UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE DOUTORADO EM BIOINFORMÁTICA

Análise da diversidade, clusterização e estudo de mecanismos evolutivos geradores de diversidade nas grandes famílias gênicas de *Trypanosoma cruzi*.

LEANDRO MARTINS DE FREITAS

Belo Horizonte
Setembro – 2011

LEANDRO MARTINS DE FREITAS

Análise da diversidade, clusterização e estudo de mecanismos evolutivos geradores de diversidade nas grandes famílias gênicas de *Trypanosoma cruzi*.

Tese apresentada ao Programa de Doutorado em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, para obtenção do Título de Doutor em Bioinformática.

Orientadora: Daniella Castanheira Bartholomeu

Belo Horizonte

Setembro – 2011

AGRADECIMENTOS

A professora Daniella Castanheira Bartholomeu pela orientação, ensinamentos, exemplo de trabalho e dedicação, servindo de modelo como professora e pesquisadora. Sempre ajudou muito para a construção desse trabalho e sendo sempre muito compreensiva.

A banca por ter aceitado o convite para avaliação desse trabalho.

A Sara Lopes dos Santos por ter contribuído de forma importantíssima na parte experimental

Aos professores do programa de Doutorado em Bioinformática e convidados que me ajudaram com conhecimento e novas idéias durante o trabalho.

A CAPES pelo apóio financeiro através de bolsa de estudos.

Aos amigos e parceiro do LIGP por sugestões e contribuições para o trabalho.

A minha amiga professora Laize Tomazi por ter me ajudado nesse último semestre assumindo as aulas da turma de Genética I para que eu pudesse dedicar mais tempo para finalização desse trabalho.

Aos amigos do Departamento de Biologia Geral da UFMG, amigos de graduação que também sempre acompanharam e torceram por mim, amigos do Instituo Multidisciplinar em Saúde da Universidade Federal da Bahia IMS/UFBA que sempre incentivaram o trabalho.

A minha família, mãe pelos incentivos e ajuda, e aos meus irmãos Leonardo, Patrícia e Haroldo pelo companheirismo e amizade.

"Não consigo me convencer de que um Deus caridoso e onipotente teria propositalmente criado vespas parasitas com a intenção expressa de alimentá-las dentro de corpos vivos de lagartas." (Charles Darwin)

SUMÁRIO

RESUMO

Com a publicação dos resultados do projeto genoma de *Trypanosoma cruzi*, abordagens que visam contribuir para um melhor entendimento da diversidade dos genes codificadores das grandes famílias de proteínas de superfície do parasito podem ser aplicadas em escala genômica. Uma das famílias multigênicas de *T. cruzi* mais bem estudadas é a trans-sialidase/sialidase-like (TcS). Apesar de quatro grupos desta família já terem sido identificados em trabalhos anteriores e apenas um deles albergar trans-sialidases ativas, todos os membros da família estão anotados em banco de dados como trans-sialidase. Neste trabalho, usando metodologias de clusterização de sequências, os 508 membros completos da família TcS foram divididos em oito grupos bem distintos, TcS I a VIII. Os oito grupos foram caracterizados baseado na presença de motivos típicos da família, localização cromossômica, expressão gênica e antigenicidade. Interessantemente, membros dos diferentes grupos apresentam um padrão distinto de localização cromossômica. Membros de TcS II, que albergam proteínas envolvidas em adesão e invasão celular, são preferencialmente localizados nas regiões subteloméricas, enquanto que membros do novo e maior grupo da família (TcS V) apresentam localização cromossômica interna. Resultados de expressão, usando RT-PCR em tempo real, mostram que os genes TcS em geral são mais expressos nas formas encontradas no hospedeiro vertebrado e existe variação de expressão entre os membros até mesmo dentro dos grupos. Novos epitopos de célula B foram identificados na família TcS em membros de grupos previamente descritos bem como em membros de novos grupos. Alguns dos peptídeos reativos foram encontrados em vários membros da família. A reatividade cruzada entre vários epitopos TcS e variabilidade de sequência da família pode contribuir na estratégia do parasito de escapar do ataque do sistema imune através da exposição simultânea de epitopos de célula B relacionados gerando respostas imunes espúrias e não neutralizantes.

Em um segunda parte deste trabalho, nós realizamos análises comparativas de diversidade das famílias gênicas que codificam proteínas de superfície de *T. cruzi*. Estas famílias apresentam diversidade altamente heterogênea, variando de níveis de baixa (DGF-1 e SAP), intermediária (RHS e mucin-like) e alta (TcMUC, TcS, GP63, MASP) diversidade. Gráficos do tipo MDS (*Multidimentional scaling*) foram gerados para representar a diversidade de cada família e o método *k-means*, para definir os grupos

intra-família. MASP e TcMUC apresentam um padrão similar de diversidade das regiões codificadora e 3' flaqueadora. As sequências codificadoras de ambas famílias apresentam um padrão contínuo de diversidade enquanto que a região 3' flanqueadora de cada família é altamente conservada. Especulamos que eventos de recombinação entre os membros de cada família possa contribuir para este padrão de diversidade, em um mecanismo similar ao previamente descrito para os genes MSP2 de *Anaplasma marginale*. De fato, nós identificamos fragmentos compartilhados entre membros da família MASP que pertencem a grupos K-means distintos. Nós também identificamos fragmentos compartilhados entre diferentes famílias gênicas de *T. cruzi*. A maioria destes fragmentos se localizam na extremidade 3' dos genes, sugerindo novamente que estas regiões possam estar envolvidas nestes possíveis mecanismos de recombinação.

ABSTRACT

With the availability of the data generated by the *Trypanosoma cruzi* genome project, it is now feasible to characterize the sequence diversity of the largest *T. cruzi* multigene families. The trans-sialidase/sialidase-like (TcS) is the largest *T. cruzi* gene family and one of the most studied of this parasite. Despite the fact that four TcS groups are well characterized and only one of them contains active trans-sialidase enzymes, all members of the family are annotated in the *T. cruzi* genome database as trans-sialidase. Here, by performing sequence clustering analysis, besides the four previously described TcS groups (TcS I to IV), we have identified four new ones (TcS V to VIII). The eight groups were characterized based on the presence of key TcS motifs, chromosomal localization, gene expression and antigenic profile. Interestingly, members of distinct TcS groups show distinctive patterns of chromosome localization. Members of the TcSgroupII, which harbor proteins involved in host cell attachment/invasion, are preferentially located in subtelomeric regions, whereas members of the largest and new TcSgroupV have internal chromosomal locations. Real-time RT-PCR confirms the expression of genes derived from new groups and shows that the pattern of expression is not similar within and between groups. We also performed B-cell epitope prediction on the family and constructed a TcS-specific peptide array, which was screened with sera from *T. cruzi*–infected mice. We demonstrated that all seven groups represented in the array are antigenic. A highly reactive peptide occurs in sixty TcS proteins including members of two new groups and may contribute to the cross-reactivity of *T. cruzi* epitopes during infection.

In a second part of this work, we have performed comparative analysis of the sequence diversity of the largest *T. cruzi* multigene families, named DGF-1, SAP, RHS, mucin-like, TcMUC, TcS, GP63, MASP. These families have distinct levels of diversity, ranging from low (DGF-1 e SAP), intermediate (RHS, mucin-like) and high (TcMUC, TcS, GP63, MASP). Multidimentional scaling (MDS) plots were generated as a visual representation of the distance matrix for each family and K-means method was used to define the intra-family groups. MASP and TcMUC have similar patterns of diversity in the coding and 3' flanking regions. Their coding regions display a gradient of diversity whereas their 3' flanking regions are highly conserved. We hypothesized that recombination might contribute to this pattern of diversity in a similar mechanism

described for the MSP2 gene family of *Anaplasma*. We have identified shared fragments among MASP members belonging to distinct k-mean groups. We have also identified shared fragments among the *T. cruzi* gene families, mainly involving TcS and MASP. The majority of these fragments is located at the 3' end of the genes, again suggesting that these regions might be involved in these recombination events.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

3'UTR – região não traduzida do mRNA localizada na extremidade 3' (3' UnTranslated Region)

5'UTR – região não traduzida do mRNA localizada na extremidade 5' (5' UnTranslated Region)

Asp-box – motivo típico de sialidase - SxDxGxWT

cDNA – DNA complementar (complement DNA) é o DNA sintetizado a partir de uma molécula de mRNA

CMP– Citidina monofosfato (Cytidine Monophosphate)

CK-18 – citoqueratina-18 (Cytokeratin 18)

DGF-1 – Dispersed Gene Family Protein 1

DIRE – degenerate ingi/L1Tc-related element

DNA – ácido desoxirribonucleico ou DNA, (em inglês: deoxyribonucleic acid)

FLY – Motivo VTVxNVFLYNR

GlcNAc - N-acetil-glicosamina (N-acetyl-D-glucosamine)

GalNAc – N-acetil-galactosamina (N-acetyl-alpha-D-galactosamine)

GPI – glicosil fosfatidil inositol (glycosilphosphatidilinositol)

GP63 – Metaloprotease GP63

indel – inserção ou deleção

kDa – Kilodalton

L1Tc – retrotransposon não LTR de *T. cruzi* (non-long terminal repeat retrotransposon)

MASP – proteína de superfície associada a mucinas (MucinAssociated Surface Protein)

MDS – escala multidimensional (Multidimensional scaling)

Mb – Mega base

MEME – Multiple Em for Motif Elicitation

mRNA – RNA mensageiro, ácido ribonucleico ou RNA, (em inglês: ribonucleic acid)

NJ – Agrupamentos de vizinhos (Neighbor Joining)

pb – pares de base

nt – nucleotídeo

HSG (high similarity group, com mais de 80% de similaridade com AAB49414)

LSG (low similarity group, similaridade entre 54 e 62% com AAB49414)

RHS – Retrotransposon Hot Spot

RT – PCR tempo real (PCR-realtime)

SAP – proteínas ricas em Serina-, Alanina-, e Prolina- (Serine-Alanine-and Proline-rich protein)

SAPA – antígeno de fase aguda (shed acute phase antigen)

SIRE – short interspersed repetitive element

TcMUC – mucina de *T. cruzi* expressa pelo parasita no estágio presente no mamífero

TcSMUG – mucina de *T. cruzi* expressa pelo parasita no estágio presente no inseto

TcS – proteínas trans-sialidase/sialidase-like de *T. cruzi*

TS – Trans-Sialidase

TSSA – pequenos de antígenos de superfície da forma tripomastigota (trypomastigote small surface antigen)

VIPER – vestigial interposed retroelement

VAR – família que codifica as proteínas variantes da superfície () do eritrócito infectado por *Plasmoduim falciparum*

VSG – Glicoproteína variável de superfície (Variable Surface Glycoprotein) de *T. brucei*

WHO – Organização Mundial de Saúde (World Health Organization)

1. Introdução

Doenças infecciosas causadas por protozoários continuam sendo a maior causa de mortalidade em países em desenvolvimento (WHO, 2011). malária, leishmaniose e tripanossomíases foram classificadas pela Organização Mundial da Saúde (WHO) como doenças negligenciadas. Esta classificação decorre do fato das companhias farmacêuticas não terem investido no desenvolvimento de novas drogas e vacinas contra estas parasitoses, uma vez que o tratamento e a prevenção destas doenças não são economicamente atrativos. Não há vacinas contra estes parasitos e as drogas atualmente disponíveis possuem eficácia limitada, efeitos colaterais, problemas relacionados à resistência e custo elevado. Devido à carência de métodos de tratamento e diagnóstico efetivos (Urbina e Docampo, 2003); WHO, 2011), novos tratamentos quimioterápicos e profiláticos necessitam ser desenvolvidos no combate a estas doenças parasitárias.

Recentemente o genoma de vários parasitos protozoários foram sequenciados (Gardner *et al.*, 2002; Carlton, 2003; Berriman *et al.*, 2005; El-Sayed *et al.*, 2005a; Gardner *et al.*, 2005; Ivens *et al.*, 2005; Loftus *et al.*, 2005; Carlton *et al.*, 2007; Franzén *et al.*, 2011). A disponibilidade dos dados gerados nestes projetos permite que novas abordagens e metodologias voltadas para o estudo da biologia e desenvolvimento de novas estratégias de prevenção, tratamento e diagnóstico das moléstias causadas pelos parasitos sejam aplicadas em escala genômica.

Dentre os genomas de parasitos já sequenciados encontra-se o genoma do *Trypanosoma cruzi* (El-Sayed *et al.*, 2005a; Franzén *et al.*, 2011), agente etiológico da doença de Chagas ou tripanossomíase americana.

1.1. O *T. cruzi* e a doença de Chagas

A distribuição do *T. cruzi* ocorre principalmente em parte da América do Norte, América Central e América do Sul. Entretanto, devido à grande mobilidade atual das populações humanas, alguns casos da doença de Chagas também são encontrados nos Estados Unidos, Canadá e

Europa (WHO, 2011). Estima-se que 10 milhões de pessoas estejam infectadas por *T. cruzi* na América Central e do Sul, 25 milhões estão sobre risco de infecção, e aproximadamente 13.000 morrem todos os anos da doença de Chagas. Ainda não existe vacina ou medicamento eficientes para a fase crônica da doença de Chagas, quando as sintomatologias clínicas mais severas podem se manifestar. Os medicamentos usados durante o tratamento são tóxicos e efetivos somente durante a fase inicial da doença, a fase aguda (WHO, 2011; Barrett *et al.*, 2003; Urbina e Docampo, 2003).

A doença de Chagas é transmitida principalmente pelo vetor invertebrado (inseto barbeiro da família *Reduviidae*), para o hospedeiro vertebrado (mamífero). Outras vias de transmissão incluem acidente de laboratório, ingestão de alimentos contaminada, transfusão de sangue, transplante de órgãos, e por via transplacentária da mãe infectada para o filho durante a gravidez (WHO, 2011; Barrett *et al.*, 2003).

O *T. cruzi* apresenta diversos "estágios evolutivos" que apresentam distintas propriedades replicativas e infectivas. As formas replicativas são epimastigota e amastigota, presentes no inseto e no mamífero, respectivamente. As formas não replicativas e infectantes são a tripomastigosta metacíclica, presente nas fezes do inseto, e a tripomastigota sanguínea presente no hospedeiro vertebrado. As formas amastigotas também apresentam propriedades infectivas (Behbehani, 1973; Nogueira e Cohn, 1976; Hudson *et al.*, 1984; Ley *et al.*, 1988; Mortara *et al.*, 2005).

O ciclo de vida do parasita, transmitido pelo inseto, começa com o protozoário, presente nas fezes do barbeiro infectado, tendo acesso à corrente sanguínea do mamífero através do ferimento provocado pelo barbeiro durante o repasto sanguíneo ou através de mucosas (boca ou olhos). O barbeiro é encontrado usualmente em casas de madeira com frestas, essas geralmente construídas na zona rural ou áreas suburbanas. Normalmente os barbeiros se escondem durante o dia, tornando-se ativos durante a noite período no qual saem para se alimentar do sangue dos mamíferos. Os barbeiros fazem o repasto sanguíneo em áreas expostas, defecando perto do ferimento causado durante o repasto. O parasito entra na corrente sanguínea na forma de tripomastigota metacíclica e invade uma variedade de tipos celulares. Na célula infectada, o parasito se diferencia na forma replicativa amastigota e passa por uma série de divisões binárias e depois se diferencia na forma tripomastigota. A forma tripomastigota que é liberada na corrente

sanguínea do mamífero após o rompimento da célula inicia um novo ciclo de invasão celular. O vetor invertebrado durante o repasto sanguíneo no hospedeiro vertebrado infectado é contaminado pela ingestão de tripomastigotas da corrente sanguínea. No intestino do invertebrado os tripomastigotas se diferenciam na forma replicativa epimastigota que se multiplica. A forma epimastigota se diferencia em tripomastigota metacíclica na porção final do intestino do barbeiro, desta forma fechando o ciclo.

A forma de controle mais eficiente da transmissão da doença de Chagas é o controle do vetor. Outra medida importante de controle da transmissão é o rastreamento dos bancos de sangue para a infecção pelo parasito.

A doença de Chagas pode ser dividida em duas fases de manifestações clínicas – fase aguda e fase crônica. A fase aguda é caracterizada pela alta parasitemia e parasitismo tecidual sendo geralmente assintomática, no entanto algumas pessoas podem manifestar dor de cabeça, febre, dores musculares, dificuldade de respirar, linfonodos inchados e dor abdominal ou no peito. Na fase crônica, os parasitos são frequentemente encontrados no músculo cardíaco ou liso ou aparelho digestivo (causando megaesôfago ou megacólon). Na fase crônica, 25 a 30% dos pacientes podem apresentar sintomatologias variadas como comprometimento cardíaco, digestivo ou neurológico, levando a debilidade, diminuição da qualidade e expectativa de vida (WHO, 2011). Não se sabe quais os fatores determinam a amplo espectro de manifestações clínicas, mas acredita-se que a variabilidade genética tanto do parasito quanto do hospedeiro possam estar envolvidas. De fato, *T. cruzi* é um táxon extremamente heterogêneo tanto genotipicamente quanto fenotipicamente. Apesar de sua reprodução ser predominantemente clonal, evidências de raros eventos envolvendo troca de material genético já foram reportadas (Machado e Ayala, 2001; Brisse *et al.*, 2003; Zingales *et al.*, 2009). Baseado em vários marcadores moleculares, o táxon *T. cruzi* foi dividido em seis linhagens filogenéticas, Tc I a VI (Zingales *et al.*, 2009). Acreditamos que com o aumento amostral das sequências nos banco de dados, outros grupos/linhagens serão identificados.

A grande divergência genética entre as linhagens de *T. cruzi* é refletida em muitos aspectos epidemiológicos e patológicos da doença de Chagas, principalmente nos que diz respeito às linhagens Tc I e Tc II, que são as mais bem estudadas. Em países do Cone Sul da América do Sul onde a doença de Chagas é mais severa, Tc I é associado ao ciclo silvestre infectando

principalmente mamíferos arbóreos, enquanto Tc II predomina nos ciclos domésticos infectando o homem e outros mamíferos terrestres (Zingales *et al.*, 1998; Coura *et al.*, 2002; Yeo *et al.*, 2005; Flores-López e Machado, 2011). Evidências epidemiológicas e de genotipagem de parasitos diretamente de tecidos humanos infectados têm demonstrado que a linhagem Tc II é o agente causal predominante da doença de Chagas nesta área. Por outro lado, Tc I predomina na bacia Amazônica e em áreas endêmicas da doença de Chagas na Venezuela (Miles *et al.*, 1981; Coura *et al.*, 2002; Zingales *et al.*, 2009).

1.2. O genoma do *T. cruzi*

O genoma de *T. cruzi* foi sequenciado em 2005 por um consórcio internacional, juntamente com o genoma de dois outros tripanossomatídeos causadores de importantes doenças tropicais, o *Trypanosoma brucei* e a *Leishmania major* (Berriman *et al.*, 2005; El-Sayed *et al.*, 2005a; Ivens *et al.*, 2005). CL Brener, representante da linhagem híbrida Tc VI, foi a cepa referência do projeto genoma de *T. cruzi* por ser bem caracterizada experimentalmente e para facilitar análises comparativas com o projeto EST de CL Brener que estava em andamento na época (Agüero *et al.*, 2004). O genoma foi sequenciado usando a estratégia "whole genome shotgun" com cobertura de 14 vezes e o tamanho do genoma diplóide foi estimado entre 106,6 – 110,7 Mb.

Devido a natureza híbrida e repetitiva do genoma de CL Brener, uma série de modificações dos parâmetros de montagem do genoma tiveram que ser implementadas. Estas modificações foram realizadas no sentido de se favorecer a montagem separada dos dois haplótipos de CL Brener, uma vez que critérios menos estringentes de montagem geraram um grande número de pseudogenes. Fortes evidências sugerem que a linhagem híbrida Tc VI, a qual pertence CL Brener, se originou de genomas ancestrais das linhagens Tc II e Tc III (De Freitas, *et al.*, 2006; Ienne *et al.*, 2010). A fim de verificar se os dois haplótipos de CL Brener foram montados separadamente, o genoma da cepa Esmeraldo, representante de um dos genomas parentais (Tc II), foi sequenciado a baixa cobertura (2,5x). Análises comparativas entre os *reads* de Esmeraldo e os *contigs* de CL Brener, permitiram a identificação de pares de alelos para cerca de metade dos genes de CL Brener. Os dois haplótipos apresentam em média 5,4% de

divergência, sendo que este valor diminui para 2,2% nas regiões codificadoras. Os dois haplótipos apresentam alta sintenia, sendo inserções e deleções as alterações mais frequentes encontradas em regiões subteloméricas e intergênicas (El-Sayed *et al.*, 2005a).

Foram identificados aproximadamente 12.000 genes por genoma haplóide e estimado um total de 22.570genes. Destes, cerca de metade corresponde a genes representados por pares de alelos derivados dos haplótipos Tc II (Esmo-like) e Tc III (non-Esmo-like). O restante dos genes não pode ser atribuído a um haplótipo específico (El-Sayed *et al.*, 2005a).

Grande parte desse genoma é composto de sequências repetitivas (~50%), tais como retrotransposons e grandes famílias de proteínas de superfície, correspondendo a 18% do total de genes codificadores de proteínas (El-Sayed *et al.*, 2005a). Esses genes ocorrem em clusters dispersos de repetições em tandem e interespaçadas, frequentemente encontradas em regiões subteloméricas e regiões não sintênicas com *T. brucei* e *L. major* (El-Sayed *et al.*, 2005a; El-Sayed *et al.*, 2005b). A montagem dos cromossomos de *T. cruzi* aconteceu posteriormente à publicação do genoma sendo proposta a existência de 41 cromossomos (Weatherly *et al.*, 2009).

O genoma de *T. cruzi* é caracterizado por uma grande expansão do número de genes das famílias de proteínas de superfície comparado com os genomas de *T. brucei* e *L. major* e alguns destes genes são específicos do parasito (El-Sayed *et al.*, 2005a). Com a anotação do genoma do *T. cruzi*, além de terem sido identificados novos membros de famílias previamente descritas, identificou-se uma nova e grande família multigênica denominada MASP (*mucin-associated surface protein*).

A maioria das proteínas das grandes famílias de superfície de *T. cruzi* possuem, em sua estrutura polipeptídica, sequências sinalizadoras conservadas com papel no direcionamento e ancoramento destas proteínas na superfície do parasita, a saber: peptídeo sinal e sequência para ancoragem na membrana através de GPI (Barry *et al.*, 2006). Outra característica compartilhada pelas famílias de proteínas de superfície de *T. cruzi* é a localização no genoma em regiões que apresentam ausência de sintenia com os genomas de *T. brucei* e *L. major*, sendo estas regiões em *T. cruzi* ricas em genes e pseudogenes das famílias MASP, TcS (trans-sialidase/sialidase-like), GP63 (glicoproteína de superfície 63 kDa), TcMUC (*T. cruzi* mucinas), DGF-1 (*dispersed gene family -1*), SAP (serine-alanine and proline rich protein), além de RHS (*retrotransposon hot spot protein*) (El-Sayed *et al.*, 2005b; Bartholomeu *et al.*, 2009). Nessas regiões também são

encontrados vários elementos similares aos retrotransposons VIPER, L1Tc, SIRE e DIRE (Baida *et al.*, 2006; Bartholomeu *et al.*, 2009). A expansão do número de genes nessas regiões, abundância de pseudogenes e retroelementos, ocorrência de gene quimeras e grupos gênicos direcionais de pequeno tamanho sugerem que estas regiões foram submetidas ou ainda sobrem intensos eventos de rearranjo gênico (Baida *et al.*, 2006, Bartholomeu *et al*., 2009).

1.3. Grandes famílias de proteínas de superfície em *T. cruzi*

1.3.1. Trans-sialidase/sialidase-like (TcS)

Trans-sialidase/sialidase-like é a maior família gênica do *T. cruzi*. Ela é composta por membros que apresentam atividade de trans-sialidase e sequências similares associadas a outras funções. Neste trabalho iremos nos referir àqueles membros com atividade trans-sialidase como "TS" e a família como um todo como "TcS".

Os eucariotos superiores e algumas bactérias são capazes de sintetizar o carboidrato ácido siálico e adicioná-lo aos glicoconjugados e/ou proteínas. A enzima sialiltransferase, presente no complexo de Golgi dos eucariotos "superiores" (deuterostômios), usa o CMP-ácido siálico como substrato doador de ácido siálico para os glicoconjugados e/ou proteínas (Frasch, 2000, Varki et al., 1999).

As espécies de protozoários do gênero *Trypanosoma* são incapazes de sintetizar ácido siálico. Apesar desta limitação, algumas espécies como *T. brucei* (agente etiológico da doença do sono em humano e nagana em animais domésticos) e *T. cruzi* expressam a enzima trans-sialidase (TS) - capaz de remover ácido siálico dos glicoconjugados e proteínas do hospedeiro e adicioná-lo a outras moléculas presentes na membrana do parasito. O parasito *T. rangeli* (espécie não patogênica em humanos) também apresenta membros da família com atividade de sialidase, mas não apresenta atividade de trans-sialidase (Buschiazzo *et al.*, 2000).

A enzima TS é expressa e ancorada na membrana externa do *T. cruzi*. A enzima TS de *T. cruzi* é capaz de retirar moléculas de ácido siálico ligados a glicoconjugados e/ou proteínas presentes na membrana plasmática das células do hospedeiro e adicioná-las a galactose beta-terminal dos glicoconjugados presentes na membrana externa da sua célula. As proteínas de

superfície da família das mucinas de *T. cruzi* (TcMUC) são as principais moléculas aceptoras de ácido siálico transferidos por TS. A família TcS é formada por 1.430 genes espalhados pelo genoma de *T. cruzi* (El-Sayed *et al.*, 2005a), estes apresentam grande variação entre suas sequências, e dentre estes foram identificados 693 pseudogenes (El-Sayed *et al.*, 2005a). A grande variação das sequências entre as proteínas da família TcS resultou na aquisição de funções adicionais à adição de ácido siálico, dentre estas funções foram relatadas função na ligação com fibronectina (Giordano *et al.*, 1994), colágeno (Velge *et al.*, 1988), citoqueratina (Magdesian *et al.*, 2001) e proteínas do complemento (Frasch, 2000; Beucher e Norris, 2008; Souza *et al.*, 2010). Os genes dessa família ainda apresentam expressão estágio dependente (Frasch, 2000; Souza *et al.*, 2010), sendo mais expressos nas fases tripomastigotas e menos expressos nas fases amastigota e epimastigota do parasito (Atwood *et al.*, 2005). TcS exibe funções críticas para infecção, persistência e patogênese da doença de Chagas, no entanto, os mecanismos moleculares associados a suas funções permanecem em grande parte desconhecidos (Tonelli *et al.*, 2010).

As diferentes funções das proteínas da família TcS foram descritas e são associadas a diferentes domínios presentes na extremidade N- ou C-terminal (Schenkman *et al.*, 1994). As funções descritas associadas à extremidade N-terminal são: trans-sialidades/atividade sialidase, regulação das proteínas do complemento, adesão celular, ligação beta-galactose, ligação a laminina. As funções associadas à extremidade C-terminal são: multimerização da proteína e modulação da resposta de anticorpos contra a região N-terminal (Schenkman *et al.*, 1994; Frasch, 2000).

Na extremidade N-terminal das proteínas da família TcS existem três motivos característicos. O motivo FRIP é o motivo mais próximo da extremidade N-terminal, sendo responsável pela ligação da proteína TS ao grupo carboxilato do ácido siálico e está presente com uma cópia na proteína TS (Todeschini *et al.*, 2000). Já o motivo SxDxGxTW, típico motivo sialidase, ocorre de uma a quatro vezes na proteína (Cross e Takle, 1993; Schenkman *et al.*, 1994). Por fim, o motivo VTVxNVfLYNR apresenta uma cópia na proteína e é característico de todos os membros da família (Cross e Takle, 1993; Schenkman *et al.*, 1994), sendo posicionado mais distante da extremidade N-terminal e é relacionado com adesão e invasão celular (Magdesian *et al.*, 2001; Magdesian *et al.*, 2007). A extremidade C-terminal das proteínas TS apresenta região de tamanho variável sendo formada por repetições de 12 aminoácidos chamadas

SAPA e também na extremidade C-terminal está presente o sítio de ancoragem glicosil de fosfatidil inositol (GPI) para ligação da proteína na membrana plasmática do parasito.

As proteínas da família TcS podem ser classificadas em quatro grupos: TcS I (TCNA, SAPA e TS-epi), TcS II (SA85 1.1, TSA, ASP, gp82 e gp90), TcS III (FL-160 e CRP) e TcS IV (Tc13) (Schenkman *et al.*, 1994). O grupo TcS I corresponde às trans-sialidases ativas (TS). Já os grupos TcS II, III e IV são chamados por alguns autores de TcS-like formando um grupo mais distante que apresenta algumas características presentes na família, no entanto, não apresentam membros com atividade trans-sialidase.

O grupo TcS I (TS) apresenta proteínas que são expressas em estágios específicos do desenvolvimento de *T. cruzi*. Alguns membros são expressos nas formas não replicativas tripomastigota metacíclica, presente no inseto vetor, e na forma tripomastigota invasiva presente na corrente sanguínea do hospedeiro mamífero. Os membros desse grupo são muito variáveis, mas apresentam as duas regiões características, que são os domínios N-terminal (FRIP, Asp-box e VTVxNVfLYNR) e a maioria possui a repetição SAPA na extremidade C-terminal. Existem proteínas do grupo I da família TcS que não apresentam a repetição SAPA na extremidade C-terminal, sendo expressas na forma replicativa epimastigota, presente no inseto. Estas também apresentam atividade trans-sialidase. A falta da repetição SAPA nas TS-epi mostra que esse motivo não é necessário para a sua atividade de trans-sialidase (Frasch, 2000). Apesar das proteínas desse grupo apresentarem ligação com a mesma especificidade à beta-galactose, elas apresentam variação na sua atividade trans-sialidase. Somente alguns membros de TcS I apresentam a região N-terminal catalítica ativa enquanto que a região N-terminal dos outros membros apresenta uma mutação na posição 342 que provocou a mudança do resíduo triptofano para histidina, que acarreta perda da atividade trans-sialidase/sialidase (Cross e Takle, 1993; Schenkman *et al.*, 1994; Cremona *et al.*, 1999; Frasch, 2000).

Foi demonstrado que a repetição SAPA de TS aumenta a meia-vida da enzima na corrente sanguínea, e acredita-se que estas repetições retardem a formação de anticorpos contra o domínio ativo de TS impedindo assim o bloqueio da atividade trans-sialidase do parasito num primeiro momento (Frasch, 2000 Colli, 1993; Cazzulo e Frasch, 1992). Possivelmente essa seria a estratégia usada pelo parasita para evitar a inibição da atividade da enzima TS durante o processo de incorporação de ácido siálico a sua membrana. TS enzimaticamente ativas, co-expressas, e que

apresentam pouca diferença na sua sequência primária, poderiam levar a um atraso e resposta imune que bloqueasse a atividade da enzima devido à presença de polimorfismos na vizinhança do sítio ativo (Ratier *et al.*, 2008). Especula-se que as TcS de outros grupos poderiam auxiliar nessa estratégia para gerar uma resposta do sistema imune atrasada e ineficaz contra a atividade de TS (Ratier *et al.*, 2008).

Os outros grupos (TcS II, III e IV) apresentam regiões protéicas similares à TcS I, mas não apresentam atividade de trans-sialidase. Estes são incluídos na família TcS por apresentarem identidade de 30-40% entre suas sequências de proteínas e as proteínas TS (Schenkman *et al.*, 1994; Frasch, 2000). Além da identidade, os membros TcS grupos II, III e IV apresentam de uma a quatro repetições do motivo SxDxGxTW (x representa qualquer aminoácido), motivo sialidase, próximo a extremidade N-terminal e motivo VTVxNVfLYNR (x representa qualquer aminoácido) característico da família, podendo ser estes motivos degenerados.

Os membros do grupo TcS II são expressos na fase de tripomastigota, amastigota e tripomastigota metacíclica, formas que interagem com hospedeiro vertebrado. Como somente o grupo TcS II possui alguns membros expressos na forma amastigota esses membros são capazes de ativar resposta T-citotóxica, enquanto as variantes expressas nas formas tripomastigotas sanguíneas, tanto do grupo TcS II como dos outros grupos, induzem a produção de anticorpos (Frasch, 2000). As proteínas desse grupo (TcS II) apresentam diversas funções, relacionadas com interação entre parasita e hospedeiro, interagindo e permitindo adesão e invasão através da ligação com laminina (Giordano *et al.*, 1994), fibronectina (Ouaissi *et al.*, 1988), colágeno (Velge *et al.*, 1988; Santana *et al.*, 1997), citoqueratina-18 (CK-18) (Magdesian *et al.*, 2001) e superfície celular (Frasch, 2000; Tonelli *et al.*, 2010). A interação com citoqueratina-18 é importante para o parasito por promover o aumento na entrada do mesmo nas células do hospedeiro mamífero (Tonelli *et al.*, 2010). Trabalhos recentes mostram que o grupo TcS II apresenta proteínas envolvidas no tropismo tecidual e interações com vários tipos de filamentos intermediários. Acredita-se que o motivo VTVxNVxLYNR do grupo TcS II esteja envolvido no tropismo tecidual de *T. cruzi*, por tecidos cardíaco, esôfago, cólon e bexiga e interação com filamentos intermediários especificamente: vimentina, CK20 e CK8 (Tonelli *et al.*, 2010).

O grupo TcS III apresenta proteínas reguladoras do complemento, CRP (*complement regulatory protein*), dentre elas FL-160. Membros deste subgrupo são expressos na fase

tripomastigota e tripomastigosta metacíclica (Schenkman *et al.*, 1994; Beucher e Norris, 2008). As proteínas CRP bloqueiam a ativação das vias clássica e alternativa do complemento através da ligação com as proteínas C3b e C4b, desta forma inibindo a montagem da C3 convertase e formação do complexo MAC (*membrane attack complex*) (Frasch, 2000; Beucher e Norris, 2008; Tonelli *et al.*, 2010).

Os membros do grupo TcS IV são encontrados nas fases tripomastigosta e tripomastigosta metacíclica. Estes membros são caracterizados por apresentarem repetições curtas de cinco aminoácidos (EPKSA) na porção C-terminal (Cross e Takle, 1993; Schenkman *et al.*, 1994), que é reconhecida pelo soro de pacientes infectados (Frasch, 2000; García *et al.*, 2003; García *et al.*, 2006). As proteínas do grupo TcS IV não têm função conhecida, mas existem evidências de que Tc13 atuaria como ligante promovendo a interação com neurotransmissores através da repetição EPKSA (García *et al.*, 2003; García *et al.*, 2006).

Outras espécies do gênero *Trypanosoma* que possuem genes ortólogos da família trans-sialidase/sialidase-like apresentam algumas diferenças com relação às TcS de *T. cruzi*. *T. brucei* apresenta menor número de cópias da família e com identidade protéica de apenas de 38% com os ortólogos de *T. cruzi* (Montagna *et al.*, 2002). Apesar desta divergência, a enzima de *T. brucei* apresenta resíduos conservados para atividade enzimática e os motivos FRIP, Asp-box e VTVxNVxLYNR, mas não apresenta repetições SAPA (Montagna *et al.*, 2002). TS de *T. brucei* é expressa somente na forma procíclica encontrada no inseto vetor (gênero *Glossina*) e o ácido siálico é transferido para proteínas prociclina presentes na forma procíclica, conferindo provavelmente um papel protetor contra as enzimas digestivas do inseto (Buschiazzo *et al.*, 1997; Montagna *et al.*, 2002). Outra característica que difere a família trans-sialidase/sialidase-like destes dois parasitos é que a família em *T. brucei* tem atividade de sialidase e trans-sialidase codificadas em genes diferentes (Montagna *et al.*, 2006). Os ortólogos de *T. rangeli* apresentam grande identidade protéica com TcS de *T. cruzi* (68,9%) e alguns dos membros desta família, assim como em *T. cruzi*, não apresentam atividade enzimática. O número de grupos também é diferente entre as três espécies, a família de *T. rangeli* é classificada por similaridade de sequência em três grupos enquanto que em *T. brucei* é classificada em oito grupos (Buschiazzo *et al.*, 1997; Montagna *et al.*, 2002; Grisard *et al.*, 2010).

1.3.2. TcMUC Mucinas

A família TcMUC é composta por proteínas de superfície, principais aceptoras de ácido siálico, que formam uma densa e contínua cobertura na forma tripomastigota que é importante na interação entre parasito e hospedeiro (Frasch, 2000; Pereira-Chioccola *et al.*, 2000; Campo *et al.*, 2006). As TcMUC são proteínas altamente glicosiladas (proteínas O-glicosiladas) em resíduos de serina e treonina (Freitas-Junior *et al.*, 1998; Frasch, 2000; Acosta-Serrano *et al.*, 2001; Buscaglia *et al.*, 2006).

A família das mucina, assim como outras famílias de proteínas de superfície, apresenta muitos membros, que constituem uma família heterogênea codificada por genes dispersos pelo genoma do parasito (Freitas-Junior *et al.*, 1998; Frasch, 2000; El-Sayed *et al.*, 2005a). Foram identificados no projeto genoma de *T. cruzi* 863 cópias de TcMUC, incluindo 201 pseudogenes (El-Sayed *et al.*, 2005a). A grande maioria dos genes TcMUC estão localizados em regiões internas dos cromossomos.

A família das mucina pode ser dividida em dois grupos bem distintos (Buscaglia *et al.*, 2006). Estes grupos se diferenciam por serem constituídos por genes que (i) apresentam diferenças de expressão durante o ciclo de vida do parasita e (ii) que codificam proteínas que apresentam diferenças entre as suas sequências. O grupo de mucinas expressas pelo parasita durante as fases presentes no hospedeiro vertebrado (tripomastigota e amastigota) codificam proteínas que possuem de 80 até 200 kDa e são chamadas de TcMUC. O segundo grupo de mucinas expressas pelo parasita durante as fases presentes no hospedeiro invertebrado (tripomastigota metacíclica e epimastigota) apresenta menor variação do peso molecular; estas proteínas apresentam de 35 a 50 kDa e são denominadas TcSMUG. As TcMUC expressas, durante a fase do hospedeiro vertebrado são maiores e mais glicosiladas do que mucinas expressas durante a fase do hospedeiro invertebrado.

A estrutura de carboidrato do tipo glicana é ligada nos resíduos de Ser/Thr das TcMUC através de N-acetil-glicosamina (GlcNAc) diferentemente das mucinas de mamíferos que se utilizam de N-acetil-galactosamina (GalNAc) (Pereira-Chioccola *et al.*, 2000; Acosta-Serrano *et al.*, 2001).

Sítios para adição de GlcNAc estão localizados em regiões de grande variabilidade das TcMUC, devido a expansões ou encurtamento dos motivos repetitivos ricos em treonina/serina (Acosta-Serrano *et al.*, 2001).

Esses oligossacarídeos são altamente imunogênicos para humanos e representam a principal alvo para anticorpo lítico anti-α-Gal em pacientes na fase aguda e crônica da doença de Chagas (Acosta-Serrano *et al.*, 2001). Demonstrou-se que anticorpo lítico anti-α-Gal produz grandes danos, desestabilizando a membrana e provocando a lise dos parasitas em poucos minutos. Por outro lado, a presença do ácido siálico nas TcMUC confere ao parasita resistência contra o anticorpo lítico anti-α-Gal (Pereira-Chioccola *et al.*, 2000).

O ácido siálico presente nas TcMUC cria uma carga negativa na membrana externa do parasita (Acosta-Serrano *et al.*, 2001) e consequentemente altera as interações parasita-hospedeiro. As proteínas TcMUC ligadas ao ácido siálico promovem reconhecimento, invasão celular e proteção da forma tripomastigota, presente na corrente sanguínea (Acosta-Serrano *et al.*, 2001; Buscaglia *et al.*, 2006). Também foi demonstrado que TcMUC são capazes de ativar a resposta de Ca^{2+} nas células hospedeiras, evento associado com estágios iniciais da invasão celular de células não fagocíticas (Acosta-Serrano *et al.*, 2001).

O grupo TcMUC, expressas pelo parasita presente no hospedeiro vertebrado, possui três subgrupos, TcMUC I a III (Buscaglia *et al.*, 2006). A subdivisão ocorre com base na presença de determinados motivos e similaridade desses motivos. As TcMUC possuem uma região central altamente variável em tamanho e sequência. Essa região central contém número variável de repetições ricas em treonina que possuem sítios de ligação GlcNAc; sendo que estas repetições podem ou não ser organizadas em série (*tandem*). As repetições são restritas a TcMUC I e II e são encontradas na região variável. A repetição $T_{(6–8)}KP_{(1–2)}$ é encontrada em TcMUC I que podem ocorrer de duas a 10 vezes. Já T_8KAP/T_8QAP ocorre de uma a duas vezes em TcMUC II. Além destes motivos ricos em treonina característicos de cada um dos grupos TcMUC I e TCMUC II, existe uma região altamente variável próximo à extremidade N-terminal que diferencia os dois grupos de TcMUC. No grupo TcMUC I esta região é curta, enquanto TcMUC II é longa (Buscaglia *et al.*, 2006).

Ao analisar os grupos TcMUC I e TcMUC II, observa-se que as diferenças entre os seus membros apresentam uma variação contínua, onde alguns membros apresentam características intermediárias entre os dois grupos (Buscaglia *et al.*, 2006).

Ainda existe o grupo TcMUC III, também conhecido como pequeno antígeno de superfície de tripomastigota, TSSA. Esse grupo apresenta proteínas com peso molecular de 20 kDa e sua região central é rica em resíduos de treonina, serina e prolina, mas diferentemente dos outros grupos TcMUC, estes resíduos não são organizados em repetições (Acosta-Serrano *et al.*, 2001; Buscaglia *et al.*, 2006).

Não existem membros TcMUC homólogos em *T. brucei*, mas existem oito genes em *L. major* que apresentam estrutura similar com TcMUC I. Estes genes de *L. major* possuem repetições T_7KP_2 (Freitas-Junior *et al.*, 1998).

Outro grupo de mucinas, TcSMUG, é expresso na forma epimastigota, onde há detecção de uma camada de glicoproteína menos densa e menos espessa devido às TcSMUG serem menores que TcMUC (Pereira-Chioccola *et al.*, 2000).

O grupo TcSMUG é composto por um conjunto de 19 genes e estes se apresentam menos polimórficos que TcMUC (Freitas-Junior *et al.*, 1998; Frasch, 2000; Acosta-Serrano *et al.*, 2001; Turner *et al.*, 2002; El-Sayed *et al.*, 2005a; Buscaglia *et al.*, 2006).

Acredita-se que as TcSMUG podem proteger os parasitas contra enzimas proteolíticas presentes no intestino do inseto e baixo pH (Acosta-Serrano *et al.*, 2001). Esta ação é evidenciada pelas TcSMUG de tripomastigotas metacíclicas serem resistentes à proteólise (Acosta-Serrano *et al.*, 2001), e resistentes a compostos oxidantes produzidos pelo hospedeiro invertebrado.

TcSMUG pode ser dividido em dois grupos baseado no tamanho do transcrito e tipos de repetições na região variável. TcSMUGL (TcSMUG large) possui longo mRNA e repetições do tipo KNT_7ST_3S(S/K)AP, enquanto que TcSMUG-S (TcSMUG-small) possui transcritos menores e repetições do tipo $DQT_{17–20}$NAPAKDT_{5-7}NAPK.

1.3.3. Mucin-associated surface protein (MASP)

MASP é a segunda maior família de proteínas de *T. cruzi* foi identificada durante o desenvolvimento do projeto genoma. Os membros dessa família são frequentemente encontrados

próximos a genes da família TcMUC e por esta razão a família recebeu o nome de "*mucin-associated surface proteins*" (MASPs) (Atwood *et al.*, 2005; El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009). As características marcantes dessa família de proteínas são: (i) extremidades N- e C-terminal altamente conservadas que codificam o peptídeo sinal e sinais de ancoragem da proteína através de GPI à membrana plasmática, respectivamente, sugerindo que a proteína seja direcionada para a membrana plasmática do parasita (Atwood, Weatherly *et al.*, 2005; El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009) e (ii) região interna altamente variável e repetitiva. A família MASP apresenta 1377 cópias no genoma de *T. cruzi,* incluindo 443 pseudogenes (El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009). Os genes MASP ainda podem ser subdivididos em dois grupos, sendo eles: (i) dos genes intactos (771 cópias), que possuem as suas extremidades N- e C-terminal conservadas; e (ii) dos genes chamados quimeras (39 cópias), que não possuem conservação na sequência de uma das suas extremidades N- ou C-terminal (Bartholomeu *et al.*, 2009).

A região 3'UTR é também muito conservada em toda sua extensão (Bartholomeu *et al.*, 2009). As extremidades N- e C-terminais conservadas de MASP apresentam similaridade com as extremidades N- e C-terminal de mucinas (TcMUC). As sequências consenso N- e C-terminal nas duas família apresentam 57% e 38% de identidade, respectivamente (Bartholomeu *et al.*, 2009).

Essa identidade entre as regiões N- e C-terminal das duas famílias (TcMUC e MASP) e a presença de um motivo degenerado *Pfam* (PF01456) característico das mucinas de diversos organismos e em alguns membros de MASP sugerem que possivelmente a família MASP evoluiu da TcMUC, seguido de diferenciação e um processo intenso de duplicação gênica e nova diferenciação (Bartholomeu *et al.*, 2009).

A região central das proteínas da família MASP é muito variável e essa variabilidade é devido a três aspectos: (i) tamanho das proteínas, que podem variar de 176 a 645 aminoácidos (ii) variação das sequências protéicas e (iii) repertório de motivos repetitivos.

Análises de expressão gênica de RNA e proteínas mostraram que os genes MASPs são expressos preferencialmente na forma tripomastigota do parasito (Atwood *et al.*, 2005; El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009). Esses resultados de expressão de MASP na forma

tripomastigota associados ao seu extenso polimorfismo sugerem o envolvimento desta família na interação entre o parasita e hospedeiro vertebrado.

T. brucei e *Plasmodium falciparum* apresentam um mecanismo sofisticado de apresentação de proteínas que permite a evasão do sistema imune através da expressão coordenada e exclusiva dos membros das famílias VSGs de *T. brucei* (Berriman *et al.*, 2005) e as VARs de *Plasmodium falciparum* (Gardner *et al.*, 2002). Esse sistema usa um repertório grande de genes com grande diversidade para evitar o reconhecimento pelo sistema imune e permanecer no hospedeiro vertebrado. Além dos genes completos, o sistema usa de recombinação com genes e pseudogenes das respectivas famílias para gerar novas cópias. Diferentemente das famílias VSGs e VAR, os genes MASP são localizados na região central dos cromossomos (Bartholomeu *et al.*, 2009). A localização dos genes MASP na região interna dos cromossomos estaria em concordância com a falta de mecanismos de variação antigênica em *T. cruzi;* que diferentemente dos genes envolvidos em variação antigênica em *T. brucei* e *P. falciparum* estão localizados em regiões subteloméricas. Os genes MASP são encontrados em regiões que não apresentam sintenia, comparando os genomas de *T. cruzi*, *T. brucei* e *L. major*. Nessa região central dos cromossomos de *T. cruzi* também são encontrados outros genes codificadores de proteína de superfície como TcS, GP63, SAP e DGF-1 e retroelementos como L1Tc, NARTc, DIRE VIPER e SIRE (El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009).

Além da localização preferencial na região interna dos cromossomos, os genes MASP também apresentam localização preferencial em cromossomos grandes (3,5 a 2,0 Mb), o que também ocorre com os genes da família TcMUC e SAP. Verificou-se ainda que os genes de MASP estão frequentemente entre genes de TcMUC II e TcS. Possivelmente esta organização gênica pode ser uma estratégia para evitar a homogeneização da família; o que poderia acontecer através da recombinação genética entre genes próximos, diluindo desta forma a diversidade genética (Bartholomeu *et al.*, 2009).

Os genes quimeras, encontrados na família MASP, possuem sequências que são compartilhadas com outros genes como TcMUC, SAP e TcS. Algumas dessas quimeras foram encontradas em bibliotecas de cDNA (El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009), o que demonstra que esses genes estão sendo expressos e que provavelmente processos de

recombinação não homóloga pode contribuir para a geração de diversidade dentro da família MASP.

Análises proteômicas mostraram que pelo menos alguns membros da família MASP são N-glicosilados (Atwood *et al.*, 2005). Análises *in silico* identificaram vários sítios potenciais de O-glicosilação e outros sítios potenciais de N-glicosilação, sendo que os sítios de N-glicosilação foram encontrados em quase todas as proteínas MASP (El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009). Acredita-se que as MASP podem sofrer outras modificações pós-traducionais, como evidenciado por seus potenciais sítios de fosforilação (Bartholomeu *et al.*, 2009).

1.3.4. Metaloprotease GP63

As proteínas de superfície do tipo metaloproteases, dentre elas a GP63, foram identificadas primeiramente em *Leishmania* (Bouvier *et al.*, 1985; Etges *et al.*, 1986; Cuevas *et al.*, 2003; Yao *et al.*, 2003). A família GP63 corresponde a 1% das proteínas produzidas na fase promastigota em *Leishmania major* e *Leishmania mexicana* (Bouvier *et al.*, 1985; Etges *et al.*, 1986).

As proteínas GP63 em *T. cruzi* formam uma família multigênica de 425 membros incluindo 251 pseudogenes (Cuevas *et al.*, 2003; Yao *et al.*, 2003; El-Sayed *et al.*, 2005a). Em *T. cruzi* esta família apresenta um número muito maior de membros quando comparado aos outros tripanosomatídeos. Esses genes estão geralmente dispersos no genoma, embora existam repetições em série (El-Sayed *et al.*, 2005a). Esta família possui alguns de seus membros com atividade de metaloprotease dependente de Zn^{2+}.

Os genes de GP63 são diferencialmente expressos durante as fases de desenvolvimento de *T. cruzi* assim como em *Leishmania spp* (Grandgenett *et al.*, 2000; Cuevas, Cazzulo *et al.*, 2003; Kulkarni, Olson *et al.*, 2009). Estudos recentes em *T. cruzi* (Yao *et al.*, 2003; Kulkarni *et al.*, 2009) identificaram as formas epimastigota, tripomastigota e amastigota expressando isoformas de 61 kDa enquanto tripomastigota metacíclica expressando isoforma com 55 kDa, ambas na cepa Y. Além da diferença no peso molecular das GP63 identificadas por Kulkarni (2009) na cepa Y também foram descritas diferenças quanto a sua glicosilação e localização celular, tendo sido encontrada GP63 de 61 kDa glicosilada na superfície do parasita e flagelo, enquanto GP63

de 51 kDa não glicosilada foi localizada no citoplasma. A família GP63 apresenta alguns de seus membros ancorados à membrana plasmática do parasita, através de âncora de GPI (Cuevas *et al.*, 2003) enquanto outros membros são localizados na superfície celular através de domínios transmembrana (Kulkarni *et al.*, 2009).

Acredita-se que o papel principal da atividade proteolítica de GP63 seja relacionada à clivagem das macromoléculas do hospedeiro, o que conferiria proteção e/ou forneceria nutrientes ao parasita. Foi demonstrado também que GP63 está envolvida na infecção das células do hospedeiro e que anticorpos contra GP63 bloqueiam parcialmente a invasão de células por tripomastigotas, demonstrando que GP63 tem um importante papel durante a adesão e/ou a invasão das células do hospedeiro por *T. cruzi* (Cuevas *et al.*, 2003; Yao *et al.*, 2003; Kulkarni *et al.*, 2009).

Os membros da família GP63 de *T. cruzi* são separados em três grupos: GP63-I, Gp63-II e GP63-III (este último formado por pseudogenes) (Cuevas *et al.*, 2003). Os membros GP63-I apresentam 543 aminoácidos, peptídeo sinal, região C-terminal hidrofóbica e sinal para ancoragem por GPI. Os membros do grupo GP63-II apresentam 566 aminoácidos e peptídeo sinal. O grupo GP63-II é o que apresenta mais membros, no entanto estes apresentam um nível menor de expressão de mRNA. As proteínas da família GP63-II apresentam dois a três sítios potenciais de N-glicosilação, suas regiões C-terminal não apresentam aminoácidos hidrofóbicos e tampouco sítio de ancoragem por GPI (Cuevas *et al.*, 2003).

As sequências protéicas de GP63 de *T. cruzi* são conservadas, apresentando identidade de 42% entre os membros do grupo GP63-I e os do grupo II (Cuevas *et al.*, 2003). Todos os membros de GP63 apresentam o motivo metaloprotease HExxH (Grandgenett, *et al.*, 2000) (x representa qualquer aminoácido). Neste os resíduos de histidina e glutamina são totalmente conservados e X representa qualquer aminoácido. Apresentam também vários resíduos de cisteína ao longo da sequência totalmente conservados (Grandgenet *et al.*, 2000; Cuevas *et al.*, 2003). A maior diferença entre os genes dos grupos de GP63-I está nas suas regiões 3'UTRs, que apresentam alta diversidade apresentando diferença de 175 bases entre GP63-Ia e GP63-Ib. Transcritos e proteínas de GP63-I foram detectados em todas as fases de infecção de *T. cruz*i, mas com maior abundância em epimastigota e amastigota, além disso o nível de mRNA apresentou diferença entre as linhagens do parasito. Baixos níveis de expressão dos transcritos do

grupo GP63-II foram detectados em todas as fases do desenvolvimento, com discreto aumento na fase amastigota (Cuevas *et al.*, 2003).

Os ortólogos de GP63 em *T. cruzi* e outros tripanosomatídeos *T. brucei*, *Leishmania guyanensis*, *Crithidia fasciculata* apresentam de 30 a 38% de identidade comparando GP63-Ia e 26 a 31% de identidade comparando GP63-II. Nestas sequências há conservação do sítio catalítico HExxH e de 18 cisteínas, indicando a preservação da estrutura da proteína (Cuevas *et al.*, 2003). Comparando GP63 entre *T. cruzi* e *T. rangeli* foram encontrados índices muito maiores de identidade, variando de 38 a 65% (Ferreira *et al.*, 2010). As diferenças entre GP63 dessas diferentes espécies (*T. cruzi, T. brucei, T. rangeli, L. guyanensis* e *Crithidia fasciculata*) se concentraram nas extremidades N- e C-terminal das proteínas.

1.3.5. Proteínas ricas em Serina-, Alanina-, e Prolina- (SAP)

A família SAP foi identificada através de análise de sequências genômicas e cDNA de formas tripomastigotas metacíclicas (Carmo *et al.*, 2001). Algumas de suas cópias são expressas nas formas amastigota e epimastigota (Baida *et al.*, 2006). As proteínas dessa família apresentam suas sequências enriquecidas com os aminoácidos serina (13.61%), alanina (13.02%), prolina (11.24%), glicina (9.47%) e leucina (8.28%) e por este motivo receberam a denominação de proteínas ricas em Serina-, Alanina-, e Prolina-, (SAP) (Carmo *et al.*, 2001; Baida, Santos *et al.*, 2006). As proteínas da família SAP localizam-se na superfície do parasito e apresentam capacidade adesiva e de invasão a células de mamíferos pela forma tripomastigota metacíclica (Baida *et al.*, 2006).

Os aminoácidos serina, alanina e prolina são encontrados em diversos tipos de repetições (AAS, AAP, AAS, APS, SAA, SSA, SPP, PAP, AAPP, SAPA, SAAP, SSPA, SSAP, SAAA, APPPP, SAAAS, PPSPP, SSPPA, AAAPP, PSAAAS, PPPPPPA, SASAASPA, SASAASAA, APPPPPPA, e SASAASSPA) (Carmo *et al.*, 2001; Baida, Santos *et al.*, 2006). A família SAP apresenta estas repetições, além de um domínio central de 55 aminoácidos, presente em todos os membros da família. A presença destas repetições mais o domínio central servem como assinatura dos membros da família (Baida *et al.*, 2006).

O peso molecular predito das proteínas da família SAP é de aproximadamente 37 kDa (Carmo *et al.*, 2001). Diferenças entre o peso molecular predito e peso molecular da proteína nativa foram detectadas, sugerindo a ocorrência de modificações pós-traducionais como sugerido pela presença de dois a cinco potenciais sítios de N-glicosilação, 17 a 41 potenciais sítios de O-glicosilação e 27 a 36 potencias sítios de fosforilação, presentes nas sequências destas proteínas (Carmo *et al.*, 2001; Baida *et al.*, 2006).

A família SAP é classificada em quatro grupos (SAP 1-4) usando características das suas regiões N- e C-terminal (Baida *et al.*, 2006). O grupo SAP 1 possui 31 membros com peso molecular predito de aproximadamente 38 kDa, apresentando extremidades N- e C-terminal conservadas, com peptídeo sinal e âncora de GPI nas extremidades N- e C-terminal, respectivamente. O grupo SAP 2 apresenta dois membros, e são menores em relação aos membros do grupo SAP 1, com peso molecular de aproximadamente 32 kDa, mas diferentemente do que ocorre no grupo SAP 1 não apresentam âncora GPI. O grupo SAP 3 é formado por cinco membros, com peso molecular predito de aproximadamente 44 kDa, estas apresentam âncora de GPI mas não apresentam peptídeo sinal. O grupo SAP 4 apresenta somente um membro, com peso molecular predito de 67 kDa, não apresenta peptídeo sinal ou sítio de adição de âncora GPI e sua extremidade N-terminal é divergente em relação a mesma região dos outros grupos. Isto ocorre por esta proteína ser descrita como uma quimera, contendo domínio da proteína gag (do retroelemento L1Tc) (Baida *et al.*, 2006).

A família SAP é formada por 39 genes completos, seis pseudogenes e quatro genes parciais (El-Sayed *et al.*, 2005a; Baida *et al.*, 2006). A família SAP é considerada espécie-específica, como indicado por busca por homólogos em outras espécies do gênero *Trypanosoma* especificamente usando sondas e iniciadores específicos para a família SAP (Baida *et al.*, 2006).

Outros organismos da classe Kinetoplastida, não pertencentes ao gênero *Trypanosoma,* também apresentam proteínas ricas em serina, alanina e prolina. Nestes organismos estas são encontrados como proteínas glicosiladas associadas à membrana. As proteínas da família SAP apresentam 30% de identidade com proteofosfoglicanas associadas à membrana de *L. major,* estas apresentam no seu domínio central de repetições APSASSSS e APSSSSSS (Carmo *et al.*, 2001). As proteínas da família SAP ainda apresentam motivos que são parcialmente similares às

repetições de serina, alanina e prolina encontradas nos proteoglicanos ppg1 e fPPG de *L. major* e ppg1, ppg2 e fPPG de *L. mexicana* (Carmo *et al.*, 2001).

Tem sido descrito proteínas SAP quiméricas possuindo sequências de aproximadamente 180 aminoácidos que apresentam identidade variando de 45 a 65% com genes da família MASP, e ainda sequências de aproximadamente 80 aminoácidos com identidade variando de 51 a 81% com proteínas da família mucina (TcMUC e TcMUC II), esta última na extremidade N-terminal da proteína (Baida *et al.*, 2006).

1.3.6. Dispersed gene family-1 (DGF-1)

A família *Dispersed Gene Family*-1 (DGF-1) é a quinta maior família de proteínas de *T. cruzi* e não apresenta genes homólogos em *T. brucei* e *L. major*. Vários membros apresentam peptídeo sinal. Outra característica da família DGF-1 é a presença em todos os membros completos de nove hélices transmembranas muito conservadas na extremidade C-terminal (Kawashita *et al.*, 2009). Novos trabalhos mostram localização de DGF-1 no meio de cultura indicando sua secreção por meio de vesículas (Lander *et al.*, 2010).

Os genes estão localizados em regiões subteloméricas (El-Sayed *et al.*, 2005a) e regiões internas dos cromossomos próximos a genes MASP e outras famílias gênicas que também codificam proteínas de superfície (Bartholomeu *et al.*, 2009). A família é composta de 565 genes, desse total, 136 são pseudogenes (El-Sayed *et al.*, 2005a).

Ainda não se conhece muito sobre o papel das proteínas DGF-1, mas existem indícios de que ela seja importante na interação parasita-hospedeiro. Existem ainda evidências de expressão das proteínas DGF-1 na superfície de tripomastigotas. As proteínas DGF-1 apresentam motivos similares com a proteína integrina beta 7 humana (Kawashita *et al.*, 2009). As proteínas integrinas são conhecidas por seu papel na adesão célula-célula e célula-matriz e também atuam na transdução de sinais entre o citoplasma e matriz extracelular. Dessa forma esses motivos similares a integrinas poderiam se ligar a componentes da matriz extracelular como fibronectina e laminina (Kawashita *et al.*, 2009).

Análises filogenéticas, usando cerca de 130 sequências, apontaram que a família poderia ser dividida em pelo menos três grupos (A, B e C-E), sendo dois grupos, A e B, com maior

número de representantes (66 e 51 sequências, respectivamente) (Kawashita *et al.*, 2009). Os três principais grupos dentro da família DGF-1 ainda podem se divididos em subgrupos (A1, A2, A3, A4, A5, B1, B2, C, D e E), mas os grupos A1 e B2 apresentaram conflitos em relação à posição em networks apresentando reticulações, mostrando que não foi possível definir com segurança as relações dos grupos (Kawashita *et al.*, 2009).

Além das várias duplicações e mutações, as reticulações encontradas na rede filogenética são evidências de que essa família passou por eventos de recombinações, hibridações e convergência gênica envolvendo genes completos e pseudogenes (Kawashita *et al.*, 2009).

1.3.7. RHS

Elementos transponíveis podem se proliferar e aumentar o número de cópias, contribuindo para evolução e rearranjos do genoma. Retrotransposons são elementos que se movem pelo genoma através de um intermediário de RNA, copiando o transcrito de RNA para DNA através de uma transcriptase reversa (RT), sendo o DNA copiado integrado ao genoma. Os retrotransposons são divididos em dois grupos: retrotransposons com *long terminal repeats* (LTR transposons) e retrotransposons sem LTR (non-LTR transposons).

Retroelementos *long terminal repeat* (LTR) e non-LTR correspondem a 5% do genoma haplóide de *T. cruzi* e 2% de *T. brucei* (El-Sayed *et al.*, 2005a). Ingis e L1Tc são Non-LTR transposons presentes em *T. brucei* e *T. cruzi,* respectivamente (Bringaud *et al.*, 2002; Barry *et al.*, 2006) e não estão dispersos aleatoriamente pelo genoma, mas sim associados a genes RHS (*retrotransposon hot spot protein*) que são realmente pontos frequentes de inserção desses retrotransposons (El-Sayed *et al.*, 2005a; Barry *et al.*, 2006).

A alta frequência de retrotransposons inseridos nos genes RHS é resultado de uma sequência consenso e (5´-AxxAxGaxxxxxtxTATGAxxxxxxxxxxx-3´) presente nos genes RHS de *T. cruzi* e usada pelo retrotranposons L1Tcpara transposição (El-Sayed *et al.*, 2005a; Barry *et al.*, 2006).

Os genes RHS estão associados a regiões subteloméricas e regiões internas dos cromossomos.

A família RHS de *T. cruzi* é composta de 752 genes, sendo 557 pseudogenes (El-Sayed *et al.*, 2005a). As proteínas RHS são altamente expressas na forma epimastigota em *T. cruzi* (Barry *et al.*, 2006) e em *T. brucei* estão localizadas na região nuclear (Bringaud *et al.*, 2002).

1.4. Evolução das famílias gênicas

As famílias gênicas podem evoluir pela duplicação de genes e posterior diferenciação. Estudos genômicos mostram que a linhagem que originou o *T. cruzi* passou por um intenso processo de expansão das famílias gênicas de proteínas de superfície, algumas famílias apresentando até milhares de cópias no genoma (El-Sayed *et al*., 2005; Barry *et al.*, 2006). Essa expansão das famílias de proteínas de superfície em *T. cruzi* aconteceu tanto para proteínas compartilhadas com outros organismos como para proteínas espécie específica (El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009; Kawashita *et al.*, 2009).

Os genes de uma família podem se diferenciar através de vários processos como mutação e recombinação. As mutações ocorrem durante a replicação, levando a mudanças na sequência de DNA. As alterações na sequência codificadora de um gene podem não afetar a sequência de proteínas (substituições sinônimas) devido à redundância do código genético. Outras substituições de bases na sequência do DNA podem levar a alterações nas sequências protéicas (substituições não sinônimas).

Mutações podem levar à formação de pseudogenes gerando códons de parada de tradução prematuros, alteração da fase de leitura devido a inserções ou deleções de bases (*indel*). As grandes famílias gênicas em protozoários como *T. cruzi* e *T. brucei* apresentam centenas de pseudogenes (Berriman *et al.*, 2005; El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009; Kawashita *et al.*, 2009), provavelmente formados por esses erros da replicação e/ou recombinações entre os membros da família.

O processo evolutivo das famílias gênicas ainda pode passar por recombinações, que permite a troca de fragmentos entre genes. Esse processo foi evidenciado em *T. cruzi* para genes da família DGF-1 onde foi produzida uma rede filogenética onde mostra que determinadas características são compartilhadas por genes filogeneticamente separados (reticulação) (Kawashita *et al.*, 2009). Esse compartilhamento de características pode ser explicado por

recombinação ou por convergência gênica. A intensidade de reticulações encontradas diminui a possibilidade de convergência evolutiva e sendo explicada através de eventos de recombinação.

Essas recombinações podem acontecer mais frequentemente entre genes que pertencem a mesma família por apresentarem sequências similares. Mutação e recombinação já foram descritos para protozoários parasitas como eventos chaves na rápida geração de variabilidade em famílias gênicas de proteínas de superfície. *T. brucei* apresenta a sua superfície VSGs responsável pela variação antigênica e permanência da infecção. Essa família é formada por centenas de genes. A análise de 806 genes de VSGs revelou que somente 57 (7%) genes são completamente funcionais enquanto que 66% são pseudogenes (Berriman *et al.*, 2005). O restante codifica proteínas funcionais atípicas (9%), talvez apresentando enovelamento anormal ou modificações pós-traducionais, ou genes incompletos (18%). *T. brucei* expressa uma única VGS por vez e possui mecanismos de recombinação entre genes da família VSG que permitem a troca de uma região codificadora completa ou partes menores no sítio de expressão, modificando assim as proteínas apresentadas na superfície (Palmer e Brayton, 2007).

Recombinações bem menos frequentes entre genes com sequências diferentes podem acontecer levando a troca de motivos entre essas sequências acelerando o processo de diferenciação gênica. Esse tipo de alteração aconteceu entre membros de diferentes famílias que codificam proteínas de superfície em *T. cruzi*, originando sequências quiméricas, contendo fragmentos de diferentes famílias gênicas. Isso pode ser observado na família MASP em que alguns genes apresentam fragmentos compartilhados com genes das famílias TcMUC, TcS e SAP. Algumas dessas quimeras foram detectados em análises de uma biblioteca de cDNA mostrando que essas cópias são transcritas (Bartholomeu *et al.*, 2009).

1.5. Medidas de distância e diversidade

Na ausência de recombinação, a distância evolutiva separando duas sequências pode ser definida como o número de substituições (alterações entre as sequências) por sítio ao longo da história evolutiva após a divergência das duas sequências (Nei, 1996; Brocchieri, 2001).

Estimativas de diversidade e distância molecular, seja usando sequências nucleotídicas ou protéicas, são importantes para contribuir no entendimento dos processos evolutivos a que foram

submetidas as sequências estudadas (Halpern e Bruno, 1998). Essas distâncias podem ser usadas para inferir árvores filogenéticas, diversidade e divergência entre sequências (Tamura e Kumar, 2002). Análises de distância também são importantes para entender padrões evolutivos em famílias multigênicas e evolução adaptativa encontradas usando dados moleculares (Nei, 1996).

Distâncias são métricas que sumarizam as diferenças em uma média geral de diferenciação entre as sequências (Kalinowski, 2002). Normalmente uma matriz de distância entre as sequências é estimada e pode ser usada para produzir gráficos que permitem a visualização das relações entre as sequências e grupos, podendo ser representado por árvores (Nei, 1996), análise de componentes principal e escala multidimensional (MDS) (Kalinowski, 2002).

A estimativa da diversidade e distância depende do número de substituições que ocorreram entre as sequências analisadas, essas medidas de diversidade seriam corretamente mensuradas conhecendo todas as substituições que aconteceram ao longo da evolução das sequências. Entretanto, o número de substituições entre as sequências não é conhecido levando ao desenvolvimento de diferentes métodos para estimar o número de substituições. Esses métodos foram propostos tanto para estudar sequências de DNA (Tajima, 1993; Nei, 1996; Kalinowski, 2002) quanto sequências de proteínas (Brocchieri, 2001), levando em consideração as características de cada tipo de sequências.

Os métodos de distância-p e Jukes-Cantor (Jukes e Cantor, 1969) apresentam melhores estimativas de distância em relação a métodos mais sofisticados quando a taxa de substituição é quase constante em todas as linhagens evolutivas analisadas. Nesses casos, o uso do método distância-p é mais adequado comparado com os métodos mais sofisticados.

Entretanto quando as taxas de substituição variam muito entre as linhagens evolutivas, se faz necessário o uso de métodos mais sofisticados (Nei, 1996).

O método Kimura (Kimura, 1980) é útil para correção de mutações reversas porque o modelo assume que a taxa de transição é diferente da taxa de transversão, apresentando melhores estimativas do número de substituições por sítio do que a simples proporção de diferenças entre as sequências (distância-p) (Nei, 1996). Entretanto essa estimativa normalmente apresenta uma variação maior em relação ao método não corrigido de distância-p ou um método simples de distância Jukes-Cantor (Nei, 1996).

Vários métodos para estimar o número de substituições por sítio não podem ser aplicados quando a distância entre as sequências se torna muito grande porque esses métodos normalmente envolvem termos logarítmicos na fórmula matemática e os argumentos do logaritmo se tornam negativos (Tajima e Nei, 1984; Nei, 1996).

Medidas de diversidade e árvores filogenéticas normalmente são construídas com distâncias par-a-par cujos valores são pequenos podendo nesses casos usar os métodos como distância-p, Kimura. Entretanto, se as distâncias se tornarem muito grandes ou houver evidências de que as taxas de substituições variam muito entre as linhagens evolutivas, os métodos mais sofisticados devem ser usados (Nei, 1996).

O método distância-p pode ser usado tanto para sequências de DNA quanto proteínas. Além dos métodos distância-p, outros foram desenvolvidos para se medir distâncias entre sequências protéicas (Zuckerkandl e Pauling, 1965; Gonnet *et al.*, 1992; Jones *et al.*, 1992; Dayhoff *et al*., 1978). Alguns levam em consideração que a taxa de substituição entre todos os sítios e aminoácidos é igual (Zuckerkandl e Pauling, 1965). Outros métodos usam de matrizes de substituição para pontuar as substituições e determinar as distâncias entre as sequências (Gonnet *et al.*, 1992; Jones *et al.*, 1992; Dayhoff *et al.*, 1978).

1.6. Agrupamento (clusterização - clustering)

O agrupamento (*clustering*) é uma técnica para análise e divisão de conjunto de dados em grupos similares. Os grupos ou conjuntos de dados similares são conhecidos como grupos (*clusters*). Análises de classificação permitem a identificação dos grupos e dos elementos que pertencem a cada um dos grupos automaticamente.

O agrupamento para segmentação de dados de acordo com categorias é geralmente usado para entender e fornecer um panorama geral, especialmente quando a quantidade de dados trabalhada é muito grande (Varshavsky *et al.*, 2008). O agrupamento de dados pode ser usado em uma grande variedade de aplicações. A grande quantidade de dados produzidos em biologia genômica, transcriptômica, proteômica e outras áreas necessitam de algoritmos eficientes de classificação (D'haeseleer, 2005; Varshavsky *et al.*, 2008).

As classificações podem ser divididas como hierárquica ou particional (Varshavsky *et al.*, 2008). Algoritmos de classificação hierárquica dividem os dados em grupos hierárquicos aninhados. Os algoritmos particionais separam os dados em grupos não aninhados.

1.6.1. Hierárquico

Os métodos hierárquicos têm duas abordagens para reconhecer grupos, repetindo ciclos de fusão de grupos menores formando grupos maiores ou dividindo grupos maiores em grupos menores, esses métodos são chamados aglomerativo e divisivo, respectivamente (Varshavsky *et al.*, 2008). A abordagem aglomerativa ("de baixo para cima", *bottom-up*) começa com cada elemento representando um cluster e os clusters vão se fundir até determinado critério seja alcançado (Jain *et al*., 1999). A abordagem divisiva ("de cima para baixo", *top-down*) inicia com todos os elementos pertencendo ao mesmo cluster (único cluster) e divisões são realizadas até determinado critério seja alcançado (Jain *et al*., 1999). As distâncias entre os elementos e clusters comumente usadas são distância euclidiana e distância Manhattan (D'haeseleer, 2005; Jain *et al*., 1999).

Os resultados de clusterização hierárquica podem ser representados por dendrogramas representando o agrupamento aninhado de padrões e níveis de similaridade ao longo do dendrograma (Varshavsky *et al.*, 2008; Jain *et al*., 1999). O dendrograma pode ser dividido em um determinado nível de similaridade, resultando em um determinado número de clusters para o conjunto de dados (Varshavsky *et al.*, 2008; Jain *et al*., 1999). A divisão do número de clusters pode variar ao longo do dendrograma.

1.6.2. Particional

Algoritmos de classificação particional (não hierárquico) geram várias partições e depois avalia o resultado de acordo com algum critério. O resultado do uso desse tipo de algoritmo é apenas um conjunto de clusters. Algoritmos particionais necessitam receber antes o número de grupos desejado, chamado de *k* (D'haeseleer, 2005). O algoritmo *kmeans* (também chamado de

K-médias) (MacQueen 1967) é um dos mais usados. Dado o número *k,* o algoritmo objetiva dividir *n* elementos do conjunto de dados entre os *k* grupos. O algoritmo analisa os dados e cria classificações, criando classes (cluster) e diz quais elementos pertencem a estas classes (Jain *et al.*, 1999).

O método inicia com um conjunto *k* fornecido pelo usuário escolhido como centróides iniciais, ou médias, dos agrupamentos. Portanto a decisão inicial de quantos clusters existe dentro do conjunto de dados é um ponto muito importante para obter resultados de qualidade. Geralmente escolhem-se os *k* primeiros elementos da tabela de dados como centróides iniciais. Cada centróide inicial é um agrupamento com apenas um único elemento. Em um segundo passo é calculado a distância entre os elementos (*n*) e cada centróide (*k*), depois os elementos são atribuídos como membro do centróide mais próximo, o centróide que está mais perto deste elemento vai incorporá-lo. Em um terceiro passo os centróides são atualizados, calculando-se a média de todos os elementos que foram atribuídos ao agrupamento correspondente. O segundo e terceiro passos são repetidos até que os centróides não sofram mais alteração (Jain *et al.*, 1999; D'haeseleer, 2005).

O resultado será uma classificação que coloca cada elemento em apenas um grupo (*k*). Existem outros algoritmos que classificam o quanto cada elemento pertence a cada um dos *k* grupos. Dessa forma existem valores para cada elemento que indica o quanto ele pertence a cada um dos grupos (*k*). Desta maneira teremos dois tipos de classificação, chamadas de *hard* e *fuzzy*. A classificação que atribui cada elemento em apenas um grupo faz uma classificação hard (*hard clustering*) uma vez que cada ponto só pode ser classificado em uma classe (Jain *et al.*, 1999). Outros algoritmos trabalham com o conceito de classificação fuzzy onde existe uma métrica que diz o quão 'dentro' de cada classe o ponto está (Jain *et al.*, 1999).

2. Justificativa

Análises comparativas dos genomas dos Tri-Tryps (os tripanosomatídeos *Trypanosoma cruzi*, *Leishmania major* e *Trypanosoma brucei*) revelam que a grande maioria das proteínas espécie-específicas corresponde àquelas localizadas na superfície dos parasitos (El-Sayed *et al.*, 2005b). Comparado a *T. brucei* e *L. major*, *T. cruzi* apresenta uma excepcional proporção de genes codificadores de proteínas de superfície (~20%). A maioria das grandes famílias de proteínas de superfície de *T. cruzi* apresenta extrema diversidade de sequência como as MASPs, Trans-sialidases e Mucinas (El-Sayed *et al.*, 2005a; Bartholomeu *et al.*, 2009). É provável que a forte pressão seletiva para diversificação destas famílias possa ser imposta pelo sistema imune do hospedeiro ou ser consequência de uma outra estratégia essencial de sobrevivência do parasito. Sabe-se que o parasito co-expressa, em um determinado momento, várias famílias e diversos membros de uma mesma família polimórfica. Especula-se que a exposição simultânea de uma grande quantidade de epitopos seja uma estratégia do parasito de induzir o sistema imune a uma série de respostas espúrias e ineficientes (Pitcovsky *et al.*, 2002). A diversidade de proteínas de superfície de *T. cruzi* pode também estar relacionada à habilidade do parasito de infectar e se replicar em uma variedade de tipos celulares, sendo esta uma estratégia essencial para sua sobrevivência. Especula-se que o polimorfismo das proteínas de superfície de *T. cruzi* seja um importante fator que pode contribuir para este fenômeno (Macedo *et al.*, 2004; Burleigh e Woolsey, 2002).

Um dos mecanismos de geração de variabilidade em famílias gênicas envolve eventos de duplicação seguido de divergência. Uma mutação que altera drasticamente a sequência de uma proteína é uma inserção ou deleção (*indel*) de nucleotídeos que não seja múltipla de três resíduos gerando uma mudança de fase de leitura do gene e modificando completamente a sequência de aminoácidos. Na maioria das vezes, *indels* geram proteínas truncadas pelo aparecimento de códons de parada prematuros na sequência nucleotídica.

Outra forma de diversificação das famílias é a troca de segmentos gênicos através da recombinação. A recombinação pode provocar a duplicação de parte do gene ou a troca de segmentos dentro e entre genes de diferentes famílias. Essa troca de segmentos em genes de

famílias de proteínas de superfície aumenta drasticamente sua diversidade, podendo gerar genes com novas funções e podendo também evitar o reconhecimento do parasito pelo sistema imune do hospedeiro. Os mecanismos que geram variação gênica em *Anaplasma marginale*, *Borrelia hermsii* e *Trypanosoma brucei* incluem a recombinação de segmentos de diferentes cópias gênicas, gerando variabilidade e permitindo ao parasita evadir do sistema imune e permanecer mais tempo no hospedeiro (Futse *et al.*, 2005; Palmer e Brayton, 2007).

A disponibilidade dos dados do genoma do *T. cruzi* permite uma melhor compreensão do real repertório das proteínas de superfície do parasito. Uma quantificação mais acurada da diversidade destas proteínas bem como a investigação dos mecanismos geradores desta variabilidade poderá contribuir para um melhor entendimento das estratégias de sobrevivência deste parasito.

3. Objetivos

3.1. Objetivo geral

O objetivo do presente trabalho é estudar a diversidade das grandes famílias gênicas de *T. cruzi* e os mecanismos geradores de variabilidade nas mesmas usando ferramentas de bioinformática para uma melhor compreensão da função e evolução das proteínas deste parasito.

3.2. Objetivos específicos

I. Analisar a diversidade da família trans-sialidase/trans-sialidase like (TcS) de *T. cruzi* através de projeção das sequências em espaço bidimensional usando *multidimensional scaling* (MDS) e análises de clusterização não hierárquica para a identificação de grupos dentro da família. Caracterizar cada grupo identificado da família TcS quanto: (i) presença de motivos característicos da família; (ii) localização genômica; (iii) expressão gênica; (iv) propriedades antigênicas.

II. Realizar uma análise comparativa das diversidades nucleotídica e protéica das grandes famílias gênicas de *T. cruzi* usando *multidimensional scaling* e clusterização não hierárquica.

III. Verificar indícios da ocorrência de mutações que provocam mudanças de fase de leitura dos genes das famílias de superfície como mecanismo gerador de diversidade.

IV. Avaliar a ocorrência de eventos de recombinação na geração de variabilidade em membros da família MASP. Avaliar a ocorrência de eventos de recombinação entre genes pertencentes a diferentes famílias gênicas de *T. cruzi*.

4. Materiais e métodos

Os materiais e métodos foram divididos em três partes (figura 1), que são:

1. Caracterização da família TcS - Identificação de novos grupos na família trans-sialidase/sialidase-like (TcS) de *T. cruzi*
2. Diversidade das grandes famílias gênicas de *T. cruzi*
3. Mecanismos evolutivos possivelmente envolvidos na geração de variabilidade

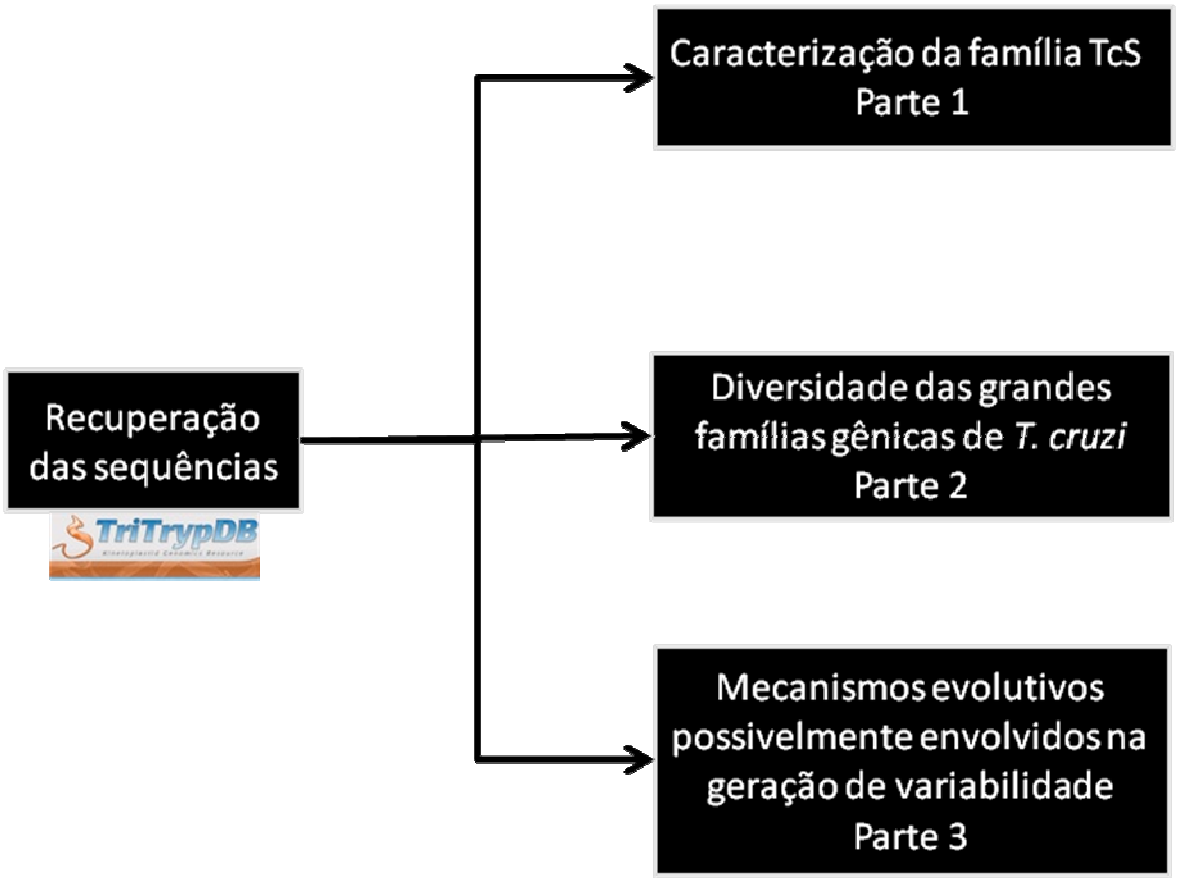


Figura 1 Fluxograma representando a divisão dos materiais e métodos em três partes.

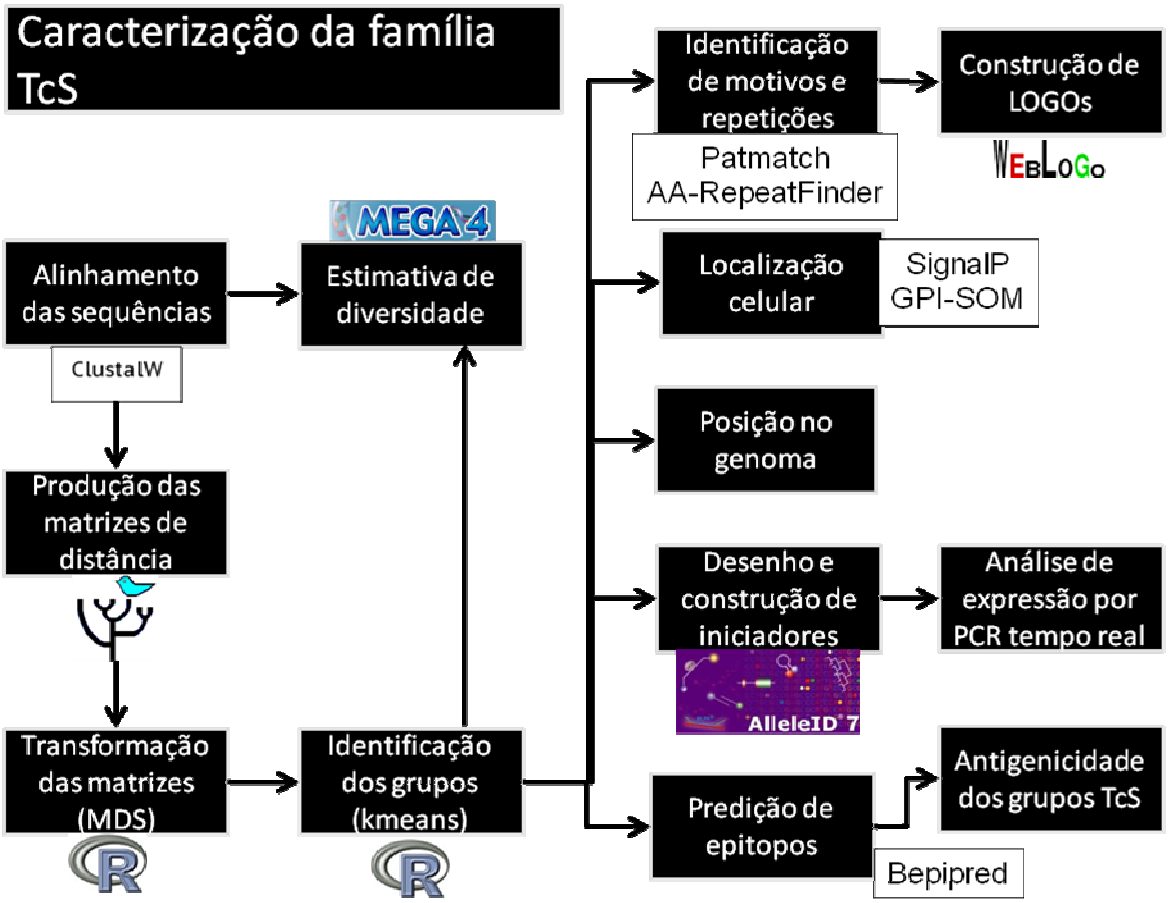


Figura 2. Detalhamento dos materiais e métodos "Caracterização da família TcS - Identificação de novos grupos na família trans-sialidase/sialidase-like (TcS) de *T. cruzi*".

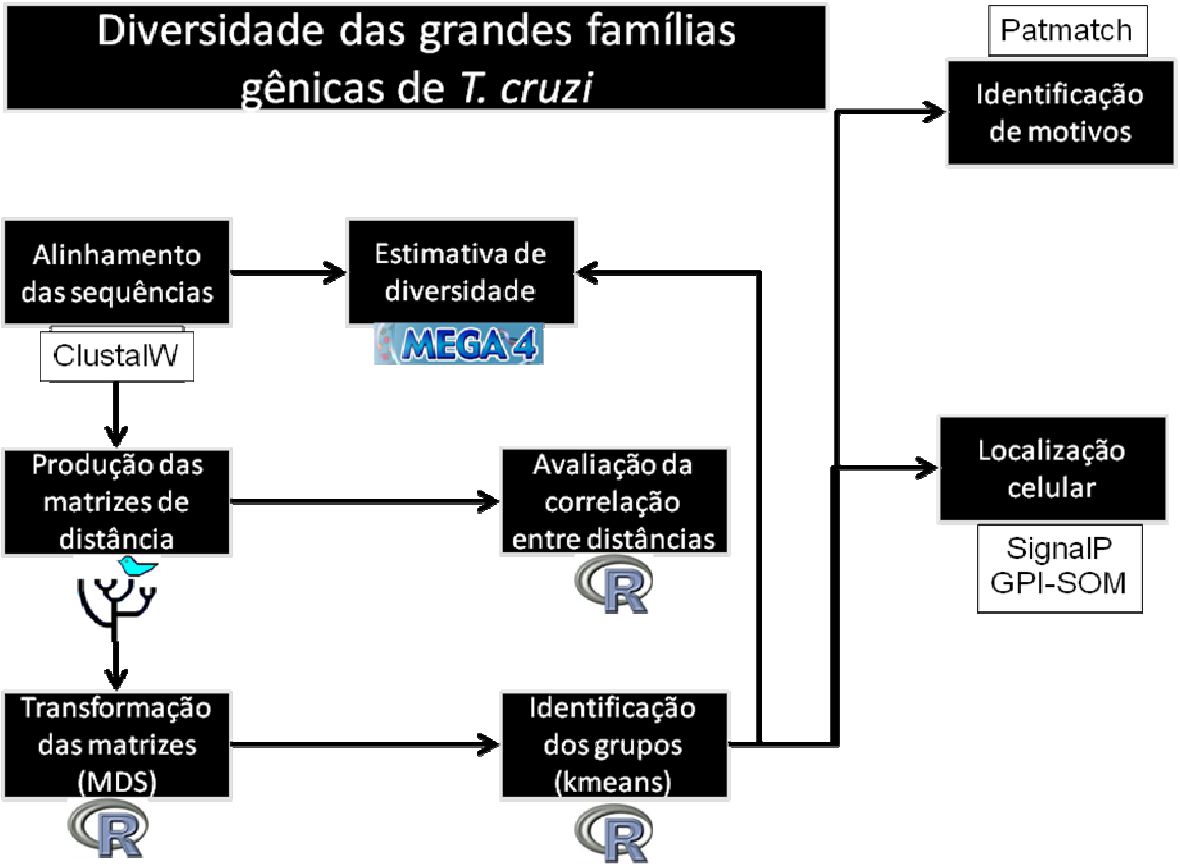


Figura 3. Detalhamento dos materiais e métodos "Diversidade das grandes famílias gênicas de *T. cruzi*".

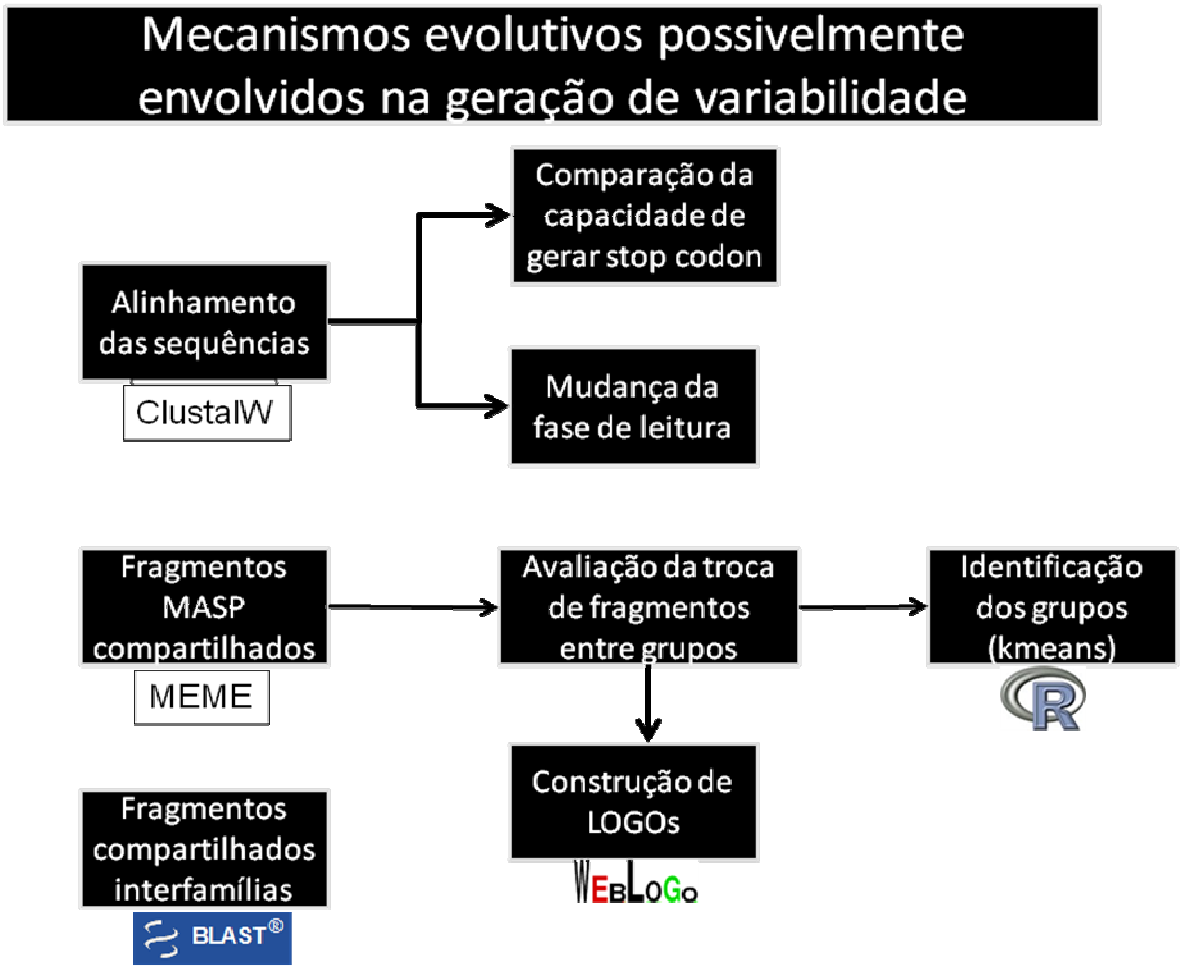


Figura 4. Detalhamento dos materiais e métodos "Mecanismos evolutivos possivelmente envolvidos na geração de variabilidade".

4.1. Sequências de DNA e proteínas das grandes famílias gênicas de *T. cruzi* - Conjunto de dados

Neste trabalho selecionou-se para estudo de diversidade oito famílias gênicas. Sete destas correspondem às maiores famílias gênicas do parasito: TcS, MASP (*mucin-associated surface protein*), mucina TcMUC, RHS (*Retrotransposon hot spot protein*), DGF-1 (*dispersed gene family*-1), GP63 (glicoproteína 63 kDa) e mucin-like. A outra família selecionada é SAP (*serine-alanine-proline rich protein*) que, apesar de não estar entre as oito maiores famílias gênicas, está clusterizada no genoma do parasito com as outras família gênicas selecionadas neste estudo (Bartholomeu *et al.*, 2009).

As sequências de DNA codificador das grandes famílias gênicas de *T. cruzi* foram extraídas do banco de dados TriTryp (http://tritrypdb.org/tritrypdb/). Foram selecionadas as sequências completas de DNA codificador das famílias selecionadas, sendo os genes parciais e pseudogenes excluídos, exceto pseudogenes de TcS que foram usados para mapeamento no genoma (ver item 4.9). As sequências protéicas de todas as famílias foram obtidas através da tradução *in silico* das sequências codificadoras. Também foi usada a sequência de 300 nucleotídeos após o códon de terminação das famílias (3' flanqueadora) com mais de 100 genes para avaliar a diversidade encontrada nessa região. A quantidade de sequências usadas em cada família encontra-se na tabela 4.

4.2. Alinhamento das sequências

Os alinhamentos múltiplo e par-a-par das sequências de DNA e proteínas foram realizados usando o programa ClustalW (Larkin *et al.*, 2007). Foram utilizados os parâmetros padrões do programa para realizar o alinhamento para todas as famílias.

4.3. Diversidade das sequências de DNA e proteínas

A diversidade das famílias e dos grupos dentro de cada família foi medida com base no alinhamento múltiplo das sequências usando três métodos. O método distância-p, tanto para medir a diversidade nas sequências de DNA (codificadora e 3' flanqueadora) quanto proteínas, Kimura-2-parâmetros somente para as sequências de DNA, enquanto a correção de Poisson foi usada somente para as sequências de proteínas. A estimativa de erro no cálculo dessas diversidades foi realizada usando o teste *bootstrap* com 1.000 réplicas.

4.4. Produção das matrizes de distância e correlação entre as distâncias nucleotídicas e protéicas

Foram produzidas quatro matrizes de distância para cada família. A primeira matriz foi produzida para avaliar a diversidade das famílias usando sequências de DNA (codificadora) através de alinhamento múltiplo usando o programa ClustalW.

As outras três matrizes foram produzidas para serem usadas na classificação das famílias em grupos, sendo usadas sequências de DNA, proteínas e 3' flanqueadoras. Estas matrizes de distância foram produzidas realizando o alinhamento par-a-par usando o programa ClustalW com seus parâmetros padrões.

Os alinhamentos foram usados para calcular as distâncias entre as sequências usando o pacote PHYLIP versão 3.69 (Felsenstein, 1989; Felsenstein, 2005), produzindo assim matrizes quadrada de dissimilaridade para todas as famílias gênicas analisadas, tanto para os alinhamentos múltiplos quanto para as matrizes produzidas com alinhamento par-a-par. As distâncias consideradas muito grandes para serem calculadas pelo programa foram determinadas com o valor -1.

As distâncias nucleotídica (região codificadora) e protéica das matrizes par-a-par foram comparadas para medir a correlação entre as duas distâncias. Foi usado o teste de correlação de Spearman (valor $p \leq 0.05$) para avaliar a correlação.

4.5. Transformação das matrizes de distância através de escala multidimensional (*Multidimensional scaling* - MDS) e classificação dos grupos

Com o objetivo de produzir um gráfico representando a distribuição espacial das sequências, avaliar quais famílias apresentam maior diversidade e fazer classificação dos grupos de cada família, foi realizada a análise de escala multidimensional *(Multidimensional scale* - MDS). As matrizes de distância foram transformadas por escala multidimensional para matrizes de duas dimensões. MDS é uma técnica estatística usada para a exploração visual dos dados com base no número de dimensões a serem exploradas. As novas matrizes contendo somente duas colunas apresentam as coordenadas para produção de projeções espaciais que representam o posicionamento das sequências em relação ao restante da família. A projeção espacial permite uma visualização global das sequências mais próximas (pontos mais próximos na projeção espacial) e das mais distantes (pontos mais afastados na projeção). Dessa forma esta análise permite avaliar a distribuição (diversidade) e a formação de grupos.

As projeções espaciais foram usadas para avaliação da diversidade das famílias usando as matrizes construídas com base nos alinhamentos múltiplos. A classificação das famílias em grupos, representando sequências mais similares, foi realizada usando as matrizes de alinhamento par-a-par. Os agrupamentos foram feitos usando o método *k-means* (Hartingan e Wong, 1979). O número de clusters iniciais foi determinado a partir do número de regiões com maior concentração de sequências. A transformação das matrizes e classificação das famílias em grupos foi realizada através do programa R-project (R Development Core Team, 2011).

4.6. Construção da árvore filogenética

A árvore da família SAP foi produzida usando alinhamento múltiplo das sequências de proteínas. O método usado para a construção das árvores foi NJ através do programa "Molecular Evolutionary Genetics Analysis 4" MEGA4 (Tamura *et al.*, 2007).

As análises de confiança dos nós (agrupamentos) nas árvores produzidas foram realizadas através de *bootstrap* com 1.000 réplicas e sendo considerados confiáveis aqueles com no mínimo de 50% das réplicas.

4.7. Procura por motivos, predição de localização celular e identificação sequências repetitivas

A presença dos motivos nas proteínas das famílias GP63, TcMUC e TcS foi pesquisada usando o programa PatMatch (Yan et al., 2005).

Os três motivos específicos de TcS de *T. cruzi* foram pesquisados. O motivo conhecido como FRIP foi pesquisado usando a sequência xRxP. Como o motivo FRIP é curto e degenerado, por isso foram adicionadas duas restrições após a pesquisa. Somente motivos FRIP próximos a extremidade N-terminal (menos de 200 aa) e/ou motivos encontrados antes de sequências Asp-box foram considerados válidos. O motivo Asp-box foi pesquisado usando a sequência SxDxGxTW para encontrar sequências que apresentavam exatamente o motivo característico da família TcS. A fim de buscar formas degeneradas do motivo Asp-box, permitiu-se um *mismatch* no consenso SxDxGxTW. Outro motivo de TcS pesquisado foi o motivo VTVxNVxLYNR que está presente em todas as proteínas TcS (Schenkman *et al.*, 1994). Também foi realizada a busca de proteínas TcS que apresentavam o motivo VTVxNVxLYNR degenerado, analisando a região do alinhamento múltiplo correspondente ao motivo VTVxNVxLYNR. Sequências repetitivas das proteínas TcS foram identificadas usando o programa AA-repeatFinder (http://gicab.decom.cefetmg.br:8080/bio-web) (Souza *et al.*, em preparação). Somente repetições com mais de 10 aminoácidos foram consideradas válidas.

O motivo relacionado com a atividade de metaloprotease presente na família GP63 corresponde a sequência HExxH. Este motivo foi pesquisado nas proteínas GP63 não permitindo *mismatches*. A busca por sequências que apresentam formas degeneradas desse motivo foi realizada através de inspeção da região contendo HExxH em alinhamentos múltiplos dos grupos encontrados na família GP63.

A família TcMUC apresenta um motivo rico em treonina. Esse motivo apresenta duas variantes que são usadas na classificação das proteínas em TcMUC I e TcMUC II juntamente com outras características (Acosta-Serrano *et al.*, 2001; Buscaglia *et al.*, 2004; Buscaglia *et al.*, 2006). As variantes do motivo pesquisado foram T_8KP_2 (presente em proteínas TcMUC I) e T_8K/QAP (presente em proteínas TcMUC II). A quantidade de treoninas presente em cada variante também foi pesquisada contendo o intervalo de 5 até 20.

A predição para localização na superfície celular foi realizada através da predição por peptídeo sinal e ancoramento na membrana por GPI nas famílias TcS e SAP. A busca por peptídeo sinal foi realizada usando o programa SignalP (Bendtsen *et al.*, 2004), e a predição para ancoramento na superfície foi realizada usando o programa GPI-SOM (Fankhauser e Mäser, 2005).

4.8. TcS ativas

A pesquisa pelas TcS que apresentam atividade trans-sialidase foi feita identificando as proteínas que apresentam aminoácidos importantes para essa atividade (tabela 1). Dentre esses aminoácidos os mais importantes são os aminoácidos Pro283 e Tyr342 que quando substituídos eliminam a atividade trans-sialidase (Cremona *et al.*, 1995; Montagna *et al.*, 2006). A identificação foi realizada usando alinhamentos das proteínas TcS separadas por grupos MDS (item 4.5 de material de métodos) e localização das regiões do alinhamento múltiplo correspondente a esses aminoácidos.

Tabela 1. Identificação dos aminoácidos importantes na enzima trans-sialidase de *T. rangeli*, *T. cruzi* e *T. brucei*. Aminoácidos que diferem em *T. rangeli* estão marcados em negrito. Adaptado de (Cremona et al., 1995; Montagna et al., 2006).

Função	*T. rangeli*	*T. cruzi*	*T. brucei*
Ligação ao grupo carboxílico do ácido siálico	Arg 36	Arg 35	Arg 106
	Arg 246	Arg 245	Arg 319
	Arg 315	Arg 314	Arg 402
Envolvido na formação da ligação de hidrogênio com o grupo 4-OH do ácido siálico	Arg 54	Arg 53	Arg 124
	Asp 97	Asp 96	Asp 168
Doador de prótons na reação	Asp 60	Asp 59	Asp 130
Ligação ao carboidrato aceptor	Tyr 120	Tyr 119	Tyr 191
Formação de ligação de hidrogênio fraca com a cadeia lateral de glicerol do ácido siálico	Trp 121	Trp 120	Trp 192
Formação do *pocket* para acomodação e Ligação ao grupo N-acetil do ácido siálico	**Met 96**	Val 95	Val 167
	Phe 114	Tyr 113	Tyr 185
Receptor de ligação	**Val 180**	Ala 179	Ala 253
Estabilidade para o estado de transição intermediária	Glu 231	Glu 230	Glu 304
Papel crucial para reação de TS	**Gln 284**	Pro 283	Pro 371
Receptor de ligação	**Asp 285**	Gly 284	Gly 372
	Cys 286	Ser 285	Ser 373
Ligação ao substrato, necessário para especificidade	Trp 313	Trp 312	Trp 400
Nucleófilo catalítico	Tyr 343	Tyr 342	Tyr 430

4.9. Mapeamento de TcS no genoma e análise de associação de grupos de TcS com regiões específicas

O posicionamento dos genes TcS nos cromossomos de *T. cruzi* foi realizado usando as informações da montagem das sequências de DNA realizada por Weatherly *et al*. (2009). Usando scripts em PERL (Practical Extraction and Report Language) e o módulo Bio::Graphics::module (www.bioperl.org), os cromossomos foram representados com todos os genes TcS mapeados e identificados de acordo com a classificação de grupos encontrados nos dados de proteínas (grupos identificados na projeção MDS e classificação usando o método *kmeans*). Esse mapeamento foi usado para identificar associação cromossomo-específica de algum grupo.

Outra análise de associação de grupos e TcS-pseudogenes com regiões cromossômicas específicas foi realizada usando a posição do primeiro códon de todas as sequências dividida pelo tamanho do cromossomo, criando valores de 0 até 1 que representam as posições relativas dos

genes TcS ao longo dos cromossomos. Esses valores das posições relativas foram usados para produção de histogramas indicando a frequência dos genes TcS, separados por grupo, ao longo dos cromossomos e permitindo observar se existe alguma associação com regiões específicas ou apresentam uma distribuição aleatória, sendo igualmente distribuídos ao longo dos cromossomos.

A presença dos genes e TcS em regiões subteloméricas também foi usada para observar evidências de associação entre regiões específicas no genoma e grupos de TcS. Foi realizada a contagem de genes TcS presentes nessas regiões e a frequências de todos os grupos de TcS em regiões subteloméricas, permitindo observar se existe uma preferência de grupos TcS localizados nas extremidades dos cromossomos.

4.10. Mecanismos evolutivos

A investigação dos mecanismos evolutivos responsáveis por alterações na sequência de proteínas se desdobrou em três partes, passando pela investigação da mudança de fase de leitura dos genes MASP, troca de motivos compartilhados entre genes da família MASP e troca de fragmentos compartilhados entre genes membros das diferentes famílias gênicas estudas neste trabalho.

4.10.1. Mudança da fase de leitura

Para a avaliação da geração de diversidade da família MASP foi elaborado um método para detecção de mutações levando a mudança de fase de leitura do gene.

Usando os alinhamentos par-a-par dos genes (sequências de DNA) da família MASP foi realizado a contagem do número de *indels* presente em cada sequência por alinhamento. Essa contagem foi dividida em dois grupos. *Indels* que são múltiplos de três e *indels* que não são múltiplos de três. A separação dos dois grupos tem como objetivo identificar *indels* que não alteraram a fase de leitura (múltiplos de três) não provocando grandes alterações na sequência primária da proteína, e identificar *indels* que alteraram a fase de leitura (não são múltiplas de três) alterando a tradução do mRNA.

O número de *indels* presente em cada par de sequências alinhadas foi comparado com a distância de proteína do par de sequências usadas. As distâncias usadas obtidas das matrizes de alinhamento par-a-par das famílias MASP foram produzidas pelo programa *protdist* do pacote PHYLIP versão 3.69 (Felsenstein, 1989; Felsenstein, 2005).

Pares de sequências que apresentam baixa distância nucleotídica e alta distância protéica também foram pesquisados nas matrizes de distância par-a-par para verificar sequências que sofreram mutações *indels* provocando grande alteração na sequência de proteínas. As distâncias pesquisadas foram: distância nucleotídica $\leq 0{,}4$ e distância protéica ≥ 0.9.

4.10.2. Número de possíveis códons de parada gerados por alteração da fase de leitura

A probabilidade de gerar códons de parada prematuros devido a alterações na leitura do mRNA devido a *indels* foi investigada. Alterações que levam a mudança na fase de leitura foram inseridas *in silico* em todos os genes das grandes famílias de superfície que apresentaram mais de 100 genes. O conjunto de todos os genes que já foram identificados e codificam proteínas no genoma de *T. cruzi* também foram usados para comparação com as grandes famílias de proteínas de superfície. As alterações foram: retirada do primeiro nucleotídeo do primeiro códon e retirada do segundo nucleotídeo também do primeiro códon. Após cada alteração no primeiro códon foi realizada uma contagem do número de códons de parada encontrado em todas as sequências.

O número médio de códons de parada de todos os grupos de sequências foi comparado com o número médio de códons de parada em sequências aleatórias geradas usando o preferência de códons (*codon usage*) de cada conjunto de sequências. Cada conjunto de sequências aleatórias possui 1000 sequências com 999 códons. As sequências aleatórias foram produzidas usando um *script* que seleciona cada códon com base na frequência encontrada em cada *codon usage*. As alterações, inserindo mudança de fase de leitura, também foram realizadas nas sequências aleatórias. O número de códons de parada do conjunto de sequências aleatórias foi verificado da mesma forma que nos genes reais de *T. cruzi*.

A posição dos códons de parada também foi verificada em cada conjunto de sequências reais e comparada com os conjuntos de sequências aleatórias correspondentes. As posições de cada códon de parada encontrado foram divididas pelo número de códons do gene para criar

regiões correspondentes entre os genes de diferentes tamanhos e as sequências aleatórias. Essa forma de correção da posição do códon de parada pelo tamanho do gene cria um mapa que varia de 0 até 1 e permite comparações de sequências de diferentes tamanhos e mapeamento da localização dos códons de parada.

4.10.3. Troca de fragmentos entre genes da família MASP

A determinação de fragmentos conservados pelos genes da família MASP foi feita usando o programa MEME versão 4.5.0 (Bailey e Elkan, 1994). Os parâmetros do programa MEME usados foram: número máximo de motivos para busca (30), busca realizada até encontrar os 30 motivos; comprimento mínimo (8 nt), comprimento máximo (50 nt), busca na direção direta e reversa do gene.

Os fragmentos que formam cada motivo MEME foram agrupados por similaridade usando o método NJ usando o programa MEGA4 (Tamura *et al.*, 2007), formando uma árvore para cada motivo MEME. Cada fragmento na árvore foi representado usando a mesma classificação do gene que possui o fragmento. A classificação usada de cada gene é a mesma produzida usando a matriz de coordenadas (MDS) aplicando o método *kmeans* para gerar os grupos (item 4.5 de material de métodos).

Alguns genes que pertencem a grupos diferentes e apresentavam o fragmento MEME muito similares foram alinhados usando o programa ClustalW com os parâmetros padrões. Esse alinhamento foi transformado em um gráfico para visualização da localização (posição) do motivo nas sequências. A possível recombinação de fragmentos de DNA entre os genes da família MASP foi verificada buscando fragmentos conservados que são compartilhados por membros de dois ou mais grupos do MDS (*clusters*).

4.10.4. Troca de fragmentos entre famílias de proteína

A procura por genes que compartilham fragmentos com genes de outras famílias foi realizada através da busca por sequências usando o programa blastn do pacote BLAST versão 2.2.23 (Altschul *et al.*, 1990), realizando comparações entre todas as famílias.

Para verificar a existência de fragmentos compartilhados entre as famílias, comparamos todos os membros das famílias TcS, MASP, TcMUC, DGF-1, GP63, RHS, SAP e Mucin-like. As comparações foram executadas usando as sequências de nucleotídeos e foram consideradas válidas as sequências que apresentaram *E-value* menor ou igual 10^{-10}. Os resultados da busca feita com pacote BLAST foram filtrados com a linguagem PERL (Practical Extraction and Report Language) e AWK para fragmentos que apresentassem identidade maior ou igual que 90% e foram selecionados cinco diferentes grupos de acordo com o tamanho do fragmento compartilhado.

Os grupos de sequências foram divididos em:

Grupo 1 → ≥ 100 e < 200 nucleotídeos
Grupo 2 → ≥ 200 e < 300 nucleotídeos
Grupo 3 → ≥ 300 e < 400 nucleotídeos
Grupo 4 → ≥ 400 e < 500 nucleotídeos
Grupo 5 → ≥ 500 nucleotídeos

Foi analisada também a posição desses fragmentos dentro dos genes. Fragmentos que se encontram total ou parcialmente dentro dos 200 primeiros nucleotídeos da região codificadora foram classificados como fragmentos de extremidade 5'. Fragmentos que se encontram total ou parcialmente dentro dos 200 últimos nucleotídeos da região codificadora foram classificados como fragmentos de extremidade 3'. Os fragmentos que não apresentam totalmente ou parcialmente entre os primeiros ou últimos 200 nucleotídeos da região codificadora foram classificados como fragmentos presentes no meio do gene.

4.11. Construção dos LOGOS

Representações gráficas da frequência de aminoácidos e nucleotídeos nos motivos encontrados na família TcS e grupos MEMEs, respectivamente, foram produzidas usando o programa WebLogo (Crooks *et al.*, 2004).

4.12. Cultura de parasitas

Epimastigotas do clone CL Brener de *T. cruzi* foram mantidos na fase de crescimento logarítmica a 28°C em meio Infusão de fígado e triptose (LIT) suplementado com 10% de soro bovino fetal. As formas amastigota e tripomastigota foram obtidas de células L6 (mioblasto de rato) infectadas crescidas em meio de Dulbecco suplementado com 5% de soro bovino fetal, a 37°C e 5% de CO_2 (Bartholomeu *et al.*, 2002). O RNA total foi extraído usando RNeasy kit (Qiagen).

4.13. PCR em tempo real

Foram desenhados iniciadores específicos para algumas sequências de TcS representativas de cada grupo usando a versão demonstrativa do programa AlleleID®. Os iniciadores foram desenhados para amplificação de fragmentos de 100 até 150pb, com temperatura de anelamento de 60°C (± 2°C) e comprimento do iniciadores variando de 18 a 24 bases. Esse conjunto de iniciadores foi usado para teste *in silico* de amplificação usando o programa e-PCR (Rotmistrosky, 2004), sendo permitido nessa amplificação 2 *mismatches* e 2 gaps. As sequências de contigs gerados no projeto genoma (El-Sayed *et al*., 2005a) e os cromossomos de *T. cruzi* (Weatherly *et al*., 2009) foram usados como molde na PCR eletrônica para realizar a amplificação *in silico*. Alguns pares de iniciadores que apresentaram regiões únicas de amplificação foram selecionados para análises de expressão por PCR em tempo real.

As reações de PCR em tempo real foram feitas no sistema de detecção ABI 7500 (Applied Biosystems). Foram realizadas triplicatas de cada reação contendo 1 mM de iniciadores direto e

reverso, SYBR® Green Supermix (Bio-Rad), e cDNA molde diluído. Curva padrão foi realizada para todas as reações para cada par de iniciadores usando diluição seriada de DNA genômico de *T. cruzi* CL Brener e foram usadas no cálculo do valor de quantidade relativa (Rq) de cada amostra. qRT-PCR para o gene GAPDH constitutivamente expresso foram realizados para normalizar a expressão dos genes TcS. Os resultados foram analisados com teste ANOVA e os gráficos construídos no GraphPad Prism 5.0 (GraphPad Inc.).

4.14. Predição de epitopos e imunoblot

Usando as 508 TcS proteínas intactas preditas do genoma do *T. cruzi*, os epitopos lineares para célula B foram preditos usando o programa Bepipred 1.0 (Larsen *et al.*, 2006). O programa avalia todos os aminoácidos da sequência dando uma pontuação baseado na hidrofilicidade (Parker, J. 1989). O programa não apresenta limitações para o número mínimo de aminoácidos para formar um epítopo, então foram selecionados peptídeos com 15 aminoácidos com pontuação igual ou maior que 1.3. Peptídeos com 70% de identidade sobre 70% do comprimento do peptídeo com outra proteína de *T. cruzi* que não fosse TcS foram excluídos. Foram escolhidos para síntese aqueles peptídeos que se apresentavam frequentes dentro de um grupo TcS e alto valor de predição como epitopo.

Os peptídeos foram sintetizados covalentemente em membrana de celulose pré-ativada de acordo com a técnica de síntese em SPOT (Frank, 1992). As membranas foram bloqueadas com BSA 5% a sacarose 4% em PBS e incubadas durante 1 hora e 30 minutos com soro de camundongos diluído (1:500) na solução de bloqueio. Depois da lavagem, a membrana foi incubada com anticorpo secundário IgG diluído 1:1200 na solução bloqueadora e, depois da segunda lavagem, revelada por ECL Plus Western blotting (GE healthcare). A membrana foi submetida às mesmas condições experimentais usando soro de camundongos não infectados. Medidas de densitometria e análise de todos os peptídeos foram realizadas usando Image Master Platinum (GE), e o limite da densidade relativa (Rd) para reação positiva foi determinado como 2,0. Gráficos foram construídos no GraphPad Prism 5.0 (GraphPad Inc.).

5. Resultados

5.1. Caracterização da família TcS - Identificação de novos grupos na família trans-sialidase/sialidase-like (TcS) de *T. cruzi*

5.1.1. Agrupamento das proteínas trans-sialidase/sialidase like de *T. cruzi*

O projeto genoma de *T. cruzi* revelou que trans-sialidase/sialidase-like (TcS) constitui a maior família gênica do parasito (El-Sayed *et al.* 2005a). Um dos representantes mais bem estudados da família são as trans-sialidases que são enzimas capazes de remover ácido siálico dos glicoconjugados e proteínas do hospedeiro e adicioná-lo a outras moléculas presentes na membrana do parasito (Cross e Takle, 1993; Schenkman *et al.*, 1994; Frasch, 2000). A família é classificada até o momento em quatro grupos, onde somente o grupo I apresenta proteínas com atividade trans-sialidase, enquanto que os outros grupos são formados por proteínas que apresentam outras funções e são denominadas trans-sialidase like. A despeito da existência destes quatro grupos, todas as sequências de TcS nos bancos de dados do genoma de *T. cruzi* estão anotadas como trans-sialidase. Como nenhuma análise sistemática sobre a diversidade desta família foi realizada após a publicação do genoma, nós nos propusemos a realizar análises de clusterização a fim de identificar todos os membros dos quatro grupos previamente definidos e eventualmente identificar novos grupos. Nossas análises se restringiram aos genes completos da família totalizando 508 membros. Portanto, pseudogenes e genes parciais não foram incluídos em nosso trabalho.

Nas análises de clusterização utilizamos a ferramenta de escala multidimensional (*Multidimensional scale* - MDS) que permite fazer uma projeção espacial da matriz de distância, onde as sequências são apresentadas por pontos e a distância entre esses pontos reflete a divergência entre as sequências. A projeção fornecida por MDS permite identificar sequências mais relacionadas, permitindo assim fazer a classificação da família em grupos.

Na projeção MDS das proteínas da família TcS ficou evidente a formação de diferentes regiões, cada uma representando sequências mais similares (Figura 5A). Essa projeção foi submetida ao método de agrupamento *kmeans* e foi possível a identificação de 10 grupos bem

distintos (Figura 5A). O número de membros em cada grupo é apresentando na (tabela 2). A separação da projeção de proteínas em mais de 10 grupos resultou na fragmentação destes grupos, sem mudança de membros entre os grupos, indicando que a divisão com 10 grupos é robusta.

Os grupos mais distantes, representados pela cor preta e marrom são formados por 1 e 2 proteínas, respectivamente. A verificação da sequência codificadora dessas proteínas permitiu a identificação de códons de parada prematuros e pontos de iniciação incorretos, levando a uma distorção da real posição dessas proteínas no MDS. Devido a esses problemas, as proteínas pertencentes a esses grupos formam excluídas das análises posteriores.

Representantes dos quatro grupos de trans-sialidase e trans-sialidase like descritos anteriormente (Cross e Takle, 1993; Schenkman *et al.*, 1994) foram mapeados na projeção MDS de proteínas. Como esperado, essas proteínas pertencem a quatro grupos diferentes (Figura 5). Todas representantes das trans-sialidases ativas TCNA, SAPA e TS-e, que pertencem ao grupo I, foram encontradas no grupo azul. As proteínas gp82, gp90, Tc85-11_SA85-1.1 e ASP-2, representantes do grupo II, pertencem ao grupo verde escuro. Enquanto FL-160 e Ts13, representantes dos grupos III e VI, pertencem aos grupos azul claro e rosa, respectivamente. Outros quatro grupos (vermelho, cinza, laranja e roxo) não apresentaram nenhum representante descrito anteriormente. A projeção do MDS mostra que esses novos grupos são completamente separados dos grupos previamente identificados.

Como a separação dos grupos na projeção é bem definida, os oito grupos identificados receberam as seguintes designações : TcSgrupoI (azul), TcSgrupoII (verde escuro), TcSgrupoIII (azul claro), TcSgrupoIV (rosa), TcSgrupoV (vermelho), TcSgrupoVI (cinza), TcSgrupoVII (laranja) e TcSgrupoVIII (roxo).

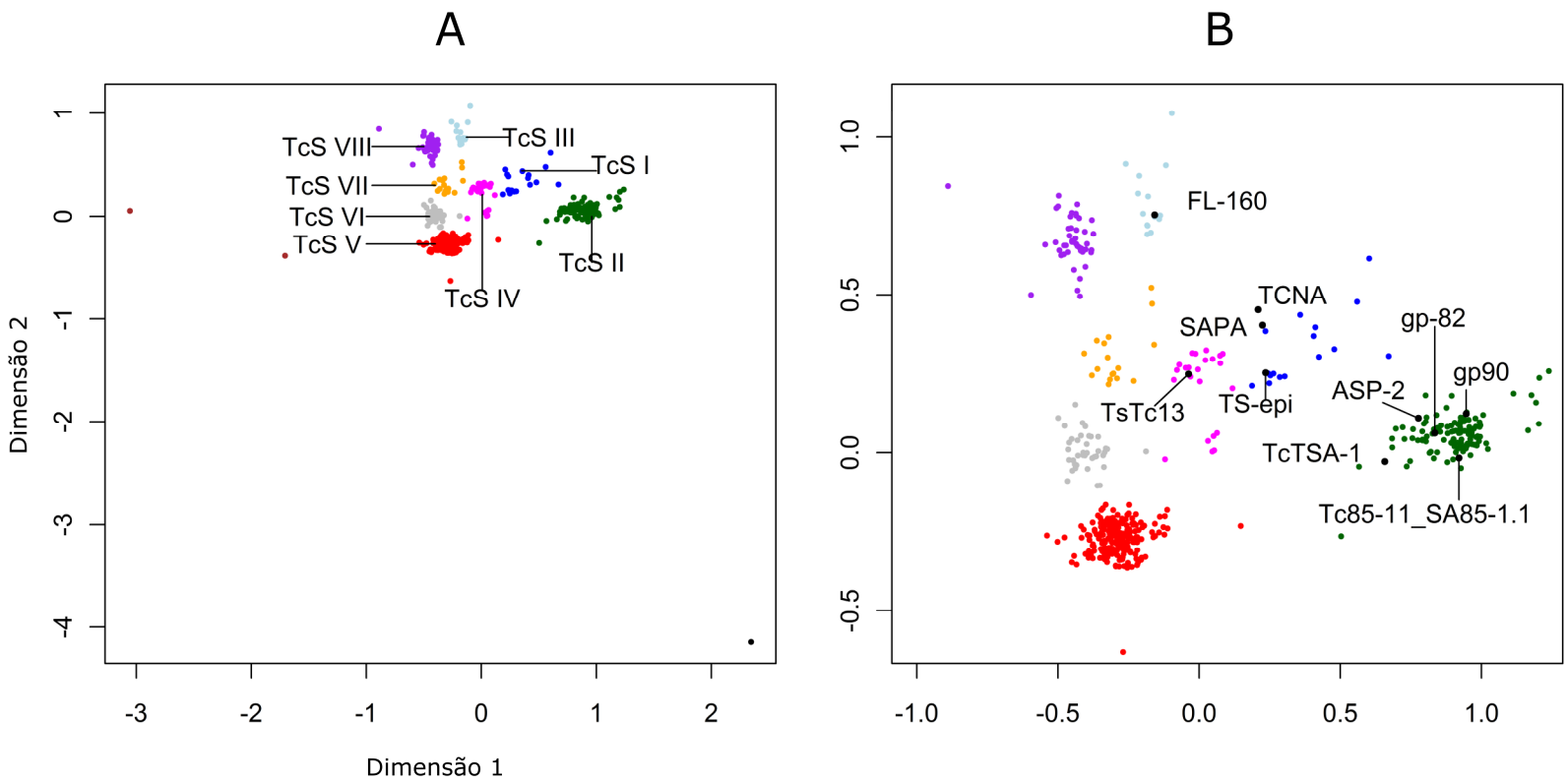


Figura 5. Projeção MDS das sequências de proteínas da família TcS de *T. cruzi*. (A) projeção e separação da família em 10 grupos identificados pelas diferentes cores usando o método *kmeans*. (B) projeção da família dos oito grupos considerados válidos e identificação dos representantes dos quatro grupos de TcS descritos por trabalhos anteriores (Cross e Takle, 1993; Schenkman *et al.*, 1994). As diferentes cores representam os grupos TcSgrupoI (azul), TcSgrupoII (verde escuro), TcSgrupoIII (azul claro), TcSgrupoIV (rosa), TcSgrupoV (vermelho), TcSgrupoVI (cinza), TcSgrupoVII (laranja) e TcSgrupoVIII (roxo). Cada ponto representa uma proteína e a distância entre os genes reflete sua dissimilaridade.

5.1.2. Diversidade da família TcS

A família TcS apresentou grande diversidade nucleotídica e protéica (tabela 2). A diversidade dentro dos grupos é menor que a diversidade total e sendo TcSgrupoI o grupo mais diverso.

A diversidade dos grupos não se mostrou relacionada ao número de membros, podendo ser observado grupos como TcSgrupoI com poucos membros mas grande diversidade, enquanto outros grupos como TcgrupoV apresenta grande número de membros (227) e níveis de diversidade similares àqueles observados para grupos menores, como por exemplo TcSgrupoVI que possui apenas 39 representantes.

A avaliação de todos os genes TcS mostra que existe uma diversidade e variação de sequências maior que aquela previamente descrita para a família, permitindo a classificação da família em oito grupos que apresentam número de membros e diversidade variáveis entre si.

Tabela 2. Diversidade nas sequências de DNA, proteínas e 3' flanqueadora da família TcS (ver item 5.1.1). Os métodos usados para os cálculos das diversidades foram: distância-p para medir a diversidade nos três tipos de sequências, Kimura-2-parâmetros foi usada somente para as sequências de DNA e para a sequência 3'flanqueadora. Para a sequência de proteínas também foi usado o método de correção de Poisson.

			DNA		**Proteína**	
Grupo	**Cor**	**Número de membros**	**Distância-p (erro)**	**K2p (erro)**	**Distância-p (erro)**	**Correção de Poisson (erro)**
TcSgrupoI	Azul	19	0,371 (0,004)	0,690 (0,029)	0,494 (0,009)	0,881 (0,027)
TcSgrupoII	Verde escuro	117	0,264 (0,004)	0,340 (0,007)	0,419 (0,010)	0,558 (0,018)
TcSgrupoIII	Azul claro	15	0,209 (0,005)	0,263 (0,008)	0,366 (0,010)	0,492 (0,023)
TcSgrupoIV	Rosa	25	0,179 (0,003)	0,226 (0,005)	0,250 (0,008)	0,320 (0,012)
TcSgrupoV	Vermelho	227	0,252 (0,004)	0,316 (0,006)	0,396 (0,009)	0,513 (0,015)
TcSgrupoVI	Cinza	39	0,246 (0,004)	0,312 (0,007)	0,394 (0,009)	0,513 (0,016)
TcSgrupoVII	Laranja	17	0,298 (0,004)	0,425 (0,009)	0,448 (0,009)	0,651 (0,020)
TcSgrupoVIII	Roxo	46	0,215 (0,004)	0,270 (0,006)	0,353 (0,009)	0,453 (0,013)
TcS total		508	0.413 (0.004)	0.662 (0.011)	0.574 (0.090)	0.912 (0.023)
3'flanqueadora (300 nt)		495	0.573 (0.007)	1.086 (0.029)	-	-

5.1.3. Busca por motivos característicos da família TcS

As proteínas TcS são caracterizadas por apresentarem o motivo VTVxNVxLYNR que está presente em todos os membros já descritos e está relacionado com adesão celular. Outros motivos típicos da família são os motivos Asp-box (SxDxGxTW - presente em múltiplas cópias na proteína), e o motivo FRIP (xRxP) que está envolvido na ligação ao grupo carboxílico do ácido siálico.

O motivo VTVxNVxLYNR, onde x representa qualquer aminoácido foi encontradas 328 das 505 sequências analisadas, e apresenta-se distribuído por todos os grupos identificados neste trabalho (Figura 6). O motivo VTVxNVxLYNR está presente na maioria dos membros de cada grupo, com exceção dos grupos TcSgrupoVII e TcSgrupoVIII, representados pela cor laranja e roxo respectivamente, que apresentam poucos membros possuindo esse motivo. Como é postulado que este motivo é característico da família (Cross e Takle, 1993; Schenkman *et al.*, 1994), nós investigamos se as 180 proteínas que não apresentaram o consenso VTVxNVxLYNR possuem uma versão degenerada deste motivo. Para isso foi realizado o alinhamento múltiplo de todas as sequências usadas neste trabalho e a região contendo o motivo VTVxNVxLYNR foi recuperada e submetida a inspeção visual, possibilitando a identificação de sequências que apresentam uma forma degenerada do motivo. Essa análise permitiu a identificação de mais 159 sequências contendo o motivo degenerado, totalizando 96% da família apresentando o motivo (Figura 6). As sequências que não apresentaram o motivo possuem códon de parada prematuro ou mudanças na fase de leitura que provocou o encurtamento da região C-terminal. Apesar de muitas sequências apresentarem uma forma degenerado do motivo, ele é conservado dentro de cada grupo (Figura 8).

A busca pelo motivo Asp-box foi realizada usando o sequência SxDxGxTW, onde x representa qualquer aminoácido. Somente 135 proteínas apresentaram o motivo Asp-box, sendo que a maioria dessas sequências pertencem a três grupos anteriormente descritos: TcSgrupoI (16 sequências, grupo azul), TcSgrupoII (95 sequências, grupo verde escuro) e TcSgrupoIV (22 sequências, grupo rosa) (Figura 6). O motivo Asp-box ainda foi encontrado nas sequências Tc00.1047053506911.30 e Tc00.1047053508365.190 que pertencem aos grupos TcSgrupoV

(vermelho) e TcSgrupoVI (cinza), respectivamente. Também foi realizada a busca pelo motivo Asp-box (SxDxGxTW) onde foi permitido a presença de 1 *mismatch*. Esta versão degenerada do motivo foi encontrada em outras proteínas TcS inclusive aquelas pertencentes a outros grupos além dos encontrados na primeira busca. Foram encontradas 383 sequências com o motivo Asp-box (Figura 6). A maioria das sequências apresentou somente 1 cópia do motivo (220 sequências), enquanto que 154 sequências apresentaram duas cópias e apenas nove sequências apresentaram três cópias. Além dos grupos identificados na primeira busca, o motivo Asp-box foi encontrado em sequências dos grupos TcSgrupoVII (1 sequência, grupo laranja) e TcSgrupoVIII (três sequências, grupo roxo). Novamente não foi encontrado o motivo Asp-box nas sequências do grupo TcSgrupoIII, sendo este o único grupo de TcS que não apresentou esse motivo. O motivo Asp-box é encontrado na maioria dos grupos TcSgrupoI (azul, 17/19 sequências), TcSgrupoII (verde escuro, 114/117), TcSgrupoIV (rosa, 24/25) e também nos novos grupos TcSgrupoV (vermelho, 188/227), TcSgrupoVI (cinza, 36/39) (Tabela 3). Os grupos TcSgrupoVII (laranja, 1/17) e TcSgrupoVIII (roxo, 3/46) apresentaram poucos sequências contendo o motivo Asp-box.

Em seguida realizamos a busca pelo motivo FRIP (xRxP, onde x representa qualquer aminoácido). Como se sabe, este motivo se localiza no N-terminal das TcS (Frasch, 2000; Todeschini *et al.*, 2000) e uma vez que é um motivo pequeno e degenerado, nós computamos apenas aquelas ocorrências que estivessem antes do primeiro motivo Asp-box (mais próximo da extremidade N-terminal) ou, na ausência do motivo Asp-box, antes do resíduo 200. O motivo FRIP foi encontrado em todos os grupos TcS, somando um total de 205 proteínas, sendo esse motivo encontrado na maioria das sequências dos grupos TcSgrupoI (azul, 68%), TcSgrupoIII (azul claro, 87%), TcSgrupoIV (rosa, 88%), TcSgrupoVII (laranja, 76%) e TcSgrupoVIII (roxo, 87%) (Figura 6 e Tabela 3)

O motivo FRIP se mostrou degenerado quando analisado todas as 505 TcS, mas dentro dos TcS grupos I, III e IV, além de se mostrar frequente, o motivo se mostrou muito conservado (Figura 7).

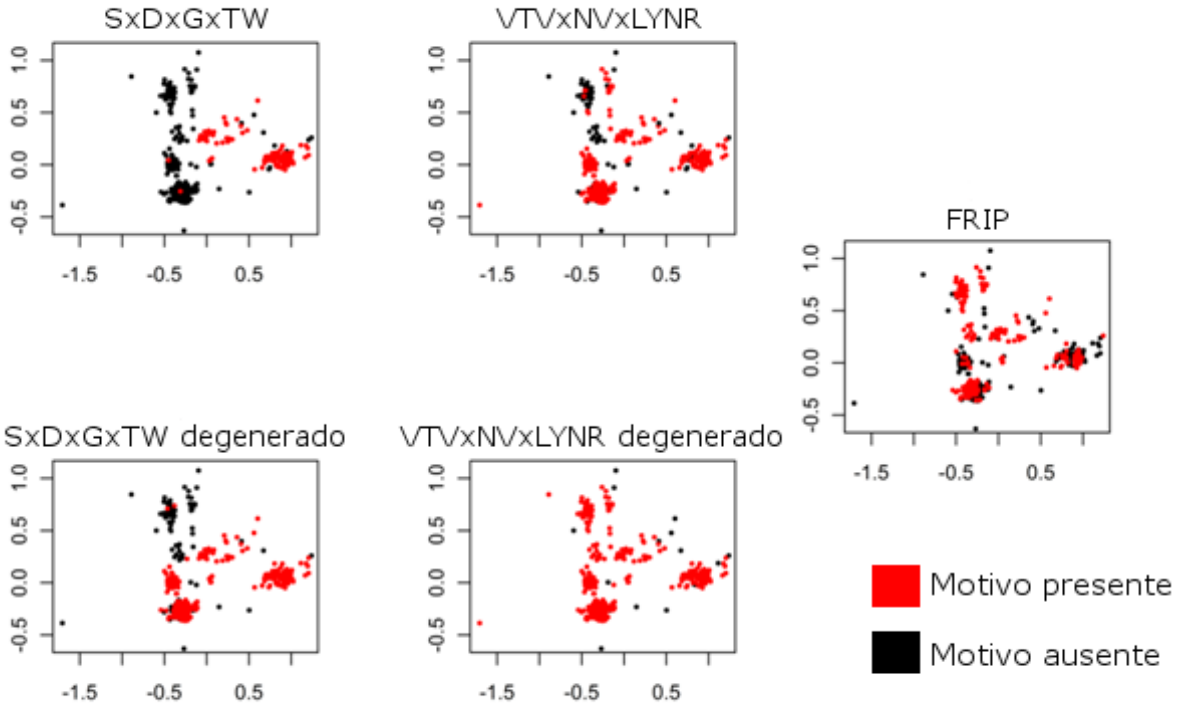


Figura 6. Projeção MDS de proteínas TcS indicando a presença de motivos. Presença do motivo SxDxGxTW; presença do motivo VTVxNVxLYNR; presença do motivo SxDxGxTW permitindo 1 degeneração; presença do motivo VTVxNVxLYNR recuperado no alinhamento de 505 TcS; Presença do motivo FRIP (xRxP). "x" representa qualquer aminoácido. As proteínas em cada painel que apresentam os motivos são representadas em vermelho.

Tabela 3. Presença dos motivos no oito grupos de TcS e frequência dos motivos FRIP e Asp-box. Está representado na tabela quantas sequências costituem cada grupo (Número de sequências); se as proteínas que costituem o grupos apresentam peptídeo sinal (Peptídeo sinal); quantas sequências apresentam o motivo FRIP e sua porcentagem (FRIP (%)); quantas proteínas apresentam o motivo Asp-box, quantos motivos estão repetidos na maioria das sequências e a varaiação encontrada detro do grupo (Asp-box (maioria) (variação)); quantas sequências apresentam o motivo VTV (VTV); se mais de 50% das proteínas do grupo apresentam reptições (Repetições). A presença de motivos e repetições é identificada por "X" e aus~encia por "-".

Grupo	**Cor do grupo nas projeções**	**Número de sequências**	**Peptídeo sinal**	**FRIP (%)**	**ASP-box (maioria) (variação)**	**VTV**	**Repetição**	**GPI**
TcSgrupoI	Azul	19	X	13 (68%)	17 (2) (0-3)	X	X	X
TcSgrupoII	Verde escuro	117	X	50 (42,7%)	114 (2) (0-3)	X	-	X
TcSgrupoIII	Azul claro	15	X	13 (87%)	0 (0) (0)	X	-	X
TcSgrupoIV	Rosa	25	X	22 (88%)	24 (2) (0-3)	X	X	X
TcSgrupoV	Vermelho	227	X	90 (39%)	188 (2) (0-3)	X	-	X
TcSgrupoVI	Cinza	39	X	8 (20%)	36 (2) (0-2)	X	-	X
TcSgrupoVII	Laranja	17	X	13 (76%)	1 (1) (0-1)	X	-	X
TcSgrupoVIII	Roxo	46	X	40 (86%)	3 (1) (0-2)	X	-	X

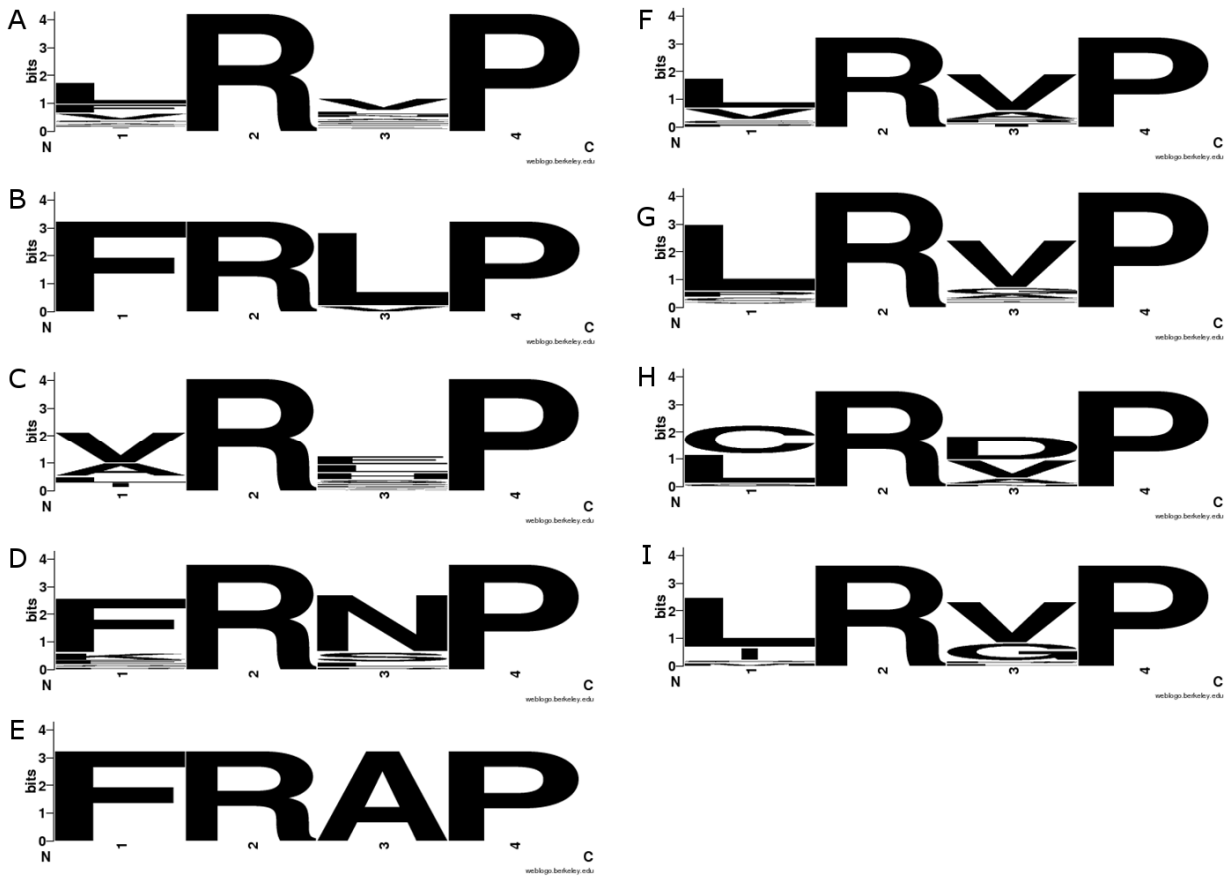


Figura 7. Conservação do motivo FRIP nas proteínas TcS. (A) conservação do motivo FRIP nas 505 proteínas analisadas, (B) TcSgrupoI, (C) TcSgrupoII, (D) TcSgrupoIII, (E) TcSgrupoIV, (F) TcSgrupoV, (G) TcSgrupoVI, (H) TcSgrupoVII, (I) TcSgrupoVIII.

Outra característica da família TcS é a ocorrência de repetições. Duas repetições já foram descritas na família, sendo conhecidas como SAPA (DSSAH(S/G)TPSTP(A/V)) e EPKSA, que são repetições longas em tandem e estão presentes em proteínas do grupo I e IV, respectivamente. A fim de verificar a existência de outras repetições na família, foi realizada a busca por repetições que apresentavam mais de 10 aminoácidos usando um programa desenvolvido pelo nosso grupo (Souza *et al.*, em preparação). As repetições encontradas foram mais frequentes nos grupos TcSgrupoI e TcSgrupoIV. De fato as maiores repetições encontradas, apresentando repetições de até 884 aminoácidos, correspondem àquelas já conhecidas DSSAH(S/G)TPSTP(A/V) e EPKSA. Entretanto, além das repetições conhecidas, novas repetições foram encontradas nesses dois grupos.

Outros grupos também apresentaram repetições, mas em menor quantidade como TcSgrupoII, TcSgrupoV, TcSgrupoVI, TcSgrupoVII e TcSgrupoVIII ocorrendo 14,5; 1; 21; 6 e 2,5%, respectivamente, das sequências desses grupos. As repetições desses grupos são curtas com aproximadamente 15 aminoácidos, exceto pelo TcgrupoVII (laranja) que apresentou um trecho de repetição de 150 aminoácidos. O grupo TcSgrupoIII não apresentou repetições.

A Figura 8 apresenta os motivos que são encontrados na maioria dos membros para todos oito grupos de TcS e a sequência representando a conservação dos motivos Asp-box e VTVxNVxLYNR.

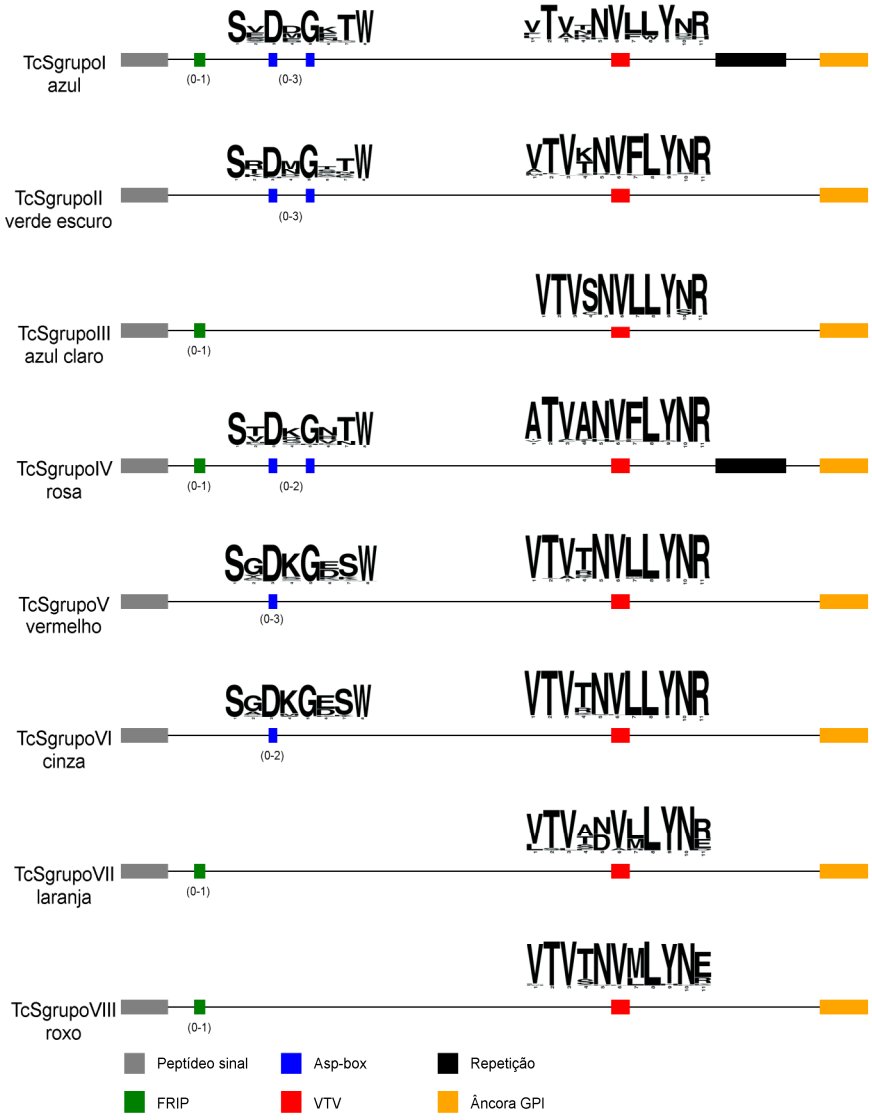


Figura 8. Representação dos motivos presentes na maioria das proteínas em cada um dos grupos. As sequências consenso dos motivos Asp-box e VTVxNVxLYNR está representando. O número entre parênteses indica a variação de quantas cópias dos motivos FRIP e ASP-box são encontradas por sequência.

O TcSgrupoI (azul) e TcSgrupoIV (rosa) têm a estrutura mais complexa, com os motivos FRIP, Asp box e VTVxNVxLYNR e repetições C-terminal, embora as sequências do motivo VTVxNVxLYNR e repetições C-terminal sejam distintas. Os TcSgrupoII (verde), TcSgrupoV

(vermelho) e TcSgrupoVI (cinza) possuem os motivos Asp Box e VTVxNVxLYNR. TcSgrupoIII (azul claro), TcSgrupoVII (laranja) e TcSgrupoVIII (roxo) possuem somente os motivos FRIP e VTVxNVxLYNR, que apresentam sequência consenso grupo específica. Este padrão de ocorrência de motivos está de acordo com a distribuição especial dos grupos TcS na MDS (Figura 5B). Os grupos TcS I e IV que apresentam todos os motivos estão na região central da projeção MDS, enquanto grupos TcS II, V e VI estão agrupados na parte inferior e grupos TcS III, VII e VIII estão agrupados na parte superior esquerda.

5.1.4. Identificação das TcS ativas

Somente 11 membros da família TcS, todos do TcSgrupoI, foram identificados apresentando todos os aminoácidos importantes para atividade trans-sialidase. A Figura 9 apresenta parte do alinhamento das 11 TcS potencialmente ativas e em destaque estão resíduos importantes para a atividade de trans-sialidase, o motivo FRIP e três motivos Asp-box. Vale ressaltar que apesar da conservação destes resíduos, existem variações nas extremidades dessas proteínas, principalmente na extremidade C-terminal, onde são localizadas as repetições encontradas no TcSgrupoI (dados não mostrados).

Nenhum dos quatro grupos previamente descritos nem os quatro novos grupos identificados neste trabalho apresentaram proteínas com todos os aminoácidos importantes para a atividade de trans-sialidase.

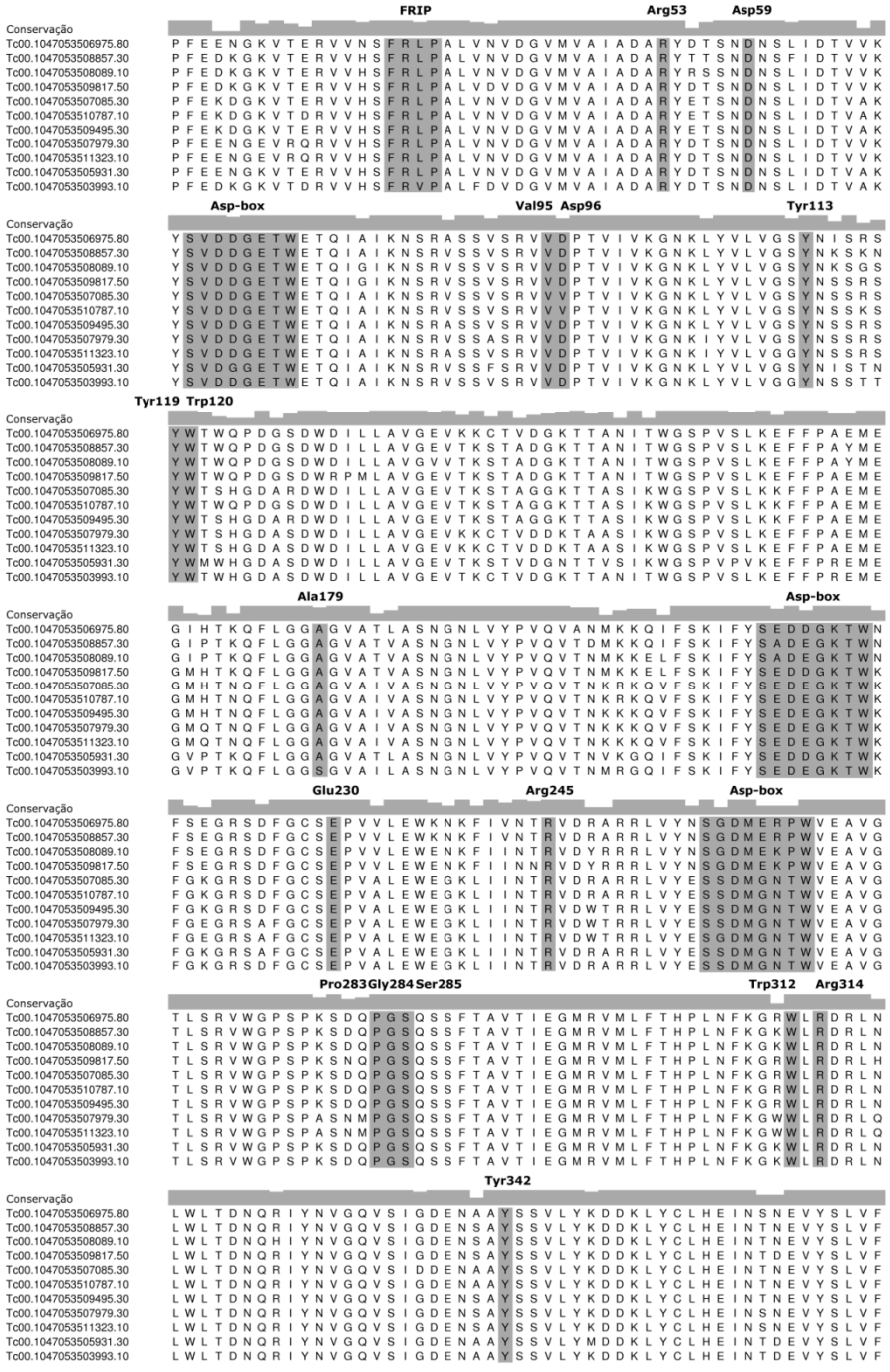


Figura 9. Alinhamento das sequências TcS potencialmente ativas. Os aminoácidos envolvidos na atividade trans-sialidase e os motivos FRIP e Asp-box estão destacados em cinza. Apenas parte do alinhamento correspondente a um segmento contínuo que contém os aminoácidos chave para a atividade de trans-sialidase é mostrado.

5.1.5. Mapeamento dos grupos TcS nos cromossomos

A fim de verificar se os genes TcS pertencentes a um mesmo grupo estão agrupados no genoma do parasito, foi feito o mapeamento da localização cromossômica dos genes da família. Como pode ser visto na Figura 10, os genes TcS estão bem espalhados no genoma de *T. cruzi* (cepa CL Brener), existindo genes TcS em praticamente todos os cromossomos, com exceção do cromossomo 1 que não apresentou representantes da família. Os genes TcS apresentaram mais cópias em cromossomos médios e grandes.

Análise visual do mapa dos cromossomos não mostrou nenhuma evidência de associação entre um cromossomo específico e um dado grupo de TcS (Figura 10). Entretanto é possível observar uma tendência de genes TcSgrupoII serem localizados nas extremidades dos cromossomos e TcSgrupoV serem localizados na região interna do cromossomo.

A fim de se quantificar a distribuição dos genes ao longo dos cromossomos, nós computamos a distância dos genes em relação às extremidades dos cromossomos. Para tanto, nós calculamos a posição relativa dos genes nos diferentes cromossomos, dividindo a coordenada inicial do gene no cromossomo (posição da primeira base do primeiro códon) pelo comprimento total do cromossomo. De fato, os grupos TcSgrupoII e TcSgrupoV foram os únicos grupos que apresentarem um viés na distribuição ao longo dos cromossomos. O grupo TcSgrupoII (verde escuro) encontra-se preferencialmente nas extremidades dos cromossomos, enquanto os genes TcSgrupoV (vermelho) encontram-se preferencialmente na região central (Figura 11A).

Também foi investigado se existe associação de grupos TcS com regiões subteloméricas. Nós definimos a região subtelomérica, como aquela que se estende desde as repetições teloméricas até a primeira sequência não repetitiva do genoma. Dentro dessas regiões foram encontradas 60 genes TcS completos, sendo 1 desses genes membro do grupo marrom que foi excluído das nossas análises como mencionado anteriormente. A maioria das sequências encontradas nas regiões subteloméricas pertencem ao grupo TcSgrupoII (verde escuro, 36 genes, 61%), enquanto os outros grupos apresentam uma quantidade menor de sequências associadas as regiões subteloméricas, sendo 7 sequências do TcSgrupoIV (rosa, 11%), 10 TcSgrupoVIII (roxo, 16,6%), 3 TcSgrupoVII (laranja, 0,05%) e 2 TcSgrupoI (azul, 0,0166%) (Figura 11B). Não foram encontradas sequências dos grupos TcSgrupoIII e TcSgrupoVI associadas às regiões

subteloméricas, enquanto que apenas 1 gene do grupo TcSgrupoV está associado a esta região (Figura 11B).

Foi também verificada a posição relativa dos pseudogenes em relação ao tamanho dos cromossomos. Como observado para aos genes do grupo TcSgrupoII (verde escuro), foi encontrado um enriquecimento de pseudogenes associados às extremidades dos cromossomos (Figura 11C).

Portanto não foi observada nenhuma associação de grupo específico com qualquer cromossomo. Porém, há claramente uma localização preferencial de genes TcSgrupoII e TcSgrupoV nas regiões subteloméricas e na parte interna dos cromossomos, respectivamente.

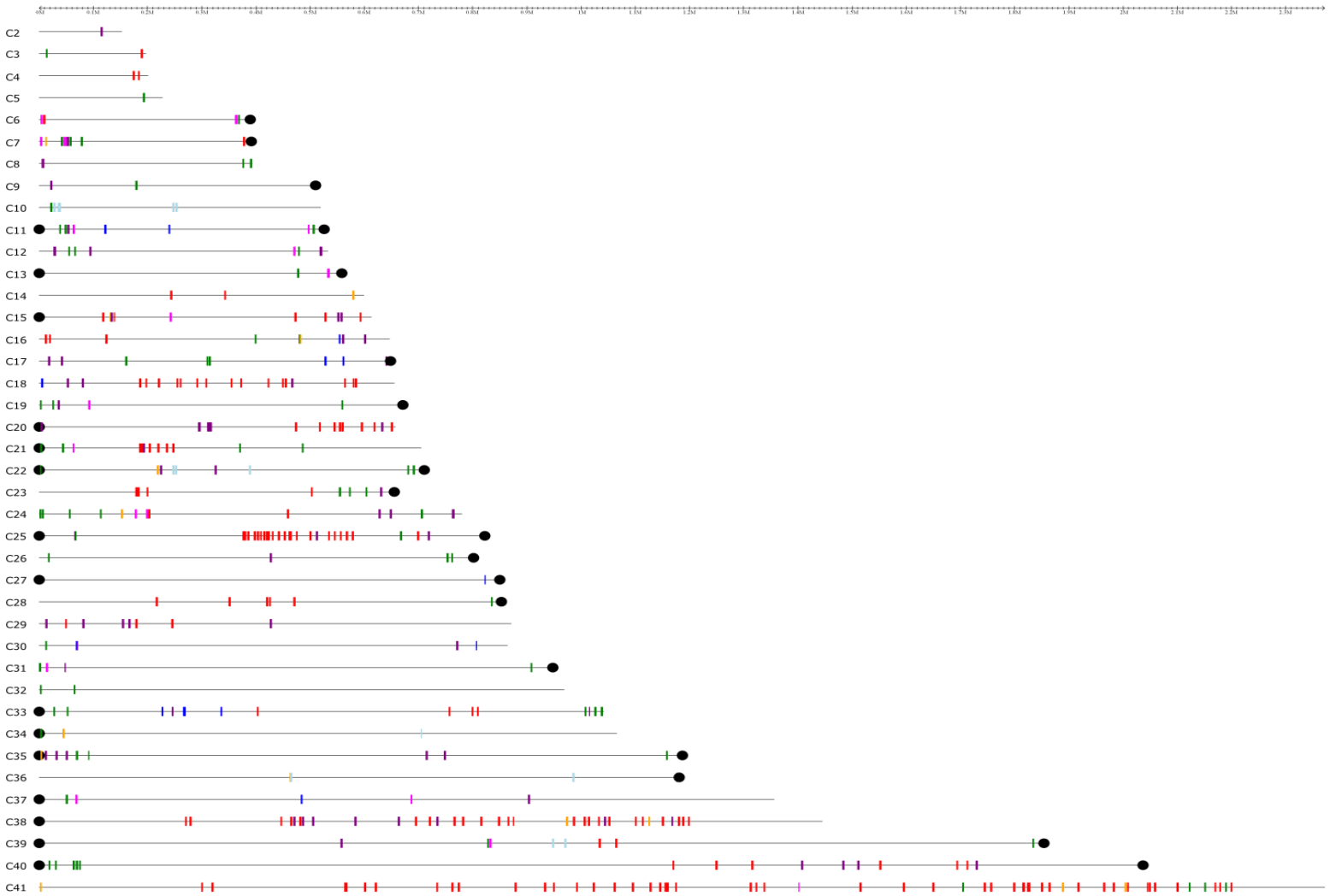


Figura 10. Mapeamento dos genes TcS nos cromossomo de *T. cruzi*. As diferentes cores representam os oito grupos de TcS. TcSgrupoI (azul), TcSgrupoII (verde escuro), TcSgrupoIII (azul claro), TcSgrupoIV (rosa), TcSgrupoV (vermelho), TcSgrupoVI (cinza), TcSgrupoVII (laranja) e TcSgrupoVIII (roxo).

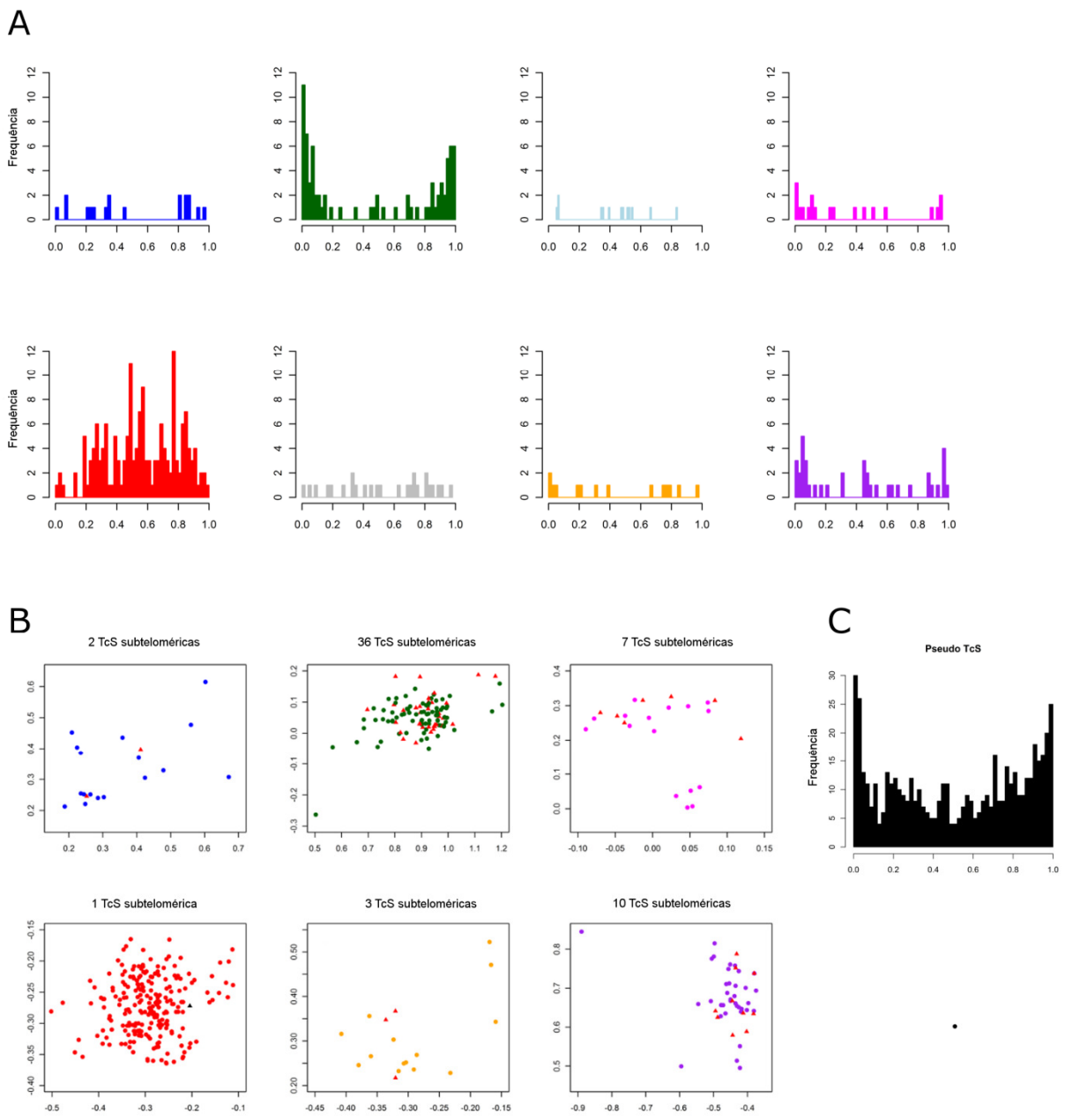


Figura 11. Distribuição dos diferentes grupos de TcS ao longo dos cromossomos. Painéis A e C apresentam a posição relativa do seis diferentes grupos de TcS e TcS-peseudogenes, respectivamente, ao longo dos cromossomos. Painel B mostra a quantidade de genes encontrados nas regiões subteloméricas e o mapeamento dos mesmos nas projeções MDS (representados em vermelho em todos os grupos, exceto o grupo TcS V sendo representado em preto).

5.1.6. Perfil de expressão dos TcS genes de diferentes subgrupos

Trabalhos anteriores mostraram que genes dos grupos TcS I, II, III e IV apresentam expressão diferencial ao longo ciclo de vida do parasita (Cross e Takle, 1993; Schenkman *et al.*, 1994; Frasch, 2000). Para investigar a expressão de genes TcS representantes dos diferentes grupos identificados neste trabalho, em colaboração com a estudante de doutorado em Parasitologia Sara Lopes dos Santos, nós realizamos análises qRT-PCR em tempo real. Através de análises *in silico* de todos os genes TcS foi possível desenhar iniciadores específicos para genes dos diferentes grupos, com exceção do grupo TcSgrupoVI. Doze genes TcS foram selecionados para avaliar a expressão nos estágios tripomastigota, amastigota e epimastigota do parasito. Os resultados mostram que os genes analisados são mais expressos nas formas tripomastigota e amastigota (Figura 12).

A expressão dos genes também se mostrou variável entre os genes que pertencem ao mesmo grupo. O gene TcS5 (verde escuro) apresentou expressão na forma tripomastigota muito acima dos outros genes analisados.

O único gene representante do TcSgrupoIII (TcS8, grupo azul claro) apresentou, como a maioria dos outros genes TcS, mais expresso na forma tripomastigota seguida da forma amastigota, sendo a diferença entre estes estágios evolutivos estatisticamente significativa. O mesmo padrão de expressão foi obtido para o gene TcS32 (TcSgrupoVII, laranja).

Os genes TcS15 e TcS21, ambos do grupos TcSgrupoV (vermelho), apresentaram baixa expressão nas três fases e não apresentam diferença significativa. Os genes TcS24 e TcS25, ambos do TcSgrupoVIII, apresentaram expressão diferente, sendo o gene Tc24 mais expresso na forma tripomastigota, enquanto o gene TcS25 apresentou expressão mais elevada na fase amastigota.

Os genes TcS analisados apresentam-se mais expressos na fase tripomastigota seguida da forma amastigota. A expressão dos genes TcS que pertencem ao mesmo grupo não é homogênea, existindo genes que pertencem ao mesmo grupo mas apresentam grande variação no nível de expressão ou na fase de desenvolvimento do parasita.

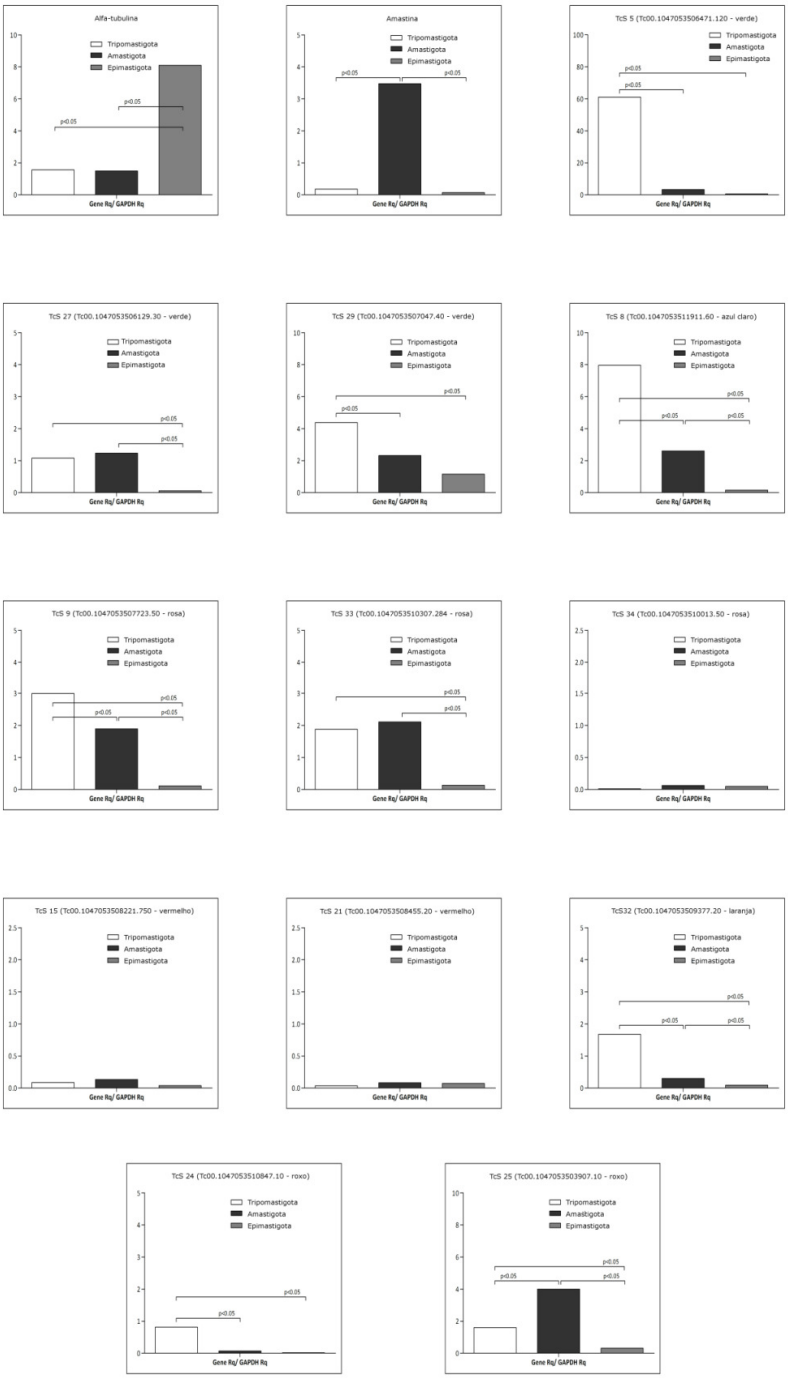


Figura 12. Perfil de expressão de genes TcS analisado qRT-PCR. Os cálculos de quantidade relativa (Rq) foram baseados em curva padrão específica para cada gene TcS. Os valores de Rq para cada amostra de cDNA (TcSRq) foram normalizados com o gene GAPDH (GAPDH Rq), um gene expresso constitutivamente durante o ciclo de vida do parasita. Alfa-tubilina e amastina foram usados como controle de genes mais expressos nos estágios epimastigota e amastigota, respectivamente.

Como não foi verificado expressão dos genes TcSgrupoV nas células tripomastigota de cultura, foi testado também a expressão na forma tripomastigota sanguínea, não sendo verificada expressão.

5.1.7. Análise da variabilidade encontrada na região 3' flanqueadora de TcS

O controle expressão em tripanosomatídeos é principalmente pós-transcricional, mediado em grande parte por motivos regulatórios presentes na região 3' não traduzida (3'UTR) do mRNA maduro que modulam a estabilidade do transcrito. Como seria inviável realizar o mapeamento das regiões 3'UTR de todos os genes TcS, nós realizamos as análises de diversidade dos 300 nucleotídeos após o códon de terminação dos genes e verificamos se há alguma associação entre a similaridade destas sequências e os 8 grupos de TcS que encontramos. Decidiu-se analisar 300 nt, uma vez que estudos prévios realizados pelo nosso grupo demonstrou ser esta a média do tamanho das regiões 3'UTRs de *T. cruzi* (Campos *et al.*, 2008). A diversidade das regiões 3' flanqueadora dos genes TcS é muito grande e maior que a diversidade nucleotídica total das sequências codificadoras e grupos de TcS (Figura 13 e Figura 14).

A comparação da projeção da matriz de distância das sequências 3' flanqueadoras com a classificação de proteínas TcS mostra que sequências 3' flanqueadoras de um mesmo grupo são muito diferentes apresentando-se dispersas, não sendo encontrada uma associação clara entre a classificação MDS de proteínas e a projeção da região 3' flanqueadora (Figura 13 e 10). Poucas foram as sequências 3' flanqueadoras que formaram pequenas regiões que pertencem ao mesmo grupo de proteínas. O TcSgrupoVIII foi o único que essa associação foi mais evidente formando uma região contendo várias sequências 3' flanqueadoras pertencentes aos genes codificadores das proteínas deste grupo, representadas pela cor roxa (Figura 13 e Figura 14). Os membros do TcSgrupoV apresentam-se bem próximos na projeção de proteínas e são muito variáveis na região 3' flanqueadora. As sequências 3' flanqueadoras que pertencem aos genes codificadores dos TcS grupo V e VI mostraram uma dispersão contínua e sobreposta (Figura 13 e Figura 14). As sequências TcSgrupoII não formaram uma região única contendo as sequências 3' flanqueadoras dos genes desse grupo, mas se concentram em três regiões principais (Figura 13 e Figura 14).

As regiões 3' flanqueadoras dos genes SAPA e TCNA, ambas TS expressas na forma tripomastigota (TcSgrupoI), estão muito próximas no MDS (Figura 14). A região 3' flanqueadora de TS-epi, TS presente na forma epimastigota e que também pertence a TcSgrupoI, está localizada distante das 3' flanqueadoras dos genes SAPA e TCNA. As regiões 3' flanqueadoras dos genes gp90 e gp82, ambos expressos na forma tripomastigota metacíclica, e ASP-2, expresso forma amastigota, são apresentadas muito próximas na projeção e pertencem ao TcSgrupoII. As regiões 3' flanqueadoras dos genes 11_SA85-1.1 e TsTc13, ambos expressos na forma tripomastigota, apresentam-se próximas nas projeção, sugerindo um mecanismo de controle de expressão similar, embora apresentem divergência suficiente na sequência de proteínas para as duas serem classificadas em grupos diferentes (Figura 14).

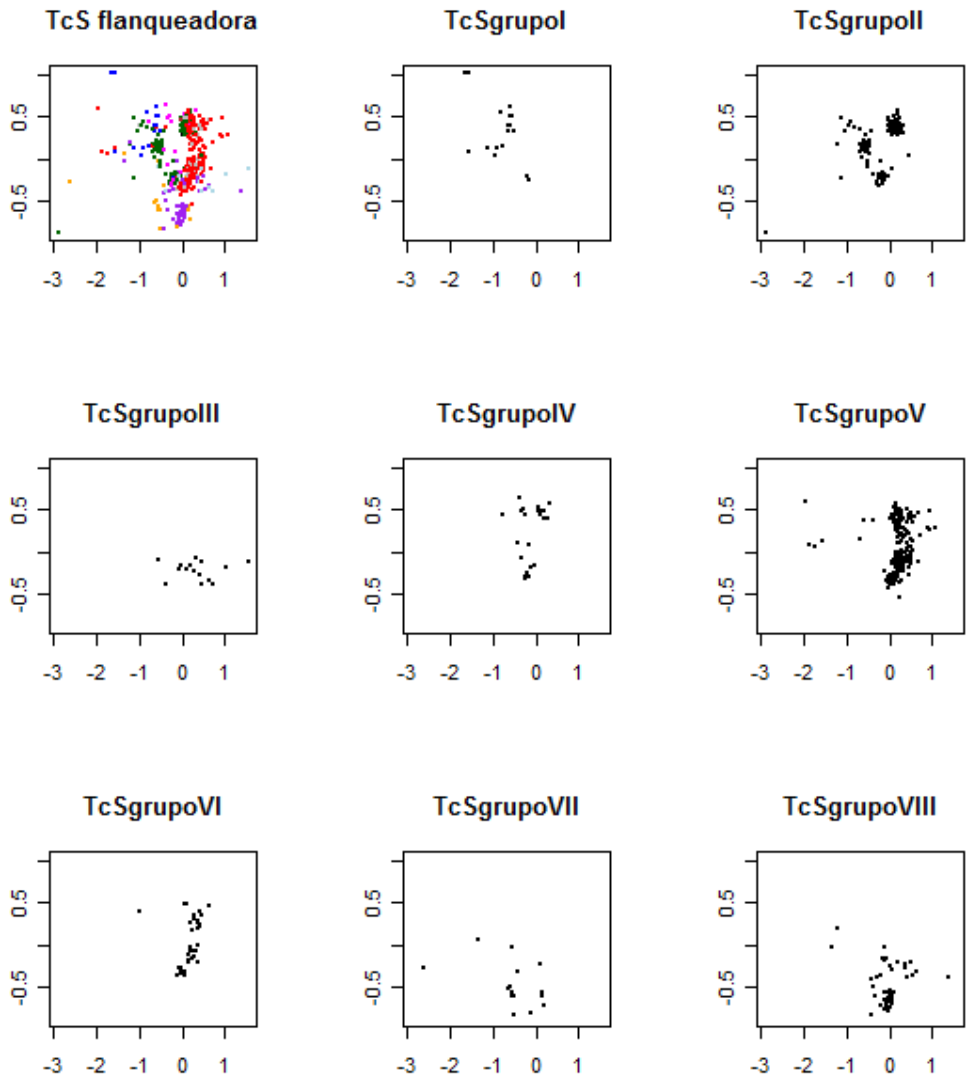


Figura 13. Projeção MDS das sequências 3' flanqueadoras da família TcS. (A) Projeção de todas as sequências 3' flanqueadoras dos genes TcS. Os outros painéis mostram a projeção de cada um dos grupos de TcS identificados. TcSgrupoI (azul), TcSgrupoII (verde escuro), TcSgrupoIII (azul claro), TcSgrupoIV (rosa), TcSgrupoV (vermelho), TcSgrupoVI (cinza), TcSgrupoVII (laranja) e TcSgrupoVIII (roxo). Cada ponto representa a região 3' flanqueadora de um gene e a distância entre os genes reflete sua dissimilaridade. Todos os gráficos estão na mesma escala.

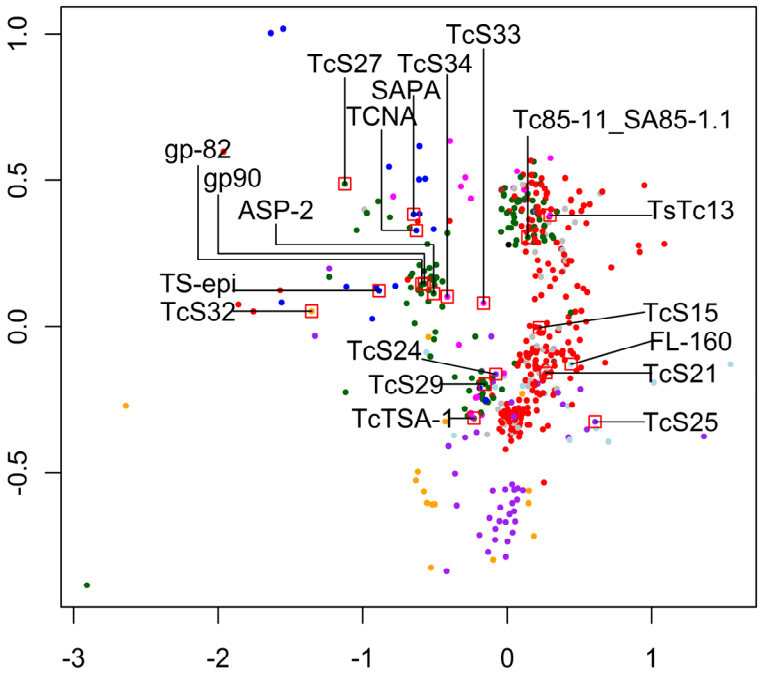


Figura 14. Projeção MDS das sequências 3'flanqueadoras da família TcS. (A) Projeção de todas as sequências 3'flanqueadoras dos genes TcS. As cores representam os grupos de TcS identificados TcSgrupoI (azul), TcSgrupoII (verde escuro), TcSgrupoIII (azul claro), TcSgrupoIV (rosa), TcSgrupoV (vermelho), TcSgrupoVI (cinza), TcSgrupoVII (laranja) e TcSgrupoVIII (roxo). Cada ponto representa a região 3' flanqueadora de um gene e a distância entre os genes reflete sua dissimilaridade.

5.1.8. Antigenicidade dos grupos TcS

Alguns trabalhos mostraram que existem regiões altamente imunogênicas repetitivas nas proteínas TcS, conhecidas como SAPA e EPKSA (Cross e Takle, 1993; Schenkman *et al.*, 1994). Grande parte da resposta imune contra o parasita durante a fase aguda da doença é contra essas regiões (Schenkman *et al.*, 1994). Para verificar a existência de outras regiões antigênicas presentes na família TcS, em colaboração com a estudante de doutorado Sara Lopes dos Santos, foi realizado o ensaio usando soro de camundongos infectados com o

parasito contra um painel de epitopos preditos derivados de diferentes grupos da família. Como previamente mostrado (Leguizamon *et al.*, 1991; Cross e Takle, 1993; Schenkman *et al.*, 1994), os peptídeos correspondentes às repetições SAPA (D5 e D8) e TsTc13 (B5) são altamente antigênicos, sendo reconhecidos pelo soro dos animais infectados (Figura 15). Além desses peptídeos, 11 peptídeos correspondendo a novas regiões antigênicas foram identificadas. Estas novas regiões antigênicas são derivadas não apenas de grupos previamente conhecidos, TcSgrupoI (spots D9 e D10), TcSgrupoIII (spots C9 e C10) e TcSgrupoIV (spot B10), como também de novos grupos TcSgrupoV (A1), TcSgrupoVI (C3), TcSgrupoVII (A10 e B4) e TcSgrupoVIII (A5 e A6).

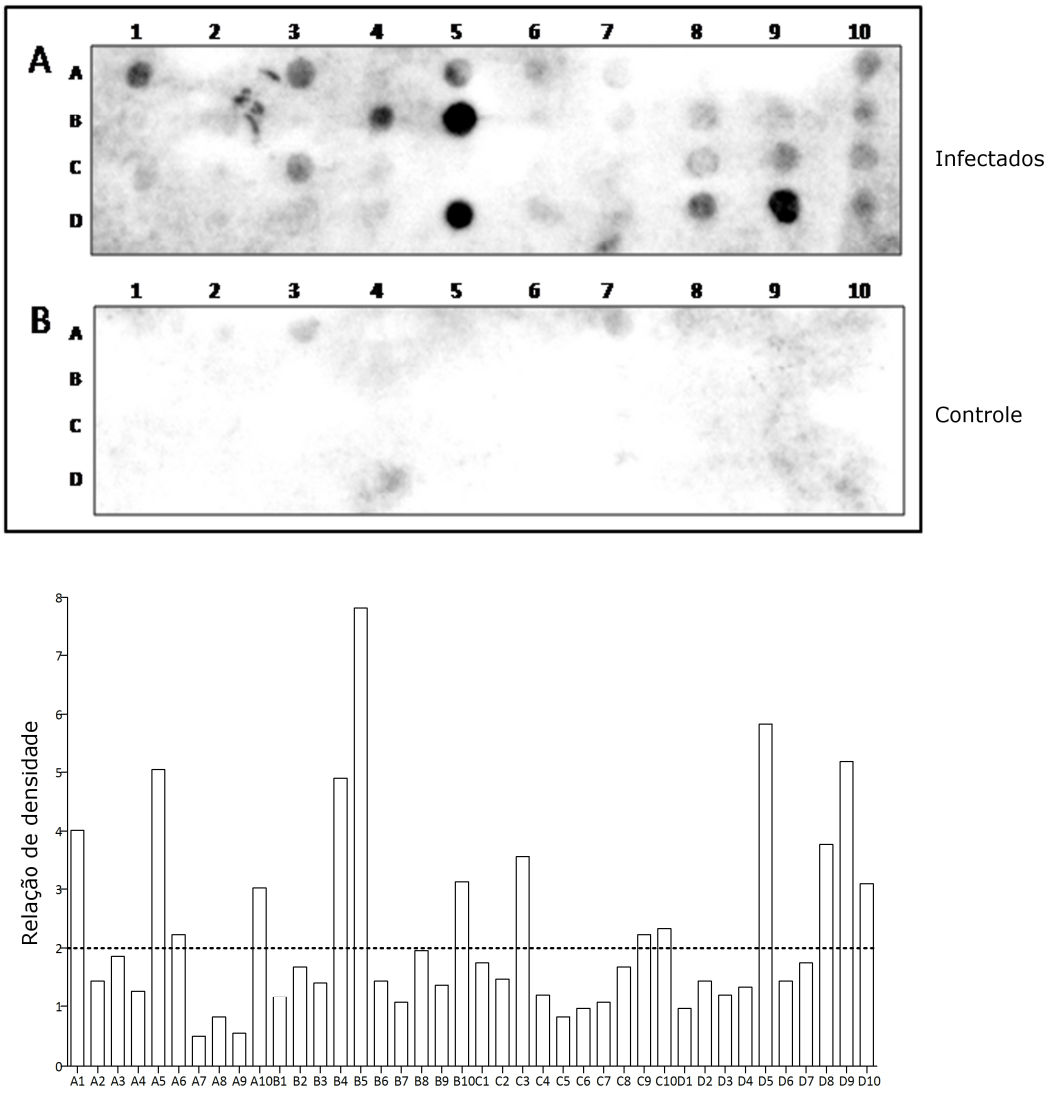


Figura 15. Perfil antigênico de peptídeos TcS. O painel superior mostra o resultado de ensaio imunoblot usando a síntese em membrana (SPOT), sendo (A) conjuntos de soros de camundongos infectados por *T. cruzi;* (B) camundongos não infectados. A reação foi revelada com anticorpo secundário anti-total IgG. O painel inferior mostra a intensidade relativa do sinal de cada spot estimada pela razão das intensidades dos sinais obtidas da reatividade com soro de camundongos infectados e soro de camundongos não infectados. O sinal foi considerado reativo quando a intensidade relativa (RI) ≥ 2. Os peptídeos analisados para cada TcS são: TcS grupo I, D5-D10; TcS grupo III, C9-D4; TcS grupo IV, B5-C1, C3; TcS grupo V, A1, C2, C3, C7; TcS grupo VI, C2-C8; TcS grupo VII, A9-B4; TcS grupo VIII, A2-A8.

5.2. Diversidade das grandes famílias gênicas de *T. cruzi*

A maioria das grandes famílias gênicas de *T. cruzi* codifica proteínas de superfície com extrema diversidade de sequência como as MASPs, Trans-sialidases e Mucinas. Com a disponibilidade dos dados do genoma deste parasito é agora possível realizar análises comparativas da diversidade destas grandes famílias gênicas, o que poderá contribuir para um melhor entendimento do papel de cada uma destas famílias na biologia do parasito. Oito família gênicas foram selecionadas neste estudo. Sete destas correspondem às maiores famílias gênicas do parasito, a saber: TcS, MASP (mucin-associated surface protein), mucina TcMUC, RHS (Retrotransposon hot spot protein), DGF-1 (dispersed gene family-1), GP63 (glicoproteína 63 kDa) e mucin-like. Selecionou-se ainda a família SAP (serine-alanine rich protein) que codifica uma proteína de superfície envolvida em invasão celular. Com exceção das RHSs, que apresenta localização nuclear, todas estas famílias codificam proteínas de superfície. Todas as famílias selecionadas estão clusterizadas no genoma do parasito em grandes regiões cromossômicas que sofrem ou sofreram intensos rearranjos (Bartholomeu *et al.*, 2009).

5.2.1. Diversidade nucleotídica e protéica

As avaliações da diversidade das grandes famílias gênicas envolveram medidas de diversidade das sequências de DNA e proteínas e visualização de projeções espaciais que representam a diversidade. Além da avaliação da diversidade foram definidos grupos dentro das famílias.

As grandes famílias gênicas estudadas neste trabalho apresentaram uma diversidade muito heterogênea entre si existindo famílias com diversidade extremamente baixa e outras com diversidade muito alta. Os resultados tanto para diversidade usando sequências de DNA quanto usando sequências de proteínas encontram-se na Tabela 4. Os quatro métodos usados para estimar a diversidade apresentaram resultados similares. A ordem de classificação das famílias com relação à diversidade foi a mesma para os quatro métodos.

Estas famílias podem ser divididas em três grupos baseado no nível de diversidade encontrado. As famílias DGF-1 e SAP apresentaram os menores valores de diversidade e estes são bem menores que as outras famílias. As famílias RHS e mucin-like apresentaram

diversidade bem maior que as famílias de baixa diversidade, formando o grupo diversidade média. As famílias TcMUC, TcS e GP63 apresentaram alta diversidade e a família MASP apresentou diversidade extremamente alta comparado com as outras famílias, dessa forma essas famílias formam o grupo de alta diversidade. A diversidade das famílias não está relacionada ao número de membros já que famílias com número de membros similares apresentam diversidades tão distintas como DGF-1 e GP63 e SAP e mucin-like.

Tabela 4. Número de sequências analisadas, diversidade nucleotídica e protéica das grandes famílias gênicas de *T. cruzi*. A diversidade nucleotídica foi estimada usando os métodos distância-p e Kimura-2-parâmetros. A diversidade protéica foi estimada usando distância-p e correção de Poisson. O erro das estimativas de diversidade foi calculado usando o método *bootstrap* com 1.000 réplicas indicado entre parênteses.

Família gênica	**Número de sequências**	**DNA**		**Proteína**	
		p-distância (erro)	**K2p (erro)**	**p-distância (erro)**	**Correção de Poisson (erro)**
DGF-1	137	0.110 (0.001)	0.122 (0.001)	0.163(0.003)	0.181(0.004)
SAP	37	0.122 (0.008)	0.135 (0.009)	0.304(0.015)	0.487(0.030)
RHS	114	0.319 (0.004)	0.434 (0.008)	0.472(0.009)	0.653(0.019)
Mucin like	24	0.324 (0.006)	0.538 (0.022)	0.499(0.012)	0.860(0.040)
TcMUC	606	0.408 (0.010)	0.612 (0.023)	0.531(0.023)	0.776(0.052)
TcS	508	0.413 (0.004)	0.662 (0.011)	0.574(0.090)	0.912(0.023)
GP63	122	0.475 (0.005)	0.879 (0.019)	0.596(0.010)	0.989(0.030)
MASP	810	0.609 (0.005)	1.375 (0.034)	0.771(0.014)	1.524(0.067)

As famílias que apresentaram menor variação no comprimento das sequências foram as famílias DGF-1 e SAP que também foram as famílias de menor diversidade nucleotídica e protéica (Figura 16, Tabela 3). Outras famílias apresentaram maior variação como mucin-like e TcS, enquanto que as famílias RHS, TcMUC, GP63 e MASP apresentaram grande variação (Figura 16). As variações nos tamanhos das sequências dentro de cada família certamente afeta o alinhamento múltiplo das sequências usado no cálculo da diversidade das famílias.

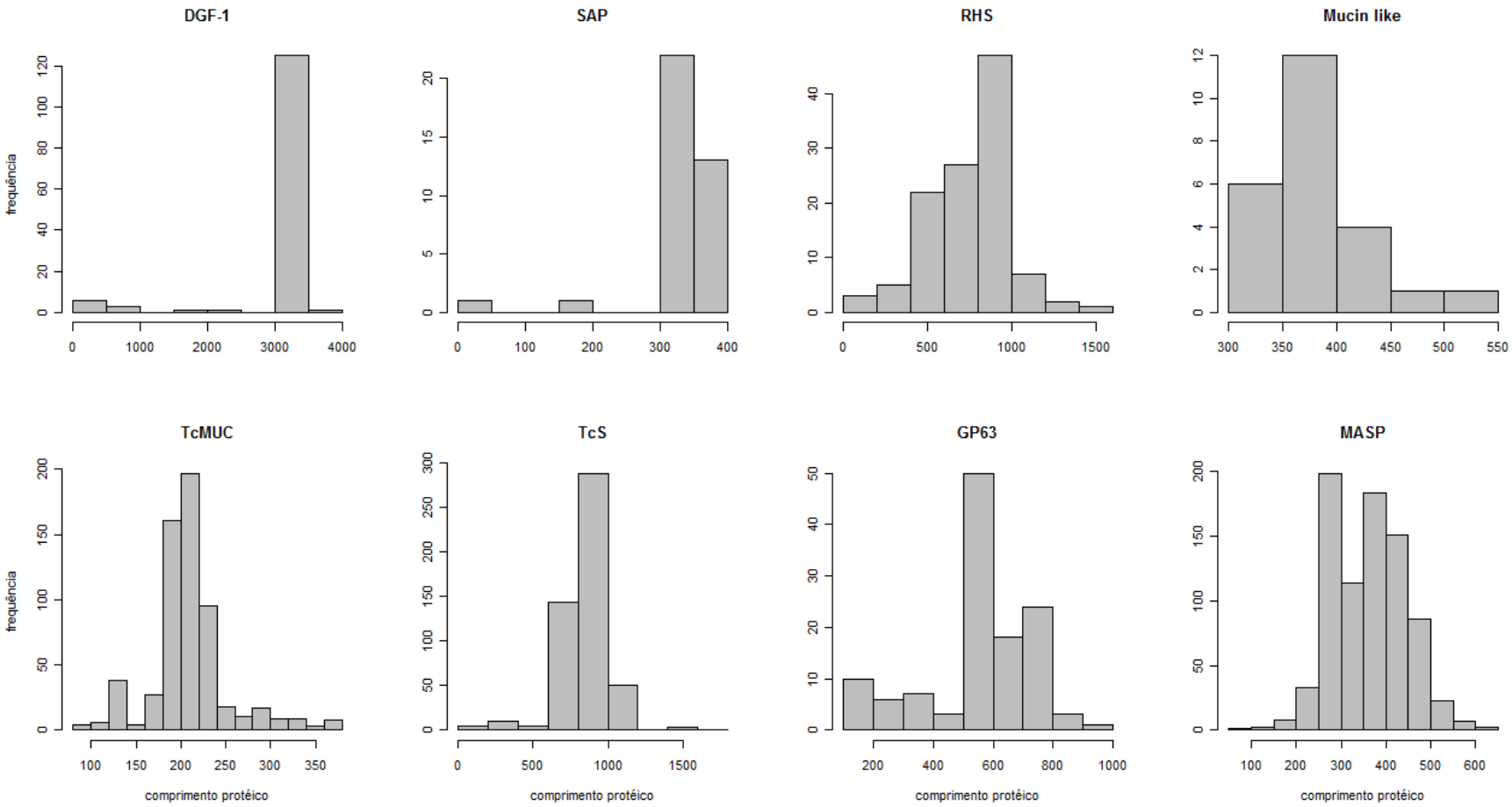


Figura 16. Variação no comprimento das proteínas das grandes famílias de *T. cruzi*. A – DGF-1; B – SAP; C – RHS; D – mucin-like; E – TcMUC; F – TcS; G – GP63 e H – MASP.

5.2.2. Projeção espacial

A projeção do MDS confirma a grande heterogeneidade de diversidade gênica das famílias (Figura 17). A família DGF-1 e SAP apresentam praticamente todos os seus genes em uma região muito pequena na projeção MDS, com somente alguns genes dispersos (Figura 17). As outras famílias apresentam os genes mais distribuídos e regiões com pontos mais próximos, formando grupos de genes mais similares.

Os resultados de projeção espacial também mostram a família MASP apresentando maior diversidade. Os resultados dos índices de diversidade nucleotídica, protéica e projeção espacial mostram e quantificam pela primeira vez a diversidade das grandes famílias gênicas de *T. cruzi* e revela que MASP é a família de maior diversidade.

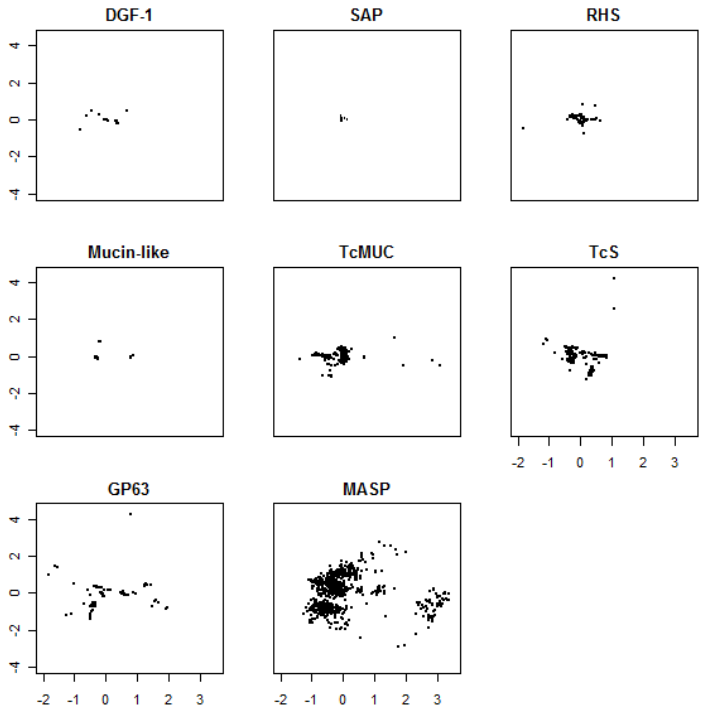


Figura 17. Projeção especial dos resultados *Multidimentional scaling* (MDS) usando a matriz de distância produzida com o alinhamento múltipo das sequências nucleotídicas de cada família. Cada ponto representa um gene e a distância entre os genes reflete sua dissimilaridade. Todos os gráficos estão na mesma escala.

5.2.3. Classificação e diversidade dos grupos de DNA e proteínas

Foram usadas matrizes de distância par-a-par para avaliar a formação de grupos de sequências mais relacionadas gerando um mapa das famílias, possibilitando observação de grupos. A projeção usando matrizes de distância par-a-par facilita observar relações próximas entre pares de sequências que poderiam apresentar relações incorretas devido a problemas enfrentados durante o alinhamento múltiplo das famílias com grande diversidade. A projeção espacial das famílias mostra a formação de regiões com maior concentração de sequências tanto para sequências de DNA quanto proteínas. Os mesmos números de grupos encontrados usando as matrizes de distância de DNA (Figura 18) foram usados nas análises das matrizes de distância de proteínas (Figura 19). Essas projeções foram usadas para determinar o número de grupos e definir os representantes de cada grupo.

Assim como a diversidade não tem relação com o número de representantes em cada família, o número de grupos não está relacionado com a diversidade. Algumas famílias que apresentaram baixa diversidade, como DGF-1 e SAP, apresentaram mesmo ou maior número de grupos comparado com algumas famílias de maior diversidade, como mucin-like e MASP. O número de grupos está relacionado com a variação da diversidade dentro de cada família. Algumas famílias apresentam baixa diversidade, mas distância suficiente entre os representantes para separação em diferentes grupos.

Após a identificação dos grupos, as sequências foram alinhadas e calculada a diversidade nucleotídica e protéica dentro dos grupos usando os métodos distância-p e kimura-2-parâmetros e correção de Poisson (Tabela 5 e 6).

Algumas famílias apresentam-se bem dispersas, refletindo a diversidade das proteínas (RHS, GP63 e MASP) e outras mais agrupadas (TcMUC e TcS). Comparando os grupos formados usando os dados de DNA e proteínas não foram observadas grandes diferenças na relação dos índices de diversidade e número de sequências formando cada grupo nas famílias DGF-1, SAP e mucin-like.

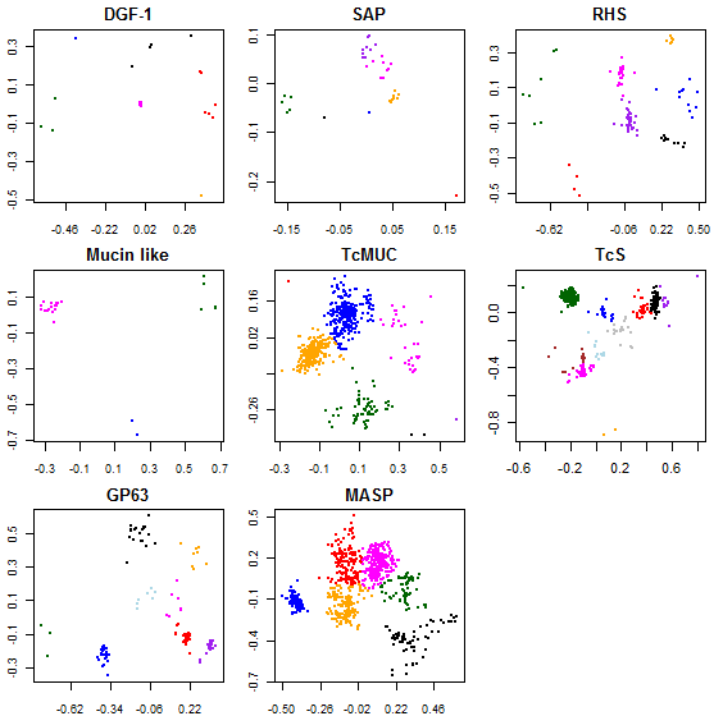


Figura 18. Projeção MDS representando o mapeamento e classificação dos genes de cada família. A projeção foi produzida baseada nas matrizes de distância de DNA com alinhamento par-a-par. Cada grupo é representado por uma cor. Cada ponto representa um gene e a distância entre os genes reflete sua dissimilaridade. Os gráficos são representados em diferentes escalas para permitir uma melhor visualização dos grupos.

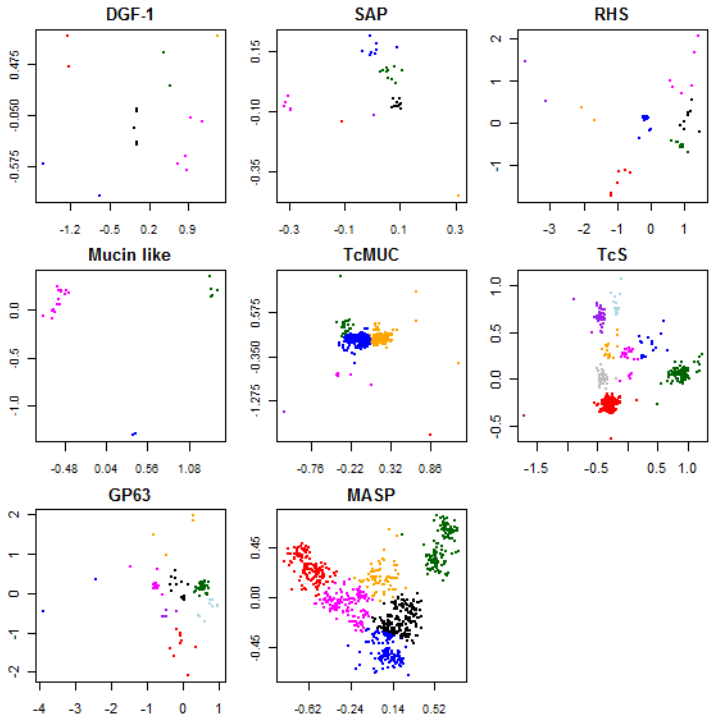


Figura 19. Projeção MDS representando o mapeamento e classificação das proteínas de cada família. A projeção foi produzida baseada nas matrizes de distância de proteínas com alinhamento par-a-par. Cada grupo é representado por uma cor. O número de grupos de cada família usando as sequências de proteínas é o mesmo daquele obtido pelo mapeamento usando as sequências de DNA. Cada ponto representa uma proteína e a distância entre os genes reflete sua dissimilaridade. Os gráficos são representados em diferentes escalas para permitir uma melhor visualização dos grupos.

Algumas famílias apresentam uma enorme variação no número de representantes dentro de cada grupo. Alguns grupos apresentam um número muito grande de genes e outros são formados por apenas uma sequência. Além da variação do número de representantes existe uma variação na diversidade entre os diferentes grupos dentro da mesma família (Tabela 5 e 6).

Tabela 5. Diversidade nucleotídica dentro dos grupos de cada família. O número de sequências formando o grupo e a diversidade nucleotídica usando os métodos distância-p e Kimura-2-parâmetros estão presentes na primeira, segunda e terceira linha, respectivamente. O erro das estimativas de diversidade foi calculado usando o método *bootstrap* com 1.000 réplicas indicado entre parênteses.

Família gênica	**C1 (azul)**	**C2 (rosa)**	**C3 (verde)**	**C4 (laranja)**	**C5 (vermelho)**	**C6 (preto)**	**C7 (roxo)**	**C8 (azul claro)**	**C9 (marrom)**	**C10 (cinza)**
DGF-1	1	122	3	1	6	4	-	-	-	-
	0	0.100(0.001)	0.442(0.007)	0	0.112(0.006)	0.221(0.005)	-	-	-	-
	0	0.108(0.002)	0.889(0.031)	0	0.125(0.007)	0.282(0.009)	-	-	-	-
SAP	1	9	5	8	1	1	12	-	-	-
	0	0.153(0.013)	0.100(0.006)	0.071(0.006)	0	0	0.118(0.007)	-	-	-
	0	0.173(0.017)	0.117(0.009)	0.075(0.007)	0	0	0.130(0.008)	-	-	-
RHS	11	30	7	6	4	13	43	-	-	-
	0.303(0.007)	0.238(0.004)	0.284(0.006)	0.118(0.005)	0.162(0.006)	0.194(0.005)	0.260(0.004)	-	-	-
	0.402(0.013)	0.296(0.007)	0.402(0.013)	0.131(0.006)	0.187(0.009)	0.238(0.008)	0.331(0.007)	-	-	-
Mucin	2	16	6	-	-	-	-	-	-	-
like	0.063(0.008)	0.126(0.006)	0.085(0.005)	-	-	-	-	-	-	-
	0.066(0.008)	0.140(0.007)	0.093(0.007)	-	-	-	-	-	-	-

Família gênica	C1 (azul)	C2 (rosa)	C3 (verde)	C4 (laranja)	C5 (vermelho)	C6 (preto)	C7 (roxo)	C8 (azul claro)	C9 (marrom)	C10 (cinza)
TcMUC	271	33	67	230	1	3	1	-	-	-
	0.414(0.011)	0.557(0.018)	0.320(0.010)	0.327(0.009)	0	0.033(0.008)	0	-	-	-
	0.621(0.027)	0.645(0.020)	0.441(0.021)	0.438(0.017)	0	0.035(0.009)	0	-	-	-
TcS	26	51	264	2	33	75	11	12	16	18
	0,197(0,004)	0,242(0,004)	0,253(0,004)	0,314(0,022)	0,269(0,005)	0,209(0,004)	0,252(0,006)	0,227(0,005)	0,310(0,005)	0,348(0,004)
	0,362(0,016)	0,467(0,015)	0,414(0,010)	0,546(0,068)	0,456(0,014)	0,309(0,008)	0,411(0,020)	0,416(0,022)	0,804(0,004)	1,395(0,114)
GP63	27	7	3	8	26	19	24	8	-	-
	0.164(0.004)	0.441(0.007)	0.209(0.009)	0.054(0.003)	0.156(0.004)	0.435(0.008)	0.161(0.005)	0.294(0.013)	-	-
	0.191(0.006)	0.811(0.030)	0.263(0.015)	0.057(0.003)	0.213(0.007)	1.017(0.071)	0.203(0.008)	0.473(0.040)	-	-
MASP	87	247	75	175	158	68	-	-	-	-
	0.232(0.007)	0.605(0.005)	0.463(0.006)	0.500(0.007)	0.522(0.005)	0.249(0.006)	-	-	-	-
	0.284(0.011)	1.351(0.030)	0.775(0.021)	0.932(0.029)	1.015(0.021)	0.329(0.012)	-	-	-	-

Tabela 6. Diversidade protéica dentro dos grupos de cada família. O número de sequências formando o grupo e a diversidade protéica usando os métodos distância-p e correção de Poisson estão presentes na primeira, segunda e terceira linha, respectivamente. O erro das estimativas de diversidade foi calculado usando o método *bootstrap* com 1.000 réplicas indicado entre parênteses.

Família gênica	**C1 (azul)**	**C2 (rosa)**	**C3 (verde)**	**C4 (laranja)**	**C5 (vermelho)**	**C6 (preto)**	**C7 (roxo)**	**C8 (azul claro)**	**C9 (marrom)**	**C10 (cinza)**
DGF-1	2	5	2	1	2	125	-	-	-	-
	0.170(0.021)	0.113(0.009)	0.245(0.041)	0	0.264(0.012)	0.160(0.003)	-	-	-	-
	0.186(0.025)	0.120(0.009)	0.281(0.054)	0	0.306(0.017)	0.177(0/003)	-	-	-	-
SAP	8	6	10	1	1	10	1	-	-	-
	0.443(0.015)	0.259(0.014)	0.280(0.015)	0	0	0.238(0.015)	0	-	-	-
	0.821(0.050)	0.305(0.019)	0.334(0.022)	0	0	0.277(0.021)	0	-	-	-
RHS	72	6	16	3	6	9	2	-	-	-
	0.449(0.008)	0.496(0.020)	0.404(0.011)	0.436(0.020)	0.450(0.013)	0.416(0.012)	0.624(0.018)	-	-	-
	0.610(0.016)	0.698(0.041)	0.548(0.022)	0.613(0.042)	0.656(0.031)	0.557(0.022)	0.979(0.050)	-	-	-
Mucin	2	16	6	-	-	-	-	-	-	-
like	0.139(0.019)	0.243(0.014)	0.168(0.013)	-	-	-	-	-	-	-
	0.150(0.022)	0.282(0.018)	0.190(0.017)	-	-	-	-	-	-	-

Família gênica	C1 (azul)	C2 (rosa)	C3 (verde)	C4 (laranja)	C5 (vermelho)	C6 (preto)	C7 (roxo)	C8 (azul claro)	C9 (marrom)	C10 (cinza)
TcMUC	331	5	28	237	1	3	1	-	-	-
	0.496(0.023)	0.551(0.016)	0.474(0.019)	0.509(0.022)	0	0.072(0.021)	0	-	-	-
	0.705(0.048)	0.989(0.059)	0.669(0.038)	0.722(0.047)	0	0.077(0.024)	0	-	-	-
TcS	19	25	117	17	227	1	46	15	1	39
	0,494(0,009)	0,250(0,008)	0,419(0,010)	0,448(0,009)	0,396(0,009)	0	0,353(0,009)	0,366(0,010)	0	0,394(0,009)
	0,881(0,027)	0,320(0,012)	0,558(0,018)	0,651(0,020)	0,51390,015)	0	0,453(0,013)	0,492(0,023)	0	0,513(0,016)
GP63	2	26	48	4	10	14	8	10		
	0.489(0.026)	0.296(0.009)	0.415(0.011)	0.495(0.019)	0.595(0.013)	0.660(0.012)	0.125(0.009)	0.362(0.010)	-	-
	0.671(0.050)	0.362(0.014)	0.602(0.025)	0.815(0.057)	1.052(0.055)	1.286(0.057)	0.138(0.011)	0.552(0.027)	-	-
MASP	134	148	132	66	146	184	-	-	-	-
	0.697(0.013)	0.729(0.015)	0.609(0.014)	0.677(0.014)	0.566(0.017)	0.729(0.014)	-	-	-	-
	1.315(0.055)	1.380(0.061)	1.005(0.040)	1.247(0.056)	0.87890.044)	1.362(0.061)	-	-	-	-

5.2.3.1. DGF-1

DGF-1 apresenta uma diversidade nucleotídica baixa (Tabela 5 e 6) e apresenta variação suficiente para a formação de seis grupos (Tabela 5 e Tabela 6, Figura 18 e Figura 19). Grande parte da família tem dispersão e diversidade muito baixa tanto para sequências de DNA quanto proteínas é devido ao fato de que maioria das sequências são muito similares. Na projeção de DNA, o grupo com maior número de sequências apresenta 122 genes, correspondendo a 89% da família, e na projeção de proteína 125 sequências, correspondendo a 91,2% (Tabela 5 e 6). A família apresenta ainda dois grupos na projeção de DNA, C1 (azul) e C4 (laranja), e 1 grupo na projeção de proteínas, C4 (laranja), com somente uma sequência.

Portanto a família DGF-1 é formada por sequências mais distantes correspondendo a 10% das sequências analisadas, enquanto os outros 90% são sequências muito similares formando um único grupo.

5.2.3.2. SAP

SAP também mostrou uma pequena dispersão confirmando sua baixa diversidade. Apesar da baixa diversidade, os genes são distantes o suficiente para serem separados em sete grupos (Figura 18 e Figura 19, Tabela 5 e Tabela 6). O número de membros nos grupos usando sequências de DNA e proteínas se mostrou similar (Figura 18 e Tabela 5). Algumas sequências estão mais distantes formando grupos com poucos membros, como o grupo rosa na projeção de proteínas (Figura 18).

A família SAP apresenta três grupos mais isolados nas sequências de proteínas. Os grupos laranja, vermelho e roxo apresentam apenas 1 sequência mais distante. A distância entre os grupos contribui para o aumento da diversidade da família.

Foi proposta uma definição de quatro grupos da família SAP de acordo com a ocorrência de motivos de endereçamento celular (Baida *et al.*, 2006). Os quatro grupos definidos por estes autores são:

(i) Presença de peptídeo sinal e âncora GPI: proteínas ancoradas na superfície;

(ii) Apenas peptídeo sinal: proteína secretada

(iii) Ausência de peptídeo sinal e âncora GPI: proteínas intracelulares

(iv) Presença de âncora sinal ou ausência de peptídeo sinal e presença de âncora GPI: localização não conhecida

Os quatro diferentes grupos obtidos por Baida *et al.* (2006) foram mapeados nas projeções de DNA e proteínas (Figura 20), mas não foi observada uma correlação entre presença e ausência destes motivos de endereçamento e os grupos obtidos usando nossa abordagem.

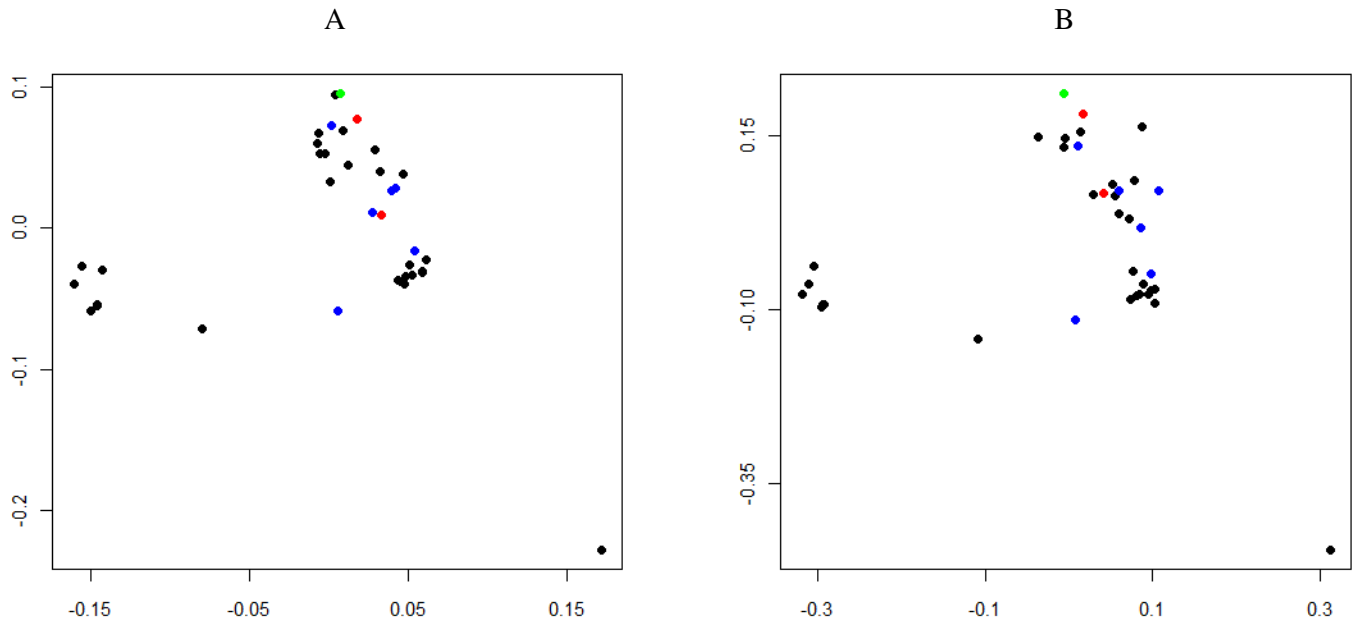


Figura 20. Identificação das proteínas da família SAP que apresentam peptídeo sinal, âncora sinal e sítio de adição de âncora GPI. (A) padrão de dispersão das sequências de DNA, (B) padrão de dispersão das sequências de proteínas. Sequências representadas pela cor preta apresentam peptídeo sinal e âncora GPI; cor azul apresenta somente peptídeo sinal; verde apresenta âncora sinal ou ausência de peptídeo sinal e presença de âncora GPI, vermelho não apresenta peptídeo sinal e âncora GPI.

Análise filogenética da família SAP apresentou vários nós que não tiveram apoio estatístico significativo, mostrando que a relação entre os mesmos não pode ser inferida (Figura 21). Além disso, não é possível a separação das proteínas em ramos monofiléticos apresentando os mesmos motivos para endereçamento de proteínas.

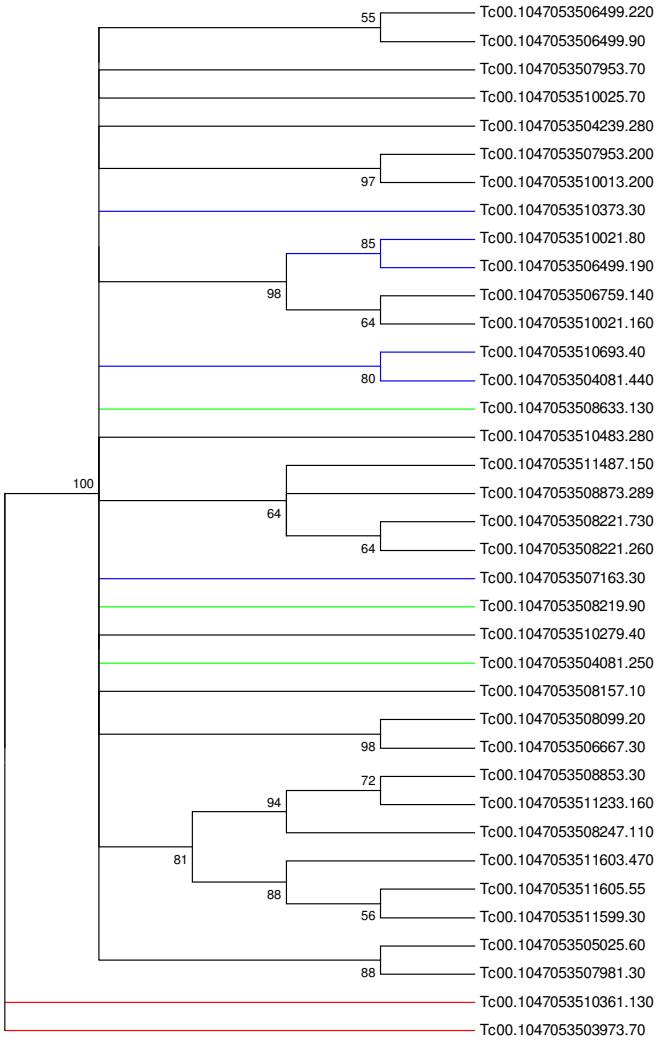


Figura 21. Árvore consenso NJ usando sequências de proteínas da família SAP. Ramos representados pela cor preta são sequências com peptídeo sinal e âncora GPI; cor azul apresenta somente peptídeo sinal; verde apresenta âncora sinal/ausência de peptídeo sinal e presença de âncora GPI; vermelho não apresenta peptídeo sinal e âncora GPI. Foi usado teste de *bootstrap* com 1.000 réplicas e mantidos os nós com mais de 50%.

5.2.3.3. RHS

RHS apresenta uma diversidade maior em relação às duas famílias anteriores e padrão de dispersão dos genes suficiente para ser classificada em sete grupos (Figura 18 e Figura 19). Alguns grupos RHS apresentam muitas sequências com menor diversidade, outros com poucas sequências e alta diversidade, tanto para grupos usando sequências de DNA quanto de proteínas (Tabela 5 e 6). Os maiores grupos de RHS na projeção de DNA são o rosa e o roxo com 30 e 43 dos genes, respectivamente, representando aproximadamente 60% dos genes da família. Ambos os grupos apresentam diversidade intermediária comparado com os outros grupos da família. Na projeção de proteínas encontramos o grupo azul com 72 representantes e sequências mais relacionadas por formarem um grupo com diversidade menor. Os outros grupos são formados por um menor número de sequências e com diversidade similar ou maior que os grupos com mais sequências. A projeção protéica de RHS apresentou ainda o grupo roxo mais distante com dois representantes e com maior diversidade, mostrando que essas proteínas também são muito divergentes entre si.

As sequências da família RHS estão concentradas em somente poucos grupos que representam 60% das sequências analisadas, sendo o restante sequências divergentes que formam grupos pequenos com grande variação.

5.2.3.4. Mucin-like

A projeção da família mucin-like mostra claramente a formação de três grupos, tanto para a projeção usando sequências de DNA como de proteína (Figura 18 e Figura 19). A família mucin-like foi a única família que apresentou os grupos sendo formados pelos mesmos membros usando sequências de DNA e proteínas, mostrando que os grupos são bem definidos. As sequências de DNA apresentaram diversidade proporcional à quantidade de sequências em cada grupo, mas os grupos de proteínas apresentam diferente número de sequências com diversidade similar (Tabela 5 e 6). Assim como a família DGF-1, mucin-like mostrou um grupo com a maioria dos genes (grupo rosa com 66,7% dos genes). A concentração dos genes em apenas um grupo contribui para a baixa diversidade da família. Os outros dois grupos também apresentam baixa diversidade. A diversidade total da família é devida, principalmente, as distâncias entre os grupos.

5.2.3.5. TcMUC

A família TcMUC apresenta uma distribuição contínua dos genes/proteínas, tendo sido formados seis grupos (Figura 18 e Figura 19). Somente alguns genes estão um pouco mais afastados dessa distribuição contínua. A maioria dos genes TcMUC da projeção de DNA pertencem aos grupos azul e laranja, com 271 (44%) e 230 (40%) genes, respectivamente, apresentando alta diversidade (Tabela 5 e 6). Na projeção de proteínas os grupos azul e laranja também apresentam alta diversidade e são formados por um grande número de sequências. Esses são os grupos que formam a maior região de continuidade encontrada na família. Esses dois grupos encontrados na projeção de DNA e proteínas apresentam diversidade similar aos outros grupos mesmo apresentando mais representantes.

A projeção da família TcMUC foi comparada com a classificação da família em grupos TcMUC I, TcMUC II e TcMUC III, que foi baseada na presença/ausência, número de determinadas repetições e sequências específicas encontradas nas proteínas (Buscaglia *et al.*, 2006) (Figura 22 e Figura 23). Existe uma grande correspondência da classificação proposta por Buscaglia e colaboradores (2006) e a distribuição espacial da projeção usando sequências de DNA. A maioria dos genes TcMUC I correspondem ao grupo C3 (verde escuro), enquanto que genes TcMUC II estão distribuídos entre todos os grupos. Entretanto, existem alguns genes cuja dispersão parece não estar de acordo com a classificação. Três genes classificados como TcMUC I foram encontrados no grupo C4 (laranja), que é formado, na sua maioria, por genes TcMUC II. O grupo C3 (verde escuro), que é formado principalmente por TcMUC I, apresenta outros 20 genes TcMUC II, três deles localizados no centro do grupo. A análise das sequências dos genes TcMUC I que estão próximos de TcMUC II e vice-versa mostrou que os genes codificam um mosaico de repetições tanto de TcMUC I como TcMUC II, além de novos motivos (repetições T_8EAP) não previamente associadas a nenhum grupo (anexo 1). Este padrão pode justificar a localização destas sequências em uma posição intermediária no MDS.

Existem somente dois genes classificados como TcMUC III e eles aparecem agrupados na projeção espacial. Existem ainda um dos genes TcMUC que não é classificado em nenhum dos três grupos propostos por Buscaglia e colaboradores (anotado como TcMUC). De fato, esse gene está localizado mais distante da maioria dos outros genes TcMUC na projeção.

As repetições que definem os grupos TcMUC I (T_8KP_2) e TcMUC II (T_8K/QAP) foram encontradas em 56 e 303 membros, respectivamente. Nesses membros ainda foram encontradas variações no número de treoninas encontradas em ambas as repetições. Nos membros TcMUC I foram encontrados motivos com as repetições $T_{6-12}KP_2$, sendo o motivo T_8KP_2 o mais frequente com 71,4%. Já nos membros TcMUC II foi encontrada uma variação maior existindo motivos com as repetições $T_{5-20}K/QAP$, sendo o motivo T_8K/QAP o mais frequente com 38%. Foi encontrado ainda um terceiro motivo em 148 sequências (T_8EAP). Esse terceiro motivo também apresentou grande variação em relação ao número de treoninas (T_5EAP até $T_{20}EAP$), sendo o motivo mais frequente T_8EAP com 36%.

Existem ainda 14 sequências do grupo TcMUC I e 71 sequências do grupo TcMUC II que apresentam também o motivo T_8EAP, e 5 sequências que apresentam tanto os motivos T_8KP_2, T_8K/QAP como T_8EAP. Então as repetições características de TcMUC I e TcMUC II não são exclusivas de cada grupo além de existir uma terceira variante (T_8EAP) muito frequente nas proteínas TcMUC II. As proteínas classificadas como TcMUC I e TcMUC II se mostraram distantes, mas foi possível observar proteínas que apresentam características dos dois grupos formando uma faixa contínua, mostrando que esses dois grupos não são completamente separados.

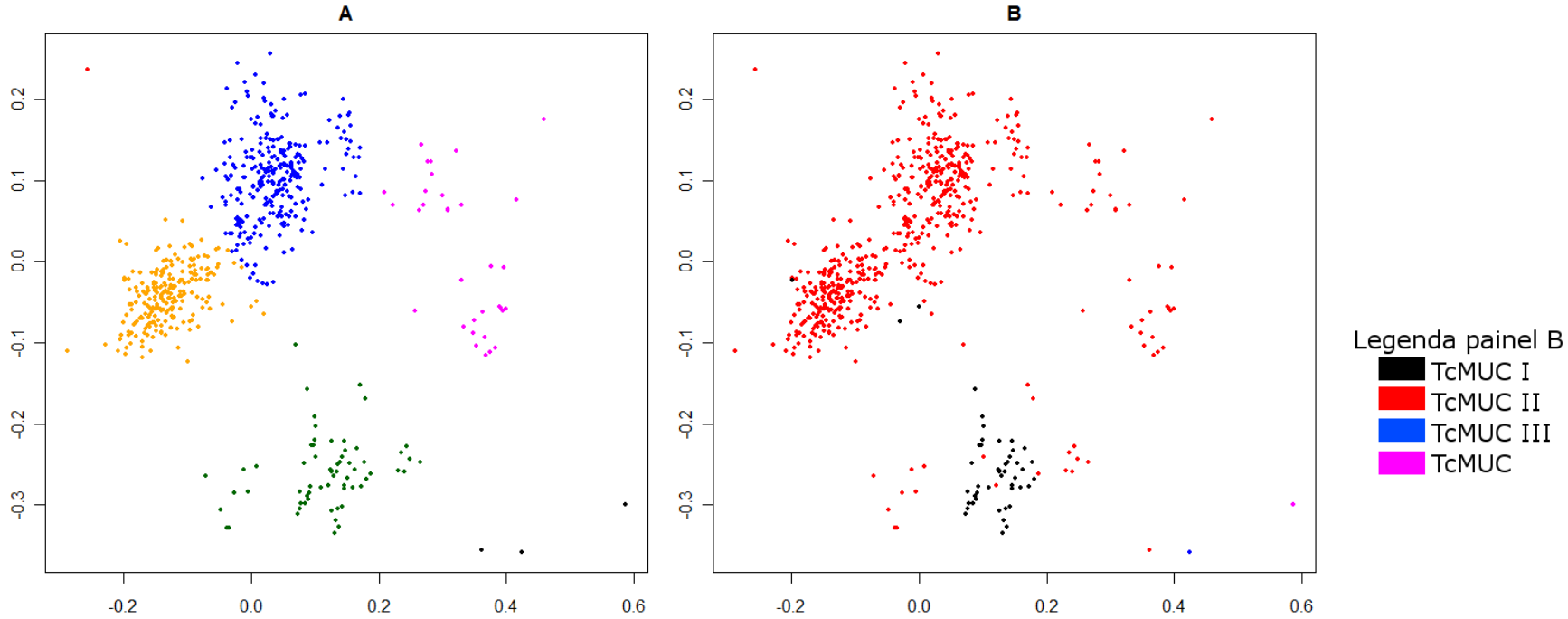


Figura 22. Projeção MDS usando sequências de DNA da família TcMUC. (A) classificação usando o método *kmeans*, (B) classificação proposta por Buscaglia e colaboradores (2006). As cores no gráfico B representam os grupos TcMUC I (preto), TcMUC II (vermelho), TcMUC III (azul) e TcMUC (verde claro).

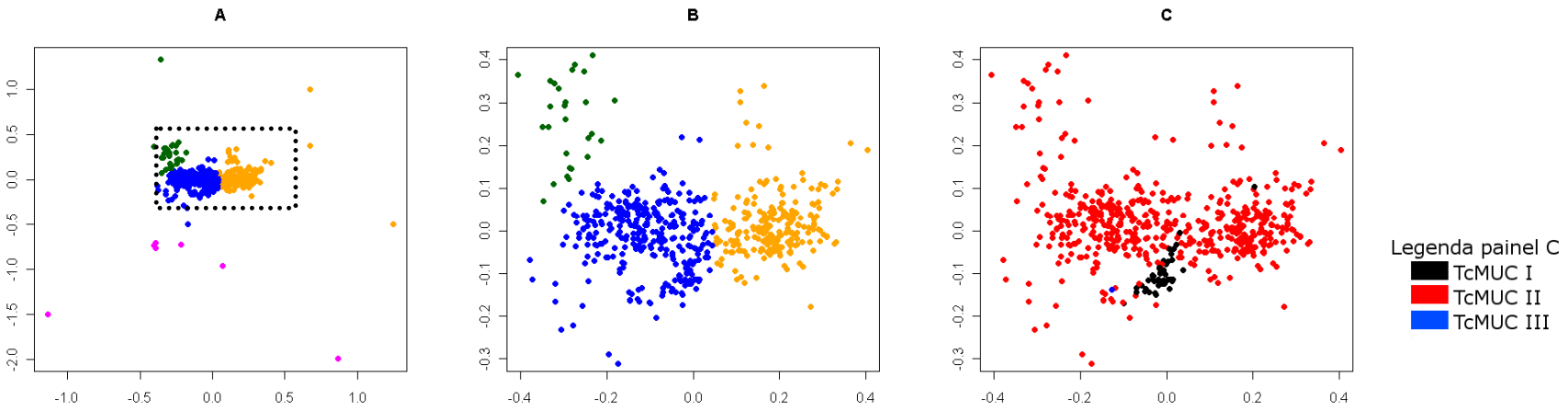


Figura 23. Projeção MDS usando sequências de proteína da família TcMUC. (A) classificação usando o método *kmeans*, (A) aproximação da região destacada no painel A, (C) aproximação da região destacada no painel A e classificação proposta por Buscaglia e colaboradores (2006). As cores no gráfico C representam os grupos TcMUC I (preto), TcMUC II (vermelho), TcMUC III (azul).

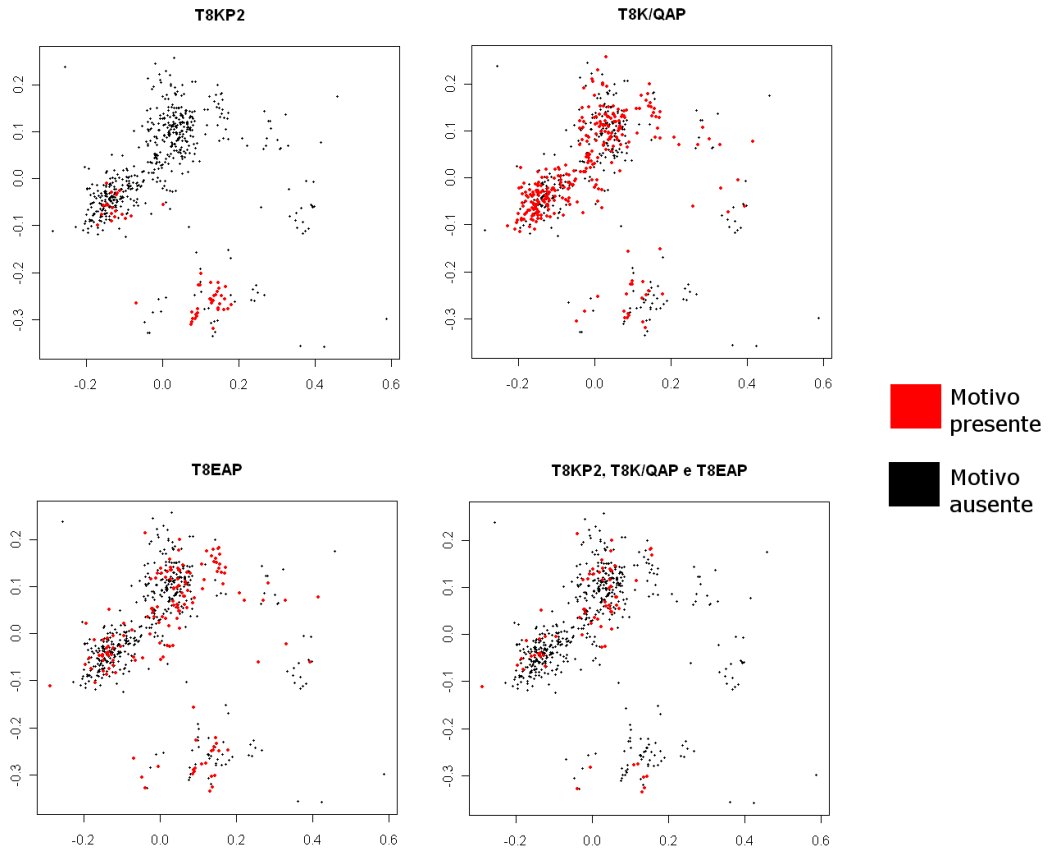


Figura 24. Projeção MDS usando sequências de DNA da família TcMUC. Sequências que apresentam os motivos T_8KP_2; T_8K/QA; T_8EAP; e T_8EAP com motivos T_8KP_2 e T_8K/QAP degenerados.

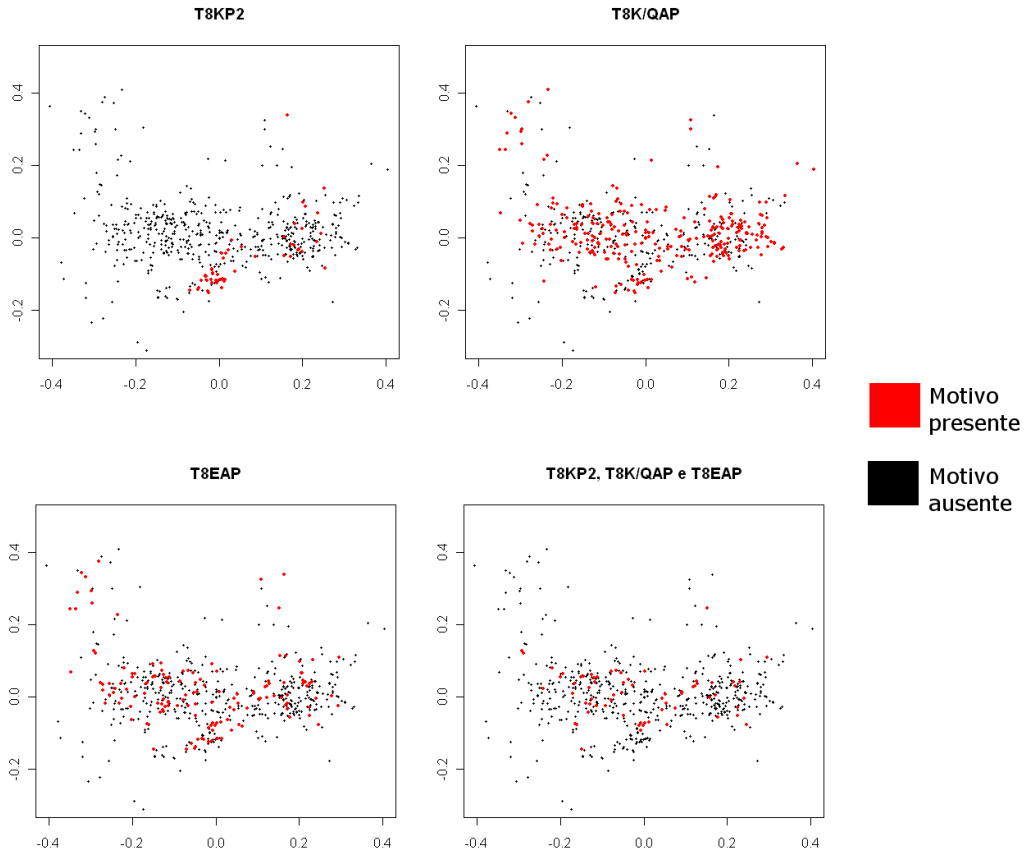


Figura 25. Projeção MDS usando sequências de proteína da família TcMUC. Sequências que apresentam os motivos T_8KP_2; T_8K/QAP; T_8EAP; e T_8EAP com motivos T_8KP_2 e T_8K/QAP degenerados. Todos os painéis são aproximações da região destacada no painel A da Figura 24.

5.2.3.6. TcS

Descrição detalhada da família se encontra na primeira parte deste trabalho.

5.2.3.7. Metaloprotease GP63

GP63 apresentou oito grupos, que são muito distantes contribuindo para a alta diversidade da família (Figura 18 e Figura 19). Além da distância entre os grupos existe a grande dispersão das sequências dentro dos grupos contribuindo para o aumento na diversidade. A família é formada por grupos pequenos com grande diversidade e também grupos com maior número de sequências que apresentam menor diversidade (Tabela 5 e 6).

O motivo HExxH, onde "x" representa qualquer aminoácido, é altamente conservado nas metaloprotease GP63 e importante para essa atividade, sendo os resíduos de histidina e ácido glutâmico essenciais para atividade catalítica de GP63, fazendo a ligação ao zinco e hidrólise de peptídeos, respectivamente (Mcgwire e Chang, 1996). O motivo HExxH foi encontrado na maioria das proteínas (74%) e grupos (Figura 26). O motivo HExxH não foi encontrado em nenhuma sequência dos grupos roxo e vermelho, que são grupos próximos. Algumas sequências não apresentaram o motivo HExxH, enquanto outras sequências, como nos grupos preto e verde escuro, apresentaram formas degeneradas do motivo HExxH. Os motivos degenerados encontrados foram HFxxH em quatro sequências do grupo preto e HExxR em 1 sequência do grupo verde escuro.

Estudos prévios propuseram a separação de membros da família GP63 de *T. cruzi* em dois grupos principais (GP63-I e GP63-II) apresentando diferenças no comprimento das sequências e na ocorrência de peptídeos para endereçamento celular (Cuevas *et al.*, 2003). A identificação das proteínas mais similares às sequências identificadas como GP63 grupo I (gi 31322788, mais similar a Tc00.1047053508611.30) e GP63 grupo II (gi 31322790, mais similar a Tc00.1047053506587.100) mostra que essas duas sequências, fazem parte do mesmo grupo na análise de projeção e agrupamento (Figura 26). Para verificar se existem sequências similares a GP63-I e GP63-II em outros grupos foram identificadas todas as sequências anotadas como GP63 que apresentavam tanto similaridade quanto cobertura mínima de 80% com cada uma destas sequências. Foram identificadas 7 e 22 sequências mais similares aos grupos I e II, respectivamente. Todas estas proteínas similares a GP63-I e GP63-

II fazem parte do mesmo grupo, mostrando que realmente a variabilidade da família é muito maior que a representada pelos dois grupos (dados não mostrados).

Membros da família GP63 analisados no nosso trabalho incluem sequências que não apresentam o motivo HExxH típico dessas proteínas, sendo essas proteínas deficientes do motivo metaloprotease possivelmente incapazes de exercer atividade metaloprotease.

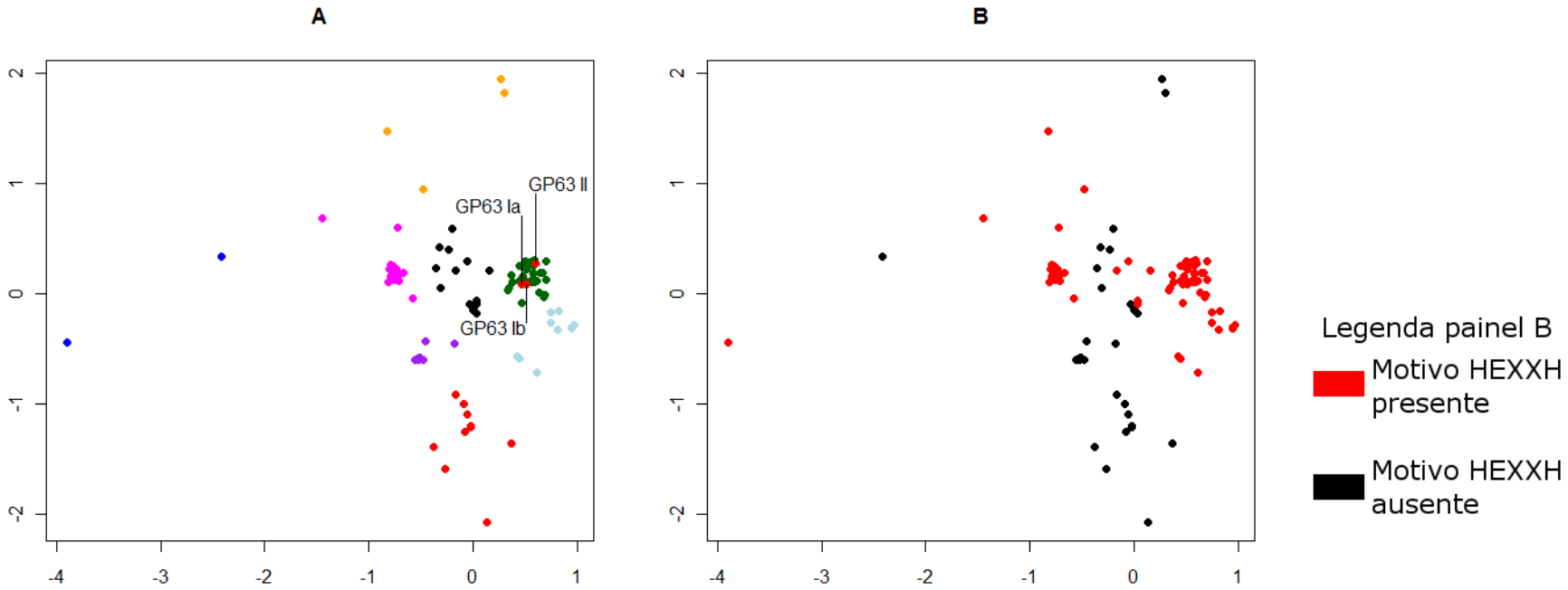


Figura 26. Identificação das proteínas que apresentam similaridade com GP63-I e GP63-II de acordo com a classificação de Cuevas e Cazzulo (2003). (A) Mapeamento das proteínas mais similares às sequências representantes dos grupos GP63-I e GP63-II na projeção MDS das proteínas da família GP63. Números identificadores (GI) das sequências usadas como referência para cada um dos grupos são: Tc00.1047053508609.10 (GP63-Ia, GI: 31322786); Tc00.1047053508611.30 (GP63-Ib, GI: 31322788), Tc00.1047053506587.100 (GP63-II, GI: 31322790). (B) Identificação de proteínas que apresentam o motivo conservado HExxH. As proteínas que apresentam o motivo HExxH estão concentradas em duas regiões correspondentes aos grupos verde escuro e rosa da projeção de proteína. Proteínas que apresentam o grupo HExxH estão representadas em vermelho e proteínas que não apresentam o motivo estão representadas em preto.

5.2.3.8. MASP

Os genes MASP apresentam dispersão e variação suficiente para a formação de seis grupos, todos eles com um número grande de genes (Figura 18 e Figura 19). Assim como a família TcMUC, a grande diversidade da família MASP é devido a grande dispersão dos genes, apresentando em algumas partes dispersão contínua, principalmente na projeção de proteínas, formando grupos com alta diversidade (Tabela 5 e 6). Esta família apresentou também a menor variação de diversidade e número de sequências entre os grupos, mas apresentou também todos os grupos com alta diversidade mostrando que existe grande divergência tanto entre os grupos quanto dentro dos grupos.

Os 39 genes quimeras que fazem parte da família MASP, possuindo extremidade N- ou C-terminal de TcMUC ou C-terminal de TcS (Bartholomeu *et al.*, 2009), pertencem aos grupos vermelho, rosa, laranja e verde da projeção usando sequências de DNA, esses grupos apresentam mais relacionados e próximos, formam uma dispersão contínua no centro do MDS (Figura 27). Aproximadamente 82% dos genes quimeras são encontrados nos grupos rosa e vermelho na dispersão de DNA.

A família MASP possui a maior diversidade dentre as proteínas de superfície, existindo um conjunto de sequências com variação contínua sem uma separação completa dos grupos formados por essas sequências. As sequências quimeras são mais frequentes nas sequências encontradas nesse conjunto de sequências que apresentam variação contínua, aumentando a variação desses grupos.

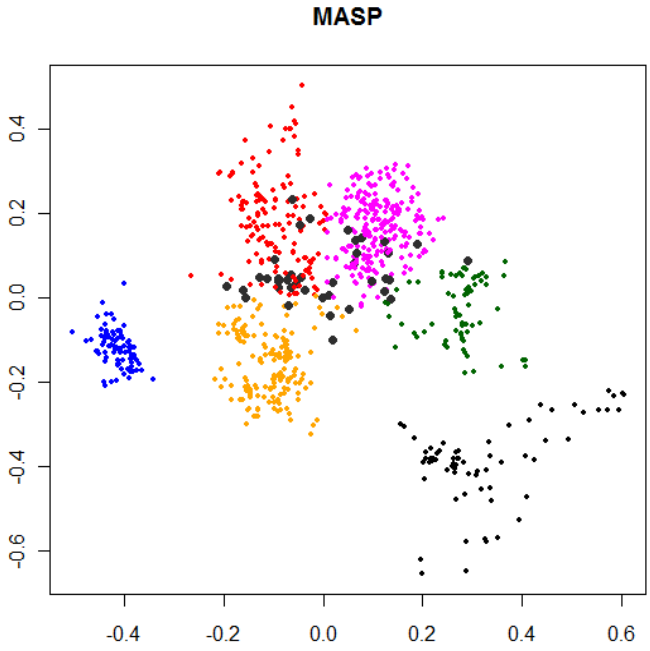


Figura 27. Identificação dos genes quiméricos de MASP na projeção MDS usando a matriz de distância par-a-par de DNA. Os genes quimeras estão destacados em cinza.

5.2.4. Correspondência entre as classificações usando as sequências de DNA e proteínas

A fim de verificar se existe uma correspondência entre a classificação de DNA e proteína para as famílias analisadas neste estudo, a cor de cada gene atribuída na classificação dos grupos usando os dados de DNA (Figura 18) foi mapeado na projeção MDS de proteína da Figura 19.

Como pode ser visto na Figura 28, a maioria das famílias mostrou boa correspondência comparando as duas projeções, com exceção da família MASP. Os grupos formados usando sequências de DNA de MASP C2 (rosa), C4 (laranja) e C5 (vermelho) foram os que apresentaram menor correspondência, mostrando um padrão de dispersão mais espalhado e misturados entre si na projeção de proteínas. Os outros grupos (azul, preto, verde escuro) apresentaram boa correspondência da classificação do DNA com a distribuição usando as distâncias de proteína (Figura 28).

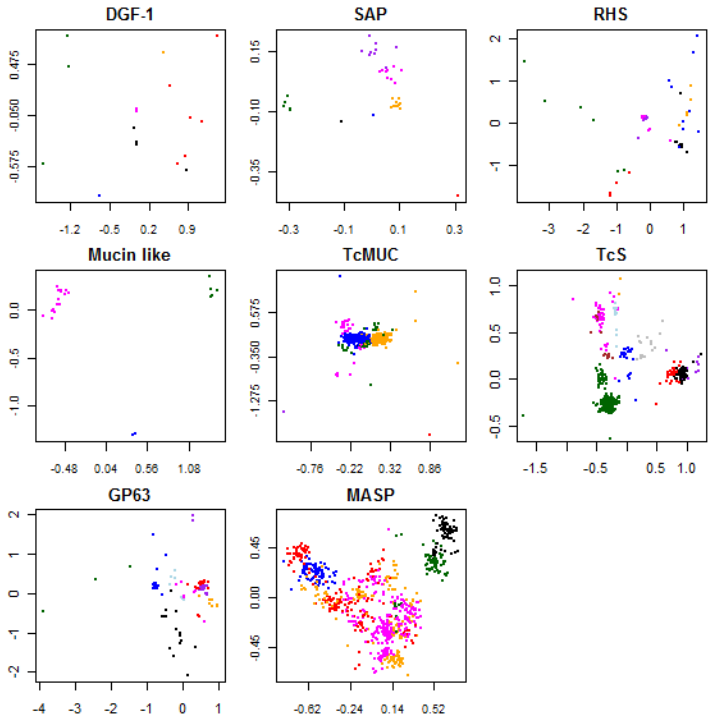


Figura 28. Projeção MDS usando matriz de distância das sequências de proteínas produzida por alinhamentos par-a-par. As cores de cada proteína foram atribuídas de acordo com a classificação das matrizes de distância nucleotídica.

5.2.5. Comparação entre as distâncias nucleotídicas e protéicas

A falta de correspondência encontrada em algumas sequências de DNA e proteínas das diferentes famílias pode ter surgido através de mutações *indel* que provocaram a mudança da fase de leitura do gene, levando a produção de proteínas que apresentam uma variação maior do que aquela encontrada comparando os pares de sequências de DNA. A fim de identificar quais famílias apresentam essa variação foi realizada a comparação de todos os pares de distâncias de DNA e proteínas. As distâncias de DNA e proteína foram comparadas para todos os pares de sequências para avaliar o efeito das mutações *indel* que teriam provocado mudança da fase de leitura e provocado alterações drásticas nas proteínas (Figura 29).

Como esperado, normalmente existe uma correlação positiva entre as duas medidas de distância. As famílias mucin-like e SAP foram as que mostraram maior correspondência das medidas para todos os pares de genes. Dentro das outras famílias essa correspondência não é observada para todos os pares de sequências (Figura 29).

Existem casos apresentando alta diversidade nucleotídica e baixa diversidade protéica indicando a ocorrência de substituições silenciosas, e também existem casos de pares com baixa diversidade nucleotídica e alta diversidade protéica, indicando que pode ter ocorrido uma mutação (*indel*) que provocou a mudança da fase de leitura de uma dessas sequências.

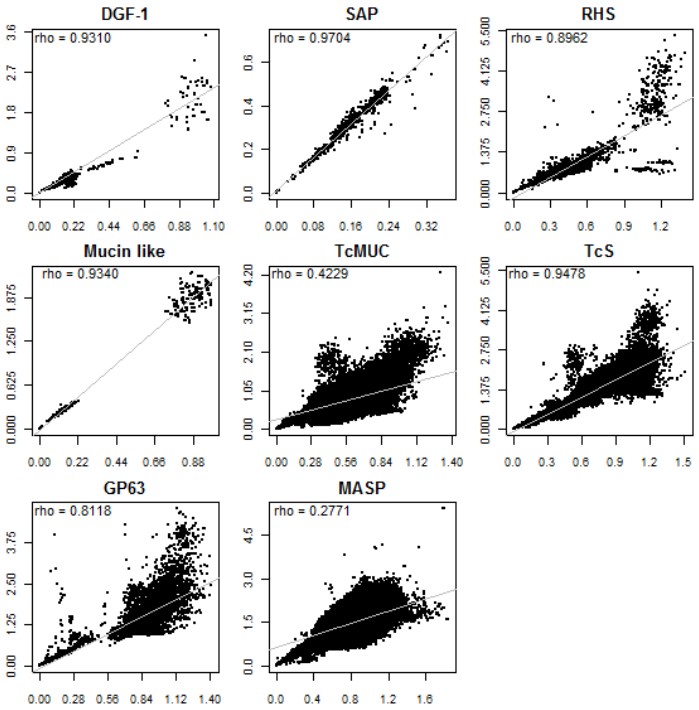


Figura 29. Comparações das distâncias de DNA e proteínas encontradas nos alinhamentos par-a-par de cada família. O eixo x representa a distância de DNA e o eixo y representa a distância protéica. A linha que melhor representa os dados é mostrada em cinza. O valor *rho* mostra a correlação entre as duas distâncias.

Como a família MASP apresentou o menor valor de correlação entre a distância nucleotídica e protéica, as mesmas análises foram realizadas para cada um dos grupos de DNA (Figura 30). As comparações entre as distâncias nos grupos da família MASP apresentam também valores menores na correlação, mostrando que dentro dos grupos existe também uma alteração da relação entre as duas medidas de distância.

As correlações nos grupos azul, rosa, laranja e preto apresentam mais similares e aproximadamente *rho*=0,36, enquanto que os grupos verde escuro e vermelho apresentaram valores menores, 0,2032 e 0,1226, respectivamente. Vale ressaltar que o

grupo vermelho que apresentam menor valor de correlação possui o maior número de sequências quiméricas, o que poderia pelo menos em parte justificar a menor correspondência entre as duas métricas neste grupo.

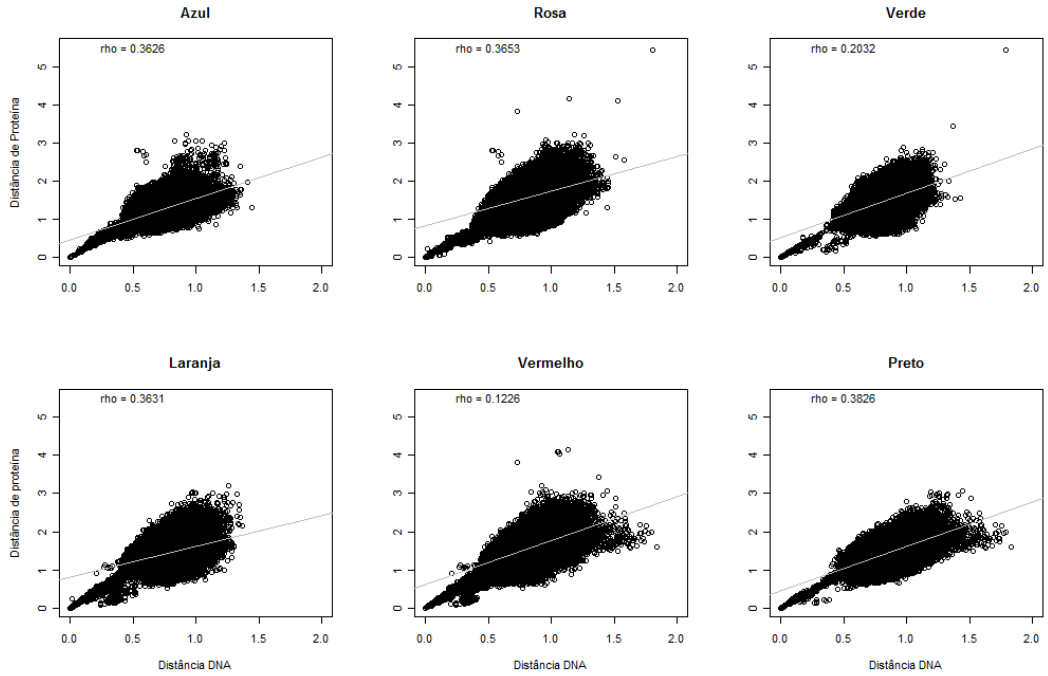


Figura 30. Comparações das distâncias de DNA e proteínas encontradas nos alinhamentos par-a-par de cada um dos grupos da família MASP. O eixo x representa a distância de DNA e o eixo y representa a distância protéica. A linha que melhor representa os dados é mostrada em cinza. O valor *rho* mostra a correlação entre as duas distâncias.

Como mostrado anteriormente, a família MASP apresenta a maior diversidade de sequência e tamanho dentre todas as famílias analisadas o que dificulta o alinhamento das sequências. As sequências da família MASP também foram separadas usando como critério o tamanho das sequências para tentar evitar falta de correspondência devido possíveis problemas encontrados no alinhamento de sequência com tamanhos tão diferentes. Os resultados da comparação das duas distâncias mostra que existe uma relação direta forte entre as duas distâncias, ou seja, aumentando a distância nucleotídica aumenta a distância protéica (Figura 31).

As distâncias e classificação encontradas na sequências de DNA e proteínas apresentam uma boa correlação exceto para a família MASP. Parte dessa falta de

correspondência da família MASP pode ser devido à dificuldade em se alinhar sequências tão diferentes ou devido a mutações de mudança de fase de leitura que provocam grandes alterações na sequências de proteínas.

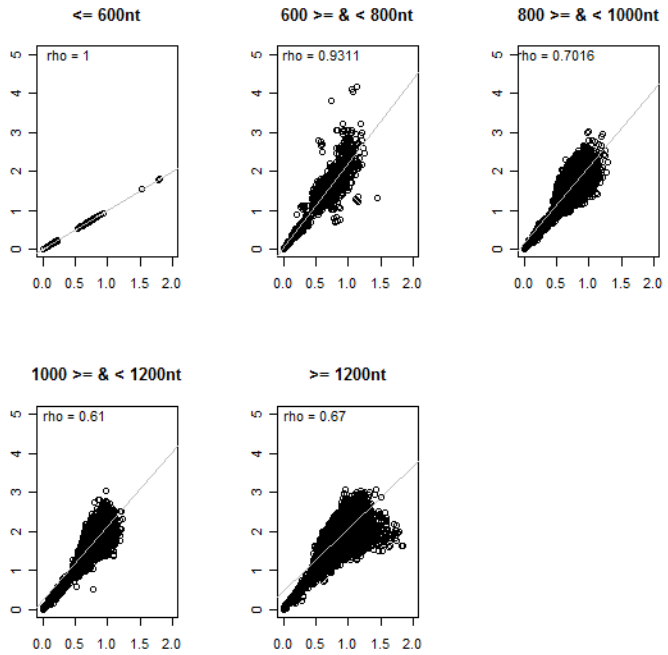


Figura 31. Comparações das distâncias de DNA e proteínas encontradas nos alinhamentos par-a-par de MASP separados por tamanho do gene. O eixo x representa a distância de DNA e o eixo y representa a distância protéica. A linha que melhor representa os dados é mostrada em cinza. O valor *rho* mostra a correlação entre as duas distâncias.

5.2.6. Diversidade das sequências 3' flanqueadoras das grandes famílias gênicas de *T. cruzi.*

Conforme mencionado anteriormente, a regulação da expressão gênica em tripanosomatídeos é pós-transcricional e feita principalmente através de motivos regulatórios presentes na sequência 3'UTR do mRNA maduro que modulam a estabilidade do transcrito. Na ausência de dados funcionais, o conhecimento sobre a diversidade dessas sequências poderia dar indícios sobre se os membros de uma mesma família estaria sujeitos a mecanismos de controle de expressão semelhantes, caso estas

sequências sejam muito similares, e/ou se diferentes grupos estariam submetidos a diferentes mecanismos de controle de expressão gênica, caso haja uma associação entre a clusterização das 3'UTRs e a classificação dos diferentes grupos proteínas da família. Como ainda não há dados de RNAseq disponíveis para *T. cruzi* que permitiriam o mapeamento em larga escala das regiões 3'UTRs dos transcritos do parasito, nós decidimos realizar as análises de diversidade dos 300 nt abaixo do códon de terminação dos genes das famílias, uma vez que nosso grupo de pesquisa demonstrou ser este o valor médio de tamanho das regiões 3'UTR de *T. cruzi* (Campos *et al.*, 2008).

Interessantemente, apesar das famílias TcMUC e MASP apresentarem valores altos de diversidade nucleotídica e protéica, as sequências 3' flanqueadoras destas famílias são as mais conservadas dentre todas as famílias analisadas (Figura 32). Ambas as famílias apresentam grupos com pouca dispersão e apresentando variação contínua. Apresentam também poucas sequências mais distantes que ficam mais afastadas dos principais grupos. Na família TcMUC ainda é possível observar a formação de duas regiões separadas com maior concentração de sequências, mostrando existir dois grupos principais.

As sequências 3' flanqueadoras das famílias DGF-1, RHS, TcS e GP63 apresentaram maior dispersão quando comparadas com o padrão de dispersão das sequências das famílias TcMUC e MASP (Figura 32). As famílias DGF-1 e RHS apresentaram regiões distantes umas das outras com concentração de sequências e continuidade da dispersão, mostrando gradiente de diversidade dentro dessas regiões. A família GP63 apresentou uma distribuição mais ampla e menos contínua comparada com as sequências 3' flanqueadoras de outras famílias. A família TcS também apresenta uma distribuição ampla e contínua em algumas partes mostrando uma grande variação nestas sequências.

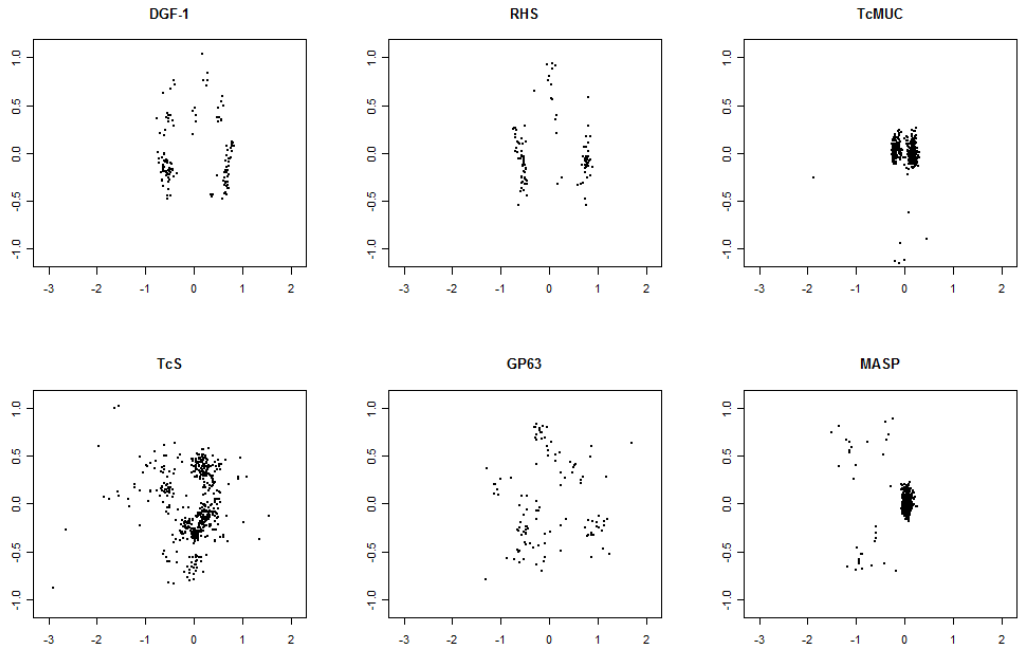


Figura 32. Projeção MDS das sequências 3' flanqueadoras (300 nt após o códon de terminação dos genes) para cada família. Todos os gráficos estão na mesma escala.

Nós em seguida mapeamos a classificação dos grupos protéicos de cada família no padrão de dispersão das sequências 3' flanqueadoras, a fim de ter algum indício de que sequências 3' flanqueadoras similares pudessem controlar a expressão de grupos específicos de cada família. A seguir estão descritas as análises para cada família.

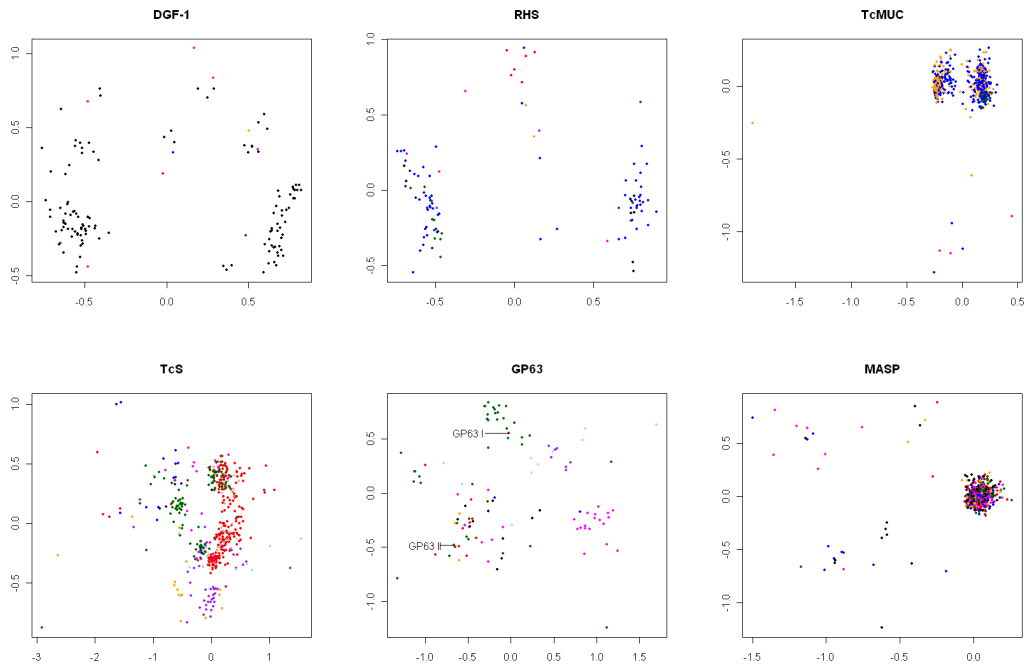


Figura 33. Projeção MDS das sequências 3' flanqueadoras (300 nt) após o códon de terminação para cada família e mapeamento da classificação dos grupos protéicos de cada família no padrão de dispersão das sequências 3' flanqueadoras. Os gráficos são representados em diferentes escalas para permitir uma melhor visualização.

5.2.6.1. DGF-1 – 3'flanqueadora

A comparação da projeção da região 3' flanqueadora com a classificação encontrada usando sequências de proteínas mostra que sequências mais distantes na classificação de proteínas apresentam-se mais afastadas das regiões com maior concentração de sequências 3' flanqueadoras (sequências mais similares) (Figura 33).

5.2.6.2. GP63 – 3'flanqueadora

A projeção mostra que os genes GP63 apresentam a região 3'flanqueadora bem diversa apresentando grande dispersão. A família GP63 possui pelo menos dois grupos, apresentando regiões 3' flanqueadoras mais similares representadas pela cores verde escuro e rosa, regiões mais próximas formando uma região quase exclusiva para as

sequências 3'flanqueadoras pertencentes os genes desses dois grupos (Figura 33). Os grupos roxo e azul claro também parecem formar uma pequena região com sequências 3'flanqueadoras mais similares, mas essa região não está clara como para as sequências pertencentes aos genes dos grupos verde escuro e rosa.

Existe ainda outra região com ampla dispersão que apresenta várias sequências 3' flanqueadoras pertencentes a diferentes grupos protéicos. Essa região mostra sequências mais próximas, entretanto possuem diversidade por estarem distantes e mostra que genes GP63 de diferentes grupos apresentam sequências 3'flanqueadoras similares.

5.2.6.3. MASP – 3'flanqueadora

De modo interessante, apesar dos índices de diversidade nucleotídica e protéicos para MASP serem os mais altos quando comparados com as outras famílias, o inverso é observado quando se analisa a região 3' flanqueadora. A região após o códon de terminação dos genes da família MASP apresenta dispersão muito baixa mostrando que essas sequências são muito similares (Figura 33). A comparação da projeção das sequências 3'flanqueadoras com a classificação encontrada usando sequências de proteínas mostra que não existe a separação das sequências 3'flanqueadoras mais similares pertencentes a um determinado grupo protéico. As sequências são tão similares que as cores dos grupos protéicos na projeção apresentaram-se totalmente misturadas (Figura 33).

5.2.6.4. TcMUC – 3'flanqueadora

A projeção das sequências 3'flanqueadoras da família TcMUC apresentou a formação de duas regiões com concentração de sequências (Figura 34). Essa projeção foi comparada com a classificação de proteínas. Essa comparação mostra que existe uma concentração de proteínas do grupo laranja em uma das regiões do MDS das sequências TcMUC-3'flanqueadora, enquanto a outra região TcMUC-3'flanqueadora apresenta maior concentração de sequências pertencentes a genes codificadores de proteínas do grupo azul. Padrão similar é encontrado quando comparamos a projeção

TcMUC-3'flanqueadora com a classificação das famílias descrita por Buscaglia *et al*. (2006), apresentando maior concentração de sequências 3'flanqueadoras de genes codificadores de proteínas TcMUC I em uma região, enquanto as sequências de genes codificadores de proteínas TcMUC II estão mais concentradas em outra região. Entretanto, não existe uma região exclusiva representada apenas por TcMUC I ou TcMUC II.

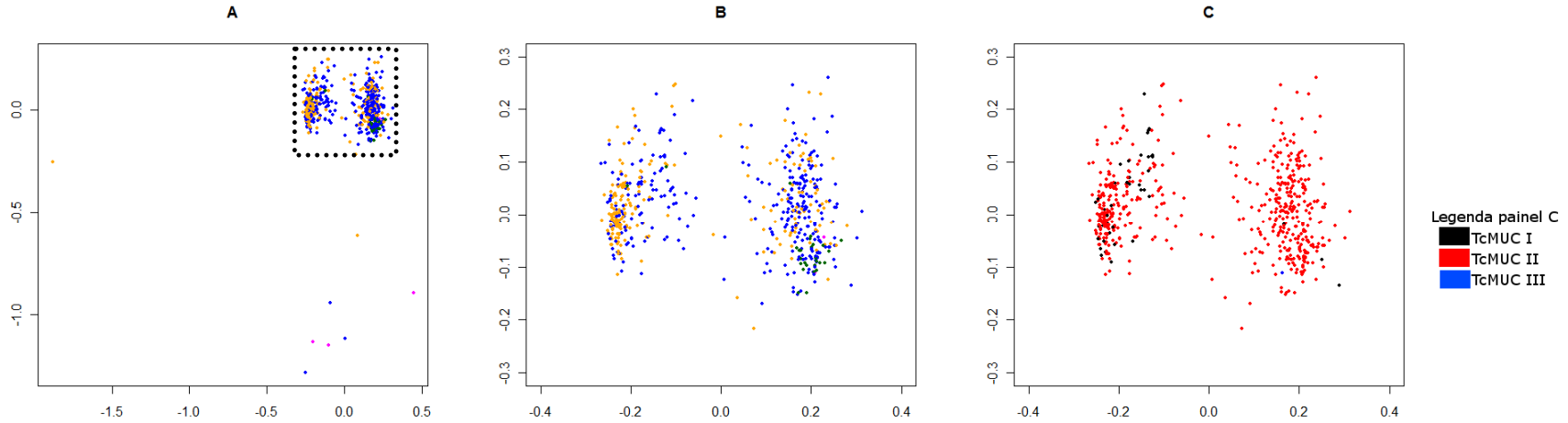


Figura 34. Projeção MDS das sequências 3'flanqueadoras (300 nt) após o códon de terminação dos genes da família TcMUC. (A) As cores representam os 7 grupos definidos na projeção MDS usando as sequências de proteínas, (B) aproximação da região destacada no painel A, (C) aproximação da região destacada no painel A, TcMUC I (preto), TcMUC II (vermelho), TcMUC III (azul) e TcMUC (verde claro).

5.2.6.5. RHS – 3'flanqueadora

A projeção das sequências 3'flanqueadoras da família RHS apresenta a formação de três regiões contínuas, duas apresentando uma região de maior continuidade (Figura 33). A comparação da projeção com a classificação encontrada usando a projeção MDS de sequências protéicas RHS mostra que os grupos de maior continuidade são representados na sua maioria por sequências do grupo azul enquanto que a região de sequências de menor continuidade apresenta-se enriquecido com sequências dos grupos vermelho e laranja, mostrando que esses dois grupos apresentam sequências 3'flanqueadoras mais similares (Figura 33).

As projeções das sequências flanqueadoras mostram que algumas famílias apresentam grande variabilidade enquanto outras, pouca diversidade, e que esta variabilidade não está associada a diversidade da sequência codificadora. Os resultados de expressão obtidos por PCR-tempo real de TcS mostram que é possível fazer algumas previsões sobre a expressão dos genes usando esses dados de variabilidade na região flanquedora 3' (Figura 33).

5.3. Mecanismos evolutivos

5.3.1. Indels e mudança da fase de leitura

A família MASP apresentou diversidade muito maior que as outras famílias além da falta de correspondência entre as distâncias dos pares de sequências de DNA e proteínas. Parte dessa alta diversidade encontrada nas proteínas e falta de correspondência pode ter surgido por mutações *indel* alterando a fase de leitura do gene. Para verificar se mutações *indel* contribuem para a geração de diversidade na família MASP foi realizado o alinhamento par-a-par dos genes MASP completos (genes que codificam ambas as extremidades conservadas, N- e C-terminal) e calculado número de *gaps* em fase e fora de fase de leitura de cada alinhamento. A distância protéica também foi calculada e comparada com o número de *gaps* em fase e fora de fase encontrados no alinhamento usando a sequências de DNA. Usando o coeficiente de correlação de Spearman foi verificada a existência de associação significativa entre essas métricas (Figura 35).

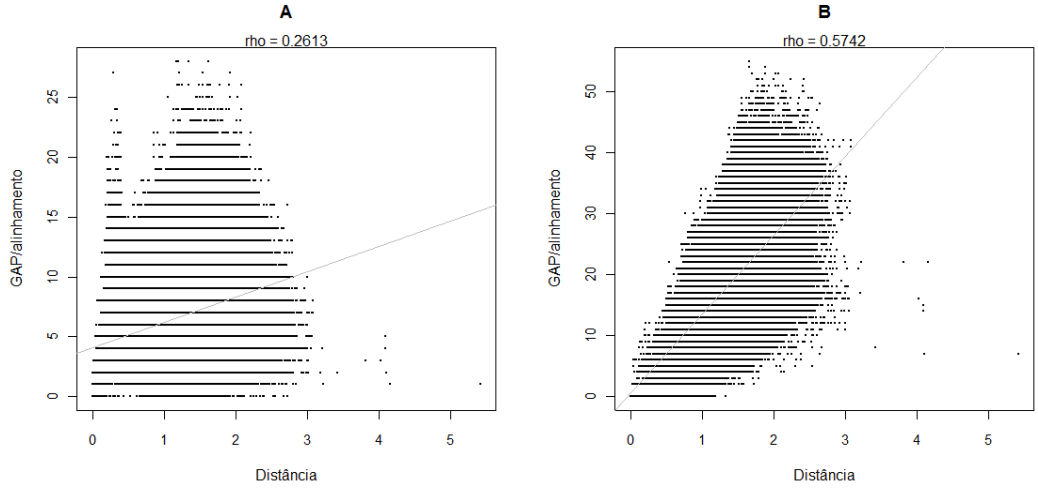


Figura 35. Associação entre número de gaps em fase e fora de fase com distância protéica na família MASP. A – distância protéica versus número de *gaps* em fase, correlação 0.2613; B – distância protéica versus número de *gaps* fora de fase, correlação 0.5742. Todas as associações foram significativa com valor $p \leq 0.025$ usando a correção de Bonferroni. A linha que melhor se ajusta aos dados é mostrada em cinza.

O número de *indels* em fase comparando com a distância protéica mostraram uma relação fraca direta e significativa (*rho* = 0,2613, $p << 0,025$). Quando se compara a relação do número de *gaps* fora de fase com a distância protéica surge uma relação direta forte e significativa (*rho* = 0,5742, $p << 0,025$).

Ambas as comparações dos dois tipos de *gaps* contra a distância protéica mostram que quanto maior a distância protéica, menor será o número de *gaps* em fase e maior aqueles fora de fase.

Outra abordagem para identificar possíveis substituições nucleotídicas que poderiam causar grandes mudanças na sequência de proteínas MASP foi a busca de pares de sequências que não apresentavam correspondências entre a distância nucleotídica e protéica. A busca foi realizada consultando as matrizes de distância par-a-par de DNA e proteína, por distância nucleotídica ≤ 0.4 e protéica ≥ 0.9.

A busca de pares de sequências que apresentavam diferença expressiva entre a distância nucleotídica e protéica resultou em uma lista de 35 casos, envolvendo 43 genes MASP. Em seguida, se verificou se os 43 genes encontrados apresentavam-se em uma região contendo no mínimo três sequências *reads* para a montagem do consenso ao longo de toda a extensão do gene. Treze genes satisfizerem este critério, totalizando sete casos (pares) em que a distância nucleotídica ≤ 0.4 e protéica ≥ 0.9.

Os alinhamentos das sequência de DNA e proteínas nos sete casos mostram que existem trechos com alta identidade entre as sequências de DNA e baixa identidade entre as sequências de proteínas (Figura 36). O alinhamento de proteínas ainda foi avaliado usando a matriz BLOSUM62 para verificar se existia similaridade nos trechos com menor identidade, mas a região de baixa identidade na sequência de DNA também não apresenta similaridade nas sequências correspondentes de proteínas. A falta de similaridade em alguns trechos da sequência de proteínas poderia ser provocada pela alteração da fase de leitura de um dos genes. Algumas dessas regiões são flanqueadas por *gaps* nas sequências de DNA o que poderia ser um indício de que ocorreu a mudança da fase de leitura.

A análise de um desses casos mostra a identidade do alinhamento das sequências de DNA e proteínas dos genes Tc00.1047053506501.110 e Tc00.1047053506965.100. As sequências de DNA apresentam trechos de alta identidade (representado por barras vermelhas na Figura 36), mas as sequências correspondentes a esses trechos nas proteínas apresentam baixa identidade. Esses trechos de pouca identidade nas sequências de proteína acontecem depois de mutações *indel* (Figura 36).

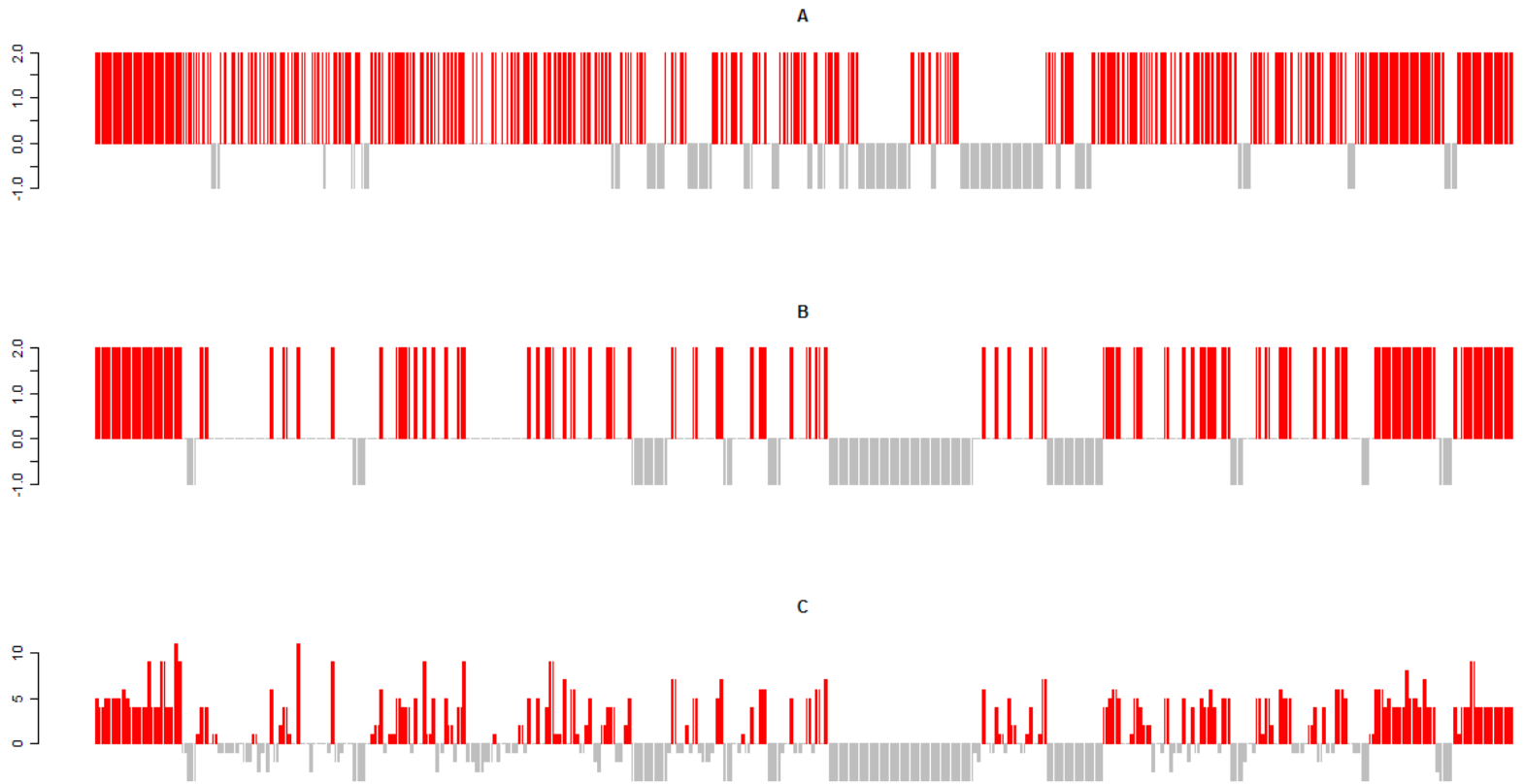


Figura 36. Representação do padrão de identidade das sequências de DNA e proteínas dos genes Tc00.1047053506501.110 e Tc00.1047053506965.100. Os painéis A e B estão representando a identidade entre as sequências de DNA e proteínas, respectivamente. As barras com valor 2 indicam *match* no alinhamento (vermelho), 0 indicam *mismatch* e -1 *gap* (cinza). O gráfico C foi produzido usando a matriz BLOSUM62 para avaliar a existência de similaridade na sequência depois da alteração da fase de leitura. Valores maiores que 1 estão indicados em vermelho e valores menores que 0, em preto.

5.3.2. Número de possíveis códons de parada

O surgimento de *indels* que alteram a fase de leitura do gene pode levar à formação de códons de parada ao longo do mesmo, interrompendo a tradução de uma sequência polipeptídica. O surgimento de mutações *indel* que alteram a fase de leitura seria vantajoso como mecanismo para geração de variabilidade para genes que apresentam menor frequência a gerar códons de parada devido a essas alterações. A frequência de uma família gênica gerar códons de parada pode ser verificada computando quantos códons de parada são gerados quando acontecem mudanças na fase de leitura logo no primeiro códon. É ainda interessante saber as posições desses códons de parada nas sequências para verificar se existem posições que seriam mais propensas a gerar diversidade por esse possível mecanismo de mutações *indel*. Foi computado o número e a posição de códons de paradas possíveis de serem formados alterando a fase de leitura dos genes das famílias com grande número de sequências e alta diversidade (GP63, TcS, RHS, TcMUC e MASP). Esses valores foram comparados com o número médio de códons de parada gerados na alteração da fase de leitura de sequências aleatórias criadas usando *codon usage* específicos calculados para cada família (ver material e métodos) (Figura 37).

Os valores médios de códons de paradas das famílias GP63, RHS, e TcS (22; 41,23 e 41,38, respectivamente), foram maiores que os valores médios encontrados nas sequências aleatórias (15,57; 29,09 e 21,78, respectivamente). Os valores médios de códons de parada encontrados para as famílias TcMUC e MASP foram aproximadamente a metade (7,2 e 10,13, respectivamente) dos valores encontrados em sequências aleatórias (15,52 e 19,81, respectivamente).

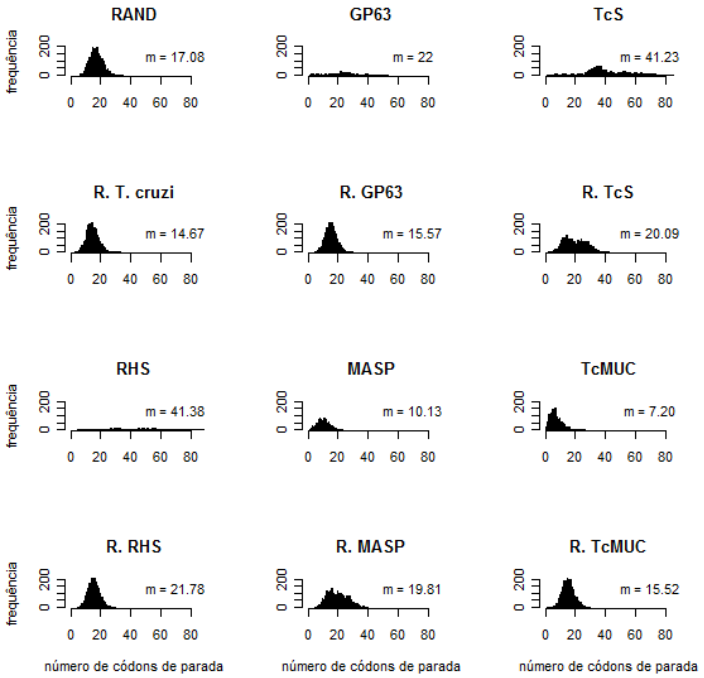


Figura 37. Frequência dos códons de parada produzidos por mudança de fase de leitura provocada no primeiro códon. RAND (R) representa 1.000 sequências codificadoras aleatórias produzidas usando *codon usage* calculados para cada família gênica.

A posição dos códons de parada em cada família gênica mostra que existe uma variação na frequência ao longo das sequências (Figura 38), enquanto que as sequências aleatórias apresentam frequência mais homogênea ao longo da sequência. Mesmo existindo essa variação, as famílias GP63, RHS e TcS não apresentam uma região com menor frequência de códons de parada comparado com os conjuntos de sequências aleatórias.

As famílias MASP e TcMUC, entretanto, apresentam as extremidades 5' e 3' com uma frequência menor de códons de parada em relação ao conjunto de sequências aleatórias. A extremidade 3' de TcMUC apresenta um longo trecho com frequência reduzida de possíveis códons de parada, correspondente a 30% do tamanho da sequência a partir da extremidade 5'.

Os genes MASP e TcMUC apresentam uma frequência menor à formação de códons de parada no gene como resultado de alterações na fase de leitura. Essa característica está presente principalmente nas extremidades 5' e 3'.

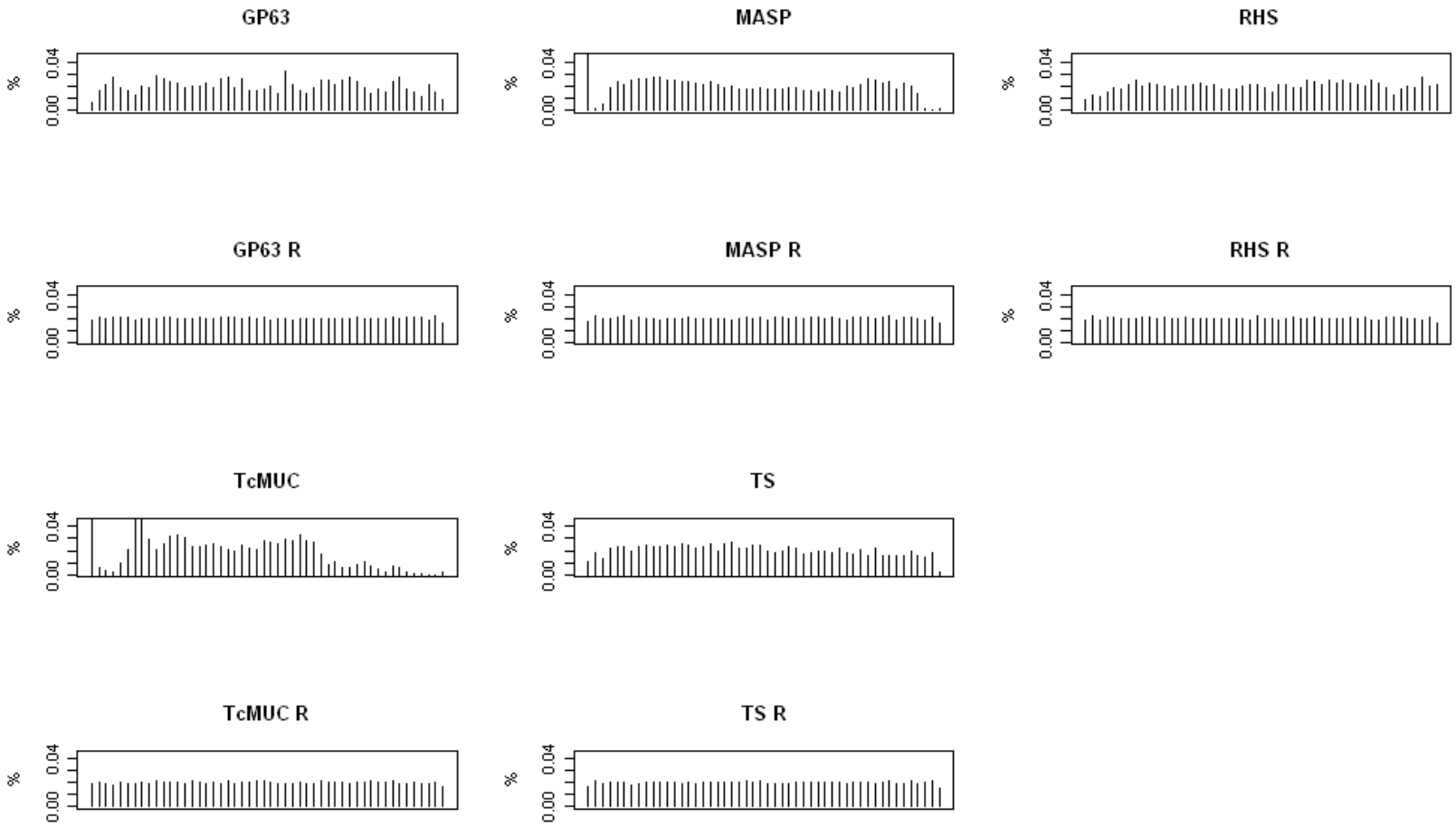


Figura 38. Frequência dos códons de parada produzidos por mudança de fase de leitura ao longo dos genes. A posição de cada códon de parada encontrado em cada gene foi dividida pelo comprimento do gene para se obter um posicionamento relativo. R representa 1.000 sequências codificadoras aleatórias produzidas usando *codon usage* de cada família.

Os resultados indicam que algumas mutações *indel* que modificaram a fase de leitura dos genes MASP podem ser responsáveis pela falta de correspondência entre distâncias medidas usando sequências de DNA e proteínas. A ocorrência de mutações alterando a fase de leitura do gene seriam favorecidas pela baixa frequência de códons de parada gerados após a mudança da fase de leitura, evitando a produção de proteínas truncadas.

5.3.3. Troca de fragmentos entre genes em *T. cruzi*

Diversos genes quiméricos formados por partes de sequências de diferentes famílias gênicas (SAP, TcMUC, TcS e MASP) foram identificados no projeto genoma do *T. cruzi*. Mais recentemente identificamos também um cDNA quimérico idêntico a um destes genes contendo parte da sequência da família TcMUC e parte da família MASP (Bartholomeu *et al.*, 2009), sugerindo que proteínas quiméricas podem ser expressas pelo parasito. O mapeamento de *reads* individuais gerados no projeto genoma neste gene mostrou que esta sequência quimérica não é um artefato da montagem do genoma ou da construção da biblioteca de cDNA, uma vez que *reads* individuais mapeam na região de junção entre as partes originadas de cada uma destas famílias gênicas. Estes dados sugerem que, possivelmente através de mecanismos de recombinação, possa ocorrer troca de fragmentos entre genes de diferentes famílias e dentro de uma mesma família e assim contribuir para um aumento da diversidade.

A pesquisa por esses fragmentos compartilhados envolveu uma busca por fragmentos compartilhados entre membros da família MASP, bem como entre membros de diferentes famílias.

5.3.4. Troca de fragmentos entre genes da família MASP

A fim de identificarmos motivos conservados entre membros da família MASP, nós utilizamos o programa MEME versão 4.5.0 (Bailey e Elkan, 1994). Este programa busca pequenos motivos conservados em um banco de dados e ordena estes motivos por ordem de significância baseado no número de sequências que contém o motivo, no tamanho e no grau de conservação do mesmo. Os motivos podem ser idênticos ou apresentarem degenerações, sendo representados como sequências consenso. Usou-se como entrada para o programa MEME as sequências nucleotídicas dos genes completos de MASP excluindo as sequências que

codificam para as regiões N- e C-terminal conservadas, de forma que os motivos encontrados sejam derivados da região central hipervariável. Um total de 30 motivos foram pesquisados. Alguns desses motivos são bem conservados e apresentam comprimento variando entre 50 e 21 bp, enquanto outros são menos conservados, mas apresentam menor variação no comprimento. Para cada motivo consenso encontrado foi produzido uma árvore mostrando a similaridade e relação dos motivos. As sequências apresentando os motivos nas árvores foram identificadas usando a classificação dos grupos MDS, mostrada na segunda parte deste trabalho de tese (Figura 18 e Figura 19).

O conjunto de fragmentos MEME pode ser dividido em três grupos. O primeiro grupo, formado por cinco MEMEs, são fragmentos que estão presentes somente em genes que pertencem ao mesmo grupo do MDS. Alguns desses motivos são conservados e outros apresentam mais variações nas suas sequências. Um representante desde tipo de MEME é mostrado na Figura 39. Esse motivo é conservado e está presente nas sequências do grupo azul. A pouca variação nesse motivo gera ramos curtos na árvore, e somente algumas sequências apresentam ramos mais longos. A separação dos grupos na árvore do MEME20 é geralmente discreta devido às sequências serem muito similares. Este grupo MEME não é informativo para a detecção de troca de fragmentos entre membros da família.

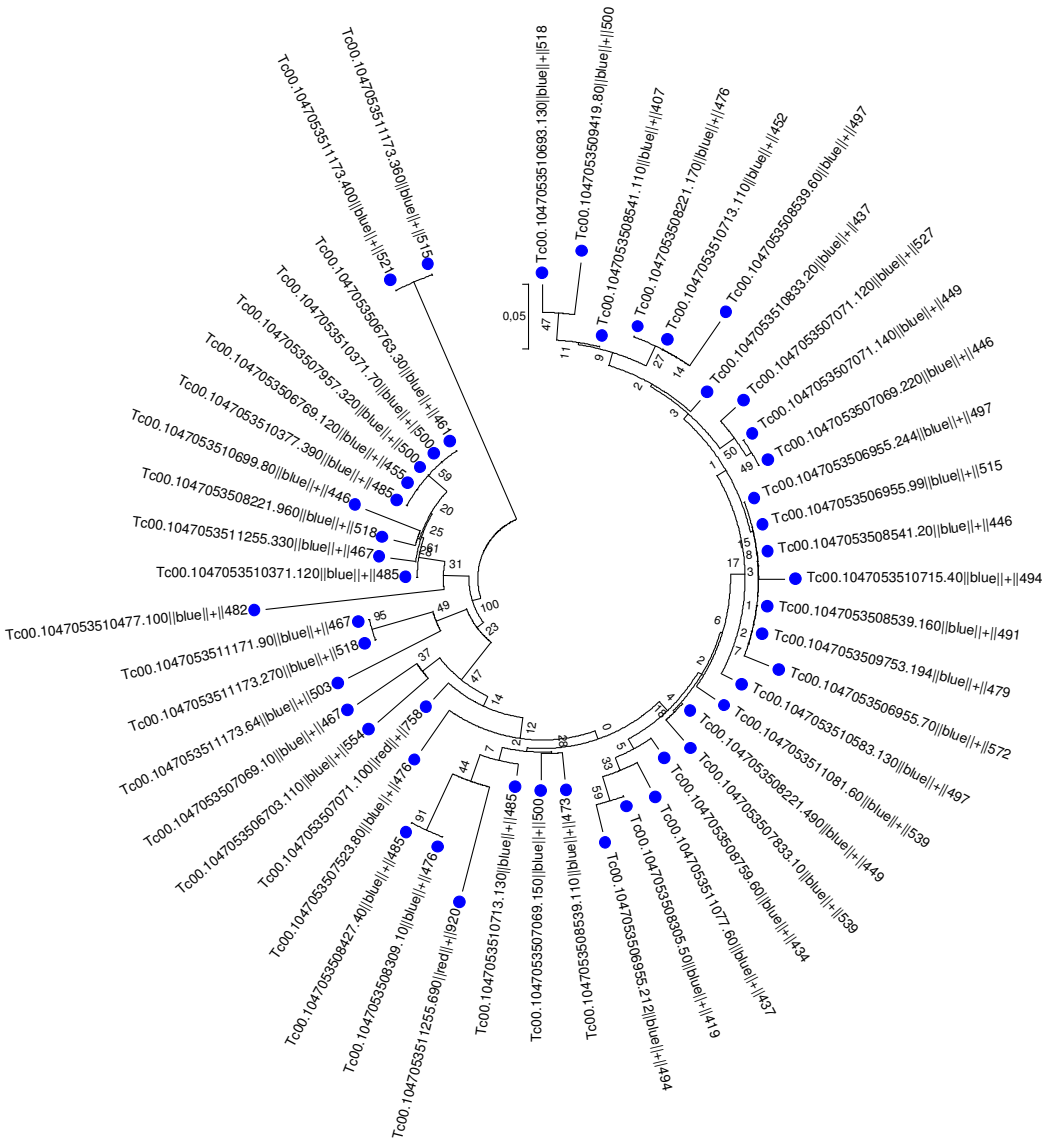


Figura 39. Árvore mostrando a relação entre os motivos dos genes contendo o MEME 6 (50 nt). Todos os genes que possuem o MEME 6 pertencem ao grupo azul.

O segundo grupo é formado por três MEMEs que apresentaram grande variação na sequência, indicado pelo comprimento dos ramos nas árvores, apresentando motivos com sequências mais divergentes (Figura 40). O MEME20 é pouco conservado e está presente nas sequências dos grupos preto, azul, vermelho, rosa, laranja. Os ramos são muito longos devido a divergências entre as sequências. Este grupo MEME também não é informativo para a detecção de troca de fragmentos entre membros da família.

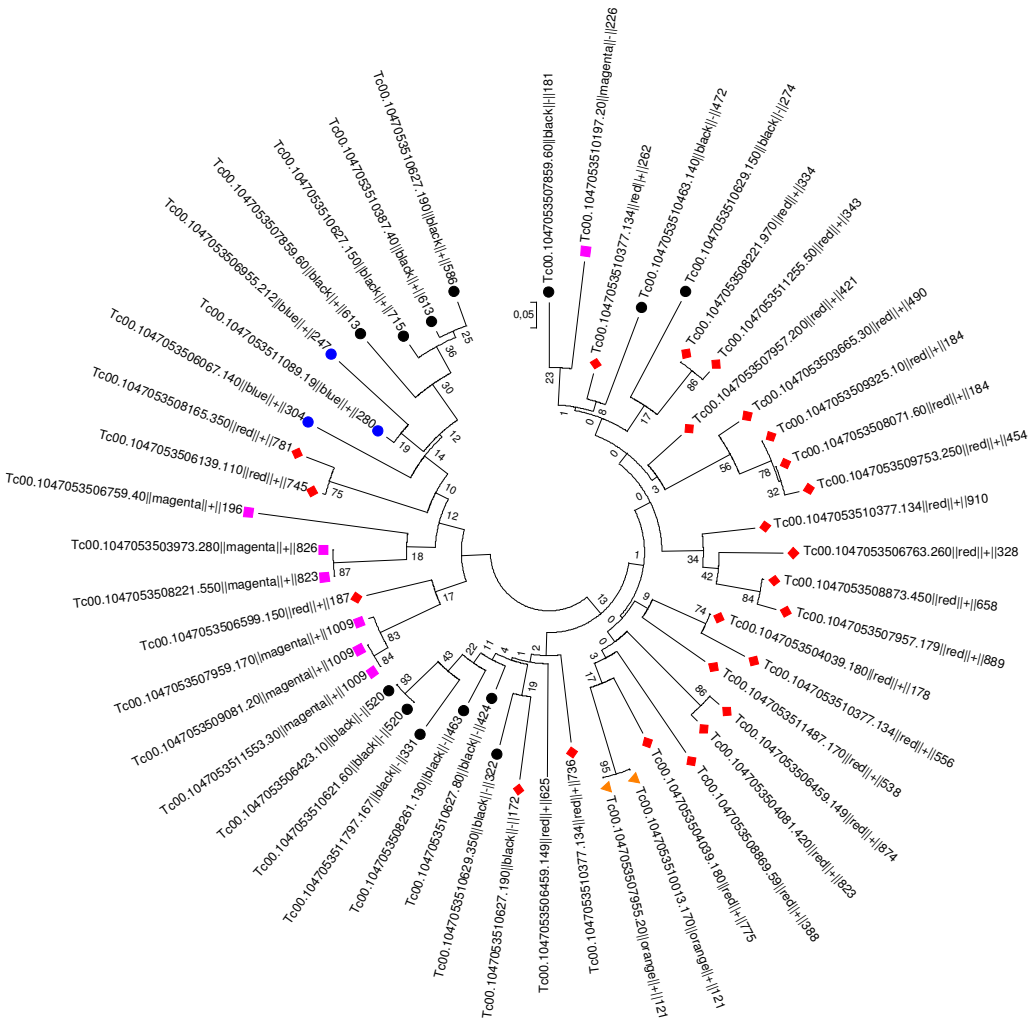


Figura 40. Árvore mostrando a relação entre os motivos do MEME 20 (50 nt). Os motivos foram encontrados em genes de diferentes grupos, entretanto se apresentaram muito divergentes.

Os outros 22 fragmentos formam o terceiro grupo de MEMEs caracterizados por apresentarem ramos compostos de motivos similares pertencentes a diferentes grupos. Alguns fragmentos que pertencem a este grupo de MEME formam, pelo menos, um ramo com pouca variação e com sequências compostas por mais de um grupo do MDS.

Foram selecionados para uma análise mais detalhada, motivos pertencentes ao terceiro grupo de MEME, que são muito similares e pertencentes a grupos MDS diferentes. O MEME4 é extremamente conservado e está presente nos grupos MDS vermelho e rosa. O motivo compartilhado por essas sequências está presente em dentro de uma região conservada e próximo a extremidade 5' do gene (Figura 41).

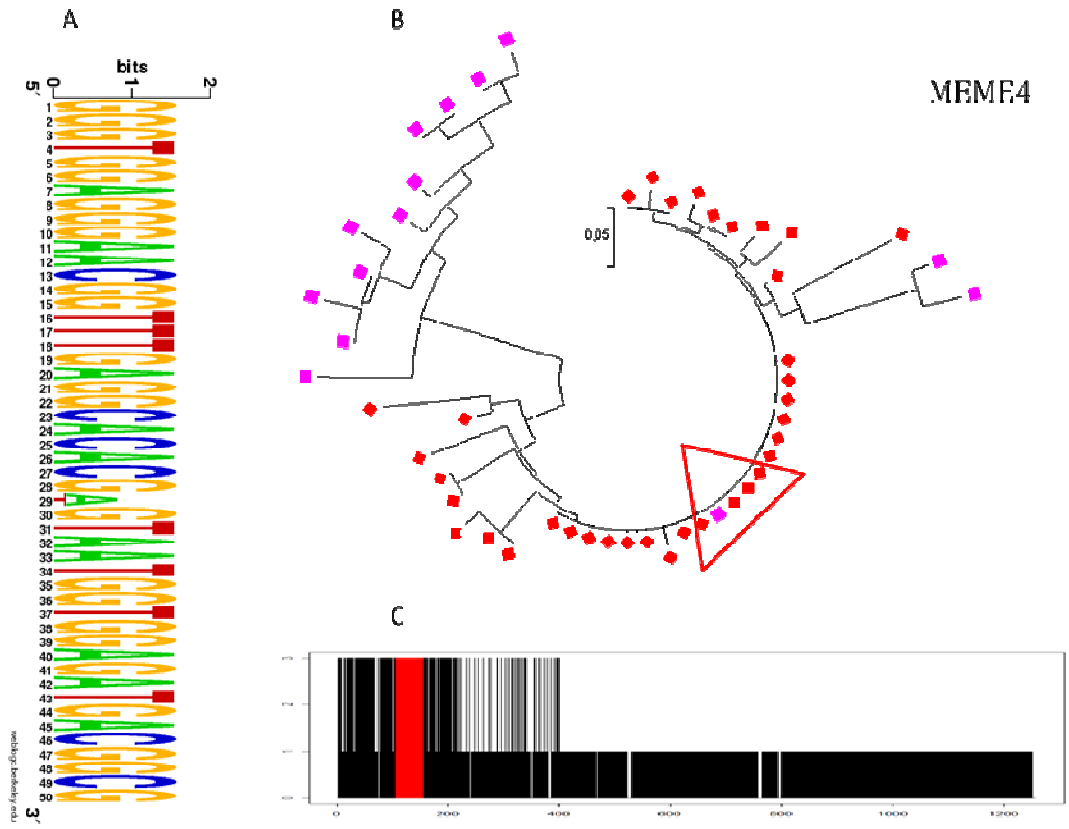


Figura 41. Árvore da relação de similaridade entre os motivos que formam o MEME 4. (A) *Logo* representando a frequência de cada nucleotídeo que compõe os motivos delimitados pelo triângulo vermelho da árvore. (B) Árvore representando a relação entre os motivos e a similaridade pelo comprimento dos ramos. (C) Alinhamento dos genes de MASP contendo os motivos delimitados pelo triângulo vermelho da árvore. As barras no alinhamento representam *match* = 3; *mismatch* = 1 e *gap* = 0. O trecho em vermelho indica a posição do motivo no alinhamento.

Em dois casos dos 22 MEMEs do terceiro grupo identificados, os motivos estavam invertidos em alguns genes (Figura 42). Os genes contendo os dois MEMEs que apresentam motivos invertidos pertencem ao grupo preto. O MEME22 está presente em sequências derivadas de todos os grupos MDS. Algumas sequências dos grupos rosa e vermelho apresentam o motivo na extremidade 3' do gene, enquanto que uma das sequências do grupo preto apresenta o motivo invertido e localizado na região interna.

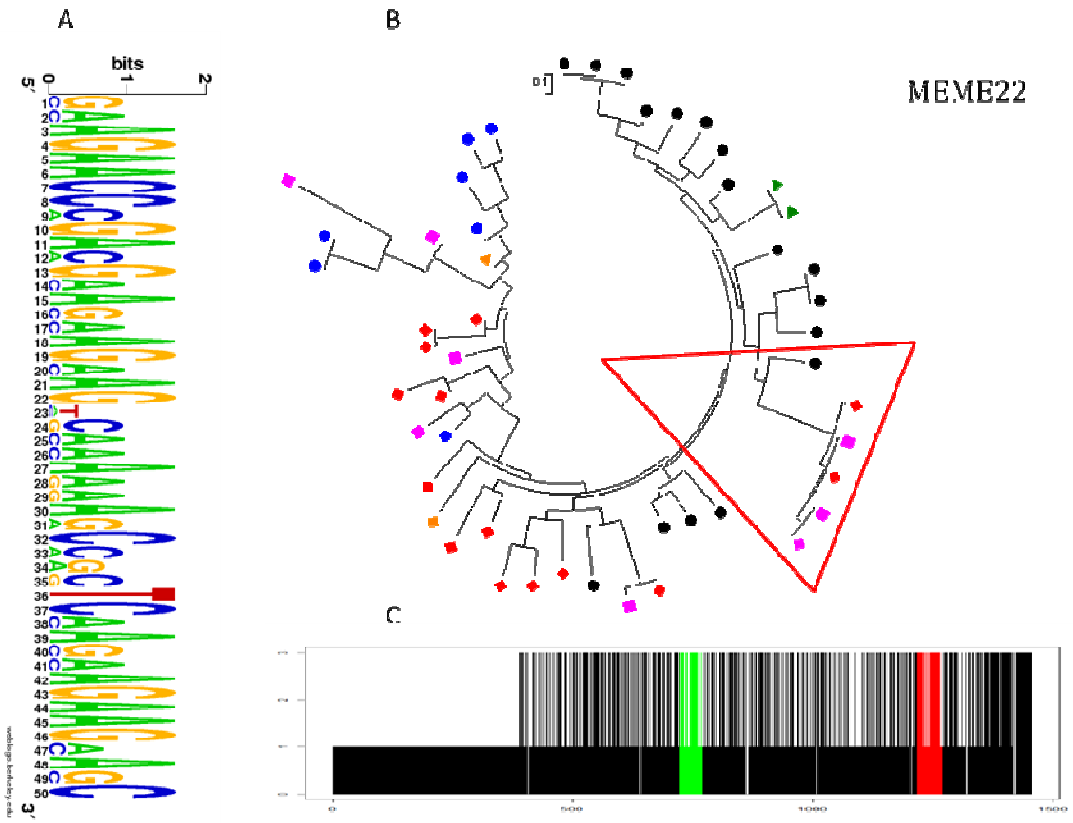


Figura 42. Árvore da relação de similaridade entre os motivos que formam o MEME 22. (A) *Logo* representando a frequência de cada nucleotídeo que compõe os motivos delimitados pelo triângulo vermelho da árvore. (B) Árvore representando a relação entre os motivos e a similaridade pelo comprimento dos ramos. (C) Alinhamento dos genes de MASP contendo os motivos delimitados pelo triângulo vermelho da árvore. As barras no alinhamento representam *match* = 3; *mismatch* = 1 e *gap* = 0, trechos em vermelho (ordem direta) e verde (ordem inversa) indicam as posições dos fragmentos no alinhamento.

Os genes de MASP que contêm os 22 MEMEs do terceiro grupo foram submetidos ao alinhamento múltiplo, e a similaridade ao longo da sequência calculada e verificado sua posição dentro dos genes (Figura 43). Estes 22 motivos quase sempre se apresentaram próximos às extremidades 5’ e 3’ conservadas do gene (19 MEMEs, 86% dos casos). Há uma frequência maior desses motivos associados à região 3’ (12 MEMEs) do que à região 5’ (7 MEMEs).

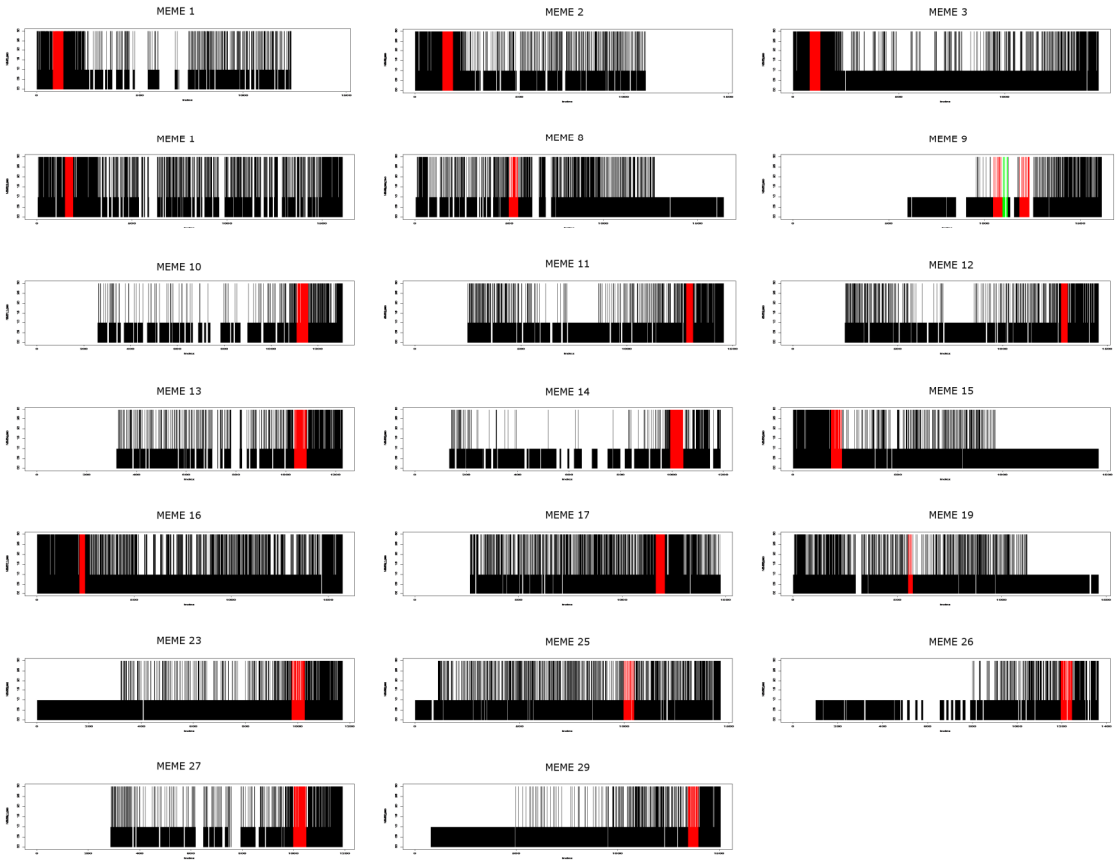


Figura 43. Alinhamento dos genes de MASP contendo os motivos MEME. As barras no alinhamento representam *match* = 3; *mismatch* = 1 e *gap* = 0. O trecho em vermelho indica a posição do motivo no alinhamento.

5.3.5. Troca de fragmentos entre famílias gênicas

Fragmentos compartilhados entre os genes de diferentes famílias foram identificados usando o programa blastn. Após a busca, os resultados foram filtrados para fragmentos com comprimento mínimo de 100 nt e 90% de identidade entre os fragmentos, sendo posteriormente separados em faixas de comprimentos (

Tabela 7). Foram encontrados vários fragmentos com diferentes comprimentos compartilhados entre as diferentes famílias gênicas que codificam proteínas de superfície. As famílias que apresentaram fragmentos compartilhados envolvendo de 100 a 200 nucleotídeos e 90% de identidade foram DGF-1, GP63, MASP, SAP, TcMUC e TcS. Todos estes fragmentos encontrados envolviam as famílias MASP e/ou TcS. Na grande maioria dos casos, os fragmentos compartilhados se localizavam na extremidade 3' dos genes.

O número de famílias compartilhando fragmentos mostra que esse pode ser um mecanismo muito importante para geração de variabilidade, principalmente para a família MASP (Tabela 7). O número de genes encontrados em cada busca também mostra uma frequência alta de genes da família MASP. Esta família apresentou 202 genes compartilhando fragmentos com 6 genes TcMUC, 10 genes apresentaram similaridade com 58 genes da família TcS e um gene apresentado similaridade com 11 genes da família SAP (Tabela 7).

A segunda família que mais apresentou fragmentos compartilhados com 100 nucleotídeos de comprimento foi a família TcS. Foram encontrados seis genes da família TcS e seis genes da família GP63 compartilhando fragmentos; também encontramos 29 genes da família TcS com fragmentos similares com 1 gene da família DGF-1.

Os fragmentos compartilhados envolvendo os comprimentos de 200 até 300 (dois fragmentos) e 300 até 400 (11 fragmentos) foram apenas nas famílias MASP e TcMUC. Os fragmentos com comprimento de 200 até 300 nucleotídeos foram encontrados no interior dos genes, enquanto que cinco fragmentos encontrados entre 300 até 400 nucleotídeos foram encontrados na extremidade 5' e outros seis na extremidade 3'. Foram encontrados 15 fragmentos compartilhados entre MASP e TcMUC com comprimento maiores que 500 nucleotídeos. Todos os fragmentos estão na região 3' dos genes. Foram encontrados 15 diferentes genes MASP e três genes TcMUC. Somente oito fragmentos foram encontrados envolvendo as famílias MASP e TcS com comprimento maior que 500 nucleotídeos. Esse fragmentos foram localizados nas extremidades 3' com média de 1260 nt. Foram encontrados oito diferentes genes MASP e 1 gene TcS.

Portanto provavelmente ocorreram trocas de fragmentos de DNA entre algumas famílias de proteínas de superfície. Essas trocas tiveram a participação principalmente da família MASP, que apresentou tanto fragmentos compartilhados entre diferentes membros da própria família (envolvendo principalmente regiões conservadas) quanto com genes de outras famílias. As trocas de fragmentos aconteceram preferencialmente na região 3' dos genes.

Tabela 7. Famílias que apresentaram fragmentos compartilhados com no mínimo de 100 nucleotídeos e 90% de identidade. Número de genes envolvidos está indicado entre parênteses.

Fragmentos similares ≥ 100 e < 200nt		**Hits**	**5'**	**Interior**	**3'**
MASP (202)	TcMUC (6)	205	4	0	201
MASP (10)	TcS (58)	112	0	0	112
MASP (1)	SAP (11)	11	0	0	11
TcS (6)	GP63 (6)	11	0	0	11
TcS (29)	DGF-1 (1)	29	0	0	0
Fragmentos similares ≥ 200 e < 300nt		**Hits**	**5'**	**Interior**	**3'**
MASP (2)	TcMUC(1)	2	0	2	0
Fragmentos similares ≥ 300 e < 400nt		**Hits**	**5'**	**Interior**	**3'**
MASP (6)	TcMUC(4)	11	5	0	6
Fragmentos similares ≥500nt		**Hits**	**5'**	**Interior**	**3'**
MASP (15)	TcMUC (3)	15	0	0	15
MASP (8)	TcS (1)	8	5	0	6

6. Discussão

Neste trabalho nos propusemos a caracterizar a diversidade das grandes famílias gênicas de *T. cruzi*, que em sua maioria codificam proteínas de superfície extremamente polimórficas. A diversidade encontrada nessas famílias pode estar relacionada com a habilidade do parasita em infectar e replicar-se em uma grande variedade de células nucleadas e/ou ter surgido devido a uma pressão seletiva imposta pelo sistema imune do hospedeiro. Estas duas hipóteses não são mutuamente exclusivas.

Os mecanismos geradores de diversidade das grandes famílias gênicas podem envolver eventos de substituição de bases, inserções ou deleções (*indels*), sendo que estas últimas poderiam alterar a fase de leitura dos genes e provocar uma grande alteração das sequências de proteínas. Além disso, a recombinação não homóloga envolvendo a troca de fragmentos entre os diferentes genes seria outra forma de aumentar a diversidade das proteínas dessas grandes famílias. Os organismos *Anaplasma marginale*, *Borrelia hermsii* e *Trypanosoma brucei*, utilizam de recombinação entre genes e pseudogenes para gerar novas variantes de proteínas de superfície envolvidas no processo de variação antigênica, permitindo a evasão do sistema imune (Futse *et al.*, 2005; Palmer e Brayton, 2007).

A diversidade das grandes famílias de proteínas de superfície foi estimada através das medidas de diversidade média tanto para as sequências de DNA quanto proteínas. Outra abordagem utilizada para avaliar a diversidade das famílias foi o uso de projeções espaciais (MDS), representando as distâncias entre as sequências de DNA, permitindo uma visualização de toda a variabilidade encontrada nas famílias.

Os gráficos MDS das sequências das famílias permitem também a visualização dos grupos formados. Usando o método *kmeans* foi realizada separação das famílias em grupos, que foram identificados e comparados com classificações prévias das famílias.

Os *indels,* que poderiam alterar a fase de leitura dos genes, foram investigados buscando pares de genes que apresentam discordância entre a distância genética e protéica. Outros indícios dessas alterações foram verificados quantificando e localizando as regiões dos possíveis códons de parada gerados após alteração da fase de leitura dos genes.

O possível mecanismo de troca de fragmentos entre os diferentes genes foi analisado através da busca e posicionamento dos fragmentos que são compartilhados entre diferentes

grupos da família MASP, e fragmentos compartilhados entre genes de diferentes famílias gênicas.

A discussão dos resultados obtidos dessas análises será dividida três partes, que são:

1. Caracterização da família TcS - Identificação de novos grupos na família trans-sialidase/sialidase-like (TcS) de T. cruzi
2. Diversidade das grandes famílias gênicas de T. cruzi
3. Mecanismos evolutivos possivelmente envolvidos na geração de variabilidade

1. Caracterização da família TcS - Identificação de novos grupos na família trans-sialidase/sialidase-like (TcS) de *T. cruzi*

TcS é a maior família multigênica de *T. cruzi*, correspondendo a 6,3% do número total de genes da cepa CL Brener (El-Sayed *et al.*, 2005b). A família TcS é classificada até o momento em quatro grupos, que apresentam várias funções e alta variabilidade de sequência. O grupo I (TcS grupo I) compreende as proteínas que apresentam atividade trans-sialidase (TS), TcSgrupoII alberga proteínas associadas a mecanismos de adesão e invasão celular, TcS grupo III, proteínas reguladoras do complemento e o grupo TcS grupo IV possui proteínas de função desconhecida, mas apresentam os motivos característico da família TcS (Cross e Takle, 1993; Schenkman *et al.*, 1994; Frasch, 2000). A despeito de extensos estudos desta família, após a publicação do projeto genoma de *T. cruzi*, nenhuma análise sobre a diversidade desta família foi realizada. Usando métodos de clusterização com todas as proteínas completas preditas no genoma de *T. cruzi*, nós identificamos quatro novos grupos. Os oito grupos identificados foram caracterizados baseado na presença de motivos característicos da família, localização cromossômica, padrão de expressão gênica e antigenicidade.

A separação e distâncias entre os grupos está de acordo com a similaridade e função descritas para a família TcS, mostrando todas as TS ativas dentro do mesmo grupo TcS (TcSgrupoI, azul). O grupo que apresenta as proteínas envolvidas na adesão e invasão celular forma o TcSgrupoII, verde escuro. O TcSgrupoIII contém as proteínas envolvidas na regulação do sistema de complemento (CRP - Complement Regulatory Protein), entre essas proteínas se encontram a FL-160 e outros membros caracterizados por Beucher e Norris (2008). Beucher e Norris identificaram genes CRP parálogos no genoma de *T. cruzi* usando a sequência com AAB49414 como referência. O grupo CRP foi dividindo no subgrupo HSG (*high similarity group*, com mais de 80% de identidade com AAB49414) e LSG (*low*

similarity group, identidade entre 54 e 61% com AAB49414). Os membros HSG e LSG (com duas exceções) estão presentes dento do TcSgrupoIII. As duas proteínas do subgrupo LSG, que não estão dentro do TcSgrupoIII, são sequências mais divergentes e se agruparam na projeção MDS dentro do TcSgrupoVII (laranja). Essas duas proteínas apresentam-se tanto próximas do grupo TcSgrupoVII (laranja) quanto do grupo TcSgrupoIII (azul claro). Os novos grupos identificados, TcS grupos V, VI, VII e VIII , apresentam-se próximos dos grupos TcSgrupoIII (azul claro) e TcSgrupoIV (rosa).

A diversidade de TcS pode estar relacionada com a produção de uma resposta imune atrasada devido a expressão coordenada de diferentes membros. A presença de diferentes TS (apresentando variações em torno do sítio catalítico) na superfície do parasita seria necessária para prevenir uma resposta específica de inativação da enzima, formando um mecanismo de evasão do sistema imune (Ratier *et al.*, 2008). Anticorpos contra TS são detectados após um longo período de infecção (Leguizamón *et al.*, 1994).

As proteínas da família TcS apresentam motivos, alguns presentes em todos os membros e outros presentes em somente algumas sequências (Cross e Takle, 1993; Schenkman *et al.*, 1994; Frasch, 2000). O motivo FRIP é o mais próximo da extremidade N-terminal está envolvido na ligação ao grupo carboxilato do ácido siálico (Gaskell *et al.*, 1995), entretanto, proteínas enzimaticamente inativas ainda preservam essa característica (Cremona *et al.*, 1999; Todeschini *et al.*, 2004). Um trabalho recente mostrou que alguns membros de trans-sialidase inativos têm a capacidade de se ligar a carboidratos de glicoconjugados somente se eles estão ligados a ácido siálico (Oppezzo *et al.*, 2011). Os autores sugerem que essas TcS inativas com capacidade de ligação a ácido siálico poderiam atuar como âncoras para fazer a ligação na superfície das células do hospedeiro e facilitar a ação das TSs. Essa conservação e alta frequência do motivo FRIP nos TcS grupos I, III e IV pode indicar que estes outros grupos também apresentam esta função.

O motivo Asp-box não foi encontrado no TcSgrupoIII, o que está de acordo com trabalhos anteriores (Beucher e Norris, 2008). Além do TcSgrupoI, membros do TcSgrupoIV apresentam tanto o motivo FRIP quanto o motivo Asp-box em mais de 68% dos membros, podendo esse grupo também apresentar propriedades de ligação a carboidratos.

O motivo VTVxNVxLYNR ou versões degeneradas do mesmo foi encontrado em todas as sequências completas da família TcS confirmando estudos prévios que sugeriram que este motivo caracteriza a família (Schenkman *et al.*, 1994; Frasch, 2000). Recentes trabalhos mostram uma variante desse motivo (VTVxNVFLYNR – chamado de FLY) pode atuar como

um fator de virulência. (Magdesian *et al.*, 2007; Tonelli, Torrecilhas *et al.*, 2011). Camundongos inoculados com o peptídeo FLY sintético são mais susceptíveis à infecção por *T. cruzi* e apresentam aumento na parasitemia e mortalidade (Scherf *et al.*, 2008; Tonelli, Torrecilhas *et al.*, 2011). O motivo FLY ainda apresentou ligação às células do endotélio cardíaco sugerindo que esse motivo pode contribuir para o tropismo celular. Uma variante do motivo FLY (ATVANVFLYNRPLN) é encontrada em 23 membros do TcSgrupoIV, podendo indicar que esse grupo também esteja relacionado com adesão/invasão celular.

As repetições foram encontradas frequentemente em TcSgrupoI (azul) e TcSgrupoIV (rosa) como descrito por outros trabalhos (Cross e Takle, 1993; Schenkman *et al.*, 1994). Além das repetições DSSAH(S/G)TPSTP(A/V) e EPKSA, encontradas nos grupos TcS I e IV respectivamente, foram encontradas novas repetições nesses grupos. Repetições menores e menos frequentes antes não descritas foram encontradas nos outros grupos.

Somente alguns membros da família TcS apresentam atividade sialidase/trans-sialidase (Cross e Takle, 1993; Schenkman *et al.*, 1994), tendo sido definido os aminoácidos importantes para exercer a atividade enzimática (Schenkman *et al.*, 1994; Cremona *et al.*, 1999). A identificação das possíveis TcS ativas foi feita através da busca destes resíduos críticos (Montagna *et al.*, 2002). Das 17 posições descritas como importantes para a atividade enzimática, as posições Tyr342 e Pro283 são consideradas posições mais críticas por fazer parte diretamente do sítio catalítico (Schenkman *et al.*, 1994; Cremona *et al.*, 1999). As TcSgrupoI inativas possuem o aminoácido histidina na posição 342 (Schenkman *et al.*, 1994; Cremona *et al.*, 1999). *T. brucei* possui trans-sialidases ativa expressa na forma presente no inseto (Engstler *et al.*, 1992). Embora nenhuma trans-sialidases tenha sido identificada em *T. rangeli*, proteínas similares a sialidase foram identificadas e apresentaram similaridade com os grupos TcS I, II e III, e vários desses membros foram identificados como sendo expressos na fase epimastigota e tripomastigota (Buschiazzo *et al.*, 1997; Grisard *et al.*, 2010).

Estudos evolutivos sugerem um gene ancestral codificando trans-sialidases ativas expressas nas formas evolutivas presentes no inseto vetor (Briones *et al.*, 1995). Posteriormente vários eventos de duplicação e diversificação teriam dado origem a trans-sialidases expressas na forma encontrada no mamífero presente em *T. cruzi* (Briones *et al.*, 1995). Estas evidências juntamente com a posição central de TcSgrupoI na projeção MDS sugerem que a expansão de genes codificadores das trans-sialidases similares a TcSgrupoI e o acúmulo de mutações teriam originado outros grupos e funções. Na linhagem evolutiva que

originou o *T. rangeli*, a enzima sofreu modificações e deixou de apresentar atividade trans-sialidase, permanecendo somente a atividade sialidase.

A grande diversidade encontrada no TcSgrupoI pode estar relacionada com a geração de uma resposta ineficaz do sistema imune contra a atividade trans-sialidase, gerando somente uma resposta atrasada para inativação da enzima (Ratier *et al.*, 2008). Essa grande diversidade dentro do grupo que apresenta todas TcS ativas poderia gerar respostas do hospedeiro incapaz de inativar a enzima TS devido variação na sequência primária dessas enzimas. Além da diversidade encontrada em TcSgrupoI, a diversidade encontrada nos outros grupos poderia contribuir na produção de uma resposta atrasada (Ratier *et al.*, 2008).

O mapeamento dos genes TcS nos cromossomos mostra que não existe uma associação clara entre um dado grupo TcS e um cromossomo em particular. De maneira interessante, verificamos que TcSgrupoII estão localizados preferencialmente nas regiões subteloméricas enquanto que TcSgrupoV está localizado no interior dos cromossomos, afastados das regiões subteloméricas. Além desses grupos, TcS pseudogenes também se apresentam preferencialmente localizados nas extremidades dos cromossomos, assim como TcSgrupoII.

T. brucei e *Plasmodium falciparum* apresentam estratégias sofisticadas para evasão do sistema imune conhecida como variação antigênica, que permite ao parasita escapar do ataque do sistema imune do hospedeiro através da exposição e troca de proteínas variantes de superfície (Cano, 2001; Scherf *et al.*, 2001; Chiurillo *et al.*, 2002; Horn e Barry, 2005; Kim *et al.*, 2005; Scherf *et al.*, 2008). Essas proteínas localizam-se preferencialmente nas extremidades dos cromossomos, que são regiões favoráveis à expansão e geração de novas variantes (Scherf *et al.*, 2001; Chiurillo *et al.*, 2002; Kim *et al.*, 2005). Como foram encontrados vários TcS pseudogenes nas extremidades dos cromossomos, *T. cruzi* pode apresentar intensos processos de recombinação nessas regiões. *T. cruzi* não apresenta sistema de variação antigênica, mas co-expressa genes que codificam proteínas polimórficas de superfície, dentre elas, as TcSs (Atwood *et al.*, 2005). Os genes codificadores de proteínas do TcSgrupoII, que estão localizados preferencialmente nas extremidades dos cromossomos, podem estar também sofrendo ação de intensos rearranjos genéticos que levariam a formação de novas variantes. Essas novas proteínas variantes formadas por rearranjos genéticos nas extremidades dos cromossomos podem ser importantes porque vários membros desse grupo estão envolvidos na adesão/invasão de células do hospedeiro e *T. cruzi* apresenta habilidade de infectar uma grande variedade de tipos celulares.

O controle de expressão gênica em tripanosomatídeos é feito principalmente por mecanismos pós-transcricionais, e em *T. cruzi* nenhuma sequência promotora de genes codificadores de proteínas foi identificada. Acredita-se que os genes nestes organismos sejam transcritos de forma constitutiva. Um dos pontos chaves no controle de expressão gênica neste grupo envolve a modulação da estabilidade dos transcritos mediada por elementos regulatórios presentes na região 3'UTR (Teixeira *et al.*, 1994; Bartholomeu *et al.*, 2002). Embora não se tenha dados de mapeamento da região 3'UTR de todos os genes TcS analisados é possível fazer uma correlação entre genes que apresentam um dado padrão de expressão ao longo do ciclo de vida do parasito e similaridade da região 3' flanqueadora. O mapeamento de TcS conhecidas na projeção MDS das sequências 3'UTR permitiu observar que alguns genes que são expressos no mesmo estágio evolutivo apresentam regiões 3' flanqueadoras similares. A região 3' flanqueadora das TcS ativas, SAPA e TCNA, que são expressas na fase tripomastigota apresentam-se muito próximas na projeção MDS e de fato estas sequências apresentam 96% de identidade. Embora o gene TS-epi também seja uma cópia ativa de TcS, a sua expressão é restrita na forma epimastigota encontrada no vetor (Cross e Takle, 1993; Schenkman *et al.*, 1994). Em concordância, a região 3' flanqueadora do gene TS-epi apresenta-se mais distante daquelas dos genes TCNA e SAPA, possuindo em média 46,5% de identidade.

Os genes gp90, gp82 (expressos na forma tripomastigota metacíclica) e ASP-2 (expressa na forma amastigota) pertencem ao grupo TcSgrupoII, e também se apresentaram próximos na projeção da região 3' flanqueadora. A identidade média entre essas sequências é de 65%, em comparação com 44 e 45,3% das sequências 3' flanqueadora de TcTSA-1 e Tc85-11_SA85-1.1 respectivamente, que pertencem também a TcSgrupoII.

A projeção TcS 3' flanqueadora também mostra que Tc85-11_SA85-1.1 e TsTc13 apresentam sequências similares, apresentando 87% de identidade. Ambas as proteínas são expressas em tripomastigota, entretanto apresentam diferenças nas sequências de proteínas suficiente para serem classificadas em grupos diferentes, TcSgrupoII e TcSgrupoIV, respectivamente. Isso sugere que os mecanismos de controle da expressão desses dos genes podem envolver elementos similares na região 3'UTR.

As sequências 3' flanqueadora do TcSgrupoVIII foram as únicas que se clusterizaram no MDS, o que poderia sugerir que esses genes apresentam níveis similares de expressão nas diferentes fases do ciclo de vida do parasita. Entretanto, este padrão não é valido para todos os membros do grupo, como por exemplo os genes TcS24 e TcS25, cujas sequências 3'

flanqueadoras estão mais afastadas no MDS apresentando apenas 49% de identidade entre si. De fato, estes genes apresentaram padrões de expressão distintos na PCR em tempo real, com TcS24 mais expressos na forma tripomastigota e TcS25 na forma amastigota. As sequências 3' flanqueadora do TcSgrupoII formaram 3 regiões com sequências mais concentradas. Essa três regiões principais, podem indicar diferentes tipos de controle regulando a expressão pós-transcricional dos genes desse grupo.

Ambas os grupos TcS V e VI apresentaram sequências 3' flanqueadora distribuídos por uma região similar na projeção MDS. Esses dois grupos de proteínas também se apresentam próximos na projeção de proteínas. Essa proximidade das sequências de proteínas e mesma variação na distribuição das sequências 3' flanqueadora pode indicar que esses genes apresentam o mesmo padrão de expressão durante o desenvolvimento de *T. cruzi* e possivelmente funções similares.

Os resultados de antigenicidade mostram as repetições DSSAH(S/G)TPSTP(A/V) e EPKSA muito reativas, como descrito em outros trabalhos (Cross e Takle, 1993; Schenkman *et al.*, 1994; Frasch, 2000). Outras repetições antigênicas foram identificadas em outros grupos, incluindo membros de novos grupos. Nove dos 14 peptídeos reativos são compartilhados, especialmente o peptídeo C3, altamente reativo (Figura 15), está presente em um grande número de proteínas (60 no total), inclusive em membros dos TcS grupos V e VI. Aproximadamente 150 membros TcS apresentam sequência similar a esse peptídeo. A reação cruzada entre vários epitopos associada à variabilidade das sequências TcS pode ser um mecanismo de evasão que leva o sistema imune a uma série de respostas ineficazes contra a infecção (Pitcovsky *et al.*, 2002). Esse número de epitopos similares poderia provocar atraso de da resposta imune como evidenciado pelas variações sutis na composição dos aminoácidos próximas ao sítio ativo de TS, impedindo a neutralização da atividade enzimática (Ratier *et al.*, 2008).

2. Diversidade das grandes famílias gênicas de *T. cruzi*

O genoma de *T. cruzi* é altamente repetitivo, sendo que parte dessas repetições é formada por grandes famílias de gênicas, que totalizam 18% dos genes codificadores de proteínas. A extraordinária variação das grandes famílias de *T. cruzi* sugere que existe uma pressão para diversificação de algumas famílias. Após a publicação do genoma não foi realizado uma quantificação e avaliação da diversidade dessas famílias. A diversidade das famílias de proteínas se mostrou muito heterogênea sendo possível a separação da diversidade

em três grupos. O grupo de baixa diversidade seria formado pelas famílias DGF-1 e SAP, média diversidade formado por RHS e mucin-like, enquanto que o grupo de alta diversidade seria formado por TcMUC, TcS, GP63 e MASP. Todas as medidas de diversidade apresentaram a mesma ordem de classificação das famílias. Essa classificação em três grupos, não poderia ser explicada por pequenas diferenças encontradas entre as sequências ou por poucas sequências excepcionalmente distantes, mostrando que a diferença de diversidade entre as famílias é suficientemente grande para todos os métodos apresentarem a mesma ordem.

Os gráficos MDS produzidos a partir dos alinhamentos múltiplos representam uma visualização da diversidade de cada família e os resultados estão de acordo com os índices de diversidade nucleotídica e protéica encontrados.

As matrizes de distância de alinhamentos par-a-par foram analisadas por MDS e foram gerados gráficos com a projeção de todas as famílias, tanto para sequências de DNA quanto para sequências de proteínas. Essa projeção funciona como um mapa de genes e proteínas mostrando a diferença entre as sequências pela distância (Figura 18 e Figura 19).

A família DGF-1 apresentou um grupo com diversidade baixa e um grande número de membros (~90%) e outros grupos com diversidade alta em poucas sequências (Figura 18 e Figura 19). O principal fator para a diversidade encontrada foi a grande quantidade de sequências muito similares encontradas em somente um grupo. As sequências mais distantes, formando grupos isolados de um único representante, apresentam sequências mais diversas do restante da família, não contribuindo de forma importante para o aumento da diversidade. Outras análises propuseram a divisão da família DGF-1 em pelos menos três grupos (Kawashita *et al.*, 2009), sendo dois desses grupos (A e B) com maior número de representantes, 66 e 51 respectivamente. Os três principais grupos ainda podem ser divididos em subgrupos (A1, A2, A3, A4, A5, B1, B2, C, D e E), mas sendo encontrados conflitos no posicionamento dos grupos A1 e B2 (Kawashita *et al.*, 2009). Isso indica que os grupos da árvore usada na classificação não são consistentes. A classificação da família DGF-1 realizada por Kawashita (2009) não leva em consideração a pequena diversidade encontrada na família. A separação dos grupos realizada por estes autores levou em consideração os agrupamentos encontrados por análises filogenéticas, mas não se considerou o comprimento dos ramos. Neste trabalho, os resultados de MDS e separação dos grupos pelo método *kmeans* indica que a família apresenta um grande grupo de baixa diversidade com aproximadamente 90% da sequências da família.

Assim como a família DGF-1, as famílias RHS e mucin-like apresentam a formação de um grupo principal, apresentando a maioria dos membros. Isso mostra que existem sequências muito similares dentro desses grupos e que a diversidade dessas famílias se deve principalmente à dispersão dos grupos menores e distância entre esses grupos e o grupo com maior quantidade de sequências.

A família SAP apresentou baixa diversidade e número de membros muito variável entre os grupos encontrados usando sequências de DNA e proteínas (tabelas 5 e 6). O número mais similar de proteínas em alguns grupos (azul, rosa verde e preto na projeção de proteínas) pode ser devido a uma distribuição mais homogênea da variação encontrada nessas sequências. Trabalhos realizados por Baida e colaboradores (2006) definiram grupos da família com base na presença de sequências de endereçamento celular (Baida *et al.*, 2006). Não encontramos entretanto uma concordância entre esta classificação e aquela proposta no nosso trabalho através do agrupamento de sequências mais similares (Figura 20 e Figura 21).

A grande diversidade da família TcMUC é devida à alta diversidade de todos os grupos que apresentam variação contínua, indicando que a separação dos grupos é artificial, exceto por aqueles grupos com poucas sequências (Figura 18 e Figura 19). A despeito disto, a separação dos grupos mostrou uma boa correspondência com a classificação da família em TcMUC I e TcMUC II previamente proposta (Acosta-Serrano *et al.*, 2001; Buscaglia *et al.*, 2004). Entretanto, alguns genes classificados como TcMUC I mostraram-se mais próximos na projeção de genes TcMUC II, o que acontece também com alguns genes TcMUC II. Esses problemas com relação à distribuição espacial e classificação em TcMUC I e TcMUC II são provavelmente devido a problemas de anotação dos genes em questão. Outras sequências, porém, apresentam características tanto de proteínas TcMUC I como TcMUC II. Isso mostra que existe uma pequena faixa de continuidade entre esses dois grupos, evidenciando que a separação entre essas sequências não é bem definida e que existe um gradiente de variação. Essa continuidade pode ser considerada um indício da evolução que aconteceu através de expansão dos genes TcMUC juntamente com eventos de mutação que provocaram a diferenciação das sequências, permanecendo algumas sequências com características tanto de TcMUC I quanto TcMUC II. Eventos de recombinação podem também ter gerado estas sequências de características intermediárias.

Trabalhos prévios utilizaram a composição dos motivos repetitivos das proteínas TcMUC como um dos critérios para a classificação da família em TcMUC I e TcMUC II (Buscaglia, *et al.*, 2006). De acordo com esta classificação, TcMUC I teria o motivo T_8KPP,

enquanto que TCMUC II, o motivo T_8[K/Q]AP. Neste trabalho, nós identificamos várias sequências TcMUC que apresentam o motivo T_8EAP, que é mais similar ao consenso da repetição de TcMUC II. Além disso, estas sequências contendo esta nova repetição se agrupam no MDS mais frequentemente com sequências TcMUC II. Desta forma, o motivo T_8EAP poderia ser usado para ajudar a definir as sequências TcMUC II formando o motivo T_8K/P/EAP.

GP63 é uma das famílias que formam o grupo de grande diversidade entre as sequências. As proteínas dessa família foram anteriormente caracterizadas por apresentarem o motivo altamente conservado HExxH, relacionado à sua atividade metaloprotease (Cuevas *et al.*, 2003). Neste trabalho, a projeção MDS evidencia a separação de regiões com proteínas que apresentam o motivo HExxH das proteínas que não o apresentam (Figura 26), sugerindo que a atividade metaloprotease foi perdida em alguns membros da família. Trabalhos prévios identificaram dois grupos da família GP63: GP63-I e GP63-II, sendo que GP63-I foi subdividida em GP63-Ia e b. Todas estas sequências mapearam no mesmo grupo na nossa projeção MDS (Figura 26). Isso mostra que a família é muito mais complexa do que a divisão atual em dois grupos.

A família MASP apresentou a maior diversidade dentre todas as grandes famílias de *T. cruzi* (Figura 17 e Tabela 4). Alguns dos grupos formados nessa família são mais distantes e homogêneos indicados pela projeção MDS de DNA (Figura 18). Outros grupos apresentam grande dispersão e formam grupos com variação contínua, indicando que esses grupos são artificiais.

Parte da diversidade das famílias de proteínas de superfície se deve à variação no comprimento das sequências e variação nas sequências primárias dos genes e proteínas. Essas variações poderiam ocorrer devido a processos de inserção e deleção de regiões dos genes e a um grande acúmulo de mutações ao longo do tempo. A diversidade encontrada nessas famílias pode ser devida a processos evolutivos como mutação, recombinação, duplicação e convergência gênica. A recombinação acontece em outras famílias de proteínas de superfície de outros protozoários parasitas promovendo um aumento drástico da diversidade (Cano, 2001; Young *et al.*, 2008). A recombinação poderia promover a troca de fragmentos entre membros da mesma família e eventualmente entre membros de diferentes famílias. De fato, recentemente foi identificado um clone de uma biblioteca de expressão das formas tripomastigotas do parasito que possui uma sequência quimérica, com a região N-terminal derivada de membros da família TcMUC e a porção central e C-terminal derivada de membros

da família MASP (Bartholomeu *et al.*, 2009). Esta sequência não é um artefato da construção da biblioteca de cDNA já que o gene quimérico e *reads* cobrindo a região de junção da quimera foram identificados nos dados do projeto genoma do parasito. Esses fragmentos compartilhados poderiam facilitar outras trocas entre genes de diferentes famílias, possivelmente facilitando rearranjos genômicos (Cano, 2001; Palmer e Brayton, 2007). Eventos que resultassem na troca de fragmentos entre genes de diferentes famílias levariam a uma aceleração da diferenciação dos genes, gerando alta diversidade. Outros genes quiméricos foram identificados no projeto genoma envolvendo membros da família MASP com genes da família SAP, TcMUC e TcS e entre membros da família SAP com TcMUC e L1Tc (El-Sayed *et al.*, 2005a; Baida *et al.*, 2006; Bartholomeu *et al.*, 2009), o que provocaria variação drástica das sequências e aumento de diversidade. Algumas dessas famílias possuem muitas cópias localizadas nas regiões subteloméricas como TcS (Cano, 2001; Kim *et al.*, 2005), o que poderia sujeitá-las a maior influência dos eventos de recombinação associadas às regiões subteloméricas. Os retrotransposons também estão associados às regiões subteloméricas e regiões internas dos cromossomos enriquecidas em genes codificadores de proteínas de superfície. Essas regiões apresentam clusters direcionais de transcrição muito mais curtos quando comparado com regiões que não apresentam estes genes e são sintênicas com os genomas de *T. brucei* e *L. major* (El-Sayed *et al.*, 2005b), sugerindo que essas regiões sofrem ou sofreram intensos rearranjos.

A comparação das projeções usando as regiões 3' flanqueadora das grandes famílias de *T. cruzi* mostrou que existem famílias que apresentam diversidade baixa (TcMUC e MASP, Figura 32) após o códon de terminação. No entanto, outras famílias apresentam uma distribuição ampla mostrando existir uma grande diversidade nestas regiões (GP63, TcS, RHS e DGF-1, Figura 32). A diversidade da região 3' flanqueadora não está associada à diversidade encontrada nas sequências de proteínas, existindo famílias que apresentaram diversidade muito baixa nas sequências protéicas, mas que apresentaram sequências 3' flanqueadora com grande diversidade (DGF-1 e RHS, Figura 32). Ainda existem famílias com grande diversidade nas proteínas, como as famílias TcMUC e MASP, mas que apresentam sequências 3' flanqueadora com baixa dispersão (Figura 32).

Análises da expressão e localização das proteínas DGF-1 mostraram que essa família é mais expressa em amastigota, mas também apresenta baixa expressão em tripomastigota seguida de epimastigota (Lander *et al.*, 2010). Essas diferenças poderiam estar relacionadas à

diversidade da região 3'UTR desses genes, que poderiam modular diferencialmente os níveis de expressão destes genes ao longo do ciclo de vida do parasito.

As sequências 3' flanqueadora dos genes TcMUC podem ser separadas em duas regiões com variação contínua (Figura 34), assim como observado nas projeções produzidas usando as sequências de DNA e proteínas da família. Existem trabalhos que mostram diferenças de expressão entre as duas formas de TcMUC (Freitas-Junior *et al.,* 1998; Acosta-Serrano *et al*., 2001; Buscaglia *et al.,* 2004; Campo *et al*., 2004; Buscaglia *et al*., 2006). Ao nível dos transcritos, TcMUC I é expressa preferencialmente nas formas presentes no hospedeiro vertebrado, enquanto que TcMUC II seria expressa em quantidades variáveis em todos os estágios do ciclo de vida de *T. cruzi* (Freitas-Junior *et al., 19*98; Acosta-Serrano *et al.*, 2001). Dados de expressão protéica, sugerem que TcMUC I seja mais expressa na forma amastigota, enquanto que TcMUC II no estágio tripomastigota (Buscaglia *et al.*, 2004; Campo, Di Noia *et al.*, 2004; Buscaglia, Campo *et al.*, 2006). Seria interessante investigar se as duas regiões encontradas na projeção usando sequências 3' flanqueadora de TcMUC poderiam estar envolvidas na modulação da expressão de TcMUC I e II.

A projeção das sequências 3' flanqueadora de GP63 mostra que esta região é bastante diversa na família (Figura 32), como demonstrado em outros trabalhos comparando a região 3'UTR de GP63-I e -II (Cuevas *et al.*, 2003). Estes autores verificaram que GP63-I e GP63-II apresentam uma expressão diferencial, sendo GP63-I mais expressa nas formas epimastigotas e amastigotas, com expressão basal em tripomastigotas, enquanto que GP63-II é mais expressa nas formas tripomastigotas e amastigotas, tendo expressão baixa na forma epimastigota. (Cuevas *et al.*, 2003). As sequências 3' flanqueadora dos representantes de GP63-I e -II apresentam-se bem distantes na projeção MDS (Figura 33). É possível, portanto que as diferenças nestas 3'UTRs possam contribuir para a expressão diferencial destes genes.

A região 3' flanqueadora dos genes da família MASP apresenta muito conservada Figura 33), diferenciando-se marcadamente das sequências codificadoras e proteínas. Esse resultado está de acordo com análises da variabilidade da região 3' flanqueadora de 771 genes completos de MASP (Bartholomeu *et al.*, 2009). Os genes MASP apresentam sua expressão preferencialmente na forma tripomastigota (Atwood *et al.*, 2005; Bartholomeu *et al.*, 2009), mas não apresentam variação na região 3' flanqueadora. A região 3'UTR conservada de MASP poderia regular essa expressão estágio-específica durante o ciclo de vida de *T. cruzi* (Bartholomeu *et al.*, 2009). Uma explicação alternativa é, o controle para expressão de MASP possivelmente se localizar em outra região e a conservação que é encontrada na região 3'

flanqueadora observada em MASP pode se relacionar com a geração de diversidade da região codificadora (Bartholomeu *et al.*, 2009). A família MSP2 de *Anaplasma marginale* envolvida em variação antigênica possui diversidade na região codificadora, mas apresenta regiões flanqueadoras conservadas que funcionam como ponto de ancoragem para recombinação e troca de segmentos entre os diferentes genes da família, mesmo que não exista grande identidade na região codificadora (Futse *et al.*, 2005; Palmer e Brayton, 2007), permitindo assim formação de novas variantes dos genes MSP2 e evasão do sistema imune. Esse mesmo mecanismo poderia estar sendo usado por *T. cruzi* para a geração de variabilidade na família MASP (Bartholomeu *et al.*, 2009).

3. Mecanismos evolutivos

A grande variação no comprimento dos membros de todas as famílias mostra que ocorreram vários eventos de inserção e deleção nesses genes. Esses eventos de *indel* provocam um grande aumento da diversidade das famílias especialmente se gerarem alteração da fase de leitura do transcrito, podendo ser uma forma de diferenciação das sequências expressas. A comparação das distâncias nucleotídica e protéica da família MASP mostrou que alguns genes apresentam aumento da distância protéica sem grande alteração da sequência nucleotídica (Figura 29 e Figura 30). Entretanto, uma correlação positiva entre estes índices de diversidade é observada quando analisamos as sequências por intervalos de tamanho (Figura 31). Isso mostra que parte da falta de correspondência entre distâncias nucleotídica e protéica foi devido a dificuldade no alinhamento, e que os grupos formados com base na projeção MDS para a família MASP são artificiais. Essa grande diversidade é devida à variação no comprimento das sequências (Figura 16) e à localização de fragmentos compartilhados, que se apresentam em posições muito diferentes entre as sequências (El-Sayed *et al.*, 2005a). Esses fragmentos compartilhados em diferentes regiões limitam a acurácia dos alinhamentos gerando muitos *mismatches* e *gaps*, aumentando os índices de diversidade. Além dessas variações existem as sequências quiméricas que contribuem para o aumento da diversidade da família e dos grupos que contêm esse tipo de sequência (Figura 27). Entretanto, mesmo nos grupos que foram formados usando como característica o tamanho das sequências, é possível identificar alguns pontos que mostram uma falta de correspondência entre as distâncias nucleotídica e protéica. Esses pares de sequências poderiam sofrer alterações *indels* que modificam a fase de leitura do gene levando a uma grande alteração na sequência de proteínas.

O resultado em que grupos de sequências similares na projeção da matriz de DNA apresentam-se diferentes e muito dispersos na projeção da matriz de proteína, observado na família MASP (Figura 28), poderia ter sido gerado por um mecanismo de mudança de fase de leitura. A falta de identidade em um trecho entre as sequências de proteínas poderia ser causada por *indels* que alteram a fase de leitura, e a continuidade da identidade nas proteínas em outros trechos poderia ser explicada por outro evento *indel* de restauração da fase de leitura (Figura 36). Outra explicação poderia se o surgimento de mutações afetando a fase de leitura em genes diferentes e a recombinação desses dois genes formaria uma variante que apresenta as duas mutações. Essa nova variante apresentaria mudança da fase de leitura em um trecho curto e promoveria aumento da diversidade na região hipervariável da proteína.

Outras famílias parecem não apresentar esse tipo de alteração. As famílias GP63, RHS, e TcS não apresentaram redução de possíveis códons de paradas que poderiam gerar proteínas truncadas quando comparadas com as sequências aleatórias (Figura 37). Já as famílias TcMUC e MASP apresentam redução do número de códons de parada gerados por *indels* quando comparados com sequências aleatórias (Figura 37). Isso poderia ser um mecanismo para favorecer a geração de diversidade através de eventos *indels* minimizando as chances de gerar uma proteína truncada.

Além de apresentarem redução do número médio de códons de parada, nas famílias TcMUC e MASP há uma menor frequência de códons de parada nas extremidades dos genes nas simulações de mudança de fase de leitura (Figura 38), indicando que os códons de parada gerados por alteração de fase de leitura são menos frequentes nas extremidades codificadoras da regiões N- e C-terminal (Figura 38). Genes que apresentem mudança de fase de leitura envolvendo essas regiões teriam menor chance de gerar proteínas truncadas, as quais podem gerar produtos não funcionais e assim um gasto de energia desnecessário. Dessa forma, se alterações de fase de leitura dos genes da família MASP representa um mecanismos de geração de variabilidade, os genes possuíssem sequências que não formam códons de parada prematuros seriam selecionados, evitando parte dos efeitos indesejáveis da tradução de RNAs truncados.

O projeto genoma de *T. cruzi* mostrou a existência de sequências quiméricas que apresentam domínio conservado N- ou C-terminal de MASP combinado com domínio N- ou C-terminal de TcMUC ou TcS (El-Sayed *et al.*, 2005a). Também foram identificadas sequências da família SAP compartilhando domínio C-terminal com membros da família TcMUC e MASP (Baida *et al.*, 2006) e identificados em uma biblioteca de cDNA sequências

MASP apresentando domínio C-terminal de TcMUC e MASP, mostrando que a sequência quiméricas são expressas (Bartholomeu *et al*., 2009). O mecanismo para geração dessas sequências quiméricas é desconhecido, sendo necessária uma avaliação das famílias que apresentam esse tipo sequência e buscar evidências para sua formação.

A fim de identificar sequências quiméricas dentro da família MASP, nós pesquisamos a ocorrência de motivos protéicos compartilhados entre membros da família derivados de diferentes grupos definidos pelo método *kmeans* (Figura 39, Figura 40, Figura 41, Figura 42 e Figura 43). Os dois primeiros tipos de MEME, representando 8 casos, não apresentam evidências de troca de fragmentos, enquanto o terceiro tipo de MEME, formado por 22 casos, apresentam muitas evidências de troca. No primeiro tipo de MEME todos os genes pertencem ao mesmo grupo, não sendo compartilhados (Figura 39). O segundo tipo de MEME apresenta muita variação (Figura 40). Esses fragmentos com muita variação, pertencendo a vários grupos, não constituem uma evidência forte da troca de fragmentos entre os genes, já que essa grande variação pode ser devida à presença do fragmento em um gene ancestral ou ainda a trocas antigas do fragmento entre os genes de diferentes grupos. O terceiro tipo de MEME poderia evidenciar trocas de fragmentos entre os membros por serem muito similares e estarem presentes em diferentes grupos (Figura 41). Verificamos que estes fragmentos similares entre genes de diferentes grupos estão, na maioria das vezes, localizados nas extremidades 3' dos genes (Figura 43). Estes dados estão de acordo com nossa hipótese que a região 3' flanqueadora conservada da família MASP pode se relacionar com a geração de diversidade da região codificadora (Bartholomeu *et al.*, 2009), em um mecanismos similar ao previamente descrito para a família MSP2 de *Anaplasma marginale* (Futse *et al.*, 2005). Os genes que codificam a família MSP2 apresentam uma estrutura semelhante a dos genes de MASP com regiões flanqueadoras conservadas e uma região central hipervariável. De acordo com o modelo de geração de variabilidade de MSP2, o grande segmento conservado na região 3' funcionaria como âncora para que essas recombinações ocorressem, sendo que um ponto de recombinação ocorreria dentro da região conservada e o outro na região central hipervariável (Futse *et al.*, 2005). Além das evidências de trocas de fragmentos entre os genes MASP, existem também evidências de trocas de fragmentos de DNA entre genes de diferentes famílias. A maior frequência de fragmentos compartilhados foi entre 100 e 200 nt (91%, 368 fragmentos) (tabela 7). A alta frequência de fragmentos compartilhados com grande identidade pode ser indício do tamanho preferencial para troca. Outros trabalhos apresentam indício de recombinação de fragmentos variando entre 100 e 200pb. Foi detectado em *T. brucei* pontos

de quebra em alinhamento de sequências indicando pontos de recombinação distantes na maioria das vezes variando entre 129 e 207pb (Young *et al.*, 2008).

A busca por fragmentos compartilhados entre os genes das famílias de proteínas de superfície mostrou que nem todas as famílias devem estar envolvidas nesse possível mecanismo para geração de variabilidade. A família MASP foi a que apresentou mais fragmentos compartilhados com outras famílias (54%, 221 fragmentos). Novamente verificamos que a maioria dos fragmentos compartilhados (≥ 100nt e identidade ≥ 90%) entre as famílias são classificados como fragmentos de extremidade 3' (89%, 362 fragmentos). *T. cruzi* poderia usar essas regiões flanqueadoras conservadas como pontos de ancoragem para recombinação e geração de variabilidade De fato, MASP e TcMUC que são as duas famílias multigênicas que apresentaram regiões 3' flanqueadoras mais conservadas, apresentam uma dispersão contínua na projeção do MDS, sugerindo que eventos de recombinação podem estar envolvidos no padrão de variabilidade encontrado para estas famílias.

7. Conclusões

Baseado nos dados apresentados nesse trabalho é possível concluir que:

1. A grande diversidade da família TcS permite sua classificação em oito grupos bem definidos e com características próprias.Quatro destes grupos (TcS I a IV) já haviam sido descritos anteriormente e albergam proteínas com atividade trans-sialidase, e envolvidas em adesão/invasão celular e bloqueio da via do complemento. Os grupos TcS V a VIII foram identificados neste trabalho. Baseado no padrão de dispersão dos oito grupos na projeção MDS e ocorrência/sequência de motivos encontrados em cada grupo, nós especulamos que os novos grupos V e VI e o grupo TcS II previamente descrito são mais relacionados entre si, quando comparado com os outros grupos. Estes são os únicos grupos que não apresentam o motivo FRIP e as sequências consenso do motivo VTVxNVxLYNR são bem similares entre estes grupos. Da mesma forma, os dados apóiam a hipótese que os novos grupos TcS VII e VIII e o grupo TcS III são mais relacionados entre si. Estes grupos compartilham o mesmo padrão de ocorrência de motivos e são clusterizados numa mesma região na projeção MDS.

2. Não existe uma associação de um grupo TcS com um cromossomo específico. Entretanto, verificamos que os grupos de TcS II e V apresentam localização preferencial ao longo dos cromossomos. O TcSgrupoII está localizado preferencialmente nas extremidades dos cromossomos e apresenta maior número de genes TcS nas regiões subteloméricas, enquanto o TcSgrupoV é localizado preferencialmente no interior dos cromossomos. Além desses grupos, TcS pseudogenes também se apresentam preferencialmente localizados nas extremidades dos cromossomos. Os outros grupos não apresentaram localização preferencial. Especulamos que os genes preferencialmente localizados nas regiões subteloméricas, sofreram ou podem sofrer ação de intensos rearranjos genéticos que são mais frequentes nestas regiões o que poderia favorecer a formação de novas variantes. Isso pode ser particularmente importante para TcS grupo II porque vários de seus membros estão envolvidos na adesão/invasão de células do hospedeiro e *T. cruzi* apresenta habilidade de infectar uma grande variedade de tipos celulares.

3. A diversidade encontrada na região 3'flanqueadora dos genes TcS é maior que a encontrada na região codificadora. Não foi encontrada uma associação clara entre conservação

de sequências 3' flanqueadoras dentro dos grupos. Entretanto, alguns genes que apresentam padrão de expressão similar apresentam regiões 3' flanqueadoras também similares, sugerindo que estas sequências possam modular a abundância dos transcritos ao longo do ciclo de vida do parasito.

4. Novos epitopos de célula B foram identificados na família TcS em membros de grupos previamente descritos bem como em membros de novos grupos. Nove dos 14 peptídeos reativos foram encontrados em mais de um membro da família. Um destes peptídeos ocorre em 60 proteínas incluindo membros dos novos grupos V and VI. Além disso, sequências similares, mas não idênticas a este peptídeo, foram encontradas em 150 membros de TcS. Especulamos que a reatividade cruzada entre vários epitopos TcS e variabilidade de sequência da família pode contribuir na estratégia do parasito de escapar do ataque do sistema imune através da exposição simultânea de epitopos de célula B relacionados gerando respostas imunes espúrias e não neutralizantes.

5. A diversidade, tanto nucleotídica quanto protéica, das famílias de proteínas de superfície de *T. cruzi* é muito heterogênea tanto entre as famílias quanto entre os grupos dentro das famílias. As estimativas de diversidade mostram as famílias DGF-1 e SAP como de baixa diversidade, mucin-like e RHS como de média diversidade, e TcMUC, TcS, GP63 e MASP como famílias de grande diversidade, sendo a família MASP a família com maior diversidade encontrada.

6. MASP e TcMUC apresentam um padrão similar de diversidade das regiões codificadora e 3' flaqueadora. As sequências codificadoras de ambas famílias apresentam, de maneira geral, um padrão contínuo de diversidade enquanto que a região 3' flanqueadora de cada família é altamente conservada. Especulamos que eventos de recombinação entre os membros de cada família possa contribuir para este padrão de diversidade, em um mecanismo similar ao previamente descrito para os genes MSP2 de *Anaplasma marginale* (Futse *et al.*, 2005).

7. A família MASP apresenta vários casos de falta de correspondência da distância nucleotídica e protéica (baixa distância nucleotídica e grande distância protéica) que poderiam

ser explicados por mutações que mudam a fase de leitura dos genes, levando a diversificação das proteínas.

8. As sequências dos genes das famílias TcMUC e MASP parecem ter evoluído de forma a minimizar as chances de formação de códons prematuros, evitando a formação de proteínas truncadas, principalmente na parte 3' dos genes.

9. A família MASP apresenta indícios de troca de segmentos de DNA entre os genes, evidenciado por fragmentos similares que são compartilhados por diferentes genes de diferentes grupos. As troca de fragmentos entre os genes MASP aconteceria preferencialmente usando regiões conservadas encontradas nas extremidades das regiões codificadoras, principalmente a região 3' dos genes.

10. As famílias gênicas que codificam proteínas de superfície DGF-1, GP63, MASP, SAP, TcMUC e TcS apresentaram indícios de troca de segmentos de DNA inter-famílias. Essa troca de fragmentos frequentemente envolveria os genes da família TcS e principalmente os genes das famílias MASP. Estes eventos aconteceriam preferencialmente nas extremidades das regiões codificadoras, principalmente a região 3' dos genes.

8. Referências

ACOSTA-SERRANO, A. *et al.* The mucin-like glycoprotein super-family of *Trypanosoma cruzi*: structure and biological roles. **Mol Biochem Parasitol,** v. 114, n. 2, p. 143-50, May 2001. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11378194 >.

AGÜERO, F. *et al.* Generation and analysis of expressed sequence tags from *Trypanosoma cruzi* trypomastigote and amastigote cDNA libraries. **Mol Biochem Parasitol,** v. 136, n. 2, p. 221-5, Aug 2004. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15478800 >.

ALTSCHUL, S. F. *et al.* Basic local alignment search tool. **J Mol Biol,** v. 215, n. 3, p. 403-10, Oct 1990. ISSN 0022-2836. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2231712 >.

ATWOOD, J. A. *et al.* The *Trypanosoma cruzi* proteome. **Science,** v. 309, n. 5733, p. 473-6, Jul 2005. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020736 >.

BAIDA, R. C. *et al.* Molecular characterization of serine-, alanine-, and proline-rich proteins of *Trypanosoma cruzi* and their possible role in host cell infection. **Infect Immun,** v. 74, n. 3, p. 1537-46, Mar 2006. ISSN 0019-9567. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16495524 >.

BAILEY, T. L.; ELKAN, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. **Proc Int Conf Intell Syst Mol Biol,** v. 2, p. 28-36, 1994. ISSN 1553-0833. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7584402 >.

BARRETT, M. P. *et al.* The trypanosomiases. **Lancet,** v. 362, n. 9394, p. 1469-80, Nov 2003. ISSN 1474-547X. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14602444 >.

BARRY, D. *et al.* **Trypanosomes After the Genome**. 1st. Horizon Bioscience, 2006. 423 ISBN 978-1-904933-27-4.

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 Res,** v. 37, n. 10, p. 3407-17, Jun 2009. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19336417 >.

BARTHOLOMEU, D. C. *et al. Trypanosoma cruzi*: RNA structure and post-transcriptional control of tubulin gene expression. **Exp Parasitol,** v. 102, n. 3-4, p. 123-33, 2002 Nov-Dec 2002. ISSN 0014-4894. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12856307 >.

BEHBEHANI, K. Developmental cycles of *Trypanosoma (Schizotrypanum) cruzi* (Chagas, 1909) in mouse peritoneal macrophages in vitro. **Parasitology,** v. 66, n. 2, p. 343-53, Apr 1973. ISSN 0031-1820. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/4595116 >.

BENDTSEN, J. D. *et al.* Improved prediction of signal peptides: SignalP 3.0. **J Mol Biol,** v. 340, n. 4, p. 783-95, Jul 2004. ISSN 0022-2836. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15223320 >.

BERRIMAN, M. *et al.* The genome of the African trypanosome *Trypanosoma brucei*. **Science,** v. 309, n. 5733, p. 416-22, Jul 2005. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020726 >.

BEUCHER, M.; NORRIS, K. A. Sequence diversity of the *Trypanosoma cruzi* complement regulatory protein family. **Infect Immun,** v. 76, n. 2, p. 750-8, Feb 2008. ISSN 1098-5522. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18070905 >.

BOUVIER, J.; ETGES, R. J.; BORDIER, C. Identification and purification of membrane and soluble forms of the major surface protein of *Leishmania* promastigotes. **J Biol Chem,** v. 260, n. 29, p. 15504-9, Dec 1985. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/4066680 >.

BRINGAUD, F. *et al.* A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*. **Eukaryot Cell,** v. 1, n. 1, p. 137-51, Feb 2002. ISSN 1535-9778. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12455980 >.

BRIONES, M. R. *et al.* Trans-sialidase genes expressed in mammalian forms of *Trypanosoma cruzi* evolved from ancestor genes expressed in insect forms of the parasite. **J Mol Evol,** v. 41, n. 2, p. 120-31, Aug 1995. ISSN 0022-2844. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7666441 >.

BRISSE, S. *et al.* Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. **Infect Genet Evol,** v. 2, n. 3, p. 173-83, Feb 2003. ISSN 1567-1348. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12797979 >.

BROCCHIERI, L. Phylogenetic inferences from molecular sequences: review and critique. **Theor Popul Biol,** v. 59, n. 1, p. 27-40, Feb 2001. ISSN 0040-5809. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11243926 >.

BUSCAGLIA, C. A. *et al.* The surface coat of the mammal-dwelling infective trypomastigote stage of *Trypanosoma cruzi* is formed by highly diverse immunogenic mucins. **J Biol Chem,** v. 279, n. 16, p. 15860-9, Apr 2004. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14749325 >.

BUSCAGLIA, C. A. *et al.*. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. **Nat Rev Microbiol,** v. 4, n. 3, p. 229-36, Mar 2006. ISSN 1740-1526. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16489349 >.

BUSCHIAZZO, A.; CAMPETELLA, O.; FRASCH, A. C. *Trypanosoma rangeli* sialidase: cloning, expression and similarity to *T. cruzi* trans-sialidase. **Glycobiology,** v. 7, n. 8, p. 1167-73, Dec 1997. ISSN 0959-6658. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9455917 >.

BUSCHIAZZO, A. *et al.* Structural basis of sialyltransferase activity in trypanosomal sialidases. **EMBO J,** v. 19, n. 1, p. 16-24, Jan 2000. ISSN 0261-4189. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10619840 >.

CAMPO, V. *et al.* Differential accumulation of mutations localized in particular domains of the mucin genes expressed in the vertebrate host stage of *Trypanosoma cruzi*. **Mol Biochem Parasitol,** v. 133, n. 1, p. 81-91, Jan 2004. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14668015 >.

CAMPO, V. A. *et al.* Immunocharacterization of the mucin-type proteins from the intracellular stage of *Trypanosoma cruzi*. **Microbes Infect,** v. 8, n. 2, p. 401-9, Feb 2006. ISSN 1286-4579. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16253534 >.

CANO, M. I. Telomere biology of trypanosomatids: more questions than answers. **Trends Parasitol,** v. 17, n. 9, p. 425-9, Sep 2001. ISSN 1471-4922. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11530354 >.

CARLTON, J. The *Plasmodium vivax* genome sequencing project. **Trends Parasitol,** v. 19, n. 5, p. 227-31, May 2003. ISSN 1471-4922. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12763429 >.

CARLTON, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. **Science,** v. 315, n. 5809, p. 207-12, Jan 2007. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17218520 >.

CARMO, M. S. *et al.* Isolation and characterisation of genomic and cDNA clones coding for a serine-, alanine-, and proline-rich protein of *Trypanosoma cruzi*. **Int J Parasitol,** v. 31, n. 3, p. 259-64, Mar 2001. ISSN 0020-7519. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11226452 >.

CAZZULO J. J., FRASCH A. C. SAPA/trans-sialidase and cruzipain: two antigens from *Trypanosoma cruzi* contain immunodominant but enzymatically inactive domains. **FASEB J.** Nov;6(14):3259-64, 1992. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1426764 >.

CHIURILLO, M. A.; PERALTA, A.; RAMÍREZ, J. L. Comparative study of *Trypanosoma rangeli* and *Trypanosoma cruzi* telomeres. **Mol Biochem Parasitol,** v. 120, n. 2, p. 305-8, Apr 2002. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11897137 >.

COLLI, W Trans-sialidase: a unique enzyme activity discovered in the protozoan *Trypanosoma cruzi*. **FASEB J** 7: 1257-1264, 1993. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8405811 >.

COURA, J. R. *et al.* Emerging Chagas disease in Amazonian Brazil. **Trends Parasitol,** v. 18, n. 4, p. 171-6, Apr 2002. ISSN 1471-4922. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11998705 >.

CREMONA, M. L. *et al.* Enzymically inactive members of the trans-sialidase family from *Trypanosoma cruzi* display beta-galactose binding activity. **Glycobiology,** v. 9, n. 6, p. 581-7, Jun 1999. ISSN 0959-6658. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10336990 >.

CREMONA, M. L. *et al.* A single tyrosine differentiates active and inactive *Trypanosoma cruzi* trans-sialidases. **Gene,** v. 160, n. 1, p. 123-8, Jul 1995. ISSN 0378-1119. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7628705 >.

CROOKS, G. E. *et al.* WebLogo: a sequence logo generator. **Genome Res,** v. 14, n. 6, p. 1188-90, Jun 2004. ISSN 1088-9051. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15173120 >.

CROSS, G. A.; TAKLE, G. B. The surface trans-sialidase family of *Trypanosoma cruzi*. **Annu Rev Microbiol,** v. 47, p. 385-411, 1993. ISSN 0066-4227. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8257103 >.

CUEVAS, I. C.; CAZZULO, J. J.; SÁNCHEZ, D. O. gp63 homologues in *Trypanosoma cruzi*: surface antigens with metalloprotease activity and a possible role in host cell infection. **Infect Immun,** v. 71, n. 10, p. 5739-49, Oct 2003. ISSN 0019-9567. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14500495 >.

DAYHOFF, M. O., Schwartz, R. M., and Orcutt, B. C. A model for evolutionary change in proteins. **Atlas of Protein Sequence and Structure**, volume 5, pages 345-352, 1978.

D'HAESELEER, P. How does gene expression clustering work? **Nat Biotechnol,** v. 23, n. 12, p. 1499-501, Dec 2005. ISSN 1087-0156. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16333293 >.

DE FREITAS, J. M. *et al.* Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. **PLoS Pathog,** v. 2, n. 3, p. e24, Mar 2006. ISSN 1553-7374. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16609729 >.

EL-SAYED, N. M. *et al.* The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science,** v. 309, n. 5733, p. 409-15, Jul 2005a. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020725 >.

EL-SAYED. *et al.* Comparative genomics of trypanosomatid parasitic protozoa. **Science,** v. 309, n. 5733, p. 404-9, Jul 2005b. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020724 >.

ENGSTLER, M.; REUTER, G.; SCHAUER, R. Purification and characterization of a novel sialidase found in procyclic culture forms of *Trypanosoma brucei*. **Mol Biochem Parasitol,** v. 54, n. 1, p. 21-30, Aug 1992. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1518530 >.

ETGES, R.; BOUVIER, J.; BORDIER, C. The major surface protein of *Leishmania* promastigotes is anchored in the membrane by a myristic acid-labeled phospholipid. **EMBO J,** v. 5, n. 3, p. 597-601, Mar 1986. ISSN 0261-4189. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/3709520 >.

FANKHAUSER, N.; MÄSER, P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. **Bioinformatics,** v. 21, n. 9, p. 1846-52, May 2005. ISSN 1367-4803. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15691858 >.

FELSENSTEIN, J. Phylogeny Inference Package (Version 3.2). ***Cladistics*****,** v. 5, p. 164-166, 1989.

FELSENSTEIN, J. **PHYLIP (Phylogeny Inference Package) version 3.6**. Free program distributed by the authors over the internet from http://evolution.genetics.washington.edu/phylip.html: *Department of Genome Sciences, University of Washington, Seattle.* 2005.

FERREIRA, K. A. *et al.* Genome survey sequence analysis and identification of homologs of major surface protease (gp63) genes in *Trypanosoma rangeli*. **Vector Borne Zoonotic Dis,** v. 10, n. 9, p. 847-53, Nov 2010. ISSN 1557-7759. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20420528 >.

FLORES-LÓPEZ, C. A.; MACHADO, C. A. Analyses of 32 Loci Clarify Phylogenetic Relationships among *Trypanosoma cruzi* Lineages and Support a Single Hybridization prior to Human Contact. **PLoS Negl Trop Dis,** v. 5, n. 8, p. e1272, Aug 2011. ISSN 1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21829751 >.

FRANK, R. Spot-synthesis: an easy technique for the position- ally addressable, parallel chemical synthesis on a membrane support. **Tetrahedron,** v. 48, n. 42, p. 9217-9232, 1992.

FRANZÉN, O. *et al.* Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio X10/1 and comparison with *T. cruzi* VI CL Brener. **PLoS Negl Trop Dis,** v. 5, n. 3, p. e984, 2011. ISSN 1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21408126 >.

FRASCH, A. C. Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*. **Parasitol Today,** v. 16, n. 7, p. 282-6, Jul 2000. ISSN 0169-4758. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10858646 >.

FREITAS-JUNIOR, L. H.; BRIONES, M. R.; SCHENKMAN, S. Two distinct groups of mucin-like genes are differentially expressed in the developmental stages of *Trypanosoma cruzi*. **Mol Biochem Parasitol,** v. 93, n. 1, p. 101-14, May 1998. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9662032 >.

FUTSE, J. E. *et al.* Structural basis for segmental gene conversion in generation of Anaplasma marginale outer membrane protein variants. **Mol Microbiol,** v. 57, n. 1, p. 212-21, Jul 2005. ISSN 0950-382X. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15948961 >.

GARCÍA, G. A. *et al.* Immunological and pathological responses in BALB/c mice induced by genetic administration of Tc 13 Tul antigen of *Trypanosoma cruzi*. **Parasitology,** v. 132, n. Pt 6, p. 855-66, Jun 2006. ISSN 0031-1820. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16478565 >.

GARCÍA, G. A. *et al. Trypanosoma cruzi*: molecular identification and characterization of new members of the Tc13 family. Description of the interaction between the Tc13 antigen from Tulahuén strain and the second extracellular loop of the beta(1)-adrenergic receptor. **Exp Parasitol,** v. 103, n. 3-4, p. 112-9, 2003 Mar-Apr 2003. ISSN 0014-4894. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12880587 >.

GARDNER, M. J. *et al.* Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. **Science,** v. 309, n. 5731, p. 134-7, Jul 2005. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15994558 >.

GARDNER, M. J. Genome sequence of the human malaria parasite *Plasmodium falciparum*. **Nature,** v. 419, n. 6906, p. 498-511, Oct 2002. ISSN 0028-0836. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12368864 >.

GASKELL, A.; CRENNELL, S.; TAYLOR, G. The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. **Structure,** v. 3, n. 11, p. 1197-205, Nov 1995. ISSN 0969-2126. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8591030 >.

GIORDANO, R. *et al. Trypanosoma cruzi* binds to laminin in a carbohydrate-independent way. **Braz J Med Biol Res,** v. 27, n. 9, p. 2315-8, Sep 1994. ISSN 0100-879X. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7787815 >.

GONNET, G. H.; COHEN, M. A.; BENNER, S. A. Exhaustive matching of the entire protein sequence database. **Science,** v. 256, n. 5062, p. 1443-5, Jun 1992. ISSN 0036-8075. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1604319 >.

GRANDGENETT, P. M. *et al.* Differential expression of GP63 genes in *Trypanosoma cruzi*. **Mol Biochem Parasitol,** v. 110, n. 2, p. 409-15, Oct 2000. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11071294 >.

GRISARD, E. C. *et al.* Transcriptomic analyses of the avirulent protozoan parasite *Trypanosoma rangeli*. **Mol Biochem Parasitol,** v. 174, n. 1, p. 18-25, Nov 2010. ISSN 1872-9428. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20600354 >.

HALPERN, A. L.; BRUNO, W. J. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. **Mol Biol Evol,** v. 15, n. 7, p. 910-7, Jul 1998. ISSN 0737-4038. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9656490 >.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. **Journal of the Royal Statistical Society, Series C (Applied Statistics)** 28 (1): 100–108, 1979. Disponível em: < http://www.jstor.org/pss/2346830 >.

HORN, D.; BARRY, J. D. The central roles of telomeres and subtelomeres in antigenic variation in African trypanosomes. **Chromosome Res,** v. 13, n. 5, p. 525-33, 2005. ISSN 0967-3849. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16132817 >.

HUDSON, L.; SNARY, D.; MORGAN, S. J. *Trypanosoma cruzi*: continuous cultivation with murine cell lines. **Parasitology,** v. 88 (Pt 2), p. 283-94, Apr 1984. ISSN 0031-1820. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/6371672 >.

IENNE, S. *et al.* Network genealogy of 195-bp satellite DNA supports the superimposed hybridization hypothesis of *Trypanosoma cruzi* evolutionary pattern. **Infect Genet Evol,** v. 10, n. 5, p. 601-6, Jul 2010. ISSN 1567-7257. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20433949 >.

IVENS, A. C. *et al.* The genome of the kinetoplastid parasite, *Leishmania major*. **Science,** v. 309, n. 5733, p. 436-42, Jul 2005. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020728 >.

JONES, D. T.; TAYLOR, W. R.; THORNTON, J. M. The rapid generation of mutation data matrices from protein sequences. **Comput Appl Biosci,** v. 8, n. 3, p. 275-82, Jun 1992. ISSN 0266-7061. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1633570 >.

JUKES T. H. e CANTOR C. R. Evolution of protein molecules. **Mammalian Protein Metabolism**, pp. 21-132, Academic Press, New York, 1969.

KALINOWSKI, S. T. Evolutionary and statistical properties of three genetic distances. **Mol Ecol,** v. 11, n. 8, p. 1263-73, Aug 2002. ISSN 0962-1083. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12144649 >.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **Journal of Molecular Evolution** 16: 111–120, 1980. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7463489 >.

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. **Mol Biochem Parasitol,** v. 165, n. 1, p. 19-31, May 2009. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19393159 >.

KIM, D. *et al.* Telomere and subtelomere of *Trypanosoma cruzi* chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of T. cruzi telomeres. **Gene,** v. 346, p. 153-61, Feb 2005. ISSN 0378-1119. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15716016 >.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **J Mol Evol,** v. 16, n. 2, p. 111-20, Dec 1980. ISSN 0022-2844. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7463489 >.

KULKARNI, M. M. *et al. Trypanosoma cruzi* GP63 proteins undergo stage-specific differential posttranslational modification and are important for host cell infection. **Infect Immun,** v. 77, n. 5, p. 2193-200, May 2009. ISSN 1098-5522. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19273559 >.

LANDER, N. *et al.* Localization and developmental regulation of a dispersed gene family 1 protein in *Trypanosoma cruzi*. **Infect Immun,** v. 78, n. 1, p. 231-40, Jan 2010. ISSN 1098-5522. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19841080 >.

LARKIN, M. A. *et al.* Clustal W and Clustal X version 2.0. **Bioinformatics,** v. 23, n. 21, p. 2947-8, Nov 2007. ISSN 1367-4811. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17846036 >.

LARSEN, J. E.; LUND, O.; NIELSEN, M. Improved method for predicting linear B-cell epitopes. **Immunome Res,** v. 2, p. 2, 2006. ISSN 1745-7580. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16635264 >.

LEGUIZAMON, M. S. *et al.* Bloodstream *Trypanosoma cruzi* parasites from mice simultaneously express antigens that are markers of acute and chronic human Chagas disease. **Parasitology,** v. 102 Pt 3, p. 379-85, Jun 1991. ISSN 0031-1820. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1907729 >.

LEGUIZAMÓN, M. S. *et al.* Mice infected with *Trypanosoma cruzi* produce antibodies against the enzymatic domain of trans-sialidase that inhibit its activity. **Infect Immun,** v. 62, n. 8, p. 3441-6, Aug 1994. ISSN 0019-9567. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8039915 >.

LEY, V. *et al.* Amastigotes of *Trypanosoma cruzi* sustain an infective cycle in mammalian cells. **J Exp Med,** v. 168, n. 2, p. 649-59, Aug 1988. ISSN 0022-1007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/3045248 >.

LOFTUS, B. *et al.* The genome of the protist parasite *Entamoeba histolytica*. **Nature,** v. 433, n. 7028, p. 865-8, Feb 2005. ISSN 1476-4687. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15729342 >.

MACHADO, C. A.; AYALA, F. J. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. **Proc Natl Acad Sci U S A,** v. 98, n. 13, p. 7396-401, Jun 2001. ISSN 0027-8424. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11416213 >.

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. **1. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.** pp. 281–297, 1967. Disponível em: < http://www.zentralblatt-math.org/zmath/en/search/?q=an:0214.46201&format=complete >

MAGDESIAN, M. H. *et al.* Infection by *Trypanosoma cruzi*. Identification of a parasite ligand and its host cell receptor. **J Biol Chem,** v. 276, n. 22, p. 19382-9, Jun 2001. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11278913 >.

MAGDESIAN, M. H. *et al.*. A conserved domain of the gp85/trans-sialidase family activates host cell extracellular signal-regulated kinase and facilitates *Trypanosoma cruzi* infection. **Exp Cell Res,** v. 313, n. 1, p. 210-8, Jan 2007. ISSN 0014-4827. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17101128 >.

MCGWIRE, B. S.; CHANG, K. P. Posttranslational regulation of a *Leishmania* HEXXH metalloprotease (gp63). The effects of site-specific mutagenesis of catalytic, zinc binding, N-glycosylation, and glycosyl phosphatidylinositol addition sites on N-terminal end cleavage, intracellular stability, and extracellular exit. **J Biol Chem,** v. 271, n. 14, p. 7903-9, Apr 1996. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8626468 >.

MILES, M. A. *et al.* Do radically dissimilar *Trypanosoma cruzi* strains (zymodemes) cause Venezuelan and Brazilian forms of Chagas' disease? **Lancet,** v. 1, n. 8234, p. 1338-40, Jun 1981. ISSN 0140-6736. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/6113312 >.

MONTAGNA, G. *et al.* The trans-sialidase from the african trypanosome *Trypanosoma brucei*. **Eur J Biochem,** v. 269, n. 12, p. 2941-50, Jun 2002. ISSN 0014-2956. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12071958 >.

MONTAGNA, G. N.; DONELSON, J. E.; FRASCH, A. C. Procyclic *Trypanosoma brucei* expresses separate sialidase and trans-sialidase enzymes on its surface membrane. **J Biol Chem,** v. 281, n. 45, p. 33949-58, Nov 2006. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16956887 >.

MORTARA, R. A. *et al.* Mammalian cell invasion and intracellular trafficking by *Trypanosoma cruzi* infective forms. **An Acad Bras Cienc,** v. 77, n. 1, p. 77-94, Mar 2005. ISSN 0001-3765. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15692679 >.

NEI, M. Phylogenetic analysis in molecular evolutionary genetics. **Annu Rev Genet,** v. 30, p. 371-403, 1996. ISSN 0066-4197. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8982459 >.

NOGUEIRA, N.; COHN, Z. *Trypanosoma cruzi*: mechanism of entry and intracellular fate in mammalian cells. **J Exp Med,** v. 143, n. 6, p. 1402-20, Jun 1976. ISSN 0022-1007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/775012 >.

OPPEZZO, P. *et al.* Crystal structure of an enzymatically inactive trans-sialidase-like lectin from *Trypanosoma cruzi*: The carbohydrate binding mechanism involves residual sialidase activity. **Biochim Biophys Acta**, Apr 2011. ISSN 0006-3002. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21570497 >.

OUAISSI, A. *et al.* Major surface immunogens of *Trypanosoma cruzi* trypomastigotes. **Mem Inst Oswaldo Cruz,** v. 83 Suppl 1, p. 502, Nov 1988. ISSN 0074-0276. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/3075691 >.

PALMER, G. H.; BRAYTON, K. A. Gene conversion is a convergent strategy for pathogen antigenic variation. **Trends Parasitol,** v. 23, n. 9, p. 408-13, Sep 2007. ISSN 1471-4922. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17662656 >.

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. **J Cell Sci,** v. 113 (Pt 7), p. 1299-307, Apr 2000. ISSN 0021-9533. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10704380 >.

PITCOVSKY, T. A. *et al.* A functional network of intramolecular cross-reacting epitopes delays the elicitation of neutralizing antibodies to *Trypanosoma cruzi* trans-sialidase. **J Infect Dis,** v. 186, n. 3, p. 397-404, Aug 2002. ISSN 0022-1899. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12134236 >.

RATIER, L. *et al.* Relevance of the diversity among members of the *Trypanosoma cruzi* trans-sialidase family analyzed with camelids single-domain antibodies. **PLoS One,** v. 3, n. 10, p. e3524, 2008. ISSN 1932-6203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18949046 >.

ROTMISTROVSKY K., JANG W., e SCHULER G. D. A web server for performing electronic PCR. (Translated from eng) **Nucleic Acids Res** 32(Web Server issue):W108-112, 2004. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/15215361.

SANTANA, J. M. *et al.* A *Trypanosoma cruzi*-secreted 80 kDa proteinase with specificity for human collagen types I and IV. **Biochem J,** v. 325 (Pt 1), p. 129-37, Jul 1997. ISSN 0264-6021. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9224638 >.

SCHENKMAN, S. *et al.* Structural and functional properties of *Trypanosoma* trans-sialidase. **Annu Rev Microbiol,** v. 48, p. 499-523, 1994. ISSN 0066-4227. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7826016 >.

SCHERF, A.; FIGUEIREDO, L. M.; FREITAS-JUNIOR, L. H. *Plasmodium* telomeres: a pathogen's perspective. **Curr Opin Microbiol,** v. 4, n. 4, p. 409-14, Aug 2001. ISSN 1369-5274. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11495803 >.

SCHERF, A.; LOPEZ-RUBIO, J. J.; RIVIERE, L. Antigenic variation in *Plasmodium falciparum*. **Annu Rev Microbiol,** v. 62, p. 445-70, 2008. ISSN 0066-4227. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18785843 >.

SOUZA, W.; CARVALHO, T. M.; BARRIAS, E. S. Review on *Trypanosoma cruzi*: Host Cell Interaction. **Int J Cell Biol,** v. 2010, 2010. ISSN 1687-8884. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20811486 >.

TAJIMA, F. Unbiased estimation of evolutionary distance between nucleotide sequences. **Mol Biol Evol,** v. 10, n. 3, p. 677-88, May 1993. ISSN 0737-4038. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8336549 >.

TAJIMA, F.; NEI, M. Estimation of evolutionary distance between nucleotide sequences. **Mol Biol Evol,** v. 1, n. 3, p. 269-85, Apr 1984. ISSN 0737-4038. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/6599968 >.

TAMURA, K. *et al.* MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. **Mol Biol Evol,** v. 24, n. 8, p. 1596-9, Aug 2007. ISSN 0737-4038. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17488738 >.

TAMURA, K.; KUMAR, S. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. **Mol Biol Evol,** v. 19, n. 10, p. 1727-36, Oct 2002. ISSN 0737-4038. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12270899 >.

TEIXEIRA, S. M. *et al.* A differentially expressed gene family encoding "amastin," a surface protein of *Trypanosoma cruzi* amastigotes. **J Biol Chem,** v. 269, n. 32, p. 20509-16, Aug 1994. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8051148 >.

TODESCHINI, A. R. *et al.* Enzymatically inactive trans-sialidase from *Trypanosoma cruzi* binds sialyl and beta-galactopyranosyl residues in a sequential ordered mechanism. **J Biol Chem,** v. 279, n. 7, p. 5323-8, Feb 2004. ISSN 0021-9258. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14634017 >.

TODESCHINI, A. R. *et al.* Trans-sialidase from *Trypanosoma cruzi* catalyzes sialoside hydrolysis with retention of configuration. **Glycobiology,** v. 10, n. 2, p. 213-21, Feb 2000. ISSN 0959-6658. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10642613 >.

TONELLI, R. R. *et al.* Role of the gp85/trans-sialidases in *Trypanosoma cruzi* tissue tropism: preferential binding of a conserved peptide motif to the vasculature in vivo. **PLoS Negl Trop Dis,** v. 4, n. 11, p. e864, 2010. ISSN 1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21072227 >.

TONELLI, R. R. *et al.* In vivo infection by *Trypanosoma cruzi*: the conserved FLY domain of the gp85/trans-sialidase family potentiates host infection. **Parasitology,** v. 138, n. 4, p. 481-92, Apr 2011. ISSN 1469-8161. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21040619 >.

TURNER, C. W.; LIMA, M. F.; VILLALTA, F. *Trypanosoma cruzi* uses a 45-kDa mucin for adhesion to mammalian cells. **Biochem Biophys Res Commun,** v. 290, n. 1, p. 29-34, Jan 2002. ISSN 0006-291X. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11779128 >.

URBINA, J. A.; DOCAMPO, R. Specific chemotherapy of Chagas disease: controversies and advances. **Trends Parasitol,** v. 19, n. 11, p. 495-501, Nov 2003. ISSN 1471-4922. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14580960 >.

VARKI, A. et al. Essentials of Glycobiology. **Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.** ISBN-10: 0-87969-559-5, 1999. Disponível em: < http://www.ncbi.nlm.nih.gov/books/NBK20709 >

VARSHAVSKY, R.; HORN, D.; LINIAL, M. Global considerations in hierarchical clustering reveal meaningful patterns in data. **PLoS One,** v. 3, n. 5, p. e2247, 2008. ISSN 1932-6203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18493326 >.

VELGE, P. *et al.* Identification and isolation of *Trypanosoma cruzi* trypomastigote collagen-binding proteins: possible role in cell-parasite interaction. **Parasitology,** v. 97 (Pt 2), p. 255-68, Oct 1988. ISSN 0031-1820. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2849081 >.

WEATHERLY, D. B.; BOEHLKE, C.; TARLETON, R. L. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. **BMC Genomics,** v. 10, p. 255, 2009. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19486522 >.

YAN, T. *et al.* PatMatch: a program for finding patterns in peptide and nucleotide sequences. **Nucleic Acids Res,** v. 33, n. Web Server issue, p. W262-6, Jul 2005. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15980466 >.

YAO, C.; DONELSON, J. E.; WILSON, M. E. The major surface protease (MSP or GP63) of *Leishmania sp*. Biosynthesis, regulation of expression, and function. **Mol Biochem Parasitol,** v. 132, n. 1, p. 1-16, Nov 2003. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14563532 >.

YEO, M. *et al.* Origins of Chagas disease: Didelphis species are natural hosts of *Trypanosoma cruzi* I and armadillos hosts of *Trypanosoma cruzi* II, including hybrids. **Int J Parasitol,** v. 35, n. 2, p. 225-33, Feb 2005. ISSN 0020-7519. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15710443 >.

YOUNG, R. *et al.* Isolation and analysis of the genetic diversity of repertoires of VSG expression site containing telomeres from *Trypanosoma brucei gambiense*, *T. b. brucei* and *T. equiperdum.* **BMC Genomics,** v. 9, p. 385, 2008. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18700033 >.

JAIN, A. K. *et al.* Data clustering: A review. **ACM Computing Surveys**, 31(3), 1999.

ZINGALES, B. *et al.* A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. **Mem Inst Oswaldo Cruz,** v. 104, n. 7, p. 1051-4, Nov 2009. ISSN 1678-8060. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20027478 >.

ZINGALES, B. *et al.* Molecular epidemiology of American trypanosomiasis in Brazil based on dimorphisms of rRNA and mini-exon gene sequences. **Int J Parasitol,** v. 28, n. 1, p. 105-12, Jan 1998. ISSN 0020-7519. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9504338 >.

ZUCKERKANDL, E.; PAULING, L. Molecules as documents of evolutionary history. **J Theor Biol,** v. 8, n. 2, p. 357-66, Mar 1965. ISSN 0022-5193. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/5876245 >.

Anexo 1 - Alinhamento de proteínas TcMUC I e TcMUC II que apresentam características de ambos os grupos ou repetições não caracterizadas.

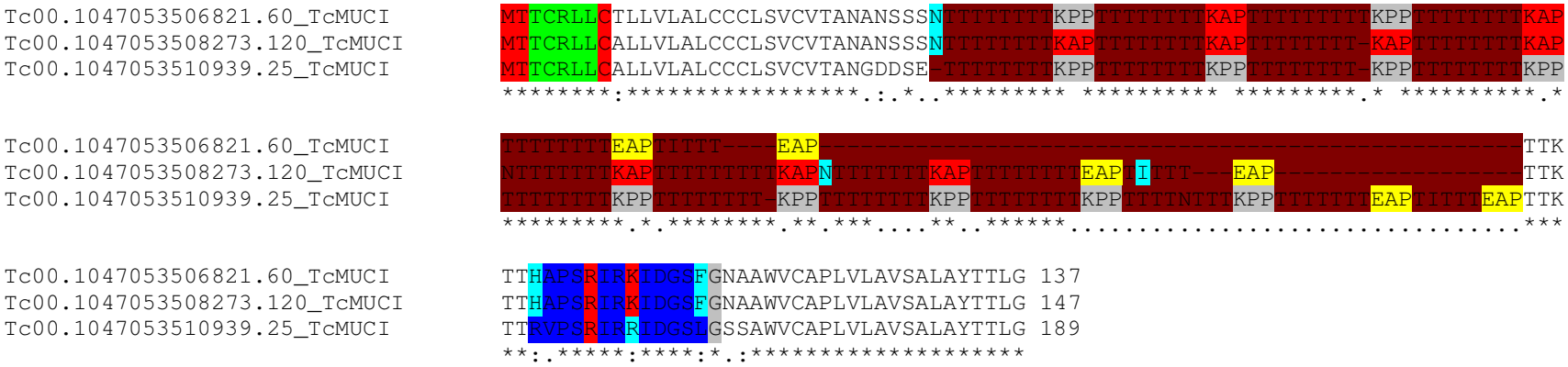
1. **Peptídeo sinal – Verde**
   MTTLTMMTCRLLY – cinza, região do peptídeo sinal típica de proteínas TcMUC I
   MMTTCRLLC – vermelho, região do peptídeo sinal típica de proteínas TcMUC II

2. **Repetições**
   T8KPP – cinza, repetição típica de proteínas TcMUC I
   T8KAP – vermelho, repetição típica de proteínas TcMUC II
   T8QAP – vermelho, repetição típica de proteínas TcMUC II
   T8EAP – amarelo, repetição não descrita

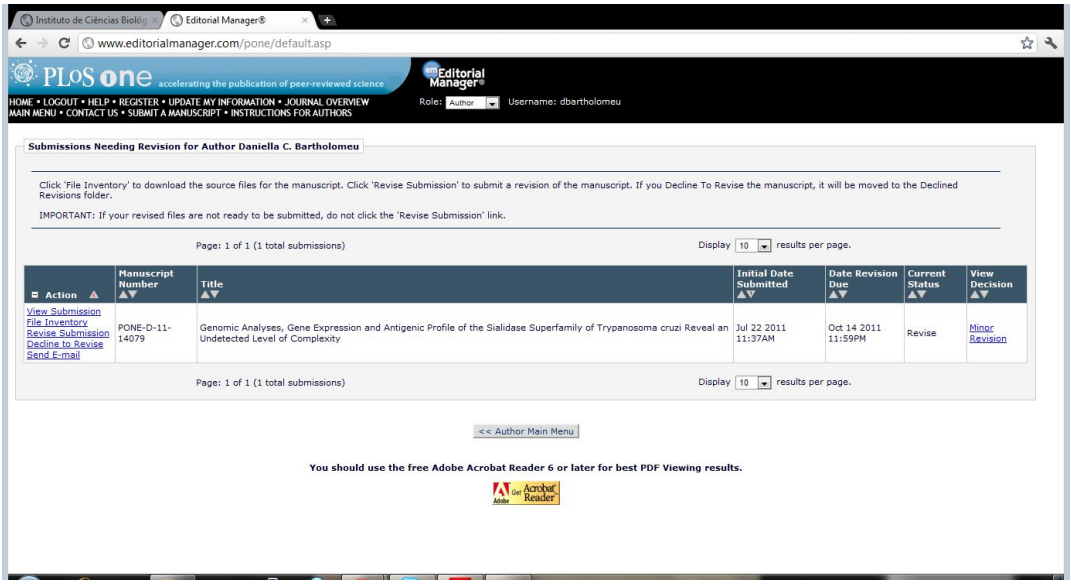
3. **Âncora GPI**
   RAPSSIRRSDGSLG – cinza, âncora GPI típica de proteínas TcMUC I
   RAPSRLRESDGSLS – vermelho, âncora GPI típica de proteínas TcMUC II

4. **Variações das repetições**

Alinhamento de duas proteínas classificadas como TcMUC I e uma proteína TcMUC II. As proteínas possuem características de TcMUC II na região do peptídeo sinal, apresentam repetições e âncora GPI com características tanto de TcMUC I quanto TcMUC II.



Anexo 2 - Artigo publicado na revista internacional “PLoS ONE”.



Genomic Analyses, Gene Expression and Antigenic Profile of the Sialidase Superfamily of *Trypanosoma cruzi* Reveal an Undetected Level of Complexity.

Freitas, Leandro M. ; dos Santos, Sara Lopes ; Rodrigues-Luiz, Gabriela F. ; Mendes, Tiago A. O. ; Rodrigues, Thiago S. ; Gazzinelli, Ricardo T. ; Teixeira, Santuza M. R. ; Fujiwara, Ricardo T. ; Bartholomeu, Daniella C. ; Rodrigues, Mauricio Martins

OPEN ACCESS Freely available online

PLoS one

Genomic Analyses, Gene Expression and Antigenic Profile of the Trans-Sialidase Superfamily of *Trypanosoma cruzi* Reveal an Undetected Level of Complexity

Leandro M. Freitas[1☯], Sara Lopes dos Santos[1☯], Gabriela F. Rodrigues-Luiz[1], Tiago A. O. Mendes[1], Thiago S. Rodrigues[2], Ricardo T. Gazzinelli[3], Santuza M. R. Teixeira[3], Ricardo T. Fujiwara[1], Daniella C. Bartholomeu[1]*

1 Departamento de Parasitologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, **2** Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Brazil, **3** Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Abstract

The protozoan parasite *Trypanosoma cruzi* is the etiologic agent of Chagas disease, a highly debilitating human pathology that affects millions of people in the Americas. The sequencing of this parasite's genome reveals that trans-sialidase/trans-sialidase-like (TcS), a polymorphic protein family known to be involved in several aspects of *T. cruzi* biology, is the largest *T. cruzi* gene family, encoding more than 1,400 genes. Despite the fact that four TcS groups are well characterized and only one of the groups contains active trans-sialidases, all members of the family are annotated in the *T. cruzi* genome database as trans-sialidase. After performing sequence clustering analysis with all TcS complete genes, we identified four additional groups, demonstrating that the TcS family is even more heterogeneous than previously thought. Interestingly, members of distinct TcS groups show distinctive patterns of chromosome localization. Members of the TcSgroupII, which harbor proteins involved in host cell attachment/invasion, are preferentially located in subtelomeric regions, whereas members of the largest and new TcSgroupV have internal chromosomal locations. Real-time RT-PCR confirms the expression of genes derived from new groups and shows that the pattern of expression is not similar within and between groups. We also performed B-cell epitope prediction on the family and constructed a TcS specific peptide array, which was screened with sera from *T. cruzi*-infected mice. We demonstrated that all seven groups represented in the array are antigenic. A highly reactive peptide occurs in sixty TcS proteins including members of two new groups and may contribute to the known cross-reactivity of *T. cruzi* epitopes during infection. Taken together, our results contribute to a better understanding of the real complexity of the TcS family and open new avenues for investigating novel roles of this family during *T. cruzi* infection.

Citation: Freitas LM, dos Santos SL, Rodrigues-Luiz GF, Mendes TAO, Rodrigues TS, et al. (2011) Genomic Analyses, Gene Expression and Antigenic Profile of the Trans-Sialidase Superfamily of *Trypanosoma cruzi* Reveal an Undetected Level of Complexity. PLoS ONE 6(10): e25914. doi:10.1371/journal.pone.0025914

Editor: Mauricio Martins Rodrigues, Federal University of São Paulo, Brazil

Received July 22, 2011; **Accepted** September 13, 2011; **Published** October 19, 2011

Funding: 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), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: daniella@icb.ufmg.br

☯ These authors contributed equally to this work.

Introduction

The protozoan parasite *Trypanosoma cruzi* is the etiologic agent of Chagas disease, a debilitating illness that is a major cause of morbidity and mortality in several Latin America countries. Approximately 10 million people carry the parasite, which causes 10,000 deaths annually [1]. During its life cycle, *T. cruzi* passes through three developmental stages. In its insect vectors, the parasite multiplies as extracellular epimastigotes, and in the hindgut it differentiates into non-dividing trypomastigotes. These infective forms are excreted in the feces after a blood meal and may contaminate the puncture site or mucous membranes of a mammalian host, where they can invade a variety of cell types. Inside host cells, trypomastigotes differentiate into amastigotes, which, after a limited number of cell divisions, differentiate into trypomastigotes that are released into circulation upon host cell rupture. This form can then infect another mammalian host cell or be taken by the insect vector during the blood meal, where it differentiates as epimastigotes.

The ability of *T. cruzi* to survive in the mammalian host is in part due to the presence of a diverse surface membrane coat. In fact, a remarkable feature of the *T. cruzi* genome is the massive expansion of genes that encode polymorphic surface proteins, which include the trans-sialidase and trans-sialidase like superfamily (hereafter called TcS), MASP (mucin-associated surface protein), and TcMUC mucins [2]. The TcS is the largest *T. cruzi* gene family, which has more than 1,400 genes, half of which are apparently functional. One of the most well-studied members of the TcS superfamily is the trans-sialidase (TcTS) enzyme. *T. cruzi* is unable to synthesize sialic acids de novo [3], a sugar modification present in

T. cruzi proteins implicated in several key aspects of the *T. cruzi*-host interaction. The sialylation of the parasite surface is possible due to the activity of a modified sialidase that, instead of hydrolyzing sialic acid, transfers alpha (2–3)-linked sialyl residues from sialoglycoconjugates and proteins from the host to the parasite cell-surface mucin proteins (TcMUC) [4–6]. The rapid sialylation of TcMUC proteins upon cell rupture confers a negatively charged coat that protects the extracellular trypomastigotes from being killed by human anti-alpha galactosyl antibodies [7].

The TcS gene family is highly polymorphic, and only a few members have critical residues necessary for catalytic activity [8]. So far, four groups of TcS have been described based on sequence similarity and functional properties. Group I contains active trans-sialidases, namely TCNA and SAPA (shed acute-phase antigen), and TS-epi proteins expressed in the trypomastigote and epimastigote forms, respectively. Group II comprises members of the gp85 surface glycoproteins TSA-1, SA85, gp90, gp82 and ASP-2, which have been implicated in host cell attachment and invasion. FL-160, a representative of group III, is a complementary regulatory protein that inhibits the alternative and classical complement pathways. TsTc13, whose function is unknown, is the representative of group IV and is included in the TcS superfamily because it contains the conserved VTVxNVxLYNR motif, which is shared by all known TcS members [8–13].

The TcS family was identified in the 1980s and, after the publication of the *T. cruzi* genome [2], no comprehensive analysis of its sequences has been performed. Here, by analyzing all the full-length predicted TcS proteins present in the *T. cruzi* genome, we identified four new groups. The TcS groups were characterized based on presence of key TcS motifs, chromosomal localization, expression profile and antigenic properties. Implications of the TcS diversity for *T. cruzi* biology are discussed.

Materials and Methods

Sequence diversity of the *T. cruzi* TcS family

Genome information and sequences were retrieved from TriTrypDB (http://TriTrypDB.org). Only complete TcS sequences totaling 508 sequences were analyzed. The DNA and the translated sequences were aligned using ClustalW 2.0 software with the default parameters [14]. These alignments were used to calculate the total (mean) nucleotide and protein diversity using MEGA4 [15] with three different methods: p-distance (nucleotide and protein sequences), Kimura-2-parameter (nucleotide sequences) and Poisson correction (protein sequences). The diversity error was estimated using bootstrap resampling with 1,000 replications.

Spatial projection and hierarchical clustering

To identify the clusters formed by the TcS protein sequences and by the 3′ sequences flanking the TcS coding regions (300 nucleotides downstream to the stop codons), we calculated the pairwise distance and generated the distance matrixes. The distances between the sequences were generated using the package PHYLIP [16,17]. To provide a visual representation of each distance matrix, we used the multidimensional scaling (MDS) plot with two dimensions (2D). The K-means method [18] was used to define ten clusters. The MDS, hierarchical clustering, statistical analyses and graphing were performed using the R software platform [19].

TcS cluster distribution on *T. cruzi* chromosomes and protein representation

To define the chromosomal distribution of the TcS groups, we used as reference the genome assembly reported in [20], where pairs of homologous chromosomes were arbitrarily built as having the same size. The chromosomal coordinates of the TcS genes, regardless from each homologous chromosome they are derived, were retrieved from the TriTrypDB (http://TriTrypDB.org) and plotted on the chromosomes. The colors of each coding region were the same as the colors used in the MDS protein clusters. The relative positions in the chromosomes were calculated by dividing the start codon coordinate of each gene by the total length of the chromosome. The values found were used to produce a histogram and to compare the distribution of each cluster and the pseudogenes on the chromosomes.

FRIP coordinates were found using the motif xRxP as a query. Only those occurrences located before the Asp-box and/or closer to the N-terminal extremity were considered. The Asp-box was found using the motif SxDxGxTW as a query, allowing up to 1 mismatch, and the TcS signature motif was searched using the VTVxNVxLYNR sequence as a query. In all query motifs, x represents any amino acid. The motifs were searched using the software PatMatch [21]. The signal peptide and the GPI anchor additional site were predicted using the software SignalP [22] and GPI-SOM [23], respectively. Repetitive sequences were identified using the AA-repeatFinder developed by our group (http://gicab.decom.cefetmg.br/bio-web). Only repeats with more than 10 amino acids were reported. The figures depicting the TcS genome distribution and the protein sequences were constructed using Perl (Practical Extraction and Report Language) scripts and the Bio::Graphics module, part of the Bioperl toolkit (http://www.bioperl.org).

Parasite cultures and RNA extraction

Epimastigotes of the CL Brener clone of *T. cruzi* were maintained in a logarithmic growth phase at 28°C in liver infusion tryptose (LIT) medium supplemented with 10% fetal bovine serum. Amastigote and trypomastigote forms were obtained from infected L6 cells grown in Dulbecco's Modified Eagle Medium supplemented with 5% fetal bovine serum, at 37°C and 5% CO_2, as described [24]. Total RNA was isolated using the RNeasy kit (Qiagen).

Real-time RT-PCR

Primers specific for each cluster were designed using Allele ID 7 (Premier Biosoft, Demo version), and the primer specificity was verified using e-PCR and the entire parasite genome as a template. The primers selected are listed in Table S1. Real-time PCR reactions were performed in an ABI 7500 sequence detection system (Applied Biosystems). Reactions in triplicate were prepared containing 1 mM forward and reverse primers, SYBR Green Supermix (Bio-Rad), and each diluted template cDNA. Standard curves were performed for each experiment for each pair of primers using serially diluted *T. cruzi* CL Brener genomic DNA and were used in the calculation of the relative quantity (Rq) values for each sample. qRT-PCRs for the constitutively expressed GAPDH gene were performed to normalize the expression of the TcS genes. Results were analyzed with an ANOVA test, and graphics were constructed in GraphPad Prism 5.0 (GraphPad Inc.).

Epitope prediction, spot peptide array and immunoblot

The 508 complete TcS proteins were submitted for linear B-cell epitope prediction using the Bepipred algorithm [25]. Peptides with 15 amino acids and with prediction scores above 1.3 were selected. Peptides with 70% identity over 70% of the peptide length with *T. cruzi* proteins other than TcS were excluded. For synthesis, we selected those peptides with higher occurrences within a group and with higher prediction scores. The peptides

synthesized are listed in Table S2. The peptides were covalently synthesized in pre-activated cellulose membranes according to the SPOT synthesis technique [26]. Membranes were blocked with 5% BSA and 4% sucrose in PBS and were incubated for one hour and 30 minutes with diluted mice sera (1:500) in blocking solution. After washing, the membrane was incubated with secondary antibody IgG (Sigma) diluted to 1:2000 in blocking solution and, after a second washing, revealed by *ECL Plus Western blotting* (GE Healthcare). The spots were visualized by fluorescence scanning. The membrane was submitted to the same experimental conditions using sera from uninfected mice. Densitometry measures and analysis of each peptide was performed using Image Master Platinum (GE), and the relative density (Rd) cut-off for positivity was determined as 2.0. Graphics were constructed in GraphPad Prism 5.0 (GraphPad Inc.).

Ethics Statement

All animal procedures were approved by the animal care ethics committee of the Federal University of Minas Gerais (Protocol # 143/2009).

Results

Sequence clustering reveals eight groups of the trans-sialidase/trans-sialidase-like superfamily (TcS) of *T. cruzi*

Despite the fact that four TcS groups were previously described [8,13,27], and only one group corresponds to the active trans-sialidase proteins, a much larger number of members of this gene family was annotated in public databases as trans-sialidases. To sort out which members correspond to the previously defined groups and to eventually identify new groups, we performed cluster analysis on all predicted TcS proteins identified in the CL Brener genome, excluding those annotated as partial and/or pseudogenes. A total of 508 TcS members were used to perform pairwise alignments resulting in a distance matrix that was used to generate a multidimensional scaling (MDS) plot (Figure 1). K-means method was then used to define ten clusters or groups (Figure 1A). Clustering with larger numbers of groups resulted in the fragmentation of previous clusters, without shuffling the members among them, indicating the robustness of the clustering of the family in ten groups (data not shown). Three members were located far from the others in the spatial distribution and therefore are the most divergent members of the family. One of them, Tc00.1047053505699.10, is the only representative of the group shown in black, and the Tc00.1047053509265.120 and Tc00.1047053507699.230 formed the brown group. Manual inspection of these three proteins revealed that their N-terminal regions are longer or shorter compared to the other TcS sequences: Tc00.1047053505699.10 contains an extra 260 amino acids at its N-terminal, whereas Tc00.1047053509265.120 and Tc00.1047053507699.230 have a deletion of approximately 160 and 450 amino acids, respectively, in their N-terminal region. The truncated sequences of these two proteins were due to the location of these genes in contig ends. Because gene prediction regarding the initial start codon could not be corrected for these three anomalous sequences, both black and brown groups were excluded from further analysis. The list of proteins belonging to each group is available in the supporting material (Table S3).

Protein and DNA sequences of the eight remaining groups were then aligned and the intra-cluster diversity was calculated using the p-distance, the Kimura-2-parameter and the Poisson correction methods, as described in the material and methods section. The groups are formed from different numbers of members and show distinct diversity indexes (Table 1). Groups labeled in red and dark green are the largest groups, with 227 and 117 members, respectively, totaling 68% of the TcS members. No clear correlation between the number of members and the diversity indexes was found. For instance, small groups (blue and orange) have similar diversity indexes of the largest ones (Table 1).

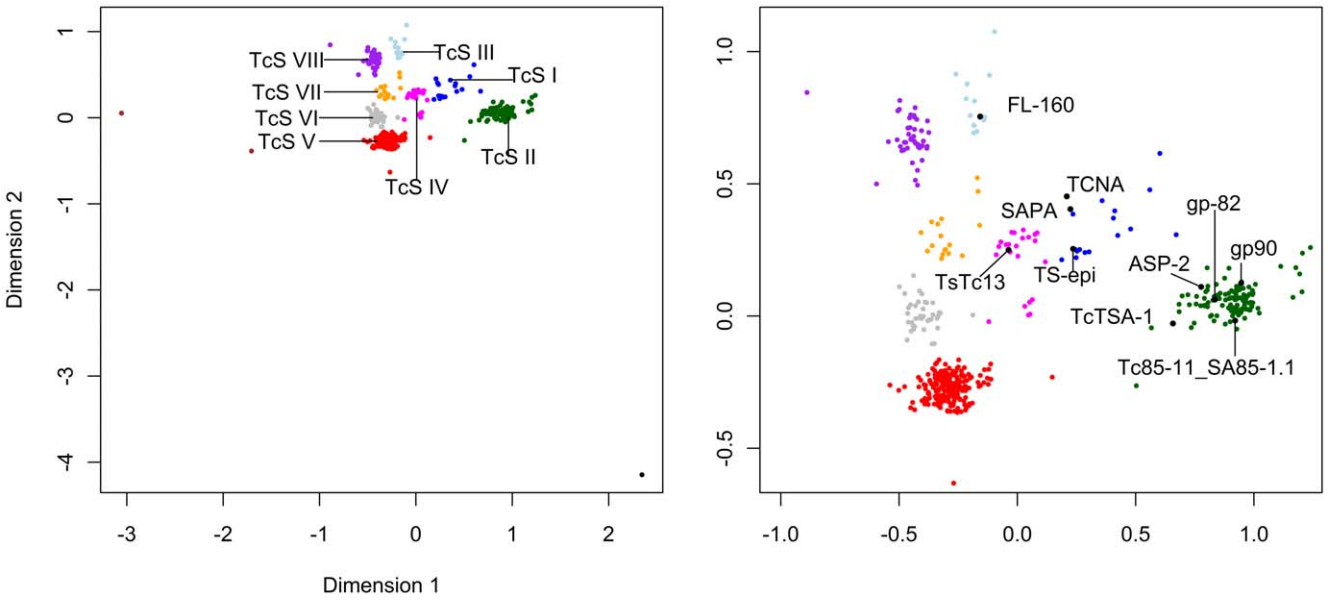


Figure 1. Multidimensional scaling (MDS) plot of the TcS protein sequences. The pairwise alignments of the 508 TcS complete members were performed and the distance matrix was used to generate a multidimensional scaling (MDS) plot. K-means method was used to define the clusters or groups. (A) Pattern of dispersion of all 508 TcS protein sequences resulting in 10 TcS groups. (B) Pattern of dispersion of 505 TcS protein sequences in eight TcS groups. Previously characterized TcS sequences were mapped on the MDS. TcSgroupI - blue; TcSgroupII - dark green; TcSgroupIII - light blue; TcSgroupIV - magenta; TcSgroupV - red; TcSgroupIV - gray; TcSgroupVII - orange and TcSgroupVIII - purple.
doi:10.1371/journal.pone.0025914.g001

Table 1. Diversity indexes of nucleotide, protein and 3′UTR sequences of the TcS family.

		DNA		Protein	
	Number of members	p-distance	K2p	p-distance	Poisson correction
TcSgroupI Blue	19	0.371/0.004	0.690/0.029	0.494/0.009	0.881/0.027
TcSgroupII Dark green	117	0.264/0.004	0.340/0.007	0.419/0.010	0.558/0.018
TcSgroupIII Light blue	15	0.209/0.005	0.263/0.008	0.366/0.010	0.492/0.023
TcSgroupIV Magenta	25	0.179/0.003	0.226/0.005	0.250/0.008	0.320/0.012
TcSgroupV Red	227	0.252/0.004	0.316/0.006	0.396/0.009	0.513/0.015
TcSgroupVI Gray	39	0.246/0.004	0.312/0.007	0.394/0.009	0.513/0.016
TcSgroupVII Orange	17	0.298/0.004	0.425/0.009	0.448/0.009	0.651/0.020
TcSgroupVIII Purple	46	0.215/0.004	0.270/0.006	0.353/0.009	0.453/0.013
TcS family	508	0.413/0.004	0.662/0.011	0.574/0.090	0.912/0.023
3′UTR	495	0.573/0.007	1.086/0.029	-	-

P-distance was used to measure the diversity of the coding regions, proteins and 3′ flanking sequences, with kimura-2-parameters and Poisson correction only to DNA coding and protein sequences, respectively.
doi:10.1371/journal.pone.0025914.t001

We next mapped on the MDS plot the TcS proteins representative from each of the four previously known groups (Figure 1B). As expected, the characterized TcS members mapped into different MDS clusters. TCNA, SAPA and TS-epi, all active trans-sialidase proteins belonging to the previously defined group I, clustered together in the blue group (hereafter named TcSgroupI). From a total of 19 TcSgroupI members, 11 have the critical catalytic residues (Figure S1). GP82, GP90, Tc85-11_SA85-1.1 and ASP-2, all representatives of the previously defined group II, mapped onto the dark green cluster (TcSgroupII). Finally, FL-160 and Ts13, which belong to sialidase groups III and IV, mapped onto the light blue (TcSgroupIII) and magenta (TcSgroupIV) clusters, respectively. None of the TcS proteins previously characterized mapped onto the clusters that are red (the largest TcS group), gray, orange or purple, hereafter named TcSgroup V, VI, VII and VIII, respectively.

Identifying key sialidase signature motifs in the eight MDS clusters

To characterize each of the eight groups, we initially searched for all the key signature motifs as they are described in the literature [8,13] and mapped them into the MDS plot (Figure S2). The canonical VTVxNVxLYNR motif was found in only 328 of 508 TcS sequences used in this study. This result prompted us to investigate whether the other proteins annotated as TcS have a degenerate form of this motif or do not have this motif at all. To this end, we performed ClustalW alignment of all 508 TcS proteins and retrieved the alignment block containing this motif. Visual inspection of this region reveals that 159 sequences have a degenerate version of this motif. Hence, 487 (96%) of the TcS sequences have the canonical or degenerate forms of the VTVxNVxLYNR motif. The remaining sequences do not contain this motif because they have a truncated C-terminal region resulting from premature stop codons and/or frameshifts. Therefore, as previously described, this motif is a signature of the TcS family that is found in all its members. As shown in Figure 2, although variations on the VTVxNVxLYNR motif are observed, the motif is highly conserved within each cluster.

We also searched for the Asp box motif found in bacterial and viral sialidases [28], using as query the SxDxGxTW sequence, where x is any amino acid. A total of 135 sequences have this motif, of which 133 belong to the previously described TcS groups I (blue), II (dark green) and IV (magenta) (Figure 2). The two other sequences having this motif belong to the TcSgroupV (red) and TcSgroupVI (gray). This result is in agreement with previous reports showing that TcSgroupIII (light blue) does not have this motif [29]. To investigate whether TcSgroupIII as well as sequences from the other four new groups have a degenerate form of the Asp box, we searched for a degenerated version of this motif, as described in the material and methods section. This search increased the number of positive Asp box sequences to 383. Only one additional degenerate position was found, resulting in the consensus SxDxGxxW. Although the majority of them have one (220) or two Asp boxes (154), a few (9) have three. Considering this new consensus motif, the Asp box is found in a large majority of the members from TcSgroupI (blue, 17 of 19 members), TcSgroupII (dark green, 114/117), and TcSgroupIV (magenta, 24/25) and is also present in the new groups TcSgroupV (red, 188/227) and TcSgroupVI (gray, 36/39). On the other hand, as previously described, it is missing in TcSgroupIII (lightblue) and has only a few occurrences in the new groups TcSgroupVII (orange, 1/17) and TcSgroupVIII (purple, 3/46) (Figure 2).

The FRIP motif was searched using the pattern xRxP (where x is any amino acid). Because this is a small and degenerate sequence, we considered only those occurrences that are before the Asp-box and/or closest to the N-terminal region [30]. A total of 205 TcS proteins contain the FRIP motif, which is found in the majority of the members of TcSgroupI (blue, 68%), TcSgroupIII

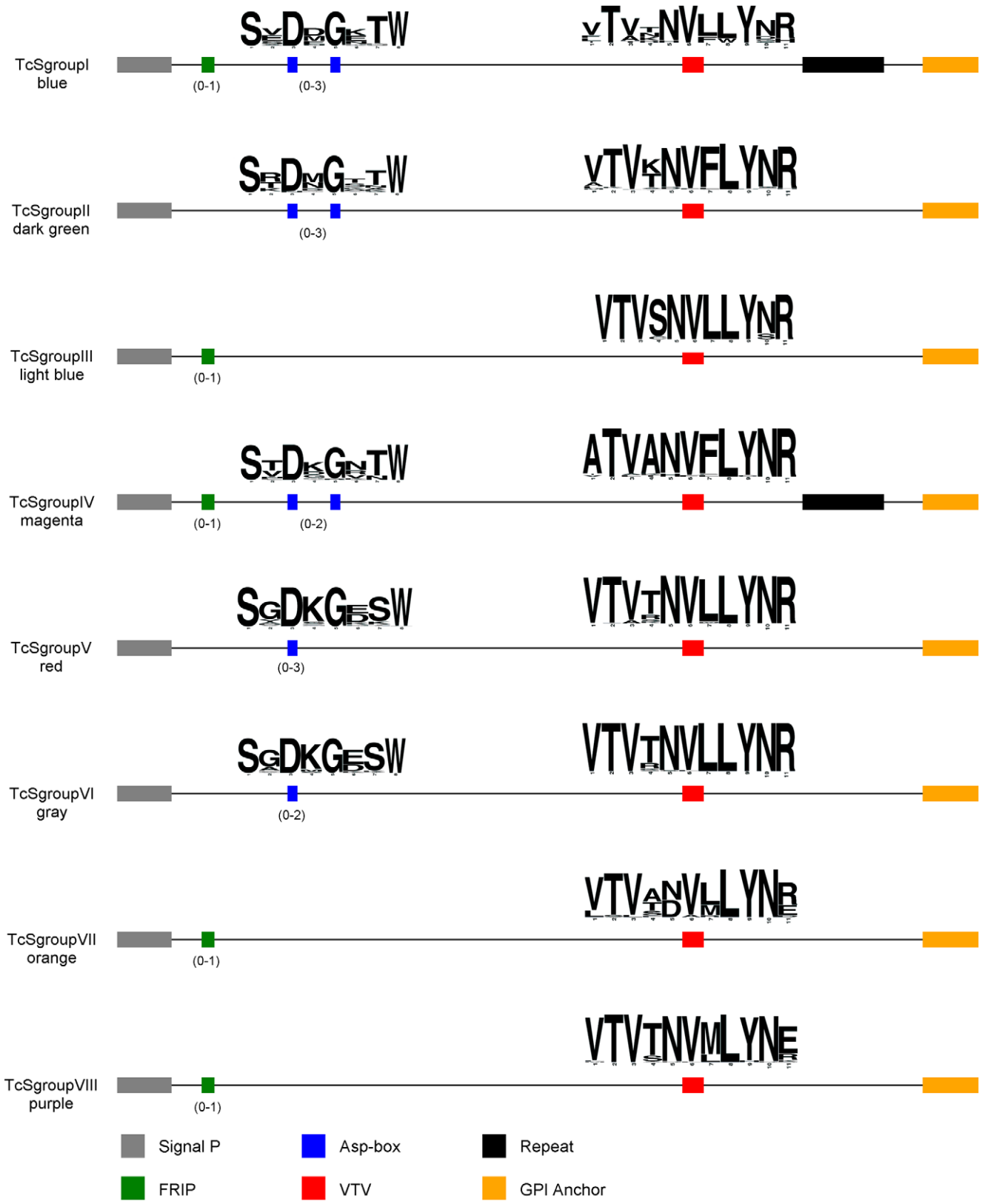


Figure 2. Prototype of each TcS group. The motifs are shown only when they occur in the majority of the proteins within the group. The Asp-box and VTVxNVxLYNR *logos* are shown above each motif. The numbers within parentheses indicate the number of occurrences of a given motif. The length of the proteins within the groups may vary. Graphical representations are not to scale.
doi:10.1371/journal.pone.0025914.g002

(light blue, 87%), TcSgroupIV (magenta, 88%), TcSgroupVII (orange, 76%) and TcSgroupVIII (purple, 87%).

To identify repetitive regions on TcS sequences, we used the AA-repeat finder program (http://gicab.decom.cefetmg.br/bioweb). Only repeats with more than 10 amino acids were considered. We found that repeats are more frequent in the TcSgroupI (blue) and TcSgroupIV (magenta) clusters. These two groups have the largest repetitive regions, which encompass up to 884 amino acids. In fact, although we identified new repeats in these two groups, the largest repeats are those corresponding to the known DSSAH(S/G)TPSTP(A/V) repeat found in TS SAPA and the TcTs13 EPKSA-repeat. On the other hand, TcSgroupV (red), TcSgroupVI (gray) and TcSgroupVII (orange) groups have only 1, 2.5 and 6% of their members, respectively, with repetitive domains whereas no repeat was found in members of the TcSgroupIII (light blue). All repeats identified in this study are shown in Table S4.

In the prototype representation of the eight TcSgroups shown in Figure 2, it is possible to identify three patterns of motif occurrence. The TcSgroupI (blue) and TcSgroupIV (magenta) clusters have the most complex structure, with the FRIP, Asp box and VTVxNVxLYNR motifs and the C-terminal repeats, although the sequences of the VTVxNVxLYNR motif and the C-terminal tandem repeats are distinct. TcSgroupII (dark green), TcSgroupV (red) and TcSgroupVI (gray) clusters contain the Asp box and VTVxNVxLYNR motifs. TcSgroupIII (light blue), TcSgroupVII (orange) and TcSgroupVIII (purple) clusters only have the FRIP and VTVxNVxLYNR motifs, which have a consensus sequence that is group-specific. This pattern of motif occurrence is in agreement with the space distribution of the TcS groups in the MDS (Figure 1). TcS groups I and IV that have all motifs are centered in the MDS, whereas TcSgroups II, V and VI are clustered in the bottom and TcSgroups III, VII and VIII are clustered in the left top region. A graphical representation for each of the 508 TcS proteins can be found in Figure S3.

Mapping the TcS groups on *T.cruzi* chromosomes

It is known that TcS genes can be found in *T. cruzi* subtelomeric regions or in internal positions in the chromosomes that are associated with other genes that encode surface proteins [2]. Subtelomeric regions are defined here as sequences extending from the telomeric hexamer repeats to the first nonrepetitive sequence. We investigated whether there is any bias on the chromosome localization of the TcS clusters. Figure 3 shows the chromosomal distribution of the TcS groups. A total of 60 complete TcS genes (not including partial or pseudogenes) can be found associated with the subtelomeric regions. One of them belongs to the brown cluster, which, as mentioned above, was excluded from our analysis. The majority of the subtelomeric TcS genes (36 members, 61%) belongs to TcSgroupII (dark green), 7 members from TcSgroupIV (magenta) and 10 from TcSgroupVIII (purple) (Figures 3 and 4). No TcSgroupIII (light blue) or TcSgroupVI (gray) genes are located at these regions. Interestingly, with one exception, all members of the largest TcS cluster (TcSgroup V, red) are at internal locations in the chromosomes (Figures 3, 4A and 4B). We have also found that the subtelomeric regions are enriched for TcS pseudogenes (Figure 4C), which is in agreement with the hypothesis that these regions were subject to intense rearrangement [54].

Expression profile of the TcS genes belonging to distinct groups

To characterize the expression profile of TcS genes belonging to distinct groups, we have performed real-time RT-PCR using member-specific primers, designed as described in the material and methods section. The expression of 12 TcS genes derived from six TcS groups was evaluated throughout the three parasite developmental stages using GAPDH mRNA levels, whose expression is constitutive throughout the parasite life cycle, for internal normalization (Figure 5). As a control, we used primers to amplify the cDNAs from the alpha-tubulin and amastin genes, whose mRNA levels we have previously shown to be up-regulated in epimastigotes and amastigotes, respectively [24,31]. The majority of the TcS transcripts are expressed in trypomastigotes and/or amastigote forms. Interestingly, within a group, the expression profile may be highly variable. For example, the TcS5 gene that belongs to TcSgroupII is highly expressed in trypomastigote forms, whereas the TcS27 from the same group shows a much lower level of expression in the trypomastigote and amastigote forms and is barely detected in epimastigotes. Also, TcS9 and TcS33 from TcSgroupIV are more expressed in trypomastigotes and amastigotes; however, TcS34, which is from the same group, is scarcely expressed in all the development stages. The new groups also display a variable expression profile. A very low level of expression was verified for the two genes analyzed from TcSgroupV in all the developmental stages (Figure 5) as well as in the blood trypomastigotes (data not shown). On the other hand, the gene TcS32 from TcSgroupVII is more expressed in the trypomastigotes. The two members of TcSgroupVIII show a variable expression profile, with TcS24 more expressed in trypomastigotes and TcS25 more expressed in amastigotes.

Analyzing sequence conservation of the 3′ flanking region of TcS groups

It is well established that, in Trypanosomatids, the 3′UTR regions are involved in post-transcriptional control mechanisms that confer stage-specific gene expression. To investigate whether the 3′ flanking sequences of TcS genes that belong to the same groups are conserved, we performed pairwise alignments of the 300 nt downstream of the stop codon of the TcSs, and the distance matrix was used to generate the MDS projection. We decided to analyze 300 nt downstream from the stop codon because this is the mean average length of the *T. cruzi* 3′UTRs [32]. The sequences were then color-coded according to the protein clusters showed in Figure 1. TcS genes already characterized as well as those genes whose expression levels were analyzed by real-time RT-PCR (Figure 5) were then mapped onto the MDS projection (Figure 6). We could not find a very clear association between the protein and the 3′ flanking region distances. For example, members of the TcSgroupV (red) form a robust cluster at the protein level and are much more variable according to the analysis of the 3′ flanking region. Also, the 3′ flanking regions of TcSgroupII (dark green) members are scattered in three MDS areas. On the other hand, the 3′ flanking regions of the TcSgroupVIII (purple) members clustered together, which suggests that similar mechanisms may control the expression of some of their genes. Interestingly, the 3′ flanking regions of SAPA and TCNA, both active trans-sialidase enzymes expressed in the trypomastigote forms (TcSgroupI), are clustered very close. Also, the 3′ flanking region of the TS-epi, an active trans-sialidase that is expressed in the epimastigote stage that also belongs to TcSgroupI, is located farther away from the SAPA and TCNA sequences. Moreover, the 3′ flanking region of gp90 and gp82, both expressed in the metacyclic trypomastigotes, and ASP-2, expressed in the amastigote stage, all belong to TcSgroupII and are very close in the MDS projection. Interestingly, although Tc85-11_SA85-1.1 and TsTc13 are expressed in the trypomastigote stage, they are divergent at the protein level (Figure 1) and belong to different TcS groups (II and

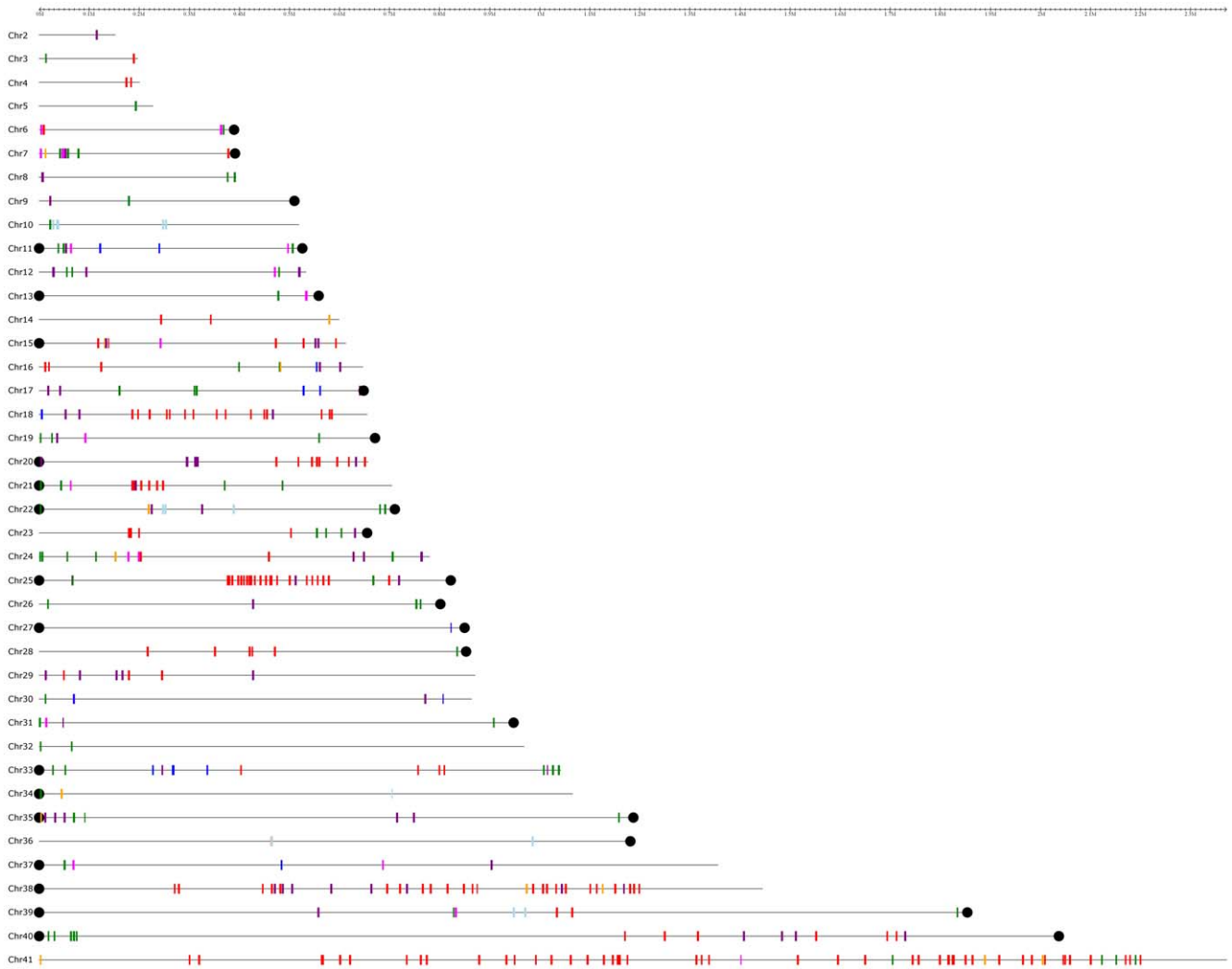


Figure 3. Mapping of TcS genes on *T. cruzi* chromosomes. Each CL Brener chromosome is comprised of 2 homologous chromosomes as proposed by [20]. The genes are color coded according to the color of the corresponding clusters of Figure 1. A total of 374 TcS genes could be mapped on the chromosomes. The remaining genes belong to contigs that could not be assigned to a specific chromosome, according to Weatherly et al., 2009, and are not represented in the figure. Only chromosomes containing TcS genes are shown. Black dots represent telomeric repeats. TcSgroupI - blue; TcSgroupII - dark green; TcSgroupIII - light blue; TcSgroupIV - magenta; TcSgroupV - red; TcSgroupIV - gray; TcSgroupVII - orange and TcSgroupVIII - purple.
doi:10.1371/journal.pone.0025914.g003

IV, respectively); they have similar 3′ flanking regions, which suggests that similar mechanisms for gene regulation may act on both genes.

Antigenicity of the TcS groups

Because the antigenicity of some members of the sialidase family was already reported [33,34], we decided to investigate whether other peptides derived from the TcS family are also antigenic. To this end, we have performed linear B-cell epitope prediction on all 508 complete members of the TcS family. A total of 40 peptides with 15 residues, high prediction scores and high occurrences within the TcS group were synthesized in a solid support by the spot synthesis technique and screened with sera from animals infected with *T. cruzi*. The list of all peptides used in this study is shown in the Table S2. As shown in Figure 7, 11 TcS peptides derived from distinct groups displayed antigenic properties based on a cut-off signal well above background. In agreement with previous studies, peptides corresponding to the SAPA (D5 and D8) [33] and to the TsTc13 repeats (B5) [34] are highly antigenic. We have also identified new epitopes specific to the previously characterized TcSgroups I and IV (D9 and D10, and B10, respectively). At least one peptide from each of the new TcSgroups -V, VI, VII and VIII - was recognized by sera of infected animals (A1, C3, A10 and B4, and A5, respectively). The peptide C3 occurs in the largest number of members (60 in total) from the new TcS groups V and VI and from the previously characterized TcS groups II and IV.

Discussion

The TcS superfamily, the largest *T. cruzi* multigene family [2], was described more than 20 years ago and, after the *T. cruzi* genome release, no comprehensive analysis of the diversity of this gene family was reported. Here, by analyzing all the 508 TcS complete genes present in the *T. cruzi* CL Brener genome [2], we demonstrated that this family displays an even greater variability

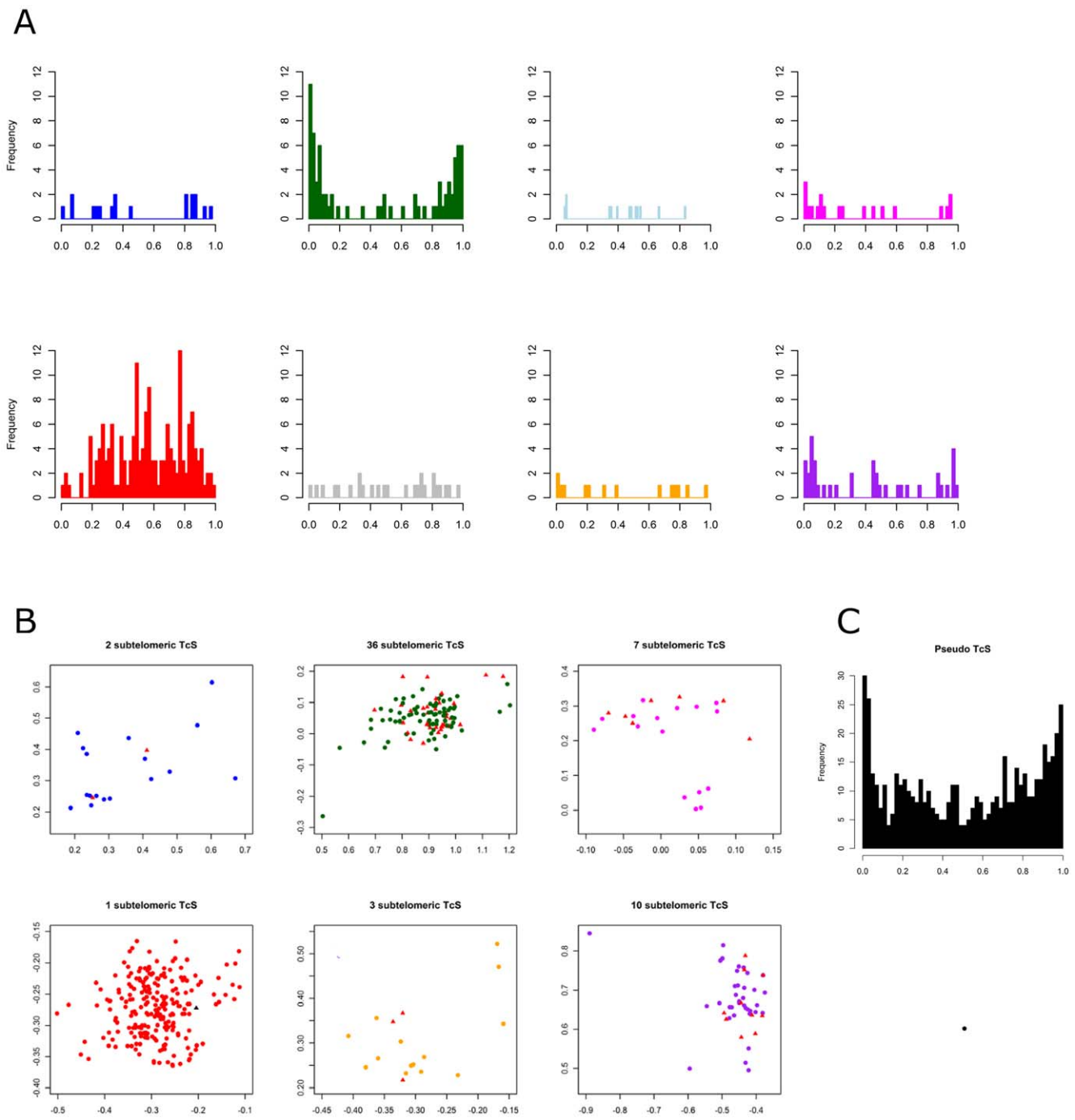


Figure 4. Distribution of each TcS group along the *T. cruzi* chromosomes. (A) Histograms showing the frequency of the TcS genes along the chromosomes. The length of all chromosomes was normalized as 1. The relative position of each gene was calculated by dividing the coordinate of the first nucleotide of the open reading frame by the length of the chromosome. (B) Representation of each group in Figure 1, showing the genes that localize in telomeric regions (black dots for the TcSgroupV and red dots for the other TcS groups). TcSgroupI - blue; TcSgroupII - dark green; TcSgroupIII - light blue; TcSgroupIV - magenta; TcSgroupV - red; TcSgroupIV - gray; TcSgroupVII - orange and TcSgroupVIII - purple. (C) Histogram showing the distribution of TcS pseudogenes along the *T. cruzi* chromosomes.
doi:10.1371/journal.pone.0025914.g004

than previously thought, as shown by means of the diversity indexes and the MDS projection. Based on their pattern of dispersion, we identified eight groups of TcS sequences, four of which were never described before (Figure 1). The distances among the clusters are consistent with the level of similarity and function of the previously described TcS sequences. All proteins that display trans-sialidase activity clustered together (TcSgroupI, blue). Another cluster was formed (TcSgroupII, dark green) from TcS proteins that have no trans-sialidase activity but that are capable of binding to β-galactose, laminin [35], fibronectin [36], collagen [37,38], and cytokeratin [39] and are involved in cell adhesion and invasion. The third TcS group encompasses proteins

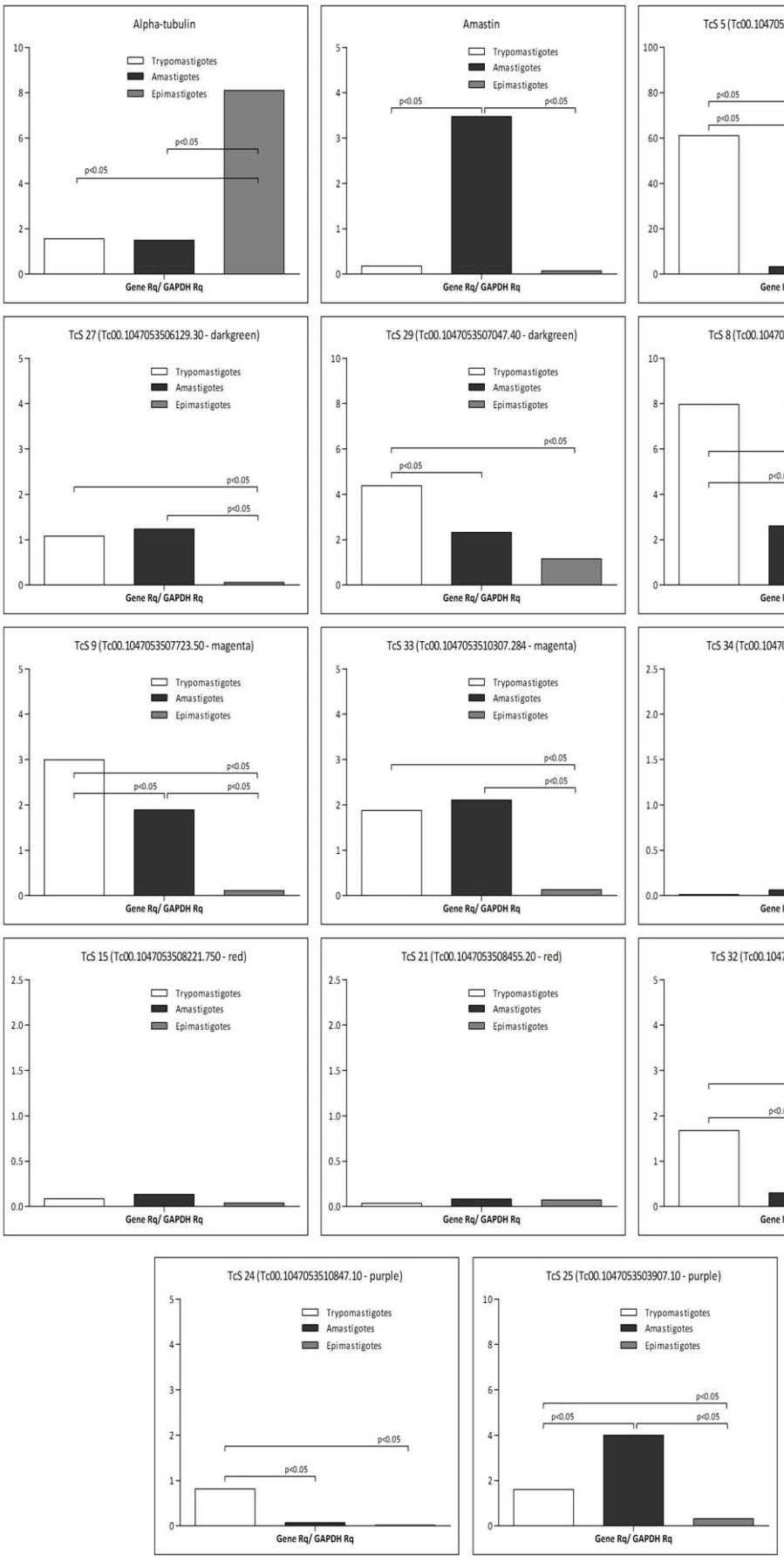

Alpha-tubulin
Amastin
TcS 5 (Tc00.1047053506471.120 - darkgreen)
TcS 27 (Tc00.1047053506129.30 - darkgreen)
TcS 29 (Tc00.1047053507047.40 - darkgreen)
TcS 8 (Tc00.1047053511911.60 - lightblue)
TcS 9 (Tc00.1047053507723.50 - magenta)
TcS 33 (Tc00.1047053510307.284 - magenta)
TcS 34 (Tc00.1047053510013.50 - magenta)
TcS 15 (Tc00.1047053508221.750 - red)
TcS 21 (Tc00.1047053508455.20 - red)
TcS 32 (Tc00.1047053509377.20 - orange)
TcS 24 (Tc00.1047053510847.10 - purple)
TcS 25 (Tc00.1047053503907.10 - purple)
Trypomastigotes
Amastigotes
Epimastigotes
p<0.05
Gene Rq/ GAPDH Rq

Figure 5. Expression profile of TcS genes by qRT-PCR. Relative quantity (Rq) calculations were based on specific standard curves for each TcS gene. Rq values of each cDNA sample (TcS Rq) were normalized with the GAPDH gene (GAPDH Rq), a gene constitutively expressed throughout the parasite life cycle. Alpha-tubulin and amastin were used as controls for genes more expressed in epimastigote and amastigote stages.
doi:10.1371/journal.pone.0025914.g005

involved in the regulation of the complement system (CRP - complement regulatory proteins). Previously characterized members of this group are the CRPs [29,40], which include the FL-160 [41]. Recently, using data from the *T. cruzi* CL Brener genome project, Beucher and Norris (2008) identified CRP paralogs based on sequence similarity with a functional characterized CRP (GenBank accession number AAB49414). Also, these authors divided the CRPs into two groups, HSG (high similarity group, with more than 80% identity with AAB49414) and LSG (low-similarity group, with sequence identity between 54 and 62% with AAB49414) [29]. Here we could verify that all HSGs, and excluding two exceptions, the LSGs, fell into TcSgroupIII (light blue). These two members, which do not belong to TcSgroupIII, were clustered within TcSgroupVII (orange). In fact, they are the two most divergent sequences of the LSG subgroup [29] and correspond to members of the TcSgroupVII that are closest to the TcSgroupIII (Figure S4). Further investigation is necessary to verify whether these two proteins as well as other members of the TcSgroupVII have complement regulatory activity. Finally, a member of the TcSgroupIV that was previously described corresponds to the TsTc13 family, whose function is unknown. Based on the pattern of dispersion of the TcS groups in the MDS projection and the occurrence and sequence of key TcS motifs, we hypothesize that the new groups V and VI and the previously described TcS groupII are more related among each other when compared to the other groups. The same is valid for the new groups VII and VIII and the TcS group III. For instance, TcS groups II, V and VI are the only ones that do not have the FRIP motif and their consensus sequences of the VTVxNVxLYNR motif are very similar. Also, TcS groups III, VII and VIII share the same pattern of motif occurrence and are clustered in a similar region in the MDS projection.

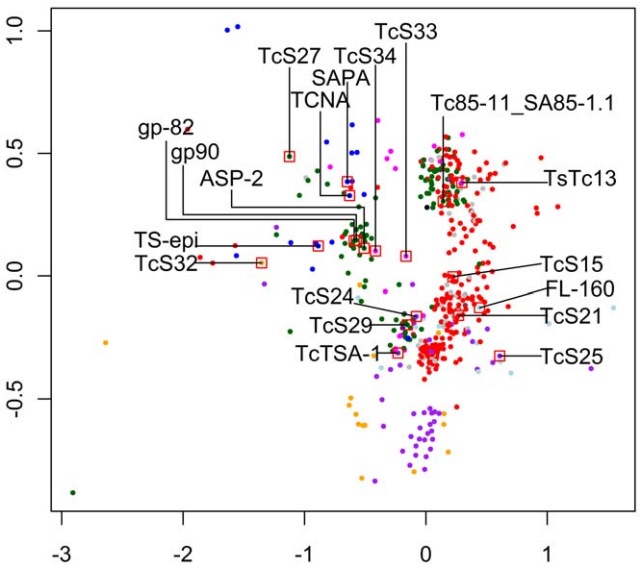
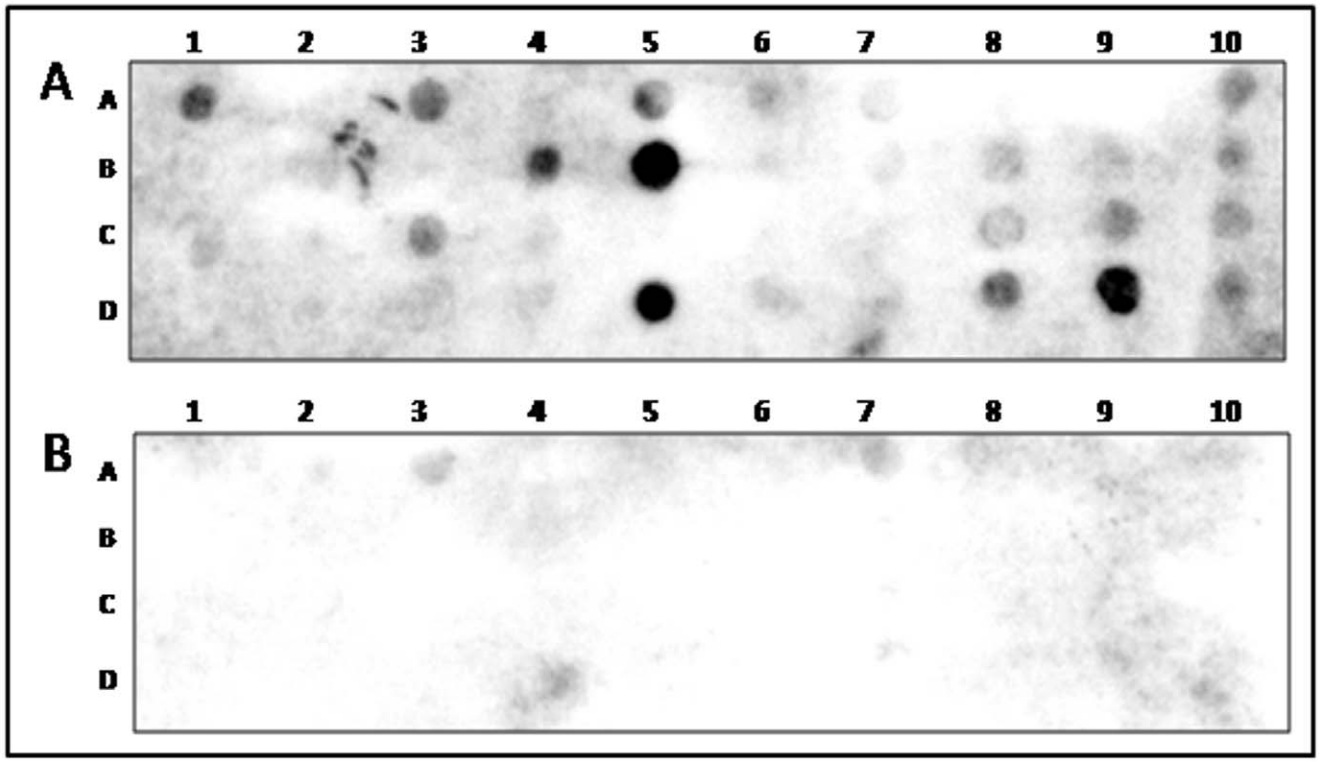


Figure 6. Multidimensional scaling (MDS) plot of the 3′ flanking regions of the TcS genes. A total of 300 nucleotides downstream from the stop codon of each gene were analyzed. Sequences smaller than 300 nucleotides were excluded. Previously characterized genes were mapped on the MDS.
doi:10.1371/journal.pone.0025914.g006

Trypanosoma brucei genome encodes active trans-sialidases expressed in the insect form of the parasite [42]. Although no active trans-sialidase was identified in *Trypanosoma rangeli*, sialidases/sialidase-like proteins similar to TcS groups I, II and III were found, and several of these members are expressed in the epimastigote and trypomastigote forms of the parasite [43,44]. The evolution of the TcS family suggests a gene ancestor encoding an active trans-sialidase expressed in insect forms of the genus *Trypanosoma* and several rounds of duplication and diversification would give rise to trans-sialidases expressed in mammalian forms [45]. Later in evolution *T. rangeli*, would have lost the active trans-sialidase, retaining the sialidase activity. These evidences along with the centered location of TcSgroupI in the MDS projection suggest that extensive expansion and sequence diversification of trans-sialidases similar to TcSgroupI would have originated other groups and functions.

Although the TcS family displays a high degree of sequence variation (Table 1), several motifs are conserved. The most conserved is the VTVxNVxLYNR motif, which is located upstream from the carboxyl terminus of all the TcS full-length members (Figure 2). Recently, it has been demonstrated that a version of this motif (VTVTNVFLYNRPLN), referred to as the FLY motif, may act as a virulence factor [46,47]. BALB/c mice administered with FLY-synthetic peptide are more susceptible to *T. cruzi* infection, displaying increased systemic parasitaemia and mortality [47]. Also, it has been shown that the FLY motif binds to endothelial cells of the heart, suggesting that it might contribute to the parasite tropism to this organ [10]. We identified the exact sequence of the FLY peptide in 28 members of TcSgroupII. Because a very similar version of this motif (**A**TV**A**NV-FLYNRPLN, in which mismatches are indicated in bold and are underlined) is also found in 23 members of TcSgroupIV, we speculate that, as several TcSgroupII members, this group may also participate in host cell attachment/invasion.

Two other motifs, FRIP (xRxP) and Asp box, can be found in various groups of the TcS family. The FRIP motif, which is closest to the N-terminal, is involved in binding the carboxylate group of sialic acid [48]. This motif is found not only in TcSgroupI, but also in the majority of the members of the TcS groups III, IV, VII and VIII (Figure 2). Although this motif is involved in binding sialic acid, it has been shown that enzymatically inactive members of the sialidase family in *T. cruzi* still preserve carbohydrate binding properties [49,50]. The Asp box follows the FRIP motif and can be repeated up to five times in the sequences of viral, bacterial, trypanosomatid and mammalian sialidases. Although its function is unknown, it is worth noting that the Asp box occurs in secreted proteins and in proteins that act on, or interact with, carbohydrates [51]. Recently, it has been shown that at least some inactive trans-sialidases act as lectin-like proteins able to interact with the carbohydrate portion of glycoconjugates, only if they are sialylated [52]. The authors hypothesized that these inactive trans-sialidase proteins could bind to host surfaces that are rich in sialyl-donor glycoconjugates (functioning as anchors), facilitating the active enzyme to more efficiently undertake the sialyl-transferring activity. Here, we have shown that, in addition to TcSgroupI, members of TcSgroupIV have both FRIP and Asp box motifs



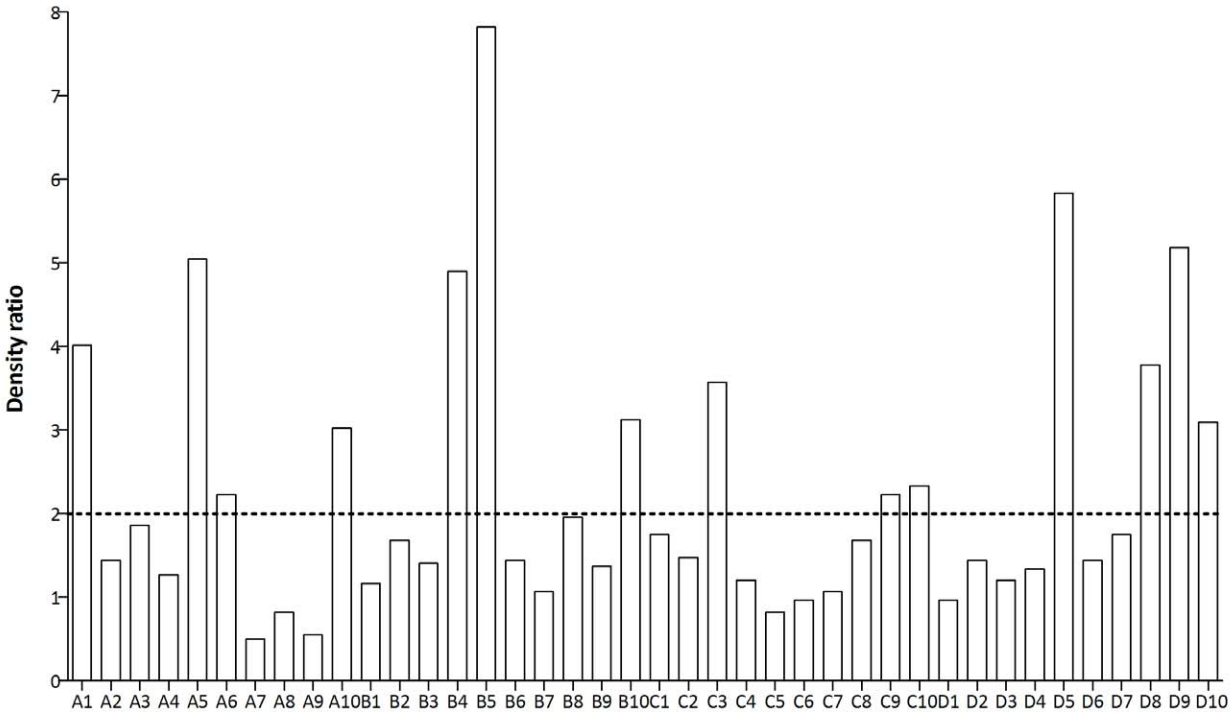


Figure 7. Antigenic profile of TcS peptides. The top panel shows a representative result of immunoblot employing a SPOT synthesis membrane and pools of sera from *T. cruzi*-infected mice (A) and from control uninfected mice (B). The reaction was revealed with secondary anti-total IgG antibody. The bottom panel shows the relative intensity of the signal of each spot estimated based on a comparison of the reactivity in immunoblots with sera from *T. cruzi*-infected mice to the background levels, determined by reactivity with sera from uninfected mice. A signal was scored as reactive when relative intensity (RI)$\geq$2. The peptides analyzed for each TcS group are as follows: TcS group I, D5–D10; TcS group III, C9-D4; TcS group IV, B5-C1, C3; TcS group V, A1, C2, C3, C7; TcS group VI, C2–C8; TcS group VII, A9-B4; TcS group VIII, A2–A8.
doi:10.1371/journal.pone.0025914.g007

(Figure 2) and therefore may also display carbohydrate binding properties.

After mapping all TcS groups on the *T. cruzi* chromosomes [20], we found no association between a group and a specific chromosomal location (Figure 3). Interestingly, we found a distinctive pattern of gene distribution along the chromosomes for members of the TcS groups II and V, with the former clearly enriched at the end of the chromosomes, whereas the latter is concentrated in the middle of the chromosomes (Figure 4). *Trypanosoma brucei* and *Plasmodium falciparum* have a sophisticated strategy for immune evasion, known as antigenic variation, which allows the parasites to adapt to the host environment through exposing and changing specific variable antigenic surface proteins [53–58]. In these parasites, the genes that encode surface proteins that are involved in antigenic variation are preferentially located at subtelomeric regions because these are favorable genomic environments that facilitate gene switching, expression, expansion and generations of new variants [53,55]. Because we found an enrichment of TcS pseudogenes within subtelomeres (Figure 4C), we speculate that these *T. cruzi* regions have also been subjected to intense rearrangement. *T. cruzi* does not undergo antigenic variation but instead co-expresses several variable surface proteins, among which is TcS [59]. Nevertheless, the subtelomeric location of TcSgroupII may facilitate the generation of new variants. In fact, *in silico* simulations suggested that both mutation and gene conversion may contribute to the generation of diversity in the TcS family [60,61]. Gene conversion may be frequent in subtelomeric regions, and therefore could promote a faster diversification of TcSgroupII. This scenario may be particularly important for this group because several of its members have been implicated in host cell attachment/invasion, and *T. cruzi* has the ability to infect a broad range of host cells. Therefore, it is possible that the large repertoire of peptides derived from TcSgroupII may contribute to this phenomenon.

Co-expression of several members of the TcS family has been described in the mammalian stages of the parasite [59]. Here, we show that the levels of expression are not homogeneous between and within the TcS groups (Figure 5). It is well known that the 3′UTRs are implicated in the control of the gene expression of several *T. cruzi* genes that are regulated during the life cycle. Although we have not mapped the 3′UTRs of the genes selected for expression analysis, for a few genes, it was possible to find a correlation between the expression profile and the sequence similarity in their 3′ flanking regions. For example, SAPA and TCNA genes, which are both active trans-sialidases expressed in trypomastigotes, have almost identical 3′ flanking regions (Figure 6). On the other hand, the 3′ flanking sequences of the genes TcS8 and TcS25 are quite similar (75% identity) despite the fact that their pattern of expression is very distinct (Figures 5 and 6). In this case, it is possible that cis-acting regulatory elements present in regions other than the 3′UTR may modulate their expression. It is also unclear what the proportion of the total TcS repertoire is expressed and whether the repertoire and/or the level of expressed genes may change during the parasite infection. High-throughput RNA sequencing approaches will clarify these questions.

We have also investigated the antigenic profile of peptides derived from distinct groups of the TcS family (Figure 7). Besides the known epitopes derived from the repetitive sequences of the SAPA and TcTS13 proteins, new B-cell epitopes were identified in members from both previously described and new TcS groups. Nine of the 14 reactive peptides are found in more than one TcS member. Specifically, the highly reactive peptide C3 (Figure 7) occurs in the largest number of proteins (60 in total) including members of the two new TcS groups V and VI. Also, similar but not identical sequences of this peptide were found in more than 150 TcS members. The cross-reaction among several epitopes and the sequence variability of the TcS family might contribute to the simultaneous presence of B-cell related epitopes during an infection. In fact, it has been proposed that cross-reactivity among the *T. cruzi* epitopes could be an evasion mechanism that drives the immune system into a series of spurious and non-neutralizing antibody responses [62]. In this regard, it has been shown that subtle differences at amino acid positions in or around the active site of the TcS proteins that have trans-sialidase activity might delay the immune response and avoid inhibiting the complete enzymatic makeup of the parasite [63]. This scenario may represent an evolutionary pressure driving the diversification of TcSgroupI, which harbors the active trans-sialidases. Whether a similar mechanism is involved in the diversification of the other TcS groups remains to be addressed.

The diversity of the TcS family may be even greater than reported here since the current assembly of the CL Brener genome is fragmented [2,20], and therefore additional TcS genes may not be part of the dataset analyzed in this study. Nevertheless, based on the nearly complete repertoire of TcS sequences, we can now design probes and antibodies specific for each group, to be employed in more assertive strategies to investigate the role of this complex family during *T. cruzi* infection.

Supporting Information

Figure S1 Partial alignment of active trans-sialidase proteins. FRIP and Asp-box motifs, and critical amino acids residues involved in trans-sialidase activity are shaded in gray. The amino acid positions are relative to the first methionine. Only N-terminal region of the active trans-sialidase proteins is shown.
(DOCX)

Figure S2 Multidimensional scaling plot of the TcS proteins indicating the presence of characteristic TcS motifs. TcS proteins with the motifs are represented by red dots. (A) SXDXGXTW motif; (B) VTVXNVXLYNR motif; (C) SXDXGXTW motif allowing 1 mismatch; (D) sequences with VTVXNVXLYNR motif found in the alignment block of the 505 TcS derived from the eight clusters identified in this study; (E) FRIP (XRXP) motif. X represents any amino acid.
(DOCX)

Figure S3 Prototype of each TcS protein. The peptide signal is represented in gray, FRIP in green, Asp-box in blue, VTVXNVXLYNR in red, repeats in black and GPI anchor addition site in orange.
(TIF)

Figure S4 Divergent CRP – complement regulatory proteins. Protein sequences involved in the regulation of complement system identified by Beucher and Norris (2008). Sequences were mapped on the MDS showed in Figure 1. HSG sequences (high similarity group) and LSG sequences (low-similarity group) are indicated by red and black squares, respectively.
(DOCX)

Table S1 Primers used in the Real-time RT-PCR reactions.
(DOC)

Table S2 TcS peptides analyzed by immunoblotting.
(DOCX)

Table S3 List of members of each TcS group.
(XLS)

Table S4 List of TcS repeats.
(XLSX)

Acknowledgments

We thank Michele Silva de Matos and Jefferson Bernardes for technical assistance.

Author Contributions

Conceived and designed the experiments: LMF SLS DCB. Performed the experiments: LMF SLS GRL TAOM. Analyzed the data: LMF SLS DCB. Contributed reagents/materials/analysis tools: TSR RTG SMRT RTF DCB. Wrote the paper: LMF SLS SMRT DCB.

References

1. World Health Organization (2002) Control of Chagas Disease. Second report of the WHO Expert Committee. WHO Technical Report Series 905.
2. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309: 409–415.
3. Previato JO, Andrade AF, Pessolani MC, Mendonça-Previato L (1985) Incorporation of sialic acid into *Trypanosoma cruzi* macromolecules. A proposal for a new metabolic route. Mol Biochem Parasitol 16: 85–96.
4. Mucci J, Risso MG, Leguizamón MS, Frasch AC, Campetella O (2006) The trans-sialidase from *Trypanosoma cruzi* triggers apoptosis by target cell sialylation. Cell Microbiol 8: 1086–1095.
5. Vercelli CA, Hidalgo AM, Hyon SH, Argibay PF (2005) *Trypanosoma cruzi* trans-sialidase inhibits human lymphocyte proliferation by nonapoptotic mechanisms: implications in pathogenesis and transplant immunology. Transplant Proc 37: 4594–4597.
6. Frasch AC (2000) Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*. Parasitol Today 16: 282–286.
7. Pereira-Chioccola VL, Acosta-Serrano A, Correia de Almeida I, Ferguson MA, Souto-Padron T, et al. (2000) Mucin-like molecules form a negatively charged coat that protects *Trypanosoma cruzi* trypomastigotes from killing by human anti-alpha-galactosyl antibodies. J Cell Sci 113(Pt 7): 1299–1307.
8. Schenkman S, Eichinger D, Pereira ME, Nussenzweig V (1994) Structural and functional properties of Trypanosoma trans-sialidase. Annu Rev Microbiol 48: 499–523.
9. Souza W, Carvalho TM, Barrias ES (2010) Review on *Trypanosoma cruzi*: Host Cell Interaction. Int J Cell Biol 2010: 1–18.
10. Tonelli RR, Giordano RJ, Barbu EM, Torrecilhas AC, Kobayashi GS, et al. (2010) Role of the gp85/trans-sialidases in *Trypanosoma cruzi* tissue tropism: preferential binding of a conserved peptide motif to the vasculature in vivo. PLoS Negl Trop Dis 4: e864.
11. Tzelepis F, de Alencar BC, Penido ML, Claser C, Machado AV, et al. (2008) Infection with *Trypanosoma cruzi* restricts the repertoire of parasite-specific CD8+ T cells leading to immunodominance. J Immunol 180: 1737–1748.
12. Rubin-de-Celis SS, Uemura H, Yoshida N, Schenkman S (2006) Expression of trypomastigote trans-sialidase in metacyclic forms of *Trypanosoma cruzi* increases parasite escape from its parasitophorous vacuole. Cell Microbiol 8: 1888–1898.
13. Cross GA, Takle GB (1993) The surface trans-sialidase family of *Trypanosoma cruzi*. Annu Rev Microbiol 47: 385–411.
14. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
15. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
16. Felsenstein J (1989) Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.
17. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Free program distributed by the authors over the internet from http://evolution.genetics.washington.edu/phylip.html: Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
18. Hartigan JA, Wong MA (1979) Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics) 28: 100–108.
19. R Development Core Team (2011) R: A language and environment for statistical computing, reference index version 2.13.0. (R Foundation for Statistical Computing, Vienna).
20. Weatherly DB, Boehlke C, Tarleton RL (2009) Chromosome level assembly of the hybrid *Trypanosoma cruzi genome*. BMC Genomics 10: 255.
21. Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, et al. (2005) PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res 33: W262–266.
22. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795.
23. Fankhauser N, Mäser P (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics 21: 1846–1852.
24. Bartholomeu DC, Silva RA, Galvão LM, el-Sayed NM, Donelson JE, et al. (2002) *Trypanosoma cruzi*: RNA structure and post-transcriptional control of tubulin gene expression. Exp Parasitol 102: 123–133.
25. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immunome Res 2: 2.
26. Frank R (1992) Spot-synthesis: an easy technique for the position- ally addressable, parallel chemical synthesis on a membrane support. Tetrahedron 48: 9217–9232.
27. Colli W (1993) Trans-sialidase: a unique enzyme activity discovered in the protozoan *Trypanosoma cruzi*. FASEB J 7: 1257–1264.
28. Roggentin P, Rothe B, Kaper JB, Galen J, Lawrisuk L, et al. (1989) Conserved sequences in bacterial and viral sialidases. Glycoconj J 6: 349–353.
29. Beucher M, Norris KA (2008) Sequence diversity of the *Trypanosoma cruzi* complement regulatory protein family. Infect Immun 76: 750–758.
30. Todeschini AR, Mendonça-Previato L, Previato JO, Varki A, van Halbeek H (2000) Trans-sialidase from *Trypanosoma cruzi* catalyzes sialoside hydrolysis with retention of configuration. Glycobiology 10: 213–221.
31. Teixeira SM, Russell DG, Kirchhoff LV, Donelson JE (1994) A differentially expressed gene family encoding "amastin," a surface protein of *Trypanosoma cruzi* amastigotes. J Biol Chem 269: 20509–20516.
32. Campos PC, Bartholomeu DC, DaRocha WD, Cerqueira GC, Teixeira SM (2008) Sequences involved in mRNA processing in *Trypanosoma cruzi*. Int J Parasitol 38: 1383–1389.
33. Pollevick GD, Affranchino JL, Frasch AC, Sánchez DO (1991) The complete sequence of a shed acute-phase antigen of *Trypanosoma cruzi*. Mol Biochem Parasitol 47: 247–250.
34. Burns JM, Shreffler WG, Rosman DE, Sleath PR, March CJ, et al. (1992) Identification and synthesis of a major conserved antigenic epitope of *Trypanosoma cruzi*. Proc Natl Acad Sci U S A 89: 1239–1243.
35. Giordano R, Chammas R, Veiga SS, Colli W, Alves MJ (1994) *Trypanosoma cruzi* binds to laminin in a carbohydrate-independent way. Braz J Med Biol Res 27: 2315–2318.
36. Ouaissi A, Cornette J, Taibi A, Velge P, Capron A (1988) Major surface immunogens of *Trypanosoma cruzi* trypomastigotes. Mem Inst Oswaldo Cruz 83 Suppl 1: 502.
37. Velge P, Ouaissi MA, Cornette J, Afchain D, Capron A (1988) Identification and isolation of *Trypanosoma cruzi* trypomastigote collagen-binding proteins: possible role in cell-parasite interaction. Parasitology 97(Pt 2): 255–268.
38. Santana JM, Grellier P, Schrével J, Teixeira AR (1997) A *Trypanosoma cruzi* secreted 80 kDa proteinase with specificity for human collagen types I and IV. Biochem J 325(Pt 1): 129–137.
39. Magdesian MH, Giordano R, Ulrich H, Juliano MA, Juliano L, et al. (2001) Infection by *Trypanosoma cruzi*. Identification of a parasite ligand and its host cell receptor. J Biol Chem 276: 19382–19389.
40. Norris KA, Bradt B, Cooper NR, So M (1991) Characterization of a *Trypanosoma cruzi* C3 binding protein with functional and genetic similarities to the human complement regulatory protein, decay-accelerating factor. J Immunol 147: 2240–2247.
41. Van Voorhis WC, Eisen H (1989) Fl-160. A surface antigen of *Trypanosoma cruzi* that mimics mammalian nervous tissue. J Exp Med 169: 641–652.
42. Engstler M, Reuter G, Schauer R (1992) Purification and characterization of a novel sialidase found in procyclic culture forms of *Trypanosoma brucei*. Mol Biochem Parasitol 54: 21–30.
43. Grisard EC, Stoco PH, Wagner G, Sincero TC, Rotava G, et al. (2010) Transcriptomic analyses of the avirulent protozoan parasite *Trypanosoma rangeli*. Mol Biochem Parasitol 174: 18–25.
44. Buschiazzo A, Campetella O, Frasch AC (1997) *Trypanosoma rangeli* sialidase: cloning, expression and similarity to *T. cruzi* trans-sialidase. Glycobiology 7: 1167–1173.
45. Briones MR, Egima CM, Eichinger D, Schenkman S (1995) Trans-sialidase genes expressed in mammalian forms of *Trypanosoma cruzi* evolved from ancestor genes expressed in insect forms of the parasite. J Mol Evol 41: 120–131.
46. Magdesian MH, Tonelli RR, Fessel MR, Silveira MS, Schumacher RI, et al. (2007) A conserved domain of the gp85/trans-sialidase family activates host cell extracellular signal-regulated kinase and facilitates *Trypanosoma cruzi* infection. Exp Cell Res 313: 210–218.
47. Tonelli RR, Torrecilhas AC, Jacysyn JF, Juliano MA, Colli W, et al. (2011) In vivo infection by *Trypanosoma cruzi*: the conserved FLY domain of the gp85/trans-sialidase family potentiates host infection. Parasitology 138: 481–492.
48. Gaskell A, Crennell S, Taylor G (1995) The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure 3: 1197–1205.
49. Cremona ML, Campetella O, Sánchez DO, Frasch AC (1999) Enzymically inactive members of the trans-sialidase family from *Trypanosoma cruzi* display beta-galactose binding activity. Glycobiology 9: 581–587.
50. Todeschini AR, Dias WB, Girard MF, Wieruszeski JM, Mendonça-Previato L, et al. (2004) Enzymatically inactive trans-sialidase from *Trypanosoma cruzi* binds sialyl and beta-galactopyranosyl residues in a sequential ordered mechanism. J Biol Chem 279: 5323–5328.

51. Copley RR, Russell RB, Ponting CP (2001) Sialidase-like Asp-boxes: sequence-similar structures within different protein folds. Protein Sci 10: 285–292.
52. Oppezzo P, Obal G, Baraibar MA, Pritsch O, Alzari PM, et al. (2011) Crystal structure of an enzymatically inactive trans-sialidase-like lectin from *Trypanosoma cruzi*: The carbohydrate binding mechanism involves residual sialidase activity. Biochim Biophys Acta 1814: 1154–1161.
53. Scherf A, Lopez-Rubio JJ, Riviere L (2008) Antigenic variation in *Plasmodium falciparum*. Annu Rev Microbiol 62: 445–470.
54. Kim D, Chiurillo MA, El-Sayed N, Jones K, Santos MR, et al. (2005) Telomere and subtelomere of *Trypanosoma cruzi* chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of *T. cruzi* telomeres. Gene 346: 153–161.
55. Horn D, Barry JD (2005) The central roles of telomeres and subtelomeres in antigenic variation in African trypanosomes. Chromosome Res 13: 525–533.
56. Chiurillo MA, Peralta A, Ramírez JL (2002) Comparative study of *Trypanosoma rangeli* and *Trypanosoma cruzi* telomeres. Mol Biochem Parasitol 120: 305–308.
57. Cano MI (2001) Telomere biology of trypanosomatids: more questions than answers. Trends Parasitol 17: 425–429.
58. Scherf A, Figueiredo LM, Freitas-Junior LH (2001) Plasmodium telomeres: a pathogen's perspective. Curr Opin Microbiol 4: 409–414.
59. Atwood JA, Weatherly DB, Minning TA, Bundy B, Cavola C, et al. (2005) The *Trypanosoma cruzi* proteome. Science 309: 473–476.
60. Azuaje FJ, Ramirez JL, Da Silveira JF (2007) *In silico*, biologically-inspired modelling of genomic variation generation in surface proteins of *Trypanosoma cruzi*. Kinetoplastid Biology and Disease 6: 6–17.
61. Azuaje F, Ramirez JL, Da Silveira JF (2007) An exploration of the genetic robustness landscape of surface protein families in the human protozoan parasite *Trypanosoma cruzi*. IEEE Transactions on Nanobioscience 6: 223–228.
62. Pitcovsky TA, Buscaglia CA, Mucci J, Campetella O (2002) A functional network of intramolecular cross-reacting epitopes delays the elicitation of neutralizing antibodies to *Trypanosoma cruzi* trans-sialidase. J Infect Dis 186: 397–404.
63. Ratier L, Urrutia M, Paris G, Zarebski L, Frasch AC, et al. (2008) Relevance of the diversity among members of the *Trypanosoma cruzi* trans-sialidase family analyzed with camelids single-domain antibodies. PLoS One 3: e3524.

Anexo 3 - Artigo publicado na revista internacional “Journal of Proteome Research”

Analysis of *Leishmania chagasi* by 2-D Difference Gel Eletrophoresis (2-D DIGE) and Immunoproteomic: Identification of Novel Candidate Antigens for Diagnostic Tests and Vaccine

Costa MM, Andrade HM, Bartholomeu DC, Freitas LM, Pires SF, Chapeaurouge AD, Perales J, Ferreira AT, Giusta MS, Melo MN, Gazzinelli RT.

Article

Analysis of Leishmania chagasi by 2-D Difference Gel Eletrophoresis (2-D DIGE) and Immunoproteomic: identification of novel candidate antigens for diagnostic tests and vaccine

Miriam Maria Costa, Hélida Monteiro Andrade, Daniella Castanheira Bartholomeu, Leandro Freitas, Simone Fonseca Pires, Alex Chapeaurouge, Jonas Perales, André T.S. Ferreira, Maria Norma Melo, Mario Silva Giusta, and Ricardo Gazzinelli

J. Proteome Res., **Just Accepted Manuscript** • Publication Date (Web): 28 February 2011

Analysis of *Leishmania chagasi* by 2-D Difference Gel Eletrophoresis (2-D DIGE) and Immunoproteomic: identification of novel candidate antigens for diagnostic tests and vaccine

Míriam M. Costa[&,], Hélida M. Andrade[#,*], Daniella C. Bartholomeu[#], Leandro M. Freitas[#], Simone F. Pires[#], Alexander D. Chapeaurouge[ψ], Jonas Perales[ψ], André T. Ferreira[ψ], Mário S. Giusta[&], Maria N. Melo [#], and Ricardo T. Gazzinelli[&,β,ζ,π]*

& - Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Departamento de Bioquímica e Imunologia, 31270-910 Belo Horizonte, Minas Gerais, Brasil;

- Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Departamento de Parasitologia, 31279-910 Belo Horizonte, Minas Gerais, Brasil;

ψ - Fundação Oswaldo Cruz, Departamento de Fisiologia e Farmacodinâmica, 20000 Rio de Janeiro, Rio de Janeiro, Brasil;

β - Centro de Pesquisas René Rachou – Fundação Oswaldo Cruz, 30190-002 Belo Horizonte, Minas Gerais, Brasil;

ζ - University of Massachusetts Medical School, Division of Infectious Diseases and Immunology, 01605-2324 Worcester, Massachusetts, USA.

* - These authors have equally contributed to the work.

π - To whom correspondence must be addressed – e-mail: ritoga@cpqrr.fiocruz.br , Phone: 55 (31) 3349-7774.

CORRESPONDING AUTHOR FOOTNOTE

Míriam M. Costa – miriamcosta@cpqrr.fiocruz.br - Phone: 55 (31) 3409-2634

Hélida M. Andrade - helida@icb.ufmg.br - Phone: 55 (31) 3409-3010

Daniella C. Bartholomeu - daniella@icb.ufmg.br - Phone: 55 (31) 3409-2825

Leandro M. Freitas – lm_freitas@yahoo.com.br - Phone: 55 (31) 3409-2825

Simone F. Pires - simonefpires@gmail.com - Phone: 55 (31) 3409-3010

Alexander D. Chapeaurouge - henk@ioc.fiocruz.br - Phone: 55 (21) 2562-1241

Jonas Perales - jperales@ioc.fiocruz.br - Phone: 55 (21) 2562-1241

André T. Ferreira - atsferreira@ioc.fiocruz.br - Phone: 55 (21) 2562-1241

Mário S. Giusta - mgiusta@yahoo.com.br - Phone: 55 (31) 3409-2634

Maria N. Melo – melo@icb.ufmg.br – Phone: 55 (31) 3409-2850

Ricardo T. Gazzinelli – ritoga@cpqrr.fiocruz.br - Phone: 55 (31) 3349-7774

ABSTRACT

Identification of novel antigens is essential for developing new diagnostic tests and vaccines. We used DIGE to compare protein expression in amastigote and promastigote forms of *Leishmania chagasi*. Nine hundred amastigote and promastigote spots were visualized. Five amastigote-specific, 25 promastigote-specific, and 10 proteins shared by the two parasite stages were identified. Furthermore, 41 proteins were identified in the Western blot employing 2-DE and sera from infected dogs. From these proteins, 3 and 38 were reactive with IgM and total IgG, respectively. The proteins recognized by total IgG presented different patterns in terms of their recognition by IgG1 and/or IgG2 isotypes. All the proteins selected by Western blot were mapped for B-cell epitopes. One hundred and eighty peptides were submitted to SPOT synthesis and immunoassay. A total of 25 peptides were shown of interest for serodiagnosis to visceral leishmaniasis. In addition, all proteins identified in this study were mapped for T cell epitopes by using the NetCTL software, and candidates for vaccine development selected. Therefore, a large-scale screening of *L. chagasi* proteome was performed to identify new B and T cell epitopes with potential use for developing diagnostic tests and vaccines.

KEYWORDS - Leishmaniasis, proteome, antigens, diagnosis, vaccine.

1. INTRODUCTION

Leishmaniasis occurs in 88 countries with approximately 12 million infected individuals and 350 million people at risk of contracting the infection (http://www.who.int/en/). There are several clinical manifestations of the disease, which differ according to the *Leishmania* species and the host immune response. These manifestations are the cutaneous, mucocutaneous, diffuse and visceral diseases. Visceral leishmaniasis (VL), the more severe form of disease, is an anthroponosis in India and Central Africa, and a zoonosis in the Mediterranean and Latin America. Infection with *L. infantum,* also named *L. chagasi* in Latin America, represents 20% of the global human cases (100,000 cases per year) of zoonotic VL, and its incidence is increasing in urban and peri-urban areas in the tropics[1]. Although available, the drugs used to treat VL lead to severe side effects, and clinical presentation of the disease is not sufficiently specific to guide treatment[2]. Moreover, the existing diagnostic tests and vaccines for VL need to be improved[3,4]. Since dogs are the main reservoirs for *Leishmania* parasites, they are important targets for control of parasite transmission in countries, where VL is a zoonosis. Hence, development of accurate serodiagnostic tests, as well as more effective vaccines for canine VL, are highly desirable to control transmission and disease dissemination in Mediterranean and Latin American countries.

The development of new serodiagnostic tests and vaccines for VL is largely hampered by insufficient knowledge about the complexity of the immune responses as well as the identification of proteins that are immunogenic for B and T lymphocytes. Indeed, only a limited number of *Leishmania* proteins have been tested in diagnostic tests and vaccine formulations for VL[2,5-8]. In this context, the identification of new immunogenic proteins derived from *Leishmania* species that cause VL, is highly desirable for the development of more sensitive/specific serodiagnostic tests, as well as effective vaccines. Thus, proteomics is a useful approach to obtain a more complete list of immunogenic proteins from *Leishmania* parasites. Currently, there are several studies in *Leishmania* proteomics, including *L. infantum*[9-

[14]. Nevertheless, this is the first proteomic study designed to identify antigens of potential use in development of diagnostic tests and vaccine for canine VL. For this purpose, we used a highly virulent strain of *L. chagasi* isolated from a VL dog in Brazil[15]. While *L. chagasi* and *L. infantum* are considered the same species[16,17], their insect vector and wild reservoirs are distinct in Europe and Latin America[18,19]. These differences in parasite life cycle may result in variation in protein expression, leading to distinct biological behavior, as is the case of parasite virulence and higher number of severe cases in Latin America[20,21].

Precisely, we performed a wide screen to search for immunogenic proteins from both promastigotes and amastigotes forms of *L. chagasi*. As criteria of antigen selection for serodiagnostic tests, we identified proteins that were recognized by antibodies present in sera from dogs with VL. We then mapped both B-cell and $CD8^+$ T cell epitopes from the immunogenic proteins. B cell epitopes were experimentally selected by a second screen employing peptide arrays and sera from infected dogs. $CD8^+$ T cell epitopes were further selected, *in silico*, based on their high affinity and ability to bind to various HLA haplotypes. Importantly, we identified 19 hypothetical as well as 7 putative proteins, which were among the antigens with highest immunogenic scores. Thus, our study allowed the identification of previously undefined *L. chagasi* proteins, which are strong candidates for developing novel immunological based diagnostic tests and vaccines for VL.

2. MATERIAL AND METHODS

Ethics Statement - Experiments with dogs were performed in accordance to guidelines of the Institutional Animal Care and Committee on Ethics of Animal Experimentation (Comitê de Ética em Experimentação Animal - CETEA) from the Universidade Federal de Minas Gerais, protocol 211/07 approved in 03/12/2008.

L. chagasi - The *L. chagasi* amastigotes (MCAN/BR/2000/BH400) were purified from spleens of hamsters 90 days post-infection. For purification of amastigotes from hamster

spleens, we used a modified version of the method described by Chang (1980) [22]. Immediately after the spleen removal, imprint smears were prepared and stained with GIEMSA for visualization of amastigotes by optic microscopy. The organ was then macerated in Schneider medium (Gibco BRL, Paisley, UK) employing a daucer, and centrifuged at 100 g for 10 min. The residual red cells in the supernatant were disrupted with the addition of 0.05% saponine w/v for 5 min, followed by centrifugation at 2,000 g for 10 min. The pellet was then resuspended in 5 ml of Schneider medium. The suspension was passed through a 26-G needle, gently added to 5 ml of Percoll solution (Sigma, St. Louis, MO), and centrifuged at 2,000 g for 40 minutes to separate the amastigote from the cellular ring. The purity of our preparation was verified by microscopic inspection, and no intact host cells were found in the amastigote suspension. The amastigote suspensions were pelleted and frozen at -70°C until use.

L. chagasi promastigotes (MCAN/BR/2000/BH400) were grown at 23°C in Schneider's medium (Gibco BRL) supplemented with 10% heat-inactivated fetal bovine serum (Sigma), 200 U of penicillin/ml (Sigma), and 100 µg of streptomycin/ml (Sigma) at pH 7.4. Promastigotes from the logarithm phase were submitted to centrifugation at 8,000 g for 20 min at 4°C and the pellet collected and stored at -70°C.

Canine Sera - For the initial screening of *Leishmania* antigens, we used a pool of sera from twenty animals per experimental group, *i.e.* acutely infected, chronically infected, and uninfected control dogs. The acutely infected dogs were challenged intravenously with 10^7 amastigotes of *L. chagasi* and blood collected 30 days post-infection. The chronically infected dogs were naturally infected with *Leishmania* in the metropolitan region from Belo Horizonte, rescued and maintained in our facility for laboratorial and clinical evaluation. VL in chronically infected dogs was certified by the presence of clinical symptoms, and parasitological tests in bone marrow cells examined by optical microscopy. The uninfected dogs were negative in parasitological as well as serological tests for VL and used as negative controls in our study. Blood was withdrawn and maintained at room temperature for 3 h to obtain serum. Individual

sera were tested for serology employing immunofluorescence and ELISA tests for anti-*Leishmania* antibodies. Animals were included in our study as acutely infected, chronically infected, and uninfected when the diagnostic tests were IgM (+), IgG (+)/ IgM (-), and IgG (-)/IgM (-), respectively. One hundred microliters from individual serum of 20 dogs selected per group were deposited in a single tube, in order to obtain a pool of sera that was representative of acutely infected, chronically infected and uninfected dogs.

Protein Extract - The parasite were suspended in lysis buffer (8 M urea, 2 M thiourea, 4% CHAPS, 65 mM dithiothreitol - DTT, 40 mM Tris base, and a protease inhibitor mix - GE Healhtcare, San Francisco, CA) in a proportion of 500 µl of lysis buffer for 10^9 parasite. Samples were incubated for 1 h at room temperature, with occasional vortexing, and then centrifuged for 15 min at 20,000 g and room temperature. The supernatant (protein extract) was kept at -70°C until analysis. The protein content was measured using the 2D-Quant kit (GE Healthcare) according to the manufacturer's instructions.

Two-Dimensional Gel Electrophoresis (2-DE) - To identify differentially expressed proteins between amastigote and promastigote forms we used Differential Gel Electrophoresis (DIGE), briefly, 150 µg of sample was labeled with 400 pmol of N-hydroxysuccinimidyl-ester-derivates of the cyanine dyes (Cy2, Cy3 and Cy5 - GE Healthcare) following the manufacturer's protocol. The reaction was quenched with 1 ml of 10 mM lysine for 10 min on ice and in the dark. A mixture of protein extracts from amastigote and promastigote forms was labeled with Cy2 as an internal standard. Protein extracts from promastigote and amastigote forms were labeled with Cy3, and Cy5, respectively. Experiments were performed with three biological replicates and a dye-swap for both parasite forms. Different labeled extracts were pooled, reduced with 2% DTT, complemented with 2% ampholytes (pH 4–7), adjusted to a final volume of 350 µL with sample buffer (7 M urea, 2 M thiourea and 4% CHAPS), and incubated for 10 min on ice and in the dark. The isoelectric focusing voltage was increased gradually to 8,000 V and run for 60,000 Vh at 20ºC and a maximum current of 50 µA/strip in

Immobilized pH Gradient (IPG) 18 cm, pH 4-7 (GE Healthcare). Focused IPG strips were incubated for 15 min in equilibration solution (50 mM Tris-HCl pH 8.8, 6 M urea, 30% glycerol, 2% SDS, 0.002% bromophenol blue and 125 mM DTT) and then alkylated for further 15 min in an equilibration solution containing 13.5 mM iodoacetamide instead of DTT. Strips were transferred onto a 12% SDS-PAGE gel and second dimensional focusing was performed at 15ºC using 20 mA/gel for 1 h, followed by 50 mA/gel, with an Ettan DALT 6 unit (GE Healthcare). Gels were scanned on a Typhoon Trio laser imager (GE Healthcare) with excitation/emission wavelengths specific for Cy2 (488/520 nm), Cy3 (532/580 nm) and Cy5 (633/670 nm).

Images were analyzed using ImageMaster 2D Platinum 6.0® software (GE Healthcare). Normalized spot volume data were log10-transformed before analysis, in order to eliminate distributional skew and improve the normal approximation for validity of p-values. Analysis of variance (ANOVA) was performed on log10-normalized spot volumes. Estimated differences between amastigotes and promastigotes were obtained from the model as differences in least-square and exponential means (linear contrasts). Significance testing was performed at the 5% level. All statistical analyses were accomplished using SASs V9.1 software. To manually remove the selected spots after scanning in Typhoon, the DIGE gels were also stained with colloidal Coomassie Brilliant Blue (CBB) G-250 following procedures described elsewhere[23].

Western blot – 2DE - For Western blot analysis, the 500 µg protein extract from promastigote forms were used and fractionated in 2-DE gel. First dimension IEF and second dimension SDS-PAGE were performed as described above, including the isoelectric focusing voltage and the IPG of 18 cm, pH 4-7 (GE Healthcare). The samples were determined in individual gels and no fluorescent dye was employed. The proteins from unstained 2D were transferred onto nitrocellulose membranes (Amersham Biosciences, Piscataway, NJ) in a trans-blot semidry transfer Unit (GE Healthcare) by applying a current of 1.6 mA/cm^2 for 2 h. The membranes were rinsed with TBS–Tween buffer (20 mM Tris, 500 mM NaCl, pH 7.4) and

incubated with blocking buffer (5% low fat milk powder in Tris-buffered saline) at 4ºC overnight. The blotted membranes were incubated with pools of canine sera (*i.e.,* acutely and chronically infected or uninfected control dogs) diluted 1:500 in blocking buffer, for 2 h at room temperature. After washing three times in 0.05% TBS–Tween for 15 min, the membranes were incubated with either anti-total dog IgG, anti-IgG1, anti-IgG2 or anti-IgM-peroxidase conjugates (Sigma) diluted 1:10,000 in blocking buffer, for 2 h at room temperature. Membranes were washed three times with TBS - Tween buffer for 5 min and three times with TBS for 5 min. Finally, nitrocellulose sheets were washed with a mixture of 6 mg of 3,3'-diaminobenzidine DAB (Sigma) in 12 ml of TBS buffer and 12 ml of a solution containing 10 ml of the phosphate-buffered saline, 10 µl of hydrogen peroxide and 2 ml of methanol (Sigma). Blots were incubated with this solution for 1–3 min. The reaction was interrupted with water, and the blots were dried with paper towels and stored at room temperature[24]. Spots that were recognized only by sera (total IgG, IgG1, IgG2 and/or IgM) from infected animals, but not from uninfected dogs were cut in gels. To select the spots, the images from membranes and gels with protein extracts were analyzed using the ImageMaster 2D Platinum 6.0® (GE Healthcare). The selected spots were removed for identification by mass spectrometry.

Identification of Proteins - MALDI-TOF/TOF MS (Matrix-Assisted Laser Desorption Ionization Time-of-Flight/Time-of-Flight Mass spectrometry) - Protein spots were manually excised from stained 2-D gels. The gel pieces were washed three times with 100 µl of 25 mM ammonium bicarbonate containing 50% v/v acetonitrile. After drying, gel pieces were rehydrated for 30 min at 4°C with 10 µl of trypsin solution (Promega, Madison, WI) containing 20 ng/µl in 25 mM ammonium bicarbonate. Excess protease solution was then removed and replaced by 20 µl of 25 mM ammonium bicarbonate. Digestion was performed for 16 h at 37°C. Peptide extraction was performed twice for 15 min with 30 µl of 50% acetonitrile/5% formic acid solution. Trypsin digests were then concentrated in a SpeedVac (Savant Instruments Inc., Farmingdale, NY) concentrator to about 10 µl and desalted using Zip-

Tip (C18 resin; P10, Millipore Corporation, Bedford, MA). Peptides were eluted from the column with 50% acetonitrile/0.1% trifluoroacetic acid.

Roughly 0.3 μl of the sample solution was mixed with an equal volume of a saturated matrix solution [10 mg/ml R-cyano-4-hydroxycinnamic acid (Aldrich, Milwaukee, WI) in 50% acetonitrile/0.1% trifluoroacetic acid] on the target plate and allowed to dry at room temperature. Raw data for the identification of proteins were obtained on the 4700 proteomics analyzer (Applied Biosystems, Foster City, CA). Both MS and MS/MS data were acquired with a neodymium-doped yttrium aluminum garnet (Nd:YAG) laser with a 200-Hz repetition rate. Typically, 1,600 shots were accumulated for spectra in the S mode, while 2,400 shots were accumulated for spectra in the MS/MS mode. Six of the most intense ion signals with a signal-to-noise ratio above 30 were selected as precursors for MS/MS acquisition, with the exclusion of common trypsin autolysis peaks and matrix ion signals. External calibration in MS mode was performed using a mixture of four peptides: des-Arg1-Bradykinin (m/z 904.468); angiotensin I (m/z 1,296.685); Glu1-fibrinopeptide B (m/z 1,570.677); and ACTH (18-39) (m/z 2,465.199). MS/MS spectra were externally calibrated using known fragment ion masses observed in the MS/MS spectrum of angiotensin I. Following data acquisition, a peak list was obtained from the raw MS/MS data using the “Peaks to Mascot” function in the 4000 Series Explorer software (Applied Biosystems).

Database Search. Uninterpreted tandem mass spectra were searched against the nonredundant protein sequence database from the National Center for Biotechnology Information (NCBI) using the Mascot (version 2.1) MS/MS ion search tool (http:// www.matrixscience.com). The search parameters were as follows: no restriction of protein molecular weight, one missed trypsin cleavage allowed, non-fixed modifications of methionine (oxidation) and cysteine (carbamidomethylation); pyroglutamate formation at the N-terminal glutamine of peptides with no other post-translational modifications being taken into account. Mass tolerance for the peptides in the searches was 0.8 Da for MS spectra and 0.6 Da for

MS/MS spectra. Peptides were considered to be identified when the scoring value exceeded the identity or extensive homology threshold value calculated by Mascot. In cases of protein identification based on a single peptide, the minimum threshold for the probability-based Mascot score was 40. Otherwise, mass spectra with lower scores, but presenting a reasonable tandem mass spectrum, were manually verified[25].

Mapping T-cell and B-cell Epitopes, Peptide Synthesis and Immunoassay - All *Leishmania* proteins identified during the course of the study were screened for potential T cell epitopes using the NetCTL algorithm (Web service for prediction of cytolitic T cell epitopes in protein sequences)[26]. NetCTL was the method of choice for cytotoxic T-lymphocyte epitope prediction because it integrates predictions for different steps involved in MHC class I presentation: proteosomal cleavage, TAP transport efficiency and MHC class I affinity. More importantly, NetCTL method was shown to have a higher predictive performance than the SYFPEITHI, the BIMAS HLA Peptide Binding Prediction, EpiJen, MAPPP, MHC-pathway, and WAPP methods using a dataset containing approximately 300 experimentally validated CTL epitopes[27]. Various studies have also used this method for CTL epitope predictions that were experimentally validated[28-33]. In the present analysis, a score cutoff of 0.75, which corresponds to a good compromise between sensitivity (0.8) and specificity (0.97) was used. A total of 10 HLA supertypes were tested.

The immunogenic proteins selected by Western blot were mapped by BEPIPRED (B cell epitope prediction) software to predict the presence and location of linear B-cell epitopes. We used the BEPIPRED method because it uses the propensity scale methods (as other linear B-cell epitope predictors) and also incorporates hidden Markov model (HMM). The combination of the two best propensity scale methods (Parker and Levitt) with HMM resulted in a performance significantly better than a number of individual tested propensity scales (Parker, Chou and Fasman, Levitt, Emini)[34]. Our group has successfully used this method in several other studies for prediction of B-cell epitopes followed by experimental validation. Moreover,

the presence of the peptide signal (http://www.cbs.dtu.dk/services/SignalP/) and the N-glycosylation sites (http://www.cbs.dtu.dk/services/NetNGlyc/) was evaluated. Peptides formed by 12 consecutive amino acids with score higher than 2.0 were selected and tested in cellulose membranes by the SPOT synthesis membrane (peptide arrays on cellulose support generated using SPOT synthesis technology).

The SPOT synthesis was employed using a method for preparation of immobilized peptides with 12 amino acids[35]. The assembly of the peptides was performed utilizing the previously-described Fmoc-chemistry[36]. The reactivity of the SPOT membrane was evaluated according to the protocol described by Soutullo et al. (2007)[37]. Sera from either chronically or acutely infected dogs diluted at 1:500 v/v were used as primary antibodies. The anti-total dog IgG or anti- IgM-alkaline phosphatase (AP) conjugates (Bethyl Laboratories, Montgomery, TX) were used as secondary antibodies, at a dilution of 1:5,000 v/v. All experiments using different combinations of primary and secondary antibodies were performed in triplicate.

3. RESULTS

Differential expression of proteins between amastigote and promastigote forms of *L. chagasi* - Differential expression of proteins between amastigotes and promastigotes forms of *L. chagasi* were analyzed by DIGE, and the representative images are presented in Figure 1. Green spots (Figure 1-A) indicate promastigote proteins, and red spots (Figure 1-B) reveal proteins from amastigote form. An overlay of Figures 1-A and 1-B is shown in Figure 1-D, and yellow spots (Figure 1-C) are mix of proteins from both, amastigotes and promastigotes. Approximately 900 spots were detected in extracts from each parasite stage. All differentially expressed proteins (spots), in addition to those that were abundant in both samples were excised from gel. A total of 113 spots were excised from the gels: (i) 56 spots only from promastigote forms; (ii) 43 spots only from amastigotes; and (iii) 14 spots present in extracts from both stages were selected and identified by MS.

The 56 spots selected from promastigote forms corresponded to 25 different proteins from *Leishmania*. In addition, we identified 10 proteins that had similar expression in both promastigote and amastigote stages (two Hypothetical proteins, Eukaryotic translation initiation factor 3 subunit, Translation elongation factor 1-beta, ATP synthase, epsilon chain, Eukaryotic initiation factor 5a, Adenosine kinase, Ribonucleoprotein p18, mitochondrial precursor, Adenosylhomocysteinase and Trypanothione reductase). From 43 spots from amastigote extract, 18 proteins were identified, and only 5 were derived from *Leishmania* (Alpha tubulin, Hypothetical protein, Phosphomannomutase, Prostaglandin f2-alpha synthase and Translation elongation factor 1-beta). The other proteins were proteins from the hamster, from which actin was by far the dominant contaminant. Thus, from the 28 proteins identified in amastigote extracts 15 (~54%), were from *Leishmania* and the remaining 13 from hamster. The criteria used to determine the origin of different proteins were the levels of homology to protein sequences deduced from genes found in the *L. chagasi* versus hamster genome. All *Leishmania* proteins identified are presented in Supplementary Table 1. Proteins that are differentially expressed in each sample are highlighted in Figures 2A and 2B, and those with similar expression in Figures 2C e 2D. Most of the 900 spots found on DIGE were common to both promastigote and amastigote stages and we assume that they are derived from *Leishmania* parasites. The Venn diagram with the most abundant proteins of amastigote an promastigote stages of *L. chagasi* identified by 2D-DIGE and mass spectrometry is presented in Figure 3.

Immunogenic proteins identified by Western Blot - 2D Gel – In order to identify additional immunogenic proteins from *L. chagasi*, we used pool of sera from animals at different stage of infection, *i.e.,* acutely infected, chronically infected, or uninfected controls. Details of selected animals and pool of sera are described in Material and Methods. We have chosen not to use individual sera, because the repertoire of immunoglobulin varies from animal to animal. Thus, using a pool of sera we should have a better representation of antibody specificity to *Leishmania* antigens, than in individual sera. This approach was shown effective,

since we found a series of novel antigens and epitopes that were recognized by some, but not other sera from infected dogs. Furthermore, particular peptides derived from the newly identified antigens were recognized by individual sera from the vast majority of dogs with VL, but not from uninfected dogs (data not shown).

The gel from promastigote forms of *L. chagasi* were stained with Coomassie Blue (Figure 4-A) or transferred to nitrocellulose membrane and incubated with sera from dogs undergoing acute leishmaniasis (30 days after infection) (Figures 4-B) or from uninfected control dogs (Figure 4-C) to identify proteins recognized by serum IgM. Three proteins (Mannose-1-phosphate guanyltransferase, heat shock protein 83-1 and α-tubulin) were identified (Supplementary Table 2).

Immunoblots were also performed to identify antigens recognized by total IgG, as well as IgG1 and IgG2 isotypes present in sera from dogs chronically infected with *L. chagasi.* The gel from promastigote forms were stained with Coomassie Blue (Figure 5-A), transferred to nitrocellulose membrane, incubated with sera from chronically infected (Figures 5-B, 5-C and 5-D) or uninfected control dogs (Figure 5-E), and probed with anti-total IgG (Fig. 5-B), anti-IgG1 (Fig. 5-C) or anti-IgG2 (Fig. 5-D) antibodies conjugated with peroxidase. A large number of antigenic spots were detected in the acidic pH range (4–7). The antigens that were specifically recognized by immune sera from infected dogs, and not by sera from control animals, were selected and further analyzed by MS-MS. Forty four spots were recognized by total IgG. From those 12 spots were recognized by both IgG1 and IgG2, 9 recognized by IgG2, and 6 recognized by IgG1 subclasses of immunoglobulin (Supplementary Table 2). Seventeen spots from *Leishmania* extract were recognized by total IgG, but not by IgG1 and IgG2 antibodies. We believe that this is a question of protein concentration and higher sensitivity of the immunoblot, when we employed a secondary antibody for total IgG, versus secondary antibodies specific for IgG1 or IgG2 isotypes. The spots identified by each subclass and the number of spots found between the subclasses of immunoglobulin are shown in Figure 6.

Peptides and immunoassay in SPOT synthesis technique as candidates for serodiagnosis - A total of 180 peptides (12 amino acids length) with a high score, were selected after BEPIPRED analysis and submitted to SPOT synthesis. The cellulose membranes containing the peptide array were incubated with pools of sera from chronically infected (positive) and uninfected (negative) dogs and developed with anti-total IgG secondary antibody. The intensity of the spots was determined by overlapping membranes incubated with positive and negative sera, as indicated by analysis employing ImageMaster software. We selected all spots (peptides) with Relative Intensity (RI) greater than 2.0 (Table 1 and Figure 7). Based on immunoassay we selected 25 peptides with no false positive according to the RI determined in the SPOT image. The 25 peptides were derived from 8 different proteins: Heat Shock Protein-83 (1 peptide), 3,2-Transenoyl CoA isomerase (2 peptides), Ribonucleoprotein p18, mitochondrial precursor (1 peptide), Aldose 1-epimerase (2 peptides) and other four hypothetical proteins presenting 2, 3, 7 and 7 peptides each (Table 1 and Figure 7).

Prediction of MHC class I binding peptides using the NetCTL software – All *Leishmania* proteins selected in the DIGE gel as well as in the immunoblot (Supplementary Tables 1 and 2) were screened for potential T cell epitopes using the NetCTL software. In our analysis we used a score cutoff of 0.75, which corresponds to a good compromise between sensitivity (0.8) and specificity (0.97)[26]. After analysis by the NetCTL software the following criteria were used to select T cell antigens/epitopes: (i) the ability of a given peptide to bind from three to five HLA haplotypes; (ii) the high score in the prediction for HLA binding; and (iii) the number of T cell epitopes present in a single protein.

We tested the ability of each peptide/protein to bind to 10 different HLA supertypes (A1, A2, A3, A24, B7, B8, B27, B44, B58 and B62) by virtual analysis. The *in silico* analysis indicate that the largest proportion of nanomers were found to bind A24 (17.5%) allele, followed by A3 (14.9%), and then B62 (11.9%), and then by the others. The majority of the nanomers were predicted to bind to only one HLA, but some seemed to be promiscuous and

bind to multiple supertypes. The largest number of supertypes to which a given peptide could bind is five. On the average, around 78% of the predicted peptides were found to bind to only one supertype, 16.6% of the peptides reacted to two HLAs, 4.3% peptides to three HLAs, 1.1% peptides to four HLAs, and 0.06% peptides to five supertypes. The sequence of peptides that bind HLA predicted to bind from three to five HLA supertypes and the name of proteins they belong to are presented in Supplementary Table 3.

Finally, we determined the number of immunogenic peptides present in *Leishmania* proteins (Supplementary Table 4). A hypothetical protein (Spot Code 64) contains the highest number of predicted peptides with score > 1.5 (166 in total). In part, this is due to the fact that this is the largest protein in the dataset (4,873 aminoacids). Nevertheless, this protein is found among the 1/3 top proteins with the highest percentage of predicted peptides (25%), which takes into account the total number of predicted peptides and the total number of amino acids in the protein. More importantly, we found that the putative protein Fatty acid elongase (Spot Code 73), which has only 299 aminoacids, presented the highest percentage of predicted peptides (33%) (Supplementary Table 4). Proteins and peptides with greater potential to be recognized by CD8+ T cell epitopes are shown in Table 2 and deserve further investigation as a vaccine candidate.

4. DISCUSSION

In this study, we performed proteomic analysis of the promastigote and amastigote stages of *L. chagasi*, aiming to identify immunogenic proteins with potential application in the development of immunodiagnostic tests and vaccine for canine VL[7,8,38,39]. To identify immunogenic B cell epitopes, we employed sera from dogs infected with *L. chagasi*. A total of 3 and 38 antigens were specifically recognized by IgM and IgG, respectively. The identification of antigens was followed by *in silico* analysis to predict B cell epitopes. One hundred and eighty putative immunogenic peptides were screened on a peptide array, leading to the selection of 25 peptides that are able to discriminate sera of infected from control

uninfected dogs. We also performed *in silico* analysis of proteins identified by DIGE or Western blot techniques to predict epitopes recognized by $CD8^+$ T lymphocytes (CTL), which are thought to be important elements in host resistance to VL. Importantly, the approach described above lead to the discovery of various hypothetical and putative proteins that are strong candidates for developing new immunological based diagnostic tests and vaccine for VL.

This study represents several novel aspects in terms of proteomic analysis and antigen discovery for *Leishmania* species. While *L. chagasi* and *L. infantum* are considered the same species[16,17], differences in their life cycle, such as the insect vector and wild reservoirs, may result in variations on protein expression, leading to different behavior[18,19]. For instance the number of cases and death due VL are much higher in Brazil than in Europe[20,21]. Here, we perform proteomic analysis of a highly virulent strain of *L. chagasi*, originally isolated from a dog with VL in Belo Horizonte, Brazil[15]. Importantly, this is the first proteomic study designed to identify antigens that are immunogenic for dogs and can be directly applied to develop new immunodiagnostic test as well as vaccine for canine VL. Furthermore, this study represents a thorough proteomic analysis, in which selection of 81 proteins by DIGE and 2D immunoblots was followed by identification of B and potential T cell epitopes, strengthening our search of antigens. Finally, we identified 19 hypothetical as well as 7 putative proteins, which were not previously described.

The lack of access to good-quality diagnostic tests for *Leishmani*a infection contributes to the enormous burden of ill health in the world. Since the clinical manifestations of VL lack specificity, confirmatory tests are required to identify dogs infected with *L. chagasi*. This is critical for control of disease, since dogs are the main reservoir of *L. chagasi,* and therefore, critical for transmission and spread of VL to humans. Several antibody-detection tests employing parasite preparations or recombinant proteins have been developed for laboratory and field diagnosis of VL. However, a diagnostic method with high specificity and sensitivity,

to guide the management and control of VL in dogs, remains to be developed. We believe that a new generation of diagnostic tests with the expected high sensitivity and specificity should be composed of various linear B-cell epitopes, which have been mapped from *Leishmania* antigens[2,34,38].

Proteomic maps have been generated for different species that cause cutaneous leishmaniasis, *i.e. L. major*[40], *L. braziliensis* [41] and *L. mexicana* [10,42,43], as well as for *L. donovani*[14,44,45] and *L. infantum*[7,9,11] that are the etiological agents of VL. Precisely, it was evaluated the differential expression of proteins in axenic amastigotes and promastigotes forms of *L. donovani*[13,14]. In addition, studies mapping *L. donovani* antigens were performed using 2-D Western blot with human sera and parasites isolated from VL patients in India[44,45]. Another study performed a high-resolution proteome analysis of *L. infantum* promastigotes and allowed the identification of immunogenic proteins recognized by a hyperimmune serum from rabbits[7].

The early diagnosis and treatment has an important role in preventing the development of long-term complications or interrupting transmission of the infectious agent. Three proteins, *i.e.*, Mannose-1-phosphate guanyltransferase, alpha tubulin and heat shock protein (HSP) 83-1, were recognized by IgM present in sera from acutely infected dogs. From these proteins, only one peptide from Mannose-1-phosphate guanyltransferase was recognized by IgM antibodies. The HSP 83-1 has a predicted site for glycosylation and carbohydrates may be the main B cell epitope recognized by IgM present in sera from acutely infected dogs.

Sera from chronically infected animals were also used to identify immunogenic proteins recognized by anti-*Leishmania* specific IgG. All immunogenic proteins were analyzed to predict B cell epitopes, and a peptide array used in an immunoblot for seeking immunogenic peptides[46-48]. Twenty five peptides from 8 proteins were recognized specifically by sera from infected dogs: HSP-83 (1 peptide), 3,2-Trans-enoyl CoA isomerase (2 peptides), Ribonucleoprotein p18, mitochondrial precursor (1 peptide), Aldose 1-epimerase (2 peptides) and another four hypothetical proteins presenting 2, 3, 7 and 7 peptides, respectively (Table 1).

A number of studies have demonstrated that different vaccine formulations can induce significant protection against infection with *Leishmania spp*. in a variety of animal models[5,49,50]. Currently, two vaccines for canine VL are commercially available in Brazil, one vaccine composed of parasite extracts[51], and one composed of a amastigote specific recombinant protein, named A2[15]. However, the efficacy of these vaccines remains partial, and it is necessary to develop a new vaccine formulation with greater efficacy. We also sought to apply bioinformatic methods to identify proteins from *L. chagasi* that contain immunogenic T cell epitopes[52,53] and are candidates for a vaccine against VL. Considering the polymorphic nature of the HLA, promiscuous T cell epitopes are of interest for vaccine design, and the first step of selection of $CD8^+$ T cell epitopes was the ability to bind at least to 3 HLAs. We also considered the affinity of peptide binding to various HLAs, and the number of T cell epitopes in a single protein. Based on these parameters, we generated a list of native proteins as well as peptides (Table 1) that could be used to generate chimera proteins for the development of a VL vaccine.

While both $CD4^+$ and $CD8^+$ T lymphocytes are important components for host resistance to VL[54], algorithms to identify $CD4^+$ T cell epitopes are more error-prone, and we plan to identified these epitopes experimentally. To predict $CD8^+$ T cell epitopes we used the NetCTL program[26] that predicts peptide binding to 10 HLA class I supertypes (A1, A2, A3, A24, B7, B8, B27, B44, B58 and B62). The NetCTL program also integrates other routines to predict proteasomal C-terminal cleavage and transport efficiency by the transporter associated with antigen processing (TAP)[55]. Because dog MHC I polymorphism is not well known and therefore not incorporated into the predictors, we used human HLA to predict potential dog MHC I peptide binders. Nevertheless, we recognize that this study can be directly applied to elaborate a vaccine for dogs, as a major overlap of dog and humans HLA has been described[56].

The amastigote is the parasite stage found in the mammalian hosts. Thus, proteins expressed in this developmental form represent potential candidates for vaccine development.

Indeed, experiments indicate that immunization with proteins expressed by amastigote stage can provide effective protection against infection[15,57]. In our DIGE we found only five proteins, which were amastigote-specific, *i.e.* phosphomannomutase, prostaglandin f2-alpha synthase, elongation factor-1a, alpha tubulin, which have been previously reported to be abundant in *Leishmania* amastigotes[9-11,13], and a newly described hypothetical protein. Importantly, from five proteins identified as amastigote-specific, four of them presented high content of T cell epitopes. Three of those proteins contained promiscuous epitopes, which potentially bind to four different supertypes and were selected as antigens with greater potential to be immunogenic for T cells and vaccine development (Table 2). Furthemore, the vast majority of antigens found in the DIGE were common to amastigotes and promastigotes, and thus, potential vaccine candidates.

Different studies have shown complex antigenic patterns in VL when accessed by Western blot technique[58-61]. Some of these antigens, such as HSP70, gp63, HSP83, several ribosomal proteins, histones, KMP11 or LACK are well characterized and have been used in either diagnostic tests and vaccination protocols[62]. Other antigens are still waiting to be identified and characterized. For example, 3,2-trans-enoyl CoA isomerase and Aldose-1-epimerase are enzymes that catalyze the geometrical or structural changes within one molecule, and the existence of proteins in *Leishmania* protozoa has so far been inferred from homology. Importantly, among the B cell as well as T cell antigens analysis, we identified various antigens previously defined as hypothetical proteins. These hypothetical proteins have conserved homology in other species of *Leishmania* (*L. infantum, L. major* and *L. braziliensis*). Some of these proteins were among the best candidates as B cell antigens for diagnostic tests (Table 1 – spot codes 46, 47, 64 and 66), and T cell antigen for vaccine formulations (Table 2 –spot codes 2, 47, 64, and 65).

In conclusion, using proteomic and *in silico* analysis, we were able to identify novel proteins that are important targets for humoral and T cell responses against *Leishmania*

parasites that cause VL. Further studies employing some of these native as well as chimera proteins shall be employed to develop an accurate diagnostic test and an effective vaccine, to identify infected hosts as well as to prevent transmission and development of canine VL.

5. ACKNOWLEDGMENT

This study was funded by Rede Mineira de Biomoléculas – Fundação de Amparo as Pesquisa do Estado de Minas Gerais (FAPEMIG), and Instituto Nacional de Ciência e Tecnologia de Vacinas (INCTV) - Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

6. SUPPORTING INFORMATION AVAILABLE

The contents of Supporting Information include four large tables, as follows: Supplementary Table 1 - detailed list of proteins expressed in either or both promastigote or amastigote stages of *L. chagasi*, as revealed by DIGE analysis; Supplementary Table 2 - detailed list of proteins revealed by immunoblot analysis; Supplementary Table 3 – sequence of peptides, derived from *L. chagasi* proteins, selected by virtual analysis employing the NetCTL and shown to bind to at least three HLA supertypes; Table 4 – list of proteins which contain high number of peptides able to bind to different HLAs with a high score, as determined by virtual analysis.

7. TABLES

Table 1. List of proteins and their respective immunogenic peptides (B cell epitopes) selected by BEPIPRED algorithm and reactivity with IgG antibodies present in sera from dogs infected with *Leishmania chagasi.*

B CELL EPITOPES			
Spot*Code	**Peptide Code**	**Protein Name**	**Peptide Sequence**
40	40A	Ribonucleoprotein p18, mitochondrial precursor	EAPSKQDKPVEN
46	46A	Hypothetical protein	TERPEGANFATP
	46B		VEGERVETT
	46C		LTMNTNQPRMPQ
47	47A	Hypothetical protein	TRGVKSSSKLPA
	47B		RDDPHKVTPSDM
49	49A	Heat shock protein 83-1	VTKEYEVQNK
51	51A	3,2-trans-enoyl-CoA isomerase,mitochondrial precursor	FQSQPPPGVPQG
	51B		RHQDTNAAPAGS
64	64A	Hypothetical protein	ANIKGVPTRAET
	64B		DSDDTEEGEDEG
	64C		EGTAGEPKPPAM
	64D		MRTSTDMPSQHI
	64E		TRQTSQEPTPVS
	64F		PLATQSYGFGSD
	64G		GVAPPGWYDPPVQ
66	66A	Hypothetical protein	PHRAGETSAAGL
	66B		SQQAPAVPPLPQ
	66C		QGMMSPGRSEEK
	66D		VPKGDKAVSSPP
	66E		GERRRGDAEDGR

	66F		SSRPSPPSKVSS
	66G		AAAAASSPSIAP
72	72A	Aldose 1-epimerase	GYPKNPEEAYAD
	72B		LPASGGPGQRYA

From 41 proteins selected by 2D gel and Western blot analysis, 180 peptides were identified with the BEPIPRED software and synthesized. Twenty five immunogenic peptides (B cell epitopes) were selected from eight proteins derived from *L. chagasi* based on their ability to discriminate sera of infected dogs from sera from uninfected dogs. These peptides present relative intensity corresponding to 2 or greater, when comparing the reactivity of sera from infected dogs to sera from uninfected dogs (see Figure 7) in a SPOT synthesis membrane.

Table 2. List of *Leismania chagasi* derived proteins and peptides with greater potential for binding with high affinity to multiple HLA supertypes.

T CELLS EPITOPES								
SELECTED PROTEINS				**SELECTED PEPTIDES**				
Spot Code	**Protein Name**	**% predicted peptides**	**No of predicted peptides with score >1.5**	**Spot Code**	**Protein Name of Selected Peptides**	**Peptide sequence a.a position**	**No of Supertypes**	**Supertypes**
73	Fatty acid elongase, putative	33.0	16	18	Putative eukaryotic initiation factor 4a	223-FMRDPVRIL-231	5	A2, A24, B7, B8, B62
53	Heat shock protein 83; HSP 90	29.5	16	47	Hypothetical protein	486-WSSQSPKSF-494	5	A1, A24, B8, B58, B62
74	Sterol 24-c-methyltransferase	29.3	22	64	Hypothetical protein	1326-RMMGVLFDY-1334	5	A2, A3, A24, B27, B62
32	Tryparedoxin peroxidase	29.3	11	72	Aldose 1-epimerase	84-FTLDGVKYY-92	5	A1, A3, A24, B58, B62
75	Peroxidoxin 1	29.1	10	3M	Mannose-1-phosphate guanyltransferase	75-WSRKLGVSF-83	5	A24, B7, B8, B58, B62
65	Hypothetical protein	28.6	9	1	Alpha tubulin	164-KSKLGYTVY-172	4	A1, A3, B58, B62
2	Hypothetical protein	27.6	17	2	Hypothetical protein	93-FVQKVMMPL-101	4	A2, B7, B8, B24
24	14-3-3 protein-like protein	27.6	12	3	Phosphomannomutase	539-GTEPKIKWY-547	4	A1, A3, B58, B62

30	IgE-dependent histamine-releasing factor	27.2	9	9	70 kDa heat shock protein	83-ITNPQSTFY-91	4	A1, A3, B58, B62
49	Heat shock protein 83-1	26.3	23	10	Actin	162-HTVPIYEGY-170	4	A1, A3, B58, B62

NetCTL algorithm was used to identify T cell epitopes. The proteins were chosen based on the number of peptides predicted to bind with high affinity to multiple HLAs. Peptides were selected based on their ability to bind with high affinity to more than three type of HLA.

8. FIGURE CAPTIONS

Figure 1. 2D-DIGE analysis of promastigote and amastigotes extracts from *Leishmania chagasi*. Proteins of promastigotes and amastigotes forms, and a pooled internal standard were labeled with CyDyes Cy3, Cy5 and Cy2, respectively, mixed and separated on a 2D gel using 18 cm pH 4-7 (left to right) strips in the first dimension and 12% PAGE-SDS gels in the second dimension. The standards for Cy2 as well as molecular weight and pI were defined in the same gel stained with Coomassie Blue and software analysis, respectively. Gels were scanned to obtain single images of (A) promastigotes (Cy3, green), (B) amastigotes (Cy5, red) or (C) the internal standard (Cy2, yellow). (D) An overlay of the two dyes (Cy3, Cy5) is shown in yellow.

Figure 2. Proteins highly expressed by promastigotes (A) and amastigotes (B) forms derived from *Leishmania chagasi*. The level of protein expression was determined by using ImageMaster 2D Platinum 6.0® software, please see details in the Material and Methods section. The "h" letter (in 2B) represents hamster-derived proteins present in amastigotes sample. Panels C and D show spots with similar expression between promastigote and amastigote, respectively. The numbers refer to the spot identification used in the Supplementary Table 1.

Figure 3. Venn diagram showing proteins expressed in promastigote and amastigote forms from *Leishmania chagasi*.

Figure 4. Proteins from promastigotes forms of *Leishmania chagasi* recognized by IgM from sera of acutely infected dogs. Promastigote extract were fractionated using 18 cm pH 4-7 (left to right) strips in the first dimension and 12% PAGE-SDS gels in the second dimension. (A) Gel stained with Coomassie Blue; (B) gel transferred to nitrocellulose membrane and incubated with sera of dogs in the acute phase of infection; or (C) uninfected dogs; and developed with anti-IgM conjugated with peroxidase.

The spots recognized only by infected animals and identified by MS are highlighted. The numbers refer to the spot identification used in the Supplementary Table 2.

Figure 5. Proteins from promastigotes forms of *Leishmania chagasi* recognized by IgG, IgG1 and IgG2 from sera of chronically infected dogs. Extracts from *Leishmania chagasi* promastigotes were fractionated using 18 cm pH 4-7 (left to right) strips in the first dimension and 12% PAGE-SDS gels in the second dimension. (A) Gel stained with Coomasssie Blue; (B, C and D) transferred to nitrocellulose membrane and incubated with sera of chronically infected, (E) or uninfected dogs. Immunoblot was developed with either anti-total IgG total (B,E), anti-IgG1 (C), or anti-IgG2 (D) conjugated with peroxidase. The spots recognized only by sera of infected animals and identified by MS are highlighted. The numbers refer to the spot identification used in the Supplementary Table 2.

Figure 6. Venn diagram showing intersections of subclasses of immunoglobulins that recognize spots identified by MS after Western blot analysis with sera from infected dogs.

Figure 7. Peptides recognized by sera from chronically infected dogs. To evaluate the antibody reactivity in SPOT synthesis membranes, the relative intensity of the signal was estimated based on comparison of reactivity in immunoblots with sera from chronically infected dogs to the background levels, determined by reactivity with sera from uninfected dogs. A signal was scored as reactive when Relative Intensity (RI) ≥ 2. Only peptides with RI ≥ 2 are shown. Insert shows representative results of immunoblot employing a SPOT synthesis membrane and pools of sera from chronically infected dogs (positive serum) as well as from control uninfected dogs (negative serum). The reaction was revealed with secondary anti-total IgG antibody. Spots 1 to 6 correspond, respectively, to peptides 46C, 40A, 47A, 64E, 66C, and 51A. Protein (GI) and the sequences of peptides are shown in Table 1.

9. REFERENCES

(1) Dye, C. The logic of visceral leishmaniasis control. *Am. J. Trop. Med. and Hyg*.**1996,** 55, 125-130.

(2) Chappuis, F.; Sundar, S.; Hailu, A.; Ghalib, H.; Rijal, S.; Peeling, R.W.; Alvar, J.; Boelaert, M. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat. Rev. Microbiol.* **2007**, 5, 873-882.

(3) Desjeux, P. Leishmaniasis: current situation and new perspectives. *Comp. Immunol. Microbiol. Infect. Dis.* **2004**, 27, 305-318.

(4) Ashford, R.W. The leishmaniases as emerging and reemerging zoonoses. *Int. J. Parasitol.* **2000**, 30, 1269-1281.

(5) de Oliveira, C.I.; Nascimento, I.P.; Barral, A.; Soto, M.; Barral-Netto, M. Challenges and perspectives in vaccination against leishmaniasis. *Parasitol. Int.* **2009**, 58, 319-324.

(6) Matlashewski, G. Leishmania infection and virulence. *Microbiol. Immunol.* **2001**, 190, 37-42.

(7) Dea-Ayuela, M.A.; Rama-Iñiguez, S.; Bolás-Fernández, F. Proteomic analysis of antigens from Leishmania infantum promastigotes. *Proteomics.* **2006**, 6, 4187-4194.

(8) Herrera-Najera, C.; Piña-Aguilar, R.; Xacur-Garcia, F.; Ramirez-Sierra, M.J.; Dumonteil, E. Mining the Leishmania genome for novel antigens and vaccine candidates. *Proteomics.* **2009**, 9, 1293-1301.

(9) El Fakhry, Y.; Ouellette, M.; Papadopoulou, B. A proteomic approach to identify developmentally regulated proteins in Leishmania infantum. *Proteomics.* **2002**, 2, 1007-1117.

(10) Nugent, P.G.; Karsani, S.A.; Wait, R.; Tempero, J.; Smith, D.F. Proteomic analysis of Leishmania mexicana differentiation. *Mol. Biochem. Parasitol.* **2004**,136, 51-62.

(11) McNicoll, F.; Drummelsmith, J.; Müller, M.; Madore, E.; Boilard, N.; Ouellette, M.; Papadopoulou, B. A combined proteomic and transcriptomic approach to the study of stage differentiation in Leishmania infantum. *Proteomics.* **2006**, 6, 3567-3581.

(12) Walker, J.; Vasquez, J.J.; Gomez, M.A.; Drummelsmith, J.; Burchmore, R.; Girard, I.; Ouellette, M. Identification of developmentally-regulated proteins in Leishmania panamensis by proteome profiling of promastigotes and axenic amastigotes. *Mol. Biochem. Parasitol.* **2006**, 147, 64-73.

(13) Leifso, K.; Cohen-Freue, G.; Dogra, N.; Murray, A.; McMaster, W.R. Genomic and proteomic expression analysis of Leishmania promastigote and amastigote life stages: the Leishmania genome is constitutively expressed. *Mol. Biochem. Parasitol*. **2007**, 152, 35-46.

(14) Morales, M.A.; Watanabe, R.; Laurent, C.; Lenormand, P.; Rousselle, J.C.; Namane, A.; Späth, G.F. Phosphoproteomic analysis of Leishmania donovani pro- and amastigote stages. *Proteomics*. **2008**, 8, 350-363.

(15) Fernandes, A.P.; Costa, M.M.S.; Coelho, E.A.; Michalick, M.S.; de Freitas, E.; Melo, M.N.; Tafuri, W.L.; Resende, D.; Hermont, M.V.; Abrantes, C.F.; Gazzinelli, R.T. Protective immunity against challenge with Leishmania (Leishmania) chagasi in beagle dogs vaccinated with recombinant A2 protein. *Vaccine.* **2008**, 26, 5888-5895.

(16) Mauricio, I.L.; Gaunt, M.W.; Stothard, J.R.; Miles, M.A. Genetic typing and phylogeny of the Leishmania donovani complex by restriction analysis of PCR amplified gp63 intergenic regions. *Parasitology.* **2001,** 122, 393-403.

(17) Lukes, J.; Mauricio, I.L.; Schönian, G.; Dujardin, J.C.; Soteriadou, K.; Dedet, J.P.; Kuhls, K.; Tintaya, K.W.; Jirků, M.; Chocholová, E.; Haralambous, C.; Pratlong, F.; Oborník, M.; Horák, A.; Ayala, F.J.; Miles, M.A. Evolutionary and geographical history of the Leishmania donovani complex with a revision of current taxonomy. *Proc Natl Acad Sci U S A*. **2007**, 104, 9375-80.

(18) Romero, G.A., Boelaert, M. Control of visceral leishmaniasis in latin america-a systematic review. PLoS Negl Trop Dis. 2010 Jan 19;4(1):e584.

(19) Schönian, G.; Mauricio, I.; Cupolillo, E. Is it time to revise the nomenclature of Leishmania? *Trends Parasitol*. **2010** 26, 466-9.

(20) Dujardin, J.C.; Campino, L.; Cañavate, C.; Dedet, J.P.; Gradoni, L.; Soteriadou, K.; Mazeris, A.; Ozbel, Y.; Boelaert, M. Spread of vector-borne diseases and neglect of Leishmaniasis, Europe. *Emerg Infect Dis*. **2008** ,14,1013-8.

(21) Gontijo, C.M.F.; Melo, M.N. Leishmaniose visceral no Brasil: quadro atual, desafios e perspectivas. *Rev. bras. epidemiol.* [online]. **2004**, 7, 338-349.

(22) Chang, K.P. Human cutaneous Leishmania in a mouse macrophage line: propagation and isolation of intracellular parasites. *Science* **1980**, 209, 1240–1242.

(23) Neuhoff, V.; Arold, N.; Taube, D.; Ehrhardt, W. Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis*. **1988**, 9, 255-262.

(24) Towbin, H.; Gordon, J. Immunoblotting and dot immunobinding-current status and outlook. *J Immunol Methods.* **1984**, 72, 313-340.

(25) Medzihradszky, K. F. Peptide sequence analysis. *Methods Enzymol.* **2005**, 402, 209–244.

(26) Larsen, M.V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Brunak, S.; Lund, O.; Nielsen, M. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* **2005**, 35, 2295-2303.

(27) Larsen, M.V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Lund, O.; Nielsen, M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics.* **2007,** 8:424.

(28) Lv, H.; Gao, Y.; Wu, Y.; Zhai, M.; Li, L.; Zhu, Y.; Liu, W.; Wu, Z.; Chen, F.; Qi, Y. Identification of a novel cytotoxic T lymphocyte epitope from CFP21, a secreted protein of Mycobacterium tuberculosis. *Immunol Lett*. **2010**, 30, 133, 94-98.

(29) Singh, S.P.; Mishra, B.N. Identification and characterization of merozoite surface protein 1 epitope. *Bioinformation.* **2009,** 17, 1-5.

(30) Larsen, M.V.; Lelic, A.; Parsons, R.; Nielsen, M.; Hoof, I.; Lamberth, K.; Loeb, M.B.; Buus, S.; Bramson, J.; Lund, O. Identification of CD8+ T cell epitopes in the West Nile virus polyprotein by reverse-immunology using NetCTL. *PLoS One*. **2010**, 5:e12697.

(31) Wang, M.; Larsen, M.V.; Nielsen, M.; Harndahl, M.; Justesen, S.; Dziegiel, M.H.; Buus, S.; Tang, S.T.; Lund, O.; Claesson, M.H. HLA Class I Binding 9mer Peptides from Influenza A Virus Induce CD4+ T Cell responses *PLoS One*. **2010**, 5:e10533.

(32) Pérez, C.L.; Larsen, M.V.; Gustafsson, R.; Norström, M.M.; Atlas, A.; Nixon, D.F.; Nielsen, M.; Lund, O.; Karlsson, A.C. Broadly immunogenic HLA class I supertype-restricted elite CTL

epitopes recognized in a diverse population infected with different HIV-1 subtypes. *Immunol*. **2008,** 180(7):5092-100.

(33) Tang, S.T.; Wang, M.; Lamberth, K.; Harndahl, M.; Dziegiel, M.H.; Claesson, M.H.; Buus, S.; Lund, O. MHC-I-restricted epitopes conserved among variola and other related orthopoxviruses are recognized by T cells 30 years after vaccination. *Arch Virol*. **2008**, 153,1833-44.

(34) Larsen, J.E.; Lund, O.; Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2006**, 2, 2-5.

(35) Frank, R.; Overwin, H. Spot-synthesis: epitope analysis with arrays of synthetic peptides prepared on cellulose membranes. In Methods in Molecular Biology. Epitope Mapping Protocols, Morris GE (ed.), Humana Press: Totowa, **1996**, 66, 149–169.

(36) Frank, R. Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support, Tetrahedron, **1992,** 48, 9217-9232.

(37) Soutullo, A.; Santi, M.N.; Perin, J.C.; Beltramini, L.M.; Borel, I.M.; Frank, R.; Tonarelli, G.G. Systematic epitope analysis of the p26 EIAV core protein. *J Mol Recogni*t. **2007**, 20, 227-237.

(38) Sundar, S.; Rai, M. Laboratory diagnosis of visceral leishmaniasis. *Clin. Diagn. Lab. Immunol.* **2002**, 9, 951-958.

(39) Rouf, M.A. ; Rahman, M.E. ; Islam, M.N. ; Ferdous, N.N. ; Hossain, M.A. Sensitivity, specificity and predictive values of immunochromatographic strip test in diagnosis of childhood kala-azar. *Mymensingh Med. J.* **2009**, 18(1 Suppl), S1-S5.

(40) Mojtahedi, Z.; Clos, J.; Kamali-Sarvestani, E. Leishmania major: identification of developmentally regulated proteins in procyclic and metacyclic promastigotes. *Exp. Parasitol.* **2008**, 119, 422-429.

(41) Cuervo, P.; de Jesus, J.B.; Junqueira, M.; Mendonça-Lima, L.; González, L.J.; Betancourt, L.; Grimaldi, G. Jr.; Domont, G.B.; Fernandes, O.; Cupolillo, E. Proteome analysis of Leishmania (Viannia) braziliensis by two-dimensional gel electrophoresis and mass spectrometry. Mol Biochem Parasitol. **2007**, 154, 6-21.

(42) Paape, D.; Lippuner, C.; Schmid, M.; Ackermann, R.; Barrios-Llerena, M.E.; Zimny-Arndt, U.; Brinkmann, V.; Arndt, B.; Pleissner, K.P.; Jungblut, P.R.; Aebischer, T. Transgenic, fluorescent Leishmania mexicana allow direct analysis of the proteome of intracellular amastigotes. *Mol. Cell Proteomics.* **2008**, 7, 1688-1701.

(43) Paape, D.; Barrios-Llerena, M.; Le Bihan, T.; Mackay, L.; Aebischer T. Gel free analysis of the proteome of intracellular Leishmania mexicana. *Mol Biochem Parasitol*. **2010**, 169, 108-114.

(44) Forgber, M.; Basu, R.; Roychoudhury, K.; Theinert, S.; Roy, S.; Sundar, S.; Walden, P. Mapping the antigenicity of the parasites in Leishmania donovani infection by proteome serology. *PLoS One*. **2006**, 1, e40.

(45) Gupta, S.K.; Sisodia, B.S.; Sinha, S.; Hajela, K.; Naik, S.; Shasany, A.K.; Dube A. Proteomic approach for identification and characterization of novel immunostimulatory proteins from soluble antigens of Leishmania donovani promastigotes. *Proteomics.* **2007**, 7, 816-823.

(46) Winkler, D.F.; Campbell, W.D. The spot technique: synthesis and screening of peptide macroarrays on cellulose membranes. *Methods Mol. Biol.* **2008**, 494, 47-70.

(47) Volkmer, R. Synthesis and application of peptide arrays: quo vadis SPOT technology. *Chembiochem.* **2009**, 10, 1431-1442.

(48) Reineke, U.; Sabat, R. Antibody epitope mapping using SPOT peptide arrays. *Methods Mol. Biol.* **2009**, 524, 145-167.

(49) Drummelsmith, J.; Brochu, V.; Girard, I.; Messier, N.; Ouellette, M. Proteome mapping of the protozoan parasite Leishmania and application to the study of drug targets and resistance mechanisms. *Mol. Cell Proteomics.* **2003**, 2, 146-55.

(50) Kedzierski, L.; Zhu, Y.; Handman, E. Leishmania vaccines: progress and problems. *Parasitology*. **2006**, 133, S87-112.

(51) Palatnik-de-Sousa, C.B.; Silva-Antunes, I.; Morgado, A. de A.; Menz, I.; Palatnik, M.; Lavor, C. Decrease of the incidence of human and canine visceral leishmaniasis after dog vaccination with Leishmune in Brazilian endemic areas. Vaccine**, 2009**, 27, 3505-3512.

(52) Collins, F.M. Vaccines and cell-mediated immunity. *Bacteriol. Rev*. 1974, 38, 371-402.

(53) Mohabatkar, H. Prediction of epitopes and structural properties of Iranian HPV-16 E6 by bioinformatics methods. *Asian Pac. J. Cancer Prev*. **2007**, *8,* 602-606.

(54) Belkaid, Y.; Von Stebut, E.; Mendez, S.; Lira, R.; Caler, E.; Bertholet, S.; Udey, M.C.; Sacks, D. CD8+ T cells are required for primary immunity in C57BL/6 mice following low-dose, intradermal challenge with Leishmania major. *J. Immunol.,* **2002,** 168, 3992-4000.

(55) Peters, B.; Bulik, S.; Tampe, R.; Endert, P.M.V.; Holzhutter, H.G. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* **2003**, 171, 1741-1749.

(56) Wagner, J.L. Molecular Organization of the Canine Major Histocompatibility Complex. *Journal of Heredity.* **2003**, 94, 23–26.

(57) McMahon-Pratt, D.; Kima, P.E.; Soong, L. Leishmania amastigote target antigens: the challenge of a stealthy intracellular parasite. *Parasitol Today*. **1998**, 14, 31-34.

(58) Mary, C.; Lamouroux, D.; Dunan, S.; Quilici, M. Western blot analysis of antibodies to Leishmania infantum antigens: potential of the 14-kD and 16-kD antigens for diagnosis and epidemiologic purposes. *Am. J. Trop. Med. Hyg*. **1992**, 47, 764–771.

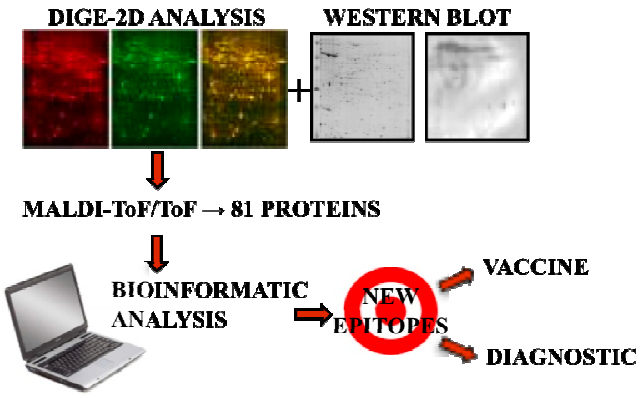
(59) Mancianti, F.; Falcone, M. L.; Giannelli, C.; Poli, A. Comparison between an enzyme-linked immunosorbent assay using a detergent-soluble Leishmania infantum antigen and indirect immunofluorescence for the diagnosis of canine leishmaniosis. *Vet. Parasitol.* **1995**, 59, 13–21.

(60) Carrera, L.; Fermin, M. L.; Tesouro, M.; Garcia, P.; Rollán, E.; González, J.L.; Méndez, S.; Cuquerella, M.; Alunda, J.M. Antibody response in dogs experimentally infected with Leishmania infantum: infection course antigen markers. *Exp. Parasitol.* **1996**, 82, 139–146.

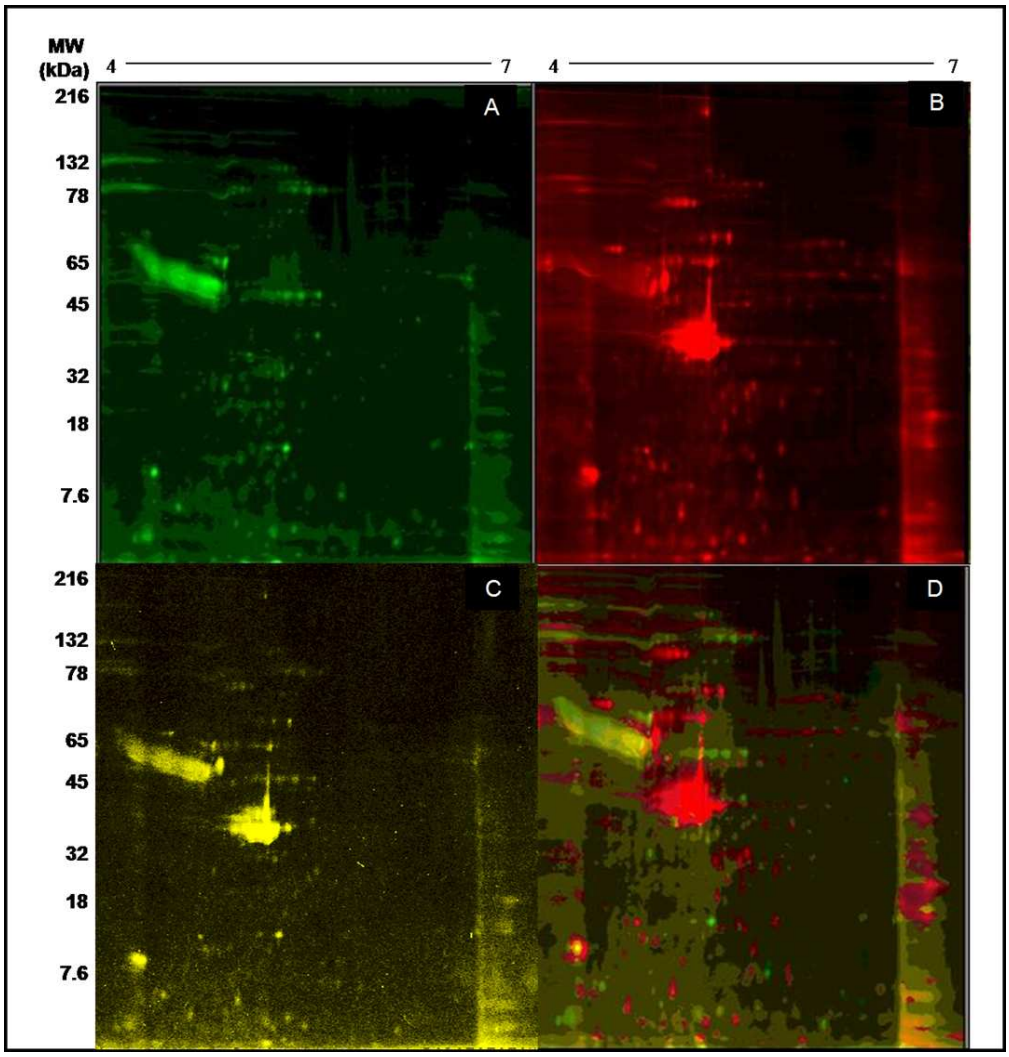
(61) Aisa, M.J.; Castillejo, S.; Gallego, M.; Fisa, R.; Riera, M.C.; de Colmenares, M.; Torras, S.; Roura, X.; Sentis, J.; Portus, M. Diagnostic potential of Western blot analysis of sera from dogs with leishmaniasis in endemic areas and significance of the pattern. *Am. J. Trop. Med. Hyg.* **1998**, 58, 154–159.

(62) Basu, R.; Roy, S.; Walden, P. HLA class I-restricted T cell epitopes of the kinetoplastid membrane protein-11 presented by Leishmania donovani-infected human macrophages. *J. Infect. Dis.* **2007**, 195, 1373-1380.

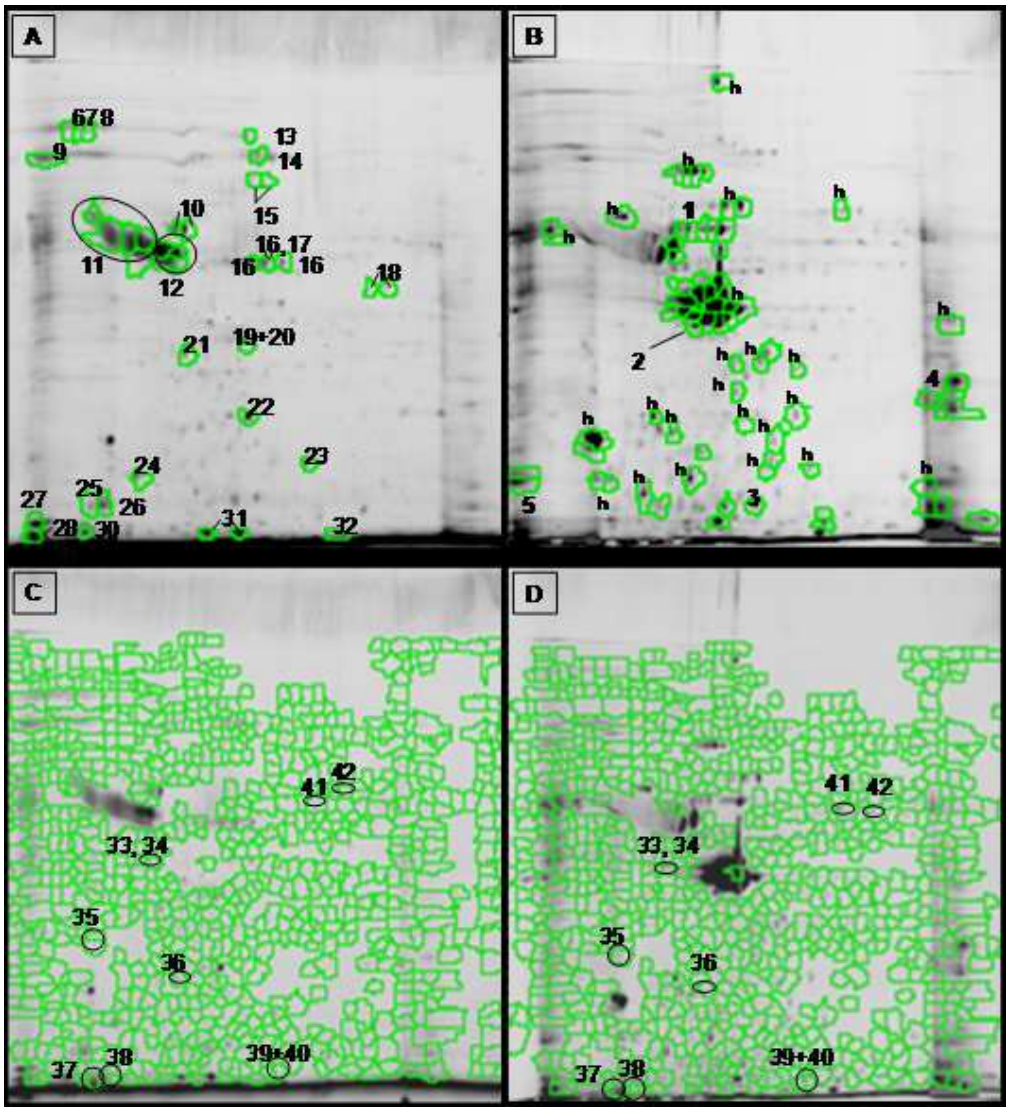
Table of Contents (TOC)- Synopsis



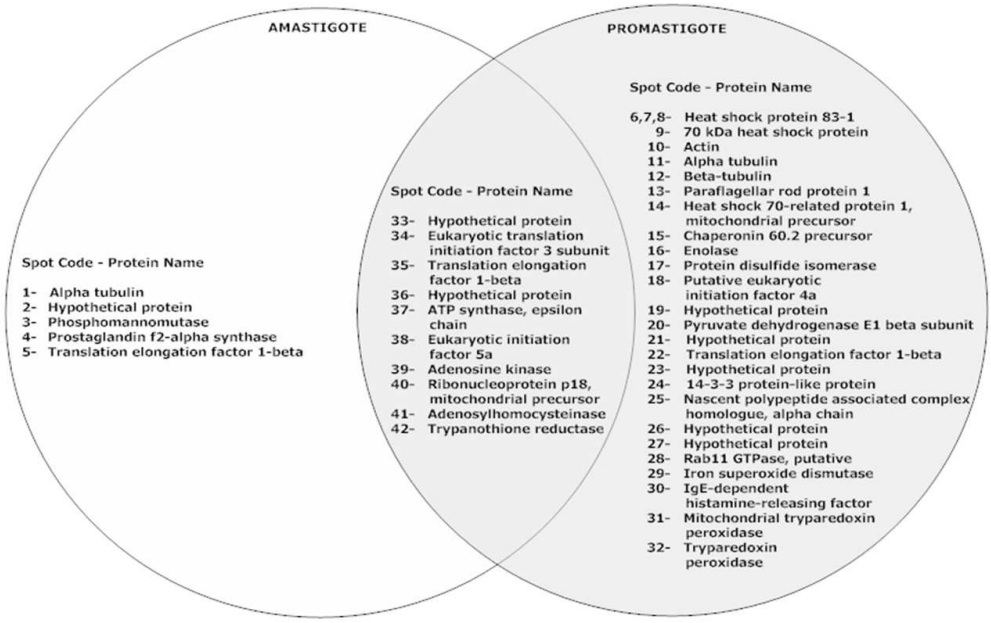
Five amastigote-, 25 promastigote-specific, and 10 common proteins were disclosed by DIGE. Furthermore, 41 proteins were indentified in an immunoblot employing 2-DE and sera from infected dogs. All proteins were mapped for B cell epitopes, and 25 peptides shown of interest to diagnostic tests for visceral leishmaniasis. In addition, various peptide/proteins were identified as potential T cell antigens. Hence, new antigens for diagnostic tests and vaccines were identified in a proteomic analysis of *L. chagasi*.



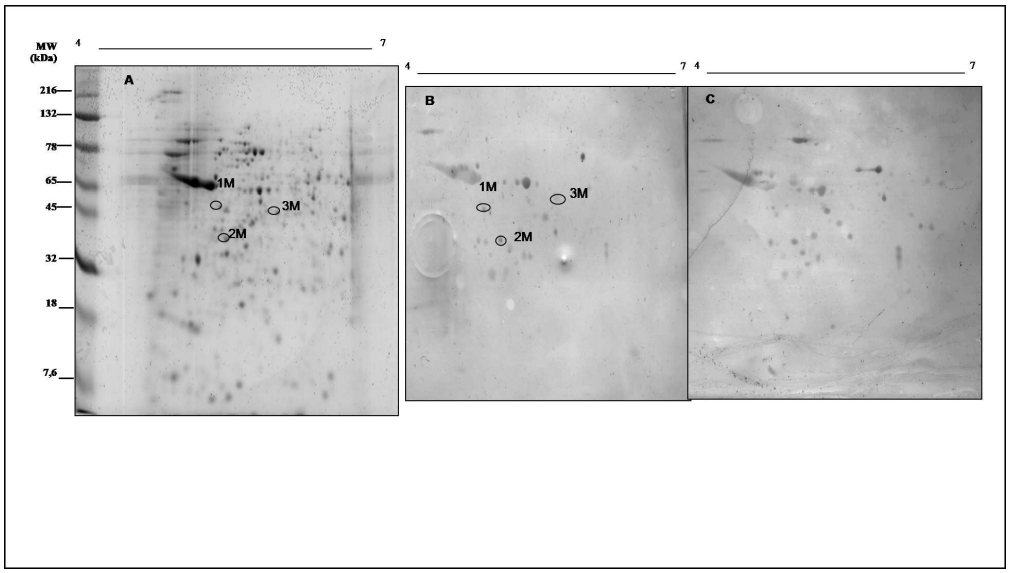
2D-DIGE analysis of promastigote and amastigotes extracts from Leishmania chagasi. Proteins of promastigotes and amastigotes forms, and a pooled internal standard were labeled with CyDyes Cy3, Cy5 and Cy2, respectively, mixed and separated on a 2D gel using 18 cm pH 4-7 (left to right) strips in the first dimension and 12% PAGE-SDS gels in the second dimension. The standards for Cy2 as well as molecular weight and pI were defined in the same gel stained with Coomassie Blue and software analysis, respectively. Gels were scanned to obtain single images of (A) promastigotes (Cy3, green), (B) amastigotes (Cy5, red) or (C) the internal standard (Cy2, yellow). (D) An overlay of the two dyes (Cy3, Cy5) is shown in yellow.
182x190mm (150 x 150 DPI)



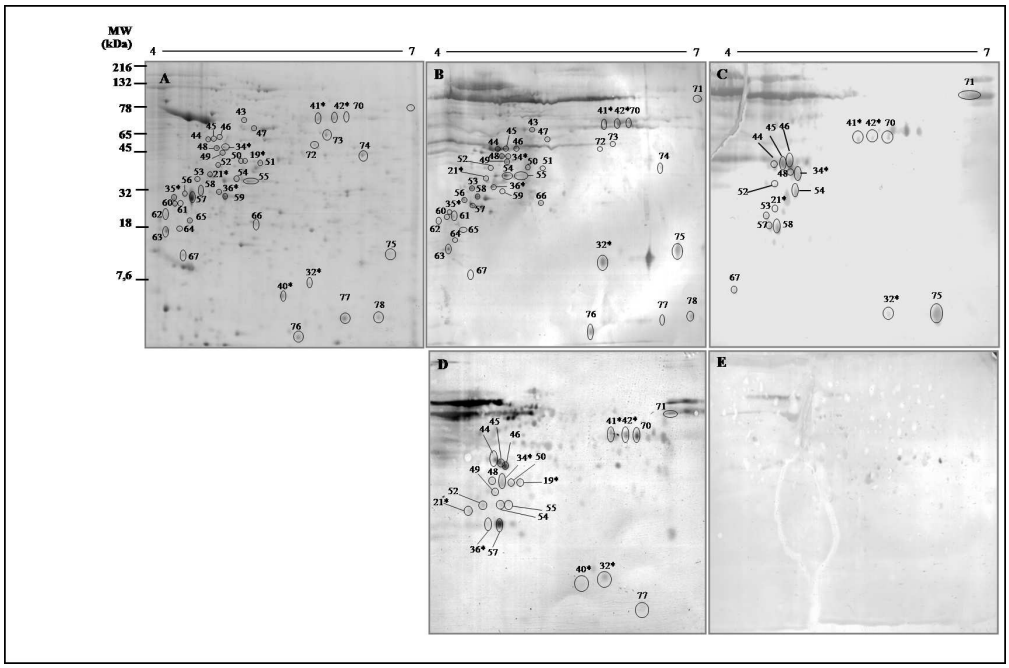
Proteins highly expressed by promastigotes (A) and amastigotes (B) forms derived from Leishmania chagasi. The level of protein expression was determined by using ImageMaster 2D Platinum 6.0® software, please see details in the Material and Methods section. The "h" letter (in 2B) represents hamster-derived proteins present in amastigotes sample. Panels C and D show spots with similar expression between promastigote and amastigote, respectively. The numbers refer to the spot identification used in the Supplementary Table 1.
151x166mm (89 x 89 DPI)



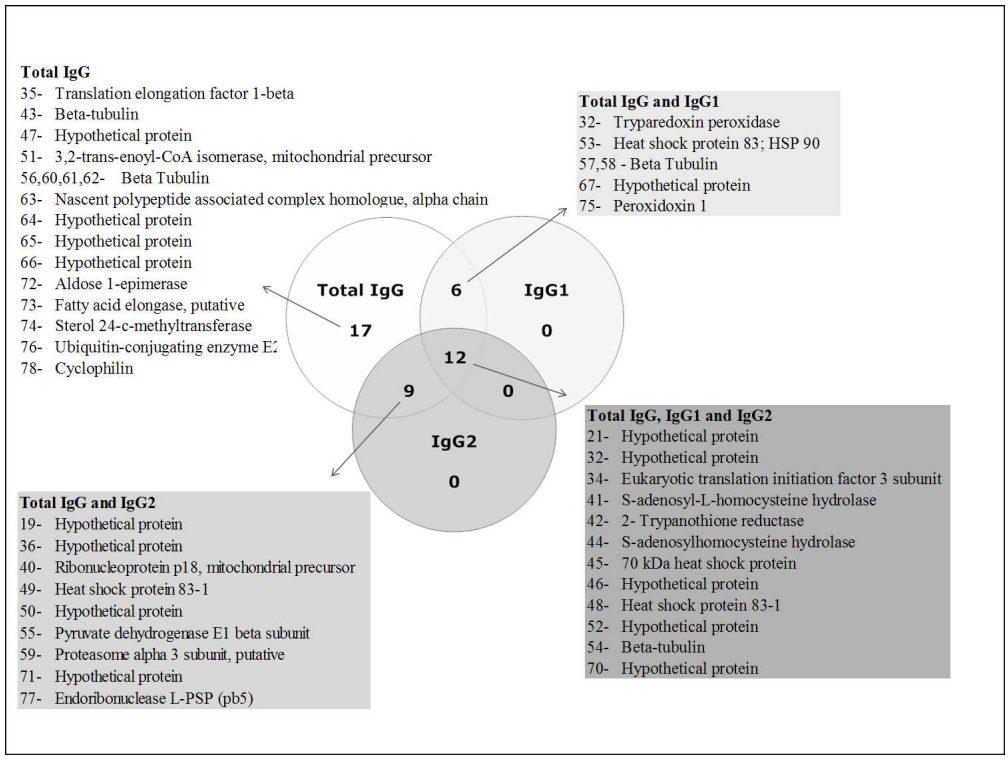
Venn diagram showing proteins expressed in promastigote and amastigote forms from Leishmania chagasi.
252x162mm (129 x 125 DPI)



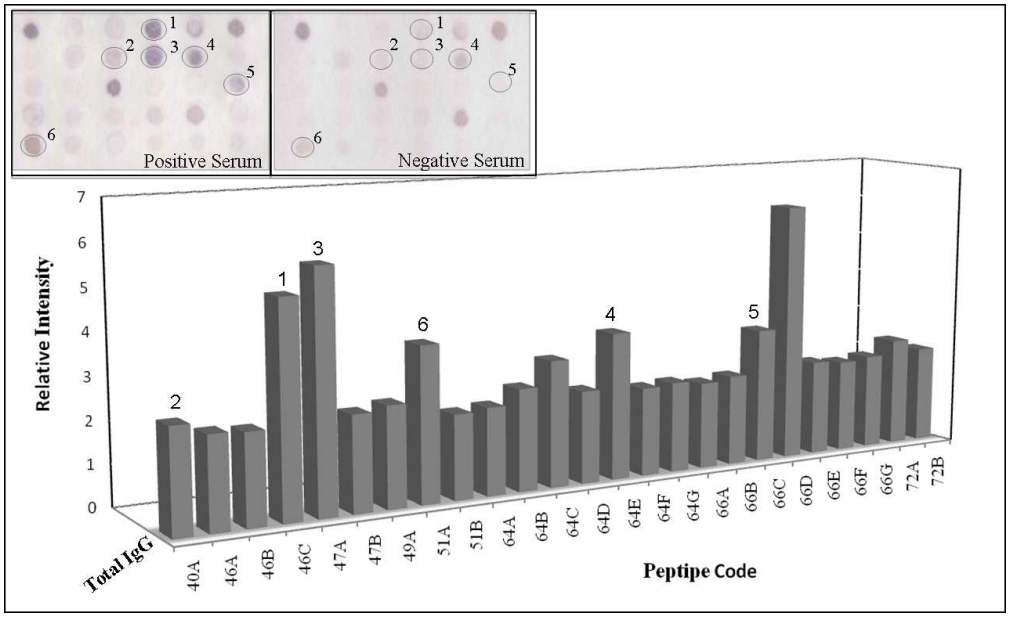
Proteins from promastigotes forms of Leishmania chagasi recognized by IgM from sera of acutely infected dogs. Promastigote extract were fractionated using 18 cm pH 4-7 (left to right) strips in the first dimension and 12% PAGE-SDS gels in the second dimension. (A) Gel stained with Coomassie Blue; (B) gel transferred to nitrocellulose membrane and incubated with sera of dogs in the acute phase of infection; or (C) uninfected dogs; and developed with anti-IgM conjugated with peroxidase. The spots recognized only by infected animals and identified by MS are highlighted. The numbers refer to the spot identification used in the Supplementary Table 2.
329x185mm (150 x 150 DPI)



Proteins from promastigotes forms of Leishmania chagasi recognized by IgG, IgG1 and IgG2 from sera of chronically infected dogs. Extracts from Leishmania chagasi promastigotes were fractionated using 18 cm pH 4-7 (left to right) strips in the first dimension and 12% PAGE-SDS gels in the second dimension. (A) Gel stained with Coomasssie Blue; (B, C and D) transferred to nitrocellulose membrane and incubated with sera of chronically infected, (E) or uninfected dogs. Immunoblot was developed with either anti-total IgG total (B,E), anti-IgG1 (C), or anti-IgG2 (D) conjugated with peroxidase. The spots recognized only by sera of infected animals and identified by MS are highlighted. The numbers refer to the spot identification used in the Supplementary Table 2.
290x191mm (150 x 150 DPI)



Venn diagram showing intersections of subclasses of immunoglobulins that recognize spots identified by MS after Western blot analysis with sera from infected dogs.
254x190mm (150 x 150 DPI)



Peptides recognized by sera from chronically infected dogs. To evaluate the antibody reactivity in SPOT synthesis membranes, the relative intensity of the signal was estimated based on comparison of reactivity in immunoblots with sera from chronically infected dogs to the background levels, determined by reactivity with sera from uninfected dogs. A signal was scored as reactive when Relative Intensity (RI) ≥ 2. Only peptides with RI ≥ 2 are shown. Insert shows representative results of immunoblot employing a SPOT synthesis membrane and pools of sera from chronically infected dogs (positive serum) as well as from control uninfected dogs (negative serum). The reaction was revealed with secondary anti-total IgG antibody. Spots 1 to 6 correspond, respectively, to peptides 46C, 40A, 47A, 64E, 66C, and 51A. Protein (GI) and the sequences of peptides are shown in Table 1.
216x131mm (150 x 150 DPI)

Anexo 4 - Manuscrito em preparação

EXPRESSION AND ANTIGENIC PROFILE OF MASP FAMILY OF *TRYPANOSOMA CRUZI* IN ACUTE PHASE OF EXPERIMENTAL INFECTION.

Lopes dos Santos, Sara 1; Freitas, Leandro Martins 1; Lobo, Francisco 1; Luiz, Gabriela 1; Mendes, Tiago Antônio de Oliveira 1; Chiari, Egler 1; Gazzineli, Ricardo Tostes 2; Teixeira, Santuza Maria. 2; Fujiwara, Ricardo Toshio 1; Bartholomeu, Daniella Castanheira 1.

1. Parasitology Department, UFMG, Belo Horizonte, Minas Gerais, Brazil.

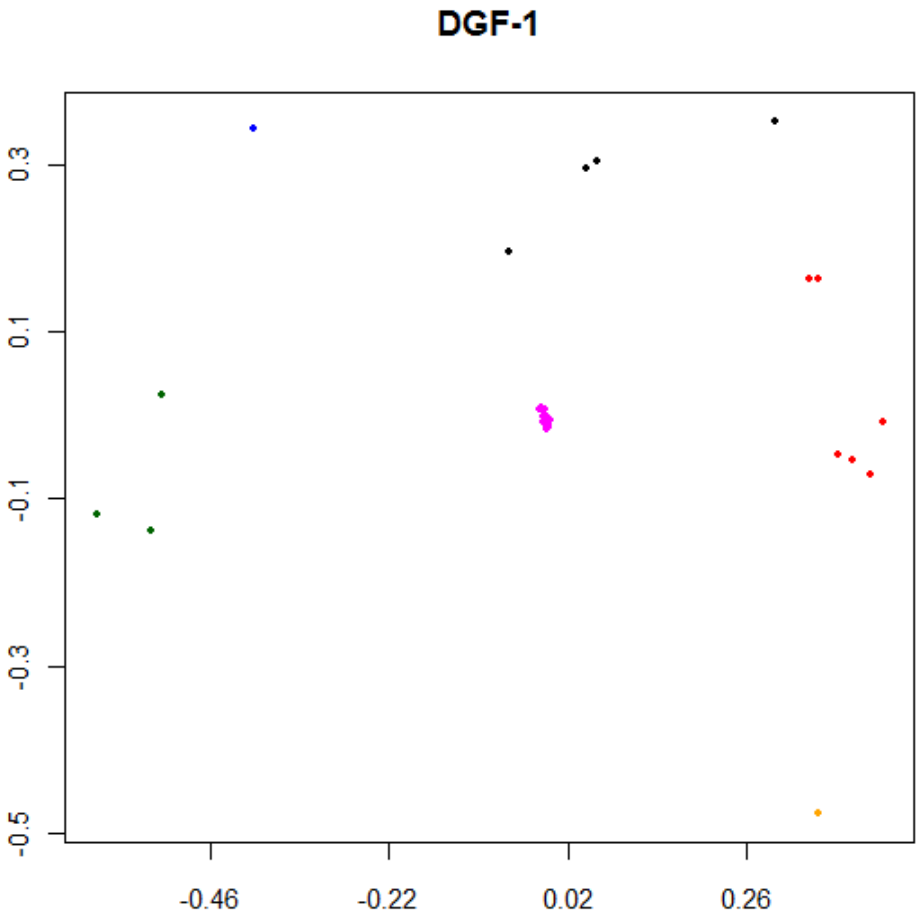
2. Biochemistry and Immunology Department, UFMG, Belo Horizonte, Minas Gerais, Brazil

ABSTRACT

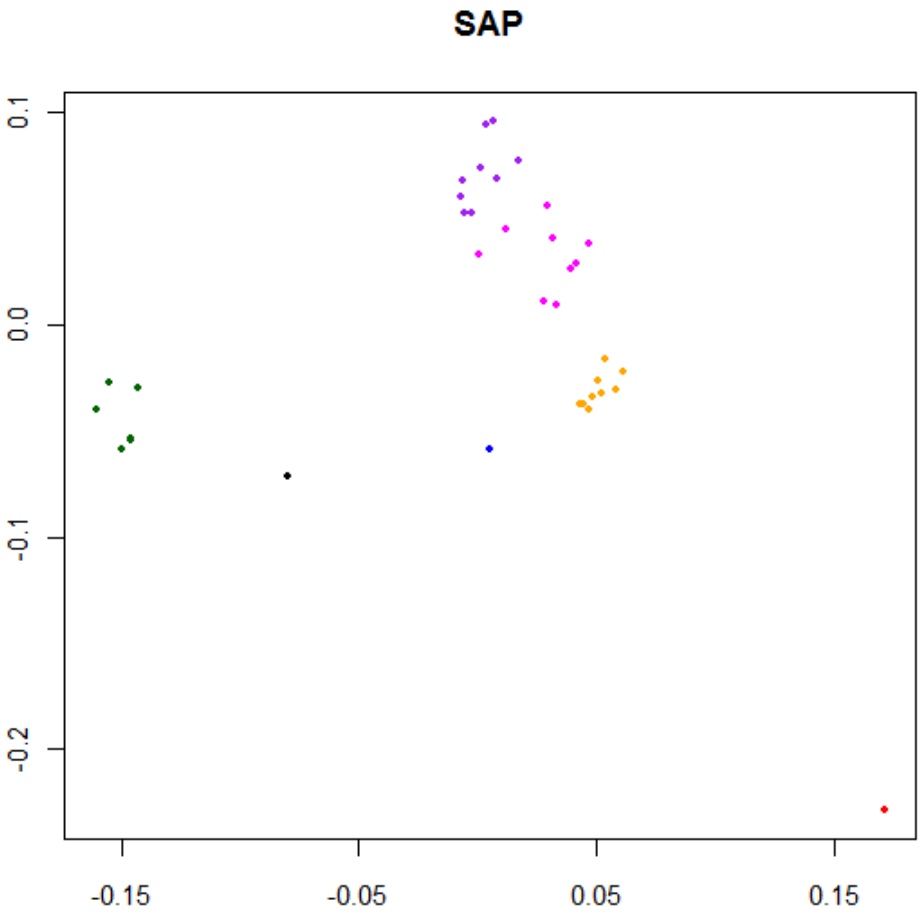
Trypanosoma cruzi is the etiological agent of Chagas disease, a debilitating illness that is a major cause of morbidity and mortality in many parts of Latin America. A major finding of the *Trypanosoma cruzi* genome project was the discovery of a novel multigene family encoding mucin-associated surface protein (MASP). Although MASP function is unknown, its extended sequence variability, repetitive nature and localization at the parasite surface suggest this family participates in parasite-host interactions such as host cell adhesion/invasion or/and immune evasion mechanisms. In order to investigate these hypotheses we have analyzed MASP expression profile of culture trypomastigotes derived from distinct host cell lineages and of bloodstream parasites after sequential passages in mice. Also, in order to investigate the MASP antigenic profile we have performed B cell linear epitope prediction on the MASP proteome and designed a SPOT peptide array with 200 putative epitopes. This peptide array was screened with sera from acutely infected mice. The results from the analysis of 7 expression libraries followed by validation by real-time RT-PCR suggested MASP expression may change depending on host cell and also after sequential passages in mice. Immunoblot experiments showed that mice IgG and IgM are reactive against several MASP peptides during acute phase and suggest differences in the antigenic profile between sequential passages. This is the first report on MASP antigenic property and profile. We speculate that variations in the large repertoire of potentially antigenic peptides derived from MASP family may favor the parasite escape of immune response during the acute phase of infection.

Anexo 5 – Painéis da figura 18. Projeções MDS para sequências nucleotídicas.

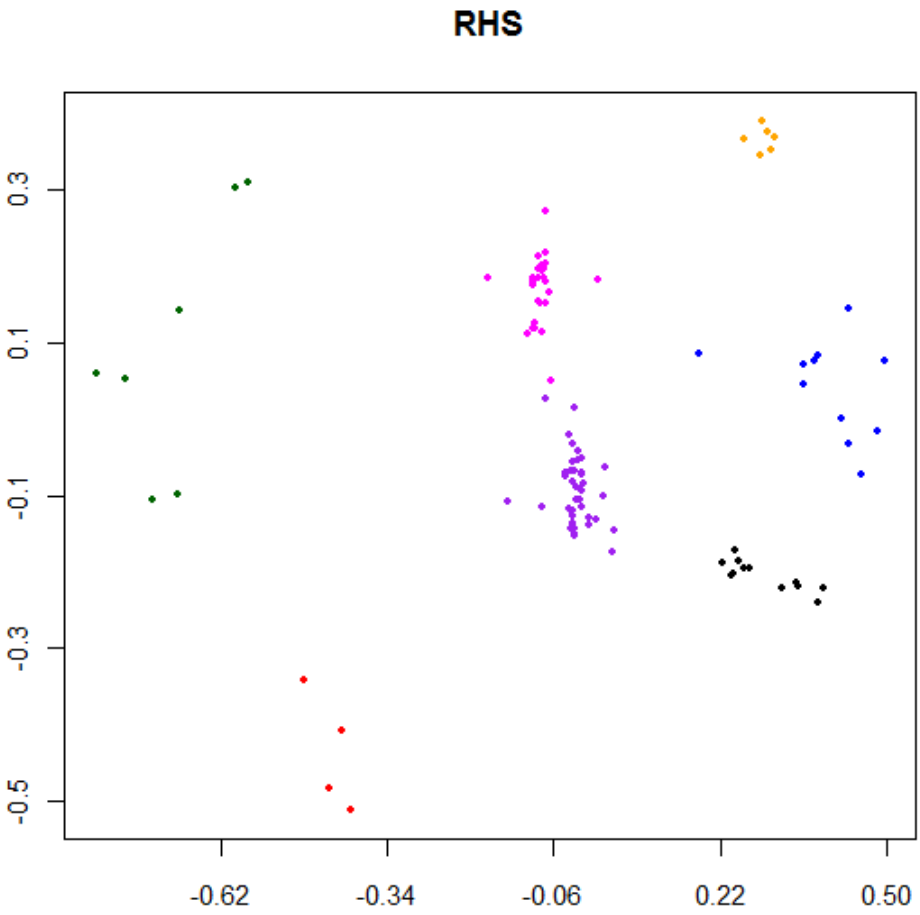
Projeção MDS para sequências nucleotídicas da família DGF-1.



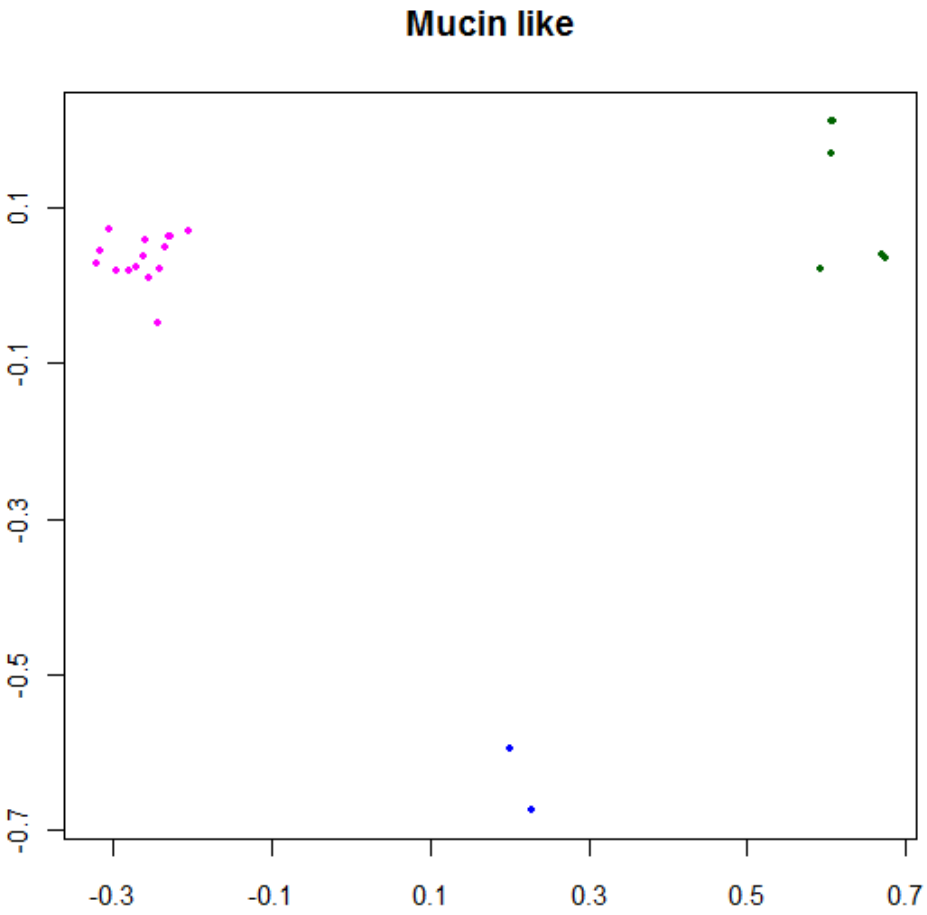
Projeção MDS para sequências nucleotídicas da família SAP.



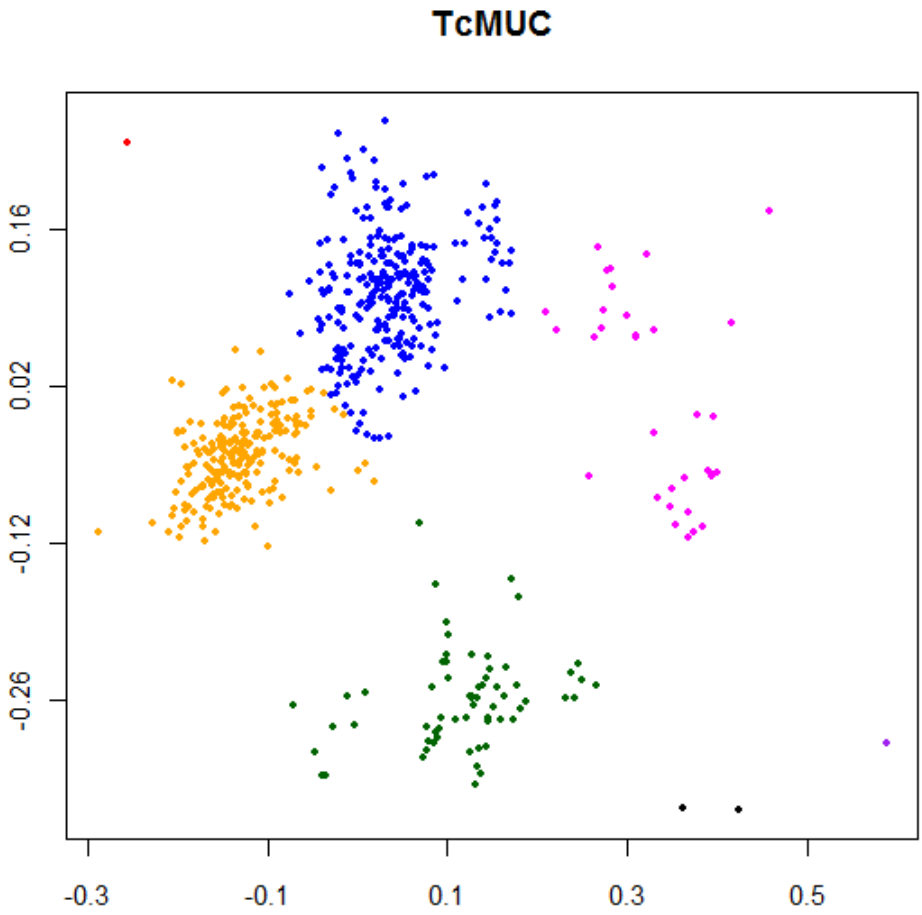
Projeção MDS para sequências nucleotídicas da família RHS.



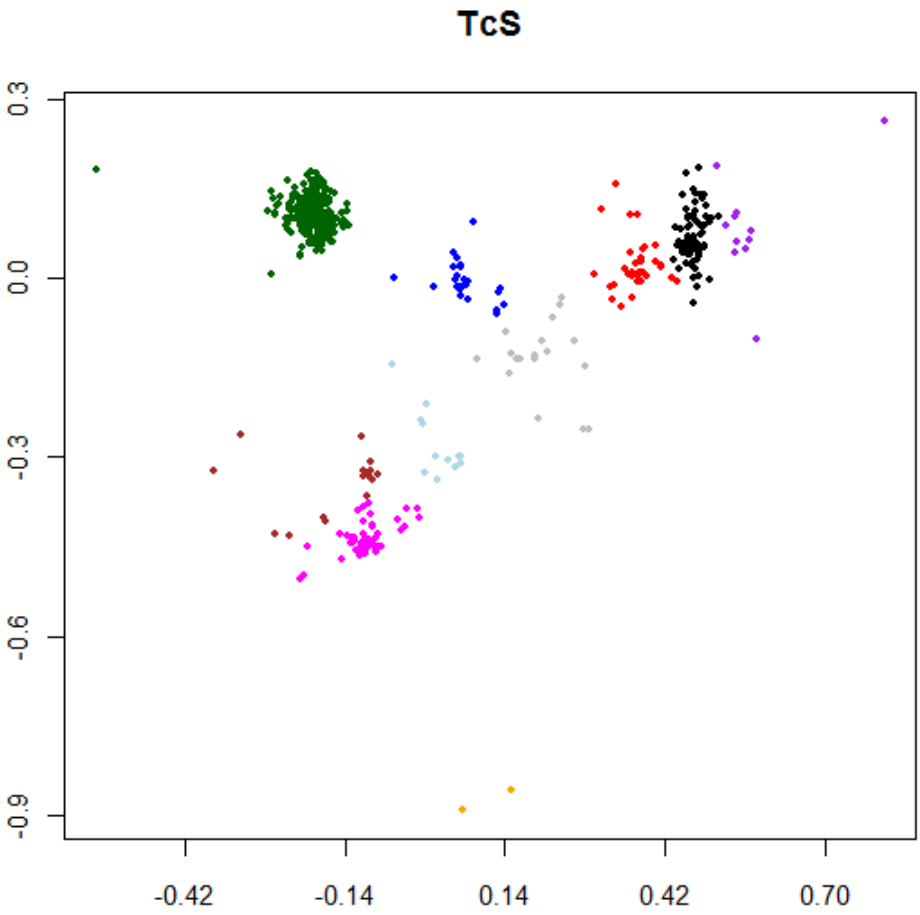
Projeção MDS para sequências nucleotídicas da família mucin like.



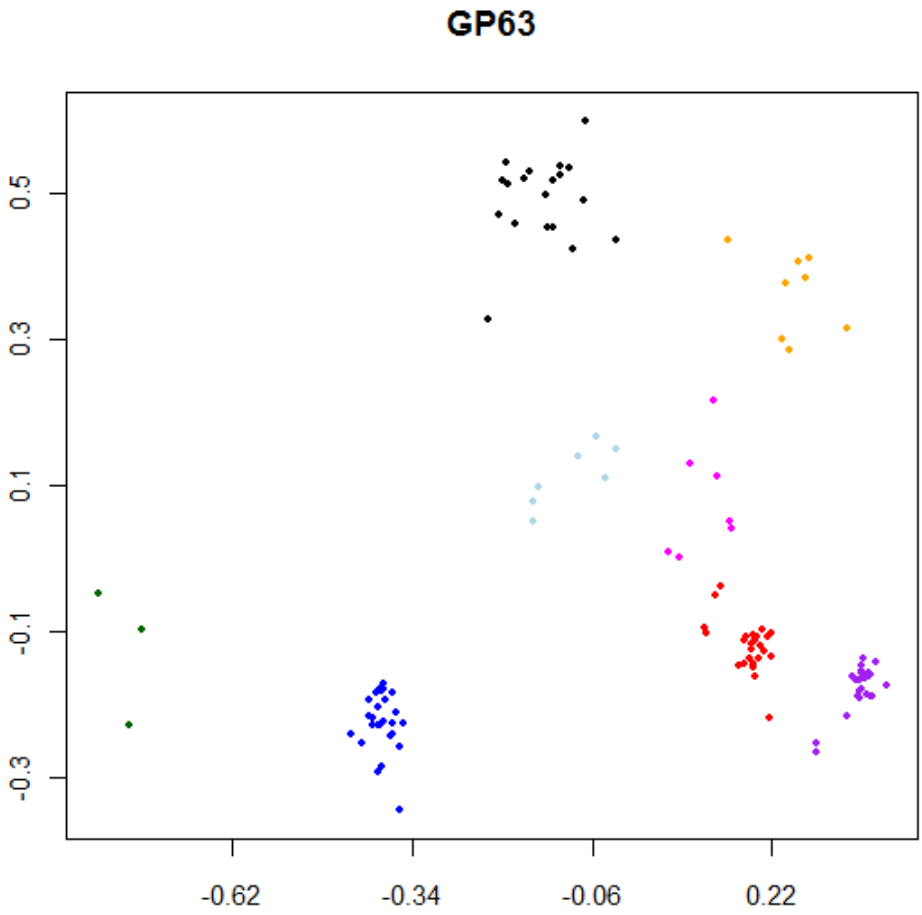
Projeção MDS para sequências nucleotídicas da família TcMUC.



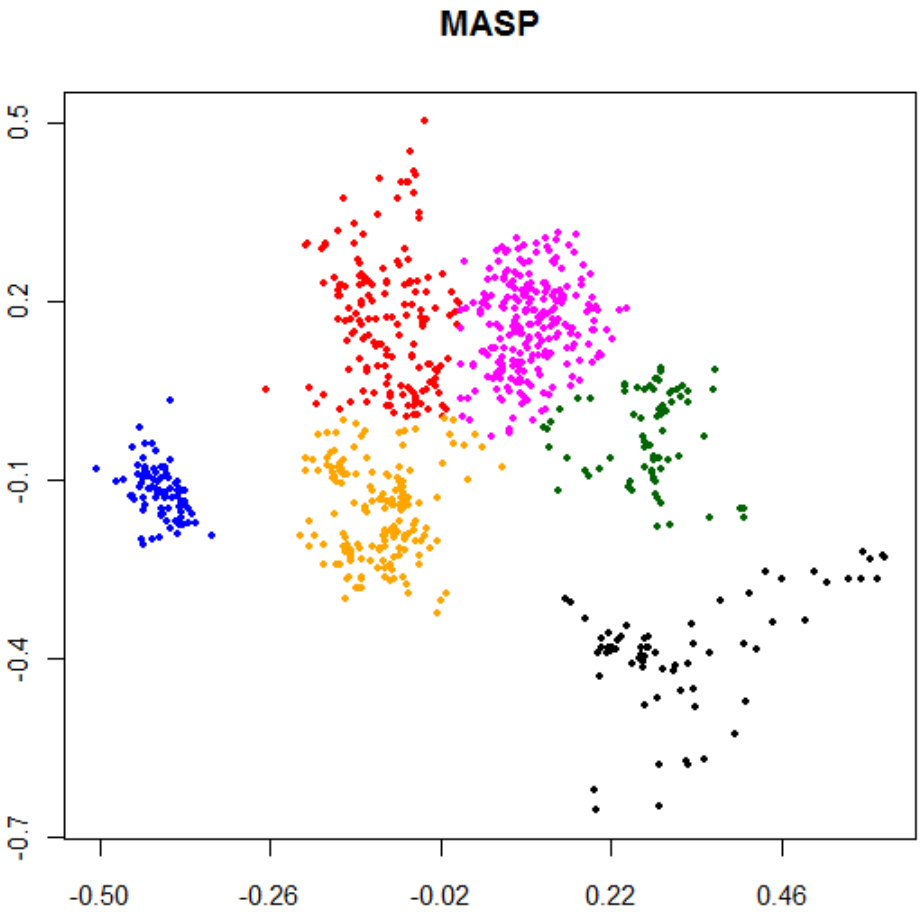
Projeção MDS para sequências nucleotídicas da família TcS.



Projeção MDS para sequências nucleotídicas da família GP63.

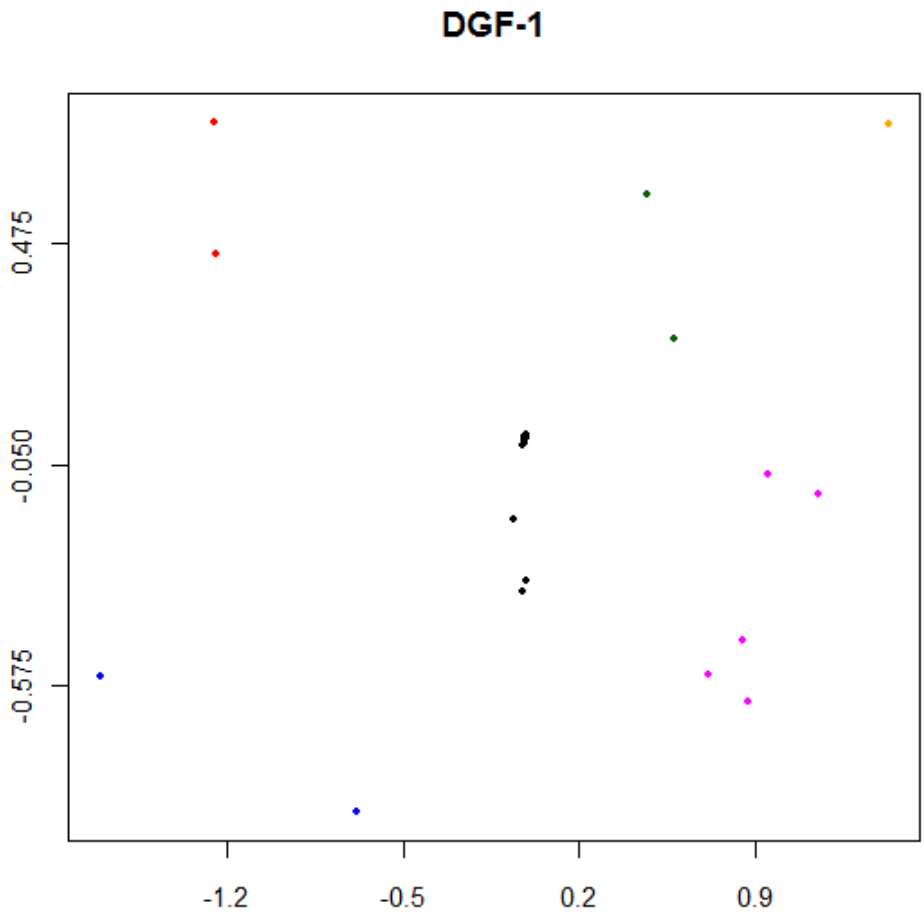


Projeção MDS para sequências nucleotídicas da família MASP.

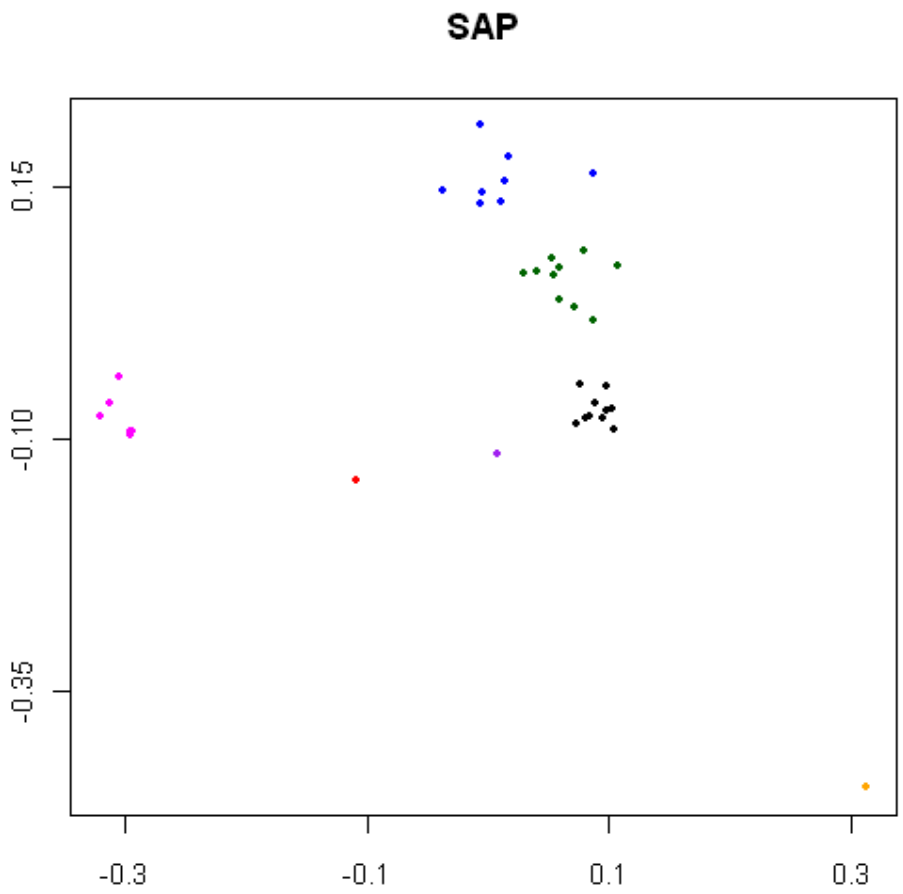


Anexo 6 – Painéis da figura 19. Projeções MDS para sequências de proteína.

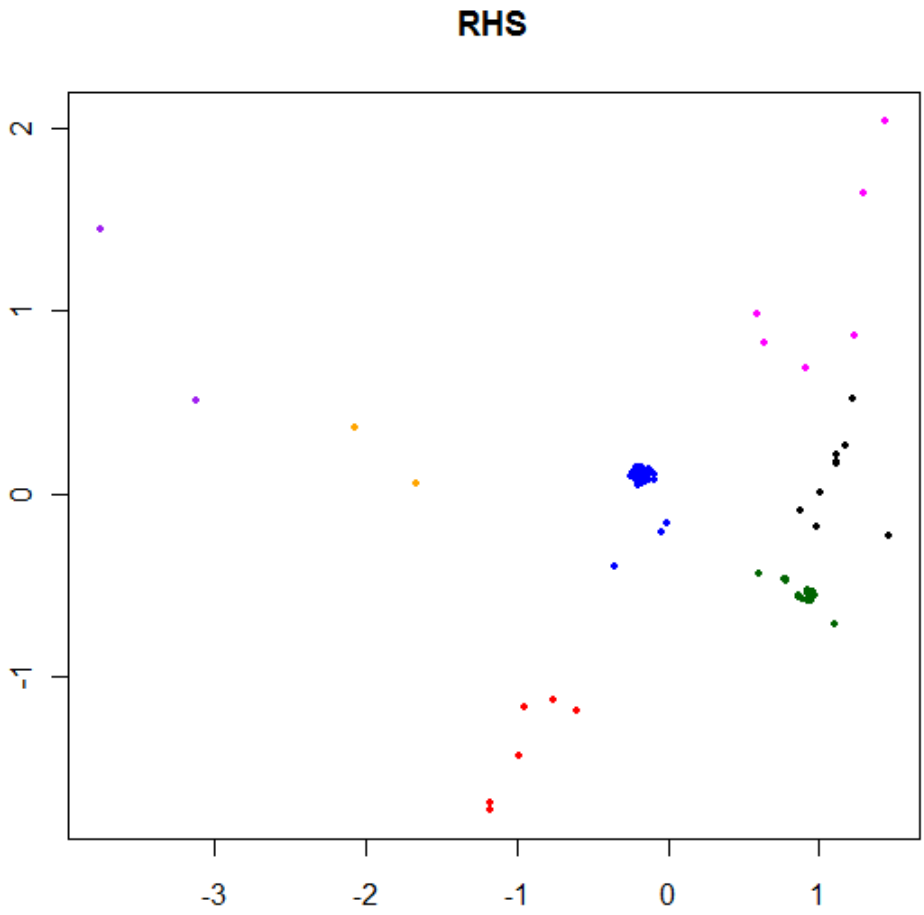
Projeções MDS para sequências de proteína da família DGF-1.



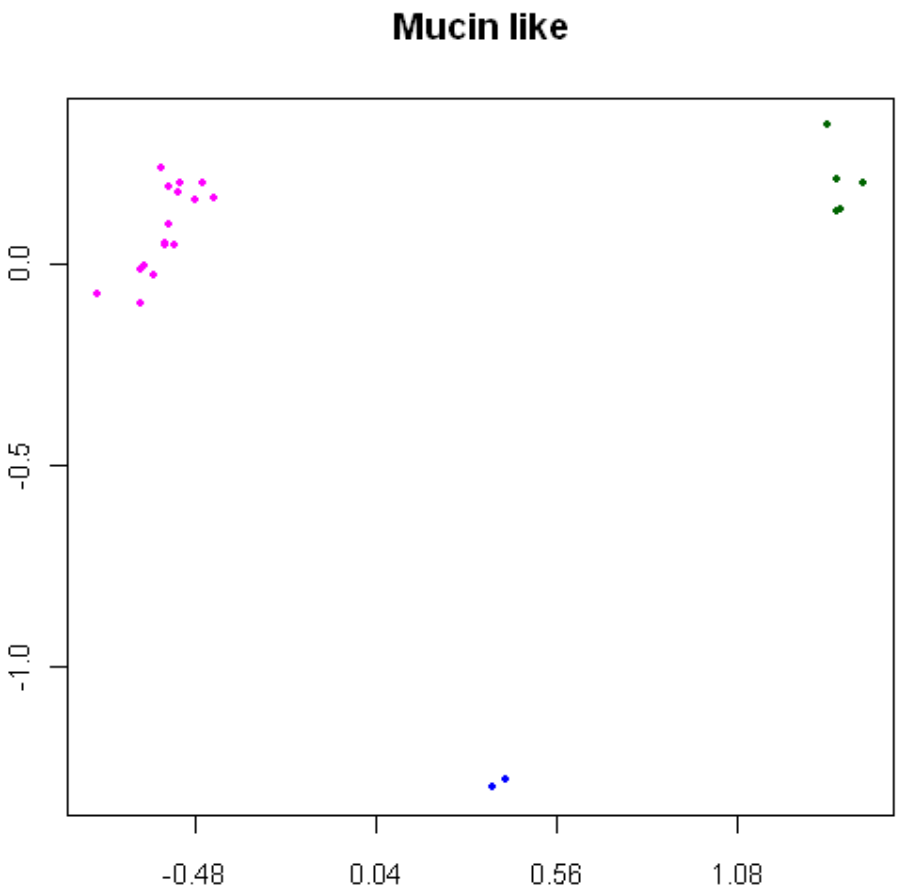
Projeções MDS para sequências de proteína da família SAP.



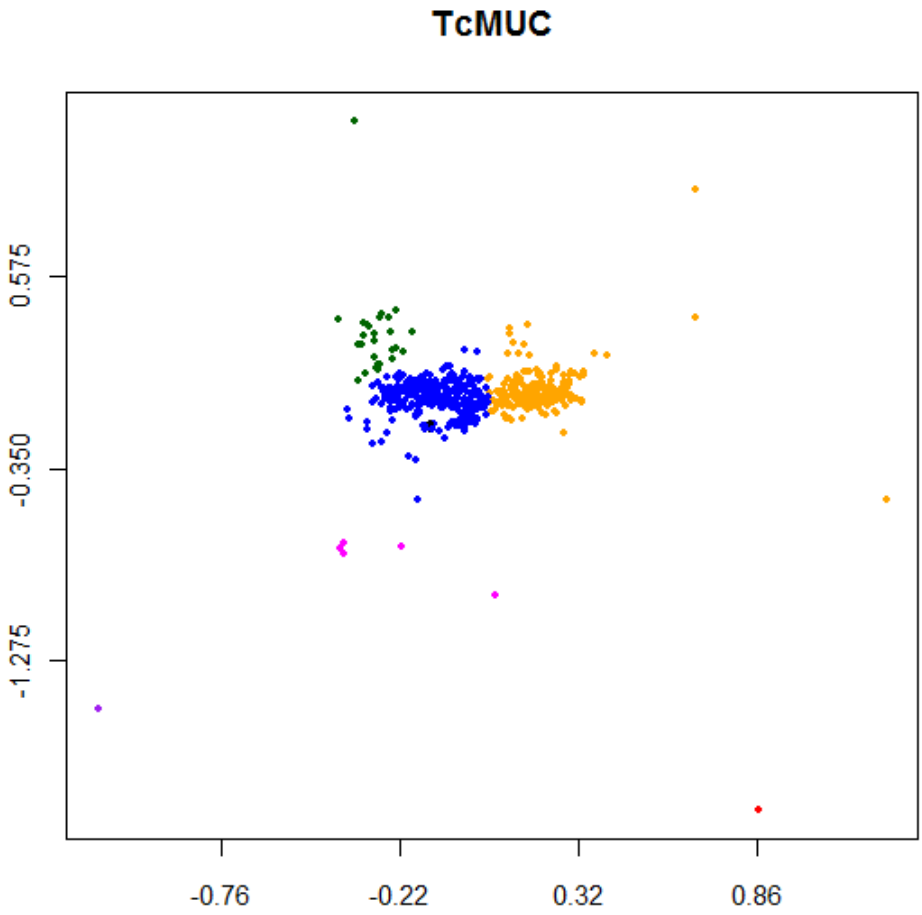
Projeções MDS para sequências de proteína da família RHS.



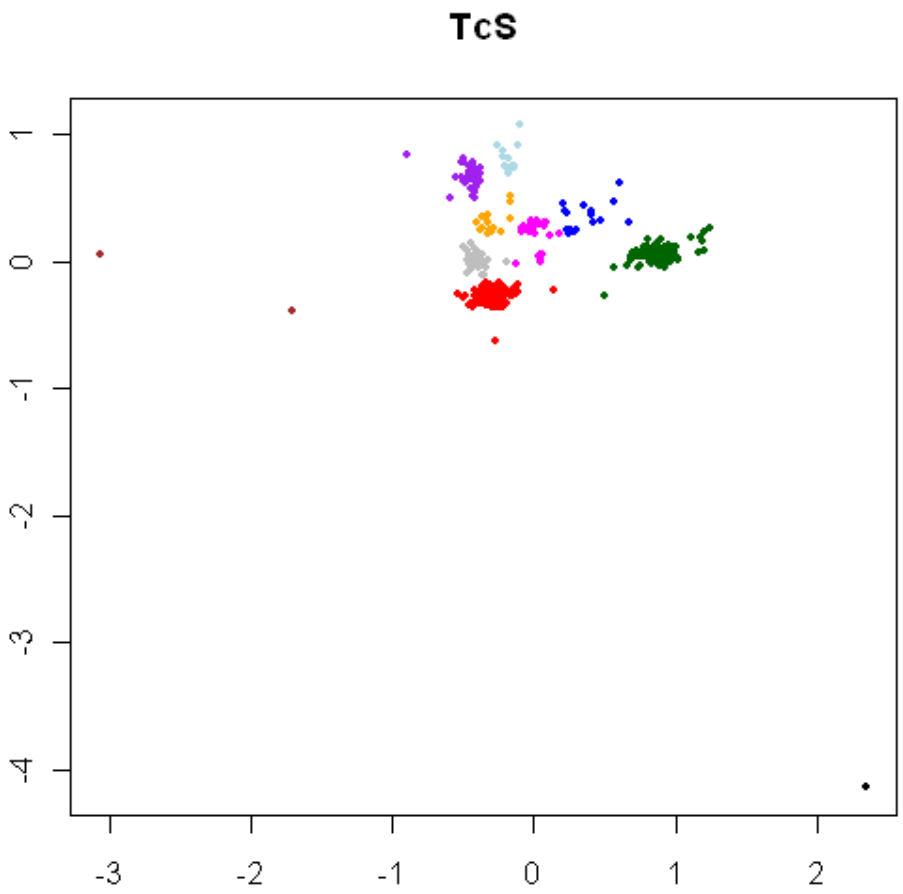
Projeções MDS para sequências de proteína da família mucin like.



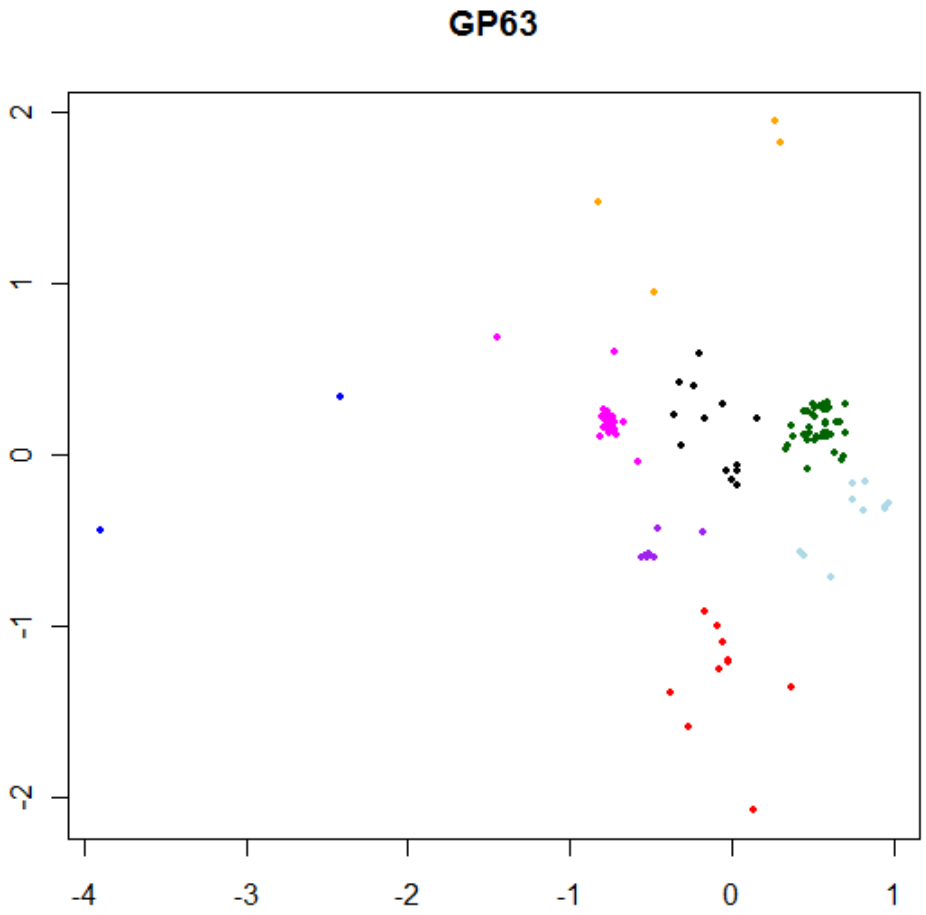
Projeções MDS para sequências de proteína da família TcMUC



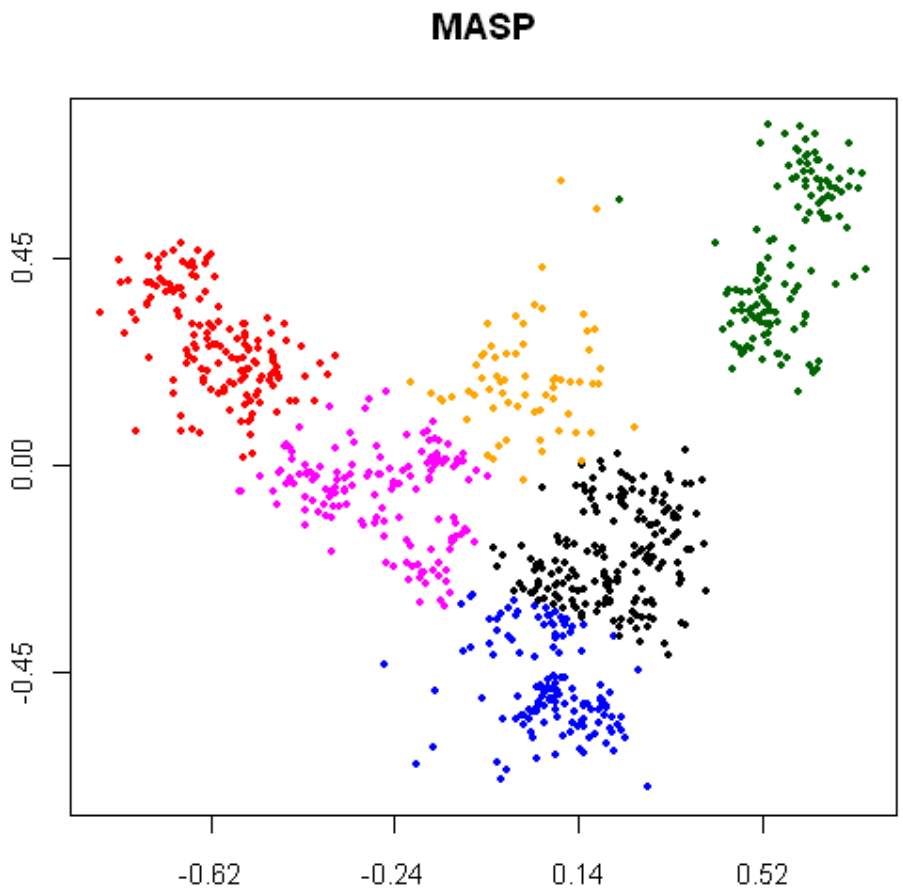
Projeções MDS para sequências de proteína da família TcS



Projeções MDS para sequências de proteína da família GP63

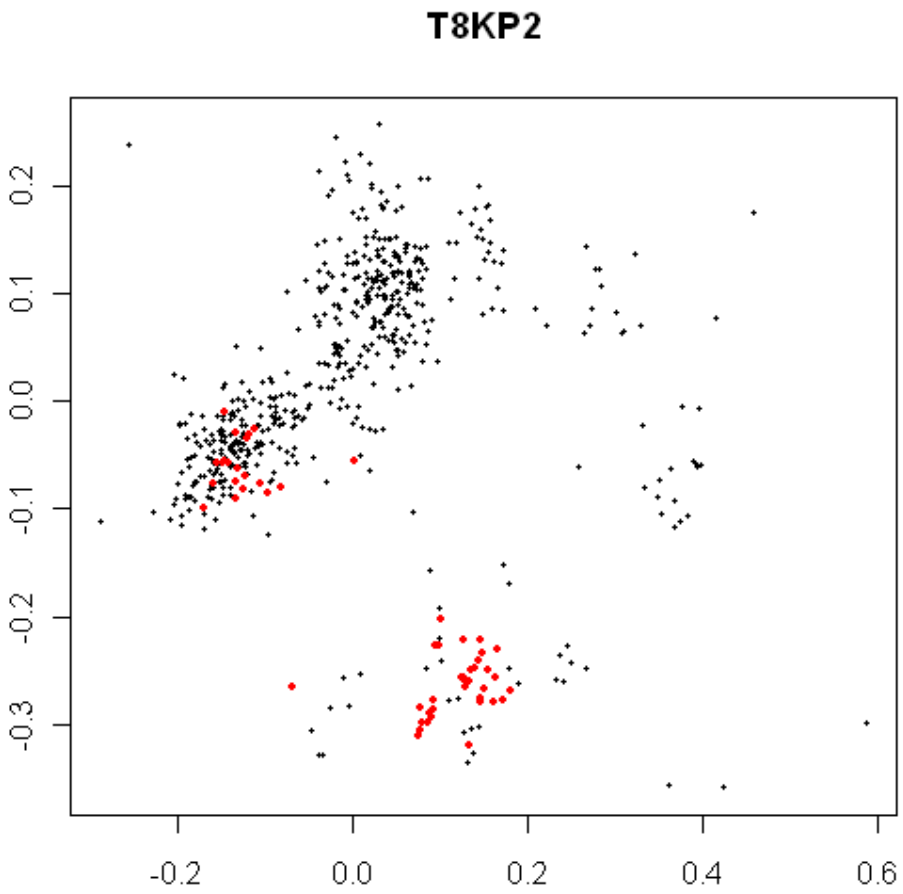


Projeções MDS para sequências de proteína da família MASP.

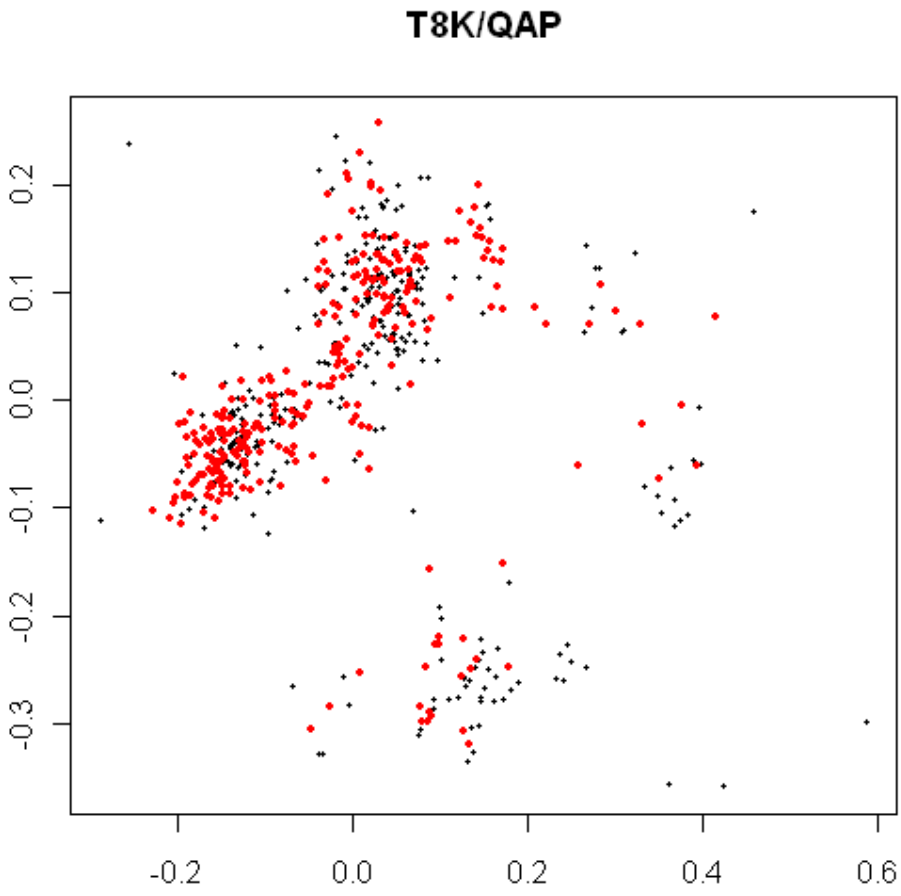


Anexo 7 – Painéis da figura 24. Projeção MDS usando sequências de DNA da família TcMUC.

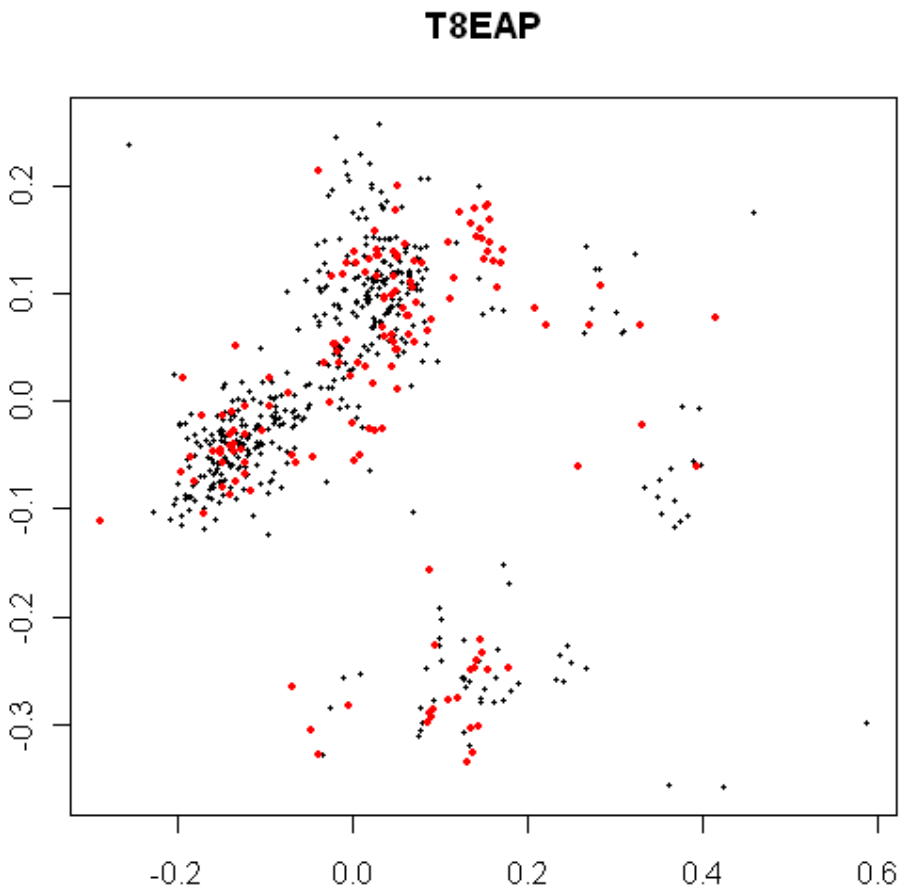
Projeção MDS usando sequências de DNA da família TcMUC. Pontos vermelhos indicam sequências que possuem o motivo T_8KP_2.



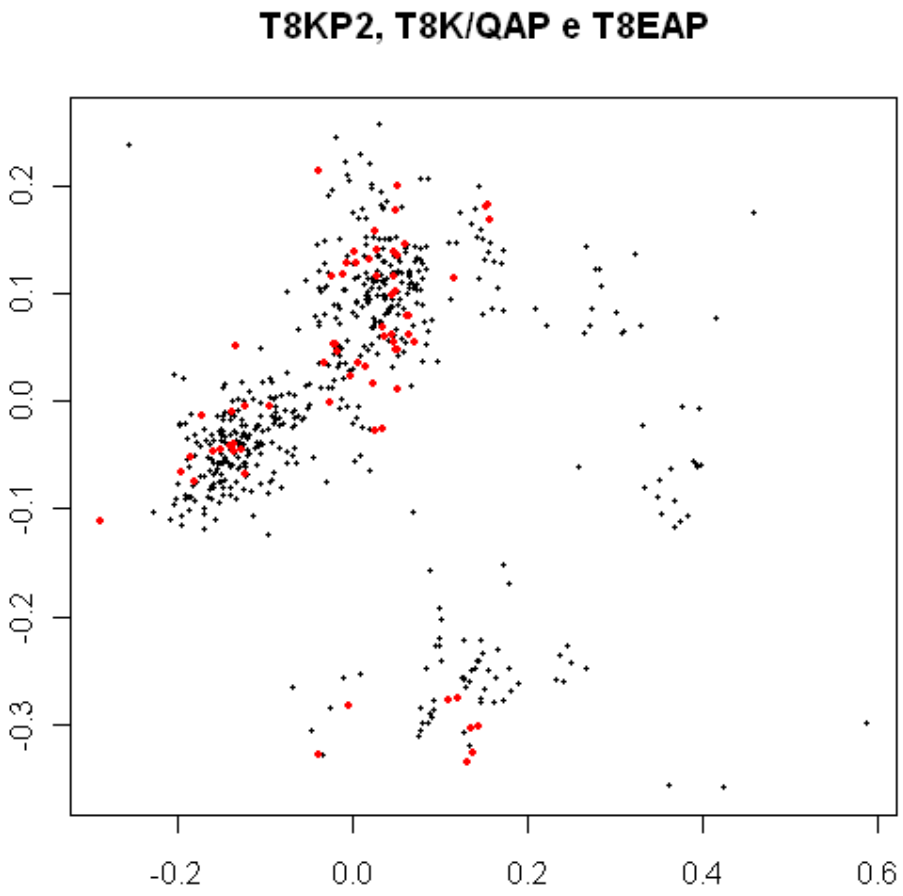
Projeção MDS usando sequências de DNA da família TcMUC. Pontos vermelhos indicam sequências que possuem o motivo T_8K/QAP.



Projeção MDS usando sequências de DNA da família TcMUC. Pontos vermelhos indicam sequências que possuem o motivo T_8EAP.

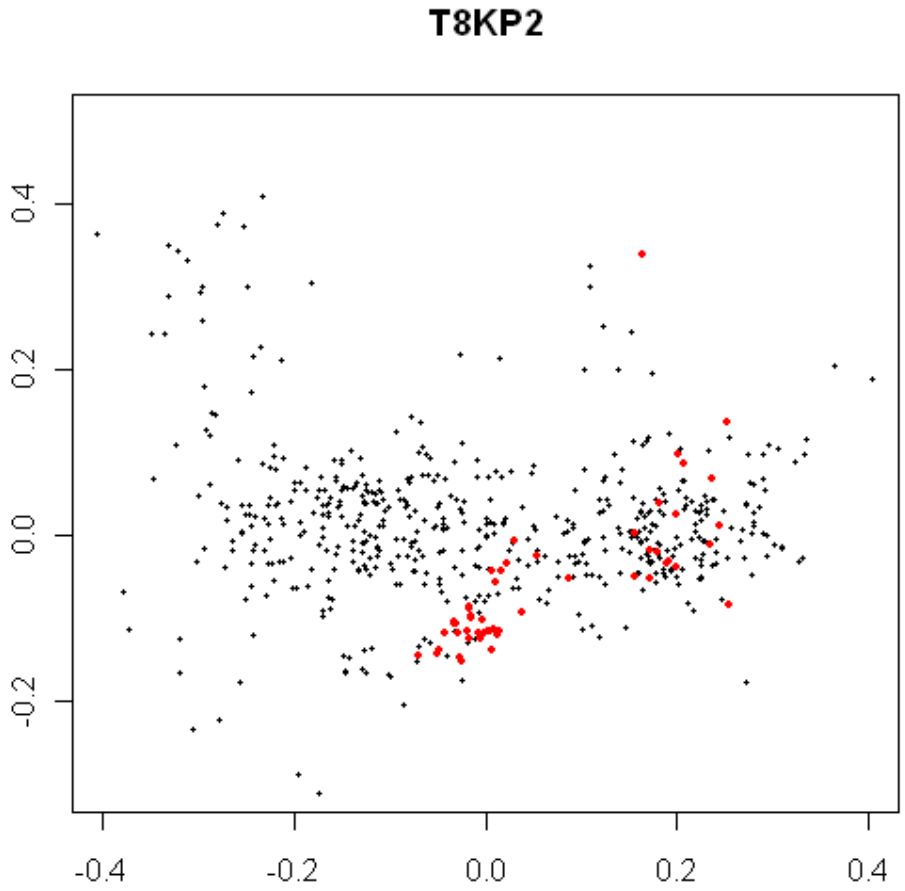


Projeção MDS usando sequências de DNA da família TcMUC. Pontos vermelhos indicam sequências que possuem os motivos T_8KP_2, T_8K/QAP e T_8EAP.

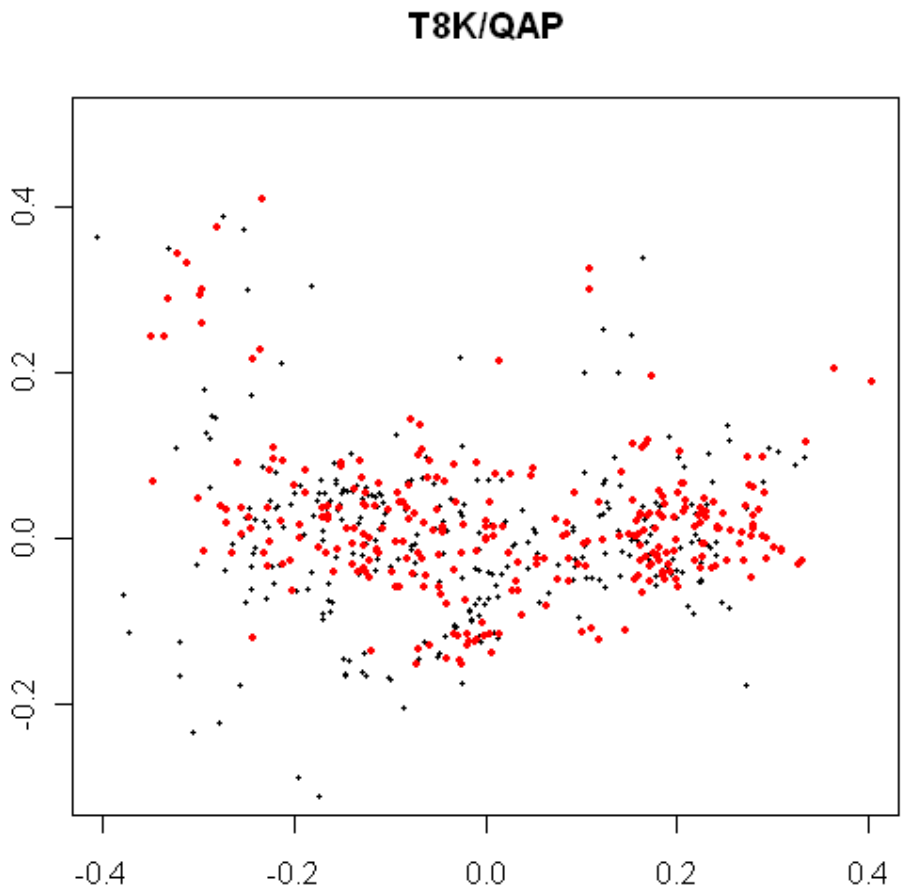


Anexo 8 – Painéis da figura 25. Projeção MDS usando sequências de proteína da família TcMUC.

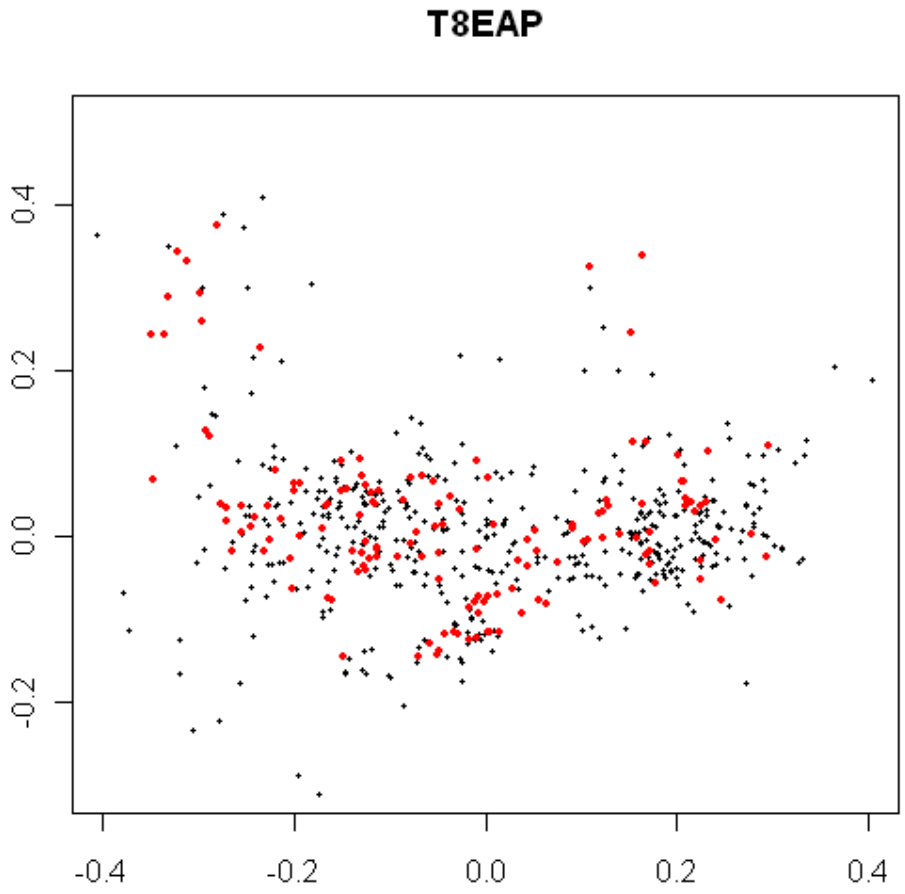
Projeção MDS usando sequências de proteína da família TcMUC. Pontos vermelhos indicam sequências que possuem o motivo T_8KP_2.



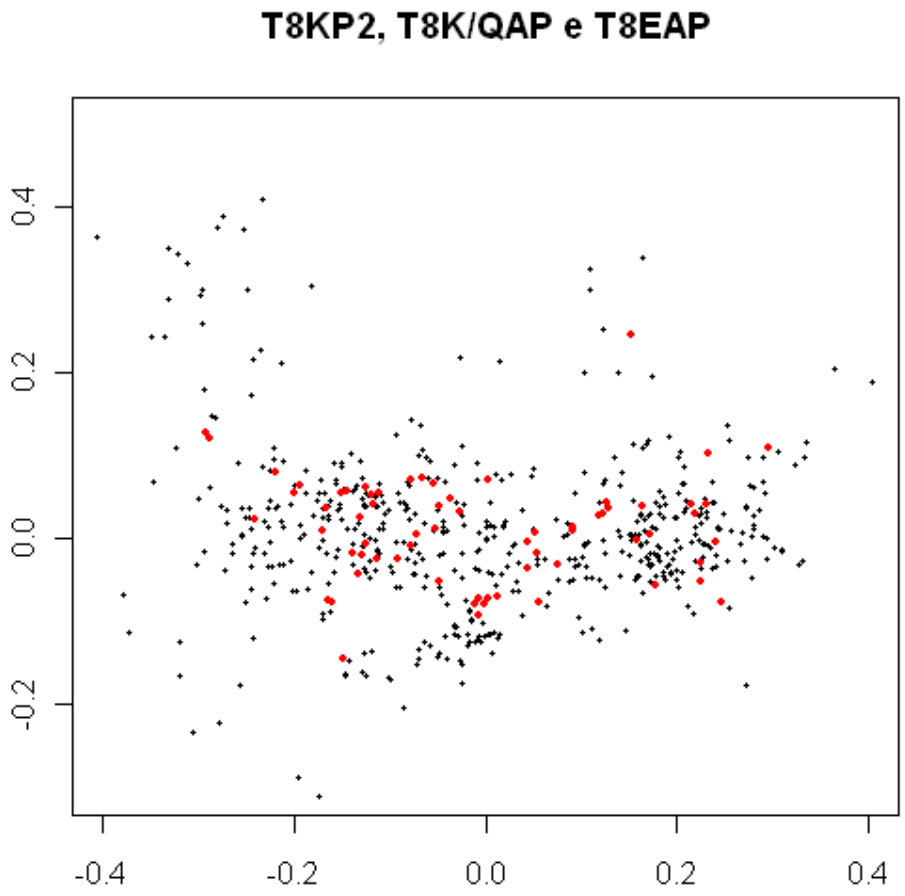
Projeção MDS usando sequências de proteína da família TcMUC. Pontos vermelhos indicam sequências que possuem o motivo T_8K/QAP.



Projeção MDS usando sequências de proteína da família TcMUC. Pontos vermelhos indicam sequências que possuem o motivo T_8EAP.

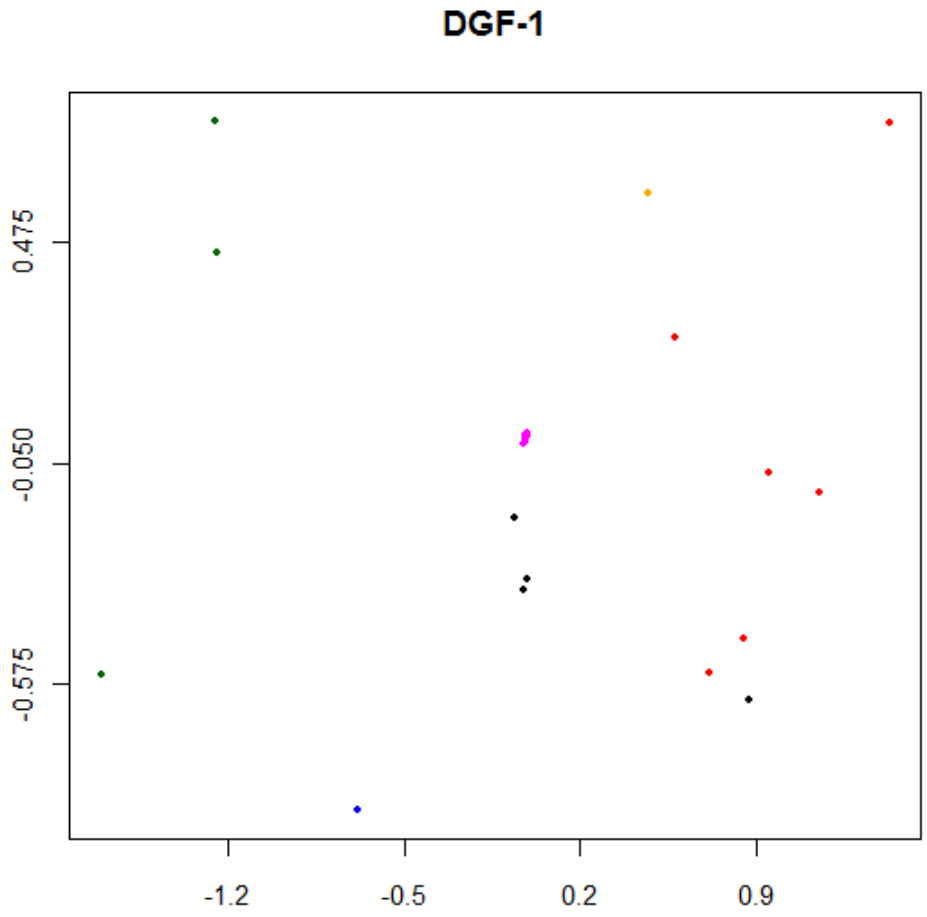


Projeção MDS usando sequências de proteína da família TcMUC. Pontos vermelhos indicam sequências que possuem os motivos T_8KP_2, T_8K/QAP e T_8EAP.

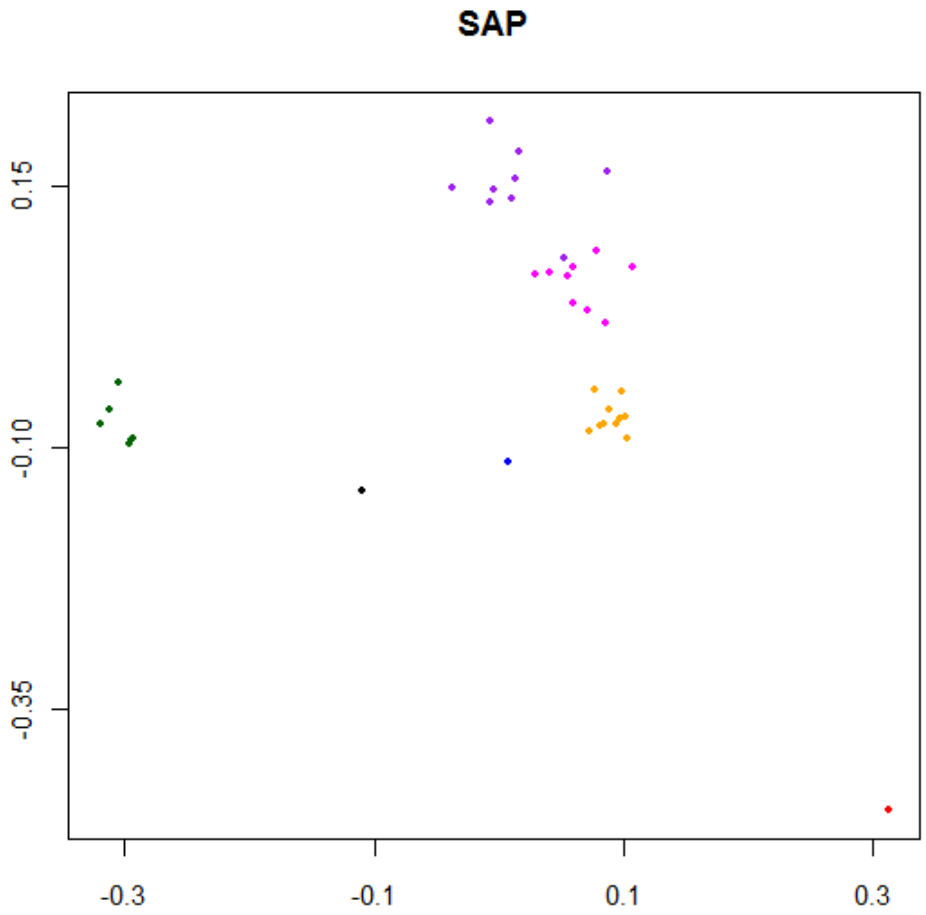


Anexo 9 – Painéis da figura 28. Projeções MDS para sequências de proteínas e classificadas de acordo com os grupos encontrados usando sequências de DNA.

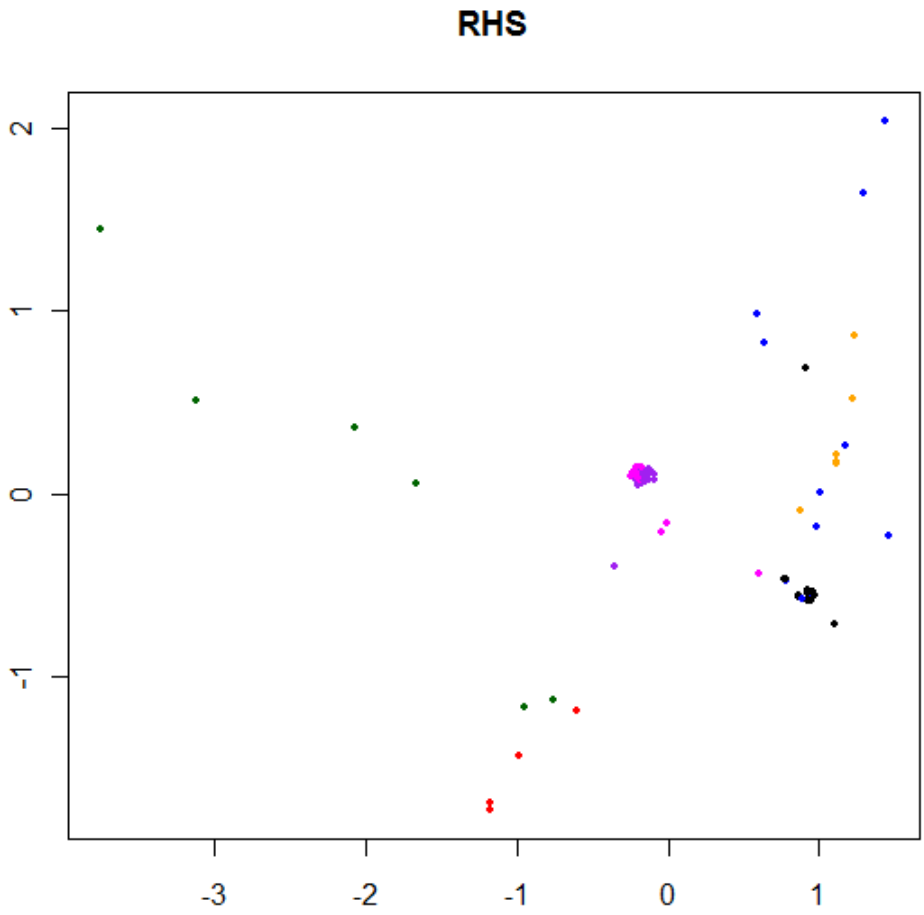
Projeções MDS para sequências de proteína da família DGF-1.



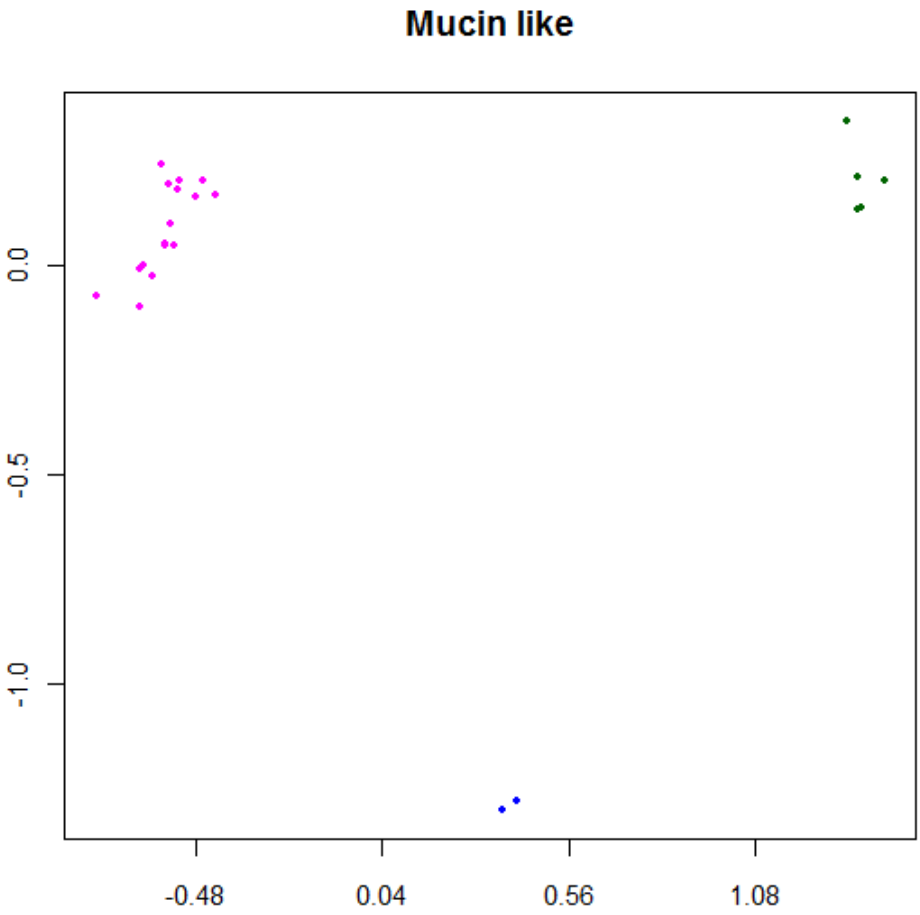
Projeções MDS para sequências de proteína da família SAP.



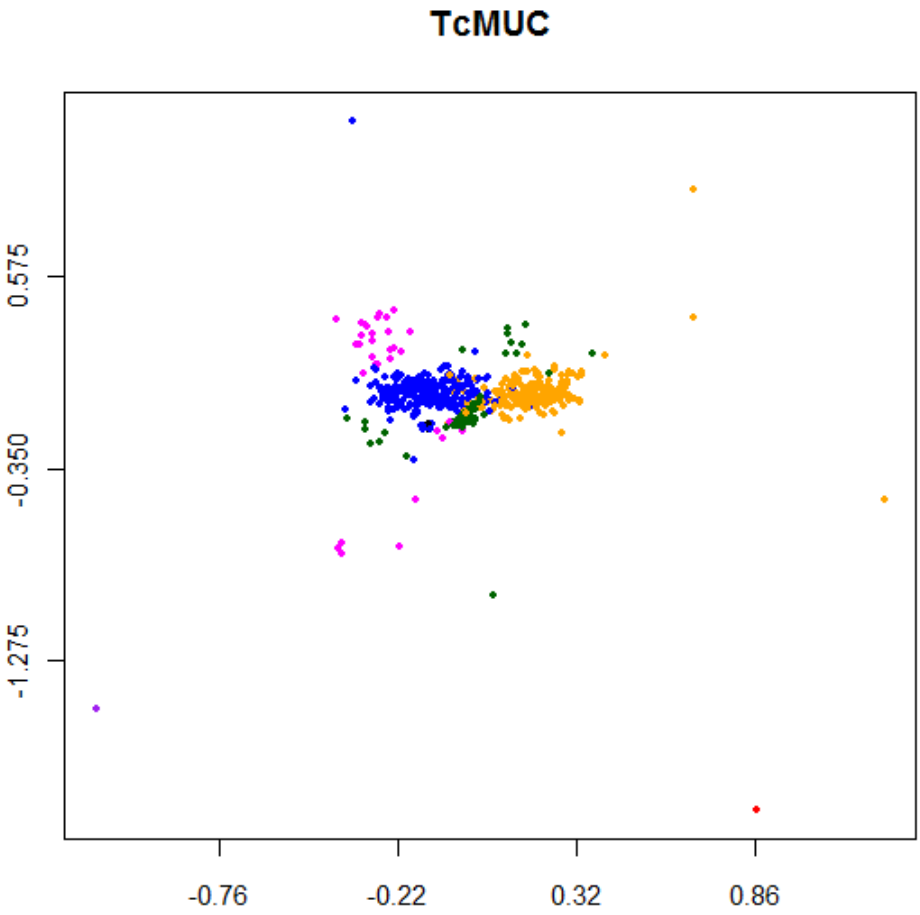
Projeções MDS para sequências de proteína da família RHS.



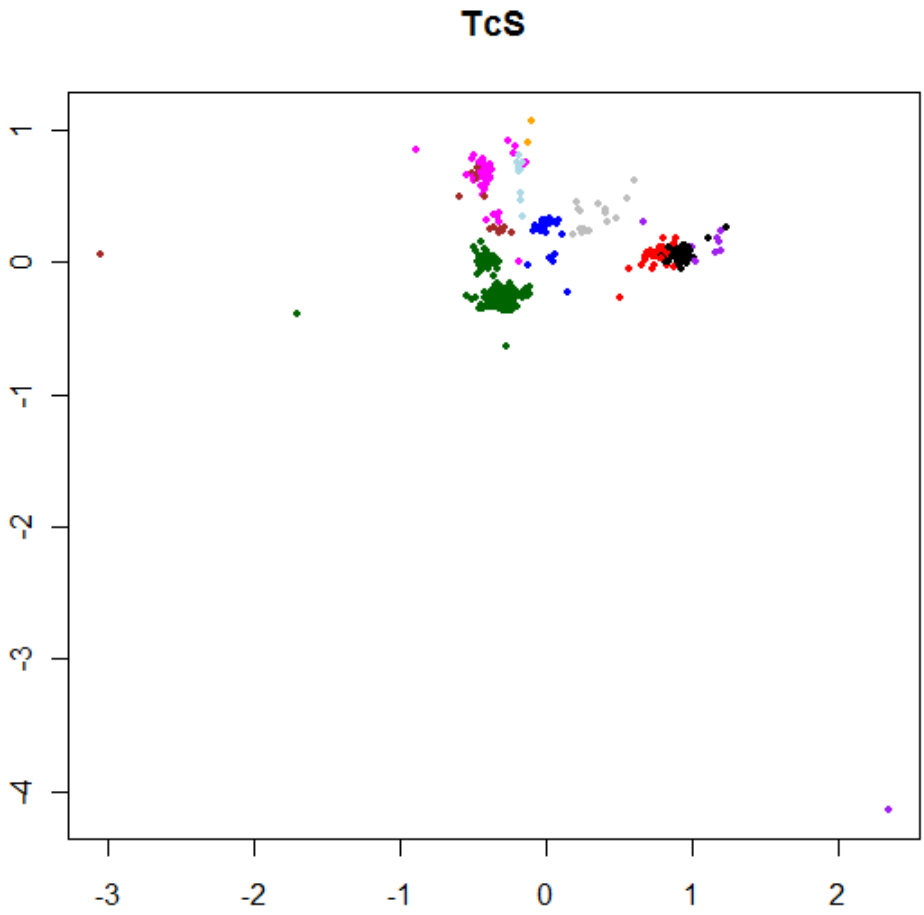
Projeções MDS para sequências de proteína da família mucin like.



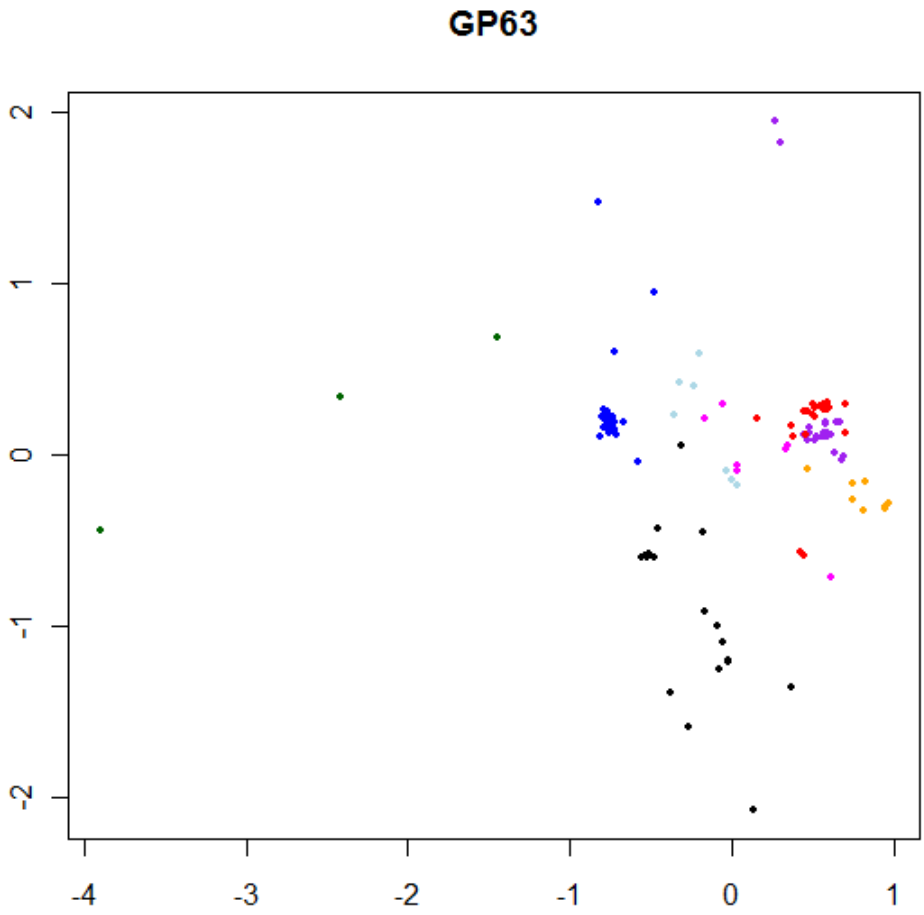
Projeções MDS para sequências de proteína da família TcMUC.



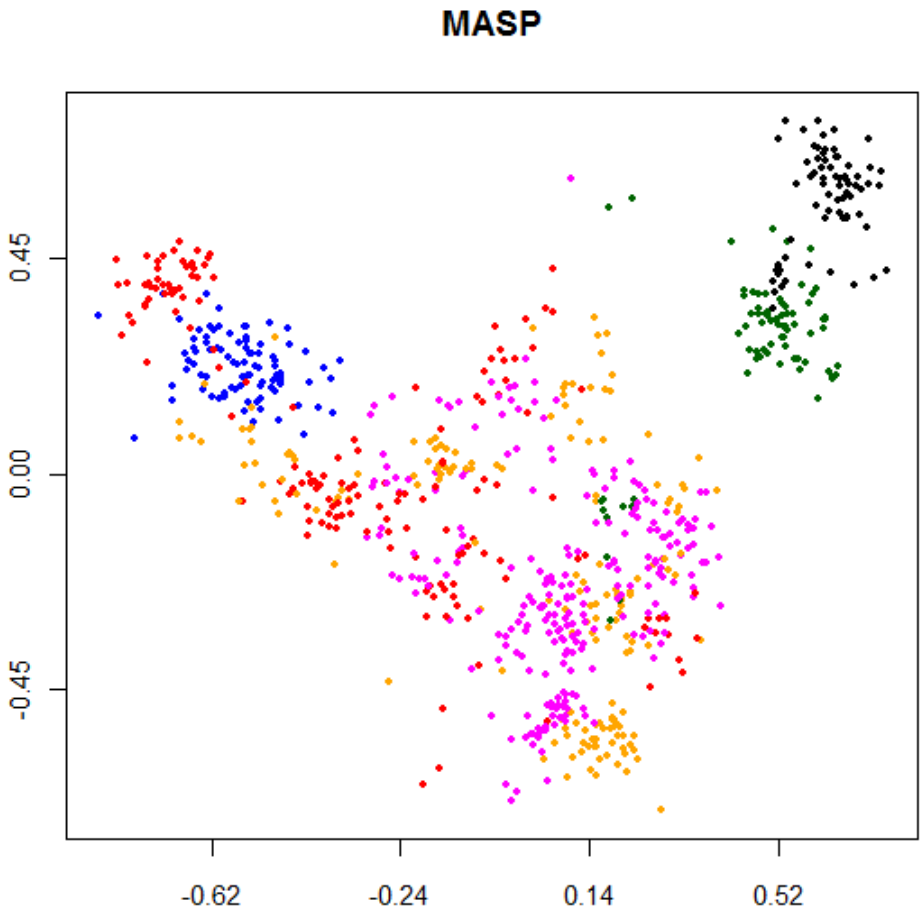
Projeções MDS para sequências de proteína da família TcS.



Projeções MDS para sequências de proteína da família GP63.

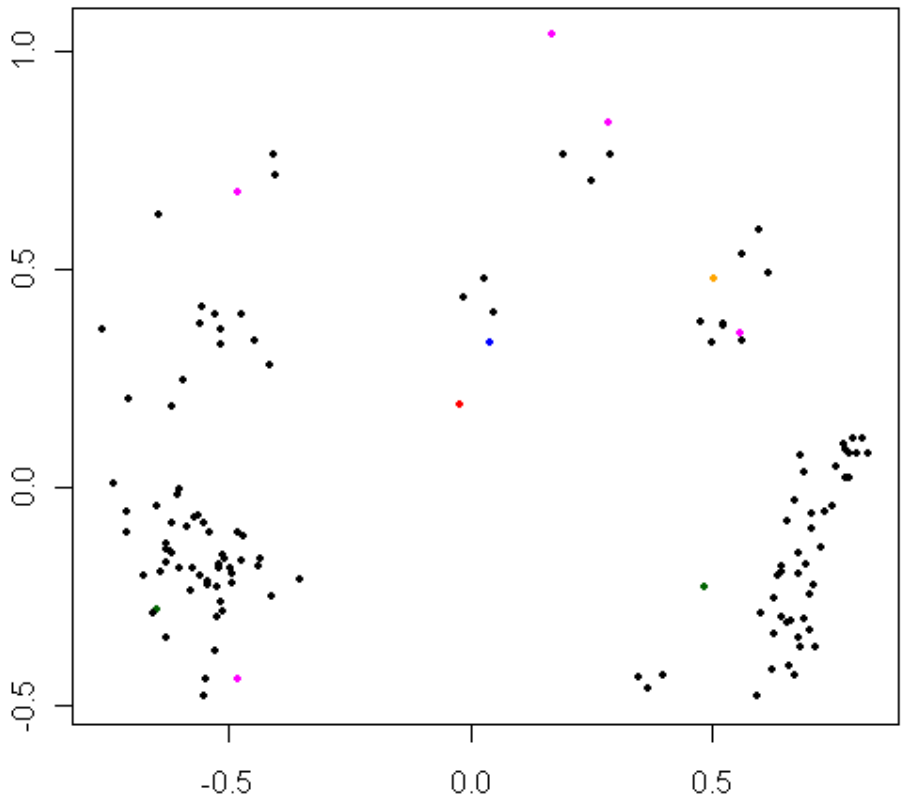


Projeções MDS para sequências de proteína da família MASP.

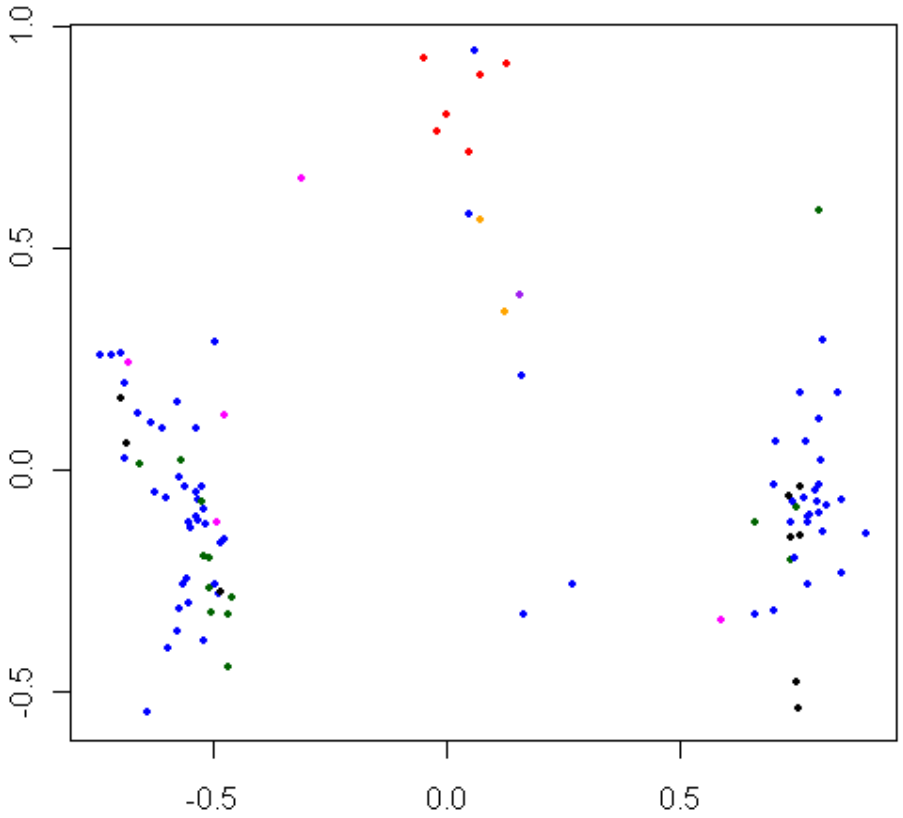


Anexo 10 – Painéis da figura 33. Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação.

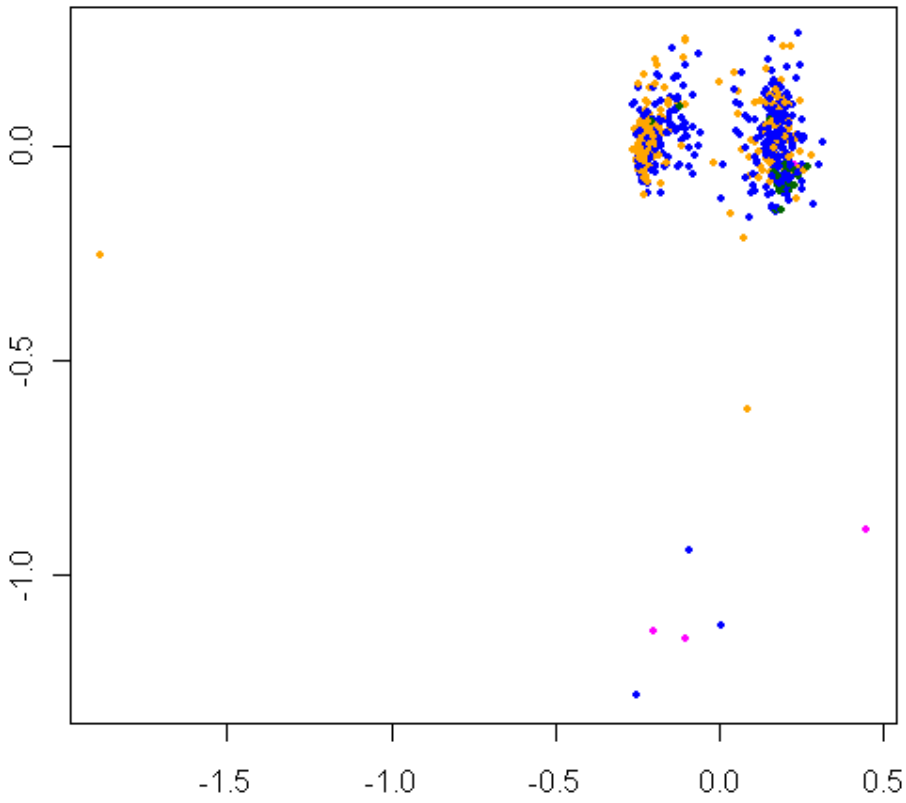
Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família DGF-1.



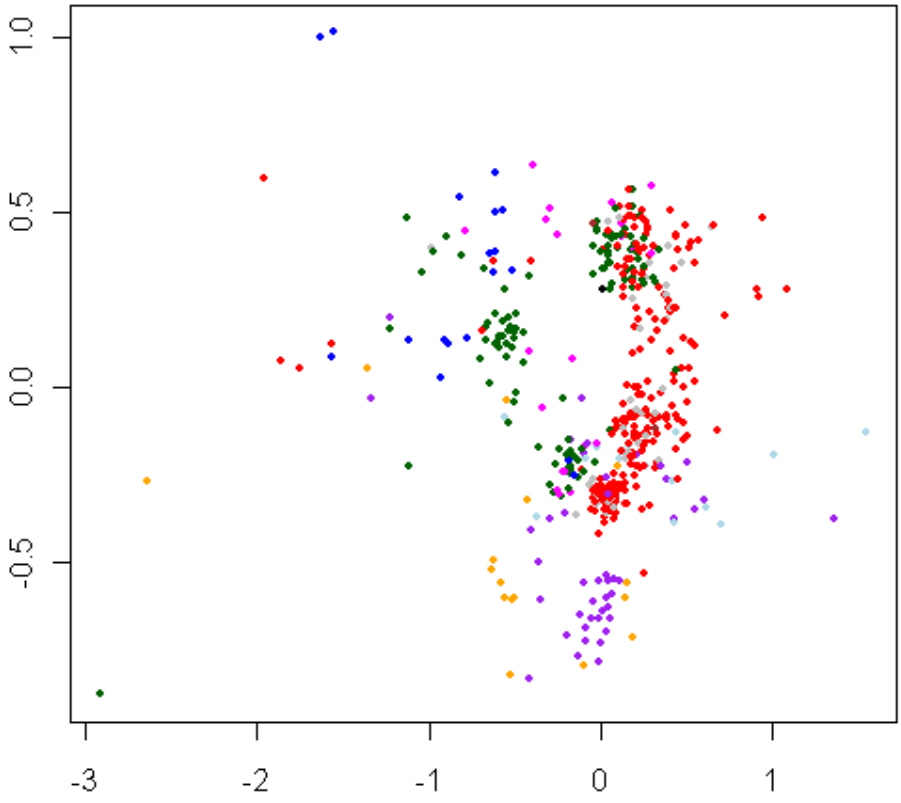
Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família RHS.



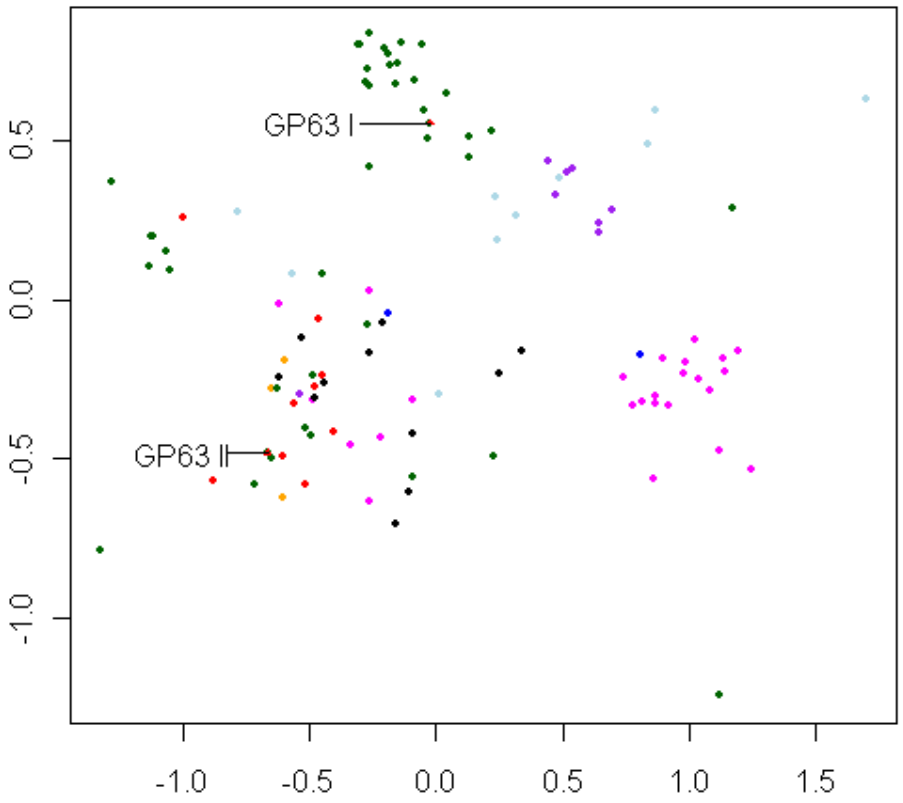
Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família TcMUC.



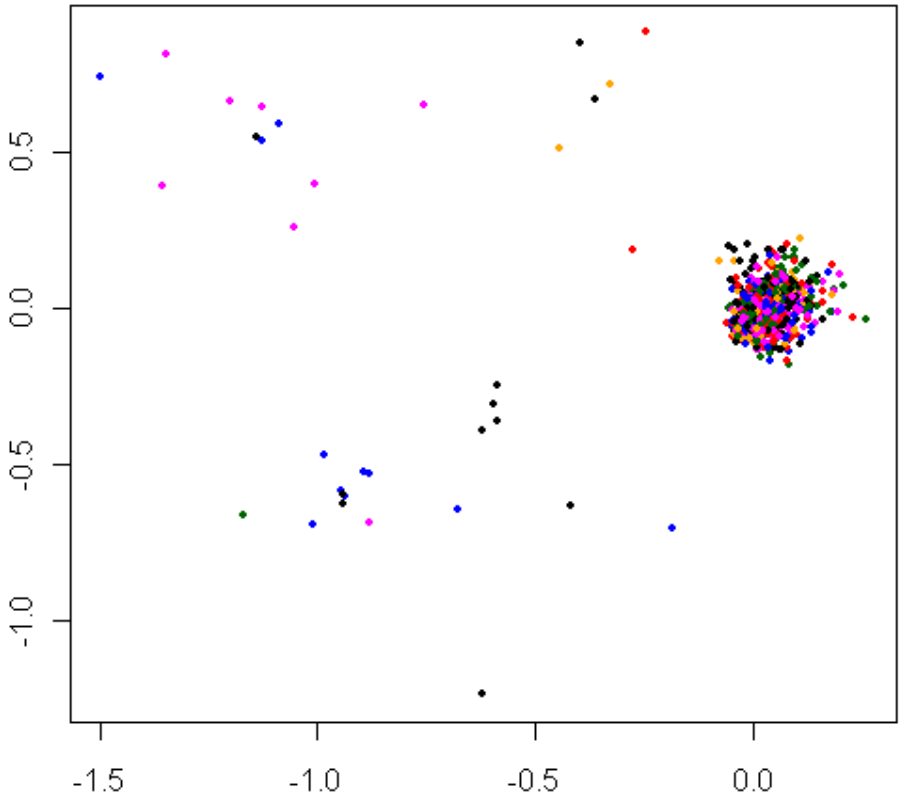
Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família TcS.



Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família GP63

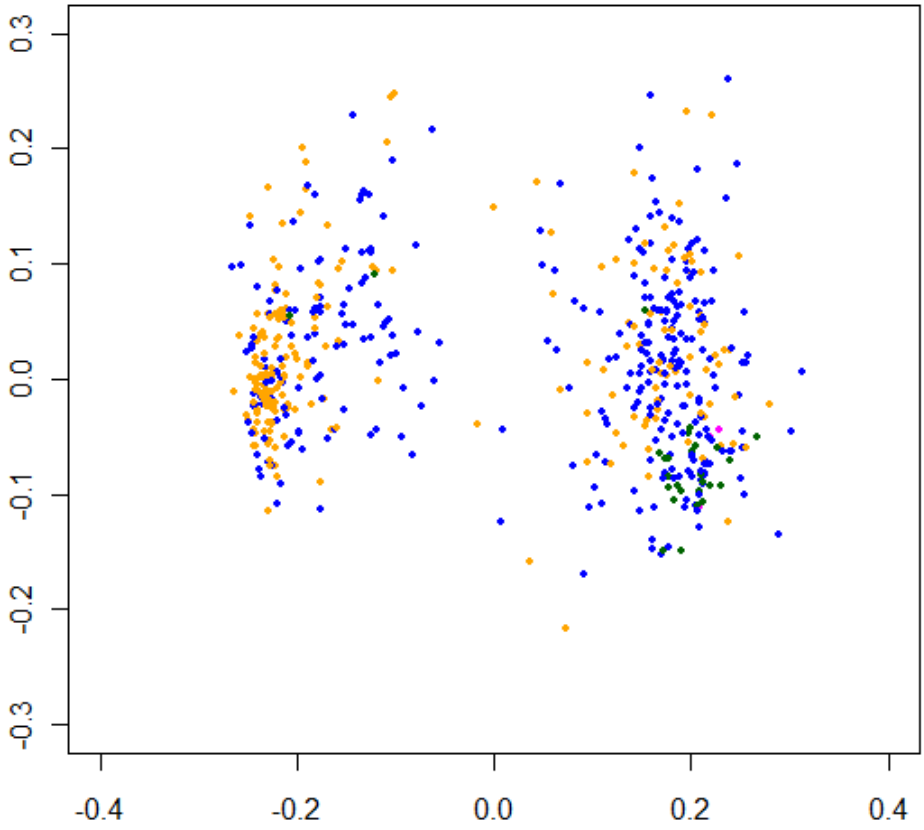


Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família MASP.



Anexo 11 – Painéis B e C da figura 34. Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família TcMUC e comparação com a classificação TcMUC I, TcMUC II e TcMUC III.

Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família TcMUC e classificação hierárquica.



Projeções MDS para sequências 3'flanqueadoras (300 nt) após o códon de terminação da família TcMUC e classificação TcMUC I (preto), TcMUC II (vermelho) e TcMUC III (azul).