in distant sites on the same chromosome, each cassette
343 with a central MASP and flanking region with high identity (Fig. 6A and B), suggesting a common origin.
344 SNPs/Indels indicate *in situ* diversification of the MASP genes, especially in the C terminus (Fig. 6B).
345 Cassette pairs I/II and III/IV share the same upstream gene and flanking sequence (mucin genes in both
346 cases) while cassette pairs I /III and II/IV shared downstream mucin and GP63 genes, respectively. In
347 addition, a recombination event was detected in the C terminus of the GP63 in cassette IV, creating
348 divergence from the GP63 C terminus in cassette IV.
349

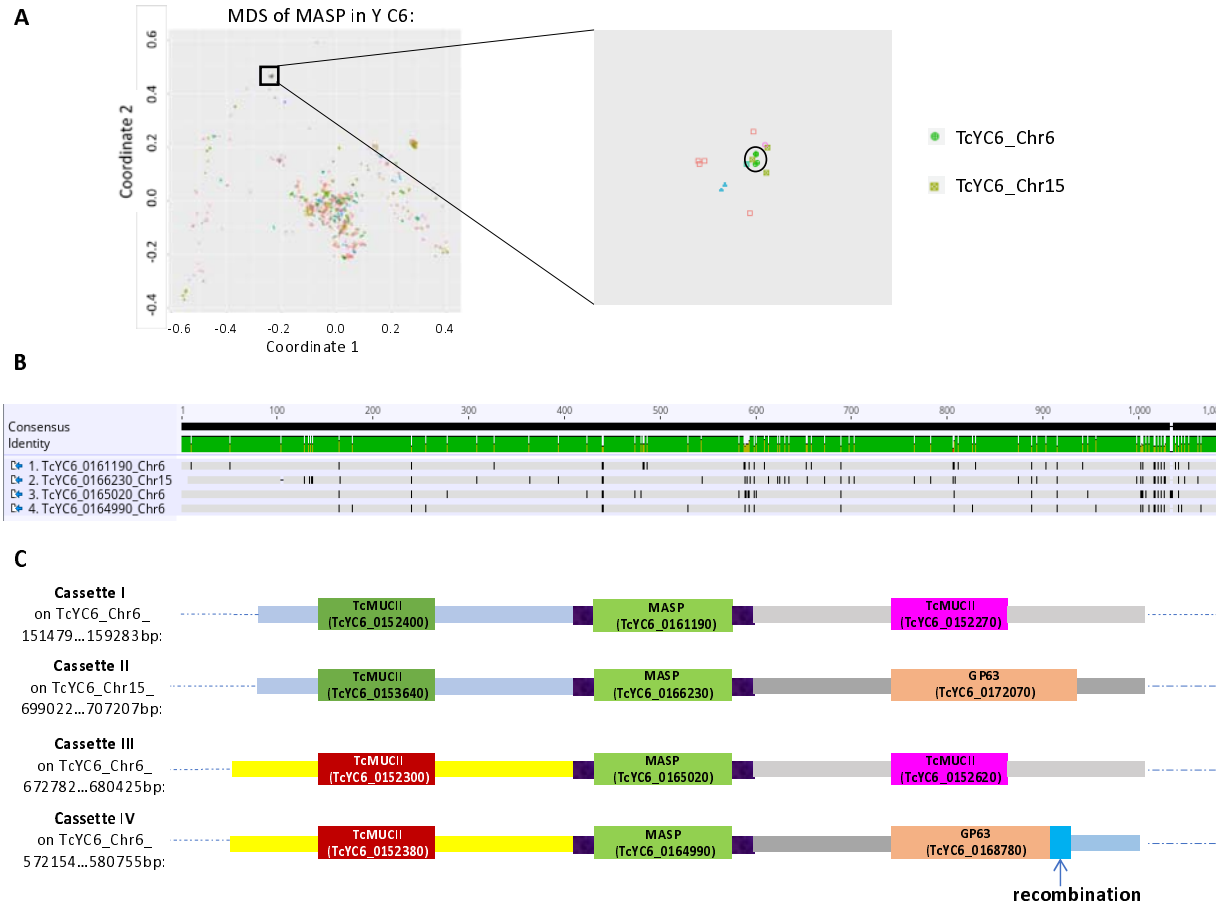


Figure 6. The combination of gene amlification, relocation, recombination and *in situ* diversification of large gene family members. (A) A tight cluster of 4 MASP genes from MDS plot are distributed on two chromosomes. (B) Alignment of the 4 MASP genes shows high identity with modest diversification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

386 than ¼ of that (29.3 kb in Brazil A4 and 33.8 kb in Y C6), indicating a disruption of the normal PTU
387 structure. Interestingly, amplified but conserved tandem gene arrays like the 'mucin + *HP' array in Y C6
388 discussed above (Fig. 4D) and the previously described TcSMUG family (Yoshida 2006; Nakayasu et al.
389 2009; Gonzalez et al. 2013) are within large PTUs containing almost no gene family members
390 (Supplemental Fig. S11A) while many other tandem arrays or apparently diverging genes reside in gene-
391 family-rich, short PTUs (Supplemental Fig. S11B). The disruption in PTU structure might also hamper the
392 preservation of transcriptional control mechanisms, in particular the tight controls on transcriptional
393 termination characterized in other kinetoplastids and mediated by base J and histone H3/4 variants
394 (Siegel et al. 2009; Cliffe et al. 2010; van Luenen et al. 2012; Reynolds et al. 2014; Reynolds et al. 2016;
395 Schulz et al. 2016; Kawasaki et al. 2017; Muller et al. 2018). To address this question, we mapped
396 strand-specific RNA-seq reads to both sense and antisense strands to assess transcriptional termination
397 relative to PTUs. Surprisingly, we found extensive antisense RNA levels throughout the genome (an
398 average of sense:antisense=114:1 in Brazil A4 and 84:1 in Y C6). Higher levels of antisense RNAs
399 occurred at the strand switch regions of long PTUs (Fig.7A and B), but in some cases, matched or
400 exceeded the sense strand transcripts in gene-family-enriched regions containing shorter PTUs (Fig. 7C).
401 Thus, unlike *T. brucei* and *Leishmania*, *T. cruzi* does not appear to regulate antisense RNA production so
402 tightly.
403

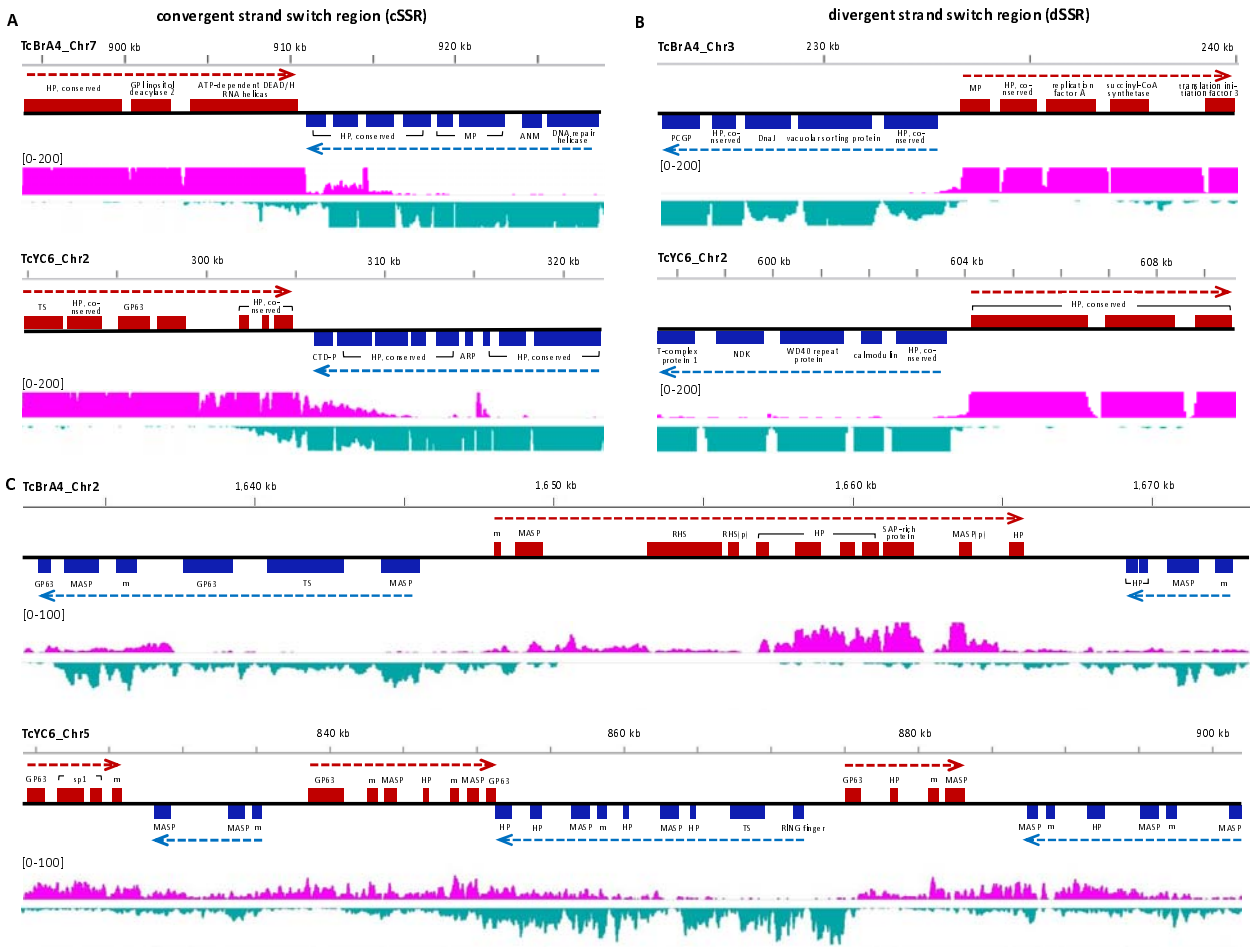


Figure 7. Antisense RNA levels in *T. cruzi* in relation to PTU structure, including both convergent strand switch regions (cSSR, A) and divergent strand switch regions (dSSR, B). (C) Gene-family-enriched regions with frequent strand switches where antisense RNA were detected in higher levels. HP, hypothetical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

415 *T. cruzi* is a highly heterogeneous specie, with at least six DTUs and with extreme variation in phenotype
416 and virulence among isolates even of the same DTU. Gaining new understanding of the genetic basis of
417 this high strain-to-strain variation in disease-causing potential in *T. cruzi* has been challenging due to the
418 lack of high-quality reference genome. The high content of repetitive sequences in the *T. cruzi* genome
419 (>50%), including multiple families of surface protein-encoding genes each with >200 members, makes
420 complete genome assembly from conventional short-read sequences impossible. This study reports very
421 high-quality genomes for *T. cruzi* strains belonging to the presumed ancestral lineages of this species,
422 TcI, represented here by the Brazil strain and TcII by the Y strain. This significantly improved resource
423 was achieved by the application of long-read sequencing techniques and proximity ligation libraries to
424 better resolve the full repertoires of gene content, thus allowing a detailed comparison of genetic
425 variation between these strains.
426
427 Although DTU-specific associations have been frequently proposed for characteristics such as virulence,
428 disease presentation, geographic distribution, and host species restrictions, many of these linkages
429 falter when more extensive sampling is done and none has been linked to DTU-specific genetic
430 differences (Revollo et al. 1998; Rassi et al. 2012; Nguyen and Waseem 2020). The current dataset
431 provides the opportunity to begin examination of representative strains of *T. cruzi* lineages that
432 diverged from each other an estimated 1-3 million years ago (Tomasini and Diosque 2015). The most
433 surprising revelations from this comparative analysis were not the variability in unique gene content
434 between these isolates, but rather the extremes of the high similarity in core gene content and the
435 comparative huge diversity in gene family-rich portions of the genomes. As anticipated based upon
436 previous strain-based screens (Ackermann et al. 2009) (Reis-Cunha et al. 2015), a considerable degree of
437 variation exists in the form of SNPs/Indels and additionally, a substantial number of strain-specific copy
438 number differences were identified. However, the core (non-gene family) genome, contains only ~20

439 strain-unique gene models, and in all cases, these are hypothetical genes encoding proteins with no
440 recognizable protein domain structures.
441
442 In very sharp contrast, the variation evident in the large gene families of *T. cruzi* is equally remarkable,
443 demonstrating vast diversity within and between strains with no perfect matches and relatively few
444 genes of the same family with even a 90% similarity. Structurally, these gene family members make up
445 ~25% of the genome and are spread widely throughout the genome, with some members on every
446 chromosome and some of the largest chromosomes being almost entirely composed of gene family
447 members. The use of synteny detection tools and Falcon-Phase validated by Hi-C methods allowed us to
448 also conservatively document heterozygosity in more than half of the chromosomes in each genome,
449 and we suspect that this heterozygosity extends to nearly all gene family-rich regions of the genome.
450 Based upon the total base count of the repeat-rich small scaffolds not assigned to chromosomes, we
451 estimate that up to 50% of all gene family members have variants on the sister chromosome.
452
453 The quality of the genome assemblies also provided the opportunity to document the continuing
454 diversification of these large gene families and to permit the beginning of an understanding of how this
455 process might work. Select members of the large gene families in *T. cruzi* have clear and critical
456 functions in parasite biology, with the best-documented example being the enzyme-active *trans*-
457 sialidases required for acquisition of sialic acid by *T. cruzi* trypomastigotes (Previato 1985) (Uemura
458 1992; Cremona et al. 1995) (Frasch 2000). However, the number, diversity, and potential for variation of
459 genes in these large gene families, and the exposure of the gene products to and response by the host
460 immune system, argue that these gene families evolve under intense immunological pressure. In this
461 respect, the three largest and most diverse gene families in *T. cruzi* (TS, MASP and mucin) are similar to
462 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

487 and mutations ultimately obscure the origins of new genes. Certainly the repeat-rich structure and
488 dense representation of retrotranposons of the *T. cruzi* genome facilitates maintenance of these
489 processes and the dispersion of gene family members throughout the genome, and interestingly not
490 restricted to chromosomes ends as is the case in *T. brucei* (El-Sayed et al. 2003) (Berriman et al. 2005)
491 (Hertz-Fowler et al. 2008) (Mugnier et al. 2016) (Muller et al. 2018). However, the specific structural
492 elements that initially established and continue to allow for these apparently constant rearrangements
493 throughout the genome but without impacting overall genome integrity, remain unidentified. From our
494 analysis, no consistent pattern of structures, such as the A/T tracks associated with gene application
495 events in *Plasmodium* (Huckaby et al. 2019) were evident.

496 The apparent high frequency and continued evolution of gene families in *T. cruzi* also create structures
497 and products unique among the kinetoplastids, including the lack of segregation of gene families to
498 chromosome ends, absence of partitioning of expression sites [as in *T. brucei* VSGs (Ersfeld et al. 1999)
499 Navarro and Gull 2001; (Navarro and Gull 2001; Hertz-Fowler et al. 2008)], tolerance for the generation
500 of short PTUs and frequent strand switching, and most surprisingly, the tolerance of antisense RNA
501 production. The latter may well explain the absence of the machinery for RNAi in *T. cruzi* (DaRocha et al.
502 2004) (Barnes et al. 2012). The presence of abundant and nearly genome-wide antisense RNAs also
503 suggests that *T. cruzi* does not adhere to the full set of rules for transcription termination as defined in
504 *T. brucei* and *Leishmania* (Reynolds et al. 2014) (Kieft et al. 2020).

505

506 Interestingly, there are several subsets of gene family members that appear to be exceptions to these
507 processes of recombination, diversification and distribution throughout the genome. The previously
508 characterized SMUG families are the best examples. Two subgroups of TcSMUG genes, TcSMUG L and S,
509 involved in development and infectivity of insect-dwelling stages

510 (Yoshida 2006; Nakayasu et al. 2009; Gonzalez et al. 2013), distribute as tandem arrays in the respective
511 subgroups within the same PTU and exhibit minimal diversification. Here we also identify an ancient,
512 lineage-specific duplication event that created a new hypothetical gene and a mucin gene in a tandem
513 array and which, like the SMUGS, has remained with minimal changes. It will be of interest to determine
514 if further diversification of this and other gene family subsets are restricted because of their location in
515 the genome, or if, like the SMUGS, this hypothetical gene/mucin tandem is under selective pressure due
516 to their unique function. One common feature of these tandem arrays is that they all locate in and are
517 flanked by large PTUs (>220 kb) containing only core genes with no members from the large gene
518 families (other than the mucins in the mucin+*HP array), suggesting that they are maintained in an
519 environment largely devoid of large-gene-family-related diversification.
520
521 An additional strain-dependent difference documented here is the higher recombination frequency in
522 Brazil A4 compared to Y and in CL Brener (Weatherly et al. 2016). The ~2X greater number of
523 recombination events in all gene families in Brazil vs Y suggests that this is an inherent property of this
524 strain and perhaps of DTUI strains in general. Alternatively, because we very conservatively call
525 recombination events which then eventually become concealed by further mutations/recombinations
526 over time, it is also possible that the Brazil A4 has been under stronger, or more recent, strong selective
527 pressure.
528
529 The apparent high levels of gene amplification/diversification readily documented in the large gene
530 families in this species also extends to a fraction of core genes as well, and represents a second major
531 source of between-strain diversity and perhaps the one primarily responsible for the broad between-
532 strain phenotypic variation in *T. cruzi*. Retention of these core gene amplifications imply a fitness

533 benefit, perhaps under certain environmental/host conditions; others may also occur regularly but
534 engender a fitness cost and thus are lost.
535
536 In summary, the careful analysis of these two *T. cruzi* strains soundly confirms the vast genetic diversity
537 of parasite lines within this species, and identifies the bulk of diversity to be represented in 3
538 compartments: 1) rapidly evolving families of genes involved in immune evasion, 2) a subset of "core"
539 genes not linked to evasion but which vary greatly in copy number and perhaps expression, and 3) SNPs
540 and Indels common to all genomes. We hypothesize that the gene family diversity is driven by immune
541 selection and that the same processes that provide for this diversity also allow for copy number
542 variation and diversification of select core genes, and this later process, rather than DTU type, accounts
543 for much of the biological diversity of *T. cruzi* lines. With these high quality genomes in hand for these
544 strains, we can now test these hypotheses by further modifying these gene sets and exposing both wild-
545 type and modified parasite lines to various levels of selection pressure and observing the genomes of
546 the lineages that emerge.
547
548
549 **Supplementary files**
550
551 **Supplemental Figure S1.** Overview of the Brazil A4 and Y C6 genomes. Tracks from outer to inner circles
552 indicate: sizes, chromosomes, gaps, gene density (window size: 20kb, range: 6-23 in Brazil A4, 1-22 for Y
553 C6), GC content (window size: 10kb, range: 0.36-0.70 in Brazil A4, 0.33-0.70 in Y C6), repetitive content
554 (window size: 10kb, range: 0-10000), heterozygous SNPs (window size: 20kb, range: 0-120 in Brazil A4,
555 1-390 in Y C6) and heterozygous Indels (window size: 20kb, range: 65-1 in Brazil A4, 129-1 in Y C6).

556 **Supplemental Figure S2.** Assembly improvement compared to CL Brener. (A) An example of filled gaps.
557 Syntenic regions between Chr1 in Brazil A4 and Chr8 in CL Brener were aligned with the Artemis
558 Comparison Tool (ACT) (Carver et al. 2005). All five gaps were filled in Brazil A4. (B) An example of
559 recovered genes. Two pieces of an adenosine monophosphate (AMP) gene were identified flanking a
560 gap, while the syntenic region in Brazil A4 shows the intact AMP gene. (C) An example of extended
561 repeats. With 8 copies of histone H4 in Chr2 of CL Brener separated by a gap, the syntenic region of
562 Brazil had the gap filled, extending the copy number of histone H4 to 41. Blocks: gaps; green bars: genes.
563 **Supplemental Figure S3.** Repetitive composition in the scaffolds. Chromosomes are calculated
564 individually, while small scaffolds are calculated by averaging a range of scaffolds as indicated on the x
565 axis.
566 **Supplemental Figure S4.** Workflow of predicting full repertoire of large gene families (taking TS as an
567 example).
568 **Supplemental Figure S5.** Distribution of large gene families and retrotransposons on the chromosomes.
569 Rings from outer to inner: chromosomes, retrotransposons, TS, MASP, mucin, GP63, RHS and DGF-1
570 gene families.
571 **Supplemental Figure S6.** Size distribution of hypothetical proteins identified in kinetoplastids. Genomes
572 of *T. brucei* TRE92 and *L. major* Friedlin were downloaded from TritrypDB database
573 (https://tritrypdb.org/tritrypdb/) release-44 (Aslett et al. 2010).
574 **Supplemental Figure S7.** Analysis of chromosome copy number. (A) Relative read depth of each
575 chromosome normalized to the mean read depth of all chromosomes at non-repetitive regions.
576 Chromosomes with more than two copies was indicated in red. (B) Allele frequency calculated by the
577 proportion of heterozygous SNPs/Indels at the non-repetitive regions of each chromosome. A diploid
578 chromosome showed the peak of allele frequency around 50% as shown in chr8 and chr31, whereas a
579 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 TcMUCI+*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 segmental duplications of large gene families identified in Brazil A4

625 and Y C6.

626

627

628 ## Methods

629

630 **Parasite cultures, DNA/RNA extraction and sequencing**

631 Epimastigotes of Brazil and Y were cultured at 26°C in supplemented liver digested- neutralized tryptose
632 (LDNT) medium as described previously (Xu et al. 2009). Single-cell clones were made for each strain by
633 depositing epimastigotes into a 96-well plate at a density of 0.5 cell/well by using a MoFlow cell sorter
634 (Dako-Cytomation, Denmark). One healthy clone that has confirmed to have cycled through all life
635 stages was chosen for sequencing for each strain. High molecular weight DNA was isolated using
636 MagAttract HMW DNA kit (Qiagen) before submitting to Duke Center for Genomic and Computational
637 Biology (GCB) for SMRT sequencing. Brazil A4 was sequenced using PacBio RS II sequencer, while PacBio
638 Sequel sequencer was used for Y C6.

639

640 Genomic DNA of the selected clone of both strains was isolated using QIAamp DNA blood mini kit
641 (Qiagen) for whole genome sequencing using Illumina HiSeq 150 PE. An RNase treatment step was
642 included to eliminate RNA in the samples. For RNA-seq sampling, extracellular amastigotes and
643 trypomastigotes isolated from infected Vero cells were pooled with epimastigotes for total RNA-
644 extraction. Following ribo-depleted RNA library construction and RNA sequencing using Illumina Nextseq
645 75PE was performed by Georgia Genomics and Bioinformatics Core (GGBC). Illumina reads from either
646 DNA or RNA sequencing with mean quality lower than 30 (Phred Score based) were removed for
647 analysis.

648 **Genome assembly**

649 The draft genome of Brazil A4 was assembled with SMRT Link v3.1, and Y C6 with SMRT Link v5.0. The
650 parameters were set at default except the expected genome size, which was set at 40 Mb for both
651 strains. Chicago and Hi-C libraries were constructed and sequenced by Dovetail Genomics, and HiRise
652 pipeline was run for scaffolding the draft assembly by incorporating data from both libraries. A gap was
653 generated whenever two contigs were joined or one contig was broken by HiRise and since the distance
654 between two contigs was unknown, all gaps were given 100 Ns.
655
656 Gap filling was performed by PBJelly (English et al. 2012) using the SMRT subreads with the minimum
657 percent identity at 85%. 122 and 4 gaps were extended for Brazil and Y, respectively. Correction of the
658 genomes using Illumina short reads was run by Pilon (Walker et al. 2014) and iCORN2 (Otto et al. 2010)
659 through multiple iterations to eliminate errors from SMRT sequencing.
660
661 **Repeat annotation**
662 RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler) was used to build a *de novo*
663 repeats library, and then used RepeatMasker v 4.0.7 (http://www.repeatmasker.org) with search engine
664 parameter as "ncbi".
665
666 **Genome annotation**
667 To develop open reading frame (ORF) in the new genome sequences, WebApollo 2.0 (Lee et al. 2013)
668 was deployed with the genome sequence and the following tracks of evidence were added:
669 1. Gene prediction from COMPANION (Steinbiss et al. 2016) using *Trypanosoma brucei* as
670 reference.
671 2. Gene prediction using AUGUSTUS (Stanke et al. 2004; Stanke and Morgenstern 2005) which was
672 self-trained by CL Brener genome.

673 3. Annotation transfer from CL Brener by Exonerate (Slater and Birney 2005).

674 4. ESTs from available EST sequencing libraries in *T. cruzi* (retrieved from

675 https://tritrypdb.org/tritrypdb/app/record/dataset/DS_6889a51dab).

676 5. Proteins from available Mass spec data for *T. cruzi* (Queiroz et al. 2013).

677 6. Strand-specific RNA-seq alignment data, the pipeline of which was followed as previously

678 described (Kieft et al. 2020).

679

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

681

682 InterProScan v5.31-70.0 (Jones et al. 2014) was used to detect protein families, domains and sites with

683 all 11 default databases. Gene Ontology (GO) term was assigned also by InterProScan based on the

684 protein domains results. Besides, BLASTP was used to search protein homology against *T. cruzi* CL

685 Brener, *T. brucei*, *Leishmania major* databases from TriTrypDB release 39

686 (https://tritrypdb.org/tritrypdb/) and RefSeq non-redundant protein database, respectively, to

687 determine the best hit for protein naming by in-house scripts. The parameter used for BLASTP was

688 evalue <1e-10, identity >70% and coverage (length of alignment/length of target protein) >70%.

689 Predicted pseudogenes were named by homology in RefSeq non-redundant nucleotide database with

690 evalue <1e-30.

691

692 **Annotation of large gene families**

693 A customized computational pipeline automated using PERL and Python scripts were developed for

694 identifying members of large gene families in the genome. First, the annotated members of each gene

695 family were searched against the Brazil A4 and Y C6 genome using BLASTN (version ncbi-blast-2.8.1+)

696 with num_alignments and max_hsps arguments set to 100, the perc_identity argument set to 85. BLAST

697 hits that have an overlap longer than 100 bp were merged if they match members from the same gene
698 family. BLAST hits that were bracketed by longer hits from the same family were removed. A minimum
699 length cutoff of 150 bp was applied to the BLAST hits. The remaining BLAST hits were considered new
700 family member gene candidates. The new candidate genes were BLASTNed against all annotated
701 transcripts from the genome of *T. cruzi* CL Brener strain (TriTrypDB release 34) (BLAST argument
702 settings: num_alignments and max_hsps set to 50, perc_identity not set). Candidate genes were
703 retained only if one of its top two best matches is a member of the candidate gene's corresponding gene
704 family.
705
706 Next, the boundaries of the candidate genes were refined by using model genes of each family in two
707 steps. (1) Extending the candidate gene boundaries to include possible segments missed by previous
708 steps. Using model gene sequences of each family to search the new genomes, and compare the
709 coordinates of the matches to that of candidate genes. If > 50% overlap was found, and the non-
710 overlapping length was < 1,000 bp, then the boundary of the candidate was extended according to the
711 genomic match of the model gene. (2) The boundary of candidate genes was next subjected to small-
712 scale trimmings. The candidate genes were BLASTed against model genes of the corresponding gene
713 family (num_alignments and max_hsps arguments set to 100, the perc_identity argument set to 85). If a
714 match was found within 100bp distance to the boundary of candidate genes, the candidate gene
715 boundary was trimmed to match that of the model gene.
716
717 The start of mucin candidate genes was further refined using a conserved signal peptide sequence (in an
718 alignment format allowing for minor variations). The signal peptide sequences were BLASTNed against
719 mucin candidate genes (BLAST argument settings: num_alignments and max_hsps set to 200,

720 perc_identity set to 65, gapopen and gapextend set to 1). Sequence upstream of signal peptide matches
721 in the candidate genes was removed.
722
723 A final trim was applied to the boundaries of all candidate genes, as many of our BLAST steps could lead
724 to inaccurate boundary identification due to 25% chance of a random matching an extra nucleotide base
725 at the boundary and 6.25% chance for two extra bases and so forth, which could obscure start and stop
726 codons. As an attempt to address this issue, we trimmed up to 10 bases which could reveal a start/stop
727 codon that is in-frame with an existing stop/start codon.
728
729 Manual corrections of boundaries for gene family members were performed when necessary.
730
731 **Hi-C contact matrix**
732 Hi-C contact matrix were analyzed by following the manual of https://github.com/hms-dbmi/hic-data-
733 analysis-bootcamp/, and then visualized in HiGlass (Kerpedjiev et al. 2018).
734
735 **Multidimensional scaling**
736 K-tuple distance between genes are calculated with Clustal-Omega 1.2.4 (Sievers et al. 2011) using
737 unaligned sequences with option parameters: “--full" and “--distmat-out". Full alignment distance
738 between genes are calculated with Clustal-Omega 1.2.4 using aligned sequences (aligned with Clustal-
739 Omega 1.2.4 using default parameters) with options parameters: --full --full-iter --distmat-out. MDS is
740 performed with the “cmdscale” function built-in R 3.6.3 with the input of a matrix of either pairwise K-
741 tuple distances or full alignment distances. The results of MDS are visualized using the Shiny package
742 1.4.0.2 in R 3.6.3.
743

744 **Phylogenetic inference**

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

751

752 ## Data access

753

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

761

762 ## Acknowledgements

763

764 We thank Dr. Todd Minning for initial contributions this project, Dr. Robert Sabatini from the University
765 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. This work was supported by funding 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.

802 Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: the Artemis
803 Comparison Tool. *Bioinformatics* **21**: 3422-3423.
804 Clayton C. 2019. Regulation of gene expression in trypanosomatids: living with polycistronic
805 transcription. *Open Biol* **9**: 190072.
806 Cliffe LJ, Siegel TN, Marshall M, Cross GA, Sabatini R. 2010. Two thymidine hydroxylases differentially
807 regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic
808 transcription units throughout the genome of *Trypanosoma brucei*. *Nucleic Acids Res* **38**: 3923-
809 3935.
810 Cremona ML, Sanchez DO, Frasch AC, Campetella O. 1995. A single tyrosine differentiates active and
811 inactive *Trypanosoma cruzi trans*-sialidases. *Gene* **160**: 123-128.
812 Cross GA. 1975. Identification, purification and properties of clone-specific glycoprotein antigens
813 constituting the surface coat of *Trypanosoma brucei*. *Parasitology* **71**: 393-417.
814 DaRocha WD, Otsu K, Teixeira SM, Donelson JE. 2004. Tests of cytoplasmic RNA interference (RNAi) and
815 construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. *Mol Biochem*
816 *Parasitol* **133**: 175-186.
817 de Freitas JM, Augusto-Pinto L, Pimenta JR, Bastos-Rodrigues L, Goncalves VF, Teixeira SM, Chiari E,
818 Junqueira AC, Fernandes O, Macedo AM et al. 2006. Ancestral genomes, sex, and the population
819 structure of *Trypanosoma cruzi*. *PLoS Pathog* **2**: e24.
820 De Pablos LM, Osuna A. 2012. Multigene families in *Trypanosoma cruzi* and their role in infectivity. *Infect*
821 *Immun* **80**: 2258-2264.
822 Diaz-Viraque F, Pita S, Greif G, de Souza RCM, Iraola G, Robello C. 2019. Nanopore sequencing
823 significantly improves genome assembly of the protozoan parasite *Trypanosoma cruzi*. *Genome*
824 *Biol Evol* **11**: 1952–1957.
825 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic*
826 *Acids Res* **32**: 1792-1797.
827 El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, Wanless D, Peterson J, Hou L, Taylor S
828 et al. 2003. The sequence and analysis of *Trypanosoma brucei* chromosome II. *Nucleic Acids Res*
829 **31**: 4856-4863.
830 El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher
831 AL, Blandin G et al. 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas
832 disease. *Science* **309**: 409-415.
833 Elbers JP, Rogers MF, Perelman PL, Proskuryakova AA, Serdyukova NA, Johnson WE, Horin P, Corander J,
834 Murphy D, Burger PA. 2019. Improving Illumina assemblies with Hi-C and long reads: An
835 example with the North African dromedary. *Mol Ecol Resour* **19**: 1015-1026.
836 English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012.
837 Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.
838 *PLoS One* **7**: e47768.
839 Ersfeld K, Melville SE, Gull K. 1999. Nuclear and genome organization of *Trypanosoma brucei*. *Parasitol*
840 *Today* **15**: 58-63.
841 Flores-Lopez CA, Machado CA. 2011. Analyses of 32 loci clarify phylogenetic relationships among
842 *Trypanosoma cruzi* lineages and support a single hybridization prior to human contact. *PLoS*
843 *Negl Trop Dis* **5**: e1272.
844 Frasch AC. 2000. Functional diversity in the *trans*-sialidase and mucin families in *Trypanosoma cruzi*.
845 *Parasitol Today* **16**: 282-286.
846 Gonzalez MS, Souza MS, Garcia ES, Nogueira NF, Mello CB, Canepa GE, Bertotti S, Durante IM, Azambuja
847 P, Buscaglia CA. 2013. *Trypanosoma cruzi* TcSMUG L-surface mucins promote development and
848 infectivity in the triatomine vector *Rhodnius prolixus*. *PLoS Negl Trop Dis* **7**: e2552.

849 Henriksson J AL, Macina RA, Franke de Cazzulo BM, Cazzulo JJ, Frasch AC, Pettersson U. 1990.
850 Chromosomal localization of seven cloned antigen genes provides evidence of diploidy and
851 further demonstration of karyotype variability in *Trypanosoma cruzi*. *Mol Biochem Parasitol* **42**:
852 213–223.
853 Henriksson J, Dujardin JC, Barnabe C, Brisse S, Timperman G, Venegas J, Pettersson U, Tibayrenc M,
854 Solari A. 2002. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is
855 evolutionarily informative. *Parasitology* **124**: 277-286.
856 Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, Cazzulo JJ, Frasch AC, Pettersson U. 1995.
857 Chromosome specific markers reveal conserved linkage groups in spite of extensive
858 chromosomal size variation in *Trypanosoma cruzi*. *Mol Biochem Parasitol* **73**: 63-74.
859 Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, Brooks K, Churcher C, Fahkro S,
860 Goodhead I et al. 2008. Telomeric expression sites are highly conserved in *Trypanosoma brucei*.
861 *PLoS One* **3**: e3527.
862 Huckaby AC, Granum CS, Carey MA, Szlachta K, Al-Barghouthi B, Wang YH, Guler JL. 2019. Complex DNA
863 structures trigger copy number variation across the *Plasmodium falciparum* genome. *Nucleic*
864 *Acids Res* **47**: 1615-1627.
865 Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A,
866 Melville SE et al. 2010. The genome sequence of *Trypanosoma brucei* gambiense, causative
867 agent of chronic human african trypanosomiasis. *PLoS Negl Trop Dis* **4**: e658.
868 Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al.
869 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-
870 1240.
871 Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from *in vivo* DNA interaction frequency.
872 *Nat Biotechnol* **31**: 1143-1147.
873 Kawasaki F, Beraldi D, Hardisty RE, McInroy GR, van Delft P, Balasubramanian S. 2017. Genome-wide
874 mapping of 5-hydroxymethyluracil in the eukaryote parasite *Leishmania*. *Genome Biol* **18**: 23.
875 Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A,
876 Kumar N et al. 2018. HiGlass: web-based visual exploration and analysis of genome interaction
877 maps. *Genome Biol* **19**: 125.
878 Kieft R, Zhang Y, Marand AP, Moran JD, Bridger R, Wells L, Schmitz RJ, Sabatini R. 2020. Identification of
879 a novel base J binding protein complex involved in RNA polymerase II transcription termination
880 in trypanosomes. *PLoS Genet* **16**: e1008390.
881 Korbel JO, Lee C. 2013. Genome assembly and haplotyping with Hi-C. *Nat Biotechnol* **31**: 1099-1101.
882 Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE.
883 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14**: R93.
884 Lewis EB. 1951. Pseudoallelism and gene evolution. *Cold Spring Harb Symp Quant Biol* **16**: 159-174.
885 Lima FM, Souza RT, Santori FR, Santos MF, Cortez DR, Barros RM, Cano MI, Valadares HM, Macedo AM,
886 Mortara RA et al. 2013. Interclonal variations in the molecular karyotype of *Trypanosoma cruzi*:
887 chromosome rearrangements in a single cell-derived clone of the G strain. *PLoS One* **8**: e63738.
888 Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:
889 1151-1155.
890 Martin DL, Weatherly DB, Laucella SA, Cabinian MA, Crim MT, Sullivan S, Heiges M, Craven SH,
891 Rosenberg CS, Collins MH et al. 2006. CD8+ T-Cell responses to *Trypanosoma cruzi* are highly
892 focused on strain-variant *trans*-sialidase epitopes. *PLoS Pathog* **2**: e77.
893 Mowatt MR, Aggarwal A, Nash TE. 1991. Carboxy-terminal sequence conservation among variant-
894 specific surface proteins of *Giardia lamblia*. *Mol Biochem Parasitol* **49**: 215-227.
895 Mugnier MR, Stebbins CE, Papavasiliou FN. 2016. Masters of Disguise: Antigenic Variation and the VSG
896 Coat in *Trypanosoma brucei*. *PLoS Pathog* **12**: e1005784.

897 Muller HJ. 1936. Bar Duplication. *Science* **83**: 528-530.
898 Muller LSM, Cosentino RO, Forstner KU, Guizetti J, Wedel C, Kaplan N, Janzen CJ, Arampatzi P, Vogel J,
899 Steinbiss S et al. 2018. Genome organization and DNA accessibility control antigenic variation in
900 trypanosomes. *Nature* **563**: 121-125.
901 Nakayasu ES, Yashunsky DV, Nohara LL, Torrecilhas AC, Nikolaev AV, Almeida IC. 2009. GPlomics: global
902 analysis of glycosylphosphatidylinositol-anchored molecules of *Trypanosoma cruzi*. *Mol Syst Biol*
903 **5**: 261.
904 Navarro M, Gull K. 2001. A pol I transcriptional body associated with VSG mono-allelic expression in
905 *Trypanosoma brucei*. *Nature* **414**: 759-763.
906 Nguyen T, Waseem M. 2020. Chagas Disease (American Trypanosomiasis). In *StatPearls*, Treasure Island
907 (FL).
908 Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides
909 (iCORN) using second generation sequencing technology. *Bioinformatics* **26**: 1704-1707.
910 Pedroso A, Cupolillo E, Zingales B. 2003. Evaluation of *Trypanosoma cruzi* hybrid stocks based on
911 chromosomal size variation. *Mol Biochem Parasitol* **129**: 79-90.
912 Pimenta PF, da Silva PP, Nash T. 1991. Variant surface antigens of *Giardia lamblia* are associated with
913 the presence of a thick cell coat: thin section and label fracture immunocytochemistry survey.
914 *Infect Immun* **59**: 3989-3996.
915 Previato J, Andrade AFB, Pessolani MCV, and Mendonça-Previato L. 1985. Incorporation of sialic acid
916 into *Trypanosoma cruzi* macromolecules. A proposal for a new metabolic route. *Mol Biochem*
917 *Parasitol* **16**: 85–96.
918 Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet
919 CW et al. 2016. Chromosome-scale shotgun assembly using an *in vitro* method for long-range
920 linkage. *Genome Res* **26**: 342-350.
921 Queiroz RM, Charneau S, Motta FN, Santana JM, Roepstorff P, Ricart CA. 2013. Comprehensive
922 proteomic analysis of *Trypanosoma cruzi* epimastigote cell surface proteins by two
923 complementary methods. *J Proteome Res* **12**: 3255-3263.
924 Rassi A, Jr., Rassi A, Marcondes de Rezende J. 2012. American trypanosomiasis (Chagas disease). *Infect*
925 *Dis Clin North Am* **26**: 275-291.
926 Reis-Cunha JL, Rodrigues-Luiz GF, Valdivia HO, Baptista RP, Mendes TA, de Morais GL, Guedes R, Macedo
927 AM, Bern C, Gilman RH et al. 2015. Chromosomal copy number variation reveals differential
928 levels of genomic plasticity in distinct *Trypanosoma cruzi* strains. *BMC Genomics* **16**: 499.
929 Revollo S, Oury B, Laurent JP, Barnabe C, Quesney V, Carriere V, Noel S, Tibayrenc M. 1998.
930 *Trypanosoma cruzi*: impact of clonal evolution of the parasite on its biological and medical
931 properties. *Exp Parasitol* **89**: 30-39.
932 Reynolds D, Cliffe L, Forstner KU, Hon CC, Siegel TN, Sabatini R. 2014. Regulation of transcription
933 termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma*
934 *brucei*. *Nucleic Acids Res* **42**: 9717-9729.
935 Reynolds D, Hofmeister BT, Cliffe L, Alabady M, Siegel TN, Schmitz RJ, Sabatini R. 2016. Histone H3
936 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of
937 siRNA Loci in *Trypanosoma brucei*. *PLoS Genet* **12**: e1005758.
938 Robert D. Denton RSK, Jacob W. Malcom, Louis Du Preez, and John H. Malone. 2018. The African Bullfrog
939 (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex
940 chromosomes. *bioRxiv* doi: https://doi.org/10.1101/329847.
941 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA,
942 Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice
943 across a large model space. *Syst Biol* **61**: 539-542.

944 Salter JF, Johnson O, Stafford NJ, 3rd, Herrin WF, Jr., Schilling D, Cedotal C, Brumfield RT, Faircloth BC.
945 2019. A Highly Contiguous Reference Genome for Northern Bobwhite (*Colinus virginianus*). *G3*
946 *(Bethesda)* **9**: 3929-3932.
947 Schreiber M, Mascher M, Wright J, Padmarasu S, Himmelbach A, Heavens D, Milne L, Clavijo BJ, Stein N,
948 Waugh R. 2020. A Genome Assembly of the Barley 'Transformation Reference' Cultivar Golden
949 Promise. *G3 (Bethesda)* **10**: 1823-1827.
950 Schulz D, Zaringhalam M, Papavasiliou FN, Kim HS. 2016. Base J and H3.V Regulate Transcriptional
951 Termination in *Trypanosoma brucei*. *PLoS Genet* **12**: e1005762.
952 Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, Wang X, Dewell S, Cross GA. 2009.
953 Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma*
954 *brucei*. *Genes Dev* **23**: 1063-1076.
955 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et
956 al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using
957 Clustal Omega. *Mol Syst Biol* **7**: 539.
958 Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC*
959 *Bioinformatics* **6**: 31.
960 Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, Pinches R, Newbold CI, Miller
961 LH. 1995. Switches in expression of Plasmodium falciparum *var* genes correlate with changes in
962 antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**: 101-110.
963 Souza RT, Lima FM, Barros RM, Cortez DR, Santos MF, Cordero EM, Ruiz JC, Goldenberg S, Teixeira MM,
964 da Silveira JF. 2011. Genome size, karyotype polymorphism and chromosomal evolution in
965 *Trypanosoma cruzi*. *PLoS One* **6**: e23042.
966 Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows
967 user-defined constraints. *Nucleic Acids Res* **33**: W465-467.
968 Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in
969 eukaryotes. *Nucleic Acids Res* **32**: W309-312.
970 Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD. 2016. Companion: a
971 web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* **44**: W29-34.
972 Sturtevant AH. 1925. The Effects of Unequal Crossing over at the Bar Locus in Drosophila. *Genetics* **10**:
973 117–147.
974 Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE. 1995.
975 The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic
976 variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82**: 89-100.
977 Theodore S. Kalbfleisch ESR, Michael S. DePriest Jr., Brian P. Walenz, Matthew S. Hestand, Joris R.
978 Vermeesch, Brendan L. O'Connell, Ian T. Fiddes, Alisa O. Vershinina, Jessica L. Petersen, Carrie J.
979 Finno, Rebecca R. Bellone, Molly E. McCue, Samantha A. Brooks, Ernest Bailey, Ludovic Orlando,
980 Richard E. Green, Donald C. Miller, Douglas F. Antczak, James N. MacLeod. 2018. EquCab3, an
981 Updated Reference Genome for the Domestic Horse. *bioRxiv* doi:
982 https://doi.org/10.1101/306928.
983 Tomasini N, Diosque P. 2015. Evolution of *Trypanosoma cruzi*: clarifying hybridisations, mitochondrial
984 introgressions and phylogenetic relationships between major lineages. *Mem Inst Oswaldo Cruz*
985 **110**: 403-413.
986 Triana O, Ortiz S, Dujardin JC, Solari A. 2006. *Trypanosoma cruzi*: variability of stocks from Colombia
987 determined by molecular karyotype and minicircle Southern blot analysis. *Exp Parasitol* **113**: 62-
988 66.
989 Uemura H, Schenkman, S., Nussenzweig, V., and Eichinger, D. 1992. Only some members of a gene
990 family in *Trypanosoma cruzi* encode proteins that express both *trans*-sialidase and
991 neuraminidase activities. *EMBO J* **11**: 3837–3844.

992 van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A,
993 Ramasamy G et al. 2012. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional
994 readthrough in *Leishmania*. *Cell* **150**: 909-921.
995 Vargas N, Pedroso A, Zingales B. 2004. Chromosomal polymorphism, gene synteny and genome size in *T.*
996 *cruzi* I and *T. cruzi* II groups. *Mol Biochem Parasitol* **138**: 131-141.
997 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young
998 SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and
999 genome assembly improvement. *PLoS One* **9**: e112963.
1000 Weatherly DB, Boehlke C, Tarleton RL. 2009. Chromosome level assembly of the hybrid *Trypanosoma*
1001 *cruzi* genome. *BMC Genomics* **10**: 255.
1002 Weatherly DB, Peng D, Tarleton RL. 2016. Recombination-driven generation of the largest pathogen
1003 repository of antigen variants in the protozoan *Trypanosoma cruzi*. *BMC Genomics* **17**: 729.
1004 Westenberger SJ, Barnabe C, Campbell DA, Sturm NR. 2005. Two hybridization events define the
1005 population structure of *Trypanosoma cruzi*. *Genetics* **171**: 527-543.
1006 Weston D, Patel B, Van Voorhis WC. 1999. Virulence in *Trypanosoma cruzi* infection correlates with the
1007 expression of a distinct family of sialidase superfamily genes. *Mol Biochem Parasitol* **98**: 105-
1008 116.
1009 Xu D, Brandan CP, Basombrio MA, Tarleton RL. 2009. Evaluation of high efficiency gene knockout
1010 strategies for *Trypanosoma cruzi*. *BMC Microbiol* **9**: 90.
1011 Yoshida N. 2006. Molecular basis of mammalian cell invasion by *Trypanosoma cruzi*. *An Acad Bras Cienc*
1012 **78**: 87-111.
1013 Zev N. Kronenberg RJH, Stefan Hiendleder, Timothy P. L. Smith, Shawn T. Sullivan, John L. Williams,
1014 Sarah B. Kingan. 2018. FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid
1015 genomes. *bioRxiv* doi: https://doi.org/10.1101/327064.
1016 Zingales B, Andrade SG, Briones MR, Campbell DA, Chiari E, Fernandes O, Guhl F, Lages-Silva E, Macedo
1017 AM, Machado CR et al. 2009. A new consensus for *Trypanosoma cruzi* intraspecific
1018 nomenclature: second revision meeting recommends TcI to TcVI. *Mem Inst Oswaldo Cruz* **104**:
1019 1051-1054.
1020 Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MM, Schijman AG, Llewellyn MS,
1021 Lages-Silva E, Machado CR et al. 2012. The revised *Trypanosoma cruzi* subspecific nomenclature:
1022 rationale, epidemiological relevance and research applications. *Infect Genet Evol* **12**: 240-253.
1023

1024