Infection, Genetics and Evolution 12 (2012) 1064–1071

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Phylogenetic evidence based on *Trypanosoma cruzi* nuclear gene sequences and information entropy suggest that inter-strain intragenic recombination is a basic mechanism underlying the allele diversity of hybrid strains

Renata C. Ferreira [a,b,c,*], Marcelo R.S. Briones [b,c]

[a] *Departamento de Medicina, Disciplina de Infectologia, Universidade Federal de São Paulo, Rua Botucatu 740, CEP 04023-900, São Paulo, SP, Brazil*
[b] *Departamento de Microbiologia, Imunologia e Parasitologia, Universidade Federal de São Paulo, Rua Botucatu 862, ECB 3 andar, CEP 04023-062, São Paulo, SP, Brazil*
[c] *Laboratório de Genômica Evolutiva e Biocomplexidade, Universidade Federal de São Paulo, Rua Pedro de Toledo 669, 4 andar, CEP 04039-032, São Paulo, SP, Brazil*

ARTICLE INFO

Article history:
Received 11 October 2011
Received in revised form 12 March 2012
Accepted 13 March 2012
Available online 23 March 2012

Keywords:
Trypanossoma cruzi
Intragenic recombination
Shannon entropy
Hybrid strains
Speciation

ABSTRACT

The diversity of *Trypanosoma cruzi* is categorized into six discrete typing units (DTUs) *T. cruzi* I to VI. Several studies indicate that *T. cruzi* I and II are ancestors of *T. cruzi* III–VI which are considered products of independent hybridization events. The individual haplotypes or alleles of these hybrids cluster in three groups, either closer to *T. cruzi* I or *T. cruzi* II or forming a midpoint clade between *T. cruzi* I and II in network phylogenies. To understand the origins of these different sets of haplotypes and test the hypothesis of a direct correlation between high entropy and positive selection, we analyzed four nuclear protein coding genes. We show that hybrid strains contain haplotypes that are mosaics probably originated by intragenic recombination. Accordingly, in phylogenies, the hybrid haplotypes are closer to one or both parentals (*T. cruzi* I and II) depending on the proportion of parental sequences composing the mosaics. In addition, Shannon entropy, used to measure sequence diversity, is highly correlated with positive selection in the four genes here analyzed. Our data on recombination patterns also support the hypothesis of two hybridization events in the hybrid structures of *T. cruzi* III–VI. Data presented and discussed here are consistent with a scenario where TcI and TcII are phylogenetically divergent forming a hybrid zone in between (*T.* cruzi III–VI). We predict that because of the quasi-random nature of *T. cruzi* I and II hybridization more DTUs, with different haplotype combinations, will be discovered in the hybrid zone.

1. Introduction

The protozoan *Trypanosoma cruzi* causes Chagas disease, still considered a public health problem in Latin America where 10 million people are infected (Rassi et al., 2010). Chagas disease has a variety of clinical manifestations affecting the heart and digestive tract although 30% of infected individuals are asymptomatic (reviewed by Macedo et al., 2004). *T. cruzi* has a significant genetic, biochemical and biological variability. Although the genetic diversity is geographically widespread throughout Latin America and in the same order of magnitude as observed between different genera of other trypanosomatids, it coalesces to two major groups or lineages, *T. cruzi* I (TcI) and *T. cruzi* II (TcII), based on a variety of makers such as random amplification of polymorphic DNA (RAPD) (Tibayrenc et al., 1993; Tibayrenc, 1995; Souto et al., 1996; Brisse et al., 1998); mini-exon gene sequences and 24S lsu rRNA sequences (Souto et al., 1996; Zingales et al., 1998; Fernandes et al., 1998); small subunit (SSU) rDNA (Briones et al., 1999; Tomazi et al., 2009); cytochrome b sequences (Brisse et al., 2003), nuclear locus (Dos Santos and Buck, 1999; Tomazi et al., 2009) and satellite DNA (Elias et al., 2005; Ienne et al., 2010). The divergence of these two groups is estimated between 88 and 37 million years ago, based on the SSU rDNA and between 16 and 3 million years ago, based on dihydrofolate reductase-thymidylate synthase (DHFR-TS) and trypanothione reductase (TR) genes (Briones et al., 1999; Kawashita et al., 2001; Machado and Ayala, 2001). Multilocus genotyping further revealed six distinct subgroups or discrete typing units (DTU) named I, IIa–IIe (Brisse et al., 2000). The current typing scheme follows an expert committee revision of nomenclature and recommended the use of *T. cruzi* I–VI to identify genotypes, or discrete typing units (Zingales et al., 2009). These DTUs correspond to the six zymodemes characterized by pioneering work of Miles and collaborators (Miles et al., 1981, 2003).

Although *T. cruzi* propagates mostly by clonal reproduction (Tibayrenc and Ayala, 1988), there is increasing evidence showing that genetic exchange between parasites may have contributed to the present population structure. This was evidenced by the discovery of hybrid strains in nature (reviewed by Sturm and

[*] Corresponding author at: Departamento de Microbiologia, Imunologia e Parasitologia, Universidade Federal de São Paulo, Rua Botucatu 862, ECB 3 andar, CEP 04023-062, São Paulo, SP, Brazil. Tel.: +55 11 5083 2980; fax: +55 11 5572 4711.
E-mail address: renata.carmona@unifesp.br (R.C. Ferreira).

1567-1348 © 2012 Elsevier B.V. Open access under the Elsevier OA license.
http://dx.doi.org/10.1016/j.meegid.2012.03.010

Campbell, 2010). Gaunt et al. (2003), by producing experimental hybrid showed a still existing capacity for genetic exchange and suggested the mechanism of nuclear fusion for the formation of hybrids. Analysis based on flow cytometry and microsatellite genotyping showed that the experimental hybrids are relatively stable sub-tetraploids while natural hybrids are largely diploid (Lewis et al., 2009).

Several lines of evidence indicate that TcI and TcII are "pure" lines with great phylogenetic divergence between them. Briones et al. (1999) proposed that *T. cruzi* I and II should be considered different species because of independent long term evolution, distinct ecological niches and epidemiological attributes and with unexpectedly high phylogenetic distance (rDNA) equivalent to the distance between genera *Crithidia* and *Endotrypanum*. The origins of the hybrids are more controversial and two scenarios have been proposed. In the first scenario the different hybrids are generated by two major independent hybridization events (Westenberger et al., 2005, 2006). The first event is the fusion of parental lineages TcI and TcII resulting in a TcI/TcII hybrid. This event was relatively ancient and TcIII and TcIV, originated by it, are homozygous for all non-satellite loci. A more recent back-cross between TcIII and its parental TcII generated the TcV and TcVI which still maintain allelic heterozygosity. The association between TcI and TcVI maxicircles coding regions is strong evidence that TcI provided the maxicircle to the progeny of TcIII and TcIV thus corroborating the two-hybridization event model and the hybris nature of TcIII (Ruvalcaba-Trejo and Sturm, 2011). The other scenario was postulated by De Freitas et al. (2006) based on microsatellite and multilocus mitochondrial genes. In this model three ancestral genotypes are proposed, TcI, TcII and TcIII. The fusion of parental lineages TcII and TcIII originated TcV and TcVI hybrids. The origin of TcIV was not addressed because only one strain from this DTU was used. Analysis using microsatellite (Venegas et al., 2009; Ienne et al., 2010) and nuclear loci (Subileau et al., 2009; Tomazi et al., 2009) supports the two-hybridization model.

A study based on molecular network phylogenetics of five nuclear genes where the two haplotypes of each hybrid strain were analyzed separately has shown that all hybrid genotypes contain two distinct types of sequence as opposed to TcI and TcII which exhibit highly similar haplotypes, invariably clustering in the same group (Tomazi et al., 2009). Because different genes, located at different chromosomes of the hybrid strains, always have one haplotype closer to TcII and the other closer to TcI or at an intermediate position in the genealogy, these authors proposed that the hybrids are polyphyletic and have different combinations of TcI and TcII haplotypes with a clear hybrid haplotype-specific group.

To understand the origins of the different *T. cruzi* hybrid haplotype groups we analyzed the same set the four coding genes alignments of Tomazi et al. (2009) using both phylogenetic (bootscan and dual multiple change-point model) and population genetic based tests (DnaSP) of recombination. *T. cruzi* diversity was measured by calculating the number of haplotypes and haplotype divergence. In an initial attempt to address the *T. cruzi* diversity problem from the perspective of Shannon information theory and thermodynamic views of evolution we address the hypothesis of a direct correlation between high entropy values and positive selection. Here we show that the "jumping" behavior of haplotypes in phylogenies of different loci could be due to intragenic recombination between the parental haplotypes originating mosaic haplotypes. Because the mosaic haplotypes have sequences that are closer to one or both parentals it could explain the existence of the hybrid haplotype-specific group with variable positions in phylogenies relative to TcI and TcII. We also show that the Shannon entropy of each position in the alignment for the four protein coding genes is highly correlated with the type of selection. Our results strongly support the proposition that *T. cruzi* III is a hybrid between TcI and II and that *T. cruzi* III–VI originated by two partially superimposed events (Westenberger et al., 2005, 2006; Ienne et al., 2010).

2. Material and methods

2.1. Trypanosoma cruzi strains

T. cruzi strains used in this work are shown in Table 1.

2.2. Sequences and alignments

Sequences of EF-1α, actin, dihydrofolate reductase-thymidylate synthase (DHFR-TS) and trypanothione reductase (TR) genes were downloaded from Genbank and the accession numbers are: (EF-1α) AY785657 to AY785706; (actin) AY785587 to AY785636; (DHFR-TS) AY785637 to AY785656, AF358956 to AF358959, AF358962 and AF358963; and (TR) AY785707 to AY785726, AF 358996 to AF358999, AF359002 and AF359003 (Machado and Ayala, 2001; Tomazi et al., 2009). We adopted the letter "A" in front of the strain name if the sequence were obtained from the cloned haplotype and the letters "AR" if the sequence were found in the direct PCR sequencing.

The alignments used here for the recombination detection are the same analyzed in Tomazi et al. (2009) and are available upon request.

2.3. Data analysis

The DnaSP version 5.0 (Librado and Rozas, 2009) was used to estimate population genetic parameters, including generation of polymorphisms table, haplotype analysis, polymorphisms statistics, recombination, linkage disequilibrium and neutrality tests. Nucleotide diversity was assessed calculating the number of haplotypes (h) and the haplotype divergency (Hd) (Nei, 1987). The confidence intervals were obtained by Monte Carlo simulations based

Table 1
Characteristics of *T. cruzi* strains used in this study.

Strain	Host / Vector	Origin	DTUs[a]
CA1	*Homo sapiens*	Argentina	*T. cruzi* I
Colombiana	*Homo sapiens*	Colombia	*T. cruzi* I
Dm 28c	*Didelphis marsupialis*	Venezuela	*T. cruzi* I
G	*Didelphis marsupialis*	Brazil	*T. cruzi* I
Honduras	*Homo sapiens*	Honduras	*T. cruzi* I
José	*Homo sapiens*	Brazil	*T. cruzi* I
Silvio X10 cl1	*Homo sapiens*	Brazil	*T. cruzi* I
Tc3014	*Didelphis albiventris*	Brazil	*T. cruzi* I
TcP06	*Didelphis albiventris*	Brazil	*T. cruzi* I
YuYu	*Triatoma infestans*	Brazil	*T. cruzi* I
Basileu	*Homo sapiens*	Brazil	*T. cruzi* II
Esmeraldo cl3	*Homo sapiens*	Brazil	*T. cruzi* II
Famema	*Homo sapiens*	Brazil	*T. cruzi* II
Hem 179	*Homo sapiens*	Brazil	*T. cruzi* II
SLU31	*Homo sapiens*	Brazil	*T. cruzi* II
SLU142	*Homo sapiens*	Brazil	*T. cruzi* II
TcVEN35	*Rhodnius neglectus*	Brazil	*T. cruzi* II
Y	*Homo sapiens*	Brazil	*T. cruzi* II
M6241 cl6	*Homo sapiens*	Brazil	*T. cruzi* III
MT3663	*Rhodnius brethesi*	Brazil	*T. cruzi* IV
MT3869	*Homo sapiens*	Brazil	*T. cruzi* IV
MT4167	*Rhodnius brethesi*	Brazil	*T. cruzi* IV
NR cl3	*Homo sapiens*	Chile	*T. cruzi* V
SC43 cl1	*Triatoma infestans*	Bolivia	*T. cruzi* V
SO3 cl5	*Triatoma infestans*	Bolivia	*T. cruzi* V
PSC-O	*Homo sapiens*	Chile	*T. cruzi* V
CLBrener	*Triatoma infestans*	Brazil	*T. cruzi* VI
Tulahuen cl2	*Homo sapiens*	Chile	*T. cruzi* VI

[a] According to Zingales et al. (2009).

on the coalescent process for a neutral infinite-sites model and assuming a large and constant population size (Kingman, 1982a,b; Hudson, 1983). Simulations were performed for 1000 independent replicates, assuming no intragenic recombination (Hudson, 1983). The recombination parameter C was estimated using the method described in Hudson and Kaplan (1985). This method is based on the minimum number of recombination events (R_m) in a DNA sample which is obtained using the four-gamete test. Estimates of R_m were used to estimate C by coalescent simulations (Hudson and Kaplan, 1985; Myers and Griffiths, 2003). The ZZ statistic (Rozas et al., 2001) was used to verify the effect of intragenic recombination on nucleotide variation by analyzing the level of linkage disequilibrium. Deviations from the neutral model of molecular evolution was acceded by Tajima's D, Fu and Li's D^* and F^* tests (Tajima, 1989; Fu and Li, 1993).

Potential recombination was inferred using phylogenetic based methods: the dual multiple change-point model implemented in the DualBrothers plugin in Geneious, and the bootscan analysis implemented at Simplot and Recombination Detection Program v3.1 (RDP3) (Lole et al., 1999; Martin and Rybicki, 2000; Martin et al., 2005a,b; Minin et al., 2005; Suchard et al., 2003). Parameters for DualBrothers analysis were: subsampling frequency = 10, window length = 200, step size = 10 and default model priors. For the bootscan analysis implemented at RDP3 we selected one sequence from each possible hybrid DTU (M6241(TcIII), MT4167 (TcIV), Nrcl3 (TcV) and CLBrener (TcVI)) and compared to the possible parental sequences G (TcI), Y (TcII) and M6241. Parameters for Bootscan analysis were: window size = 200; step size = 20; bootstrap replicates = 10,000; cutoff percentage = 70; use neighbor joining trees, calculate binomial *p*-value; model option=Jin and Nei, 1990 and closest relative scan. The program Simplot was used to compare groups of sequences belonging to the different DTUs. Sliding windows of 200 bp and step size of 20 bp were used. Neighbor-joining trees were estimated using the modeltest parameters obtained by Tomazi et al. (2009) for each of the four genes, and bootstrap values were obtained from 1000 pseudoreplicates. The threshold for assignment of parenthood was set 70%. For actin and EFIα, parental sequences were TcI and TcII with unrelated group being *T. brucei* when TcIII and TcIV strains were used as queries. When TcV strain was used as query for the EF1α, parental sequences were TcII and TcIII with unrelated group being *T. brucei*; for DHFR-TS and TR genes parental sequences were TcI and TcII with *T. cruzi marinkellei* and *T.* rangeli, respectively, as unrelated groups.

The Shannon entropy (Shannon, 1948) was calculated for each position in the complete amino-acid alignments using BioEdit software (Hall, 1999) for the four genes. The entropy calculated is given in nits. Conversion to bits is made by simply multiplying the entropy in nits by $\log_2 e$ (approx. 1.4427). From a random protein sequence alignment, with the same number of sequences of the real alignment, we calculate the average and the standard error associated with the entropy measurement for each amino-acid position.

For each codon in the nucleotide alignment we compared the difference between non-synonymous and synonymous substitutions with the entropy value to test whether or not variable positions under positive selection are correlated with high Shannon entropy values. We used a codon estimate selection test implemented in Mega v.5 (Tamura et al., 2011). The statistical method used was maximum likelihood with the initial tree generated by neighbor-joining. Substitution model was the General Time Reversible. Null hypothesis was H0: dN-dS = 0 (neutral evolution).

3. Results and discussion

3.1. Sequence diversity and neutrality tests

The number of polymorphic sites, haplotypes (h) and the haplotype diversity (Hd) for all genes are shown in Table 2. The overall haplotype diversity is extremely high varying between 0.878 and 0.984. Haplotypes 2, 28 and 2 have the greatest frequency (7/52; 9/52; 9/28) in the EF1α, actin and TR alignments, respectively. DHFR-TS have two haplotypes, 3 and 16, with the greatest frequency (3/28). The high diversity becomes very evident when comparing the number of haplotype and the number of sequences in each of the genes.

To investigate whether the observed pattern of nucleotide variation observed by Tomazi et al. (2009) is compatible to what is expected under neutrality, we performed three tests that compare different estimations of θ (Table 3) (Kimura, 1983; Tajima, 1989; Fu and Li, 1993). These tests are global analyses (use the whole gene sequence and not position-*per*-position). The neutral equilibrium hypothesis was not rejected in all tests.

Shannon entropy can be used to quantify the diversity of a system (Shannon, 1948). We calculated the entropy for each position in the amino-acid alignment for the four genes (Table 4). For actin, only 11 of 376 (2.93%) positions have positive values of entropy. DHFR-TS has 14 of 508 (2.76%), EF1α has 17 of 449 (3.79%) and TR has 18 of 444 (4.05%). All entropy values were above the standard error calculated from an alignment of random protein sequences indicating that at least for these sites, entropy values are undeniably positive. Adding new sequences to the alignments is expected to maintain or increase the entropy values for these sites. We cannot rule out the possibility of new positive entropy sites with the addition of novel sequences to the alignment. A different set of sequences may or may not have the same positions with positive entropy value but, if the alignments were large enough we

Table 3
Tests for neutrality in actin, DHRF-TS, EF1α and TR.

Gene	Tajima's D	Fu & Li's D*	Fu & Li's F*
Actin	−0.25446 (NS)	−2.37782 (NS)	−1.91221 (NS)
DHFR-TS	0.51670 (NS)	−0.30001 (NS)	−0.04249 (NS)
EF1α	−1.04910 (NS)	−1.90902 (NS)	−1.89951 (NS)
TR	0.86558 (NS)	−0.19435 (NS)	0.17434 (NS)

DHFR-TS, dihydrofolate reductase-thymidylate synthase; EF1α, elongation factor 1-alpha; TR, trypanothione reductase; NS, not significant.

Table 2
Population genetic parameters for actin, DHRF-TS, EF1α and TR genes.

Gene	Number of sequences in the alignment	Total number of sites (bp)	Polymorphic sites	h[a]	Hd[b]	R_m[c]
Actin	52	1128	41	30	0.948	1
DHFR-TS	28	1473	44	24	0.984	7
EF1α	52	1347	53	44	0.983	12
TR	28	1289	44	14	0.878	3

DHFR-TS, dihydrofolate reductase-thymidylate synthase; EF1α, elongation factor 1-alpha; TR, trypanothione reductase.

[a] Number of haplotypes.

[b] Haplotype diversity.

[c] Minimum number of recombination events.

Table 4
Information entropy values (in bits) for each variable position in the amino-acid alignment for the four genes, actin, DHRF-TS, EF1α and TR.

Gene	Position in alignment	Informationentropy (bits)	Standard error
Actin	11, 21, 28, 97, 138, 172, 203, 243, 354 and 364	0.13710	0.09499
Actin	13	0.24365	0.09499
DHFR-TS	30	0.99403	0.15991
DHFR-TS	53 and 135	0.97603	0.15991
DHFR-TS	147 and 303	0.43950	0.15991
DHFR-TS	177, 213, 246, 264, 265, 270, 318, 346 and 446	0.26677	0.15991
EF1α	30	0.45366	0.09499
EF1α	58, 63, 131, 212, 235, 253, 331, 339, 342 and 430	0.13710	0.09499
EF1α	269, 320, 328 and 381	0.23519	0.09499
EF1α	288	0.99038	0.09499
EF1α	433	0.97316	0.09499
TR	31, 43, 124, 188, 228, 259, 280, 290, 300 and 442	0,26677	0.15991
TR	77 and 332	0,43950	0.15991
TR	95 and 424	0,99403	0.15991
TR	139, 336, 385 and 386	0,97603	0.15991

DHFR-TS, dihydrofolate reductase-thymidylate synthase; EF1α, elongation factor 1-alpha; TR, trypanothione reductase.

expect the result to be stable. All other sites in the amino-acid alignments have entropy values equal to zero indicating that these sites are conserved among all haplotypes of all strains analyzed here. *T. cruzi* actin gene has three copies per diploid genome while EF1α, DHFR-TS and TR are single copy genes (Sullivan and Walsh, 1991; Reche et al., 1994; Billaut-Mulot et al., 1996; Cevallos et al., 2003). Because these genes have very low copy numbers and are important proteins for parasite metabolism, mutations will tend to have drastic effects on the parasite's ability to survive. For example, Eid and Sollner-Webb (1991) have shown that a reduction in the copy number of calmodulin genes is associated with a slow growth phenotype in *T. brucei*. High degrees of conservation found in these genes may indicate either functional or structural importance (Capra and Singh, 2007).

To verify if there is a direct correlation between high entropy and positive selection, estimate by the dN *minus* dS (ΔdNdS), for each codon in *T. cruzi* sequence alignments. Positions where ΔdNdS = 0 were excluded from the analysis because they represent invariant sites leaving only positions where the null hypothesis, i.e. neutral evolution, was rejected (Supplementary Table S1). We observed a high correlation between entropy values and positive selection. Of 181 codons under selection analyzed from all genes, 56 were under positive and 125 under negative selection. All of the codons under positive selection have entropy values $H > 0$. For the 125 codons under negative selection, 121 have entropy value equal to zero. Only four out of 125 present entropy value $H > 0$.

All the 44 sites analyzed for DHFR-TS were highly correlated. Fourteen sites had positive selection and entropy values > zero, while the 30 remaining sites had negative selection and entropy values equal zero. For actin, from the 40 analyzed sites, 10 had positive selection with entropy values higher than zero while 29 sites had negative selection with entropy equal zero. Only one site (a.a./codon 203) had negative selection with entropy higher than zero. This site corresponds to a non-conservative amino-acid change (T/K). EF1α has 51 sites analyzed with 15 showing positive selection and entropy values above zero and 34 positions with negative selection with entropy equals zero. Two sites have negative selection with entropy values higher than zero. Residue in the position 320 may be an aspartic acid or a glutamic acid (D/E) while in position 331 may be a lysine or a glutamine (K/Q). For TR, 46 sites were analyzed with 17 showing positive selection and entropy values above zero and 28 positions with negative selection with entropy equals zero. Again only one position (a.a. 188) has negative selection with entropy value higher than zero. Residue in this position may be an alanine or a valine.

The information at DNA and protein levels in most biological systems is where the second law of thermodynamics operates to generate biological complexity and order (Wiley and Brooks, 1982; Brooks and Wiley, 1986). Several studies show a correspondence between thermodynamic entropy and information entropy which is a very tempting perspective for the analysis of biomolecules, such in the case of *T. cruzi* proteins, that have both thermodynamic properties and, at the same time, mutual information (Landauer, 1961; Adami, 2004). Differently to what was observed by Kawashita et al. (2009) our findings are in agreement with the hypothesis of a direct correlation between high entropy values and positive selection. This incongruence may be due to the type of sequences used in both works. Here we used highly conserved, very low copy genes whereas Kawashita et al. (2009) used DGF-1 gene family that had undergone several events of gene duplication. These duplicated genes tend to evolve at different rates. Despite the fact that actin has three copies in the genome, i.e. duplicated, we were still able to find a correlation between positive selection and entropy. The correlation is probably observed because actin is an essential gene under strong functional constraints.

3.2. Recombination detection and linkage disequilibrium

The effect of intragenic recombination on nucleotide sequence variation was estimated using the ZZ statistic and the minimum number of recombination events (R_m) implemented in the DnaSP version 5.0 package. Because linkage disequilibrium decays with physical distance due to intragenic recombination, the ZZ statistic is expected to have larger positive values with increasing recombination (Rozas et al. (2001)). No correlation was found between the levels of linkage disequilibrium and polymorphic sites for the four genes (data not shown). On the other hand, the R_m estimated by means of the four-gamete test reveled recombination sites for all genes analyzed here, suggesting intragenic recombination between the strains (Table 2). Not all recombination events in the history of the sample are revealed by the four-gamete test. This means that for a typical sample many more recombinant events probably took place, indicating that the R_m values found for the genes are underestimated (Hudson and Kaplan, 1985).

A recombination detection based on a dual multiple change-point model (MCP) implemented in the DualBrothers plugin in Geneious Pro 5.5.3 was performed for the four genes (Suchard et al., 2003; Minin et al., 2005). This method assumes that the evolutionary process is not independent and identically distributed at each site of the sequence, which would produce spatial phylogenetic variation along the sequences. In this model, an alignment is partitioned into an unknown number of segments. When appropriate sequences representing the parental are included, the topology describing a putative recombination can vary appearing as changes in the most probable topology between the segments. Recombination breakpoints were found for the four analyzed genes (data not shown).

To test whether these recombination patterns were consistent, we compared groups of sequences belonging to each *T. cruzi* DTU using the Simplot program (Lole et al. (1999)). When TcI and TcII were used as parental sequences, we detected recombination in TcIII and TcIV strains for actin and α genes; and in TcV for DHFR-TS and TR genes. When TcII and TcIII were used as parental, recombination was detected in TcV strains for EF1α. No recombination was detected neither in actin for TcV and TcVI nor in EF1α for TcVI with TcII and TcIII as parental. The same happened with TcVI strains for DHFR-TS and TR when TcI and TcII were the parental (Supplementary Fig. 1).

To test for intragenic recombination within each sequence we analyzed each haplotype separately from the four possible hybrids DTUs for the four protein encoding genes using the RDP3 software. The possible recombinant strains used were: M6241 (TcIII), MT4167 (TcIV), NRcl3 (TcV) and CLBrener (TcVI). Depending on the gene, sequences from strains G (TcI), Y (TcII) and M6241 were used as parental.

The M6241 and MT4167 strain were analyzed using G and Y as a parental. Both actin sequence haplotypes of strain M6241 have the same intragenic recombination patterns with regions that are close to TcI and others that are close to TcII (Fig. 1 A). The recombination pattern of EF1α sequences of the same strain is slightly different between the two haplotypes (Fig. 1 B, C). Both loci seem to be homozygous, because they have the same hybridization pattern for both haplotypes. The low bootstrap support in some regions is probably due to the deep divergence time between strains. If strain M6241 is representative of an ancestral group and not a hybrid we would expect to see regions close to TcI and II with the same frequency in both genes and with low bootstrap support in the entire alignment, but this is not true. Actin haplotypes possess larger regions closer to TcI than EF1α haplotype and regions with bootstrap support of 100%. This result suggests that this strain is a hybrid and not an ancestral group as suggested by de Freitas et al. (2006).

For MT4167, one actin haplotype are closer to TcI with a high bootstrap support while the other has regions at the beginning and end of the sequence that are closer to TcII (Fig. 1 D, E). The haplotypes of EF1α have small differences but the overall pattern is that the sequences are closer to TcII (Fig. 1 F). To test whether the TcIII strain is a possible parental sequence of TcIV we reanalyzed the data using M6241 and Y as ancestrals. Actin haplotypes have regions with similarity with the TcIII strain but the bootstrap support is lower than the values obtained when strain G was used as parental. Again, both EF1α haplotypes are closer to TcII with high bootstrap values. Taken together, these results favor the hypothesis that TcIV is a hybrid of TcI and TcII. This hybridization event should be as old as the event that gave birth to TcIII since all non satellite loci are expected to be homozygous (Westenberger et al., 2005). Our data suggest that this hybrization is not as old as the TcIII hybridization event, since both haplotypes for the genes, especially for actin, have different recombination patterns hence are different alleles.

For actin and EF1α genes, NRcl3 and CLBrener strains were analyzed using TcII and III as parental. Actin sequences for both strains showed one haplotype closer to TcIII and the other closer to TcII (Fig. 2A–D). EF1α also have distinct recombination patterns between haplotypes of the two strains but they are all closer to TcII. Sequences for DHFR-TS and TR genes were not available for TcIII, so TcI sequences were used as parental. Recombination detection for NRcl3 DHFR-TS gene shows that one haplotype is entirely close to TcII with high bootstrap value where the other has regions closer to TcI and TcII (Fig. 2 E, F). The pattern for TR gene shows one haplotype almost entirely closer to TcI while the other has portions of the sequence closer to TcII. CLBrener recombination pattern is the same for DHFR-TS and TR genes with one haplotype entirely close to TcII while the other is almost entirely TcI, both with high bootstrap values (Fig. 2 G, H). Results reinforce the hybrid nature of these strains. The use of a TcI strain as parental for for TcV and TcVI

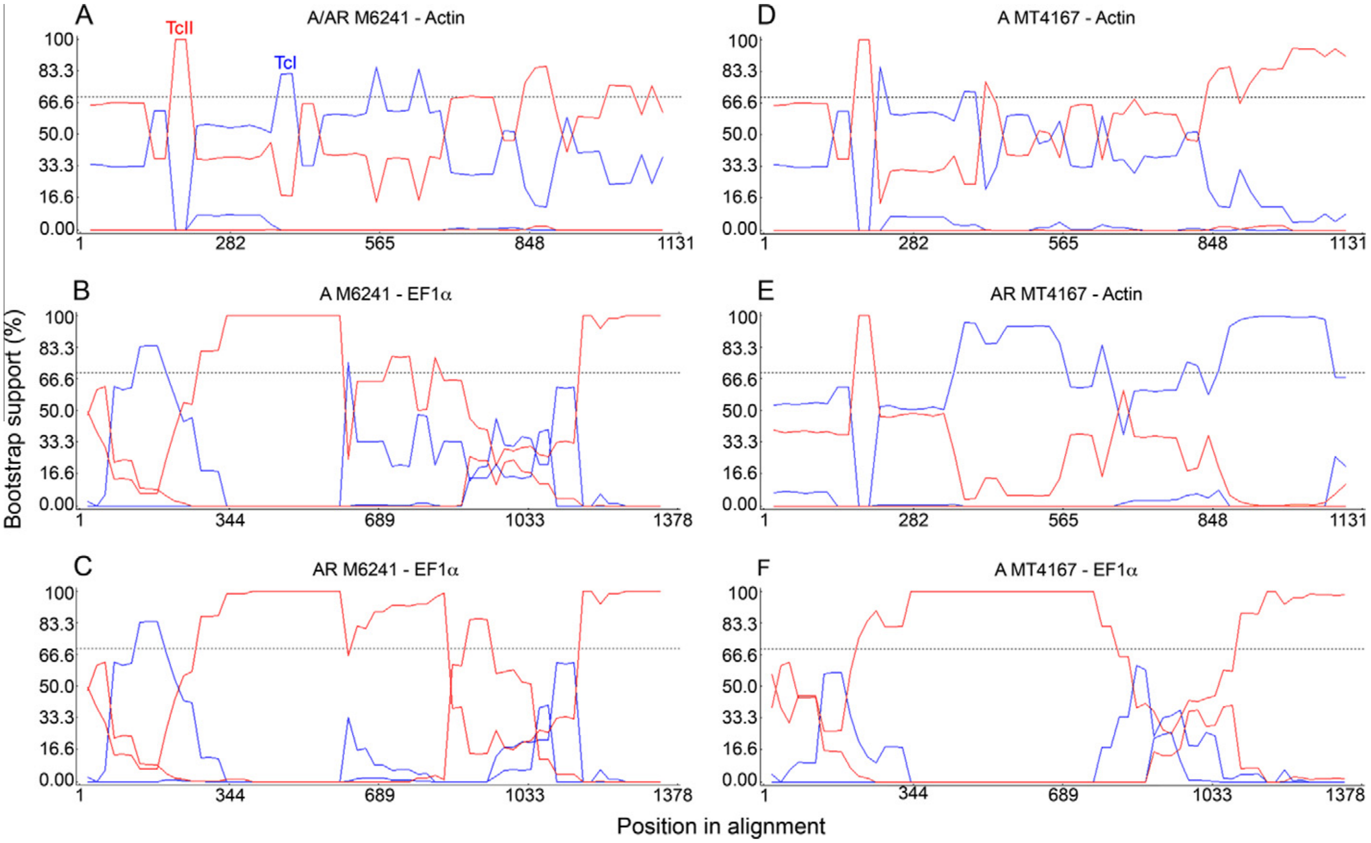

Fig. 1. Bootscan/RDP analysis of strains M6241 (TcIII) and MT4167 (TcIV). G (TcI – blue line) and Y(TcII – red line) sequences were used as parental. (A) A/AR_M6241 – actin (B) A_M6241 - EF1α (C) AR_M6241- EF1α (D) A_MT4167 – actin (E) AR_MT4167 – actin and (F) A_MT4167 – EF1α. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

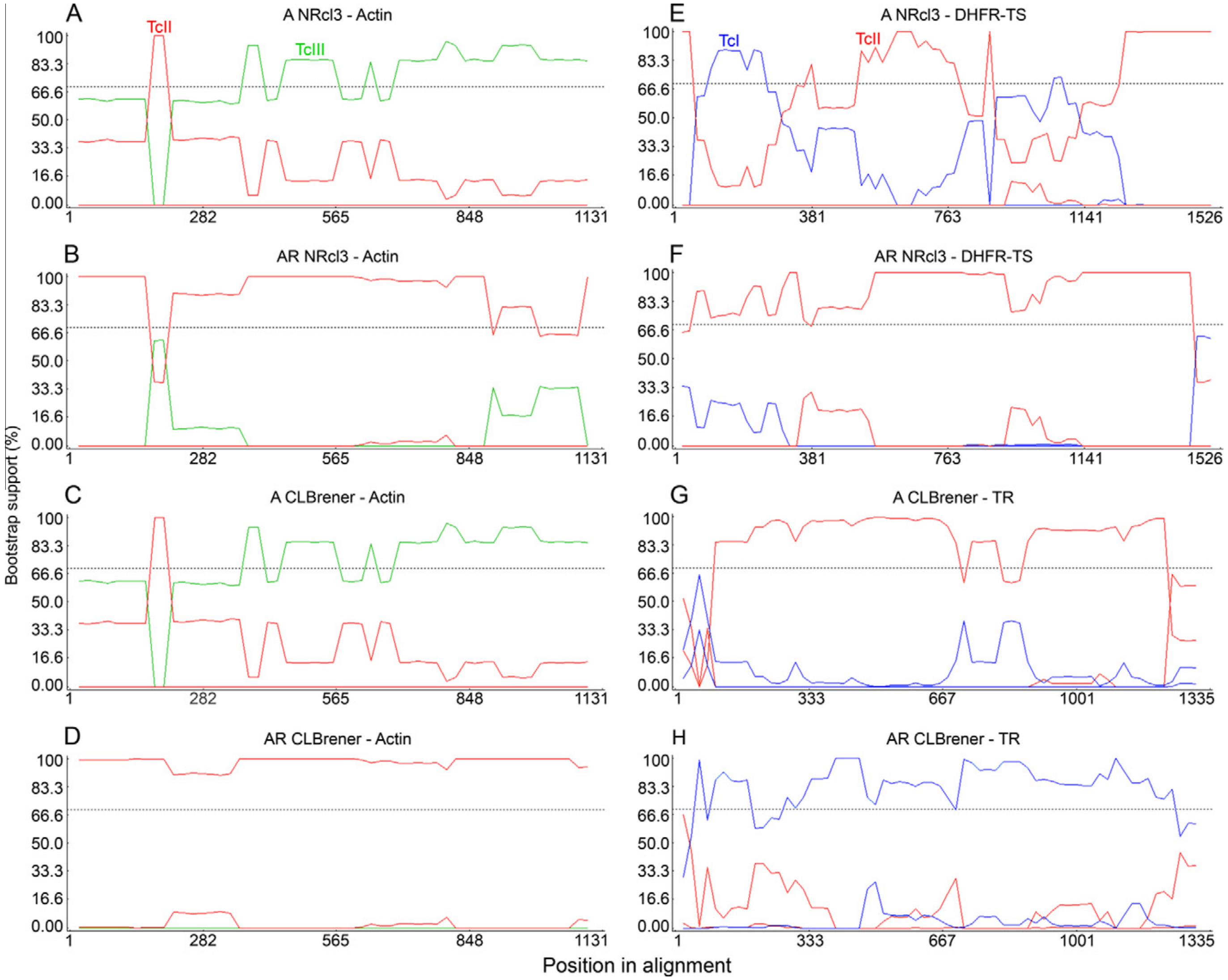

Fig. 2. Bootscan/RDP analysis of strains NRcl3 (TcV) and CLBrener (TcVI). G (TcI – blue line), Y(TcII – red line) and M6241 (green line – TcIII) sequences were used as parental. (A) A_NRcl3 – actin (B) AR_NRcl3 – actin (C) A_CLBrener – actin (D) AR_CLBrener – actin (E) A_NRcl3 – DHFR-TS (F) AR_NRcl3 – DHFR-TS (G) A_CLBrener – TR and (H) AR_CLBrener – TR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

might be a problem because the hybridization event that generated TcIII, the putative parental, is old and the portions of sequence that were closer to TcI could have been lost due to mutation. We observed high portions of the hybrids being closer to TcI so some genetic fingerprint of this parental is left on TcIII and thus in TcV and TcVI strains.

The most common methods for detecting intragenic recombination probably miss events that occurred a long time ago, as expected in *T. cruzi* hybridization (Posada, 2002). Posada and Crandall (2001) have shown that detection of recombination is rarely due to false positives, despite that the power of these methods is not good at low recombination frequencies. This means that for the four datasets analyzed here, recombination is not likely an isolated or extremely rare event. Because recombination detection power is also dependent on the number of sequences, addition of other DTUs would allow a higher power on detecting recombination, regardless of increased or decreased genetic diversity. Homoplasies produced by very rapid substitution rates could also result in false-positives. For TR and DHFR-TS the estimated rate of mutation is very low and fall within the range of silent rates estimated for mammalian and *Drosophila* nuclear genes, thus attesting that intragenic recombination found here is not a false positive result (Machado and Ayala, 2001).

The mosaic structure of the hybrid genes is likely caused by intragenic recombination. This may be one of the reasons for the "jumping" behavior of some haplotypes in phylogenies and why in some phylogenies one haplotype is closer to TcI clade while in others, closer to TcII. Depending on the degree of recombination there is no phylogenetic resolution to cluster these hybrids with their ancestral groups which could explain the hybrid clade identified in Tomazi et al. (2009). Despite the fact that we only have one TcIII strain in our dataset, our bootscan results suggest that TcIII is a hybrid. This finding is consistent with the two hybridization event hypothesis and contradict the proposition that TcIII is an ancestral group (Westenberger et al., 2005; de Freitas et al., 2006; Ienne et al., 2010). Taken together, the recombination data, the genealogies presented in Tomazi et al. (2009) and the maxicircle pedigree showed by Ruvalcaba-Trejo and Sturm (2011), provides strong evidence of the TcI parental contribution to the first hybridization event and therefore to the *T. cruzi* population structure.

Acknowledgements

We thank Prof. Dr. Francisco Bosco for the valuable assistance with entropy discussions. RCF is a postdoctoral fellow supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil. MRSB received grants from FAPESP, Conselho Nacional de Pesquisas Ceintíficas e Tecnológicas (CNPq), Brazil, and the international program of Howard Hughes Medical Institute.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2012.03.010.

References

Adami, C., 2004. Information theory in molecular biology. Phys. Life Rev. 1, 3–22.

Billaut-Mulot, O., Fernandez-Gomez, R., Loyens, M., Ouaissi, A., 1996. *Trypanosoma cruzi* elongation factor 1-alpha: nuclear localization in parasites undergoing apoptosis. Gene 174, 19–26.

Briones, M.R.S., Souto, R.P., Stolf, B.S., Zingales, B., 1999. The evolution of two *Trypanosoma cruzi* subgroups inferred from rRNA genes can be correlated with the interchange of American mammalian faunas in the Cenozoic and has implications to pathogenicity and host specificity. Mol. Biochem. Parasitol. 104, 219–232.

Brisse, S., Barnabé, C., Tibayrenc, M., 1998. *Trypanosoma cruzi*: how many relevant phylogenetic subdivisions are there? Parasitol. Today 14, 178–179.

Brisse, S., Barnabé, C., Tibayrenc, M., 2000. Identification of six *Trypanosoma cruzi* phylogenetic lineages by random amplified polymorphic DNA and multilocus enzyme eletrophoresis. Int. J. Parasitol. 30, 34–44.

Brisse, S., Henriksson, J., Barnabé, C., Douzery, E.J.P., Berkvens, D., Serrano, M., de Carvalho, M.R.C., Buck, G.A., Dujardin, J.C., Tibayrenc, M., 2003. Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. Infect. Genet. Evol. 2, 173–183.

Brooks, D.R., Wiley, E.O., 1986. Evolution as Entropy: Toward a Unified Theory of Biology. University of Chicago Press, Chicago.

Capra, J.A., Singh, M., 2007. Predicting functionally important residues from sequence conservation. Bioinformatics 23, 1875–1882.

Cevallos, A.M., López-Villasenor, I., Espinosa, N., Herrera, J., Hernández, R., 2003. *Trypanosoma cruzi*: allelic comparisons of the actin genes and analysis of their transcripts. Exp. Parasitol. 103, 27–34.

De Freitas, J.M., Augusto-Pinto, L., Pimenta, J.R., Bastos-Rodrigues, L., Gonçalves, V.F., Teixeira, S.M.R., Chiari, E., Junqueira, A.C.V., Fernandes, O., Macedo, A.M., 2006. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. PLoS Pathog. 2 (e24), 226–235.

Dos Santos, W.G., Buck, G.A., 1999. Polymorphisms at the topoisomerase II gene locus provides more evidence for the partition of *Trypanosoma cruzi* into two major groups. J. Eukar. Microbiol. 46, 17–23.

Eid, J.E., Sollner-Webb, B., 1991. Homologous recombination in the tandem calmodulin genes of *Trypanosoma brucei* yields multiple products: compensation for deleterious deletions by gene amplification. Genes Dev. 5, 2024–2032.

Elias, M.C.Q.B., Vargas, N., Tomazi, L., Pedroso, A., Zingales, B., Schenkman, S., Mriones, M.R.S., 2005. Comparative analysis of genomic sequences suggests that *Trypanosoma cruzi* CL Brener contains two set of non-intercalated repeats of satellite DNA that correspond to *T. cruzi* I and *T. cruzi* II types. Mol. Biochem. Parasitol. 140, 221–227.

Fernandes, O., Souto, R.P., Castro, J.A., Pereira, J.B., Fernandes, N.C., Junqueira, A.C., Naiff, R.D., Barrett, T.V., Degrave, W., Zingales, B., Campbell, D.A., Coura, J.R., 1998. Brazilian isolates of *Trypanosoma cruzi* from humans and triatomines classified into two lineages using mini-exon and ribosomal RNA sequences. Am. J. Trop. Med. Hyg. 58, 807–811.

Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.

Gaunt, M.W., Yeo, M., Frame, I.A., Stothard, J.R., Carrasco, H.J., Taylor, M.C., Mena, S.S., Veazey, P., Miles, G.A.J., Acosta, N., Arias, A.R., Miles, M.A., 2003. Mechanism of genetic exchange in American trypanosomes. Nature 421, 936–939.

Hall, T.A., 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41, 95–98.

Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23, 183–201.

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–164.

Ienne, S., Pedroso, A., Ferreira, R.C., Briones, M.R.S., Zingales, B., 2010. Network genealogy of 195-bp satellite DNA supports the superimposed hybridization hypothesis of *Trypanosoma cruzi* evolutionary pattern. Infect. Gent. Evol. 10, 601–606.

Jin, L., Nei, M., 1990. Limitation of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7, 82–102 (erratum in Mol. Biol. Evol. 1990. 7, 201).

Kawashita, S.Y., Sanson, G.F.O., Fernandes, O., Zingales, B., Briones, M.R.S., 2001. Maximum-likelihood divergence date estimates based on rRNA gene sequences suggest two scenarios of *Trypanosoma cruzi* intraspecific evolution. Mol. Biol. Evol. 18, 2250–2259.

Kawashita, S.Y., Silva, C.V., Mortara, R.A., Burleigh, B.A., Briones, M.R.S., 2009. Homology, paralogy and function of DGF-1, a highly dispersed *Trypanosoma cruzi* specific gene family and its implications for information entropy of its encoded proteins. Mol. Biochem. Parasitol. 165, 19–31.

Kimura, M., 1983. The Neutral Theory of Evolution. Cambridge University Press, Cambridge.

Kingman, J.F.C., 1982a. The coalescent. Stochastic Proc. Appl. 13, 235–248.

Kingman, J.F.C., 1982b. On the genealogy of large populations. J. Appl. Probab. 19A, 27–43.

Landauer, R., 1961. Irreversibility and heat generation in the computing process. IBM J. Res. Devel. 5, 183–191.

Lewis, M.D., Llewellyn, M.S., Gaunt, M.W., Yeo, M., Carrasco, H.J., Miles, M.A., 2009. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanossoma cruzi* populations and expose contrasts between natural and experimental hybrids. Int. J. Parasitol. 39, 1305–1317.

Librado, P., Rozas, J., 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451–1452.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73, 152–160.

Macedo, A.M., Machado, C.R., Oliveira, R.P., Pena, S.D., 2004. *Trypanosoma cruzi*: genetic structure of populations and relevance of genetic variability to the pathogenesis of Chagas disease. Mem. Inst. Oswaldo Cruz 99, 1–12.

Machado, C.A., Ayala, F.J., 2001. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. Proc. Natl. Acad. Sci. USA 98, 7396–7401.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563.

Martin, D.P., Crandall, K.A., Willianson, C., 2005a. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Re. Hum. Retroviruses 21, 98–102.

Martin, D.P., Williamson, C., Posada, D., 2005b. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21, 260–262.

Miles, M.A., Cedillos, R.A., Povoa, M.M., de Souza, A.A., Prata, A., Macedo, V., 1981. Do radically dissimilar *Trypanosoma cruzi* strains cause Venezuelan and Brazilian forms of Chagas' disease? Lancet 1, 1338–1340.

Miles, M.A., Yeo, M., Gaunt, M.W., 2003. Genetic diversity of *Trypanosoma cruzi* and the epidemiology of Chagas disease. In: Kelly, J.M. (Ed.), Molecular Mechanisms of Pathogenesis in Chagas Disease. Kluwer Academic, New York, pp. 1–15.

Minin, V.N., Dorman, K.S., Fang, F., Suchard, M.A., 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21, 3034–3042.

Myers, S.R., Griffiths, R.C., 2003. Bounds on the minimum number of recombination events in a sample history. Genetics 163, 375–394.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York, NY.

Posada, D., 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19, 708–717.

Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA 98, 13757–13762.

Rassi Jr., A., Rassi, A., Marin-Neto, J.A., 2010. Chagas disease. The Lancet 375, 1388–1402.

Reche, P., Arrebola, R., Olmo, A., Santi, D.V., Gonzalez-Pacanowska, D., Ruiz-Perez, L.M., 1994. Cloning and expression of the dihydrofolate reductase-thymidylate synthase gene from *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 65, 247–258.

Rozas, J., Gullaud, M., Blandin, G., Aguadé, M., 2001. DNA variation at the*rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158, 1147–1155.

Ruvalcaba-Trejo, L.I., Sturm, N.R., 2011. The *Trypanosoma cruzi* Sylvio X10 strain maxicircle sequence: the third musketeer. BMC Genomics 12, 58–70.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Tech. J. 27 (379–423), 623–656.

Souto, R.P., Fernandes, O., Macedo, A.M., Campbell, D.A., Zingales, B., 1996. DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 83, 141–152.

Sturm, N.R., Campbell, D.A., 2010. Alternative lifestyles: the population structure of *Trypanosoma cruzi*. Acta Trop. 115, 35–43.

Subileau, M., Barnabé, C., Douzery, E.J.P., Diosque, P., Tibayrenc, M., 2009. *Trypanosoma cruzi*: new insights on ecophylogeny and hydridization by multigene sequencing of three nuclear and one maxicircle genes. Exp. Parasitol. 122, 328–337.

Suchard, M.A., Weiss, R.E., Dorman, K.S., Sinsheimer, J.S., 2003. Inferring spatial variation along nucleotide sequences: a multiple changepoint model. J. Am. Stat. Assoc. 98, 427–437.

Sullivan, F.X., Walsh, C.T., 1991. Cloning, sequencing, overproduction and purification of trypanothione reductase from *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 44, 145–147.

Tajima, F., 1989. Statistical method for testing the neutral mutation hyphotesis by DNA polymorphisms. Genetics 123, 585–595.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Tibayrenc, M., Neubauer, K., Barnabé, C., Guerrini, F., Skarecky, D., Ayala, F.J., 1993. Genetic characterization of six parasitic protozoa: parity between random-primer DNA typing and multilocus enzyme electrophoresis. Proc. Natl. Acad. Sci. USA 90, 1335–1339.

Tibayrenc, M., 1995. Population genetics of parasitic protozoa and other microorganisms. Adv. Parasitol. 36, 47–115.

Tibayrenc, M., Ayala, F.J., 1988. Isoenzyme variability in *Trypanosoma cruzi*, the agent of Chagas' disease: genetical, taxonomical and epidemiological significance. Evolution 42, 277–292.

Tomazi, L., Kawashita, S.Y., Pereira, P.M., Zingales, B., Briones, M.R.S., 2009. Haplotype distribution of five nuclear genes based on network genealogies and Bayesian inference indicates that *Trypanosoma cruzi* hybrid strains are polyphyletic. Genet. Mol. Res. 8, 458–476.

Venegas, J., Coñoepan, W., Pichuantes, S., Miranda, S., Jercic, M.I., Gajardo, M., Sánchez, G., 2009. Phylogenetic analysis of microsatellite further supports the two hybridization events hyphothesis as the origin of the *Trypanosoma cruzi* lineages. Parasitol. Res. 105, 191–199.

Westenberger, S.J., Barnabé, C., Campbell, D.A., Sturm, N.R., 2005. Two hybridization events define the population structure of *Trypanosoma cruzi*. Genetics 171, 527–543.

Westenberger, S.J., Cerqueira, G.C., El-Sayed, N.M., Zingales, B., Campbell, D.A., Sturm, N.R., 2006. *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. BMC Genomics 7, 60–77.

Wiley, E.O., Brooks, D.R., 1982. Victims of historyb a non-equilibrium approach to evolution. Syst. Zool. 31, 1–24.

Zingales, B., Souto, R.P., Mangia, R.H., Lisboa, C.V., Campbell, D.A., Coura, J.R., Jansen, A., Fernandes, O., 1998. Molecular epidemiology of American trypanosomiasis in Brazil based on dimorphisms of rRNA and mini-exon gene sequences. Int. J. Parasitol. 28, 105–112.

Zingales, B., Andrade, S.G., Briones, M.R.S., Campbell, D.A., Chiari, E., Fernandes, O., Guhl, F., Lages-Silva, E., Macedo, A.M., Machado, C.R., Miles, M.A., Romanha, A.J., Sturm, N.R., Tibayrenc, M., Schijman, A.G., 2009. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem. Inst. Oswaldo Cruz 104, 1051–1054.