UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Variação de número de cópias revela extensa plasticidade genômica no parasito *Trypanosoma cruzi*

João Luís Reis Cunha

Belo Horizonte
2017

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

Variação de número de cópias revela extensa plasticidade genômica no parasito *Trypanosoma cruzi*

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

Aluno: João Luís Reis Cunha

Orientadora: Daniella Castanheira Bartholomeu

Co-Orientador: Gustavo Coutinho Cerqueira

Belo Horizonte -MG
Junho - 2017

“I am among those who think that science has great beauty.

A scientist in his laboratory is not only a technician:

he is also a child placed before natural phenomena

which impress him like a fairy tale”

Marie Curie

Meus sinceros agradecimentos

À Daniella Castanheira Bartholomeu, que me orientou desde minha primeira iniciação científica, em 2007, até o término de meu doutorado. Muito obrigado por todas as chances de crescimento científico e pessoal que você me proporcionou nestes 10 anos de aprendizado. Se hoje consigo formular perguntas interessantes e meios de responde-las sem a timidez de errar, é porque você sempre teve tempo para ouvir todos os meus questionamentos e corrigi-los quando necessário. Muito obrigado por sua disponibilidade, bom humor e humildade, características importantes e raras atualmente. Você sempre será para mim um exemplo de pessoa e de profissional.

Ao Gustavo Cerqueira, que não só me deu a oportunidade de estudar um ano no *Broad Institute* (onde cresci muito cientificamente) como também me recebeu em sua casa durante minha estadia em Boston. Ainda sinto falta de "rabiscar a parede toda" para montarmos os esquemas mais adequados para nossos projetos. A partir de seus ensinamentos, hoje fundamento todos os parâmetros que uso em minhas análises e uso o café como um dos principais combustíveis para passar o dia.

A Gabi "Oracle" Luiz, por todos os ensinamentos e companhia nas manhãs. Devo muito do que sei hoje em bioinformática a paciência que você teve em me ensinar tudo o que precisei, sempre com um sorriso. "Sem *strict,* sem ajuda".

A Laila, meu braço direito no laboratório. Por sua dedicação, comprometimento e ética com o trabalho. Obrigado por sempre escutar as coisas que eu ensinei e também por sempre questionar o que não acha correto. Sei que você terá uma carreira científica brilhante.

A Mari, por ser meu braço esquerdo, sempre pronta para ajudar e a providenciar amostras biológicas com urgência durante minha estadia fora do país.

Ao Hugo Vavá e ao Rodrigo Baptista, por seus ensinamentos sobre a análises e montagem de genomas, assim como pelas excelentes colaborações. Vocês fazem muita falta!

Ao Robson e ao Fernando por toda a ajuda com analises de Bioinformática, principalmente no uso do R.

A Michelle "Mimi" Mattos, por tomar conta de nós e evitar diariamente que o laboratório pegue fogo. Obrigado por seu comprometimento e dedicação, assim como por sempre "cuidar dos seus meninos".

A todos os membros do LIGP, o nosso grupo de pesquisa pelas discussões científicas de excelente qualidade. Vocês fazem a minha "segunda-feira" tão prazerosa quanto um sábado. Obrigado pelo respeito, carinho e também pelos momentos de descontração. Um obrigado

especial aos membros do “*Dani’s Folks*”, o seleto grupo de pessoas loucas que decidiram trabalhar com famílias multigênicas de *T. cruzi*.

A minha querida Turma do pepino, em especial aos membros sempre presentes, Erica, Brunas e Gatti. Fazendo minhas as palavras do Brunão: “Todo mundo tem uma turma marcante, seja da escola ou da universidade. No meu caso, foi a turma do mestrado”.

A Sheila e a todo o pessoal da secretaria do programa de pós-graduação em bioinformática, por sempre me ajudarem com prontidão em todos os parâmetros burocráticos do doutorado. Vocês brilham!

A todos os professores e ao programa de pós-graduação em bioinformática, por todo o aprendizado ao longo destes quatro anos de doutorado.

A Santuza e ao “pessoal da Santuza” que me receberam em seu laboratório durante meu “exílio” devido a obras em nosso laboratório.

Aos professores Lúcia Galvão, Egler Chiari, Bjorn Anderson e Ana Tereza Vasconcelos por cederam amostras ou realizar sequenciamentos genômicos necessários para o desenvolvimento deste trabalho.

Aos membros da banca. Vocês foram os primeiros nomes escolhidos para a discussão deste trabalho. Fiquei muito feliz e honrado com o aceite de todos.

Às agências de fomento e instituições que permitiram que este trabalho fosse realizado: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Organização Mundial de Saúde (OMS).

Á minha família, que me apoiou na minha escolha de ser um cientista e sempre providenciaram para mim todos as condições necessárias para chegar onde eu cheguei. Esta conquista é também de vocês. Em especial, quero agradecer a minha mãe Patrícia Amélia, meu Pai João Bosco e meu Irmão Thiago.

Aos meus amigos, próximos ou distantes, sei que estão sempre comigo.

E, finalmente, a Ana “Baguncinha” Rita, minha esposa e companheira de vida. Muito obrigado pela paciência, carinho, cuidado e principalmente pelo amor.

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

A	Amastigota
AS	Ácido-sialico
Au	*approximately unbiased*
BAC	Cromossomo artificial de bactéria
BiNGO	*Biological Networks Gene Ontology tool*
BLAST	Basic Local Alignment Search Tool
BLASTn	BLAST entre sequências de nucleotídeos
BLASTp	BLAST entre sequências proteicas
BLASTx	BLAST de tradução de nucleotídeos contra banco de proteínas
BLAT	*BLAST-like alignment tool*
CCNV	Variação no número de cópias cromossômicas
CDS	Sequência codificadora
CNV	Variação no número de cópias
CZAR	Retrotransposon associado a cruzi
DALYs	Anos de vida perdidos ajustados por incapacidade
DB	*Data base*
DGCs	Clusters gênicos direcionais
DGF-1	*Disperse Gene Familly* 1
DIRE	Elemento degenerado relacionado a ingi/L1Tc
DNA	Ácido Desoxirribonucleico
DTU	*Discrete typing units*
ESAG	Gene associado a sítios de expressão
FISH	Hibridização fluorescente *in situ*
GATK	*Genome Analysis Tool Kit*
gDNA	DNA genômico
GFF	*Generic feature format*
GPI	*Glicosilfosfatidilinositol*
GTR	*Generalized Time Reversible*
HV	Região Hipervariável
kDNA	DNA do Kinetoplasto
L1Tc	Longa e autônoma repetição terminal
LIT	*Liver Infusion Tryptose*
LNCC	Laboratório Nacional de Computação Científica
MASP	Proteína de Superfície Associada a Mucina
MB	Mega Base
MCL	*Markov clustering algorithm*
MEME	*Multiple EM for Motif Elicitation*
MMR	*Missmatch repair*
mRNAs	RNA mensageiro
MSP	*Major Surface Protease*
MUSCLE	*MUltiple Sequence Comparison by Log-Expectation*
NARTc	Pequeno e não autônoma repetição terminal
NCBI	*National Center for Biotechnology Information*
NGS	Sequenciamento de nova geração

Nr	Não redundante
ORF	Janela Aberta de Leitura
Pb	Pares de base
PBS	Tampão fosfato salino
PCA	Análise de componente principal
PCR	Reação em cadeia da polimerase
PDNF	Fator neurotrófico derivado do parasito
PTU	Unidade de transcrição policistrônica
RDC	Profundidade de cobertura reads
RHS	*Retro-Transposon Hotspot proteins*
RNA	Ácido Ribonucleico
SAPA	*Shed acute phase antigen*
SCoPE	*Single Copy Ploidy Estimation*
SIRE	Pequeno elemento repetitivo intercalado
SMC-3	*Structural maintenance of chromosomes protein 3*
SNP	Polimorfismo de nucleotídeo único
SP	*Peptídeo sinal*
SRA	Arquivo de Reads Sequenciamento
T	Tripomastigota
TcMUC	Mucinas de *T. cruzi*
Thr-rich	Regiões ricas em treonina
Tritryps	*Trypanosoma cruzi*, *Leishmania major* e *T. brucei*
TS	Trans-sialidases
TSSA	*Trypomastigote small surface antigen*
UDP-GlcNAc	UDP-Glicosil-N-acetil transferase
UGCDFA	Unidade de Genômica Computacional Darcy Fontoura de Almeida
UTR	Região não traduzida
VCF	*Variant Call Format*
VIPER	Retroelemento interposto vestigial
VSG	Glicoproteínas variantes de superfície
WCPE	*Whole Chromosome Ploidy Estimation*
WGS	Sequenciamento genômico completo
YAC	Cromossomo artificial de leveduras

Resumo

Trypanosoma cruzi é o agente etiológico da doença de Chagas, uma doença crônica que afeta ~5-8 milhões de pessoas em todo o mundo. Este parasito apresenta uma grande variabilidade genômica, com aneuploidias e uma massiva expansão de famílias multigênicas repetitivas que codificam proteínas de superfície relacionadas a processos de interação parasito-hospedeiro. Apesar de aneuploidias serem usualmente associadas a fenótipos debilitantes em eucariotos multicelulares, dados recentes revelam que elas podem aumentar o fitness de eucariotos unicelulares em situação de estresse e promover resistência a drogas. Apesar de estudos apontarem instabilidade cariotípica em *T. cruzi*, a extensão da variação no número de cópias cromossômicas (CCNV) neste parasito ainda não foi elucidada. Para identificar CCNV em *T. cruzi*, nós sequenciamos genomas de cepas mantidas em longos períodos em cultivo celular e cepas recentemente isoladas de pacientes, pertencentes aos subgrupos relacionados a infecção humana e estimamos a sua ploidia. O padrão de CCNV varia entre e dentro dos subgrupos de *T. cruzi*, mas aparenta ser constante dentro de uma mesma população. O cromossomo 31 foi o único consistentemente supranumerário em todas as cepas de *T. cruzi* avaliadas, sendo enriquecido com genes cujo produto proteico está relacionado a processos de glicosilação. Entre eles, destaca-se a enzima UDP-dependente-glicosil-n-acetil-transferase, envolvida nos passos iniciais da glicosilação de mucinas, que recobrem a superfície do parasito e estão relacionadas a processos de invasão celular e escape do sistema imune. Além de aneuploidias, *T. cruzi* também utiliza a expansão de famílias multigênicas para gerar variabilidade de sequências e promover adaptação a novos ambientes. Apesar de altamente polimórficas, estas famílias apresentam motivos conservados entre membros distintos, resultando em uma estrutura em mosaico que favorece a geração de variabilidade através do rearranjo de blocos definidos, por recombinação. Nós investigamos a abundancia destes motivos entre cepas de *T. cruzi*, provendo considerações sobre a evolução do parasito e abrindo novos caminhos para potenciais alvos vacinais e marcadores de diagnóstico para a doença de Chagas.

Abstract

Trypanosoma cruzi is the etiologic agent of Chagas disease, a chronic illness that affects ~5-8 million people worldwide. This parasite has extreme genotypic and phenotypic variability, with several aneuploidies and a massive expansion of repetitive multigene families enrolled in host-parasite interactions. Although aneuploidies are usually associated with debilitating phenotypes in superior eukaryotes, recent data showed that it could also provide increased fitness in stress conditions and generate drug resistance in unicellular eukaryotes. Even though studies point toward karyotype variability in *T. cruzi*, the extent of chromosomal copy number variation (CCNV) in this parasite has not been determined yet. To identify CCNV in *T. cruzi*, we sequenced the genomes of lab-derived and field-isolated strains from subgroups usually associated to human infections, and estimated the ploidy of each chromosome based on read depth coverage. We have shown that the pattern of CCNV varies among and within *T. cruzi* subgroups, but seems stable inside a given population. Chromosome 31, the only supernumerary chromosome in all *T. cruzi* samples, is enriched with genes related to glycosylation pathways, such as the enzyme UDP-GlcNAc-dependent glycosyltransferase involved in the initial steps of mucins glycosylation, which is a protein that covers the entire parasites surface and is enrolled in the adhesion and cellular invasion of mammalian cells. Besides aneuploidies, *T. cruzi* also relies on the expansion of multigene families to generate variability and to adapt to new environments. Although these families are highly polymorphic, they also present motifs shared among distinct members, resulting in a mosaic structure that favors the generation of sequence variability by rearrangement of defined blocks, through recombination. We have estimated the relative abundance of these conserved motifs among *T. cruzi* strains, providing insights into the evolution of these gene families and opening new avenues for identifying new potential vaccine and diagnosis targets for Chagas disease.

1. Introdução

1.1 *Trypanosoma cruzi* e Doença de Chagas

A doença de Chagas, também denominada Tripanossomíase americana, tem como agente etiológico o protozoário *Trypanosoma cruzi*, classificado na ordem Kinetoplastida e família Trypanosomatidae, a qual pertencem também outros parasitos de interesse médico como o *Trypanosoma brucei* e as várias espécies do gênero *Leishmania*. Esta doença foi descrita pelo mineiro Carlos Chagas no ano de 1909, enquanto ele investigava supostos casos de malária em trabalhadores situados na cidade de Lassance, Minas Gerais (CHAGAS, 1909). Chagas identificou não só o agente etiológico da doença e o nomeou *Schizotrypanum cruzi*, posteriormente reclassificado como *Trypanosoma cruzi*, como também caracterizou as manifestações clínicas, descreveu insetos vetores, o modo de transmissão e reservatórios naturais do parasito (CHAGAS, 1909; LEWINSOHN, 1981).

A região endêmica para a doença de Chagas coincide com a distribuição de seus insetos vetores, englobando desde o sul da Argentina até o sul dos Estados Unidos. Estima-se que existam 8-10 milhões de pessoas infectadas com a doença e um montante de 90 milhões em áreas de risco (COURA, 2009; WHO, 2012). Apesar de não apresentar alta letalidade, a doença de Chagas apresenta grande morbidade, o que resulta em aproximadamente 586.000 anos de vida perdidos ajustados por incapacidade (DALYs) (MATHERS, 2006; WHO, 2012). Com a erradicação da transmissão vetorial pelo *Triatoma infestans* no Brasil, a via transfusional se tornou uma das mais importantes formas de transmissão desta parasitose até a implantação de triagem para esta infecção em bancos de sangue no país (COURA; BORGES-PEREIRA, 2012; COURA; DIAS, 2009). Além da via transfusional, surtos de transmissão oral do parasito decorrente da ingestão de alimentos contaminados com triatomíneos infectados com *T. cruzi* já foram notificados no Brasil, México, Bolívia, Colômbia e Venezuela (COURA et al., 2002; COURA; BORGES-PEREIRA, 2012; COURA; DIAS, 2009; DÍAZ-BELLO et al., 2014; YOSHIDA, 2009). Nos casos de transmissão oral do parasito, a variabilidade em moléculas de superfície já foi descrita como um fator relacionado à eficiência de infecção, onde a presença das glicoproteínas de superfície gp82 e gp30, assim como da variante de gp90 susceptível a digestão por pepsina agravam o quadro clínico da doença (BARRETO-DE-ALBUQUERQUE et al., 2015; YOSHIDA, 2008). Devido à migração de pacientes chagásicos para países não endêmicos somada a uma tardia implantação de triagens para esta moléstia em bancos de sangue, novos casos desta doença vêm sendo reportados em países da América do Norte, Europa e Oceania, se tornando um relevante problema de saúde também nestas localidades (Figura 1) (ANGHEBEN et al., 2015;

BENJAMIN et al., 2012; BERN et al., 2011; BERN; MONTGOMERY, 2009; JACKSON, YVES; PINTO, ANGIE; PETT, 2014). Esta transmissão da doença em regiões não endêmicas reforça a importância de vias não vetoriais de transmissão, como as vias transfusional, transplacentária (congênita) e transplantes de órgãos.

Figura 1: Distribuição mundial da doença de Chagas. Retirado de Ribeiro 2012 (RIBEIRO et al., 2012).

Em seu ciclo de vida, o parasito *T. cruzi* alterna entre barbeiros da família Reduviidae e uma grande variedade de hospedeiros mamíferos, que incluem o homem (COURA; DIAS, 2009). Esta alternância entre hospedeiros submete o parasito a diferentes pressões seletivas, que, somada a sua reprodução predominantemente clonal, resultam em extensa diversidade genética intraespecífica.

O ciclo de vida do parasito começa quando um triatomíneo susceptível realiza repasto sanguíneo em um hospedeiro mamífero infectado, ingerindo formas tripomastigotas sanguíneas localizadas na corrente sanguínea periférica. Tripomastigotas se diferenciam em formas replicativas extracelulares denominadas epimastigotas no intestino médio do vetor, multiplicando por divisão binária. Durante o trânsito para a ampola retal do inseto, estímulos como diminuição de pH e escassez de nutrientes, levam a diferenciação do parasito em tripomastigota metacíclico, que é liberado junto com as fezes e urina após o repasto sanguíneo. As formas tripomastigotas metacíclicas infectam o hospedeiro mamífero ao entrar em contato com mucosas ou regiões de descontinuidade do epitélio, podendo invadir qualquer célula nucleada. A invasão celular ocorre através da formação de um vacúolo parasitóforo, pelo

recrutamento de lisossomos para a membrana da célula hospedeira dependente de liberação intracelular de Ca^{2+} (ANDRADE; ANDREWS, 2005; BURLEIGH; ANDREWS, 1995; RODRÍGUEZ et al., 1996; TARDIEUX et al., 1992), ou por penetração ativa seguida de posterior fusão do vacúolo com lisossomos (WOOLSEY; BURLEIGH, 2004). A acidificação do vacúolo parasitóforo decorrente da fusão com lisossomos induz a expressão pelas formas tripomastigotas de uma proteína formadora de poro, denominada TcTOX, que leva a destruição do vacúolo parasitóforo e escape do parasito para o citoplasma, onde se diferencia em formas replicativas com o flagelo interiorizado, denominadas amastigotas. Após sucessivos ciclos de divisão, a forma amastigota se diferencia em tripomastigota sanguíneo, que é liberado com a lise da célula hospedeira, podendo levar a invasão de novas células no local ou escape para a circulação sanguínea e dispersão para outros órgãos. Parasitos circulantes podem ser ingeridos pelos insetos vetores durante o repasto sanguíneo, fechando o ciclo (Figura 2) (MACEDO; OLIVEIRA; PENA, 2002).

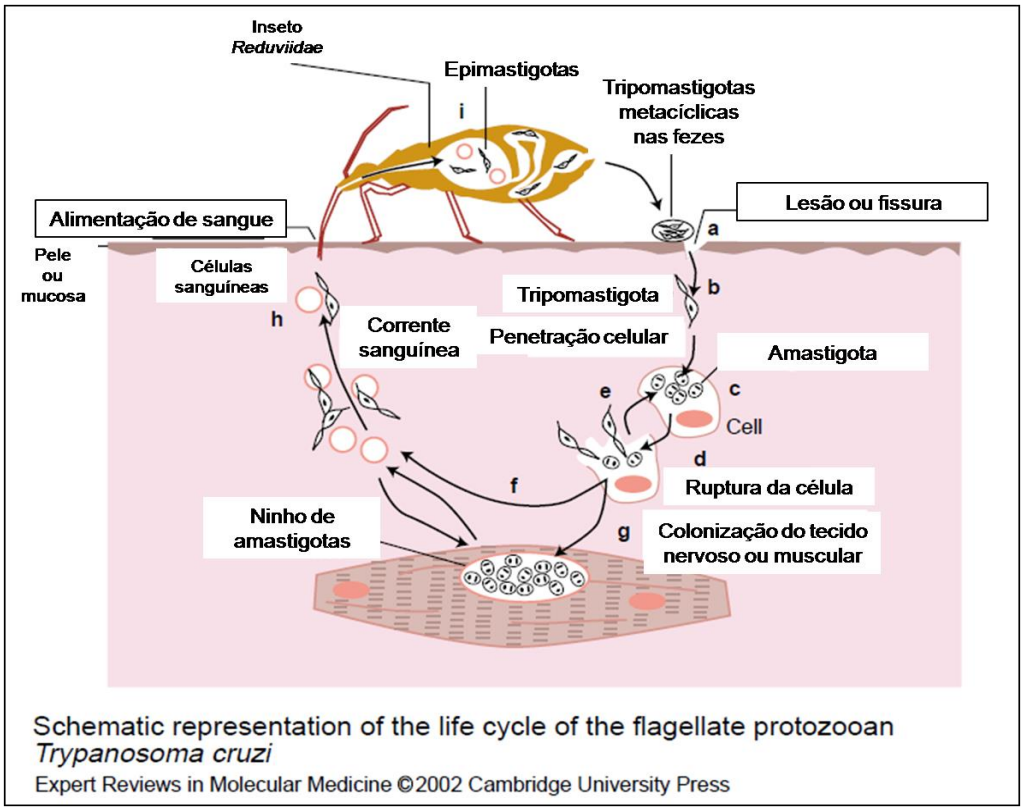


Figura 2: Ciclo de vida do *T. cruzi*. Retirado e modificado de Macedo 2002 (MACEDO; OLIVEIRA; PENA, 2002).

1.2 Sintomatologia clínica e fases da doença

A sintomatologia da doença de Chagas apresenta duas fases características, denominadas “fase aguda” e “fase crônica”. A fase aguda ocorre durante os primeiros 90 dias após a infecção, sendo caracterizada por uma intensa parasitemia e parasitismo tecidual. Esta fase é usualmente assintomática ou apresenta sintomas inespecíficos como febre, dor muscular, dor nas juntas e distúrbios respiratórios, o que dificulta o diagnóstico e leva a uma subnotificação desta moléstia. Em alguns casos, a fase aguda da doença pode apresentar sintomas específicos como o sinal de Romanã, um edema bipalpebral unilateral, ou o chagoma de inoculação, uma intensa reação inflamatória no local de penetração do parasito (JUNQUEIRA et al., 2010; TEIXEIRA; NASCIMENTO; STURM, 2006). Após 12 semanas, com a redução do número de parasitos circulantes mediada por uma resposta imune eficiente pelo hospedeiro, tem-se início a fase crônica indeterminada, que é assintomática e pode perdurar por toda a vida do paciente. Em aproximadamente 30% dos pacientes, o quadro indeterminado evolui para manifestações cardíacas, digestivas ou mistas, que podem levar a morte por insuficiência cardíaca, obstrução do esôfago ou cólon (JUNQUEIRA et al., 2010; RASSI, 2010; TEIXEIRA; NASCIMENTO; STURM, 2006; ZINGALES et al., 2009, 2012) (Figura 3).

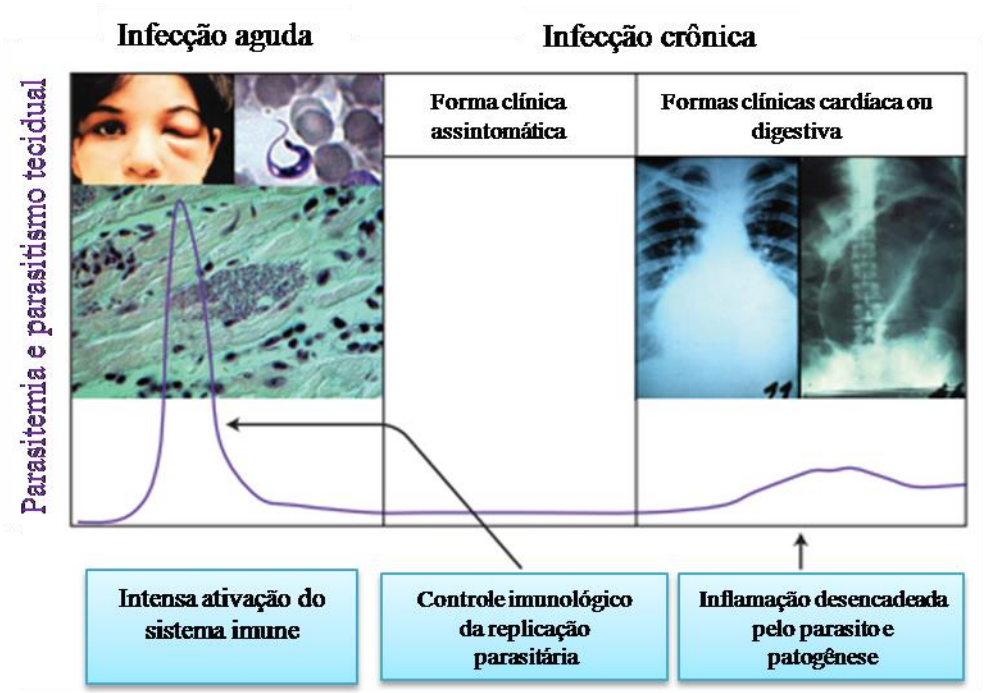


Figura 3: Fases clínicas da doença de Chagas. Retirado e modificado de (JUNQUEIRA et al., 2010).

1.3 Variabilidade genética e classificação

O táxon *T. cruzi* é atualmente dividido em 6 subgrupos ou "*Discrete typing units*" (DTU's), denominados *T. cruzi* I–VI (ZINGALES et al., 2009, 2012). Esta separação visa uniformizar as diferentes classificações previamente propostas, como os Zimodemas (MILES et al., 1977), linhagens e os antigos DTUs (BARNABÉ; BRISSE; TIBAYRENC, 2000; BRISSE; BARNABE; TIBAYRENC, 2000; BRISSE; DUJARDIN; TIBAYRENC, 2000; TIBAYRENC, 2003). A divisão em subgrupos é necessária devido a grande variabilidade genética apresentada por este parasito, que confere diferentes características fenotípicas como virulência, tropismo tecidual, resistência a drogas e adaptação a vetores (ANDRADE et al., 2010; ANDRADE; MAGALHÃES, 1996; CAETANO et al., 2010; DE LANA et al., 1998; DE ORNELAS TOLEDO et al., 2003; STURM et al., 2003; STURM; CAMPBELL, 2010a). Atualmente, três distintos modelos evolutivos baseados em marcadores mitocondriais e nucleares foram propostos para explicar a evolução intra-específica de *T. cruzi*: o "modelo de duas hibridizações" (WESTENBERGER et al., 2005), o "modelo de três ancestrais" (DE FREITAS et al., 2006) e, mais recentemente, o "modelo de dois ancestrais" (BURGOS et al., 2013) (Figura 4). No "modelo de duas hibridizações", um evento de hibridização entre os DTUs TcI e TcII originou TcIII e TcIV, enquanto um segundo evento entre TcII e a TcIII originou TcV e TcVI (Figura 4a) (WESTENBERGER et al., 2005). No "modelo de três ancestrais", TcI, TcII e TcIII são DTUs ancestrais, e eventos de hibridização, entre TcII e TcIII originaram TcV e TcVI (Figura 4b) (DE FREITAS et al., 2006). Finalmente, no "modelo de dois ancestrais", apenas TcI e TcII são DTUs ancestrais, TcI origina TcIII e TcIV e, novamente, uma hibridização entre TcII e TcIII originou TcV e TcVI (Figura 4c) (BURGOS et al., 2013). Apesar de discordarem na historia evolutiva, os três modelos concordam que os DTUs TcV e TcVI são híbridos, derivados de populações ancestrais de TcII e TcIII.

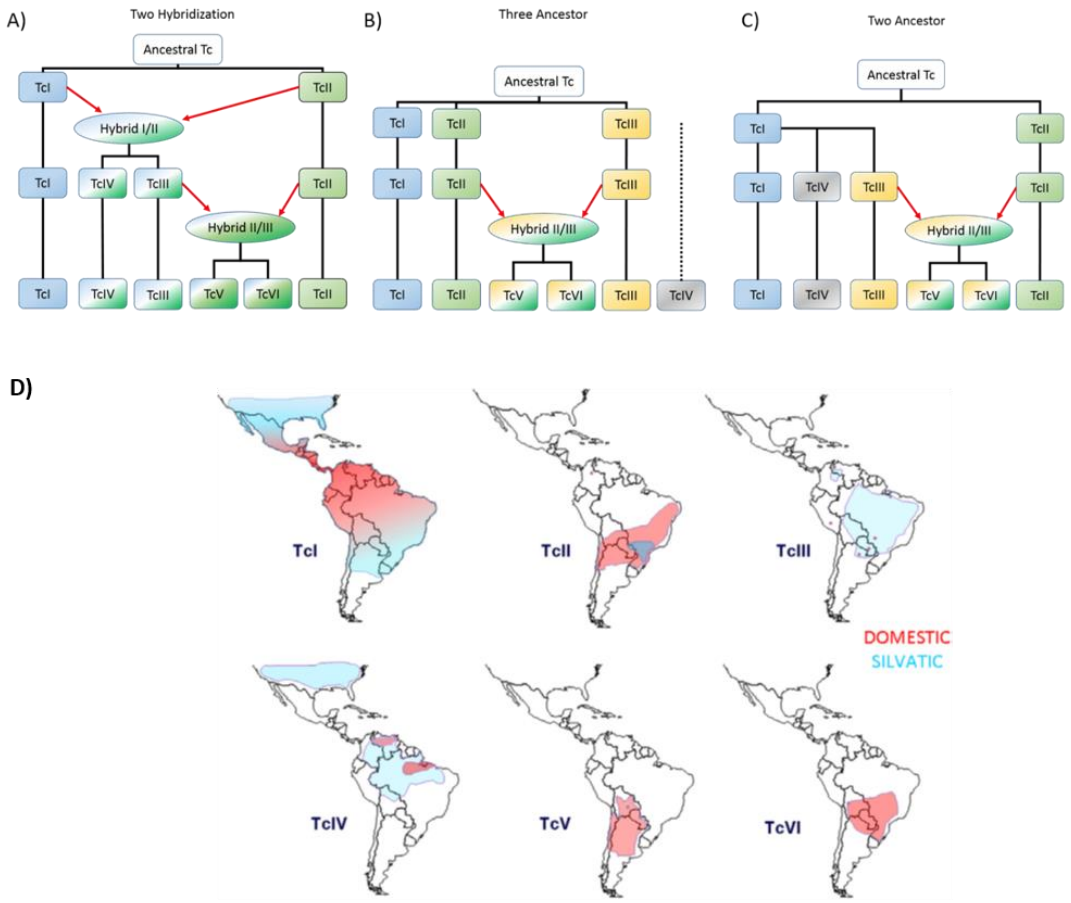


Figura 4: Modelos de evolução e distribuição geográfica dos DTUs do parasito *T. cruzi*. Retirado de Reis-Cunha 2016 e (ZINGALES et al., 2012).

O subgrupo TcI é o agente predominante da doença de Chagas humana na região da Amazônia e em países ao Norte da América do Sul, como a Venezuela, Colômbia e Peru, onde esta DTU está relacionada ao ciclo domiciliar ou peridomiciliar (MILES et al., 2009). Em contrapartida, na região central e Sul do Brasil, este subgrupo está diretamente relacionado ao ciclo silvestre (MILES et al., 2009; ZINGALES et al., 2012). Infecções crônicas sintomáticas por cepas pertencentes a este subgrupo estão usualmente relacionadas à forma cardíaca da doença de Chagas, sendo raros casos apresentando comprometimento do sistema digestório (MILES et al., 2009). Cepas pertencentes a este subgrupo apresentam uma menor variabilidade intra-genômica, mas apresentam consideravel variação de sequência entre cepas (CERQUEIRA et al., 2008). Alguns fatores que podem estar associados a uma menor variabilidade em TcI são a melhor eficiência no sistema de reparo de erro de pareamento (*Miss match repair*) neste DTU quando comparada a cepas da DTU TcII (AUGUSTO-PINTO et al., 2003; MACHADO et al., 2006a), assim como uma menor expansão de famílias multigênicas nesta DTU quando comparado com

DTUs híbridas como o TcVI (CERQUEIRA et al., 2008; FRANZÉN et al., 2011, 2012; VARGAS; PEDROSO; ZINGALES, 2004).

O subgrupo TcII é o mais prevalente em infecções humanas no Brasil e no cone Sul da América do Sul, sendo responsável pela doença de Chagas aguda grave e por manifestações clíncas com acometimento cardíaco, digestivo ou misto (MILES et al., 2009; ZINGALES et al., 2012). Este subgrupo está diretamente relacionado ao ciclo doméstico e peridoméstico da doença de Chagas, sendo menos frequentemente encontrado em ambiente silvestre (MILES et al., 2009; ZINGALES et al., 2012).

Os subgrupos TcIII e TcIV estão usualmente relacionados ao ciclo silvestre do parasito. A distribuição do TcIII vai desde o norte da América do Sul até a Argentina e Venezuela. A distribuição do TcIV é pouco compreendida, mas aparentemente apresenta um dispersão semelhante a TcIII, porém atingindo até a região sul dos Estados Unidos (MILES et al., 2009; ZINGALES et al., 2012). Recentemente, diversos casos e surtos de infecção por TcIII vem sendo reportados (MARTINS et al., 2015; MONTEIRO et al., 2010), reforçando a nescessidade de nova avaliação da real importância desta DTU na infecção humana.

Os dois DTUs confirmadamente híbridos, TcV e TcVI, são responsáveis pela maioria dos casos de doença de Chagas aguda grave no Grande Chaco e países vizinhos, como Bolívia, Chile, Paraguai, norte da Argentina e Sul do Brasil (MILES et al., 2009; ZINGALES et al., 2012). Um representante da DTU VI é o clone CL Brener, utilizado como referência do projeto genoma do parasito (EL-SAYED, 2005a). Estudos recentes descrevem uma distribuição geográfica mais ampla para estas duas DTUs, onde cepas híbridas foram encontradas do sul da América do Sul até a Colômbia, sugerindo uma maior importância para estas DTUs nas infecções humanas do que previamente estimado (MESSENGER et al., 2016).

Um potencial sétimo genótipo restrito a hospedeiros morcegos denominado TcBat, com semelhanças genéticas a TcI foi descrito na América Central e do Sul (LIMA et al., 2015; MARCILI et al., 2009; PINTO et al., 2012). O estudo deste parasito, juntamente com outros Tripanosomatídeos relacionados, como: *T. cruzi marinkellei*, *T. dionisii*, *T. rangeli*, *T. livingstonei* e *T. lewisi* podem auxiliar na compreensão da evolução do parasitismo neste grupo de eucariotos (COTTONTAIL et al., 2014; FRANZÉN et al., 2012; LIMA et al., 2012, 2013b, 2015; MAEDA et al., 2012; STOCO et al., 2014)

1.4 Estrutura genômica e expressão gênica

Por fazer parte do clado Excavata, um dos primeiros ramos a divergir dos eucariotos (ADL et al., 2012; HAAG; O'HUIGIN; OVERATH, 1998), os tripanossomatídeos apresentam

algumas características peculiares como: (I) Transcrição policistrônica da maioria de seus genes (CLAYTON, 2016; MARTÍNEZ-CALVILLO et al., 2010); (II) Genes organizados em densos clusters direcionais (EL-SAYED, 2005b); (III) Transcrição de genes codificadores de proteínas pela RNA polimerase I (CLAYTON, 2016; GÜNZL et al., 2003; LEE; VAN DER PLOEG, 1997; ZOMERDIJK; KIEFT; BORST, 1991); (IV) Trans-splicing de RNAs policistrônicos para produzir moléculas maduras de mRNA (LEBOWITZ et al., 1993; LIANG; HARITAN; ULIEL, 2003); (V) Regulação da expressão gênica principalmente através de mecanismos pós-transcricionais (CLAYTON, 2013; CLAYTON; SHAPIRA, 2007; TEIXEIRA et al., 2012); (VI) Edição de RNA mitocondrial para produzir mRNAs funcionais (APHASIZHEVA; APHASIZHEV, 2016; FEAGIN, 1990; LANDWEBER, 1992; STUART, 1991) e (VII) tolerância a aneuploidias (DOWNING et al., 2011; MINNING et al., 2011; ROGERS et al., 2011). Os genes em tripanossomatídeos são organizados em grandes clusters gênicos direcionais (DGCs), que compreendem de poucos a centenas de genes não sobrepostos, codificados na mesma fita de DNA e com a mesma orientação, separados por regiões de troca de fita codificadora (IVENS, 2005; TEIXEIRA et al., 2012). Estes DGCs são transcritos de forma policistrônica, de modo similar ao que ocorrem em operons de procariotos, porém são rapidamente processados no núcleo, gerando mRNAs monocistrônicos maduros através dos processos de trans-splicing e poliadenilação. O trans-splicing leva a adição de uma sequência espécie específica de 39 nucleotídeos, denominado miniexon ou *spliced leader*, a região 5'. Concomitantemente, ocorre a adição da cauda poli-A na região 3' de cada gene transcrito, permitindo a individualização do transcrito em unidades monocistrônicas e regulação diferencial de transcritos inicialmente presentes no mesmo policistron (GÜNZL, 2010; PREUSSE; JAÉ; BINDEREIF, 2012) (Figura 5). Genes derivados de um mesmo policistron podem não possuir funções biológicas relacionadas e apresentar níveis de mRNA bem discordantes. Estes processos reforçam a importância de mecanismos de regulação pós-transcricional para o controle da expressão gênica, que incluem o processamento e estabilidade de mRNA, eficiência de tradução e estabilidade do produto proteico (BRINGAUD et al., 2007; CLAYTON; SHAPIRA, 2007; MCNICOLL et al., 2005; MÜLLER et al., 2010).

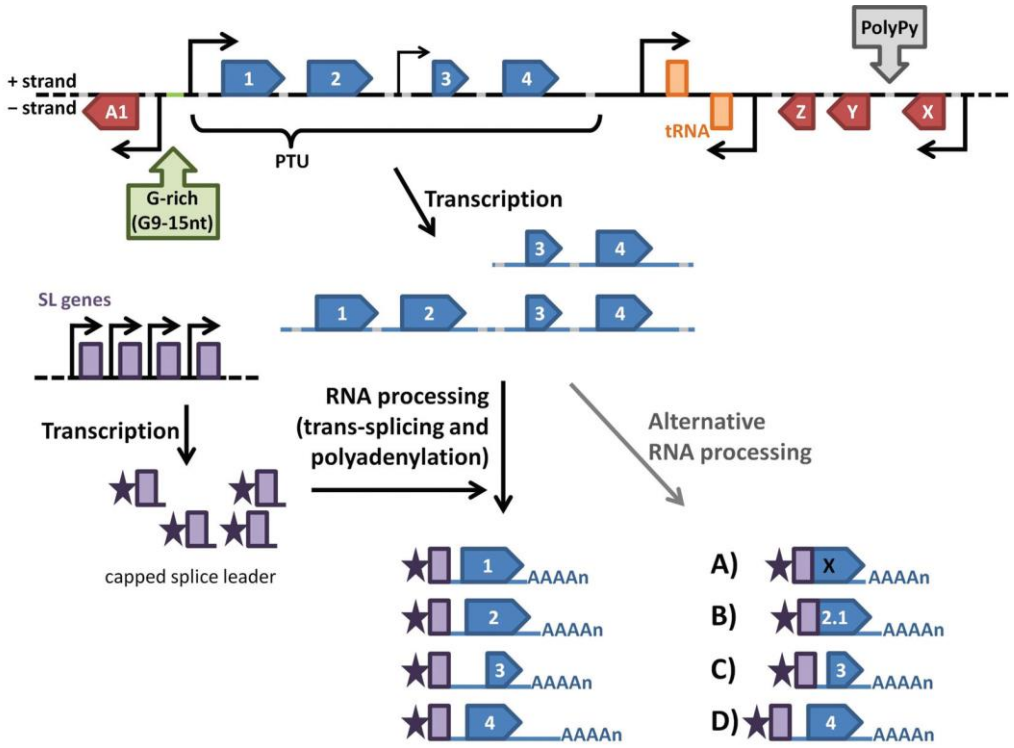


Figura 5: Clusters gênicos direcionais e expressão gênica em Tripanossomatídeos. Clusters direcionais de genes não relacionados estão organizados em unidades de transcrição policistrônica (PTU), representados por setas azuis quando localizados na fita + e em vermelho quando localizados na fita - . A região de início de transcrição geralmente está localizada *upstream* do primeiro gene da PTU. Regiões de troca de fita estão usualmente localizadas em clusters de RNAs transportadores. O *policistron* representado pelas setas azuis numeradas de 1 a 4 é individualizado em RNAs monocistrônicos após a adição de um *spliced leader* capeado através do processo de trans-splicing, o qual é acoplado à poliadenilação. RNAs policistrônicos podem sofrer processamento alternativo, gerando RNAs monocistrônicos com redução/aumento de UTRs ou mudança do sítio de iniciação da tradução , através da mudança do sítio de processamento e adição do *spliced leader* ou da cauda poli-A. Retirado de Teixeira 2012 (TEIXEIRA et al., 2012).

Outra peculiaridade do genoma destes parasitos é a dicotomia na distribuição genômica de genes relacionados ao metabolismo basal e genes relacionados a interação com os seus hospedeiros. Quando analisado de uma forma funcional, o genoma dos tripanossomatídeos pode ser dividido em duas regiões: (I)- Regiões contendo genes sintênicos e conservados entre as diferentes espécies, composta por genes relacionados ao metabolismo basal do parasito, e (II)- clusters de genes espécie-específicos que codificam proteínas de superfície diretamente relacionadas a interações parasito-hospedeiro (BARTHOLOMEU et al., 2009; EL-SAYED, 2005b). Estes clusters são compostos por principalmente glicoproteínas variantes de superfície (VSGs) e genes associados a sítios de expressão (ESAGs) em *T. brucei* (BERRIMAN; GHEDIN; HERTZ-FOWLER, 2005), e por trans-sialidases, TcMUCs e Proteínas de Superfície Associadas a Mucinas

(MASPs) em *T. cruzi*, *T. c. marinkelei* e *T. rangeli* (EL-SAYED, 2005a; STOCO et al., 2014; WEATHERLY; BOEHLKE; TARLETON, 2009). A família MASP de *T. cruzi* foi descoberta apenas durante o sequenciamento do genoma de *T. cruzi* (BARTHOLOMEU et al., 2009; EL-SAYED, 2005a), reforçando a importância de estudos genômicos para uma melhor compreensão da biologia dos organismos.

Atualmente, três isolados de *T. cruzi* tiveram o seu genoma sequenciado e montado: o clone CL Brener (TcVI), como um genoma estimado de 55 Mb (EL-SAYED, 2005a) e duas cepas do DTU TcI, Sylvio X10 e Dm28c, com tamanhos de genomas estimados em respectivamente 44Mb e 27 Mb (FRANZÉN et al., 2011, 2012; GRISARD et al., 2014). O genoma do clone da cepa CL (TcVI), CL Brener, foi o primeiro genoma de *T. cruzi* a ser sequenciado, confirmando a natureza híbrida desta DTU (EL-SAYED, 2005a). Comparações pós-montagem dos contigs gerados com reads da cepa Esmeraldo (TcII) permitiram a identificação de dois haplótipos compondo CL Brener: O haplótipo "*Esmeraldo-Like*", derivado de um ancestral TcII; e o haplótipo "*non-Esmeraldo-like*", derivado de um ancestral TcIII (EL-SAYED, 2005a). Estes haplótipos apresentam um alto grau de sintenia, especialmente em clusters gênicos direcionais que codificam genes "*Housekeeping*", relacionados ao metabolismo basal do parasito, chegando a 97,8% de similaridade (EL-SAYED, 2005a; FRANZÉN et al., 2011). Em 2010, o sequenciamento da cepa Sylvio X10 (TcI) revelou que a maioria dos genes deste isolado é também encontrado em CL Brener (EL-SAYED, 2005a; FRANZÉN et al., 2011, 2012). Ademais, Sylvio apresenta uma maior similaridade nucleotídica média com o haplótipo *non-Esmeraldo-like* (98.2%) de CL Brener, do que com o haplótipo *Esmeraldo-like* (97.5%) (FRANZÉN et al., 2011). A similaridade entre as DTUs de *T. cruzi* é menor do que a encontrada entre subespécies do parasito *T. brucei* (99.2%) (JACKSON et al., 2010), e maior do que entre as espécies do gênero *Leishmania* (94%), o que suporta a classificação das DTUs de *T. cruzi* em uma mesma espécie.

A sintenia encontrada entre os haplótipos de CL Brener e também Sylvio X10 é interrompida por clusters compostos por elementos transponíveis e famílias multigênicas que codificam para proteínas de superfície, como: MASPs, Mucinas, Trans-Sialidases, GP63, *Disperse Gene Familly* 1 (DGF-1) e *Retro-Transposon Hotspot proteins* (RHS) (BARTHOLOMEU et al., 2009; BUSCAGLIA et al., 2006; DC-RUBIN; SCHENKMAN, 2012; DE PABLOS; OSUNA, 2012; EL-SAYED, 2005a; SCHENKMAN et al., 1994). Nestas regiões, genes de cada família multigênica não estão organizados em tandem, e sim se alternam com genes de outras famílias em uma disposição não ordenada, constantemente mudando a fita codificadora (BARTHOLOMEU et al., 2009; EL-SAYED, 2005a; WEATHERLY; BOEHLKE; TARLETON, 2009) (Figura 6). Estes clusters podem se estender a até 600 kb e variam em sequência e tamanho entre os haplótipos de CL Brener, assim

como entre as DTUs de *T. cruzi.* Estas regiões estão mais expandidas em CL Brener quando comparado com a cepa Sylvio X10, correspondendo a diferença de tamanho de 5.9 MB entre o tamanho dos genomas destas duas cepas (EL-SAYED, 2005a; FRANZÉN et al., 2011). Esta variação no número de cópias (CNV) - o ganho ou perda de material genômico - entre diferentes cepas de *T. cruzi* pode resultar em impacto fenotípico, alterando o *fitness* do parasito.

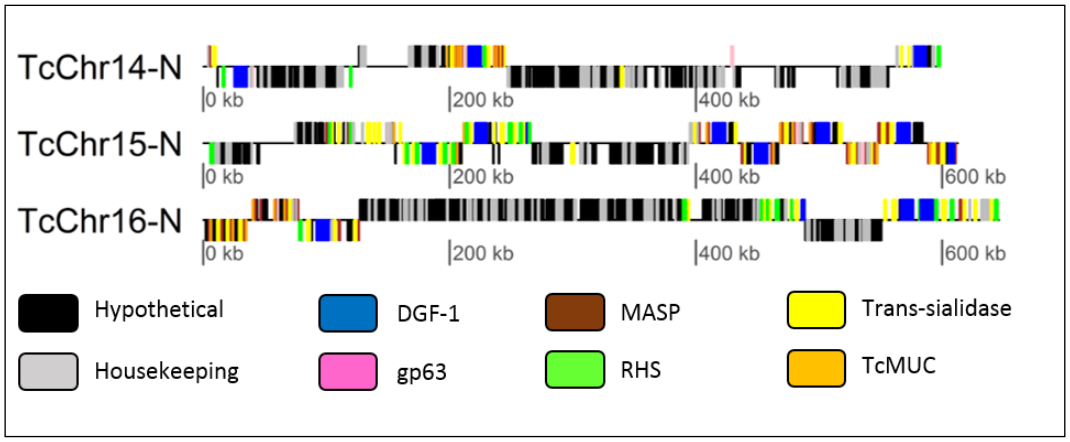


Figura 6: Estrutura cromossômica do haplótipo Non-Esmeraldo de CL Brener. Distribuição gênica dos cromossomos 14, 15 e 16 do haplótipo *non-Esmeraldo-like* de da cepa CL Brener de *T. cruzi*. Famílias multigênicas estão representadas como caixas coloridas, enquanto genes hipotéticos e *housekeeping* são representados por caixas pretas e cinzas respectivamente. Os genes e o cromossomo estão desenhados em proporção a seu tamanho, onde genes acima da linha se encontram na fita codificadora positiva, enquanto genes abaixo estão na fita codificadora negativa (Retirado de Reis-Cunha 2017) (REIS-CUNHA; VALDIVIA; BARTHOLOMEU, 2017).

1.5 Variação no número de cópias gênicas

É bem aceito pela comunidade científica que a variação no número de cópias gênicas pode alterar o fitness de um organismo (DOWNING et al., 2011; HENRICHSEN; CHAIGNAT; REYMOND, 2009; ISKOW; GOKCUMEN; LEE, 2012; JACKSON, 2007a, 2007b; LAFFITTE et al., 2016; MANNAERT et al., 2012; ROGERS et al., 2011; SEBAT et al., 2004). Eventos de CNV podem levar a alterações na expressão gênica e aumento de variabilidade de sequências através da formação de genes parálogos (BARTHOLOMEU et al., 2009; CLAYCOMB; ORR-WEAVER, 2005; EL-SAYED, 2005a; ISKOW; GOKCUMEN; LEE, 2012; JACKSON, 2007a, 2007b; STRANGER et al., 2007). Estes genes parálogos podem apresentar uma acelerada taxa evolutiva (JACKSON, 2007b; JIANG et al., 2007; WANG et al., 2005), resultando em possíveis consequências, como: (I) Neofuncionalização de genes duplicados; (II) Subfuncionalização de genes duplicados com mais de uma função pela segregação de suas funções em dois genes com apenas uma função; e (III) Expressão diferencial

de genes duplicados em tecidos ou estágios de desenvolvimento distintos pela mudança em suas regiões regulatórias (FORCE et al., 1999; ISKOW; GOKCUMEN; LEE, 2012; JACKSON, 2007b; ROTH et al., 2007; TÜMPEL et al., 2006; ZHANG, 2000). Além disso, eventos de CNV são especialmente interessantes em tripanossomatídeos, visto que transcrevem todos os seus genes, regulando a expressão gênica por mecanismos pós-transcricionais e pós-traducionais, através de estabilidade/instabilidade de moléculas de RNA-mensageiro (mRNA), distribuição dos transcritos em diferentes compartimentos celulares, alterações em taxas de tradução e meia- vida das proteínas (CLAYTON, 2002; MARTÍNEZ-CALVILLO et al., 2010).

As pressões seletivas decorrentes da adaptação e alternância entre hospedeiros vertebrados e invertebrados resultaram em um série de duplicações gênicas e alta plasticidade genômica em tripanossomatídeos (JACKSON, 2007a, 2007b; JACKSON et al., 2016; VALDIVIA et al., 2015a). Esta plasticidade genômica em resposta a mudanças ambientais pode afetar cromossomos inteiros, gerando aneuploidias, ou ser restrita a regiões genômicas específicas (JACKSON, 2007b; LAFFITTE et al., 2016). Este fenômeno também ocorre em fungos e células cancerosas, onde mudanças no número de cópias pode alterar a virulência, proliferação celular e susceptibilidade a drogas (GORDON; RESIO; PELLMAN, 2012; PFAU; AMON, 2012; SHELTZER et al., 2011). Entre os genes provavelmente perdidos durante a evolução dos tripanossomatídeos estão diversas enzimas envolvidas na degradação de polissacarídeos, como β-glicosidases e glicoamilases, requeridas para o processamento de organismos de origem bacteriana (JACKSON et al., 2016). Com relação ao ganho de função, a adaptação à vida parasitária levou a evolução de genes que permitem a sobrevivência do protozoário dentro de seu hospedeiro, levando a formação de diversas famílias multigênicas relacionadas a interações especificas de cada parasito com seus hospedeiros (JACKSON et al., 2016).

T. cruzi apresenta a maior expansão de famílias multigênicas entre os tripanossomatídeos, o que resulta em seu maior tamanho genômico ~55MB, quando comparado aos genomas de *Leishmania major* ~33MB e *T. brucei* ~25MB (EL-SAYED, 2005b). Na montagem atual do genoma da cepa CL Brener de *T. cruzi*, estes clusters de famílias multigênicas estão localizados tanto em regiões sub-teloméricas quanto internas do cromossomo (EL-SAYED, 2005a; WEATHERLY; BOEHLKE; TARLETON, 2009), diferente do que é encontrado em *T. brucei* onde os clusters de VSGs e ESAGs estão localizados em regiões sub-teloméricas dos megacromossomos (BERRIMAN; GHEDIN; HERTZ-FOWLER, 2005; EL-SAYED, 2005b). Como muitas destas famílias codificam proteínas de superfície, como MASP, TcMUC e trans-sialidases, a sua expansão em *T. cruzi* pode ser uma consequência da maior variabilidade de hospedeiros mamíferos que este parasito pode infectar, assim como da sua habilidade única entre os

tripanossomatídeos de infectar qualquer célula nucleada (ANDRADE; ANDREWS, 2005; BURLEIGH; ANDREWS, 1995; EPTING; COATES; ENGMAN, 2010; FERNANDES; ANDREWS, 2012). Este vasto número de hospedeiros/células expõe o parasito a diferentes pressões seletivas, o que pode contribuir para a diversificação destas sequências (FLORES-LOPEZ; MACHADO, 2015). De fato, uma significante fração de genes codificadores de proteínas em *T. cruzi* apresenta evidências de estar submetido à seleção positiva para diversificação quando comparado a genes de *Leishmania* (FLORES-LOPEZ; MACHADO, 2015).

1.6 Famílias Multigênicas de *T. cruzi*

Os clusters de famílias multigênicas em *T. cruzi* são compostos principalmente por MASPs, TcMUCs, trans-sialidases, GP63s, *Dispersed Gene Family*-1 (DGF-1), *Retrotransposon Hotspot proteins* (RHS) e elementos transponíveis, como L1Tc, NARTc, SIRE, DIRE, CZAR e VIPER (BARTHOLOMEU et al., 2009; BRINGAUD et al., 2006; EL-SAYED, 2005a; GHEDIN et al., 2004; LORENZI; ROBLEDO; LEVIN, 2006; WEATHERLY; BOEHLKE; TARLETON, 2009; WICKSTEAD; ERSFELD; GULL, 2003). As proteínas codificadas por estes genes são usualmente aderidas à superfície do parasito por âncoras de glicosilfosfatidilinositol (GPI), podendo ser liberadas no meio através da clivagem do GPI pela enzima fosfatidilinositol fosfolipase C (BARTHOLOMEU et al., 2014; BUSCAGLIA et al., 2006; DE PABLOS; OSUNA, 2012; ROMANO et al., 2012). As funções das principais famílias multigênicas de *T. cruzi* estão descritas abaixo.

Trans-sialidases e sialidases: O genoma de CL Brener contém aproximadamente 1400 genes que codificam para membros da família das Trans-sialidases, divididos em 8 subgrupos com base em similaridade de sequência proteica (EL-SAYED, 2005a; FREITAS et al., 2011). Trans-sialidases catalisam a transferência do ácido sialico de glicoconjugados do hospedeiro para resíduos de β-galactopiranose localizados em proteínas mucinas do parasito, gerando uma superfície carregada negativamente que fornece proteção contra a via alternativa do complemento e opsonização por anticorpos (BUSCAGLIA et al., 2006; DC-RUBIN; SCHENKMAN, 2012; FREIRE-DE-LIMA et al., 2012; SCHENKMAN et al., 1994). Esta família também está envolvida em processos de adesão e invasão de células epiteliais (BUTLER et al., 2013), neurais e células da glia (DE MELO-JORGE; PEREIRAPERRIN, 2007), promovendo resistência a apoptose e facilitando a reparação neural ao se ligarem a receptores TrK em células neuronais no hospedeiro mamífero. Por este motivo ela é também nomeada de fator neurotrófico derivado do parasito (PDNF) (CHUENKOVA; PEREIRAPERRIN, 2011, 2009). No inseto vetor, proteínas desta família expressas na forma epimastigota promovem a adesão/liberação a células endoteliais do intestino do barbeiro, assim como protegem o parasito da ação de enzimas glicolíticas (DC-RUBIN;

SCHENKMAN, 2012) (Figura 7). Proteínas da família das trans-sialidases também estão relacionadas a uma modulação do ciclo de vida do parasito, visto que tripomastigotas derivadas de infecção em células de mamífero expressam e liberam uma maior quantidade destas proteínas do que metacíclicas axênicas, e a super-expressão de trans-sialidases leva a um escape prévio dos parasitos do vacúolo parasitóforo para o citoplasma da célula hospedeira (RUBIN-DE-CELIS et al., 2006). A superfície sializada de *T. cruzi* promove um efeito imunossupressor em células dendríticas, ao interagir com a lectina *Siglec-E* localizada na superfície destas células, inibindo a produção de IL-12, uma das principais citocinas responsáveis por mediar a resposta imune protetora contra o parasito (ERDMANN et al., 2009). Membros da família das trans-sialidases encontradas no parasito *T. rangeli*, apatogênico para hospedeiro mamífero, não possuem a ação de *Trans*, tendo apenas a atividade de sialidase (AMAYA et al., 2003). Estas proteínas de *T. rangeli*, assim como as expressas na forma epimastigota de *T. cruzi*, também não apresentam a repetição C-terminal de 12 aminoácidos denominada *shed acute phase antigen* (SAPA), necessária para a oligomerização e estabilidade da enzima no parasito (FRASCH, 1994; STOCO et al., 2014).

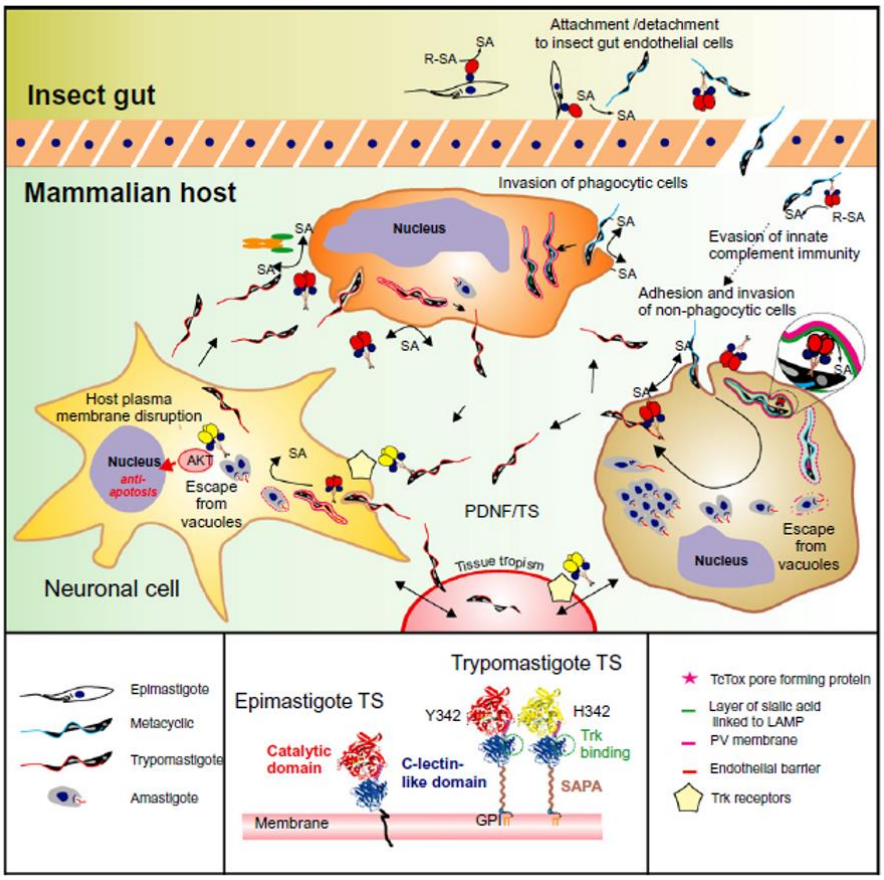


Figura 7: Principais funções da família trans-sialidase (Retirado de DC-Rubin 2012) (DC-RUBIN; SCHENKMAN, 2012) – Esta figura denota as principais funções da família trans-sialidase ao longo

do ciclo de vida do *T. cruzi*. O painel superior mostra o envolvimento das trans-sialidases com a interação da parede endotelial do intestino do barbeiro vetor. O painel central denota as diferentes interações de proteínas desta família com células de mamíferos, onde as setas seguidas por R-SA ou -SA para SA representam ações de transferase de ácido sialico ou sialidase mediadas pela enzima. A figura inferior esquerda representa a legenda das ilustrações das formas evolutivas do parasito *T. cruzi*, enquanto a figura inferior direita contém os fatores chaves envolvidos na interação parasito-hospedeiro, representadas na figura central. A figura central inferior possui a representação da estrutura de trans-sialidases das formas epimastigota e tripomastigota de *T. cruzi*, com o sítio catalítico representado em vermelho para a variante ativa ou amarela para a inativa. Neste mesmo painel, o domínio de ligação a lecitina C está representado em azul, as repetições SAPA em marrom e o domínio de ligação a Trk em verde. (TS) Trans-sialidases; (AS) Ácido-sialico; (PDNF) Fator neutotrófico derivado do parasito;

Mucinas: Mucinas são as principais glicoproteínas encontradas na superfície do *T. cruzi*, recobrindo a face externa da membrana do parasito (BUSCAGLIA et al., 2006; SERRANO et al., 1995). Esta família é responsável por proteger o parasito da resposta imune tanto do vetor quanto do hospedeiro mamífero, assim como promover a adesão e invasão celular (DE PABLOS; OSUNA, 2012; YOSHIDA et al., 1997). A cepa CL Brener de *T. cruzi* possui cerca de 860 genes que codificam para mucinas, cuja estrutura consiste em uma região N-terminal hipervariável, seguida por uma região repetitiva rica em resíduos de Treonina e Serina (BUSCAGLIA et al., 2006). Estas repetições são os sítios aceptores para oligossacarídeos, que conferem uma natureza hidrofílica à membrana do parasito (BUSCAGLIA et al., 2006). Mucinas são subdivididas em três grupos: TcMUC I, TcMUC II e *Trypomastigote small surface antigen* (TSSA) (BUSCAGLIA et al., 2006). TcMUC I é primariamente expressa no estágio amastigota do parasito, apresenta uma região hipervariável menor e uma longa região repetitiva. TcMUC II é mais expressa no estágio tripomastigota, apresenta uma maior região hipervariável e uma menor região repetitiva (BUSCAGLIA et al., 2006). Genes da sub-família TSSA codificam um pequeno peptídeo, sem a região repetitiva, expressos na forma tripomastigota do parasito. Mucinas expõem epítopos terminais de Gal(α1,3)Gal, que são os principais alvos de anticorpos em pacientes chagásicos (ALMEIDA et al., 1994). A carga negativa resultante da sialização destes resíduos por trans-sialidases neutralizam a lise induzida pelo complemento a partir de anticorpos anti-α-galactosil, porém também reduz a eficiência de infecção do parasito (ACOSTA-SERRANO et al., 2001; PEREIRA-CHIOCCOLA et al., 2000). Mucinas purificadas de tripomastigotas metacíclicas se ligam e promovem a mobilização de Ca^{2+} em células hospedeiras, ambos processos necessários para a invasão celular (BUSCAGLIA et al., 2006; RODRÍGUEZ, A., MARTINEZ, I., CHUNG, A., BERLOT, C.H., AND ANDREWS, 1999; RUIZ et al., 1993).

Family	Group	Host	Parasite stage	Structural features	Schematic representation
TcMUC	TcMUC I	Mammal	A>T	2–10 *O*-glycosylated T_8KP_2 repeats, short N-terminal HV region, Thr-rich mature C terminus, 60–200-kDa mature products	
	TcMUC II	Mammal	T>A	1–2 T_8KAP/T_8QAP repeats, long N-terminal variable region, Thr-rich mature C terminus, 60–200-kDa mature products	
	TSSA	Mammal	T/T*	No repeats, 20-kDa mature product	

Figura 8: Estrutura das subfamílias de TcMUC (Retirado e modificado de (BUSCAGLIA et al., 2006)) – Representação esquemática da estrutura gênica das subfamílias TcMUC I, TcMUC II e TSSA. (A) Amastigota; (T) Tripomastigota; (SP) Peptídeo sinal; (HV) Região Hipervariável; (Thr-rich) Regiões ricas em treonina;

MASPs: A família multigênica MASP foi descoberta durante o sequenciamento do genoma de CL Brener, constituída por ~ 1400 genes (771 genes completos), que correspondem a ~6% do genoma diploide do parasito (BARTHOLOMEU et al., 2009; EL-SAYED, 2005a). MASPs são caracterizadas por regiões N- e C-terminais conservadas, que codificam, respectivamente, para o peptídeo sinal e sítio de adição de âncora de GPI, que flanqueiam uma região central hipervariável (BARTHOLOMEU et al., 2009) (Figura 9). Ambas regiões conservadas são clivadas da proteína madura, de modo que apenas a região hipervariável é exposta na superfície do parasito.

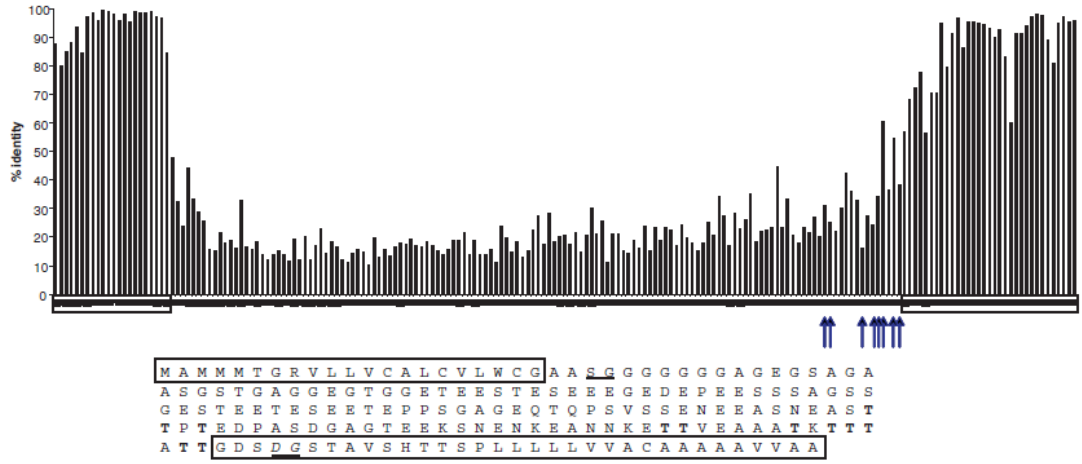


Figura 9: Estrutura proteica consenso da família MASP (retirado de (BARTHOLOMEU et al., 2009)**.** Todas as 771 sequências proteicas completas de MASP contendo ambas sequências N- e C- terminais foram alinhadas, onde uma sequência consenso foi gerada computando a porcentagem de identidade de cada posição ao longo de toda a sequência. As regiões conservadas N- e C- terminal, que correspondem respectivamente ao peptídeo sinal e sítio de adição de âncora de GPI estão destacados em caixas de linha preta. Os aminoácidos flanqueando o peptídeo sinal estão sublinhados, enquanto os aminoácidos flanqueando o sítio de adição de

GPI estão em itálico e sublinhado. As setas indicam resíduos de treonina próximos a região C-terminal.

Membros desta família são mais expressos na forma tripomastigota do parasito, e são diferencialmente expressas na população (EL-SAYED, 2005a; SECO-HIDALGO; DE PABLOS; OSUNA, 2015). Passagens consecutivas do parasito em camundongos levaram a drásticas mudanças nos membros de MASPs que eram mais expressos, sugerindo que esta família pode estar envolvida em estratégias de evasão da resposta imune (BARTHOLOMEU et al., 2009; DOS SANTOS et al., 2012). MASPs também podem estar envolvidas em processos de invasão celular, visto que passagens consecutivas em células L6 ou Llcmk2 levaram à seleção de diferentes membros da família (DOS SANTOS et al., 2012). Apesar de ser extremamente variável, diferentes membros da família MASP compartilham motivos derivados da região central da proteína, um indicativo de geração de variabilidade através de rearranjo de blocos definidos, ou módulos (Figura 10). O algoritmo MEME (BAILEY et al., 2006) foi utilizado para identificar motivos conservados e abundantes entre as 771 sequências proteicas completas de MASP identificadas no projeto genoma do parasito (EL-SAYED, 2005a). Um total de 20 motivos foram identificados, variando de 8 a 50 aminoácidos, onde a maioria dos motivos apresenta ~15-20 aminoácidos. Cada um destes motivos está representado por uma cor diferente na figura 10, onde os motivos 1 (vermelho) e 2 (azul) correspondem respectivamente às regiões conservadas N- e C- terminal de MASP. Desta forma, diversos genes de MASP são formados pelo rearranjo de blocos conservados, conferindo uma estrutura de mosaico para a família (Figura 10) (EL-SAYED, 2005a).

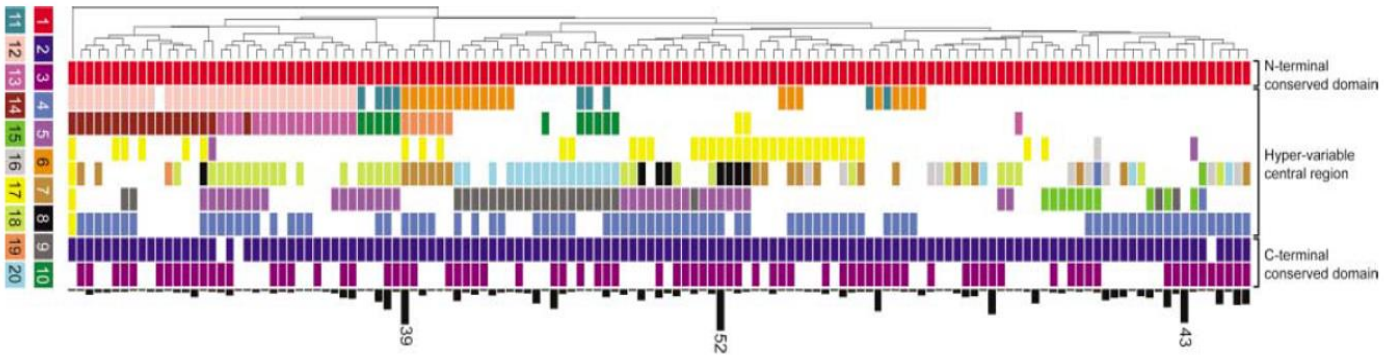


Figura 10: Representação esquemática da variabilidade e estrutura proteica da família MASP de *T. cruzi*, Retirado de (EL-SAYED, 2005a)**.** O algoritmo MEME foi utilizado para identificar motivos compartilhados entre membros da família MASP. Cada coluna corresponde a uma combinação de MEMEs, onde o número de genes de MASP que possuem esta combinação está representado pelo histograma na parte inferior de cada coluna. Os motivos consenso estão numerados de acordo com o número de ocorrência e conservação de sequência, de 1 a 20, e

codificados por cor. As regiões conservadas N-, C- terminais, assim como a região hipervariável estão identificados à direita da figura.

GP-63: Também conhecida como *Major Surface Protease* (MSP) apresenta aproximadamente 450 cópias no genoma da cepa CL Brener de *T. cruzi*. Esta família está relacionada a diversos processos biológicos em *Leishmania* sp., como inibição da resposta antiparasitária em macrófagos, resistência a peptídeos antimicrobianos e promoção da sobrevivência de amastigotas dentro do vacúolo parasitóforo em macrófagos (YAO, 2010). Em *T. cruzi*, homólogos de gp63 também parecem estar envolvidos em interações parasito-hospedeiro, visto que alguns de seus membros estão ligados à superfície do parasito por âncora de GPI, e anticorpos anti-MSP bloqueiam parcialmente a infecção de células de mamífero em cultura e *in vivo* (CUEVAS et al., 2003; KULKARNI et al., 2009; YAO, 2010).

DGF-1: Em CL Brener, a família multigênica DGF-1 é composta por 565 genes e 136 pseudogenes, onde grande parte de seus membros estão localizados em regiões sub-teloméricas (MORAES BARROS et al., 2012). Apesar de apresentar grande variabilidade de sequência, esta família ainda não foi implicada em interações parasito-hospedeiro. A sua localização citoplasmática e presença de domínios como de fatores de crescimento epidérmicos e o tri peptídeo arginina-glicina-ácido aspártico (RGD) sugerem uma ação semelhante a integrinas para os membros desta família (DE PABLOS; OSUNA, 2012; KAWASHITA et al., 2009).

1.7 Variação no número de cópias cromossômicas

Aneuploidias, a presença de um número anormal de cromossomos em uma célula, resulta em letalidade ou em graves anomalias em eucariotos superiores multicelulares, reduzindo grandemente seu fitness (TORRES; WILLIAMS; AMON, 2008). Dentre os casos mais comuns de aneuploidias em humanos estão a trissomia do cromossomo 21 na síndrome de Down, números aberrantes de cromossomos sexuais (como o triplo X, Klinefelter e a síndrome XYY) e tumorigênese (HASSOLD; HUNT, 2001; LV et al., 2012; STANKIEWICZ; LUPSKI, 2010). Porém, alguns organismos unicelulares como *Saccharomyces cerevisiae* e *Candida albicans* parecem utilizar aneuploidias como mecanismo para permitir uma rápida adaptação a mudanças ambientais, sugerindo que variações no número de cromossomos pode resultar em um fitness positivo em condições de estresse, como promover resistência a drogas (ABBEY et al., 2011; PFAU; AMON, 2012; SELMECKI; FORCHE; BERMAN, 2010; SHELTZER et al., 2011). Um

grande número de aneuploidias tem sido documentado em células de organismos eucariotos, como hepatócitos de camundongos (DUNCAN et al., 2010), leveduras (ABBEY et al., 2011; SELMECKI; FORCHE; BERMAN, 2010; SHELTZER et al., 2011), células de câncer de pulmão (DOUBRE et al., 2005) e tripanossomatídeos do gênero *Leishmania* (DOWNING et al., 2011; LACHAUD et al., 2014; MANNAERT et al., 2012; ROGERS et al., 2011; STERKERS et al., 2011, 2014; VALDIVIA et al., 2015b). Alterações de ploidia podem ser adquiridas rapidamente através de uma falha na segregação ou replicação estocástica de cromossomos, podendo alterar as dosagens gênicas em uma única geração e permitindo ao parasito se adaptar rapidamente a novos ambientes (MANNAERT et al., 2012). De modo similar, um estado de aneuploidias mosaico pode gerar um pool de fenótipos na mesma população proporcionando um aumento de adaptabilidade, visto que distintas combinações de CCNV podem ser vantajosos em ambientes diferentes. Se a polissomia for estável por longos períodos evolutivos, ela pode aumentar a variabilidade de sequência, visto que genes em cromossomos duplicados podem mutar sem perda de função, já que a função ancestral seria mantida na cópia original do cromossomo. A recombinação homóloga poderia então embaralhar genes em cromossomos aneuploides, gerando uma variedade de diferentes fenótipos. Como visto em leveduras, mudanças deletérias encontradas em cromossomos monosômicos podem ser rapidamente eliminados (OTTO; GERSTEIN, 2008; ZEYL; VANDERFORD; CARTER, 2003), enquanto mutações benéficas seriam mantidas ou até expandidas para estados triplóides e tetraplóides (STERKERS et al., 2011, 2014). Por estes motivos, a presença de aneuploidias em mosaico favorece a rápida adaptação do parasito a novos ambientes, condições de estresse, a transição entre hospedeiros vertebrados e invertebrados, e combina as vantagens de monossomias e polissomias na mesma população (Figura 11).

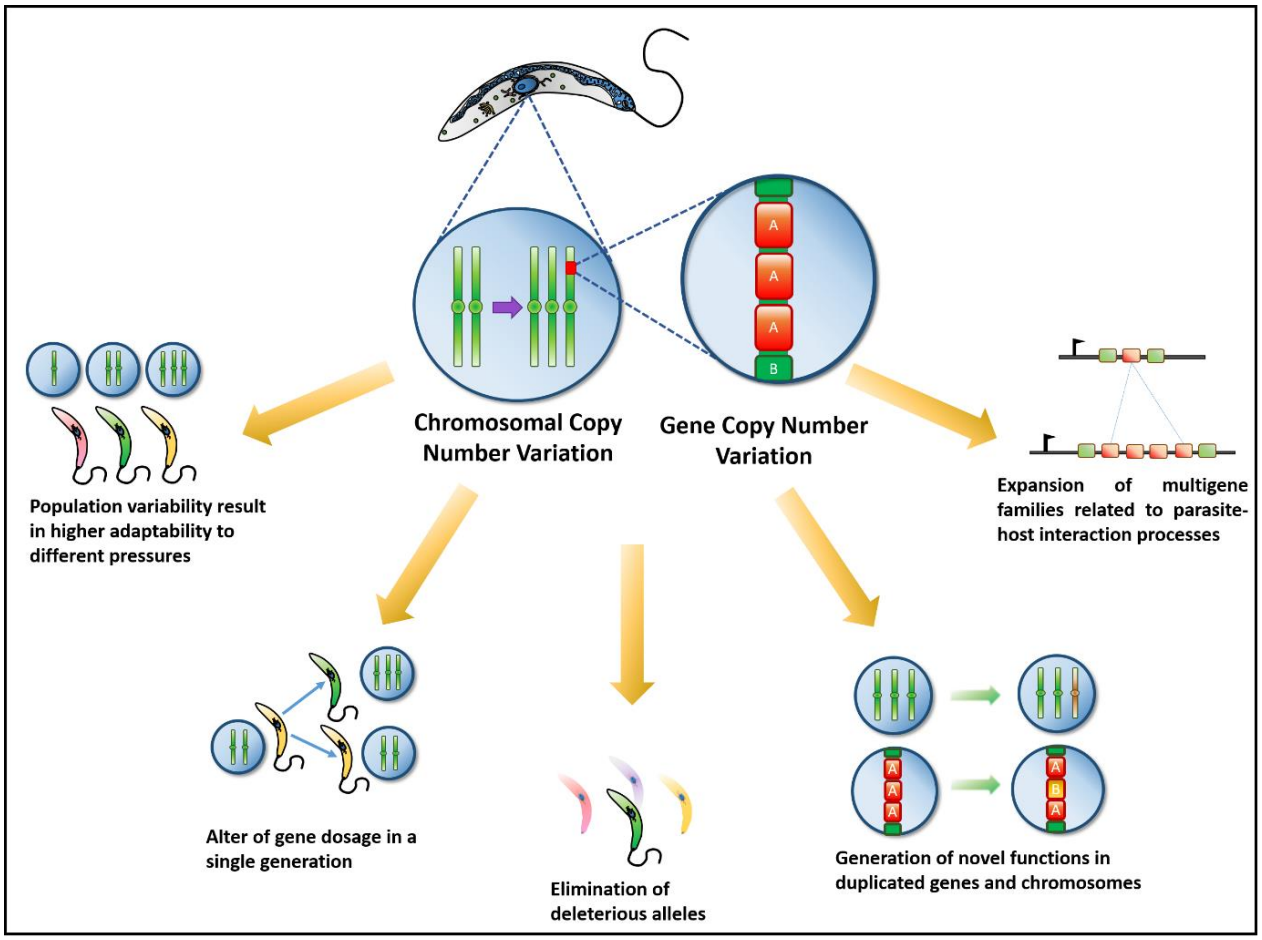


Figura 11: Principais implicações de variações no número de cópias gênicas e cromossômicas (Retirado de Reis-Cunha 2017 – *In press*).

Como a ausência de condensação dos cromossomos durante a divisão celular inviabiliza a utilização de análises citogenéticas em tripanossomatídeos, existem dois principais métodos alternativos e complementares para estimar aneuploidias nestes organismos: Hibridização fluorescente *in situ* (FISH) (LACHAUD et al., 2014; STERKERS et al., 2011, 2014), e sequenciamento genômico (WGS) seguido de análise de profundidade de cobertura reads (RDC) e frequência alélica (DOWNING et al., 2011; ROGERS et al., 2011; VALDIVIA et al., 2015b). Enquanto FISH permite a simultânea identificação de aneuploidias em cada célula de uma população, ela restringe a análise a apenas alguns cromossomos (STERKERS et al., 2011, 2014). Por outro lado, WGS seguido por análises de RDC permite a comparação simultânea do padrão de variação no número de cópias cromossômicas (CCNV) de todos os cromossomos em uma população, assim como a determinação de genótipos e frequências alélicas. Porém, esta análise não possui a resolução de célula única fornecida pelo FISH e necessita de genoma de referência para ser realizado (DUJARDIN et al., 2014; STERKERS et al., 2014).

Em 2011 a presença de CCNV entre diversas espécies do gênero *Leishmania* foi descrita através de análises de RDC e frequência alélica (DOWNING et al., 2011; ROGERS et al., 2011). Análises de FISH revelaram a ocorrência de CCNV entre indivíduos de uma mesma cultura celular em *Leishmania*, onde o padrão de aneuploidias entre elas era variável, sugerindo que a formação diferencial de aneuploidias é um evento frequente no parasito (STERKERS et al., 2011, 2014). Trabalhos recentes sugerem que esta aneuploidia mosaico pode também ser encontrada no parasito *T. cruzi* (MINNING et al., 2011), porém a frequência de sua ocorrência e importância biológica ainda necessitam maiores estudos.

1.8 Estado atual do genoma de referência de *T. cruzi*

A natureza híbrida e o grande conteúdo repetitivo do genoma de CL Brener dificultaram a montagem do genoma em cromossomos completos, levando a uma montagem inicial que continha ~5.000 scaffolds/contigs em 2005 (EL-SAYED, 2005a). Em 2009, com o auxílio de *BAC-end sequences* e mapas de sintenia com *T. brucei* e *L. major*, os contigs de CL Brener foram montados em 41 cromossomos hipotéticos (WEATHERLY; BOEHLKE; TARLETON, 2009). Porém, esta montagem ainda apresenta grandes regiões de gaps, e 9.8 MB de sequências codificadoras que não foram incorporadas nos 41 cromossomos hipotéticos, superando em mais de três vezes a extensão do maior cromossomo predito de *T. cruzi*, que possui 2.3MB (WEATHERLY; BOEHLKE; TARLETON, 2009). Estas sequencias não incorporadas nos cromossomos correspondem, em sua maioria, a grande parcela das famílias multigênicas que codificam para proteínas de superfície de *T. cruzi* (WEATHERLY; BOEHLKE; TARLETON, 2009). Desta forma, a montagem atual do genoma de CL Brener permite análises do conteúdo estrutural de genes conservados e *housekeeping* do parasito, mas inviabiliza análises focadas na estrutura e organização genômica das regiões que codificam para famílias multigênicas de *T. cruzi*.

2. Estrutura da tese

No presente trabalho nós exploramos as regiões genômicas conservadas e variáveis de *T. cruzi* para fornecer uma avaliação global da variabilidade genômica intra- e inter-DTUs deste parasito. No primeiro capítulo, é proposta uma nova metodologia denominada SCoPE para predizer a presença de CCNV entre DTUs de *T. cruzi* baseada em WGS seguido por RDC de genes de cópia simples conservados do parasito. Esta metodologia é menos enviesada por regiões repetitivas e gaps, sendo adequada para a atual montagem do genoma de referência de *T. cruzi*. No segundo capítulo, a variação genômica intra-DTU TcII é avaliada com base em amostras de campo recém isoladas do estado de Minas Gerais, onde foram apresentados resultados de distância geográfica, filogenia mitocondrial, nuclear e CCNV. Finalmente, o terceiro capitulo explora os clusters de famílias multigênicas de *T. cruzi*, onde avalia a variabilidade na abundancia de motivos gênicos das famílias MASP, TcMUC e Trans-sialidase intra- e inter-DTUs, através de nova metodologia independente de mapeamento membro específico e montagem *de novo* de *reads*. As justificativas, objetivos gerais e específicos estão descritos em cada capítulo.

CAPÍTULO 1: Variação no número de cópias cromossômicas revela distintos níveis de plasticidade em diferentes cepas de *T. cruzi*

3. Justificativa:

O cariótipo de *T. cruzi* não foi ainda completamente elucidado devido a inabilidade de se realizar análises citogenéticas, pela fraca condensação de seus cromossomos durante a divisão celular (HENRIKSSON et al., 2002; MINNING et al., 2011; SOUZA et al., 2011). Análises de gel de eletroforese de campo pulsátil e citometria de fluxo têm mostrado uma variação significativa em tamanho e número de cromossomos entre diferentes cepas do parasito (HENRIKSSON et al., 2002; LIMA et al., 2013a; PEDROSO; CUPOLILLO; ZINGALES, 2003; SOUZA et al., 2011; TRIANA et al., 2006; VARGAS; PEDROSO; ZINGALES, 2004), assim como diferentes graus de similaridade do perfil eletroforético e massa de DNA por célula (CERQUEIRA et al., 2008; LEWIS et al., 2009; SOUZA et al., 2011; VARGAS; PEDROSO; ZINGALES, 2004). Para melhor caracterizar o cariótipo de *T. cruzi*, o genoma da cepa CL Brener foi recentemente montado em 41 cromossomos hipotéticos, baseado nos contigs gerados em 2005 por El-Sayed e colaboradores, assim como sequências BAC-end e mapas de sintenia com *T. brucei* e *L. major* (WEATHERLY; BOEHLKE; TARLETON, 2009). Apesar de ainda apresentar regiões fragmentadas, em especial nos clusters de famílias multigênicas que codificam para proteínas de superfície, as regiões cópias simples do genoma representadas em ambos haplótipos Esmeraldo-like e Non-Esmeraldo-like estão mais completas. Estas regiões podem, portanto, ser utilizadas como marcadores cromossômicos para regiões únicas do genoma, permitindo a normalização e comparação das estimativas de número de cópias cromossômicas entre diferentes cepas/isolados.

A ocorrência de aneuploidias cromossômicas em *T. cruzi* já foi sugerida anteriormente por experimentos de *tiling arrays* genômicos e hibridizações competitivas (MINNING et al., 2011). Porém, este tipo de análise não permite a detecção de expansões/deleções cromossômicas presentes simultaneamente em CL Brener e na cepa avaliada. Neste capítulo, visamos a identificação de CCNV em cepas de *T. cruzi* originadas de diferentes DTUs, baseado na profundidade da cobertura de *reads* (*read depth* coverage ou RDC) mapeadas contra genes cópia simples presentes nos 41 cromossomos de CL Brener. Por ser baseada no conteúdo haploide de CL Brener, nossa metodologia permite a detecção de CCNV também em cromossomos aneuploides na cepa de referência. A identificação de variações no padrão de

aneuploidia entre diferentes cepas e DTUs de *T. cruzi* pode explicar algumas das peculiaridades em relação a aspectos da biologia deste parasito, como o diferente grau de similaridade intragenômica e massa de DNA por célula, e pode fornecer ao parasito uma rápida adaptabilidade a variações ambientais.

4. Objetivos

4.1 Objetivo Geral

O principal objetivo deste capítulo é identificar padrões diferenciais de expansões ou deleções cromossômicas entre cepas de *T. cruzi* pertencentes a diferentes DTUs.

4.2 Objetivos específicos

1-Obter *whole genome shotgun* reads das cepas Arequipa (TcI), Colombiana (TcI) Sylvio (TcI), Y (TcII), Esmeraldo (TcII) e 231 (TcIII).

2-Determinar através de metodologias de mapeamento competitivo qual conjunto de cromossomos de CL Brener, Esmeraldo-like ou Non-Esmeraldo-like, corresponde a melhor sequência de referência para o mapeamento de cada uma das bibliotecas de reads.

3-Desenvolver uma metodologia independente de montagem para determinar e comparar a ploidia de cromossomos em diferentes DTUs de *T. cruzi*.

4-Determinar a ploidia geral de cada cepa de *T. cruzi* com base em frequência alélica.

5. Metodologia

5.1 Cultivo e clonagem de epimastigotas de *T. cruzi*

Epimastigotas provenientes das cepas Arequipa (TcI), Colombiana (TcI), Y (TcII) e 231 (TcIII) de *T. cruzi* foram cultivados em meio LIT suplementado com 10% de soro fetal bovino a 28°C (CAMARGO, 1964) acrescido de 100 µg de estreptomicina/mL e 100 unidades de penicilina/mL. As culturas foram repicadas em intervalos de 3-5 dias mediante inóculos de $1x10^6$ células/mL até a fase logarítmica de crescimento, quando a densidade celular foi de aproximadamente $1x10^7$ células/mL. Para a clonagem das cepas Arequipa e 231, 10^3 epimastigotas foram plaqueados em meio semi-sólido (agarose de baixa temperatura de fusão 0.75%, infusão de fígado e cérebro 48.4%, infusão de fígado e triptose (LIT) 48.4%, 2.5% sangue desfibrinado, e 250 µg/mL penicilina/estreptomicina) e incubados a 28°C por 35 dias. Clones isolados foram obtidos e transferidos para garrafas de cultura com 5 mL de LIT acrescido de 10% de soro fetal bovino.

5.2 Extração de DNA

Um total de $1x10^8$ parasitos de cada uma das cepas/clones previamente mencionados foi centrifugada a 3000 g. Os parasitos foram lavados três vezes com PBS gelado, ressuspendidos em PBS com 100 µg/mL de proteinase K e incubados a 25 °C por dez minutos. O DNA genômico foi extraído utilizando o kit *Wizard® Genomic DNA Purification Kit* (Promega), seguindo as recomendações do fabricante. A integridade do DNA foi avaliada por eletroforese em gel de agarose. As amostras de DNA foram submetidas ao protocolo de genotipagem de acordo com Souto 1996 (SOUTO et al., 1996), de-Freitas 2006 (DE FREITAS et al., 2006) e Burgos 2007 (BURGOS et al., 2007) para confirmar seu DTU de origem.

5.3 Sequenciamento genômico

A construção das bibliotecas de sequenciamento de genoma completo (WGS) por *shotgun* e o sequenciamento genômico das cepas Arequipa (TcI), Colombiana (TcI) e Y (TcII) foram realizados na Unidade de Genômica Computacional Darcy Fontoura de Almeida (UGCDFA), do Laboratório Nacional de Computação Científica (LNCC) (Petrópolis RJ, Brasil), utilizando as tecnologias de sequenciamento 454 GS-FLX *Titanium* e *Ion Proton*™. Para o sequenciamento por 454-GS-FLX *Titanium*, cada biblioteca não pareada foi construída utilizando 5µg de DNA genômico (gDNA), seguindo as recomendações do protocolo da série GS FLX

Titanium. Todas as titulações, emulsões, reações em cadeia da polimerase (PCRs), e sequenciamentos foram realizados de acordo com o protocolo do fabricante. Uma *PicoTiterPlate* cheia foi utilizada para o sequenciamento de cada biblioteca. Para os sequenciamentos por *Ion Proton*™, um total de 1µg de DNA foi utilizado para a preparação das bibliotecas de sequenciamento. Todos os passos do sequenciamento foram realizados com base nas recomendações do fabricante. As reads genômicas de 454 da cepa Sylvio de *T. cruzi* foram gentilmente cedidas pelo Dr. Bjorn Andersson (Karolinska Institutet). O sequenciamento da cepa 231 (TcIII) foi obtida utilizando a plataforma Illumina Hiseq 2000 NGS (Baptista et. al., em preparação). As bibliotecas de reads genômicas de Esmeraldo (reads de Illumina e 454) foram obtidas no Arquivo de Reads Sequenciamento (SRA) do *National Center for Biotechnology Information* (NCBI). A tabela com todos os códigos de depósito das reads se encontra abaixo (Tabela 1):

Cepa	Plataforma	Origem	Identificador
Arequipa (TcI)	Ion Proton™	LNCC	SRS838181
	454 GS FLX Titanium	LNCC	SRS838181
Colombiana (TcI)	Ion Proton™	LNCC	SRS841912
	454 GS FLX Titanium	LNCC	SRS841912
Sylvio (TcI)	454 GS FLX Titanium	*Karolinska Institutet*	-
Esmeraldo (TcII)	454 GS FLX Titanium	NCBI	SRR833799
	Illumina Genome Analyzer II	NCBI	SRR833800
	454 GS FLX Titanium	NCBI	SRR058517
	454 GS FLX Titanium	NCBI	SRR058509
	454 GS FLX Titanium	NCBI	SRR058520
	454 GS FLX Titanium	NCBI	SRR058518
	454 GS FLX Titanium	NCBI	SRR058519
	454 GS FLX Titanium	NCBI	SRR058515
	454 GS FLX Titanium	NCBI	SRR058516
	454 GS FLX Titanium	NCBI	SRR058513
	454 GS FLX Titanium	NCBI	SRR058514
	454 GS FLX Titanium	NCBI	SRR058510
	454 GS FLX Titanium	NCBI	SRR058511
	454 GS FLX Titanium	NCBI	SRR058512
Y (TcII)	Ion Proton™	LNCC	SRS842149
	454 GS FLX Titanium	LNCC	SRS842149
231 (TcIII)	Illumina Hiseq 2000 NGS	UFMG	PRJEB9129

Tabela 1: Identificação de todas as bibliotecas de reads genômicas utilizadas no presente trabalho.

5.4 Pré-processamento das reads

A qualidade do *base calling* das reads foi avaliada utilizando o programa FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads menores que 30 nucleotídeos e com uma qualidade média de Phred menor do que 20 (podendo apresentar um erro a cada 100 bases) (EWING et al., 1998; EWING; GREEN, 1998) foram removidas das bibliotecas utilizando o programa *fast_quality_trimmer*, do pacote *fastx* toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html).

5.5 Mapeamento genômico

O mapeamento competitivo foi utilizado para selecionar o haplótipo de referência de CL Brener mais relacionado às bibliotecas de WGS de reads das cepas Arequipa (TcI), Colombiana (TcI), Sylvio (TcI), Esmeraldo (TcII), Y (TcII) e 231 (TcIII). Para isso, cada uma das bibliotecas de reads de WGS foi simultaneamente mapeada nos 41 cromossomos dos haplótipos Esmerado-like e Non-Esmeraldo-like do genoma de CL Brener versão 6 (http://tritrypdb.org/tritrypdb/), utilizando o programa Bowtie2 (LANGMEAD; SALZBERG, 2012) (parâmetros "*very sensitive*" com a permissão de 1 *mismatch* na *seed*). Reads mapeadas foram filtradas por um cutoff mínimo de qualidade de mapeamento 30 (alta confiabilidade do posicionamento da read em determinada região genômica) utilizando o programa SAMtools v1.1 (LI et al., 2009). Como as reads foram mapeadas simultaneamente nos dois haplótipos, elas tenderão a mapear nos cromossomos do haplótipo de CL Brener mais relacionado à cepa em questão. No caso de regiões conservadas entre os dois haplótipos, a read poderia mapear com confiabilidade em regiões dos dois haplótipos. Se isso ocorrer, o mapeador não consegue ter confiabilidade para assinalar a read a uma região específica, fornecendo um baixo valor de qualidade de mapeamento. Desta forma, a remoção de reads com qualidade de mapeamento menor que 30 de um mapeamento competitivo, leva a recuperação de apenas as reads que mapearam com maior confiabilidade em um haplótipo de CL Brener em detrimento de outro. O número de reads que mapeou preferencialmente em cada haplótipo foi contabilizado e utilizado como critério de proximidade evolutiva entre eles. Estas análises foram realizadas utilizando scripts em *PERL* e *BEDtools genomecov* v2.16.2 (QUINLAN; HALL, 2010). As imagens foram geradas utilizando o programa *GraphPad Prism V5.01* e scripts em PERL e R. Após a seleção do haplótipo mais adequado para o mapeamento das *reads* de cada cepa, cada uma das bibliotecas de WGS foi mapeada nos 41 cromossomos do haplótipo de CL Brener selecionado, utilizando o programa Bowtie2

(LANGMEAD; SALZBERG, 2012). Reads mapeadas foram filtradas por um cutoff mínimo de qualidade de mapeamento 30 utilizando o programa SAMtools v1.1 (LI et al., 2009)

5.6 Genes cópia simples e cobertura genômica

A profundidade de cobertura média do genoma (o número de reads que mapeou em cada posição do genoma) foi estimada com base na cobertura de 1563 genes de cópias simples, apresentando ortólogos em ambos os haplótipos Esmeraldo-like e Non-Esmeraldo-like (ortólogos 1:1). Estes genes foram identificados através do programa OrthoMCL v2.0 (LI; STOECKERT; ROOS, 2003), baseado em uma combinação de maior similaridade recíproca e "*Markov clustering algorithm*" (MCL). Inicialmente um BLASTp (pacote 2.2.21) de todas as sequências proteicas dos haplótipos Esmeraldo-like e Non-Esmeraldo-like, com um cutoff de e-*value* $1e^{-5}$ foi realizado. Os valores de e-*value* foram convertidos em escala log para criar uma matriz de similaridade, onde um MCL com parâmetro de inflação de 1.5 foi utilizado para selecionar os clusters de ortólogos. Um total de 1563 genes com apenas uma cópia (sem parálogos) em cada um dos haplótipos de CL Brener foram identificados e assumidos como genes de cópia simples no genoma haploide de CL Brener. A lista contendo estes genes pode ser encontrada no "***Additional file 1: Table S1***", no paper Reis-Cunha 2015 (REIS-CUNHA et al., 2015).

A cobertura média destes 1.563 genes em cada haplótipo foi assumida como a cobertura média do genoma haploide do parasito, estimada como 47x para Arequipa, 28x para colombiana, 9x para Sylvio, 52x para Esmeraldo 34x para Y e 76x para 231.

5.7 Conteúdo de Single-Nucleotide Polymorphisms (SNPs)

Para a identificação de SNPs nas reads de cada uma das 6 cepas de *T. cruzi* com relação aos cromossomos de CL Brener selecionados como referência, o programa SAMtools, função mpileup, foi utilizado (LI et al., 2009). Este programa foi escolhido pois apresenta em seu arquivo de saída uma coluna que contém as contagens de reads que apresentam o mesmo alelo da referência, assim como o número de reads que apresenta o alelo mutado, permitindo a contabilização da proporção de reads de cada alelo. Para ser considerado um SNP confiável, foi utilizado um *cutoff* mínimo de cobertura de 10 reads para validar a posição, com um mínimo de

5 reads suportando cada variante, assim como um *cutoff* mínimo *mapping quallity* 30 para a posição.

5.8 Ploidia cromossomal

No caso de reads de WGS, a cobertura de reads é proporcional ao número de cópias de um segmento de DNA no genoma de origem. Por isso, o número de cópias de um motivo em um genoma pode ser estimado através da divisão da profundidade de reads do motivo pela cobertura média do genoma (Figura 12).

Devido ao grande conteúdo repetitivo do genoma de CL Brener, a utilização da média do mapeamento de reads em todas as posições de um cromossomo levaria a uma incorreta estimativa de sua ploidia. Como o mapeador não consegue alocar estas reads repetitivas com confiabilidade em apenas uma posição, as reads destas regiões recebem uma pontuação de qualidade de mapeamento baixa e são removidas na etapa de trimagem. Portanto, utilizar a cobertura média de todas as posições de um cromossomo para cálculo de sua ploidia pode levar a resultados subestimados, principalmente nos cromossomos ricos em famílias multigênicas como o 18 e o 41. Por outro lado, o não uso de um cutoff de mapeamento alto permite que reads erroneamente mapeadas sejam contabilizadas, enviesando a contagem. Por estes motivos, nós utilizamos um cutoff mínimo de qualidade de mapeamento 30 e a cobertura média dos genes de cópia simples presentes em cada cromossomo como marcadores de sequências únicas do genoma haplóide, assumindo que a média de cobertura de todos os genes de cópia simples em um cromossomo corresponde à cobertura média do cromossomo. Nós denominamos esta metodologia de SCoPE, do inglês *Single Copy Ploidy Estimation*. Desta forma, a RDC média de todos os genes cópia simples em cada cromossomo foi gerada com base em scripts em PERL, normalizada pela cobertura média de todos os genes cópia simples presentes em cromossomos diploides do genoma e assumida como a ploidia predita do cromossomo. As imagens do padrão de CCNV foram gerados através de scripts em PERL, R e o programa *GraphPad Prism* V5.01.

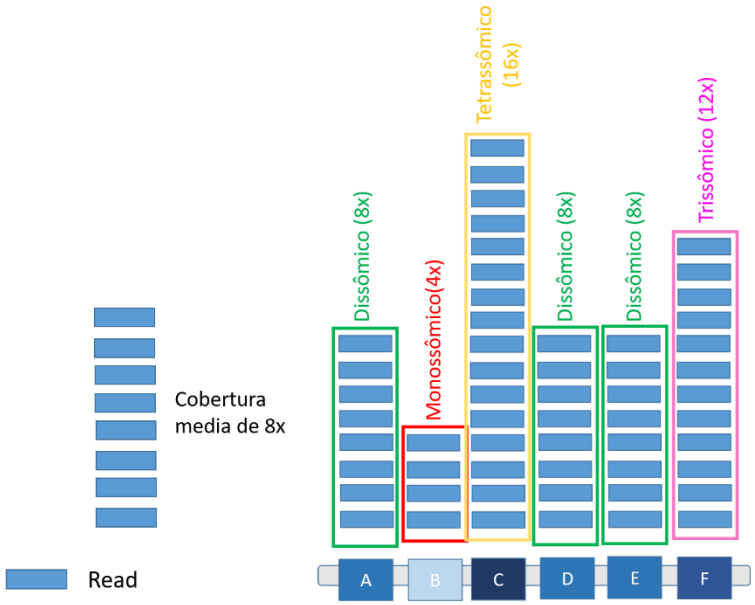


Figura 12: Estimativa do número de cópias de região genômica com base em RDC. Na figura, cada box em azul corresponde a uma read. A linha cinza corresponde a um cromossomo, que contém regiões nomeadas de "A até F". No caso de um sequenciamento genômico com cobertura média de 8x (8 reads por posição), as regiões "A, D e E" que possuem 8 reads seriam dissômicas, a região "B" que possui apenas 4 reads seria monossômica, a região "C" com 16 reads seria tetrassômica e a região "F" que possui 12 reads seria trissômica.

Uma outra forma de avaliar a ploidia de uma região genômica é através do padrão de SNPs heterozigóticos. Considerando apenas SNPs heterozigóticos presentes em genes de cópia simples, se ele estiver em um cromossomo diploide, a proporção de cada variante do SNP será de 50% (Figura 13 A). Porém, se ele estiver em cromossomo triploide, o número total de cópias desta região sobe para 3. Desta forma, a proporção de um alelo será ~66% (duas ocorrências) e do outro ~33% (uma ocorrência) (Figura 13 B). No caso de cromossomos tetraploides, o padrão fica um pouco mais complexo, com proporções de 25%, 50% e 75% (Figura 13 C).

A metodologia acima descrita foi aplicada para cada cromossomo em cada amostra de *T. cruzi* avaliada, onde a proporção de alelos em cada posição dos SNP heterozigóticos foi obtida e arredondada para a primeira casa decimal. As frequências das proporções dos SNPs foram agrupadas em 10 categorias, variando de 0,1 até 1, e a distribuição aproximada de suas frequências de variantes, para cada cromossomo, foram plotadas em R. Para obter a predição da ploidia geral de cada genoma, a mesma metodologia foi aplicada, mas utilizando simultaneamente as regiões heterozigóticas de todas as sequências codificadoras (CDSs) de todos os cromossomos. As imagens foram geradas em R.

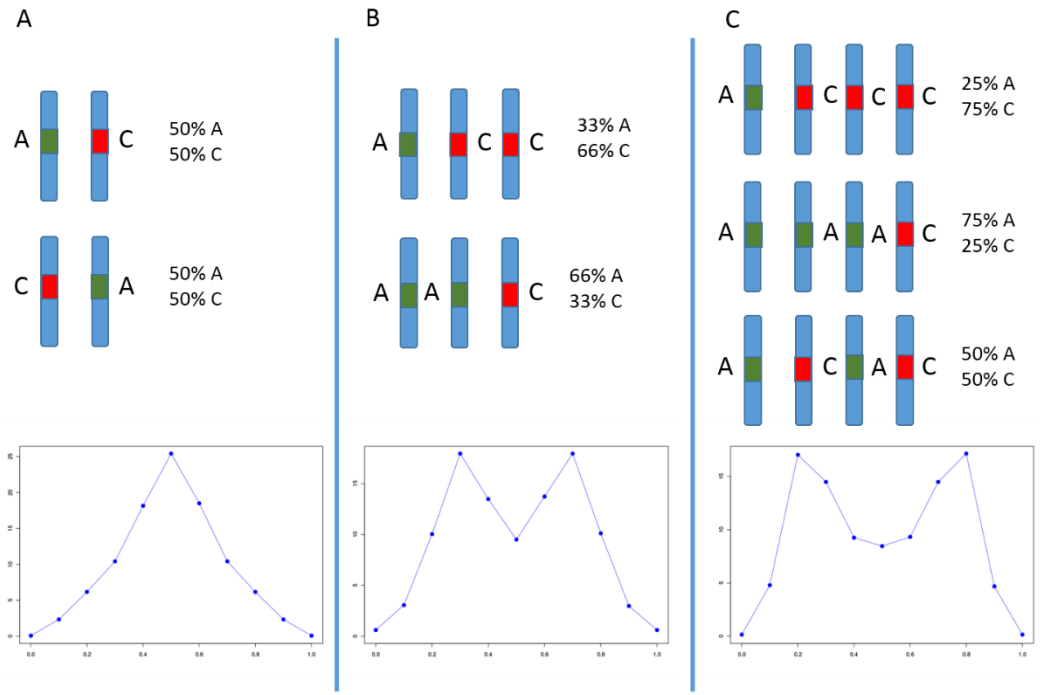


Figura 13: Proporções de SNPs heterozigóticos. (A) em cromossomos diploides, a proporção esperada entre alelos de SNPs heterozigóticos, no caso "A" e "C" é de 50% para cada variante. Desta forma ao se obter a proporção de cada variante em cada posição heterozigótica do cromossomo, o pico da frequência destes valores é de 0.5 (linha azul), que corresponde a 50% de cada variante. **(B)** Em caso de cromossomos triploides a proporção entre os SNPs heterozigóticos será de 66% para uma das variantes e 33% para a outra, gerando picos em 0,33, 0.66 ou em ambos. **(C)** Já no caso de cromossomos tetraploides, o padrão é mais complexo, podendo ser de 25% de uma variante e 75% de outra ou de 50% de cada variante. Esta proporção é utilizada para confirmar as predições de ploidia baseada em RDC.

5.9 Dendrograma de clusterização

A análise de clusterização hierárquica baseada nas distâncias euclidianas da ploidia predita de cada cromossomo de todas as seis cepas de *T. cruzi* avaliadas foi realizada usando o pacote Pvclust, da plataforma R. Dois métodos foram utilizados: o "*approximately unbiased*" (au) e a "*bootstrap probability*" (bp). Ambos os métodos foram calculados com 10.000 iterações. A imagem foi gerada em R.

5.10 Ontologia Gênica

As análises de ontologia gênica para genes presentes no cromossomo 31 de *T. cruzi* foram realizadas utilizando a análise de distribuição hipergeométrica no programa BiNGO (MAERE; HEYMANS; KUIPER, 2005), com a correção de falsa descoberta de Benjamini e Hochberg (BENJAMINI; YEKUTIELI, 2001).

6. Resultados:

6.1 Mapeamento competitivo e conteúdo de SNPs

Para selecionar o haplótipo de CL Brener mais adequado para servir como referência para o mapeamento das bibliotecas de WGS das 6 cepas de *T. cruzi*, foi realizado um mapeamento competitivo das bibliotecas de reads nos haplótipos Esmeraldo-Like e Non-Esmeraldo-like (Figura 13 A). Como esperado baseado na filogenia de *T. cruzi*, as reads das cepas de TcII (Esmeraldo e Y) mapearam preferencialmente com os cromossomos do haplótipo Esmeraldo-like, enquanto as read da cepa de TcIII (231) mapearam preferencialmente com cromossomos do haplótipo Non-Esmeraldo-like. Já as reads das cepas pertencentes ao DTU TcI (Arequipa, Colombiana e Sylvio) apresentaram uma melhor cobertura com o haplótipo Non-Esmeraldo-like do que no Esmeraldo-like (Figura 14 B).

Para confirmar a seleção do haplótipo de CL Brener para ser utilizado como referência para o mapeamento, as bibliotecas de reads de cada cepa foram mapeadas separadamente com cada um dos dois haplótipos de CL Brener, e o número de SNPs/1kb em cada um dos 1563 genes de cópia simples foi estimado (Figura 14 C). Os números de SNPs/1kb entre cada biblioteca de reads e os haplótipos de CL Brener Esmeraldo-like e Non-Esmeraldo-like foram, respectivamente, 17,50 e 3,10 para Arequipa; 32,74 e 14,37 para Colombiana; 52,41 e 25,36 para Sylvio; 5,66 e 18,07 para Esmeraldo; 4,39 e 16,06 para Y e 17,89 e 5,72 para 231 (Figura 14 C). Baseado nestes resultados, o haplótipo de Non-Esmeraldo-Like foi selecionado como referência para o mapeamento das reads das cepas Arequipa, Colombiana, Sylvio e 231, enquanto o haplótipo Esmeraldo-like foi escolhido como referência para o mapeamento das cepas Esmeraldo e Y.

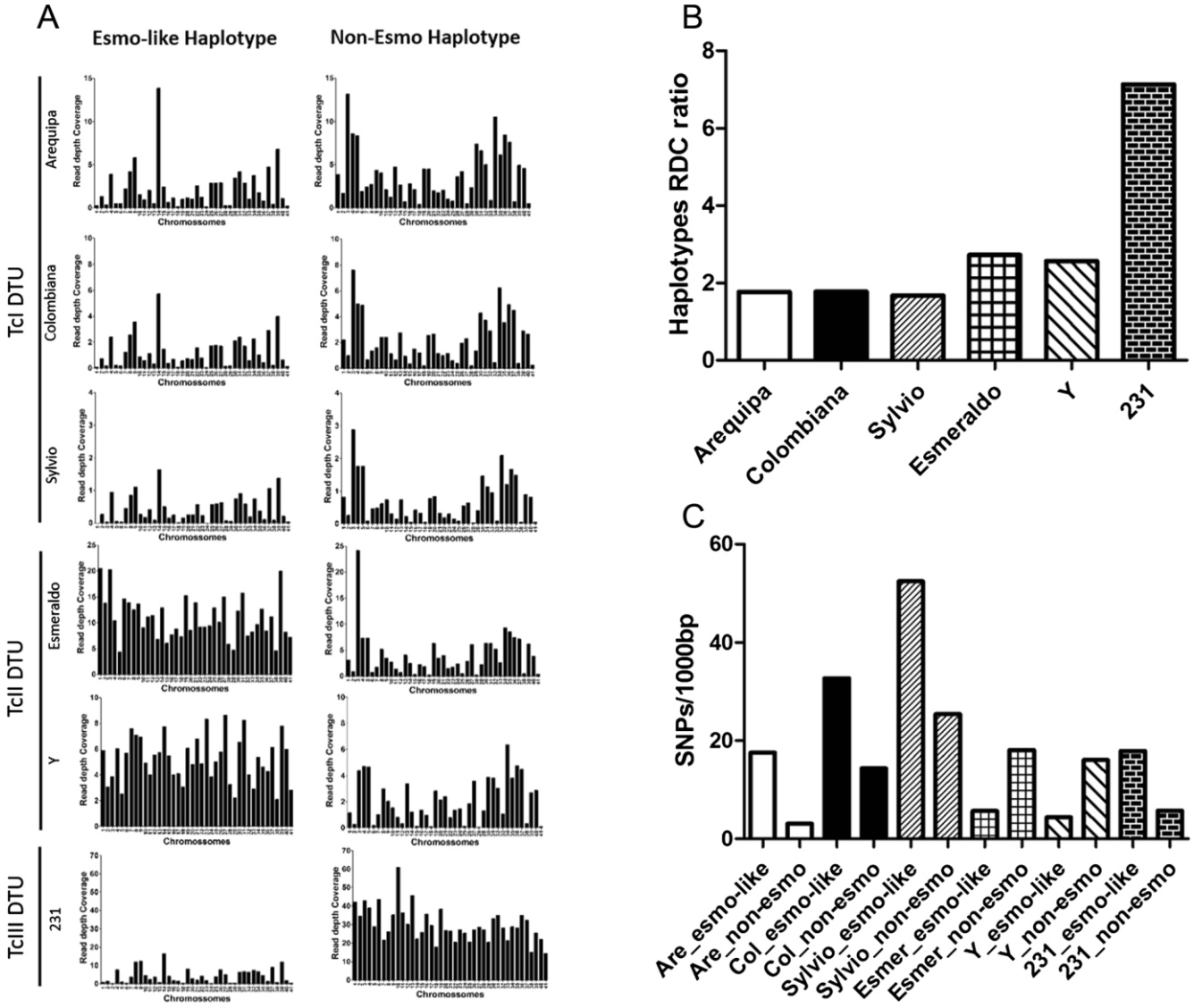


Figura 14: Mapeamento Competitivo e conteúdo de SNPs das cepas de *T. cruzi* de diferentes DTUs. No mapeamento competitivo, reads de WGS das cepas Arequipa, Colombiana, Sylvio, Esmeraldo, Y e 231 foram simultaneamente mapeados nas 41 sequências cromossômicas dos haplótipos Esmeraldo-like e Non-Esmeraldo-like de CL Brener, onde a RDC de cada cromossomo foi estimada. **(A)** As barras pretas em cada painel correspondem a RDC de cada um dos 41 cromossomos de CL Brener baseados no mapeamento preferencial de reads, onde painéis à esquerda correspondem aos cromossomos do haplótipo Esmeraldo-like e painéis à direita correspondem aos cromossomos do haplótipo Non-Esmeraldo-like. A cepa e o DTU referente a cada painel estão descritos à esquerda. **(B)** Para cada cepa de *T. cruzi*, a RDC média do haplótipo com o maior RDC no mapeamento competitivo (Non-Esmeraldo para Arequipa, Colombiana, Sylvio e 231; Esmeraldo-like para Esmeraldo e Y) foi dividida pela RDC média do haplótipo de CL Brener com o menor RDC (Esmeraldo-like para Arequipa, Colombiana, Sylvio e 231; Non-Esmeraldo-like para Esmeraldo e Y). **(C)** Densidade de SNPs em cada uma das cepas de *T. cruzi* nos 1563 genes cópia simples.

6.2 Metodologia para estimar o número de cópias cromossômicas nas cepas de *T. cruzi*

Aproximadamente 50% do genoma de *T. cruzi* é composto por regiões repetitivas, o que dificulta a confiabilidade do mapeamento de reads a uma região específica. Além disso, a representação atual dos 41 cromossomos dos haplótipos Esmeraldo-like e Non-Esmeraldo de CL Brener apresentam grandes *gaps* internas em cromossomos (WEATHERLY; BOEHLKE; TARLETON, 2009), o que pode comprometer a acurácia da estimativa de ploidia baseada em RDC.

Para avaliar a melhor metodologia para estimar a ploidia de cada cromossomo em cada cepa de *T. cruzi*, duas metodologias foram comparadas, a *Whole Chromosome Ploidy Estimation* (WCPE) e a *Single Copy Ploidy Estimation* (SCoPE). Na metodologia de WCPE, a predição de ploidia é baseada na razão entre a RDC média de cada posição no cromossomo e a cobertura média do genoma (Figura 15 A). Esta abordagem considera todas as posições em cromossomo para estimar sua ploidia, incluindo regiões repetitivas e *gaps*. Em contrapartida, a metodologia SCoPE estima o número de cópias cromossômicas para cada cromossomo baseado na razão entre a cobertura média de todos os genes de cópia simples presentes em um cromossomo, dividido pela cobertura média do genoma (Figura 15 B). Esta abordagem infere o número de cópias de cada cromossomo baseado apenas na RDC dos 1563 genes de cópia simples, ortólogos 1:1 entre os haplótipos Esmeraldo-like e Non-Esmeraldo-like de CL Brener. Cromossomos ricos em famílias multigênicas, repetições ou gaps, como os cromossomos 18, 28, 38 e 41 (Figura 15 C), apresentaram uma menor ploidia estimada baseada na metodologia WCPE quando comparada com o SCoPE. Como a metodologia de SCoPE é menos susceptível a vieses de alteração na ploidia por sequências repetitivas, esta metodologia foi a selecionada para estimar a ploidia de todos os cromossomos em cada cepa utilizada neste estudo.

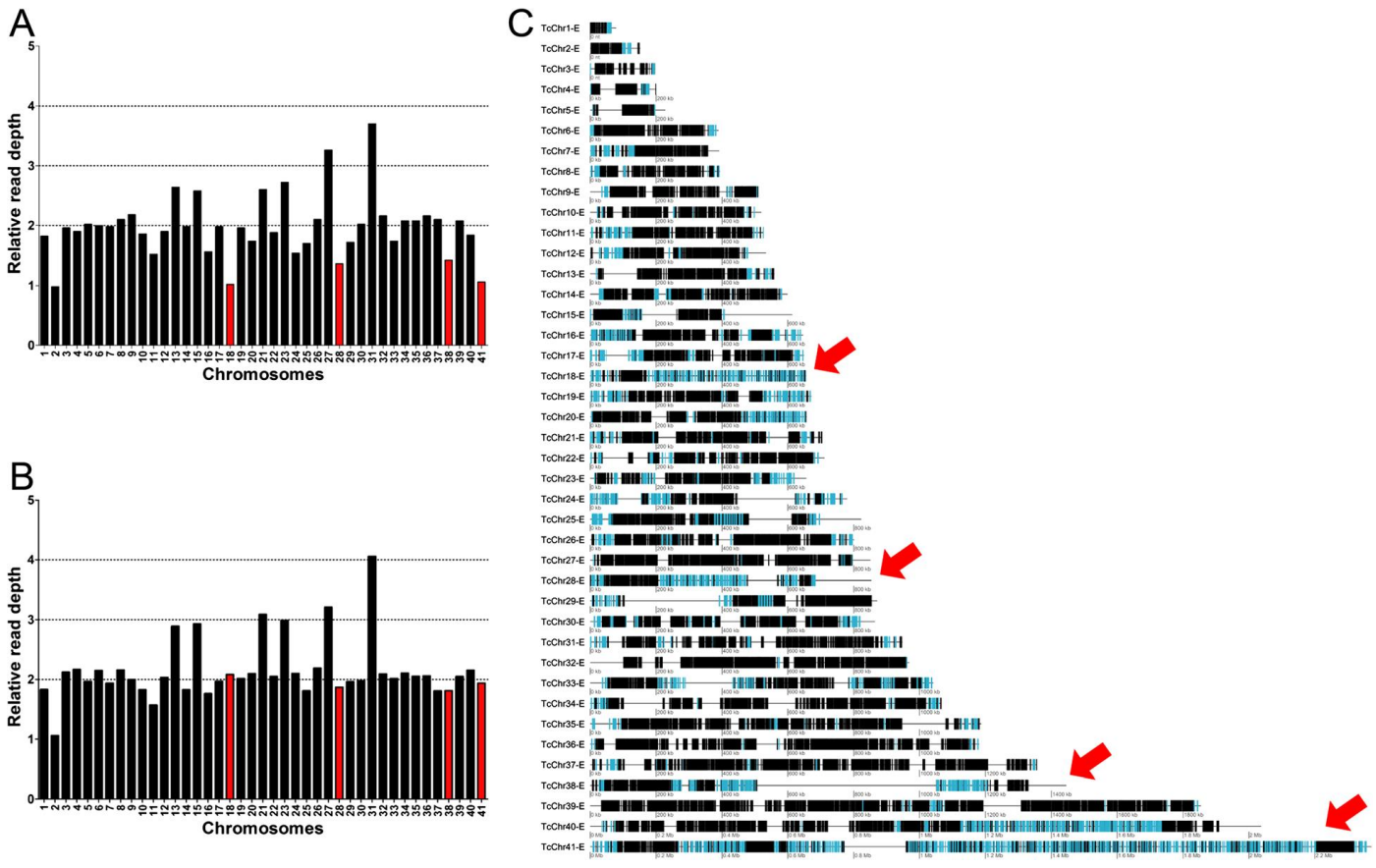


Figura 15: Metodologias para determinar o número de cópias cromossômicas em cada cepa de *T. cruzi*. Para estimar a ploidia dos cromossomos nas cepas de *T. cruzi* baseada em RDC das sequências de CL Brener, duas metodologias foram comparadas. **(A)** A "*Whole Chromosome Ploidy Estimation"* (WCPE) prediz a ploidia com base na cobertura média ao longo de toda a sequência cromossômica. Cada barra preta numerada de 1 a 41 corresponde a um cromossomo de CL Brener, onde a altura da barra remete a ploidia do cromossomo. Nestas barras, o valor de 2 corresponde a um cromossomo diploide. **(B)** A "*Single Copy Ploidy Estimation"* (SCoPE) prediz a ploidia cromossomal com base na média de cobertura de genes cópia simples presentes em cada cromossomo. **(C)** Distribuição gênica dos 41 cromossomos de CL Brener. Cada linha corresponde a um cromossomo desenhado em proporção ao seu tamanho, onde genes pertencentes a famílias multigênicas estão representadas por caixas azuis, e genes hipotéticos e *housekeeping* por caixas pretas. As setas em vermelho denotam os cromossomos enriquecidos com famílias multigênicas ou com grandes regiões de gap. Estes cromossomos apresentam uma cobertura enviesada com base na predição por WCPE quando comparada ao método SCoPE.

6.3 Variação no número de cópias cromossômicas entre as cepas de *T. cruzi*

A metodologia SCoPE foi utilizada para estimar a ploidia cromossômica das 6 cepas de *T. cruzi* utilizadas neste trabalho (Figura 16). Os valores de RDCs de cada gene cópia simples, assim como as imagens das razões da frequência alélica de cada cromossomo estão respectivamente resumidas nos arquivos "***Additional file* 2: *Table* S2**" e "***Additional file* 3: *Figure* S1**"do paper Reis-cunha 2015 (REIS-CUNHA et al., 2015). Para determinar a ploidia geral do

genoma de cada cepa de *T. cruzi*, foram contabilizadas as frequências alélicas em todas as posições heterozigotas com duas e apenas duas variantes em CDSs. Para este fim, a RDC de um alelo em posições heterozigóticas era dividida pelo somatório do total de reads que mapearam nos dois alelos da posição e arredondado para a primeira casa decimal. Baseado nesta estimativa, um cromossomo diploide apresenta usualmente uma razão de ~0.5, um triploide de 0.3 e 0.7 e um tetraploide apresenta um padrão mais complexo com uma combinação de ~0.5, 0.2 e 0.8. Como a maioria dos SNPs heterozigóticos apresentou uma proporção de ~0.5, a ploidia genômica geral foi assumida como diploide para todas as cepas de *T. cruzi* (Figura 16).

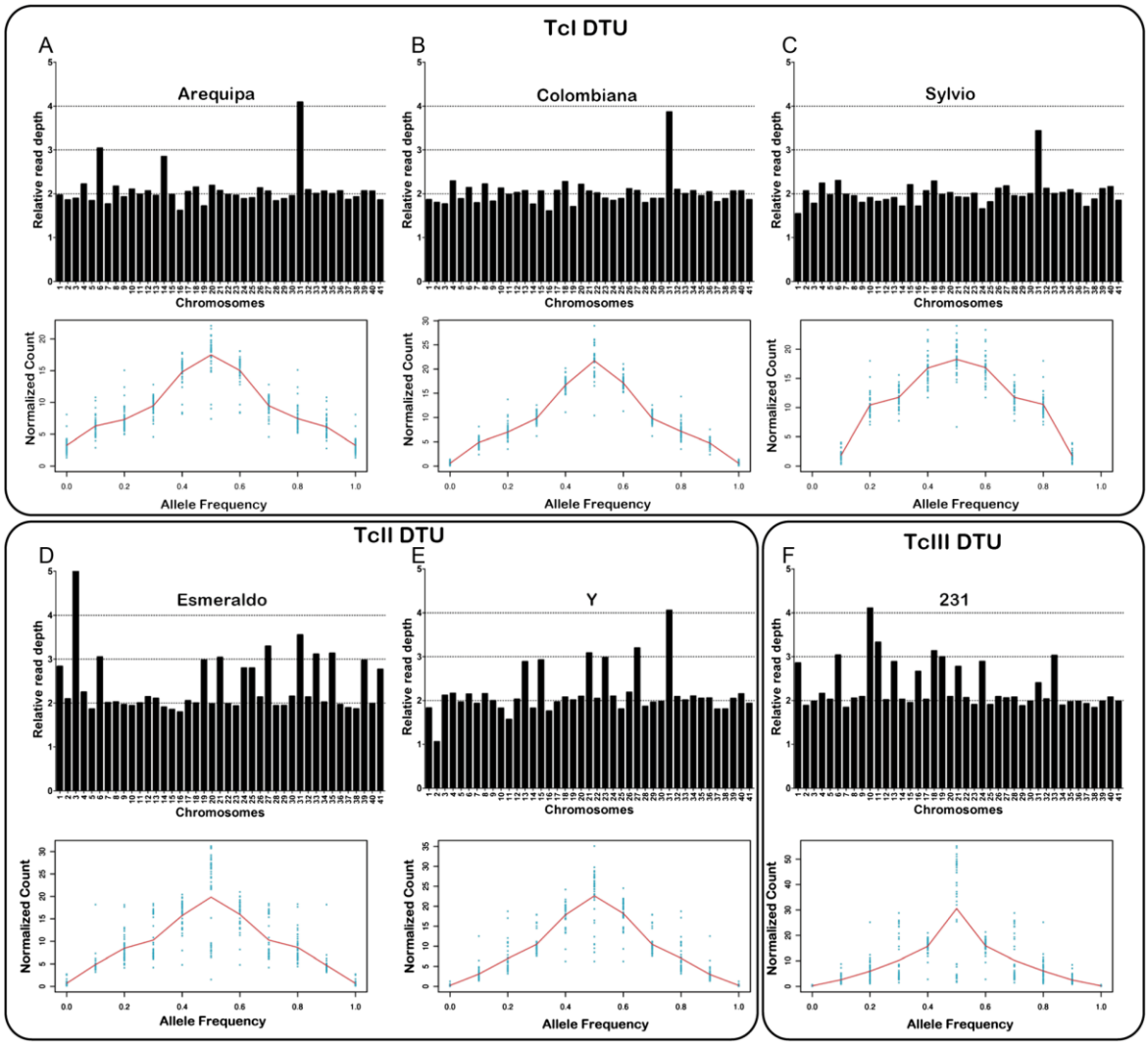


Figura 16: Predição de ploidia das seis cepas de *T. cruzi*. A ploidia predita de cada cromossomo das cepas de *T. cruzi* **(A)** Arequipa, **(B)** Colombiana e **(C)** Sylvio do DTU TcI; **(D)** Y e **(E)** Esmeraldo do DTU TcII; e **(F)** 231 do DTU TcIII, utilizando os 41 cromossomos de CL Brener como referência foi estimado pela metodologia de SCoPE. Cada barra preta corresponde a razão entre a média da RDC dos genes de cópia simples presentes no cromossomo e a cobertura do genoma, onde a altura da barra corresponde a ploidia predita do cromossomo. Abaixo de cada painel, a ploidia geral do genoma, baseada na proporção da frequência alélica está representada por uma linha vermelha, onde um pico ~0.5 corresponde a um genoma predominantemente diploide.

A análise de CNV cromossômica revelou grandes diferenças entre as cepas de *T. cruzi* de diferentes DTUs e algumas diferenças entre membros do mesmo DTU. Aparentemente cepas do DTU TcI possuem um cariótipo mais estável, quando comparados a cepas dos DTUs TcII e TcIII (Figura 16). Cepas do DTU TcI apresentaram aneuploidias apenas no cromossomo 31, com exceção da cepa Arequipa, que apresentou também trissomias dos cromossomos 6 e 14. As cepas do DTU TcII apresentaram um cariótipo mais plástico, com várias predições de duplicação cromossômica e uma monosomía. A cepa Esmeraldo apresentou o cromossomo 3 como pentassômico, cromossomos 27, 31, 33 e 35 entre trissômicos e tetrassômicos; cromossomos 6, 19, 21 e 39 como trissômicos; e cromossomos 1, 24, 25, e 41 como intermediária entre dissômicos e trissômicos. A cepa Y, também pertencente ao DTU TcII apresentou uma tetrassomia do cromossomo 31, os cromossomos 21 e 27 entre trissômicos e tetrassômicos, os cromossomos 13, 15 e 23 como trissômicos e uma monosomía do cromossomo 2. O representante do DTU TcIII, 231, apresentou uma tetrassomia no cromossomo 10; cromossomos 11 e 18 entre triploide e tetraploide, trissomia dos cromossomos 6 e 19; e cromossomos 1, 13, 16, 21, 24 e 31 com uma ploidia intermediaria entre dissômico e trissômico (Figura 16).

Para confirmar as predições de ploidia, a distribuição das frequências alélicas entre posições de SNP heterozigóticos entre todas as CDSs nos 41 cromossomos das seis cepas de *T. cruzi* foi estimado (*Additional file* 3: *Figure* S1 do paper (REIS-CUNHA et al., 2015)). Os resultados desta análise estão de acordo com os valores de CCNV preditos com base na cobertura de genes cópia simples para todos os cromossomos, exceto os cromossomos 20 e 23 da cepa Sylvio e cromossomo 7 da cepa Esmeraldo, preditos como tetrassômicos nas análises de frequência alélica e como dissômicos pelo método SCoPE, assim como os cromossomos 6 e 14 de Arequipa, estimados como tetrassômicos com base em frequência alélica e trissômicos pela análise de SCoPE.

Para avaliar se essas predições de ploidia eram resultantes de ganho/perda de um cromossomo inteiro ou se eram oriundos de duplicações segmentais ou perda parcial de cromossomos, a RDC normalizada de cada posição ao longo de cada cromossomo das seis cepas de *T. cruzi* foi estimada, e pode ser visualizada em "*Additional file* 4: *Figure* S2"do paper (REIS-CUNHA et al., 2015). A figura 17 resume um exemplo comparativo da ploidia predita pela média da cobertura dos genes cópia simples em cada cromossomo (Figura 17 A), a RDC em casa posição do cromossomo (Figura 17 B) e a ploidia estimada por frequência alélica (Figura 17 C), para cromossomos dissômicos, trissômicos e tetrassômicos (Figura 17). Como esperado, com exceção

das regiões compostas por famílias multigênicas repetitivas ou gaps, a predição de ploidia ao longo do cromossomo inteiro está em concordância com a predição de ploidia por SCoPE e frequência alélica. Este resultado sugere que as aneuploidias são provavelmente um resultado de perda e ganho de cromossomos inteiros.

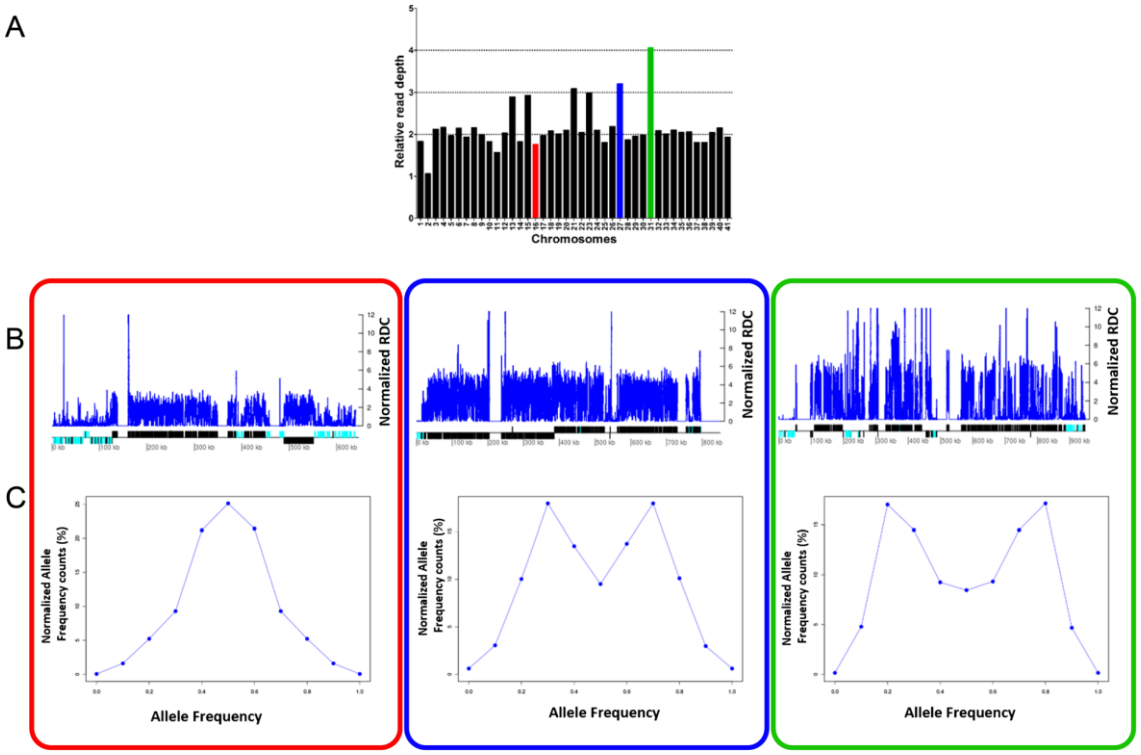


Figura 17: Correspondência entre a ploidia predita pela metodologia de SCoPE, frequência alélica e RDC ao longo de todo cromossomo. A correspondência entre a ploidia cromossômica predita por SCoPE e a RDC normalizada de cada posição dos cromossomos 16 (caixa vermelha), 27 (caixa azul) e 31 (caixa verde) da cepa Y de *T. cruzi* é mostrada. **(A)** Predição da ploidia pela metodologia de SCoPE. **(B)** RDC em cada posição do cromossomo. A linha azul corresponde a RDC normalizada em cada posição do cromossomo, estimada como a razão entre a RDC e a cobertura genômica. Abaixo, os genes codificadores de proteína estão ilustrados como retângulos desenhados em proporção ao seu tamanho, e a sua fita codificadora está representada pela posição acima (fita +) ou abaixo (fita -) da linha central. Retângulos em ciano ou preto representam, respectivamente, famílias multigênicas e genes hipotéticos/*housekeeping*. Gaps são representados por regiões com ausência de genes, sem cobertura de reads. O cromossomo 16 foi predito como dissômico, 27 como trissômico e 31 como tetrassômico. As regiões com baixa cobertura correspondem a clusters de famílias multigênicas ou gaps. **(C)** A ploidia predita com base em frequência alélica, para os cromossomos 16, 27 e 31. O pico em 0.5 classifica o cromossomo 16 como dissômico, os picos de 0.3 e 0.6 classificam o cromossomo 27 como trissômico e os picos de 0.2 e 0.8 classificam o cromossomo 31 como tetrassômico.

Para determinar se o perfil de ploidia das seis cepas de *T. cruzi* avaliadas estava de acordo com a filogenia, uma análise de clusterização hierárquica foi realizada utilizando como

entrada a ploidia predita de cada cromossomo. A análise de clusterização foi baseada nas distancias euclidianas par-a-par, utilizando o método de *Complete Linkage* (Figura 18). Nesta análise, todas as cepas pertencentes ao DTU TcI agruparam juntas, com as cepas Colombiana e Sylvio mais próximas entre si do que Arequipa. Isto ocorreu provavelmente pelas aneuploidias exclusivas de Arequipa nos cromossomos 6 e 14. Ambas as cepas de TcII apresentaram um padrão diferente de aneuploidias, resultando em uma menor robustez de agrupamento do que o visto para TcI. A cepa de TcIII, 231, apresentou o padrão mais discordante de aneuploidias entre as cepas de *T. cruzi* avaliadas (Figura 18).

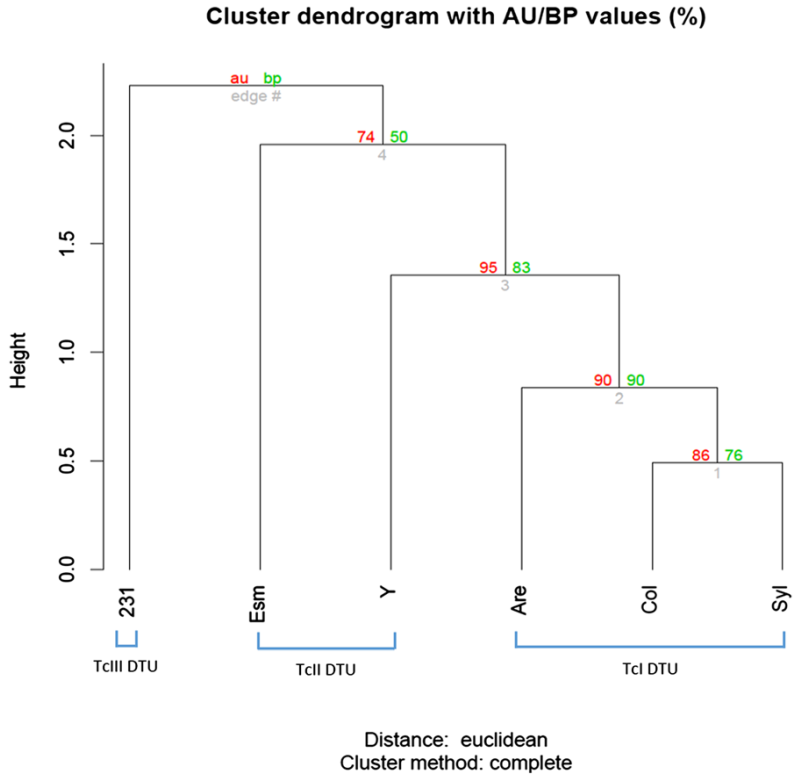


Figura 18: Dendrograma da análise hierárquica da ploidia predita das cepas de *T. cruzi*. A clusterização hierárquica baseada nas distâncias euclidianas dos padrões de aneuploidias de cada cromossomo das cepas de *T. cruzi* Arequipa, Colombiana, Sylvio, Esmeraldo Y e 231 foi realizada utilizando o pacote Pvclust, na plataforma R. Dois métodos de reamostragem foram utilizados para avaliar a incerteza do método de clusterização: *approximately unbiased* (au) em vermelho e a probabilidade de *bootstrap* (bp) em verde.

6.4 Cromossomo 11 da cepa Y de *T. cruzi*

O cromossomo 11 da cepa Y apresentou uma predição de ploidia intermediária entre monossômico e dissômico. Uma avaliação da RDC ao longo de toda a sequência deste cromossomo revelou que os primeiros 248kb de sequência apresentaram uma cobertura menor

do que a apresentada no resto de sua sequência (Figura 19 A). Para confirmar esta observação, a RDC média dos genes de cópia simples localizados *upstream* e *downstream* da posição de 248kb foi estimada. Os genes de cópia simples localizados *upstream* da posição de 248kb apresentaram uma RDC não normalizada de 20, enquanto os *downstream* da posição apresentaram uma RDC de 40 (Figura 19 b). Esta redução de RDC na região 5' do cromossomo sugere uma perda segmental da região inicial de uma das cópias do cromossomo na cepa Y. De modo alternativo, este padrão pode ser resultante de diferenças na estrutura do cromossomo 11 na população de parasitos, visto que a cepa Y não foi clonada antes do sequenciamento.

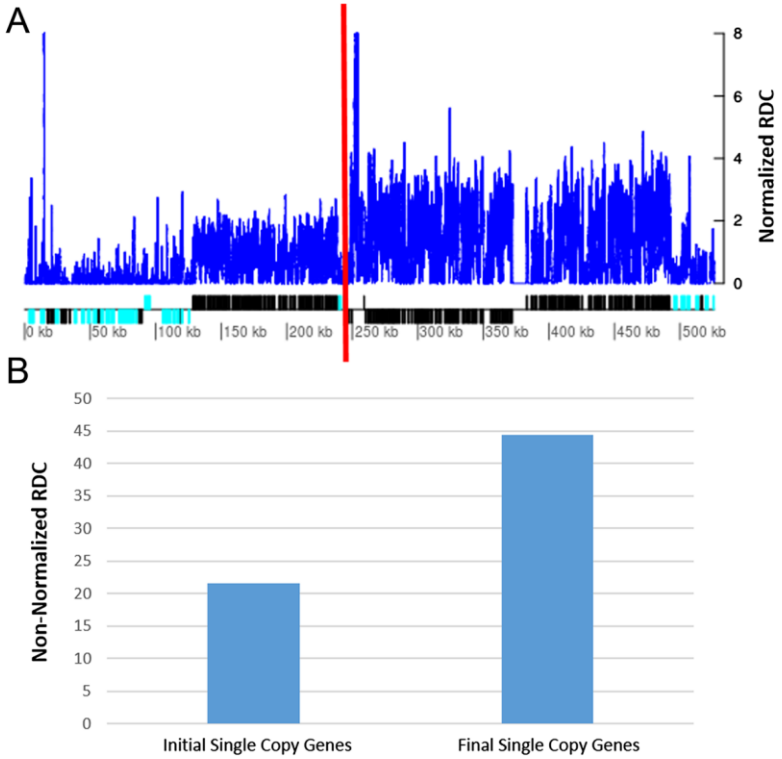
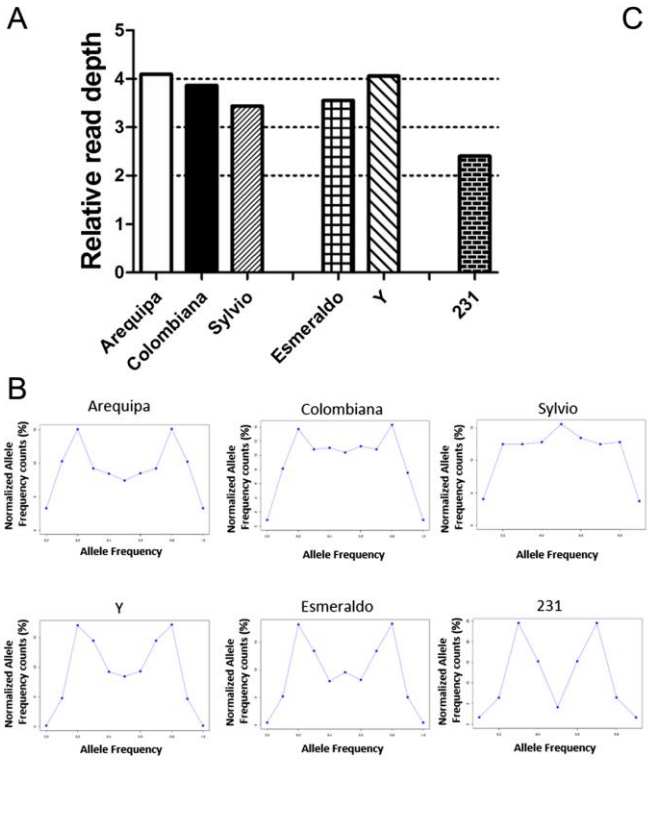


Figura 19: Ploidia predita do cromossomo 11 da cepa Y. (A) A linha azul corresponde a RDC normalizada em cada posição do cromossomo 11. Abaixo, os genes codificadores de proteína estão representados por retângulos desenhados em proporção ao seu tamanho, onde a fita codificadora é indicada pelo posicionamento acima (fita +) ou abaixo (fita -) da linha central. Retângulos em ciano ou preto correspondem, respectivamente, a famílias multigênicas ou genes hipotéticos/*housekeeping*. Gaps são representados por regiões sem genes, com ausência de cobertura. A linha vermelha corresponde a posição 248kb no cromossomo, que separa as regiões de baixa e alta RDC. **(B)** RDC média não normalizada dos genes de cópia simples localizados upstream ou downstream da posição de 248kb do cromossomo 11.

6.5 Análise de enriquecimento gênico por ontologia do cromossomo 31

Dentre os 41 cromossomos de CL Brener, o cromossomo 31 foi o único supranumerário em todas as seis cepas de *T. cruzi* avaliadas por ambas as metodologias SCoPE (Figura 20 A) e

frequência alélica (Figura 20 B). Análises de enriquecimento gênico por ontologia foram utilizadas para identificar funções biológicas enriquecidas neste cromossomo quando comparado a todo o genoma (Figura 20 C, "*Additional file* 5:*Table* 3" (REIS-CUNHA et al., 2015). Esta análise revelou um enriquecimento de genes envolvidos em processos de glicosilação e síntese de glicoproteínas em ambos os haplótipos de CL Brener.



C

Non-Esmo - Molecular		
GO-ID	Description	p-value
8378	Galactosyltransferase activity	6,55E-06
16758	Transferase activity, transferring hexosyl groups	3,41E-04
16298	Lipase activity	1,86E-03
4806	Triglyceride lipase activity	1,86E-03
16757	Transferase activity, transferring glycosyl groups	5,30E-03
3824	Catalytic activity	1,64E-02
52689	Carboxylic ester hydrolase activity	3,00E-02
4198	Calcium-dependent cysteine-type endopeptidase activity	4,13E-02

Non-Esmo - Biological		
GO-ID	Description	p-value
43413	Macromolecule glycosylation	3,23E-05
6486	Protein glycosylation	3,23E-05
9101	Glycoprotein biosynthetic process	3,23E-05
9100	Glycoprotein metabolic process	3,23E-05
70085	Glycosylation	7,22E-05
6629	Lipid metabolic process	3,78E-03
1901137	Carbohydrate derivative biosynthetic process	1,96E-02
2643	Single-organism carbohydrate metabolic process	3,64E-02
46394	Carboxylic acid biosynthetic process	3,64E-02
16053	Organic acid biosynthetic process	3,64E-02
6633	Fatty acid biosynthetic process	3,64E-02
72330	Monocarboxylic acid biosynthetic process	3,64E-02

Esmo - Molecular		
GO-ID	Description	p-value
8378	Galactosyltransferase activity	1,77E-06
16758	Transferase activity, transferring hexosyl groups	4,78E-05
4806	Triglyceride lipase activity	4,86E-04
16757	Transferase activity, transferring glycosyl groups	5,73E-04
16298	Lipase activity	5,73E-04
4198	Calcium-dependent cysteine-type endopeptidase activity	8,18E-03
8777	Acetylornithine deacetylase activity	1,01E-02
50897	Cobalt ion binding	1,01E-02
52689	Carboxylic ester hydrolase activity	1,16E-02
4197	Cysteine-type endopeptidase activity	1,93E-02
19213	Deacetylase activity	2,16E-02
4148	Dihydrolipoyl dehydrogenase activity	3,90E-02
8234	Cysteine-type peptidase activity	3,90E-02

Esmo - Biological		
GO-ID	Description	p-value
43413	Macromolecule glycosylation	6,24E-06
6486	Protein glycosylation	6,24E-06
9101	Glycoprotein biosynthetic process	6,24E-06
9100	Glycoprotein metabolic process	6,24E-06
70085	Glycosylation	7,07E-06
1901137	Carbohydrate derivative biosynthetic process	5,34E-03
44723	Single-organism carbohydrate metabolic process	9,92E-03
6526	Arginine biosynthetic process	1,86E-02
5975	Carbohydrate metabolic process	3,80E-02
6525	Arginine metabolic process	4,40E-02

Figura 20: Ploidia e análise de enriquecimento gênico por ontologia no cromossomo 31. Dos 41 cromossomos de CL Brener o cromossomo 31 foi o único supranumerário em todas as cepas de *T. cruzi* avaliadas. **(A)** Estimativa ploidia baseada em SCoPE. **(B)** Estimativa de ploidia baseada em frequência alélica. **(C)** Ontologias gênicas enriquecidas no cromossomo 31.

7. Discussão:

A disponibilidade da representação dos 41 cromossomos dos haplótipos Esmeraldo-like (TcII) e Non-Esmeraldo-like (TcIII) de CL Brener (WEATHERLY; BOEHLKE; TARLETON, 2009) forneceu um arcabouço para análises comparativas de ploidia entre cepas de *T. cruzi* baseada em reads de sequenciamento de nova geração. Porém, a natureza repetitiva do genoma de *T. cruzi*, assim como a presença de grandes regiões de gaps, comprometem estimativas de ploidia baseadas na média da RDC de todas as posições dos cromossomos. Para eliminar estas limitações, nosso grupo de estudos propôs a metodologia de SCoPE, baseada na RDC de genes cópia simples e conservados entre os dois haplótipos de CL Brener. Além disso, para permitir a comparação entre bibliotecas de reads com profundidades de cobertura diferentes, a cobertura média do genoma foi estimada e utilizada como normalizador. Estas análises revelaram grandes diferenças nos números de cópias cromossômicas entre as cepas e DTUs de *T. cruzi* (Figura 15).

O parasito *T. cruzi* apresenta um perfil distinto de bandas cromossômicas em experimentos de eletroforese de campo pulsátil, onde o padrão e número de bandas varia entre os DTUs. Estas diferenças foram atribuídas principalmente a expansões e retrações de clusters de famílias multigênicas ou a eventos de fusão ou quebra cromossômica, ocorridos durante a evolução do parasito (BRANCHE et al., 2006; PEDROSO; CUPOLILLO; ZINGALES, 2003; SOUZA et al., 2011; TRIANA et al., 2006; VARGAS; PEDROSO; ZINGALES, 2004). De fato, aproximadamente 50% do genoma de CL Brener corresponde a sequências repetitivas, muitas das quais representam elementos transponíveis e famílias multigênicas que codificam para proteínas de superfície (EL-SAYED, 2005a; WEATHERLY; BOEHLKE; TARLETON, 2009). Estes clusters são extremamente variáveis em conteúdo gênico e tamanho, correspondendo a regiões de perda de sintenia entre os haplótipos de CL Brener (EL-SAYED, 2005b), assim como entre CL Brener e Sylvio (FRANZÉN et al., 2011). Por outro lado, genes localizados em clusters direcionais compostos por genes *housekeeping* e hipotéticos são geralmente conservados e sintênicos entre as cepas de *T. cruzi* (EL-SAYED, 2005b; FRANZÉN et al., 2011; MINNING et al., 2011; SOUZA et al., 2011), e, portanto, representam uma fonte adequada para normalizações de RDC e análises de CCNV.

A ocorrência de aneuploidias em *T. cruzi* já foi previamente sugerida com base em *Tiling arrays* de oligonucleotídeos e hibridização genômica competitiva entre 16 cepas de *T. cruzi* utilizando a cepa CL Brener como referência (MINNING et al., 2011). Porém, estas análises de hibridização comparativa não permitiram a detecção de duplicações/deleções cromossômicas presentes simultaneamente nas cepas de referência e teste (MINNING et al., 2011). Como a nossa metodologia utiliza o conteúdo haploide dos cromossomos de CL Brener como referência

para o mapeamento das reads, ele permite a detecção de CCNVs presentes simultaneamente em CL Brener e na cepa de *T. cruzi* avaliada em questão.

7.1 Mapeamento competitivo

Para identificar o haplótipo de CL Brener mais adequado a ser usado como referência para o mapeamento das reads de cada cepa de *T. cruzi*, cada uma das bibliotecas de reads foi simultaneamente mapeada nas sequências cromossômicas dos haplótipos Esmeraldo-like e Non-Esmeraldo-like de CL Brener, através de uma abordagem competitiva (Figura 13 A). Como o valor de cutoff utilizado foi muito estringente, as reads que mapearam em regiões conservadas nos dois haplótipos de CL Brener foram excluídas, e apenas reads que mapearam preferencialmente com um dos haplótipos foram recuperadas. Como esperado, as reads das cepas de TcII mapearam preferencialmente com o haplótipo Esmeraldo-like, enquanto as reads de TcIII mapearam preferencialmente com o haplótipo Non-Esmeraldo-like (Figura 13 B). Todas as cepas do DTU TcI apresentaram um mapeamento preferencial com o haplótipo Non-Esmeraldo-like, sugerindo uma maior proximidade entre os DTUs TcI e TcIII, como previamente descrito na literatura (DE FREITAS et al., 2006; FRANZÉN et al., 2011; PANUNZI; AGÜERO, 2014; WESTENBERGER et al., 2005). Comparações nos padrões de SNPs dos haplótipos Esmeraldo-like e Non-Esmeraldo também sugeriram uma maior proximidade entre TcI e TcIII (Figura 13 C). Resultados semelhantes foram encontrados recentemente por Tomasini e Diosque em 2017, com base em comparações das montagens genômicas dos haplótipos Esmeraldo-like (TcII), Non-Esmeraldo-like (TcIII) de CL Brener e Sylvio (TOMASINI; DIOSQUE, 2017). Interessantemente, todas as cepas de TcI apresentaram uma maior RDC para o cromossomo 14 do haplótipo Esmeraldo-like do que para o cromossomo 14 do haplótipo Non-Esmeraldo-like. Uma avaliação criteriosa da RDC ao longo destes cromossomos revelou que este padrão se devia a uma região pontual com alta RDC no cromossomo Esmeraldo-like, que corresponde ao gene Tc00.10470535007099.80 (ABC *transporter putative*), do qual a região sintênica no haplótipo Non-Esmeraldo-like corresponde a um gap. Desta forma, é provável que durante o mapeamento as reads correspondentes aos dois haplótipos foram mapeadas no haplótipo Esmeraldo-like. De modo semelhante, o cromossomo 3 do haplótipo Non-Esmeraldo like apresentou uma maior RDC no mapeamento das reads de Esmeraldo do que o cromossomo 3 do haplótipo Non-Esmeraldo-like. Este resultado também se deve a uma região pontual com alta cobertura no haplótipo Non-Esmeraldo-like, que corresponde ao gene Tc00.1047053508533.10 (*hypothetical protein conserved*). Aproximadamente 1.500 nucleotídeos da região 5' deste gene do haplótipo

Non-Esmeraldo-like estavam ausentes no alelo correspondente do haplótipo Esmeraldo-like (Tc00.1047053404001.20), de forma que as reads correspondentes a esta região em ambos haplótipos mapearam no haplótipo Non-Esmeraldo-like. Estes genes não estão incluídos nos 1.563 genes de cópia simples utilizado na metodologia SCoPE para estimar CCNV.

7.2 Variação no número de cópias cromossômicas

Todas as seis cepas de *T. cruzi* avaliadas neste capítulo apresentaram um genoma predominantemente diploide, baseado nas análises de SCoPE e frequência alélica quando todo o genoma foi avaliado (Figura 15). Este resultado está de acordo com estimativas prévias de ploidia em *T. cruzi* (ACKERMANN et al., 2012; BRISSE et al., 2003; MINNING et al., 2011) e espécies de *Leishmania* que foram classificadas como predominantemente diploides, com a exceção da cepa M2904 de *L. braziliensis* que apresentou um genoma primariamente triplóide (DOWNING et al., 2011; ROGERS et al., 2011; STERKERS et al., 2011).

Com relação a predição de ploidia em cromossomos individuais por SCoPE, as cepas do DTU TcI, Arequipa, Colombiana e Sylvio apresentaram um baixo número de duplicações cromossômicas (1, 1 e 3 respectivamente), enquanto as cepas dos DTUs TcII, Esmeraldo e Y; e TcIII, 231, apresentaram um grande número destas expansões (13, 6 e 12 respectivamente). Estas predições foram confirmadas pela razão da frequência alélica de posições heterozigotas, que estimou as mesmas aneuploidias, com a exceção dos cromossomos 20 e 23 de Sylvio, cromossomo 7 de Esmeraldo e cromossomos 20 e 23 de Arequipa, que foram preditos como tetrassômicos baseados na razão dos SNPs heterozigóticos. Para avaliar se as predições de CCNV não foram grandemente impactadas por duplicações segmentais ou perdas parciais de cromossomos, a RDC de todas as posições de todos os 41 cromossomos das seis cepas de *T. cruzi* foi estimada (Figura 16; "*Additional file* 4: *Figure* S2"(REIS-CUNHA et al., 2015)). Excluindo gaps cromossômicos, clusters de famílias multigênicas, e regiões de troca de fita codificadora, a RDC ao longo dos cromossomos está de acordo com as predições pela metodologia de SCoPE, validando esta abordagem (Figura 16).

A relevância da variação no número de cópias gênicas como um importante mecanismo para alterar a expressão gênica e aumentar a variabilidade genômica já foi extremamente documentado não apenas em *T. cruzi*, mas também em *T. brucei* e espécies do gênero *Leishmania* (BARTHOLOMEU et al., 2009, 2014; DE PABLOS; OSUNA, 2012; EL-SAYED, 2005a, 2005b; MARTÍNEZ-CALVILLO et al., 2010; MINNING et al., 2011; TIBAYRENC et al., 1986). CCNV corresponde a um novo nível de CNV, expandindo grandes blocos gênicos simultaneamente e

em uma geração. A ocorrência de CCNV já foi relacionada a um aumento de fitness em condições de estresse em *Saccharomyces cerevisiae*, *Candida albicans* e *Leishmania* sp. (ABBEY et al., 2011; LEPROHON et al., 2009; RANCATI et al., 2008; SHELTZER et al., 2011; UBEDA et al., 2008). A plasticidade de cariótipo encontrada em *T. cruzi* pode também representar um arcabouço para a seleção natural de fenótipos favoráveis, como uma maior expressão de fatores de virulência e um aumento na diversidade de sequência, resultando em um aumento de adaptabilidade para o parasito e facilitando alterações de expressão gênica na troca de hospedeiros.

Os mecanismos por trás da geração de aneuploidias em *T. cruzi* ainda não são conhecidos. Apesar de evidencias de reprodução sexuada já terem sido descritos em *T. brucei* (AKOPYANTS et al., 2009; PEACOCK et al., 2011), e eventos de troca de material genético já terem sido observados em *T. cruzi* (BAPTISTA et al., 2014; BRISSE et al., 2003; DE FREITAS et al., 2006; MACHADO; AYALA, 2001; MESSENGER; MILES, 2015; STURM et al., 2003; WESTENBERGER et al., 2005), a ocorrência de reprodução sexuada neste parasito ainda não foi demonstrada. Porém, um processo independente de meiose já foi proposto para explicar a geração de híbridos entre DTUs de *T. cruzi* (GAUNT et al., 2003). De acordo com este modelo, o núcleo de duas células diploides é fusionado, gerando uma prole poliploide seguido de eventos de recombinação entre seus alelos. Esta célula poliploide pode perder algumas cópias de seus cromossomos supranumerários, eventualmente retornando ao estado basal diploide, de modo semelhante ao ciclo parasexual de *Candida albicans* (BENNETT, 2015; SELMECKI; FORCHE; BERMAN, 2010; STERKERS et al., 2014). Este pode ser um dos mecanismos por trás da geração de CCNV em *T. cruzi*, visto que algumas cópias de cromossomos extras podem ser mantidas pelo parasito mesmo após a redução genômica. Análises de FACs revelaram que cepas híbridas de *T. cruzi* apresentam um aumento em seu conteúdo de DNA, quando comparado as suas cepas parentais (LEWIS et al., 2009). Além disso, um prolongado cultivo destas cepas híbridas experimentais levou um gradual e progressivo declínio no conteúdo de DNA da célula (LEWIS et al., 2009; MESSENGER; MILES; BERN, 2015), suportando o ciclo parasexual como um modelo para a geração de aneuploidias em *T. cruzi*.

A ocorrência de cromossomos supranumerários também já foi observada em várias espécies do gênero *Leishmania*, onde o padrão de aneuploidias varia entre espécies, dentro de espécies, cepas e até mesmo dentro de uma mesma população de parasitos (LACHAUD et al., 2014; STERKERS et al., 2011, 2014). Para explicar o mecanismo por trás desta aneuploidia em mosaico em *Leishmania* sp., modelos baseados em falha na segregação de cromossomos durante a replicação celular ou falhas na maquinaria de replicação de cromossomos foram propostos (STERKERS et al., 2011). Erros na segregação de cromossomos em *T. brucei* nocautes para SMC-3 (componente central de um complexo que une as cromátides irmãs durante a

mitose) resultaram em números totais pares de cromossomo após a divisão celular, onde as células resultantes apresentaram 3 e 1 ou 4 e 0 cópias de um determinado cromossomo. Porém, o número total de cromossomos assimétricos em *Leishmania* sp. é geralmente ímpar: 1 e 2, ou 2 e 3 ou 3 e 4, sugerindo que falhas na duplicação de cromossomos tenham maior relevância para a geração de aneuploidias neste parasito (STERKERS et al., 2011, 2014).

De todos os DTUs avaliados em *T. cruzi*, TcI apresentou o menor número de expansões cromossômicas, exibindo o cariótipo mais estável (Figura 15). Este reduzido número de aneuploidias pode explicar alguns fatores encontrados neste subgrupo de *T. cruzi*, como uma reduzida heterozigosidase e uma menor massa de DNA por célula. O baixo número de cromossomos duplicados, somados a uma melhor eficiência da maquinaria de *missmatch repair* (MMR) em TcI são fatores adicionais para explicar a redução de heterozigosidase neste DTU (AUGUSTO-PINTO et al., 2003; FRANZÉN et al., 2011, 2012; LLEWELLYN et al., 2009; MACHADO et al., 2006a). Enquanto a maquinaria de MMR corrige grande parte das mutações, o baixo número de duplicações cromossômicas impede que regiões mutem sem perda de função, reduzindo o potencial de geração de heterozigosidase. Análises de citometria de fluxo, assim como por intensidade de fluorescência mostraram que as cepas de TcI apresentam o menor tamanho genômico dentre os DTUs de *T. cruzi* (LEWIS et al., 2009; SOUZA et al., 2011), o que foi associado a uma redução no conteúdo de famílias multigênicas e retrotransposons neste DTU (FRANZÉN et al., 2011, 2012; SOUZA et al., 2011). Com base em nossos resultados, acreditamos que a redução na massa do genoma pode também estar associada ao menor número de cromossomos neste grupo, visto que alterações no número de cromossomos também altera a massa de DNA da célula. De modo similar, Rogers e colaboradores encontraram uma correlação positiva entre duplicações cromossômicas e conteúdo de DNA por célula em *Leishmania* sp. (ROGERS et al., 2011). Apesar de TcI possuir uma menor heterogeneidade intragenômica, este DTU possui uma grande variação de sequências entre suas cepas, maior do que o encontrado entre cepas pertencentes aos DTUs TcII e TcVI (CERQUEIRA et al., 2008; MINNING et al., 2011). Isto pode ser uma consequência de um menor número de eventos de hibridização intra-DTU em TcI, o que levaria cada indivíduo em uma população a se comportar de forma semelhante a um clone. Desta forma, a grande variabilidade de sequências entre as cepas de TcI pode representar uma redução em eventos de trocas gênicas entre cepas deste DTU, e consequentemente, uma menor chance de eventos de variações de ploidia.

Entre as cepas do DTU TcII, Esmeraldo apresentou 13 expansões cromossômicas, enquanto Y apresentou 6 expansões e uma perda, sugerindo que existem grandes diferenças no padrão de aneuploidias dentro do DTU TcII (Figura 15). Recentemente, eventos de recombinação e troca gênica entre cepas do DTU TcII que coexistem na mesma região geográfica

foram demonstrados, baseado em marcadores nucleares e mitocondriais (BAPTISTA et al., 2014). Uma das possíveis explicações por trás do maior número de expansões cromossômicas em Esmeraldo, seria a ocorrência de um evento de hibridização recente intra-DTU TcII para gerar a cepa Esmeraldo. Os haplótipos que constituem Esmeraldo são mais divergentes entre si do que os que constituem Y (DE FREITAS et al., 2006), o que sugere a ocorrência de um evento mais recente de troca gênica em Esmeraldo, tornando-o mais susceptível a aneuploidias (DE FREITAS et al., 2006; STURM; CAMPBELL, 2010b). O padrão de aneuploidias em *Leishmania* sp. também varia entre espécies (ROGERS et al., 2011), dentro de uma mesma espécie (DOWNING et al., 2011; ROGERS et al., 2011), e até mesmo dentro de uma mesma população (STERKERS et al., 2011, 2014). De modo semelhante ao encontrado no DTU TcI neste estudo, diferentes cepas de *L. major* (Friedlin e LV39) apresentaram o mesmo padrão de CCNV. Por outro lado, como observado em TcII, cepas das espécies *L. mexicana* e *L. donovani* apresentaram grandes variações em seu padrão de CCNV (BRANCHE et al., 2006; ROGERS et al., 2011). Este grande número de eventos de aneuploidias em *T. cruzi* e *Leishmania* requer cuidado na seleção de marcadores para estudos genéticos baseados em hipótese de diploidia, como a geração de nocautes e analises filogenéticas.

Análises de clusterização hierárquica baseada na ploidia predita de cada cromossomo de *T. cruzi* agrupou todas as cepas de TcI juntas com elevada pontuação de confiança, (Figura 18). Por outro lado, as cepas de TcII e TcIII apresentaram um padrão de CCNV muito variável, o que sugere que exista uma maior plasticidade genômica entre membros destes DTUs. É interessante ressaltar que as duas cepas de TcII, Esmeraldo e Y, apresentaram um padrão muito discordante de aneuploidias, sugerindo que eventos duplicação ou perda cromossômica variáveis entre cepas são comuns neste DTU. A análise de um maior número de cepas poderia auxiliar na correta estimativa da taxa na qual os eventos de CCNV ocorrem em *T. cruzi*.

7.3 Cromossomo 11 da cepa Y de *T. cruzi*

Nos primeiros 248 kb do cromossomo 11 da cepa Y, foram identificados 22 genes de cópia simples que apresentaram uma RDC média que era a metade da RDC encontrada nos outros genes de cópia simples neste cromossomo (Figura 19). Esta grande variação de RDC começa em uma região de mudança de fita codificadora, que são regiões usualmente relacionadas a rearranjos e trocas genicas em genomas de tripanossomatídeos (EL-SAYED, 2005b; GHEDIN et al., 2004). Uma possível explicação para este resultado é a perda de braço em uma das cópias do cromossomo 11, de forma que um cromossomo apresenta o braço completo

e o outro um braço truncado. Outra possível hipótese seria a de que o cromossomo 11 de Y é dividido em 2 cromossomos distintos, 11a e 11b, na posição de 248kb de CL Brener, onde 11a é monossômico e 11b é dissômico. Eventos de quebras ou fusões cromossômicas como o possivelmente ocorrido em Y, podem parcialmente explicar os variáveis padrões de banda em géis de eletroforese de campo pulsátil encontrados entre as cepas de *T. cruzi* (BRANCHE et al., 2006; LIMA et al., 2013a; PEDROSO; CUPOLILLO; ZINGALES, 2003; SOUZA et al., 2011; TRIANA et al., 2006; VARGAS; PEDROSO; ZINGALES, 2004). De modo alternativo, como a amostra de Y sequenciada no presente trabalho não corresponde a uma população clonada, uma parte da população de células pode apresentar a estrutura completa do cromossomo 11 em homozigose enquanto uma outra parte da população apresentaria apenas o cromossomo truncado.

7.3 Cromossomo 31 de *T. cruzi*

O cromossomo 31 foi o único supranumerário em todas as seis cepas de *T. cruzi* avaliadas neste estudo (Figura 20). Análises de enriquecimento gênico por ontologia revelaram que este cromossomo apresenta um enriquecimento de genes relacionados à biossíntese de glicoproteínas e a processos de glicosilação (Figura 20 C). O genoma da cepa CL Brener apresenta aproximadamente 100 genes anotados como glicotransferases putativos, que estão envolvidos na síntese de uma gama de glicoconjugados abundante e diferencialmente expressos em todos os estágios do ciclo do *T. cruzi* (DE LEDERKREMER; AGUSTI, 2009). Destes 100 genes, 54 correspondem a glicosiltransferases dependentes de UDP-GlcNAc. O cromossomo 31 apresenta 9 das 27 cópias deste gene no haplótipo Esmeraldo-like e 13 das 27 cópias deste gene no haplótipo Non-Esmeraldo-like. A glicosiltransferase dependente de UDP-GlcNAc está envolvida na transferência de N-acetilgicosamina (GlcNAc) de um precursor UDP-GlcNAc para um grupo hidroxila em resíduos de serina e treonina, resultando em oligossacarídeos O-linked nas mucinas de *T. cruzi* (BUSCAGLIA et al., 2006). Mucinas são glicoproteínas de superfície de *T. cruzi* altamente glicosiladas, nas quais o conteúdo de glicanos pode corresponder a até 60% do peso molecular da mucina (ACOSTA-SERRANO et al., 2001; BUSCAGLIA et al., 2006). Estas glicoproteínas são o componente mais abundante na superfície do parasito, e estão diretamente envolvidas em processos de adesão e invasão celular e evasão do sistema imune (BUSCAGLIA et al., 2006; DE PABLOS; OSUNA, 2012). Por este motivo, a expansão do cromossomo 31 em todas as cepas avaliadas pode ser uma consequência da necessidade de glicosilar o grande número de mucinas simultaneamente expressas na superfície do parasito. Análises de hibridização genômica competitiva entre cepas de *T. cruzi* também identificaram CNV em outro gene envolvido na síntese de glicanos em mucinas de *T. cruzi*, o gene beta-galactofuranosil

transferase (JONES et al., 2004; MINNING et al., 2011), reforçando a importância deste processo biológico para a sobrevivência do parasito. Apesar do cromossomo 31 também estar expandido em diversas espécies de *Leishmania* (ROGERS et al., 2011), não foram encontramos grandes regiões de sintenia entre o cromossomo 31 deste parasito com o de *T. cruzi*, sugerindo que a expansão deste cromossomo em *Leishmania* sp. é decorrente de diferentes pressões seletivas.

7.4 Conclusões parciais do capítulo 1

- A metodologia SCoPE proposta neste trabalho é mais adequada do que a metodologia WCPE para a estimativa de ploidia de cromossomos com alto conteúdo repetitivo. Esta metodologia apresentou resultados mais precisos nos cálculos de ploidia cromossomal em *T. cruzi*, em especial nos cromossomos ricos em famílias multigênicas como os cromossomos 18, 28, 38 e 41.

-O padrão de aneuploidias varia entre os DTUs TcI, TcII e TcIII de *T. cruzi*, onde TcI apresentou um menor número de expansões cromossômicas quando comparado aos outros dois DTUs avaliados. Estes resultados estão de acordo com a menor massa de DNA por célula e também a menor variabilidade intragenômica encontrada na maioria das cepas do DTU TcI de *T. cruzi*. Por outro lado, as duas cepas do DTU TcII, Esmeraldo e Y apresentaram um padrão diferente de cromossomos duplicados, sugerindo que eventos de CCNV sejam comuns neste DTU.

-O cromossomo 31 foi o único cromossomo expandido em todas as seis cepas de *T. cruzi* avaliadas. Análises de enriquecimento gênico com base em ontologia deste cromossomo revelaram um enriquecimento de genes ligados a glicosilação e biossíntese de glicanos, como enzimas relacionadas à glicosilação de proteínas de superfície, reforçando a importância deste processo biológico para a sobrevivência do parasito.

CAPÍTULO 2: Variabilidade genômica entre isolados de campo do DTU TcII de *T. cruzi*.

8. Justificativa:

A maioria das infecções humanas por *T cruzi* em países do cone Sul da América do Sul, como Brasil, Argentina e Chile, ocorrem por cepas pertencentes ao DTU TcII, sendo responsáveis por casos agudos graves e também por sintomatologia crônica com cardiomegalia chagásica, megaesôfago/megacólon ou mista (MILES et al., 2009; ZINGALES et al., 2012). Recentemente, eventos de recombinação e trocas gênicas substanciais entre cepas do grupo TcII que coexistem na mesma região foram demonstradas através de genotipagem por microssatélites (BAPTISTA et al., 2014). Porém, uma comparação baseada em toda a extensão dos genomas mitocondriais e nucleares entre isolados de campo de TcII ainda não foi realizado. No presente capítulo, nós sequenciamos e analisamos os genomas nucleares e mitocondriais de sete cepas de TcII isoladas de pacientes chagásicos nos estágios indeterminado e cardíaco da doença de Chagas. Estes isolados foram originados das regiões central e nordeste do estado de Minas Gerais, Brasil, regiões endêmicas para a infecção por cepas de *T. cruzi* do DTU TcII. A filogenia destas amostras de campo foi determinada, juntamente com de cepas de *T. cruzi* dos DTUs TcI, TcII, TcIII e TcVI previamente sequenciadas, com base em genes conservados nucleares ou mitocondriais. Finalmente, foram realizadas análises de RDC e frequência alélica para estimar CCNV nucleares e heteroplasmia mitocondrial entre estas amostras de campo de TcII.

9. OBJETIVOS

9.1 Objetivo Geral

Determinar a variabilidade genômica de isolados de campo do DTU TcII de *T. cruzi* e correlacionar com a sua distribuição geográfica.

9.2 Objetivos específicos

1-Obter reads de WGS referentes a 7 amostras de *T. cruzi* do DTU TcII recém isoladas de campo.

2-Determinar a classificação filogenética e eventos de recombinação das amostras de campo de *T. cruzi* com base em genes ortólogos conservados.

3-Determinar a ocorrência de CCNV entre isolados de campo do grupo TcII de *T.* cruzi

4- Correlacionar as distâncias geográficas com a filogenia e com o padrão de aneuploidias encontradas no DTU TcII.

10. Metodologia

10.1 Cepas do parasito e sequenciamento

Um total de 19 bibliotecas de sequenciamento genômico por *shotgun* (WGS) de *T. cruzi* de cepas dos DTUs TcI, TcII, TcIII, TcV e TcVI foram utilizadas neste trabalho. Destas 19, 11 amostras correspondem a novos sequenciamentos genômicos nucleares e mitocondriais. Destes 11, sete consistem em parasitos de TcII isolados por hemocultura, obtidos de pacientes da região central (S15 e S162) e nordeste (S11, S23b, S44a, S92a e S154a) do estado de Minas Gerias, Brasil, gentilmente cedidos pela prof. Lúcia M. Galvão e Prof. Egler Chiari. Outras três amostras correspondem a clones da população de Y (Ycl2, Ycl4 e Ycl6), clonadas a partir da população de Y descrita no capítulo 1. A última destas 11 amostras corresponde a cepa CL Brener (TcVI) de *T. cruzi*. Estas amostras foram sequenciadas pela empresa Macrogen (Seoul, Coréia do Sul), com ~60x de cobertura em sequenciador Hiseq2000, gerando pair-end reads de 100pb com inserto de 350pb. As outras oito bibliotecas de reads de cepas de *T. cruzi* foram obtidas no Arquivo de Sequencias de Reads (SRA) do *National Center for Biotechnology Information* (NCBI), consistindo em amostras do DTU TcI (Arequipa, Colombiana e Sylvio); TcII (Esmeraldo e Y não clonada); TcIII (231); TcV (9280) e TcVI (Tulahuen). A descrição detalhada de cada biblioteca de reads está na tabela 2.

A qualidade de cada biblioteca de reads foi avaliada utilizando o programa FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), e filtrada utilizando o programa *Trimmomatic* (BOLGER; LOHSE; USADEL, 2014). O cutoff de filtragem das reads foi um valor médio de phred acima de 20 para as bibliotecas de 454 e de 30 para as de Illumina, assim como um tamanho mínimo de 50 nucleotídeos para ambas.

A sequência nucleotídica dos cromossomos dos haplótipos Esmeraldo-like, Non-Esmeraldo-like e contigs não utilizados na montagem de CL Brener versão 26 foram obtidos a partir do banco de dados TriTrypdb (ASLETT et al., 2009).

Cepa	Plataforma de sequenciamento	DTU	Número de acesso	Forma clínica do paciente	Local de isolamento
Arequipa	454/Ion Torrent	TcI	SRS838181	-	Arequipa/Peru
Colombiana	454/Ion Torrent	TcI	SRS841912	-	Colômbia
Sylvio	454	TcI	-	-	Pará/Brasil
S11	Illumina	TcII	-	Indeterminado	Itaipé/Brasil
S15	Illumina	TcII	-	Indeterminado	Felixlândia/Brasil
S154a	Illumina	TcII	-	Indeterminado	Itaipé/Brasil
S162a	Illumina	TcII	-	Indeterminado	Congonhas do Norte/Brasil
S23b	Illumina	TcII	-	Cardíaco	Porteirinha/Brasil
S44a	Illumina	TcII	-	Indeterminado	Turmalina/Brasil
S92a	Illumina	TcII	-	Cardíaco	Teófilo Otoni/Brasil
Esmeraldo	454/Illumina	TcII	SRR833799/ SRR833800/ SRR058517/ SRR058509/ SRR058520/ SRR058518/ SRR058519/ SRR058515/ SRR058516/ SRR058513/ SRR058514/ SRR058510/ SRR058511/ SRR058512	-	Bahia/Brasil
Y-população	454/Ion Torrent	TcII	SRS842149	-	São Paulo/Brasil
Y-cl2	Illumina	TcII	-	-	São Paulo/Brasil
Y-cl4	Illumina	TcII	-	-	São Paulo/Brasil
Y-cl6	Illumina	TcII	-	-	São Paulo/Brasil
231	Illumina	TcIII	ERR864236	Indeterminado	Minas Gerais/Brasil
9280	Illumina	TcV	SAMN03568097	-	Santa Cruz/Bolívia
Tulahuen	Illumina	TcVI	SRX268895/ SRX268893	-	Tulahuen/Chile
CL Brener	Illumina	TcVI	-	-	Rio Grande do Sul/Brasil

Tabela 2: Descrição das bibliotecas de reads utilizadas neste capítulo.

10.2 Cultivo, clonagem e extração de DNA genômico e mitocondrial de *T. cruzi*

A cepa Y de *T. cruzi* foi cultivada e clonada de acordo com o protocolo descrito no **subitem 5.1** do capítulo 1 desta tese. O DNA genômico nuclear e mitocondrial de três clones de Y, denominados Ycl2, Ycl4 e Ycl6 foram extraídos de acordo com o protocolo no **subitem 5.2** do mesmo capítulo.

10.3 Montagem do genoma nuclear

Inicialmente, os genomas nucleares de 16 das 19 amostras de *T. cruzi* que não possuíam o genoma montado (7 amostras de campo de TcII, três clones de Y, população Y não clonada, Arequipa, Colombiana, 231, 9280 e Tulahuen) foram montadas *de novo* utilizando o programa *Velvet optimizer* versão 1.2.10 (ZERBINO; BIRNEY, 2008) para as bibliotecas de reads de Illumina; e Celera 8.3 (MILLER et al., 2008; MYERS, 2000) para as bibliotecas de reads de 454. Os contigs das montagens genômicas da cepa Sylvio (versão 29) e dos haplótipos Esmeraldo-like (versão 26), Non-Esmeraldo-like (versão 26) de CL Brener foram obtidos a partir do TritrypDB, enquanto os contigs de Esmeraldo foram obtidos a partir do *European Nucleotide Archive*. Os links para a obtenção destas montagens estão listados na tabela 3.

Genoma	Link
Esmeraldo	http://www.ebi.ac.uk/ena/data/view/ANOX01000001-ANOX01020187
Sylvio	http://tritrypdb.org/common/downloads/release-29/TcruziSylvioX10-1/fasta/data/TriTrypDB-29_TcruziSylvioX10-1_Genome.fasta
CL Brener haplótipo Esmeraldo-like	http://tritrypdb.org/common/downloads/release-26/TcruziCLBrenerEsmeraldo-like/fasta/data/TriTrypDB-26_TcruziCLBrenerEsmeraldo-like_Genome.fasta
CL Brener haplótipo Non-Esmeraldo	http://tritrypdb.org/common/downloads/release-26/TcruziCLBrenerNon-Esmeraldo-like/fasta/data/TriTrypDB-26_TcruziCLBrenerNon-Esmeraldo-like_Genome.fasta

Tabela 3: Links para os genomas montados utilizados neste capítulo.

10.4 Montagem e comparação de sequências de maxicírculo

Para a montagem da sequência de maxicírculo das cepas de TcI (Arequipa e Colombiana), TcII (S11, S15, S23b, S44a, S92a, S154a, S162a, população não clonada de Y, clones de Y), TcIII (231) e TcVI (Tulahuen), incialmente todas as reads mitocondriais das bibliotecas WGS foram recuperadas. Para tanto, realizou-se um mapeamento simultâneo contra as três montagens de maxicírculo atualmente disponíveis, as quais foram usadas como referências:

Sylvio (TcI), Esmeraldo (TcII) e CL Brener (TcVI - maxicírculo derivado de ancestral TcIII), utilizando o mapeador BWA-mem (LI, 2013; LI; DURBIN, 2009). Os links para acesso das referências mitocondriais são: TcI Sylvio (Número de acesso do NCBI: FJ203996.1), TcII Esmeraldo (Número de acesso do NCBI: DQ343646.1) and TcVI CL Brener (Número de acesso do NCBI: DQ343645.1). A sequência de referência de maxicírculo que apresentou a maior cobertura neste mapeamento competitivo com cada uma das bibliotecas de reads foi selecionada como guia para a montagem baseada em referência do maxicírculo de cada cepa. Esta montagem foi realizada através do mapeamento de cada biblioteca de reads com a referência selecionada utilizando o programa BWA-mem, seguido pela submissão do arquivo resultante do mapeamento de cada cepa com cada referência a um pipeline que utiliza *SAMtools* mpileup, *BCFtools* vcfutils.pl e seqtl (DANECEK et al., 2011; LI et al., 2009). Para visualização dos padrões de similaridade entre os maxicírculos de todas as cepas avaliadas, análises de similaridade par-a-par utilizando o programa BLASTn (ALTSCHUL et al., 1990; MORGULIS et al., 2008) foram realizadas, utilizando um cutoff mínimo de *e-value* de $1e^{-20}$, onde o resultado desta análise foi submetido para o pacote Circoletto, do programa Circos (DARZENTAS, 2010).

10.5 Análises filogenéticas

A filogenia baseada em marcadores nucleares de 17 das 19 amostras de *T. cruzi* foi estimada com base nos 1563 genes de cópia simples conservados entre os haplótipos Esmeraldo-like e Non-Esmeraldo-like de CL Brener, descritos no **subitem 5.6** do capítulo 1. As cepas Tulahuen (TcVI) e 9280 (TcV) foram excluídas desta análise, pois a sua natureza híbrida comprometeu a qualidade das montagens *de novo*. Estas duas amostras foram incluídas na filogenia baseada em marcadores mitocondriais, que foi, portanto, realizada com as 19 amostras.

Para a filogenia nuclear, as sequências dos 1563 genes de cópia simples descritos no **subitem 5.6** do capítulo 1 foram recuperadas de cada uma das montagens utilizando BLAT (KENT, 2002), onde apenas os genes encontrados em todas as montagens previamente mencionadas foram utilizados em análises filogenéticas. Já na filogenia baseada em marcadores mitocondriais, todos os genes codificadores de proteínas foram utilizados. Em ambas as filogenias, cada um dos genes recuperados de cada cepa foi alinhado utilizando o programa MUSCLE (EDGAR, 2004), onde as regiões mal alinhadas ou ausentes foram removidas utilizando o programa Gblocks (CASTRESANA, 2000; TALAVERA; CASTRESANA, 2007). O melhor modelo de substituição nucleotídica para as análises filogenéticas foi determinado através do programa Jmodeltest (POSADA, 2008). As árvores filogenéticas foram geradas por máxima verossimilhança

utilizando o programa PhyML (GUINDON et al., 2009, 2010), com o modelo *Generalized Time Reversible* (GTR), com 1000 replicadas de bootstrap, proporção de sítios invariáveis de 0.9 e distribuição gama de 0.93 para a filogenia nuclear e de 0.27 para o genoma mitocondrial. As imagens das árvores filogenéticas finais foram geradas utilizando o programa FigTree v.1.4.2 (RAMBAUT, 2009) (http://tree.bio.ed.ac.uk/software/figtree/). Um Tanglegrama comparativo baseado na filogenia dos genes nucleares e mitocondriais foi gerado utilizando o programa Dendroscope (HUSON et al., 2007).

10.6 Estrutura populacional das amostras de campo de TcII

A estrutura populacional da sete amostras de campo de TcII foi estimada com base em SNPs presentes nos 1563 genes de cópia simples nucleares, através do programa ADMIXTURE (ALEXANDER; NOVEMBRE, 2009). Inicialmente, um total de 24.172 posições polimórficas foram obtidas utilizando os pacotes do *Genome Analysis Tool Kit* (GATK) (MCKENNA et al., 2010) (https://software.broadinstitute.org/gatk/,): *SelectVariants*, *CombineVariants* assim como o programa Bedtools (QUINLAN; HALL, 2010). O arquivo *Variant Call Format* (VCF) foi convertido para o formato ped e posteriormente map utilizando vcftools_0.1.12b (DANECEK et al., 2011) e PLINK (PURCELL et al., 2007). Para estimar o número de populações ancestrais que deu origem aos sete isolados de campo de TcII, o programa ADMIXTURE foi utilizado com valores de K (populações ancestrais) variando de 1 a 9, onde o valor de K que apresentou o menor valor de erro de validação cruzada foi 7. A imagem da estrutura populacional da saída do ADMIXTURE foi gerada em R (www.r-project.org, R Development 2010).

10.7 Análise de Componente Principal (PCA)

Para estimar a distância de sequência entre as sete amostras de campo de TcII baseado no padrão de SNPs, uma sequência genômica nuclear consenso foi gerada para cada amostra, utilizando o mapeamento das reads nos cromossomos do haplótipo Esmeraldo-like de CL Brener juntamente com o programa GATK *FastaAlternateReferenceMaker* (https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_fasta_FastaAlternateReferenceMaker.php). Após, uma matriz de distâncias baseada em SNPs diferenciais entre as amostras foi gerada e utilizada como entrada para o pacote do R "caret" (http://topepo.github.io/caret/index.html), para realizar a plotagem do PCA.

10.8 Variação no número de cópias cromossômicas

A avaliação de CCNV de cada uma das cepas de *T. cruzi* utilizadas neste trabalho foi realizada utilizando a metodologia SCoPE, descrita no **subitem 5.8** do capítulo 1, baseada na cobertura média dos genes de cópia simples em cada um dos 41 cromossomos de CL Brener como uma estimativa do seu número de cópias. Inicialmente, as sequências de referência dos cromossomos de CL Brener versão 26 foram obtidos do Tritrypdb (http://tritrypdb.org/tritrypdb/). Após, as bibliotecas de reads de TcI e TcIII foram mapeados nos cromossomos do haplótipo Non-Esmeraldo-like, enquanto reads das cepas de TcII, TcV e TcVI foram mapeadas nos cromossomos do haplótipo Esmeraldo-like (WEATHERLY; BOEHLKE; TARLETON, 2009) utilizando o algoritmo BWA-MEM (LI, 2013). As reads mapeadas foram filtradas por qualidade de mapeamento 30 utilizando SAMtools v1.1 (LI et al., 2009), e a RDC de cada posição em cada cromossomo foi determinada com o programa BEDtools genomecov v2.16.2 (QUINLAN; HALL, 2010) e scripts em Perl i*n house*. Finalmente, a RDC média de todos os genes cópia simples em um determinado cromossomo foi assumido como a RDC média do cromossomo.

10.9 Conteúdo de SNPs e frequência alélica

SNPs provenientes do mapeamento das reads de todas as cepas de *T. cruzi* avaliadas foram obtidas utilizando a função mpileup do programa SAMtools (LI et al., 2009), onde apenas posições com profundidade de reads de ao menos 10, sendo 5 reads em cada variante foram avaliadas. Para cada cromossomo, a proporção de profundidade de reads em cada alelo predito em posições heterozigóticas foi obtida e arredondada para a segunda casa decimal. As frequências das proporções dos SNPs foram agrupadas em 100 categorias, variando de 0.01 até 1, e a distribuição aproximada de suas frequências de variantes, para cada cromossomo, foi plotada em R. Para estimar a ploidia predominante em cada genoma, a mesma metodologia foi aplicada, mas as posições heterozigóticas em todas as posições de todas as CDSs foram utilizadas simultaneamente.

10.10 Clusterização hierárquica do padrão de CCNV entre as cepas de *T. cruzi*

A análise de clusterização hierárquica baseada nas predições de CCNV de todas as 19 cepas de *T. cruzi* foi realizada com base no pacote Pvclust (SUZUKI; SHIMODAIRA, 2006) implementado em R (www.r-project.org, R Development 2010). Uma matriz baseada na distância euclidiana dos padrões de CCNV baseado em RDC das 19 cepas foi utilizada para

construir um dendrograma, pelo método de *complete linkage*. Para avaliar a incerteza em análises de clusterização hierárquica, dois métodos de reamostragem foram implementados no Pvclust: *bootstrap probabilitty* (BP), o método de reamostragem normal do bootstrap; e o *approximately unbiased* (AU), de reamostragem de bootstrap multiescalar. Ambos os métodos foram calculados com 10.000 iterações.

11. Resultados:

Para avaliar a diversidade genômica no DTU TcII de *T. cruzi*, nós sequenciamos, montamos e comparamos os genomas mitocondrial e nuclear de sete isolados de campo de *T. cruzi* isolados por hemocultura de pacientes da região central (S15 e S162) e nordeste (S11, S23b, S44a, S92a e S154a) do estado de Minas Gerais, Brasil (Figura 21).

11.1 Filogenia nuclear e mitocondrial de *T. cruzi*

Do total de 1563 genes nucleares de cópia simples, conservados entre os haplótipos Esmeraldo-like e Non-Esmeraldo-like de CL Brener, 961 foram parcialmente recuperados da montagem de novo do genoma nuclear das sete amostras de campo de TcII. Estas sequências conservadas foram utilizadas para estimar a filogenia destas cepas por máxima verossimilhança. Para melhor classificar as cepas de TcII, amostras dos DTUs TcI (Arequipa, Colombiana e Sylvio), TcII (população de Y não clonada, Ycl2, Ycl4, Ycl6 e Esmeraldo), TcIII (231) e TcVI (CL Brener haplótipos Esmeraldo-like e Non-Esmeraldo-like) foram também incluídas nesta análise (Figura 21 B). Todas as amostras de TcII agruparam juntas; em um cluster separado de cepas de TcI e TcIII. Como esperado, o haplótipo Esmeraldo-like de CL Brener que é derivado de um ancestral TcII agrupou com as amostras de TcII. De modo similar, o haplótipo de CL Brener Non-Esmeraldo-like, derivado de um ancestral TcIII agrupou com a cepa 231 (TcIII). Em relação às amostras de campo de TcII, dois pares de amostras de regiões geográficas próximas: S15-S162 e S11-S92; também agruparam nas análises filogenéticas, sugerindo que existe uma relação entre diversidade genômica e distância geográfica dentro do DTU TcII. Por outro lado, as cepas S11 e S154 que foram isoladas da mesma localidade apresentaram uma grande distância filogenética, sugerindo que apesar de existir um viés geográfico na variabilidade genômica em TcII, existem diferentes cepas coexistindo em uma mesma região (Figura 21 B). Uma análise de componente principal (PCA) de todos os SNPs nucleares apresentados pelas sete amostras de campo de TcII também clusterizou S15-S162 e S11b-S92a, e classificou S154a como a cepa mais divergente entra as cepas avaliadas (Figura 21 C).

Para estimar a estrutura populacional dentro das amostras de campo de TcII, o programa ADMIXTURE (ALEXANDER; NOVEMBRE, 2009) foi utilizado (Figura 21 D). O número de populações ancestrais (K) dentro das amostras de campo de TcII foi estimado como sete, com base na análise do erro de validação cruzada utilizando valores de K variando de 1 a 9 (Figura 21 E). Todas as cepas de TcII avaliadas são compostas por duas ou mais populações ancestrais, o que sugere que eventos de troca de material genético foram frequentes durante a história de TcII em Minas Gerais. As populações ancestrais, numeradas de 1 a 7, correspondem, respectivamente, a: 13,42%, 21,55%, 8,35%, 17,08%, 15,40%, 13,55% e 10,65% das populações atuais. Interessantemente, as amostras S15 e S162 apresentaram uma proporção similar de cada população ancestral, reforçando a proximidade evolutiva entre elas. A amostra S154a foi a única composta por 5 populações ancestrais, reforçando a complexidade genômica deste isolado (Figura 21 D).

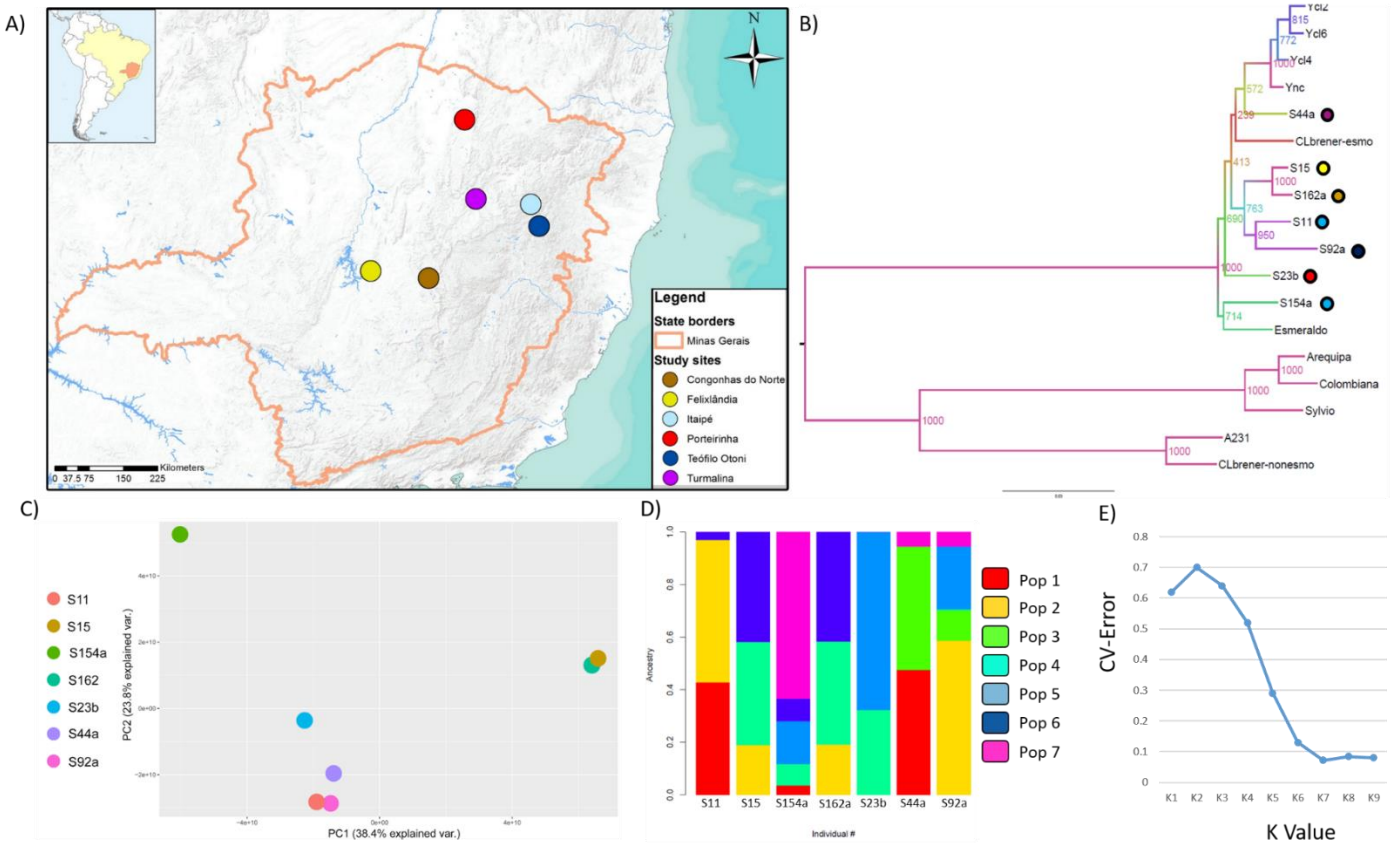
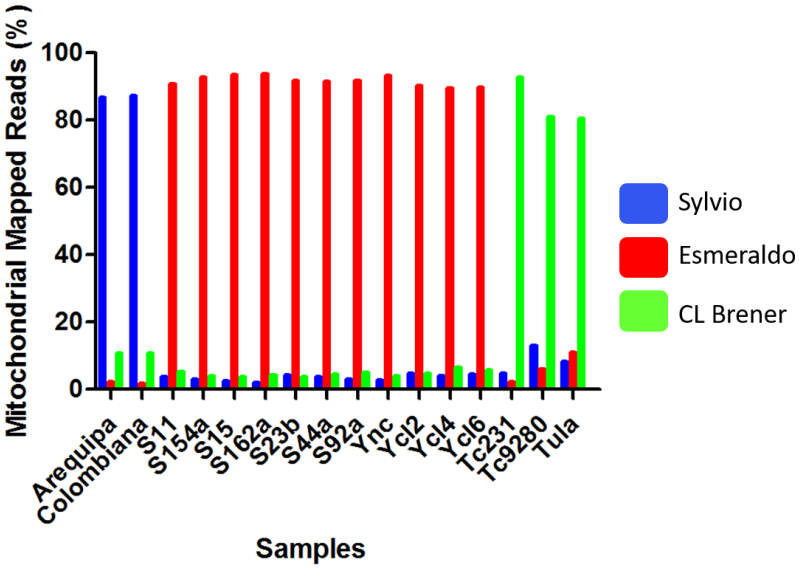


Figura 21: Distribuição geográfica e análise filogenética com base em marcadores nucleares das amostras de campo do DTU TcII de *T. cruzi*. (A) Das sete amostras de campo de TcII avaliadas neste estudo, cinco foram isoladas da região nordeste de Minas Gerais: S11a (Itaipé), S154a (Itaipé), S44a (Turmalina), S23b (Porteirinha) and S92a (Teófilo Otoni); enquanto duas foram obtidas da região central do estado: S15 (Felixlândia) and S162a (Congonhas do Norte). **(B)** Análise filogenética por máxima verossimilhança das amostras de campo de TcII juntamente com amostras previamente sequenciadas dos DTUs TcI, TcII, TcIII, TcV e TcVI, baseada em 961

genes nucleares de cópias simples conservados. A análise filogenética foi realizada com 1.000 replicatas de bootstrap, onde uma escala de cores entre rosa e marrom correspondem, respectivamente, a valores altos e baixos de bootstrap. Os círculos coloridos correspondem a localização geográfica de origem das amostras. **(C)** Plot de PCA baseado no padrão de SNPs presentes em cada amostra de campo de TcII. **(D)** Estimativa da proporção de cada população ancestral em cada isolado de campo de TcII, onde a proporção de cada cor corresponde a porcentagem da população ancestral presente no isolado atual. **(E)** O número de populações ancestrais K=7 foi estimado com base no menor valor de erro de validação cruzada encontrado.

Para selecionar a referência de maxicírculo mais adequada para a montagem com base em referência do genoma mitocondrial de cada uma das cepas de *T. cruzi* avaliadas neste trabalho, cada biblioteca de reads foi mapeada de forma competitiva nas três referências de maxicírculo atualmente disponíveis: Sylvio (TcI), Esmeraldo (TcII) e CL Brener (TcVI com maxicírculo derivado de ancestral TcIII). A referência com maior cobertura neste mapeamento competitivo foi selecionada como modelo para a montagem com base em referência de cada cepa. Desta forma, a sequência de maxicírculo de Sylvio foi selecionada para as cepas Arequipa e Colombiana; a de Esmeraldo foi selecionada como referência para todas as amostras de campo de TcII clones e população de Y; e a de CL Brener foi selecionada como referência para as cepas 231, 9280 e Tulahuen (Figura 22).



	Silvio	Esmeraldo	CLBrener
Arequipa	86,781780	2,361214	10,857010
Colombiana	87,198490	1,900988	10,900530
S11	3,825319	90,811740	5,362947
S154a	3,069480	92,822590	4,107932
S15	2,592189	93,506440	3,901375
S162a	1,987306	93,797330	4,215364
S23b	4,353747	91,849620	3,796633
S44a	3,810072	91,635120	4,554806
S92a	3,085690	91,827770	5,086535
Ync	2,688027	93,282170	4,029797
Ycl2	4,867468	90,269400	4,863130
Ycl4	4,127257	89,441100	6,431643
Ycl6	4,526820	89,789840	5,683338
Tc231	4,759498	2,361373	92,879130
Tc9280	13,064180	5,981914	80,953900
Tula	8,411653	11,133640	80,454700

Figura 22: Mapeamento competitivo das reads mitocondriais nas três referências de maxicírculo disponíveis. Para a determinação da sequência de maxicírculo mais adequada para a montagem com base em referência, cada biblioteca de reads foi mapeada simultaneamente com os três modelos de maxicírculo disponíveis, onde a porcentagem de reads preferencialmente mapeadas no maxicírculo de Sylvio (TcI) estão em azul, de Esmeraldo (TcII) em vermelho e de CL Brener (TcVI) em verde. Os valores brutos das porcentagens das reads que mapearam com cada referência podem ser vistos na tabela ao lado.

Uma comparação entre a conservação de sequência gênica entre as 19 sequências de maxicírculo montadas utilizando como base de comparação o maxicírculo de CL Brener revelou diferenças entre as sequências de kDNA entre as cepas de *T. cruzi* (Figura 23 A). A sequência de maxicírculo de CL Brener apresenta uma maior similaridade com as sequências de Tulahuen, 231 e 9280; valores de identidade intermediários com cepas de TcI (Arequipa, Colombiana e Sylvio) e uma menor identidade com cepas de TcII (S15, S23b, S44a, S154a, S162a, S11, S92a, Esmeraldo, clones de Y clones e população de Y). Em seguida, uma comparação da variabilidade de sequência do maxicírculo dentro do DTU TcII utilizando como base de comparação a sequência de Esmeraldo (Figura 23 B), revelou uma maior proximidade de Esmeraldo com amostras de campo S44a, S154a e S23b, e uma menor proximidade em relação às amostras da cepa Y. Por outro lado, quando os maxicírculos de TcI foram avaliados, todos apresentaram uma grande conservação de sequências, sugerindo que pode haver uma maior divergência de sequência entre os kDNAs de TcII do que entre os kDNAs de TcI (Figura 23 C). Análises filogenéticas das 19 cepas de *T. cruzi* baseadas nas sequências codificadoras do maxicírculo separou os DTUs de forma similar à filogenia com base em marcadores nucleares, com clusters de TcI e TcII bem definidos. Porém, as análises de maxicírculo clusterizaram as cepas de TcV e TcVI juntamente com TcIII, reforçando que a mitocôndria dos DTUs híbridos foi originada do ancestral TcIII (Figura 23 D). Os pares de cepas de TcII de regiões geográficas próximas, S15-S162 e S11-S92, também clusterizaram na filogenia baseada na sequência de maxicírculo, enquanto as amostras da mesma localidade, S11 e S154a, também apresentaram uma maior divergência (Figura 23 D), suportando os resultados prévios baseados em marcadores nucleares.

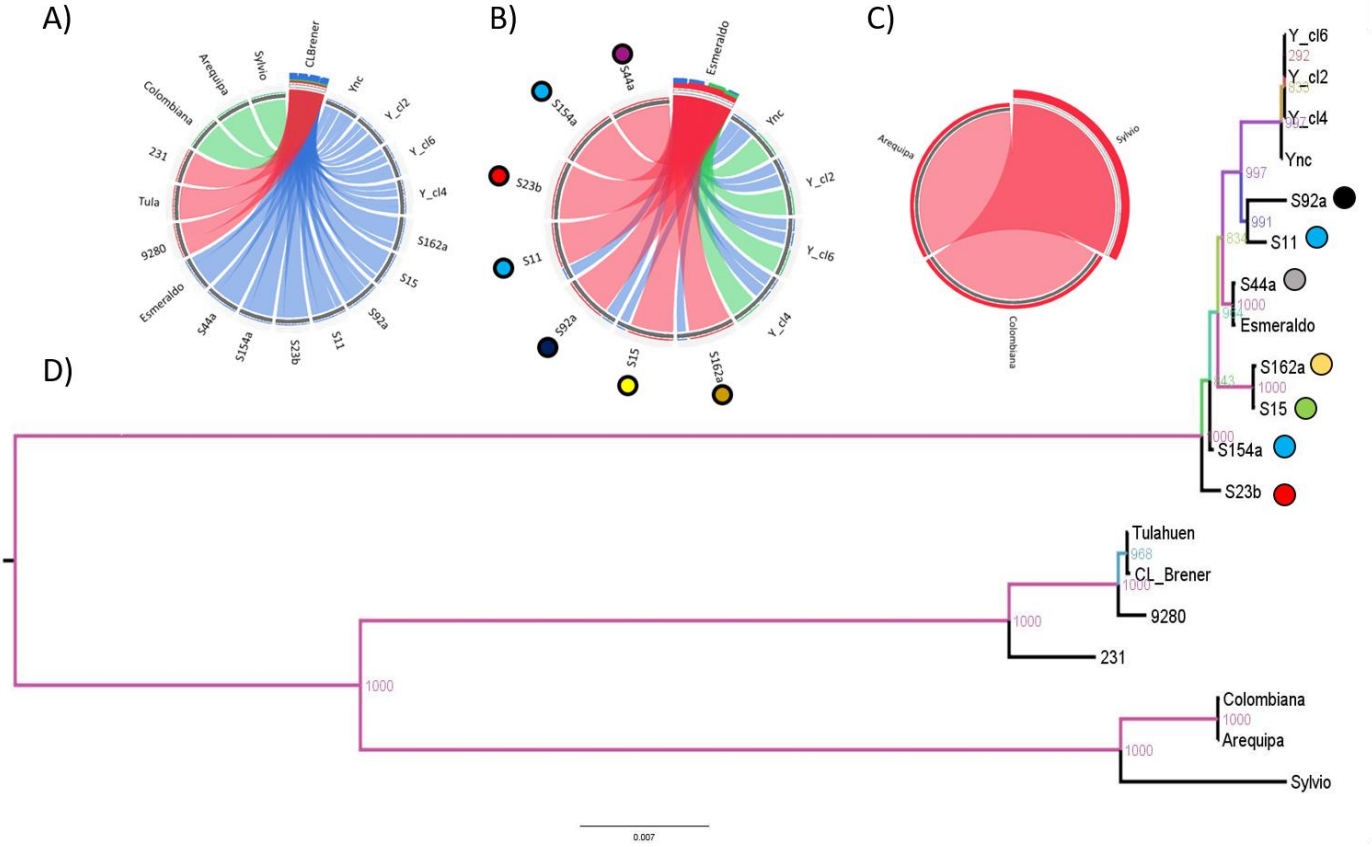


Figura 23: Filogenia baseada em sequências de maxicírculo. Gráfico de *Circus* baseado na similaridade ao longo de toda a sequência montada dos maxicírculos de: **(A)** Todas as 19 cepas de *T. cruzi*, utilizando como raiz o maxicírculo de CL Brener; **(B)** Apenas as sequências de maxicírculo do DTU TcII utilizando a sequência de Esmeraldo como raiz; **(C)** Apenas as sequências de maxicírculo do DTU TcI, utilizando a sequência de Sylvio como raiz. Nestes gráficos, a porcentagem de similaridade está representada por um padrão de cores, onde vermelho corresponde a uma alta similaridade; verde, similaridade intermediária; e azul, baixa similaridade. **(D)** Filogenia estimada por máxima verossimilhança da sequência dos maxicírculos em todas as 19 amostras de *T. cruzi* avaliadas. Os círculos coloridos correspondem a localização geográfica na qual cada cepa foi isolada, como visto na figura 21 A.

Um Tanglegrama comparativo entre a filogenia baseada em marcadores nucleares e mitocondriais mostrou grandes correspondências na maioria dos ramos em ambas as metodologias (Figura 24). Porém, algumas discordâncias também foram observadas. Com base na filogenia baseada nas sequências de maxicírculos, o grupo irmão da cepa Y é o clado S11-S92, enquanto na filogenia baseada em marcadores nucleares, o grupo irmão de Y corresponde a cepa S44. A cepa Esmeraldo clusterizou com a S154a na filogenia baseada em marcadores nucleares, tendo clusterizado com S44a (longe de S154a) na filogenia mitocondrial. Finalmente, os dois pares de cepas de TcII, S15-S162 e S92a-S11 clusterizaram próximos na filogenia nuclear, mas distantes na filogenia dos maxicírculos. Estas discordâncias entre as sequências nucleares e

mitocondriais sugerem taxas evolutivas diferentes entre o genoma nuclear e o mitocondrial, ou podem ser o resultado de uma "introgressão mitocondrial", onde uma linhagem do parasito pode herdar a mitocôndria de outra linhagem sem alterações em seu genoma nuclear, resultando em filogenias nucleares e mitocondriais discordantes.

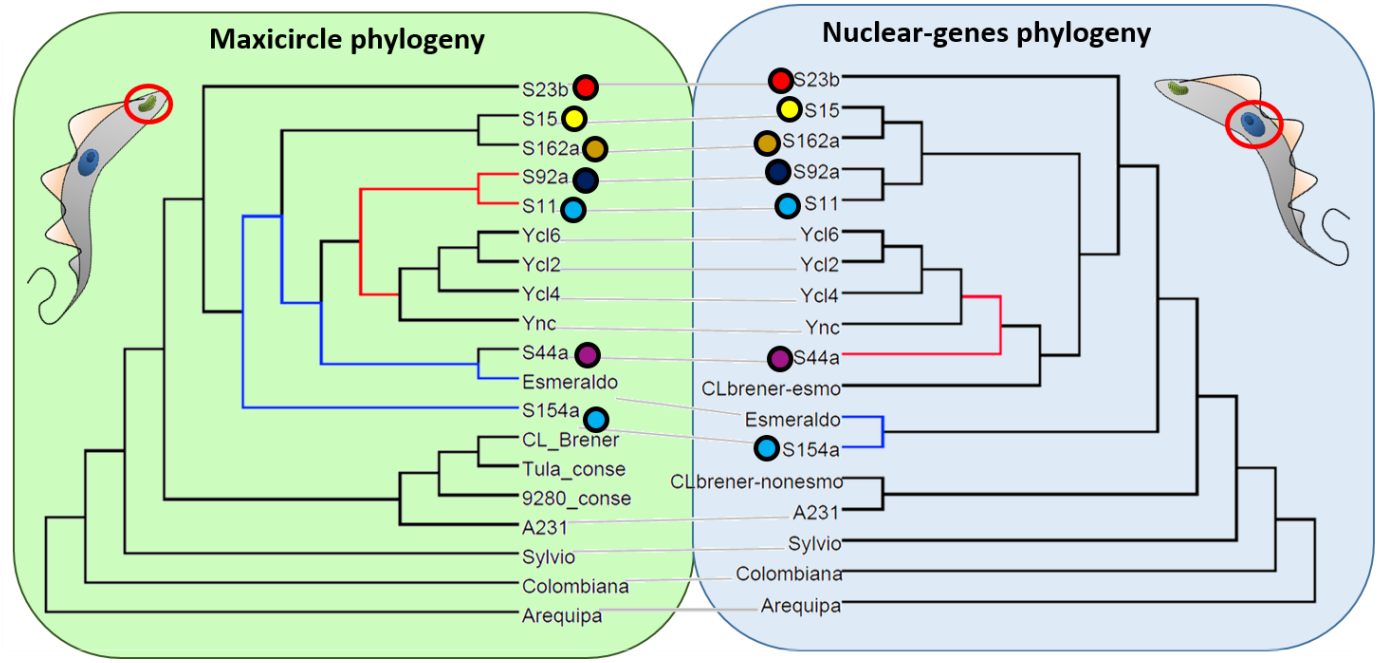


Figura 24: Tanglegrama das filogenias baseadas em maxicírculo ou em marcadores nucleares. Esta imagem representa a comparação entre a filogenia baseada em genes de cópia simples nucleares e a filogenia baseada na sequência do maxicírculo. Ambas as árvores foram enraizadas com base na cepa Arequipa. Os ramos das árvores foram reorganizados para facilitar as comparações entre as árvores. O grupo irmão de Y está destacado em vermelho, enquanto a distância entre Esmeraldo e S154a está destacado em azul para ambas as árvores. Os círculos coloridos representam ao local de isolamento de cada amostra, como descrito na figura 21 A.

Para avaliar possíveis evidências de heteroplasmia mitocondrial dentro das amostras de campo de TcII, nós re-mapeamos as reads de kDNA de cada amostra de campo de TcII em sua sequência de maxicírculo montada e procuramos por SNPs heterozigóticos (Figura 25). Um total de 38, 46, 33, 29, 40, 27, 36, 16, 17 e 21 SNPs heterozigóticos foram encontrados, respectivamente, nas cepas S11, S154a, S15, S162, S23b, S44a, S92a, Ycl2, Ycl4, Ycl6 (Figura 25 A). Porém, a maioria destes SNPs estava localizado em regiões não-codificadoras ou repetitivas (Figura 25), e desta forma não são indícios robustos para a ocorrência de heteroplasmia mitocondrial nestes isolados.

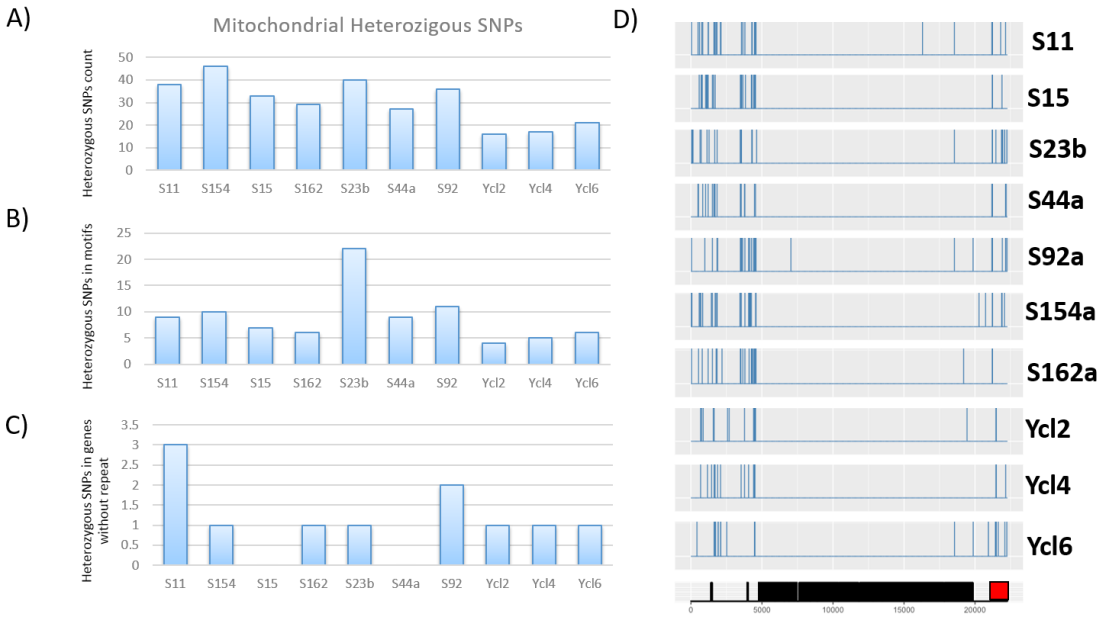


Figura 25: SNPs heterozigóticos nas sequências de maxicírculo. Para avaliar a ocorrência de heteroplasmia mitocondrial, a ocorrência de SNPs heterozigóticos por toda a sequência de maxicírculos nas sete amostras de campo de TcII foi estimada. **(A)** Número total de SNPs ao longo de toda a sequência do maxicírculo. **(B)** SNPs encontrados em genes ou regiões repetitivas do maxicírculo **(C)** SNPs localizados apenas em genes codificadores de proteínas do maxicírculo. **(D)** Distribuição dos SNPs ao longo da sequência de maxicírculo. Em cada caixa cinza, as linhas azuis correspondem às posições de SNP, enquanto a linha preta abaixo corresponde a toda a sequência do maxicírculo, de 0 até 22292 pb. Nesta linha, cada gene é representado por uma caixa preta, onde a região repetitiva é representada por uma caixa vermelha.

11.2 Variação no número de cópias cromossômicas

Para avaliar a ocorrência de CCNV entre as sete amostras de campo de *T. cruzi* do DTU TcII, a metodologia de SCoPE foi utilizada (**Subitem 5.8**) (Figura 26). Esta análise revelou grandes diferenças entre o padrão de duplicação/perda cromossômica entre os isolados de campo de TcII, destacando a extensa variabilidade de ploidia encontrada neste DTU. As cepas S11 (Figura 26 A), S154a (Figura 26 B) e S162a (Figura 26 C) apresentaram >5 aneuploidias, enquanto S92a (Figura 26 D) e S15 (Figura 26 E) apresentaram entre 3 e 5 aneuploidias e S23b (Figura 26 F) e S44a (Figura 26 G) apresentaram < 3 duplicações/perdas cromossômicas. De modo interessante, as amostras filogeneticamente relacionadas S15 e S162 apresentaram um padrão similar de aneuploidias, especialmente nos cromossomos iniciais, como 2, 3 e 7; mas apresentaram um padrão variável em outros, como os cromossomos 22 e 29, sugerindo que eventos de duplicação/perda de cromossomos são eventos frequentes em TcII. Para determinar a ploidia geral de cada amostra de campo de TcII, as razões de frequência alélica de SNPs heterozigóticos foram estimadas, resultando em um pico em 0.5 para todas as amostras, reforçando a predominância de cromossomos diploides em *T. cruzi* (Figura 26 H).

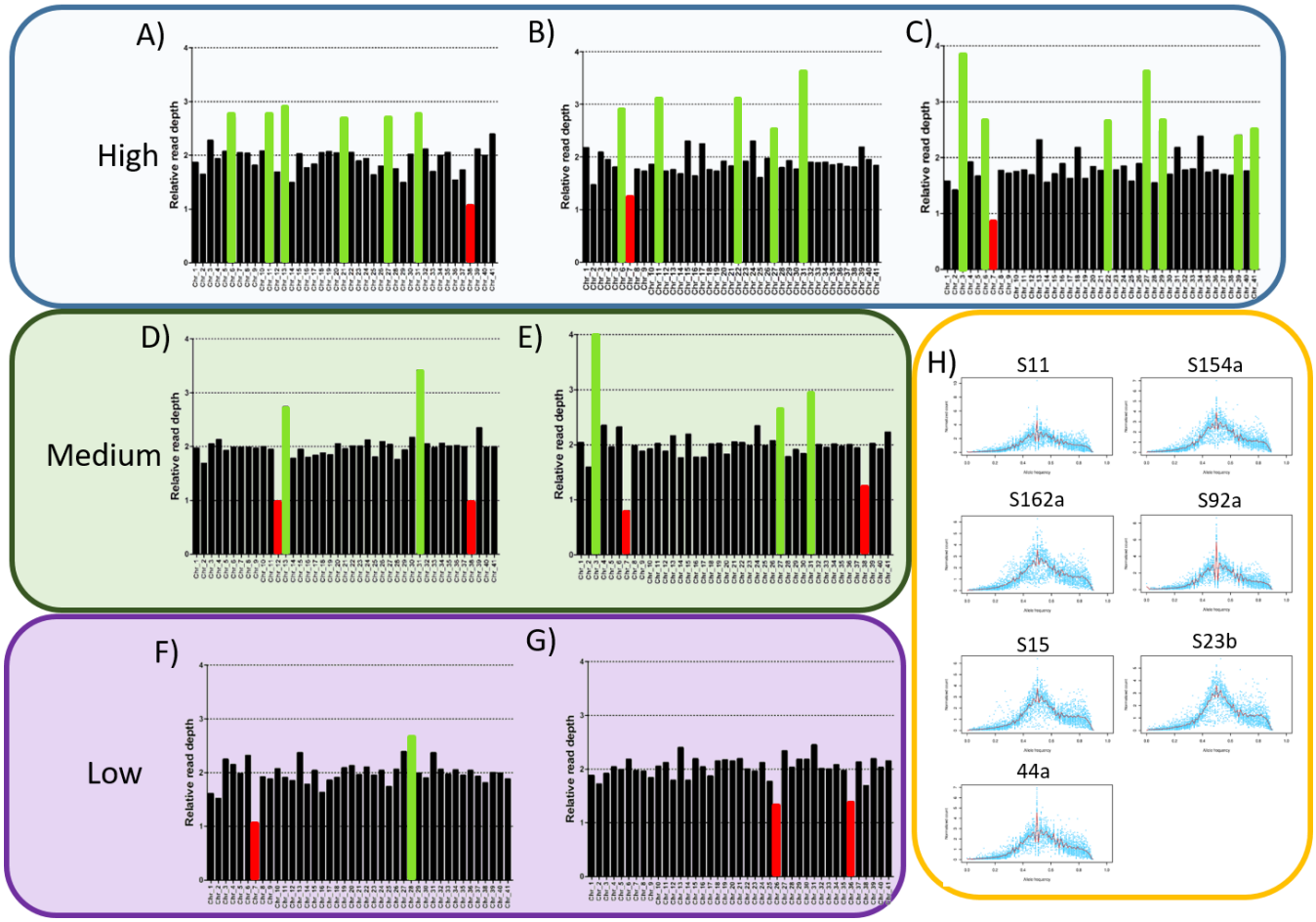


Figura 26: Ploidia das amostras de campo de *T. cruzi* do DTU TcII. A ploidia predita de cada cromossomo das amostras de campo *de T. cruzi* **(A)** S11b, **(B)** S154a, **(C)** S162a, **(D)** S92a, **(E)** S15, **(F)** S23b, **(G)** S44a, utilizando como referência os 41 cromossomos do haplótipo Esmeraldo-like de CL Brener foi estimado com base na metodologia SCoPE. Cada barra corresponde a razão entre a RDC dos genes cópias simples e a cobertura do genoma, representando a ploidia predita do cromossomo, onde um valor de 2 denota um cromossomo diploide. Barras pretas correspondem a cromossomos diploides ou próximos de diploide, enquanto barras verdes correspondem a cromossomos triploides ou tetraploides e barras vermelhas correspondem a cromossomos haploides. **(H)** A ploidia genômica geral de cada cepa foi estimada pela proporção alélica de posições heterozigóticas simultaneamente em todos os cromossomos, onde uma tendência de 0.5 representou uma predominância a diploidia em todas as cepas avaliadas.

Para comparar o padrão de CCNV das amostras de campo de TcII com outros DTUs, os eventos de duplicação/perda cromossômica em todas as 19 cepas de *T. cruzi* abordadas neste estudo foi estimado, revelando padrões distintos dentro e entre DTUs, com alguns cromossomos sendo consistentemente perdidos ou ganhos (Figura 27). O cromossomo 31 apresentou um padrão supranumerário na grande maioria das cepas dos cinco DTUs avaliados, enquanto os cromossomos 6, 13 e 27 também apresentaram uma tendência para polissomia,

mas em menor extensão do que o cromossomo 31. Existe também evidencia para uma perda cromossômica, como nos cromossomos 2, 7, 25 e 38, que apresentaram padrão monossômico em várias cepas de *T. cruzi*. Para avaliar se o padrão de CCNV varia dentro de uma população, nós comparamos o padrão de aneuplodias na população não clonada de Y, assim como nos 3 clones derivados desta população: Ycl2, Ycl4 e Ycl6. Todos os clones de Y apresentaram o mesmo padrão de CCNV entre si e também com a população não clonada. Apesar do pequeno número de clones avaliados, estes dados sugerem uma conservação no padrão de variação cromossômica intra-populacional em *T. cruzi* (Figura 27).

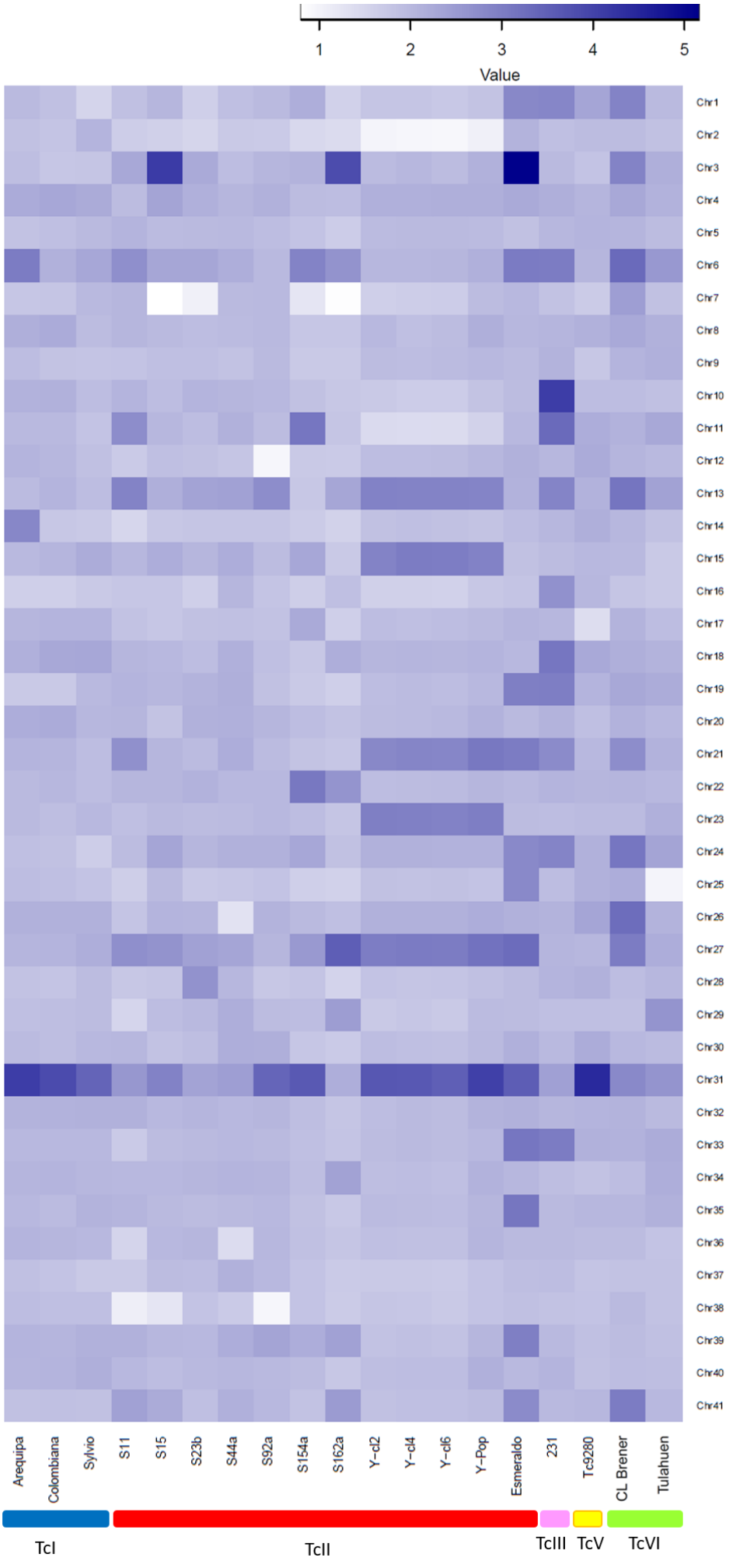

1
2
3
4
5
Value
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
Chr23
Chr24
Chr25
Chr26
Chr27
Chr28
Chr29
Chr30
Chr31
Chr32
Chr33
Chr34
Chr35
Chr36
Chr37
Chr38
Chr39
Chr40
Chr41
Arequipa
Colombiana
Sylvio
S11
S15
S23b
S44a
S92a
S154a
S162a
Y−cl2
Y−cl4
Y−cl6
Y−Pop
Esmeraldo
231
Tc9280
CL Brener
Tulahuen
TcI
TcII
TcIII
TcV
TcVI

Figura 27: Heatmap representativo do padrão de CCNV entre os DTUs de *T. cruzi* e amostras de campo. A estimativa do número de cópias de cada um dos 41 cromossomos em cada amostra está representada. Cada coluna representa uma cepa de *T. cruzi.* Cada linha corresponde a um cromossomo de *T. cruzi*, numerado de 1 a 41. Neste heatmap, o número de cópias cromossômicas é representado por um gradiente indo de branco até azul escuro, onde uma caixa branca corresponde a ~1 cópia cromossômica e uma caixa azul escuro a ~5 cópias.

Como visto no **subitem 6.4,** a população não clonada de Y apresentou uma drástica mudança de RDC começando na posição 248kb do cromossomo 11, resultando em uma região inicial monossômica e terminal dissômica neste cromossomo. Para investigar se esta diferença é um resultado de uma variação populacional, onde parte da população apresenta um cromossomo 11 completo e parte apresenta o cromossomo truncado, ou ainda se cada célula de Y apresenta uma cópia completa e uma truncada do cromossomo 11 nós comparamos a RDC ao longo de todo o cromossomo entre os clones e a população não clonada de Y. Os 248 kb iniciais dos três clones de Y apresentaram metade da RDC do restante do cromossomo nos três clones, de forma similar ao que foi encontrado na população não clonada (Figura 28). Isto sugere que em Y, os genes localizados nos 248kb iniciais do cromossomo 11 de CL Brener apresentam um estado haploide, enquanto os genes localizados após esta região estão em estado diploide.

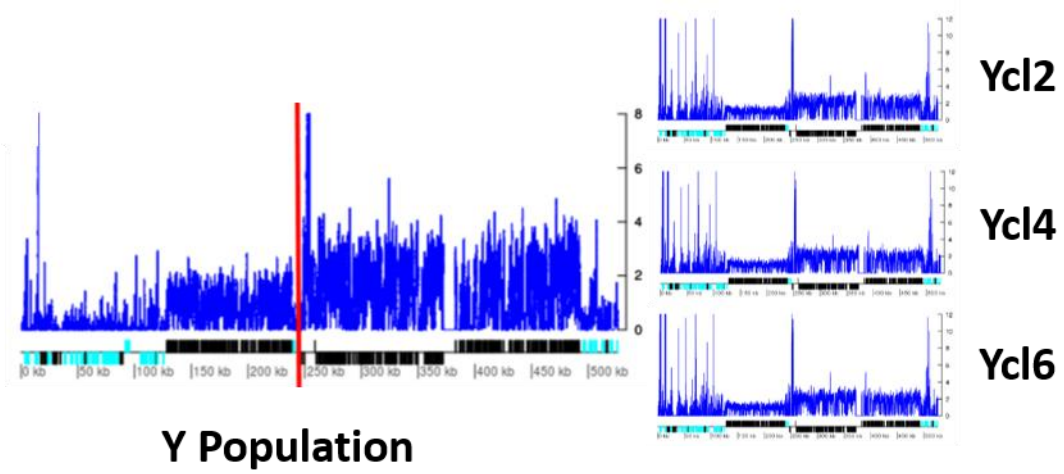


Figura 28: RDC ao longo de toda a extensão do cromossomo 11 dos clones e população de Y. Nesta figura, a linha azul corresponde a RDC normalizada pela cobertura do genoma para cada posição do cromossomo 11. Abaixo, os genes codificadores de proteína estão representados como retângulos desenhados em proporção ao seu tamanho, onde a fita codificadora a que eles pertencem está representada pela orientação acima da linha central (fita +) ou abaixo da linha (fita -). Caixas em ciano ou preto correspondem, respectivamente, a famílias multigênicas e clusters de genes hipotéticos/*housekeeping*. Os 248kb iniciais neste cromossomo apresentam um RDC menor do que a parte seguinte, tanto na população não clonada de Y quanto nos 3 clones de Y avaliados.

Uma análise de clusterização hierárquica baseada na distância Euclidiana da ploidia predita de cada cromossomo nas 19 cepas de *T. cruzi* avaliadas mostrou que o padrão atual de CCNV entre amostras de *T. cruzi* não apresenta uma distribuição semelhante à filogenia, visto que cepas de TcI, TcV e TcVI se intercalam com cepas de TcII (Figura 29). A única cepa de TcIII, 231, apresentou o padrão mais divergente entre as cepas avaliadas, porém mais cepas do DTU TcIII são necessárias para avaliar a extensão da variação de cópias cromossômicas neste DTU. As cepas de TcII, S15-S162 e S11-S92, que foram agrupadas na filogenia com base em marcadores nucleares e mitocondriais também agruparam com base no perfil de CCNV, reforçando a proximidade evolutiva destas amostras.

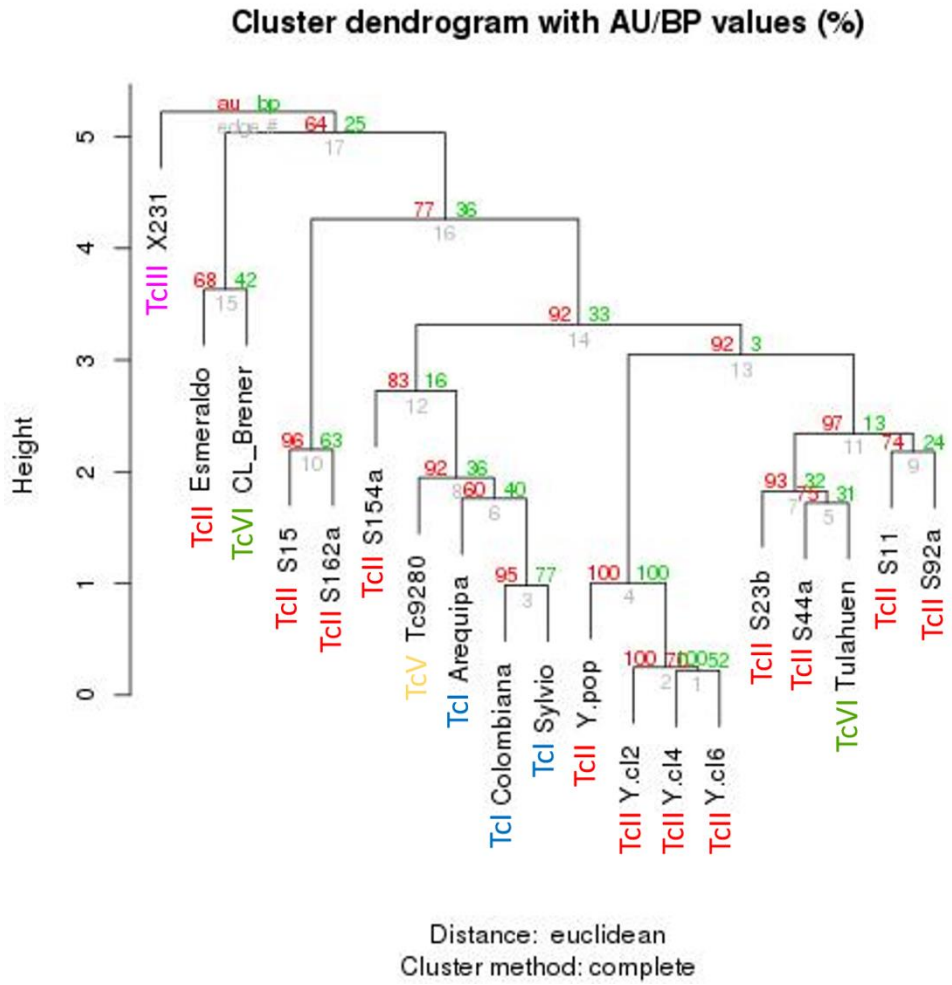


Figura 29: Dendrograma da clusterização hierárquica baseada no padrão de aneuploidias das cepas de *T. cruzi*. A análise de clusterização hierárquica baseada nas distâncias euclidianas da ploidia predita de cada cromossomo das 19 cepas de *T. cruzi* foi realizada utilizando o pacote do R Pvclust. Duas metodologias de reamostragem de bootstrap foram utilizados para avaliar a incerteza da análise de clusterização hierárquica: *bootstrap probabilitty* (BP), o método de reamostragem normal do bootstrap; e o *approximately unbiased* (AU), de reamostragem de bootstrap multi-escalar. Ambos os métodos foram calculados com 10.000 iterações.

11.3 CCNV na cepa CL Brener de *T. cruzi*

O estudo do padrão de aneuploidias na cepa CL Brener de *T. cruzi* forneceu alguns resultados interessantes com relação a proporção do genoma corresponde ao haplótipo Esmeraldo-like e ao haplótipo Non-Esmeraldo-like (Figura 30). A comparação das predições de ploidia de CL Brener pela metodologia de SCoPE utilizando os genes de cópia simples do haplótipo Esmeraldo-like (Figura 30 A) ou Non-Esmeraldo-like (Figura 30 B), mostrou que análises de CCNV mapeando as reads em cada um dos haplótipos de CL Brener levaram ao mesmo resultado, com 10 aneuploidias. Este resultado reforça a robustez da metodologia SCoPE, mostrando que ela não é influenciada pelo uso dos haplótipos Esmeraldo-like ou Non-Esmeraldo-like como referência para o mapeamento das reads (Figura 30 A e B).

Como a cepa CL Brener apresenta dois haplótipos em seu genoma, foi determinado qual dos haplótipos apresentaria uma cópia extra de cromossomo no caso de cromossomos supranumerários, com base nas análises de frequência alélica. Em nossos scripts para a análise de frequência alélica (Figura 12), é realizada a avaliação da proporção de reads que apresentam o alelo igual ao da referência pelo total de reads que mapearam na posição. Desta forma, um valor de razão de 0.66 significa que 66% das reads apresentam o alelo igual ao da referência e 33% apresentam o alelo mutado (diferente da referência). De forma similar, um resultado de 0.33 significa que 33% das reads apresentam o alelo igual ao da referência e 66% apresentam o alelo mutado, enquanto um valor de 0.5 corresponde a 50% das reads mapearem em cada uma das duas variantes. Quando avaliamos a frequência alélica nos cromossomos supranumerários de CL Brener, mapeando as reads no haplótipo Esmeraldo-like ou Non-Esmeraldo-like nós vimos que para todos os casos, quando um dos haplótipos apresentava um valor de razão de 0.33, o outro haplótipo apresentava um valor de 0.66, e vice-versa (Figura 30 C). Isso quer dizer que, dos 10 cromossomos supranumerários de CL Brener estimados como triploides, os cromossomos 1, 6, 13, 21, 24 e 41 apresentaram duas cópias do cromossomo derivado do haplótipo Non-Esmeraldo-like, enquanto os cromossomos 3, 26, 27 e 31 apresentaram uma cópia extra do cromossomo do haplótipo Esmeraldo-like (Figura 30 C). De modo comparativo, a frequência alélica do cromossomo 17 (dissômico) apresentou uma razão de 0.5 tanto para o mapeamento na referência Esmeraldo-like quanto na referência Non-Esmeraldo-like. Resultados semelhantes ao cromossomo 17 foram obtidos para todos os outros cromossomos dissômicos.

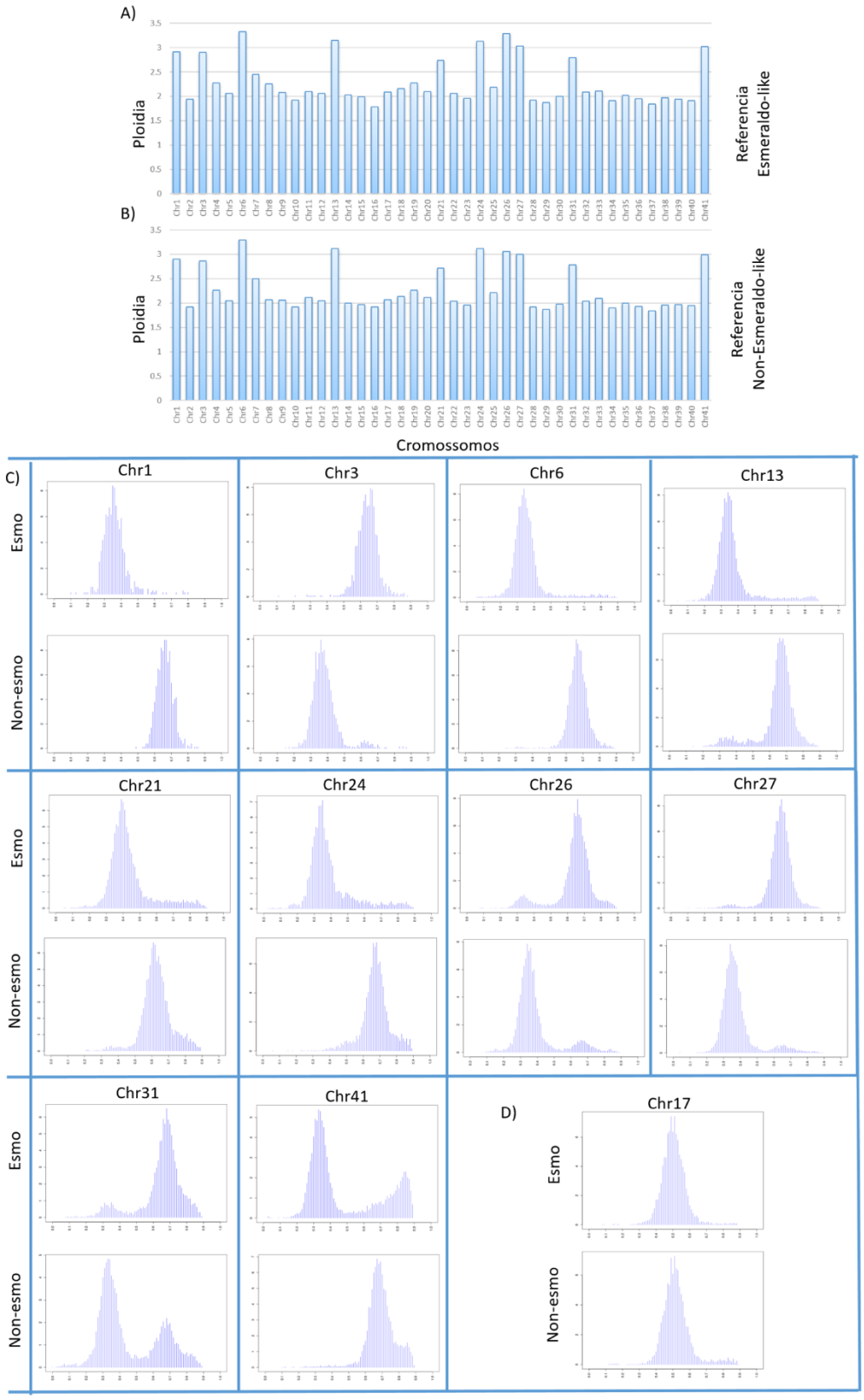


Figura 30: Padrão de CCNV na cepa CL Brener de *T. cruzi*. Avaliação de ploidia de todos os cromossomos de CL Brener utilizando o haplótipo **(A)** Esmeraldo-like; ou **(B)** Non-Esmeraldo-like. Nestas figuras, cada barra azul corresponde a ploidia predita de cada um dos 41 cromossomos de CL Brener. A altura da barra corresponde a ploidia, onde 1 representa um cromossomo monossômico, 2 um dissômico e 3 um trissômico. **(C)** Resultados da razão da

frequência alélica de cada um dos 10 cromossomos trissômicos de CL Brener. Em cada uma das figuras neste bloco, o eixo X contém valores de 0,01 a 1,00 representado a razão entre o número de reads que apresentou a variante igual à referência pelo total de reads que mapeou na posição. O eixo Y corresponde ao número de SNPs heterozigóticos que apresentou esta razão. Desta forma, picos nas posições 0.33, 0.5 e 0.66 do eixo X significam, respectivamente, que a variante igual da referência apresentou 33%, 50% ou 66% do mapeamento das reads. Como exemplo, o cromossomo trissômico 1 apresentou um pico em 0.33% (uma cópia) para o haplótipo Esmeraldo-like e um pico em 0.66% (duas cópias) para o haplótipo Non-Esmeraldo-like. **(D)** Resultados da razão da frequência alélica de um cromossomo dissômico, o cromossomo 17.

12. Discussão

A maioria das análises de genética de população de *T. cruzi* realizadas até hoje são baseadas em sequenciamento multilocus de marcadores nucleares ou mitocondriais específicos, restringindo a comparação da variabilidade do parasito a um número restrito de regiões genômicas (BAPTISTA et al., 2014; LIMA et al., 2014a, 2014b; LLEWELLYN et al., 2009; MESSENGER et al., 2012, 2015, 2016; MESSENGER; MILES, 2015; MILES et al., 2009). Por outro lado, o uso de sequenciamento de todo o genoma do parasito através de metodologias de sequenciamento de nova geração fornece uma avaliação mais completa da variabilidade entre amostras, permitindo não só uma comparação das divergências baseada em maior número de genes, mas também a correlação entre CCNV com a filogenia do parasito. A maioria dos estudos de ecologia molecular em *T. cruzi* são focados no DTU TcI, a mais antiga e dispersa linhagem do parasito, que é responsável pela maioria das infecções humanas na região norte da América Latina (BARNABE et al., 2013; BARNABÉ et al., 2011; LIMA et al., 2014b; MESSENGER et al., 2012; ZINGALES et al., 2012). Deste modo, existem menos estudos avaliando a variabilidade, distribuição e trocas gênicas no DTU TcII, um dos subgrupos de *T. cruzi* mais relevantes em relação à infecção humana no cone sul da América do Sul (BAPTISTA et al., 2014; CÂMARA et al., 2010; D'ÁVILA et al., 2009; DA CÂMARA et al., 2013; ZINGALES et al., 2012). Apesar da correspondência entre DTU e manifestações clínicas da doença de Chagas ainda ser incerto, em algumas regiões geográficas, evidencias apontam para uma associação de TcII com manifestações clínicas graves, onde pacientes apresentam a sintomatologia cardíaca e digestiva (LUQUETTI et al., 1986; MESSENGER; MILES; BERN, 2015; ZINGALES et al., 2012). Para uma melhor avaliação da variabilidade genética de amostras de campo de TcII, nós sequenciamos o genoma nuclear e mitocondrial de sete cepas de TcII, isoladas de pacientes de regiões geográficas próximas do estado de Minas Gerais (Figura 21), e as comparamos entre si e com cepas de outros DTUs.

12.1 Filogenia nuclear e mitocondrial de *T. cruzi*

O primeiro passo para estimar as relações entre evolução, duplicação/perda cromossômica e recombinação dentro do DTU TcII de *T. cruzi* foi a avaliação de sua filogenia baseada em um conjunto de genes de cópia simples nucleares e em todos os genes do maxicírculo. O uso de genes cópia simples não repetitivos como marcadores para estimar a filogenia é mandatório em análises baseadas em reads curtas de Illumina, pois a montagem de novo ou o mapeamento destas reads em regiões repetitivas (como microssatélites) pode

resultar em erros de mapeamento, gerando falsos SNPs que poderiam comprometer as inferências filogenéticas. Apesar destes genes de cópia simples não mutarem na mesma taxa que microssatélites, o uso de um grande número de genes proveu resolução suficiente para separar TcII de outros DTUs, assim como para estimar a distância filogenética intra-DTU TcII (Figura 21 B).

Uma comparação da filogenia baseada nos genes de cópia simples nucleares ou nos genes mitocondriais revelou que a maioria dos seus ramos são compartilhados pela filogenia por máxima verossimilhança nos dois marcadores (Figura 24). As duas amostras da região central de Minas Gerais, S15 e S162, agruparam com alto valor de confiança em ambas as análises baseadas em marcadores nucleares ou mitocondriais, assim como na análise por PCA baseada nos SNPs genômicos (Figura 21 B, C; Figura 22 D). Estas duas amostras estão separadas das outras cinco pela Serra do Espinhaço, uma cadeia de montanhas cuja extensão vai desde o centro de Minas Gerais até a região norte da Bahia. Esta barreira geográfica pode ter restringido o trânsito de insetos vetores, separando as populações de *T. cruzi* da região central e do norte/nordeste de Minas Gerais.

Análises de grupos populacionais por ADMIXTURE (ALEXANDER; NOVEMBRE, 2009) estimaram que a as populações originais de TcII avaliadas foram geradas a partir de 7 populações ancestrais (Figura 21 D e E). Todos os sete isolados de campo de TcII apresentara uma origem mista, composta por frações de duas a cinco populações ancestrais, sugerindo que diversos eventos de trocas gênicas ocorreram durante a história evolutiva de TcII em Minas Gerais (Figura 21 C), de acordo com o que foi predito por Freitas 2006 e Baptista 2014. As populações da região central de Minas Gerais foram predominantemente formadas pelas populações ancestrais 4 (ciano) e 6 (azul escuro). Estas duas populações ancestrais também constituem uma menor parte as cepas S11, S154 e S23b, que foram isoladas de pacientes da região nordeste do estado. De modo similar, as amostras da região central apresentaram uma pequena proporção da população ancestral 2 (amarelo), que é a maior constituinte das cepas S11 e S92. Estes resultados sugerem que ocorreram trocas gênicas entre cepas da região central e nordeste do estado durante a evolução do parasito, ou que as populações ancestrais como pop2, 4 e 6 apresentaram uma grande distribuição geográfica, que englobava tanto a região central quanto a nordeste de Minas Gerais. De modo interessante, a amostra S154a foi a única a apresentar regiões genômicas derivadas de 5 populações ancestrais. Esta amostra também foi a mais divergente de todas as amostras de campo na filogenia dos genes nucleares, agrupando com Esmeraldo (Figura 21 B), apresentou o padrão de SNPs mais divergente baseado no PCA de todos os SNPs (Figura 21 C) e apresenta discordâncias entre a filogenia dos genes nucleares e

mitocondriais. Estes resultados sugerem que S154a pode ter sofrido vários eventos de recombinação durante sua história, sendo composta por um maior número de populações ancestrais dentre as cepas de *T. cruzi* avaliadas. Esta hipótese é também suportada pelo fato que S154a apresentou um alto número de aneuploidias (Figura 26), o que pode ter sido uma herança do maior número de eventos de recombinação que esta linhagem do parasito sofreu durante a sua história evolutiva. Atualmente, a maioria dos resultados com amostras de campo de localidades próximas suporta que *T. cruzi* não é estritamente clonal, e que recombinação é um evento não-obrigatório, mas comum (BAPTISTA et al., 2014; LEWIS et al., 2011; LIMA et al., 2014b; MESSENGER et al., 2012; MESSENGER; MILES, 2015; RAMÍREZ et al., 2012). A ocorrência de recombinação entre populações de *T. cruzi* foi documentada em amostras de TcI da Bolívia (BARNABE et al., 2013), Colômbia (RAMÍREZ et al., 2012) e Brasil (LIMA et al., 2014b), assim como entre cepas TcII no Brasil (BAPTISTA et al., 2014). Cepas de *T. cruzi* foram também capazes de realizar recombinação genética em laboratório, apresentando fusão de genótipos parentais, perda de alelos, recombinação homologa e herança uniparental do genoma mitocondrial (GAUNT et al., 2003).

Apesar da alta concordância entre a filogenia usando marcadores nucleares e mitocondriais entre as cepas de *T. cruzi* avaliadas neste trabalho, algumas discordâncias são evidentes (Figura 24). O grupo irmão do clado de Y foi a cepa S44a com base na filogenia nuclear e S11/S92a baseado nos marcadores mitocondriais (Figura 24, linha vermelha). De modo similar, a cepa Esmeraldo clusterizou com a amostra S154a baseado em marcadores nucleares, mas com S44a com base em marcadores do maxicírculo (Figura 24, linha azul). Esta diferença na filogenia nuclear e do kinetoplasto pode ser resultado de uma introgressão mitocondrial ou a consequência de diferentes taxas evolutivas entre os genomas nuclear e mitocondrial (LIMA et al., 2014b; MESSENGER et al., 2012). Eventos de introgressão mitocondrial foram documentados em cepas de *T. cruzi*, isoladas nas Américas do Norte e do Sul (DE FREITAS et al., 2006; LIMA et al., 2014b; MACHADO; AYALA, 2001; MESSENGER et al., 2012, 2016), sugerindo que este é um evento comum na biologia do parasito. Apesar das implicações decorrentes da introgressão mitocondrial ainda não serem conhecidas, a sua ocorrência reforça inferências de recombinação entre cepas de *T. cruzi*. Para checar eventos de heteroplasmia mitocondrial, a presença de genomas mitocondriais heterogêneos em uma célula individual, nós re-mapeamos as reads de maxicírculo de cada cepa de TcII em sua referência montada *de novo*, e avaliamos a ocorrência de SNPs heterozigóticos, como realizado por Messenger 2012 (MESSENGER et al., 2012). Apenas poucos SNPs heterozigóticos foram identificados no genoma mitocondrial das cepas de TcII, onde a maioria estava localizado em regiões repetitivas ou não codificadores, não sendo,

portanto, evidências robustas da ocorrência de heteroplasmia (Figura 25). Até hoje, evidências de ocorrência de heteroplasmia mitocondrial em *T. cruzi* são escassos (MESSENGER et al., 2012, 2016). Heteroplasmia já foi observada no clone Sylvio X10/1 (TcI), baseado no re-mapeamento das reads na sequência montada de seu maxicírculo, resultando em 74 SNPs em oito genes e três regiões intergênicas (MESSENGER et al., 2012). Em nossa análise, nós encontramos apenas 1-3 SNPs em genes codificadores, não obtendo suporte suficiente para corroborar a ocorrência de heteroplasmia. Porém, a ausência de heteroplasmia nas amostras de TcII pode ser um resultado da baixa cobertura (~60X), quando comparados a cobertura utilizada em Sylvio X10/1 por Messenger 2012 (~163X), visto que apenas 12,2% das reads corresponderam ao SNP variante menor em Sylvio X10/1 (MESSENGER et al., 2012).

12.2 Variação no número de cópias cromossômicas

No presente trabalho foi realizada a comparação do padrão de aneuploidias das sete amostras de campo de TcII (Figura 26), assim como entre cepas previamente sequenciadas dos DTUs TcI, TcIII, TcV e TcVI (Figura 27) utilizando como base os 41 cromossomos hipotéticos da cepa CL Brener de *T. cruzi*. Apesar desta ser a melhor referência atual do genoma de *T. cruzi* publicada, mapas físicos gerados através da hibridização de clones de YAC em marcadores cromossomo específico e analises de sintenia com *T. brucei* sugerem que existem algumas incongruências nestas montagens (SOUZA et al., 2011), como por exemplo a junção dos cromossomos hipotéticos TcChr4 e TcChr37 em um cromossomo único. Em nossas análises não encontramos nenhuma discordância entre as predições de ploidia de ambos os cromossomos, que se apresentaram em estado dissômico para as 19 as cepas de *T. cruzi* avaliadas, o que está de acordo com esta predição.

O dendrograma baseado na clusterização hierárquica do padrão de CCNV das 19 cepas de *T. cruzi* avaliadas agrupou cepas de TcI, TcV e TcVI dentro de clusters de cepas de TcII, mostrando que eventos de duplicação/deleção cromossômica não seguem a filogenia baseada em genes cópia simples nucleares nem em marcadores mitocondriais (Figura 29). De fato, as amostras de campo de TcII apresentaram um padrão diferente de aneuploidias com um número baixo (S23b, S44a), médio (S15, S92a) ou alto (S11, S154a, S162a) de aneuploidias (Figura 26), o que está de acordo com a alta divergência entre as cepas Y e Esmeraldo, previamente observada no capítulo 1. Isto sugere que eventos de ganho e perda cromossômicas são frequentes em *T. cruzi*, e ocorrem em uma taxa maior do que os eventos de segregação dos DTUs (MINNING et al., 2011; REIS-CUNHA et al., 2015). Essa hipótese é suportada pelo fato de que as amostras S15 (Figura 26 E) e S162 (Figura 26 C), que são agrupadas na filogenia baseada em marcadores

nucleares e mitocondriais (Figura 21 B, Figura 23 D) e tem a mesma ancestralidade baseada em análises de ADMIXTURE (Figura 21 D) apresentaram um padrão discordante de aneuploidias. O padrão de aneuploidias também varia entre populações de *Leishmania donovani* isoladas de regiões geográficas próximas (DOWNING et al., 2011), reforçando que ambos os parasitos são naturalmente aneuploides (DUJARDIN et al., 2014). Baseado em análises de FISH, eventos de CCNV foram identificados entre células de uma mesma população em *Leishmania* (STERKERS et al., 2011, 2014). Por esta razão, foi sugerido que *Leishmania* apresentam uma "aneuploidia em mosaico", onde células de uma mesma população apresentam diferentes padrões de aneuploidias, e o genótipo mais prevalente em uma população foi estimado como ~10% das contagens celulares (Sterkers 2011, Sterkers 2014). Para avaliar se o padrão de aneuploidias em *T. cruzi* varia em uma taxa similar ao que ocorre em *Leishmania*, nós clonamos a população de Y e sequenciamos o genoma de três clones, para estimar o seu padrão de CCNV baseado em RDC. Todos os três clones apresentaram exatamente o mesmo padrão de aneuploidias entre si, assim como o mesmo padrão da população de Y não clonada (Figura 27) sugerindo que, diferente do que foi observado em análises de FISH para *Leishmania*, o padrão de CCNV seja estável dentro de uma mesma população em *T. cruzi*. Este resultado está de acordo com experimentos em gel de eletroforese de campo pulsátil utilizando vários clones da cepa G (TcI), que mostraram que o padrão de bandeamento cromossômico de clones de uma mesma população é constante (LIMA et al., 2013a). Além disso, os perfis de cariótipos da cepa G, e de seu clone D11 foram semelhantes e estáveis quando mantidos em cultura por diversos anos, sugerindo que existe estabilidade no padrão de CCNV apresentado pelo parasito (LIMA et al., 2013a). De modo similar, análises de RDC revelaram que a cepa BPK282/0cl4 de *Leishmania donovani* apresentou um padrão estável de aneuploidias por 32 passagens, inferindo que estimativas na taxa de mudanças no padrão de CCNV por FISH podem ter sido superestimadas (DOWNING et al., 2011). Por outro lado, esta constância no padrão de ploidia estimado por RDC pode ser uma consequência da soma normalizada da ploidia populacional, mascarando o padrão de ploidia de células únicas. Esta hipótese é corroborada pelo fato de que diversas cepas de *T. cruzi* apresentaram ploidia intermediária (como por exemplo valores entre 2 e 3), o que pode ser uma consequência de uma mistura populacional contendo células dissômicas e trissômicas para um determinado cromossomo (REIS-CUNHA et al., 2015).

Eventos de CCNV parecem ser bem tolerados nos tripanossomatídeos *T. cruzi* (MINNING et al., 2011; REIS-CUNHA et al., 2015) e *Leishmania* (DOWNING et al., 2011; LEPROHON et al., 2009; ROGERS et al., 2011; STERKERS et al., 2011, 2014, VALDIVIA et al., 2015b, 2017), porém parecem estar ausente em *T. brucei* (WEIR et al., 2016) (**Almeida et al., em preparação**). A

avaliação de aneuploidias por RDC em 85 amostras de campo de *T. gambiense* grupo 1, isoladas do leste e oeste africano não revelaram nenhuma ocorrência de aneuploidias (WEIR et al., 2016). Resultados similares foram obtidos para as outras duas subespécies de *T. brucei*; *T. b. brucei* e *T. b. rhodesiense* (**Almeida et al., em preparação**). Esta ausência de aneuploidias em *T. brucei* não é uma consequência de reprodução sexuada, visto que *T. b. gambiense* grupo 1 aparenta ser estritamente clonal (WEIR et al., 2016). Como os parasitos *T. cruzi* e *Leishmania* apresentam o seu genoma dividido em um grande número de pequenos cromossomos com tamanhos variando entre ~100k – 2.5 Mb (34 a 47 cromossomos) (BERRIMAN; GHEDIN; HERTZ-FOWLER, 2005; EL-SAYED, 2005a, 2005b; IVENS, 2005; MINNING et al., 2011), alterações no número de cópias de cromossomos específicos irão alterar a dosagem de um número restrito de genes (LAFFITTE et al., 2016; MANNAERT et al., 2012). Por outro lado, o parasito exclusivamente diploide *T. brucei* apresenta o seu genoma dividido em cromossomos de tamanhos no intervalo de 1Mb- 4.5Mb (BERRIMAN; GHEDIN; HERTZ-FOWLER, 2005; EL-SAYED, 2005b; WEIR et al., 2016), sugerindo que aneuplodias podem ser mais bem suportadas em organismos que apresentam seu genoma dividido em um maior número de cromossomos de menor tamanho. A avaliação de CCNV em um maior número de eucariotos unicelulares com tamanhos diferentes de cromossomos é necessária para confirmar esta hipótese. Outro processo biológico que pode ser responsável por desencadear aneuploidias em tripanossomatídeos seria a ocorrência de um "*clash*" entre a maquinaria de replicação de DNA e a maquinaria de transcrição. Como ambos processos apresentam uma grande sobreposição funcional e ocorrem em regiões de mudança de fita codificadora (LOMBRAÑA et al., 2016; MARQUES et al., 2015; TIENGWE et al., 2012), um atraso da maquinaria de transcrição na região de início de replicação poderia impedir o começo da síntese de DNA, impedindo a replicação de um cromossomo. *T. brucei* apresenta mais de uma origem de replicação em seus cromossomos (TIENGWE et al., 2012), de forma que se ocorrer este "*clash*" entre a maquinaria de replicação e transcrição em uma das forquilhas de replicação, a replicação iniciada em outra forquilha simultânea poderia permitir a correta replicação do cromossomo. Por outro lado, apesar de ser sugerido que *Leishmania* sp. também apresenta mais de uma origem de replicação por cromossomo (LOMBRAÑA et al., 2016), analises de RDC Indicam que possa haver um sítio ou sítios preferenciais próximos de começo de replicação de DNA (MARQUES et al., 2015). Portanto, um "*clash*" da maquinaria de replicação e transcrição neste sítio poderia resultar em uma falha de replicação cromossomal em *Leishmania*. Desta forma, cópias extras de cromossomos podem ser importantes para assegurar que ao menos uma cópia de cada cromossomo seja duplicada, evitando haploidias.

Apesar de variantes estruturais e aneuploidias estarem geralmente associadas a fenótipos desvantajosos em eucariotos complexos (HASSOLD; HUNT, 2001; LV et al., 2012; STANKIEWICZ; LUPSKI, 2010), alguns eucariotos unicelulares utilizam aneuploidias para permitir uma rápida adaptação a mudanças ambientas, sugerindo que a variação no número de cópias cromossômicas pode também aumentar o *fitness* em situações de estresse (ABBEY et al., 2011; DOUBRE et al., 2005; SELMECKI; FORCHE; BERMAN, 2010). Variação no número de cópias é um mecanismo bem documentado para alterar a expressão gênica e aumentar a variabilidade, especialmente em organismos como os tripanossomatídeos que regulam a sua expressão gênica por mecanismos pós-transcricionais (CLAYTON, 2013, 2002; LEIFSO et al., 2007; MYUNG et al., 2002; TEIXEIRA et al., 2012). Mecanismos baseados em falha de segregação cromossômica, falhas na divisão cromossômica e o ciclo parasexual já foram propostos como possíveis causadores destas aneuploidias, podendo alterar simultaneamente a dosagem de um grande número de genes em uma ou poucas gerações, permitindo a parasitos heteroxênicos se adaptarem rapidamente a transição entre os hospedeiros mamíferos e invertebrados (MANNAERT et al., 2012; MESSENGER; MILES, 2015). De fato, a avaliação do padrão de aneuploidias em *L. donovani* após passagens em cultura, no inseto vetor e no hospedeiro mamífero por RDC mostraram que existe uma seleção do padrão de aneuploidias durante esta troca de hospedeiros, e que esta seleção não é aleatória (DUMETZ et al., 2017).

Alternativamente, se mantido por longos períodos, um estado polissômico pode incorporar mutações e consequentemente permitir a evolução de novas funções em genes duplicados visto que a cópia ancestral ainda estaria presente no genoma (MANNAERT et al., 2012). Ganhos e perdas cromossômicas já foram associados a um aumento de fitness em condições de estresse e a resistência a drogas em *S. cerevisiae*, *C. albicans* e células de câncer de pulmão (ABBEY et al., 2011; DOUBRE et al., 2005; LEPROHON et al., 2009; SHELTZER et al., 2011), e podem também ser explorada pelo parasito para a seleção de fenótipos positivos.

Como visto no **capítulo 1**, o cromossomo 31 é encontrado em estado supranumerário em praticamente todas as 19 cepas de *T. cruzi* avaliadas (Figura 27). O enriquecimento de genes relacionados a glicosilação neste cromossomo reforçam a importância deste processo biológico para a sobrevivência do parasito.

Comparações na avaliação do número de aneuploidias nos cromossomos da cepa CL Brener de *T. cruzi* pela metodologia de SCoPE, utilizando tanto o haplótipo Esmeraldo-like (Figura 30 A) quanto o Non-Esmeraldo-like (Figura 30 B) como referência apresentaram o mesmo resultado. Este resultado confirma que as estimativas de CCNV não são enviesadas pelo

haplótipo de CL Brener usado como referência, de modo que comparações de mapeamentos realizados somente no haplótipo Esmeraldo-like (como as cepas de TcII) podem ser comparados com as mapeadas exclusivamente no haplótipo Non-Esmeraldo-like (como a cepas de TcI). Interessantemente, no caso de CL Brener, a razão das frequências alélicas em posições heterozigóticas pode ser utilizada para estimar, em um cromossomo trissômico, qual dos haplótipos, Esmeraldo-like ou Non-Esmeraldo-like, apresenta um cromossomo duplicado (Figura 30 C). Em caso de cromossomos triplóides um valor de 0.33 (33%) das reads corresponde a uma cópia e um valor de 0.66% das reads corresponde a duas cópias (0.33 + 0.66 = 0.99 que é aproximadamente 1, ou seja 100% das reads). Desta forma, CL Brener apresentou 6 cópias extras de cromossomos do haplótipo Non-Esmeraldo-like (1, 6, 13, 21, 24 e 41), e 4 cópias extras de cromossomos do haplótipo Esmeraldo-like (3, 26, 27 e 31). Este resultado mostra que um número semelhante de cromossomos de ambos haplótipos de CL Brener foram amplificados, sem mostrar uma predileção por aneuploidias em um haplótipo. A avaliação de um maior número de cepas dos DTUs híbridos TcV e TcVI, assim como de outros isolados de CL Brener pode fornecer informações sobre a existência de predileção de amplificação de cromossomos específicos de um dos haplótipos parentais.

12.3 Conclusões do capítulo 2

-Existe uma grande variabilidade genética entre cepas de *T. cruzi* do DTU TcII isoladas de pacientes localizados no estado de Minas Gerais, caracterizada por divergências no padrão de populações ancestrais, SNPs e aneuploidias.

-Misturas populacionais moldaram a evolução de TcII em Minas Gerais, onde os genomas das cepas atuais apresentam contribuição de 7 populações ancestrais.

-A cepa S154a apresentou o padrão mais complexo de ancestralidade, com marcadores de 5 populações ancestrais. Ela também apresentou o padrão mais divergente de SNPs entre as 7 amostras de campo de TcII avaliadas em análises de PCA e apresenta discordâncias entre as filogenias dos genes nucleares e mitocondriais. Estes resultados sugerem que a cepa S154a pode ter sofrido um maior número de eventos de recombinação dentre as sete cepas de *T. cruzi* do DTU TcII avaliadas.

-A cepa S154a foi isolada da cidade de Itaipé, a mesma cidade de isolamento da cepa S11. Porém, ambas as cepas são classificadas em posições distantes nas análises filogenéticas com base em marcadores nucleares e mitocondriais, sugerindo que diferentes populações de *T. cruzi* circulam na mesma área.

-As cepas S15 e S162a de *T. cruzi* isoladas da região central de Minas Gerais agruparam em todas as análises realizadas neste capítulo. Estas duas amostras estão separadas das outras 5 amostras de campo de *T. cruzi* do DTU TcII pela Serra do Espinhaço, uma barreira geográfica que pode ter afetado o trânsito de insetos vetores durante a evolução do DTU TcII no estado.

-O padrão de aneuploidias em *T. cruzi* varia entre DTUs, como visto pelo padrão discordante de CCNV entre as cepas de TcI, TcII, TcIII, TcV e TcVI; e varia também intra-DTU, onde todas os 7 isolados de campo de TcII apresentaram um padrão de aneuploidias diferente, o que está de acordo com o observado em espécies de *Leishmania* sp (DOWNING et al., 2011; ROGERS et al., 2011).

-Diferente do que é encontrado em *Leishmania* sp. (DOWNING et al., 2011; ROGERS et al., 2011), o padrão de aneuploidias parece ser constante dentro de uma população de *T. cruzi*, visto que os três clones da cepa Y apresentaram o mesmo padrão de aneuploidia, assim como com a população de Y não clonada.

-Análises de distância euclidiana do padrão de aneuploidias agrupam cepas de TcI, TcV e TcVI entre cepas de TcII, reforçando a hipótese que os eventos que levam a formação de aneuploidias ocorrem em maior frequência daqueles que definem a filogenia das DTUs de *T. cruzi*.

CAPÍTULO 3: Variabilidade modular de famílias multigênicas de *T. cruzi* que codificam para proteínas de superfície

13. Justificativa:

Entre os TriTryps, o parasito *T. cruzi* apresenta a maior expansão de famílias multigênicas que codificam para proteínas de superfície, muitas delas sabidamente envolvidas em interações parasito-hospedeiro (DE PABLOS; OSUNA, 2012; EL-SAYED, 2005a, 2005b). O alto conteúdo repetitivo destas regiões genômicas, a natureza híbrida, bem como a presença de vários cromossomos supranumerários como verificado nesta tese dificultaram a montagem do genoma da cepa de referência CL Brener (EL-SAYED, 2005a), resultando em uma sequência genômica distribuída em 5.489 scaffolds. Uma segunda tentativa de produzir um genoma menos fragmentado foi realizada em 2009 por Weatherly e colaboradores, com base nos contigs originais, informações adicionais de mate-pairs de BACs e mapas de sintenia com *T. brucei* e *Leishmania major*, resultando em 41 cromossomos hipotéticos (WEATHERLY; BOEHLKE; TARLETON, 2009). Porém, a montagem final destes cromossomos apresenta grandes gaps e 9.8 Mb de sequências codificadores em contigs e scaffolds não incorporadas nos 41 cromossomos, o que supera em mais de três vezes o maior cromossomo de *T. cruzi* que possui 2.3 Mb. Estes contigs não incorporados na montagem são compostos principalmente por genes que codificam para proteínas de superfície como MASPs, Mucinas e Trans-sialidases, que estão diretamente envolvidas em processos de interação parasito-hospedeiro. Estas regiões são variáveis em tamanho e sequência, tanto intra-DTU quanto entre DTUs, resultando em grande parte dos 5.9 MB de diferença entre os genomas de CL Brener (TcVI) e Sylvio (TcI). Análise de variação no conteúdo destas famílias entre amostras é dificultada por seu caráter repetitivo, o qual reduz a confiabilidade do mapeamento membro específico e também montagens *de novo*. Porém, as famílias multigênicas de *T. cruzi* apresentam motivos conservados, um indicativo da geração de variabilidade através de rearranjos de blocos definidos, ou módulos, como visto por análises de MEMEs na família MASP de CL Brener (EL-SAYED, 2005a). Desta forma, a comparação entre a abundância de diferentes módulos em cada família multigênica entre diferentes isolados de *T. cruzi* baseada em reads pode revelar o padrão de diversidade intraespecífico destas famílias, independentemente de montagens *de novo* ou mapeamento membro específico.

14. Objetivos

14.1 Objetivo Geral

Identificar variações na abundância de motivos (módulos) das famílias MASP, Mucinas e Trans-Sialidase entre diferentes isolados de *T. cruzi* a partir de reads não montadas destes genomas.

14.2 Objetivos específicos

1-Desenvolver uma metodologia independente de montagem e independente de mapeamento membro específico que permita estudar variações modulares nas famílias multigênicas de *T. cruzi*.

2-Avaliar variações na abundância destes motivos entre diferentes isolados/cepas de *T. cruzi* pertencentes a diferentes DTUs.

3-Comparar estas distribuições com a filogenia dos isolados/cepas.

Desenho Experimental:

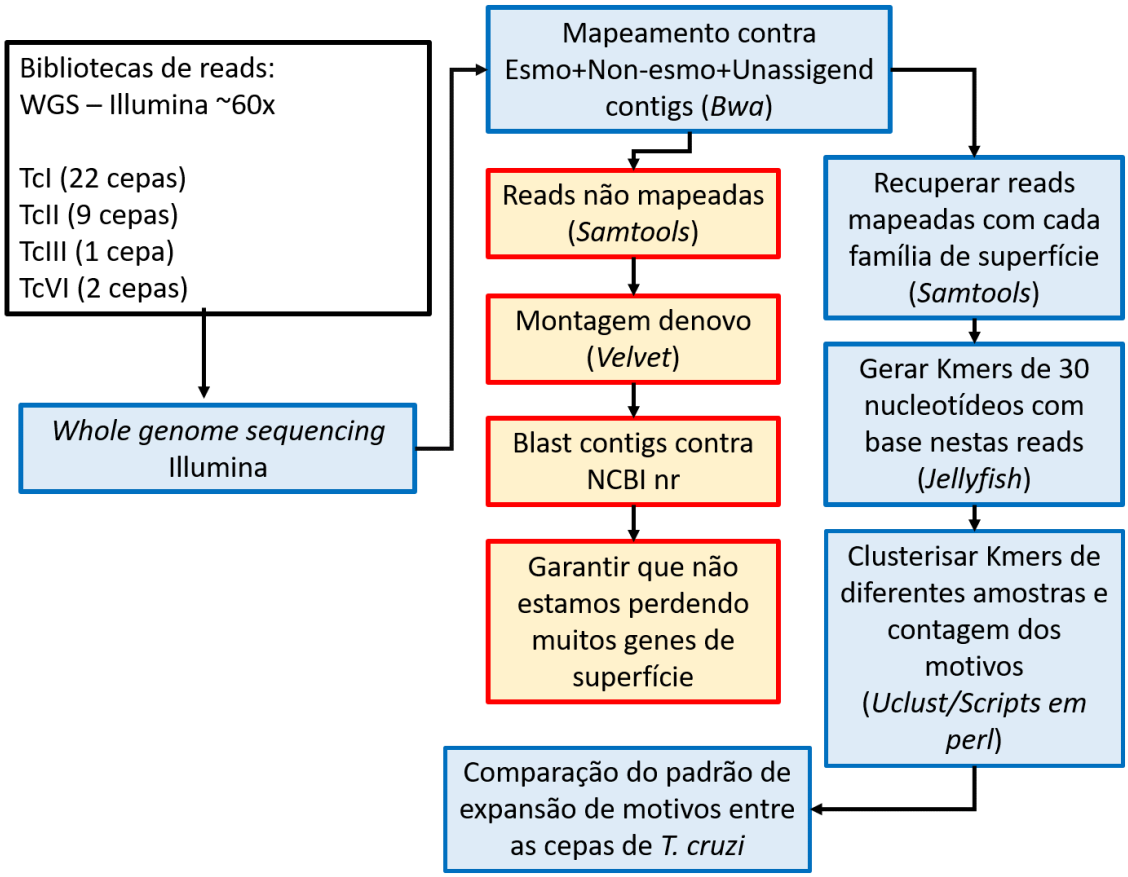


Figura 31: Mapa conceitual das análises utilizadas para realizar as contagens relativas dos motivos das famílias multigênicas MASP, Mucinas ou Trans-sialidases de *T. cruzi*. Caixas azuis correspondem a metodologia para se obter os resultados da abundância diferencial de motivos de famílias multigênicas entre diferentes cepas/isolados. Já as caixas em vermelho correspondem a metodologia para avaliar a origem das reads que não mapearam com o genoma de referência. Em todas as caixas coloridas, os termos entre parênteses correspondem aos programas utilizados nas análises.

15. Metodologia

15.1 Cepas do parasito e sequenciamento de DNA genômico

Ao todo, neste trabalho, foram utilizadas 34 bibliotecas de *whole genome sequencing* de amostras de *T. cruzi*. Destas amostras, 22 correspondem a cepas de TcI obtidas do SRA do NCBI: SRR3676264, SRR3676265, SRR3676267, SRR3676268, SRR3676272, SRR3676273, SRR3676274, SRR3676275, SRR3676276, SRR3676277, SRR3676278, SRR3676279, SRR3676282, SRR3676283, SRR3676309, SRR3676311, SRR3676312, SRR3676313, SRR3676315, SRR3676316, SRR3676317, SRR3676319. Outras 9 correspondem a cepas de TcII utilizadas e descritas no

capítulo II: S11, S154, S15, S162a, S23b S44a, Ycl2, Ycl4 e Ycl6. As ultimas 3 amostras correspondem a uma cepa de TcIII (231) e duas cepas de TcVI (CL Brener e Tulahuen).

Todas as amostras utilizadas neste trabalho foram sequenciadas utilizando sequenciadores Illumina, gerando reads de 100 bases. Mesmo quando presente, não foram utilizadas as informações de *pair-ends*, sendo que as duas coleções de reads foram agrupadas em um único arquivo tratado como *single-end*. Como a montagem das regiões que codificam proteínas de superfície na atual versão do genoma pode apresentar problemas decorrentes, por exemplo, da quebra dos contigs gerando genes truncados, a utilização de informação de *pair-end* poderia prejudicar o mapeamento das reads. As reads das 34 amostras de *T. cruzi* utilizadas neste trabalho foram filtradas com o uso do programa Trimmomatic (BOLGER; LOHSE; USADEL, 2014), utilizando um *cutoff* de *base quallity* médio de phred 30 em uma janela deslizante de 5 nucleotídeos, e tamanho mínimo de 50 bases.

15.2 Mapeamento genômico

Para a obtenção das *reads* referentes a todos os membros das famílias multigênicas TcMUC, MASP e Trans-sialidase, cada biblioteca de reads foi mapeada nos 41 cromossomos hipotéticos de CL Brener derivados dos haplótipos Esmeraldo-like e Non-Esmeraldo-like, assim como nos contigs não utilizados na montagem dos cromossomos de CL Brener (*unassigned contigs*), utilizando o programa BWA-mem (LI, 2013; LI; DURBIN, 2009). A utilização do template Esmo+Non-Esmo+unassigned é importante pois um grande número de genes das famílias em questão estão nos contigs não utilizados na montagem dos cromossomos de CL Brener, e o não uso destas regiões levaria a uma subestimativa do número de reads que mapearam em cada família. Scripts em Perl utilizando o programa SAMtools view (LI et al., 2009) e arquivos GFFs contendo as coordenadas dos genes de cada família foram utilizados para a recuperação das reads que mapearam preferencialmente em MASP, TcMUCs e Trans-sialidases. A cobertura do genoma de cada amostra foi estimada com base na cobertura de todos os 1563 genes de cópia simples, previamente descritos no subitem 4.7 “Genes cópia simples e cobertura genômica” (Tabela 4).

Cepa	DTU	Cobertura Genômica	Local de origem
SRR3676264	TcI	76.54	Colômbia
SRR3676265	TcI	73.5	Colômbia
SRR3676267	TcI	34.12	Equador
SRR3676268	TcI	45.41	Equador
SRR3676272	TcI	73.44	Texas/EUA
SRR3676273	TcI	93.89	Texas/EUA
SRR3676274	TcI	65.12	Venezuela
SRR3676275	TcI	63.53	Venezuela
SRR3676276	TcI	104.96	Colômbia
SRR3676277	TcI	51.52	Panamá
SRR3676278	TcI	40.45	Panamá
SRR3676279	TcI	35.07	Panamá
SRR3676282	TcI	43.49	Panamá
SRR3676283	TcI	40.21	Panamá
SRR3676309	TcI	39.57	Panamá
SRR3676311	TcI	48.29	Colômbia
SRR3676312	TcI	40.1	Panamá
SRR3676313	TcI	46.44	Panamá
SRR3676315	TcI	63.44	Panamá
SRR3676316	TcI	80.56	Colômbia
SRR3676317	TcI	155.61	Colômbia
SRR3676319	TcI	64.83	Colômbia
S11	TcII	34.33	Minas Gerais/Brasil
S154a	TcII	47.67	Minas Gerais/Brasil
S15	TcII	51.58	Minas Gerais/Brasil
S162a	TcII	49.57	Minas Gerais/Brasil
S23b	TcII	56.58	Minas Gerais/Brasil
S44a	TcII	36.34	Minas Gerais/Brasil
Ycl2	TcII	68.81	São Paulo/Brasil
Ycl4	TcII	75.24	São Paulo/Brasil
Ycl6	TcII	72.02	São Paulo/Brasil
231	TcII	75.42	Minas Gerais/Brasil
TcCLB	TcVI	89.64	Rio Grande do Sul/Brasil
Tulahuen	TcVI	77.23	Tulahuen/Chile

Tabela 4: Cobertura média genômica e local de isolamento das 34 cepas utilizadas nas estimativas da abundância dos motivos presentes em genes que codificam para proteínas de superfície.

15.3 Análises de reads não mapeadas

Para avaliar a ocorrência de perda de reads de famílias multigênicas por falha no mapeamento, as reads não mapeadas no template Esmo-Nonesmo-unassigned foram recuperadas, utilizando o programa *SAMtools view*. Estas reads foram montadas em *contigs*

utilizando o programa *Velvet Optimizer*, parte do pacote *Velvet* versão 1.2.10 (ZERBINO; BIRNEY, 2008). Os *contigs* gerados foram utilizados como *query* em um BLASTn contra o banco de dados NR do Genbank, para a identificação da unidade taxonômica a que pertenciam. Quando a unidade taxonômica encontrada foi Kinetoplastidae, a região genômica de *match* foi extraída com um *script* em Perl e submetida a um BLASTn contra as sequências codificadoras de proteínas de *T. cruzi* para a identificação de sua identidade.

15.4 Obtenção dos kmers

Para a obtenção dos motivos presentes em cada família multigênica para cada amostra de *T. cruzi* avaliada, a contagem de kmers de 30 nucleotídeos presentes nas reads que mapearam com cada uma das famílias MASP, Mucina ou Trans-sialidase foi gerada utilizando o programa *Jellyfish* (MARÇAIS; KINGSFORD, 2011). Este programa gera uma tabela com duas colunas, onde a primeira corresponde a sequência do kmer, e a segunda ao número de sua ocorrência nas reads. Para remover kmers oriundos de erros de sequenciamento foram inicialmente removidos todos os kmers com cobertura menor que 10. Para permitir a comparação entre as diferentes amostras de *T. cruzi*, a ocorrência de cada kmer foi normalizada dividindo a sua cobertura pela cobertura média de cada genoma (tabela 4), onde um cutoff mínimo de cobertura de kmers de 30% da cobertura diploide do genoma foi utilizado.

Para validar a metodologia, inicialmente as sequências nucleotídicas dos kmers de MASP de cepas dos DTUs TcII e TcIII foram alinhados contra as sequências proteicas dos MEMEs de MASP da cepa CL Brener previamente descritos do paper do genoma do parasito (EL-SAYED, 2005a) utilizado o programa BLASTx. Apenas matches com identidade maior que 80% e alinhamento mínimo de 9 aminoácidos (o máximo seria 10) foram recuperados. O MEME 3 foi excluído por possuir apenas 8 aminoácidos, sendo proporcionalmente menor do que os kmers e não satisfazendo os requisitos mínimos de similaridade utilizados no BLASTx. A cobertura de todos os kmers com o melhor hit de BLASTx com um determinado MEME de MASP foram somadas e assumidas como a cobertura do MEME, utilizando scripts em Perl. Os heatmaps da cobertura estimada de cada MEME foram gerados em R, utilizando a função heatmap.2 e a bibliotecas gplots.

15.5 Clusterização

Para cada uma das famílias multigênicas, os kmers únicos presentes em todas as cepas de *T. cruzi* foram clusterizados utilizando a ferramenta UCLUST com os parâmetros --optimal e -

-nofastalign (modo mais estringente e com o mínimo de heurística que garantem que todo kmer seja mapeado contra todos os centroides – ver abaixo), e com a flag --rev (para procurar matches nas orientações *foward* e *reverse*) (EDGAR, 2010). Este programa recebe como entrada um arquivo no formato fasta com as sequências dos kmers e um valor de cutoff mínimo de identidade de sequência para o alinhamento global. O programa UCLUST realiza uma clusterização com base em alinhamento global par-a-par de forma "*greedy*". De modo geral, o programa utiliza o primeiro kmer do arquivo como o primeiro centroide. Depois ele realiza um alinhamento global do segundo kmer com o único centroide atual, o primeiro kmer. Se o alinhamento apresentar uma similaridade acima do cutoff, o kmer dois é assinalado ao cluster do kmer um. Caso ele tenha uma similaridade abaixo do cutoff, o kmer dois passa a ser um centroide também. Depois, o programa alinha o terceiro kmer, par-a-par, com todos os centroides existentes e assinala o kmer 3 ao cluster do kmer que apresentar a maior similaridade. Se nenhum centroide tiver uma similaridade maior do que o *cutoff* com o kmer 3, ele passa a ser um centroide (Figura 32). O programa repete este processo até que todos os kmers do arquivo tenham sido avaliados. Foram testados valores de cutoff de identidade variando entre 75-95%, sendo que o cutoff de 75% foi escolhido para análises posteriores. O conjunto de um centroide juntamente com todos os kmers que agruparam com ele foi denominado de Motivo ou Cluster. As somas das coberturas dos kmers de cada motivo foram obtidas com scripts em Perl. As coberturas de cada motivo foram utilizadas para gerar heatmaps correspondentes a abundância de cada motivo, utilizando as bibliotecas *gplots*, *RColorBrewer* e o *heatmap.2* em R.

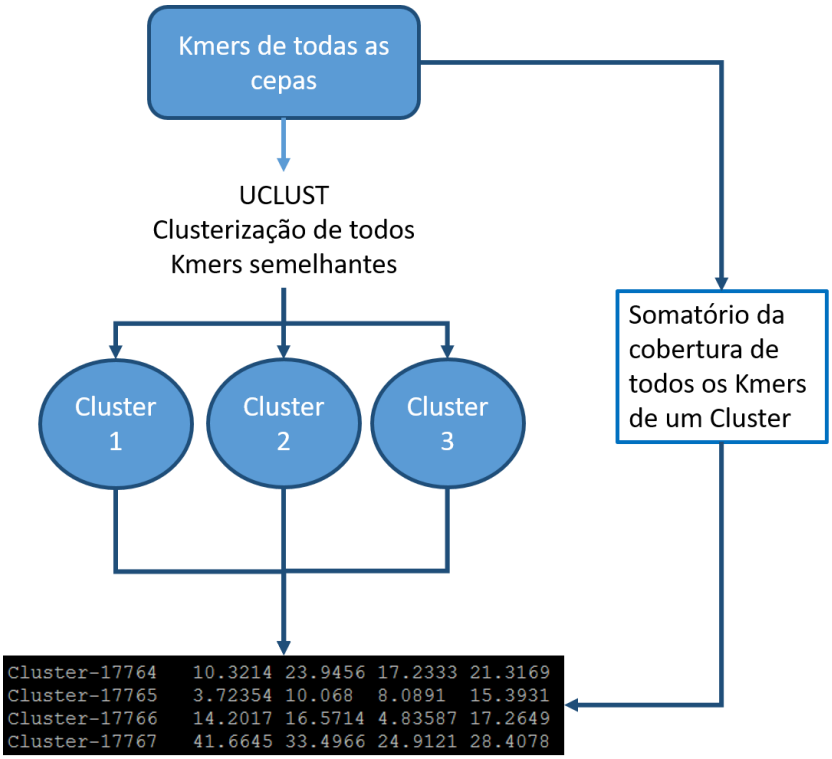


Figura 32: Mapa conceitual das análises de clusterização dos kmers. Em primeiro lugar, os kmers são gerados utilizando o programa Jellyfish (MARÇAIS; KINGSFORD, 2011). As sequências fastas de todos os kmers de todas as cepas são utilizadas como entrada para o programa UCLUST, que realiza a sua clusterização com base em similaridade de sequência global. Scripts em Perl recebem como entrada a informação de clusterização e a contagem dos kmers, realizando a soma das contagens de cada kmer em cada cluster. No final, é gerada uma tabela onde em cada linha está o número de ocorrências de um cluster e em cada coluna está uma cepa do parasito.

15.6 Análise de Componente Principal

A análise de componente principal foi realizada com base na tabela contendo as coberturas de cada motivo em cada uma das 34 cepas de *T. cruzi*, utilizando a função PCA do R e a biblioteca “rgl”, que permite a manipulação do ângulo de visualização do gráfico. Para uma melhor visualização da separação dos grupos, 3 dimensões foram utilizadas para plotar o gráfico.

Analises similares foram realizadas utilizando apenas as cepas do grupo TcI ou TcII, para a avaliação da relação entre padrão de amplificação de motivos e distribuição geográfica, assim como avaliar a variação intra-DTU.

15.7 Filogenia baseada em marcadores nucleares das cepas de *T. cruzi*

A filogenia de 33 das 34 cepas de *T. cruzi* foi estimada com base em genes nucleares de cópia simples, como previamente descrito no subitem **10.5 Análises filogenéticas**, onde um total de 760 genes de cópia simples foi utilizado. Por pertencer a um DTU híbrido, a cepa Tulahuen foi excluída das análises devido à dificuldade da montagem *de novo* de um genoma híbrido usando reads curtas, que resulta no colapso de alelos Esmeraldo-like e Non-Esmeraldo. Para CL Brener foram utilizados os genes derivados dos haplótipos Esmeraldo-like e Non-Esmeraldo-like da montagem do genoma presente no Tritryp DB versão 26. Os genomas montados das cepas Sylvio (TcI), Colombiana (TcI) Arequipa (TcI) e Esmeraldo (TcII) foram incluídos para servirem como marcadores dos DTUs TcI e TcII. Foi adicionado à análise filogenética como grupo externo os genes de *T. rangeli* das cepas Choachi e SC-58. O modelo de distribuição nucleotídica utilizado foi o GTR, com proporção de posições invariáveis de 0.53 e distribuição gama de 0.71, parâmetros estimados pelo programa Jmodeltest

15.8 Diagrama de Venn do compartilhamento de motivos entre os DTUs de *T. cruzi*

A avaliação do número de motivos específicos ou compartilhados entre os 4 DTUs de *T. cruzi* foi realizado com base em dois diagramas de Venn. No primeiro, para que um motivo fosse considerado como presente em um DTU, todas as cepas do referido DTU deveriam apresentar o motivo. Desta forma, para que um motivo fosse considerado presente no DTU TcI, ele deveria estar presente em todas as 22 cepas de TcI avaliadas. De modo semelhante, para o motivo ser considerado compartilhado apenas entre TcI e TcII, ele precisaria estar presente em todas as cepas de TcI (22) e TcII (9) e estar ausentes em todas as cepas de outros DTUs.

No segundo diagrama de Venn, para que o motivo fosse considerado como presente em um determinado DTU, ao menos uma das cepas deste DTU deveria apresentar o motivo. Desta forma, para que o motivo fosse considerado presente no DTU TcI, ele deveria estar presente em ao menos uma (podendo ser mais de uma) das 22 cepas de TcI avaliadas. De modo semelhante, para o motivo ser considerado compartilhado apenas entre TcI e TcII, ele precisaria estar presente em ao menos uma das cepas de TcI (22), estar presente em ao menos uma cepa de TcII (9) e estar ausente em todas as cepas de outros DTUs.

16. Resultados

16.1 Resultados do mapeamento das reads utilizando o template Esmo+Nonesmo+unassigned

Apesar de existirem mais bibliotecas de reads de WGS de *T. cruzi* sequenciadas disponíveis no SRA do NCBI, foram utilizadas no presente trabalho apenas as bibliotecas sequenciadas por Illumina, com aproximadamente 100 bases de extensão, que apresentaram 75% ou mais de suas reads mapeando com a referência Esmo+Nonensmo+unassigned (Figura 33). Desta forma, 34 bibliotecas de reads foram selecionadas para este trabalho.

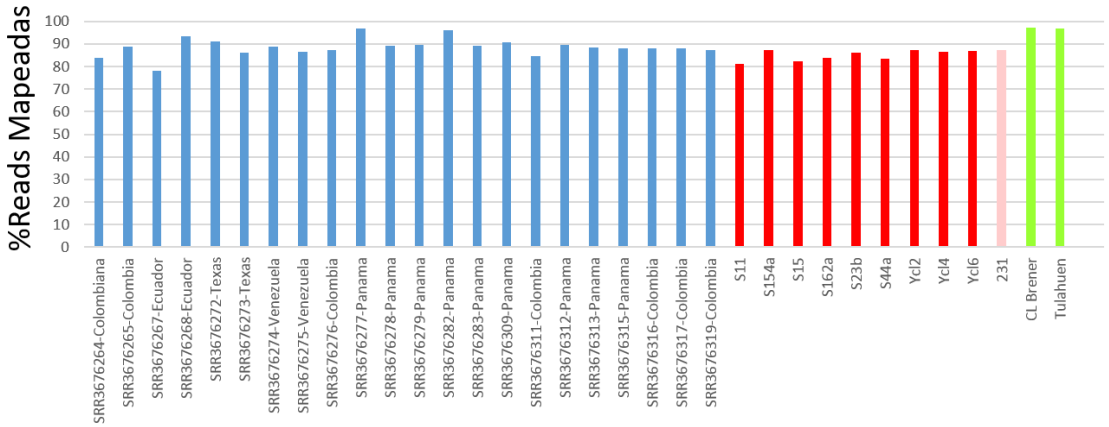


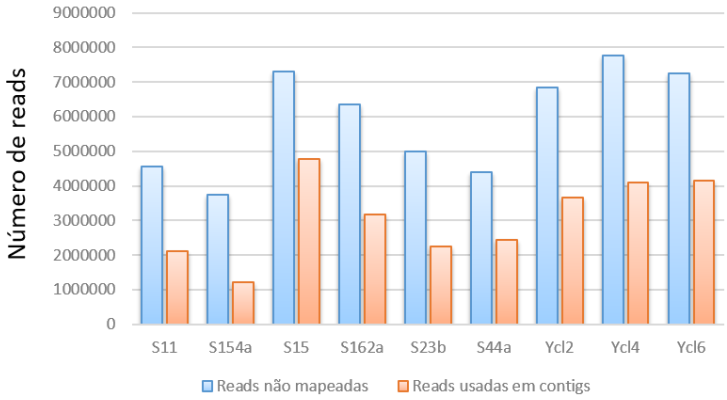
Figura 33: Porcentagem das bibliotecas de reads que mapearam com a referência Esmo+Nonesmo+unassigned. Barras em azul, vermelho, rosa e verde correspondem respectivamente a cepas do DTU TcI, TcII, TcIII e TcVI, onde a altura da barra representa a porcentagem de reads que mapeou com a referência.

16.2 Estudo das reads não mapeadas

Para assegurar a recuperação da maior quantidade de reads que correspondem às famílias multigênicas de *T. cruzi*, e garantir que não estamos perdendo quantidades relevantes destas reads por falhas no mapeamento ou por variabilidade entre os isolados, nós analisamos as reads não mapeadas das cepas de TcII de *T. cruzi*. Para isso, as reads não mapeadas foram montadas em contigs e a identidade destes contigs foi investigada através de BLASTn contra Genbank *non-reduntant database* do NCBI. Aproximadamente metade das reads não mapeadas em cada amostra foi utilizada na montagem dos contigs (Figura 34 A). A maioria destes contigs apresentou matches com regiões de *Homo sapiens*, *Gallus gallus* e membros da família Kinetoplastidae (Figura 34 B). Análises dos 204 contigs que apresentaram melhor match com

sequências de Kinetoplastidae revelaram que, 199 correspondem a hits com regiões do maxicírculo ou minicírculo, o DNA mitocondrial do parasito (Figura 34 C). O melhor match dos outros 5 contigs foram com mRNA parciais de membros da família MASP (XM_812058.1, XM_802267.1, XM_800009.1, XM_800009.1 e XM_815292.1). Desta forma, apesar da perda de algumas reads correspondentes a famílias multigênicas no mapeamento, esta perda é pequena para causar grandes impactos na análise de variabilidade destas famílias entre diferentes isolados.

A)



B)

	S11	S154a	S15	S162a	S23b	S44a	Ycl2	Ycl4	Ycl6
Contigs Total	108	168	61	104	78	190	130	129	92
Contigs Mapped	107	168	60	103	78	190	128	128	92
Homo sapiens	54	78	29	52	27	95	68	69	47
Gallus gallus	33	50	18	27	27	65	41	37	30
Mus musculus	0	0	0	0	0	0	1	1	1
Kinetoplastidae	20	40	13	24	24	30	18	21	14

C)

	S11	S154a	S15	S162a	S23b	S44a	Ycl2	Ycl4	Ycl6
Contigs kDNA	20	38	13	23	24	30	16	21	14
MASP	0	2	0	1	0	0	2	0	0

Figura 34: Origem dos contigs montados *de novo* a partir das reads não mapeadas com o template Esmo+Nonesmo+unassigned. **(A)** Número de reads que não mapearam com a referência e o número destas reads que foi utilizada na montagem *de novo* dos contigs. **(B)** Número de contigs cujos melhores matches foram atribuídos a diferentes grupos taxonômicos **(C)** Discriminação dos contigs que tiveram match coma família Kinetoplastidae.

16.3 Seleção de valor mínimo de cutoff para validação dos kmers

Para a inclusão de um dado kmer nas nossas análises, era necessário apresentar um valor mínimo de cobertura de 10 reads. Essa escolha foi feita para evitar incorporar kmers

espúrios gerados por erros de sequenciamento e que, portanto, não contribuem para a variabilidade da família. Inicialmente foram avaliados diferentes valores de cutoff de cobertura: 1 read, ou 20; 30; 40; 50; 60; 70; 80; 90 e 100% da cobertura média do genoma. O número de kmers que apresentava valores de cobertura acima de cada um destes valores de cutoffs de cobertura foi estimado para a família MASP na cepa CL Brener (Figura 35 A). A utilização de 1 read adiciona um grande número de sequencias não confiáveis à análise, visto que erros de sequenciamento seriam assumidos como polimorfismos reais. Portanto, para aumentar a confiabilidade de todos os kmers avaliados, utilizamos inicialmente um cutoff mínimo de cobertura de 10 reads. Desta forma, o número de kmers com cobertura maior ou igual a 10 reads foi assumido como o universo total de cada cepa (100%), e a porcentagem dos kmers que apresentavam cobertura acima dos cutoffs 20%, 30% e 80% da cobertura do genoma foi estimado para todas as 34 cepas avaliadas (Figura 35 B). Apesar do cutoff de 20% apresentar uma maior proximidade ao valor de 100% (mínimo de 10 reads) do que 30%, em cepas como a SRR3676267, SRR3676268, SRR3676283, S11 e S44a que apresentaram cobertura baixa, 10 reads é um valor de cutoff mais estringente do que 20% da cobertura do genoma, de modo que o cutoff precisava ser aumentado ou estas amostras deveriam ser excluídas. Optamos por elevar o cutoff para 0.3 (30%) de cobertura para que estas amostras fossem utilizadas nas análises posteriores.

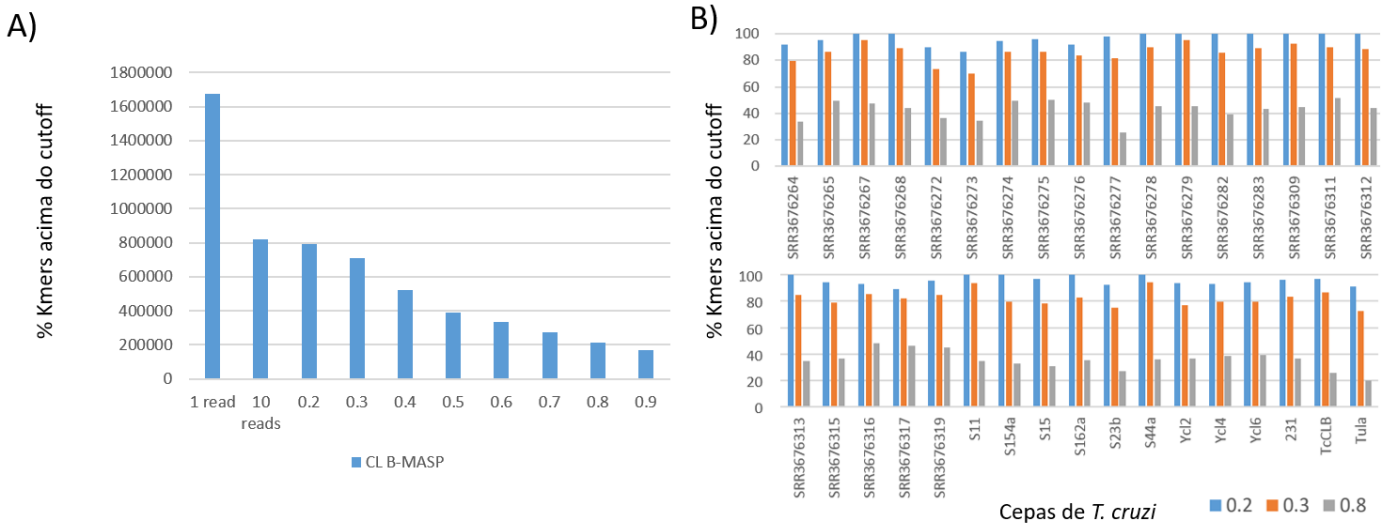


Figura 35: Analise da cobertura dos kmers. (A) Número de kmers de MAPS com cobertura acima de cutoffs de 1 read até 0.9 (90%) da cobertura do genoma haplóide de CL Brener. **(B)** Porcentagem de kmers de MASP com cobertura acima dos cutoffs 0.2 (20%), 0.3 (30%) e 0.8 (80%) da cobertura média do genoma para as 34 cepas de *T. cruzi* avaliadas, onde 100% corresponde a todos os kmers com cobertura acima de 10 reads.

16.4 Validação da metodologia: análise da abundância diferencial de motivos MEME de MASP entre as amostras de *T. cruzi*

Análises prévias sobre a estrutura e variabilidade da família MASP de CL Brener foram apresentadas no paper do genoma do parasito (EL-SAYED, 2005a). Nestas análises, todas as 771 proteínas completas de MASP (contendo as regiões conservadas N- e C-terminal, que correspondem ao peptídeo sinal e sítio de adição de ancora GPI, respectivamente) foram submetidas ao algoritmo MEME (BAILEY et al., 2006). Este programa busca pequenos motivos conservados nas sequências fornecidas pelo usuário e ranqueia estes motivos baseado no número de ocorrências e conservação dos motivos, sendo o motivo 1 o de maior frequência/conservação nas sequências, o motivo 2 apresenta a segunda maior frequência/conservação e assim por diante. Nestas análises publicadas no paper do genoma, buscou-se 20 motivos (MEMEs) derivados da família MASP. Estes motivos foram usados para validar a metodologia aqui proposta para análise da variabilidade das famílias multigênicas de *T. cruzi* baseado na abundância de kmers. Para tanto, o número e cobertura de kmers que apresentaram melhor match com cada MEME de MASP foram computados, utilizando como grupo teste os kmers de TcII e TcIII. Para isso, os kmers gerados a partir das reads que mapearam em genes da família MASP foram utilizados como *query* em *BLASTx* contra 19 dos 20 MEMEs de MASPs previamente identificados. O MEME 3 foi excluído por possuir apenas 8 aminoácidos. A cobertura dos kmers com matches significativos de BLASTX com cada MEME foi somada e a sua distribuição pode ser visualizada em heatmaps (Figura 36). Nesta análise, como esperado, os MEMEs 1 e 2, que correspondem respectivamente às regiões conservadas N- e C- terminal presentes na maioria dos genes de MASP (EL-SAYED, 2005a), foram as que apresentaram maior cobertura, validando a metodologia (Figura 36 A). Para uma melhor avaliação da variação intra-DTU TcII, um novo heatmap foi gerado, excluindo a amostra do grupo TcIII (231) assim como os MEMEs mais conservados 1 e 2 (Figura 36 B). Esta análise permitiu a observação de padrões conservados de amplificações em TcII, como os MEMEs 15 e 16, assim como variações intra-DTU, como os padrões encontrados para os MEMEs 6 e 8 (Figura 36 B).

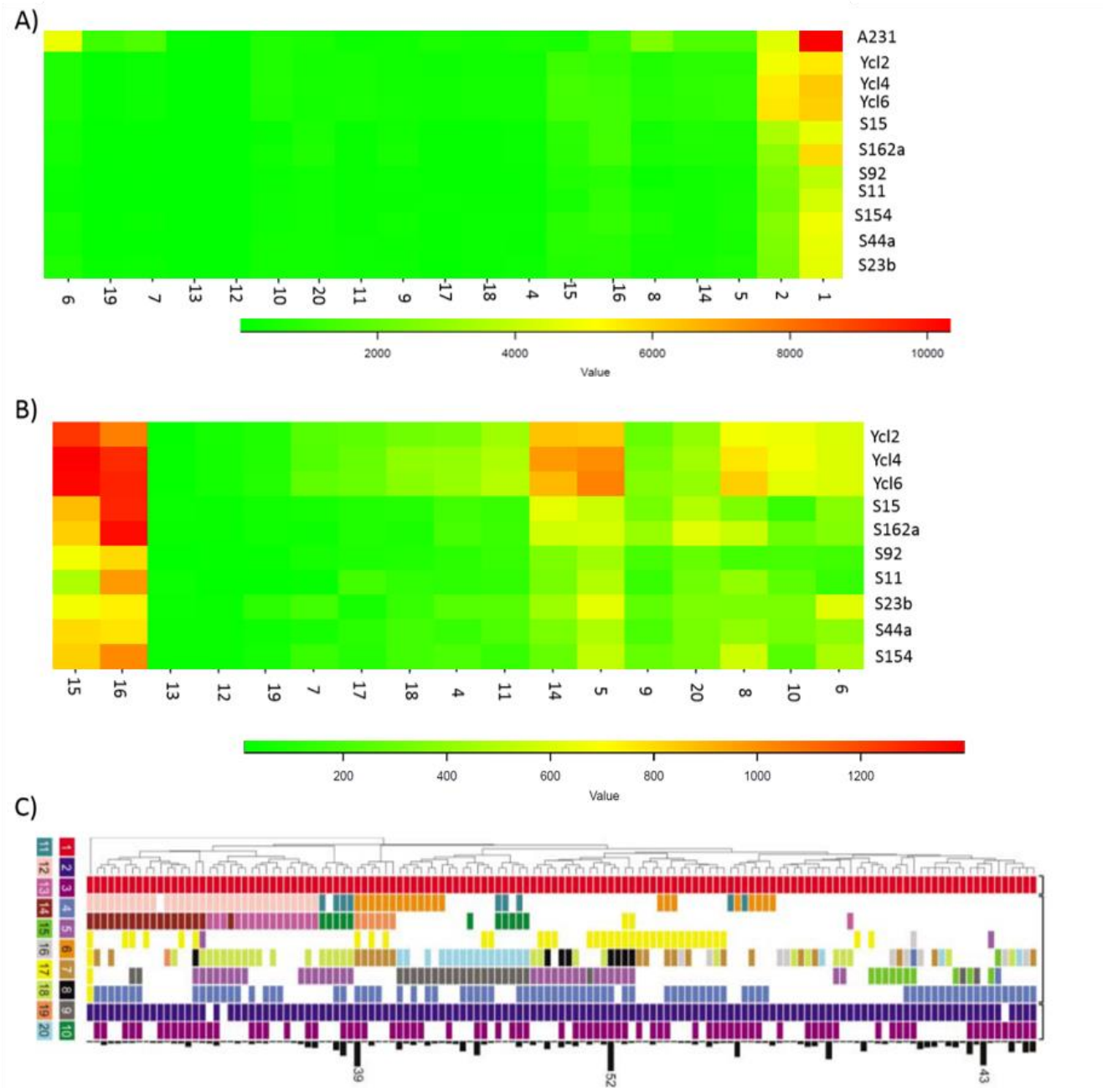


Figura 36: Padrão de variação de abundância nos motivos MEME de MASP. (A) Heatmap dos motivos MEME de MASP com base em amostras de TcII e TcIII. Neste heatmap, cada coluna corresponde a um motivo, numerado de 1 a 20, e cada linha corresponde a uma cepa de *T. cruzi*. Valores em vermelho e verde correspondem respectivamente à alto e baixo número de cópias. **(B)** Heatmap específico de MEMEs de MASP sem a amostra de TcIII e sem os MEMEs conservados 1 e 2. **(C)** Motivos MEMEs derivados da família MASP de CL Brener previamente identificados por El-Sayed et al., 2005. Os 20 MEMEs de MASP estão representados por caixas coloridas, onde cada coluna corresponde a uma combinação de motivos, formando os genes de MASP. O histograma abaixo representa o número de membros de MASP que apresentam a combinação correspondente de motivos. O motivo 1 corresponde à região conservada N-terminal (peptídeo sinal) e motivos 2 e 3, à região conservada C-terminal (sítio de adição de âncora GPI) (Retirado de (EL-SAYED, 2005a)).

Para permitir uma melhor visualização de variações na abundância dos MEMEs entre as amostras de *T. cruzi*, os valores de cobertura foram normalizados por MEME, ou seja, por cada coluna (Figura 37). Desta forma, a escala do gráfico está em Z-score, onde valores próximos de zero correspondem a valores perto da média, valores positivos denotam o número de desvios padrões que a amostra está acima da média, e valores negativos o número de desvios padrões que a amostra está abaixo da média. É importante ressaltar que nesta distribuição, por normalizar cada coluna (MEME) separadamente, as comparações de abundância entre diferentes MEMEs não devem ser realizadas, sendo então as análises limitadas a comparação de cada MEME entre amostras. Para a visualização do padrão de variação entre as amostras de campo de TcII, foi gerado um heatmap excluindo a amostra 231 (TcIII) e os clones de Y, revelando grandes diferenças no padrão de MEMEs expandidos ou retraídos entre as amostras de campo de TcII (Figura 37 B). Esta análise revelou um padrão altamente variável de distribuição das abundâncias relativas entre todas as amostras, em especial para os MEMEs 5, 10, 13, 15 e 16 (Figura 37 B).

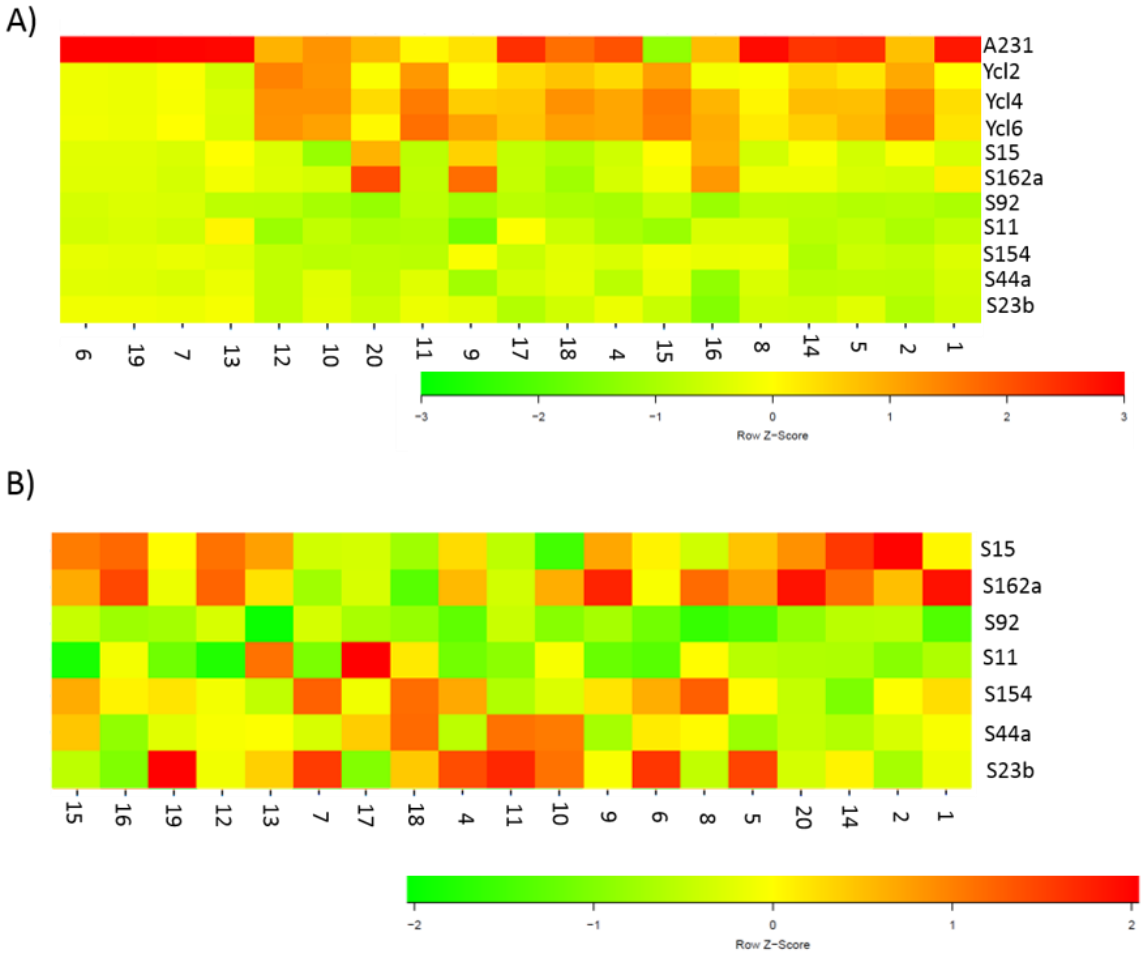


Figura 37: Variação de abundância relativa nos motivos MEME de MASP previamente descritos (EL-SAYED, 2005a) entre diferentes isolados de *T. cruzi*. Heatmaps gerados a partir da normalização das abundâncias relativas por MEME, por *Z-score*. **(A)** Heatmap dos motivos

MEME de MASP presentes nas amostras de TcII e TcIII. Neste heatmap, cada coluna corresponde a um motivo, numerado de 1 a 20, e cada linha corresponde a uma cepa de *T. cruzi*. Valores em verde e vermelho correspondem respectivamente à alto e baixo número de cópias. **(B)** Heatmap específico das amostras de campo de TcII.

16.5 Otimização dos parâmetros de clusterização

Após a validação dos parâmetros mínimos de cobertura para a inclusão dos kmers nas análises, o próximo passo foi a determinação dos parâmetros de clusterização. Como sequências que variam em apenas um SNP geram kmers diferentes, para permitir uma comparação mais correta entre motivos representativos e presentes nas diferentes amostras, os diferentes kmers de cada família foram clusterizados utilizando o programa UCLUST (EDGAR, 2010). Este programa realiza clusterizações "*greedy*" com base no alinhamento global par-a-par entre sequências e centroides de clusters. Um exemplo de todos os kmers agrupados em um cluster pode ser visualizado na Figura 38.

Figura 38: Exemplo de alinhamento múltiplo de kmers em um cluster. Todos os kmers atribuídos a um cluster específico utilizando o programa UCLUST, com um valor mínimo de similaridade de 0.75 foram alinhados utilizando o programa MUSCLE. Kmers oriundos de janelas de leitura consecutivas, assim como apresentando um ou poucos SNPs são agrupados juntos.

Para avaliar a eficiência do programa UCLUST, um dataset menor correspondendo aos kmers de MASP da cepa Ycl2 foi utilizado. Como critério de seleção, foram analisados apenas os kmers que possuíam matches de BLASTx contra os MEMEs de MASP com identidade maior ou igual a 80%, uma extensão de alinhamento de ao menos 9 aminoácidos (máximo seria 10) e que a soma de *missmatches* e gaps não fosse maior que 3. O resultado do BLASTx deste dataset restrito foi importante para funcionar como um padrão ouro, permitindo a determinação de cada um destes kmers a um MEME específico antes da clusterização. Com base neste dataset, um total de 13.664 kmers que tiveram matches com um dos 19 MEMEs de MASP foram utilizados como input no UCLUST, com valores de cutoff de identidade variando de 0.75 até 0.95 (75% a 95% de identidade). A eficiência de clusterização foi avaliada com base em 3 parâmetros:

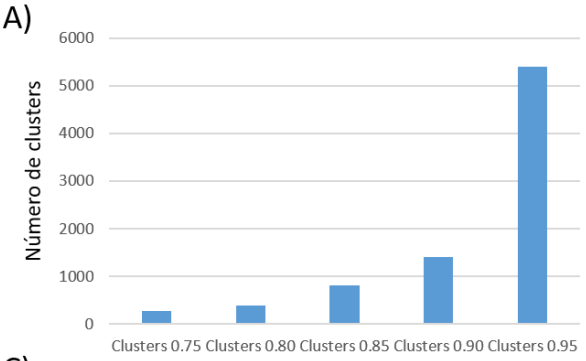
I) Obtenção do menor número possível de clusters;

II) Todos os kmers de um determinado cluster deveriam apresentaram melhor match contra o mesmo MEME de MASP;

III) O coeficiente de variação (razão do desvio padrão da cobertura de todos os kmers de um mesmo cluster pela média de cobertura) da RDC de kmers em um mesmo cluster que vieram de uma mesma amostra biológica deve ser semelhante, visto que grande parte dos kmers em um mesmo cluster são janelas de leitura diferentes da mesma sequência;

Entre os cutoffs de identidade avaliados, 0.75 apresentou o menor número de clusters (277), quando comparados aos outros cutoffs 0.80 (383), 0.85 (817), 0.90 (1.405) e 0.95 (5.388) (Figura 39 A e B). O número de cluster com 1 kmer, com menos de 5 kmers e com menos de 10 kmers também foram avaliados, mostrando que o cutoff de 0.75 de identidade levou a uma condensação maior da informação do que os outros cutoffs (Figura 39 C). Quanto à especificidade, para todos os valores de cutoff, todos os kmers de um mesmo cluster tiveram o melhor resultado de BLAST com o mesmo MEME, mostrando que o UCLUST não agrupa sequências divergentes, mesmo utilizando um cutoff permissivo de 0.75 (dados não mostrados). Finalmente, o coeficiente de variação da RDC de todos os kmers em cada cluster para o cutoff de 0.75 foram calculados (Figura 39 D). A maioria dos cluster apresentou um valor de coeficiente de variação baixo, enquanto poucos apresentaram um valor alto. A presença de um valor baixo de variação é esperada, devido à grande parte dos cluster serem formados por kmers de janelas de leitura consecutivas. Porém, em alguns casos podem ser agrupados kmers de regiões com grande similaridade de sequência originadas de diferentes regiões genômicas, e que apresentam um número de cópias diferentes, o que poderia acarretar em um aumento do coeficiente de variação. A ocorrência de baixos valores de coeficiente de variação para a maioria dos clusters, assim como o menor número de clusters e a ausência de kmers de diferentes MEMEs em um mesmo cluster suportam o uso de 0.75 como cutoff para clusterização em análises posteriores.

A)



B)

Cutoff de Similaridade	Numero de clusters
Clusters 0.75	277
Clusters 0.80	383
Clusters 0.85	817
Clusters 0.90	1405
Clusters 0.95	5388

C)

	1 Kmer	< 5 Kmers	< 10 Kmers
Clusters 0.75	2	23	46
Clusters 0.80	5	47	87
Clusters 0.85	16	176	377
Clusters 0.90	65	530	956
Clusters 0.95	1103	5271	5385

D)

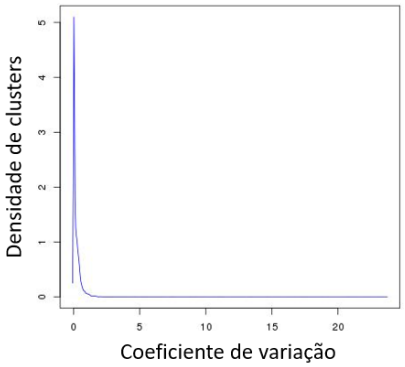


Figura 39: Otimização dos parâmetros de clusterização. Avaliação da clusterização utilizando os kmers de Ycl2 que mapearam nos MEMEs de MASP como padrão ouro. **(A e B)** Número de clusters gerados ao se alterar o cutoff de identidade do UCLUST. **(C)** Número de clusters que apresentaram RDC de 1 kmer, menor que 5 kmers ou menor que 10 kmers nas análises baseadas em cada um dos cutoffs de identidade avaliados. **(D)** Coeficiente de variação das RDC dos kmers em cada cluster obtidos pelo UCLUST utilizando um cutoff de similaridade de 0.75.

16.6 Obtenção dos kmers e clusterização das famílias multigênicas MASP, TcMUC e Trans-sialidase

Uma vez que a metodologia foi validada com motivos previamente identificados na família MASP e o cutoff mínimo de similaridade do UCLUST foi selecionado como 75%, as análises foram expandidas para todo o repertório de kmers derivado das famílias multigênicas MASP, TcMUC e trans-sialidase de todas as cepas de *T. cruzi* dos 4 DTUs avaliados.

O número total de kmers com sequência única com cobertura acima de 30% da cobertura média do genoma gerados para as famílias MASP (Figura 40 A), TcMUC (Figura 40 B) e Trans-sialidases (Figura 40 C) foi contabilizado. As cepas híbridas do DTU TcVI, CL Brener e Tulahuen, apresentam um maior número de kmers diferentes do que as cepas dos outros DTUs para todas as três famílias multigênicas avaliadas. De maneira interessante, a amostra SRR3676277 classificada como TcI apresentou uma contagem muito superior às outras cepas de TcI, com valores comparáveis as cepas híbridas de TcVI.

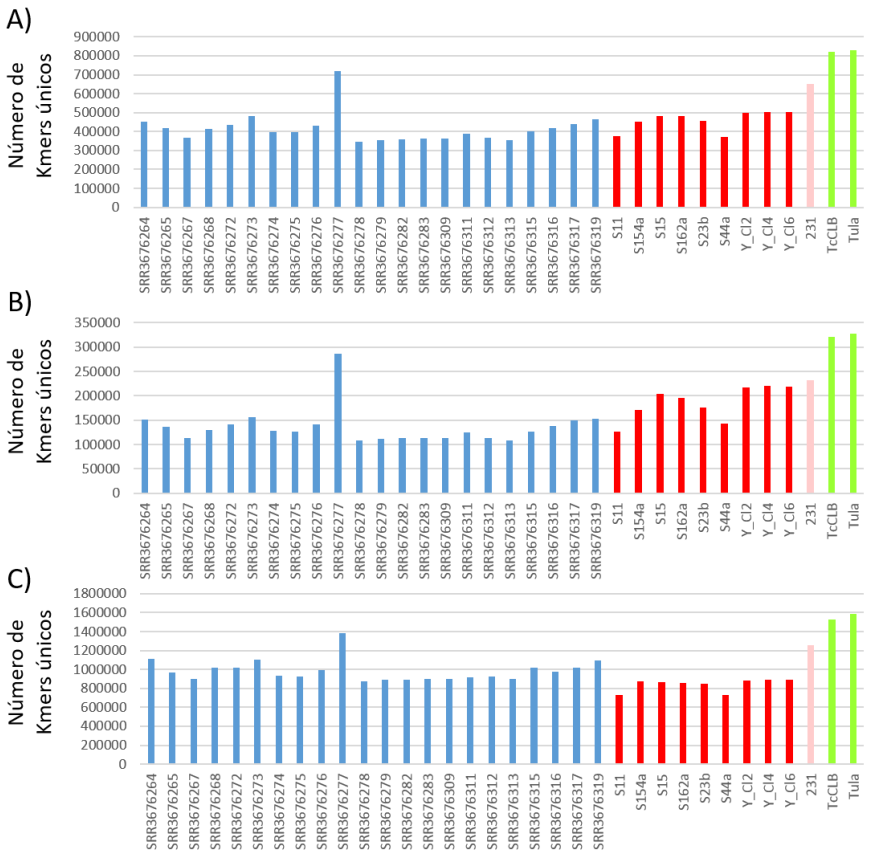


Figura 40: Kmers únicos das famílias multigênicas presentes em cada cepa de *T. cruzi*. O número de kmers únicos gerados com base nas reads que mapearam nas famílias multigênicas **(A)** MASP, **(B)** TcMUC e **(C)** Trans-sialidase. Barras em azul, vermelho, rosa e verde correspondem, respectivamente, a cepas pertencentes a DTU TcI, TcII, TcIII e TcVI avaliadas.

A partir das reads que mapearam com cada família multigênica em todas as 34 cepas avaliadas foram gerados 2.147.142, 807.025 e 4.347.190 kmers únicos, respectivamente, para as famílias MASP, TcMUC e Trans-sialidase, que foram agrupados em 40.794 clusters para MASP, 14.934 para TcMUC e 53.057 para trans-sialidase. A avaliação do número de clusters diferentes de TcMUC, MASP e trans-sialidase em cada cepa mostrou que as cepas do grupo TcVI apresentaram um maior número de motivos diferentes para as três famílias multigênicas, assim como a amostra SRR3676277, classificada no banco de dados como TcI. Para a família TcMUC, as cepas de TcII e TcIII apresentaram um número intermediário e cepas de TcI apresentaram um menor número de motivos. Para a família das trans-sialidases, um número comparável de motivos entre cepas de TcI e TcII foi observado, sendo que TcIII apresentou um número maior de motivos (Figura 41). Tanto para a família MASP, quanto trans-sialidase, TcIII apresentou um número maior de motivos em relação à TcII e menor em relação à TcVI.

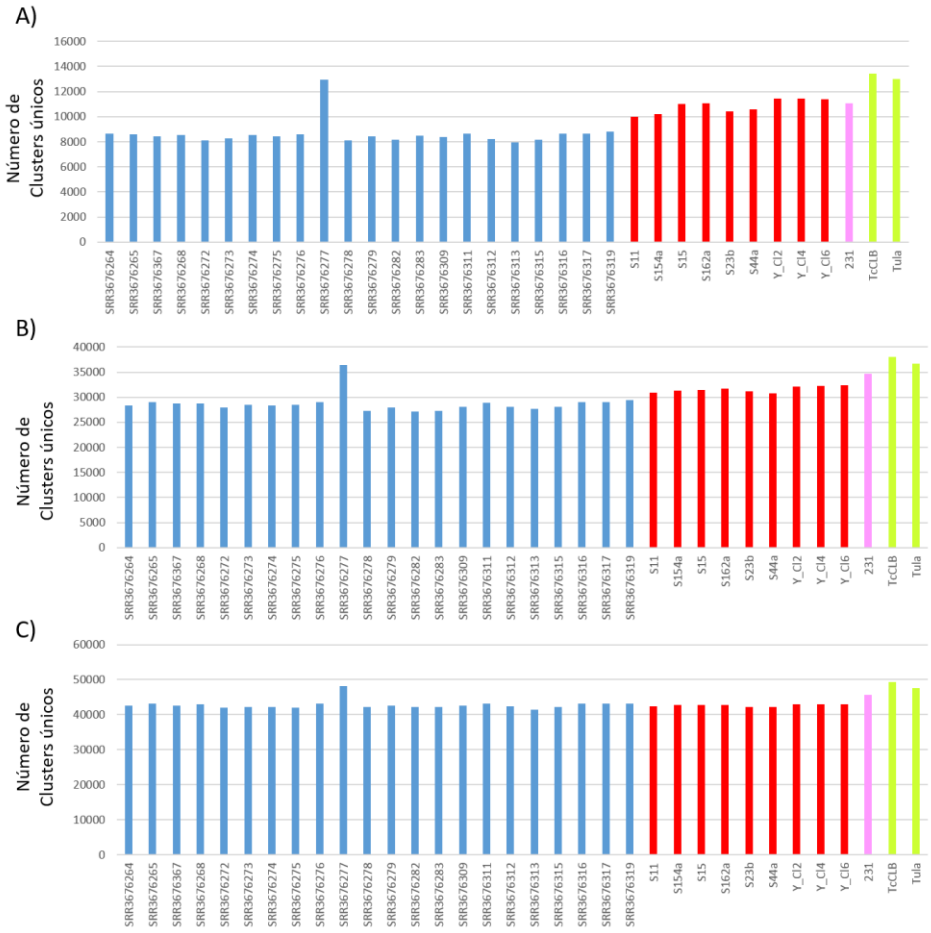


Figura 41: Clusters únicos das famílias multigênicas presentes em cada cepa de *T. cruzi*. O número de clusters únicos gerados com base na clusterização dos kmers das famílias multigênicas **(A)** TcMUC, **(B)** MASP e **(C)** trans-sialidase. Barras em azul, vermelho, rosa e verde correspondem, respectivamente, a cepas pertencentes a DTUs TcI, TcII, TcIII e TcVI avaliadas.

Para ilustrar a variabilidade de motivos presentes em cada família multigênica, a abundância relativa de cada cluster para MASP, TcMUC e trans-sialidase foi representada em heatmaps normalizados por cluster (Z-score) como previamente realizado para os MEMEs de MASP, porém desta vez cada coluna corresponde a uma cepa e cada linha a um cluster. Esta análise mostrou que existe uma grande variabilidade entre os motivos que apresentam grande número de cópias entre os DTUs de *T. cruzi* para as três famílias multigênicas MASP, TcMUC e Trans-sialidase (figuras 42, 43 e 44, respectivamente). Interessantemente, os heatmaps das três famílias apresentam padrões semelhantes. É possível visualizar grandes diferenças entre os motivos expandidos e retraídos nos DTUs TcI e TcII, onde a presença de alto número de cópias em um, corresponde a um pequeno número de cópias no outro. Isso sugere a ocorrência de expansões de motivos típicas de cada uma destes DTUs, o que pode ser um resultado da longa

divergência evolutiva e possível ausência ou baixa frequência de recombinação entre DTUs TcI e TcII. A única amostra de TcIII, 231, apresentou um padrão destoante de TcI e TcII, mas com vários motivos compartilhados com TcVI. As amostras do grupo TcVI, CL Brener e Tulahuen, apresentaram o maior número de clusters diferentes, englobando motivos presentes em TcII e TcIII, assim como motivos únicos, o que pode ser um reflexo da natureza híbrida deste DTU. Os três clones da cepa Y, Ycl2, Ycl4 e Ycl6 apresentaram um padrão extremamente similar de amplificação de motivos, o que reforça a robustez da metodologia e sugere uma baixa variabilidade intra-populacional da cepa Y. É importante ressaltar, entretanto, que pequenas diferenças entre os três clones de Y são observadas. Outro padrão observado nos heatmaps de ambas as famílias MASP e TcMUC foi o obtido pela amostra SRR3676267. Apesar desta amostra estar descrita no banco de dados do NCBI como pertencente ao DTU TcI, ela apresenta um perfil de motivos discordantes com outras amostras deste DTU, agrupando com CL Brener com base no padrão de clusters amplificados/reduzidos de mucinas (Figura 42 B) e trans-sialidases (Figura 44 B), o que sugere que esta cepa possa ter uma origem híbrida. Resultados semelhantes foram observados nas análises dos números de kmers e clusters únicos (Figuras 40 e 41), reforçando esta hipótese.

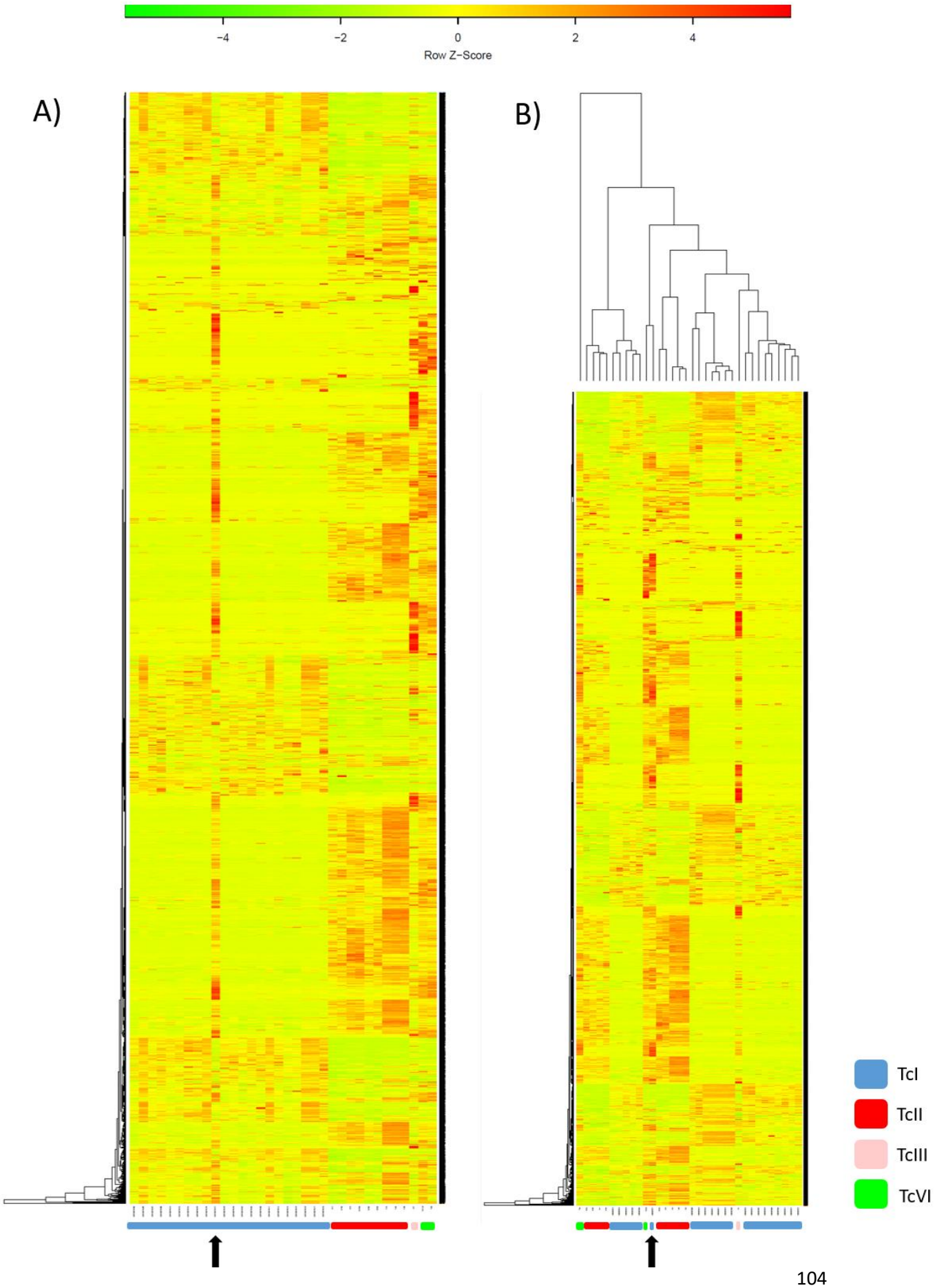

−4
−2
0
2
4
Row Z−Score
A)
B)
TcI
TcII
TcIII
TcVI

Figura 42: Heatmap da amplificação/deleção de motivos da família TcMUC de *T. cruzi*. Nesta figura, cada linha corresponde a um motivo diferente de TcMUC, e cada coluna a uma cepa/isolado de *T. cruzi*. Motivos em vermelho correspondem a um alto número de cópias, enquanto em verde correspondem a baixo número de cópias, com base em valores de Z-score. Abaixo dos gráficos, as linhas coloridas denotam os DTUs de *T. cruzi*, onde azul corresponde a amostras de TcI, vermelho a TcII, rosa a TcIII e verde a TcVI. A seta preta está apontando para a amostra SRR3676277, que apresenta um padrão destoante das amostras de TcI. **(A)** Heatmap organizado para respeitar os DTUs, onde a ordem das amostras, da esquerda para a direita é: SRR3676264, SRR3676265, SRR3676267, SRR3676268, SRR3676272, SRR3676273, SRR3676274, SRR3676275, SRR3676276, SRR3676277, SRR3676278, SRR3676279, SRR3676282, SRR3676283, SRR3676309, SRR3676311, SRR3676312, SRR3676313, SRR3676315, SRR3676316, SRR3676317, SRR3676319, S11, S154, S15, S162a, S23b, S44a, Ycl2, Ycl4, Ycl6, 231, TcCLB e Tulahuen. **(B)** Heatmap onde a reordenação da ordem das cepas de *T. cruzi* pelo programa heatmap.2 do R, com base valor médio dos clusters de cada coluna (cepa) foi permitida. Neste gráfico, a ordem das amostras da esquerda para a direita é: Tulahuen, S44a, S23b, S11, S154a, SRR3676313, SRR3676315, SRR3676278, SRR3676282, SRR3676283, TcCLB, SRR3676277, S162a, S15, Ycl2, Ycl6, Ycl4, SRR3676268, SRR3676319, SRR3676276, SRR3676311, SRR3676317, SRR3676316, SRR3676265, 231, SRR3676275, SRR3676274, SRR3676267, SRR3676279, SRR3676309, SRR3676264, SRR3676312, SRR3676272, SRR3676273.

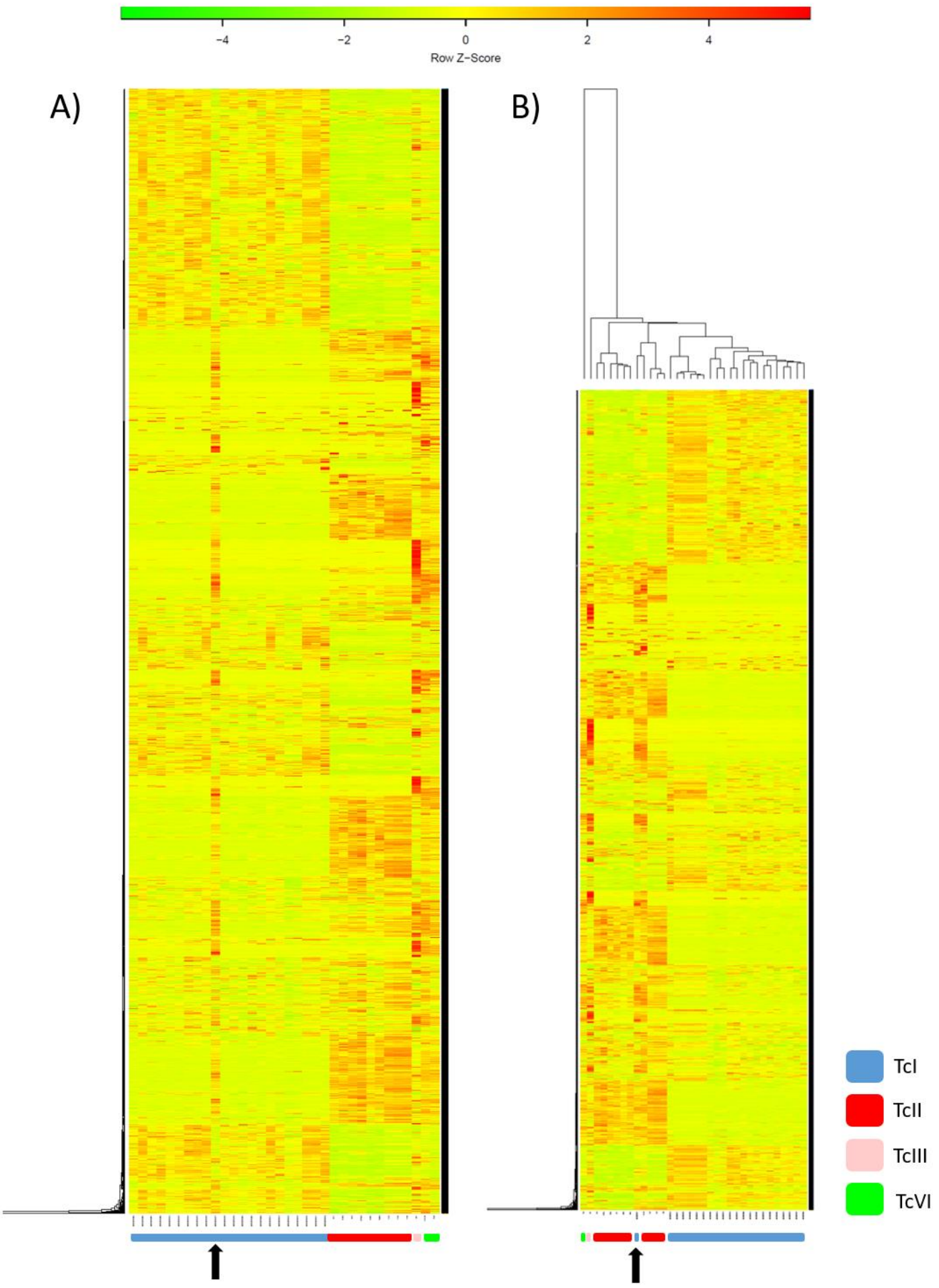

−4
−2
0
2
4
Row Z−Score
A)
B)
TcI
TcII
TcIII
TcVI

Figura 43: Heatmap da amplificação/deleção de motivos da família MASP de *T. cruzi*. Nesta figura, cada linha corresponde a um motivo diferente de MASP, e cada coluna a uma cepa de *T. cruzi* diferente. Motivos em vermelho correspondem a um alto número de cópias, enquanto em verde correspondem a baixo número de cópias, com base em valores de Z-score. Abaixo dos gráficos, as linhas coloridas denotam os DTUs de *T. cruzi*, onde azul corresponde a amostras de TcI, vermelho a TcII, rosa a TcIII e verde a TcVI. A seta preta está apontando para a amostra SRR3676277, que apresenta um padrão destoante das amostras de TcI. **(A)** Heatmap organizado para respeitar os DTUs, onde a ordem das amostras, da esquerda para a direita é: SRR3676264, SRR3676265, SRR3676267, SRR3676268, SRR3676272, SRR3676273, SRR3676274, SRR3676275, SRR3676276, SRR3676277, SRR3676278, SRR3676279, SRR3676282, SRR3676283, SRR3676309, SRR3676311, SRR3676312, SRR3676313, SRR3676315, SRR3676316, SRR3676317, SRR3676319, S11, S154, S15, S162a, S23b, S44a, Ycl2, Ycl4, Ycl6, 231, TcCLB e Tulahuen. **(B)** Heatmap onde a reordenação da ordem das cepas de *T. cruzi* pelo programa heatmap.2 do R, com base valor médio dos clusters de cada coluna (cepa) foi permitida. Neste gráfico, a ordem das amostras da esquerda para a direita é: Tulahuen, 231, S15, S162a, S154, S11, S23b, S44a, SRR3676277, SRR3676277, TcCLB, Ycl2, Ycl4, Ycl6, SRR3676319, SRR3676311, SRR3676276, SRR3676265, SRR3676317, SRR3676316, SRR3676282, SRR3676313, SRR3676315, SRR3676275, SRR3676274, SRR3676268, SRR3676312, SRR3676267, SRR3676279, SRR3676309, SRR3676264, SRR3676272, SRR3676273, SRR3676278, SRR3676283.

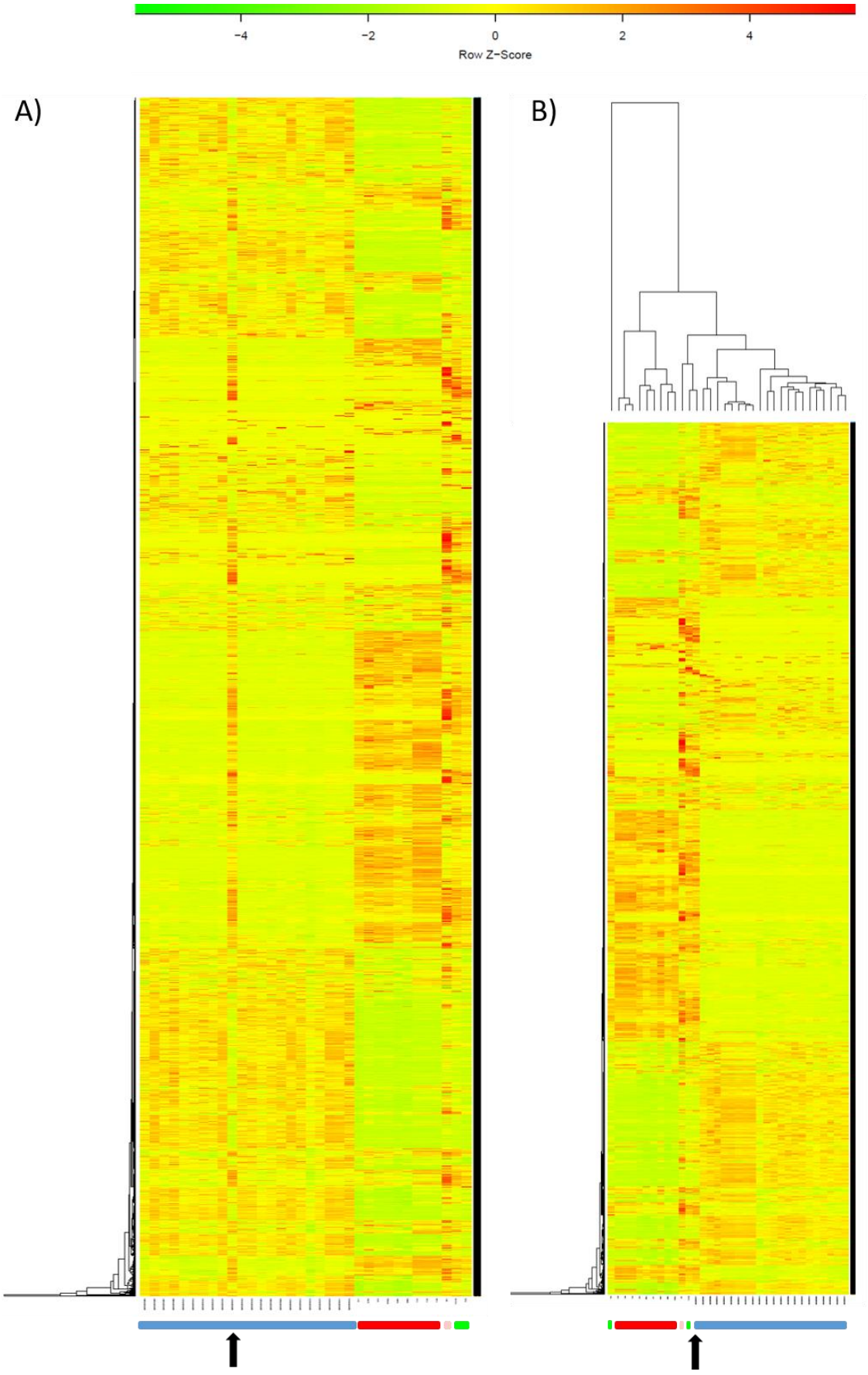

−4
−2
0
2
4
Row Z−Score
A)
B)

Figura 44: Heatmap da amplificação/deleção de motivos da família trans-sialidase de *T. cruzi*. Nesta figura, cada linha corresponde a um motivo diferente de trans-sialidase, e cada coluna a uma cepa de *T. cruzi* diferente. Motivos em vermelho correspondem a um alto número de cópias, enquanto em verde correspondem a baixo número de cópias, com base em valores de Z-score. Abaixo dos gráficos, as linhas coloridas denotam os DTUs de *T. cruzi*, onde azul corresponde a amostras de TcI, vermelho a TcII, rosa a TcIII e verde a TcVI. A seta preta está apontando para a amostra SRR3676277, que apresenta um padrão destoante das amostras de TcI. **(A)** Heatmap organizado para respeitar os DTUs, onde a ordem das amostras, da esquerda para a direita é: SRR3676264, SRR3676265, SRR3676267, SRR3676268, SRR3676272, SRR3676273, SRR3676274, SRR3676275, SRR3676276, SRR3676277, SRR3676278, SRR3676279, SRR3676282, SRR3676283, SRR3676309, SRR3676311, SRR3676312, SRR3676313, SRR3676315, SRR3676316, SRR3676317, SRR3676319, S11, S154, S15, S162a, S23b, S44a, Ycl2, Ycl4, Ycl6, 231, TcCLB e Tulahuen. **(B)** Heatmap onde a reordenação da ordem das cepas de *T. cruzi* pelo programa heatmap.2 do R, com base valor médio dos clusters de cada coluna (cepa) foi permitida. Neste gráfico, a ordem das amostras da esquerda para a direita é: Tulahuen, Ycl2, Ycl6, Ycl4, S15, S44a, S11, S154, S23b, S162a, 231, TcCLB, SRR3676277, SRR3676267, SRR3676268, SRR3676319, SRR3676316, SRR3676265, SRR3676311, SRR3676276, SRR3676317, SRR3676313, SRR3676275, SRR3676274, SRR3676312, SRR3676309, SRR3676278, SRR3676279, SRR3676315, SRR3676282, SRR3676283, SRR3676264, SRR3676272, SRR3676273.

Para determinar a diferença entre os padrões de amplificação de motivos entre as 34 cepas de *T. cruzi* avaliadas, uma análise de componente principal (PCA) foi realizada. Como os valores dos 5 maiores componentes principais para TcMUC foram de: PC1 = 38%; PC2 = 17%; PC3 = 8%; PC4 = 5% e PC5 = 4%; para MASP: PC1 = 36,8%; PC2 =17,2%; PC3 = 7,6%; PC4 = 5,2% e PC5 = 3,9%; e para trans-sialidase: PC1 = 44,3%; PC2 = 16,8%; PC3 = 6,0%; PC4 = 4,3% e PC5 = 3,8%; a utilização de 3 dimensões já englobaria ~62% para TcMUC, ~61,6% para MASP e 67,1% para trans-sialidase da divergência entre as sequências. Por isso, gráficos com três dimensões do PCA foram gerados. Para as três famílias, a distribuição espacial dos isolados mostrou que a metodologia aqui proposta apresenta uma resolução bastante satisfatória no tocante à separação dos diferentes DTUs, os quais formaram clusters distantes (Figura 45). As amostras de TcI (em azul) apresentaram uma maior separação do que as amostras de TcII (em vermelho). As duas amostras de TcVI (em verde) agruparam em posição entre as amostras de TcII e de TcIII novamente reforçando o caráter híbrido deste DTU. A amostra SRR3676277 (em preto) está localizada próxima as cepas de TcVI, reforçando o fato de que ela não pertence a TcI, que pode ser possivelmente um representante de TcVI ou é o resultado de uma hibridização entre cepas de outras DTUs de *T. cruzi*.

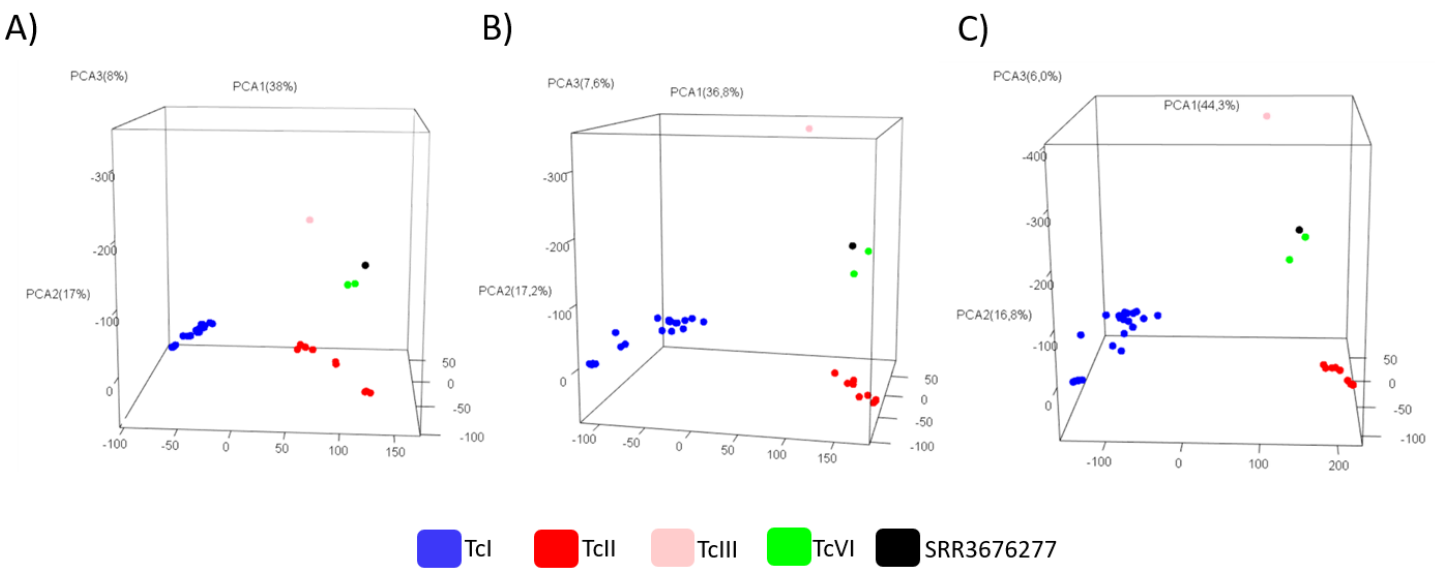


Figura 45: Análise de componente principal da distribuição dos Cluster de TcMUC, MASP e Trans-sialidase entre DTUs. Representação tridimensional do padrão de diversidade dos motivos de **(A)** TcMUC, **(B)** MASP e **(C)** trans-sialidase entre as 34 cepas/isolados de *T. cruzi* por PCA. Nesta imagem, cada ponto corresponde a uma cepa/isolado, onde pontos em azul, vermelho, rosa e verde correspondem, respectivamente, a representantes dos DTUs TcI, TcII, TcIII e TcVI. A amostra SRR3676277, depositada no NCBI como pertencente ao DTU TcI, está representada em preto. A distância entre os pontos é proporcional à distância entre os padrões de amplificações/deleções de motivos entre as amostras.

Após avaliar o padrão de variação de amplificações de motivos entre DTUs, o próximo passo foi a avaliação desta variação dentro dos DTUs TcI e TcII (Figura 46). Inicialmente as 21 (excluindo a SRR3676277) cepas de TcI, que correspondem a 7 amostras da Colômbia, 2 do Equador, 2 do Texas, 2 da Venezuela e 8 do Panamá foram utilizadas em uma análise de PCA. Os valores de PC1, PC2 e PC3 para TcMUC (Figura 46 A), MASP (Figura 46 B) e trans-sialidase (Figura 46 C) foram de respectivamente 26,2, 24,5 e 22,2%; 11,7, 10,5 e 13,7% e 9,0, 9,4 e 10,2%. Para as três famílias, uma certa relação entre distribuição geográfica e distância no gráfico de PCA pode ser observada. Das 7 amostras da Colômbia, cinco clusterizaram próximas entre si, enquanto a amostra SRR3676264 agrupou com maior proximidade às amostras do Panamá e a amostra SRR3676319 ficou distante das demais amostras estudadas. Em seguida, foram avaliadas as cepas de TcII, que correspondem a 2 amostras da região central (S15 e S162a) e 4 amostras da região norte/nordeste de Minas Gerais, assim como 3 clones da cepa Y. Os valores de PC1, PC2 e PC3 para TcMUC (Figura 46 D), MASP (Figura 46 E) e trans-sialidase (Figura 46 F) entre as cepas de TcII foram de respectivamente 43,6, 31,7 e 35.3%; 18,8, 21,7 e 18.8% e 10,1, 14,7 e 16,0%. Novamente uma relação entre a dispersão das amostras e a sua origem geográfica

pode ser realizada. As amostras S23B e S11 apresentaram um padrão extremamente próximo, comparável ao daquele dos clones de Y, para a família MASP (Figura 46 E), mas não para as famílias TcMUC e trans-sialidase.

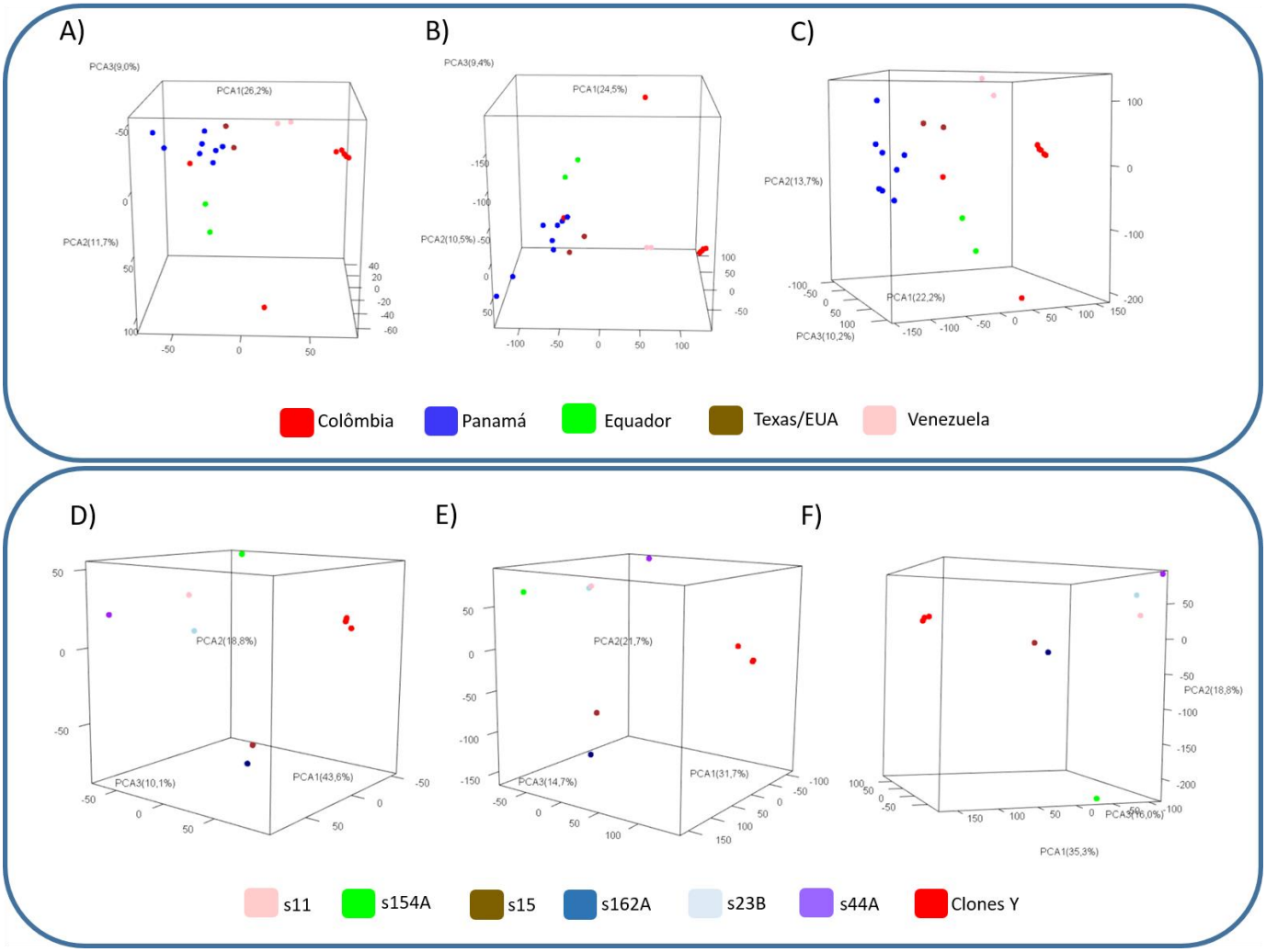


Figura 46: Análise de componente principal da distribuição dos cluster de MASP, TcMUC e trans-sialidase entre amostras do mesmo DTU. Representação tridimensional do padrão de diversidade entre as 21 cepas de TcI, com base no padrão de amplificação dos motivos de **(A)** TcMUC, **(B)** MASP e **(C)** trans-sialidase por PCA. Representação tridimensional do padrão de diversidade entre os 9 isolados de TcII, com base no padrão de amplificação dos motivos de **(D)** TcMUC, **(E)** MASP e **(F)** trans-sialidase por PCA. A distância entre os pontos é proporcional a distância entre os padrões de amplificações/deleções de motivos entre as amostras.

16.7 Filogenia de genes nucleares da amostra SRR3676277 de *T. cruzi*

Devido ao padrão discordante em relação ao número e expansão de motivos apresentado pela amostra SRR3676277, uma análise filogenética baseada nos genes de cópia simples das amostras de *T. cruzi* foi realizada. Foram incluídas nas amostras as sequências de Colombiana (TcI), Arequipa (TcI), Sylvio (TcI) e Esmeraldo (TcII) para auxiliar na identificação dos DTUs e isolados de *T. rangeli* das cepas SC-58 e Choachí, como grupos externos (Figura 47). A amostra SRR3676277 foi classificada entre os clusters de TcII e TcIII, sugerindo que ela pode ser uma cepa híbrida entre estes dois DTUs, possivelmente pertencendo aos DTUs TcV ou TcVI, porém mais estudos são necessários para a correta classificação desta amostra. Desta forma, devido a não identificação do DTU de origem da amostra SRR3676277, ela foi excluída das análises de compartilhamento de motivos intra- e inter-DTUs, que serão descritas a seguir.

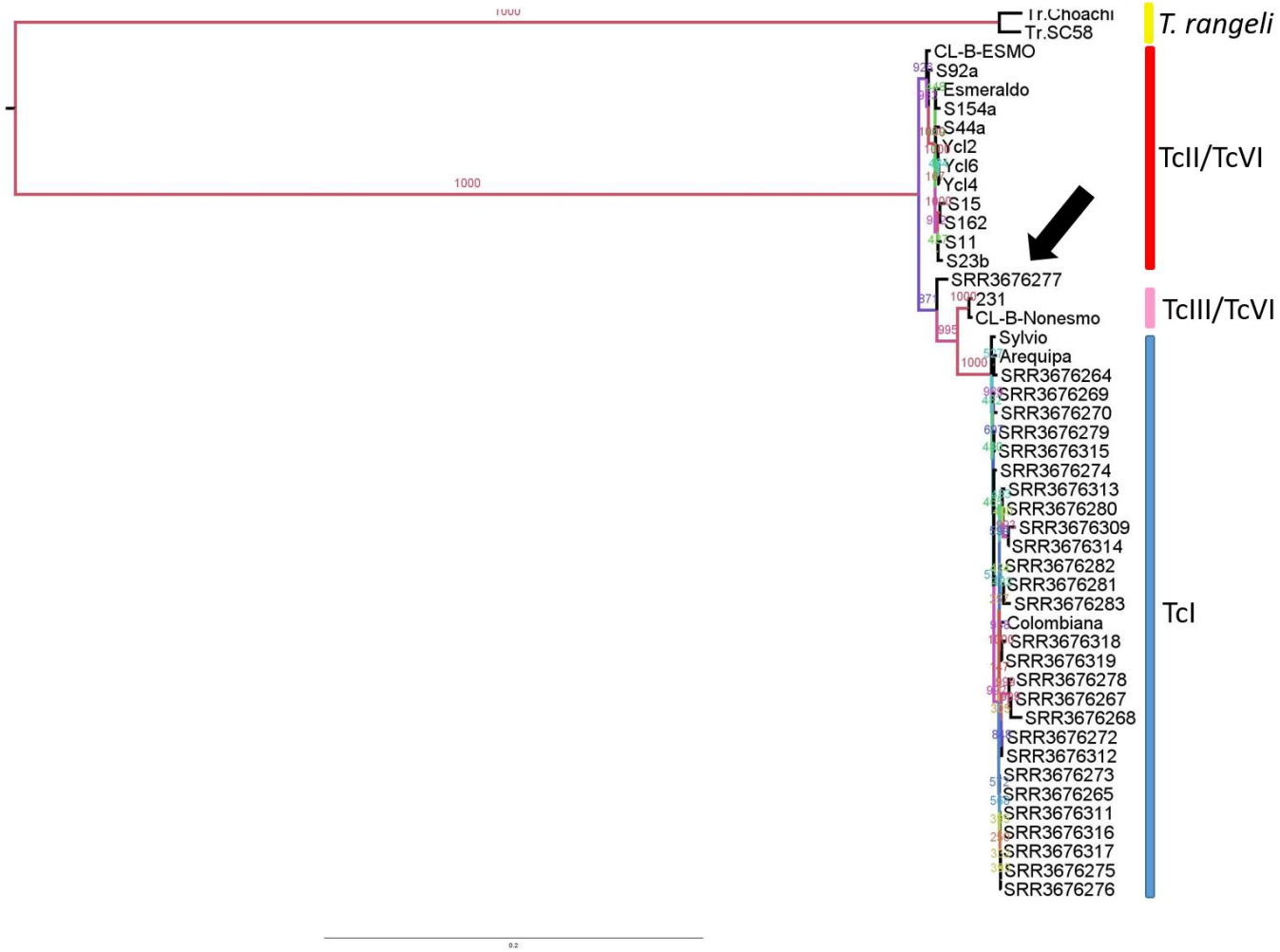


Figura 47: Análise filogenética com base em marcadores nucleares das cepas de *T. cruzi* dos 4 DTUs avaliados. Análise filogenética por máxima verossimilhança das amostras de *T. cruzi*

baseada em 760 genes nucleares de cópias simples. A análise filogenética foi realizada com 1.000 replicatas de bootstrap, onde uma escala de rosa a marrom denota, respectivamente, a valores altos e baixos de bootstrap. As cepas de *T. rangeli* SC-58 e Choachi foram utilizadas como grupo externo. As amostras de cada DTU estão assinaladas por barras coloridas, onde azul, rosa e vermelho correspondem respectivamente aos DTUs TcI, TcIII/TcVI (haplótipo Non-Esmeraldo-like), e TcII/TcVI (haplótipo Esmeraldo-like); e a barra amarela corresponde às cepas de *T. rangeli*. A seta em preto destaca a amostra SRR3676277.

16.8 Identificação de motivos específicos e compartilhados entre os DTUs de *T. cruzi*

Para avaliar o número de motivos específicos ou compartilhados por DTU, dois diagramas de Venn foram gerados para cada uma das famílias TcMUC, MASP e trans-sialidase. No primeiro, foram contabilizados apenas os motivos presentes em todas as cepas analisadas de um determinado DTU, enquanto no segundo foram contabilizados os motivos presentes em ao menos um representante de cada DTU. Como mencionado anteriormente, como a filogenia da amostra SRR3676277 ainda não foi completamente estabelecida, ela foi excluída da análise de compartilhamento de motivos entre DTUs, que foi então realizada com apenas 33 cepas.

Ao analisar os motivos de TcMUC presentes em todas as cepas de cada DTU (Figura 48 A), vimos que dos 14.934 motivos desta família, 3.742 (25%) estavam presentes nas 33 cepas de *T. cruzi* avaliadas. Nesta análise foram também identificados motivos exclusivos de cada DTU. Foram identificados 108 motivos exclusivos para TcI, 20 para TcII, 98 para TcIII e 212 para TcVI. É importante ressaltar que apesar de terem sido encontrados 98 motivos em TcIII, como apenas uma cepa (231) foi avaliada, estes motivos não representam TcIII e sim motivos exclusivos da cepa 231 entra as amostras avaliadas. Foram encontrados mais motivos compartilhados entre todas as cepas de TcVI e TcII (684) do que entre TcVI e TcIII (612), o que pode ser devido e uma maior retenção de sequências de mucinas do ancestral Esmo-Like do que do Non-Esmeraldo-like. Este resultado não é um viés de ter sido utilizada somente uma amostra de TcIII, visto que a inclusão de mais amostras deste DTU só poderia no máximo manter o número de motivos compartilhados com TcVI em 612, onde o esperado seria uma redução deste número por variabilidade intra-DTU TcIII.

Por outro lado, ao analisarmos os motivos de TcMUC presentes em ao menos uma cepa de cada DTU (Figura 48 B), vimos que dos 14.934 motivos de TcMUC 7.411 (49,6%) estão presentes em ao menos uma cepa de todos os DTUs avaliados. Este resultado sugere que pode haver variação na abundância de motivos conservados que não foram recuperados em algumas

cepas com cobertura mínima 30% da cobertura diploide do genoma, ou a hipótese menos provável de que existam motivos compartilhados entre cepas de diferentes DTUs, mas não dentro do mesmo DTU.

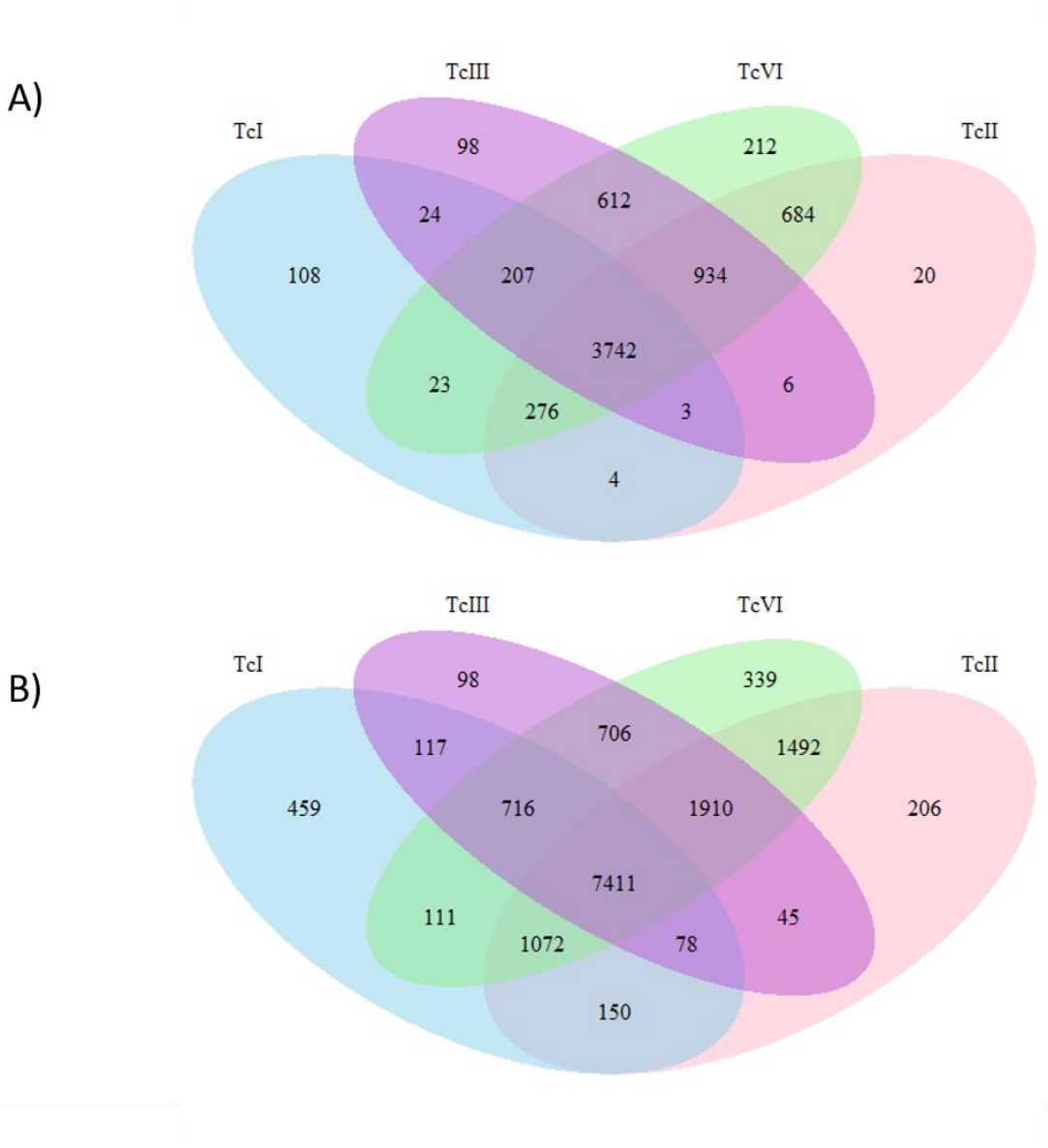


Figura 48: Diagrama de Venn representativo do compartilhamento de motivos da família TcMUC de *T. cruzi* entre diferentes DTUs. (A) Número de motivos presentes em todas as cepas de um determinado DTU. **(B)** Número de motivos presentes em ao menos uma cepa de cada DTU.

De modo semelhante, ao analisar os motivos de MASP presentes em todas as cepas de cada DTU (Figura 49 A), vimos que dos 40.793 motivos desta família, 14.852 (36,4%) estavam presentes nas 33 cepas de *T. cruzi* avaliadas. Nesta análise foram também identificados motivos

exclusivos de cada DTU. Foram identificados 183 motivos exclusivos para TcI, 30 para TcII, 239 para TcIII e 156 para TcVI. Diferentemente do que foi encontrado para as TcMUCs, foram encontrados mais motivos de MASP compartilhados entre os DTUs TcIII e TcVI (1.232) do que entre TcII e TcVI (1.016), porém isso pode caracterizar viés pelo uso de apenas uma cepa de TcIII. Quando foram avaliados os motivos de MASP presentes em ao menos uma cepa de cada DTU (Figura 49 B), 25.846 (63,35%) motivos foram encontrados em todos os DTUs. Desta vez, um número ligeiramente maior de motivos compartilhados entre TcVI e TcII (1.676) foram encontrados, quando comparado com aqueles compartilhados entre TcVI e TcIII (1.427). Porém, devido à presença de apenas uma cepa de TcIII, as contagens de clusters compartilhados entre este DTU e TcVI podem estar subestimadas.

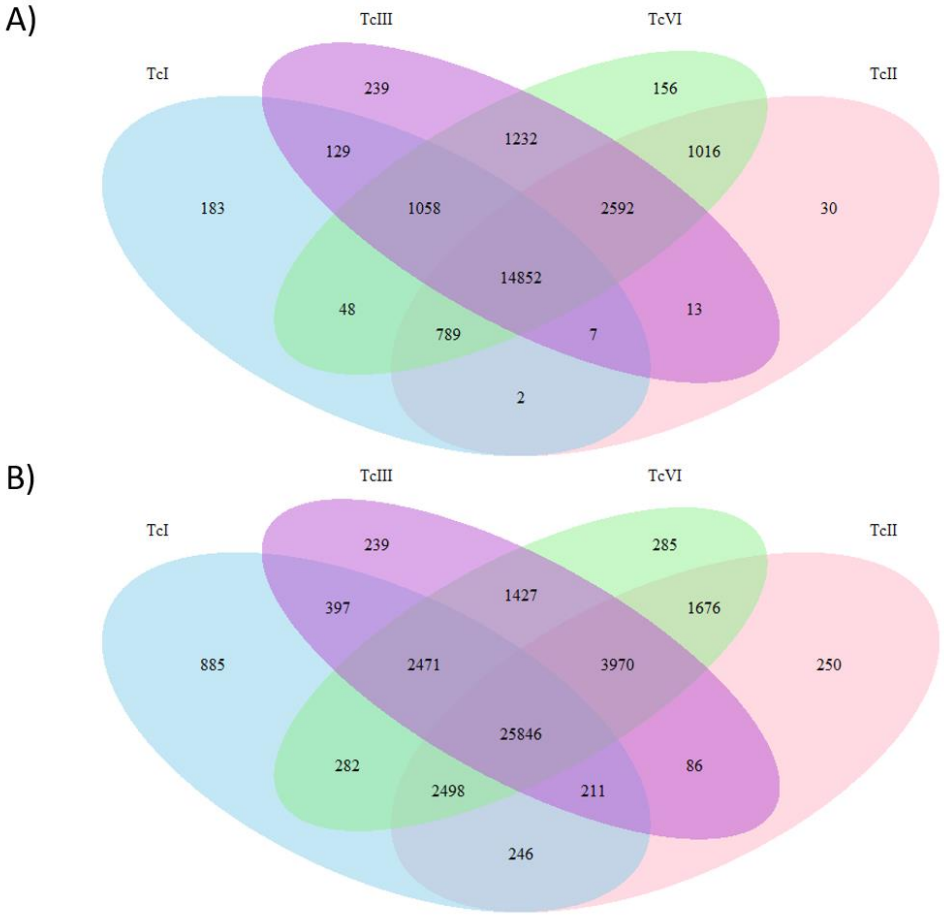


Figura 49: Diagrama de Venn representativo do compartilhamento de motivos da família MASP de *T. cruzi* entre diferentes DTUs. (A) Número de motivos presentes em todas as cepas de um determinado DTU. **(B)** Número de motivos presentes em ao menos uma cepa de cada DTU.

Finalmente, quando os motivos de trans-sialidases presentes em todas as cepas de cada DTU foram avaliados (Figura 50 A), foi observado que dos 53.056 motivos únicos, 29.033 (54,7%) estavam presentes em todas as cepas avaliadas, tornando esta família a mais conservada entre as três avaliadas. Além disso, foram identificados 504 motivos exclusivos a TcI, 33 a TcII, 176 a TcIII e 154 a TcVI. De modo semelhante ao observado para as mucinas, foram encontrados mais motivos de trans-sialidases compartilhados entre TcVI e TcII (1233) do que entre TcVI e TcIII (880). Quando foram analisados os motivos presentes em ao menos uma cepa de cada DTU (Figura 50 B), 38.318 (72.2%) motivos foram encontrados em todos os DTUs.

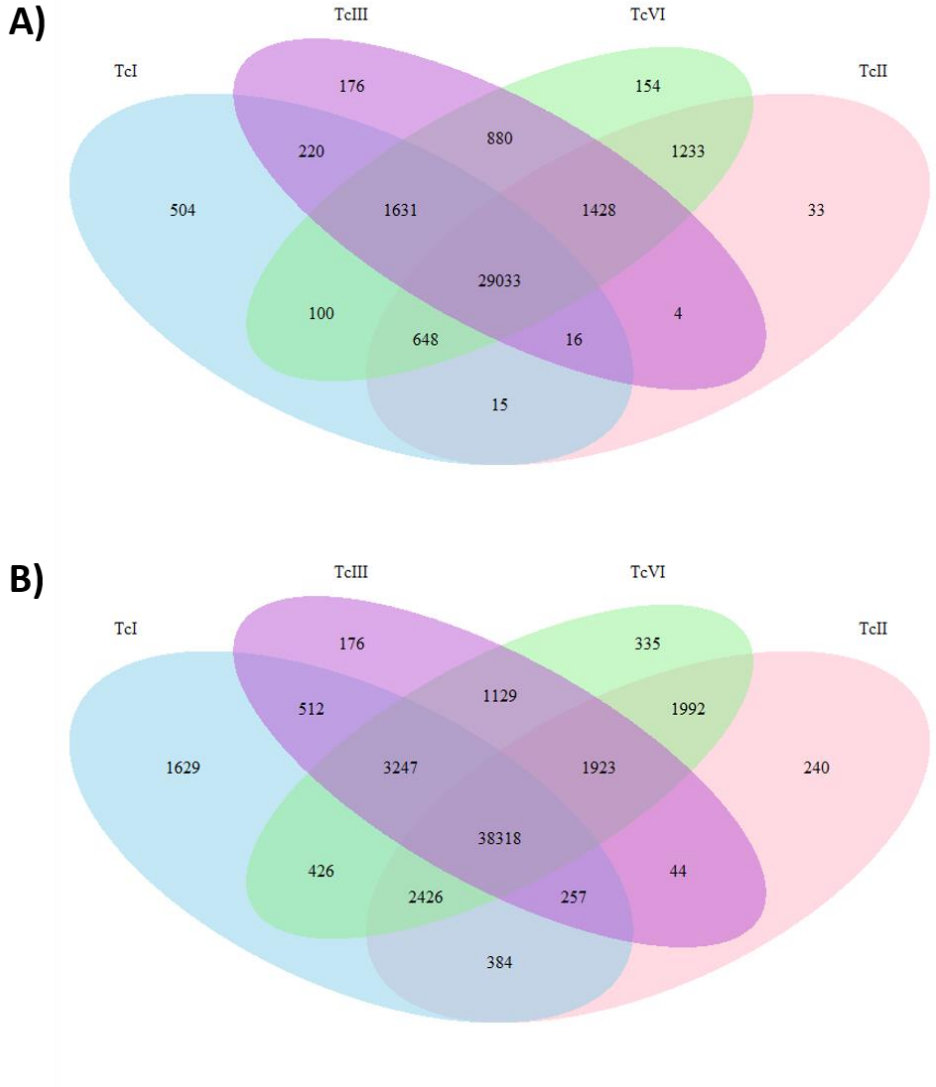


Figura 50: Diagrama de Venn representativo do compartilhamento de motivos da família trans-sialidase de *T. cruzi* entre diferentes DTUs. (A) Número de motivos presentes em todas as cepas de um determinado DTU. **(B)** Número de motivos presentes em ao menos uma cepa de cada DTU.

17. Discussão:

O estudo comparativo em escala genômica da variabilidade de famílias multigênicas entre DTUs de *T. cruzi* é dificultado pelo grande número de genes e conteúdo altamente repetitivo destas famílias. Regiões repetitivas complicam o mapeamento de *reads* pequenas (~100 bases) a um gene específico visto que muitas vezes uma determinada *read* pode mapear com confiança em dois ou mais genes reduzindo a qualidade de mapeamento. Por este motivo, é possível determinar que a *read* mapeou em um dos membros de uma família multigênica com confiança, mas não em qual membro ela mapeou. De modo semelhante, a montagem de novo de *reads* destas regiões resulta em "bolhas", regiões que podem apresentar mais de uma solução possível para a montagem de um gene em regiões repetitivas, onde os programas montadores *de novo* como o Velvet (ZERBINO; BIRNEY, 2008) e o Celera (MILLER et al., 2008) não conseguem resolver qual é a sequência correta, levando a quebra da sequência ou ao colapso de mais de uma variante em uma sequência única.

Desta forma, o presente trabalho propôs o estudo de variabilidade de famílias multigênicas de *T. cruzi* por metodologia independente de mapeamento membro específico assim como sem a necessidade de montagem *de novo* das *reads*, eliminando os vieses mencionados acima. Esta metodologia baseia-se na recuperação de *reads* que mapearam em qualquer membro de uma família multigênica específica, seguido pela determinação de todos os kmers de 30 nucleotídeos gerados a partir destas *reads* e remoção de redundância por clusterização com base em similaridade de sequência. Esta metodologia permitiu a identificação de grandes diferenças na abundância de motivos de famílias multigênicas entre diferentes DTUs do parasito *T. cruzi*.

17.1 Mapeamento genômico e estudo de *reads* não mapeadas.

Para otimizar a recuperação de *reads* correspondentes às famílias multigênicas MASP, TcMUC e Trans-sialidase de *T. cruzi*, um arquivo contendo as sequências dos 41 cromossomos hipotéticos dos haplótipos Esmeraldo-like e Non-Esmeraldo-like, juntamente como os contigs não utilizados na montagem destes cromossomos foi utilizado como referência no mapeamento de todas as bibliotecas de *reads*. Apesar de já existirem outros genomas de *T cruzi* parcialmente montados como das cepas Sylvio (FRANZÉN et al., 2011, 2012) e Dm28c (GRISARD et al., 2014), a anotação incompleta e ausência de curadoria manual destes genomas poderia comprometer

a qualidade das análises devido a anotação incorreta ou não predição de genes destas famílias multigênicas, o que levaria a uma subestimativa da contagem total de genes/motivos.

Um importante parâmetro avaliado para a seleção das cepas a serem utilizadas nesse trabalho foi a porcentagem de mapeamento das *reads*, ou seja, a proporção de *reads* que mapeou na referência utilizada em relação a proporção de *reads* não mapeadas. A seleção de um ponto de corte mínimo de 75% de mapeamento visa minimizar a utilização de bibliotecas de *reads* extremamente divergentes a CL Brener (Figura 33), e também excluir amostras que estejam possivelmente contaminadas com *reads* de outros organismos. Algumas das amostras disponíveis no SRA e não utilizadas neste trabalho, apresentaram grande número de *reads* derivados de sequências de bactérias, o que poderia comprometer as estimativas de abundância de motivos de famílias multigênicas de *T. cruzi* (dados não mostrados).

Para avaliar o conteúdo gênico que estava sendo perdido por não mapear na referência de CL Brener, as *reads* não mapeadas foram montadas de novo e a sua identidade foi avaliada. Aproximadamente metade das reads não mapeadas foram utilizadas na montagem de novo de contigs, onde em sua maioria a identidade pode ser determinada. Entre estes contigs, a maioria era composta por sequencias de *H. sapiens* e *G. gallus*, sugerindo leve contaminação das *reads* com DNA exógeno ou mistura de *reads* de outros organismos no centro de sequenciamento. Como a sequência do DNA mitocondrial (maxicírculo e minicírculo) não foi adicionada na referência de CL Brener utilizada no mapeamento das reads, 199 dos 204 dos contigs que tiveram melhor match com sequências de kinetoplastídeos corresponderam a sequências do kDNA. Os outros 5 contigs corresponderam a matches parciais com genes de MASP de CL Brener (Figura 34). Desta forma, dos 1.201 contigs gerados com as reads não mapeadas, apenas cinco corresponderam a matches com famílias multigênicas, confirmando que não foram perdidas grande quantidade de *reads* destas famílias no mapeamento.

17.2 Validação da análise de kmers.

Após a seleção das *reads* que mapearam em cada família multigênica, o próximo passo foi estimar a contagem diferencial dos motivos presentes nestas *reads* como estimativa da variabilidade presente em cada uma das famílias MASP, TcMUC e Trans-sialidase de modo semelhante ao que foi realizado na análise de MEMEs para a família MASP em 2005 (EL-SAYED, 2005a). O programa MEME (BAILEY et al., 2006), que identifica motivos conservadas e com alto número de cópias, foi utilizado para identificar 20 motivos de MASP em CL Brener com base na sequência proteica codificada pelos seus 771 genes completos por EL-Sayed e colaboradores em

2005. Este programa foi desenvolvido para trabalhar com um pequeno número de sequências fasta no arquivo de entrada, não sendo adequando para a utilização de um *input* como *reads* de NGS que apresentam milhares a milhões de sequências. Por este motivo, nós desenvolvemos uma nova metodologia para identificar motivos presentes em bibliotecas de *reads*, através da contagem diferencial e clusterização de todos os kmers de 30 nucleotídeos presentes no *dataset*. A escolha do tamanho de 30 nucleotídeos corresponde a pouco menos do tamanho médio dos menores MEMEs de MASP (EL-SAYED, 2005a) que apresentaram por volta de 15 aminoácidos (45 nucleotídeos).

As *reads* foram previamente submetidas a análise e trimagem por qualidade de Phred 30, ou seja, aceitava apenas 1 erro de sequenciamento a cada 1.000 bases. Porém, como aproximadamente 200.000 – 900.000 *reads* de 100b mapearam com cada família, eram esperados ~20.000 – 90.000 kmers únicos resultantes de erros de sequenciamento. Para evitar este viés, foram utilizados apenas kmers com cobertura acima de 10 reads. Como as bibliotecas de *read* utilizadas apresentaram diferentes profundidades de sequenciamento (Tabela 4), a RDC de cada kmer foi normalizada pela cobertura do genoma, para permitir a comparação entre cepas. Um segundo cutoff de cobertura mínima da read de 30% da cobertura do genoma foi utilizado para permitir a inclusão de amostras com baixa cobertura onde o cutoff inicial de 10 reads seria mais estringente do que 20% da cobertura do genoma (Figura 34). Este valor de 30% foi usado para capturar kmers derivados de regiões haploides do genoma, o que ocorre com frequência em regiões que codificam proteínas de superfície. Além disso, este cutoff mais permissivo permite incorporar kmers de baixa frequência decorrente de uma variabilidade intrínseca intra-populacional.

Para avaliar se a frequência de kmers gerados poderia ser relacionado à abundância real de sequências no genoma de origem, a proporção de kmers correspondentes aos motivos MEME de MASP previamente identificados foi estimado para as bibliotecas de *reads* de TcII e TcIII (Figuras 36 e 37). Genes da família MASP apresentam regiões N- e C- terminal conservadas, que flanqueiam uma região central hipervariável (Figura 9) (BARTHOLOMEU et al., 2009). Desta forma, é esperado que motivos correspondentes às regiões conservadas flanqueadores apresentem um maior número de cópias no genoma de origem do que motivos da região variável, o que foi confirmado pela presença dos MEMEs 1 e 2 em praticamente todos os genes de MASP (Figura 36) (EL-SAYED, 2005a). De modo semelhante, uma maior cobertura dos motivos MEME 1 e MEME 2 foram obtidos a partir do alinhamento e contagem do número de matches dos kmers gerados a partir das *reads* de MASP, confirmando que a recuperação das *reads* e geração dos kmers pelo programa *Jellyfish* mantiveram a proporção real de *reads* das famílias

multigênicas de *T. cruzi*. Interessantemente, a única cepa do DTU TcIII incluída nas análises (cepa 231) apresentou alta cobertura do motivo MEME 6, o que não foi observado nas outras amostras pertencentes ao grupo TcII (Figura 36), sugerindo que a presença deste motivo em CL Brener pode ter se originado do ancestral TcIII. Apesar de ser apenas uma validação do método, estes resultados sugerem que ocorrem grandes variações na abundância dos motivos de MASP entre os DTUs TcII e TcIII. Estas variações resultariam em diferentes disponibilidades de blocos gênicos para a construção de MASPs, levando potencialmente a grande variabilidade na estrutura gênica da família entre estes DTUs.

17.3 Clusters de famílias multigênicas.

No presente trabalho, foram encontradas grandes variações no padrão de amplificação de motivos gênicos das famílias MASP TcMUC e trans-sialidase entre diferentes DTUs do parasito *T. cruzi*. Como estas famílias estão relacionadas a processos de interação parasito-hospedeiro, como invasão celular e evasão do sistema imune, estas variações podem ter consequências relevantes na virulência e adaptação do parasito (BARTHOLOMEU et al., 2009; BUSCAGLIA et al., 2006; DE PABLOS; OSUNA, 2012; DOS SANTOS et al., 2012; SECO-HIDALGO; DE PABLOS; OSUNA, 2015).

Para avaliar o grau de conservação destas duas famílias multigênicas entre os DTUs de *T. cruzi*, duas modalidades de diagrama de Venn foram utilizados. No primeiro, foram contabilizados apenas os motivos presentes em todas as cepas analisadas de um determinado DTU. Nesta metodologia, motivos identificados como compartilhados entre DTUs possuem alta confiabilidade, porém o número de motivos específicos de cada DTU pode estar subestimado por alguma de suas cepas não ter apresentado a RDC necessária para a sua identificação. Um total de 3.742 (25%) motivos de TcMUC, 14.852 (36,4%) motivos de MASP e 29.033 (54.8%) motivos de trans-sialidases estão presentes em todas as cepas avaliadas (Figuras 48 A, 49 A e 50 A). Estes resultados sugerem que durante a evolução do parasito aproximadamente 30-50% dos motivos de famílias multigênicas em *T. cruzi* se mantiveram conservados, possivelmente porque codificam regiões “essenciais” de cada família, necessárias para sua funcionalidade e sobrevivência do parasito, enquanto os outros 50-70% expandiu de forma diferente entre as cepas de *T. cruzi*. A presença de ~50% dos motivos de trans-sialidase em todas as cepas avaliadas sugere que esta família pode estar sofrendo uma menor pressão seletiva para diversificação, quando comparada às famílias MASP e TcMUC. O mapeamento destes motivos ao longo das

sequências proteicas poderá identificar regiões críticas para a estrutura/função destas três famílias multigênicas.

No segundo diagrama de Venn, foram contabilizados os motivos presentes em ao menos um representante de cada DTU. Esta avaliação foi utilizada para fornecer um universo mais completo da variabilidade de sequências em cada DTU, assim como minimizar problemas de subestimativas da abundância de motivos devido à baixa RDC em algumas cepas avaliadas (Figura 48 B, 49 B e 50 B). Desta forma a real contagem de motivos compartilhados entre os DTUs deve estar em um valor intermediário entre estes dois diagramas.

Uma grande dicotomia no padrão de amplificação/deleção de motivos entre as cepas pertencentes aos DTUs TcI e TcII pode ser observada para as três famílias MASP, TcMUC e trans-sialidase, onde motivos com alto número de cópias em um DTU usualmente apresentam baixo número de cópias no outro (Figuras 42, 43 e 44). Estes resultados estão de acordo com o baixo número de motivos compartilhados exclusivamente entre estes dois DTUs, que foi de 6 para TcMUC, 13 para MASP e 4 para trans-sialidase (Figura 48 A, 49 A e 50 A). O DTU que apresentou o maior número de motivos compartilhados com TcI foi TcIII, correspondendo a 24 motivos para TcMUC (Figura 48 A), 129 para MASP (Figura 49 A) e 220 para trans-sialidase (Figura 50 A), o que sugere uma maior proximidade evolutiva entre TcI e TcIII. Apesar destes resultados poderem estar superestimados devido à presença de apenas uma cepa de TcIII, TcI ainda apresentou um maior número de motivos compartilhados com TcIII para a família MASP quando foram contabilizados os motivos presentes em ao menos uma das cepas do DTU (Figura 49 A). Por outro lado, TcI apresentou um maior número de motivos compartilhados com TcII para a família TcMUC quando foram contabilizados os motivos presentes em ao menos uma das cepas do DTU, resultado que pode ter sido enviesado pelo grande número de amostras destes dois DTUs.

Entre as amostras do DTU TcII, o padrão de motivos de MASP, TcMUC e trans-sialidase expandidos entre os três clones de Y é extremamente semelhante, o que mostra que a metodologia é capaz de agrupar amostras de origem similar (Figuras 42 43 e 44). As amostras S15 e S162a, originadas da região central de Minas Gerais, apresentaram um padrão de amplificação de motivos semelhante, corroborando a proximidade evolutiva entre estas cepas, como visto no Capítulo 2. Como esperado, TcII apresentou um maior número de motivos conservados com TcVI com base em todas as análises, reforçando que corresponde a um dos DTUs ancestrais, distantes de TcI e TcIII (BURGOS et al., 2013; DE FREITAS et al., 2006; WESTENBERGER et al., 2005).

As amostras de TcVI apresentam um mosaico de motivos compartilhados principalmente com cepas de TcII e a cepa de TcIII, o que está de acordo estudos prévios que mostram que a hibridização entre representantes de TcII e TcIII originou cepas do DTU TcVI (BARNABÉ; BRISSE; TIBAYRENC, 2000; BRISSE et al., 2003; EL-SAYED, 2005a; GAUNT et al., 2003; MACHADO; AYALA, 2001)(Figuras 42, 43 e 44). Por outro lado, TcVI apresenta um menor número de motivos compartilhados com TcI para ambas famílias, mesmo quando foram avaliados motivos presentes em ao menos uma cepa de cada DTU. De modo similar, uma maior variabilidade de kmers e motivos de todas as famílias foram encontradas em cepas do DTU TcVI quando comparadas a cepas dos DTUs TcI, TcII e TcIII (Figura 40 e Figura 41). Estes resultados estão de acordo com dados prévios da literatura, que mostraram que a expansão de membros de famílias multigênicas em CL Brener (TcVI) correspondem a grande parte da diferença de 5.9 MB entre o seu genoma e aquele da cepa Sylvio X10 (TcI) (FRANZÉN et al., 2011). Porém, este aumento da variabilidade de motivos em TcVI pode ser uma consequência do processo de seleção das reads através do mapeamento no genoma de referência de CL Brener (TcVI), visto que motivos específicos de outros DTUs como TcII, TcIII e especialmente TcI (que não faz parte das DTUs que deram origem a TcVI) podem não estar representados nas nossas análises.

A amostra SRR3676277 apresentou um padrão diferente de amplificação de motivos de famílias multigênicas (Figura 42, 43 e 44), assim como maior variabilidade de sequência (Figura 41), quando comparada a outras cepas de TcI. Nossos resultados sugerem que esta cepa não pertence ao DTU TcI, podendo ter sido originada de uma hibridização entre cepas de diferentes DTUs. Por outro lado, a existência de cepas de TcI com um padrão genômico aberrante já foi descrita na literatura (SOUZA et al., 2011). A cepa Tc1161 (José-IMT) isolada de um paciente com estágio grave de manifestações cardíacas apresentou um aumento no tamanho de suas bandas cromossômicas quando comparada a outras cepas de TcI, apresentando um padrão de bandeamento semelhante a cepas de TcII e TcIII (SOUZA et al., 2011).

A análise de componente principal da abundância dos motivos das famílias MASP, TcMUC e trans-sialidase separou claramente os três DTUs de *T. cruzi* em clusters distintos, onde as duas amostras de TcVI foram alocadas em pontos intermediários entre as amostras de TcII e a amostra de TcIII (Figura 45). Em ambos os gráficos a maior dispersão das amostras de TcI está de acordo com o padrão variável deste DTU observado nos heatmaps para TcMUC (Figura 42), MASP (Figura 43) e trans-sialidase (Figura 44), sugerindo uma maior distância evolutiva entre cepas deste DTU quando comparado com TcII, como visto antes por Cerqueira e colaboradores em 2008 (CERQUEIRA et al., 2008). Este resultado sugere que possa ser necessária uma nova revisão do DTU TcI, e possivelmente a sua subdivisão, como proposto na literatura (RAMÍREZ;

HERNÁNDEZ, 2017). Porém, a maior variabilidade neste DTU pode ser uma consequência da maior distância geográfica das amostras de TcI, que foram isoladas da Colômbia, Equador, Texas, Venezuela e Panamá, enquanto todas as cepas de TcII foram isoladas de dois estados do Brasil, Minas Gerais (amostras de campo) e São Paulo (cepa Y e seus clones) (capítulo 2). De fato, quando foram analisadas apenas as mostras de TcI, uma forte influência de componente geográfico pode ser observada (Figura 46 A, B e C). Este componente geográfico também foi observado entre cepas de TcII, onde as amostras da região central de Minas Gerais, S11 e S162 sempre agruparam e ficaram distantes dos outros isolados do Estado (Figura 46 D, E e F). As amostras S11 e S23b apresentaram um padrão divergente para as famílias TcMUC e trans-sialidase, mas um padrão extremamente semelhante para a família MASP, o que pode sugerir diferentes pressões seletivas na diversificação destas famílias. Esta separação entre os DTUs de *T. cruzi*, juntamente com a observação do caráter geográfico na separação das amostras sugerem que esta metodologia também pode ser utilizada para auxiliar em análises filogenéticas do parasito.

A maior distância entre as cepas de *T. cruzi* com base no padrão de amplificação de motivos das famílias MASP e trans-sialidase quando comparado a família TcMUC está de acordo com a maior variabilidade intragenômica destas famílias encontradas em CL Brener (Figura 45 e 46), e pode também ser uma consequência do maior número de genes presentes nestas famílias (EL-SAYED, 2005a).

17.4 Outras aplicabilidades da metodologia.

Além de permitir a avaliação do padrão de variabilidade entre as famílias multigênicas, a metodologia descrita no presente capítulo também pode auxiliar na descoberta de alvos para o diagnóstico sorológico e molecular da doença de Chagas, assim como identificar potenciais candidatos vacinais. Em primeiro lugar, peptídeos baseados na sequência proteica codificada pelos motivos conservados e presentes em todas as 34 cepas de *T. cruzi* avaliadas podem ser sintetizados e utilizados para o diagnóstico específico da doença de Chagas. Uma avaliação global da predição de epítopos de célula B em todo o proteoma predito da cepa CL Brener de *T. cruzi* revelou que MASPs correspondem a mais de 500 entre as 1000 proteínas com maior *score* de predição de epítopos de célula B (Almeida et al., em preparação). Desta forma, a combinação de conservação de sequência entre DTUs de *T. cruzi*, com os resultados de predições de epítopos de célula B e a ausência desta família em parasitos do gênero *Leishmania* (um dos principais responsáveis por reações cruzadas em testes sorológicos para o diagnóstico da doença de Chagas) pode levar a identificação de alvos específicos e sensíveis para o diagnóstico desta

parasitose. Como estas famílias multigênicas estão envolvidas em interação parasito-hospedeiro, a identificação de motivos presentes em todas as cepas de *T. cruzi* com alto número de cópias e que apresente grande quantidade de epítopos de célula $TcD8^+$ pode também constituir um promissor candidato vacinal. Diversos candidatos vacinais baseados em sequencias de trans-sialidase já foram propostos para esta doença e apresentaram resultados promissores (BONTEMPI et al., 2015; HOFT et al., 2007; MACHADO et al., 2006b). Porém, uma constante crítica a estes trabalhos consiste na possibilidade desta reposta protetora ser cepa ou DTU específica, visto que os candidatos utilizados podem não estar presentes em outras linhagens do parasito. A utilização de nossa metodologia para a seleção de alvos vacinais permite a previa identificação da conservação destes motivos entre diferentes cepas do parasito, removendo este viés. Finalmente, motivos restritos a um DTU e que apresentem um grande número de cópias podem ser utilizados como alvos sensíveis para a tipagem molecular por PCR, otimizando a correta identificação das linhagens do parasito, e removendo a necessidade de uma reação de digestão enzimática, como nos métodos mais utilizados atualmente (BURGOS et al., 2007; DE FREITAS et al., 2006; SOUTO et al., 1996).

17.5 Conclusões do capítulo 3.

-A metodologia proposta neste capítulo permite a avaliação de variação no número de cópias entre regiões repetitivas utilizando reads curtas de NGS. Esta metodologia é independente de montagem *de novo* e também de mapeamento membro-específico, removendo os principais problemas associados ao mapeamento de reads NGS em regiões repetitivas do genoma.

-O cutoff de similaridade mínimo de 75% entre o Kmer e o centroide para a sua clusterização em um motivo apresentou os melhores resultados da condensação da informação, sem levar ao agrupamento errôneo de kmers.

-O padrão de amplificação de motivos das famílias multigênicas MASP, TcMUC e Trans-sialidase de *T. cruzi* varia entre os DTUs. Esta diferença é especialmente observada entre os DTUs TcI e TcII, onde motivos expandidos em um DTU geralmente estão retraídos no outro.

-O padrão de expansão de motivos pode ser utilizado para separar cepas de diferentes DTUs de *T. cruzi* com base em análises de PCA.

-Aproximadamente 25, 35 e 55% dos motivos, respectivamente, das famílias TcMUC, MASP e Trans-sialidase estão presentes em todas as cepas dos DTUs TcI, TcII, TcIII e TcVI avaliadas, constituindo-se em assinaturas destas famílias multigênicas.

-A maior conservação de motivos de trans-sialidase em todas as cepas avaliadas sugere que esta família sofreu uma menor pressão para diversificação quando comparadas as famílias MASP e TcMUC.

18 Considerações finais

Durante a última década, grandes avanços na determinação de relevância biológica de variações no número de cópias gênicas e cromossômicas em tripanossomatídeos foram alcançados, contribuindo para um melhor entendimento dos mecanismos associados à evolução do parasitismo neste grupo de organismos. Estas análises foram impulsionadas pelo desenvolvimento de técnicas de sequenciamento genômico de nova geração (NGS), que permitiram sequenciamentos genômicos em larga escala, com alta cobertura e a um preço acessível (METZKER, 2010). Análises baseadas em *reads* genômicas por NGS permitem a simultânea avaliação da filogenia, presença de aneuploidias, variações no genótipo e frequências alélicas em uma mesma amostra, fornecendo uma visão global da variabilidade genômica do parasito (DUJARDIN et al., 2014; REIS-CUNHA et al., 2015; ROGERS et al., 2011). No presente trabalho, estas reads foram utilizadas para estimar a variabilidade genômica apresentada pelo parasito *T. cruzi*, avaliando tanto a região conservada do genoma para a identificação de aneuploidias, quanto a região variável composta por famílias multigênicas para a identificação de expansão diferenciais de motivos gênicos.

Durante sua evolução, *T. cruzi* apresentou uma massiva expansão de famílias multigênicas, relacionadas a processos de invasão celular e evasão do sistema imune, como TcMUCs, MASPs e Trans-sialidases. A ausência de ortólogos das famílias TcMUC e MASP em *T. brucei e Leishmania*, assim como a substancial redução no número de cópias destas três famílias no parasito não patogênico para mamíferos *T. rangeli*, bem como das trans-sialidases em *T. brucei* reforça a importância destas expansões para o estabelecimento de uma infecção produtiva no hospedeiro mamífero pelo parasito *T. cruzi* (EL-SAYED, 2005b; STOCO et al., 2014). A identificação de grandes variações no número de motivos expandidos ou retraídos entre os DTUs de *T. cruzi* reforça a hipótese que estas famílias estão em constante mudança. Porém o achado de que aproximadamente 30-50% dos motivos são compartilhados entre todas as cepas denota que alguns membros (ou motivos) destas famílias são mantidos e possivelmente necessários para manutenção da estrutura e função destas famílias e sobrevivência do parasito.

A metodologia desenvolvida neste trabalho para a comparação dos motivos de famílias multigênicas de *T. cruzi* pode também ser utilizada para a avaliação da variabilidade de regiões repetitivas de qualquer genoma, desde que uma referência para a seleção de *reads* derivadas destas repetições esteja disponível. Esta metodologia pode também ser empregada para a identificação de motivos exclusivos a um grupo de organismos, e está sendo atualmente utilizada para a identificação de sequências específicas de bactérias resistentes a antibióticos.

Além de expansão diferencial de motivos de famílias multigênicas, eventos de CCNV aparentam ser bem tolerados pelo parasito *T. cruzi*, onde o seu padrão também varia entre DTUs (capitulo I), dentro de DTUs (capítulo I e II), mas parece ser constante dentro de uma população de parasitos (capitulo II). Porém, para confirmar a ausência de variação no padrão de CCNV intra-populacional obtido com os clones de Y, análises de FISH para estimar o número de cópias cromossômicas do parasito como realizado para *Leishmania* devem ser realizados (STERKERS et al., 2011, 2014). Interessantemente, o cromossomo 31 foi o único expandido em todas as cepas de *T. cruzi* avaliadas. Este cromossomo está enriquecido com genes relacionados a glicosilação e síntese de glicanos, reforçando a importância deste processo biológico para a sobrevivência do parasito. A metodologia de SCoPE proposta para estimar as aneuploidias no presente trabalho reduz os vieses da não confiabilidade de mapeamento em regiões repetitivas e da presença de grandes regiões de gaps nas montagens das sequências cromossômicas da cepa CL Brener de *T. cruzi*, atualmente a melhor referência genômica do parasito. A discordância entre a filogenia e o padrão de CCNV sugere que eventos de ganho e perda de cromossomos são frequentes na evolução do parasito e são governados por pressões seletivas diferentes daquelas que definem a filogenia do grupo. CCNV pode ser explorado pelo parasito para uma rápida seleção de fenótipos favoráveis em condições ambientais diversas. Estes resultados estão de acordo com os dados obtidos por Dowins e colaboradores, que demonstraram que o padrão de aneuploidias varia entre 17 cepas de *L. donovani* isoladas de pacientes na Índia, onde todos os isolados apresentaram um cariótipo diferente, enquanto apresentavam apenas 0,011% de variação nucleotídica, sugerindo que eventos de CCNV ocorrem também frequentemente durante a evolução deste parasito (DOWNING et al., 2011). De modo semelhante, mudanças no padrão de aneuploidias ocorridas durante a transição entre os hospedeiros mamíferos e inseto em *Leishmania* sugerem que diferentes combinações de CCNV podem auxiliar a infecção de distintos hospedeiros (DUMETZ et al., 2017).

A grande mistura populacional e o padrão altamente variável de CCNV entre amostras de campo de *T. cruzi* isoladas de regiões geográficas próximas sugerem que um ciclo parasexual pode ser um dos principais mecanismos por trás de trocas gênicas e aneuploidias neste parasito (MESSENGER; MILES, 2015; MESSENGER; MILES; BERN, 2015). Porém, falhas na segregação ou replicação estocástica de cromossomos, como proposto para *Leishmania* (STERKERS et al., 2011, 2014), podem também estar envolvidos na geração de CCNV em *T. cruzi*.

Apesar de ser proposto para *T. cruzi* e *Leishmania*, não existem evidências de aneuploidias em *T. brucei* (WEIR et al., 2016). O genoma de *T. brucei* é dividido em 11 cromossomos com megabases de tamanho, que variam de 1 a 6 Mbp, enquanto os genomas de *T. cruzi* e *Leishmania* estão distribuídos em ~ 34-47 cromossomos, com tamanhos variando entre

0.3-3Mbp (BERRIMAN; GHEDIN; HERTZ-FOWLER, 2005; BRITTO et al., 1998; EL-SAYED, 2005a; PEACOCK et al., 2007; WEATHERLY; BOEHLKE; TARLETON, 2009; WINCKER et al., 1996). Por este motivo, mudanças no número de cópias cromossômicas em *T. cruzi* e *Leishmania* alteram a dosagem de um set restrito de genes, reduzindo as consequências deletérias de alterações gênicas em larga escala (MANNAERT et al., 2012). Isto sugere que aneuploidias seriam melhores suportadas em organismos que apresentam seu genoma dividido em um grande número de pequenos cromossomos. A avaliação da ocorrência e padrão de CCNV em outros tripanossomatídeos basais como *T. rangeli*, *T. c. marinkelei*, *T. livingstonei* e *T. lewisi* poderiam esclarecer a origem das ocorrências de CCNV durante a evolução dos tripanossomatídeos, assim como a sua implicação para a biologia destes parasitos.

Uma das limitações de nossa abordagem de predição de CCNV é a inviabilidade de investigar a estrutura cariótipica de cada cepa, sendo que a análise se limitou a comparar as diferenças entre as cepas de *T. cruzi* baseadas nos cromossomos preditos de CL Brener, visto que não existem ainda outras referências cromossômicas de *T. cruzi* publicadas na literatura. Devido ao extenso conteúdo repetitivo do genoma do parasito, o sequenciamento genômico de moléculas únicas que produz *reads* longas é necessário não só para preencher os gaps nos cromossomos de CL Brener e corrigir problemas da montagem atual, mas também para gerar sequências cromossômicas confiáveis dos outros 5 DTUs de *T. cruzi*, permitindo a resolução de seus cariótipos e melhores predições de CCNV. Nós estamos atualmente sequenciando e realizando a montagem de uma nova versão dos cromossomos de CL Brener, baseado em *reads* longas de terceira geração. Genomas de referência de outros DTUs de *T. cruzi* também estão sendo produzidas em outros institutos de pesquisa por metodologias semelhantes. Desta forma, a avaliação de CCNV entre as cepas de *T. cruzi* poderá ser revisitada tão logo estes genomas mais completos estejam disponíveis. Estas novas referências também permitirão a comparação direta do conteúdo repetitivo das famílias multigênicas do parasito, visto que a estrutura gênica e os clusters que contém estas famílias poderão ser melhor resolvidos.

19 Perspectivas

O presente trabalho permitiu uma melhor compreensão da variabilidade genômica do parasito *T. cruzi*. Porém, várias outras importantes perguntas ainda precisam ser respondidas. Portanto, as principais perspectivas deste trabalho são:

1- Realizar a montagem e anotação do genoma de CL Brener, baseado em uma combinação de *reads* longas de terceira geração, reads curtas de Illumina de alta qualidade e reads de Sanger de BACs;
2- Avaliar o padrão de CCNV entre cepas de *T. cruzi* com a nova referência de CL Brener e de SylvioX10 produzida e comparar com os resultados obtidos no presente trabalho;
3- Estimar o padrão de CCNV intra-populacional de *T. cruzi* com base em análises de FISH;
4- Avaliar variações de expressão gênica em cromossomos trissômicos ou tetrassômicos quando comparados a cromossomos dissômicos, para avaliar se aneuploidias resultariam em impactos na expressão gênica global de cromossomos em *T. cruzi*, como visto para *Leishmania*.
5- Avaliar a taxa de ocorrência de CCNV em *T. cruzi* quando submetido a estresse replicativo ou outro tipo de estresse.
6- Avaliar a ocorrência/expansão de motivos das famílias multigênicas em tripanossomatídeos basais como *T. rangeli*, *T. c. marinkelei*, *T. livingstonei* e *T. lewisi*, permitindo o estudo da evolução das famílias multigênicas nestes parasitos;
7- Encontrar motivos específicos de cada DTU de *T. cruzi*, que poderão ser usados como alvo para o diagnóstico molecular destes subgrupos em apenas um experimento de PCR multiplex.
8- Busca de motivos compartilhados entre as diferentes DTUs do parasito, que poderão ser candidatos a alvos vacinais.

20 Lista de arquivos em anexo:

Anexo 1: Artigo publicado:

Reis-Cunha, J. L.; Rodrigues-Luiz, G. F.; Valdivia, H. O.; Baptista, R. P.; Mendes, T. A. O.; de Morais, G. L.; Guedes, R.; Macedo, A. M.; Bern, C.; Gilman, R. H.; et al. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains. *BMC Genomics* **2015**, *16* (1), 499.

21 Outras publicações e patentes obtidas durante o doutorado:

Artigos:

Cerqueira, G. C.; Earl, A. M.; Ernst, C. M.; Grad, Y. H.; Dekker, J. P.; Feldgarden, M.; Chapman, S. B.; **Reis-Cunha, J. L**.; Shea, T. P.; Young, S.; et al. Multi-institute analysis of carbapenem resistance reveals remarkable diversity, unexplained mechanisms, and limited clonal outbreaks. *Proc. Natl. Acad. Sci.* **2017**, *114* (5), 1135–1140.

Valdivia, H. O.; Almeida, L. V.; Roatt, B. M.; **Reis-Cunha, J. L**.; Pereira, A. A. S.; Gontijo, C.; Fujiwara, R. T.; Reis, A. B.; Sanders, M. J.; Cotton, J. A.; et al. Comparative genomics of canine-isolated *Leishmania* (*Leishmania*) *amazonensis* from an endemic focus of visceral leishmaniasis in Governador Valadares, southeastern Brazil. *Sci. Rep.* **2017**, *7*, 40804.

Silva, L. A.; **Reis-Cunha, J. L**.; Bartholomeu, D. C.; Vitor, R. W. A. Genetic polymorphisms and phenotypic profiles of sulfadiazine-resistant and sensitive *Toxoplasma gondii* isolates obtained from newborns with congenital toxoplasmosis in minas gerais, Brazil. *PLoS One* **2017**, *12* (1), 1–14.

Cardoso, M. S.*; **Reis-Cunha, J. L**.*; Bartholomeu, D. C. Evasion of the immune response by trypanosoma cruzi during acute infection. *Frontiers in Immunology*. 2016. ***Co-autoria**

Valdivia, H. O.; **Reis-Cunha, J. L.**; Rodrigues-Luiz, G. F.; Baptista, R. P.; Baldeviano, G. C.; Gerbasi, R. V; Dobson, D. E.; Pratlong, F.; Bastien, P.; Lescano, A. G.; et al. Comparative genomic analysis of *Leishmania* (*Viannia*) *peruviana* and *Leishmania* (*Viannia*) *braziliensis*. *BMC Genomics* **2015**, *16* (1), 715.

Menezes-Souza, D.; De Mendes, T. A. O.; De Gomes, M. S.; **Reis-Cunha, J. L**.; Nagem, R. A. P.; Carneiro, C. M.; Coelho, E. A. F.; Da Cunha Galvão, L. M.; Fujiwara, R. T.; Bartholomeu, D. C. Epitope mapping of the HSP83.1 protein of *Leishmania braziliensis* discloses novel targets for immunodiagnosis of tegumentary and visceral clinical forms of leishmaniasis. *Clin. Vaccine Immunol.* **2014**, *21* (7), 949–959.

Capítulo de livro:

Reis-Cunha, J. L.; Valdivia, H. V.; Bartholomeu, D. C.; Trypanosomatid Genome Organization and Ploidy. Frontiers in Parasitology, **2016**, Vol. 1, 61-103

Artigo de revisão aceito para publicação:

Reis-Cunha, J. L.; Valdivia, H. V.; Bartholomeu, D. C.; Gene and Chromosomal Copy Number Variations as na Adaptative Mechanism Towards a Parasitic Lifestyle in Trypanosomatids. Current Genomics. **2017**

Patentes:

Leão, A. C.; BARTHOLOMEU, DANIELLA CASTANHEIRA; **REIS-CUNHA, JOÃO LUÍS**; MENDES, T. A. O. ; CARDOSO, M. S. ; FUJIWARA, R. T. PROTEÍNA QUIMÉRICA, MÉTODO E KIT PARA DIAGNÓSTICO DA DOENÇA DE CHAGAS E USO. 2016, Brasil.
Patente: Privilégio de Inovação. Número do registro: BR1020160026970, título: "PROTEÍNA QUIMÉRICA, MÉTODO E KIT PARA DIAGNÓSTICO DA DOENÇA DE CHAGAS E USO", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial, Depositante (s): BARTHOLOMEU, DANIELLA CASTANHEIRA, Depósito: 05/02/2016

SILVA, A. L. T.; BUENO, L. L.; CARDOSO, M. S.; FUJIWARA, R. T.; BARTHOLOMEU, D. C.; **Reis-Cunha, J.L.**; LOBO, F. P. MÉTODO, KIT PARA DIAGNÓSTICO DE LEISHMANIOSES E USO. 2017, Brasil.
Patente: Privilégio de Inovação. Número do registro: BR10201700513, título: "MÉTODO, KIT PARA DIAGNÓSTICO DE LEISHMANIOSES E USO", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial, Depositante (s): João Luís Reis Cunha; Daniella C Bartholomeu; Ricardo Toshio Fujiwara; Francisco Pereira Lobo; Lilian L. Bueno; Mariana S. Cardoso; Ana Luiza T. Silva, Depósito: 14/03/2017

SILVA, A. L. T.; BUENO, L. L.; CARDOSO, M. S.; FUJIWARA, R. T.; Bartholomeu DC; **Reis-Cunha, J.L.** ; LOBO, F. P. . PROTEÍNA RECOMBINANTE, MÉTODO E KIT PARA TRIAGEM DE TRIPANOSSOMATÍDEOS E USO. 2017, Brasil.
Patente: Privilégio de Inovação. Número do registro: BR1020170050688, título: "PROTEÍNA RECOMBINANTE, MÉTODO E KIT PARA TRIAGEM DE TRIPANOSSOMATÍDEOS E USO", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial, Depositante (s): **João Luís Reis Cunha**; Daniella C. Bartholomeu; Ricardo Toshio Fujiwara; Francisco Pereira Lobo ;Lilian L. Bueno; Mariana S. Cardoso; Ana Luiza T. Silva, Depósito: 14/03/2017

22 Referências:

ABBEY, D. et al. High-Resolution SNP/CGH Microarrays Reveal the Accumulation of Loss of Heterozygosity in Commonly Used *Candida albicans* Strains. **G3 (Bethesda, Md.)**, v. 1, n. 7, p. 523–30, 2011.

ACKERMANN, A. A. et al. A genomic scale map of genetic diversity in *Trypanosoma cruzi*. **BMC Genomics**, v. 13, p. 1, 2012.

ACOSTA-SERRANO, A. et al. The mucin-like glycoprotein super-family of *Trypanosoma cruzi*: Structure and biological roles. **Molecular and Biochemical Parasitology**, v. 114, n. 2, p. 143–150, 2001.

ADL, S. M. et al. The revised classification of eukaryotes. **Journal of Eukaryotic Microbiology**, v. 59, n. 5, p. 429–493, 2012.

AKOPYANTS, N. S. et al. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. **Science**, v. 324, n. 5924, p. 265–8, 2009.

ALEXANDER, D. H.; NOVEMBRE, J. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. p. 1655–1664, 2009.

ALMEIDA, I. C. et al. Lytic anti-alpha-galactosyl antibodies from patients with chronic Chagas' disease recognize novel O-linked oligosaccharides on mucin-like glycosyl-phosphatidylinositol-anchored glycoproteins of *Trypanosoma cruzi*. **The Biochemical journal**, v. 304 (Pt 3, p. 793–802, 1994.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403–10, 1990.

AMAYA, M. F. et al. The high resolution structures of free and inhibitor-bound *Trypanosoma rangeli* sialidase and its comparison with *T. cruzi* trans-sialidase. **Journal of Molecular Biology**, v. 325, n. 4, p. 773–784, 2003.

ANDRADE, L. O. et al. Differential tissue tropism of *Trypanosoma cruzi* strains: An in vitro study. **Memorias do Instituto Oswaldo Cruz**, v. 105, n. 6, p. 834–837, 2010.

ANDRADE, L. O.; ANDREWS, N. W. The *Trypanosoma cruzi*-host-cell interplay: location, invasion, retention. **Nature reviews. Microbiology**, v. 3, n. 10, p. 819–823, 2005.

ANDRADE, S. G.; MAGALHÃES, J. B. Biodemes and zymodemes of *Trypanosoma cruzi* strains:

correlations with clinical data and experimental pathology. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 30, n. 1, p. 27–35, 1996.

ANGHEBEN, A. et al. Chagas disease and transfusion medicine: A perspective from non-endemic countries. **Blood Transfusion**, v. 13, n. 4, p. 540–550, 2015.

APHASIZHEVA, I.; APHASIZHEV, R. U-Insertion/Deletion mRNA-Editing Holoenzyme: Definition in Sight. **Trends in Parasitology**, v. 32, n. 2, p. 144–156, 2016.

ASLETT, M. et al. TriTrypDB: A functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research**, v. 38, n. SUPPL.1, p. 457–462, 2009.

AUGUSTO-PINTO, L. et al. Single-nucleotide polymorphisms of the *Trypanosoma cruzi* MSH2 gene support the existence of three phylogenetic lineages presenting differences in mismatch-repair efficiency. **Genetics**, v. 164, n. 1, p. 117–126, 2003.

BAILEY, T. L. et al. MEME: Discovering and analyzing DNA and protein sequence motifs. **Nucleic Acids Research**, v. 34, n. WEB. SERV. ISS., p. 369–373, 2006.

BAPTISTA, R. DE P. et al. Evidence of substantial recombination among *Trypanosoma cruzi* II strains from Minas Gerais. **Infection, Genetics and Evolution**, v. 22, p. 183–191, 2014.

BARNABE, C. et al. Putative panmixia in restricted populations of *Trypanosoma cruzi* isolated from wild triatoma infestans in Bolivia. **PLoS ONE**, v. 8, n. 11, 2013.

BARNABÉ, C. et al. *Trypanosoma cruzi* discrete typing units (DTUs): Microsatellite loci and population genetics of DTUs TcV and TcI in Bolivia and Peru. **Infection, Genetics and Evolution**, v. 11, n. 7, p. 1752–1760, 2011.

BARNABÉ, C.; BRISSE, S.; TIBAYRENC, M. Population structure and genetic typing of *Trypanosoma cruzi*, the agent of Chagas disease: a multilocus enzyme electrophoresis approach. **Parasitology**, v. 120 (Pt 5), p. 513–526, 2000.

BARRETO-DE-ALBUQUERQUE, J. et al. *Trypanosoma cruzi* Infection through the Oral Route Promotes a Severe Infection in Mice: New Disease Form from an Old Infection? **PLoS neglected tropical diseases**, v. 9, n. 6, p. e0003849, 2015.

BARTHOLOMEU, D. C. et al. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen *Trypanosoma cruzi*. **Nucleic Acids Research**, v. 37, n. 10, p. 3407–3417, 2009.

BARTHOLOMEU, D. C. et al. Unveiling the Intracellular Survival Gene Kit of Trypanosomatid

Parasites. **PLoS Pathogens**, v. 10, n. 12, 2014.

BENJAMIN, R. J. et al. *Trypanosoma cruzi* infection in North America and Spain: Evidence in support of transfusion transmission (CME). **Transfusion**, v. 52, n. 9, p. 1913–1921, 2012.

BENJAMINI, Y.; YEKUTIELI, D. The control of the false discovery rate in multiple testing under depencency. **The Annals of Statistics**, v. 29, n. 4, p. 1165–1188, 2001.

BENNETT, R. J. The parasexual lifestyle of *Candida albicans*. **Current Opinion in Microbiology**, v. 28, p. 10–17, 2015.

BERN, C. et al. *Trypanosoma cruzi* and chagas' disease in the united states. **Clinical Microbiology Reviews**, v. 24, n. 4, p. 655–681, 2011.

BERN, C.; MONTGOMERY, S. P. An estimate of the burden of Chagas disease in the United States. **Clinical infectious diseases : an official publication of the Infectious Diseases Society of America**, v. 49, n. 5, p. e52-4, 2009.

BERRIMAN, M.; GHEDIN, E.; HERTZ-FOWLER, C. The genome of the African trypanosome, *Trypanosoma brucei*. **Science**, v. 309(5733), n. 2005, p. 416–422, 2005.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BONTEMPI, I. A. et al. Efficacy of a trans-sialidase-ISCOMATRIX subunit vaccine candidate to protect against experimental Chagas disease. **Vaccine**, v. 33, n. 10, p. 1274–1283, 2015.

BRANCHE, C. et al. Comparative karyotyping as a tool for genome structure analysis of *Trypanosoma cruzi*. **Molecular and Biochemical Parasitology**, v. 147, n. 1, p. 30–38, 2006.

BRINGAUD, F. et al. The *Trypanosoma cruzi* L1Tc and NARTc non-LTR retrotransposons show relative site specificity for insertion. **Molecular Biology and Evolution**, v. 23, n. 2, p. 411–420, 2006.

BRINGAUD, F. et al. Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. **PLoS Pathogens**, v. 3, n. 9, p. 1291–1307, 2007.

BRISSE, S. et al. Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. **Infection, Genetics and Evolution**, v. 2, n. 3, p. 173–183, 2003.

BRISSE, S.; BARNABE, C.; 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.; DUJARDIN, J. C.; TIBAYRENC, M. Identification of six *Trypanosoma cruzi* lineages by sequence-characterised amplified region markers. **Molecular and Biochemical Parasitology**, v. 111, n. 1, p. 95–105, 2000.

BRITTO, C. et al. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes. **Gene**, v. 222, n. 1, p. 107–117, 1998.

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. 9–11, 2013.

BURLEIGH, B. A.; ANDREWS, N. W. The mechanisms of *Trypanosoma cruzi* invasion of Mammalian Cells. **Annual Reviews in Microbiology**, v. 49, n. 1, p. 175–200, 1995.

BUSCAGLIA, C. A et al. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. **Nature reviews. Microbiology**, v. 4, n. 3, p. 229–236, 2006.

BUTLER, C. E. et al. Trans-sialidase Stimulates Eat Me Response from Epithelial Cells. **Traffic**, v. 14, n. 7, p. 853–869, 2013.

CAETANO, L. C. et al. *Trypanosoma cruzi*: Do different sylvatic strains trigger distinct immune responses? **Experimental Parasitology**, v. 124, n. 2, p. 219–224, 2010.

CÂMARA, A. C. J. et al. Genetic analyses of *Trypanosoma cruzi* isolates from naturally infected triatomines and humans in northeastern Brazil. **Acta Tropica**, v. 115, n. 3, p. 205–211, 2010.

CAMARGO, E. Growth and differentiation in *Trypanosoma cruzi*. I. Origin of metacyclic trypanosomes in liquid media**. Rev Inst Med São Paulo**, 1964.

CASTRESANA, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. **Molecular Biology and Evolution**, v. 17, n. 4, p. 540–552, 2000.

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, 2008.

CHAGAS, C. Nova tripanozomiaze humana: estudos sobre a morfolojia e o ciclo evolutivo do *Schizotrypanum cruzi* n. gen., n. sp., ajente etiolojico de nova entidade morbida do homem. **Memórias do Instituto Oswaldo Cruz**, 1909.

CHUENKOVA, M. V.; PEREIRAPERRIN, M. *Trypanosoma cruzi*-Derived Neurotrophic Factor: Role in Neural Repair and Neuroprotection. **Journal of Neuroparasitology**, v. 1, p. 55–60, 2011.

CHUENKOVA, M. V; PEREIRAPERRIN, M. *Trypanosoma cruzi* targets Akt in host cells as an intracellular antiapoptotic strategy. **Science signaling**, v. 2, n. 97, p. ra74, 2009.

CLAYCOMB, J. M.; ORR-WEAVER, T. L. Developmental gene amplification: Insights into DNA replication and gene expression. **Trends in Genetics**, v. 21, n. 3, p. 149–162, 2005.

CLAYTON, C. The Regulation of Trypanosome Gene Expression by RNA-Binding Proteins. **PLoS Pathogens**, v. 9, n. 11, p. 9–12, 2013.

CLAYTON, C. E. Life without transcriptional control? From fly to man and back again. **EMBO Journal**, v. 21, n. 8, p. 1881–1888, 2002.

CLAYTON, C. E. Gene expression in Kinetoplastids. **Current Opinion in Microbiology**, v. 32, p. 46–51, 2016.

CLAYTON, C.; SHAPIRA, M. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. **Molecular and Biochemical Parasitology**, v. 156, n. 2, p. 93–101, 2007.

COTTONTAIL, V. M. et al. High local diversity of *Trypanosoma* in a common bat species, and implications for the biogeography and taxonomy of the *T. cruzi* clade. **PLoS ONE**, v. 9, n. 9, 2014.

COURA, J. R. et al. Emerging Chagas disease in Amazonian Brazil. **Trends in Parasitology**, v. 18, n. 4, p. 171–176, 2002.

COURA, J. R.; BORGES-PEREIRA, J. Chagas disease: What is known and what should be improved: a systemic review. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 45, n. 3, p. 286–296, 2012.

COURA, J. R.; DIAS, J. C. P. Epidemiology, control and surveillance of Chagas disease: 100 years after its discovery. **Memórias do Instituto Oswaldo Cruz**, v. 104, n. i, p. 31–40, 2009.

CUEVAS, I. C. et al. gp63 Homologues in *Trypanosoma cruzi* : Surface Antigens with Metalloprotease Activity and a Possible Role in Host Cell Infection. **Infection and Immunity**, v. 71, n. 10, p. 5739–5749, 2003.

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.

DA CÂMARA, A. C. J. et al. Homogeneity of *Trypanosoma cruzi* I, II, and III populations and the overlap of wild and domestic transmission cycles by *Triatoma brasiliensis* in northeastern Brazil. **Parasitology Research**, v. 112, n. 4, p. 1543–1550, 2013.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156–2158, 2011.

DARZENTAS, N. Circoletto: Visualizing sequence similarity with Circos. **Bioinformatics**, v. 26, n. 20, p. 2620–2621, 2010.

DC-RUBIN, S. S. C.; SCHENKMAN, S. *Trypanosoma cruzi* trans-sialidase as a multifunctional enzyme in chagas' disease. **Cellular Microbiology**, v. 14, n. 10, p. 1522–1530, 2012.

DE FREITAS, J. M. et al. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. **PLoS Pathogens**, v. 2, n. 3, p. 0226–0235, 2006.

DE LANA, M. et al. *Trypanosoma cruzi*: compared vectorial transmissibility of three major clonal genotypes by *Triatoma infestans*. **Experimental parasitology**, v. 90, n. 90, p. 20–25, 1998.

DE LEDERKREMER, R. M.; AGUSTI, R. **Chapter 7 Glycobiology of *Trypanosoma cruzi***. v. 62

DE MELO-JORGE, M.; PEREIRAPERRIN, M. The Chagas' Disease Parasite *Trypanosoma cruzi* Exploits Nerve Growth Factor Receptor TrkA to Infect Mammalian Hosts. **Cell Host and Microbe**, v. 1, n. 4, p. 251–261, 2007.

DE ORNELAS TOLEDO, M. J. et al. Chemotherapy with benznidazole and itraconazole for mice infected with different *Trypanosoma cruzi* clonal genotypes. **Antimicrobial Agents and Chemotherapy**, v. 47, n. 1, p. 223–230, 2003.

DE PABLOS, L. M.; OSUNA, A. Multigene families in *Trypanosoma cruzi* and their role in infectivity. **Infection and Immunity**, v. 80, n. 7, p. 2258–2264, 2012.

DÍAZ-BELLO, Z. et al. *Trypanosoma cruzi* genotyping supports a common source of infection in a school-related oral outbreak of acute Chagas disease in Venezuela. **Epidemiology and infection**, v. 142, n. 1, p. 156–62, 2014.

DOS SANTOS, S. L. et al. The MASP Family of *Trypanosoma cruzi*: Changes in Gene Expression and Antigenic Profile during the Acute Phase of Experimental Infection. **PLoS Neglected Tropical**

Diseases, v. 6, n. 8, 2012.

DOUBRE, H. et al. Multidrug resistance-associated protein (MRP1) is overexpressed in DNA aneuploid carcinomatous cells in non-small cell lung cancer (NSCLC). **International Journal of Cancer**, v. 113, n. 4, p. 568–574, 2005.

DOWNING, T. et al. Whole genome sequencing of multiple *Leishmania* donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. **Genome Research**, v. 21, p. 2143–2156, 2011.

DUJARDIN, J. C. et al. Mosaic aneuploidy in *Leishmania*: The perspective of whole genome sequencing. **Trends in Parasitology**, v. 30, n. 12, p. 554–555, 2014.

DUMETZ, F. et al. Modulation of Aneuploidy in *Leishmania* donovani during Adaptation to Different In Vitro and In Vivo Environments and Its Impact on Gene Expression. **mBio**, v. 8, n. 3, p. 1–14, 2017.

DUNCAN, A. W. et al. The ploidy conveyor of mature hepatocytes as a source of genetic variation. **Nature**, v. 467, n. 7316, p. 707–10, 2010.

EDGAR, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, v. 26, n. 19, p. 2460–2461, 2010.

EL-SAYED, N. M. The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease. **Science**, v. 309, n. 5733, p. 409–415, 2005a.

EL-SAYED, N. M. Comparative Genomics of Trypanosomatid Parasitic Protozoa. **Science**, v. 309, n. 5733, p. 404–409, 2005b.

EPTING, C. L.; COATES, B. M.; ENGMAN, D. M. Molecular mechanisms of host cell invasion by *Trypanosoma cruzi*. **Experimental Parasitology**, v. 126, n. 3, p. 283–291, 2010.

ERDMANN, H. et al. Sialylated ligands on pathogenic *Trypanosoma cruzi* interact with Siglec-E (sialic acid-binding Ig-like lectin-E). **Cellular Microbiology**, v. 11, n. 11, p. 1600–1611, 2009.

EWING, B. et al. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. **Genome Research**, v. 8, n. 3, p. 175–185, 1998.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error

probabilities. **Genome Research**, v. 8, n. 3, p. 186–194, 1998.

FEAGIN, J. E. RNA editing in kinetoplastid mitochondria. **Journal of Biological Chemistry**, v. 265, n. 32, p. 19373–19376, 1990.

FERNANDES, M. C.; ANDREWS, N. W. Host cell invasion by *Trypanosoma cruzi*: A unique strategy that promotes persistence. **FEMS Microbiology Reviews**, 2012.

FLORES-LOPEZ, C. A.; MACHADO, C. A. Differences in inferred genome-wide signals of positive selection during the evolution of *Trypanosoma cruzi* and *Leishmania* spp. lineages: A result of disparities in host and tissue infection ranges? **Infection, Genetics and Evolution**, v. 33, p. 37–46, 2015.

FORCE, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. **Genetics**, v. 151, n. 4, p. 1531–1545, 1999.

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. 1–9, 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, 2012.

FRASCH, A. C. C. Trans-sialidase, SAPA amino acid repeats and the relationship between *Trypanosoma cruzi* and the mammalian host. **Parasitology**, v. 108, n. Supplement S1, p. S37–S44, 1994.

FREIRE-DE-LIMA, L. et al. Sialic acid: A sweet swing between mammalian host and *Trypanosoma cruzi*. **Frontiers in Immunology**, v. 3, n. NOV, p. 1–12, 2012.

FREITAS, L. M. et al. Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of *Trypanosoma cruzi* reveal an undetected level of complexity. **PLoS ONE**, v. 6, n. 10, 2011.

GAUNT, M. W. et al. Mechanism of genetic exchange in American trypanosomes. **Nature**, v. 421, n. February, p. 936–939, 2003.

GHEDIN, E. et al. Gene synteny and evolution of genome architecture in trypanosomatids. **Molecular and Biochemical Parasitology**, v. 134, n. 2, p. 183–191, 2004.

GORDON, D. J.; RESIO, B.; PELLMAN, D. Causes and consequences of aneuploidy in cancer. **Nature reviews. Genetics**, v. 13, n. 3, p. 189–203, 2012.

GRISARD, E. C. et al. *Trypanosoma cruzi* Clone Dm28c Draft Genome Sequence. **Genome announcements**, v. 2, n. 1, p. 2–3, 2014.

GUINDON, S. et al. Estimating maximum likelihood phylogenies with PhyML. **Methods in Molecular Biology**, v. 537, p. 113–137, 2009.

GUINDON, S. et al. New Algorithms and Mehtods to Estimate Maximum-Likelihood Phylogenies: Asessing the Performance of PhyML 2.0. **Systematic Biology**, v. 59, n. 3, p. 307–321, 2010.

GÜNZL, A. et al. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. **Eukaryotic Cell**, v. 2, n. 3, p. 542–551, 2003.

GÜNZL, A. The pre-mRNA splicing machinery of trypanosomes: Complex or simplified? **Eukaryotic Cell**, v. 9, n. 8, p. 1159–1170, 2010.

HAAG, J.; O'HUIGIN, C.; OVERATH, P. The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. **Molecular and Biochemical Parasitology**, v. 91, n. 1, 1998.

HASSOLD, T.; HUNT, P. To err (meiotically) is human: the genesis of human aneuploidy. **Nat Rev Genet**, v. 2, n. 4, p. 280–291, 2001.

HENRICHSEN, C. N.; CHAIGNAT, E.; REYMOND, A. Copy number variants, diseases and gene expression. **Human Molecular Genetics**, v. 18, n. R1, 2009.

HENRIKSSON, J. et al. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is evolutionarily informative. **Parasitology**, v. 124, n. Pt 3, p. 277–286, 2002.

HOFT, D. F. et al. Trans-Sialidase Recombinant Protein Mixed with CpG Motif-Containing Oligodeoxynucleotide Induces Protective Mucosal and Systemic *Trypanosoma cruzi* Immunity Involving CD8+ CTL and B Cell-Mediated Cross-Priming. **The Journal of Immunology**, v. 179, n. 10, p. 6889–6900, 2007.

HUSON, D. H. et al. Dendroscope: An interactive viewer for large phylogenetic trees. **BMC bioinformatics**, v. 8, p. 460, 2007.

ISKOW, R.; GOKCUMEN, O.; LEE, C. Exploring the role of copy number variants in human adaptation. **Trends in Genetics**, v. 28, n. 6, p. 245–257, 2012.

IVENS, A. C. The Genome of the Kinetoplastid Parasite, *Leishmania major*. **Science**, v. 309, n. 5733, p. 436–442, 2005.

JACKSON, YVES; PINTO, ANGIE; PETT, S. Chagas disease in Australia and New Zealand: risks and

needs for public health interventions. **Tropical Medicine & International Health**, v. 19, n. 2, p. 212–218, 2014.

JACKSON, A. P. Tandem gene arrays in *Trypanosoma brucei*: comparative phylogenomic analysis of duplicate sequence variation. **BMC evolutionary biology**, v. 7, p. 54, 2007a.

JACKSON, A. P. Evolutionary consequences of a large duplication event in *Trypanosoma brucei*: Chromosomes 4 and 8 are partial duplicons. **BMC Genomics**, v. 8, n. 1, p. 432, 2007b.

JACKSON, A. P. et al. The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African Trypanosomiasis. **PLoS Neglected Tropical Diseases**, v. 4, n. 4, 2010.

JACKSON, A. P. et al. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. **Current Biology**, v. 26, n. 2, p. 161–172, 2016.

JIANG, H. et al. Rapid evolution in a pair of recent duplicate segments of rice. **Journal of Experimental Zoology Part B: Molecular and Developmental Evolution**, v. 308, n. 1, p. 50–57, 2007.

JONES, C. et al. Heterogeneity in the biosynthesis of mucin O-glycans from *Trypanosoma cruzi* Tulahuen strain with the expression of novel galactofuranosyl-containing oligosaccharides. **Biochemistry**, v. 43, n. 37, p. 11889–11897, 2004.

JUNQUEIRA, C. et al. The endless race between *Trypanosoma cruzi* and host immunity: lessons for and beyond Chagas disease. **Expert reviews in molecular medicine**, v. 12, p. e29, 2010.

KAWASHITA, S. Y. et al. Homology, paralogy and function of DGF-1, a highly dispersed *Trypanosoma cruzi* specific gene family and its implications for information entropy of its encoded proteins. **Molecular and Biochemical Parasitology**, v. 165, n. 1, p. 19–31, 2009.

KENT, W. J. BLAT — The BLAST -Like Alignment Tool. **Genome Research**, v. 12, p. 656–664, 2002.

KULKARNI, M. M. et al. *Trypanosoma cruzi* GP63 proteins undergo stage-specific differential posttranslational modification and are important for host cell infection. **Infection and Immunity**, v. 77, n. 5, p. 2193–2200, 2009.

LACHAUD, L. et al. Constitutive mosaic aneuploidy is a unique genetic feature widespread in the *Leishmania* genus. **Microbes and Infection**, v. 16, n. 1, p. 61–66, 2014.

LAFFITTE, M.-C. N. et al. Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance. **F1000Research**, v. 5, p. 2350, 2016.

LANDWEBER, L. F. The evolution of RNA editing in kinetoplastid protozoa. **BioSystems**, v. 28, n. 1–3, p. 41–45, 1992.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods**, v. 9, n. 4, p. 357–359, 2012.

LEBOWITZ, J. H. et al. Coupling of poly(A) site selection and trans-splicing in *Leishmania*. **Genes and Development**, v. 7, n. 6, p. 996–1007, 1993.

LEE, M. G.; VAN DER PLOEG, L. H. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. **Annual review of microbiology**, v. 51, p. 463–489, 1997.

LEIFSO, K. et al. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: The *Leishmania* genome is constitutively expressed. **Molecular and Biochemical Parasitology**, v. 152, n. 1, p. 35–46, 2007.

LEPROHON, P. et al. Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. **Nucleic Acids Research**, v. 37, n. 5, p. 1387–1399, 2009.

LEWINSOHN, R. Carlos Chagas and the discovery of Chagas' disease (American trypanosomiasis). **Journal of the Royal Society of Medicine**, v. 74, n. 6, p. 451–5, 1981.

LEWIS, M. D. et al. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. **International Journal for Parasitology**, v. 39, n. 12, p. 1305–1317, 2009.

LEWIS, M. D. et al. Recent, independent and anthropogenic origins of *Trypanosoma cruzi* hybrids. **PLoS Neglected Tropical Diseases**, v. 5, n. 10, 2011.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LI, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **arXiv**. V1 p. 3, 2013.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 2009.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. **Genome Research**, v. 13, n. 9, p. 2178–2189, 2003.

LIANG, X.; HARITAN, A.; ULIEL, S. trans and cis Splicing in Trypanosomatids : Mechanism , Factors , and Regulation. **Eukaryotic cell**, v. 2, n. 5, p. 830–840, 2003.

LIMA, F. M. et al. Interclonal Variations in the Molecular Karyotype of *Trypanosoma cruzi*: Chromosome Rearrangements in a Single Cell-Derived Clone of the G Strain. **PLoS ONE**, v. 8, n. 5, 2013a.

LIMA, L. et al. Evolutionary Insights from Bat Trypanosomes: Morphological, Developmental and Phylogenetic Evidence of a New Species, *Trypanosoma* (*Schizotrypanum*) *erneyi* sp. nov., in African Bats Closely Related to *Trypanosoma* (*Schizotrypanum*) *cruzi* and Allied Species. **Protist**, v. 163, n. 6, p. 856–872, 2012.

LIMA, L. et al. *Trypanosoma livingstonei*: a new species from African bats supports the bat seeding hypothesis for the *Trypanosoma cruzi* clade. **Parasites & vectors**, v. 6, n. 1, p. 221, 2013b.

LIMA, L. et al. Genetic diversity of *Trypanosoma cruzi* in bats, and multilocus phylogenetic and phylogeographical analyses supporting Tcbat as an independent DTU (discrete typing unit). **Acta Tropica**, v. 151, n. 1, p. 166–177, 2015.

LIMA, V. D. OS S. et al. Expanding the knowledge of the geographic distribution of *Trypanosoma cruzi* TcII and TcV/TcVI genotypes in the Brazilian Amazon. **PloS one**, v. 9, n. 12, p. e116137, 2014a.

LIMA, V. S. et al. Wild *Trypanosoma cruzi* I genetic diversity in Brazil suggests admixture and disturbance in parasite populations from the Atlantic Forest region. **Parasites & vectors**, v. 7, n. 1, p. 263, 2014b.

LLEWELLYN, M. S. et al. Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. **PLoS Pathogens**, v. 5, n. 5, 2009.

LOMBRAÑA, R. et al. Transcriptionally Driven DNA Replication Program of the Human Parasite *Leishmania major*. **Cell Reports**, v. 16, n. 6, p. 1774–1786, 2016.

LORENZI, H. A.; ROBLEDO, G.; LEVIN, M. J. The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. **Molecular and Biochemical Parasitology**, v. 145, n. 2, p. 184–194, 2006.

LUQUETTI, A. O. et al. *Trypanosoma cruzi*: zymodemes associated with acute and chronic Chagas' disease in central Brazil. **Trans R Soc Trop Med Hyg**, v. 80, n. 3, p. 462–470, 1986.

LV, L. et al. Tetraploid cells from cytokinesis failure induce aneuploidy and spontaneous transformation of mouse ovarian surface epithelial cells. **Cell Cycle**, v. 11, n. 15, p. 2864–2875, 2012.

MACEDO, A. M.; OLIVEIRA, R. P.; PENA, S. D. J. Chagas disease: role of parasite genetic variation in pathogenesis. **Expert reviews in molecular medicine**, v. 4, n. 5, p. 1–16, 2002.

MACHADO, C. A; 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. 7396–7401, 2001.

MACHADO, C. R. et al. DNA metabolism and genetic diversity in Trypanosomes. **Mutation Research - Reviews in Mutation Research**, v. 612, n. 1, p. 40–57, 2006a.

MACHADO, A. V. et al. Long-Term Protective Immunity Induced Against *Trypanosoma cruzi* Infection After Vaccination with Recombinant Adenoviruses Encoding Amastigote Surface Protein-2 and *Trans*- Sialidase. **Human Gene Therapy**, v. 17, n. 9, p. 898–908, 2006b.

MAEDA, F. Y. et al. Mammalian cell invasion by closely related *Trypanosoma* species *T. dionisii* and *T. cruzi*. **Acta Tropica**, v. 121, n. 2, p. 141–147, 2012.

MAERE, S.; HEYMANS, K.; KUIPER, M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. **Bioinformatics**, v. 21, n. 16, p. 3448–3449, 2005.

MANNAERT, A. et al. Adaptive mechanisms in pathogens: Universal aneuploidy in *Leishmania*. **Trends in Parasitology**, v. 28, n. 9, p. 370–376, 2012.

MARÇAIS, G.; KINGSFORD, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. **Bioinformatics**, v. 27, n. 6, p. 764–770, 2011.

MARCILI, A. et al. A new genotype of *Trypanosoma cruzi* associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA. **Parasitology**, v. 136, n. 6, p. 641–655, 2009.

MARQUES, C. A. et al. Genome-wide mapping reveals single-origin chromosome replication in *Leishmania*, a eukaryotic microbe. **Genome Biology**, v. 16, n. 1, p. 230, 2015.

MARTÍNEZ-CALVILLO, S. et al. Gene expression in trypanosomatid parasites. **Journal of Biomedicine and Biotechnology**, v. 2010, 2010.

MARTINS, K. et al. *Trypanosoma cruzi* III causing the indeterminate form of Chagas disease in a

semi-arid region of Brazil. **International Journal of Infectious Diseases**, v. 39, p. 68–75, 2015.

MATHERS CD, LOPEZ AD, M. C. The Burden of Disease and Mortality by Condition: Data, Methods, and Results for 2001. In: **Global Burden of Disease and Risk Factors**. 2006

MCKENNA, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, v. 20, p. 1297–1303, 2010.

MCNICOLL, F. et al. Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*. **Journal of Biological Chemistry**, v. 280, n. 42, p. 35238–35246, 2005.

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, 2012.

MESSENGER, L. A. et al. Ecological host fitting of *Trypanosoma cruzi* TcI in Bolivia: Mosaic population structure, hybridization and a role for humans in Andean parasite dispersal. **Molecular Ecology**, v. 24, n. 10, p. 2406–2422, 2015.

MESSENGER, L. A. et al. Importation of hybrid human-associated *Trypanosoma cruzi* strains of southern South American origin, Colombia. **Emerging Infectious Diseases**, v. 22, n. 8, p. 1452–1455, 2016.

MESSENGER, L. A.; MILES, M. A. Evidence and importance of genetic exchange among field populations of *Trypanosoma cruzi*. **Acta Tropica**, v. 151, n. 1, p. 150–155, 2015.

MESSENGER, L. A.; MILES, M. A.; BERN, C. Between a bug and a hard place: *Trypanosoma cruzi* genetic diversity and the clinical outcomes of Chagas disease. **Expert review of anti-infective therapy**, v. 13, n. 8, p. 995–1029, 2015.

METZKER, M. L. Sequencing technologies - the next generation. **Nature reviews. Genetics**, v. 11, n. 1, p. 31–46, 2010.

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. 217–225, 1977.

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, 2009.

MILLER, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. **Bioinformatics**, v. 24, n. 24, p. 2818–2824, 2008.

MINNING, T. A et al. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization. **BMC genomics**, v. 12, n. 1, p. 139, 2011.

MONTEIRO, W. M. et al. *Trypanosoma cruzi* TcIII/Z3 genotype as agent of an outbreak of Chagas disease in the Brazilian Western Amazonia: Short Communication. **Tropical Medicine and International Health**, v. 15, n. 9, p. 1049–1051, 2010.

MORAES BARROS, R. R. et al. Anatomy and evolution of telomeric and subtelomeric regions in the human protozoan parasite *Trypanosoma cruzi*. **BMC genomics**, v. 13, p. 229, 2012.

MORGULIS, A. et al. Database indexing for production MegaBLAST searches. **Bioinformatics**, v. 24, n. 16, p. 1757–1764, 2008.

MÜLLER, M. et al. Rapid decay of unstable *Leishmania* mRNAs bearing a conserved retroposon signature 3'-UTR motif is initiated by a site-specific endonucleolytic cleavage without prior deadenylation. **Nucleic Acids Research**, v. 38, n. 17, p. 5867–5883, 2010.

MYERS, E. W. A Whole-Genome Assembly of *Drosophila*. **Science**, v. 287, n. 5461, p. 2196–2204, 2000.

MYUNG, K. S. et al. Comparison of the post-transcriptional regulation of the mRNAs for the surface proteins PSA (GP46) and MSP (GP63) of *Leishmania chagasi*. **Journal of Biological Chemistry**, v. 277, n. 19, p. 16489–16497, 2002.

OTTO, S. P.; GERSTEIN, A. C. The evolution of haploidy and diploidy. **Current Biology**, v. 18, n. 24, 2008.

PANUNZI, L. G.; AGÜERO, F. A Genome-Wide Analysis of Genetic Diversity in *Trypanosoma cruzi* Intergenic Regions. **PLoS Neglected Tropical Diseases**, v. 8, n. 5, 2014.

PEACOCK, C. S. et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. **Nature genetics**, v. 39, n. 7, p. 839–47, 2007.

PEACOCK, L. et al. Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 9, p. 3671–6, 2011.

PEDROSO, A.; CUPOLILLO, E.; ZINGALES, B. Evaluation of *Trypanosoma cruzi* hybrid stocks based

on chromosomal size variation. **Molecular and Biochemical Parasitology**, v. 129, n. 1, p. 79–90, 2003.

PEREIRA-CHIOCCOLA, V. L. et al. Mucin-like molecules form a negatively charged coat that protects *Trypanosoma cruzi* trypomastigotes from killing by human anti-alpha-galactosyl antibodies. **Journal of cell science**, v. 113 (Pt 7, p. 1299–1307, 2000.

PFAU, S. J.; AMON, A. Chromosomal instability and aneuploidy in cancer: from yeast to man. **EMBO reports**, v. 13, n. 6, p. 515–27, 2012.

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**, v. 12, n. 6, p. 1328–1332, 2012.

POSADA, D. jModelTest: Phylogenetic model averaging. **Molecular Biology and Evolution**, v. 25, n. 7, p. 1253–1256, 2008.

PREUSSE, C.; JAÉ, N.; BINDEREIF, A. MRNA splicing in trypanosomes. **International Journal of Medical Microbiology**, v. 302, n. 4–5, p. 221–224, 2012.

PURCELL, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. **The American Journal of Human Genetics**, v. 81, n. 3, p. 559–575, 2007.

QUINLAN, A. R.; HALL, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, n. 6, p. 841–842, 2010.

RAMBAUT, A. FigTree, a graphical viewer of phylogenetic trees. **Institute of Evolutionary Biology University of Edinburgh**, 2009.

RAMÍREZ, J. D. et al. Contemporary cryptic sexuality in *Trypanosoma cruzi*. **Molecular Ecology**, v. 21, n. 17, p. 4216–4226, 2012.

RAMÍREZ, J. D.; HERNÁNDEZ, C. Acta Tropica *Trypanosoma cruzi I* : Towards the need of genetic subdivision ?. **Acta Tropica**, n. February, p. 0–1, 2017.

RANCATI, G. et al. Aneuploidy Underlies Rapid Adaptive Evolution of Yeast Cells Deprived of a Conserved Cytokinesis Motor. **Cell**, v. 135, n. 5, p. 879–893, 2008.

RASSI, A. Predicting prognosis in patients with Chagas disease: Why are the results of various studies so different? **International Journal of Cardiology**, v. 145, n. 1, p. 64–65, 2010.

REIS-CUNHA, J. L. et al. Chromosomal copy number variation reveals differential levels of

genomic plasticity in distinct *Trypanosoma cruzi* strains. **BMC genomics**, v. 16, n. 1, p. 499, 2015.

REIS-CUNHA, J. L.; VALDIVIA, H. O.; BARTHOLOMEU, D. C. Trypanosomatid Genome Organization and Ploidy. In: SILVA, M. S.; CANO, M. I. (Eds.) . **Molecular and Cellular Biology of Pathogenic Trypanosomatids**. 1. ed. Frontiers in Parasitology, 2017. p. 61–103.

RIBEIRO, A. L. et al. Diagnosis and management of Chagas disease and cardiomyopathy. **Nature Reviews Cardiology**, v. 9, n. 10, p. 576–589, 2012.

RODRÍGUEZ, A., MARTINEZ, I., CHUNG, A., BERLOT, C.H., AND ANDREWS, N. W. cAMP Regulates Ca2+ -dependent Exocytosis of Lysosomes and Lysosome-mediated Cell Invasion by Trypanosomes. **Biochemistry**, v. 274, n. 24, p. 16754–16759, 1999.

RODRÍGUEZ, A. et al. Host cell invasion by trypanosomes requires lysosomes and microtubule/kinesin-mediated transport. **Journal of Cell Biology**, v. 134, n. 2, p. 349–362, 1996.

ROGERS, M. B. et al. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. **Genome Res.**, v. 21, p. 2129–2142, 2011.

ROMANO, P. S. et al. Molecular and cellular mechanisms involved in the *Trypanosoma cruzi*/host cell interplay. **IUBMB Life**, v. 64, n. 5, p. 387–396, 2012.

ROTH, C. et al. Evolution after gene duplication: Models, mechanisms, sequences, systems, and organisms. **Journal of Experimental Zoology Part B: Molecular and Developmental Evolution**, v. 308, n. 1, p. 58–73, 2007.

RUBIN-DE-CELIS, S. S. C. et al. Expression of trypomastigote trans-sialidase in metacyclic forms of *Trypanosoma cruzi* increases parasite escape from its parasitophorous vacuole. **Cellular Microbiology**, v. 8, n. 12, p. 1888–1898, 2006.

RUIZ, R. D. C. et al. The 35/50 kDa surface antigen of *Trypanosoma cruzi* metacyclic trypomastigotes, an adhesion molecule involved in host cell invasion. **Parasite Immunology**, v. 15, n. 2, p. 121–125, 1993.

SCHENKMAN, S. et al. Structural and Functional Properties of Trans-Sialidase. **Annual Review of Microbiology**, v. 48, p. 499–523, 1994.

SEBAT, J. et al. Large-scale copy number polymorphism in the human genome. **Science**, v. 305, n. 5683, p. 525–8, 2004.

SECO-HIDALGO, V.; DE PABLOS, L. M.; OSUNA, A. Transcriptional and phenotypical heterogeneity of *Trypanosoma cruzi* cell populations. **Open biology**, v. 5, n. 12, p. 150190, 2015.

SELMECKI, A.; FORCHE, A.; BERMAN, J. Genomic plasticity of the human fungal pathogen *Candida albicans*. **Eukaryotic Cell**, v. 9, n. 7, p. 991–1008, 2010.

SERRANO, A. A. et al. The lipid structure of the glycosylphosphatidylinositol-anchored mucin-like sialic acid acceptors of *Trypanosoma cruzi* changes during parasite differentiation from epimastigotes to infective metacyclic trypomastigote forms. **Journal of Biological Chemistry**, v. 270, n. 45, p. 27244–27253, 1995.

SHELTZER, J. M. J. et al. Aneuploidy drives genomic instability in yeast. **Science**, v. 333, n. 6045, p. 1026–30, 2011.

SOUTO, R. P. et al. DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. **Molecular and Biochemical Parasitology**, v. 83, n. 2, p. 141–152, 1996.

SOUZA, R. T. et al. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. **PLoS ONE**, v. 6, n. 8, 2011.

STANKIEWICZ, P.; LUPSKI, J. R. Structural Variation in the Human Genome and its Role in Disease. **Annual Review of Medicine**, v. 61, n. 1, p. 437–455, 2010.

STERKERS, Y. et al. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. **Cellular Microbiology**, v. 13, n. 2, p. 274–283, 2011.

STERKERS, Y. et al. Parasexuality and mosaic aneuploidy in *Leishmania*: Alternative genetics. **Trends in Parasitology**, v. 30, n. 9, p. 429–435, 2014.

STOCO, P. H. et al. Genome of the Avirulent Human-Infective Trypanosome - *Trypanosoma rangeli*. **PLoS Neglected Tropical Diseases**, v. 8, n. 9, 2014.

STRANGER, B. E. et al. Expression Phenotypes. **Recherche**, v. 315, n. February, p. 848–853, 2007.

STUART, K. RNA editing in kinetoplastid protozoa. **Current Opinion in Genetics and Development**, v. 1, n. 3, p. 412–416, 1991.

STURM, N. R. et al. Evidence for multiple hybrid groups in *Trypanosoma cruzi*. **International Journal for Parasitology**, v. 33, n. 3, p. 269–279, 2003.

STURM, N. R.; CAMPBELL, D. A. Alternative lifestyles: The population structure of *Trypanosoma cruzi*. **Acta Tropica**, v. 115, n. 1–2, p. 35–43, 2010a.

STURM, N. R.; CAMPBELL, D. A. Alternative lifestyles: The population structure of *Trypanosoma cruzi*. **Acta Tropica**, v. 115, n. 1–2, p. 35–43, 2010b.

SUZUKI, R.; SHIMODAIRA, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. **Bioinformatics**, v. 22, n. 12, p. 1540–1542, 2006.

TALAVERA, G.; CASTRESANA, J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. **Systematic Biology**, v. 56, n. 4, p. 564–577, 2007.

TARDIEUX, I. et al. Lysosome recruitment and fusion are early events required for trypanosome invasion of mammalian cells. **Cell**, v. 71, n. 7, p. 1117–1130, 1992.

TEIXEIRA, A. R. L.; NASCIMENTO, R. J.; STURM, N. R. Evolution and pathology in Chagas disease - A Review. **Memorias do Instituto Oswaldo Cruz**, v. 101, n. 5, p. 463–491, 2006.

TEIXEIRA, S. M. et al. Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases. **Genetics and Molecular Biology**, v. 35, n. 1, p. 1–17, 2012.

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, p. 115–119, 1986.

TIBAYRENC, M. Genetic subdivisions within *Trypanosoma cruzi* (Discrete Typing Units) and their relevance for molecular epidemiology and experimental evolution. **Kinetoplastid biology and disease**, v. 2, p. 12, 2003.

TIENGWE, C. et al. Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of *Trypanosoma brucei*. **Cell Reports**, v. 2, n. 1, p. 185–197, 2012.

TOMASINI, N.; DIOSQUE, P. Phylogenomics of *Trypanosoma cruzi* : Few evidence of TcI/TcII mosaicism in TcIII challenges the hypothesis of an ancient TcI/TcII hybridization. **Infection, Genetics and Evolution**, v. 50, p. 25–27, 2017.

TORRES, E. M.; WILLIAMS, B. R.; AMON, A. Aneuploidy: Cells losing their balance. **Genetics**, v. 179, n. 2, p. 737–746, 2008.

TRIANA, O. et al. *Trypanosoma cruzi*: Variability of stocks from Colombia determined by molecular karyotype and minicircle Southern blot analysis. **Experimental Parasitology**, v. 113, n. 1, p. 62–66, 2006.

TÜMPEL, S. et al. Evolution of cis elements in the differential expression of two Hoxa2

coparalogous genes in pufferfish (Takifugu rubripes). **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 14, p. 5419–5424, 2006.

UBEDA, J.-M. et al. Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy. **Genome biology**, v. 9, n. 7, p. R115, 2008.

VALDIVIA, H. O. et al. The *Leishmania* metaphylome: a comprehensive survey of *Leishmania* protein phylogenetic relationships. **BMC genomics**, v. 16, n. 1, p. 887, 2015a.

VALDIVIA, H. O. et al. Comparative genomic analysis of *Leishmania* (*Viannia*) *peruviana* and *Leishmania* (*Viannia*) *braziliensis*. **BMC genomics**, v. 16, n. 1, p. 715, 2015b.

VALDIVIA, H. O. et al. Comparative genomics of canine-isolated *Leishmania* (*Leishmania*) *amazonensis* from an endemic focus of visceral leishmaniasis in Governador Valadares, southeastern Brazil. **Scientific Reports**, v. 7, n. September 2016, p. 40804, 2017.

VARGAS, N.; PEDROSO, A.; ZINGALES, B. Chromosomal polymorphism, gene synteny and genome size in *T. cruzi* I and *T. cruzi* II groups. **Molecular and Biochemical Parasitology**, v. 138, n. 1, p. 131–141, 2004.

WANG, H. et al. Molecular evidence for asymmetric evolution of sister duplicated blocks after cereal polyploidy. **Plant Molecular Biology**, v. 59, n. 1, p. 63–74, 2005.

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

WEIR, W. et al. Population genomics reveals the origin and asexual evolution of human infective trypanosomes. **eLife**, v. 5, n. JANUARY2016, p. 1–14, 2016.

WESTENBERGER, S. J. et al. Two hybridization events define the population structure of *Trypanosoma cruzi*. **Genetics**, v. 171, n. 2, p. 527–543, 2005.

WHO. Research priorities for Chagas disease, human African trypanosomiasis and leishmaniasis. **World Health Organization technical report series**, n. 975, p. v–xii, 1-100, 2012.

WICKSTEAD, B.; ERSFELD, K.; GULL, K. Repetitive Elements in Genomes of Parasitic Protozoa Repetitive Elements in Genomes of Parasitic Protozoa. **Society**, v. 67, n. 3, p. 360–375, 2003.

WINCKER, P. et al. The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. **Nucleic Acids Research**, v. 24, n. 9, p. 1688–1694, 1996.

WOOLSEY, A. M.; BURLEIGH, B. A. Host cell actin polymerization is required for cellular retention of *Trypanosoma cruzi* and early association with endosomal/lysosomal compartments. **Cellular Microbiology**, v. 6, n. 9, p. 829–838, 2004.

YAO, C. Major surface protease of trypanosomatids: One size fits all? **Infection and Immunity**, v. 78, n. 1, p. 22–31, 2010.

YOSHIDA, N. et al. Removal of sialic acid from mucin-like surface molecules of *Trypanosoma cruzi* metacyclic trypomastigotes enhances parasite-host cell interaction. v. 84, p. 57–67, 1997.

YOSHIDA, N. *Trypanosoma cruzi* infection by oral route. How the interplay between parasite and host components modulates infectivity. **Parasitology International**, v. 57, n. 2, p. 105–109, 2008.

YOSHIDA, N. Molecular mechanisms of *Trypanosoma cruzi* infection by oral route. **Memórias do Instituto Oswaldo Cruz**, v. 104 Suppl, p. 101–7, 2009.

ZERBINO, D. R.; BIRNEY, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, n. 5, p. 821–829, 2008.

ZEYL, C.; VANDERFORD, T.; CARTER, M. An evolutionary advantage of haploidy in large yeast populations. **Science**, v. 299, n. 2003, p. 555–558, 2003.

ZHANG, P. A Segmental Gene Duplication Generated Differentially Expressed myb-Homologous Genes in Maize. **the Plant Cell Online**, v. 12, n. 12, p. 2311–2322, 2000.

ZINGALES, B. et al. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: Second revision meeting recommends TcI to TcVI. **Memorias do Instituto Oswaldo Cruz**, v. 104, n. 7, p. 1051–1054, 2009.

ZINGALES, B. et al. The revised *Trypanosoma cruzi* subspecific nomenclature: Rationale, epidemiological relevance and research applications. **Infection, Genetics and Evolution**, v. 12, n. 2, p. 240–253, 2012.

ZOMERDIJK, J. C.; KIEFT, R.; BORST, P. Efficient production of functional mRNA mediated by RNA polymerase I in *Trypanosoma brucei*. **Nature**, v. 353, n. 6346, p. 772–775, 1991.

Anexo 1:

Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi*

BMC Genomics, 2015

Reis-Cunha et al. BMC Genomics (2015) 16:499
DOI 10.1186/s12864-015-1680-4

BMC Genomics

RESEARCH ARTICLE Open Access

Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains

João Luís Reis-Cunha[1], Gabriela F. Rodrigues-Luiz[1], Hugo O. Valdivia[1], Rodrigo P. Baptista[1], Tiago A. O. Mendes[1], Guilherme Loss de Morais[2], Rafael Guedes[2], Andrea M. Macedo[3], Caryn Bern[4], Robert H. Gilman[5,6], Carlos Talavera Lopez[7], Björn Andersson[7], Ana Tereza Vasconcelos[2] and Daniella C. Bartholomeu[1*]

Abstract

Background: *Trypanosoma cruzi*, the etiologic agent of Chagas disease, is currently divided into six discrete typing units (DTUs), named TcI–TcVI. CL Brener, the reference strain of the *T. cruzi* genome project, is a hybrid with a genome assembled into 41 putative chromosomes. Gene copy number variation (CNV) is well documented as an important mechanism to enhance gene expression and variability in *T. cruzi*. Chromosomal CNV (CCNV) is another level of gene CNV in which whole blocks of genes are expanded simultaneously. Although the *T. cruzi* karyotype is not well defined, several studies have demonstrated a significant variation in the size and content of chromosomes between different *T. cruzi* strains. Despite these studies, the extent of diversity in CCNV among *T. cruzi* strains based on a read depth coverage analysis has not been determined.

Results: We identify the CCNV in *T. cruzi* strains from the TcI, TcII and TcIII DTUs, by analyzing the depth coverage of short reads from these strains using the 41 CL Brener chromosomes as reference. This study led to the identification of a broader extent of CCNV in *T. cruzi* than was previously speculated. The TcI DTU strains have very few aneuploidies, while the strains from TcII and TcIII DTUs present a high degree of chromosomal expansions. Chromosome 31, which is the only chromosome that is supernumerary in all six *T. cruzi* samples evaluated in this study, is enriched with genes related to glycosylation pathways, highlighting the importance of glycosylation to parasite survival.

Conclusions: Increased gene copy number due to chromosome amplification may contribute to alterations in gene expression, which represents a strategy that may be crucial for parasites that mainly depend on post-transcriptional mechanisms to control gene expression.

Keywords: Chromosome copy number variation, *Trypanosoma cruzi*, Genomic plasticity

Background

American trypanosomiasis is a neglected tropical disease, caused by the protozoan *Trypanosoma cruzi*, a highly polymorphic parasite that belongs to the order Kinetoplastida and family Trypanosomatidae. The distribution of this disease ranges from southern Argentina to the southern United States of America, where it affects eight million people and accounts for 662,000 disability-adjusted life years [1–4].

The *T. cruzi* taxa is currently subdivided into six discrete typing units (DTUs), named TcI – TcVI, due to its high genotypic and phenotypic heterogeneity. The major *T. cruzi* DTUs involved in the domestic cycle of Chagas disease are TcI, TcII, TcV and TcVI [5–11]. The distinct DTUs are differently distributed in the Americas, with TcI prevalent in Central America and in the northern region of South America, while TcII, TcV and TcVI are more common in the Southern cone of South America [10].

* Correspondence: daniella@icb.ufmg.br
[1]Laboratório de Imunologia e Genômica de Parasitos, Departamento deParasitologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

T. cruzi replication is usually clonal [12, 13], but there is evidence of natural hybridization and genetic exchange between the strains [14–19]. The hybrid nature of *T. cruzi* DTU TcVI was confirmed during the whole-genome sequencing of the TcVI CL Brener clone [20]. Post-assembly comparisons of CL Brener contigs with reads from the *T. cruzi* Esmeraldo TcII strain allowed the differentiation of the two CL Brener haplotypes, named Esmeraldo-like, derived from a TcII ancestor, and non-Esmeraldo-like, derived from a TcIII ancestor [20].

Apparently, TcI has the smallest genome of all the *T. cruzi* DTUs [6, 21, 22], and it seems to have less intra-genomic heterogeneity than TcII and TcVI [23]. However, sequences from different TcI strains may present more sequence variability between each other than the variability within the TcII and TcVI strains [8, 23].

Copy number variation (CNV)—the gain or loss of genomic material—may have a phenotypic impact by altering the fitness of an organism. CNV creates paralog genes that may evolve differently than the progenitor gene or that may alter the expression level of a gene or genomic region [24, 25]. In CL Brener, at least 50 % of the genome consists of repetitive sequences, represented primarily by large multigene families that encode surface proteins, retrotransposons, and telomeric and satellite repeats [20, 26]. The CL Brener genome contains approximately 1000 paralogous clusters with more than two genes, encompassing over 8000 genes. Several of these clusters are represented by surface protein-encoding genes that account for 18 % of the total of protein-encoding genes of CL Brener [20]. An increased gene copy number due to chromosomal amplification may contribute to alterations in gene expression, providing a strategy for organisms, such as *T. cruzi,* that depend mainly on post-transcriptional mechanisms to control gene expression [27, 28].

The *T. cruzi* karyotype has not been completely elucidated owing to the inability to perform cytogenetic analysis because there is no apparent chromosome condensation during the parasite cell cycle [8, 22, 29]. Using pulse-field gel electrophoresis (PFGE), various studies have shown a significant variation in chromosome size and content between different *T. cruzi* strains and even between clones of the same strain [5, 29–32]. To better characterize the *T. cruzi* karyotype, the genome sequence of the CL Brener strain was recently assembled into 41 putative chromosomes based on scaffold information, BAC-end sequences, and synteny maps with *Trypanosoma brucei* and *Leishmania major* [8, 33].

Read depth coverage (RDC) analysis allows the identification of extensive variations in the copy number of chromosomes among different species of *Leishmania* [34]. However, the chromosomal copy number variation (CCNV) among *T. cruzi* strains as determined by read depth coverage analysis has not yet been reported. In the present work, we sought to identify the CCNV in *T. cruzi* strains belonging to different DTUs, based on read depth coverage of the 41 CL Brener chromosomes. Identifying the CCNV will lead to explanations of some of the genome structural peculiarities of these DTUs. This analysis also led to the identification of a broader extent of CCNV in *T. cruzi* than previously speculated, especially in strains from the TcII and TcIII DTUs.

Results

Competitive mapping and SNP content

To select the CL Brener haplotype most suitable as a reference in the mapping of the reads from the distinct *T. cruzi* strains, their quality-filtered reads were mapped simultaneously to the 41 chromosomal sequences from the CL Brener Esmeraldo-like and non-Esmeraldo-like haplotypes (Fig. 1a). As expected, the reads from the TcII strains (Esmeraldo and Y) mapped preferentially with chromosomes from the Esmeraldo-like haplotype, and the reads from the TcIII strain (231) mapped preferentially with chromosomes from the non-Esmeraldo-like haplotype. The reads from the strains from TcI DTU (Arequipa, Colombiana and Sylvio) mapped slightly better to the non-Esmeraldo-like than to the Esmeraldo-like CL Brener haplotype (Fig. 1b).

To further confirm the selection of the haplotype to be used as a reference for read mapping, the filtered reads from each strain were mapped separately to both CL Brener haplotypes, and then the number of SNPs/1000 bp in the CL Brener single-copy genes for each combination of strain-haplotype was estimated (Fig. 1c). The numbers of SNPs/1000 bp between each *T. cruzi* strain and the CL Brener Esmeraldo-like and non-Esmeraldo-like haplotypes were, respectively, 17.50 and 3.10 for Arequipa; 32.74 and 14.37 for Colombiana; 52.41 and 25.36 for Sylvio; 5.66 and 18.07 for Esmeraldo; 4.39 and 16.08 for Y; and 17.89 and 5.72 for 231.

Based on these results, the CL Brener non-Esmeraldo-like haplotype was selected as a reference for the mapping of Arequipa, Colombiana, Sylvio and 231 reads, and the CL Brener Esmeraldo-like haplotype was selected as a reference for the mapping of Esmeraldo and Y strain reads.

Methodology to estimate chromosome copy number in *T. cruzi* strains

Approximately 50 % of the *T. cruzi* genome corresponds to repetitive sequences, including multigene families that account for much of the differences in the gene content of the assembled genomes of CL Brener (TcVI) and Sylvio (TcI) [20, 21]. The chromosomal sequence representation of the CL Brener non-Esmeraldo-like and Esmeraldo-like haplotypes [33] also contains large internal gap regions

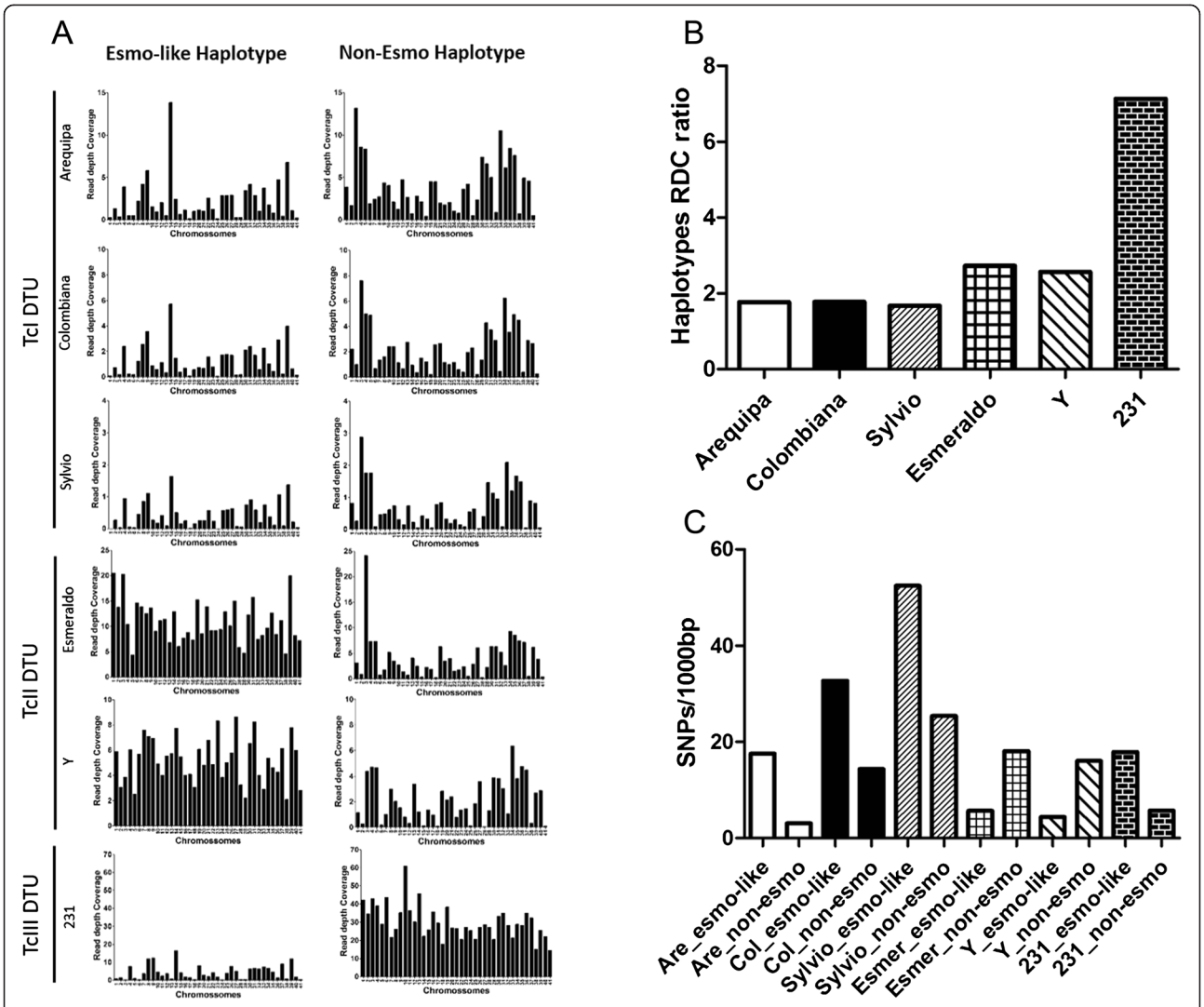


Fig. 1 Competitive mapping and SNP content of the *T. cruzi* strains from different DTUs. In competitive mapping, whole-genome shotgun reads from the *T. cruzi* Arequipa, Colombiana, Sylvio, Esmeraldo, Y and 231 strains were simultaneously mapped to the 41 chromosomal sequences from Esmeraldo-like (Esmo-like) and non-Esmeraldo-like (Non-Esmo) CL Brener haplotypes, and their mean read depth coverage (RDC) was estimated. Each black bar corresponds to the mean RDC of each of the 41 CL Brener chromosomes (**a**). For each *T. cruzi* strain, the mean RDC of the CL Brener haplotype with the highest RDC in the competitive mapping (non-Esmeraldo for Arequipa, Sylvio and 231; Esmeraldo-like for Esmeraldo and Y) was divided by the mean RDC of the CL Brener haplotype with the lower RDC (Esmeraldo-like for Arequipa, Sylvio and 231; non-Esmeraldo for Esmeraldo and Y) (**b**). The SNP density of each *T. cruzi* strain and the 1563 Esmeraldo-like/non-Esmeraldo-like single-copy genes are shown (**c**)

that may reduce the accuracy of the predicted ploidy based on RDC. To reveal the best methodology to determine the *T. cruzi* chromosome ploidy based on RDC, two approaches were evaluated. In the Whole Chromosome Ploidy Estimations (WCPE) approach, the chromosomal ploidy prediction for each chromosome was estimated based on the ratio between the mean RDC of each chromosome position and the genome coverage (Fig. 2a). This approach accounts for the coverage of all positions in a given chromosome to estimate its copy number, including repetitive and gap regions. In the single-copy genes ploidy estimations (SCoPE) approach, estimations of the chromosomal ploidy for each chromosome were based on the ratio between the mean coverage of all single-copy genes in a given chromosome and the genome coverage (Fig. 2b). This approach infers the copy number for each chromosome based only on the RDC of the 1563 1:1 orthologs between CL Brener Esmeraldo-like and non-Esmeraldo-like haplotypes, which were assumed to be single-copy genes in the haploid CL Brener genome content (Additional file 1: Table S1). As shown in Fig. 2, chromosomes that are rich in multigene families, repetitive sequences and gaps, such as chromosomes 18, 28, 38 and 41 (Fig. 2c), tend to have a lower predicted ploidy as determined using WCPE methodology when compared to the SCoPE approach using the Y strain reads. Similar results

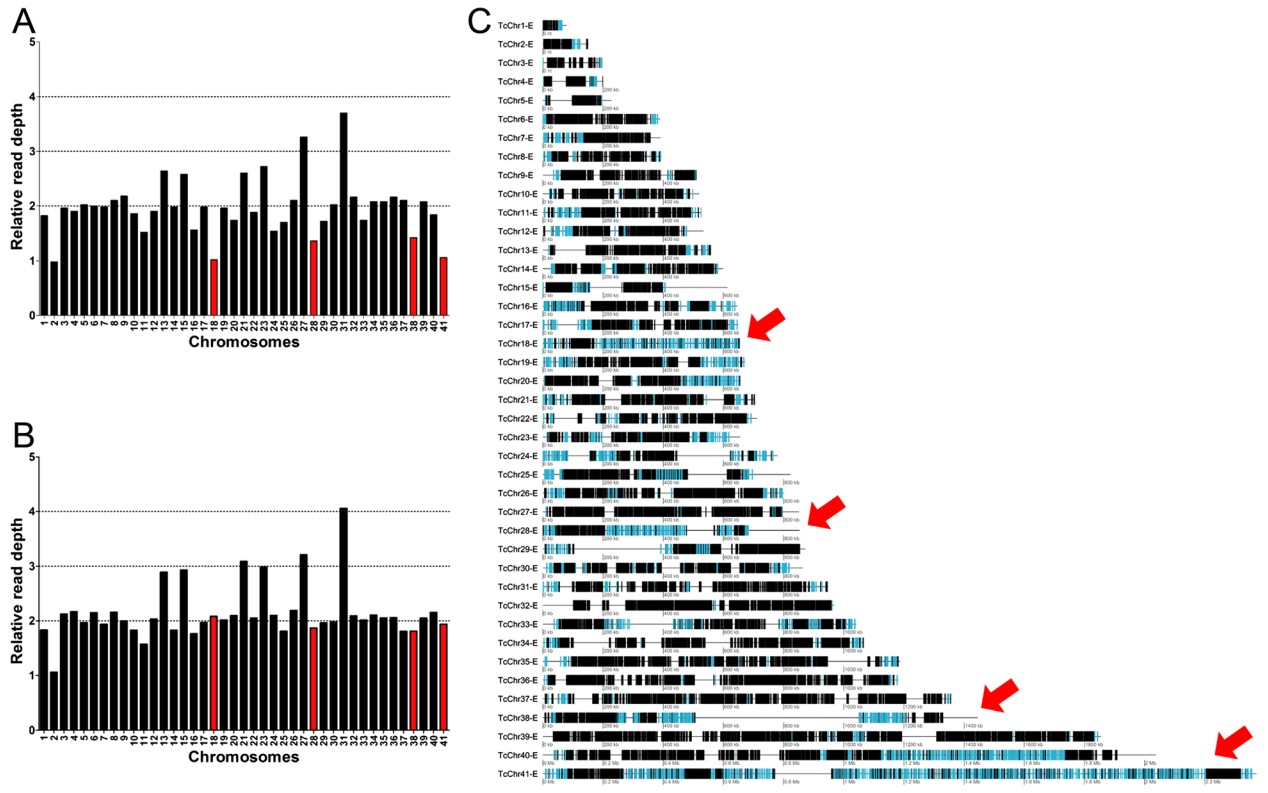


Fig. 2 Methodologies to determinate the chromosomal copy number in *T. cruzi* strains. To estimate the chromosome ploidy of *T. cruzi* strains, two approaches were evaluated by mapping reads from the Y strain on the 41 CL Brener chromosomes. **a** The Whole Chromosome Ploidy Estimations (WCPE) predicts the ploidy based on the mean coverage of the whole-chromosome sequence. **b** The single-copy genes ploidy estimations (SCoPE) predict the chromosomal ploidy based only on the coverage of the 1563 Esmo-like/Non-Esmo single-copy genes present in each chromosome. The read depth was scaled to give a value of 2 for disomic chromosomes. **c** The 41 *T. cruzi* chromosome gene distribution is shown. Multicopy gene families are represented as blue boxes and the hypothetical and housekeeping genes as black boxes. The red arrows highlight the chromosomes with a higher content of multigene families or with the largest proportion of gap regions. Chromosomes with a higher density of multigene families or that present large gap regions tend to have a biased lower predicted ploidy in the WCPE when compared to the SCoPE method. Similar results were obtained for all the other strains evaluated

were obtained by mapping reads from the other *T. cruzi* strains on the 41 CL Brener chromosomes (data not shown). As the SCoPE approach is less prone to bias toward chromosomal repetitive content, this methodology was chosen to estimate the chromosomal ploidy for each of the strains used in this study.

Chromosome copy number variation in *T. cruzi* strains

The SCoPE approach was used to estimate the chromosome ploidy of the *T. cruzi* Arequipa, Colombiana, Sylvio, Esmeraldo, Y and 231 strains. Initially, based on the mean RDC of 1563 CL Brener single-copy genes, the genome coverage was estimated for each strain: 47× for Arequipa, 28× for Colombiana, 9× for Sylvio, 52× for Esmeraldo, 34× for Y and 76× for 231. The normalized read depth coverage and the percentage of length coverage of each single-copy gene in each chromosome are provided in Additional file 2: Table S2. To determine the overall chromosome ploidy of each *T. cruzi* strain, the allele frequencies were estimated for each predicted heterozygous site. To this end, the proportion of each allele in the heterozygous sites divided by the total read depth for the site was determined and rounded to the first decimal place. Based on this estimation, a diploid chromosome usually has a tendency of 0.4-0.5 and a triploid of 0.3 and 0.6. A tetraploid chromosome has a more complex pattern, which can be 0.4-0.5, 0.2 and 0.8 or a combination of both. As the majority of the heterozygous SNPs show a proportion of 0.4-0.5, the overall ploidy of all the strains was assumed to be diploid (Fig. 3).

The chromosome CNV analysis revealed large differences between the *T. cruzi* strains from different DTUs and some differences between members of the same DTU. Apparently, strains from the TcI DTU have a more stable karyotype when compared to strains from the TcII and TcIII DTUs (Fig. 3). Strains from the TcI DTU usually only have an aneuploidy in chromosome 31, with

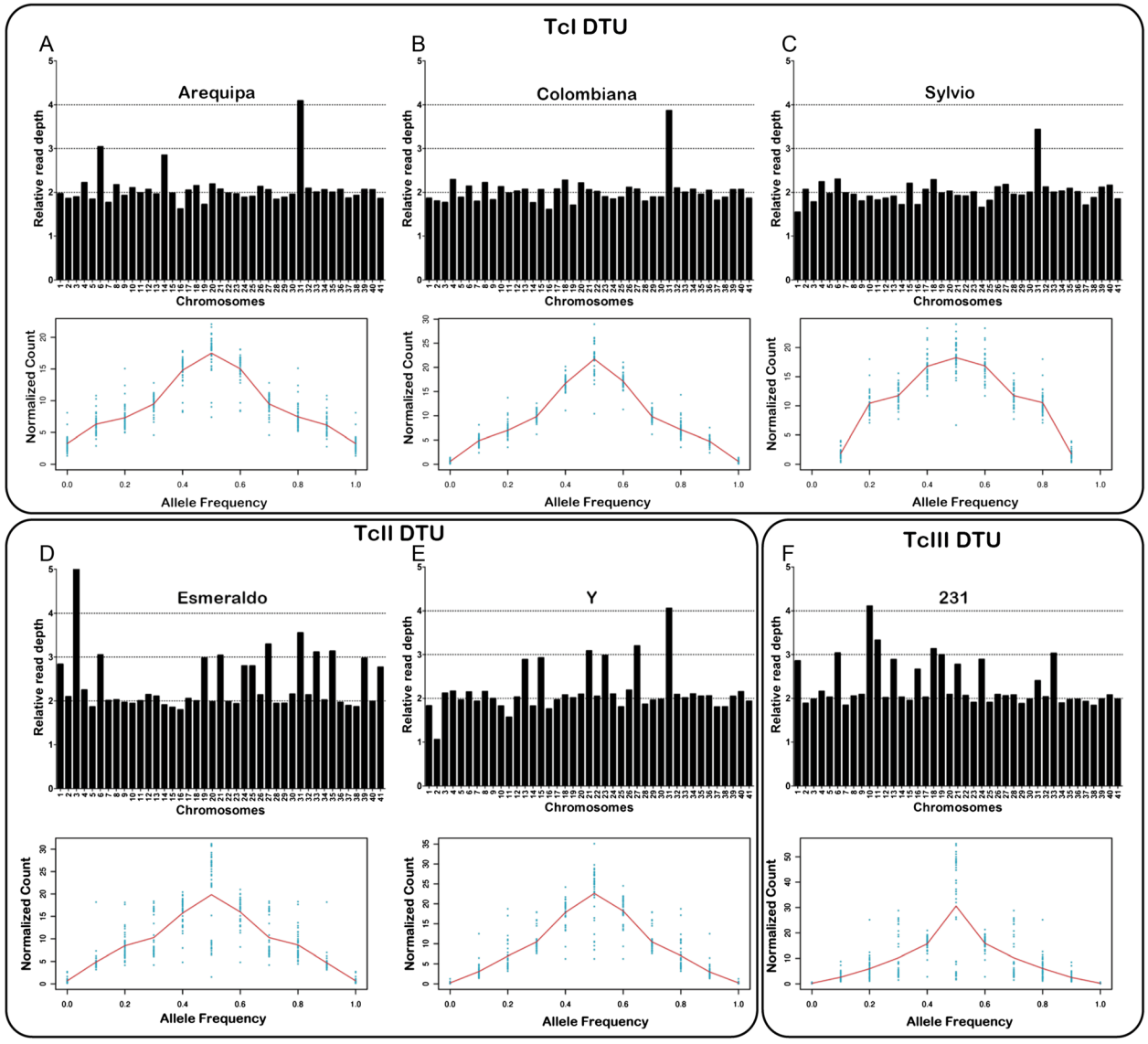


Fig. 3 Predicted ploidy of the six *T. cruzi* strains. The predicted ploidy of each chromosome from the *T. cruzi* strains Arequipa (**a**), Colombiana (**b**) and Sylvio (**c**) from the TcI DTU; Y (**d**) and Esmeraldo (**e**) from the TcII DTU; and 231 (**f**) from the TcIII DTU, using as a reference the 41 CL Brener chromosome sequences, was estimated with the SCoPE approach. Each black bar corresponds to the ratio between the single-copy gene mean RDC in the chromosome and the single-copy gene mean RDC of the whole genome, representing its predicted chromosome copy number. The overall ploidy of each strain was estimated by the proportion of the alleles in heterozygous SNP positions, where a tendency of 0.4-0.5 represents a diploid chromosome and a proportion of 0.3 and/or 0.7 corresponds to a triploid chromosome (red line)

the exception of the Arequipa strain, which also has a trisomy in chromosomes 6 and 14. The strains from the TcII DTU have a more plastic karyotype with several predicted supernumerary chromosomes and a monosomy case. The Esmeraldo strain has chromosome 3 as pentasomic; chromosomes 27, 31, 33 and 35 range from trisomic to tetrasomic; chromosomes 6, 19, 21 and 39 as trisomic; and chromosomes 1, 24, 25 and 41 range from disomic to trisomic. The Y strain, also from TcII DTU, has tetrasomy of chromosome 31; chromosomes 21 and 27 range from trisomic to tetrasomic; trisomy of chromosomes 13, 15, and 23; and monosomy of chromosome 2. The representative of TcIII DTU, 231, has tetrasomy of chromosome 10; chromosomes 11 and 18 range from trisomic to tetrasomic; trisomy of chromosomes 6 and 19; and chromosomes 1, 13, 16, 21, 24 and 31 have a ploidy ranging from disomic to trisomic (Fig. 3).

To further confirm the chromosomal ploidy, the distribution of base frequencies between the heterozygous SNP positions among all the CDSs in the 41 chromosomes of the six *T. cruzi* strains were estimated (Additional file 3: Figure S1). This analysis is in agreement with the CCNV

results predicted by the SCoPE methodology. The only exceptions were chromosomes 20 and 23 from Sylvio and chromosome 7 from Esmeraldo, which were predicted as tetraploid by the heterologous SNP proportion and as diploid in the SCoPE analysis, and chromosomes 6 and 14 from Arequipa that were tetraploid by the heterologous SNP proportion and triploid in the SCoPE analysis.

To evaluate if these predicted aneuploidies were produced by the gain or loss of a whole chromosome, or if they result from segmental duplication or loss of partial fragments from these chromosomes, the normalized read depth coverage of each position along each chromosome of the six *T. cruzi* strains was estimated (Additional file 4: Figure S2). Figure 4 represents the read depth coverage along each position of the predicted disomic, trisomic and tetrasomic chromosomes of the Y strain, as well as the base frequency distribution between the heterozygous SNP positions. As expected, with the exception of the regions that are rich in multigene families and gaps, the predicted ploidy along the entire chromosome is in agreement with the predicted ploidy based on the SCoPE and SNP analyses. This finding suggests that these aneuploidies are probably a result of a whole chromosomal duplication/loss.

To determine whether the ploidy profile of the six *T. cruzi* strains is in agreement with their phylogeny, a hierarchical clustering analysis was performed, using as input the predicted ploidy of each chromosome of each strain based on the SCoPE approach. The clustering analysis was based on their pairwise Euclidean distances using the Complete Linkage method (Fig. 5). In this analysis, all strains belonging to the TcI DTU clustered together, with Colombiana and Sylvio strains being closer to each other than to the Arequipa strain. This is likely

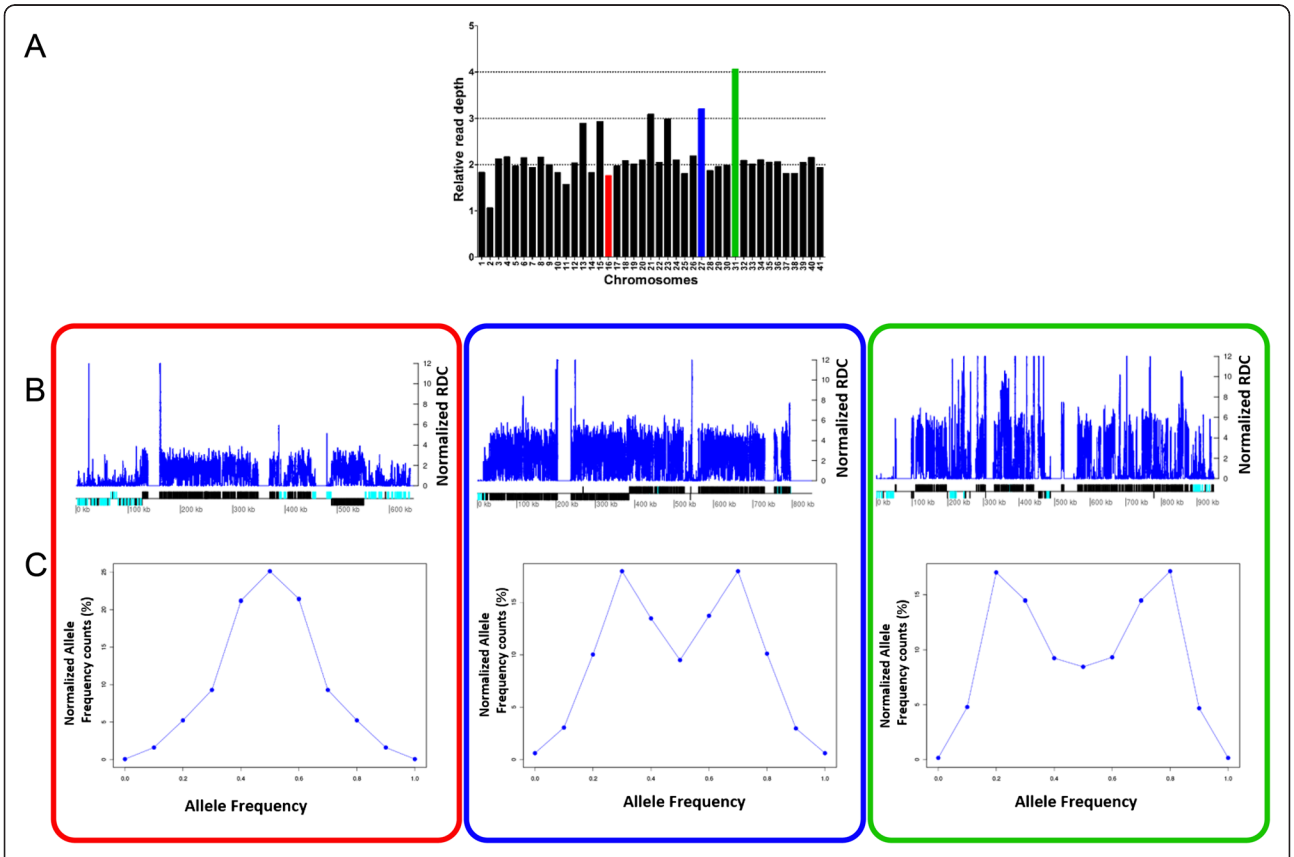


Fig. 4 Correspondence between the ploidy predicted by the SCoPE approach and the normalized read depth coverage (RDC) along the entire chromosome sequences. The correspondence between the chromosomal ploidy and the normalized read depth coverage of chromosomes 16 (red box), 27 (blue box) and 31 (green box) of the *T. cruzi* Y strain is shown. **a** Chromosomal ploidy predicted by the SCoPE approach. **b** The blue line corresponds to the normalized RDC of each position, estimated by the ratio between the RDC and the genome coverage. Below, the protein-coding genes are depicted as rectangles drawn as proportional to their length, and their coding strand is indicated by their position above (top strand) or below (bottom strand) the central line. Cyan and black rectangles represent multigene families and hypothetical/housekeeping genes, respectively. Gaps are represented by gene-less regions with no read coverage. Chromosome 16 was predicted as disomic, 27 as trisomic and 31 as tetrasomic. The regions of low coverage correspond to regions rich in multigene families or chromosomal gap regions. **c** The predicted ploidy based on the proportion of the alleles in the heterozygous SNP positions for chromosomes 16, 27 and 31 is shown. The peak of 0.5 classifies chromosome 16 as diploid, the peaks of 0.3 and 0.6 classify chromosome 27 as triploid, and the peaks of 0.2-0.8 and 0.5 classify chromosome 31 as tetraploid

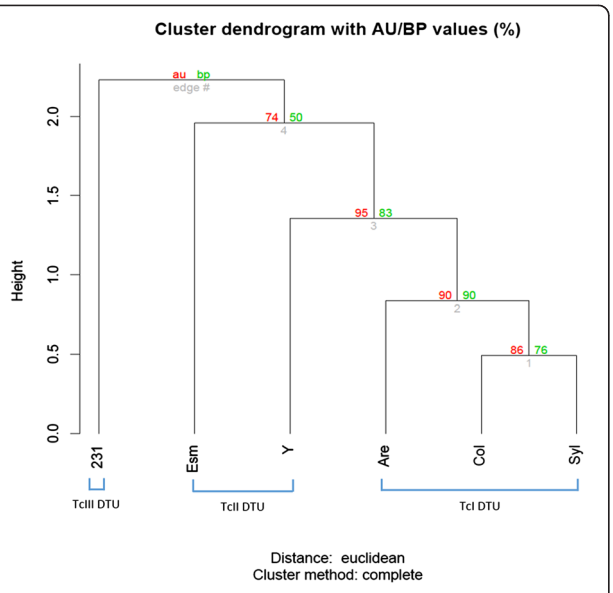


Fig. 5 Dendrogram of the hierarchical analysis of the *T. cruzi* strains' predicted ploidy. The hierarchical clustering analysis based on Euclidian distances of the predicted ploidy of each chromosome from the *T. cruzi* Arequipa, Colombiana, Sylvio, Esmeraldo, Y and 231 strains was performed using the Pvclust package and the R software platform. Two bootstrap resampling methods were used to assess the uncertainty in the hierarchical cluster analysis: the approximately unbiased (au) in red, and the bootstrap probability (bp) in green

due to the presence of exclusive aneuploidies in chromosomes 6 and 14 from Arequipa. Both TcII DTU strains showed a different pattern of chromosomal aneuploidies from each other and showed a smaller group consistency than the strains from the TcI DTU. The 231 TcIII DTU strain shows the most different pattern among the *T. cruzi* strains evaluated.

Y Chromosome 11

Chromosome 11 from the Y strain has some interesting features. As shown in Fig. 3, this chromosome displayed a smaller ploidy than disomic but a greater ploidy than monosomic. This led us to investigate whether this pattern occurs due to the loss of a chromosomal region instead of the loss of a chromosome copy. As showed in Fig. 6, there is a drastic change in the RDC starting at the 248-kb position in this chromosome that corresponds to a strand switch region. The first 248 kb of this chromosome has a smaller predicted ploidy when compared to the rest of the chromosome sequence (Fig. 6a). To further confirm this chromosomal region loss, the mean RDC of the single-copy genes located upstream and downstream of the 248-kb coordinate were estimated. The single-copy genes that were upstream of the 248-kb position had a non-normalized mean RDC of 20, while the single-copy genes downstream had an approximate mean RDC of 40 (Fig. 6b). This reduction of the RDC in the 5′ region when compared to the rest of the chromosome and the estimated genome coverage of 34× suggest a segmental loss of the 248 kb at the 5′ region in one copy of this chromosome in the Y strain. Alternatively, this pattern may be due to differences in the structure of chromosome 11 among the different cells of the parasite population because Y is not a cloned line.

Chromosome 31 gene ontology

From all the 41 CL Brener chromosomes, chromosome 31 was the only one that was supernumerary in all the strains analyzed, in both SCoPE (Fig. 7a) and heterozygous SNP analyses (Fig. 7b). To identify gene functions that were overrepresented in this chromosome when compared to the whole genome, a gene ontology analysis of both Esmeraldo-like and non-Esmeraldo-like CL Brener chromosome 31 was performed (Fig. 7c, Additional file 5: Table S3). This analysis shows that this chromosome is enriched in genes involved in glycosylation and glycoprotein biosynthetic processes in both CL Brener haplotypes.

Discussion

The availability of the sequence representation of the 41 putative chromosomes from both Esmeraldo-like (TcII) and non-Esmeraldo (TcIII) CL Brener haplotypes [33] allowed a comparative analysis of *T. cruzi* strain ploidy using the unassembled next generation sequencing reads. However, the repetitive nature of the *T. cruzi* genome along with the presence of gap regions in the sequence representation of CL Brener chromosomes hampers the effectiveness of CCNV comparisons based on the read depth coverage (RDC) of whole-chromosome sequences. To overcome these limitations, we propose the SCoPE approach that is based on the RDC of single-copy genes that are conserved between Esmeraldo-like and non-Esmeraldo-like haplotypes as well as in the six *T. cruzi* strains evaluated in this study (Additional file 2: Table S2). These single-copy genes were used as chromosomal markers for unique genomic sequences to normalize the CCNV estimations, revealing large differences in chromosomal copy number between *T. cruzi* strains (Fig. 3).

It is well known that *T. cruzi* strains present a distinct profile of chromosomal bands, which present different sizes and numbers in Pulse Field Gel Electrophoresis, suggesting a variable karyotype among the DTUs. These differences were mainly attributed to a differential repetitive content in the genomes of distinct *T. cruzi* strains, or to chromosomal fusion/break events during the parasite evolution [5, 22, 30, 31, 35]. In fact, approximately 50 % of the CL Brener genome corresponds to repetitive sequences, many of them clustered into regions containing multigene families encoding surface proteins and transposable elements [20, 33]. These clusters are extremely

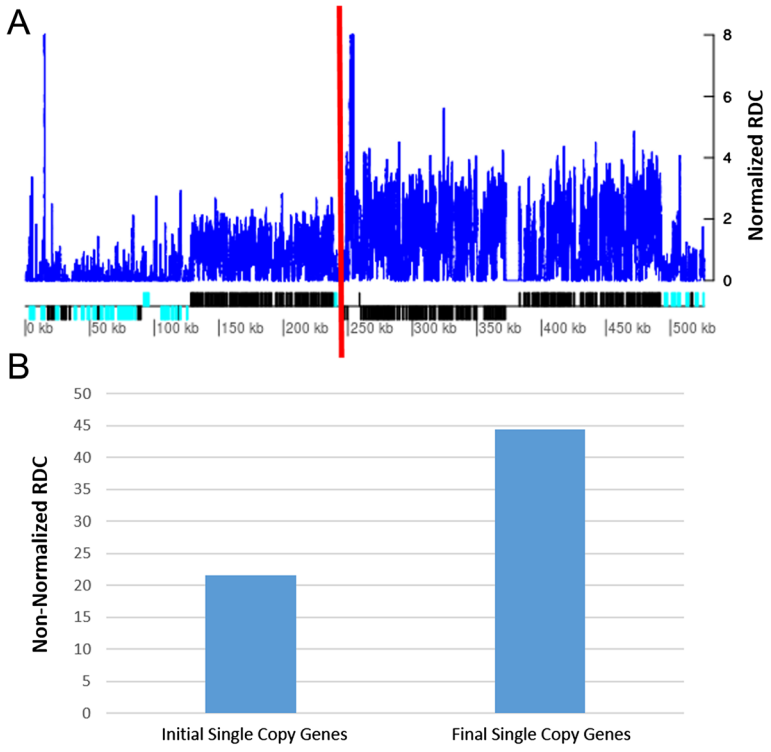


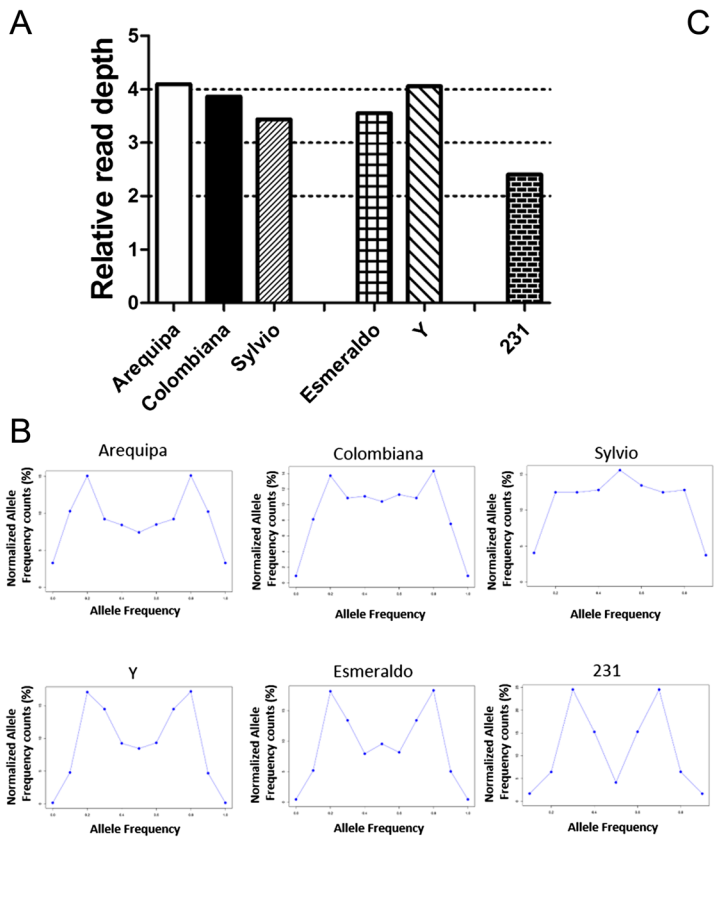
Fig. 6 Predicted ploidy of chromosome 11 of the *T. cruzi* Y strain. The blue line corresponds to the normalized RDC of each position of chromosome 11, estimated by the ratio between the RDC and the genome coverage. Below, the protein-coding genes are depicted as rectangles drawn in proportion to their length, and their coding strand is indicated by their position above (top strand) or below (bottom strand) the central line. Black and cyan rectangles represent the housekeeping/hypothetical and multicopy gene families, respectively. Gaps are represented by gene-less regions with no read coverage. The red line corresponds to the 248-kb position in this chromosome, which separates the regions of low RDC and high RDC (**a**). The non-normalized mean RDC estimated based on the SCoPE approach of the genes that were upstream and downstream of the 248-kb position was estimated (**b**)

variable in their gene content and length and are regions of synteny loss when comparing the two CL Brener haplotypes (El-Sayed 2005a) and the CL Brener and Sylvio genomes [21]. Despite these variations, single-copy genes are, in general, highly conserved and syntenic among the *T. cruzi* strains [8, 21, 22, 36] (Baptista in preparation) and therefore represent an adequate source for sequence normalization in comparative CCNV analyses.

The occurrence of whole chromosome aneuploidy in *T. cruzi* was previously suggested based on whole genome oligonucleotide tiling arrays and competitive genomic hybridization between 16 *T. cruzi* strains and CL Brener clone as reference [8]. However, the absolute number of individual chromosomes in each strain was not estimated [8]. Also, the comparative genomic hybridization analysis does not allow the detection of chromosomal expansion/deletions that are present simultaneously in CL Brener and in the tested strain [8]. Our approach, however, used the haploid content of CL Brener chromosomes as reference for read mapping, allowing the detection of chromosomal copy number variations that may be also present in CL Brener.

Competitive mapping

To identify the CL Brener haplotype suitable to be used as a reference, each *T. cruzi* strain read collection from the TcI (Arequipa, Colombiana and Sylvio), TcII (Esmeraldo and Y) and TcIII (231) strains was simultaneously mapped to both Esmeraldo-like and non-Esmeraldo-like CL Brener haplotypes, using a competitive approach (Fig. 1a). As the mapping quality cutoff was very stringent, the conserved regions that mapped simultaneously to both CL Brener haplotypes were excluded, and only reads that mapped preferentially to one haplotype were considered. As expected, reads from the TcII strains, namely Esmeraldo and Y, mapped better to the Esmeraldo-like haplotype (CL Brener haplotype derived from a TcII ancestor genome), and reads from TcIII mapped better to the non-Esmeraldo haplotype (derived from the TcIII ancestor) (Fig. 1b). All strains belonging to the TcI DTU—Arequipa, Colombiana and Sylvio—mapped slightly better to the non-Esmeraldo haplotype, suggesting a closer proximity to TcI compared to TcIII as previously described [11, 14, 15, 21]. Interestingly, all TcI strains had a higher RDC for the Esmeraldo-like chromosome 14 than for the same chromosome from



C

Non-Esmo - Molecular

GO-ID	Description	p-value
8378	Galactosyltransferase activity	6,55E-06
16758	Transferase activity, transferring hexosyl groups	3,41E-04
16298	Lipase activity	1,86E-03
4806	Triglyceride lipase activity	1,86E-03
16757	Transferase activity, transferring glycosyl groups	5,30E-03
3824	Catalytic activity	1,64E-02
52689	Carboxylic ester hydrolase activity	3,00E-02
4198	Calcium-dependent cysteine-type endopeptidase activity	4,13E-02

Non-Esmo - Biological

GO-ID	Description	p-value
43413	Macromolecule glycosylation	3,23E-05
6486	Protein glycosylation	3,23E-05
9101	Glycoprotein biosynthetic process	3,23E-05
9100	Glycoprotein metabolic process	3,23E-05
70085	Glycosylation	7,22E-05
6629	Lipid metabolic process	3,78E-03
1901137	Carbohydrate derivative biosynthetic process	1,96E-02
2643	Single-organism carbohydrate metabolic process	3,64E-02
46394	Carboxylic acid biosynthetic process	3,64E-02
16053	Organic acid biosynthetic process	3,64E-02
6633	Fatty acid biosynthetic process	3,64E-02
72330	Monocarboxylic acid biosynthetic process	3,64E-02

Esmo - Molecular

GO-ID	Description	p-value
8378	Galactosyltransferase activity	1,77E-06
16758	Transferase activity, transferring hexosyl groups	4,78E-05
4806	Triglyceride lipase activity	4,86E-04
16757	Transferase activity, transferring glycosyl groups	5,73E-04
16298	Lipase activity	5,73E-04
4198	Calcium-dependent cysteine-type endopeptidase activity	8,18E-03
8777	Acetylornithine deacetylase activity	1,01E-02
50897	Cobalt ion binding	1,01E-02
52689	Carboxylic ester hydrolase activity	1,16E-02
4197	Cysteine-type endopeptidase activity	1,93E-02
19213	Deacetylase activity	2,16E-02
4148	Dihydrolipoyl dehydrogenase activity	3,90E-02
8234	Cysteine-type peptidase activity	3,90E-02

Esmo - Biological

GO-ID	Description	p-value
43413	Macromolecule glycosylation	6,24E-06
6486	Protein glycosylation	6,24E-06
9101	Glycoprotein biosynthetic process	6,24E-06
9100	Glycoprotein metabolic process	6,24E-06
70085	Glycosylation	7,07E-06
1901137	Carbohydrate derivative biosynthetic process	5,34E-03
44723	Single-organism carbohydrate metabolic process	9,92E-03
6526	Arginine biosynthetic process	1,86E-02
5975	Carbohydrate metabolic process	3,80E-02
6525	Arginine metabolic process	4,40E-02

Fig. 7 Ploidy and gene ontology analysis of chromosome 31. From the 41 CL Brener chromosomes, chromosome 31 was the only one that was supernumerary in all the studied strains based on the SCoPE analysis (**a**). The ploidy of chromosome 31 from each strain, estimated by the ratio of the alleles in the heterozygous SNP positions, classifies Arequipa, Colombiana, Sylvio, Esmeraldo and Y as tetraploid and 231 as triploid (**b**). Gene ontology analysis shows that this chromosome is enriched in genes involved in the glycosylation and glycoprotein biosynthetic processes (**c**)

the non-Esmeraldo haplotype. Another example is the Esmeraldo chromosome 3, which had a slightly better RDC for the non-Esmeraldo than for the Esmeraldo-like. A careful inspection of the RDC along the entire sequences of these chromosomes revealed that this unexpected profile was due to an extremely high RDC of a specific region in both chromosomes. In the case of chromosome 14, the genomic region encompassing the gene Tc00.10470535007099.80 (ABC transporter putative) in the Esmeraldo-like haplotype displays a very high RDC, and this region is absent in the non-Esmeraldo-like haplotype. Therefore, it is likely that this is a repetitive sequence where reads from both CL Brener haplotypes were collapsed in the assembled Esmeraldo-like haplotype. Likewise, in the case of chromosome 3, approximately 1500 nucleotides of the 5′ end of the gene Tc00.1047053508533.10 (hypothetical protein conserved) from the non-Esmeraldo-like haplotype are missing in the corresponding allele (Tc00.1047053404001.20) of the Esmeraldo-like haplotype. These genes are not included in the 1563 copy genes used by SCoPE to estimate the CCNV, further validating this methodology.

Chromosomal copy number variation and strain ploidy

All six *T. cruzi* strains evaluated in this study had an overall genome ploidy predicted as diploid, based on both the SCoPE approach and the allele proportion in heterozygous positions, when the whole genome was evaluated (Fig. 3). This result is in agreement with previous estimations for *T. cruzi* [8, 9, 37] and *Leishmania* species that classified these parasites as diploid, with the exception of *L. braziliensis* (M2904) that had an overall trisomic chromosomal pattern [34].

The CCNV among *T. cruzi* strains was initially estimated by the SCoPE approach (Fig. 3). The strains from the TcI DTU (Colombiana, Sylvio and Arequipa) had a low number of whole chromosome expansions (1, 1 and 3, respectively), while the TcII strains (Esmeraldo and Y)

and the TcIII strain (231) had a larger number of these expansions (13, 6 and 12, respectively). These predictions were further confirmed by the heterologous SNP analysis, which predicted the same aneuploidies, with the exception of chromosomes 20 and 23 from Sylvio, 7 from Esmeraldo and 6 and 14 from Arequipa that were predicted as tetrasomic. In these chromosomes, the heterozygous SNP analysis predicted an even broader chromosomal expansion than those estimated by the SCoPE approach. To evaluate if the CCNV predicted using the single-copy genes was not impacted by segmental duplications or loss of partial fragments from these chromosomes, we estimated the RDC of all the positions from all 41 chromosomes of the six *T. cruzi* strains (Fig. 4, Additional file 4: Figure S2). Excluding chromosomal gaps and regions containing clusters of multigene families, the RDC of the whole chromosome is in agreement with the SCoPE CCNV prediction, validating this approach (Fig. 4).

Gene copy number variation is well documented as an important mechanism to enhance gene expression and variability not only in *T. cruzi*, *but also in T. brucei and Leishmania* [8, 20, 28, 36, 38–40]. The CCNV arises as a new level of CNV, expanding whole blocks of genes simultaneously [34]. The occurrence of CCNV was already associated to increased fitness in stress conditions and to drug resistance in *Saccharomyces cerevisiae* and *Candida albicans* [41–43]. The karyotype plasticity and CNV found in *T. cruzi* could also represent a framework for the natural selection of favorable phenotypes, such as higher expression of virulence-factors and increased diversity, resulting in an enhanced adaptability of the parasite [8, 44].

The mechanisms involved in the generation of CCNV in *T. cruzi* are still unknown. Even though it has been previously proposed that meiosis events occur in *Trypanosoma brucei* [45] and *Leishmania* [46], this process has not been demonstrated in *T. cruzi*. However, a non-meiotic hybridization model has been proposed to explain the generation of hybrid *T. cruzi* DTUs [47]. According to this model, during the mammalian stage of the parasite, the nucleus of two diploid cells fuse, resulting in a polyploidy progeny that can undergo recombination between alleles. This new polyploid cell may lose some of its supernumerary chromosomal copies, eventually returning to the diploid state [48]. This could be one of the mechanisms related to the generation of CCNV in *T. cruzi*, as some of these extra chromosomal copies may be maintained by the parasite.

For all the *T. cruzi* DTUs evaluated, TcI had the fewest number of chromosomal expansions, exhibiting the most stable karyotype (Fig. 3). The reduced number of chromosomal duplications could be a contributing factor for the low levels of heterozygosity found in TcI [21, 49]. It has been demonstrated that the mismatch repair machinery, which repairs base miss-incorporation and erroneous insertions and deletions during DNA recombination and replication, is more efficient in TcI than in the other DTUs and is a known factor in the reduction of heterozygosity in TcI [50, 51]. TcI strains also have the smallest genome size compared to the other *T. cruzi* DTUs [6], which were associated with the reduction of multigene family clusters within this DTU [21]. Based on our results, we propose that this reduction in genome size is also associated with the lower number of chromosomal duplications in TcI when compared to other DTUs. Similarly, data from Rogers 2011 showed a correlation between CCNV based on RDC analysis and DNA content estimated using flow cytometry in different *Leishmania* species [34]. Variation in DNA content among *T. brucei* species and isolates was also observed [52] and could be related with some level of CCNV in this parasite, which so far has not been formally demonstrated. Although TcI strains have less intragenomic heterogeneity, sequences belonging to different strains are more distant from each other within TcI than within TcII and TcVI [8, 23]. This increased variation suggests a reduced number of genetic exchange events among strains from the TcI DTU, which could result in a lower rate of chromosomal ploidy variations.

For the TcII DTU, Esmeraldo had 13 chromosomal expansions while Y had 6, suggesting that there are extensive differences in chromosomal duplications within TcII DTU (Fig. 3). Recently, substantial recombination and genetic exchange among strains from the TcII DTU that coexist in the same geographical area was proposed based on microsatellite genotype data [19]. The broader chromosomal expansion in Esmeraldo may be explained by the fact that haplotypes that constitute this strain are more distant from each other than the ones that constitute Y [15], which suggests that Esmeraldo suffered more recombination events than Y, making it more susceptible to acquiring aneuploidies [15, 48]. Leishmania isolates also display a broad range of CCNV. As TcI DTU strains evaluated in this study, different strains of *L. major* (Friedlin and LV39) had the same CCNV pattern. On the other hand, as observed in TcII DTUs, strains from both *L. mexicana* and *L. donovani* had extensive CCNV within the species [34]. It will be interesting to compare the efficiency of processes associated with the maintenance of genomic stability in *T. cruzi* DTUs and *Leishmania* species, and investigate the occurrence of CCNV in *T. brucei*, which could help to elucidate the mechanisms behind the CCNV in these parasites. Widespread aneuploidy found in *T. cruzi* and *Leishmania*, implies that caution should be taken when selecting markers for population genetic studies based on the hypothesis of diploidy [13]. In this case, it would be imperative selecting markers from genomic regions known to be diploid.

Hierarchical clustering analysis based on the predicted ploidy of each *T. cruzi* chromosome clustered all TcI strains together with high confidence scores, further confirming the genome structural stability within DTUI (Fig. 5). Aside from the TcI DTU, strains from the TcII and TcIII DTUs had a variable CCNV pattern, suggesting a higher genome plasticity between and within these DTUs. It is interesting that the two strains from the TcII DTU (Esmeraldo and Y) had a different pattern of chromosomal ploidy, suggesting that chromosomal expansions are highly variable and may have originated several times during the evolution of the TcII DTU. The analysis of a broader number of strains would be required to correctly estimate the ratio of CCNV within and between the *T. cruzi* DTUs.

Y chromosome 11

In the first 248 kb of the Y chromosome 11, we identified 22 single-copy genes, which had half the mean RDC compared with the remaining chromosome sequence, which contains 15 single-copy genes (Fig. 6). This large change in RDC starts in a strand switch region, which is frequently associated with rearrangements in Trypanosomatid genomes [36, 53]. This finding suggests that Y chromosome 11 could have an arm loss. Another possibility is that the CL Brener chromosome 11 is divided into two distinct chromosomes in Y, chromosomes 11a and 11b, at the 248-kb position, where 11a is haploid and 11b is diploid. Events of chromosomal break or fusion may explain the variable band pattern in PFGE among the *T. cruzi* strains [5, 22, 30–32, 35]. These events are easily detected in RDC analysis when there are large aneuploidies between fragments of the same chromosome. Alternatively, because Y is not a cloned population, this result may represent a mosaic structure of the parasite population, where some cells may have the entire chromosome 11 sequence while other cells may have an arm loss.

Chromosome 31

Chromosome 31 was the only one that was supernumerary in all six *T. cruzi* samples evaluated in this study (Fig. 7). Gene ontology analysis showed that this chromosome has an enhanced number of genes related to glycoprotein biosynthesis and glycosylation processes (Fig. 7c). CL Brener has approximately 100 genes annotated as putative glycosyltransferase genes, which are involved in the synthesis of a variety of glycoconjugates and are abundantly and differentially expressed in all *T. cruzi* stages [54]. From these 100 genes, 54 are UDP-GlcNAc-dependent glycosyltransferases. Chromosome 31 has 9 of the 27 UDP-GlcNAc-dependent glycosyltransferase gene copies in the CL Brener Esmeraldo-like haplotype and 13 of the 27 copies in the non-Esmeraldo haplotype. This enzyme is involved in the transfer of N-acetylglucosamine (GlcNAc) from the UDP-GlcNAc precursor to the hydroxyl group of serine and threonine residues, resulting in O-linked oligosaccharides in *T. cruzi* mucins [55]. Comparative genomic hybridization among *T. cruzi* strains also shows gene CNV in another gene involved in the synthesis of glycans on *T. cruzi* mucins, the beta-galactofuranosyl transferase genes [8, 56]. Mucins are heavy glycosylated glycoproteins, and their glycan content may account for up to 60 % of the total mucin weight [55, 57]. These proteins are the most abundant component on the *T. cruzi* surface, covering the whole parasite with approximately 2×10^6 copies per cell [39, 55]. Mucins are responsible for protecting the parasite from both the vector and the mammal defensive mechanisms and ensure the anchorage point and invasion of specific cells and tissues [55]. One of the forces driving the expansion of chromosome 31 in all the *T. cruzi* strains may be the need to glycosylate this large number of proteins that cover the parasite surface and are directly involved in parasite survival in both invertebrate and vertebrate hosts. Although chromosome 31 was also expanded in several *Leishmania* species [34], we found no large syntenic regions between *T. cruzi* and *Leishmania* chromosome 31, suggesting that the chromosome 31 expansion in *Leishmania* is driven by different evolutionary pressures.

Conclusions

It is well known that *T. cruzi*, as well as *Leishmania* and *T. brucei*, relies on gene duplication to increase the expression levels of key genes and to allow the generation of novel genes without loss of function [8, 20, 28, 36, 38–40]. Our study highlights the genome-wide CCNV in *T. cruzi* as a new level of gene expansion mechanism, allowing the rapid generation of diversity within the parasite. The estimation of chromosomal aneuploidies based on the RDC of single-copy genes comes as a new approach to evaluate the CCNV in *T. cruzi*, reducing the bias of repetitive and gap regions in the analysis and improving chromosomal comparisons between DTUs. As previously observed in *Leishmania* [34], aneuploidy appears to be well tolerated in trypanosomatids, due to their predominantly asexual replication mechanism. The chromosome copy number can vary considerably between strains from different *T. cruzi* DTUs and even within the same DTU. TcI appears to be more stable, and TcII had large differences between its strains, suggesting that this mechanism is widely used by the parasite to expand groups of genes. One of the limitations of our approach is that we are not able to investigate the karyotype structure of each strain, and the analysis is limited to comparing their differences based on the CL Brener predicted chromosomes because there are no other *T. cruzi* chromosomal sequences already

published in the literature. Due to the extensive repetitive content, third-generation long read single-molecule sequencing is required not only to close the gaps in CL Brener chromosomes, but also to generate reliable chromosomal sequences of the other five DTUs allowing better CCNV estimations. Only three DTUs (TcI, II and III) were evaluated in this work. Therefore, it would be interesting to evaluate CCNV in the other *T. cruzi* DTUs, to better investigate this variation in the parasite. Finally, the expansion of chromosome 31, which is enriched with genes related to glycosylation pathways in all six strains evaluated, highlights the importance of this biochemical process to the parasite's survival.

Methods

Parasite cloning in a semi-solid medium

For cloning the *T. cruzi* Arequipa (TcI) and 231 (TcIII) strains, 10^3 epimastigotes were plated into a semi-solid medium (low-melting agarose 0.75 %, brain heart infusion 48.4 %, liver infusion tryptose (LIT) 48.4 %, 2.5 % defibrinated blood, and 250 μg/mL penicillin/streptomycin) and incubated at 28 °C for 35 days. Single clones were obtained and transferred to 25-cm^3 culture flasks with 5 mL of LIT medium and 10 % fetal bovine serum.

Parasite culture and DNA isolation

T. cruzi epimastigotes from Arequipa (TcI), Colombiana (TcI) and Y (TcII) strains were cultured in LIT medium supplemented with 10 % fetal bovine serum. A total of 1×10^8 parasites from each strain were centrifuged at 3000 g in an Eppendorf 5804 Centrifuge. The parasites where washed three times with ice-cold PBS, suspended in PBS with 100 μg/mL proteinase K and incubated at 25 °C for 10 min. The genomic DNA was obtained with the Wizard® Genomic DNA Purification Kit (Promega) by following the manufacturer instructions. The DNA integrity was evaluated by agarose gel electrophoresis. The DNA samples were submitted to a genotyping protocol according to Souto et al., 1996 [58], de-Freitas et al., 2006 [15] and Burgos et al., 2007 [59].

Genome sequencing

A whole-genome shotgun library (WGSG) and sequencing of the *T. cruzi* Arequipa (TcI), Colombiana (TcI) and Y (TcII) strains were performed at the Computational Genomics Unity Darcy Fontoura de Almeida (UGCDFA) of the National Laboratory of Scientific Computation (LNCC) (Petrópolis, RJ, Brazil). For the 454 GS-FLX Titanium sequencing, each unpaired library was constructed using 5 μg of genomic DNA (gDNA) and by following the GS FLX Titanium series protocols. All titrations, emulsions, PCR, and sequencing steps were carried out according to the manufacturer's protocol. One full PicoTiterPlate (PTP) was used for sequencing each library. In addition to the 454 sequencing, the Ion Proton™ was also used for unpaired sequencing. A total of 1 μg of gDNA was used to prepare the deep sequencing libraries. All steps were also performed according to manufacturer's protocol. Reads from Sylvio (TcI) were kindly provided by Dr. Bjorn Andersson (Karolinska Institut). The 231 (TcIII) sequences were obtained using the Illumina Hiseq 2000 NGS platform (Baptista et al., in preparation). The Illumina and 454 read libraries from the Esmeraldo strain were downloaded from the National Center for Biotechnology Information (NCBI) (Additional file 6: Table S4).

Preprocessing of reads

The reads were checked for quality using the FASTQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads smaller than 30 nt and with a Phred score lower than 20 [60, 61] were removed from the libraries using the fast_quality_trimmer, from the fastx toolkit ("http://hannonlab.cshl.edu/fastx_toolkit/index.html).

Mapping and competitive mapping

Whole genome shotgun reads from the *T. cruzi* Arequipa, Colombiana, Sylvio, Esmeraldo, Y, and 231 strains were mapped to the 41 chromosomes from both Esmeraldo-like and non-Esmeraldo-like CL Brener haplotypes, version 6, downloaded from Tritrypdb (http://tritrypdb.org/tritrypdb/). Mapping for each read library was performed simultaneously for both CL Brener haplotypes for competitive mapping, or separately for single mapping, using Bowtie 2 [62]. To account for the divergence between strains, the Bowtie2 preset "very sensitive" parameter was used, with the mismatch parameter changed to 1. The competitive mapping was used to select the CL Brener haplotype closely related to each read library, and single mapping was used to estimate the CCNV in each strain. Mapped reads of each strain were filtered using a mapping quality threshold of 30 using SAMtools v1.1 [63]. The read depth coverage (RDC) for each position of each chromosome of the Esmeraldo-like and non-Esmeraldo-like haplotypes for each *T. cruzi* strain was obtained by an in-house PERL script and BEDtools genomecov v2.16.2 [64]. The competitive mapping graphs were generated using GraphPad Prism V5.01 and scripts developed in PERL and R.

Single-copy genes and chromosomal ploidy

Ortholog genes between Esmeraldo-like and non-Esmeraldo-like CL Brener haplotypes were identified using OrthoMCL v2.0 [65], based on a combined approach of "reciprocal best hits" and a "Markov clustering algorithm" (MCL). Initially, an "all vs. all" local alignment using the BLASTp package 2.2.21 with an E-value of 1e-5 as a cut-off was performed. The E-values were

converted into log base to create the similarity matrix. An MCL with a 1.5 inflation parameter was applied to produce the ortholog clusters. A total of 1563 1:1 orthologs were selected and assumed to be single-copy genes in the haploid CL Brener genome (Additional file 1: Table S1). These single copy genes were selected to be used as chromosomal markers for unique genomic sequences, allowing CCNV estimations without the bias of repetitive regions. The mean RDC of the single-copy genes in each of the 41 CL Brener chromosomes, based on the mapped reads from the Arequipa, Colombiana, Sylvio and 231 strains to the non-Esmeraldo-like haplotype, and the Esmeraldo and Y to the Esmeraldo-like haplotype, were generated by PERL scripts. The mean RDC of all single-copy genes in all chromosomes of each strain was assumed to be the genome coverage. The predicted copy number of each chromosome was determined based on the mean RDC of the single-copy genes in a given chromosome and normalized by the genome coverage. The genome coverage was estimated as 47× for Arequipa, 28× for Colombiana, 9× for Sylvio, 52× for Esmeraldo, 34× for Y and 76× for 231. The CCNV graphs were generated with GraphPad Prism V5.01 software and scripts in PERL and R.

SNP content

Single-nucleotide polymorphisms (SNPs) of the mapped reads from the single-copy genes in the *T. cruzi* Arequipa, Colombiana, Sylvio, Esmeraldo, Y and 231 strains to the CL Brener Esmeraldo-like and non-Esmeraldo-like haplotypes were obtained using the SAMtools function mpileup [63]. To reduce the chance of incorrectly identifying a SNP due to sequencing artifacts, we set the minimum number of mapped reads to 10. To reduce the bias of collapsed regions, the maximum number of reads mapped in a SNP position was set to double the genome coverage of the corresponding genome. The SNP density for each strain was calculated and plotted using GraphPad Prism V5.01.

Heterozygous SNPS

Heterozygous SNPs between the CL Brener chromosome and the mapped reads for the six *T. cruzi* stains were obtained from the filtered SAMtools mpileup results [63]. To be considered as a reliable SNP, the position RDC must be at least 10. To reduce the bias of collapsed regions, the position RDC must also be lower than twice the genome coverage. For each chromosome, the proportion of the alleles in each predicted heterozygous site was obtained and rounded to the first decimal place. Base frequencies were rounded in ten categories, ranging from 0.1 to 1, and an approximate distribution of base frequencies for each chromosome was plotted in R. To estimate the overall ploidy of each genome, the same methodology was applied, but the heterozygous positions from all CDSs from all chromosomes were employed simultaneously.

Cluster dendrogram

A hierarchical clustering analysis of all predicted *T. cruzi* chromosomal ploidy was performed using the Pvclust package [66] implemented in R (www.r-project.org) (R Development 2010). First, a distance matrix was built with pairwise Euclidean distances between the strains and the dendrogram was generated by the complete linkage method. To assess the uncertainty in hierarchical cluster analysis, we used the two bootstrap resampling methods implemented in Pvclust: bootstrap probability (BP) by ordinary bootstrap resampling and the approximately unbiased (AU) probability from multiscale bootstrap resampling, which provides better estimations than BP. Both methods were calculated with 10,000 iterations.

Gene ontology

Gene ontology categories that were significantly overrepresented in the genes of the CL Brener chromosome 31 were detected using the hypergeometric distribution analysis in BiNGO [67] with Benjamini and Hochberg false discovery rate correction.

Availability of supporting data

The Sequence Read Archives (SRA) supporting the results of this article are available in the NCBI GenBank repository, accession numbers: Arequipa(SRS838181), Colombiana(SRS841912), Esmeraldo(SRR833799, SRR833800, SRR058517, SRR058509, SRR058520, SRR058518, SRR058519, SRR058515, SRR058516, SRR058513, SRR058514, SRR058510, SRR058511, SRR058512) and Y(SRS842149). The 231 strain SRA is available ate the European Nucleotide Archive repository, by the accession number: PRJEB9129.

Additional files

Additional file 1: Table S1. *T. cruzi* CL Brener single copy genes.

Additional file 2: Table S2. Normalized RDC of the single copy genes from all six *T. cruzi* strains evaluated.

Additional file 3: Figure S1. Heterozygous SNPs proportion predicted ploidy of all the chromosomes from the *T. cruzi* strains: Arequipa, Colombiana, Sylvio, Esmeraldo, Y and 231.

Additional file 4: Figure S2. Normalized RDC of each position on the 41 CL Brener chromosomes by the reads from the *T. cruzi* strains: Arequipa, Colombiana, Sylvio, Esmeraldo, Y and 231.

Additional file 5: Table S3. Gene ontology analysis of the Esmeraldo-like and Non-Esmeraldo chromosome 31.

Additional file 6: Table S4. *Trypanosoma cruzi* strains sequencing information.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions
JLRC conceived the design of the study, performed all analyses, and draft the manuscript; GFRL, HRV, performed gene ontology analysis and participated in the chromosome copy number estimation analysis; TAOM performed the clustering analysis; RPB, GLM, RG, AMM, CB, RHG, CTL, BA, ATV participated in its design and helped to draft the manuscript; DCB conceived the study, its design and coordination and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This study was funded by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Instituto Nacional de Ciência e Tecnologia de Vacinas (INCTV)—Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). DCB, AMM, ATV are CNPq research fellows. JLRC, GFRL, HRV and TAOM received scholarships from CAPES and RPB received a scholarship from CNPq. We thank Michele Silva de Matos, Alexandra Lehmkuhl Gerber and Laila Viana de Almeida for the technical support.

Author details
[1]Laboratório de Imunologia e Genômica de Parasitos, Departamento deParasitologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2]Laboratório Nacional de Computação Científica, Petrópolis, Rio de Janeiro, Brazil. [3]Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [4]University of California San Francisco, San Francisco, CA, USA. [5]Universidad Cayetano Heredia, Lima, MD, Peru. [6]Johns Hopkins University, Baltimore, MD, USA. [7]Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden.

Received: 3 March 2015 Accepted: 1 June 2015
Published online: 04 July 2015

References

1. Hotez PJ, Bottazzi ME, Franco-Paredes C, Ault SK, Periago MR. The neglected tropical diseases of Latin America and the Caribbean: a review of disease burden and distribution and a roadmap for control and elimination. PLoS Negl Trop Dis. 2008;2(9):e300.
2. Coura JR, Dias JC. Epidemiology, control and surveillance of Chagas disease: 100 years after its discovery. Mem Inst Oswaldo Cruz. 2009;104 Suppl 1:31–40.
3. Martins-Melo FR, Alencar CH, Ramos Jr AN, Heukelbach J. Epidemiology of mortality related to Chagas' disease in Brazil, 1999–2007. PLoS Negl Trop Dis. 2012;6(2):e1508.
4. WHO. Research Priorities for Chagas Disease, Human African Trypanosomiasis and Leishmaniasis. Technical Report of the TDR Disease Reference Group on Chagas Disease, Human African Trypanosomiasis and Leishmaniasis. (Technical report series; no. 975). 2012. Availabe at: http://apps.who.int/iris/bitstream/10665/77472/1/WHO_TRS_975_eng.pdf.
5. Vargas N, Pedroso A, Zingales B. Chromosomal polymorphism, gene synteny and genome size in *T. cruzi* I and *T. cruzi* II groups. Mol Biochem Parasitol. 2004;138(1):131–41.
6. Lewis MD, Llewellyn MS, Gaunt MW, Yeo M, Carrasco HJ, Miles MA. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. Int J Parasitol. 2009;39(12):1305–17.
7. Zingales B, Andrade SG, Briones MR, Campbell DA, Chiari E, Fernandes O, Guhl F, Lages-Silva E, Macedo AM, Machado CR, et al A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem Inst Oswaldo Cruz. 2009;104(7):1051–4.
8. Minning TA, Weatherly DB, Flibotte S, Tarleton RL. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization. BMC Genomics. 2011;12:139.
9. Ackermann AA, Panunzi LG, Cosentino RO, Sanchez DO, Aguero F. A genomic scale map of genetic diversity in *Trypanosoma cruzi*. BMC Genomics. 2012;13:736.
10. Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MM, Schijman 627 AG, Llewellyn MS, Lages-Silva E, Machado CR, et al The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. Infection, genetics and evolution. 2012;12(2):240–53.
11. Panunzi LG, Aguero F. A genome-wide analysis of genetic diversity in *Trypanosoma cruzi* intergenic regions. PLoS Negl Trop Dis. 2014;8(5):e2839.
12. Tibayrenc M, Kjellberg F, Ayala FJ. 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 U S A. 1990;87(7):2414–8.
13. Tibayrenc M, Ayala FJ. How clonal are *Trypanosoma* and *Leishmania*? Trends Parasitol. 2013;29(6):264–9.
14. Westenberger SJ, Barnabe C, Campbell DA, Sturm NR. Two hybridization events define the population structure of *Trypanosoma cruzi*. Genetics. 2005;171(2):527–43.
15. de Freitas JM, Augusto-Pinto L, Pimenta JR, Bastos-Rodrigues L, Goncalves VF, Teixeira SM, Chiari E, Junqueira AC, Fernandes O, Macedo AM, et al. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. PLoS Pathog. 2006;2(3):e24.
16. Machado CA, Ayala FJ. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. Proc Natl Acad Sci U S A. 2001;98(13):7396–401.
17. Brisse S, Henriksson J, Barnabe C, Douzery EJ, Berkvens D, Serrano M, De Carvalho MR, Buck GA, Dujardin JC, Tibayrenc M. Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. Infection, genetics and evolution. 2003;2(3):173–83.
18. Sturm NR, Vargas NS, Westenberger SJ, Zingales B, Campbell DA. Evidence for multiple hybrid groups in *Trypanosoma cruzi*. Int J Parasitol. 2003;33(3):269–79.
19. Baptista RD, D'Avila DA, Segatto M, Do Valle IF, Franco GR, Valadares HMS Gontijo ED, Galvao LMD, Pena SDJ, Chiari E, et al Evidence of substantial recombination among *Trypanosoma cruzi* II strains from Minas Gerais. Infect Genet Evol. 2014;22:183–91.
20. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN Ghedin E, Worthey EA, Delcher AL, Blandin G, et al The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science. 2005;309(5733):409–15.
21. Franzen O, Ochaya S, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, Andersson B. Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio X10/1 and comparison with *T. cruzi* VI CL Brener. PLoS Negl Trop Dis. 2011;5(3):e984.
22. Souza RT, Lima FM, Barros RM, Cortez DR, Santos MF, Cordero EM, Ruiz JC, Goldenberg S, Teixeira MM, da Silveira JF. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. PLoS One. 2011;6(8):e23042.
23. Cerqueira GC, Bartholomeu DC, DaRocha WD, Hou L, Freitas-Silva DM, Machado CR, El-Sayed NM, Teixeira SM. Sequence diversity and evolution of multigene families in *Trypanosoma cruzi*. Mol Biochem Parasitol. 2008;157(1):65–72.
24. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848–53.
25. Iskow RC, Gokcumen O, Lee C. Exploring the role of copy number variants in human adaptation. Trends in genet. 2012;28(6):245–57.
26. Martins C, Baptista CS, Ienne S, Cerqueira GC, Bartholomeu DC, Zingales B. Genomic organization and transcription analysis of the 195-bp satellite DNA in *Trypanosoma cruzi*. Mol Biochem Parasitol. 2008;160(1):60–4.
27. Clayton CE. Life without transcriptional control? From fly to man and back again. EMBO J. 2002;21(8):1881–8.
28. Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE. Gene expression in trypanosomatid parasites. J Biomed Biotechnol. 2010;2010:525241.
29. Henriksson J, Dujardin JC, Barnabe C, Brisse S, Timperman G, Venegas J, Pettersson U, Tibayrenc M, Solari A. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is evolutionarily informative. Parasitology. 2002;124(Pt 3):277–86.
30. Pedroso A, Cupolillo E, Zingales B. Evaluation of *Trypanosoma cruzi* hybrid stocks based on chromosomal size variation. Mol Biochem Parasitol. 2003;129(1):79–90.
31. Triana O, Ortiz S, Dujardin JC, Solari A. *Trypanosoma cruzi*: variability of stocks from Colombia determined by molecular karyotype and minicircle Southern blot analysis. Exp Parasitol. 2006;113(1):62–6.
32. Lima FM, Souza RT, Santori FR, Santos MF, Cortez DR, Barros RM, Cano MI, Valadares HM, Macedo AM, Mortara RA, et al Interclonal variations in the molecular karyotype of *Trypanosoma cruzi*: chromosome rearrangements in a single cell-derived clone of the G strain. PLoS One. 2013;8(5):e63738.

33. Weatherly DB, Boehlke C, Tarleton RL. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. BMC Genomics. 2009;10:255.
34. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H, et al Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. Genome Res. 2011;21(12):2129–42.
35. Branche C, Ochaya S, Aslund L, Andersson B. Comparative karyotyping as a tool for genome structure analysis of *Trypanosoma cruzi*. Mol Biochem Parasitol. 2006;147(1):30–8.
36. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al Comparative genomics of trypanosomatid parasitic protozoa. Science. 2005;309(5733):404–9.
37. Tibayrenc M, Ward P, Moya A, Ayala FJ. Natural populations of *Trypanosoma cruzi*, the agent of Chagas disease, have a complex multiclonal structure. Proc Natl Acad Sci U S A. 1986;83(1):115–9.
38. Bartholomeu DC, Cerqueira GC, Leao AC, daRocha WD, Pais FS, Macedo C, Djikeng A, Teixeira SM, El-Sayed NM. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen *Trypanosoma cruzi*. Nucleic Acids Res. 2009;37(10):3407–17.
39. De Pablos LM, Osuna A. Multigene Families in *Trypanosoma cruzi* and Their Role in Infectivity. Infect Immun. 2012;80(7):2258–64.
40. Bartholomeu DC, de Paiva RM, Mendes TA, DaRocha WD, Teixeira SM. Unveiling the intracellular survival gene kit of trypanosomatid parasites. PLoS Pathog. 2014;10(12):e1004399.
41. Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, Humpton TJ, Brito IL, Hiraoka Y, Niwa O, Amon A. Aneuploidy drives genomic instability in yeast. Science. 2011;333(6045):1026–30.
42. Abbey D, Hickman M, Gresham D, Berman J. High-Resolution SNP/CGH Microarrays Reveal the Accumulation of Loss of Heterozygosity in Commonly Used *Candida albicans* Strains. G3 (Bethesda). 2011;1(7):523–30.
43. Rancati G, Pavelka N, Fleharty B, Noll A, Trimble R, Walton K, Perera A, Staehling-Hampton K, Seidel CW, Li R. Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. Cell. 2008;135(5):879–93.
44. Farrer RA, Henk DA, Garner TW, Balloux F, Woodhams DC, Fisher MC. Chromosomal copy number variation, selection and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity. PLoS Genet. 2013;9(8):1003703.
45. Peacock L, Ferris V, Sharma R, Sunter J, Bailey M, Carrington M, Gibson W. Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. Proc Natl Acad Sci U S A. 2011;108(9):3671–6.
46. Akopyants NS, Kimblin N, Secundino N, Patrick R, Peters N, Lawyer P, Dobson DE, Beverley SM, Sacks DL. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. Science. 2009;324(5924):265–8.
47. Gaunt MW, Yeo M, Frame IA, Stothard JR, Carrasco HJ, Taylor MC, Mena SS, Veazey P, Miles GA, Acosta N, et al Mechanism of genetic exchange in American trypanosomes. Nature. 2003;421(6926):936–9.
48. Sturm NR, Campbell DA. Alternative lifestyles: the population structure of *Trypanosoma cruzi*. Acta Trop. 2010;115(1–2):35–43.
49. Llewellyn MS, Miles MA, Carrasco HJ, Lewis MD, Yeo M, Vargas J, Torrico F, Diosque P, Valente V, Valente SA, et al Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. PLoS Pathog. 2009;5(5):e1000410.
50. Augusto-Pinto L, Teixeira SMR, Pena SDJ, Machado CR. Single-nucleotide Polymorphisms of the *Trypanosoma cruzi* MSH2 gene support the existence of three phylogenetic lineages presenting differences in mismatch-repair efficiency. Genetics. 2003;164(1):117–26.
51. Machado CR, Augusto-Pinto L, McCulloch R, Teixeira SM. DNA metabolism and genetic diversity in Trypanosomes. Mutat Res. 2006;612(1):40–57.
52. Kanmogne GD, Bailey M, Gibson WC. Wide variation in DNA content among isolates of *Trypanosoma brucei* ssp. Acta Trop. 1997;63(2–3):75–87.
53. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, Andersson B, Bontempi E, Eisen J, Angiuoli S, et al Gene synteny and evolution of genome architecture in trypanosomatids. Mol Biochem Parasitol. 2004;134(2):183–91.
54. de Lederkremer RM, Agusti R. Glycobiology of *Trypanosoma cruzi*. Adv Carbohydr Chem Biochem. 2009;62:311–66.
55. Buscaglia CA, Campo VA, Frasch ACC, Di Noia JM. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. Nat Rev Microbiol. 2006;4(3):229–36.
56. Jones C, Todeschini AR, Agrellos OA, Previato JO, Mendonca-Previato L. Heterogeneity in the biosynthesis of mucin O-glycans from *Trypanosoma cruzi* tulahuen strain with the expression of novel galactofuranosyl-containing oligosaccharides. Biochemistry. 2004;43(37):11889–97.
57. Acosta-Serrano A, Almeida IC, Freitas-Junior LH, Yoshida N, Schenkman S. The mucin-like glycoprotein super-family of *Trypanosoma cruzi*: structure and biological roles. Mol Biochem Parasitol. 2001;114(2):143–50.
58. Souto RP, Fernandes O, Macedo AM, Campbell DA, Zingales B. DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. Mol Biochem Parasitol. 1996;83(2):141–52.
59. Burgos JM, Altcheh J, Bisio M, Duffy T, Valadares HM, Seidenstein ME, Piccinali R, Freitas JM, Levin MJ, Macchi L, et al. Direct molecular profiling of minicircle signatures and lineages of *Trypanosoma cruzi* bloodstream populations causing congenital Chagas disease. Int J Parasitol. 2007;37(12):1319–27.
60. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment genome res. 1998;8(3):175–85.
61. Richterich P. Estimation of errors in "raw" DNA sequences: a validation study. Genome Res. 1998;8(3):251–9.
62. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
65. Li L, Stoeckert Jr CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
66. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2.
67. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448–9.