Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

Rodrigo de Paula Baptista

Aplicação de abordagens bioinformáticas no estudo da estrutura populacional de *Trypanosoma cruzi*

Belo Horizonte, 2014

Rodrigo de Paula Baptista

Aplicação de abordagens bioinformáticas no estudo da estrutura populacional de *Trypanosoma cruzi*

Tese apresentada ao Programa de Pós-graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para obtenção do titulo de Doutor em Bioinformática.

Orientadora: Profa. Dra. Andrea Mara Macedo
Co-Orientadora: Daniella Castanheira Bartholomeu

Belo Horizonte, 2014

“Só falha aquele que não acredita”

(Yoda- adaptado)

AGRADECIMENTOS

À minha orientadora Profa. Andréa Mara Macedo por ter acreditado em mim, pela amizade, compreensão e paciência durante os anos de iniciação científica e doutorado, pela orientação sensata e objetiva indispensável para a minha formação.

À minha co-orientadora Profa. Daniella Castanheira Bartholomeu pela compreensão, orientação, entre outras várias colaborações, indispensáveis para a finalização deste trabalho.

Ao Prof. Carlos Renato Machado, pela amizade, paciência, e por ter sido um dos principais idealizadores de parte deste trabalho.

À Profa. Glória Regina Franco por ter sempre me dado apoio, sendo até mesmo confundida como minha orientadora por muitas vezes e por suas ideias e sugestões que contribuíram para o desenvolvimento deste trabalho.

Ao Dr. Helder Magno Silva Valadares, meu Pai científico, que me passou os primeiros ensinamentos em Biologia Molecular os quais despertaram em mim o gosto pela pesquisa.

À Dra. Jessica Kissinger por ter me recebido em seu laboratório nos Estados Unidos, me providenciando toda a infraestrutura para a realização de boa parte deste projeto, pelas ideias, ensinamentos e pela amizade construída.

Aos colaboradores deste trabalho Dr. Egler Chiari, Dra. Maria Lúcia Galvão, Dra. Daniella Dávila, Dra. Eliane Gontijo, Dra. Silvane Murta e Dr. Jeremy DeBarry, pelas amostras biológicas, idéias e informações.

Às agências financiadoras WHO, NIH, CAPES, CNPq.

À Neuzinha pela amizade e carinho, sempre disponível e interessada em ajudar os alunos do laboratório em quaisquer dificuldades.

À Katita pela amizade e pelo carinho de mãe, pelo seu trabalho no ALF que foi essencial para a obtenção dos dados indispensáveis para este trabalho.

Aos amigos atuais e antigos do Laboratório de Genética-Bioquímica: Grupo da Andréa (Marcela Segatto, Ítalo do Valle, Viviane, Guilherme, Claudiney, Isabella, Priscilla, Wagson, Cláudia, Ludmila, Joice, José Ricardo, Tiago Bruno, Polly e Kamilla). Grupo do Carlos Renato (Pedrão, Bruno, Carol, Matheus, Mari Babys, Dani GPS, Michele Barbi, Ceres, Isabela, Bruno Repolês, Paula, Sabrina, Helida, Anna, Jarbas, Hugo, Selma, Egídio e Chris), Grupo da Glória (Priscila Grynberg, Marcela Drummond, Marina, Chico Lobo, Daiane, Maíra Rodrigues, Jane, João Pedro, Heron, Michele, Mari Bobs, Thiago Mafra, Leonardo Carnevalli, Dani Chame, Dani Durso, Carol, Thomaz, Mari Costa, Elizângela, Mainá, Carlos, André, Silvia, Magui e Jessica) e Grupo do Sérgio (Lucas, Ferdi, Higgor, Clarice, Vanessa, Pricila, Natália, Maíra, Grazi e Raony) pela amizade, pelos momentos alegres e tristes, por terem aturado as minhas piadas sem

graça e mesmo assim rindo delas, pelo companheirismo, pela discussão e ajuda na execução de experimentos, enfim, por terem feito parte da minha vida.

Aos amigos da UGA, principalmente à Betsy pela amizade, confiança e todo suporte que tornaram minha passagem por terras gringas perfeita.

Ao Dr. Adhemar Zerlotini, a Dra. Laila Nahum e ao Dr. Guilherme Oliveira por me ajudarem no contato com a Jessica para o Doutorado Sanduíche.

Aos novos companheiros e amigos do LIGP, principalmente ao João Cunha e a Mariana, pelas ideias, ajuda, colaborações e pela nova etapa.

À Dra. Leonor Guerra por ter acreditado, incentivado e me indicado para a iniciação científica, porta principal para eu estar finalizando o doutorado.

Aos demais amigos do curso de pós-graduação, principalmente Giordano, Rondon, Rodrigo Kato, Dr. Bráulio e Dr. Marcos pelo companheirismo e cooperação na discussão e prática de experimentos.

Aos amigos João Gabriel Machado, Thiago Tavares, grupo OKIMI, SN e Alluminium pelo companheirismo e amizade sincera.

À Camila de Araújo Andrade por todo amor, paciência, ensinamentos, e por sempre estar ao meu lado.

À minha família que sempre esteve e sempre estará ao meu lado onde quer que eu esteja. Principalmente meus PAIS que me deram todo suporte para que fosse possível realizar esse sonho.

À Deus, por tudo que fez acontecer em minha vida.

RESUMO

A doença de Chagas, causada pelo protozoário *Trypanosoma cruzi*, apresenta diferentes manifestações clínicas, resultantes de uma interação complexa entre fatores ambientais e genéticos tanto do hospedeiro quanto do parasito. Com relação aos fatores genéticos do parasito, uma variedade de estudos biológicos e moleculares revelou uma substancial variabilidade gênica entre populações de *T. cruzi.* De fato, em 2009, O táxon *T. cruzi* foi subdividido em seis linhagens ou DTUs principais (TcI-VI). Os aspectos epidemiológicos associados às diferentes linhagens ainda não estão bem definidos, mas já há evidências de associação, pelo menos parcial, com a distribuição geográfica, ciclos de transmissão e aspectos clínicos da doença. Todavia vários aspectos da estrutura populacional do parasito ainda necessitam esclarecimentos. Por exemplo, o modo de reprodução, a origem das seis linhagens e as relações com os seus aspectos epidemiológicos. No presente trabalho procurou-se desenvolver ou aperfeiçoar procedimentos moleculares, e computacionais que permitissem investigar e dissecar a complexidade da estrutura populacional de *T. cruzi*. Para tanto, foram propostos os seguintes objetivos (i) investigar a estratégia de reprodução do *T. cruzi*; (ii) buscar novas metodologias de caracterização molecular capazes de diferenciar as diferentes linhagens de *T. cruzi* em futuros ensaios multiplex de PCR; (iii) sequenciar o genoma de uma linhagem ainda não disponível nos bancos de dados públicos e realizar uma analise comparativa com seus dados genômicos a fim de resolver a origem ancestral das seis linhagens filogenéticas propostas; (iv) e, por fim, desenvolver uma metodologia computacional de agrupamento capaz de reconhecer padrões responsáveis pela divisão das linhagens e gerar regiões candidatas para o desenho de iniciadores para novos ensaios de caracterização molecular.

Baseados em marcadores nucleares e mitocondriais e em parâmetros de genética populacional, analisamos dois grupos de isolados de *T. cruzi*: um proveniente diferentes regiões da América Latina e outros obtidos exclusivamente de pacientes do Estado de Minas Gerais. Nossos resultados, ao contrário daqueles majoritariamente encontrados na literatura, não suportam uma estrutura de população essencialmente clonal para *T. cruzi*, ao menos para TcII coexistindo em uma mesma área geográfica, sugerindo que as trocas genéticas entre cepas estreitamente relacionadas geograficamente são mais frequentes do que inicialmente esperado.

Baseado na técnica de caracterização alelo específica BI-PASA, nós selecionamos uma variação no gene NAD desidrogenase subunidade I (ND1) de *T. cruzi*, e nossos resultados demostraram que esta metodologia foi eficaz, sendo rápida e mais barata, tornando-se uma boa candidata a ser utilizada em futuros ensaios de caracterização destes haplótipos.

O genoma do clone da cepa 231 pertencente à linhagem TcIII foi sequenciado e para a montagem foi desenvolvida uma abordagem de montagem combinada de diferentes metodologias capaz de melhorar a acurácia da montagem deste genoma de conteúdo altamente repetitivo. Posteriormente, nós realizamos uma análise comparativa entre o

genoma da cepa 231 montado e os demais disponíveis em bancos de dados e selecionamos genes altamente conservados, que foram utilizados em análises filogenéticas. Os resultados obtidos apontam para uma nova hipótese de origem das seis linhagens filogenéticas de *T. cruzi* e nos permitiram concluir que a linhagem TcIII, diferentemente do que foi proposto por diversos trabalhos, não é uma linhagem híbrida.

Sequências gênicas codificadoras de proteínas altamente repetitivas e variáveis, como a amastina, de diferentes linhagens do parasito foram analisadas em um script criado por nosso grupo baseado em decomposição de valores singulares e K-means juntamente com uma análise de regressão logística, resultando na geração de regiões candidatas responsáveis pela divisão destas sequências em grupos, que foram similares à divisão filogenética esperada. Esta metodologia poderá ser aplicada futuramente a outros grupos gênicos para elaboração de novas metodologias de caracterização molecular. As metodologias desenvolvidas e os resultados obtidos neste trabalho contribuem para um melhor entendimento da complexidade da estrutura populacional de *T. cruzi*.

ABSTRACT

Chagas disease, caused by the protozoan parasite *Trypanosoma cruzi*, has different clinical manifestations, that result from complex host-parasite interactions and involves both genetic and environmental factors. Regarding the parasite genetic factors, a variety of biological and molecular studies revealed an extensive genetic variability among *T. cruzi* strains, which were recently classified into six major lineages or DTUs (TcI-VI). Epidemiological aspects associated with these different lineages are not well defined, but there are some evidences of association with geographic distribution, transmission cycles and clinical aspects of this disease. Thus, in order to thoroughly investigate the population structure of this parasite, the efforts of our research group have been focused on developing new methodologies for *T. cruzi* characterization and study of the Chagas disease molecular epidemiology. In this present work, we seek to develop molecular and computational procedures enabling us to elucidate and dissect the complexity of population structure of *T. cruzi*. To this end, we proposed (i) to investigate the mode of reproduction of *T. cruzi* and search new molecular markers able to differentiate all six *T. cruzi* phylogenetic lineages and that could be used to develop multiplex assays; (ii) sequencing the genome of 231 strain, a representative of TcIII, a DTU not yet sampled in *T. cruzi* genome projects, and perform comparative genomic analyses to better resolve the ancestral origin of these six lineages (iii) and develop a sequence clustering methodology capable of recognizing sequences patterns responsible for the division of *T. cruzi* DTUs and based on this analysis identify novel candidate regions for molecular characterization assays.

Using population genetic approaches, we analyzed both nuclear and mitochondrial markers in two groups of *T. cruzi*: one using strains isolated from different regions of Latin America and another using strains isolated exclusively from the Minas Gerais state, in Brazil. Our results, unlike those mostly found in literature, does not support an clonal population structure for *T. cruzi*, at least for TcII coexisting in the same geographic area, suggesting that genetic exchange between geographically closely related strains are more frequent than initially expected.

The perspective is to incorporate the selected target in a multiplex PCR assay suitable for the study of *T. cruzi* populations present in tissues of chronic chagasic patients. Based on the Bi-PASA technique of allele specific characterization, we select a SNP variation in the *T. cruzi* ND1 mitochondrial gene, and our results showed that this methodology was effective, faster and cheaper than the RFLP-PCR array, making it a good candidate to be used in future mitochondrial haplotypes characterization tests.

The genome of the 231 clone, which belongs to the TcIII lineage, was sequenced and for the assembly we have developed a combined approach using different methodologies to improve the assembly accuracy of this highly repetitive genome. Subsequently, we performed a comparative analysis between 231 genome and the other *T. cruzi* genomes available in public databases to selected highly conserved genes across the different strains to perform phylogenetic analyses. Based on our results, we

propose a new hypothesis for the origin of the six *T. cruzi* DTUs and conclude that TcIII, unlike what has been proposed by several studies, is not a hybrid lineage.

Surface protein gene sequences, which are highly repetitives and variable, such as amastin, from different strains of the parasite were analyzed using an in-house script based on singular value decomposition and K-means along with a logistic regression analysis. Using this approach, we identified candidate regions that recapture the *T. cruzi* phylogenetic division. This methodology can be applied to other gene groups for the development of novel molecular characterization assays. The methodologies developed during this study and our results contributed to better understand the complexity of *T. cruzi* population structure.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AIC: Akaike information criterion
ALF: Automated Laser-Fluorescence
aLRT: Approximated Likelihood ratio test
BHI: Brain heart infusion
BIC: Bayesian information criterion
BI-PASA: Bidirectional PCR amplification of a specific Alelle
CDS: Coding sequence (sequência codificadora)
COII: Citocromo oxidase subunidade 2
CR4: C rich region 4 (região 4 rica em C)
DNA: Deoxyribonucleic (Ácido desoxirribonucleico)
dNTP: Deoxynucleotide triphosphate
DTUs: Discrete typing or taxonomic units
fg: Fentograma
gRNA: RNA guia
GTR: Generalised time reversible substitution model
HW: Hardy-Weinberg
ICORN: Iterative Correction of Reference Nucleotide
IDT: Integrated DNA technologies
IMAGE: Iterative Mapping and Assembly for Gap Elimination
Indel: Insersion and/or deletion (inserção e/ou deleção)
ITS: Internal transcribed spacer
JTT: Jones-Taylor-Thornton substitution model
Kb: Kilobases (1000 bases)
kDNA: Kinetoplast DNA (DNA do cinetoplasto)
LD: Linkage desequilibrium (desequilíbrio de ligação)
LIT: Liver infusion triptose
Ln: log likelihood
Mb: Mega bases (1,000,000 bases)
MG: Minas Gerais
ML: Maximum likelihood (Máxima verossimilhança)
mRNA: RNA mensageiro
mya: million years ago (milhões de anos atrás)
NCBI: National Center for Biotechnology Information
ND1: NADH desidrogenase subunidade 1
ND4: NADH desidrogenase subunidade 4
ND7: NADH desidrogenase subunidade 7
NGS: Next Generation Sequence
NJ: Neighbor joining
ORF: Open Reading Frame
Pb: Pares de bases
PCR: Polimerase chain reaction
RAPD: Random Amplified Polymorphic DNA
RATT: Rapid Annotation Transfer Tool
RFLP-PCR: Restriction fragment length polymorphism
RNA: Ribonucleic acid (Ácido ribonucleico)
rRNA: RNA ribossômico
SNP: Single-nucleotide polymorphism (polimorfismo de nucleotídio único)
SVD: Singular value decomposition (decomposição singular de valores)
Tc: *T. cruzi*
TcMCA: *T. cruzi* most recent common ancestor (ancestral recente mais comum)
UFMG: Universidade Federal de Minas Gerais

SUMÁRIO

INTRODUÇÃO

I. INTRODUÇÃO

1) *Trypanosoma cruzi* e doença de Chagas

A tripanossomíase americana ou doença de Chagas, causada pelo protozoário *Trypanosoma cruzi*, foi descoberta pelo pesquisador Carlos Ribeiro Justiniano Chagas no Rio de Janeiro, em 1909. As descobertas científicas feitas desde o descobrimento da doença de Chagas têm se processado de uma maneira gradativa, fazendo-se ainda necessário, a elucidação de muitos processos envolvidos na transmissão e tratamento da doença, assim como, no entendimento das diferentes manifestações clínicas e na entrada do parasito nas células do hospedeiro vertebrado.

T. cruzi, agente etiológico da doença de Chagas, é um protozoário flagelado, digenético, pertencente à ordem Kinetoplastidae, família Trypanosomatidae. Os membros desta ordem se caracterizam pela presença de uma mitocôndria única e alongada que alberga uma estrutura denominada de cinetoplasto. Em *T. cruzi*, o cinetoplasto apresenta-se volumoso e contém uma fonte de DNA extranuclear denominada de DNA do cinetoplasto ou kDNA (MYLER, 1993). Nele encontramos dois tipos de moléculas circulares, denominados de minicírculos e maxicírculos. O tamanho, a forma e a posição relativa do cinetoplasto são variáveis em diferentes estágios do desenvolvimento do parasito (BRENER, 1992).

O ciclo de vida de *T. cruzi* é bastante complexo, onde vários estágios morfologicamente distintos estão envolvidos (RODRIGUES, GODINHO e SOUZA, 2014). Eles estão presentes tanto no hospedeiro invertebrado, triatomíneos da família Reduviidae, quanto em hospedeiros vertebrados, mamíferos de diferentes ordens.

Nos triatomíneos, os parasitos se multiplicam ao longo do trato digestivo sob a forma epimastigota e se diferenciam para a forma infectante tripomastigota metacíclica quando chegam ao intestino posterior. Durante seu repasto sanguíneo nos hospedeiros vertebrados, os triatomíneos infectados liberam junto com as fezes ou urina as formas tripomastigotas metacíclicas, que por meio de uma lesão na pele ou diretamente pela mucosa penetram no hospedeiro vertebrado, invadem um número variado de células e iniciam seus ciclos intracelulares. No citoplasma das células, os parasitos diferenciam-se em amastigotas, o estágio proliferativo, que se multiplicam por divisões binárias sucessivas e diferenciam-se para a forma tripomastigota. A célula hospedeira se rompe e as formas tripomastigotas são liberadas na corrente sanguínea, podendo tanto infectar

novas células quanto serem ingeridas pelo hospedeiro invertebrado, completando assim o seu ciclo de vida (Figura 1).

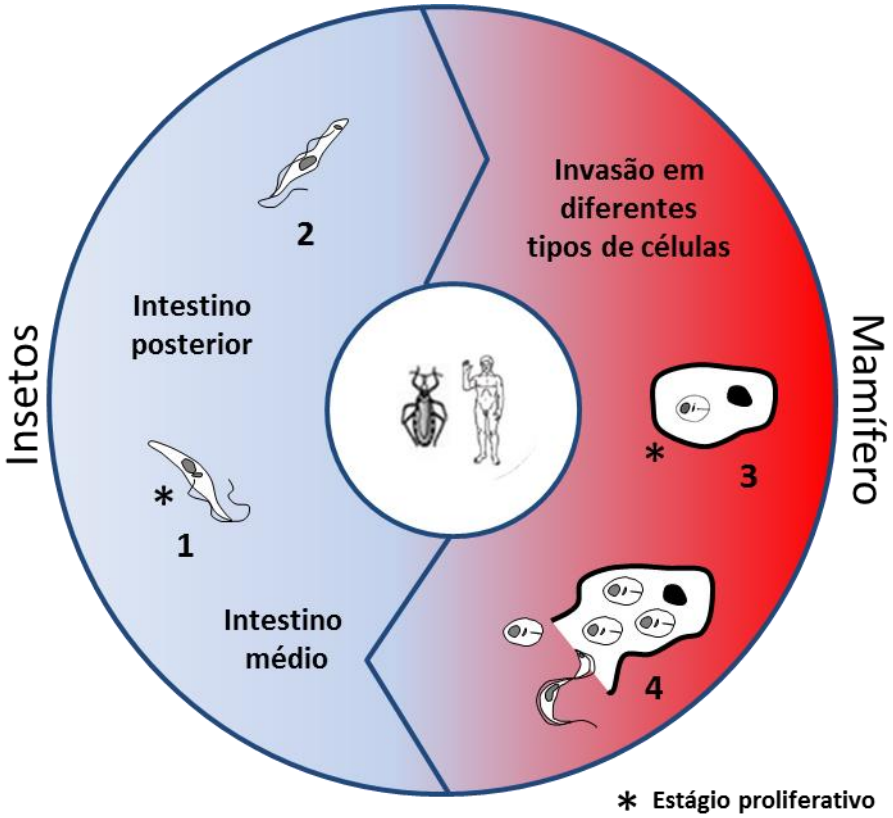

Figura 1. Ciclo de vida de *T. cruzi*. A figura mostra as várias formas tanto em hospedeiros invertebrados (insetos - Triatomíneos) quanto em vertebrados (mamíferos). 1- forma epimastigota; 2- forma tripomastigota metacíclica (forma infectante); 3- forma amastigota; 4- tripomastigotas.

1.1) Epidemiologia da doença de Chagas

A dispersão do parasito *T. cruzi* é bastante ampla no continente americano, estendendo-se desde a região sul dos EUA, até regiões meridionais do Chile e Argentina. Por sua vez, a doença de Chagas é mais restrita, limitando-se basicamente às áreas em que, por diferentes circunstâncias bioecológicas e sociais, ocorreu em algum momento a domiciliação de seus vetores invertebrados: os triatomíneos (VINHAES e DIAS, 2000). Apesar de haver mais de 100 espécies de triatomíneos, somente algumas espécies têm importância epidemiológica como origem regular de infecção humana (SCHMUNIS e YADON, 2010).

Neste aspecto, a incidência, prevalência e severidade da doença de Chagas têm sido altas na América Latina, principalmente nas áreas onde hemípteros da família Reduviidae, como *Triatoma infestans* e *Rhodnius prolixus*, são os principais vetores.

Durante o ciclo de vida de *T. cruzi*, hospedeiros de diferentes espécies podem ser infectados caracterizando os ciclos doméstico e silvestre da doença de Chagas.

O ciclo silvestre de *T. cruzi* envolve a interação de vetores e hospedeiros silvestres em biótopos naturais do continente americano. No âmbito silvestre, tem-se registrado mais de 100 espécies de pequenos mamíferos infectados naturalmente com *T. cruzi*, uma relação aparentemente muito antiga que proporciona um real equilíbrio entre parasitos e hospedeiros, em geral sem dano para as espécies (BARRETO, 1979; RYCKMAN, 1986). Num plano mais prático, maior importância é dada aos reservatórios capazes de aproximar-se dos seres humanos, caracterizando o chamado ciclo peridoméstico.

No ciclo peridoméstico intervêm animais domésticos que atuam como reservatórios de *T. cruzi* e triatomíneos silvestres atraídos às casas pela luz e pelo alimento. Esse ciclo tem grande importância epidemiológica, pois serve de ligação entre os ciclos doméstico e silvestre (SCHMUNIS e YADON, 2010).

O ciclo doméstico é considerado o de maior importância epidemiológica, já que perpetua a infecção em seres humanos (SCHMUNIS e YADON, 2010). Esse ciclo é resultante do contato entre o homem e o vetor envolvendo a colonização de biótopos artificiais pelos triatomíneos. Normalmente, essa colonização é resultante de modificações sociais e ecológicas no meio ambiente provocadas pelo homem (VINHAES e DIAS, 2000).

Mais de 100 anos após a sua descrição por Carlos Chagas, a doença de Chagas continua sendo um problema de saúde importante na América Latina, acometendo endemicamente cerca de 18 países e afetando 8 milhões de pessoas (SCHOFIELD, JANNIN e SALVATELLA, 2006). No Brasil, o número de pessoas infectadas por *T. cruzi* situa-se em torno de 3,5 milhões, destes, aproximadamente, 600 mil no Estado de Minas Gerais (DIAS, 2001). O custo social da doença também é expressivo. Cerca de 75.000 trabalhadores chagásicos apresentam cardiopatia grave, incapacitando-os para o trabalho e gerando absenteísmo de 2.250.000 dias de trabalho/ano (DIAS, 2001).

As ações de controle, centradas no combate ao vetor domiciliado através do tratamento químico das habitações infestadas, alcançaram toda a área endêmica a partir de 1983. O impacto sobre a transmissão vetorial foi evidente, obtendo-se amplos e

efetivos resultados no controle do vetor com a eliminação de *T. infestans* das casas e do ambiente peridomiciliar em áreas endêmicas, o que levou a Comissão Intergovernamental dos Países do Cone Sul considerar livre da transmissão vetorial por essa espécie os estados de Minas Gerais, Goiás, Mato Grosso, Mato Grosso do Sul, Paraíba, Pernambuco, Rio de Janeiro e São Paulo (VINHAES, 2002). Mesmo assim, a transmissão da doença de Chagas está longe de ser considerada um problema do passado. Em alguns países da América Latina a transmissão vetorial continua sendo um problema, e outros mecanismos de transmissão, como transplante de órgãos, acidentes de laboratório, transmissão por via oral e congênita vêm despontando como as principais responsáveis pela transmissão da endemia nos dias atuais.

1.2) Aspectos clínicos

A doença de Chagas apresenta um curso clínico bastante variado. Duas fases diferentes são consideradas ao longo do desenvolvimento da doença humana: a fase aguda e a fase crônica.

Na fase aguda da doença de Chagas, o parasito é encontrado em abundância no sangue e fluido cérebro-espinhal dos pacientes. Uma das características marcantes nessa fase da doença, e que pode ser observada em alguns pacientes, é a porta de entrada do parasito, que pode ser evidenciada pelo inchaço unilateral das pálpebras (sinal de Romaña), quando os parasitos penetram pela mucosa ocular.

As manifestações clínicas, nessa fase podem incluir febre, aumento do volume dos linfonodos, hepatomegalia, esplenomegalia, náuseas, vômitos, diarreia, anorexia e irritação das meninges (TANOWITZ et al., 1992). A parasitemia, durante a fase aguda é relativamente alta e a taxa de mortalidade nesta fase pode alcançar 10 – 15% em certas regiões, e parece estar relacionada com o inóculo e a diversidade da população de parasitos infectantes (SCHOFIELD, 1994).

A fase crônica da doença de Chagas compreende toda a vida do indivíduo depois dos primeiros 1-2 meses da infecção aguda, onde se observa parasitemia sub-patente (WENDEL et al., 1992). Seguindo um padrão de sintomas relacionados, esta fase pode ser subdividida em três ou quatro formas clínicas principais: indeterminada, cardíaca, digestiva, e em alguns casos, a forma cárdio-digestiva.

Na forma cardíaca os pacientes podem apresentar arritmias, falhas cardíacas, ou tromboembolismo. Nessa forma, o coração encontra-se muitas vezes dilatado e

hipertrofiado, havendo aneurisma apical e uma miocardite progressiva e fibrosante, o que caracteriza a cardiomiopatia chagásica. A forma digestiva é usualmente manifestada como megaesôfago e/ou megacólon. Na forma indeterminada, do ponto de vista clínico, não há comprometimento cardíaco ou digestivo. Pacientes portadores desta forma têm prognósticos melhores do que os pacientes sintomáticos, entretanto 2% a 5% dos pacientes com a forma indeterminada da doença convertem às formas cardíaca e/ou digestiva a cada ano (PRATA, 2001; UMEZAWA et al., 2001).

Embora ambas as formas cardíaca e digestiva apresentem aspectos patológicos similares, como inflamação e denervação, elas diferem quanto à prevalência e distribuição geográfica que variam entre países e, no interior de cada país, entre distintas áreas endêmicas (DIAS, 1992). No Brasil, cerca de 50 a 60% dos pacientes chagásicos crônicos apresentam a forma indeterminada da doença, 20 a 30% apresentam a forma cardíaca, 8 a 10% apresentam a forma digestiva e menos de 2% apresentam as formas clínicas combinadas. A ocorrência da forma digestiva é bastante desigual variando ainda entre as diversas regiões do Brasil, sendo predominante na região central (DIAS, 1992; LUQUETTI et al., 1986). Ainda permanece inexplicável por que diferentes pacientes desenvolvem as formas clínicas cardíaca, digestiva, cárdio-digestiva ou permanecem na forma indeterminada. Atualmente, acredita-se que uma das principais causas dessas diferentes manifestações clínicas está diretamente relacionada à variação genética de *T. cruzi*, não podendo descartar, dentro deste panorama, um possível papel para os aspectos ambientais, nutricionais, genéticos e imunológicos do hospedeiro (MACEDO et al., 2004).

2) Estrutura populacional em *T. cruzi*

2.1) Variabilidade intra-específica e linhagens filogenéticas

Vários estudos têm demonstrado que *T. cruzi* é um táxon muito heterogêneo (BUSCAGLIA e NOIA, 2003; DEVERA, FERNANDES e COURA, 2003; MACEDO e PENA , 1998; MACEDO et al., 2001). Devido a esta grande heterogeneidade e uma pressuposta ausência de reprodução sexuada em *T. cruzi*, conceitos básicos como espécie e cepa não são facilmente aplicáveis (MOREL, DEANE e GONÇALVES, 1986; TIBAYRENC, KJELLBERG e AYALA, 1990; TIBAYRENC et al., 1986, 1991). O termo cepa tem sido utilizado por vários autores para designar um isolado já estudado do parasito obtido a partir de um hospedeiro vertebrado ou invertebrado e representa um

conceito útil para os pesquisadores, em vista desta grande heterogeneidade observada entre os isolados de *T. cruzi*.

Durante muitos anos, vários grupos de pesquisadores, ao observarem a diversidade genética de diferentes populações de *T. cruzi*, procuravam agrupá-los com base em marcadores moleculares distintos. Vários nomes foram dados a estes grupos, de acordo com o marcador ou técnica utilizados, como Zimodemas - grupo de cepas que tem o mesmo perfil de isoenzimas (Z1, Z2 e Z3) (MILES et al., 1977, 1978, 1980) e Esquizodemas - padrões obtidos por polimorfismos de fragmentos de restrição do minicírculo (MOREL, et al., 1980), dentre outros.

Em abril de 1999, durante o Simpósio Internacional comemorativo dos 90 anos da descoberta da doença de Chagas, no Rio de Janeiro, a subdivisão da espécie *T. cruzi* em dois grupos ou linhagens principais foi reconhecida, tendo resultado na nomenclatura de *T. cruzi* I, primariamente associada com o ciclo silvestre da infecção, correspondendo ao grupo de izoenzimas Z1 e *T. cruzi* II relacionadas ao ciclo doméstico correspondente ao grupo Z2. Algumas cepas não puderam ser classificadas adequadamente em nenhum desses dois grupos, tendo permanecido sem linhagem definida correspondente ao grupo Z3. Pelo menos parte dessas cepas apresentavam características híbridas ou incongruentes com a nomenclatura definida de maneira que naquela época sua definição deveria ser decidida após estudos adicionais (ANNONIMOUS, 1999).

Estudos epidemiológicos subsequentes demonstraram evidências da associação preferencial da linhagem *T. cruzi* II com a infecção humana pelo menos no Brasil e Argentina (FREITAS et al., 2005; NOIA et al., 2002). Poucos foram os casos de infecção humana por *T. cruzi* I já descritos nesses países, e menos ainda, casos com sintomas clínicos (BUSCAGLIA e NOIA, 2003; COURA et al., 2002; TEIXEIRA et al., 2006). Por outro lado, Añes et al. (2004) demonstraram a ocorrência de infecção humana por cepas da linhagem *T. cruzi* I em pacientes venezuelanos com severas e fatais manifestações clínicas de doença de Chagas aguda.

Com o crescente desenvolvimento e aprimoramento das técnicas de biologia molecular, diferentes metodologias baseadas na PCR utilizando marcadores multilocais surgiram e foram utilizadas para reforçar a variabilidade genética de *T. cruzi* e buscar melhor definir a posição taxonômica das cepas que não puderam ser classificadas como *T. cruzi* I ou *T. cruzi* II, em 1999. Por exemplo, usando tipagens por isoenzimas e RAPD, foi proposto que as cepas de *T. cruzi* poderiam ser subdivididas não em duas,

mas em seis linhagens filogenéticas discretas referidas como DTUs I, IIa, IIb, IIc, IId e IIe (BRISSE, BARNABÉ e TIBAYRENC, 2000; BRISSE, VERHOEF e TIBAYRENC, 2001). O esquema DTU (*Discrete Typing Unit*) foi definido como uma coleção de cepas que são geneticamente relacionadas e que são identificadas por marcadores moleculares comuns. As cepas de *T. cruzi* I, neste sistema de classificação, permaneceram como uma linhagem única (DTU I), as cepas de *T. cruzi* II corresponderam a sublinhagem DTU IIb. No DTU II, além de reunir cepas da linhagem *T. cruzi* II, foram reunidas as demais cepas não classificadas no encontro de 1999 (DTUs IIa, IIc, IId e IIe).

Em uma análise posterior de sequências do gene mini-éxon, a fim de se descobrir o porquê de algumas cepas do parasito não apresentarem produto de amplificação em um ensaio de caracterização desenvolvido por Souto e cols. (1996), indicou-se a presença de um terceiro grupo caracterizado por "*indels*" de aproximadamente 50 pares de bases na região espaçadora destes genes representado por cepas caracterizadas por isoenzimas como Z3 ou DTU IIc (FERNANDES, et al., 1998). Assim, conjuntos de iniciadores foram desenhados para diferenciar esse novo grupo de cepas (BURGOS et al., 2007; FERNANDES et al., 2001), tornando-se assim possível descrever novos aspectos da estrutura populacional de *T. cruzi*, especialmente, no que diz respeito à caracterização de uma terceira linhagem ancestral, denominada então de *T. cruzi* III (FREITAS et al., 2006). Nesse trabalho, essa nova linhagem apresentou-se constituída por cepas pertencentes ao grupo, até então, não definido de *T. cruzi*, as quais foram correspondentes à sublinhagem IIc proposta por Brisse et al. (2000 e 2001).

Visto o crescente número de novos achados sobre a estrutura da população de *T. cruzi*, em 2009, durante a reunião científica em comemoração ao centenário da descoberta da doença de Chagas, propôs-se nova classificação em que *T. cruzi* foi subdividido em seis unidades taxonômicas discretas (DTUs), foi reconhecida e denominada de TcI a TcVI (ZINGALES et al., 2009). A relevância clínica e epidemiológica dessa subdivisão, os aspectos filogenéticos das relações entre os membros destes grupos e entre os diferentes grupos ainda estão sob investigação, mas já foi observada associação parcial das linhagens com o seu ciclo de transmissão e a sua distribuição geográfica (MILES et al., 2009; ZINGALES et al., 2012)

Várias técnicas de tipagem molecular estão sendo desenvolvidas, para a caracterização das seis linhagens principais de *T. cruzi*, a fim de se esclarecer os aspectos epidemiológicos e patogênicos associados especificamente a cada uma delas.

No entanto, nenhum dos marcadores já descritos isoladamente, permite uma resolução completa das seis linhagens. Ademais, o recurso de um único marcador é até o mesmo desaconselhável devido à consequente perda de resolução resultante de troca potencial de material genético entre algumas linhagens (LEWIS et al., 2009). Por outro lado, a necessidade do uso de diversos marcadores e metodologias torna a caracterização complexa, e por isso padronizar nova estratégia de genotipagem simples, de baixo custo e reprodutiva aplicável em qualquer laboratório básico de trabalho em *T. cruzi* é desejada.

2.2) Genética de populações de *Trypanosoma cruzi*

Como já mencionado, existe uma grande heterogeneidade intra-específica em *T. cruzi*. E, mesmo com os consideráveis progressos da biologia celular, molecular e da genética evolutiva, desde 1990, o debate sobre a estrutura populacional e modo reprodutivo de *T. cruzi* está longe de ser resolvido (PINTO et al., 2012; ZINGALES et al., 2012). Nesse contexto, duas hipóteses coexistem: reprodução clonal com trocas genéticas eventuais e a reprodução sexual mais frequente contribuindo para a definição da estrutura populacional do parasito. Análises de genética de população, baseadas em polimorfismos genéticos, a partir de amostras oriundas de localidades geográficas dispersas, têm demonstrado que o número observado de genótipos em *T. cruzi* é bastante inferior ao das combinações teóricas esperadas para cada *locus*, caracterizando um grande desvio dos valores esperados para organismos se reproduzindo de maneira aleatória, ou seja, em equilíbrio de Hardy-Weinberg (HW). Além disso, foi observado um forte desequilíbrio de ligação (LD) entre marcadores genéticos independentes nesses mesmos estudos. Esses achados levaram a proposição de um modelo de estrutura predominantemente clonal para as populações de *T. cruzi* (TIBAYRENC e AYALA, 2002).

Esse modelo tem sido amplamente aceito pela comunidade científica, pois se baseia na ausência de panmixia nas populações estudadas, na reprodução predominante por divisões binárias, a presença de desvios no equilíbrio de Hardy-Weinberg, no desequilíbrio de ligação e no conceito de que os descendentes são geneticamente idênticos aos pais observados (BRISSE, BARNABÉ e TIBAYRENC, 2000; FREITAS et al, 2006a; MACEDO et al., 2001; OLIVEIRA et al., 1998; ROUGERON et al., 2010;

TIBAYRENC e AYALA, 1987; TIBAYRENC, 1995; TIBAYRENC et al., 1986, 1993).

Apesar de o modelo clonal ter sido amplamente aceito pela comunidade cientifica, o *T. cruzi* ainda é considerado como um paradigma para microorganismos patogênicos eucarióticos classificados como clonais. Por exemplo, a ocorrência de cepas híbridas em populações naturais de *T. cruzi* sugere que eventos de reprodução sexuada ocorreram, pelo menos no passado recente, e contribuíram de forma significativa para a estrutura genética das atuais populações de *T. cruzi*. Quando essas evidências foram inicialmente descritas, concordou-se que o modelo de reprodução assexuada determinado por estudos de genética de população, não excluíam a presença de eventos de recombinação genética ocasional, os quais foram fortemente reforçados em vários estudos subsequentes (LEWIS et al., 2011a). Dentre esses achados merece destaque a detecção de um excesso de homozigotos em detrimento dos heterozigotos, verificado nas diferentes populações deste parasito, que contrasta com a expectativa esperada para populações essencialmente assexuais (OLIVEIRA et al., 1998). Em organismos assexuais os dois alelos evoluem independentemente por acúmulo de mutações e tendem, portanto, a um aumento da proporção de indivíduos heterozigotos na população. Esse efeito, conhecido como efeito Meselson, é o que ocorre, por exemplo, nos rotíferos que se reproduzem assexualmente por divisão binária (revisado por MACEDO et al., 2004; MACHADO et al., 2006).

Com base nessas observações a hipótese clonal para *T. cruzi* tem sido contestada por alguns autores, entre eles o nosso grupo de pesquisa. A ausência de panmixia identificada por alguns autores para todo o táxon, não exclui a ocorrência de recombinação genética, visto os diferentes níveis de heterozigozidade observados em cada população (BASTIEN, BLAINEAU e PAGES, 1992). Naturalmente para que uma troca de material genético possa ser observada em uma população natural, deve haver a oportunidade de interação entre os diferentes indivíduos. Isso já foi observado, por exemplo, em relação à presença de homozigotos e de heterozigotos correspondentes nos estoques e clones de *T. cruzi* provenientes de um mesmo hospedeiro circulante em uma mesma área em populações naturais, proporcionando provas convincentes para a possibilidade da troca genética (BOGLIOLO, LAURIA-PIRES e GIBSON, 1996). O uso frequente de diferentes cepas isoladas de diferentes espécies de hospedeiros e provenientes de diferentes localidades nos trabalhos de estudos de determinação da estrutura populacional de *T. cruzi*, é, portanto, um fator importante a ser observado.

Em 2003, Gaunt *et al*. mostraram a presença de troca de material genético em *T. cruzi* produzindo clones híbridos *in vitro* através da mistura de duas populações de parasitos, cada uma carregando diferentes marcadores de resistência a drogas. Depois de alguns ciclos de passagem em cultura de células, observou-se a presença de um novo clone de *T. cruzi* duplo-resistente, mostrando a fusão de genótipos, perda de alelos, recombinação homóloga e herança uni-parental do maxicírculo. Isto corrobora com observações anteriores sobre a ocorrência de linhagens híbridas naturais, tais como o clone CL Brener, objeto do primeiro projeto genoma da espécie, que aparentemente se originou de uma fusão entre cepas pertencentes às linhagens TcII e TcIII (BRISSE et al., 2003; ELIAS et al., 2005; FLORES-LÓPEZ e MACHADO, 2011; FREITAS et al., 2006b; GAUNT et al., 2003; MACHADO e AYALA, 2001; WESTENBERGER et al., 2005).

Demonstrada a ocorrência de eventos de recombinação no táxon *T. cruzi*, um grande esforço tem sido dispensado no sentido de identificar as cepas parentais que originaram as cepas híbridas em *T. cruzi*. Entretanto, dependendo do marcador genético empregado, diferentes histórias evolutivas foram neste contexto reconstruídas (BURGOS et al., 2013; FLORES-LÓPEZ e MACHADO, 2011; FREITAS, DE et al., 2006b; STURM et al., 2003; WESTENBERGER et al., 2005). Duas hipóteses principais predominam: a primeira sugerindo a existência de dois eventos de hibridação, um deles envolvendo cepas das linhagens TcI e TcII e outro envolvendo TcII e TcIII (WESTENBERGER et al., 2005). A segunda hipótese, também sugerindo a ocorrência de dois eventos de hibridação, mas ambos entre cepas das linhagens TcII e TcIII (FREITAS et al., 2006). Mais recentemente, outras evidências reforçaram a hipótese da ocorrência de dois eventos de hibridização entre TcII e TcIII, com diferença somente em como essas linhagens divergiram ao longo dos anos. Essas evidências, diferentemente da hipótese proposta por Freitas *et al*. (2006), demonstraram que TcI e TcII poderiam ser os únicos ancestrais de todas outras linhagens, sendo TcIII e TcIV derivadas de TcI (BURGOS et al., 2013). Essas incongruências entre as hipóteses apresentadas, principalmente em relação a quais são as cepas ancestrais de todas as linhagens de *T. cruzi*, poderiam ser resolvidas através de estudos focados em cepas pertencentes às linhagens TcIII e TcIV, cujos genomas ainda não foram publicados (Figura 2). Uma observação interessante é que todas as cepas sabidamente resultantes de eventos de hibridização entre TcII e TcIII apresentaram DNA mitocondrial

exclusivamente originado de TcIII (FREITAS et al., 2006). Sugerindo que esta cepa seria a doadora de DNA mitocondrial nos híbridos.

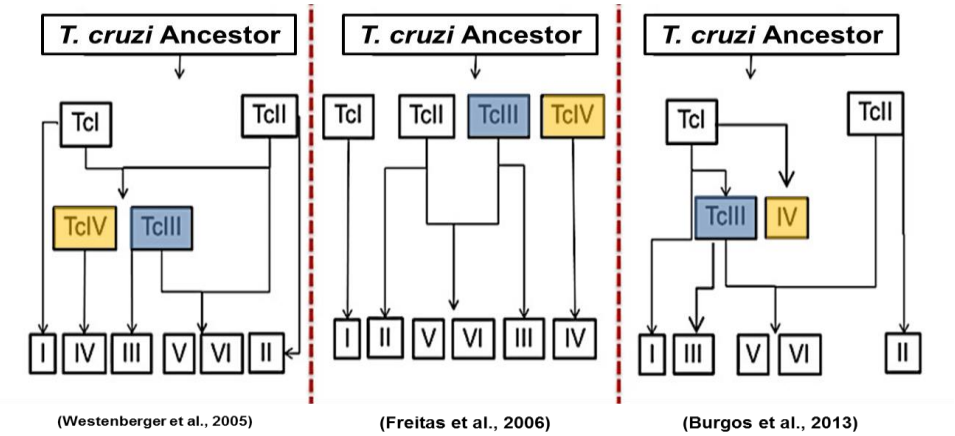

Figura 2. As três principais hipóteses existentes explicando a origem das seis linhagens de *T. cruzi*. Como pode-se observar existem incogruências entre elas, principalmente em relação à origem das linhagens TcIII e TcIV.

Mais recentemente, Lewis *et al*. (2011), através de uma reconstrução filogenética usando máxima verossimilhança e inferência bayesiana, dataram os principais eventos evolutivos do clado *T. cruzi*, incluindo o aparecimento das linhagens híbridas TcV e TcVI, que foi estimado ter ocorrido nos últimos 60.000 anos. Eles também encontraram evidências para a recente troca genética entre TcIII e TcIV e entre TcI e TcIV, dados, até então, não observados por outros grupos.

3) O genoma do *T. cruzi*

3.1) Genoma mitocondrial

Trypanosoma cruzi, assim como os demais tripanossomatídeos, apresentam uma única e grande mitocôndria contendo uma rede complexa de moléculas circulares de DNA denominada de cinetoplasto ou kDNA. O kDNA é composto por dois tipos de moléculas circulares denominadas de minicírculo e maxicírculo que diferem tanto no tamanho como na função (JUNQUEIRA, DEGRAVE e BRANDÃO, 2005; SILVEIRA, 2000).

Os minicírculos possuem cerca de 1.400 pares de base (pb) e estão presentes em torno de 10.000 a 20.000 cópias por célula. Neles estão contidas as regiões que transcrevem os pequenos RNAs, denominados RNAs guia (gRNAs), envolvidos no

processo de editoração (adição ou deleção de uridinas) dos mRNAs das enzimas mitocondriais desses organismos (HAJDUK e SABATINI, 1996; JUNQUEIRA, DEGRAVE e BRANDÃO, 2005; SILVEIRA, 2000; STUART, 1995).

Os maxicírculos, por sua vez, possuem cerca de 22 kb de tamanho e o número de cópias por célula varia de 20 a 50. Os genes de algumas proteínas da cadeia respiratória (ATPases, complexo do citocromo oxidase, NADH desidrogenase) e rRNAs mitocondriais estão localizados nessa molécula. Portanto, pode-se considerar o maxicírculo como um homólogo ao DNA mitocondrial dos demais eucariotos (JUNQUEIRA, DEGRAVE e BRANDÃO, 2005; SILVEIRA, 2000; WESTENBERGER et al., 2006). Uma característica importante dos maxicírculos, além da presença de vários genes, é a falta de alguns elementos-chave para a sua tradução, como códons de iniciação ou janelas abertas de leitura, o que é resolvido através da adição e/ou remoção de uridinas feito pós-transcricionalmente através do mecanismos de editoração de genes presentes nos minicírculos (SIMPSON, 1987; WESTENBERGER et al., 2006).

Já estão disponíveis em banco de dados públicos, como o do NCBI, as sequências completas de DNA da região codificante do maxicírculo de três cepas de *T. cruzi* sendo elas, as dos clones CL Brener (TcVI), Sylvio x10 (TcI) e Esmeraldo (TcII). Além das regiões codificantes, os maxicírculos de *T. cruzi* também apresentam regiões não codificantes variando de tamanho entre 4 a 6 kb, bem como a presença de um elemento apresentando sequência conservada que parece servir como a origem de replicação do maxicírculo e uma região variável e repetitiva não conservada entre cepas (RUVALCABA-TREJO e STURM, 2011; WESTENBERGER et al., 2006).

Estudos preliminares conduzidos pelo grupo da Prof.ª Bianca Zingales haviam demonstrado um aumento da expressão do gene NADH desidrogenase subunidade 7 (ND7), em cepas isoladas de pacientes apresentando a forma cardíaca da doença de Chagas em relação às cepas isoladas de pacientes com a forma indeterminada, sugerindo, um possível marcador para uso na via de diagnóstico e prognóstico da doença (BAPTISTA et al., 2006). Todavia, estudos posteriores do nosso grupo em colaboração com o grupo da Profa. Bianca Zingales não confirmaram aqueles primeiros achados (CARRANZA et al., 2009). Nesse ultimo estudo ficou demonstrado que, do ponto de vista genômico, havia três diferentes subtipos mitocondriais em cepas de *T. cruzi* II isoladas pacientes chagásicos independentemente da forma clinica apresentada, embora todos eles pertencentes ao haplogrupo ou clado mitocondrial C, específico de

cepas da linhagem TcII. Esses subtipos eram resultantes da presença ou não de deleções em genes das subunidades 4 e 7 da NADH desidrogenase (ND4/CR4 e ND7), (CARRANZA et al., 2009) (Figura 3).

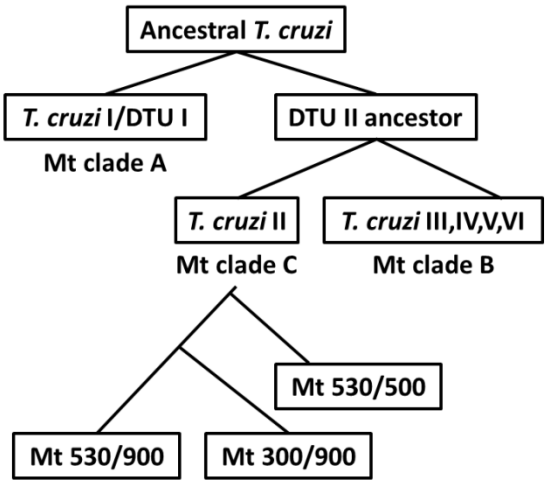

Figura 3. Cepas de *T. cruzi* II pertencentes ao clado mitocondrial C. É possível observar a existência de três perfis subtipos mitocondriais dependendo da presença ou ausência de deleções nos marcadores ND4/CR4 (300/530 pb) e ND7 (500/900 pb). Esquema modificado de Carranza *et al*. (2009).

3.2) Genoma nuclear

O conteúdo de DNA total por célula em *T. cruzi* varia de 125 a 330 fg sendo 16 a 30% pertencente ao DNA do cinetoplasto (HENRIKSSON, ASLUND e PETTERSSON, 1996; MCDANIEL e DVORAK, 1993).

A existência de poliploidia e um intenso polimorfismo cromossômico tanto em número quanto em tamanho entre cepas e clones *de T. cruzi* pode ser responsável por essas variações de conteúdo de DNA (DVORAK, 1993; HENRIKSSON, ASLUND e PETTERSSON, 1996; WAGNER e SO, 1990).

A organização da cromatina dos tripanossomas difere em vários aspectos daquela encontrada nos eucariotos superiores e mesmo em outros protistas. Devido à deficiência de proteínas histônicas semelhantes à H1, os cromossomos não se condensam durante a divisão celular, dificultando sua análise pelos métodos convencionais de citogenética (ASLUND et al., 1994; TORO e GALANTI, 1988). Em 2009, se propôs a existência de 41 pares de cromossomos para o clone CL Brener,

organismo modelo utilizado no primeiro Projeto Genoma de *T. cruzi* (WEATHERLY, BOEHLKE e TARLETON, 2009).

Estudos envolvendo a ploidia em *T. cruzi* têm apoiado a ideia de uma constituição predominantemente diploide para o genoma deste parasito. O primeiro genoma nuclear completamente sequenciado de *T. cruzi*, representado pelo clone CL Brener, foi publicado em 2005, tendo sido estimado um tamanho entre 106,4 e 110,7 Mb. Cerca de 60% (60,4 Mb) do genoma foi montado e anotado, sendo preditos 22.570 produtos gênicos dos quais pouco mais da metade (12.570) representam cópias alélicas. Pelo menos 50% deste genoma diploide é composto por sequências repetitivas, consistindo principalmente por grandes famílias de genes que codificam para proteínas de superfície, retrotransposons e repetições subteloméricas. (EL-SAYED, et al., 2005).

A comparação dos *contigs* de CL Brener com as *reads* provenientes do sequenciamento do genoma do clone Esmeraldo, permitiu distinguir dois haplótipos diferentes em CL Brener, os quais foram chamados de haplótipos Esmeraldo *like* e Non-Esmeraldo *like* com um grau de divergência de 5.4%. Estes dois haplótipos mostraram um maior nível de conservação nas regiões codificadoras de proteínas (diferença de 2.2%) do que nas regiões intergênicas, mostrando que a maior parte das diferenças foi devida a inserções/deleções em regiões intergênicas e subteloméricas e/ou amplificação de sequências repetitivas (EL-SAYED et al., 2005).

Mais recentemente, o mesmo estudo foi feito com o genoma do clone Sylvio X10/1 de *T. cruzi*, e não foram visualizadas diferenças significativas no conteúdo gênico entre os dois genomas, com exceção de seis ORFs (Open Reading Frames) presentes em CL Brener e ausentes em Sylvio X10/1 e uma diferença no tamanho do genoma do mesmo, que é menor que o de CL Brener (FRANZÉN et al., 2011). Além dessas sequências, outros genomas de *T. cruzi* já foram sequenciados, como é o caso do clone JR (TcI), que apesar de já ter sido sequenciado, ainda não teve seu estudo de comparação genômica publicado, e o clone CL14 (TcVI), que apesar de ser um clone distinto também foi isolado da cepa parental CL. O clone CL14 possui uma característica peculiar de ser um clone não virulento para animais diferentemente do clone CL Brener (ATAYDE et al., 2004). Os genomas dos clones derivados das cepas Arequipa (TcI), Colombina (TcI), Y (TcII) e CANIII (TcIV) também já estão sendo montados por outros grupos de pesquisa.

4) Justificativa

Ainda há muito a se discutir sobre a estrutura populacional de *T. cruzi*. Embora importantes avanços tenham sido feitos, muitas perguntas ainda precisam ser respondidas a fim de verificar a ocorrência e a frequência de recombinação entre parasitos, como *T. cruzi*. Uma grande dificuldade de avaliar este parâmetro se encontra nos delineamentos experimentais até então utilizados envolvendo a comparação de populações isoladas geograficamente. Nossa hipótese é que se há uma grande distância geográfica ou qualquer isolamento físico que impeça o livre encontro entre as populações analisadas, dependendo da dinâmica migratória, a chance de recombinação entre elas é pequena. Assim, em populações de mesma linhagem e pertencentes a uma única ou próxima região geográfica, haveria uma maior probabilidade de encontro destas populações e consequentemente maior frequência de eventos de recombinação entre elas. Ainda tendo em vista a possibilidade de ocorrência de recombinação entre populações naturais de *T. cruzi*, torna-se interessante buscar estudos mais aprofundados sobre as linhagens ancestrais deste táxon. Entre as linhagens já apontadas como possivelmente ancestrais, a linhagem TcIII encontra-se até então sem genoma sequenciado. O seu sequenciamento torna-se muito relevante para melhor caracterização da importância da mesma na origem e evolução de *T. cruzi*.

Para a realização de estudos populacionais temos à disposição um grande volume de dados que se encontram atualmente dispersos em diferentes bancos de dados como o *GeneBank*, *TriTrypDB*, entre outros. Desta forma, mesmo com as inúmeras sequências de genes e outros marcadores moleculares já disponíveis para *T. cruzi*, muitas dessas informações são subutilizadas ou erroneamente utilizadas em estudos populacionais do parasito. Para o melhor aproveitamento dessas informações torna-se necessário, além da geração de novos dados, desenvolver e refinar ferramentas experimentais e computacionais que facilitem a extração destas informações e que ajudem a melhor esclarecer aspectos da estrutura populacional, ecoepidemiologia e evolução de *T. cruzi*.

Para a geração de novos dados, outro problema encontrado hoje em dia é uma metodologia ideal para montagem de genomas altamente repetitivos, como o de *T. cruzi*, utilizando sequenciamento de nova geração. Em trabalhos anteriores, a estratégia escolhida para a base da montagem de genomas de *T. cruzi*, como uma tentativa de se minimizar este problema, optou-se pelo sequenciamento com a plataforma 454 que

produz *reads* mais longas, juntamente com a plataforma Illumina que produzem *reads* de maior qualidade, porém mais curtas (FRANZÉN et al., 2011). Apesar de bons resultados com esta metodologia híbrida, além de tornar o processo de sequenciamento caro e trabalhoso, está com os dias contados tendo em vista que recentemente foi anunciada a descontinuidade da plataforma 454. Torna-se então necessária a padronização de uma nova metodologia para montar estes genomas somente utilizando *reads* curtas obtidas em plataformas de sequenciamento de última geração.

Com os novos dados, a criação de uma ferramenta capaz de clusterizar e encontrar padrões para procura de alvos a serem usados para ensaios de caracterização molecular tornam-se necessária. Até porque mesmo existindo metodologias de caracterização das seis linhagens de *T. cruzi*, a mais eficaz delas até agora, faz uso de enzimas de restrição (*Alu*I), que aumentam muito o valor e o tempo de execução do ensaio. Com isso torna-se necessário otimizar e encontrar novos marcadores para estas metodologias experimentais de caracterização de linhagens filogenéticas que futuramente sejam utilizadas no estudo de populações presentes em tecidos dos pacientes chagásicos crônicos.

OBJETIVOS

II. OBJETIVOS

O objetivo geral deste projeto é refinar os estudos da estrutura populacional e evolução das linhagens principais de *Trypanosoma cruzi* através de abordagens moleculares e computacionais alternativas e mais contemporâneas. Para alcançarmos este objetivo no presente projeto propomos os seguintes objetivos específicos:

1- Avaliar a ocorrência de recombinação sexuada entre cepas distintas de *T. cruzi* da linhagem TcII oriundas de diferentes regiões geográficas.

2- Otimizar metodologias experimentais multiplex de caracterização de linhagens filogenéticas de *T. cruzi* adequadas ao estudo de populações presentes em tecidos dos pacientes chagásicos crônicos.

3- Sequenciar e desenvolver uma metodologia de montagem com dados provenientes da plataforma de nova geração Illumina, para a obtenção do genoma completo de um clone da cepa 231 da linhagem *T. cruzi* III.

4- Realizar um estudo de genômica comparativa entre o genoma de TcIII e os dados genômicos disponíveis de outras linhagens para resolver o papel e origem da linhagem TcIII meio às seis linhagens atualmente propostas.

5- Desenvolver novos algoritmos de predição de candidatos tetranucleotídicos responsáveis pela divisão de grupos gênicos em populações de *T. cruzi*.

MATERIAL E MÉTODOS

III. MATERIAIS E MÉTODOS

Neste trabalho a metodologia foi agrupada em função dos objetivos específicos para facilitar o acompanhamento posterior dos resultados.

1) Avaliação a presença de recombinação entre cepas de *T. cruzi* II

1.1) Cepas de *T. cruzi* utilizadas neste trabalho

Os isolados dos parasitos caracterizados neste projeto foram obtidos de 60 pacientes na fase crônica da doença residentes em diferentes cidades do estado de Minas Gerais, gentilmente cedidas pelo Prof. Dr. Egler Chiari, Laboratório de Biologia do *T. cruzi*, Departamento de Parasitologia, UFMG. Além dessas cepas foram também utilizadas outras 28 cepas de *T. cruzi* isolados de pacientes provenientes de diferentes estados brasileiros e alguns pertencentes a outros países da América Latina, já caracterizados em trabalhos anteriores (FREITAS et al., 2006a). Esses últimos isolados, obtidos de residentes de fora de Minas Gerais, foram utilizados como grupo controle para comparações nos cálculos de desequilíbrios de HW e LD. Todas as cepas utilizadas no presente projeto foram obtidas com o consentimento esclarecido dos pacientes seguindo procedimentos aprovados pelo Comitê de ética da UFMG 087/99, Belo Horizonte MG, Brasil.

Os parasitos foram obtidos por hemocultura (CHIARI et al., 1979) e para minimizar a seleção dos parasitos, as culturas foram mantidas em tubos individuais no laboratório num curto período de tempo e cultivadas por apenas duas passagens sucessivas em meio LIT (Liver Infusion Triptose). Após o crescimento, estas culturas foram lavadas 3 vezes com Krebs Ringer-Tris, pH 7.3, centrifugando a 2.000xg por 15 min a 4ºC e estocadas até a extração de DNA à -20ºC. A Tabela 1 mostra a relação das cepas de *T. cruzi* II obtidas de pacientes humanos utilizadas neste estudo e sua origem geográfica.

Tabela 1 – Cepas e clones de *Trypanosoma cruzi* II analisados na primeira etapa do trabalho.

Cepa/isolado[1]	Origem	Cepa/isolado[1]	Origem
002 B	MG/Brasil	**154 a**	MG/Brasil
003 a	MG/Brasil	**162 a**	MG/Brasil
005 B	MG/Brasil	**128 a**	BA/Brasil
007 B	MG/Brasil	**38**	MG/Brasil
009 B	MG/Brasil	**013 a**	MG/Brasil
010 B	MG/Brasil	**016 B**	MG/Brasil
011 B	MG/Brasil	**192 a**	MG/Brasil
012 a	MG/Brasil	**JG**	MG/Brasil
012 B	MG/Brasil	**022 b**	BA/Brasil
013 B	MG/Brasil	**003 B**	MG/Brasil
019 a	MG/Brasil	**002 a**	MG/Brasil
020 B	MG/Brasil	**005 a**	MG/Brasil
023 B	MG/Brasil	**006 a**	MG/Brasil
024 B	MG/Brasil	**021 B**	MG/Brasil
025 B	MG/Brasil	**097 a**	MG/Brasil
026 B	MG/Brasil	**188 a**	MG/Brasil
029 a	MG/Brasil	**Be62**[2]	MG/Brasil
031 a	MG/Brasil	**Esmeraldo**[2]	MG/Brasil
037 a	MG/Brasil	**GMS**[2]	MG/Brasil
044 a	MG/Brasil	**Ig539**[2]	MG/Brasil
045 a	MG/Brasil	**Mas1 cl1**[2]	MG/Brasil
050 b	MG/Brasil	**MPD**[2]	MG/Brasil
053 a	MG/Brasil	**Tula cl2**[2]	MG/Brasil
055 a	MG/Brasil	**84**[2]	MG/Brasil
058 a	MG/Brasil	**169/1**[2]	MG/Brasil
065 b	MG/Brasil	**200pm**[2]	MG/Brasil
067 a	MG/Brasil	**209**[2]	MG/Brasil
079 a	MG/Brasil	**239**[2]	MG/Brasil
083 a	MG/Brasil	**803**[2]	MG/Brasil
085 a	MG/Brasil	**1005**[2]	MG/Brasil
090 a	MG/Brasil	**1014**[2]	MG/Brasil
092 a	MG/Brasil	**1043**[2]	MG/Brasil
094 a	MG/Brasil	**1931**[2]	MG/Brasil
103 a	MG/Brasil	**GOCH**[2]	GO/Brasil
105 a	MG/Brasil	**577**[2]	GO/Brasil
109 a	MG/Brasil	**578**[2]	GO/Brasil
110 a	MG/Brasil	**580**[2]	GO/Brasil
115 b	MG/Brasil	**183744**[2]	GO/Brasil
116 a	MG/Brasil	**CPI95/94**[2]	PI/Brasil
120 a	MG/Brasil	**OPS27/94**[2]	PI/Brasil
129 a	MG/Brasil	**GLT564**[2]	RJ/Brasil
132 a	MG/Brasil	**GLT593**[2]	RJ/Brasil
138 a	MG/Brasil	**Y**[2]	SP/Brasil
146 a	MG/Brasil	**CPI11/94**[2]	Colombia

[1] A linhagens destas cepas de *T. cruzi* foram previamente determinadas pela amplificação do gene rDNA 24Sα (Souto *et al.*, 1996), do espaçador intergênico dos genes mini-exon (Burgos et al.,2007) e pelo RFLP do gene mitocondrial Citocromo Oxidase II (Freitas *et al*., 2006), e caracterizadas como *T. cruzi* II, de acordo com Zingales *et al*. (2009) realizada este trabalho. [2] Informações obtidas das amostras previamente caracterizadas por Freitas et al. (2006).

1.2) Extração de DNA

Para a extração de DNA 60 cepas de *T. cruzi* isoladas por hemoculturas foram submetidas a um processo de extração de DNA total com fenol:clorofórmio:álcool isoamílico (25:24:1) de acordo com Macedo *et al*. (1992). O DNA total foi precipitado usando 10% acetato de sódio 3M e etanol absoluto. Em seguida, o DNA foi ressuspendido em 20μL de água Milli-Q estéril e quantificado utilizando um gel de agarose 1% e padrões de concentração de DNA para comparação. Além disto, as amostras foram dosadas no Nanodrop (Thermo Scientific) para confirmar a dosagem em gel. Após a dosagem, uma alíquota da amostra concentrada foi diluída para 1ng/μL e o restante estocado para estoque. Para cada ensaio de PCR, foram usados 3μL de cada amostra diluída à 1ng/μL.

1.3) Caracterização das linhagens filogenéticas das amostras de *T. cruzi*

Para a determinação das linhagens filogenéticas das amostras de *T. cruzi* utilizadas neste trabalho, foi utilizado um protocolo constituído de três ensaios subsequentes, como proposto por Macedo e colaboradores (D'ÁVILA et al., 2009). Esta triagem se baseia incialmente pela análise por PCR-RFLP do gene Citocromo Oxidase subunidade II (COII) constituída da amplificação de um fragmento de aproximadamente 375pb do gene, seguida pela digestão com a enzima de restrição *Alu* I (FREITAS et al., 2006a), que nos permite discriminar as linhagens *T. cruzi* I e *T. cruzi* II das demais linhagens (*T. cruzi* III-VI). O segundo passo é realizado utilizando o polimorfismo de tamanho do espaçador intergênico dos genes mini-exon (BURGOS et al., 2007) aplicado neste caso para dividir as linhagens não classificadas na primeira etapa em dois subgrupos, um formado por *T. cruzi* III e *T. cruzi* IV e o outro por *T. cruzi* V e *T. cruzi* VI. A terceira e última etapa consiste na caracterização do polimorfismo de tamanho do gene rDNA 24Sα (SOUTO et al., 1996) que diferencia as amostras contidas em cada um desses subgrupos. O esquema dessa triagem pode ser conferido na Figura 4.

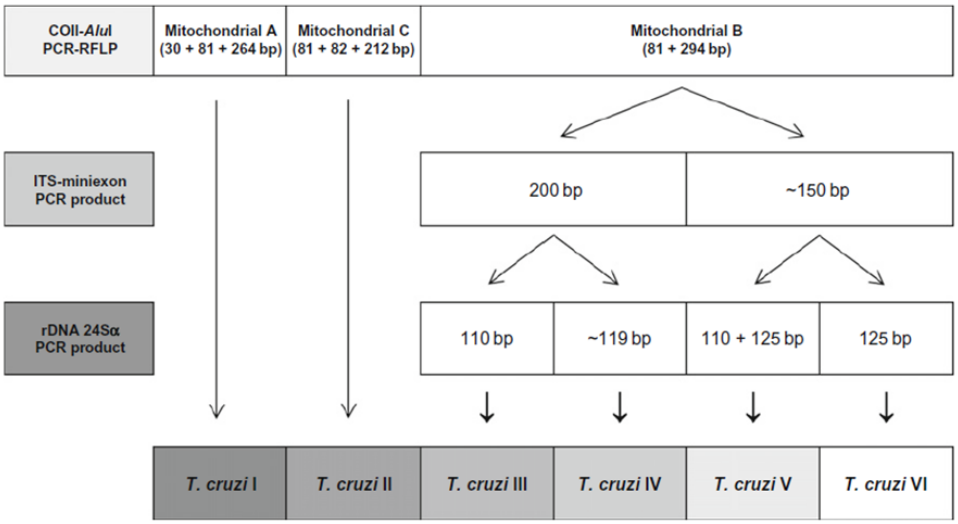

Figura 4. Ensaio triplo, tal como recomendado por Macedo e colaboradores, para discriminar as linhagens de *T. cruzi* (MACEDO e SEGATTO, 2010). Inicialmente, os isolados devem ser analisados por PCR-RFLP do gene COII (FREITAS et al., 2006a), permitindo a discriminação de TcI e TcII das demais linhagens (Tc III-VI). Um segundo passo compreendendo uma PCR do espaçador intergênico ITS do gene de mini-éxon (BURGOS et al., 2007) é aplicado aos isolados não classificados na primeira etapa, resultando em dois grupos distintos, um formado por TcIII e TcIV e um outro por TcV e TcVI. O passo final consiste na análise do polimorfismo de rDNA 24Sα por PCR (SOUTO et al., 1996). Os tamanhos dos fragmentos esperados para cada linhagem é apresentado em pares de base (pb).

1.3.1) Análises de polimorfismo do gene da subunidade II do citocromo oxidase de *T. cruzi*

Para a amplificação da região gênica que compreende a subunidade II da enzima citocromo oxidase mitocondrial (COII) de *T. cruzi*, o DNA das amostras do parasito foi adicionado a uma solução de PCR composta de 10mM Tris-HCl pH 8,8, 25mM KCl, 3,5mM $MgCl_2$, 1U *Taq DNA Polimerase Platinum* (Invitrogen), 250µM de cada dNTP, 0,3µM dos iniciadores TcMit10 e TcMit21 (Tabela 2), 3uL de DNA (1ng/µL) e quantidade de H_2O Milli-Q estéril suficiente para 15uL. O programa utilizado para a amplificação consiste de uma desnaturação inicial a 94ºC por 5 minutos, seguida de 40 ciclos constituídos de um passo de desnaturação por 45 segundos, anelamento à 45ºC por 45 segundos e extensão à 72ºC por 1 minuto.

Os produtos amplificados foram resolvidos em gel de agarose 2%. Posteriormente, 10µL da PCR foram submetidos a uma digestão empregando-se a enzima de restrição *Alu*I (Fermentas), conforme instrução do fabricante. A análise dos polimorfismos de tamanho dos fragmentos de restrição para este gene, foi realizada em

gel de poliacrilamida 6% corado pela prata. Como padrões de comparação dos RFLP, foram utilizados DNA de cepas e clones característicos das linhagens de *T. cruzi* I (Haplótipo A) e *T. cruzi* II (Haplótipo C). Entretanto, esse marcador não diferencia as demais linhagens, pois elas apresentam o mesmo perfil de restrição para o gene COII (Haplótipo B). Como padrão para os haplótipos A, B e C, foram utilizados respectivamente, o clone Col 17G2, o clone híbrido CL Brener e a cepa JG (FREITAS, et al., 2006a). O perfil de bandas esperado para cada linhagem é de 30/81/264 pb para amostras do haplótipo A, 81/82/212 pb para o haplótipo B e 81/294 pb para o haplótipo C.

1.3.2) Análise do polimorfismo do espaçador intergênico ITS do gene de mini-éxon de *T. cruzi*

Para a amplificação da região intergênica ITS do gene de mini-éxon, o DNA das amostras de *T. cruzi* foi adicionado a uma solução de PCR composta de 20mM Tris-HCl pH 8,4, 50mM KCl, 3mM $MgCl_2$, 250 µM de cada dNTP, 3µM de cada iniciador (TcIII e UTCC – Tabela 2), 1U *Taq DNA Polymerase Platinum* (Invitrogen), 1µL de DNA total (1ng/µL) e quantidade de H_2O Milli-Q estéril suficiente para 15µL. Os iniciadores utilizados (Tabela 2) reconhecem posições entre 368-386 pb e 546-570 pb da unidade repetitiva do mini-éxon de *T. cruzi* (BURGOS et al., 2007), quando bandas de aproximadamente 200 pb para amostras de *T. cruzi* III e *T. cruzi* IV e de 150 pb para as demais linhagens.

Os ciclos de amplificação da PCR consistiram de uma etapa de desnaturação inicial de 3 minutos a 94°C, seguida de 3 ciclos constituídos de uma fase anelamento à 68°C por 1 minuto, extensão dos iniciadores à 72°C e desnaturação à 94ºC por 1 minuto. A cada três ciclos a temperatura de anelamento foi sequencialmente diminuída para 66, 64, 62 e 60ºC. A esta última temperatura o número de ciclos foi aumentado para 35, sendo seguido por uma extensão final dos iniciadores a 72ºC por 10 minutos. A análise dos produtos amplificados, em gel de poliacrilamida 6% corado pela prata, permite distinguir a linhagens de *T. cruzi* III e IV (estas apresentam um *amplicon* de 200pb) das cepas das demais linhagens *T. cruzi* I, II,V e VI (estas apresentam um *amplicon* de 150-157pb) (BURGOS et al., 2007). Como padrão de comparação foram utilizados DNA de cepas ou clones característicos de cada linhagem, os clones 231 e CANIII foram utilizados como representantes do grupo com *amplicon* de tamanho de

200 pb e os clones Colombiana 17G2, JG, SO3Cl5 e CL Brener como representantes do grupo com *amplicon* de tamanho de 150-157 pb.

1.3.3) Análises do polimorfismo do gene rDNA 24sα de *T. cruzi*

Para a tipagem do rDNA 24S**α**, o DNA das amostras de *T. cruzi* foi adicionado a uma solução de PCR composta de 10mM Tris-HCl pH 9,0, 50mM KCl, 0,1% Triton X-100 (Buffer B, Promega), 3,5mM $MgCl_2$, 1U *Taq DNA Polymerase Platinum* (Invitrogen), 200μM de cada dNTP, 0,25μM de cada iniciador (D71 e D72 – descritos na Tabela 2), 2μL de DNA (1ng/μL) e quantidade de H_2O Milli-Q estéril suficiente para 12,5μL.

Os ciclos de amplificação consistiram de uma desnaturação inicial à 94ºC por 1 minuto, seguido de 30 ciclos constituídos de um passo de desnaturação por 30 segundos, um passo de anelamento à 60ºC por 30 segundos e um passo de extensão à 72ºC por 30 segundos (SOUTO et al., 1996).

Após a amplificação, uma alíquota de 5μL dos produtos da PCR foi analisada em gel de acrilamida 6% corado pela prata para a visualização dos amplicons. Como controles das amplificações foram utilizados DNA de cepas ou clones padrão que representam as linhagens de *T. cruzi* I (tamanho esperado de 110), *T. cruzi* II (tamanho esperado de 125pb) e *T. cruzi V* (tamanhos esperados de 110 e 125pb).

1.4) Caracterização dos genes mitocondriais ND4 e ND7 das amostras de *T. cruzi*

Para a análise de polimorfismos do DNA mitocondrial, amplificações por PCR da sequência de ND7 e de ND4/CR4 foram realizadas utilizando os iniciadores ND7-direto e ND7-reverso e ND4-direto e ND4-reverso (Tabela 2). Os ciclos de amplificação dos genes ND7 e ND4/CR4 consistiram de uma desnaturação inicial à 94ºC por 2 minutos, seguido de 30 ciclos constituídos de um passo de desnaturação à 94ºC por 1 minuto, um passo de anelamento à 60ºC (ND7) e à 55ºC (ND4/CR4) e por 1 minuto e um passo de extensão à 72ºC por 1 minuto, seguidos de um passo final de extensão de 10 minutos à 72ºC (BAPTISTA et al., 2006; CARRANZA et al., 2009). Os produtos de PCR de ND7 e ND4/CR4 genes foram separados em géis de agarose a 1% e corados com brometo de etídio.

1.5) Caracterização de microssatélites nas amostras de *T. cruzi*

O DNA de cepas de *T. cruzi* provenientes somente de hospedeiros humanos foi submetido a ensaios de PCR com cinco *loci* de microssatélites compostos por repetições de dinucleotídeos, três loci de trinucleotídeos e um locus de tetranucleotídeo. Totalizando para esse estudo nove marcadores polimórficos, alguns fisicamente ligados, sendo eles SCLE11, SCLE10, MCLE01, MCLF10, MCLG10, TCAAT8, TCATT14, TCTAT20 e TCAAAT6 (OLIVEIRA, RIVA P et al., 1998; VALADARES et al., 2008). As sequências dos iniciadores utilizados para a amplificação desses *loci* de microssatélites no genoma de *T. cruzi* estão descritas na Tabela 2.

Os ciclos de amplificação para todos os *loci* microssatélites consistiram de uma etapa de desnaturação inicial a 94ºC por 5 minutos seguida de 5 ciclos constituídos de uma etapa de anelamento a 58ºC por 30 segundos, extensão a 72ºC por 1 minuto e desnaturação a 94ºC por 30 segundos. A cada cinco ciclos a temperatura de anelamento foi diminuída para 55, 53, 51 e 48ºC. A esta última temperatura, o número de ciclos foi aumentado para 15, sendo seguido por uma extensão final dos iniciadores a 72ºC por 10 minutos.

Tabela 2 - Sequências dos iniciadores utilizados neste projeto.

Iniciador	Sequência
ND4-direto	5’-AAACTCTATCTTTCGAAAACCC-3’
ND4-Reverso	5’-GGGAAAAATAGACTTTCAAAAAGTATC-3’
ND7-direto	5’-AAGAAAAGAGGGGACAAACG-3’
ND7-reverso	5’-AAAAATCCCCTTCCAAAAGC-3’
SCLE10-direto	5’-GATCCCGCAATAGGAAAC-3'
SCLE10-reverso	5’-FluoresceínaGTGCATGTTCCATGGCTT-3'
SCLE11-direto	5’-FluoresceínaACGACCAAAGCCATCATT-3'
SCLE11-reverso	5’-GATGCTAACTGCTCAAGTGA-3'
MCLF10-direto	5’-FluoresceínaGCGTAGCGATTCATTTCC-3'
MCLF10-reverso	5’-ATCCGCTACCACTATCCAC-3'
MCLG10-direto	5’-FluoresceínaAGGAGTCAAATATAATGAGGCA-3’
MCLG10-reverso	5’-ACGTGTGAAAGGCATCTATC-3'
MCLE01-direto	5’-FluoresceínaCTGCCATGTTTGATCCCT-3'
MCLE01-reverso	5’-CGTGTACATATCGGCAGTG -3'
TCAAT8-direto	5'-FluoresceínaACCTCATCGGTGTGCATGTC-3'
TCAAT8-reverso	5'-TATTGTCGCCGTGCAATTTC-3'
TCATT14-direto	5'-FluoresceínaTTATGGATGGGGTGGGTTTG-3'
TCATT14-reverso	5'-AGCAATAATCGTATTACGGC-3'
TCTAT20-direto	5'-FluoresceínaGATCCTTGAGCAGCCACCAA-3'
TCTAT20-reverso	5'-CAAATTCCCAACGCAGCAGC-3'
TCAAAT6-direto	5'-FluoresceínaGCCGTGTCCTAAAGAGCAAG-3'
TCAAAT6-reverso	5'-GGTTTTAGGGCCTTTAGGTG-3'
D71-direto	5’-AAGGTGCGTCGACAGTGTGG-3´
D72-reverso	5’-TTTTCAGAATGGCCGAACAGT-3´
TcMit10	5’-CCATATATTGTTGCATTATT-3’
TcMit21	5’-TTGTAATAGGAGTCATGTTT-3’
TcIII-direto	5’-CTCCCCAGTGTGGCCTGGG-3’
UTCC-reverso	5’-CGTACCAATATAGTACAGAAACTG-3’

1.5.1) Determinação do tamanho dos alelos de microssatélites

Para a determinação do tamanho dos alelos dos microssatélites, uma alíquota de 1 a 3µl do produto da PCR, utilizando iniciadores fluorescentes, foi desnaturada a 90°C por 3 minutos e, depois, submetida a uma eletroforese em gel de poliacrilamida 6% desnaturante (8M uréia) no sequenciador automático de DNA A.L.F (GE Healthcare, Milwaukee, Wisconsin, USA). A corrida foi feita por 10 horas, a 45°C.

Como padrão externo das corridas eletroforéticas, foram utilizadas escadas alélicas contendo uma mistura de fragmentos fluorescentes de pesos moleculares variando de 50 em 50pb, (sizer 50-500, GE Healthcare), na concentração de 5fmol/ fragmento. Para cada canaleta aplicada, incluindo aquelas contendo os padrões externos, foram acrescentados fragmentos fluorescentes de tamanhos conhecidos (75, 155, 210 e 320pb) que foram utilizados como padrões internos nas corridas.

Os resultados obtidos em forma de cromatogramas foram analisados com o auxílio do programa *Allelelocator* versão 1.03 (GE Healthcare). Os fragmentos de tamanhos conhecidos, presentes nos padrões externo e interno da corrida, foram corretamente alinhados e os tamanhos dos alelos para os *loci* de microssatélites foram calculados para cada amostra.

1.6) Avaliação de parâmetros de genética de populações

As análises de alguns parâmetros de genética de populações foram feitas baseando-se nos dados obtidos com as amplificações dos *loci* de microssatélites em diferentes cepas de *T. cruzi*. A partir desses dados de genótipos obtidos das populações, haplótipos foram estimados pelo programa PHASE (STEPHENS, SMITH e DONNELLY, 2001) e estes foram projetados em uma rede de haplótipos utilizando software Network 4.6 (FORSTER et al., 2000).

Subestruturação populacional foi inferida pelo programa STRUCTURE (PRITCHARD, STEPHENS, e DONNELLY, 2000). A estimativa do número apropriado de subpopulações presentes no plano amostral foi determinada por dois métodos: um utilizando os valores log-likelihood (Ln) para K 1-10 e o outro com a equação de ΔK (EVANNO, REGNAUT e GOUDET, 2005).

Utilizando o programa ARLEQUIN v3.1 (EXCOFFIER, LAVAL e SCHNEIDER, 2007) foram calculadas, para cada *locus* de microssatélite descrito, a estimativa do equilíbrio de Hardy-Weinberg (HW) e as heterozigosidades observada e esperada, com seu correspondente intervalo de confiança em 95%. Com o programa GENEPOP (ROUSSET e RAYMOND, 1995) investigamos o nível de associação entre alelos de diferentes *loci*, aplicando testes de desequilíbrio de ligação (LD) entre os *loci* de microssatélites estudados, em todas as possíveis comparações par a par.

2) Padronização de uma nova técnica de caracterização molecular: amplificação alelo específica por PCR bidirecional (Bi-PASA)

2.1) Amostras

Foram utilizadas, neste projeto, aproximadamente 200 sequências parciais dos genes da citocromo oxidase subunidade 2 (COII) concatenado ao da subunidade 1 da NADH desidrogenase (ND1) de *T. cruzi* depositadas no GenBank por diferentes grupos

de pesquisa (LEWIS et al., 2011b; MACHADO e AYALA, 2001; MESSENGER et al., 2012; OCAÑA-MAYORGA et al., 2010; SUBILEAU et al., 2009). Esse gene foi escolhido, pois além de ter muitas sequências disponíveis para análise, estudos realizados por Freitas et al. (2006) demonstraram através de uma abordagem filogenética de *Neighbor-Joining* (NJ) que a sequência do gene ND1 de *T. cruzi* servia como um marcador para separar as cepas de *T. cruzi* em três agrupamentos filogenéticos TcI, TcII e TcIII-VI. As sequências utilizadas neste trabalho foram originadas de isolados e clones pertencentes às seis linhagens ou DTUs de *T. cruzi*. Todas as sequências utilizadas foram inseridas em um arquivo MultiFASTA e alinhadas para a busca e identificação de sítios polimórficos adequados que permitissem o desenho de iniciadores capazes de diferenciar as cepas das linhagens filogenéticas de *T. cruzi* já caracterizadas.

Com o alinhamento das sequências de DNA foi verificado que na posição 798 das sequências existiam polimorfismos de nucleotídeo único (SNPs), característicos das linhagens filogenéticas em *T. cruzi*: o desoxirribonucleotídeo T em cepas da linhagem *T. cruzi* I; o desoxirribonucleotídeo A em cepas da linhagem *T. cruzi* II; e o desoxirribonucleotídeo C típica de cepas das linhagens *T. cruzi* III a VI (Figura 3).

2.2) Bi-PASA

A técnica de Bi-PASA (*Bidirectional PCR Amplification of Specific Alleles*) foi primeiramente descrita por Liu *et al*. (1997) e permite o diagnóstico de qualquer diferença alelo específica conhecida no DNA genômico de forma simples, rápida, confiável e de baixo custo. É uma adaptação da PCR onde dois ou mais iniciadores "*forward*" são desenhados com tamanhos diferentes e extremidades 3'-OH complementares aos alelos específicos e um iniciador reverse comum e totalmente complementar a todos os alelos, resultando assim em fragmentos de tamanhos distintos para cada alelo. A técnica é baseada no princípio de que a *Taq* DNA polimerase não apresenta atividade 3'→ 5' exonuclease, de modo que o mal pareamento entre a extremidade 3' do iniciador e o DNA molde resulta na impossibilidade de amplificação.

O desenho dos iniciadores foi realizado com a ajuda do programa OligoAnalyzer (IDT – Integrated DNA Technologies). Foram desenhados três iniciadores diretos específicos para cada linhagem filogenética contendo os SNPs característicos na extremidade 3´-OH e apenas um iniciador reverso comum a todas as linhagens. À

extremidade 5’ dos iniciadores diretos foram adicionadas caudas de tamanhos distintos com sequências de DNA oriundas do vetor de clonagem pUC18. Estas caudas têm como objetivo aumentar diferencialmente o tamanho dos produtos obtidos após os ensaios de PCR e com isso garantir uma melhor separação destes produtos em gel de poliacrilamida corado pela prata (Figura 5). Os iniciadores utilizados nessa técnica estão descritos na Tabela 3.

Tabela 3 - Sequências dos iniciadores utilizados neste projeto.

Iniciador	Sequência
ND11F-direto	5’-GTAATACGGTTATCCACAGAATCAGGTTAAATTTACTTTATTTATAATTGGT -3’
ND12F- direto	5’-TTGCTTTATTTGTAATTGGA -3’
N13HF-direto	5’-CGAGCGGTATCAGCTCTTCACTTTATTTATAATTGGC -3’
ND1R-reverso	5’-TCAAAAAGAATAATAAATCCTAA-3’

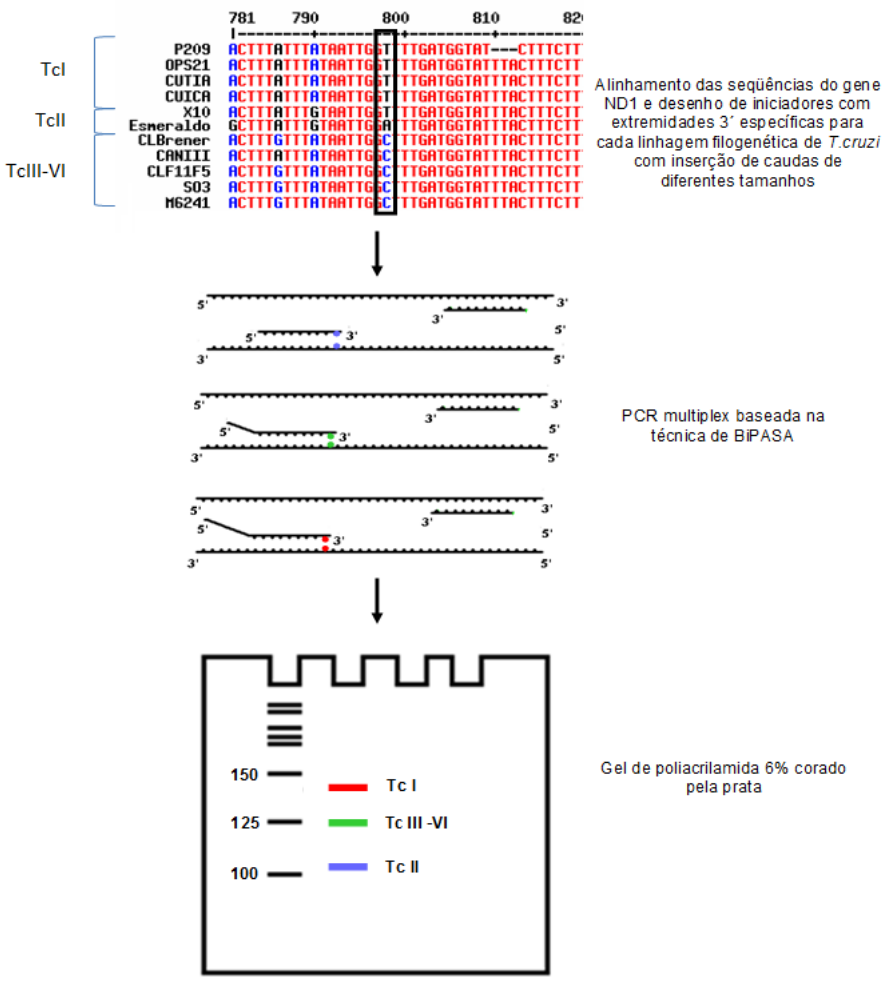

Figura 5. Esquema representativo do protocolo de caracterização molecular de *T. cruzi* baseado na técnica de Bi-PASA envolvendo o gene mitocondrial ND1 do parasito.

Na padronização dos ensaios de Bi-PASA-ND1 PCR foram avaliadas as seguintes variáveis:

- Tampões para PCR (concentração de NaCl, KCl, $MgCl_2$, Triton X-100 e pH);
- Concentração de iniciadores e de dNTPs;
- Marca e quantidade da *Taq DNA Polymerase*;
- Condições dos ciclos da PCR.

Os ensaios para a padronização do Bi-PASA-ND1 foram realizados inicialmente utilizando 1ng de DNA de cepas de *T. cruzi* pertencentes a diferentes linhagens filogenéticas, bem como misturas de DNA das cepas. Após a obtenção de uma reação em que todos os fragmentos foram amplificados satisfatoriamente, foram realizados testes de especificidade para as linhagens.

3) Genômica comparativa entre genomas diferentes linhagens de *T. cruzi* para estudos evolutivos:

3.1) A cepa 231

A cepa 231 de *T. cruzi* foi isolada por hemocultura de um paciente, na fase crônica da doença de Chagas, residente na região endêmica de Bambuí em Minas Gerais (CHIARI et al., 1979). Essa cepa, que faz parte do acervo de cepas criopreservadas em nitrogênio líquido no Laboratório de Biologia do *Trypanosoma cruzi* da UFMG, foi cedida gentilmente pelo professor Egler Chiari para a realização de nossos estudos.

3.2) Clonagem da cepa

Para o desenvolvimento deste trabalho, a cepa 231 de *T. cruzi*, pertencente à linhagem TcIII, e que será utilizada no sequenciamento completo, foi clonada em meio de cultura seguindo protocolo previamente por descrito Gomes et al., (1991). Neste protocolo foi colocado em cada placa meio LIT 48.4% (Liver Infusion Triptose), BHI 48.4% (Brain Heart Infusion), agarose low melting 0.75% e sangue desfibrinado e com o complemento inativado 2.5%. Após dois dias de incubação, para verificar se houve alguma contaminação, 100µL da cultura de parasito a ser clonado foi adicionada em cada placa e incubada por 30 dias à 28ºC.

Após os 30 dias, as colônias de parasitos foram capturadas com a ponta de uma ponteira e mantidas em meio LIT para crescimento, quando atingiu um número ideal para extração, em geral 10^8 parasitos/ml, e teve seu DNA extraído pela metodologia de fenol:clorofórmio:álcool isoamílico (25:24:1) (MACEDO et al., 1992). A clonagem foi feita para minimizar a chance de mais de uma população de parasitos estarem presentes na cultura.

3.3) Sequenciamento do genoma: montagem e anotação

O sequenciamento foi realizado utilizando a plataforma de sequenciamento de nova geração Illumina Hiseq 2000 pela empresa Fasteris, gerando uma biblioteca com insertos com um tamanho aproximado de 300 pb. Deste sequenciamento foram gerados cerca de 55,031,792 *pair-end reads* com um tamanho igual a 100 pb cada. Estas *reads* foram trimadas para uma qualidade mínima de 35, na escala Illumina 1.8+ Phred+33,

utilizando o programa Trimommatic (BOLGER, LOHSE e USADEL, 2014). Estas *reads* trimadas foram usadas em nosso *pipeline* de montagem (Figura 6). Nosso *pipeline* é composto principalmente por duas diferentes abordagens. Na primeira abordagem, nós realizamos uma montagem guiada por um genoma referência que foi realizado utilizando o programa Bowtie (LANGMEAD et al., 2009), SAMtools (LI et al., 2009) e o pacote VCFtools (DANECEK et al., 2011). Esta abordagem, apesar de não aproveitar regiões altamente polimórficas de nosso genoma em relação ao genoma referência, ajuda a mapear regiões altamente repetitivas, que são um desafio em outras abordagens. Para esta abordagem de montagem, nós usamos a referência de *T. cruzi* CL Brener (TcVI) do haplótipo Non-Esmeraldo *Like* que é descrita como mais próximo de TcIII (90% de identidade) do que das demais linhagens já sequenciadas (aproximadamente 55% de identidade).

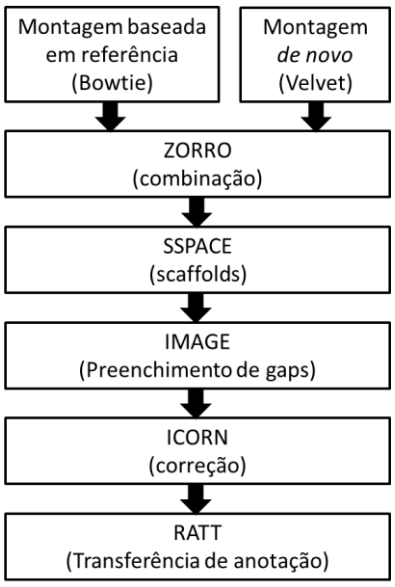

Figura 6. Esquema resumido do *pipeline* de montagem combinada utilizado neste trabalho.

Na segunda abordagem, as *reads* foram submetidas à uma montagem *de novo* utilizando o programa Velvet (ZERBINO e BIRNEY, 2008). A montagem foi realizada com uma janela de nucleotídeos deslizante (K-mers) de 51, obtida por um teste através do script VelvetOptimizer disponibilizado junto ao pacote do programa Velvet, nele diferentes tamanhos de janelas são testados e com base em estatísticas de cobertura e qualidade das montagens ele seleciona a melhor opção para o usiário. Esta abordagem foi utilizada principalmente para recuperar informações que são específicas de nosso

genoma e não são recuperadas em uma montagem baseada em referência, no entanto, essa abordagem *de novo* não consegue resolver bem a montagem de regiões altamente repetitivas. Os contigs obtidos foram submetidos a um programa de mapeamento interativo e montagem para eliminação de Gaps (Iterative Mapping and Assembly for Gap Elimination software - IMAGE) (TSAI, OTTO e BERRIMAN, 2010) e depois a um programa de correção interativa de referência nucleotídica (Iterative Correction of Reference Nucleotide software - ICORN) (OTTO et al., 2010).

As montagens obtidas por cada abordagem de nosso *pipeline* foram combinadas utilizando o programa ZORRO, que junta os dois conjuntos de contigs pré-montados em um conjunto mais contíguo e consistente, recuperando o que é normalmente perdido em cada metodologia. Por serem complementares esta abordagem permite recuperar tanto informações específicas do genoma estudado quanto seu conteúdo repetitivo. Os contigs obtidos pela combinação foram colocados em *scaffolds* utilizando o programa SSPACE (BOETZER et al., 2011), e depois tratados novamente pelos programas IMAGE e ICORN. O resultado final da nossa metodologia foi comparado aos resultados obtidos pelas metodologias *de novo* e baseadas em referência, quando realizadas separadamente, e visualizadas utilizando o programa Mummer (KURTZ et al., 2004). A montagem final também foi comparada com as outras montagens de genoma de *T. cruzi* já publicadas e disponibilizadas em banco de dados públicos.

A anotação, neste trabalho, foi realizada utilizando uma ferramenta de transferência de anotação rápida (Rapid Annotation Transfer Tool - RATT) (OTTO et al., 2011). Esta abordagem foi escolhida, pois além de mais rápida e prática, a alta identidade de nosso genoma com a referência, possibilitou um alto aproveitamento das anotações já disponíveis nos banco de dados TriTryp. Esta abordagem gerou além do arquivo de anotação, um segundo arquivo contendo as anotações que não foram transferidas. Essas foram manualmente inspecionadas para verificar se houve alguma perda muito relevante durante a transferência ou se este conjunto de dados corresponde a regiões repetitivas, como é comumente esperado.

A qualidade da montagem e anotação foi avaliada verificando o percentual de *open reading frames* (ORFs) que foram completamente montadas e anotadas no genoma final, as montagens parciais que não cobrissem toda a ORF foram incluídas. A mesma verificação foi realizada para a montagem gerada pela abordagem *de novo* de nosso *pipeline* para investigar se houveram diferenças na qualidade da montagem final quando comparada com a inicial.

Para estimar o conteúdo repetitivo do genoma da cepa TcIII 231, as sequências foram analisadas pelos programas Repeat-scout (PRICE, JONES e PEVZNER, 2005) e Repeat-Masker (SMIT, HUBLEY e GREEN, 2010). Ambos os programas disponibilizaram uma estimativa do percentual do genoma de 231 constituído por repetições simples, como microssatélites, elementos transponíveis e famílias multigênicas. Estes percentuais foram comparados ao percentual de regiões repetitivas contidas no arquivo de anotações não transferidas tanto da montagem combinada proposta neste trabalho quanto da montagem *de novo* para verificar o quão bem as repetições foram recuperadas na abordagem combinada.

3.4) Seleção de genes nucleares específicos para a análise evolutiva

A análise de genômica comparativa baseia-se em encontrar diferenças globais no genoma completo dos organismos estudados e utilizá-las como informações para relacionar e compreender os processos da função evolutiva e que atuaram sobre eles. Como nosso organismo alvo sofreu ao longo de sua história evolutiva, uma multiplicidade de eventos que deram origem a diversas linhagens individuais, esse tipo de estudo se torna necessário.

Com a sequência do genoma do clone da cepa 231 devidamente montada e anotada, será possível proceder às análises de comparação do genoma obtido contra as demais linhagens, cujos genomas já se encontram disponíveis: TcI (Sylvio X10 - parental), TcII (Esmeraldo - parental) e TcVI (CL Brener - hibrida). Para evitar um viés de seleção inadequada de genes, a estratégia escolhida para as análises evolutivas foi de concentrar o trabalho nas regiões de sequências de cópias únicas relativamente conservadas entre todas as linhagens utilizadas na análise.

Para a seleção do melhor conjunto de dados a ser utilizado para as análises filogenéticas, inicialmente foram identificadas quais *reads* de 231 conservadas e especificas para cada linhagem utilizada de referência. Selecionando e separando estas *reads* de acordo com suas correspondências aos genomas referência de *T. cruzi* utilizados, o conjunto de dados obtido a partir do genoma 231 foi subdividido em sete diferentes partições denominadas no presente projeto como subsets (figura 7). O próximo passo foi montar cada set usando o programa Bowtie, e depois recuperando a montagem consenso no formato fasta utilizando o programa VCFtools. Cada subset foi submetido a um BLASTX contra o proteoma do *T. cruzi* pertencente ao UniprotDB

(#5693) para a identificação das sequências correspondentes a proteínas presentes em cada um deles. Todas as sequências correspondentes a proteínas identificadas foram agrupadas em grupos ortólogos utilizando a ferramenta do OrthoMCLDB (FISCHER et al., 2011). Todos os genes parálogos e pseudogenes foram removidos para obter cada subset de ortólogos. Finalmente para realizar a análise filogenética, os ortólogos que estavam somente presentes no subset que representava proteínas que são altamente conservadas entre todas as linhagens analisadas foram selecionadas (Subset 7)(Figura 8).

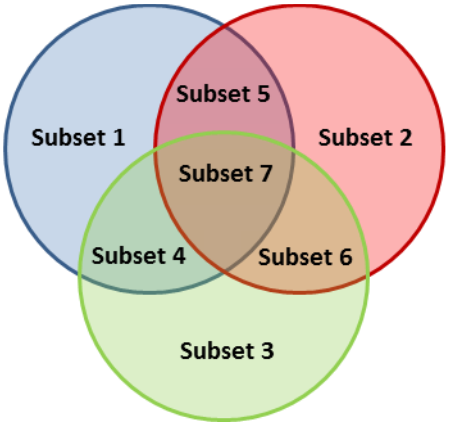

Figura 7. Diagrama de Venn mostrando os sete subsets diferentes selecionados pelas nossas *reads* conservadas em todos genomas referências de *T. cruzi* utilizadas. Em azul está representado a referência de TcI, em vermelho a de TcII e em verde a referência do haplotipo Non-Esmeraldo *like* da cepa CL Brener de *T. cruzi* TcVI.

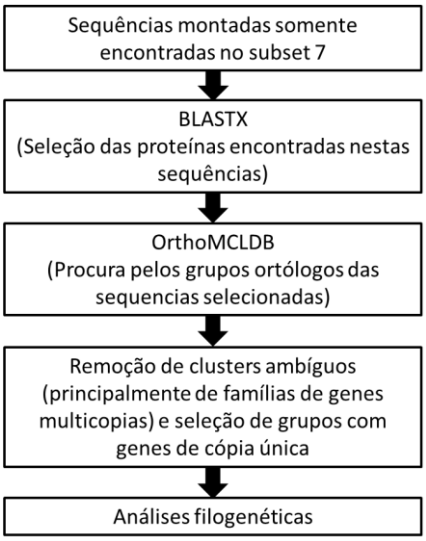

Figura 8. Fluxograma mostrando como foi feita a seleção dos genes para a análise filogenética neste trabalho.

3.5) Análise Filogenética e Estimativa de tempo de divergência de *T. cruzi*

Utilizando os dados nucleares selecionados, a análise filogenética foi realizada em diferentes etapas. Inicialmente cada grupo de sequência selecionado foi alinhado com sequências de outras linhagens de *T. cruzi*, *Trypanosoma brucei* e *Leishmania major* disponíveis em bancos de dados. As sequências foram alinhadas utilizando 3 programas diferentes: MUSCLE v3.8 (EDGAR, 2004), MAFFT v7.0 (KATOH e STANDLEY, 2013) e ClustalW v2.0 (LARKIN et al., 2007). Os três alinhamentos obtidos foram então combinados em um alinhamento consenso usando o programa M-Coffee (WALLACE et al., 2006). Essa abordagem, por utilizar diferentes métodos, minimiza o erro de alinhamento entre as amostras. O alinhamento combinado foi submetido a uma *trimagem* com o programa trimAl v1.4 (CAPELLA-GUTIÉRREZ, SILLA-MARTÍNEZ e GABALDÓN, 2009), com um corte de consistência de 0,1667 e uma nota de corte de gap de 0,1, para remover regiões de difícil alinhamento, que poderiam interferir na qualidade das análises (CAPELLA-GUTIERREZ, KAUFF e GABALDÓN, 2014).

Para as análises filogenéticas, nós utilizamos o programa PhyML onde árvores foram reconstruídas utilizando estimativas de máxima verossimilhança (ML). Para a obtenção destas árvores, duas diferentes etapas foram realizadas. Na primeira, todos os grupos ortólogos alinhados foram selecionados separadamente para se verificar a hipótese de estar em relógio molecular, esta verificação foi realizada através do teste aproximado de relação de máxima verossimilhança (aLRT). Na segunda etapa, a reconstrução foi feita a partir da concatenação de todos os grupos ortólogos alinhados que se encontraram dentro do relógio molecular, para gerar uma árvore referência. A capacidade de reconstruir a filogenia de referência foi utilizada para classificar e comparar posteriores reconstruções filogenéticas individuais da espécie.

Todas as árvores de ML tiveram seus melhores modelos de substituição de aminoácidos selecionados através do programa ProtTest 3 (DARRIBA et al., 2011), contido no pacote ModelTest, de acordo com os valores obtidos pelo Critério de Informação Akaike (AIC) e pelo Critério de Informação Bayesiano (BIC). Com essas informações, as árvores foram reconstruídas usando o programa PhyML v3.0 (GUINDON et al., 2010).

Para cada árvore ML obtida dos grupos ortólogos, foi realizado um teste de razão de verossimilhança (Likelihood Ratio Tests - LRT), para avaliar se a hipótese nula que cada locus utilizado teria evoluído sob relógio molecular (HORDIJK e GASCUEL, 2005). Todos *loci*, para os quais o relógio molecular não foi rejeitado, e que tiveram um homólogo com as espécies do grupo externo (*Trypanosoma brucei e Leishmania major*), foram concatenados para as análises de divergência. As datas de divergência foram estimadas utilizando o programa de análise bayesiano BEAST v.2 (BOUCKAERT et al., 2014). Ambos modelos de relógio, strict e relaxed lognormal, foram usados para estimar o tempo de divergência nuclear da espécie *T. cruzi*.

Além das análises de sequências nucleares, também foi utilizada toda a região codificante recuperada da mitocôndria para a reconstrução filogenética de uma árvore de verossimilhança mitocondrial e para a análise de tempo de divergência, utilizando o melhor modelo de substituição de nucleotídeo obtido pelo programa JModeltest (DARRIBA et al., 2012), também contido no pacote ModelTest.

As análises de tempo de divergência foram feitas separadamente para as sequências nucleares e mitocondriais, visto que análises anteriores obtiveram estimativas de datas bem diferentes para cada tipo de dado (FLORES-LÓPEZ e MACHADO, 2011). Todas as análises foram conduzidas sem nenhuma restrição topológica, usando o melhor modelo de substituição selecionado pelo programa do pacote ModelTest, com 4 categorias gama, assim como o particionamento de códons em 3 posições. Todos os antecedentes foram definidos com os valores padrão do programa BEAST, exceto para o processo de especiação Yule como árvore antecedente. A estimativa de divergência entre *T. cruzi* e *T. brucei*, foi fixado, em 100 milhões de anos (mya) em uma distribuição normal, com 10 mya como o desvio padrão. Esses valores embora passíveis de algumas críticas (HAMILTON, TEIXEIRA e STEVENS, 2012), têm sido utilizados pela maioria dos trabalhos de filogenia de tripanossomas, e levam em conta a separação dos continentes Africano e Sul-Americano, como fator responsável pela diferenciação das duas espécies (FLORES-LÓPEZ e MACHADO, 2011; LEWIS et al., 2011; MACHADO e AYALA, 2001)

Todos os resultados nucleares obtidos foram comparados com os obtidos para a mitocôndria a fim de explicar se houve algum evento de recombinação durante a evolução do *T. cruzi* ou melhor explicar algo sobre a segregação das linhagens do parasito.

4) Algoritmo de predição de agrupamentos gênicos de *T. cruzi*

4.1) Conjunto de dados utilizado

Foram utilizadas, nesta parte do projeto, 294 sequências do gene de amastina de *T. cruzi* depositadas no GenBank por Cerqueira e colaboradores (2008) ou obtidos no genoma sequenciado neste trabalho da cepa TcIII 231. O gene da amastina codifica uma glicoproteína de superfície de cerca de 174 aminoácidos, muito hidrofóbica expressa nas células amastigotas intracelulares de *T. cruzi* (ROCHETTE et al., 2005). As sequências utilizadas no presente trabalho foram originadas de clones pertencentes a seis diferentes cepas de *T. cruzi* representantes de três linhagens ou DTUs (Tabela 4).

Tabela 4 - Cepas de *T. cruzi* com suas respectivas linhagens e o número de sequências utilizadas neste trabalho.

Cepas	Nº de Sequências	Linhagens
Sylvio X10 cl1[1]	43	TcI
Colombiana[1]	43	TcI
DM28[1]	43	TcI
Esmeraldo cl3[1]	43	TcII
JG[1]	43	TcII
231[2]	36	TcIII
CL Brener[1]	43	TcVI

[1] Sequências obtidas por Cerqueira et al 2008
[2] Sequências obtidas neste trabalho

Todas as sequências de amastina utilizadas foram inseridas em um arquivo MultiFASTA e curadas seguindo duas etapas. Inicialmente as sequências foram alinhadas pelo alinhador global de sequências MUSCLE 3.6 (EDGAR, 2004). As regiões das sequências não alinhadas foram identificadas e removidas usando o pacote MEGA 5 (TAMURA et al., 2011) e o programa trimAL (CAPELLA-GUTIÉRREZ, SILLA-MARTÍNEZ e GABALDÓN, 2009). Este segundo passo é essencial para minimizar interferências, como variações como erros de sequenciamento, ou diferenças no tamanho das sequências obtidas, que poderiam gerar dados de variação falsos na análise do algoritmo.

4.2) O algoritmo

A fim de se encontrar agrupamentos que permitissem diferenciar as linhagens ou grupos do parasito a partir do conjunto de dados de amastina, foi construído um

algoritmo, com o pipeline esquematizado na Figura 9, aplicado na linguagem de programação MATLAB.

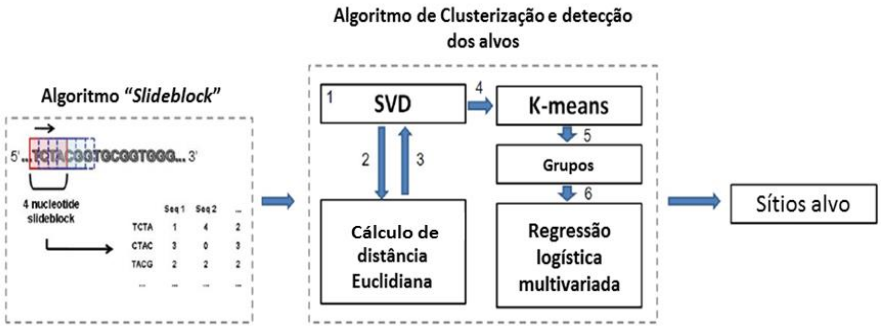

Figura 9. Pipeline proposto neste trabalho para a montagem do algoritmo de clusterização.

Inicialmente, os dados de sequências nucleotídicas pré-editadas foram transformadas em uma matriz numérica usando um algoritmo que foi denominado por nós de "*Slideblock*". Esse algoritmo cria uma janela de tetranucleotídeos deslizante ao longo das sequências da população estudada e calcula, para cada bloco de tetranucleotídeos, a frequência encontrada. As sequências do gene analisado, neste caso da amastina de *T. cruzi*, são registradas como vetores em $\Re^b$, onde $\Re$ é o número de possíveis caracteres nucleotídicos (A, C, T ou G) e b é o tamanho do bloco ($\Re^b = 4^4 = 256$). Desta forma, toda a base de dados é transformada em uma matriz M, de dimensão mxn, onde m é o número de blocos encontrados ($m = \Re^b$) e n é o número de sequências estudadas. Cada elemento da matriz indica a frequência do bloco de tetranucleotídeos observada em cada sequência do gene analisado. Assim, cada x_{ij} é a frequência (0, 1, 2, ...) do tetranucleotídeo *i* na sequência do gene no vetor *j*.

$$M = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Após a aplicação do algoritmo "*Slideblock*", a matriz M foi decomposta utilizando o cálculo do SVD (Decomposição de Valores Singulares). Essa decomposição foi utilizada, pois constitui uma abordagem capaz de revelar um padrão de clusterização com o mínimo de perda de informações, permitindo excelente

visualização dos dados multivariados com uma alta dimensionalidade. O SVD executa uma fatoração de uma matriz real ou complexa, $M = USV^T$, utilizada aqui para normalizar M, reduzindo a dimensionalidade do conjunto de dados e capturando as "características" que podem ser usadas para comparar as sequências de amastina (DEERWESTER et al., 1990; GOLUB e KAHAN, 1965). Assim, U é a matriz ortogonal de mxr, tendo os vetores singulares da esquerda de M como suas colunas, V^T é a matriz ortogonal transposta rxn, tendo os vetores singulares de M como suas colunas, e S é a matriz diagonal de rxr com os valores singulares $1 \geq 2 \geq 3 \cdot \cdot \cdot \geq$ r de M em ordem, ao longo da sua diagonal (r é o posto de M, que é o número de colunas ou linhas linearmente independentes de M). Os valores singulares colocados em ordem decrescente ao longo da diagonal principal de S estão diretamente relacionados com as características independentes, dentro do dataset utilizado (BERRY, DUMAIS e O'BRIEN, 1995; DEERWESTER et al., 1990; ELDÉN, 2006).

$$\underset{m \times n}{M} = \underset{m \times r}{U} \; \underset{r \times r}{S} \; \underset{r \times n}{V^{\mathsf{T}}}$$

Após a representação por meio das frequências dos tetranucleotídeos e da matriz tratada com SVD, uma nova matriz foi obtida aplicando-se o cálculo da distância euclidiana. Essa metodologia permite uma melhor representação na plotagem das amostras, visto que ela calcula a distância entre os pontos, aqui representados pelas sequências dos genes. Essa nova matriz foi também tratada com SVD a fim de se remover as redundâncias obtidas.

Para a fase de clusterização do dataset, foi utilizado um algoritmo chamado k-Means. Este algoritmo inicializa, de forma aleatória, posições centroides k, onde o número de k é designado pelo usuário, em um espaço (HAN e KAMBER, 2000). O problema sobre como escolher o melhor número de centroides k foi resolvido, usando o número de valores singulares selecionado pela análise do gráfico de valores relativos dos 10 primeiros valores singulares de M gerado pela fatoração do SVD, como número de k centroides. Cada iteração do algoritmo k-Means consiste em duas etapas: a atribuição de cluster e atualização centroide.

A atribuição de cluster e as etapas da atualização do centroide são realizadas iterativamente para se atingir um número máximo de iterações ou até que um mínimo de

modificação de posição dos centroides seja encontrado (ZAKI e MEIRA JUNIOR, 2011). Assim, quando os centroides param de mudar de posição para as próximas iterações o algoritmo supõe que a distribuição dos grupos já convergiu, e os grupos mais prováveis encontrados são mostrados. Como o k-Means é um algoritmo não determinístico, este método foi reiterado 100 vezes para gerar um gráfico em três dimensões pela função plot do MATLAB que representa consenso dos aglomerados mais prováveis, possibilitando um apoio estatístico aos resultados e minimizando a susceptibilidade do algoritmo a *outliers* que poderiam influenciar a posição dos centroides.

4.3) Validação do algoritmo

Para estimar a qualidade dos clusters obtidos pela metodologia do k-means, os mesmos foram comparados com aqueles obtidos em uma análise filogenética utilizando as mesmas sequências de amastina curadas. As distâncias médias entre cada isolado foram calculadas utilizando o pacote MEGA 5. As distâncias foram computadas utilizando o modelo de substituição de 2 parâmetros de Kimura com 1000 réplicas de bootstrap. As árvores filogenéticas foram geradas também no pacote MEGA 5, usando o método de reconstrução *neighbor-join*, e representadas no visualizador de árvores Figtree (MORARIU et al., 2008). Essa metodologia utilizada como base de validação foi escolhida, pois ela é baseada em um método de distância assim como o algoritmo proposto neste trabalho. O método de reconstrução baseado em distância é uma das metodologias mais utilizadas em estudos de filogenia, e inclusive foi o método utilizado no estudo onde obtivemos parte do conjunto de dados de sequências utilizado no presente trabalho (CERQUEIRA et al., 2008).

4.4) Regressão Logística

O modelo de regressão logística é normalmente utilizado para prever a probabilidade de ocorrência de um evento, pelo ajuste dos dados, a uma função em uma curva logística. Aqui ela foi aplicada para cada *cluster* encontrado no conjunto de dados de sequências de amastina, permitindo a seleção de recursos para identificar quais blocos são estatisticamente significativos para a determinação dos *clusters*. Em adição, o modelo logístico também pode ser utilizado para prever a probabilidade (π) de uma

sequência pertencer a um cluster específico, com base numa combinação dos blocos k selecionados pelo modelo:

$$\pi = \frac{e^{\beta_0 + \sum \beta_i x_i}}{1 + e^{\beta_0 + \sum \beta_i x_i}}$$

A equação acima dá a probabilidade de uma sequência pertencer a um cluster específico. Nesta equação, i corresponde ao coeficiente de regressão para cada bloco (i = 1, 2, 3, ..., k), sendo k o número de blocos de tetranucleotídeos significativamente selecionados para o modelo. A estratégia de construção de modelos para a seleção dos blocos tetranucleotídicos foi uma regressão "*forward stepwise*" logística automática realizada pelo pacote de análise SPSS (SPSS Inc., 2008). O conjunto de dados foi dividido ao acaso em dois grupos para a análise multivariada: 212 sequências foram utilizadas durante as análises, como conjunto treinamento, e as 82 sequências restantes foram utilizadas para validação da regressão realizada.

Antes de executar a regressão logística, uma análise univariada foi realizada utilizando teste *t-student* para cada um dos 256 blocos de sequências analisadas (ALTMAN, 2008). Apenas os blocos com valores de p menor ou igual a 0,05 obtidos pelo teste *t* foram utilizados na análise multivariada. A sensibilidade e a especificidade da presença de blocos nos clusters obtidos foram também calculadas para avaliar a qualidade discriminante dos modelos logísticos na classificação de grupos de sequências de desconhecidos.

RESULTADOS

IV. RESULTADOS

1) Avaliação da ocorrência de recombinação entre cepas de *T. cruzi* II

1.1) Diversidade mitocondrial dos genes ND4 e ND7

Antes de se avaliar a diversidade mitocondrial entre as cepas de *T. cruzi* II, todos isolados de TcII foram primeiramente submetidos a um ensaio de caracterização com a metodologia de RFLP-PCR do gene COII para confirmar se o haplótipo mitocondrial obtido seria equivalente ao haplótipo C, perfil mitocondrial específico para amostras desta linhagem. Após o ensaio todos isolados de TcII analisados exibiram o haplótipo mitocondrial C, confirmando como esperado, que todas amostras eram pertencentes a linhagem TcII.

Em seguida a fim de se determinar a diversidade mitocondrial dos genes ND4 e ND7 entre as cepas de *T. cruzi* II, foi realizado um ensaio de caracterização molecular para os genes mitocondriais ND4 e ND7 de *T. cruzi*. Por meio da caracterização de 47 das 88 amostras de TcII, provenientes de pacientes de MG, foi observada a presença de diferentes tamanhos *amplicons*: 300 e 530 para ND4/CR4 e 500 e 900 para ND7, o menor tamanho de cada marcador corresponde aos alelos com deleções (CARRANZA et al., 2009). Três diferentes possíveis haplótipos ND7/ND4 foram encontradas nas amostras mineiras como esperado: sete amostras apresentaram o perfil do haplótipo C1 (300/900pb), 24 amostras apresentaram o perfil do haplótipo C2 (530/500pb) e 16 amostras apresentaram o perfil do haplótipo C3 (530/900pb) (Tabela 5). As 41 amostras restantes não tiveram o seu haplótipo mitocondrial determinado.

Para investigar se os *indels* analisados foram simples ou tiveram múltiplas origens (homoplasia), o que poderia interferir em nossas análises, os limites do *indel* do gene ND7 de cepas de TcII polares foram sequenciados. O alinhamento das sequências confirmou que o *indel* era aparentemente livre de homoplasia e, portanto, adequado para análises genéticas de populações aqui realizados (Figura 10).

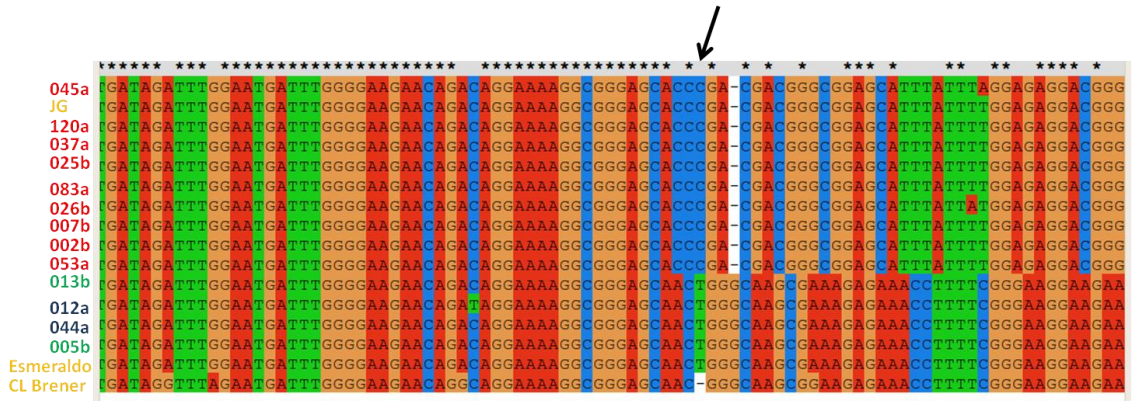

Figura 10. Alinhamento das sequências obtidas do gene ND7 de cepas de TcII polares utilizando o programa ClustalX. As sequências alinhadas possuem representantes de cada um dos haplótipos encontrados. O inicio da região do *indel* está indicada pela seta. As cores dos nomes representam os seguintes haplótipos mitocondriais: azul – C1; vermelho – C2; verde – C3; amarelo – C* (não determinado).

1.2) Caracterização dos microssatélites nucleares

A análise de nove *loci* polimórficos de microssatélites revelou uma grande diversidade de padrões de repetições em diferentes populações de *T. cruzi*, resultando em uma identidade estável, reprodutível e única para cada isolado. Genótipos dos microssatélites das cepas, com base no número de repetições destes *loci*, estão resumidos na Tabela 5. Para os *loci* SCLE11, MCLE01, AAT8 e TAT20 foi observada uma predominância de perfis heterozigotos enquanto que para os *loci* SCLE10, MCLF10, MCLG10, ATT14 e AAAT6 perfis homozigotos foram mais encontrados entre as amostras analisadas. Um perfil homozigoto é detectado pela presença de um único pico no cromatograma do sequenciador automático, enquanto que um perfil heterozigoto é detectado pela presença de dois picos em posições diferentes. Três ou mais picos, sugestivos de mistura de populações heterozigotas ou aneuploidia, não foram observadas entre as amostras utilizadas neste trabalho.

Tabela 5 - Genótipos mitocondriais (ND4/7) e nucleares dos nove *loci* de microssatélites obtidos para as cepas de *T. cruzi* II utilizadas neste trabalho.

Isolados	Haplótipos ND4/7	Haplótipos Nucleares dos Microssatélites								
		SCLE10	SCLE11	MCLE01	MCLG10	MCLF10	TCAAT8	ATT14	TCTAT20	TCAAAT6
002B	C2	26\26	14\14	13\15	8\8	8\8	10\17	13\13	15\7	8\9
003ª	C1	26\28	15\17	11\7	7\11	8\8	13\18	11\11	18\11	9\8
005B	C3	25\36	16\18	11\14	8\8	8\8	12\10	8\12	18\21	9\5
007B	C2	27\27	16\16	13\13	10\10	8\8	9\17	12\12	13\18	5\9
09B	C3	28\36	11\14	9\8	8\12	8\8	13\10	13\13	18\15	8\5
010B	C2	25\25	13\16	11\11	9\9	8\8	14\14	13\13	15\15	5\8
011B	C3	27\36	13\17	11\8	8\8	8\8	10\10	13\13	21\21	5\5
012ª	C1	27\27	14\16	7\8	10\10	8\8	13\13	11\11	15\10	9\8
012B	C2	26\26	13\17	15\11	8\8	8\8	17\18	12\12	7\15	9\9
013B	C3	26\26	14\16	19\11	9\9	8\8	12\10	12\12	15\10	5\9
019ª	C1	27\36	13\17	8\7	8\8	8\8	12\18	12\12	10\11	9\8
020B	C3	28\28	17\17	13\19	12\12	8\8	18\18	12\12	18\15	9\9
023B	C3	26\36	14\14	8\8	8\8	8\8	9\12	8\12	15\10	5\9
024B	C3	27\27	13\14	7\7	8\8	10\10	14\14	11\11	18\11	8\9
025B	C2	26\26	13\14	11\19	11\11	8\8	9\9	11\11	15\17	8\6
026B	C2	25\28	11\15	9\13	8\13	8\8	12\10	12\12	18\15	8\8
029ª	C2	26\28	11\13	13\13	7\7	8\8	17\17	12\12	19\16	9\9
031ª	C3	26\38	17\17	8\11	10\10	9\9	13\14	12\12	15\16	9\9
037ª	C2	28\28	14\15	19\11	11\11	8\8	8\8	11\11	13\11	9\9
044ª	C1	27\36	13\13	12\17	8\11	8\8	5\5	11\11	18\12	8\9
045ª	C2	25\25	14\16	19\11	8\8	8\8	14\14	12\12	10\15	8\8
050b	C2	24\24	16\16	11\13	8\12	9\9	13\17	12\12	11\19	8\9
053ª	C2	25\25	11\13	9\12	7\7	9\9	10\13	12\12	18\19	8\8

Isolados	Haplótipos ND4/7	Haplótipos Nucleares dos Microssatélites								
		SCLE10	SCLE11	MCLE01	MCLG10	MCLF10	TCAAT8	ATT14	TCTAT20	TCAAAT6
055ª	C3	27\27	14\21	11\11	8\8	8\8	7\7	12\12	16\9	5\8
058ª	C2	25\36	16\18	11\15	8\8	8\8	12\9	8\12	18\21	9\5
065b	C2	26\28	14\16	19\11	8\10	8\8	12\9	12\12	15\10	5\9
067ª	C2	26\26	13\13	11\15	8\8	8\8	17\17	12\12	15\7	9\9
079ª	C2	24\24	13\15	13\18	9\9	8\8	18\18	11\11	16\14	8\8
083ª	C2	26\26	12\15	9\9	8\13	8\8	13\10	11\11	19\11	8\8
085ª	C2	25\28	11\14	7\9	10\10	8\8	17\17	12\12	18\10	9\9
090ª	C3	28\36	11\18	9\9	10\10	8\8	13\9	9\12	18\15	8\5
092ª	C2	26\26	11\11	11\13	8\8	8\8	9\9	12\12	15\15	9\5
094ª	C3	25\25	11\14	18\18	10\10	8\8	14\14	12\12	15\10	9\8
103ª	C3	26\28	10\13	9\18	5\12	8\8	12\12	11\11	19\19	8\8
105ª	C2	28\32	14\16	20\11	8\8	8\8	12\9	11\11	10\10	9\9
109ª	C2	25\28	16\16	11\13	7\7	8\8	12\12	11\11	19\16	9\8
110ª	C2	27\27	14\16	11\11	10\10	10\10	13\10	11\11	17\10	8\9
115b	C1	28\28	15\15	11\11	8\8	8\8	2\2	11\11	8\21	7\7
116ª	C1	26\36	11\13	9\7	7\12	8\9	13\10	12\12	18\12	8\8
120ª	C2	31\31	13\14	9\12	8\8	8\8	9\8	13\13	10\26	9\9
129ª	C1	25\25	14\14	7\11	8\8	8\8	13\13	11\11	15\16	9\8
132ª	C3	27\27	12\13	9\20	6\7	8\8	18\18	11\11	11\7	9\8
138ª	C3	25\26	14\15	11\11	6\8	8\8	22\22	11\11	13\22	8\8
146ª	C2	26\26	13\14	12\8	8\8	8\8	13\9	9\12	19\15	8\5
154ª	C3	28\28	12\19	21\19	5\5	8\9	10\10	11\11	7\13	9\9
162ª	C3	26\26	14\16	19\11	8\8	8\8	12\9	12\12	15\10	5\9
128ª	C2	26\26	14\14	9\9	8\8	8\8	14\14	12\12	7\10	9\9
38	C*	28\28	14\14	9\19	10\10	8\8	12\12	12\12	10\18	9\8
013ª	C*	26\28	14\14	10\19	10\10	8\8	12\10	12\12	11\16	9\6
016B	C*	26\26	14\16	19\11	9\9	9\9	12\10	12\12	15\10	5\9

Isolados	Haplótipos ND4/7	Haplótipos Nucleares dos Microssatélites								
		SCLE10	SCLE11	MCLE01	MCLG10	MCLF10	TCAAT8	ATT14	TCTAT20	TCAAAT6
192ª	C*	27\27	13\14	19\9	8\8	8\8	18\18	12\12	7\10	8\9
JG	C*	27\28	11\13	11\12	8\8	8\8	13\13	11\11	10\19	9\8
022b	C*	25\25	13\16	7\11	9\9	9\9	14\14	8\12	18\15	8\5
003B	C*	NA	14\16	19\9	NA	8\8	12\10	NA	15\10	5\9
002ª	C*	NA	13\13	7\11	NA	NA	14\14	NA	18\15	8\9
005ª	C*	NA	14\14	7\12	NA	NA	12\10	NA	10\11	9\8
006ª	C*	NA	14\14	9\9	NA	NA	17\17	NA	10\18	9\8
021B	C*	27\27	13\14	11\9	9\9	8\8	14\14	NA	15\18	5\9
097ª	C*	28\31	15\17	11\11	8\8	8\8	9\12	NA	26\10	9\9
188ª	C*	NA	14\16	9\13	NA	NA	18\18	NA	10\17	9\9
Esmeraldo	C*	28\34	14\18	6\12	8\9	7\8	10\9	NA	12\7	7\9
Be62	C*	27\28	16\16	12\13	8\8	8\8	ND	ND	ND	ND
GMS	C*	27\33	13\17	7\11	8\8	8\8	ND	ND	ND	ND
Ig539	C*	31\31	14\15	10\19	6\9	9\9	ND	ND	ND	ND
Mas1 cl1	C*	23\34	14\14	8\8	8\8	9\9	ND	ND	ND	ND
MPD	C*	24\26	10\15	13\13	8\10	7\7	ND	ND	ND	ND
Tula cl2	C*	35\35	13\14	9\9	8\8	8\8	ND	ND	ND	ND
84	C*	24\26	15\16	7\11	8\8	8\8	ND	ND	ND	ND
169/1	C*	27\28	13\13	8\13	8\8	7\9	ND	ND	ND	ND
200pm	C*	29\29	15\16	12\12	8\9	9\9	ND	ND	ND	ND
209	C*	25\28	15\16	10\10	8\8	8\8	ND	ND	ND	ND
239	C*	26\28	14\17	11\19	8\8	8\8	ND	ND	ND	ND
803	C*	26\31	11\11	12\20	8\11	9\9	ND	ND	ND	ND
1005	C*	28\28	15\16	9\22	8\9	8\8	ND	ND	ND	ND
1014	C*	25\28	13\13	14\15	8\8	8\8	ND	ND	ND	ND
Y	C*	27\27	15\15	9\9	8\8	8\9	ND	ND	ND	ND
1043	ND	28\31	17\17	8\11	8\9	7\7	ND	ND	ND	ND

Isolados	Haplótipos ND4/7	Haplótipos Nucleares dos Microssatélites								
		SCLE10	SCLE11	MCLE01	MCLG10	MCLF10	TCAAT8	ATT14	TCTAT20	TCAAAT6
1931	ND	28\28	9\15	14\14	10\10	9\9	ND	ND	ND	ND
GOCH	ND	26\31	15\18	8\19	6\6	10\10	ND	ND	ND	ND
577	ND	27\31	14\16	20\21	6\8	10\10	ND	ND	ND	ND
578	ND	31\31	13\16	9\18	8\8	10\10	ND	ND	ND	ND
580	ND	30\30	13\13	9\11	8\8	9\9	ND	ND	ND	ND
183744	ND	31\31	13\13	9\11	8\8	9\9	ND	ND	ND	ND
CPI95/94	ND	29\29	13\16	6\10	8\8	9\9	ND	ND	ND	ND
OPS27/94	ND	34\39	13\18	10\12	8\8	9\9	ND	ND	ND	ND
GLT564	ND	28\28	14\14	12\12	8\10	8\8	ND	ND	ND	ND
GLT593	ND	26\33	14\14	12\12	9\11	7\7	ND	ND	ND	ND
CPI11/94	ND	29\39	15\15	9\9	8\8	9\9	ND	ND	ND	ND

Os haplótipos mitocondriais foram determinados pela combinação dos perfis obtidos para cada amostra pelos marcadores das subunidades 4 e 7 da NADH desidrogenase. Haplótipo C1 (300/900pb), o haplótipo C2 (530/500pb) ou haplótipo C3 (530/900pb). Os perfis de microssatélites são representados pelo número de repetições obtidos para cada amostra. C*: haplótipos de ND4/7 não determinados para estes isolados, mas como esperado para cepas de TcII, todos obtiveram o perfil de haplotipo C no ensaio de caracterização para o marcador de COII. NA: não amplificada; ND: não determinado.

1.3) Análise da estrutura populacional intra-linhagem TcII

Os genótipos de microssatélites encontrados como descrito no item anterior foram convertidos em 108 haplótipos pelo programa PHASE e estes dados foram utilizados para construir uma rede de haplótipos pelo programa Network (Figura 11). Essa rede mostra as distâncias entre os dois haplótipos alélicos obtidos para cada isolado com base no número de passos mutacionais necessários para transformar um haplótipo em outro. Assim haplótipos com números semelhantes são mais similares entre si. Considerando o excesso de homozigose geralmente observada no genoma de *T. cruzi*, era esperado em uma população clonal, detectar pequenas diferenças no número de passos mutacionais entre os haplótipos diplóides dos diferentes isolados. No entanto estes números variaram amplamente: o número mínimo observado de passos mutacionais foi de três para os dois haplótipos da cepa 128a, e o maior foi de 67 passos para os dois haplótipos da cepa 012B, indicando dentro desta população que existem isolados que podem ter sido resultados de eventos de hibridização relativamente recentes.

Considerando a grande diversidade de haplótipos nucleares observada entre as cepas analisadas, a próxima etapa foi determinar se haveria variação genética suficiente para suportar uma subestruturação dentro da população de 47 amostras de TcII de Minas Gerais caracterizadas para ND4 e ND7. O número apropriado de subpopulações no teste da amostra foi determinado usando o programa STRUCTURE, por comparação dos valores de *log-likelihood* (Ln) para K 1-10 e do cálculo de ΔK. Ambas as análises indicaram a existência de três subgrupos dentro da nossa população, K = 3, nomeados de N1, N2 e N3 (Figura 12). Também foi calculado valores de F_{ST} e F_{IS} para cada subgrupo apontado pelo STRUTURE. Estes valores foram utilizados como uma medida da diferenciação da população, ou seja, a distância genética entre grupos de indivíduos, baseada em dados de polimorfismo genético, tais como microssatélites.

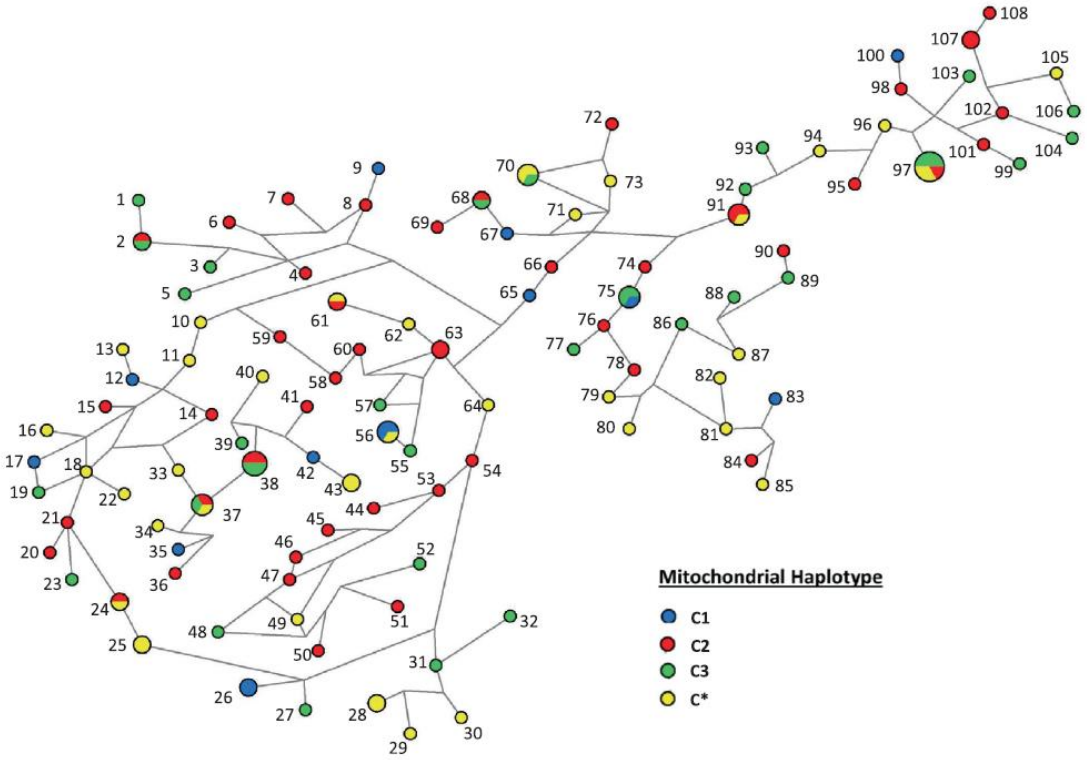

Figura 11. Rede haplotípica de marcadores nucleares gerada pelo software PHASE, indicando as distâncias entre os diferentes haplótipos de cepas da linhagem TcII. O tamanho dos círculos é proporcional ao número de haplótipos idênticos observados. Os números designam os diferentes haplótipos encontrados e, pela forma com que são gerados, números mais próximos designam haplótipos mais similares entre si. As diferentes cores se referem aos diferentes haplótipos de mitocondriais de ND4 e ND7: Azul – C1 (300/900 bp), vermelho – C2 (530/500 bp), verde – C3 (530/900 bp) e amarelo – C* (não determinado).

O F_{ST} é uma medida da diferenciação população devido à sua estrutura genética. Ele se baseia na variação de frequências alélicas entre as populações. Valores de F_{ST} abaixo de 0.05 indicam que não há divergência entre as populações enquanto maiores indicam que há uma diferenciação entre elas. O $F_{IS,}$ ou coeficiente de endogamia, é a proporção da variância na subpopulação contido em um indivíduo. Valores positivos do F_{IS} implicam um considerável grau de endogamia na população.

Os cálculos de F_{ST} e F_{IS} foram feitos de duas formas para o conjunto de dados estudados. O primeiro cálculo foi realizado para cada subgrupo nuclear separadamente (N1-3), a fim de se verificar semelhanças e evidências de recombinações entre as cepas dentro do agrupamento identificado. Nestas análises, os valores de F_{ST} foram maiores que 0.05 para os três subgrupos (F_{ST} = 0.0661-0.1786), o que é consistente com diferenciação genética significante entre os isolados de cada grupo do STRUCTURE.

No entanto, valores positivos de F_{IS} (F_{IS} = 0.0540-0,8455) indicaram que a quantidade de descendentes heterozigotos nestas subpopulações foi menor do que o esperado, provavelmente devido à endogamia. De fato, na presença de endogamia significativa, a reprodução é considerada não aleatória, e parentes próximos reproduzem entre si, e porque estes parentes têm provavelmente genes semelhantes, a descendência tem tendência a apresentar excesso de homozigoze. Assim, tomadas em conjunto os valores de F_{ST} quanto os valores positivos F_{IS} observados para cada subpopulação observada pelo STRUCTURE (Tabela 6) indicam a recombinação entre as cepas analisadas TcII de Minas Gerais.

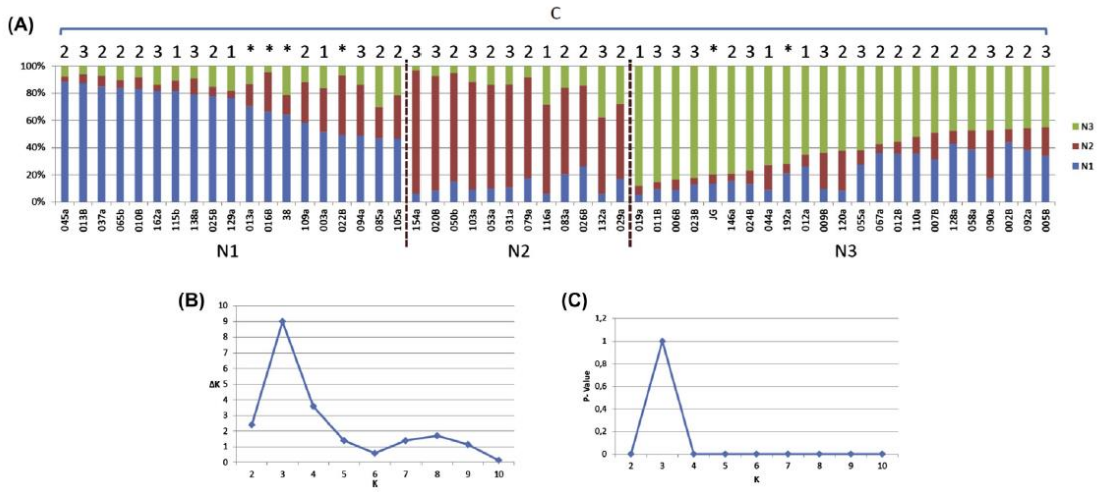

Figura 12. Determinação de subgrupos na população de *T. cruzi* II proveniente de MG baseados na análise de marcadores nucleares e mitocondriais. O número de subpopulações baseadas em dados nucleares, foi determinado usando o programa STRUCTURE onde (A) é o gráfico "*barploting*" particionado em três segmentos de cor (K = 3), e a determinação de K identificada pela correlação entre os valores de log-likelihood (Ln) para K variando de 1 a 10 (B) ou pelo cálculo do ΔK (C). As letras e números acima e abaixo do gráfico "*barploting*" indicam falta de correlação entre os marcadores nucleares e mitocondriais. Todos os três haplótipos mitocondriais (C1, C2 e C3) foram dispersos entre os três subgrupos nucleares diferentes (N1, N2 e N3), demonstrando que não existe nenhuma correlação entre os dois marcadores. Os haplótipos caracterizados com C* são de amostras onde o haplótipo ND4/7 não foi determinado.

1.4) Evidenciação da presença de recombinação em TcII

Com base nos dados de genótipos de microssatélites dois parâmetros foram utilizados para estimar a taxa de recombinação genética dentro das populações do parasito: desequilíbrio de HW e LD (Tabela 7). Inicialmente, esses parâmetros foram avaliados considerando todo o conjunto de dados da população disponível (isolados pertencentes a MG e outras regiões da América Latina), e, em seguida, apenas para as amostras procedentes de MG. Para esta análise foram utilizados inicialmente um subconjunto de cinco *loci* localizados em diferentes cromossomos. A hipótese nula investigada foi que as populações de TcII estavam em equilíbrio de Hardy-Weinberg e que os *loci* de microssatélites segregavam independentemente. Para o conjunto completo de dados, incluindo amostras pertencentes às diferentes localidades da América Latina, a hipótese nula foi rejeitada: já que a maioria dos *loci* analisados se encontrava em desequilíbrio de HW (5/5) e LD (4/10), resultados com valores $p < 0,05$, calculados pelo teste de Chi quadrado de Pearson, que indicaram desvios significativos dos valores esperados para os organismos com reprodução sexual. Estes achados indicaram que os genótipos observados poderiam ser passados em blocos de gerações, apoiando a ideia prevalente de clonalidade para esses parasitos (TIBAYRENC e AYALA, 2002). Todavia, quando apenas as amostras oriundas de MG foram analisadas, a hipótese nula não pode ser rejeitada visto que 4/5 *loci* encontravam-se em equilíbrio de HW e 9/10 pares *loci* estavam em equilíbrio de ligação (valores de $p > 0,05$). Embora alguns dos *loci* só marginalmente passaram no teste, a tendência para atingir o equilíbrio foi significativamente maior quando amostras apenas de MG foram analisadas, sugerindo a ocorrência de recombinação dentro da população de parasitos analisados (Tabela 7).

Como foi detectada a presença de três diferentes haplótipos para os marcadores mitocondriais (C1, C2 e C3) e também três subgrupos para os marcadores nucleares (N1, N2, N3), conforme determinado pelo STRUCTURE para os parasitos isolados de Minas Gerais, foi possível investigar se haveria correlação na herança destes dois conjuntos de marcadores genéticos. A presença de uma associação entre esses conjuntos supostamente independente de marcadores poderia indicar que eles estavam sendo herdados em bloco como é esperado para numa população baseada em reprodução clonal (ZHANG, TIBAYRENC e AYALA, 1988). No entanto, não foi observada nenhuma correlação entre os subgrupos de TcII identificadas pelos marcadores

nucleares e mitocondriais, como demonstrado na Figura 12B: os três haplótipos mitocondriais foram randomicamente dispersos entre os três subgrupos nucleares.

Devido aos marcadores polimórficos utilizados anteriormente, todos eles localizados em cromossomos diferentes, os dados moleculares até agora obtidos indicam a ocorrência de segregação independente de cromossomos em cepas de TcII isoladas de Minas Gerais, mas não permitiu identificar a ocorrência de recombinação homóloga. Para investigar a existência de recombinação, o desvio de HW e LD foram calculados usando o marcador MCLF10 e quatro *loci* microssatélites adicionais (TCAAT8, TCTAT20, TCAAAT6 e ATT14), todos eles localizados no cromossomo 6 (Figura 13). Embora no mesmo cromossomo, seis dos 10 pares de *loci* combinados comparados mostraram estar em equilíbrio de ligação (Tabela 8), sugerindo que além de segregação de cromossomos homólogos, eventos de recombinação homóloga podem também estar envolvidos. Além disso, quando comparamos o valor-p de LD para os 10 pares de *loci* e sua posição relativa no cromossomo 6 foi observada uma boa correlação entre esses dois elementos. Por exemplo, o locus MCLF10 que está relativamente localizado distante dos outros marcadores está em equilíbrio de ligação com todas elas. Por outro lado, os *loci* TCAAAT6, ATT14 e TCAAT8 que estão mais próximos uns dos outros são claramente herdados em blocos. Alguns desequilíbrios de ligação foram detectados entre este bloco e o locus TCTAT20 localizado um pouco distante (Figura 13).

Tabela 6 - Estimativa multilocus com análises F_{ST} e F_{IS} para os dados diploides.

Locus	F_{IS}	F_{ST}	F_{IT}
SCLE11	0.0729	0.0970	0.1628
MCLE01	0.0540	0.0661	0.1165
SCLE10	0.5062	0.0943	0.5528
MCLF10	0.8455	0.1063	0.8619
MCLG10	0.7276	0.1786	0.7763
Todos juntos	0.3660	0.1074	0.4341

Tabela 6. Análises F_{ST} e F_{IS} foram realizadas separadamente para cada subgrupo nuclear identificado pelo STRUCTURE (N1-3) para estimar as semelhanças e recombinações entre as cepas em cada subpopulação observada. Os valores de $F_{ST} > 0,05$ foram consistentes com a diferenciação genética entre os isolados de cada grupo do STRUCTURE. Valores positivos de F_{IS} indicam que a quantidade de descendência heterozigótica nestas subpopulações estava inferior ao esperado.

Tabela 7 - Testes estatísticos de genética de população.

A

Teste de Hardy - Weinberg		
Loci	Valor P (world)	Valor P (MG)
SCLE11	0.00057	0.05150
SCLE10	0.00000	0.06232
MCLE01	0.00000	0.06100
MCLF10	0.00000	0.00000
MCLFG10	0.00000	0.05110

B

Teste de Desequilíbrio de Ligação				
Par de *Loci*			Valor P (world)	Valor P (MG)
SCLE11	**&**	**MCLE01**	0.000000	0.172377
SCLE11	**&**	**SCLE10**	0.17168	0.344099
MCLE01	**&**	**SCLE10**	0.062400	0.142935
SCLE11	**&**	**MCLF10**	0.250320	0.548089
MCLE01	**&**	**MCLF10**	0.044430	0.420419
SCLE10	**&**	**MCLF10**	0.000000	0.528246
SCLE11	**&**	**MCLG10**	0.000000	0.026943
MCLE01	**&**	**MCLG10**	0.273930	0.869349
SCLE10	**&**	**MCLG10**	0.538440	0.151469
MCLF10	**&**	**MCLG10**	0.060340	0.066145

Tabela 7. Testes estatísticos populacionais observados para os dois conjuntos de dados analisados. As amostras "*world*" representam todas as amostras utilizadas no trabalho pertencendo a variadas localizações geográficas e as "MG" representam as amostras provenientes somente do estado de Minas Gerais. (A) Apresenta os valores-p para o teste de equilíbrio de HW para os nossos dois conjuntos amostrais, com a hipótese nula de que a população está em equilíbrio Hardy-Weinberg. (B) Apresenta os valores-p para o teste de LD para cada par de *loci*. Em ambos os testes para os valores P menores que 0,05, a hipótese nula é rejeitada e os dados apresentam desvios ou desequilíbrios em relação aos valores esperados.

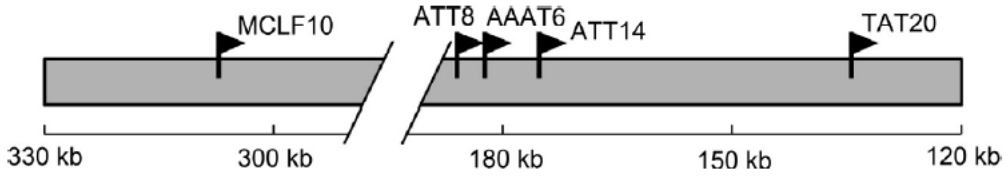

Figura 13. Representação esquemática de um fragmento do cromossomo seis de *T. cruzi* com base no TriTrypDB.org. As bandeiras negras marcam a posição relativa dos cinco *loci* microssatélites utilizados localizados neste cromossomo: MCLF10, ATT8, AAAT6, ATT14 e TAT20.

Tabela 8. Teste de desequilíbrio de ligação para os cinco *loci* microssatélites localizados no cromossomo 6 de *T. cruzi*.

Teste de desequilíbrio de ligação			
Par de *Loci*			**Valor P**
TCAAT8	&	**TCTAT20**	0.00416
TCAAT8	&	**TCAAAT6**	0.00346
TCTAT20	&	**TCAAAT6**	0
TCAAT8	&	**ATT14**	0.11851
TCTAT20	&	**ATT14**	0.25128
TCAAAT6	&	**ATT14**	0.00058
TCAAT8	&	**MCLF10**	0.86894
TCTAT20	&	**MCLF10**	0.61134
TCAAAT6	&	**MCLF10**	0.95698
ATT14	&	**MCLF10**	0.49585

2) Padronização do Bi-PASA

A fim de desenhar os iniciadores específicos para a técnica Bi-PASA, 200 sequências dos genes COII e ND1 provenientes de diferentes linhagens de *T. cruzi* foram alinhadas. Esses genes foram escolhidos por possuírem polimorfismos bem caracterizados que diferenciam as linhagens de *T. cruzi* em três haplótipos mitocondriais e também já serem amplamente utilizados em ensaios de caracterização das linhagens do parasito.

Regiões conservadas entre as sequências foram selecionadas para o desenho dos iniciadores. Em cada caso, os iniciadores diretos possuíam nucleotideos complementares aos SNPs dos alelos específicos de seus haplótipos mitocondriais na extremidade 3´-OH e os iniciadores reversos eram comuns e totalmente complementares a todos os alelos, resultando assim em uma amplificação alelo específica para cada grupo mitocondrial. Além disso, foi adicionada à extremidade 5’ dos iniciadores diretos caudas de tamanhos distintos com sequências de DNA oriundas do vetor de clonagem pUC18, para possibilitar a diferenciação dos tamanhos dos produtos obtidos após os ensaios de PCR.

Para a padronização da técnica Bi-PASA, inicialmente as reações de PCR foram submetidas a um ensaio de gradiente de temperaturas a fim de se esclarecer a melhor temperatura de anelamento a ser utilizada para todas as reações (figura 14). Com base neste experimento a temperatura de 49ºC foi escolhida por ser aquela mais alta na qual as três reações apresentaram *amplicons* em quantidades e tamanho desejado. Após essa

análise, experimentos para determinação dos tamanhos das bandas obtidas pelos iniciadores para as amostras de cada linhagem filogenética foram realizados. Três tamanhos esperados de *amplicons* foram observados pelo sistema de PCR desenvolvido para amostras padrões de linhagens de *T. cruzi* I, II e III (Figura 15). Para garantir que o teste é funcional para caracterização de diferentes populações de *T. cruzi*, um novo ensaio foi realizado com as 12 cepas, mostrando que o padrão esperado para cada linhagem é mantido (Figura 16).

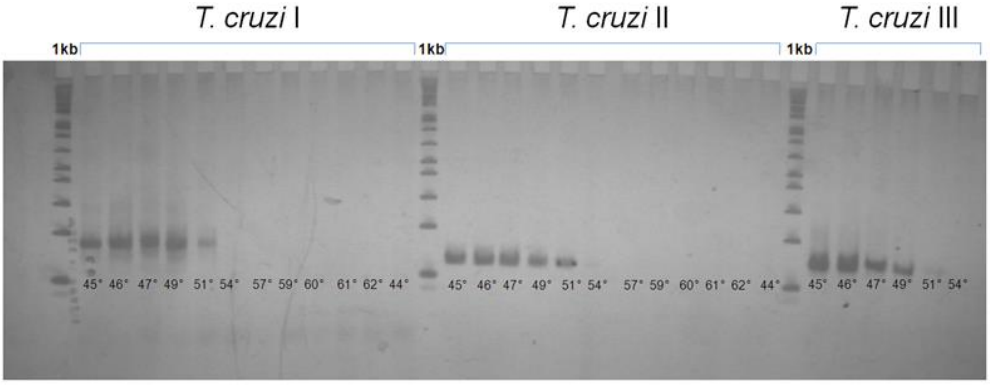

Figura 14. Teste de PCR com gradiente de temperatura. A determinação da temperatura de anelamento de 49ºC foi escolhida a ser utilizada para cada marcador.

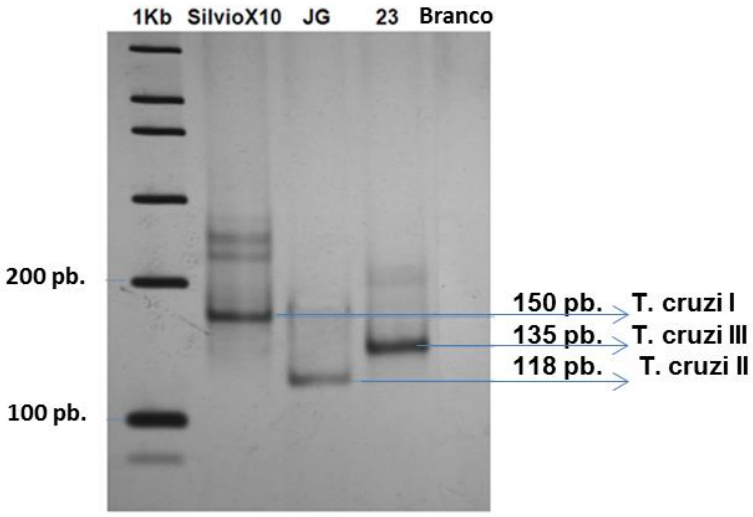

Figura 15. Avaliação do tamanho das bandas obtidas com os iniciadores alelo específicos para populações de *T. cruzi*, pertencentes a diferentes linhagens filogenéticas, em gel de polyacrilamida 6% corado com prata. Por se tratar de uma etapa de avaliação dos tamanhos, a poliacrilamida foi utilizada por sua melhor resolução. As canaletas contêm: (1) padrão de peso molecular 1kb *plus* (Invitrogen); (2) clone padrão representante de TcI – Silvio X10; (3) cepa padrão representante de TcII - JG; (4) cepa padrão representante de TcIII - 231 e (5) branco (controle negativo).

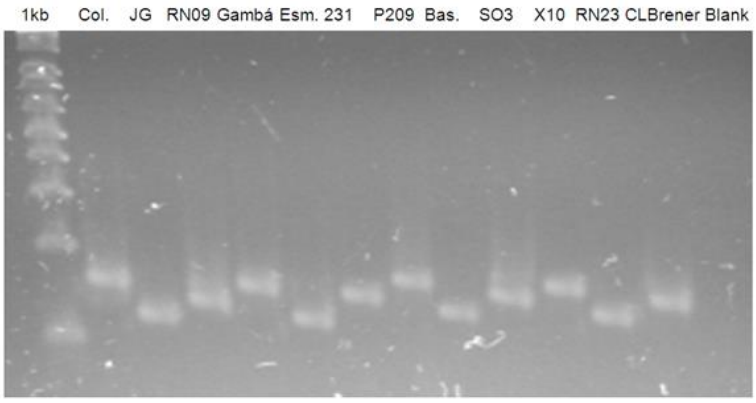

Figura 16. Teste do BI-PASA com populações de *T. cruzi* de diferentes linhagens (TcI - VI) resolvida em gel de agarose 3% corado com brometo de etídio. Visto a boa diferenciação vista anteriormente dos tamanhos dos *amplicons* no gel de polyacrilamida, nós testamos a metodologia na agarose devido a sua maior rapidez. As populações utilizadas como representantes das linhagens TcI (Col 1.7G2, Gambá, P209 e X10), TcII (JG, Esmeraldo, Basileu e RN23) e TcIII-VI (RN09, 231, SO3 e CL Brener), foram amplificados com iniciadores diretos linhagem-específicos para cada linhagem especifica a ser caracterizada (ND11F, ND12F e ND13HF, respectivamente).

Com o intuito de se avaliar, em uma reação multiplex, o comportamento e especificidade dos iniciadores, misturas de DNA de populações de diferentes linhagens (Sylvio X10 – TcI; Esmeraldo – TcII; 231 – TcIII) foram amplificadas utilizando a mistura de três iniciadores diretos e um reverso comum (Figura 17). Neste experimento foram observados perfis bem característicos para cada linhagem.

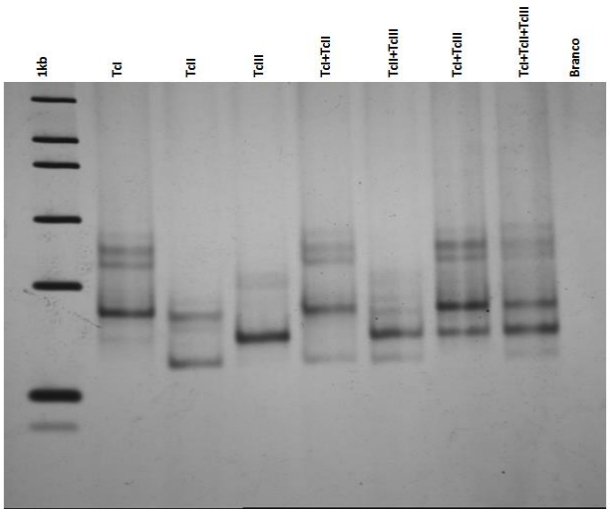

Figura 17. Teste de PCR multiplex do Bi-PASA-ND1. Esse ensaio foi realizado utilizando diferentes misturas de populações de diferentes de linhagens utilizando um

mix de PCR com os três iniciadores diretos linhagem-específicos e o iniciador reverso em comum. Para este experimento foram utilizadas as seguintes populações de *T. cruzi*: Sylvio X10 – TcI; Esmeraldo – TcII; 231 – TcIII.

3) Genômica comparativa entre cepas de *T. cruzi* para análise evolutiva

3.1) Sequenciamento e análise comparativa de um clone da cepa 231 (TcIII)

Para a obtenção de um clone puro da cepa 231 de TcIII, foi realizado um ensaio de clonagem da cepa. Após a clonagem conseguimos isolar nove clones da cepa, que foram submetidos a testes de crescimento e de contaminação, principalmente por micoplasma. Assim, o clone 231 r6 foi selecionado para ser usado para o sequenciamento.

Para o sequenciamento do genoma do clone da cepa 231 foi utilizada a plataforma NGS Illumina HiSeq 2000 para produzir um conjunto de reads para serem montadas. Para estimar a cobertura e o tamanho do genoma de 231, foram utilizados scripts em Perl escritos pelo nosso grupo. Resumindo, a abordagem utilizada calcula a cobertura de cada nucleotídeo derivada de todos os 1.594 genes de cópia única de *T. cruzi* para estimar a cobertura do genoma. Usando esta abordagem, a cobertura do genoma foi estimada em 41,7x. Para estimar o tamanho esperado do genoma 231, o número total de nucleotídeos utilizados na montagem (2.823.893.082 nt) foi divido pela cobertura estimada do genoma, resultando num tamanho aproximado do genoma diplóide do parasito de 67,71 Mb.

A montagem do genoma de 231 foi inicialmente feita por duas diferentes abordagens: montagem *de novo* e baseada em um genoma de referência. Para a montagem *de novo*, foram utilizadas as *pair-end reads* trimadas pelo script VelvetOptimizer.pl que resultou em um *best K-mers* de tamanho 51. Após o preenchimento dos gaps com o programa IMAGE e a correção no programa ICORN, o tamanho do genoma haploide obtido foi de 28,41 Mb , com aproximadamente 13.482 *scaffolds*. O tamanho obtido por esta estratégia de montagem não atingiu aquele estimado pela cobertura, o que sugere que regiões repetitivas podem não ter sido resolvidas na montagem *de novo*. Como segunda estratégia, foi realizada a montagem baseada em genomas de referência. As *reads* do clone 231 foram mapeadas inicialmente a todos os genomas disponíveis de *T. cruzi* a fim de se selecionar o melhor genoma referência para a montagem do genoma de 231. Dentre os genomas referência utilizados o que resultou na melhor cobertura com alta similaridade ao genoma de 231 foi o

genoma do haplótipo Non-Esmeraldo *like* da cepa CL Brener. Assim, usando os programas Bowtie, SAMtools e VCFtools, foi obtido um genoma montado de aproximadamente 32,27 Mb, ou 24,98 Mb sem as regiões de Ns, para o genoma de 231, o que é muito próximo do tamanho esperado, cobrindo cerca de 90% do tamanho do genoma referência usado. Na montagem por referência, entretanto, o genoma apresentou-se muito fragmentado, com cerca de 21.464 *contigs*, montados em *scaffolds* com grandes regiões cobertas por "N", isso ocorreu devido a polimorfismos entre o genoma do clone de 231 e o do genoma da referência.

Para contornar as limitações das duas estratégias escolhidas, uma abordagem de montagem alternativa, combinando as estratégias de montagem baseada em referência e a montagem *de novo.* A partir dessa estratégia combinada, foi obtido uma montagem final com 13.576 *contigs* (N50 = 5.300 pb), 8.471 *scaffolds* (N50 = 14.202), e um genoma haplóide de aproximadamente 35,36 Mb, tamanho este que foi perto do tamanho de 33,85 Mb estimado anteriormente. Esta abordagem combina os melhores elementos da montagem *de novo* (informação específica do genoma sequenciado) com o da montagem de baseada em uma referência (conteúdo repetitivo melhor resolvido), contornando as limitações inerentes a cada estratégia de montagem isolada.

Para avaliar a eficiência da estratégia combinada de montagem proposta no presente trabalho, as métricas da montagem combinada final com as obtidas com a montagem *de novo* e com a montagem baseada em referência, bem como, com as métricas de montagem obtidas para o genoma de outras populações de *T. cruzi* disponíveis em bancos de dados públicos (banco de GeneBank e TriTrypDB). Como mostrado na Tabela 9, a estratégia de montagem combinada resultou numa melhoria de todas as métricas em comparação com a montagem *de novo* e métricas mais próximas àquelas obtidas para as outras cepas de *T. cruzi* que usaram plataformas com *reads* mais longas, principalmente as que usaram a plataforma NGS Roche 454.

Quando analisados os tamanhos dos genomas de populações pertencentes a diferentes linhagens foram observadas diferenças entre os seus tamanhos como esperado e previamente demonstrado (EL-SAYED, MYLER, BLANDIN, et al., 2005; FRANZÉN et al., 2012). Como esperado, as cepas de CL Brener e Tula cl2, ambas TcVI, apresentam tamanho de genoma próximos entre si e relativamente maior do que representantes de outras linhagens devido à natureza híbrida de TcVI.

As três montagens resultantes das estratégias usadas neste trabalho - *de novo*, baseadas em referência e combinada - foram alinhadas e mapeadas nos cromossomos de

CL Brener para visualização da cobertura global de seus *scaffolds* ao longo dos cromossomos, usando o pacote Mummer. Como mostrado na Figura 18, regiões ricas em genes *house-keeping* de cópia única foram cobertas adequadamente, enquanto que a região rica em repetições é mais fragmentada, como esperado, devido às dificuldades técnicas associadas com a montagem correta e completa destas repetições. Todavia, a estratégia de montagem combinada resultou em uma melhor cobertura de regiões altamente repetitivas, quando comparado com a montagem *de novo*, além de preencher também algumas regiões divergentes perdidas na montagem baseada em referência (Figura 18).

Tabela 9 - Comparação entre todas as montagens disponíveis de *T. cruzi* em bancos de dados públicos, mostrando uma melhoria da metodologia proposta no presente trabalho.

Organismo	Tamanho* (MB)	GC%	#scaffolds	Scaffold N50	#contigs	Contig N50	Plataforma
***T. cruzi* CL Brener**	89.94	51.7	29,495	88,624	32,746	14,669	Sanger
***T. cruzi* JR cl.4**	41.48	51.3	15,312	83,591	18,103	7,407	Roche 454
***T. cruzi* Tula cl.2**	83.51	51.4	45,711	7,772	53,083	2,193	Roche 454
***T. cruzi* Esmeraldo**	38.08	50.9	15,803	66,229	20,187	5,353	Roche 454
***T. cruzi* Sylvio X10**	38.59	51.1	-	-	27,019	2,307	Roche 454 + Illumina
***T. cruzi* TcIII-231**							
Montagem *de novo*	28.41	50.0	13,482	3,745	16,684	2,242	
Montagem baseada em Referência	24.98	50.7	NA	NA	21,464	3,239	
Método de montagem *de novo* + baseado em referência	35.35	48.7	8,471	14,202	13,576	5,300	Illumina

*Os tamanhos dos genomas foram obtidos pela contagem de nucleotídeos no genoma sem a presença de "Ns" e está disponível no site: http://www.ncbi.nlm.nih.gov/genome/genomes/25. O genoma de Sylvio X10 ainda não possui os dados de *scaffolds*, somente de *contigs*. NA - não se aplica devido a referência estar montada em um arquivo fasta com os 41 cromossomos gerando um consenso com 41 *scaffolds*.

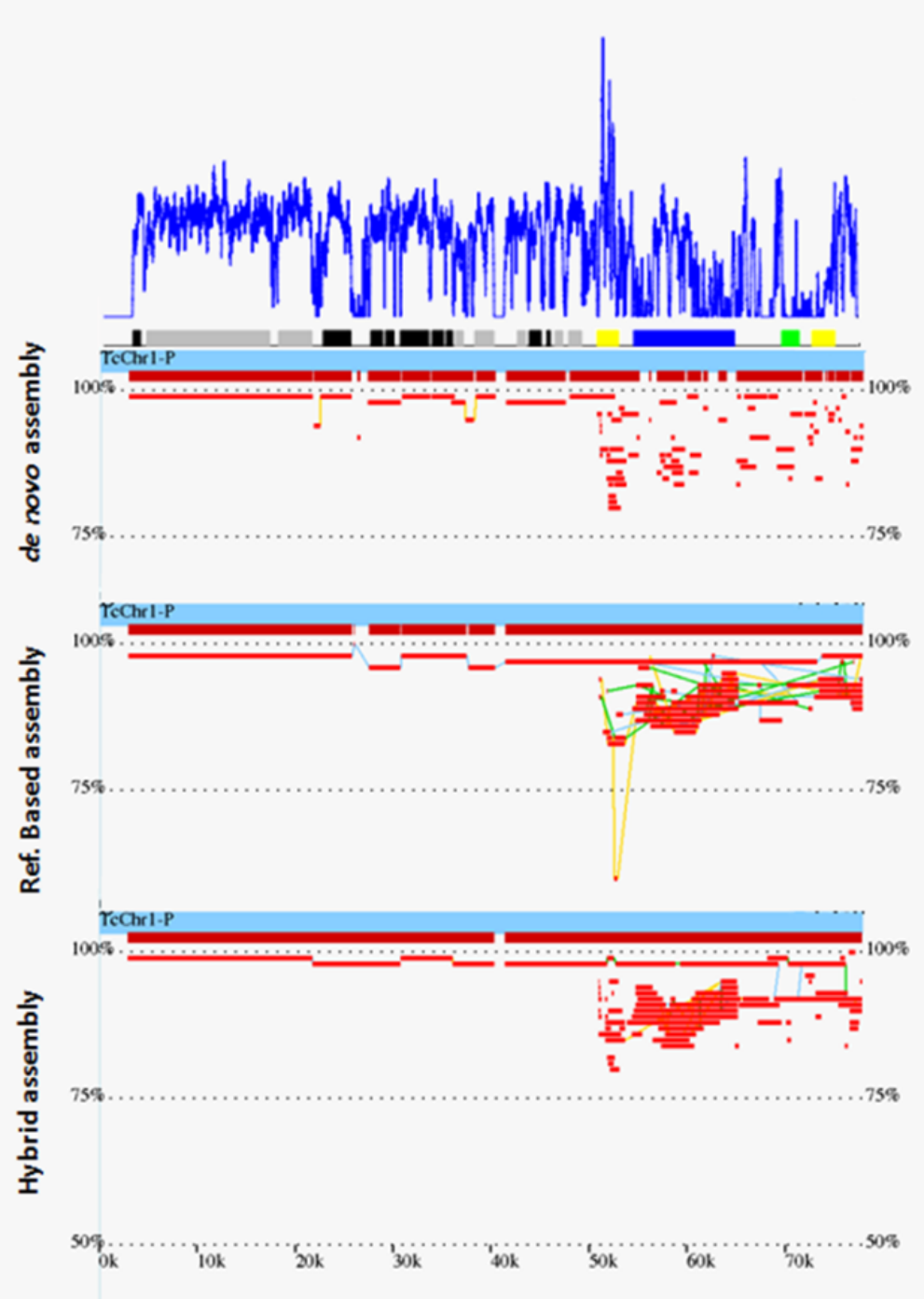

Figura 18. MapView das três estratégias de montagem utilizadas neste trabalho: montagem de novo, baseada em referência e combinada, destacando as diferenças entre a cobertura e alinhamento do genoma montado através da nossa abordagem combinada, em comparação com a montagem *de novo* e a montagem baseada em referência do cromossomo um de *T. cruzi* de Non-Esmeraldo (TcChr1-P) (vermelho). No alinhamento pode-se observar o alto grau de identidade dos *scaffolds* com a referência, sendo próxima a 100% nas regiões de genes de *house keeping* e acima de 75% nas regiões repetitivas. No desenho esquemático do cromossomo um, as cores pretas representam proteínas hipotéticas, em cinza, genes *house keeping*, e em amarelo, azul escuro e verde, genes pertencentes a famílias multigênicas. O gráfico em azul representa a frequência de alinhamentos das *reads* ao longo do cromossomo um obtido durante a montagem baseada em referência.

3.2) Resultados da transferência de anotação

O genoma resultante da montagem combinada foi anotado usando o programa RATT. Esse programa faz a transferência de anotação com base na similaridade e sintenia da sequência não anotada, no caso, o genoma do clone de 231, com as sequências anotadas de um genoma referência no caso, o genoma de CL Brener. Usando essa abordagem de anotação, de um total de 34.803 elementos anotados no genoma de referência, 31.182 foram transferidos para o genoma de 231. Desses elementos transferidos, 29.916 foram completamente transferidos e 1.266 foram divididos em diferentes *scaffolds* (dois ou mais *scaffolds* foram necessários para completar o elemento anotado). Nenhuma transferência parcial, aqui caracterizada por transferência de genes não totalmente cobertos (incompletos no genoma montado), foi encontrada e 3.621 elementos não puderam ser transferidos, sendo estes últimos elementos principalmente constituídos por repetições e proteínas hipotéticas. Um total de 10.592 sequências codificantes (CDS) foram transferidas corretamente a partir dos 10.833 modelos de referências e nenhuma CDS foi parcialmente transferida.

Os conjuntos de dados montados utilizando tanto a abordagem *de novo* quanto a combinada foram avaliados comparando as percentagens dos elementos anotados presentes do genoma de referência transferidos. Avaliando a percentagem de anotações de elementos que não foram transferidos, a montagem combinada teve uma média de 19% de não sintenia com a referência, enquanto a montagem *de novo* teve 32%. Isso demonstra que com a abordagem combinada temos uma melhoria na recuperação de alguns genes antes perdidos usando a montagem *de novo*. A maioria dos elementos anotados não transferidos para as montagens de 231 foram compostos por elementos repetitivos, como elementos transponíveis, famílias de genes multicópia, repetições em tandem, e proteínas hipotéticas.

3.3) Estimativa do conteúdo repetitivo do genoma montado do clone de 231

Para estimar o conteúdo repetitivo do genoma montado do clone de 231 (TcIII), as montagens *de novo* e combinada das sequências foram submetidas aos programas RepeatScout e RepeatMasker. Todos os elementos repetitivos encontrados foram filtrados para remoção de elementos de baixa complexidade e de elementos de repetição

com menos de cinco repetições ao longo do genoma. Em 32% do genoma do clone 231 foram encontradas repetições.

Curiosamente, como mencionado anteriormente, também cerca de 32% das anotações do genoma de referência não foram transferidos para o genoma de 231 montado pela abordagem *de novo*. A diminuição do número de anotações não transferidas usando nossa nova abordagem para a montagem, de 32% para 19%, o que condiz com o aproveitamento de algumas regiões repetitivas, mostrando que a recuperação das informações de conteúdo repetitivo é parcialmente melhorada comparada à montagem *de novo* adotando-se a montagem combinada final proposta neste trabalho.

3.4) Seleção dos genes

Ao longo da evolução dos organismos, cada gene sofre diferentes pressões evolutivas que acabam contando uma história diferente, o que interfere muito quando se quer reconstruir não a história gênica, mas a história do organismo estudado. Por isso, um dos principais fatores responsáveis para uma análise evolutiva de qualidade, é a seleção de um conjunto de sequências significativas e relativamente conservadas, que provavelmente sofreram menos pressão externa ao longo da evolução do parasito.

Para selecionar o melhor conjunto de dados para realizar a análise evolutiva, inicialmente foi avaliada a semelhança do genoma do clone 231 com genomas de populações de diferentes linhagens de *T. cruzi*, mapeando e comparando as *reads* obtidas do clone de 231 com as com os genomas de referência. A comparação das *reads* de 231 mapeadas com os genomas de Sylvio (TcI), Esmeraldo (Tc II), e CL Brener Non-Esmeraldo *like* (TcIII-*like*) mostrou que a amostra é altamente semelhante ao genoma de referência TcIII-*like* com aproximadamente 90% de semelhança, enquanto para as linhagens TcI e TcII só obtivemos uma semelhança de aproximadamente 50% (Figura 19). Os valores foram bem menores para TcI e TcII, pois estes genomas são mais divergentes em comparação ao genoma TcIII-*like*, e como nosso foco nesta etapa foi recuperar regiões bem conservadas, só foi aceita a tolerância de um máximo dois *mismatches* para o alinhamento.

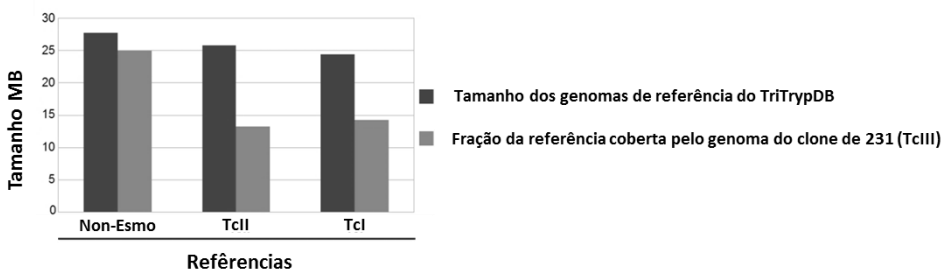

Figura 19. Comparação da similaridade do genoma de 231 com os genomas de referência CL Brener Non-Esmeraldo-*like* (90%) e TcII e TcI (> 55%).

Para avaliar a possibilidade de TcIII, aqui representada pela cepa 231, ser uma linhagem híbrida originada de um evento ancestral de hibridização entre TcI e TcII, as *reads* originadas do genoma do clone de 231 foram identificadas como específicas de TcIII ou compartilhadas com as linhagens TcI e TcII, sete diferentes sets ou conjuntos de *reads* foram identificados e estão esquematizados na figura 7. Após a identificação dos sets das *reads* as mesmas foram montadas em *scaffolds* usando o melhor genoma referência para cada conjunto de *reads*. Para selecionar as melhores sequências candidatas para a análise evolutiva, os *scaffolds* que mapearam para as regiões compartilhadas entre todas as linhagens de referência de *T. cruzi* (subset 7) foram selecionados, visto que as mesmas são bem conservadas entre todas as linhagens. Essa abordagem possibilitaria a diminuição de um viés para falsos agrupamentos de genes que tiveram uma história evolutiva muito diferente. Para a identificação das proteínas presentes no subset 7 foram utilizados o BLASTX contra o proteoma de *T. cruzi* e depois o OrthoMCLDB agrupando estas proteínas em clusters de ortólogos .

Com esta estratégia um total de 6.082 agrupamentos de genes ortólogos de alta qualidade foram inicialmente identificados, mas após a remoção de pseudo-genes e famílias de genes de múltiplas cópias, filogeneticamente pouco viáveis devido à presença de variabilidade, um total de 136 grupos ortólogos foram utilizados para a análise evolutiva. Destes 136 grupos ortólogos, foram selecionados 43 genes presentes em todos os genomas analizados e que atenderam os seguintes requisitos: serem genes de cópia única, terem um alinhamento com o BLAST cobrindo > 95% das sequências de referência de *T.cruzi* com um *e-value* $<1x10^{-30}$, e estarem também presentes em outros representantes evolucionariamente próximos do táxon, como os tripanosomatídeos *T. brucei* e *Leishmania major*, para serem utilizados como raízes nas árvores filogenéticas (Figura 20).

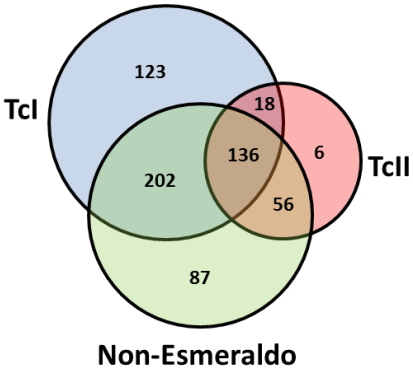

Figura 20. Diagrama de Venn das sequências de 231 (TcIII) mapeadas contra as sequências dos genomas de referência. Os valores indicam o número de grupos ortólogos que são específicos de genes de cópia única e compartilhados entre TcI, TcII e Non-Esmeraldo-*like* (TcIII-*like*). Das 136 sequências conservadas entre as três linhagens, 43 apresentaram alinhamento de BLAST com outgroups (*T. brucei* e *L. major*) acima de 95%, com *e-value* de corte de 1e-10.

3.5) Análise filogenética e estimativa de tempo de divergência de *T. cruzi*

O alto grau de conservação nos marcadores selecionados, permitiu o uso destas sequências proteicas de genes nucleares para reconstruir filogenias intra-específicas. Os 43 *loci* nucleares analisados estão distribuídos aleatoriamente no genoma, mais precisamente, eles estão localizados em 18 dos 41 cromossomos previstos para CL Brener (Figura 21). Para a realização das análises filogenéticas, primeiramente nós alinhamos e trimamos cada um dos 43 grupos de sequências proteicas ortólogas das diferentes linhagens de *T. cruzi* e também concatenamos todos esses *loci* curados para a construção de uma árvore ML referência (CAPELLA-GUTIERREZ, KAUFF e GABALDÓN, 2014).

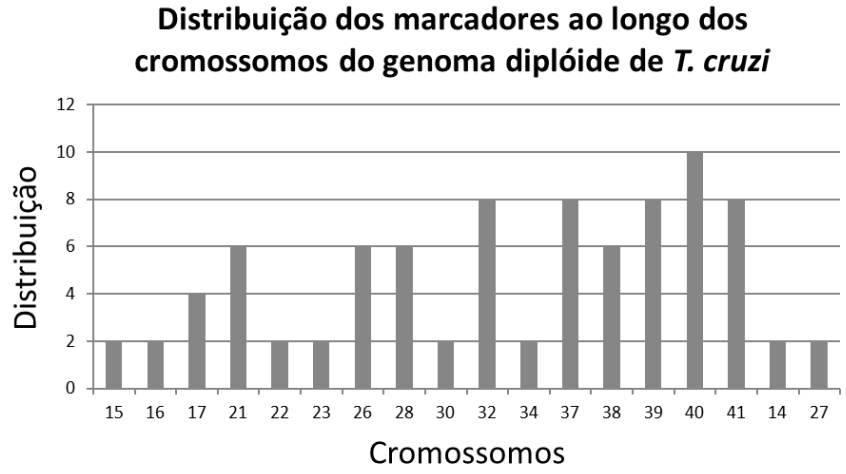

Figura 21. Distribuição dos 43 *loci* nucleares utilizados ao longo do genoma diploide de 231. Eles estão localizados em 18 dos 41 cromossomos previstos de *T. cruzi*.

A análise evolutiva dos 43 *loci* selecionados produziram árvores filogenéticas individuais e onde todas tiveram topologias próximas da árvore ML referência. Alguns desses genes já haviam sido utilizados em trabalhos prévios e a topologia observada foi similar (FLORES-LÓPEZ e MACHADO, 2011) (Figura 22). Esta topologia é consistente com uma história de divergência, em que as cepas de *T. cruzi* II apresentam em determinado momento uma história evolutiva separada das demais linhagens analisadas.

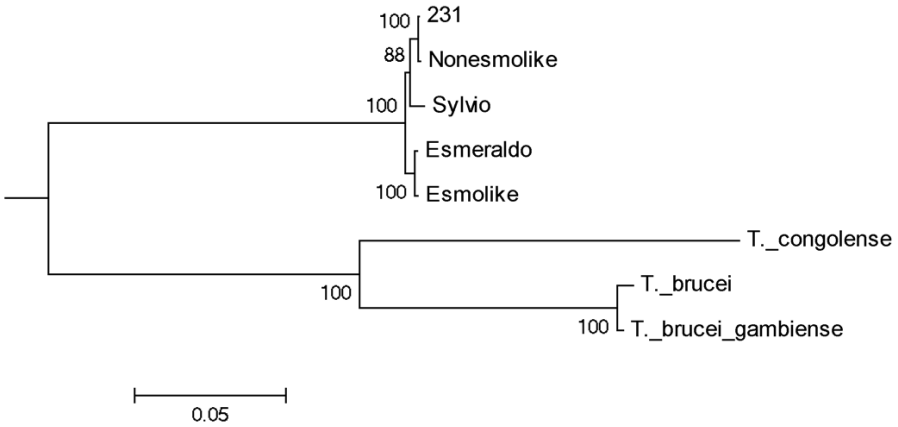

Figura 22. Árvore de ML obtida a partir das sequências com os 43 genes nucleares concatenados. Observa-se uma relação mais próxima entre TcIII e o haplótipo Non-Esmeraldo-*Like* de TcVI (TcIII-*like*), depois para TcI e, tendo TcII como a primeira linhagem a divergir do ancestral comum das linhagens de *T. cruzi*). Para a seleção do melhor modelo de substituição de aminoácidos foi utilizado o programa ProtTest (modelo JTT + G) e na reconstrução foram utilizadas 1.000 amostragens de bootstrap.

Curiosamente, 38 dos 43 *loci* nucleares selecionados, quando caracterizados de acordo com OrthoMCLDB, foram identificados como proteínas hipotéticas somente encontradas em espécies do filo Euglenozoa. Isso significa que esses genes, ainda hipotéticos, poderiam ter um papel importante na evolução deste filo e constituem potenciais genes candidatos para serem utilizados como *barcodes* em ensaios de caracterização. Como ver na árvore ML nuclear, a linhagem TcIII é mais estreitamente relacionada com TcI do que a TcII. A mesma observação pode ser feita analisando o diagrama de Venn anteriormente descrito na qual se observou 325 genes de 231 mais similares a TcI contra 62 mais similares a TcII (Figura 20).

Além do genoma nuclear do clone 231 (TcIII), foi montado o genoma mitocondrial usando 129.292 *reads pair-end* que foram recuperadas durante a filtragem feita para remover estas *reads* do conjunto de dados de sequência bruto a ser utilizado para a montagem do genoma nuclear. Após a montagem do genoma mitocondrial, foi obtido um genoma de 20,23 kpb. Todavia, excluindo-se as regiões mais repetitivas que não foram bem recuperadas devido ao grande número de nucleotídeos identificados por "Ns", apenas 17,17 kpb do genoma mitocondrial foram recuperados incluindo todas as regiões codificantes mitocondriais. A região codificadora do genoma mitocondrial foi alinhada com os genomas mitocondriais de outros tripanosomatídeos utilizados nas analises filogenéticas nucleares, a fim de se comparar as topologias de ambas as árvores. Depois da curadoria feita trimando-se regiões com alinhamento de baixa qualidade das sequencias proteicas alinhadas usando o programa trimAL, uma árvore filogenética ML baseada no genoma mitocondrial foi construída (Figura 23).

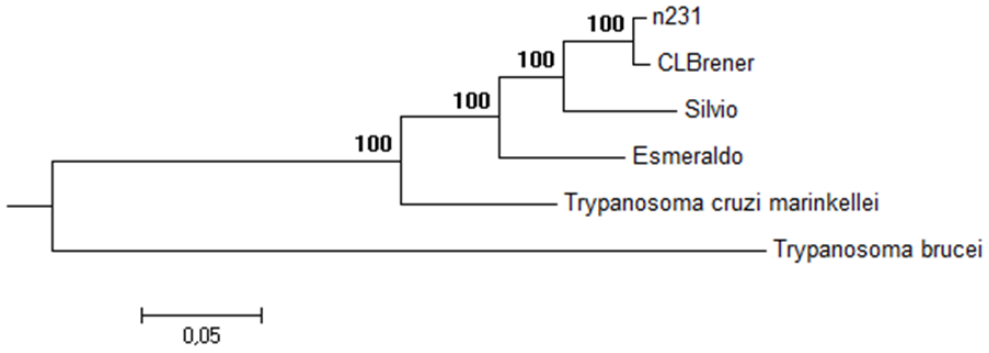

Figura 23. Árvore de ML obtida a partir de sequências do genoma mitocondrial de diferentes tripanosomatídeos, usando o modelo de substituição GTR + G, calculado pelo programa JmodelTest, e com bootstrap de 1000 amostragens.

Como pode ser visto, o padrão observado na árvore construída com base nos dados mitocondriais é semelhante ao da árvore feita para os genes nucleares, o que sugere que a evolução destes genomas é semelhante. Como os cinetoplastídeos têm apenas uma única mitocôndria herdada, seguindo uma herança uniparental, a grande semelhança da mitocôndria do clone de 231 (TcIII) com a do clone CL Brener (TcVI) sugere que TcIII é a linhagem parental doadora do DNA mitocondrial durante o evento de hibridização que originou TcVI. Esses dados em conjunto com os achados nucleares, reforçam a hipótese de um evento de hibridação entre TcII e TcIII como origem de TcVI (FLORES-LÓPEZ e MACHADO, 2011; FREITAS, DE et al., 2006).

Para todas as árvores filogenéticas nucleares obtidas foi realizado um cálculo aproximado de LRT (Likelihood Ratio Test) para avaliar se todos os *loci* utilizados evoluíram sob relógio molecular. Assim, dos 43 *loci* analisados, 30 tiveram um valor de ramo paramétrico baseado em chi-quadrado perto de um, o que indica que os *loci* utilizados estão sob relógio molecular. Com base nesse achado todos esses 30 *loci*, considerados sob relógio molecular, foram concatenados para reconstruir uma árvore filogenética de ML referência.

Adicionalmente, para os 30 *loci* cuja função de verossimilhança não pôde rejeitar o relógio molecular, foi investigada a existência de um homólogo em *T. brucei*, genoma evolutivamente mais próximo de *T. cruzi* e disponível para ser usado como *outgroup*. Assim, os 30 *loci* que passaram no teste paramétrico foram selecionados para fazer parte de um conjunto de dados concatenados adequado para executar uma análise bayesiana de tempo de divergência.

O tempo de divergência foi estimado para os dois conjuntos de dados obtidos: toda região codificadora mitocondrial e para os 30 *loci* nucleares. Como pode ser observado na tabela 10, o tempo de divergência estimado a partir do conjunto de dados mitocondriais diferiu significativamente das estimativas obtidas a partir dos *loci* concatenados nucleares. O tempo estimado para o ancestral comum mais recente (TcMRCA) usando os dados mitocondriais sugere que as linhagens de *T. cruzi* divergiram durante o Mioceno (TcMRCA = 13.36 mya), estimativas estas que são semelhantes aos apresentados por outros trabalhos usando dados mitocondriais (FLORES-LÓPEZ e MACHADO, 2011; MACHADO e AYALA, 2001). Por outro lado, as datas estimadas com os dados concatenados dos 30 *loci* nucleares apontam no sentido de uma origem do *T. cruzi* entre o Pleistoceno e o Plioceno (TcMRCA = 2.74 mya (*strict*); TcMRCA = 3.05 mya (*relaxed*)) (Tabela 10, Figura 24). Essas datas

também são semelhantes aos tempos de divergência anteriormente estimados utilizando diferentes *loci* nucleares (Machado: TcMRCA = 3,91 mya e Flores-López: 2.18 mya) (FLORES-LÓPEZ e MACHADO, 2011; MACHADO e AYALA, 2001). Estimativas de tempo divergência muito semelhantes foram obtidas no conjunto concatenado dos *loci* nucleares que tiveram um homólogo em *T. brucei* (43 *loci*), mesmo incluindo na análise genes que rejeitaram a hipótese de relógio molecular (TcMRCA = 3,38 mya).

Tabela 10 - Estimativas bayesianas de tempo de divergência (em mya) para as diferentes linhagens de *T. cruzi*.

Modelo de relógio	T.cruzi[a]	TcI-TcIII[b]
Loci nuclear[c] (30 loci)		
Strict	2.74	1.84
Relaxed lognormal	3.38	1.87
Genoma Mitocondrial		
Strict	13.36	11.84

T.cruzi[a]- Ancestral comum de todas as linhagens de *T. cruzi*;
TcI-TcIII[b]- Ancestral comum de TcI e TcIII.
[c] – nesta análise foram usados apenas os 30 loci nucleares que se apresentaram sob relógio molecular.

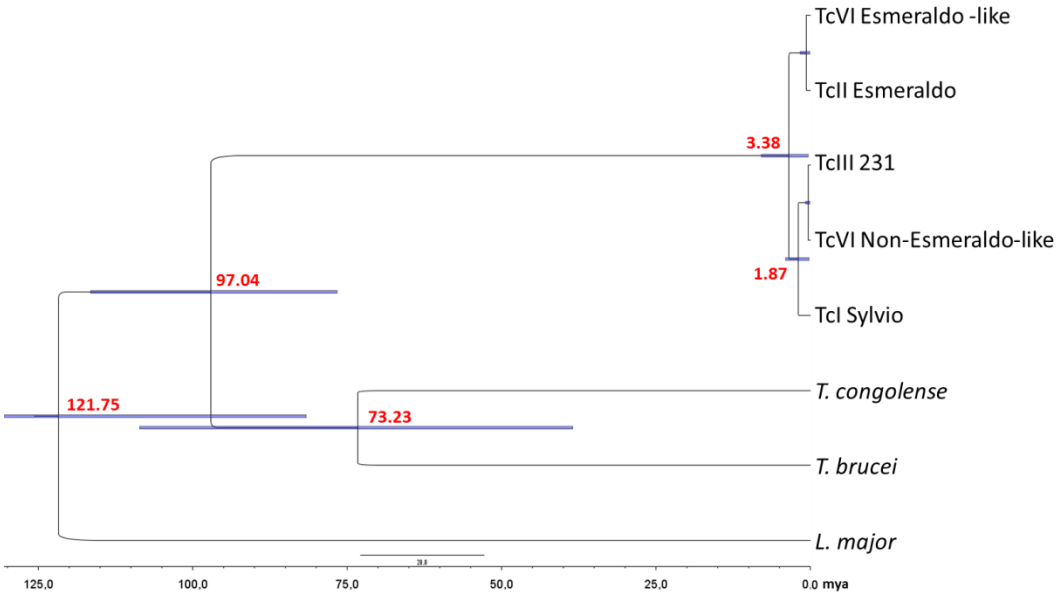

Figura 24. Tempos médios de divergência para as principais linhagens de *T. cruzi* utilizadas no trabalho, utilizando os *loci* nucleares com o modelo de relógio relaxado (*relaxed lognormal*). Conjunto de dados consiste de um alinhamento de 30 *loci* nucleares concatenadas onde o relógio molecular não foi rejeitado, e que tiveram pelo menos um homólogo em *T. brucei*. Barra de escala utilizada está em milhões de anos (mya).

4) Aplicação do algoritmo de predição de clusterização em populações naturais de *T. cruzi*,.

4.1) Decomposição de valores singulares (SVD)

Para o desenvolvimento de uma nova ferramenta de clusterização baseada em sequências gênicas, o gene da amastina foi escolhido como gene alvo, uma vez que um número bastante grande de sequências deste gene encontrava-se disponível nos bancos de dados de sequência. Após a transformação das sequências do gene amastina, que foram previamente alinhadas e curadas, em uma matriz M numérica pelo algoritmo "*SlideBlock*" (SB), a frequência de blocos de tetranucleotídeos presentes em cada sequência foi calculada, com a qual foi possível realizar o cálculo do SVD do *dataset*. Como pode ser visualizado na Figura 25, o espectro de valor singular obtido a partir da análise do SVD na matriz M mostra quantos grupos podem ser escolhidos no banco de dados, sendo estes classificados com base no princípio da Curva-L (BRAGA, 2001), ou seja, baseados na conversão dos seus valores relativos na população no gráfico (COUTO, SANTORO e SANTOS, 2011). Esse princípio diz que após o nivelamento da curva, onde o valor no eixo Y, representado pelos valores relativos, começa a ficar constante, as análises tendem a gerar mais ruídos.

De acordo com o gráfico de SVD e com base no princípio da Curva-L (LAWSON e HANSON, 1995), os valores singulares três e quatro foram escolhidos como ideais para serem utilizados nas análises. Os dados obtidos foram então traçados em um gráfico tridimensional testando ambos valores singulares, e ao correlacionarmos as duas topologias obtidas a variação foi mínima. Como não havia diferença significativa entre três e quatro valores singulares, foi escolhido o menor valor singular convergido inicialmente, por permitir uma análise mais rápida e simples. Esse valor foi utilizado para definir o número de grupos k a serem definidos em nosso método de agrupamento *k-Means* (k = 3). Ademais, neste caso, o valor k = 3 também coincide com o número de perfis esperados, já que temos três linhagens filogenéticas não híbridas, das quais as sequências de amastina foram analisadas.

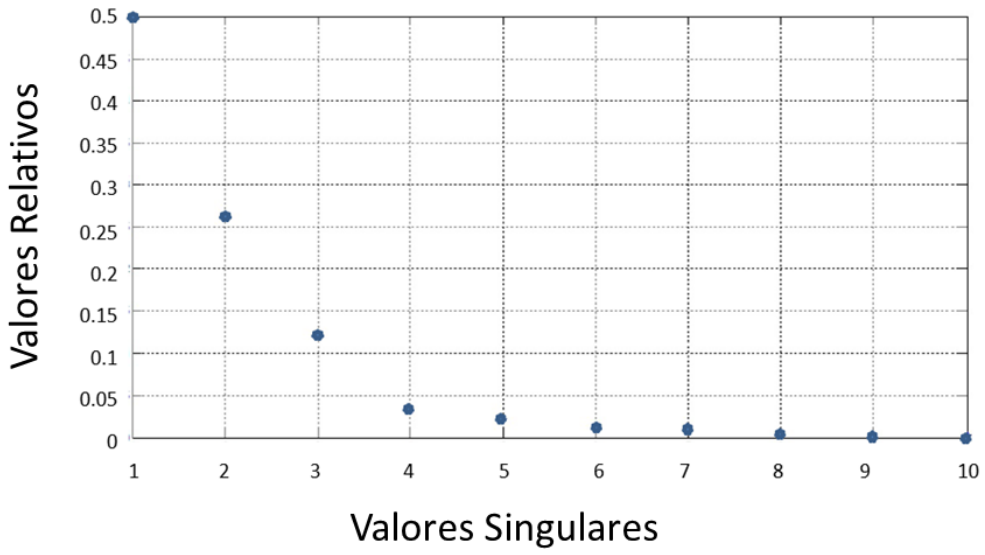

Figura 25. Gráfico de valores relativos de SVD: são mostrados os dez primeiros valores singulares de M obtidos pela fatoração do SVD traçados em ordem decrescente. O eixo X corresponde ao índice de valores singulares e o eixo Y aos valores relativos. Observa-se que a Curva-L ocorre entre os valores de k = 3 e k = 4.

4.2) Clusterização

A partir do valor de k selecionado com base nos valores SVD, como descrito no item anterior, os grupos obtidos com dois métodos de agrupamento diferentes usados para clusterizar as sequências de amastina foram comparados. O primeiro método avaliado, que constituiu a estratégia proposta neste trabalho, foi usar o valor k, obtido pelo SVD a partir da matriz gerada pelo cálculo da distância euclidiana, no algoritmo *k-Means*. Após rodar o algoritmo com k=3, reiterado 100 vezes, foi gerado um gráfico consenso com três grupos bastante distintos (Figura 26A).

Todas as 294 sequências de amastina foram classificadas em um dos três grupos identificados pelo SVD como mostrado na Figura 26A. Foi observado que as sequências derivadas de cepas pertencentes à linhagem TcI (Colombiana, Sylvio e Dm28) se encontravam no grupo SVD1 (verde), as sequências derivadas das cepas pertencentes à linhagem TcII (Esmeraldo e JG) foram agrupadas no SVD2 (azul escuro) e as sequências derivadas do clone da cepa TcIII 231 agruparam-se no grupo SVD3 (Azul Claro). Como esperado para uma linhagem híbrida parte das sequências derivadas do clone de CL Brener (TcVI) foram agrupadas no grupo SVD2 e parte no SVD3.

Para avaliar se a estratégia proposta neste trabalho é uma boa abordagem para ser utilizada para clusterização de sequências gênicas como as de amastina, os resultados obtidos foram comparados com um segundo método de clusterização por distância, uma das abordagens de agrupamento filogenéticos mais utilizadas na literatura. Usando o modelo de substituição Kimura-2-parâmetros, e o método de reconstrução *Neighbour-Joining*, o comportamento da clusterização foi muito similar como pode ser observado na topologia da árvore gerada (Figura 26B), sugerindo que o algoritmo proposto funciona bem para análises de agrupamento. Para cada grupo obtido, uma variável binária de saída foi criada a partir da matriz de frequência já decomposta com seus valores singulares (SVD1, SVD2 e SVD3), e utilizada em um modelo de regressão logística construído para prever a possibilidade de um bloco pertencer a um determinado grupo.

4.3) Regressão Logística

Esse é um teste padrão que pode ser usado para identificar padrões que são responsáveis pelo agrupamento de cada grupo SVD. Na Tabela 11 são mostrados os blocos de tetranucleotídeos com os coeficientes de regressão dos blocos com valores significativos. Todas as análises foram feitas considerando como nível significativo o valor P de 0,05 após a regressão logística, usando o método de "*forward stepwise*". A razão de chances para cada tetranucleotídeo é calculada pelo $e^{\beta i}$, onde β_i é o coeficiente de regressão, que resume o sentido e a importância da frequência de cada tetranucleotídeo para caracterizar um determinado grupo SVD. Se esta razão for maior do que 1,0 ($\beta_i > 0$), o bloco pertence a um grupo SVD. Se a razão é inferior a 1,0 ($\beta_i < 0$), o bloco não pertence a um grupo SVD específico.

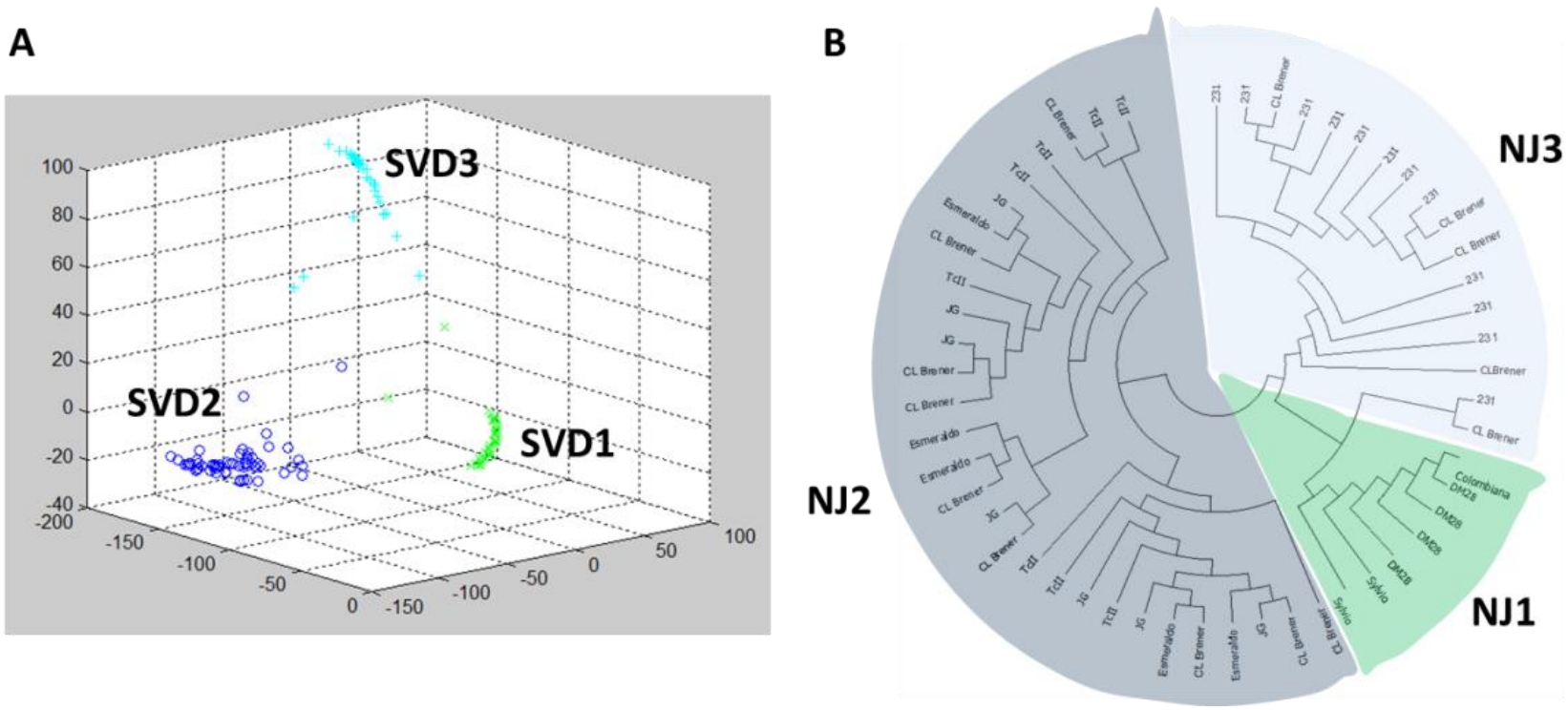

Figura 26. Gráficos obtidos com dois métodos de clusterização diferentes para as 294 sequências de amastina analisadas. (A) Gráfico SVD gerado pelo algoritmo de clusterização proposto neste trabalho. O cálculo de distância euclidiana foi aqui utilizado com o cálculo de SVD para a plotagem. Para o cálculo do k-means foi utilizado (k=3). (B) Cladograma não enraizado obtido por *Neighbor-Joining* (NJ), mostrando a clusterização baseada em distância de Kimura 2-Parâmetros. Em ambos os métodos parte das sequências de amastina do clone CL Brener pertencente à linhagem híbrida TcVI foi agrupada em SVD2 (NJ2) e parte em SVD3 (NJ3). Na árvore de NJ os ramos caracterizados com TcII significam que tanto Esmeraldo quanto JG estavam presentes nos mesmos. As sequências derivadas de cepas pertencentes à linhagem TcI (Colombiana, Sylvio e Dm28) se encontravam no grupo SVD1 (verde), as sequências derivadas das cepas pertencentes à linhagem TcII (Esmeraldo e JG) foram agrupadas no SVD2 (azul escuro) e as sequências derivadas do clone da cepa TcIII 231 agruparam-se no grupo SVD3 (Azul Claro).

Como pode ser visto na Tabela 11, apenas dois ou três dos 256 tipos de blocos identificados foram importantes para caracterizar cada cluster. A presença dos tetranucleotídeos GCGG, AAGT e CTGT, com coeficientes de regressão (β_i) positivos, aumentam a probabilidade de a sequência pertencer ao grupo SVD1. Para a identificação do padrão do grupo SVD2 apenas dois blocos também com coeficientes de regressão positivos foram considerados relevantes (CTGC e TGCT). E finalmente, para a identificação do grupo SVD3, dois blocos foram significativos (CCAC e CCGC), com dois tipos de coeficiente de regressão: um positivo, cuja presença indica o aumento da probabilidade da sequência pertencer ao grupo SVD3, e um negativo, que indica que sua presença diminui esta probabilidade (Tabela 11).

Tabela 11 - Análise de regressão logística por bloco de tetranucleotídeos considerado relevante para caracterizar os três clusters identificados.

Clusters	Bloco de Tetranucleotídeo	Coeficiente de regressão (β_i)	Desvio padrão	Valor P
SVD1 (β_0= -32.11)				
	GCGG	3.64	1.26	0.0163
	AAGT	4.80	1.71	0.045
	CTGT	2.60	1.30	0.0436
SVD2 (β_0= -32.09)				
	CTGC	4.94	1.59	0.0000
	TGCT	3.46	1.07	0.0498
SVD3 (β_0= -13.22)				
	CCAC	-2.98	1.60	0.05
	CCGC	3.36	1.02	0.01

Para ilustrar o efeito da presença dos blocos de tetranucleotídeos relevantes na sequência de amastina para identificação de cada cluster, a figura 27 ilustra o comportamento do modelo logístico para o grupo SVD1, quando a frequência do bloco AAGT aumenta em uma sequência de amastina, e as frequências dos demais blocos do mesmo grupo são mantidas constantes. Observa-se que a presença de quatro ou mais cópias deste bloco de tetranucleotídeo na sequência aumentam fortemente a chance da sequência de amastina pertencer ao grupo SVD1.

Para mostrar o quanto a diferença destes blocos tetranucleotídeos foi significativa entre os grupos SVD preditos, a tabela 12 mostra a frequência média dos blocos de tetranucleotídeos identificados por meio de regressão logística para cada

grupo SVD, sendo eles comparados pelas frequências de um grupo SVD contra os demais. Os modelos de regressão logística foram validados utilizando 82 sequências de amastina que foram randomicamente selecionadas a partir do nosso conjunto de dados analisado. Sensibilidade e especificidade foram superiores a 90% para a utilização desses blocos de tetranucleotídeos selecionados como determinantes para a diferenciação dos três grupos SVD avaliados (Figura 28).

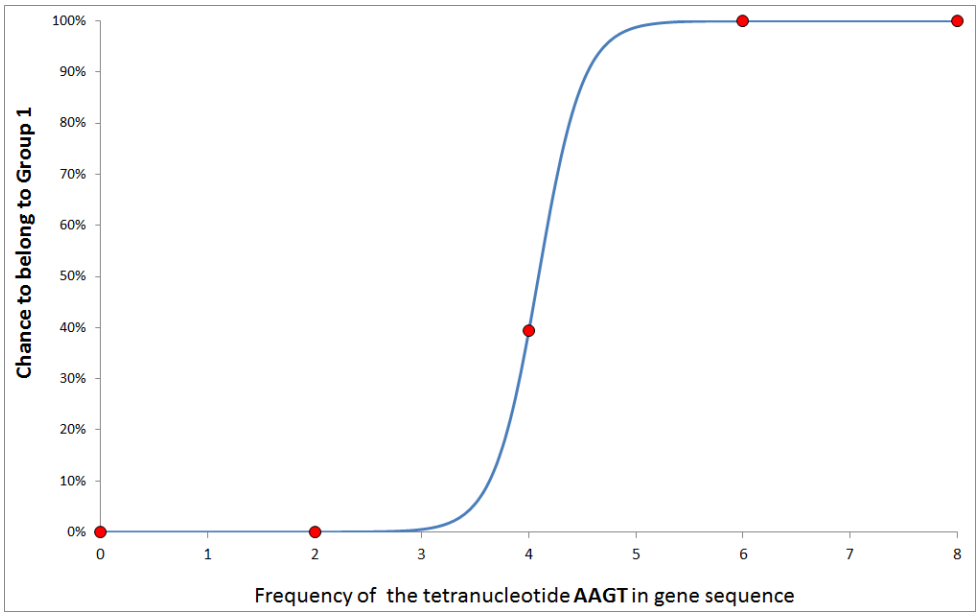

Figura 27. Comportamento do modelo logístico para o grupo SVD1 em função da frequência do bloco AAGT. Mantidas as frequências dos demais blocos de tetranucleotídeos relevantes (GCGG e CTGT), pequenas mudanças na frequência do bloco AAGT aumentam as chances de a sequência pertencer ao grupo SVD1.

Tabela 12 - Média de frequência dos blocos de tetranucleotídeos de grupos SVD específicos contra os demais grupos.

Média de frequência dos Blocos de Tetranucleotídeos			
Grupo SVD1 *versus* outros grupos			
Bloco de Tetranucleotídeo	Grupo SVD1	Outros Grupos	Valores P
GCGG	2,8 +- 0,86	0,3 +- 0,52	<0,001
AAGT	4,4 +- 0,56	2,0 +- 0,25	<0,001
CTGT	4,4 +- 0,66	3,0 +- 0,20	<0,001
Grupo SVD2 *versus* outros grupos			
Bloco de Tetranucleotídio	Grupo SVD2	Outros Grupos	Valores P
CTGC	3,3 +- 0,52	1,1 +- 0,34	<0,001
TGCT	7,0 +- 0,09	5,6 +- 0,54	<0,001
Grupo SVD3 *versus* outros grupos			
Bloco de Tetranucleotídio	Grupo SVD3	Outros Grupos	Valores P
CCAC	1,1 +- 0,29	2,4 +- 0,65	<0,001
CCGC	6,7 +- 0,55	3,2 +- 0,95	<0,001

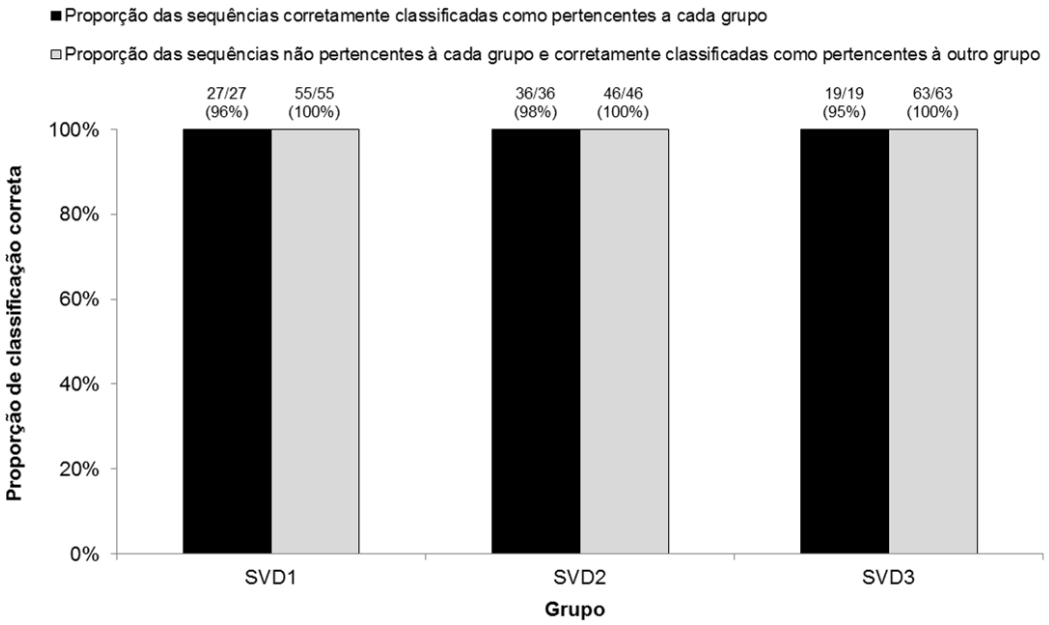

Figura 28. Teste de sensibilidade e especificidade dos blocos para determinação dos grupos SVD representados pelas taxas de verdadeiros positivos e negativos para cada modelo logístico na classificação das sequências: se a chance de uma sequência de pertencer um grupo é superior a 90%, então ela é classificada no grupo modelado. Caso contrário, ela é classificada como pertencente a outro grupo.

O número de sítios significativos obtidos pela regressão logística para cada grupo SVD foi de três, dois e dois para SVD1, SVD2 e SVD3, respectivamente. A baixa presença ou ausência de blocos específicos nas sequências de amastina de outros grupos SVD mostra que a estratégia proposta neste trabalho foi eficiente. Esses resultados foram reunidos em um manuscrito em fase final de elaboração.

DISCUSSÃO

IV. DISCUSSÃO

Embora as sequências completas dos genomas de cinco clones de *T. cruzi*, pertencentes a diferentes linhagens, tenham sido publicadas, muitos aspectos da estrutura populacional e evolução deste protozoário ainda permanecem obscuros. A espécie *T. cruzi* apresenta uma grande heterogeneidade tanto genotípica quanto fenotípica, provavelmente resultante do seu modo de reprodução e o isolamento geográfico das de suas populações, que causam um acúmulo de mutações e características próprias. Muitos trabalhos têm sido realizados na tentativa de evidenciar as características peculiares de cada isolado de *T. cruzi*, sendo que variações de forma, virulência, dinâmica de crescimento, sequências gênicas e até mesmo número de cromossomos foram demonstradas.

Com relação aos fatores genéticos do parasito, uma variedade de estudos biológicos e moleculares revelou uma substancial variabilidade genética entre as cepas de *T. cruzi*. De fato, em 2009, foram classificadas em seis linhagens ou DTU's principais (TcI-VI) para este parasito. Os aspectos epidemiológicos associados às diferentes linhagens ainda não estão bem definidos, mas já há evidências de associação com a distribuição geográfica, ciclos de transmissão e aspectos clínicos da doença. Com isso, a fim de se discutir profundamente a estrutura populacional desse parasito, os esforços do nosso grupo de pesquisa têm sido focados no desenvolvimento de metodologias para a caracterização do *T. cruzi* e sua aplicação no estudo da epidemiologia molecular da doença de Chagas.

A estrutura populacional do parasito ainda é bastante discutida, em parte devido à falta de dados genômicos de todas as linhagens de *T. cruzi*, o que torna os estudos limitados. Atualmente, há disponível em bancos de dados genomas de apenas três linhagens filogenéticas de *T. cruzi* (Sylvio e JR – TcI; Esmeraldo – TcII; CL Brener e Tula – TcVI), e verifica-se a concentração de estudos evolutivos somente em poucos grupos gênicos bem definidos, que nem sempre são os ideais para esse tipo de estudo (FREITAS, DE et al., 2006a; PINTO et al., 2012; ZINGALES et al., 1999). Assim, torna-se necessária a geração de novos dados genômicos de *T. cruzi*, bem como o desenvolvimento de novas estratégias capazes de identificar os melhores genes candidatos para as análises evolutivas, visando ao refinamento dos estudos sobre a estrutura populacional do parasito.

Assim, neste trabalho nos propusemos a desenvolver e aperfeiçoar procedimentos moleculares, e computacionais que permitissem investigar e dissecar a complexidade da estrutura populacional de *T. cruzi*. Para tanto, foram nossos objetivos neste trabalho: (i) investigar se as trocas genéticas entre cepas de *T. cruzi* são mais frequentes que o esperado; (ii) desenvolver novas metodologias de caracterização molecular capazes de diferenciar as diferentes linhagens de *T. cruzi* em futuros ensaios multiplex de PCR; (iii) sequenciar o genoma de um representante da linhagem TcIII, ainda não disponível nos bancos de dados públicos e realizar uma análise comparativa com seus dados genômicos, a fim de se resolver a origem ancestral das seis linhagens filogenéticas e (iv) desenvolver uma metodologia computacional de agrupamento capaz de reconhecer padrões responsáveis pela divisão das linhagens e gerar regiões candidatas para o desenho de iniciadores para ensaios de caracterização molecular.

1) As trocas genéticas entre cepas de *T. cruzi* são mais frequentes que o esperado

Embora a reprodução sexual já tenha sido indubitavelmente demonstrada para *T. cruzi* (GAUNT et al., 2003; RAMÍREZ et al., 2012), muitas questões relativas à frequência e importância destes eventos e sua contribuição para a determinação da estrutura populacional e biologia destes parasitos continuam muito debatidas. Em especial, merece destaque o debate sobre o modo reprodutivo do parasito se essencialmente clonal com trocas genéticas eventuais, ou se essas trocas são mais frequentes, caracterizando uma reprodução sexual. No presente estudo foi empregada uma estratégia inovadora para se avaliar a dinâmica populacional TcII simultaneamente em escalas geográficas local e abrangente usando conjuntos de marcadores polimórficos nucleares e mitocondriais. As abordagens experimentais utilizadas neste estudo foram projetadas para se investigar melhor a clonalidade anteriormente assumida como o mecanismo predominante para a reprodução de *T. cruzi* (TIBAYRENC e AYALA, 2002).

Dados contrastantes na literatura, tais como a presença de homozigotos e de heterozigotos correspondentes nos estoques e clones de *T. cruzi* provenientes de um mesmo hospedeiro circulante em uma mesma área em populações naturais, assim como a confirmação de eventos de hibridização *in vitro*, e o uso frequente de isolados de linhagens, ciclos de transmissão e localidades diferentes, sinalizavam que o pressuposto

de uma estrutura essencialmente clonal para o táxon poderia não refletir a realidade (BOGLIOLO, LAURIA-PIRES e GIBSON, 1996; GAUNT et al., 2003).

Assim, neste trabalho, foram utilizados alguns dos parâmetros de genética populacional mais comuns descritos na literatura, como equilíbrio de HW e LD, para investigar a estrutura populacional de dois conjuntos de populações de TcII: um conjunto que incluiu cepas isoladas a partir de uma ampla região geográfica (lugares diferentes na América Latina) e um conjunto menor, incluindo cepas isoladas de uma área mais restrita (Minas Gerais, Brasil). A ideia principal desta abordagem foi avaliar a potencial influência de barreiras geográficas nos aparentes desequilíbrios de HW e LD que normalmente são detectadas em estudos populacionais, desses parasitos e que levou à proposição da estrutura essencialmente clonal de *T. cruzi* (AYALA, 1993; OLIVEIRA, R P et al., 1998; TIBAYRENC e AYALA, 1987, 2002; WAHLUND, 1928).

No presente estudo, a ocorrência de LD e desvios de HW foram detectados quando cepas de *T. cruzi* isoladas de regiões geograficamente distantes foram analisadas, o que poderia ser incorretamente interpretado como indicativo de reprodução clonal predominante. No entanto, esse cenário foi completamente modificado quando foram analisados apenas isolados que circulavam em uma mesma área geográfica. Neste último caso, o equilíbrio de ligação e de HW foram restaurados para a maioria dos *loci* analisados, indicando que as distâncias geográficas e/ou barreiras físicas podem ser fatores importantes na redução de oportunidades para a reprodução sexual, logo, afetando aparentes as taxas de recombinação entre as cepas de *T. cruzi*.

Curiosamente, esses eventos de recombinação não ocorreram apenas por segregação independente, mas também por recombinação alélica. Isso foi evidenciado pela presença de recombinação aparentemente mais frequente entre genes localizados longe uns dos outros em um mesmo cromossomo (cromossomo 6).

Assim, os resultados obtidos neste trabalho são consistentes com uma subestruturação da população conduzida provavelmente pelo efeito Wahlund (um fenômeno comum observado quando os indivíduos de uma população são provenientes de subpopulações geneticamente segregadas), em que os desequilíbrios dos marcadores genéticos contribuíram para uma subestimativa da ocorrência de sexo e recombinação na população de *T. cruzi*. Esses achados foram particularmente relevantes, pois a

maioria dos estudos anteriores, utilizaram cepas isoladas de áreas geograficamente muito distantes para este tipo de análise.

A estratégia escolhida neste estudo para analisar cepas de *T. cruzi* isoladas do mesmo local (Minas Gerais) e da mesma linhagem TcII para investigar a extensão da recombinação homóloga e da segregação alélica aleatória entre os parasitos sugere que as trocas genéticas entre essas linhagens foram mais frequentes do que o inicialmente esperado. Se isto é uma característica específica de algumas populações de *T. cruzi* ou uma característica geral do táxon *T. cruzi* ainda falta ser esclarecido. Todavia, Llewellyn *et al*. (2009), utilizando 48-49 *loci* microssatélites para investigar a estrutura da população das linhagens TcI e TcIII, observou um excesso de homozigose, um achado incongruente com os modelos extremos de evolução clonal de longo prazo em organismos diplóides.

Da mesma forma, Ocaña-Mayorga *et al*. (2010), utilizando 10 *loci* microssatélites e 81 isolados de TcI provenientes de populações de triatomíneos e pequenos mamíferos de 16 comunidades da província de Loja, no sul do Equador, também identificaram frequências alélicas de equilíbrio de HW e equilíbrio de ligação mesmo entre os *loci* fisicamente ligados.

Ao analisar um gene nuclear (glicose fosfato isomerase - GPI) e um mitocondrial (NADH desidrogenase subunidade 1 – ND1) de 60 isolados de TcI e 15 cepas de referência pertencentes às seis linhagens de *T. cruzi*, Barnabé e Brenière (2012) identificaram indícios de introgressão trans-linhagem mitocondrial. Este tipo de evento, que é o fluxo de genes de uma linhagem para o acervo genético de outra através de repetidos retrocruzamentos entre um híbrido e sua original geração progenitora. Este tipo de evento inicialmente considerado raro em tripanosomatídeos foi observado pelo menos em dez casos independentes: duas vezes entre as linhagens TcII e TcIII (FREITAS, DE et al., 2006a) e oito vezes entre TcI e TcIII ou TcIV (LEWIS et al., 2011; MACHADO e AYALA, 2001; MESSENGER et al., 2012).

Finalmente, ao contrário da visão predominante de que eventos de hibridização em *T. cruzi* são antigos e de pouca importância epidemiológica, Lewis *et al*. (2011) dataram eventos evolutivos chave do táxon, incluindo o surgimento de linhagens híbridas TcV e TcVI, como ocorridos nos últimos 60 mil anos, o que indica que a recombinação ainda é ativa no taxon.

Concluindo, tomados em conjunto esses resultados sugerem que as trocas genéticas entre cepas de *T. cruzi* ocorrem com mais frequência do que inicialmente

esperadas e essa informação pode ser importante para moldar e definir propriedades biológicas das linhagens filogenéticas deste taxon. No entanto, a relevância destes eventos de recombinação para todo o taxon em ambas as escalas evolutivas e de gerações continuaram a ser estabelecidas (TIBAYRENC e AYALA, 2013).

Esses resultados foram utilizados para a redação de um artigo "Evidence of substantial recombination among *Trypanosoma cruzi* II strains from Minas Gerais" publicado no periódico "Infection, Genetics and Evolution" (Anexo 1).

2) Nova técnica de caracterização de três linhagens de *T. cruzi* baseada em um gene mitocondrial

Como mencionado anteriormente, apesar de já existirem diversas técnicas de caracterização das diferentes linhagens de *T. cruzi*, hoje poucos ensaios são simples e conseguem diferenciar as seis linhagens.

A perspectiva inicial quase utópica de se encontrar um único marcador capaz distinguir as seis linhagens de *T. cruzi*, mostrou-se dependente de mais dados genômicos de linhagens cujos genomas ainda não foram sequenciados, para uma análise global dos polimorfismos.

Neste presente trabalho nos propusemos simplificar a metodologia proposta por Macedo & Segatto (2010) baseada em um ensaio triplo (Figura 4). Dentre as etapas dessa proposta, uma das técnicas utilizadas é a PCR RFLP do gene mitocondrial COII de *T. cruzi*. Apesar de ser muito eficaz para diferenciar os três haplótipos mitocondriais A (TcI), C (TcII) e B (TcIII-VI), o uso de enzimas de restrição (*Alu*I) aumenta muito o trabalho e o tempo de execução do ensaio, diminuindo sua sensibilidade e capacidade de automação.

Nossa proposta foi de desenvolver uma abordagem que removesse a etapa de digestão de maneira a permitir a sua futura utilização em ensaios multiplex com outros marcadores. Para isso, padrões de polimorfismos condizentes com os três haplótipos mitocondriais foram localizados ao longo dos genes mitocondriais do parasito. Assim, através de um SNP encontrado no gene ND1, foi possível desenvolver uma técnica de detecção alelo específica com iniciadores de tamanhos diferentes e temperaturas de anelamento próximas para serem utilizados no ensaio multiplex.

Foi possível demonstrar com a técnica proposta neste trabalho que a estratégia escolhida para substituir a etapa de RFLP-PCR do gene COII, permite diferenciar bem os haplótipos mitocondriais A, B e C, sem o uso de enzimas de restrição, tornando o ensaio mais rápido, mais simples e viabilizando a sua posterior automatização.

Todavia, embora o resultado obtido tenha se mostrado encorajador, a presença de algumas bandas inespecíficas, mostra que ainda é possível melhorar alguns parâmetros da estratégia proposta para aumentar a especificidade da tipagem. Além disso, pretendemos avaliar a sensibilidade desta metodologia, para aplicações diretamente em tecidos humanos infectados, o que permitiria, por exemplo, a caracterização precoce das linhagens em de casos de reinfecção em pacientes transplantados, como já foi visto em outros trabalhos desenvolvidos pelo nosso grupo (SEGATTO, 2013).

3) Análise do genoma de TcIII 231

3.1) Nova estratégia de montagem de genomas altamente repetitivos

Como se sabe, a montagem de regiões repetitivas em genomas complexos representa um grande desafio na bioinformática, especialmente em projetos que utilizam *reads* muito curtas como as obtidas em plataformas de sequenciamento de última geração. Uma das abordagens mais usadas para minimizar este problema é a utilização de *reads* obtidas de diferentes plataformas de sequenciamento, que geram *reads* também variáveis curtas e longas (FRANZÉN et al., 2012). Um exemplo clássico foi a montagem do genoma do clone de Sylvio (TcI), que utilizou *reads* geradas pela plataforma Illumina e Roche 454 (FRANZÉN et al., 2011). Todavia, há que se destacar ainda a escassez de algoritmos e estratégias de montagem de genomas repetitivos, mesmo para os casos contendo *reads* originárias o que demanda o desenvolvimento de novas alternativas e estratégias de montagem.

O pipeline de montagem combinada proposto nesse trabalho (combinação da estratégia de montagem de novo e baseada em um genoma de referência) melhora a qualidade da montagem de um genoma com um conteúdo altamente repetitivo usando apenas *reads* curtas geradas pela plataforma NGS Illumina de uma forma acessível e rápida. A única exigência do uso dessa estratégia de montagem combinada é a

disponibilidade de um genoma de referência relativamente próximo evolutivamente do organismo a ser montado.

Esta metodologia pode ser um passo importante para futuros projetos de montagem de genomas altamente repetitivos, principalmente por eliminar a dependência do uso de plataformas de última geração que produzem de *reads* longas, mas que são em geral de menor eficiência, como o descontinuado Roche 454 ("Roche Shutting Down 454 Sequencing Business" 2013). Existe ainda uma infinidade de novos genomas sendo sequenciados, a maioria em plataformas que geram *reads* curtas, o que torna a estratégia de montagem proposta neste trabalho uma boa opção para a montagem de genomas, especialmente se os mesmos tiverem um alto conteúdo repetitivo.

O tamanho do genoma das diferentes linhagens de *T. cruzi* varia bastante entre si (FRANZÉN et al., 2011, 2012). Para mostrar como o pipeline proposto pode ser estendido para genomas de outras linhagens de *T. cruzi*, ou mesmo de outros organismos, a mesma abordagem foi utilizada na montagem de genoma de outro tripanosomatideo, no caso, o genoma escolhido foi o de *Leishmania peruviana* para a montagem do qual se utilizou o genoma de *Leishmania braziliensis* como referência, que são espécies diferentes, portanto mais divergentes entre si do que entre linhagens de uma mesma espécie como foi o caso de *T. cruzi*. Os resultados obtidos foram compatíveis com os achados neste trabalho (Valdivia *et al*., em preparação), demonstrando que, mesmo usando um genoma de uma espécie próxima, mas não idêntica como genoma de referência, a abordagem de montagem proposta por nós é uma metodologia alternativa atrativa. Essa abordagem é objeto de um manuscrito está em fase final de elaboração para a publicação (Anexo 2).

3.2) Uma nova hipótese para a origem das linhagens de *T.cruzi*

Um dos principais problemas em trabalhos que visam à reconstrução da história evolutiva de organismos é a seleção de um conjunto de dados ideal. Para *T. cruzi* o principal problema encontrado é que, como demonstrado, poucos representantes de suas linhagens filogenéticas estão disponíveis, dificultando a escolha de um conjunto de dados de genes que tenha uma histórias evolutivas condizentes com a do parasito. Com isso torna-se interessante reconstruir histórias evolutivas de conjuntos genes

mitocondriais e nucleares, que apresentem uma relativa conservação entre todas as linhagens analisadas.

Para reconstruir a história evolutiva das principais linhagens do táxon *T. cruzi*, foram escolhidas as regiões codificadoras do genoma mitocondrial e 43 *loci* nucleares independentes e relativamente conservados entre as diferentes linhagens do parasito disponíveis. Isso possibilitou comparar as histórias evolutivas mitocondriais e nucleares do parasito. Para as análises dos *loci* nucleares, observamos que temos mais genes ortólogos similares entre TcI e TcIII do que entre TcII e TcIII. Para a análise do genoma mitocondrial, a reconstrução filogenética mostra que mesmo sendo diferente de TcI, o genoma tem maior similaridade com o genoma mitocondrial de TcI do que com o de TcII. Com estes resultados mostram que o TcIII não é um híbrido de TcI e TcII como alguns grupos anteriormente haviam proposto (WESTENBERGER et al., 2005), mostrando que sua diversificação nas atuais linhagens existentes é recente. Podemos também afirmar que TcIII é mais estreitamente relacionado com a linhagem TcI do que com TcII, refutando a classificação original de *T. cruzi* em dois grandes grupos, *T. cruzi* I (TcI) e *T. cruzi* II (TcII) (BURGOS et al., 2013; FLORES-LÓPEZ e MACHADO, 2011; FREITAS, DE et al., 2006).

Com base nos resultados obtidos nesse trabalho associados a outros achados anteriores obtidos pelo nosso e outros grupos (FLORES-LÓPEZ e MACHADO, 2011; FREITAS, DE et al., 2006), é razoável propormos que a linhagem TcII divergiu primeiro em relação as demais linhagens inclusive em relação a TcI e TcIII (Figura 29). Hipótese semelhante foi também proposta por Flores-López e Machado (2011), que utilizaram sequências nucleotídicas de 32 *loci* não ligados para reconstruir a história evolutiva da linhagem TcIV, não utilizada no nesse presente trabalho.

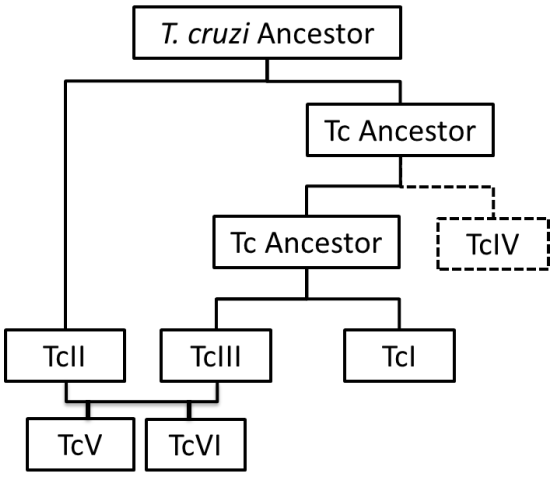

Figura 29. Desenho esquemático da hipótese de como teria ocorrido a divergência das linhagens de *T. cruzi*, utilizando os dados obtidos no presente trabalho junto com os dados de TcIV obtidos do testudo de Flores-Lopez et al. (2011) e TcV obtidos de Freitas et al (2006), representados na figura por linhas pontilhadas.

Um perfil evolutivo muito similar foi obtido entre as árvores reconstruídas tendo como bases apenas com os genes nucleares e aquelas com o genoma mitocondrial, onde principalmente a linhagem TcVI se mostrou hibrida de TcII e TcIII. Essa observação associada com a maior similaridade do genoma mitocondrial do clone da cepa 231 (TcIII) ao genoma mitocondrial do híbrido TcVI, mostra que durante o evento de hibridação que deu origem à linhagem TcVI, a mitocôndria de TcIII foi transferida unilateralmente, suportando a hipótese de que TcIII foi a "cepa mãe" do híbrido TcVI (do Valle et al., em preparação). A congruência entre as topologias das árvores nucleares e mitocondriais e o fato de que as cepas híbridas estão em regiões geográficas onde suas ancestrais se encontram compartilhando o mesmo ciclo de transmissão (Figura 30), reforça que o efeito Wahlund (presença de barreiras) deve ser um fator a ser considerado na proposição de hipóteses sobre os eventos sexuais naturais nestes parasitos. E também apoia a validade das comparações feitas em anteriores de nosso grupo comparando marcadores mitocondriais e nucleares (BAPTISTA et al., 2014; CARRANZA et al., 2009).

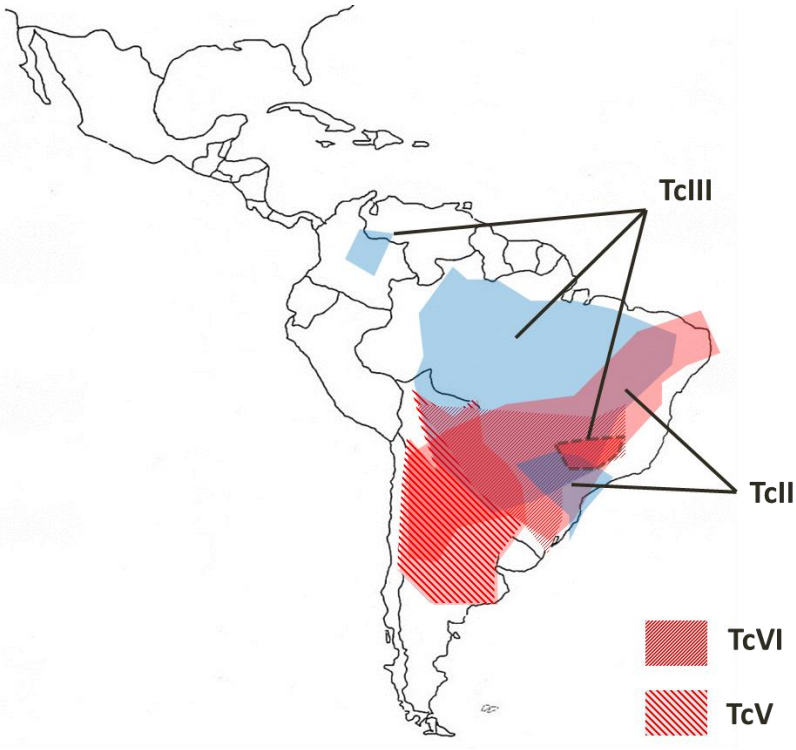

Figura 30. Mapa da distribuição geográfica das linhagens de *T. cruzi*. O mapa mostra que, principalmente na região sudeste do Brasil, as cepas de linhagens hibridas TcV e TcVI encontrados na natureza estão concentrados basicamente em regiões que as linhagens TcII e TcIII são sobrepostas em um mesmo ciclo de transmissão; silvestre (azul) e ciclos de transmissão (vermelho) domésticos.

Obviamente, que estudos que relatam análises de uma ou de algumas cepas não abrangem todo o quadro possível de variabilidade genética do táxon. Assim, futuros estudos filogenéticos multilocus com outras linhagens e uma quantidade maior de amostras devem também conduzir uma maior profundidade de amostragem para melhor refletir as diferenças entre estas linhagens e melhor resolver a história evolutiva do *T. cruzi*.

4) Nova abordagem para o agrupamento de sequências e identificação de sítios candidatos específicos para caracterização das linhagens.

Atualmente um dos principais desafios no desenvolvimento de metodologias de caracterização molecular a serem utilizados em estudos subpopulações em qualquer conjunto de organismos, é a detecção de sítios polimórficos significantes e específicos para essa estruturação. Como já foi mencionado, o *T. cruzi*, revelou uma substancial variabilidade genética entre suas cepas, o que culminou, em 2009, com a subdivisão da

população desse parasito em seis linhagens ou DTUs principais (TcI-VI). Um dos desafios atuais é a detecção de marcadores que permitam de maneira simples a caracterização das seis linhagens. Com o aumento da geração de dados genômicos do parasito, a detecção de novos sítios específicos responsáveis pela diferenciação dessas linhagens de *T. cruzi* se torna importante ferramenta para encontrar novos marcadores, visando, por exemplo, a investigar melhor eventuais associações com aspectos epidemiológicos e clínicos.

Para isto propusemos desenvolver uma estratégia de identificação de sítios candidatos usando a abordagem de decomposição de Valores Singulares ou SVD. Com essa metodologia obtivemos três clusters no gráfico SVD a partir de sequências de amastina de quatro linhagens (TcI, TcII, TcIII e TcVI). Embora o número clusters obtidos pelo gráfico SVD (três) não tenha coincidido com o número de linhagens filogenéticas utilizadas para a geração de sequências de amastina (quatro), houve uma clusterização das sequências da quarta linhagem de acordo com as linhagens supostamente parentais desta população, visto o caráter híbrido da linhagem TcVI como já demonstrado em trabalhos prévios (BRISSE et al., 2003; EL-SAYED, MYLER, BARTHOLOMEU, et al., 2005; FREITAS, DE et al., 2006) e nesse trabalho.

Isso pode ser constatado, visto que as sequências de amastina derivadas do clone CL Brener (TcVI) se dividiram em dois grupos SVD diferentes: grupo SVD3 (TcIII) e SVD2 (TcII). Estes achados estão de acordo com o esperado visto que o clone CL Brener é um clone híbrido derivado do evento de hibridização de duas linhagens TcII e TcIII (BURGOS et al., 2013; FLORES-LÓPEZ e MACHADO, 2011; FREITAS, DE et al., 2006). Os alelos de CL Brener são comumente chamados de Esmeraldo-*like*, que em nossa análise foi agrupado no grupo SVD2, e alelo Non-Esmeraldo-*like*, que em nossa análise foi agrupada no grupo SVD3, (EL-SAYED et al., 2005). Visto que a curva L observada no gráfico SVD se encontra entre três e quatro valores singulares, foi realizado também um teste com o valor de k=4 no k-Means para verificar se haveria uma separação significante de um quarto grupo entre as sequências de amastina usadas, mas como a quarta linhagem utilizada tem um caráter híbrido não houve uma divisão adequada de um novo grupo equivalente a uma nova linhagem filogenética. Em análises preliminares, sem a inclusão das sequências de amastina de TcIII, foi observado que as sequências de amastina pertencentes à população TcI apresentaram menor variação em suas sequências intracepas (genes parálogos) e maiores diferenças entre as sequências

pertencentes a diferentes cepas (genes ortólogos), tendo sido divididas em dois grupos SVD diferentes.

A comparação entre o método de clusterização baseado em SVD proposto aqui e o da reconstrução de uma árvore de distância filogenética, revelou topologias de agrupamentos muito similares, sugerindo que a metodologia clusterização proposta é robusta e eficaz. Um dos pontos mais positivos da metodologia de clusterização por SVD é que ela possui uma baixa complexidade computacional podendo ser, em princípio, usada com grandes conjuntos de dados, gerando resultados com rapidez e confiabilidade. Essa abordagem pode ser uma boa ferramenta de previsão de sítios candidatos para utilização de enzimas de restrição, assim como para desenhos de iniciadores, visto que blocos nucleotídeos significativos e específicos para cada cluster pode ser identificado. Assim o algoritmo de clusterização por SVD além de ser uma ferramenta importante para ensaios de taxonomia, também identifica sítios polimórficos disponíveis para estudos filogenéticos e para diagnósticos moleculares linhagem específicos. Esta metodologia é objeto de um manuscrito em fase final de elaboração (Anexo 2).

CONSIDERAÇÕES FINAIS E PERSPECTIVAS

V. CONSIDERAÇÕES FINAIS E PERSPECTIVAS

Com os resultados obtidos no presente trabalho foi possível atingir os objetivos propostos, tendo sido demonstrado que existe uma frequência maior de recombinação entre populações de *T. cruzi* circulando em áreas geograficamente próximas. Esses dados em conjunto com o sequenciamento do genoma do clone da cepa 231, uma linhagem ainda não disponível nos bancos de dados públicos, e a realização da análise de genômica comparativa com esses novos dados genômicos, propiciaram a resolução da origem ancestral das linhagens filogenéticas do parasito, principalmente em relação às linhagens híbridas. Nossa perspectiva é adicionar ao banco de dados outros genomas de diferentes linhagens, apresentando diferentes aspectos epidemiológicos, para que mais informações relevantes sobre aspectos clínicos e evolutivos desse parasito sejam recuperadas.

Igualmente, o sequenciamento e a montagem do genoma de um clone de 231 uma cepa da linhagem TcIII, permitirão diversas análises complementares àquelas já realizadas no presente trabalho. Um exemplo é a análise de variação no número de cópias (CNV) de regiões gênicas entre as linhagens com genomas disponíveis, bem como a análise de SNP's. Ademais, devido à grande quantidade de proteínas hipotéticas identificadas no genoma de referência e, por conseguinte também no genoma de 231, faz-se necessária a adoção de uma segunda estratégia de anotação do genoma de *T. cruzi*.

Como detalhado no presente trabalho, a abordagem inicial escolhida para a anotação do genoma de 231 foi a transferência de anotação por sintenia, utilizando o programa RATT e a anotação de um genoma de referência (CLBrener). Essa abordagem apesar de mais rápida, tem algumas limitações. Com isso, o genoma de 231 manteve um grande número de proteínas hipotéticas e não teve suas regiões não sintênicas anotadas, tornando necessárias novas análises visando à melhoria da anotação do genoma do clone da cepa 231 (TcIII), através de um método de predição gênica.

Será ainda necessário caracterizar computacionalmente as principais proteínas hipotéticas conservadas entre todas as linhagens filogenéticas de *T. cruzi*, como aquelas utilizadas para as análises filogenéticas neste trabalho, buscando revelar as funções das mesmas. A principal contribuição dessa caracterização computacional será encontrar características dinâmicas e estruturais de cada uma das proteínas hipotéticas de forma a auxiliar na determinação de suas possíveis funções. Conhecer mais sobre essas

proteínas hipotéticas conservadas no táxon pode ser o caminho para se determinar novos alvos de drogas, diagnóstico e até mesmo fornecer mais dados sobre a evolução da espécie.

Em relação à metodologia Bi-PASA, o próximo passo será aplicá-la em ensaios multiplex de PCR em tecidos de pacientes chagásicos crônicos. Essa aplicação é importante, pois dados obtidos com outros marcadores indicam que a utilização de ensaios de PCR de genes multicópia diretamente em tecidos de pacientes, como em biópsias cardíacas, é uma boa estratégia para o diagnóstico precoce da reativação da infecção pelo *T. cruzi*, com potencial para auxiliar os médicos nas decisões de tratamento antes do início da reativação clínica da doença de Chagas.

Por fim, é nossa intenção utilizar a metodologia de agrupamento desenvolvida neste trabalho, e que se mostrou capaz de reconhecer padrões linhagem-específicos, com dados de sequências das demais linhagens, a fim de se encontrar as melhores regiões gênicas candidatas para a caracterização de cada uma das linhagens. Além disto separadamente pretendemos utilizá-la como metodologia de agrupamento de famílias hipervariáveis de proteínas de superfície de *T. cruzi*, visto que como ela não depende necessariamente de um alinhamento, diminui a chance de haver algum tipo de viés por alinhamento errôneo e facilita a análise de sequências quiméricas e ricas em repetições.

REFERÊNCIAS

REFERÊNCIAS

ALTMAN, E. Default recovery rates and lgd in credit risk modelling and practice: An updated review of the literature and empirical evidence. *Advances in credit risk modeling and corporate bankruptcy prediction*. [S.l: s.n.], 2008. p. 175– 206.

AÑEZ, N.;; CRISANTE, G. e ROJAS, A. Update on Chagas disease in Venezuela--a review. *Memórias do Instituto Oswaldo Cruz*, v. 99, n. 8, p. 781–7, doi:/S0074-02762004000800001, 2004.

ANNONIMOUS. Recommendations from a satellite meeting. *Memórias do Instituto Oswaldo Cruz*, v. 94 Suppl 1, p. 429–32, 1999.

ASLUND, L. et al. A gene family encoding heterogeneous histone H1 proteins in Trypanosoma cruzi. *Molecular and biochemical parasitology*, v. 65, n. 2, p. 317–30, 1994.

ATAYDE, V. D. et al. Molecular basis of non-virulence of Trypanosoma cruzi clone CL-14. *International journal for parasitology*, v. 34, n. 7, p. 851–60, doi:10.1016/j.ijpara.2004.03.003, 2004.

AYALA, F. J. Trypanosoma and Leishmania have clonal population structures of epidemiological significance. *Biological research*, v. 26, n. 1-2, p. 47–63, 1993.

BAPTISTA, C. S. et al. Differential transcription profiles in Trypanosoma cruzi associated with clinical forms of Chagas disease: Maxicircle NADH dehydrogenase subunit 7 gene truncation in asymptomatic patient isolates. *Molecular and Biochemical Parasitology*, v. 150, n. 2, p. 236–248, 2006.

BAPTISTA, R. de P. et al. Evidence of substantial recombination among Trypanosoma cruzi II strains from Minas Gerais. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, v. 22, p. 183–91, doi:10.1016/j.meegid.2013.11.021, 2014.

BARNABÉ, C. e BRENIÈRE, S. F. Scarce events of mitochondrial introgression in Trypanosoma cruzi: new case with a Bolivian strain. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, v. 12, n. 8, p. 1879–83, doi:10.1016/j.meegid.2012.08.018, 2012.

BARRETO, M. *Trypanosoma cruzi e Doença de Chagas*. 2. ed. RJ: Koogan, Guanabara, 1979. p. 89–151

BASTIEN, P.;; BLAINEAU, C. e PAGES, M. Leishmania: sex, lies and karyotype. *Parasitology today (Personal ed.)*, v. 8, n. 5, p. 174–7, 1992.

BERRY, M.;; DUMAIS, S. e O'BRIEN, G. Using linear algebra for intelligent information retrieval. *SIAM Review*, v. 4, n. 37, p. 575–595, 1995.

BOETZER, M. et al. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics (Oxford, England)*, v. 27, n. 4, p. 578–9, doi:10.1093/bioinformatics/btq683, 2011.

BOGLIOLO, A. R.;; LAURIA-PIRES, L. e GIBSON, W. C. Polymorphisms in Trypanosoma cruzi: evidence of genetic recombination. *Acta tropica*, v. 61, n. 1, p. 31–40, 1996.

BOLGER, A. M.;; LOHSE, M. e USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, doi:10.1093/bioinformatics/btu170, 2014.

BOUCKAERT, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, v. 10, n. 4, p. e1003537, doi:10.1371/journal.pcbi.1003537, 2014.

BRAGA, J. P. Numerical comparison between Tikhonov regularization and singular value decomposition methods using the L curve criterion. *Journal of Mathematical Chemistry*, v. 29, n. 2, 2001.

BRENER, Z. Trypanosoma cruzi: taxonomy, morphology and life cicle. . [S.l: s.n.], 1992.

BRISSE, S. et al. Evidence for genetic exchange and hybridization in Trypanosoma cruzi based on nucleotide sequences and molecular karyotype. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, v. 2, n. 3, p. 173–83, 2003.

BRISSE, S.;; BARNABÉ, C. e TIBAYRENC, M. Identification of six Trypanosoma cruzi phylogenetic lineages by random amplified polymorphic DNA and multilocus enzyme electrophoresis. *International Journal for Parasitology*, v. 30, n. 1, p. 35–44, 2000.

BRISSE, S.;; VERHOEF, J. e TIBAYRENC, M. Characterisation of large and small subunit rRNA and mini-exon genes further supports the distinction of six Trypanosoma cruzi lineages. *International Journal for Parasitology*, v. 31, n. 11, p. 1218–1226, 2001.

BURGOS, J. M. et al. Direct molecular profiling of minicircle signatures and lineages of Trypanosoma cruzi bloodstream populations causing congenital Chagas disease. *International Journal for Parasitology*, v. 37, n. 12, p. 1319–1327, 2007.

BURGOS, J. M. et al. Differential distribution of genes encoding the virulence factor trans-sialidase along Trypanosoma cruzi Discrete typing units. *PloS one*, v. 8, n. 3, p. e58967, doi:10.1371/journal.pone.0058967, 2013.

BUSCAGLIA, C. A. e NOIA, J. M. DI. Trypanosoma cruzi clonal diversity and the epidemiology of Chagas' disease. *Microbes and infection / Institut Pasteur*, v. 5, n. 5, p. 419–27, 2003.

CAPELLA-GUTIERREZ, S.;; KAUFF, F. e GABALDÓN, T. A phylogenomics approach for selecting robust sets of phylogenetic markers. *Nucleic acids research*, v. 42, n. 7, p. e54, doi:10.1093/nar/gku071, 2014.

CAPELLA-GUTIÉRREZ, S.;; SILLA-MARTÍNEZ, J. M. e GABALDÓN, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, v. 25, n. 15, p. 1972–3, doi:10.1093/bioinformatics/btp348, 2009.

CARRANZA, J. C. et al. Trypanosoma cruzi maxicircle heterogeneity in Chagas disease patients from Brazil. *International Journal for Parasitology*, v. 39, n. 9, p. 963–973, 2009.

CERQUEIRA, G. C. et al. Sequence diversity and evolution of multigene families in Trypanosoma cruzi. *Molecular and biochemical parasitology*, v. 157, n. 1, p. 65–72, doi:10.1016/j.molbiopara.2007.10.002, 2008.

CHIARI, E. et al. Hemocultures for the parasitological diagnosis of human Chagas disease in the chronic phase. In: CONGRESSO INTERNACINAL DA DOENÇA DE CHAGAS. *Anais*... [S.l: s.n.], 1979.

COURA, J. R. et al. Emerging Chagas disease in Amazonian Brazil. *Trends in parasitology*, v. 18, n. 4, p. 171–6, 2002.

COUTO, B.;; SANTORO, M. e SANTOS, M. Singular value decomposition (SVD) and BLAST: quite different methods achieving similar results,. 2011.

D'ÁVILA, D. A. et al. Probing Population Dynamics of Trypanosoma cruzi during Progression of the Chronic Phase in Chagasic Patients. *Journal of Clinical Microbiology*, v. 47, n. 6, p. 1718–1725, 2009.

DANECEK, P. et al. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, v. 27, n. 15, p. 2156–8, doi:10.1093/bioinformatics/btr330, 2011.

DARRIBA, D. et al. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, v. 27, n. 8, p. 1164–5, doi:10.1093/bioinformatics/btr088, 2011.

DARRIBA, D. et al. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, v. 9, n. 8, p. 772, doi:10.1038/nmeth.2109, 2012.

DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, v. 6, n. 41, p. 391–407, 1990.

DEVERA, R.;; FERNANDES, O. e COURA, J. R. Should Trypanosoma cruzi be called "cruzi" complex? a review of the parasite diversity and the potential of selecting population after in vitro culturing and mice infection. *Memórias do Instituto Oswaldo Cruz*, v. 98, n. 1, p. 1–12, 2003.

DIAS, J. C. [Chagas disease, environment, participation, and the state]. *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública*, v. 17 Suppl, p. 165–9, 2001.

DIAS, J. C. P. Epidemiology of Chagas disease (american Trypanosomiasis): It's impact on transfusion and clinical medicine. . [S.l: s.n.], 1992.

DVORAK, J. A. Analysis of the DNA of parasitic protozoa by flow cytometry. *Methods in molecular biology (Clifton, N.J.)*, v. 21, p. 191–204, doi:10.1385/0-89603-239-6:191, 1993.

EDGAR, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, v. 5, p. 113, doi:10.1186/1471-2105-5-113, 2004.

ELDÉN, L. Numerical linear algebra in data mining. *Acta Numerica*, p. 327{384, 2006.

ELIAS, M. C. Q. B. et al. Comparative analysis of genomic sequences suggests that Trypanosoma cruzi CL Brener contains two sets of non-intercalated repeats of satellite DNA that correspond to T. cruzi I and T. cruzi II types. *Molecular and biochemical parasitology*, v. 140, n. 2, p. 221–7, doi:10.1016/j.molbiopara.2004.12.016, 2005.

EL-SAYED, N. M.;; MYLER, P. J.;; BLANDIN, G.;; et al. Comparative genomics of trypanosomatid parasitic protozoa. *Science (New York, N.Y.)*, v. 309, n. 5733, p. 404–9, doi:10.1126/science.1112181, 2005.

EL-SAYED, N. M.;; MYLER, P. J.;; BARTHOLOMEU, D. C.;; et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science (New York, N.Y.)*, v. 309, n. 5733, p. 409–15, doi:10.1126/science.1112631, 2005.

EVANNO, G.;; REGNAUT, S. e GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, v. 14, n. 8, p. 2611–2620, 2005.

EXCOFFIER, L.;; LAVAL, G. e SCHNEIDER, S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*, v. 1, n. 1, p. 47–50, 2007.

FERNANDES, O. et al. Brazilian isolates of Trypanosoma cruzi from humans and triatomines classified into two lineages using mini-exon and ribosomal RNA sequences. *The American journal of tropical medicine and hygiene*, v. 58, n. 6, p. 807–11, 1998.

FERNANDES, O. et al. A mini-exon multiplex polymerase chain reaction to distinguish the major groups of Trypanosoma cruzi and T. rangeli in the Brazilian Amazon. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 95, n. 1, p. 97–9, 2001.

FISCHER, S. et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, v. Chapter 6, p. Unit 6.12.1–19, doi:10.1002/0471250953.bi0612s35, 2011.

FLORES-LÓPEZ, C. A. e MACHADO, C. A. Analyses of 32 loci clarify phylogenetic relationships among Trypanosoma cruzi lineages and support a single hybridization

prior to human contact. *PLoS neglected tropical diseases*, v. 5, n. 8, p. e1272, doi:10.1371/journal.pntd.0001272, 2011.

FORSTER, P. et al. A short tandem repeat-based phylogeny for the human Y chromosome. *The American Journal of Human Genetics*, v. 67, n. 1, p. 182–196, 2000.

FRANZÉN, O. et al. Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener. *PLoS neglected tropical diseases*, v. 5, n. 3, p. e984, doi:10.1371/journal.pntd.0000984, 2011.

FRANZÉN, O. et al. Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei. *BMC genomics*, v. 13, p. 531, doi:10.1186/1471-2164-13-531, 2012.

FREITAS, J. M. et al. Real time PCR strategy for the identification of major lineages of Trypanosoma cruzi directly in chronically infected human tissues. *International journal for parasitology*, v. 35, n. 4, p. 411–7, doi:10.1016/j.ijpara.2004.10.023, 2005.

FREITAS, J. M. DE et al. Ancestral Genomes, Sex, and the Population Structure of Trypanosoma cruzi. *PLoS Pathogens*, v. 2, n. 3, p. 10, 2006.

GAUNT, M. W. et al. Mechanism of genetic exchange in American trypanosomes. *Nature*, v. 421, n. 6926, p. 936–939, doi:10.1038/nature01393.1.2.3.4.5.6.7.8.9.10.11.12.13.14.15.16.17.18.19.20.21.22.23.24.25.26.Walther, 2003.

GOLUB, G. e KAHAN, W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, v. 2, n. 2, p. 205, doi:10.1137/0702016, 1965.

GOMES, M. L.;; ARAUJO, S. M. e CHIARI, E. Trypanosoma cruzi: growth of clones on solid medium using culture and blood forms. *Memórias do Instituto Oswaldo Cruz*, v. 86, n. 1, p. 131–2, 1991.

GUINDON, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, v. 59, n. 3, p. 307–21, doi:10.1093/sysbio/syq010, 2010.

HAJDUK, S. e SABATINI, R. RNA editing: post-transcriptional restructuring of genetic information. *Molecular Biology of Parasitic Protozoa*. [S.l: s.n.], 1996. p. 135–158.

HAMILTON, P. B.;; TEIXEIRA, M. M. G. e STEVENS, J. R. The evolution of Trypanosoma cruzi: the “bat seeding” hypothesis. *Trends in parasitology*, v. 28, n. 4, p. 136–41, doi:10.1016/j.pt.2012.01.006, 2012.

HAN, J. e KAMBER, M. Data Mining: Concepts and Techniques. 2000.

HENRIKSSON, J.;; ASLUND, L. e PETTERSSON, U. Karyotype variability in Trypanosoma cruzi. *Parasitology today (Personal ed.)*, v. 12, n. 3, p. 108–14, 1996.

HORDIJK, W. e GASCUEL, O. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics (Oxford, England)*, v. 21, n. 24, p. 4338–47, doi:10.1093/bioinformatics/bti713, 2005.

JUNQUEIRA, A. C. V;; DEGRAVE, W. e BRANDÃO, A. Minicircle organization and diversity in Trypanosoma cruzi populations. *Trends in parasitology*, v. 21, n. 6, p. 270–2, doi:10.1016/j.pt.2005.04.001, 2005.

KATOH, K. e STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, v. 30, n. 4, p. 772–80, doi:10.1093/molbev/mst010, 2013.

KURTZ, S. et al. Versatile and open software for comparing large genomes. *Genome biology*, v. 5, n. 2, p. R12, doi:10.1186/gb-2004-5-2-r12, 2004.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, v. 10, n. 3, p. R25, doi:10.1186/gb-2009-10-3-r25, 2009.

LARKIN, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, v. 23, n. 21, p. 2947–8, doi:10.1093/bioinformatics/btm404, 2007.

LAWSON, C. L. e HANSON, R. J. Solving Least Squares Problems. 1995.

LEWIS, M. D. et al. Genotyping of Trypanosoma cruzi: systematic selection of assays allowing rapid and accurate discrimination of all known lineages. *The American journal of tropical medicine and hygiene*, v. 81, n. 6, p. 1041–9, doi:10.4269/ajtmh.2009.09-0305, 2009.

LEWIS, M. D. et al. Recent, independent and anthropogenic origins of Trypanosoma cruzi hybrids. *PLoS neglected tropical diseases*, v. 5, n. 10, p. e1363, doi:10.1371/journal.pntd.0001363, 2011.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, v. 25, n. 16, p. 2078–9, doi:10.1093/bioinformatics/btp352, 2009.

LIU, Q. et al. Overlapping PCR for bidirectional PCR amplification of specific alleles: a rapid one-tube method for simultaneously differentiating homozygotes and heterozygotes. *Genome research*, v. 7, n. 4, p. 389–98, 1997.

LLEWELLYN, M. S. et al. Trypanosoma cruzi IIc: phylogenetic and phylogeographic insights from sequence and microsatellite analysis and potential impact on emergent Chagas disease. *PLoS neglected tropical diseases*, v. 3, n. 9, p. e510, doi:10.1371/journal.pntd.0000510, 2009.

LUQUETTI, A. O. et al. Trypanosoma cruzi: zymodemes associated with acute and chronic Chagas' disease in central Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 80, n. 3, p. 462–70, 1986.

MACEDO, A. M. et al. DNA fingerprinting of Trypanosoma cruzi: a new tool for characterization of strains and clones. *Molecular and Biochemical Parasitology*, v. 55, n. 1-2, p. 147–153, 1992.

MACEDO, A. M. et al. Usefulness of microsatellite typing in population genetic studies of Trypanosoma cruzi. *Memórias do Instituto Oswaldo Cruz*, v. 96, n. 3, p. 407–13, 2001.

MACEDO, A. M. et al. Trypanosoma cruzi: genetic structure of populations and relevance of genetic variability to the pathogenesis of chagas disease. *Memorias do Instituto Oswaldo Cruz*, v. 99, n. 1, p. 1–12, 2004.

MACEDO, A. M. e PENA, S. D. Genetic Variability of Trypanosoma cruzi:Implications for the Pathogenesis of Chagas Disease. *Parasitology today (Personal ed.)*, v. 14, n. 3, p. 119–24, 1998.

MACHADO, C. A. e AYALA, F. J. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of Trypanosoma cruzi. *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 13, p. 7396–401, doi:10.1073/pnas.121187198, 2001.

MACHADO, C. R. et al. DNA metabolism and genetic diversity in Trypanosomes. *Mutation research*, v. 612, n. 1, p. 40–57, doi:10.1016/j.mrrev.2005.05.001, 2006.

MCDANIEL, J. P. e DVORAK, J. A. Identification, isolation, and characterization of naturally-occurring Trypanosoma cruzi variants. *Molecular and biochemical parasitology*, v. 57, n. 2, p. 213–22, 1993.

MESSENGER, L. A. et al. Multiple mitochondrial introgression events and heteroplasmy in trypanosoma cruzi revealed by maxicircle MLST and next generation sequencing. *PLoS neglected tropical diseases*, v. 6, n. 4, p. e1584, doi:10.1371/journal.pntd.0001584, 2012.

MILES, M. A. et al. The identification by isoenzyme patterns of two distinct strain-groups of Trypanosoma cruzi, circulating independently in a rural area of Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 71, n. 3, p. 217–25, 1977.

MILES, M. A. et al. Isozymic heterogeneity of Trypanosoma cruzi in the first autochthonous patients with Chagas' disease in Amazonian Brazil. *Nature*, v. 272, n. 5656, p. 819–21, 1978.

MILES, M. A. et al. Further enzymic characters of Trypanosoma cruzi and their evaluation for strain identification. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 74, n. 2, p. 221–37, 1980.

MILES, M. A. et al. The molecular epidemiology and phylogeography of Trypanosoma cruzi and parallel research on Leishmania: looking back and to the future. *Parasitology*, v. 136, n. 12, p. 1509–28, doi:10.1017/S0031182009990977, 2009.

MORARIU, V. et al. Automatic online tuning for fast Gaussian summation. *Advances in Neural Information Processing Systems*, 2008.

MOREL, C. et al. Strains and clones of Trypanosoma cruzi can be characterized by pattern of restriction endonuclease products of kinetoplast DNA minicircles. *Proceedings of the National Academy of Sciences of the United States of America*, v. 77, n. 11, p. 6810–4, 1980.

MOREL, C. M.;; DEANE, M. P. e GONÇALVES, A. M. The complexity of Trypanosoma cruzi populations revealed by schizodeme analysis. *Parasitology today (Personal ed.)*, v. 2, n. 4, p. 97–101, 1986.

MYLER, P. J. Molecular variation in trypanosomes. *Acta tropica*, v. 53, n. 3-4, p. 205–25, 1993.

NOIA, J. M. DI et al. A Trypanosoma cruzi small surface molecule provides the first immunological evidence that Chagas' disease is due to a single parasite lineage. *The Journal of experimental medicine*, v. 195, n. 4, p. 401–13, 2002.

OCAÑA-MAYORGA, S. et al. Sex, subdivision, and domestic dispersal of Trypanosoma cruzi lineage I in southern Ecuador. *PLoS neglected tropical diseases*, v. 4, n. 12, p. e915, doi:10.1371/journal.pntd.0000915, 2010.

OLIVEIRA, R. P. et al. Probing the genetic population structure of Trypanosoma cruzi with polymorphic microsatellites. *Proceedings of the National Academy of Sciences of the United States of America*, v. 95, n. 7, p. 3776–3780, 1998.

OTTO, T. D. et al. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics (Oxford, England)*, v. 26, n. 14, p. 1704–7, doi:10.1093/bioinformatics/btq269, 2010.

OTTO, T. D. et al. RATT: Rapid Annotation Transfer Tool. *Nucleic acids research*, v. 39, n. 9, p. e57, doi:10.1093/nar/gkq1268, 2011.

PINTO, C. M. et al. TcBat a bat-exclusive lineage of Trypanosoma cruzi in the Panama Canal Zone, with comments on its classification and the use of the 18S rRNA gene for lineage identification. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, v. 12, n. 6, p. 1328–32, doi:10.1016/j.meegid.2012.04.013, 2012.

PRATA, A. Clinical and epidemiological aspects of Chagas disease. *The Lancet infectious diseases*, v. 1, n. 2, p. 92–100, doi:10.1016/S1473-3099(01)00065-2, 2001.

PRICE, A. L.;; JONES, N. C. e PEVZNER, P. A. De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)*, v. 21 Suppl 1, p. i351–8, doi:10.1093/bioinformatics/bti1018, 2005.

PRITCHARD, J. K.;; STEPHENS, M. e DONNELLY, P. Inference of population structure from multilocus genotype data. *Genetics*, v. 155, p. 945–959, 2000.

RAMÍREZ, J. D. et al. Natural and emergent Trypanosoma cruzi I genotypes revealed by mitochondrial (Cytb) and nuclear (SSU rDNA) genetic markers. *Experimental parasitology*, v. 132, n. 4, p. 487–94, doi:10.1016/j.exppara.2012.09.017, 2012.

ROCHETTE, A. et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp. *Molecular and biochemical parasitology*, v. 140, n. 2, p. 205–20, doi:10.1016/j.molbiopara.2005.01.006, 2005.

RODRIGUES, J. C. F.;; GODINHO, J. L. P. e SOUZA, W. DE. Biology of human pathogenic trypanosomatids: epidemiology, lifecycle and ultrastructure. *Sub-cellular biochemistry*, v. 74, p. 1–42, doi:10.1007/978-94-007-7305-9_1, 2014.

ROUGERON, V. et al. Everything You Always Wanted to Know about Sex (but Were Afraid to Ask)" in Leishmania after Two Decades of Laboratory and Field Analyses. *PLoS Pathogens*, v. 6, n. 8, p. 5, 2010.

ROUSSET, F. e RAYMOND, M. Testing Heterozygote Excess and Deficiency. *Genetics*, v. 140, n. 4, p. 1413–1419, 1995.

RUVALCABA-TREJO, L. I. e STURM, N. R. The Trypanosoma cruzi Sylvio X10 strain maxicircle sequence: the third musketeer. *BMC genomics*, v. 12, p. 58, doi:10.1186/1471-2164-12-58, 2011.

RYCKMAN, R. The vertebrate hosts of the Triatominae of North and Central America and the West Indies (Hemiptera: Reduviidae: Triatominae). *Bull. Soc. Vect. Ecol.*, v. 11, p. 221–241, 1986.

SCHMUNIS, G. A. e YADON, Z. E. Chagas disease: a Latin American health problem becoming a world health problem. *Acta tropica*, v. 115, n. 1-2, p. 14–21, doi:10.1016/j.actatropica.2009.11.003, 2010.

SCHOFIELD, C. J. Triatominae, Biología y Control. *eurocommunica publications*, 1994.

SCHOFIELD, C. J.;; JANNIN, J. e SALVATELLA, R. The future of Chagas disease control. *Trends in Parasitology*, v. 22, n. 12, p. 583–588, 2006.

SEGATTO, M. *FERRAMENTAS MOLECULARES PARA A DETECÇÃO E GENOTIPAGEM DO Trypanosoma cruzi: APLICAÇÃO EM ESTUDOS POPULACIONAIS DO PARASITO E DA EPIDEMIOLOGIA DA DOENÇA DE CHAGAS CARDÍACA*. Universidade Federal de Minas Gerais - [S.l.]. 2013.

SILVEIRA, J. Biologia Molecular do Trypanosoma cruzi. *Trypanosoma cruzi e Doença de Chagas*. [S.l: s.n.], 2000. p. 127–152.

SIMPSON, L. The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution. *Annual review of microbiology*, v. 41, p. 363–82, doi:10.1146/annurev.mi.41.100187.002051, 1987.

SMIT, A.;; HUBLEY, R. e GREEN, P. *RepeatMasker*. Disponível em: <http://www.repeatmasker.org>.

SOUTO, R. P. et al. DNA markers define two major phylogenetic lineages of Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, v. 83, n. 2, p. 141–52, 1996.

STEPHENS, M.;; SMITH, N. J. e DONNELLY, P. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, v. 68, n. 4, p. 978–989, 2001.

STUART, K. *RNA editing: an overview, status report, and personal perspective.* . [S.l: s.n.], 1995.

STURM, N. R. et al. Evidence for multiple hybrid groups in Trypanosoma cruzi. *International journal for parasitology*, v. 33, n. 3, p. 269–79, 2003.

SUBILEAU, M. et al. Trypanosoma cruzi: new insights on ecophylogeny and hybridization by multigene sequencing of three nuclear and one maxicircle genes. *Experimental parasitology*, v. 122, n. 4, p. 328–37, doi:10.1016/j.exppara.2009.04.008, 2009.

TAMURA, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, v. 28, n. 10, p. 2731–9, doi:10.1093/molbev/msr121, 2011.

TANOWITZ, H. B. et al. Chagas' disease. *Clinical microbiology reviews*, v. 5, n. 4, p. 400–19, 1992.

TEIXEIRA, M. M. G. et al. Short communication: Trypanosoma cruzi lineage I in endomyocardial biopsy from a north-eastern Brazilian patient at end-stage chronic Chagasic cardiomyopathy. *Tropical medicine & international health : TM & IH*, v. 11, n. 3, p. 294–8, doi:10.1111/j.1365-3156.2006.01575.x, 2006.

TIBAYRENC, M. et al. Natural populations of Trypanosoma cruzi, the agent of Chagas disease, have a complex multiclonal structure. *Proceedings of the National Academy of Sciences of the United States of America*, v. 83, n. 1, p. 115–9, 1986.

TIBAYRENC, M. et al. Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proceedings of the National Academy of Sciences of the United States of America*, v. 88, n. 12, p. 5129–33, 1991.

TIBAYRENC, M. et al. Genetic characterization of six parasitic protozoa: parity between random-primer DNA typing and multilocus enzyme electrophoresis. *Proceedings of the National Academy of Sciences of the United States of America*, v. 90, n. 4, p. 1335–9, 1993.

TIBAYRENC, M. Population genetics of parasitic protozoa and other microorganisms. *Advances in parasitology*, v. 36, p. 47–115, 1995.

TIBAYRENC, M. e AYALA, F. J. Trypanosoma cruzi populations: more clonal than sexual. *Parasitology today (Personal ed.)*, v. 3, n. 6, p. 189–90, 1987.

TIBAYRENC, M. e AYALA, F. J. The clonal theory of parasitic protozoa: 12 years on. *Trends in parasitology*, v. 18, n. 9, p. 405–10, 2002.

TIBAYRENC, M. e AYALA, F. J. How clonal are Trypanosoma and Leishmania? *Trends in parasitology*, v. 29, n. 6, p. 264–9, doi:10.1016/j.pt.2013.03.007, 2013.

TIBAYRENC, M.;; KJELLBERG, F. e AYALA, F. J. A clonal theory of parasitic protozoa: the population structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and Trypanosoma and their medical and taxonomical consequences. *Proceedings of the National Academy of Sciences of the United States of America*, v. 87, n. 7, p. 2414–8, 1990.

TORO, G. C. e GALANTI, N. H 1 histone and histone variants in Trypanosoma cruzi. *Experimental cell research*, v. 174, n. 1, p. 16–24, 1988.

TSAI, I. J.;; OTTO, T. D. e BERRIMAN, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome biology*, v. 11, n. 4, p. R41, doi:10.1186/gb-2010-11-4-r41, 2010.

UMEZAWA, E. S. et al. Chagas' disease. *Lancet*, v. 357, n. 9258, p. 797–9, doi:10.1016/S0140-6736(00)04174-X, 2001.

VALADARES, H. M. S. et al. Genetic profiling of Trypanosoma cruzi directly in infected tissues using nested PCR of polymorphic microsatellites. *International Journal for Parasitology*, v. 38, n. 7, p. 839–850, 2008.

VINHAES, M. C. Os Programas nacionais de controle na fase avançada de controle e os novos desafios estratégicos, políticos e epidemiológicos. *Revista de Patologia tropical*, v. 31, p. 124–129, 2002.

VINHAES, M. C. e DIAS, J. C. [Chagas disease in Brazil]. *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública*, v. 16 Suppl 2, p. 7–12, 2000.

WAGNER, W. e SO, M. Genomic variation of Trypanosoma cruzi: involvement of multicopy genes. *Infection and immunity*, v. 58, n. 10, p. 3217–24, 1990.

WAHLUND, S. Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, v. 11, p. 65–106, 1928.

WALLACE, I. M. et al. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic acids research*, v. 34, n. 6, p. 1692–9, doi:10.1093/nar/gkl091, 2006.

WEATHERLY, D. B.;; BOEHLKE, C. e TARLETON, R. L. Chromosome level assembly of the hybrid Trypanosoma cruzi genome. *BMC Genomics*, v. 10, n. 1, p. 255, 2009.

WENDEL, S. et al. Chagas disease (American Trypanosomiasis): It's impact on transfusion and clinical medicine. . [S.l: s.n.], 1992.

WESTENBERGER, S. J. et al. Two hybridization events define the population structure of Trypanosoma cruzi. *Genetics*, v. 171, n. 2, p. 527–43, doi:10.1534/genetics.104.038745, 2005.

WESTENBERGER, S. J. et al. Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. *BMC genomics*, v. 7, p. 60, doi:10.1186/1471-2164-7-60, 2006.

ZAKI, M. e MEIRA JUNIOR, W. *Fundamentals of Data Mining Algorithms*. Disponível em: <http://www.dcc.ufmg.br/miningalgorithms/>.

ZERBINO, D. R. e BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, v. 18, n. 5, p. 821–9, doi:10.1101/gr.074492.107, 2008.

ZHANG, Q.;; TIBAYRENC, M. e AYALA, F. J. Linkage disequilibrium in natural populations of Trypanosoma cruzi (flagellate), the agent of Chagas' disease. *The Journal of protozoology*, v. 35, n. 1, p. 81–5, 1988.

ZINGALES, B. et al. Epidemiology, biochemistry and evolution of Trypanosoma cruzi lineages based on ribosomal RNA sequences. *Memórias do Instituto Oswaldo Cruz*, v. 94 Suppl 1, p. 159–64, 1999.

ZINGALES, B. et al. A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. MEMORIAS DO INSTITUTO OSWALDO CRUZ., 2009.

ZINGALES, B. et al. The revised Trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, v. 12, n. 2, p. 240–53, 2012.

Roche Shutting Down 454 Sequencing Business. Disponível em: <http://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business>.

ANEXOS

ANEXO 1

Artigo *Evidence of substancial recombination among Trypanosoma cruzi II strains from Minas Gerais*

Infection, Genetics and Evolution 22 (2014) 183–191

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Evidence of substantial recombination among *Trypanosoma cruzi* II strains from Minas Gerais

Rodrigo de Paula Baptista [a], Daniella Alchaar D'Ávila [b], Marcela Segatto [a], Ítalo Faria do Valle [a], Glória Regina Franco [a], Helder Magno Silva Valadares [c], Eliane Dias Gontijo [d], Lúcia Maria da Cunha Galvão [b,e], Sérgio Danilo Junho Pena [a], Egler Chiari [b], Carlos Renato Machado [a,*], Andréa Mara Macedo [a,*]

[a] Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[b] Departamento de Parasitologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[c] Universidade Federal de São João del-Rey, Campus Centro-Oeste Dona Lindu, Divinópolis, Minas Gerais, Brazil
[d] Departamento de Medicina Preventiva e Social, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[e] Centro de Ciências da Saúde, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil

ARTICLE INFO

Article history:
Available online 1 December 2013

Keywords:
Trypanosoma cruzi II
Sex reproduction
Clonality
Recombination
Population substructuring
Wahlund effect

ABSTRACT

Due to the scarcity of evidence of sexuality in *Trypanosoma cruzi*, the causative agent of Chagas disease, it has been general accepted that the parasite reproduction is essentially clonal with infrequent genetic recombination. This assumption is mainly supported by indirect evidence, such as Hardy–Weinberg imbalances, linkage disequilibrium and a strong correlation between independent sets of genetic markers of *T. cruzi* populations. However, because the analyzed populations are usually isolated from different geographic regions, the possibility of population substructuring as generating these genetic marker imbalances cannot be eliminated. To investigate this possibility, we firstly compared the allele frequencies and haplotype networks using seven different polymorphic loci (two from mitochondrial and five from different nuclear chromosomes) in two groups of TcII strains: one including isolates obtained from different regions in Latin America and the other including isolates obtained only from patients of the Minas Gerais State in Brazil. Our hypothesis was that if the population structure is essentially clonal, Hardy–Weinberg disequilibrium and a sharp association between the clusters generated by analyzing independent markers should be observed in both strain groups, independent of the geographic origin of the samples. The results demonstrated that the number of microsatellite loci in linkage disequilibrium decreased from 4 to 1 when only strains from Minas Gerais were analyzed. Moreover, we did not observed any correlation between the clusters when analyzing the nuclear and mitochondrial loci, suggesting independent inheritance of these markers among the Minas Gerais strains. Besides, using a second subset of five physically linked microsatellite loci and the Minas Gerais strains, we could also demonstrate evidence of homologous recombination roughly proportional to the relative distance among them. Taken together, our results do not support a clonal population structure for *T. cruzi*, particularly in TcII, which coexists in the same geographical area, suggesting that genetic exchanges among these strains may occur more frequently than initially expected.

1. Introduction

Trypanosoma cruzi, the casual agent of Chagas disease, is a heterogeneous species as extensively demonstrated in several biological, biochemical, and molecular studies (Devera et al., 2003; Macedo and Pena, 1998; Macedo et al., 2001). Since 2009, to standardize the taxon nomenclature, the *T. cruzi* strains have been divided into six discrete taxonomic units (DTUs), TcI–TcVI (Zingales et al., 2009). The epidemiological relevance of these DTUs or even the subdivision of some of them or inclusion of a new one (Tcbat) is still under debate (Zingales et al., 2012; Pinto et al., 2012). However, TcII and its direct derived hybrids, TcV and TcVI, seem to be associated with more severe cases of Chagas disease in Southern Cone countries.

Despite considerable progress in cellular and molecular biology and in evolutionary genetics within recent decades, the debate

* Corresponding authors. Address: Laboratório de Genética Bioquímica, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Brazil. Tel.: +55 31 3409 2641; fax: +55 31 3409 2984 (A. M. Macedo and C. R. Machado).
E-mail addresses: crmachad@icb.ufmg.br (C.R. Machado), andrea@icb.ufmg.br (A.M. Macedo).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.11.021

regarding the population structure and reproductive mode of *T. cruzi* is far from being settled. The current theory, which is known as "the clonal theory of parasitic protozoa", was proposed by Tibayrenc and Ayala in 1990, which stated that *T. cruzi* and other protozoans undergo predominant clonal propagation with very little, if any, sexuality (Tibayrenc et al., 1990). For many years, the primary basis of this theory has been the observation of large deviations from Hardy–Weinberg (H–W) expectation and the apparent linkage disequilibrium (LD) between genotypes at different loci, some of which involved genomic separate compartments, such as the nucleus and mitochondria (Tibayrenc et al., 1986, 1991; Tibayrenc and Ayala, 2002).

Although prevalent in the literature, the clonal hypothesis has been challenged by many authors (Bastien et al., 1992; Bogliolo et al., 1996; Gaunt et al., 2003; Freitas et al., 2006). For instance, conflicting findings, obtained by our and other groups questioned the extent of this theory. Using microsatellite markers to study *T. cruzi* populations, we observed systematic deviations between the genotypic and allelic frequency expected by Hardy–Weinberg equilibrium. However, unlike an excess of heterozygosity naturally expected in populations in which recombination events are rare, an excess of homozygosity was observed (Oliveira et al., 1998; Balloux et al., 2003; Bengtsson, 2003). These findings were inconsistent with a population of parasites operating under an essentially clonal reproduction because asexual organisms exhibit the Meselson Effect (Mark Welch and Meselson, 2000).

Moreover, the occurrence of *T. cruzi* hybrid strains presenting evidence of sorting of homologous chromosomes and homologous recombination has been also identified (Gaunt et al., 2003; Freitas et al., 2006; Tomazi et al., 2009; Venegas et al., 2009; Westenberger et al., 2005; Carranza et al., 2009; Brisse et al., 2003; Carrasco et al., 1996; Sturm et al., 2003). And, at least for in vitro experiments, the recombination events in these parasites involve fusion of nuclear genotypes, homologous recombination, allelic loss, and uniparental inheritance of mitochondrial genotypes (Gaunt et al., 2003).

Despite these observations, many questions remain regarding the occurrence and frequency of recombination between current assumed asexual parasites such as *T. cruzi*. Indeed, most studies on population structures of *T. cruzi* have compared parasite populations isolated from distant geographical locations or at different times of isolation, which may have failed in detecting genetic recombination events among the strains due to bias of the sample selections. It is well known that if two or more subpopulations have different allele frequencies, the heterozygosity of the overall population is reduced, despite the subpopulations being in Hardy–Weinberg equilibrium. This is easily achieved, for example, by the existence of geographical barriers that prevent meetings and thus, the flow of genes among subpopulations, leading to a phenomenon known as the Wahlund effect (Poulin and Morand, 1999).

To verify this hypothesis, herein we compared the molecular data obtained from TcII strains isolated from both distant geographic areas throughout Latin America and only Minas Gerais, a Brazilian Southern state. In this study, our aim was to investigate whether geographical proximity would increase the probability of detecting evidence of sexual recombination among the strains. Our results demonstrated that when only TcII isolates from Minas Gerais were analyzed, four of the five microsatellite loci entered into H–W equilibrium, suggesting that the apparent deviation observed in previous studies could be due to geographical isolation of the analyzed strains.

2. Methods

2.1. T. cruzi populations analyzed

A total of 88 *T. cruzi* isolates were analyzed in this study (Table 1). Sixty of them were recently isolated from chronic patients, 58 from natives or residents in Minas Gerais (Southeast Brazil) and two from Bahia (Northeast Brazil). In addition, genotyping data from 28 other *T. cruzi* isolated in different times and different places in Latin America (Freitas et al., 2006), were also used for the calculation of HW and LD imbalances.

Table 1
T. cruzi II[a] strains and clones analyzed in this study.

Strain/clone[a]	Origin	Strain/clone[a]	Origin
002 B	MG/Brazil	154 a	MG/Brazil
003 a	MG/Brazil	162 a	MG/Brazil
005 B	MG/Brazil	128 a	BA/Brazil
007 B	MG/Brazil	38	MG/Brazil
009 B	MG/Brazil	013 a	MG/Brazil
010 B	MG/Brazil	016 B	MG/Brazil
011 B	MG/Brazil	192 a	MG/Brazil
012 a	MG/Brazil	JG	MG/Brazil
012 B	MG/Brazil	022 b	BA/Brazil
013 B	MG/Brazil	003 B	MG/Brazil
019 a	MG/Brazil	002 a	MG/Brazil
020 B	MG/Brazil	005 a	MG/Brazil
023 B	MG/Brazil	006 a	MG/Brazil
024 B	MG/Brazil	021 B	MG/Brazil
025 B	MG/Brazil	097 a	MG/Brazil
026 B	MG/Brazil	188 a	MG/Brazil
029 a	MG/Brazil	**Be62**[b]	MG/Brazil
031 a	MG/Brazil	Esmeraldo[b]	MG/Brazil
037 a	MG/Brazil	**GMS**[b]	MG/Brazil
044 a	MG/Brazil	**Ig539**[b]	MG/Brazil
045 a	MG/Brazil	Mas1 cl1[2]	MG/Brazil
050 b	MG/Brazil	**MPD**[2]	MG/Brazil
053 a	MG/Brazil	Tula cl2[b]	MG/Brazil
055 a	MG/Brazil	**84**[b]	MG/Brazil
058 a	MG/Brazil	**169/1**[b]	MG/Brazil
065 b	MG/Brazil	**200pm**[b]	MG/Brazil
067 a	MG/Brazil	**209**[b]	MG/Brazil
079 a	MG/Brazil	**239**[2]	MG/Brazil
083 a	MG/Brazil	**803**[b]	MG/Brazil
085 a	MG/Brazil	**1005**[b]	MG/Brazil
090 a	MG/Brazil	**1014**[b]	MG/Brazil
092 a	MG/Brazil	**1043**[b]	MG/Brazil
094 a	MG/Brazil	**1931**[b]	MG/Brazil
103 a	MG/Brazil	**GOCH**[2]	GO/Brazil
105 a	MG/Brazil	**577**[2]	GO/Brazil
109 a	MG/Brazil	**578**[2]	GO/Brazil
110 a	MG/Brazil	**580**[2]	GO/Brazil
115 b	MG/Brazil	183744[b]	GO/Brazil
116 a	MG/Brazil	CPI95/94[b]	PI/Brazil
120 a	MG/Brazil	OPS27/94[b]	PI/Brazil
129 a	MG/Brazil	GLT564[b]	RJ/Brazil
132 a	MG/Brazil	GLT593[b]	RJ/Brazil
138 a	MG/Brazil	**Y**[b]	SP/Brazil
146 a	MG/Brazil	CPI11/94[b]	Colombia

[a] *T. cruzi* strains were identified as *T. cruzi* II as described (D'Avila et al., 2009).
[b] The data of these strains were obtained from Freitas et al. (2006).

T. cruzi isolation from patients and all of the procedures were performed with the informed consent of the participants and approved by the Ethics Committee 087/99 of UFMG, Belo Horizonte, MG, Brazil.

The parasites DNA was extracted using the phenol:chloroform protocol (Macedo et al., 1992) and used as a template for the PCR assays.

2.2. T. cruzi DTU genotyping

All of the 88 isolates of *T. cruzi* were genotyped as TcII, using the triple assay as recommended by Macedo and colleagues (D'Avila et al., 2009) to discriminate the six *T. cruzi* DTUs. Initially, the strains were analyzed using PCR-RFLP of the COII gene followed by digestion with *Alu* I (de Freitas et al., 2006). The second step consists of a PCR of internally transcribed spacer leader (ITS) gene

(Burgos et al., 2007). The final step consisted of rDNA 24Sα PCR (Souto et al., 1996).

Using this typing strategy, TcII strains were identified as those presenting the follow genotype: COII haplotype C, ITS of 150 bp and rDNA 24Sα of 125 bp.

2.3. Mitochondrial DNA InDels characterization

Some TcII strains presented insertion–deletions (InDels) polymorphisms in the NADH dehydrogenase subunit 7 (ND7) and subunit 4 (ND4/CR4) mitochondrial genes. Thus, three different mitochondrial haplotypes can be identified among these parasites: haplotype C1 corresponding to a ~230 bp deletion in the ND4/CR4 gene, C2 corresponding to a 400 bp deletion in the ND7 gene region and C3 corresponding to wildtype genes (without any deletion) (Freitas et al., 2006; Macedo et al., 1992). In this present work, the InDels were typed in 47 of the 88 TcII samples by PCR amplifications of the ND7 and of the ND4/CR4 genes as previously described (Baptista et al., 2006; Carranza et al., 2009).

2.4. Nuclear microsatellite characterization

All of the 88 *T. cruzi* DNA samples, of which, 84 samples were isolated from human hosts and 4 samples were obtained from vectors (1005, 1024, GLT564 and GLT593), were subjected to a microsatellite assay using five dinucleotide repeats-based polymorphic loci – SCLE10, SCLE11, MCLE01, MCLG10 and MCLF10 – located on chromosomes 26, 30, 33, 39 and 6, respectively. Besides, 58 of these 88 samples were also analyzed for four extra polymorphic microsatellite loci (three trinucleotide and one tetranucleotide repeats-based loci – AAT8, ATT14, TAT20, and AAAT6), all of them located on chromosome 6 (Oliveira et al., 1998; Valadares et al., 2008). PCR amplifications and microsatellite analyses were conducted using an automated laser fluorescent DNA sequencer (GE Healthcare, Milwaukee, Wisconsin, USA) as previously described (Valadares et al., 2008). To allow the correct determination of the microsatellite allele sizes all electrophoresis runs were calibrated using internal fluorescent DNA fragments of 50–500 bp and standard strains such as Silvio X10, Colombiana (TcI), JG (TcI), Esmeraldo (TcII) and CLBrener (TcVI). The resulting chromatograms were analyzed using Allelelocator software (GE Healthcare).

2.5. Population genetics parameters

Herein, only five–nine polymorphic microsatellite markers were used to genotype TcII populations for the genetic analyses. However, previous long term work using the same set of markers have demonstrated their usefulness in differentiating any single stains and even clones of *T. cruzi* (Macedo et al., 2009). Using these markers, two expected microsatellite haplotypes for each *T. cruzi* population were initially estimated using PHASE (Stephens et al., 2001) and employed to reconstruct a haplotype network using the Network 4.6 software (Forster et al., 2000). Population substructuring was also inferred using STRUCTURE (Pritchard et al., 2000) and the appropriate number of subpopulations was determined using the log-likelihood values (*Ln*) for *K* 1–10 and Δ*K equation* (Evanno et al., 2005). A Hardy–Weinberg (H–W) estimation and the observed versus expected heterozygosity for each microsatellite locus with a corresponding confidence interval of 95% were performed using ARLEQUIN v3.1 (Excoffier et al., 2005). To investigate the association level between the different loci, linkage disequilibrium (LD) indices between all of the potential microsatellite loci pairwise comparisons were tested separately for the two subsets of markers – one involving five microsatellite loci located on five different chromosomes (SCLE10, SCLE11, MCLE01, MCLG10 and MCLF10) and another involving also five microsatellite loci, but located on the same chromosome 6 (AAT8, ATT14 TAT20, AAAT6 and MCLF10) – using GENEPOP (Raymond and Rousset, 1995).

3. Results

3.1. Mitochondrial ND7 and ND4 InDels polymorphisms

All of the analyzed TcII isolates exhibited the mitochondrial haplotype C, as revealed using PCR-RFLP analysis of the COII gene. However, the ND7 and ND4/CR4 gene characterization performed in 47 of the 88 TcII samples revealed the presence of the three ND4/ND7 haplotypes: haplotype C1 (corresponding to amplicons of 300/900 bp) was observed in 7 samples, haplotype C2 (corresponding to amplicons of 530/500 bp) was observed in 24 samples and haplotype C3 (corresponding to amplicons of 530/900 bp) was observed in 16 samples (Table 2).

To investigate whether the analyzed InDels were simple or had multiple origins (homoplasy), which might interfere with our analyses, InDel boundaries of the ND7 gene of polar TcII strains were sequenced (Supplementary Material). The alignment of the sequences confirmed that the InDel was apparently free of homoplasy and therefore suitable for population genetic analyzes performed herein (Fig. 1S).

3.2. Nuclear microsatellite characterization

By typing the five–nine microsatellite loci, we obtained the stable, reproducible and unique identity genotype for each analyzed strain (Table 2). For the SCLE11, MCLE01, AAT8 and TAT20 loci, we observed a predominance of heterozygous profiles. In contrast, for the SCLE10, MCLF10, MCLG10, ATT14 and AAAT6 loci, homozygous profiles were mainly found. Three or more peaks, which indicated a mixture of heterozygous strains or aneuploidy, were not observed.

3.3. TcII population structure analysis

The microsatellite profiles obtained for the 88 analyzed strains were converted into 108 allele haplotypes using the PHASE program, and these data were used to construct a haplotype network (Fig. 1). This network represents the distance between both allele haplotypes of each strain on the basis of the number of mutational steps. Considering the excesses of homozygosis generally observed for *T. cruzi* genome, we expected to detect slight differences in the number of mutational steps between the diploid haplotypes from each strain. However, they varied widely: the minimum number of mutational steps observed was 3 for the 128a strain, and the highest was 67 for the 012B strain, indicating that at least some of these isolates may have resulted from relatively recent hybridization events.

On the basis of the diversity observed in these strains, we investigated whether there was sufficient genetic variation to support a substructuring within the 47 Minas Gerais TcII populations typed for ND4/ND7. The appropriate number of subpopulations in the evaluated parasitic samples was determined using the STRUCTURE program by analyzing the log-likelihood values (*Ln*) for *K* 1–10 and the Δ*K* parameters. Both analyses indicated the presence of three subgroups ($K = 3$) within this population, named N1, N2 and N3 (Fig. 2A).

The F_{ST} and F_{IS} values were also estimated for the same TcII strains, as a measure of the genetic distance and mating among the strains. First, F_{ST} and F_{IS} analyses were separately performed for each nuclear subgroup identified by STRUCTURE (N1-3) to estimate the similarities and mating among the strains within each

Table 2
Mitochondrial ND4/7 haplotypes and microsatellite profiles of the *T. cruzi* II strains.

Strains/clones	ND4/7 haplotypes	Nuclear microsatellite haplotypes								
		SCLE10	SCLE11	MCLE01	MCLG10	MCLF10	TCAAT8	ATT14	TCTAT20	TCAAAT6
002 B	C2	26\26	14\14	13\15	8\8	8\8	10\17	13\13	15\7	8\9
003 a	C1	26\28	15\17	11\7	7\11	8\8	13\18	11\11	18\11	9\8
005 B	C3	25\36	16\18	11\14	8\8	8\8	12\10	8\12	18\21	9\5
007 B	C2	27\27	16\16	13\13	10\10	8\8	9\17	12\12	13\18	5\9
009 B	C3	28\36	11\14	9\8	8\12	8\8	13\10	13\13	18\15	8\5
010 B	C2	25\25	13\16	11\11	9\9	8\8	14\14	13\13	15\15	5\8
011 B	C3	27\36	13\17	11\8	8\8	8\8	10\10	13\13	21\21	5\5
012 a	C1	27\27	14\16	7\8	10\10	8\8	13\13	11\11	15\10	9\8
012 B	C2	26\26	13\17	15\11	8\8	8\8	17\18	12\12	7\15	9\9
013 B	C3	26\26	14\16	19\11	9\9	8\8	12\10	12\12	15\10	5\9
019 a	C1	27\36	13\17	8\7	8\8	8\8	12\18	12\12	10\11	9\8
020 B	C3	28\28	17\17	13\19	12\12	8\8	18\18	12\12	18\15	9\9
023 B	C3	26\36	14\14	8\8	8\8	8\8	9\12	8\12	15\10	5\9
024 B	C3	27\27	13\14	7\7	8\8	10\10	14\14	11\11	18\11	8\9
025 B	C2	26\26	13\14	11\19	11\11	8\8	9\9	11\11	15\17	8\6
026 B	C2	25\28	11\15	9\13	8\13	8\8	12\10	12\12	18\15	8\8
029 a	C2	26\28	11\13	13\13	7\7	8\8	17\17	12\12	19\16	9\9
031 a	C3	26\38	17\17	8\11	10\10	9\9	13\14	12\12	15\16	9\9
037 a	C2	28\28	14\15	19\11	11\11	8\8	8\8	11\11	13\11	9\9
044 a	C1	27\36	13\13	12\17	8\11	8\8	5\5	11\11	18\12	8\9
045 a	C2	25\25	14\16	19\11	8\8	8\8	14\14	12\12	10\15	8\8
050 b	C2	24\24	16\16	11\13	8\12	9\9	13\17	12\12	11\19	8\9
053 a	C2	25\25	11\13	9\12	7\7	9\9	10\13	12\12	18\19	8\8
055 a	C3	27\27	14\21	11\11	8\8	8\8	7\7	12\12	16\9	5\8
058 a	C2	25\36	16\18	11\15	8\8	8\8	12\9	8\12	18\21	9\5
065 b	C2	26\28	14\16	19\11	8\10	8\8	12\9	12\12	15\10	5\9
067 a	C2	26\26	13\13	11\15	8\8	8\8	17\17	12\12	15\7	9\9
079 a	C2	24\24	13\15	13\18	9\9	8\8	18\18	11\11	16\14	8\8
083 a	C2	26\26	12\15	9\9	8\13	8\8	13\10	11\11	19\11	8\8
085 a	C2	25\28	11\14	7\9	10\10	8\8	17\17	12\12	18\10	9\9
090 a	C3	28\36	11\18	9\9	10\10	8\8	13\9	9\12	18\15	8\5
092 a	C2	26\26	11\11	11\13	8\8	8\8	9\9	12\12	15\15	9\5
094 a	C3	25\25	11\14	18\18	10\10	8\8	14\14	12\12	15\10	9\8
103 a	C3	26\28	10\13	9\18	5\12	8\8	12\12	11\11	19\19	8\8
105 a	C2	28\32	14\16	20\11	8\8	8\8	12\9	11\11	10\10	9\9
109 a	C2	25\28	16\16	11\13	7\7	8\8	12\12	11\11	19\16	9\8
110 a	C2	27\27	14\16	11\11	10\10	10\10	13\10	11\11	17\10	8\9
115 b	C1	28\28	15\15	11\11	8\8	8\8	2\2	11\11	8\21	7\7
116 a	C1	26\36	11\13	9\7	7\12	8\9	13\10	12\12	18\12	8\8
120 a	C2	31\31	13\14	9\12	8\8	8\8	9\8	13\13	10\26	9\9
129 a	C1	25\25	14\14	7\11	8\8	8\8	13\13	11\11	15\16	9\8
132 a	C3	27\27	12\13	9\20	6\7	8\8	18\18	11\11	11\7	9\8
138 a	C3	25\26	14\15	11\11	6\8	8\8	22\22	11\11	13\22	8\8
146 a	C2	26\26	13\14	12\8	8\8	8\8	13\9	9\12	19\15	8\5
154 a	C3	28\28	12\19	21\19	5\5	8\9	10\10	11\11	7\13	9\9
162 a	C3	26\26	14\16	19\11	8\8	8\8	12\9	12\12	15\10	5\9
128 a	C2	26\26	14\14	9\9	8\8	8\8	14\14	12\12	7\10	9\9
38	C*	28\28	14\14	9\19	10\10	8\8	12\12	12\12	10\18	9\8
013 a	C*	26\28	14\14	10\19	10\10	8\8	12\10	12\12	11\16	9\6
016 B	C*	26\26	14\16	19\11	9\9	9\9	12\10	12\12	15\10	5\9
192 a	C*	27\27	13\14	19\9	8\8	8\8	18\18	12\12	7\10	8\9
JG	C*	27\28	11\13	11\12	8\8	8\8	13\13	11\11	10\19	9\8
022 b	C*	25\25	13\16	7\11	9\9	9\9	14\14	8\12	18\15	8\5
003 B	C*	NA	14\16	19\9	NA	8\8	12\10	NA	15\10	5\9
002 a	C*	NA	13\13	7\11	NA	NA	14\14	NA	18\15	8\9
005 a	C*	NA	14\14	7\12	NA	NA	12\10	NA	10\11	9\8
006 a	C*	NA	14\14	9\9	NA	NA	17\17	NA	10\18	9\8
021 B	C*	27\27	13\14	11\9	9\9	8\8	14\14	NA	15\18	5\9
097 a	C*	28\31	15\17	11\11	8\8	8\8	9\12	NA	26\10	9\9
188 a	C*	NA	14\16	9\13	NA	NA	18\18	NA	10\17	9\9
Esmeraldo	C*	28\34	14\18	6\12	8\9	7\8	10\9	NA	12\7	7\9
Be62	C*	27\28	16\16	12\13	8\8	8\8	ND	ND	ND	ND
GMS	C*	27\33	13\17	7\11	8\8	8\8	ND	ND	ND	ND
Ig539	C*	31\31	14\15	10\19	6\9	9\9	ND	ND	ND	ND
Mas1 cl1	C*	23\34	14\14	8\8	8\8	9\9	ND	ND	ND	ND
MPD	C*	24\26	10\15	13\13	8\10	7\7	ND	ND	ND	ND
Tula cl2	C*	35\35	13\14	9\9	8\8	8\8	ND	ND	ND	ND
84	C*	24\26	15\16	7\11	8\8	8\8	ND	ND	ND	ND
169/1	C*	27\28	13\13	8\13	8\8	7\9	ND	ND	ND	ND
200pm	C*	29\29	15\16	12\12	8\9	9\9	ND	ND	ND	ND
209	C*	25\28	15\16	10\10	8\8	8\8	ND	ND	ND	ND
239	C*	26\28	14\17	11\19	8\8	8\8	ND	ND	ND	ND
803	C*	26\31	11\11	12\20	8\11	9\9	ND	ND	ND	ND

Table 2 (*continued*)

Strains/clones	ND4/7 haplotypes	Nuclear microsatellite haplotypes								
		SCLE10	SCLE11	MCLE01	MCLG10	MCLF10	TCAAT8	ATT14	TCTAT20	TCAAAT6
1005	C*	28\28	15\16	9\22	8\9	8\8	ND	ND	ND	ND
1014	C*	25\28	13\13	14\15	8\8	8\8	ND	ND	ND	ND
Y	C*	27\27	15\15	9\9	8\8	8\9	ND	ND	ND	ND
1043	ND	28\31	17\17	8\11	8\9	7\7	ND	ND	ND	ND
1931	ND	28\28	9\15	14\14	10\10	9\9	ND	ND	ND	ND
GOCH	ND	26\31	15\18	8\19	6\6	10\10	ND	ND	ND	ND
577	ND	27\31	14\16	20\21	6\8	10\10	ND	ND	ND	ND
578	ND	31\31	13\16	9\18	8\8	10\10	ND	ND	ND	ND
580	ND	30\30	13\13	9\11	8\8	9\9	ND	ND	ND	ND
183744	ND	31\31	13\13	9\11	8\8	9\9	ND	ND	ND	ND
CPI95/94	ND	29\29	13\16	6\10	8\8	9\9	ND	ND	ND	ND
OPS27/94	ND	34\39	13\18	10\12	8\8	9\9	ND	ND	ND	ND
GLT564	ND	28\28	14\14	12\12	8\10	8\8	ND	ND	ND	ND
GLT593	ND	26\33	14\14	12\12	9\11	7\7	ND	ND	ND	ND
CPI11/94	ND	29\39	15\15	9\9	8\8	9\9	ND	ND	ND	ND

The mitochondrial ND4/7 haplotypes: C1 (300/900 bp), C2 (530/500 bp) or C3 (530/900 bp). The microsatellite profiles were determined by the number of repetitions for each allele at each one of the five analyzed loci. C*: ND4/7 haplotype have not been determined for these isolates, but as expected for the *T. cruzi* II strains all of the mitochondrial COII profiles were haplotype C. NA, not amplified; ND, not determined.

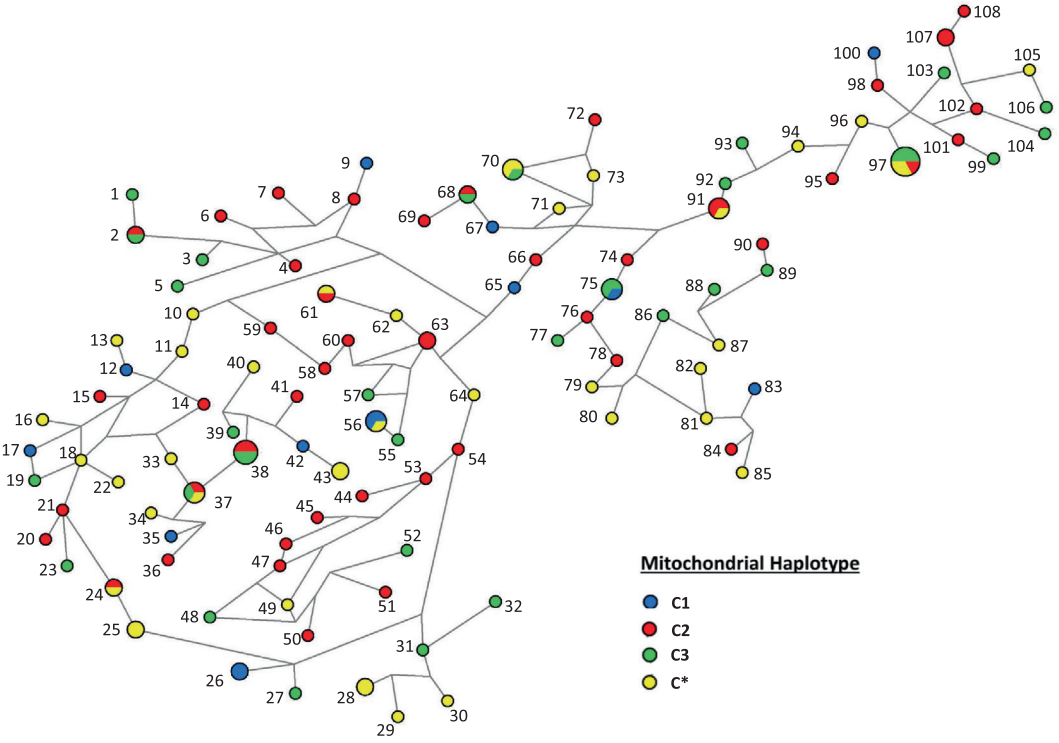

Fig. 1. Haplotype network basing on the PHASE data, indicating the distance between different nuclear haplotypes from TcII strains. The circle size was proportional to the number of identical haplotypes observed. Different colors referee to different ND4/7 haplotypes: blue – C1 (300/900 bp), red – C2 (530/500 bp), green – C3 (530/900 bp) and yellow – C* (not determined).

observed subpopulation. In this case, F_{ST} values >0.05 for the three subgroups (F_{ST} = 0.0661–0.1786) were consistent with significant genetic differentiation among the isolates from each STRUCTURE group. However, positive F_{IS} values (F_{IS} = 0.0540–0.8455) indicated that the amount of heterozygous offspring in these subpopulations was lower than expected, most likely due to inbreeding. In fact, in the presence of significant inbreeding, the mating is nonrandom, and close relatives reproduce among themselves, and because these relatives likely have similar genes, the offspring are likely to be homozygous. Thus, taken together, both the high F_{ST} values and positive F_{IS} values observed for each STRUCTURE subpopulation (Table 3) indicated free interbreeding among the analyzed TcII strains of Minas Gerais.

3.4. Compelling evidence for substantial recombination within TcII

On the basis of the microsatellite genotype data, we employed two parameters to estimate the presence of genetic recombination within the parasite populations: H–W imbalance and LD (Table 4). Initially, these parameters were evaluated on the basis of the entire dataset population (strains belonging to Minas Gerais and other regions of Latin American), and then with only samples isolated from Minas Gerais, using the subset of five loci located on the different chromosomes. Our null hypothesis was that these TcII populations were in Hardy–Weinberg equilibrium and that the microsatellite loci segregate independently. For the entire dataset, most of the analyzed loci

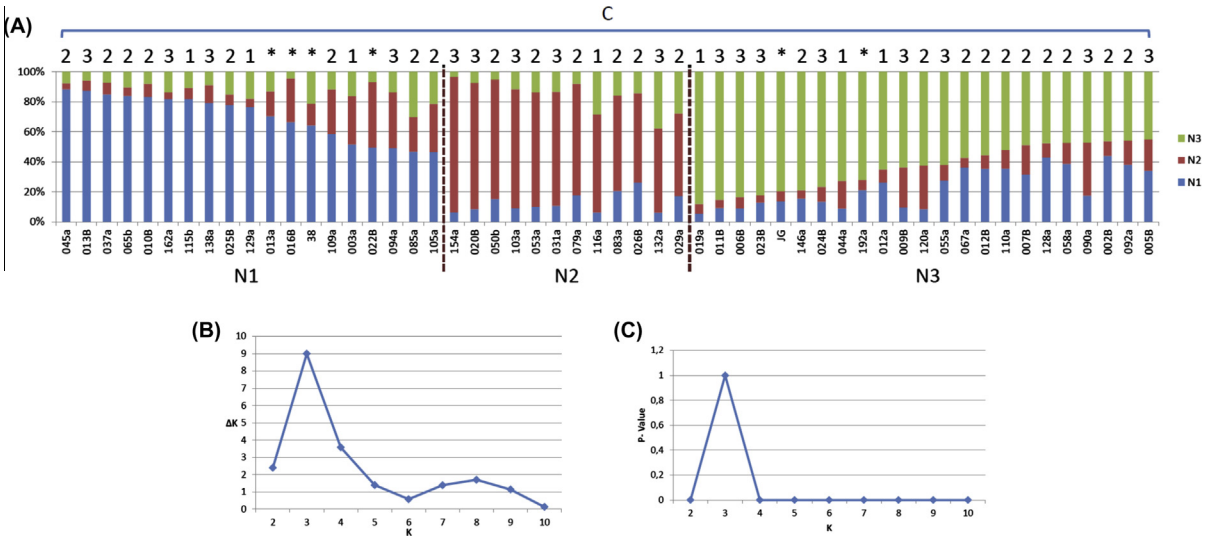

Fig. 2. Subgroup determination on the basis of the nuclear and mitochondrial markers. The number of subsets in our test with MG samples, basing on nuclear data, were determined using STRUCTURE, where (A) is the "bar plotting" graph that has been partitioned into three-color segments ($\Delta K = 3$), and K was determined by the correlation between the log-likelihood values (Ln), where K varied from 1 to 10 (B) or by the ΔK calculation (C). Letters above and below the "bar plotting" graph indicate lack of correlation between nuclear and mitochondrial markers. All three mitochondrial haplotypes (C1, C2 and C3) were dispersed among the three different nuclear subgroups (N1, N2 and N3), demonstrating that there is no correlation between the two markers. C*: ND4/7 haplotypes have not been determined.

Table 3
Multilocus estimates for the diploid data.

Locus	*F*is	*F*st	Fit
SCLE11	0.0729	0.0970	0.1628
MCLE01	0.0540	0.0661	0.1165
SCLE10	0.5062	0.0943	0.5528
MCLF10	0.8455	0.1063	0.8619
MCLG10	0.7276	0.1786	0.7763
All together	0.3660	0.1074	0.4341

F_{ST} and F_{IS} analyses were separately performed for each nuclear subgroup identified by STRUCTURE (N1-3) to estimate the similarities and mattings among the strains within each observed subpopulation. F_{ST} values >0.05 were consistent with genetic differentiation among the isolates from each STRUCTURE group. Positive F_{IS} values indicate that the amount of heterozygous offspring in these subpopulations was lower than expected.

Table 4
Statistical tests for population genetics.

A. Hardy–Weinberg test

Loci	*p*-value (World)	*p*-value (MG)
SCLE11	0.00057	0.05150
SCLE10	0.00000	0.06232
MCLE01	0.00000	0.06100
MCLF10	0.00000	0.00000
MCLFG10	0.00000	0.05110

B. Linkage disequilibrium test

Loci pair			*p*-value (World)	*p*-value (MG)
SCLE11	&	MCLE01	0.000000	0.172377
SCLE11	&	SCLE10	0.17168	0.344099
MCLE01	&	SCLE10	0.062400	0.142935
SCLE11	&	MCLF10	0.250320	0.548089
MCLE01	&	MCLF10	0.044430	0.420419
SCLE10	&	MCLF10	0.000000	0.528246
SCLE11	&	MCLG10	0.000000	0.026943
MCLE01	&	MCLG10	0.273930	0.869349
SCLE10	&	MCLG10	0.538440	0.151469
MCLF10	&	MCLG10	0.060340	0.066145

Population statistical tests observed for the two data sets. The "World" included isolates from Pan-America locations. "MG" included only the isolates from Minas Gerais. (A) H–W *p*-values for the two data sets and (B) LD *p*-values for each pair of loci. For both methods, *p*-values less than 0.05 represented deviations and imbalances.

presented H–W imbalances (5/5) and LD (4/10) results with *p*-values <0.05 as calculated using the Pearson's Chi square test, which indicated significant deviations from the expected values for sex-reproducing organisms. These findings suggested that the genotypes were passed in blocks from generations, supporting the idea of predominant clonality in these parasites. However, when only samples from Minas Gerais were analyzed, 4/5 loci were in H–W equilibrium and 9/10 loci pairs in Linkage equilibrium (*p*-values >0.05), although some of them only narrowly passed the test (Table 4).

Because we had detected the presence of three haplotypes for mitochondrial markers (haplotypes C1, C2 and C3) and also three subgroups for nuclear markers (N1, N2, N3) as determined by STRUCTURE for the parasites isolated from Minas Gerais, we further investigated whether there was a correlation between these two independent sets of genetic markers. The presence of an association between these sets of markers might indicate that they were inherited together as expected in a clonal-based population (Zhang et al., 1988). However, no correlation was observed between the TcII subgroups identified by the nuclear or mitochondrial markers, as demonstrated in Fig. 2B: the three mitochondrial haplotypes were dispersed among the three nuclear subgroups.

Due to the previously used polymorphic markers, all of them located on different chromosomes, the molecular data so far obtained indicated the occurrence of independent segregation of chromosomes in TcII strains isolated from Minas Gerais, but failed to identify the occurrence of homologous recombination. To investigate that, we further calculated H–W deviation using the MCLF10 and four additional microsatellite loci (TCAAT8, TCTAT20, TCAAAT6 and ATT14), all of them located on chromosome 6 (Fig. 3). Even though on the same chromosome, six of the 10 loci pairwise comparisons showed to be in linkage equilibrium (Table 5), suggesting that besides segregation of chromosomes homologous recombina-

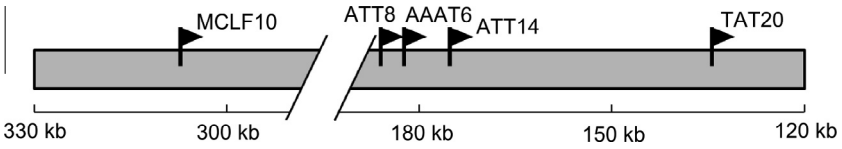

Fig. 3. Schematic representation of a fragment of *T. cruzi* chromosome 6 based on tritrypdb.org. The black flags mark the relative position of the five used microsatellites loci located on this chromosome: MCLF10, ATT8, AAAT6, ATT14 and TAT20.

Table 5
Linkage disequilibrium test for the five microsatellite loci located on chromosome 6.

Linkage disequilibrium Test			
Loci pair			*p*-value
TCAAT8	&	TCTAT20	0.00416
TCAAT8	&	TCAAAT6	0.00346
TCTAT20	&	TCAAAT6	0
TCAAT8	&	ATT14	0.11851
TCTAT20	&	ATT14	0.25128
TCAAAT6	&	ATT14	0.00058
TCAAT8	&	MCLF10	0.86894
TCTAT20	&	MCLF10	0.61134
TCAAAT6	&	MCLF10	0.95698
ATT14	&	MCLF10	0.49585

tion may also be involved. Moreover, when we compared the *p*-value of LD for the 10 loci pairwise and their relative position in the chromosome 6 we observed a good correlation between these two elements. For instance, the locus MCLF10 that is located far from the others is in linkage equilibrium with all of them. On the other hand, the TCAAAT6, ATT14 and TCAAT8 loci that are close to each other are clearly inherited in bloc. Some linkage disequilibrium was detected between this bloc and the locus TCTAT20 located a bit apart (Fig. 3).

4. Discussion

Although sexual reproduction has being indubitably demonstrated for *T. cruzi* (Gaunt et al., 2003; Ramirez et al., 2012), many questions concerning the frequency and importance of these events for the population structure and biology of these parasites are still under debate. The present study employs an innovative strategy to evaluate the TcII population dynamic simultaneously at local and broad scales using nuclear and mitochondrial sets of polymorphic markers. The experimental approaches used in this study were designed to better understand why clonality was previously assumed as the predominant mechanism for *T. cruzi* reproduction (Tibayrenc and Ayala, 2002).

Thus, some of the most common parameters (Hardy–Weinberg: H–W and Linkage Disequilibrium; LD) in the literature were used to demonstrate the clonal structure of protozoa in two TcII populations: one population, which included strains isolated from a broad geographical region (different places in Latin America) and a smaller population including strains isolated from a much more restricted area (Minas Gerais, Brazil). The rational for this approach was to evaluate the potential influence of geographical barriers in the LD and H–W imbalances that are usually detected for these parasites (Oliveira et al., 1998; Wahlund, 1928).

In this study, we detected the occurrence of LD and H–W deviations when *T. cruzi* isolated from distant geographic regions were analyzed, which could be interpreted as indicative of predominant clonal reproduction (Tibayrenc et al., 1991). However, this scenario was completely modified when we analyzed only *T. cruzi* circulating in a small geographic area. In this latter case, the Linkage and H–W equilibrium were restored for most of the loci analyzed, indicating that geographic distances and/or physical barriers may be important factors in reducing opportunities for sexual reproduction and thereby recombination rates among the *T. cruzi* strains.

Thus, our results are consistent with subpopulation structure leading to the Wahlund effect (a common phenomenon observed when individuals of a putative population are from genetically segregated subpopulations), in which the genetic marker imbalances contributed to the underestimation of the occurrence of sex and recombination in *T. cruzi*. This was particularly relevant because most previous studies have analyzed strains isolated from different areas (Wahlund, 1928).

The strategy selected in this study of analyzing *T. cruzi* strains isolated from the same location (Minas Gerais) and from the same lineage TcII to investigate the extension of homologous recombination and random allele segregation among the parasites suggests that genetic exchanges among these strains were more frequent than initially expected. Whether this is a specific characteristic of some *T. cruzi* populations or a general feature of the *T. cruzi* taxon remain to be clarified, but Llewellyn et al. (2009a,b), using 48–49 microsatellite loci to investigate population structure of TcI and TcIII strains, observed excess homozygosity a finding incongruent with extreme models of long-term clonal evolution in diploids.

Similarly, Ocaña-Mayorga et al. (2010), using 10 microsatellite loci and 81 isolates of TcI populations isolated from triatomines and small mammals of 16 communities in Loja Province, southern Ecuador, identified of H–W equilibrium allele frequencies and linkage equilibrium even among physically linked loci.

By analyzing a nuclear (glucose phosphate isomerase) and a mitochondrial (NADH dehydrogenase subunit 1) gene of 60 TcI and 15 reference strains belonging to the six DTUs, Barnabé and Brenière (2012) identified evidence of trans-lineage mitochondrial introgression. This kind of event, initially thought to be rare, but has been seen at least in nine independent cases, twice between TcII and TcIII strains (Freitas et al., 2006) and eight times between TcI and TcIII or TcIV (Machado and Ayala, 2001; Lewis et al., 2011; Messenger et al., 2012).

Finally, unlike the prevailing view that hybridization events in *T. cruzi* are ancient and of little epidemiological importance, Lewis et al. (2011) dated key evolutionary events in the taxon, including the emergence of hybrid lineages TcV and TcVI, within the last 60,000 years, indicating that recombination still active in the taxon.

In conclusion, taken together, these findings suggest that genetic exchanges among *T. cruzi* strains occur more frequently than initially expected and may be important to shape and define biological properties of DTUs. Nevertheless, the relevance of these recombination events for the whole parasite taxon in both evolutionary and generation scales remains to be established (Tibayrenc and Ayala, 2013).

5. Author contributions

Conceived and designed the experiments: R.P.B., D.A.D., G.R.F., C.R.M. and A.M.M. Performed the experiments: R.P.B., D.A.D., M.S.C. and I.F.V. Analyzed the data: R.P.B., M.S.C., I.F.V., C.R.M. and A.M.M. Contributed to the collecting of samples and writing

of the manuscript: R.P.B., D.A.D., M.S.C., I.F.V., H.M.S.V., E.D.G., L.M.C.G., S.D.J.P., E.C., C.R.M. and A.M.M.

Acknowledgments

We thank Fernanda Kehdy and Giordano Souza from the Laboratório de Diversidade de Genética Humana, Universidade Federal de Minas Gerais for their help in part of the population analyses. We also thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) for supporting this study.

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.2013.11.021.

References

Balloux, F., Lehmann, L., de Meeus, T., 2003. The population genetics of clonal and partially clonal diploids. Genetics 164, 1635–1644.
Baptista, C.S., Vencio, R.Z., Abdala, S., Carranza, J.C., Westenberger, S.J., Silva, M.N., Pereira, C.A., Galvao, L.M., Gontijo, E.D., Chiari, E., Sturm, N.R., Zingales, B., 2006. Differential transcription profiles in *Trypanosoma cruzi* associated with clinical forms of Chagas disease: maxicircle NADH dehydrogenase subunit 7 gene truncation in asymptomatic patient isolates. Mol. Biochem. Parasitol. 150, 236–248.
Barnabé, C., Brenière, S.F., 2012. Scarce events of mitochondrial introgression in *Trypanosoma cruzi*: new case with a Bolivian strain. Infect. Genet. Evol. 12, 1879–1883.
Bastien, P., Blaineau, C., Pages, M., 1992. *Leishmania*: sex, lies and karyotype. Parasitol. Today 8, 174–177.
Bengtsson, B.O., 2003. Genetic variation in organisms with sexual and asexual reproduction. J. Evol. Biol. 16, 189–199.
Bogliolo, A.R., Lauria-Pires, L., Gibson, W.C., 1996. Polymorphisms in *Trypanosoma cruzi*: evidence of genetic recombination. Acta Trop. 61, 31–40.
Brisse, S., Henriksson, J., Barnabe, C., Douzery, E.J., Berkvens, D., Serrano, M., De Carvalho, M.R., 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.
Burgos, J.M., Altcheh, J., Bisio, M., Duffy, T., Valadares, H.M., Seidenstein, M.E., Piccinali, R., Freitas, J.M., Levin, M.J., Macchi, L., Macedo, A.M., Freilij, H., Schijman, A.G., 2007. Direct molecular profiling of minicircle signatures and lineages of *Trypanosoma cruzi* bloodstream populations causing congenital Chagas disease. Int. J. Parasitol. 37, 1319–1327.
Carranza, J.C., Valadares, H.M., D'Avila, D.A., Baptista, R.P., Moreno, M., Galvao, L.M., Chiari, E., Sturm, N.R., Gontijo, E.D., Macedo, A.M., Zingales, B., 2009. *Trypanosoma cruzi* maxicircle heterogeneity in Chagas disease patients from Brazil. Int. J. Parasitol. 39, 963–973.
Carrasco, H.J., Frame, I.A., Valente, S.A., Miles, M.A., 1996. Genetic exchange as a possible source of genomic diversity in sylvatic populations of *Trypanosoma cruzi*. Am. J. Trop. Med. Hyg. 54, 418–424.
D'Avila, D.A., Macedo, A.M., Valadares, H.M., Gontijo, E.D., de Castro, A.M., Machado, C.R., Chiari, E., Galvao, L.M., 2009. Probing population dynamics of *Trypanosoma cruzi* during progression of the chronic phase in chagasic patients. J. Clin. Microbiol. 47, 1718–1725.
de Freitas, J.M., Augusto-Pinto, L., Pimenta, J.R., Bastos-Rodrigues, L., Goncalves, V.F., Teixeira, S.M., Chiari, E., Junqueira, A.C., Fernandes, O., Macedo, A.M., Machado, C.R., Pena, S.D., 2006. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. PLoS Pathog. 2, e24.
Devera, R., Fernandes, O., Coura, J.R., 2003. Should *Trypanosoma cruzi* be called "cruzi" complex? a review of the parasite diversity and the potential of selecting population after in vitro culturing and mice infection. Mem. Inst. Oswaldo Cruz 98, 1–12.
Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611–2620.
Excoffier, L., Laval, G., Schneider, S., 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online 1, 47–50.
Forster, P., Rohl, A., Lunnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C., Brinkmann, B., 2000. A short tandem repeat-based phylogeny for the human Y chromosome. Am. J. Hum. Genet. 67, 182–196.
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., Acosta, N., de Arias, A.R., Miles, M.A., 2003. Mechanism of genetic exchange in American trypanosomes. Nature 421, 936–939.
Lewis, M.D., Llewellyn, M.S., Yeo, M., Acosta, N., Gaunt, M.W., Miles, M.A., 2011. Recent, independent and anthropogenic origins of *Trypanosoma cruzi* hybrids. PLoS Negl. Trop. Dis. 5, e1363.
Llewellyn, M.S., Miles, M.A., Carrasco, H.J., Lewis, M.D., Yeo, M., Vargas, J., Torrico, F., Diosque, P., Valente, V., Valente, S.A., Gaunt, M.W., 2009a. Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. PLoS Pathog. 5, e1000410.
Llewellyn, M.S., Lewis, M.D., Acosta, N., Yeo, M., Carrasco, H.J., Segovia, M., Vargas, J., Torrico, F., Miles, M.A., Gaunt, M.W., 2009b. *Trypanosoma cruzi* IIc: phylogenetic and phylogeographic insights from sequence and microsatellite analysis and potential impact on emergent Chagas disease. PLoS Negl. Trop. Dis. 1, e510.
Macedo, A.M., Pena, S.D., 1998. Genetic variability of *Trypanosoma cruzi*: implications for the pathogenesis of Chagas disease. Parasitol. Today 14, 119–124.
Macedo, A.M., Martins, M.S., Chiari, E., Pena, S.D., 1992. DNA fingerprinting of *Trypanosoma cruzi*: a new tool for characterization of strains and clones. Mol. Biochem. Parasitol. 55, 147–153.
Macedo, A.M., Pimenta, J.R., Aguiar, R.S., Melo, A.I., Chiari, E., Zingales, B., Pena, S.D., Oliveira, R.P., 2001. Usefulness of microsatellite typing in population genetic studies of *Trypanosoma cruzi*. Mem. Inst. Oswaldo Cruz 96, 407–413.
Macedo, A.M., Rodrigues, C.M., Oliveira, R.P., Franco, G.R., Machado, C.R., Pena, S.D.J., Valadares, H.M.S., 2009. Contribution of *Trypanosoma cruzi* polymorphic microsatellite analyses in refining epidemiology aspects of Chagas disease. Rev. Soc. Bras. Med. Trop. 42, 80–86.
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.
Mark Welch, D.B., Meselson, M., 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. Science 288, 1211–1215.
Messenger, L.A., Llewellyn, M.S., Bhattacharyya, T., Franzén, O., Lewis, M.D., Ramírez, J.D., Carrasco, H.J., Andersson, B., Miles, M.A., 2012. Multiple mitochondrial introgression events and heteroplasmy in *Trypanosoma cruzi* revealed by maxicircle MLST and next generation sequencing. PLoS Negl. Trop. Dis. 6, e1584.
Ocaña-Mayorga, S., Llewellyn, M.S., Costales, J.A., Miles, M.A., Grijalva, M.J., 2010. Sex, subdivision, and domestic dispersal of *Trypanosoma cruzi* lineage I in southern Ecuador. PLoS Negl. Trop. Dis. 4, e915.
Oliveira, R.P., Broude, N.E., Macedo, A.M., Cantor, C.R., Smith, C.L., Pena, S.D., 1998. Probing the genetic population structure of *Trypanosoma cruzi* with polymorphic microsatellites. Proc. Natl. Acad. Sci. USA 95, 3776–3780.
Pinto, C.M., Kalko, E.K., Cottontail, I., Wellinghausen, N., Cottontail, V.M., 2012. TcBat a bat-exclusive lineage of *Trypanosoma cruzi* in the Panama Canal Zone, with comments on its classification and the use of the 18S rRNA gene for lineage identification. Infect. Genet. Evol. 12, 1328–1332.
Poulin, R., Morand, S., 1999. Geographical distances and the similarity among parasite communities of conspecific host populations. Parasitology 119 (Pt 4), 369–374.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure from multilocus genotype data. Genetics 155, 945–959.
Ramirez, J.D., Duque, M.C., Montilla, M., Cucunuba, Z., Guhl, F., 2012. Natural and emergent *Trypanosoma cruzi* I genotypes revealed by mitochondrial (Cytb) and nuclear (SSU rDNA) genetic markers. Exp. Parasitol. 132, 487–494.
Raymond, M., Rousset, F., 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. 86, 248–249.
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.
Stephens, M., Smith, N.J., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68, 978–989.
Sturm, N.R., Vargas, N.S., Westenberger, S.J., Zingales, B., Campbell, D.A., 2003. Evidence for multiple hybrid groups in *Trypanosoma cruzi*. Int. J. Parasitol. 33, 269–279.
Tibayrenc, M., Ayala, F.J., 2002. The clonal theory of parasitic protozoa: 12 years on. Trends Parasitol. 18, 405–410.
Tibayrenc, M., Ayala, F.J., 2013. How clonal are *Trypanosoma* and *Leishmania*? Trends Parasitol. 29, 264–269.
Tibayrenc, M., Ward, P., Moya, A., Ayala, F.J., 1986. Natural populations of *Trypanosoma cruzi*, the agent of Chagas disease, have a complex multiclonal structure. Proc. Natl. Acad. Sci. USA 83, 115–119.
Tibayrenc, M., Kjellberg, F., Ayala, F.J., 1990. A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. Proc. Natl. Acad. Sci. USA 87, 2414–2418.
Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Breniere, S.F., Darde, M.L., Ayala, F.J., 1991. Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. Proc. Natl. Acad. Sci. USA 88, 5129–5133.
Tomazi, L., Kawashita, S.Y., Pereira, P.M., Zingales, B., Briones, M.R., 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.
Valadares, H.M., Pimenta, J.R., de Freitas, J.M., Duffy, T., Bartholomeu, D.C., Oliveira Rde, P., Chiari, E., Moreira Mda, C., Filho, G.B., Schijman, A.G., Franco, G.R., Machado, C.R., Pena, S.D., Macedo, A.M., 2008. Genetic profiling of *Trypanosoma cruzi* directly in infected tissues using nested PCR of polymorphic microsatellites. Int. J. Parasitol. 38, 839–850.

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

Wahlund, S., 1928. Zusammensetzung von population und korrelationserscheinung vom standpunkt der vererbungslehre aus betrachtet. Hereditas 11, 65–106.

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

Zhang, Q., Tibayrenc, M., Ayala, F.J., 1988. Linkage disequilibrium in natural populations of *Trypanosoma cruzi* (flagellate), the agent of Chagas' disease. J. Protozool. 35, 81–85.

Zingales, B., Andrade, S.G., Briones, M.R., 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.

Zingales, B., Miles, M.A., Campbell, D.A., Tibayrenc, M., Macedo, A.M., Teixeira, M.M.G., Schijman, A.G., Llewellyn, M.S., Lages-Silva, E., Machado, C.R., Andrade, S.G., Sturm, N.R., 2012. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. Infect. Genet. Evol. 12, 240–253.

ANEXO 2

Lista de manuscritos em fase final de elaboração

LISTA DOS MANUSCRITOS EM FASE FINAL DE ELABORAÇÃO

Título: Solving the population structure of *Trypanosoma cruzi*: a look at the non-Esmo parent

Autores: Rodrigo P. Baptista, Glória R. Franco, Carlos R. Machado, Jessica C. Kissinger, Daniella C. Bartholomeu e Andréa M. Macedo

Título: Assembling high repetitive trypanosomatids genomes using short reads: *Trypanosoma cruzi* III 231 strain

Autores: Rodrigo P. Baptista, João Luis R. Cunha, Jeremy DeBarry, Jessica C. Kissinger, Daniella C. Bartholomeu and Andréa M. Macedo

Título: Target sites detection in genomic sequences for phylogenetic studies: a *T. cruzi* amastin study case

Autores: Rodrigo P. Baptista, Rodrigo B. Kato, Bráulio RGM Couto, Marcos A. Santos, Gisele L. Papa, Daniella C. Bartholomeu, Jadson C. Belchior e Andréa M. Macedo.