UNIVERSIDADE FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

NEWTON DE MEDEIROS VIDAL

ANÁLISE COMPUTACIONAL DA REGULAÇÃO DA EXPRESSÃO GÊNICA EM *Trypanosoma cruzi*

CURITIBA
2012

NEWTON DE MEDEIROS VIDAL

ANÁLISE COMPUTACIONAL DA REGULAÇÃO DA EXPRESSÃO GÊNICA EM *Trypanosoma cruzi*

Tese apresentada ao Programa de Pós-Graduação em Biologia Celular e Molecular da Universidade Federal do Paraná, como requisito para a obtenção do título de Doutor em Biologia Celular e Molecular.

Orientador: Christian Macagnan Probst
Co-orientador: Marco Aurélio Krieger

CURITIBA
2012

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

Departamento de Biologia Celular e Departamento de Fisiologia
Setor de Ciências Biológicas - Universidade Federal do Paraná
Instituto Carlos Chagas (ICC/FIOCRUZ)

PARECER

A banca examinadora, instituída pelo colegiado do Programa de Pós-Graduação em Biologia Celular e Molecular, do Setor de Ciências Biológicas, da Universidade Federal do Paraná, composta por:

Prof. Dr. Christian Macagnan Probst
Orientador e presidente da banca
Instituto Carlos Chagas - ICC

Profa. Dra. Angela Kaysel Cruz
Universidade de São Paulo – USP-FMRP

Prof. Dr. Jerônimo Conceição Ruiz
FIOCRUZ-MG

Profa. Dra. Andréa Rodrigues Ávila
Instituto Carlos Chagas - ICC

Prof. Dr. Leonardo Magalhães Cruz
Universidade Federal do Paraná - UFPR

E tendo como suplentes,

Profa. Dra. Daniela Parana Pavoni
Instituo Carlos Chagas – ICC

Prof. Dr. Stênio Perdigão Fragoso
Instituto Carlos Chagas - ICC

Após arguir o candidato ***Newton de Medeiros Vidal,*** em relação ao seu trabalho intitulado: "Análise computacional da regulação da expressão gênica em *Trypanosoma cruzi*", são de parecer favorável à aprovação do acadêmico, habilitando-o ao título de DOUTOR em Biologia Celular e Molecular, área de concentração Biologia Celular e Molecular.
A obtenção do título está condicionada à implementação das correções sugeridas pelos membros da banca examinadora, bem como ao cumprimento integral das exigências estabelecidas no Regimento Interno deste Programa de Pós-Graduação.

Curitiba, 31 de Agosto de 2012

Prof. Dr. Edvaldo da Silva Trindade
Coordenador do Programa de Pós-Graduação
em Biologia Celular e Molecular - UFPR
Matr. 185795

À Laura, com amor.

AGRADECIMENTOS

Agradeço a todas as pessoas que contribuíram para que esse trabalho se tornasse realidade.

“Não está morto quem peleia!!
Dizia uma ovelha no meio de dez lobos.”

Dito popular gaúcho

RESUMO

Trypanosoma cruzi é o agente etiológico da Doença de Chagas, um sério problema de saúde pública na América Latina. Além disso, divergiu muito cedo da linhagem principal dos eucariotos e possui características moleculares muito peculiares, como transcrição policistrônica, processamento dos transcritos por trans-splicing e a regulação da expressão gênica pós-transcricional. O estudo sistemático e amplo das questões relacionadas à expressão gênica em *T. cruzi,* principalmente sobre uma perspectiva ômica, é essencial para que possamos entender esses aspectos, identificando novos mecanismos, e proporcionando alvos para a definição de melhores terapias. O presente trabalho visa estabelecer uma plataforma computacional para o projeto "Reguloma de *T. cruzi"*, utilizando técnicas bioinformáticas para aprofundar o entendimento sobre a regulação da expressão gênica desse parasita. Para isso, foi implementado um banco computacional para armazenamento de dados ômicos de tripanossomatídeos, que comporte a quantidade e a complexidade dos dados que estão sendo gerados pelo Projeto Reguloma do Instituto Carlos Chagas. A esse banco foram incorporadas informações relativas a genoma, transcriptoma, ribonoma, proteoma, fosfoproteoma, interatoma, primariamente de resultados que estão sendo produzidos pelo nosso grupo. Associado a isso, as informações relativas a metabolismo (KEGG), domínios protéicos (PFAM) e anotação gênica funcional (Gene Ontology) foram também incorporados. Um aspecto central é a natureza redundante dos dados existentes sobre as regiões codificadoras de *T. cruzi,* as quais foram organizadas. Para exemplificar a utilização dessa plataforma, mostramos a sua aplicação em dados transcriptômicos obtidos pela técnica de RNA-Seq do ciclo de vida de *T. cruzi,* abrangendo as quatro formas principais do parasita, e a sua integração com dados metabólicos, de domínios proteicos e identificação de motivos presentes na região 3'UTR que influenciam os padrões de expressão analisados. Para esse último tópico, é essencial a correta definição dos limites dos mRNAs de *T. cruzi,* e para isso utilizamos todos os dados existentes atualmente em nosso grupo (2,6 bilhões de leituras) para predizer bioinformaticamente as regiões de adição do mini-éxon e da cauda poli-A nos mRNAs de *T. cruzi.* Os dados resultantes dessa análise possibilitam uma avaliação mais correta e poderosa sobre os determinantes dos padrões de expressão existentes no projeto Reguloma. Para a caracterização de motivos regulatórios candidatos, fizemos a definição de amostras controles, sendo que a definição de condições opostas como um possível controle é uma abordagem inédita na análise de elementos reguladores. Finalmente, também demonstramos a aplicação dessa plataforma na integração dinâmica de dados de RNA-Seq com o metabolismo do organismo, o que cria um nível adicional de complexidade nas análises passíveis de serem realizadas. Portanto, o presente trabalho representa a construção de um arcabouço sobre o qual as futuras análises do projeto Reguloma poderão ser feitas, visando obter os elementos e as relações entre os mesmos, construindo uma rede biológica regulatória. Com a incorporação de um grande número de dados referentes às mais diversas caracterizações ômicas, que estão sendo geradas primariamente por nosso grupo, vislumbra-se que a infraestrutura bioinformática apresentada nesse trabalho viabilizará a realização de análises sistêmicas, promovendo o avanço do conhecimento científico sobre a regulação da expressão gênica de *T. cruzi.*

Palavras-chave: *Trypanosoma cruzi,* expressão gênica, redes metabólicas, transcriptômica, elementos regulatórios, bioinformática

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Trypanosoma cruzi E SEU CICLO DE VIDA

Trypanosoma cruzi (CHAGAS, 1909) é um eucarioto unicelular e flagelado. É um protozoário parasita pertencente ao supergrupo Excavata (MOREIRA *et al*., 2007; YOON *et al*., 2008), filo Euglenozoa, ordem Kinetoplastida, família Trypanosomatidae (CAVALIER-SMITH, 1993; CAVALIER-SMITH, 2010).

Organismos da ordem Kinetoplastida possuem uma mitocôndria única, estrutura especializada chamada de cinetoplasto. É uma estrutura em forma de bastão, que contém o genoma mitocondrial (kDNA). O kDNA é uma rede de DNA circular com várias cópias do genoma mitocondrial, formada por maxi e mini-círculos concatenados (SHAPIRO & ENGLUND, 1995; TELLERIA *et al*., 2006).

A família Tripanosomatidae é composta por diversos gêneros. Dentre eles, *Trypanosoma* e *Leishmania* estão entre os mais importantes, pois incluem espécies causadoras de diversas doenças de importância médica e veterinária. Por exemplo, em humanos, *Trypanosoma brucei rhodesiense* e *Trypanosoma brucei gambiense* são causadores da Doença do Sono (GARCIA *et al*., 2006), também chamada de Tripanossomíase Africana Humana; e *Trypanosoma cruzi* é causador da Doença de Chagas (ou Tripanossomíase Americana Humana). As subespécies de *Trypanosoma brucei: T. b. brucei, T. b. evansi* e *T. b. equiperdum,* causam a doença chamada Nagana (ou Tripanossomíase Animal), infectando bois, cavalos, camelos e búfalos (HIDE, 1999; LAI *et al*., 2008). Existem cerca de 30 espécies de *Leishmania* que causam leishmaniose em humanos. Essa doença possui três formas: leishmaniose cutânea, leishmaniose mucocutânea e leishmaniose visceral, causadas, por exemplo, pelas espécies *Leishmania major*, *Leishmania braziliensis* e *Leishmania infantum*, respectivamente (PEACOCK *et al*., 2007).

Apesar de evoluir por divisão clonal, *Trypanosoma cruzi* possui grande heterogeneidade intraespecífica. Devido à evolução paralela, as diferentes linhagens e cepas desse parasita possuem diferenças genotípicas e fenotípicas quanto à capacidade de infecção e à virulência (ZINGALES *et al*., 1999; STURM *et al*., 2003).

As populações de *T. cruzi* podem ser divididas em dois grupos: *T. cruzi* I e *T.*

cruzi II (ZINGALES *et al*., 1999). O grupo *T. cruzi* I está associado ao ciclo silvático de transmissão da doença, enquanto o grupo *T. cruzi* II está associado ao ciclo doméstico (YEO *et al*., 2005). O segundo grupo é subdividido em 5 subgrupos: IIa, IIb, IIc, IId e IIe (BRISSE *et al*., 2000). Posteriormente esses grupos passaram a se chamar DTU (do inglês *Discrete Typing Unit*) I e II (TIBAYRENC, 2003). Um novo consenso acerca da nomenclatura classifica as diferentes linhagens evolutivas de *T. cruzi* em 6 subpopulações, chamados DTUs TcI a TcVI. A correspondência entre as duas nomenclaturas é a seguinte: TcI-I, TcII-IIb, TcIII-IIc, TcIV-IIa, TcV-IId, TcVI-IIe (ZINGALES *et al*., 2009; ZINGALES *et al*., 2012).

A distância evolutiva entre as DTUs de *T. cruzi* é tão grande quanto a distância entre espécies com ciclos de vida bastante diferentes, como as dos gêneros *Leishmania* (dixênica, que alterna entre dois hospedeiros) e *Crithidia* (monoxênica, que possui apenas um hospedeiro) (JUNQUEIRA *et al*., 2005). Por esse motivo, alguns autores sugerem que essas diferentes linhagens evolutivas fossem consideradas um complexo de espécies, ou que sejam divididas em espécies ou subespécies diferentes (DEVERA *et al*., 2003; JUNQUEIRA *et al*., 2005).

Por ser um parasita digenético, no seu ciclo de vida (FIGURA 1), *Trypanosoma. cruzi* alterna entre dois hospedeiros, um inseto vetor hematófago e um hospedeiro mamífero. Durante o repasto sanguíneo, o triatomíneo libera suas excretas próximo à ferida da picada. As formas tripomastigotas metacíclicas do *T. cruzi* presentes nas excretas infectam o hospedeiro através da ferida quando este se coça ou através das membranas mucosas (1). Na corrente sanguínea, os parasitas invadem células do hospedeiro e se transformam em amastigotas (2). Os amastigotas se multiplicam por fissão binária (3) e se diferenciam em tripomastigotas sanguíneos, que então são liberados na corrente sanguínea, podendo infectar outras células ou serem ingeridos pelo triatomíneo (5). No intestino médio do inseto os tripomastigotas se transformam em epimastigotas (6) e se multiplicam por fissão binária (7). No intestino posterior se diferenciam em tripomastigotas metacíclicos (8), fechando o ciclo. O *T. cruzi* alterna formas intra e extracelulares ao longo do seu ciclo de vida (http://www.dpd.cdc.gov/dpdx/HTML/TrypanosomiasisAmerican.htm).

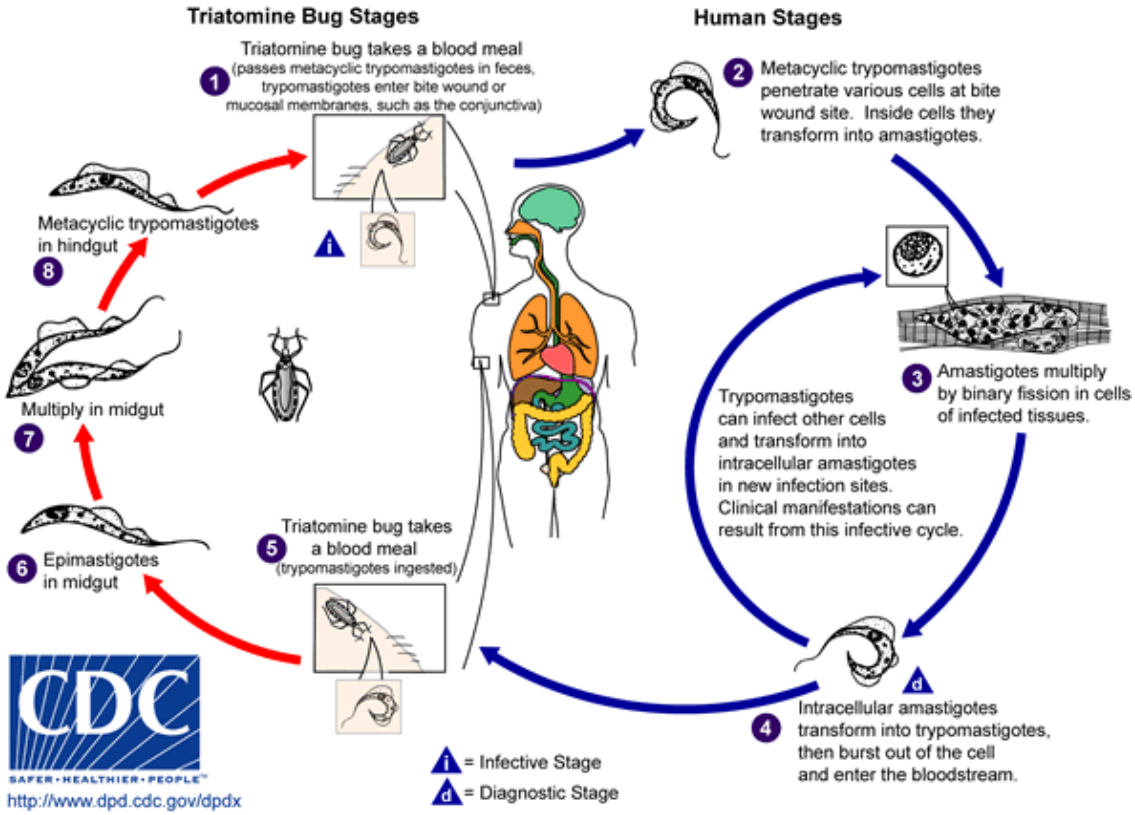


FIGURA 1 – Ciclo de vida do *Trypanosoma cruzi*.

FONTE: http://www.dpd.cdc.gov/dpdx/HTML/TrypanosomiasisAmerican.htm

1.2 DOENÇA DE CHAGAS

Trypanosoma cruzi é o agente etiológico da Doença de Chagas, uma das doenças com maior distribuição no continente americano e um sério problema de saúde pública. O parasita infecta cerca de 10 a 15 milhões de pessoas, mata cerca de 14 mil pessoas por ano e estima-se que cerca de 120 milhões de pessoas estejam sob o risco de contaminação. A Doença de Chagas não possui vacina nem medicamentos para tratamento efetivo (Assembléia Mundial de Saúde, Genebra, 2010).

A doença é transmitida por um inseto vetor hematófago, popularmente conhecido como barbeiro. Ele pertence à ordem Hemiptera, família Reduviidae, subfamília Triatomidae, por isso também chamado de triatomíneo. As espécies mais comuns pertencem aos gêneros *Triatoma*, *Rhodnius* e *Panstrongylus*. Os triatomíneos picam os mamíferos selvagens, reservatórios naturais do *T. cruzi*, para alimentar-se de seu sangue (ciclo silvático) (NEVES, 1991; TEIXEIRA *et al*., 2006).

Com a destruição de seus habitats naturais e pela ocupação dessas áreas pelo homem, o barbeiro encontrou nas habitações humanas um abrigo seguro e com abundante oferta de alimento (ciclo doméstico) (DIAS, 2000, TEIXEIRA *et al*., 2006).

A FIGURA 2 mostra o parasita, o vetor e os hospedeiros no seu ciclo silvático e doméstico.

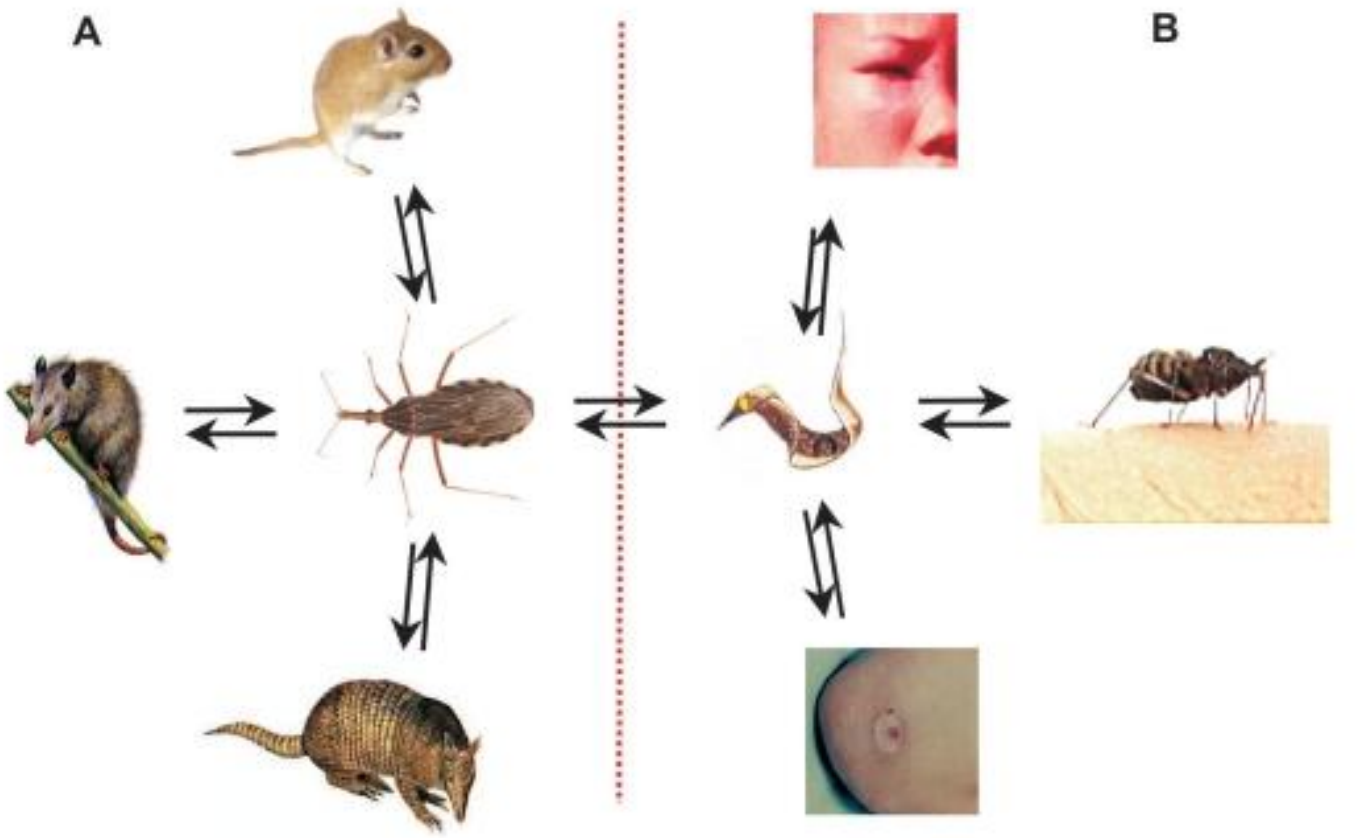


FIGURA 2 – Ciclo de vida silvático e doméstico de *Trypanosoma cruzi.*

A: inseto hematófago contaminado ou sendo contaminado pelo parasita ao se alimentar de hospedeiros silvestres. Rato (acima), gambá (à esquerda) e tatu (abaixo). B: Triatomíneo inicia o ciclo doméstico ao se alimentar do homem (à direita). Marcas da entrada do parasita: sinal de Romaña (acima) e chagoma (abaixo).

FONTE: TEIXEIRA *et al*., 2006

A transmissão da doença ocorre principalmente por: entrada do parasita através de mucosas ou lesões na pele (transmissão vetorial), como a causada pela picada do barbeiro; transfusão com sangue contaminado; congênita (transmissão vertical); ou ingestão oral acidental de alimentos contaminados (NEVES, 1991; TEIXEIRA *et al*., 2006). A transmissão vetorial foi a principal responsável pela maioria dos casos e pela expansão da Doença de Chagas, mas a transmissão

transfusional também adquiriu importância. Recentemente, com medidas de combate ao inseto vetor e o maior controle da qualidade das bolsas de sangue para transfusão no Brasil, a infecção por via oral tem ganhado mais atenção e tem sido a forma mais freqüente de transmissão da doença (COURA, 2006). Surtos de Doença de Chagas, através da ingestão de comida contaminada, têm ocorrido em diversas regiões do Brasil. Na região Amazônica, 50% dos casos registrados da doença aguda são atribuídos a essa forma de transmissão (COURA *et al*., 2002).

A Doença de Chagas possui duas fases: a aguda e a crônica.

A fase aguda desenvolve-se logo após a infecção e breve período de incubação, podendo ser assintomática. Quando sintomática, o indivíduo infectado apresenta alta parasitemia, febre, mal-estar, cefaléia, edema subcutâneo, disfunção cardíaca, hepatomegalia e esplenomegalia, o que pode levar à morte por insuficiência cardíaca ou meningoencefalite (COURA, 2007). A infecção por via oral normalmente gera casos agudos com sintomas mais fortes do que as outras formas de transmissão, porque a quantidade de parasitas ingeridos e em contato com a mucosa gástrica é muito maior (YOSHIDA, 2009).

O parasita persiste indefinidamente no organismo, mesmo em níveis mínimos. A resposta imunológica do organismo ao parasita resulta no acúmulo de lesões que podem gerar alterações morfológicas e funcionais nos tecidos afetados (COURA & DE CASTRO, 2002; ANDRADE & ANDREWS, 2005).

Depois de anos ou décadas, a doença evolui para um quadro crônico, geralmente assintomático, mas que entre 20 e 50% dos infectados sofrem debilitação severa que pode levar à morte (COURA & DE CASTRO, 2002). Nesta fase, diversos quadros clínicos podem ser evidenciados: distúrbios cardíacos (megacárdio), digestivos (megaesôfago e megacólon) e neurológicos (BRENER *et al*., 2000). Esta variedade de sintomas está relacionada com a variabilidade genética do parasita e do hospedeiro, que influencia o curso natural da infecção (STURM & CAMPBELL, 2009).

Duas drogas têm sido usadas no tratamento da Doença de Chagas: Nifurtimox, um derivado de nitrofuran; e Benzonidazol, um derivado de nitroimidazol. Os resultados obtidos com ambas as drogas variam de acordo com a fase da doença, o período de tratamento e a idade e origem geográfica do paciente (COURA & DE CASTRO, 2002). O efeito tripanocida dessas drogas se dá através da formação de radicais livres e metabólitos eletrolíticos que atacam o parasita. De

modo geral, quando administrados na fase aguda da doença produzem resultados satisfatórios, com até 80% de eficácia, principalmente em crianças. Já na fase crônica, os resultados não são satisfatórios, com apenas 10 a 20% de pacientes curados.

A utilização de Nifurtimox e Benzonidazol pode causar toxicidade sistêmica e diversos efeitos colaterais: anorexia, náusea, vômitos, cefaléia, depressão do sistema nervoso central ou sintomas psicóticos, dermatites e parestesia (MAYA *et al*., 2007). Além dos efeitos colaterais, outro fator que contriubui para a baixa efetividade clínica das mesmas é que as diferentes cepas do parasita possuem diferente suscetibilidade a essas drogas (FILARDI & BRENER, 1987; BRENER *et al*, 1993).

A comercialização de Nifurtimox no Brasil foi descontinuada e atualmente apenas o Benzonidazol é encontrado. Além da falta de medicamentos efetivos contra a Doença de Chagas, não existem vacinas eficazes (DUMONTEIL, 2009).

1.3 ORGANIZAÇÃO GENÔMICA

O material genético dos tripanossomatídeos está localizado em duas estruturas celulares: o núcleo e a mitocôndria. Uma das características mais peculiares dos tripanossomatídeos é a organização de seu genoma mitocondrial, que é uma complexa rede de moléculas de DNA chamada cinetoplasto (kDNA), presente em sua mitocôndria única. Essa rede consiste em centenas de minicírculos e dezenas de maxicírculos concatenados (OPPERDOES & MICHELS, 2007). Os maxicírculos codificam genes de rRNA e do complexo respiratório, que necessitam ser editados para produzirem mRNAs funcionais. O processo de edição de RNA, que é a adição e/ou remoção de uridinas nos transcritos dos maxicírculos, é outra característica peculiar dos tripanossomatídeos. A ampla edição de uma variedade de transcritos dos maxicírculos é guiada pelos RNAs guias (gRNAs) que são codificados pelos minicírculos (OPPERDOES & MICHELS, 2007).

Os cromossomos dos tripanossomatídeos não se condensam durante o ciclo celular, o que dificulta a análise de seus cariótipos (VICKERMAN & TETLEY, 1977). Linhagens distintas de *T. cruzi* são polimórficas quanto ao tamanho do genoma, quantidade de DNA repetitivo, e tamanho e número de cromossomos. Apesar dessas

diferenças, grandes grupos sintênicos são conservados (SOUZA *et al*., 2011) entre as diferentes linhagens.

Os primeiros genomas de tripanossomatídeos a serem publicados foram: *Trypanosoma brucei* 927(BERRIMAN *et al*., 2005), *T. cruzi* CL Brener (EL-SAYED *et al*., 2005a), *Leishmania major* Friedlin (IVENS *et al*., 2005), *L. braziliensis* M2904 e *L. infantum* JPCM5 (PEACOCK *et al*., 2007).

A cepa de *T. cruzi* sequenciada foi a CL Brener (DTU TcVI), por ser uma das mais estudadas experimentalmente (ZINGALES *et al*., 1997). Devido ao alto conteúdo repetitivo deste genoma (cerca de 50%) e a natureza híbrida, resultante de hibridização entre as DTUs TcII e TcIII, da cepa escolhida, e conseqüentemente à alta variação alélica, foram modificados os parâmetros padrões de montagem do genoma. Após a montagem inicial, também foi sequenciada em uma cobertura de 2,5 vezes a cepa Esmeraldo (DTU TcII), para permitir a distinção entre os dois haplótipos (Esmeraldo e Não-Esmeraldo). Anteriormente o número de cromossomos da cepa CL Brener foi estimado em 64, com tamanho de 87 Mb, utilizando eletroforese por campo pulsado (PFGE, do inglês Pulse Field Gel Electrophoresis) e hibridização com sondas teloméricas (CANO *et al*., 1995).

No entanto, os resultados do projeto de sequenciamento do genoma estimam que o genoma diplóide esteja entre 106,4 e 110,7 Mb. Desse total, 67 Mb foram selecionados para anotação, sendo que o restante do genoma não selecionado provavelmente contém uma grande quantidade de genes codificadores de moléculas de RNA, como rRNA, SL-RNA e snoRNA, organizados em tandem.

A montagem totalizando 67 Mb consiste de 5489 *scaffolds,* contendo 8740 *contigs*, os quais representam muito bem o estado extremamente fragmentado da atual versão do genoma. A análise da fração anotada, que perfaz 60,7 Mb, demonstra que 30,5 Mb são compostos de seqüências encontradas pelo menos duas vezes, o que sugere que elas são oriundas dos diferentes haplótipos que compõe o genoma da cepa CL Brener. Foram anotados 22.570 genes codificadores de proteína e estima-se que o genoma haplóde contenha ~12.000 genes. Pelo menos 50% do genoma é constituído de seqüências repetidas, retrotransposons e repetições sub-teloméricas. As maiores famílias gênicas de *T. cruzi* codificam proteínas de superfície como trans-sialidade, proteínas de superfície associadas a mucina (MASP), mucinas e protease de superfície GP63, que correspondem a ~18% dos genes codificadores de proteína do genoma.

O genoma de outra cepa de *T. cruzi*, Sylvio X10/1, foi parcialmente sequenciado (FRANZÉN *et al*., 2011). Essa cepa pertencente à DTU TcI, que não é híbrida, apresenta um menor grau de heterozigosidade, e seu genoma é menor e menos repetitivo, o que possibilitaria uma montagem de maior resolução. O genoma foi montado em 7.092 contigs (24Mbp) e foi comparado com a cepa referência CL Brener. O conteúdo gênico entre as duas linhagens foi muito semelhante, mas possuem grandes diferenças no conteúdo de repetições, o que pode ter implicações na diferença funcional e epidemiológica entre as linhagens, Sylvio X10/1 é associada so ciclo silvático da doença, e CL Brener associada ao ciclo doméstico da doença.

Considerando o genoma haplóide de CL Brener, ele possui 5,9Mbp a mais de sequências relacionadas às famílias gênicas de trans-sialidade, MASP, mucina, GP63, RHS e DGF1 do que a cepa Sylvio10/1. Apesar desse maior número de cópias de genes das famílias multigênicas por causa da diferença no tamanho dos dois genomas, *T. cruzi* Sylvio X10/1 possui uma proporção maior de cópias das famílias GP63, trans-sialidade, mucina e RHS se comparado o número de cópias em relação ao tamanho do respectivo genoma. A identidade nucleotídica entre Sylvio X10/1 e Non-Esmeraldo (DTU TcIII) foi de 98,2% e entre Sylvio X10/1 e Esmeraldo (DTU TcII) foi de 97,5%. A identidade nucleotídica média entre os dois haplótipos de CL Brener Esmeraldo e Non-Esmeraldo é de 97,8%.

Essa comparação em escala genômica evidencia que a linhagem DTU TcI (Sylvio X10/1) é mais próxima filogeneticamente a DTU TcIII (Non-Esmeraldo) do que com DTU TcII (Esmeraldo). As divergência entre essas três linhagens de *T. cruzi* é maior do que a divergência entre as sub-espécies de *T. brucei*, *T. brucei brucei* e *T. brucei gambiense* (99,2%) (JACKSON *et al*., 2010). Mas é menor do que a diferença entre *L. major* e *L. infantum* (94%) (PEACOCK *et al*., 2007).

O genoma haplóide de *T. brucei* foi montado em 11 cromossomos, constituído por 9.068 genes codificadores de proteína. Para a maioria dos cromossomos foi possível montar a região sub-telomérica, que corresponde a mais de 20% dos genes codificados no genoma, sendo a maioria genes específicos de *T. brucei*, relacionados à capacidade de variação antigênica do parasita na forma sanguínea presente no hospedeiro mamífero.

Os genomas das três espécies de *Leishmania* foram montados em uma única seqüência contígua para cada um de seus cromossomos, 36 para *Leishmania major* e *L. infantum* e 35 para *L. braziliensis*. Os três genomas apresentam

aproximadamente o mesmo tamanho, cerca de 34 Mb. Foram anotados 9.155 (8.264), 8.185 (7.992) e 8.127 (7.896) genes para *L. major*, *L. infantum* e *L. braziliensis*, respectivamente, sendo o número entre parênteses o número de genes codificadores de proteína (http://www.ncbi.nlm.nih.gov/sites/entrez).

Uma análise comparativa foi feita entre os genomas de *T. brucei*, *T. cruzi* e *L. major*, colapsando os genes parálogos proximamente relacionados, utilizando o melhor resultado mútuo de BlastP entre o proteoma das três espécies, e agrupando em grupos de genes ortólogos (COGs, do inglês Clusters of Orthologous Genes) (EL-SAYED *et al*., 2005b). Essa análise revelou um conjunto de 6.158 COGs compartilhados por essas três espécies e 1.014 COGs compartilhados por duas dessas espécies (

FIGURA 3). A identidade entre uma grande amostra dos COGs compartilhados entre os três organismos é de 57% entre *T. brucei* e *T. cruzi* e 44% entre *L. major* e os outros dois tripanossomatídeos, refletindo a relação filogenética esperada entre eles. Os parasitas intracelulares, *T. cruzi* e *L. major*, apresentam ligeiramente mais COGs compartilhados apenas entre as duas espécies (n=482) do que entre *T. brucei* e *T. cruzi* (n=458) e consideravelmente mais do que *T. brucei* e *L. major* (n=74). O restante do proteoma é formado por membros espécie-específicos, onde *T. brucei* (32%) e *T. cruzi* (26%) possuem uma proporção de genes espécie-específicos no genoma bem maior do que *L. major* (12%). Apesar da divergência de 200 a 300 milhões de anos entre *T. brucei* e *T. cruzi* e de 400 a 600 milhões de anos entre os gêneros *Trypanosoma* e *Leishmania* (OVERATH *et al*., 2001), seus genomas apresentam um grau muito grande de sintenia. De todos os genes de *T. brucei* e *L. major*, 68 e 75%, respectivamente, permanecem no mesmo contexto genômico. Além disso, quase todos (94%) COGs compartilhados entre as três espécies, que formam o cerne do proteoma, estão em regiões de sintenia conservada (EL-SAYED *et al*., 2005b).

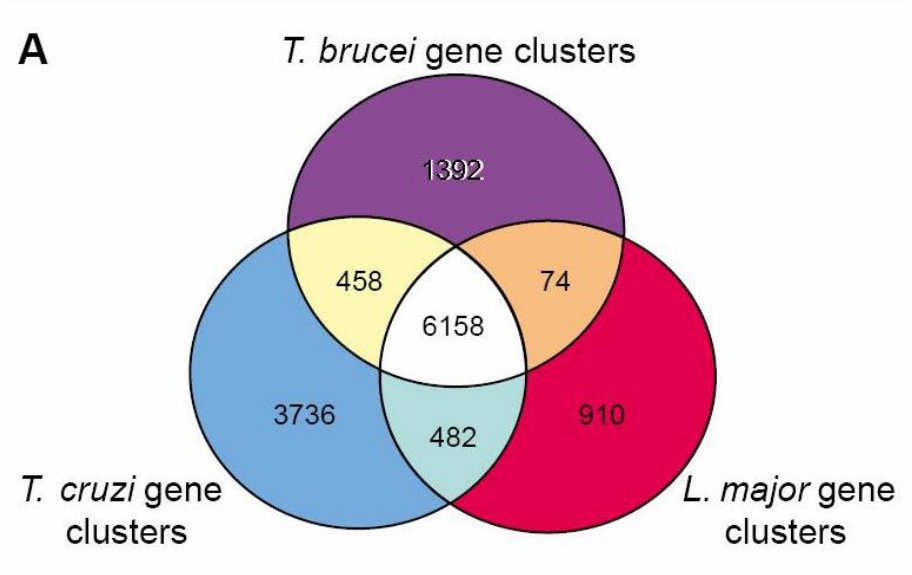


FIGURA 3 – Distribuição de COGs entre *T. brucei, T. cruzi e L. major*.

FONTE: EL-SAYED *et al*., 2005b

A comparação entre os genomas das três espécies de *Leishmania* mostrou uma conservação na sintenia de mais de 99%. A conservação das regiões codificadoras também é alta: a média de identidade nas seqüências nucleotídicas e aminoacídicas são, respectivamente, 94 e 92% entre *L. major* e *L. infantum*, 82 e 77% entre *L. major* e *L. braziliensis*, e 81 e 77% entre *L. infantum* e *L. braziliensis*. Enquanto *L. major* e *L. infantum* possuem 36 cromossomos, *L. braziliensis* possui 35 devido a uma aparente fusão dos cromossomos 20 e 34 (BRITTO *et al*., 1998). Com base na similaridade de seqüência e na arquitetura cromossômica, *L. braziliensis* claramente é a espécie menos similar entre as três, consistente com a classificação de subgênero. *L. braziliensis* pertence ao subgênero *Leishmania* e as outras duas espécies ao subgênero *Viannia*. Apesar das grandes diferenças entre os fenótipos dos diferentes tipos de leishmanioses, foram encontrados poucos genes espécie-específicos nas espécies de *Leishmania*. *L. major*, *L. infantum* e *L. braziliensis* possuem 5, 26 e aproximadamente 47 genes espécie-específicos, respectivamente. Pelo tempo de divergência dentro do gênero *Leishmania* ser de 20 a 100 milhões de anos, esse pequeno número de diferenças espécie-específicas no conteúdo gênico não era esperado. Na procura por genes que estivessem evoluindo sob pressão de

seleção positiva, como indicador de que eles estivessem envolvidos na interação parasita-hospedeiro, a maioria são genes envolvidos em processos biológicos não conhecidos (PEACOCK *et al*., 2007).

O genoma de *L. tarentolae* Parrot-TarII, que não é patogênica a humanos e utilizada como modelo de estudo de *Leishmania* sp., foi recentemente sequenciado (RAYMOND *et al*., 2012). A cobertura final foi de 16 vezes, e inicialmente montado em 773 scaffolds e 2.499 contigs, totalizando 34,4 Mbp. Utilizando hibridização de fragmentos cromossômicos com as outras três espécies de *Leishmania*, foi possível montar o genoma de *L. tarentolae* em 36 cromossomos. *L. tarentolae* possui 8.201 genes codificadores de proteína, dos quais, mais de 90% são compartilhados com as outras espécies de *Leishmania*. Quando comparada às outras três espécies de *Leishmania*s já seqüenciadas, *L. tarentolae* possui 95 sequências únicas e 250 que estão presentes apenas nas outras espécies.

Vários dos genes que não estão presentes em *L. tarentolae* são expressos preferencialmente nas formas amastigotas das espécies patogênicas. Isso poderia explicar em parte porque *L. tarentolae* é menos adaptada a infectar macrófagos humanos e porque geralmente são reportados como organismos de vida livre em lagartos. Embora *L. tarentolae* possua identidade nucleotídica similar a *L. infantum* e *L. major*, ela possui uma sintenia maior com *L. major*. Em alguns longos trechos, *L. tarentolae* e *L. infantum* são sintênicas, mas a ordem dos genes é diferente.

O genoma de *L. donovani*, que causa leishmaniose visceral, foi sequenciado. Além disso, foram estudados aspectos estruturais, populacionais e de resistência à droga em 17 isolados diferentes. O genoma foi sequenciado com uma cobertura de 22 vezes, e comparado contra o genoma referência de *L. infantum.* Foi produzida uma montagem de alta qualidade de 32,4Mbp contituídos por 2.154 contigs. 8.252 genes foram transferidos para *L. donovani*, oriundos da anotação de *L. infantum*. Os diferentes isolados de *L. donovani*, provenientes de pacientes com leishmaniose cutânea com diferentes respostas ao tratamento com antimônio, foram seqüenciados com uma cobertura de 66 vezes cada um. A comparação entre os diferentes isolados mostrou evidência de seleção atuando sobre uma série de genes associados a genes de superfície e genes de transporte, incluindo genes associados à resistência à droga. A análise da estrutura do genoma mostrou que a resistência à drogas contendo antimônio surgiu múltiplas vezes nessas linhagens.

1.4 TRANSCRIÇÃO POLICISTRÔNICA

De uma maneira geral, os genes de eucariotos são constituídos por uma região reguladora (promotor), regiões não traduzidas (UTRs) 5’ e 3’, íntrons e éxons. A transcrição de um gene é controlada pelo seu promotor, iniciando na UTR 5’, passando pelos íntrons e éxons e terminando na UTR 3’. Os genes codificadores de proteína são transcritos pela RNA polimerase II e o transcrito primário deve passar por três modificações co-transcricionais para que sejam funcionais: adição do *cap* na extremidade 5’, que é uma base modificada, metil-guanosina-trifosfato (m7Gppp ou *cap*); processamento (“splicing”), que é a excisão dos íntrons e união dos éxons, que por se tratar da mesma molécula também é chamado de “cis-splicing”; e poliadenilação, que é a adição de uma cauda poli-A na extremidade 3’ (WRAY *et al*., 2003).

Os três tipos de RNA polimerase, I, II e III, descritos em eucariotos mais derivados foram identificados em tripanossomatídeos. Alguns dos genes que codificam as subunidades dessas enzimas foram clonados e caracterizados. A RNA polimerase I transcreve genes ribossomais (28S, 5,8S, e 18S rRNAs) e os genes codificadores das proteínas variantes de superfície (VSGs) e da prociclina em *T. brucei*. A RNA polimerase II transcreve os demais genes codificadores de proteína (mRNAs), o gene do mini-éxon (ou seqüência líder, SL-RNA) e alguns pequenos RNAs nucleolares (snoRNAs). Já a RNA polimerase III transcreve o gene ribossomal 5S (5S rRNA), os RNAs de transferência (tRNAs) e os pequenos RNAs nucleares (snRNAs) (CLAYTON, 2002; CAMPBELL *et al*., 2003).

No entanto, os tripanossomatídeos apresentam diferenças com relação ao perfil de transcrição dos demais eucariotos. De modo geral, os genes codificadores de proteína em tripanossomatídeos não apresentam promotores canônicos (CAMPBELL *et al*., 2003) e íntrons (MAIR, *et al*., 2000), embora existam exceções. Como por exemplo: os genes das proteínas variantes de superfície (VSGs) e da prociclina em *T. brucei* possuem seus respectivos promotores; o gene codificador da poli-A polimerase em *T. cruzi* e *T. brucei* (MAIR, *et al*., 2000) possui dois éxons e um íntron que é removido por “cis-splicing”.

Os genes dos tripanossomatídeos estão dispostos nos cromossomos como

grupos de genes direcionais (DGCs, “directional gene clusters”), que formam longos arranjos de genes sem função relacionada na mesma fita do DNA. Os DGCs alternam entre as duas fitas dos cromossomos podendo convergir ou divergir quanto à sua direção de transcrição (HALL *et al*., 2003; IVENS *et al*., 2005). Essa organização gênica e as regiões de mudança na fita codificadora são conservadas em grande parte no genoma de *T. brucei* e *L. major*, que são os tripanossomatídeos com a montagem do genoma de melhor qualidade (EL-SAYED *et al*., 2005b).

A transcrição nos tripanossomatídeos produz unidades policistrônicas que são processadas por *trans-splicing* para produzir unidades monocistrônicas traduzíveis. O *trans-splicing* envolve duas moléculas diferentes, uma doadora (chamada precursora do mini-éxon) e uma aceptora (transcrito derivado da unidade policistrônica), gerando um transcrito monocistrônico com o mini-éxon (seqüência de 39 pb espécie-específica) na extremidade 5’ e uma cauda poli-A na extremidade 3’ (CLAYTON, 2002).

1.5 REGULAÇÃO DA EXPRESSÃO GÊNICA

Os eucariotos utilizam diversos mecanismos para regular a expressão gênica, incluindo condensação da cromatina, metilação do DNA, controle de início da transcrição, processamento alternativo do mRNA, exportação do mRNA do núcleo para o citoplasma, estabilidade do mRNA, controle de início da tradução, várias formas de modificações pós-traducionais de proteínas, tráfico intracelular e degradação da proteína. Desses diversos níveis, o ponto mais comum de controle da expressão gênica é a taxa de início da transcrição. Desse modo, o controle transcricional parece ser o determinante primário, ou um dos determinantes mais importantes, do perfil geral da expressão gênica, em especial o controle do início da transcrição pela RNA polimerase II, já que os genes codificadores de proteína compreendem a grande maioria dos genomas eucarióticos e apresentam as mais diversas funções (WRAY *et al*., 2003).

A regulação da expressão gênica em tripanossomatídeos apresenta diversas características peculiares, como a cromatina menos condensada, a edição dos RNAs mitocondriais, a disposição dos genes em, a transcrição policistrônica, o *trans-splicing*, a transcrição por RNA polimerase I de alguns genes codificadores de

proteína e a ausência de regiões promotoras típicas para a transcrição dos genes codificadores de proteína pela RNA polimerase II (BELLI, 2000, CLAYTON, 2002; CAMPBELL *et al*., 2003).

A transcrição policistrônica e a ausência de promotores da RNA polimerase II para a transcrição dos genes codificadores de proteína sugere que a regulação da expressão gênica em tripanossomatídeos seja principalmente pós-transcricional. Deste modo, a regulação pode ocorrer nos seguintes pontos: processamento do transcrito primário, exportação do núcleo para o citoplasma, estabilidade do mRNA, controle de início da tradução, modificações pós-traducionais e degradação da proteína (FIGURA 4).

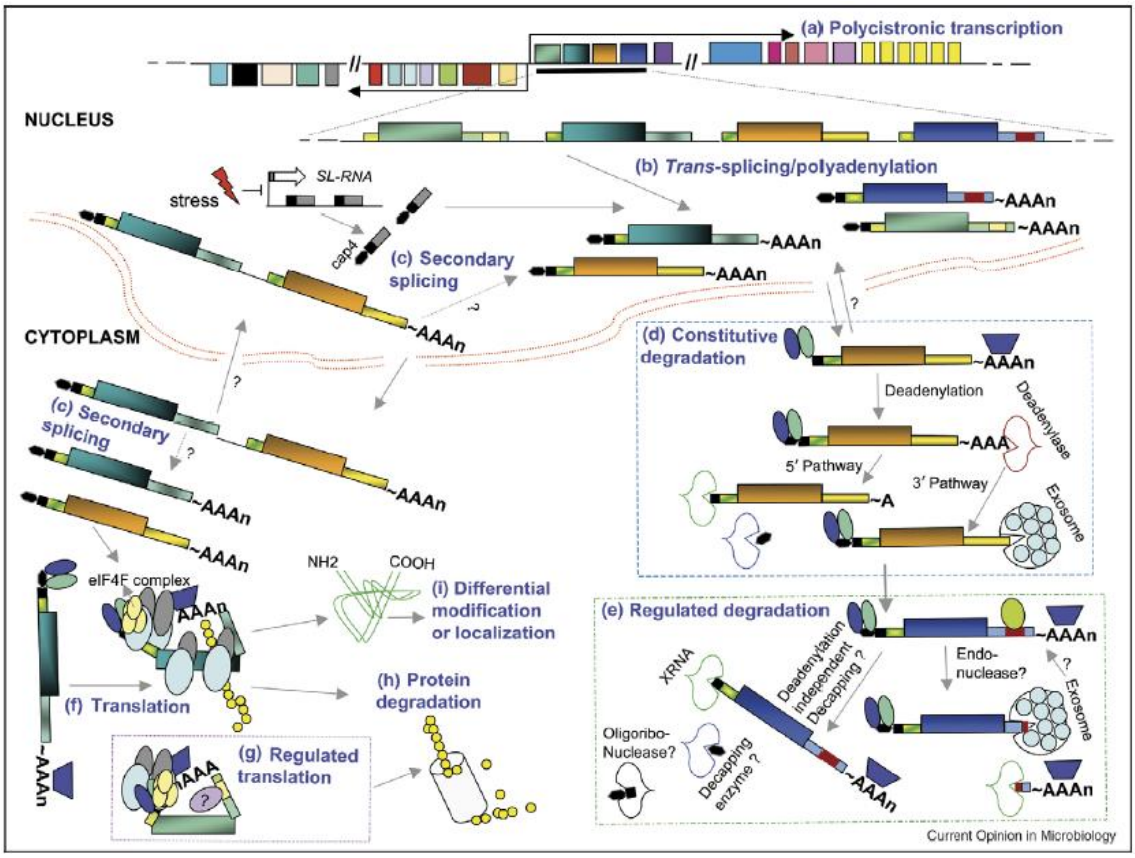


FIGURA 4 – Diferentes mecanismos de controle da expressão gênica em tripanossomatídeos

. (a) Transcrição policistrônica. (b) *trans-splicing* e poliadenilação acoplados durante a transcrição. (c) Processamento de mRNAs dicistrônicos que são armazenados e posteriormente, novamente processados gerando mRNAs monocistrônicos. (d) Degradação do mRNA dependente de desadenilação pela extremidade 5' e/ou 3'. (e) Degradação do mRNA independente de desadenilação. (f) Início da tradução. (g) Alongamento da tradução. (h) Degradação de proteínas. (i) Modificações pós-traducionais de proteínas ou direcionamento da localização da proteína. ? indica questões não confirmadas experimentalmente.

FONTE: HAILE & PAPADOPOLOU, 2007

1.5.1 Regulação Pós-transcricional

Foi sugerido que o fato dos tripanossomatídeos possuírem transcrição policistrônica, o trans-splicing e a poliadenilação ocorrerem ao mesmo tempo, e que a maioria dos genes que são co-transcritos não possuem função relacionada fizesse a regulação do processamento do transcrito policistrônico ser pouco provável (HAILE & PAPADOPOLOU, 2007). Porém, dados recentes mostram que o “trans-splicing alternativo”, ou seja, a existência de diferentes sítios de adição do mini-éxon e cauda poli-A nas UTRs de um transcrito, pode ser um ponto de regulação da expressão gênica (KOLEV *et al*., 2010; SIEGEL *et al*., 2010).

Siegel e colaboradores analisaram o transcriptoma das formas procíclicas e sanguíneas de *T. brucei* utilizando RNA-Seq. Além da análise de expressão gênica para mais de 90% dos genes, foi possível identificar os sítios de adição de mini-éxon e de cauda poli-A para 6.959 e 5.948 genes, respectivamente. A maioria dos genes continha entre um e três sítios aceptores de mini-éxon e o número de sítios de adição de poli-A foi bastante variável e também difícil de determinar devido a baixa complexidade das sequências das regiões 3’UTR. Mas para os genes com mais de 20 tags mapeados, o número médio de sítios de poliadenilação foi 10, o que demonstra certa “promiscuidade” na poliadenilação (SIEGEL *et al*., 2010).

Kolev e colaboradores utilizaram RNA-seq de bibliotecas enriquecidas para a captura das extremidades 5’ e 3’ dos transcritos. Essa estratégia permitiu a delimitação das extremidades 5’ e 3’ de 8.960 transcritos (KOLEV *et al*., 2010). Para os 8.592 transcritos com mais de 10 tags do mini-éxon, apenas 926 (11%) possuíam apenas um sítios de adição de mini-éxon, já 5.327 transcritos (62%) tinham entre dois e quarto sítios, e 2.339 (27%) tinham 5 ou mais sítios. Os sítios de poliadenilação se mostraram ainda mais heterogêneos.

Essas novas evidências demonstram a importância do processamento diferencial dos transcritos como um importante ponto da regulação da expressão gênica em tripanossomatídeos, uma vez que diferentes variantes do trasncrito podem conter diferentes elementos regulatórios nas UTRs e determinar a regulação gênica diferencial.

A exportação do mRNA do núcleo ainda é pouco estudada em tripanossomatídeos (HAILE & PAPADOPOLOU, 2007). Foi sugerido que um transcrito longo poderia ser processado de maneira regulada depois de ser

exportado do núcleo (FIGURA 4), mas a evidência é circunstancial (JÄGER *et al*., 2007). Utilizando genômica comparativa, Serpeloni e colaboradores mostraram que a via de exportação de mRNAs é a menos conservada em Excavata, e sugerem que esta via de exportação seja bastante diferente da via dos demais grupos de eucariotos (SERPELONI *et al*., 2011a). Em outro estudo, Serpeloni e colaboradores caracterizaram funcionalmente o gene Sub2, gene do complexo TREX de exportação de mRNAs, previamente identificado como um gene conservado em Excavata (incluindo *T. cruzi* e *T. brucei*). Em *T. cruzi*, utilizando microscopia de fluorescência, a proteína desse gene foi localizada no núcleo. E por microscopia eletrônica de varredura a proteína foi localizada na interface entre as áreas densa e não densa da cromatina, áreas tipicamente associadas à transcrição/processamento de mRNA. Além disso, ensaios de incorporação de BrUTP mostraram co-localização da proteína Sub2 com sítios de transcrição de RNA polimerase II, o que sugere a participação dessa proteína no metabolismo de mRNA no núcleo. Em experimentos de interferência de RNA do gene Sub2 em *T. brucei* foi observado o acúmulo de mRNA no núcleo e a diminuição dos níveis de tradução, confirmando essa proteína como componente da via de exportação/transcrição de mRNAs em tripanossomatídeos (SERPELONI *et al*., 2011b).

O atual modelo de degradação de mRNA em tripanossomatídeos envolve duas vias: a via dependente de desadenilação e a via independente de desadenilação (FIGURA 4; CLAYTON & SHAPIRA, 2007).

Na primeira, a via começa com a desadenilação do mRNA, que pode levar a duas situações: (1) a degradação do mRNA por exonucleases 3', a maioria associada ao complexo exossomo; (2) ou à estimulação da remoção do *cap* e subseqüente degradação do mRNA pela extremidade 5'. Para a maioria das enzimas de levedura e mamíferos associadas à degradação de mRNAs, foram encontrados homólogos em tripanossomatídeos: exossomo, exonucleases 3', exonucleases 5', enzimas associadas à via de degradação de mRNAs com códons de parada prematuros (NMD, do ingês *Nonsense-mediated Decay*), enzimas que se ligam ao *cap* e enzimas de desadenilação. Embora a atividade de enzimas que removem o *cap* foi encontrada em extratos de tripanossomatídeos (Milone *et al*., 2002), as proteínas envolvidas ainda não foram identificadas. Essa via é constitutiva e possui uma cinética lenta de degradação dos mRNA estáveis e provavelmente uma subpopulação dos mRNAs instáveis (CLAYTON & SHAPIRA, 2007; HAILE &

PAPADOPOULOU, 2007).

Por outro lado, os mRNAs também podem ser degradados pela ação de enzimas que removem o *cap* e subseqüente ação de exonucleases 5’ ou endonucleases. Essa via, independente da desadenilação, é mais rápida e degrada os mRNA mais instáveis (CLAYTON & SHAPIRA, 2007; HAILE & PAPADOPOULOU, 2007).

A estabilidade de uma molécula de mRNA pode ser determinada por elementos cis (na própria molécula de mRNA) e elementos trans (outras moléculas), como por exemplo, regiões não traduzidas 3’ (UTRs 3’) e proteínas que se ligam a RNA (RBPs, do inglês RNA-binding Proteins), respectivamente.

A maioria dos elementos cis reguladores identificados até o momento encontram-se na UTR 3’, embora também possam ser encontrados na UTR 5’ e regiões intergênicas (CLAYTON, 2002). Elementos ricos em AU (AREs, do inglês AU-rich Elements) foram identificados na UTR 3’ dos genes “small mucine gene” (SMUG) e “EP-rich procyclin” (EP1) em *T. cruzi* e *T. brucei*, respectivamente (D’ORSO & FRASCH, 2001; IRMER & CLAYTON, 2001). Por outro lado, elementos ricos em G (GREs, do inglês G-rich Elements) também foram encontrados na UTR 3’ do gene SMUG. AREs e GREs atuam de maneira diferente, eles desestabilizam e estabilizam os mRNAS, respectivamente, de maneira estágio-específica (D’ORSO & FRASCH, 2001; D’ORSO *et al*., 2003).

Vasconcelos e colaboradores utilizaram uma abordagem in silico em escala genômica para identificação de possíveis elementos regulatórios nas regiões 5’ e 3’ UTR de *Leishmania*. Eles fizeram uso de genômica comparativa para identificar seqüências intergênicas conservadas entre os genomas de *L. major, L. infantum* e *L. braziliensis.* As seqüências conservadas foram filtradas para remover elementos repetitivos e agrupadas para a obtenção de uma sequência comum ao conjunto de genes de cada grupo. Para cada sequência comum foram preditas as extremidades dos transcritos. Parte dos grupos de seqüências conservadas foi utilizada como training set para posterior busca por possíveis elementos regulatórios nos grupos de sequências conservadas obtidos. Foram encontrados motivos regulatórios previamente descritos, e conservados nas três espécies de *Leishmania*. Além disso, também foram encontrados novos possíveis elementos regulatórios nas UTRs de *Leishmania*.

Os genomas de *Trypanosoma* e de *L. major* apresentam um grande número

de RBPs. Eles possuem proteínas que possuem os domínios de interação com RNA: RRM (n>75), Pumilio (PUF, n=13) e dedo de zinco CCCH (n>50) (IVENS, *et al*.; 2005, De GAUDENZI *et al*.; 2005). Apesar disso, poucas RBPs foram verificadas experimentalmente como tendo função regulatória. Uma maneira de se verificar isso é usando técnicas de ribonômica, que buscam identificar os mRNAs que interagem com uma determinada proteína. Por exemplo, as proteínas TcUBP1 e TcUBP2 de *T. cruzi* estão envolvidas na regulação estágio-específica de mRNAs de SMUG (D'ORSO *et al*., 2003). Assim como em *T. brucei*, alterações na expressão dos genes TbUBP1 e TbUBP2 afetam a expressão de diversos mRNAs de genes envolvidos no ciclo celular (HARTMANN *et al*., 2007). A ribonômica é uma das abordagens que está sendo utilizada pelo Instituto Carlos Chagas no esforço de entender os mecanismos de regulação pós-transcricional. Os mRNAs associados à proteína TcPUF6 foram identificados (DALLAGIOVANNA *et al*., 2008) e outras proteínas estão sendo avaliadas.

1.5.2 Regulação Pós-traducional

Em eucariotos a tradução inicia com a ligação do complexo ternário eIF4F (formado por eIF4E, eIF4A e eIF4G) ao *cap* do mRNA através da subunidade eIF4E (SONENBERG & DEVER, 2003). Pouco se sabe em relação aos mecanismos de tradução em tripanossomatídeos, mas eles possuem homólogos dos três genes que formam o complexo eIF4F, portanto, este é um potencial ponto de regulação da expressão gênica (CLAYTON & SHAPIRA, 2007, HAILE E PAPADOPOULOU, 2007).

Pouco se sabe sobre a regulação da expressão gênica pós-traducional em tripanossomatídeos, mas seus genomas contêm um conteúdo relativamente alto de quinases (PARSONS *et al*., 2005). Em *Crithidia fasciculata*, a fosforilação de proteínas que se ligam a RNA (RBPs, "RNA binding proteins") está relacionada a estabilidade de mRNAs durante o ciclo celular (Mittra e Ray, 2004). Deste modo, acredita-se que as modificações pós-traducionais de proteínas possa desempenhar um papel importante na regulação gênica (CLAYTON & SHAPIRA, 2007; HAILE & PAPADOPOULOU, 2007).

A proteína EP1, uma prociclina (proteína de superfície abundantemente expressa na forma procíclica de *T. brucei*), é pouco expressa na forma sanguínea do

parasita. No entanto, quando este sofre um estresse por baixa temperatura, a proteína é produzida, mas ao invés de ser direcionada para a superfície da membrana, é retida na via biossintética-secretora. Este controle no direcionamento da proteína até “o último minuto” permite que o parasita reaja quase instantaneamente quando ele se encontrar abruptamente dentro do inseto vetor (ENGSTLER & BOSHART, 2004).

1.6 Projeto Reguloma de *Trypanosoma cruzi*

Apesar da relevância do tópico de regulação da expressão gênica em tripanossomatídeos e a quantidade relativamente grande de trabalhos sobre o tema, esse é um tópico ainda pouco entendido. Com o advento das técnicas em larga escala, dentro da perspectiva de análises ômicas, é possível avaliar esse tópico extremamente relevante de uma forma inédita.

Nosso grupo, bem como outros pesquisadores, vem utilizando sistematicamente ferramentas como microarranjo, sequenciamento de nova geração ou proteômica para caracterizar o conteúdo global da expressão gênica em tripanossomatídeos (DIEHL *et al.,* 2002; SAXENA *et al.,* 2003; MINNING *et al.,* 2003; BAPTISTA *et al.,* 2004; BREMS *et al.,* 2005; MURTA *et al.,* 2006; HOLETZ *et al.,* 2006; McNICOLL *et al.;* 2006; LUU *et al.,* 2006; SAXENA *et al.,* 2007; SINGH *et al.,* 2007; SRIVIDYA *et al.,* 2007; DALLAGIOVANNA *et al.,* 2008; ROCHETTE *et al.,* 2008; KOUMANDOU *et al.,* 2008; MINNING *et al.,* 2009; KABANI *et al.,* 2009; QUEIROZ *et al.,* 2009; DA SILVA *et al.,* 2009; VEITCH *et al.,* 2010; SIEGEL *et al.,* 2010; HOLETZ *et al.,* 2010; ALCOLEA *et al.,* 2010; KOLEV *et al.,* 2010; FENG *et al.,* 2011; DO MONTE-NETO *et al.,* 2011; MICHAELI *et al.,* 2011;ARCHER *et al.,* 2011; MANFUL *et al.,* 2011; MARCHINI *et al.,* 2011; GRYNBERG *et al.,* 2012; GODOY *et al.,* 2012).

A grande quantidade de trabalhos já publicados relacionado a análises ômicas em tripanossomatídeos demonstra a capacidade dessas técnicas em evidenciar padrões gerais de expressão gênica nesses organismos. Porém, em sua grande maioria, esses artigos focam em questões pontuais de interesse de sua biologia ou, quando sistêmico, se restringem a avaliar questões tradicionais. Esses dados podem ser utilizados para uma abordagem computadorizada, de mineiração

de dados, para a identificação de módulos de expressão em tripanossomatídeos. No entanto, uma abordagem mais sistemática é necessária.

Nesse sentido, nosso grupo propôs recentemente o projeto “reguloma de *T. cruzi”* cujo objetivo maior é identificar globalmente os padrões de co-regulação de seu transcriptoma e proteoma, bem como os elementos em cis e trans que os determinam.

O projeto Reguloma de *T. cruzi* se apoia em sub-projeto principais, os quais são:

Avaliação extensa e sistemática do transcriptoma de *T. cruzi:* a capacidade de se identificar módulos de co-regulação está diretamente relacionada à análise bioinformática de co-expressão, a qual necessita, para ser realizada com profundidade, de uma grande quantidade de dados do transcriptoma na mais ampla gama de situações. Isso se deve pelo fato de que os padrões observados podem ser causados pelos mais diversos motivos e, para a correta definição de módulos de co-expressão uniformes, é necessário ampliar o número de condições para que as diferentes situações nos quais os genes apresentam o mesmo padrão de expressão sejam fragmentadas, restando ao final módulos de co-expressão “puros”. Esse raciocínio é explicitado na FIGURA 5.

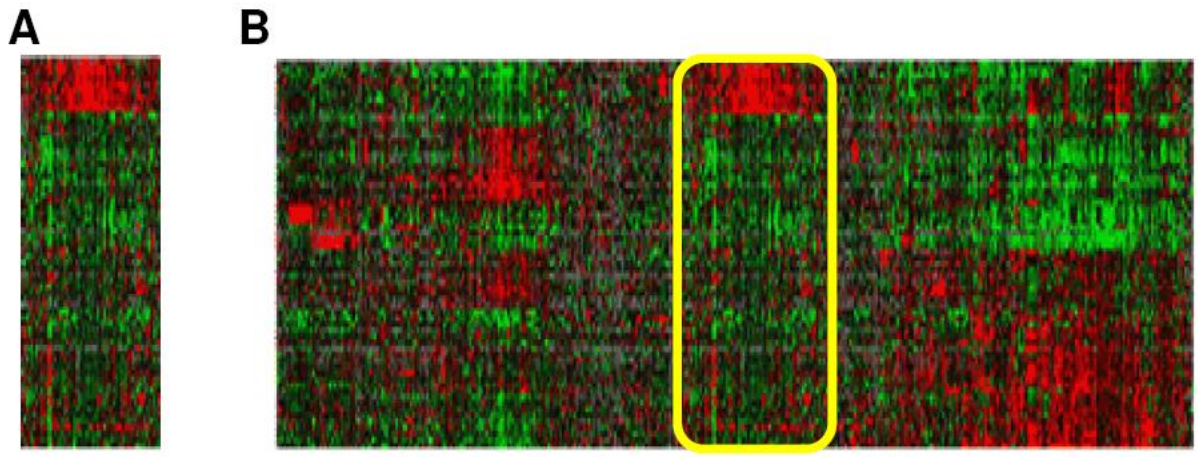


FIGURA 5 - Demonstração da validade da ampliação da avaliação do transcriptoma.

Heatmaps demonstrando o padrão de expressão de um conjunto de genes em diversas situações, sendo que as linhas representam genes e as colunas representam situações biológicas distintas; em A, é possível evidenciar dois grupos de genes co-expressos, um com padrão aumentado nas situações avaliadas (células em vermelho; grupo 1) e outro com padrão diminuído (células em verde; grupo 2), os quais podem ser utilizados para análises subsequentes visando identificar elementos em cis determinadores do padrão observado. No entanto, ao ampliarmos a avaliação do transcriptoma para um conjunto muito mais amplo de situações (B), podemos observar que o grupo 1 continua apresentando forte co-expressão, porém o grupo 2

pode ser separado em pelo menos mais 4 subgrupos. É provável que a identificação dos elementos reguladores seja muito mais factível na segunda análise. A caixa em amarelo representa a posição das condições visualizadas em A.

Portanto, para o sub-projeto de avaliação do transcriptoma, estamos estudando a resposta de *T. cruzi* a um grande conjunto de situações:

- Ciclo de vida: As fases principais do ciclo de vida de *T. cruzi,* nominalmente amastigotas, epimastigotas e tripomastigotas sanguíneos e metacíclicos foram avaliados (PAVONI *et al.,* em preparação).
- Diferenciação: *T. cruzi* é um excelente modelo de diferenciação, e estamos avaliando os seguintes processos:
  - Metaciclogênese *in vitro*;
  - Amastigogênese *in vitro* a partir de tripomastigota metacíclico;
  - Amastigogênese *in vitro* a partir de tripomastigota de cultivo;
  - Epimastigogênese *in vitro* a partir de tripomastigota metacíclico;
  - Epimastigogênese *in vitro* a partir de tripomastigota de cultivo;
  - Metaciclogênese *in vivo;*
  - Amastigogênese *in vivo;*
- Resposta a drogas: Estimular o transcriptoma com drogas diversas que afetam a viabilidade do parasita:
  - Inibidores da biossíntese de ergosterol (KESSLER *et al.,* submetido; KESSLER *et al.,* em preparação);
  - Inibidores de quinase (diversos);
  - Outros fármacos.
- Resposta a estímulos ambientais:
  - Estresses nutricionais: meio TAU (urina artificial de triatomíneo), PBS, TAU3AAG (TAU acrescido de glicose, ácido aspártico, glutâmico e prolina).
  - Estresses por temperatura: 10oC, 16oC, 37oC, 41oC.
  - Estresses por pH: pH 4, pH 5,8, pH 8,5.
  - Estresses oxidativos: 20 µM e 200 µM de H_2O_2
  - Estresses osmóticos: hipo e hiperosmótico
- Pertubações gênicas:
  - Nocautes gênicos em *T. cruzi:* estamos desenvolvendo uma

plataforma computacional e laboratorial para a realização de nocautes gênicos sistemáticos em *T. cruzi* (LORUSSO *et al.,* em preparação).

- Modificações metabólicas:
  - Modificação de meio quimicamente definido: inicialmente, avaliamos a eficiência do meio AR103 (ROITMAN & AZEVEDO, 1984), sem sucesso. Finalmente, conseguimos crescer *T. cruzi* em meio HX25 modificado, com ou sem soro bovino fetal (LIMA *et al.,* em preparação), o que possibilitou a realização de ensaios metabólicos com maior qualidade. Nosso plano em relação ao projeto reguloma é avaliar sistematicamente a resposta da expressão gênica de *T. cruzi* à retirada de elementos do meio.
- Ciclo celular
- Interação com células hospedeiras

- **Identificação das associações entre proteínas:** nessa etapa do processo, estamos sistematicamente avaliando o interatoma de *T. cruzi* utilizando o sistema de duplo-híbrido de leveduras. Para o projeto reguloma, o foco primário são as proteínas ligadoras de RNA (RBPs), as quais tem um papel essencial na regulação pós-transcricional da expressão gênica (PRETI *et al.,* em preparação). Um passo essencial para esse sub-projeto foi a produção de um ORFeoma de *T. cruzi,* o qual consiste na clonagem em um vetor apropriado de todas as regiões codificadoras completas de *T. cruzi.*
- **Identificação das associações entre RNA e proteína:** a partir do ORFeoma de *T. cruzi,* inserimos uma etiqueta específica às extremidades amino e carbóxi-terminal de RBPs, individualmente, e as transfectamos em *T. cruzi.* Utilizando um anticorpo específico para a etiqueta, é feito a imunoprecipitação dos complexos RBP-mRNA e a subsequente identificação dos mRNAs por RNA-Seq, dentro do que é chamado ribonômica.
- **Predição bioinformática dos motivos presentes em cis no mRNA:** com os dados obtidos nas duas etapas anteriores, podemos organizar os genes de *T. cruzi* em grupos de potencial co-regulação, tanto pelo padrão de co-expressão quanto pela associação com proteínas regulatórias específicas. Nesses grupos, aplicamos algoritmos de predição de motivos (MEME,

TIMOTHY *et al.,* 1984; FIRE, ELEMENTO *et al.,* 2007), para identificar possíveis elementos reguladores. Esses motivos são seqüências primárias, mas a interação com as proteínas reguladoras em trans geralmente ocorre por motivos secundários ou tridimensionais. Estamos avaliando metodologias para introduzir essa etapa no projeto, porém sua aplicação ainda é restrita (GOODARZI *et al.,* 2012).

- **Identificação laboratorial de motivos presentes em cis no mRNA:** utilizando o mesmo procedimento descrito para a ribonômica, faz-se a ligação cruzada (*crosslinking*) dos mRNAs e proteínas e subsequente digestão do mRNA por endonucleases. Após essa etapa, extrai-se o mRNA associado, o qual é submetido ao sequenciamento para identificação e quantificação dos sítios de ligação (CLIP, *UV crosslinking and immunoprecipitation,* ULE *et al.,* 2003; PAR-CLIP, *Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation,* HAFNER *et al.,* 2010).

Todos esses elementos em conjunto constituem o projeto reguloma. Como todas as abordagens são em larga escala, a forma como esses dados serão analisados será prioritariamente quantitativa e sob a ótica de redes biológicas de regulação gênica. Obviamente, os resultados pontuais, tanto referente às caracterizações biológicas de cada um dos estudos realizados quanto relacionadas aos elementos de regulação da expressão gênica, também poderão ser identificados e analisados.

Associado a esses sub-projetos, há outros elementos que estão acoplados ao projeto Reguloma. A quantificação proteômica das mesmas situações analisadas pelo transcriptoma é muito importante, e está sendo realizada. No entanto, a produtividade relacionada à proteômica é bem menor que a da transcriptômica e por isso ela não está categorizada conjuntamente. Em relação à transcriptômica, estamos também avaliando as taxas de transcrição e de degradação, que são relacionados diretamente com as redes de regulação gênica. Finalmente, postulamos que a comparação evolutiva será muito importante para auxiliar na definição computadorizada dos elementos reguladores em tripanossomatídeos; nesse sentido, estamos também realizando experimentos de genômica, visando sequenciar o genoma de outros tripanossomatídeos, incluindo cepas de *T. cruzi,* para identificar regiões de conservação alta nas UTRs desses genes.

O escopo e abrangência do projeto reguloma é grande, com diversas

iniciativas distintas mas convergentes sendo realizadas por nosso grupo. Obviamente, os trabalhos realizados por outros grupos também serão incorporados e utilizados, no entanto o aspecto principal reforçado aqui é a filosofia da abordagem científica, cuja aplicação é essencial para a obtenção de dados apropriados à identificação dos elementos reguladores. É evidente que tal iniciativa necessita de uma estrutura bioinformática adequada, possibilitando o armazenamento de uma grande quantidade de dados das mais diversas fontes, bem como possibilitando a realização de análises computacionais sobre esses elementos. É nesse contexto que se insere o presente trabalho.

2 OBJETIVOS

2.1 OBJETIVO GERAL

O objetivo geral deste trabalho é primariamente bioinformático, ao construir uma plataforma computadorizada para a realização do projeto Reguloma de *Trypanosoma cruzi.* Essa base é amplamente fundamentada na criação de um banco de dados computadorizado que permite armazenar a grande quantidade de dados gerada, em suas mais variadas formas. Associada a essa base, incorporar métodos analíticos para, nesse momento, realizar análises básicas relacionadas à regulação da expressão gênica, visando aprofundar a compreensão da regulação da expressão gênica em *Trypanosoma cruzi*, utilizando primariamente redes metabólicas e dados de expressão gênica do ciclo de vida do parasita numa abordagem computacional, com ênfase na identificação de elementos regulatórios em regiões não traduzidas.

2.2 OBJETIVOS ESPECÍFICOS

- Implementar um banco de dados para armazenamento, visualização, integração e análise de dados genômicos e pós-genômicos de tripanossomatídeos;
- Representar as vias metabólicas de *T. cruzi* do KEGG em redes metabólicas e identificar genes vizinhos na rede que sejam diferencialmente expressos, e sua possível importância para a regulação da expressão gênica;
- Identificar genes marcadores das fases do ciclo de vida do *T. cruzi* e buscar elementos regulatórios nas UTRs que sejam comuns aos transcritos de cada fase;
- Delimitar os sítios de inserção do mini-éxon e da cada poli-A (extremidades 5’ e 3’) dos transcritos de *T. cruzi*.

3 JUSTIFICATIVA

A importância do estudo de *Trypanosoma cruzi* se baseia em dois aspectos principais.

Primeiro, porque ele é o agente etiológico da Doença de Chagas, uma das doenças com maior distribuição no continente americano e um sério problema de saúde pública. O parasita infecta cerca de 10-15 milhões de pessoas, mata cerca de 14 mil pessoas por ano e estima-se que cerca de 120 milhões de pessoas estejam sob o risco de contaminação. A Doença de Chagas não possui vacina nem medicamentos para tratamento efetivo. Devido ao seu grande impacto sócio-econômico, a pesquisa da biologia básica do parasita que possa levar ao desenvolvimento de quimioterápicos é muito importante.

Segundo, porque os tripanossomatídeos divergiram muito cedo na linhagem evolutiva dos eucariotos e possuem características biológicas extremamente peculiares: transcrição policistrônica, *trans*-splicing, presença de uma sequência comum a todos os mRNAs (mini-éxon), e algumas características únicas, como a ausência de promotores canônicos para os genes codificadores de proteínas e a regulação da expressão gênica majoritariamente pós-transcricional. Além disso, possui um ciclo de vida complexo, alternando entre hospedeiros distantes filogeneticamente, onde passa por processos de diferenciação em diferentes formas. Dessa forma, o estudo dos tripanossomatídeos pode aprofundar o nosso entendimento de mecanismos básicos de biologia celular e molecular, de evolução, de diferenciação celular e da interação parasita-hospedeiro. Além de ser um excelente modelo para o estudo da regulação pós-transcricional da expressão gênica.

Dentro do Projeto Reguloma de *Trypanosoma cruzi* do Instituto Carlos Chagas, que busca identificar e caracterizar as redes de regulação gênica em *T. cruzi*, estão sendo utilizadas diversas abordagens para a obtenção de dados em larga escala de diferentes aspectos da biologia do parasita: análise genômica da cepa referência CL Brener; sequenciamento do genoma da cepa Dm28; análise transcriptômica das fases do ciclo de vida, de processos de diferenciação celular (metaciclogênese e epimastigogênese), da resposta a diferentes tipos de estresses, da meia-vida do mRNA, da resposta à infecção de diferentes tipos celulares, de parasitas nocaute; análise ribonômica (identificação do conjunto de mRNAs

associados a mesma proteína); análise proteômica do ciclo de vida, da metaciclogênese, de proteínas fosforiladas e ubiquitiladas; construção de vetores para caracterização molecular, detecção de interações protéicas e nocaute de genes em larga escala; construção do ORFeoma de *T. cruzi* (biblioteca de clones do conjunto total de genes codificadores de proteínas); análise do interatoma (conjunto de interações proteína-proteína pelo sistema duplo-híbrido em levedura e interações *in vivo*);

Nesse contexto, o presente trabalho se encaixa como um sub-projeto do Projeto Reguloma de *Trypanosoma cruzi* do Instituto, contribuindo para a implementação inicial de uma base de dados para mineração de dados futura, agregando informações das redes metabólicas de *T. cruzi* associadas à expressão gênica, identificando possíveis elementos regulatórios nas regiões não traduzidas de genes marcadores das fases do ciclo de vida do parasita e delimitando as extremidades 5’ e 3’ dos transcritos.

4 MATERIAL E MÉTODOS

4.1 CONSTRUÇÃO DO BANCO DE DADOS

Todos os programas utilizados para a implementação do banco de dados são de código livre, gratuitos e com código fonte disponível. O banco de dados foi desenvolvido utilizando a coleção de programas e ferramentas desenvolvidas pelo projeto GMOD (do inglês Generic Model Organism Database project) (http://gmod.org/wiki/Main_Page). Os dados foram organizados no “schema” Chado (MUNGALL *et al*., 2007, componente do GMOD) utilizando um banco de dados relacional implementado em PostgreSQL (http://www.postgresql.org).

A ferramenta implementada para a visualização dos dados genômicos foi o programa GBrowse (STEIN *et al*., 2002, componente do GMOD). A integração entre os diferentes componentes do GMOD e as ferramentas implementadas para visualização foi feita utilizando “scripts” “in-house” com a linguagem de programação Perl 5 (http://www.perl.org) e o conjunto de módulos BioPerl (http://www.bioperl.org/wiki/Main_Page). O conteúdo do banco de dados foi disponibilizado utilizando o servidor web Apache (http://www.apache.org).

Os dados genômicos das nove espécies de tripanossomatídeos sequenciadas até o momento (*T. brucei*, *T. congolense, T. cruzi*, *L. major*, *L. infantum*, *L. braziliensis, L. mexicana* e *L. tarentolae*) foram utilizados para povoar o banco de dados. As seqüências nucleotídicas e aminoacídicas, bem como a anotação e outras informações desses genomas, foram obtidas do banco de dados TriTrypDB (ASLETT *et al*., 2010).

Os dados do projeto ORFeoma de *Trypanosoma cruzi* do Instituto Carlos Chagas também foram inseridos no banco, mas para isso modificações no “schema” pré-existente foram necessárias, as quais serão comentadas na seção de resultados.

A ontologia utilizada para organizar os dados e a relação entre foi a adotada pela comunidade científica biomédica, “Open Biomedical Ontology” (OBO), incluindo seus projetos, como “Gene Ontology” (GO), “Sequence Ontology” (SO), entre outros (http://www.obofoundry.org).

4.2 ORGANIZAÇÃO DOS GENES DE *Trypanosoma cruzi*

A cepa utilizada para como referência para o projeto “Genoma de *T. cruzi”,* CL Brener, é híbrida. Além disso, o genoma da espécie *T. cruzi* é extremamente repetitivo. Esses dois fatores influenciaram fortemente na montagem e anotação do genoma, o qual se apresenta truncado: foram 8.740 contigs, organizados em 5.489 scaffolds, sendo que os contigs utilizados para a anotação totalizaram 4.008 (EL-SAYED *et al.,* 2005). Foram preditos 23.216 modelos gênicos, sendo que em média um organismo tripanossomatídeo tem entre 9.000 e 10.000 genes; os autores do trabalho estimaram que o conteúdo gênico haploide de *T. cruzi* seria de ~12.000 genes. Muitos desses genes são componentes de famílias multi-gênicas com um grande número de cópias, geralmente mais do que 100, chegando a 1.430 membros para a família das trans-sialidades.

Para organizar essa informação, clusterizamos os modelos gênicos de *T. cruzi* utilizando o algoritmo MCL (VAN DONGEN, 2000), a partir da comparação entre todos elementos usando o *software* BLAST. Para a clusterização, utilizamos o valor de inflação de 5,0. Após a análise, os 23.216 modelos gênicos de *T. cruzi* foram organizados em 9.107 “super-genes” que estão disponíveis na plataforma computadorizada TrypanosOmics (http://www.icc.fiocruz.br/omics/). Esses super-genes foram utilizados para todas as análises subsequentes.

4.3 ANÁLISE DE GENES ADJACENTES DIFERENCIALMENTE EXPRESSOS NAS REDES METABÓLICAS DE *Trypanosoma cruzi*

4.3.1 Representação das vias metabólicas em redes metabólicas

A representação das vias metabólicas de *Trypanosoma cruzi* em formato XML foram obtidas do banco de dados KEGG (do inglês *Kyoto Encyclopedia os Genes and Genomes*).

Foram obtidos 66 arquivos correspondentes a 66 vias metabólicas referência do KEGG que possuem genes de *T. cruzi* mapeados.

Foram utilizados scripts em Perl para, a partir das informações codificadas

nos arquivos XML, obter a lista de quais genes interagem com quais genes, através de seus compostos correspondentes, produzindo no final a lista de pares de genes de *T. cruzi* que estão adjacentes nas diferentes vias metabólicas de *T. cruzi*. Através da informação dos compostos foi possível ligar um gene de uma via metabólica à outra via metabólica.

Para melhor manipular os dados, os mesmos foram inseridos num banco de dados *in house* utilizando PostgreSQL.

Os identificadores (IDs) dos genes de *T. cruzi* foram convertidos em identificadores dos supergenes (SGIDs) de *T. cruzi*, gerados conforme descrito acima.

A visualização da rede metabólica de *T. cruzi*, representando genes "vizinhos" foi feita utilizando o programa Cytoscape (SHANNON *et al*., 2003).

4.3.2 Análise de genes diferencialmente expressos

Cinco amostras biológicas diferentes (réplicas) de cada uma das quatro fases do ciclo de vida de *T. cruzi* (formas epimastigotas, tripomastigotas metacíclicos, amastigotas e tripomastigotas sanguíneos) foram utilizadas para a identificação de genes diferencialmente expressos (DEG, do inglês *Differentially Expressed Genes*). A partir dessas amostras foi extraído mRNA total e convertido em cRNA. Esse material foi processado e sequenciado na plataforma SOLiD (dados de RNA-seq do ciclo de vida de *T. cruzi*, PAVONI *et al*., em preparação).

Por simplificação, as quatro formas de *T. cruzi*: epimastigotas, tripomastigotas metacíclicos, amastigotas e tripomastigotas sanguíneos, durante o texto serão abreviadas como: epi, met, ama e trp, respectivamente.

Foi feita análise para determinar os DEGs de seis comparações diferentes: epi versus met, epi versus ama, epi versus trp, met versus ama, met versus trp e ama versus trp. Os dados de expressão gênica das cinco réplicas biológicas de cada forma foram normalizados e os genes diferencialmente expressos foram determinados usando o pacote edgeR (ROBINSON *et al*., 2010) do ambiente de análises estatísticas R (www.r-project.org). Os DEGs foram selecionados usando um critério de FDR de 10%.

Os DEGs das diferentes comparações 2 a 2 foram mapeados na rede

metabólica contruída e, visando identificar padrões de co-regulação adjacente nessa rede, foram contados os pares de genes adjacentes que eram diferencialmente expressos nas seguinttes situações: ambos aumentados (+/+), ambos diminuídos (-/-), a montante aumentado e a jusante diminuído (+/-), a montante diminuído e a jusante aumentado (-/+), bem como aumentado e diminuído em reações reversíveis, nos quais a direcionalidade da modulação não é relevante.

Para avaliar a significância estastística dos padrões observados nas análises descritas no parágrafo anterior realizamos um procedimento semelhante ao *bootstrap.* Utilizando a mesma estrutura de rede, com o mesmo número de DEGs e incorporando a direção de sua modulação (aumentado ou diminuído), eram selecionados aleatoriamente nós da rede metabólica, aos quais eram atribuídos os padrões de modulação observados. Nessa rede cujo padrão de modulação era aleatório, utilizamos o mesmo algoritmo de contagem empregado para obter os dados observados para avaliar a probabilidade de ocorrência das situações observadas. Foram criadas 10.000 redes aleatórias que foram analisadas e a porcentagem de ocorrência das diferentes situações (+/+, -/-. +/-, -/+) foram computadas, as quais indicam qual é a probabilidade por acaso de sua ocorrência. Esses valores foram comparados com os observados para determinar a significância estatística dos achados.

Também foi feita uma análise estatística baseada em teste exato de Fisher para avaliar se genes modulados, em geral ou em uma das comparações realizadas, tendem a realizar reações enzimáticas irreversíveis.

4.4 GENES MARCADORES E ELEMENTOS REGULATÓRIOS

4.4.1 Genes marcadores

Para a identificação dos genes marcadores do ciclo de vida de *T. cruzi*, utilizamos os mesmo dados de expressão gênica descritas acima. Mas dessa vez analisamos 15 categorias de genes diferentes, todas as combinações possíveis entre as 4 formas do parasita (Epi, Met, Ama, Trp, EpiMeta, EpiAma, EpiTrp, MetaAma, MetTrp, AmaTrp, EpiMetAma, EpiMetTrp, EpiAmaTrp, MetAmaTrp), além da categoria *housekeeping*.

Para um gene ser considerado marcador de uma dessas categorias ele deveria ter uma razão de mudança (*fold change*) maior de 1,5 em relação às demais formas. Para genes serem considerados *housekeeping* eles não poderiam ter uma diferença maior do que 1,5 entre o menor e maior valor de expressão dentre as 20 amostras analisadas (4 formas do parasita, cinco réplicas biológicas por forma).

4.4.2 Identificação de elementos regulatórios na 3'UTR

Os dados obtidos da versão 3.3 do genoma de *Trypanosoma cruzi* disponível no TriTrypDB (ASLETT *et al*., 2010), conforme descrito acima, foram utilizados para a análise.

As coordenadas das regiões intergênicas a jusante de cada um dos 23.311 genes codificadores de proteína foram extraídas a partir do arquivo GFF, utilizando scripts em Perl. Com essas coordenadas, foram obtidas, a partir dos arquivos fasta do genoma, as sequências 3' intergênicas de todos os genes de *T. cruzi*.

As sequências 3'UTR de cada gene utilizadas nesse trabalho foram selecionadas segundo os seguintes critérios: 1) o gene correspondente deveria possuir códon de parada; 2) o gene correspondente não poderia ser pseudogene; 3) foi pego um tamanho máximo de 300 nucleotídeos, caso a sequência não possuísse Ns (nucleotídeos não determinados); 4) Se a sequência possuísse N, a UTR seria considerada até a região do N, desde essa região fosse maior do que 100 nucleotídeos. Os genes cujos padrões não passaram nesse critério foram excluídos da análise.

Utilizamos a unidade super-gene (SG) para análise dos elementos reguladores na região 3'UTR. Como um SG pode ser composto por diversos genes, nosso critério de definição da região 3'-UTR a ser utilizada nas análises subsequentes foi a maior sequência de 3' UTR obtida segundo os critérios estabelecidos acima. Caso houvesse dois ou mais genes com o mesmo tamanho, um gene era selecionado aleatoriamente.

A busca por motivos na região 3'UTR foi feita separadamente para cada conjunto de seqüências, de cada uma das 15 categorias dos genes marcadores do ciclo de vida, conforme explicado no item 4.4.1.

Para a busca de motivos foi utilizado o programa MEME (do inglês *Multiple*

EM for Motif Elicitation), que faz uma busca por motivos sem intervalos (*gaps*) em um conjunto de seqüências não alinhadas (TIMOTHY *et al*., 1994). Foram utilizados os seguintes parâmetros: -nostatus -dna -p 8 -mod zoops -nmotifs 50 -evt 0.001 -maxsize 100000000 -minw 6 -maxw 15.

A opção -nostatus é utilizada para que a análise não gere relatórios durante a sua realização; a opção -dna informa ao programa que as sequências utilizadas são de DNA; a opção -mod com o parâmetro zoops indica ao programa que cada seqüência do conjunto analisado tem no máximo uma ocorrência do motivo; a opção -nmotifs que limita os resultados da análise aos 50 motivos mais significativos; a opção -evt que limita a busca a motivos com um valor mínimo de e-value de $1x10^{-3}$, a opção -maxsize para permitir que o programa utilize mais memória para processar a análise; e as opções -minw e -maxw que determinam o intervalo a ser procurado para o tamanho do motivo. Para esses dois últimos parâmetros realizamos duas análises, uma limitando o intervalo entre 6 e 15 nucleotídeos e outra no intervalo entre 15 e 25 nucleotídeos.

Para a análise da significância do enriquecimento dos motivos encontrados na categoria em questão, os mesmos motivos encontrados foram procurados utilizando o programa FIMO (do inglês *Find Individual Motif Occurences*), que busca um conjunto de sequência por um motivo conhecido (CHARLES *et al*., 2011). Os motivos encontrados em cada categoria foram rodados contra quatro conjuntos de sequências diferentes: 1) o próprio conjunto de sequências que deu origem ao motivo (a categoria em questão), para avaliar a capacidade do programa FIMO em achar os motivos no conjunto origem; 2) todas as regiões 3'-UTRs do genoma de *T. cruzi* selecionadas para a análise conforme os critérios acima; 3) o conjunto das regiões 3'-UTRs da categoria *housekeeping*; 4) e o conjunto das regiões 3'-UTRs da comparação considerada inversa.

O controle inverso de cada categoria é a sua contra-parte, como por exemplo, para a categoria Epi é a categoria MetAmaTrp; para a categoria EpiMeta é a categoria AmaTrp, para a categoria EpiMetAma é a categoria Trp; para a categoria *housekeeping*, utilizamos o conjunto total de DEGs das outras 14 categorias.

A partir do resultado do FIMO, criamos uma matriz 2 por 2 contendo o número de regiões 3'UTR da comparação em destaque (por exemplo, Epi) que continham (a) ou não continham (b) o motivo analisado, e o número de regiões 3'-UTR do controle que continham (c) ou não continham (d) o motivo analisado. Foram

realizadas três comparações, com controles distintos (os conjuntos supra-citados: todas as regiões 3'-UTR, as regiões 3'-UTR da categoria *housekeeping* e as regiões 3'-UTR do controle inverso). Aplicou-se à matriz o teste exato de Fisher para avaliar se o motivo identificado na categoria analisada estava enriquecido significativamente.

4.5 DEFINIÇÃO DAS EXTREMIDADES DOS mRNAs

Para a definição das extremidades dos mRNAs, isto é, os limites das UTRs, utilizamos como dado inicial o conjunto total de leituras que foram geradas nas avaliações sistemáticas do transcriptoma de *T. cruzi* desenvolvidas no Instituto Carlos Chagas nos últimos dois anos. Nesse sentido, foram utilizadas 2.690.467.573 leituras oriundas do seqüenciador SOLiD (2,7 bilhões) que representam uma cobertura não-enviesada do transcriptoma desse organismo.

Para a identificação do limite 5' do mRNA a seqüência representativa do mini-éxon de *T. cruzi* (AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG) foi utilizada como isca em sua representação em espaço de cores (*color space*). A busca foi feita de forma seqüencial, do tipo força bruta, iniciando-se a comparação com os últimos cinco nucleotídeos do mini-éxon comparados aos 5 primeiros nucleotídeos da leitura e registrando-se a proporção de pareamentos (*matches*) entre as duas sub-sequencias. Posteriormente, essa janela de comparação foi aumentada gradativamente (por exemplo, seis últimos nucleotídeos do mini-éxon comparados com os seis primeiros nucleotídeos da leitura) até o tamanho máximo do mini-éxon (39 nucleotídeos). Posteriormente, essa janela contendo a sequencia completa do mini-éxon foi deslocada pelo restante da leitura até chegar ao seu final (posições 22 a 50 da leitura), e o processamento inverso ao início foi realizado, isto é, os primeiros 38 nucleotídeos do mini-éxon foram comparados com os últimos 38 nucleotídeos da leitura e assim, sucessivamente, até que os cinco primeiros nucleotídeos do mini-éxon foram comparados com os últimos cinco nucleotídeos da leitura. A comparação cujo nucleotídeo inicial estava mais à esquerda e que a proporção de pareamento foi a maior foi selecionada como o sítio mais provável de localização do mini-éxon na seqüência, e o grau de similaridade foi registrado para análises posteriores.

Após a identificação de leituras que potencialmente poderiam conter o mini-

éxon, a sub-sequencia da qual foi retirada a região de similaridade com o mini-éxon foi mapeada contra o genoma de *T. cruzi* utilizando o software SHRiMP (RUMBLE *et al.,* 2009), com os seguintes parâmetros: -o 500 –s w10 –n 1 –v 50% -h 50% --strata. Esses parâmetros servem para aumentar a probabilidade de se realizar o pareamento mesmo com seqüências curtas.

As leituras foram consideradas como possivelmente contendo um mini-éxon quando:

- O grau de similaridade do pareamento foi maior do que 70%;
- A posição inicial de pareamento do mini-éxon na leitura foi o seu primeiro nucleotídeo, isto é, leituras nos quais o mini-éxon pareou internamente ou seja havia nucleotídeos pré-mini-éxon foram considerados como uma estimativa de identificação espúria, ou falso-positivos, e eliminados das análises posteriores.
- Após a realização do mapeamento, caso a região sugestiva de conter o mini-éxon tivesse também similaridade com a região genömica, essa leitura foi considerada como falso-positivo e eliminada das análises futuras. Para seqüências pequenas, esssa etapa é problemática e o critério utilizado foi que o grau de pareamento com a seqüência de mini-éxon deveria ser 30% maior do que o grau de pareamento com o genoma. Por exemplo, caso uma sequencia de cinco nucleotídeos tivesse um pareamento perfeito com o mini-éxon (100% de similaridade), ela teria que ter pelo menos 2 não-pareamentos com o genoma (60% de similaridade), pois com um pareamento a diferença seria de somente 20%.

Para a identificação da extremidade 3’ do mRNA uma abordagem similar foi utilizada, restringindo-se no entanto o mapeamento somente à porção direita da leitura. Nesse caso, identificou-inicialmente as leituras cuja codificação em espaço de cores dos últimos cinco nucleotídeos seqüenciado fosse sugestiva de cauda poli-A (codificada como 00000). Essa codificação é somente sugestiva pois também pode representar os outros 3 homopolímeros, isto é CCCCC, GGGGG ou TTTTT. Para as leituras nas quais essa semente teve um pareamento com zero ou um mal-pareamento, ou seja, até 80% de similaridade mínima, procedeu-se à extensão da região de pareamento, utilizando-se a seguinte regra para encerrar o pareamento: duas bases adjacentes eram codificadas por um caracter distinto da semente (ou

seja, codificada por 1, 2 ou 3). Após a finalização dessa etapa, o número de mal-pareamentos com a cauda poli-A (000...000) e a posição mais à esquerda da região da cauda poli-A foram armazenados para posterior processamento.

As leituras foram consideradas como possivelmente contendo uma cauda poli-A quando:

- O tamanho da sub-sequencia remanescente após a retirada da região putativa de conter a cauda poli-A fosse maior do que 12 nucleotídeos.
- O grau de similaridade da região putativa de conter a cauda poli-A fosse superior a 70%.
- Após a realização do mapeamento, caso a região sugestiva de conter a cauda poli-A tivesse também similaridade com a região genömica, essa leitura foi considerada como falso-positivo e eliminada das análises futuras. O critério utilizado foi idêntico ao do mini-éxon, isto é, o grau de similaridade com a sequência da cauda poli-A teria que ser 30% superior ao grau de similaridade com a sequência genômica.

As leituras que passaram nos critérios acima foram utilizadas para as análises subseqüentes. As regiões de inserção do mini-éxon e cauda poli-A foram identificadas utilizando-se como critérios a definição clara de picos de mapeamento no genoma de *T. cruzi.* O software Integrative Genome Browser (IGV) foi utilizado para a visualização dos mapeamentos.

Para a identificação em larga escala dos sítios de adição de mini-éxon ou cauda poli-A, o mapeamento no genoma foi transformado em contagem por posição do genoma. Uma janela deslizante de 10 nucleotídeos foi aplicada a esses dados, selecionando-se as regiões que apresentaram uma contagem superior a 10 leituras como sendo fortes candidatas a sítios de inserção.

5 RESULTADOS

5.1 BANCO DE DADOS

Foi implementado um banco de dados local, chamado KinetoDB, para armazenamento das informações genômicas de nove espécies de tripanossomatídeos com genoma disponibilizado até o momento: *T. brucei, T. congolense, T. cruzi, T. vivax, L. major, L. braziliensis, L. infantum, L. mexicana* e *L. tarentolae*. Por se tratar de um conjunto muito grande de informações e que possuem relações complexas entre si, utilizamos o “schema” Chado. Chado é um esquema para representação complexa de dados biológicos, que foi desenvolvido inicialmente pelo Flybase e constantemente melhorado pela comunidade científica em geral. Vários outros bancos de dados importantes também utilizam o Chado ou ferramentas do projeto GMOD.

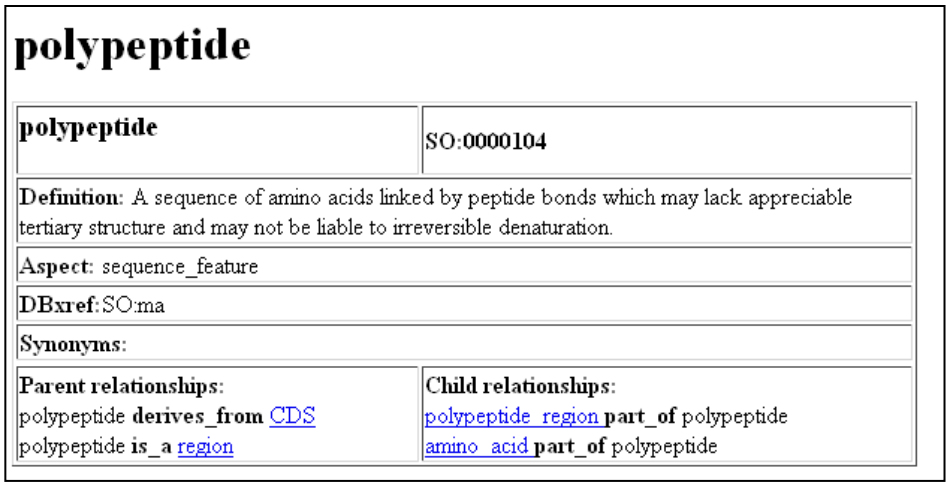
A informação dos nove genomas obtidas dos arquivos em formato FASTA e GFF do TriTrypDB foram utilizadas para serem formatadas e inseridas no Chado. Para que os dados pudessem ser inseridos no banco foi necessária uma série de edições nas informações dos genomas e adição de novos termos de ontologia no Chado.

Por exemplo, o nível máximo de organização genômica para inserção no banco deve ser o mesmo para todas as espécies, para poderem ser representadas ao mesmo tempo no GBrowse. Como *T. cruzi* estava com seu genoma montado em supercontigs e as demais espécies em cromossomos, foi necessário chamar os contigs de *T. cruzi* de cromossomos para não criar incompatibilidades na visualização dos dados.

O Chado é um banco estruturado pela ontologia de modo que termos presentes na informação dos genomas tiveram que ser mudados ou inseridos nas tabelas correspondentes para que os dados pudessem ser inseridos no banco. A FIGURA 6 mostra a relação entre termos parentais e termos filhos da característica (*feature*) polipeptídeo, segundo o *Sequence Ontology.* E na FIGURA 7 podemos observar a relação entre diversas características diferentes.

A principal tabela do Chado é a tabela chamada ***feature***, onde estão todas

as entidades presentes em um genoma.



polypeptide

polypeptide	SO:0000104
Definition: A sequence of amino acids linked by peptide bonds which may lack appreciable tertiary structure and may not be liable to irreversible denaturation.	
Aspect: sequence_feature	
DBxref: SO:ma	
Synonyms:	
Parent relationships: polypeptide **derives_from** CDS polypeptide **is_a** region	**Child relationships**: polypeptide_region **part_of** polypeptide amino_acid **part_of** polypeptide

FIGURA 6 – Termos parentais e termos filhos da feature polipeptídeo

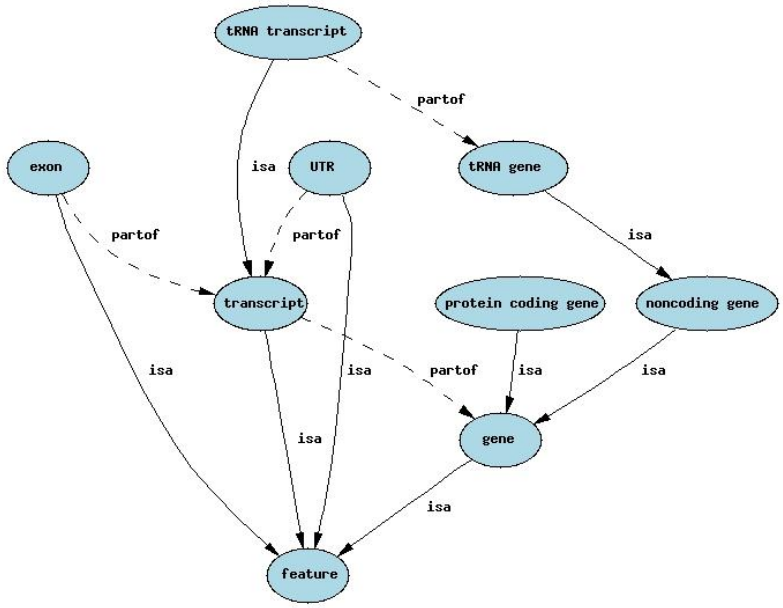


FIGURA 7 – Termos parentais e termos filhos de diferentes features ligados pelo termo do vocabulário controlado.

Na TABELA 1 podemos observar o número de características das espécies mais estudadas e cujo genoma está melhor anotado, dentre as que foram inseridas no banco de dados. Observem o grande número de pseudogenes em *T. cruzi*. Algumas características não foram anotadas em alguns desses genomas, como por exemplo, tRNA em *L. braziliensis* e *L. infantum*; unidades de repetição em geral**,** etc. As diferenças no grau de anotação dos diferentes genomas é um grande desafio

bioinformático, sendo que essas diferenças representam um problema na correta utilização subsequente dos dados.

TABELA 1 – Características utilizadas para inserção no banco de dados.

Característica	*T. brucei*	*T. cruzi*	*L. major*	*L. infantum*	*L. braziliensis*
CDS	8.712	19.607	8.265	7.992	7.896
chromosome	12	32.746	36	36	35
exon	9.284	21.604	9.175	8.083	8.002
Gap	30	0	1	427	975
gene	9.276	21.601	9.112	7.992	7.896
mRNA	8.712	19.607	8.265	7.992	7.896
ncRNA	359	1.466	700	5	2
processed_transcript	29	194	60	3	4
promoter	2	0	0	0	0
pseudogene	916	3.609	43	193	231
pseudogenic_region	46	270	153	1	231
pseudogenic_transcript	0	0	1	0	0
region	6.577	35.688	15.448	0	16
repeat_region	7.830	34.465	11.970	0	152
repeat_unit	219	0	3	3	
residues	26.075.494	89.612.356	32.816.778	32.153.582	31.379.198
residues (tr)	4.394.689	9.755.524	5.216.227	4.941.602	4.910.389
rRNA	117	219	63	6	0
sequence_variant	959	0	0	0	0
STS	37	0	46	0	0
three_prime_UTR	1	0	0	0	0
tRNA	65	115	83	0	0

Implementamos o programa de visualização genômica GBrowse, integrado ao Chado. Utilizando esse programa podemos visualização o contexto genômico de diferentes características, como genes, transcritos, proteínas, etc. Essa visualização é possível para as nove espécies, o que permite a comparação entre os diferentes contextos genômicos.

Na FIGURA 8 pode ser observado o contig AAHK01000001 de *T. cruzi* e algumas de suas características. Ao clicar sobre a característica, somos direcionados a outra página com informações detalhadas, como pode ser visto na FIGURA 9.

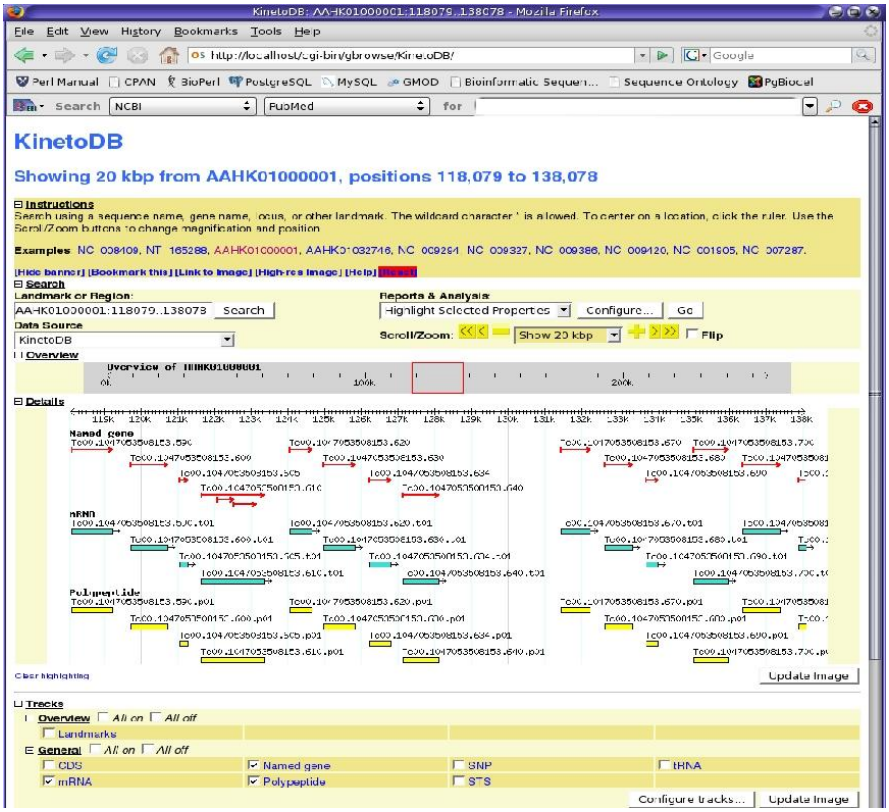


FIGURA 8 – Visualização de um contig do genoma de *T. cruzi* no programa Gbrowse.

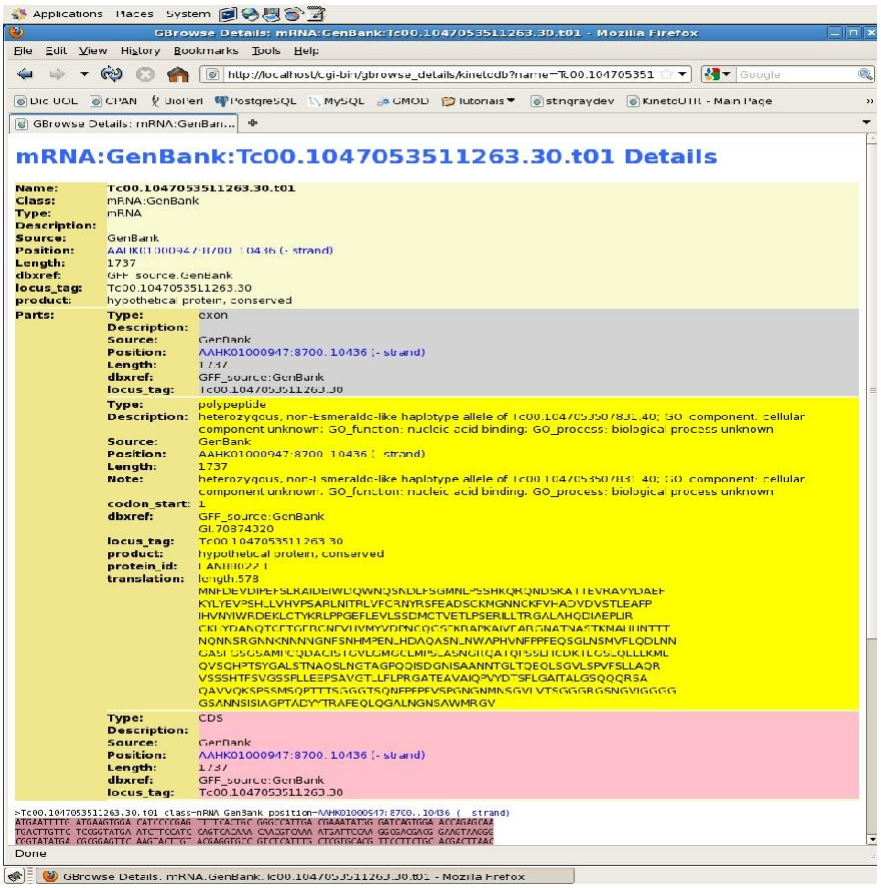


FIGURA 9 – Visualização das informações de um gene de *T. cruzi* no programa Gbrowse.

Para permitir o gerenciamento do Projeto ORFeoma de *T. cruzi* do Instituto Carlos Chagas, tivemos que criar e inserir três tabelas no “schema” Chado, conforme mostrado na FIGURA 10. Uma das tabelas contém os primers utilizados no projeto ORFeoma (tabela primer), que está relacionada a um resultado de BlastN contra os genes de *T. cruzi* (tabela feature_primer), que por sua vez está conectada à tabela feature, a tabela central do Chado. A tabela restante conecta cada gene de *T. cruzi* com os clones reais do ORFeoma, isto é, a amplificação do material e sua inserção na biblioteca.

Para podermos fazer análises comparativas entre os nove genomas, também incorporamos a informação dos genes ortólogos entre as espécies. Para comportar essa informação criamos e associamos ao Chado mais uma tabela, associando o número do grupo ortólogo a que pertence cada gene na tabela feature, de acordo com as classificação do OrthMCL Database (tabela feature_og).

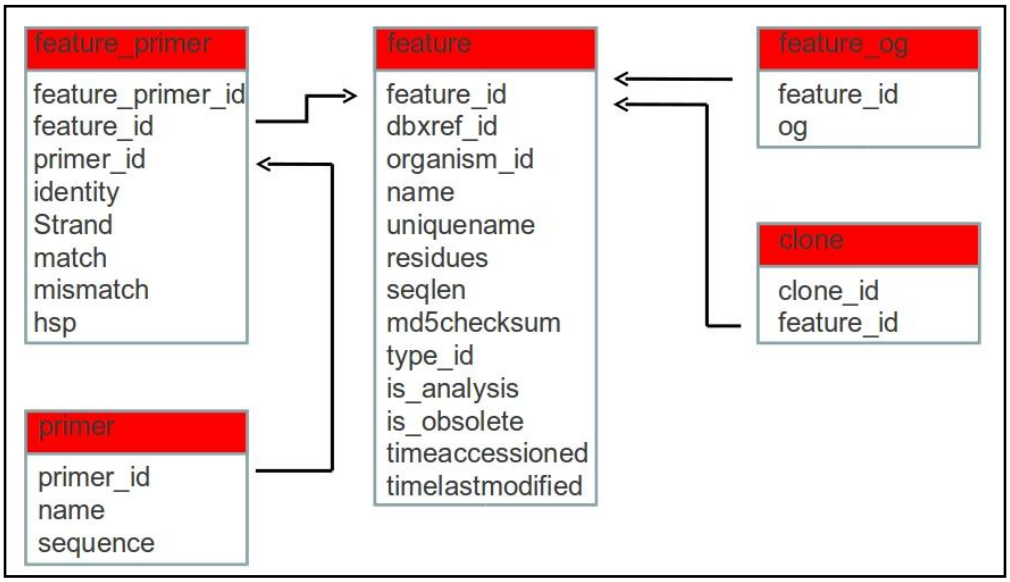


FIGURA 10 – Tabelas inseridas no Chado para comportar o Projeto ORFema de *T. cruzi*.

5.2 ANÁLISE DE GENES ADJACENTES DIFERENCIALMENTE EXPRESSOS NAS REDES METABÓLICAS DE *Trypanosoma cruzi*

As vias metabólicas do KEGG são representadas por diagramas, como por exemplo a via de síntese e degradação de corpos cetônicos, que é a menor via metabólica representada no KEGG (FIGURA 11). Essa figura é interpretada e gerada

a partir de um arquivo XML, que contém todas as informações, inclusive de como desenhar a figura.

Os 66 arquivos em XML contendo as informações das conexões por compostos metabólicos dos genes de *T. cruzi* foram parseados. Esses dados foram inseridos num banco de dados desenvolvido no Laboratório de Biologia Computacional e Bioinformática do Instituto Carlos Chagas.

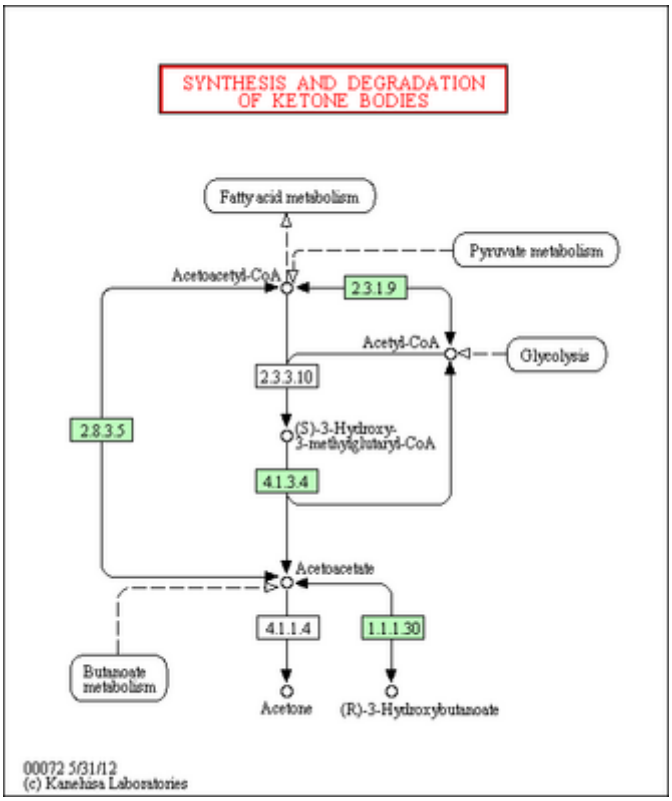


FIGURA 11 – Diagrama representando a via de síntese e degradação de corpos cetônicos.

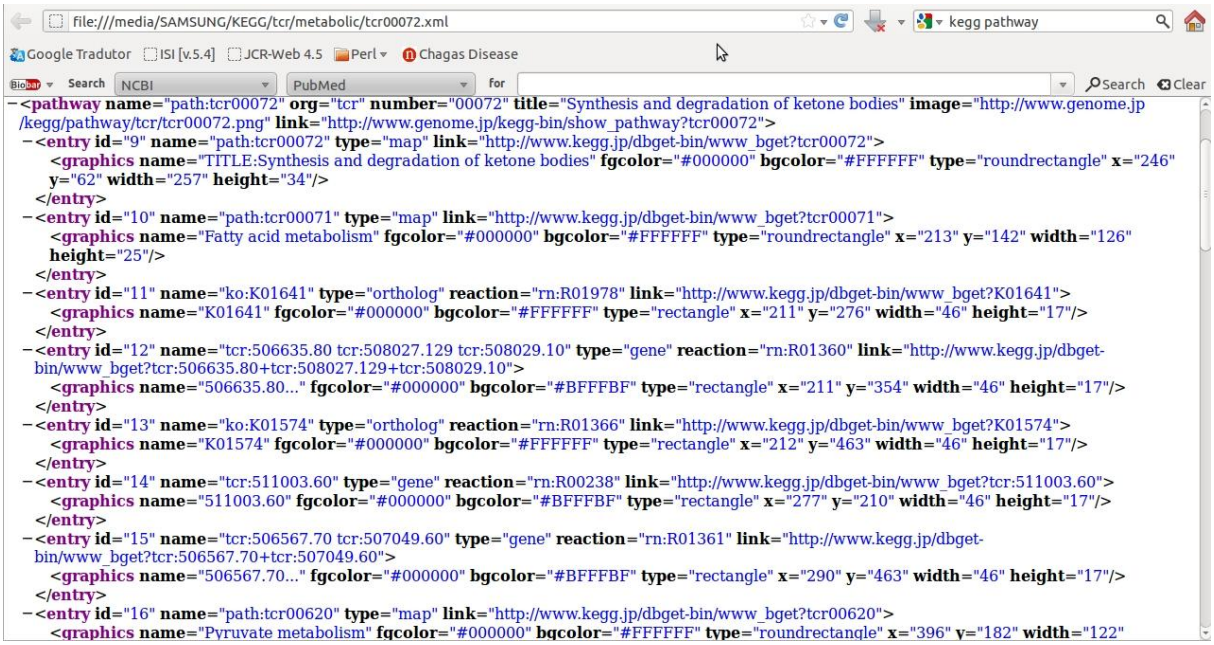


FIGURA 12 – Arquivo XML que codifica a informação para a representação gráfica da via de síntese e degradação de corpos cetônicos mostrada na FIGURA 11.

Com a informação completa dos genes de *T. cruzi* disponíveis no KEGG foi gerada a rede metabólica de *T. cruzi* representando a relação entre os genes (FIGURA 18). A rede é formada por 326 nós (genes) e 1.520 arestas (relação de ligação entre os genes, ou compostos). Como também poder ser observado a rede é bastante conectada. Na TABELA 2 está a lista de compostos e o número de genes com que estes interagem.

Avaliamos a rede do KEGG sobre duas abordagens distintas: a primeira considera que as arestas (reações) entre os nós (genes) não apresenta direcionamento; embora uma rede metabólica não apresenta essa característica, é uma forma usual de se trabalhar com grafos; a segunda, denominada direcionada, considera o sentido da reação enzimática, isto é, se ela é reversível (o gene pode catalisar ou metabolizar o produto) ou irreversível (o fluxo metabólico segue somente uma direção).

Na TABELA 3, vemos os parâmetros relacionados à rede geral (rede 1), nas situações direcionadas e não-direcionada.

O coeficiente de clusterização representa o grau com os quais os nós do grafo tendem a se clusterizar e é calculado pela proporção de triângulos que são conectados totalmente entre si em relação a todos os triângulos da rede. De maneira geral, o grau de clusterização obtido indica que a rede não é aleatória, pois nesse

caso esperaríamos que o coeficiente de clusterização fosse menor. Além disso, a rede não-direcionada apresenta um coeficiente de clusterização ainda maior, coerente com a maior conectividade da mesma.

TABELA 2 – Lista de compostos com mais de 3 conexões.

ID	Nome	Conexões
C00024	Acetyl-CoA	13
C05345	beta-D-Fructose 6-phosphate	9
C00014	NH3	9
C00025	L-Glutamate	8
C01194	1-Phosphatidyl-D-myo-inositol	8
C00668	alpha-D-Glucose 6-phosphate	7
C00096	GDP-mannose	7
C00037	Glycine	7
C00118	D-Glyceraldehyde 3-phosphate	7
C00097	L-Cysteine	6
G00143	(GlcNAc)1 (Ino-P)1	6
C00101	Tetrahydrofolate	6
C00022	Pyruvate	6
C00049	L-Aspartate	6
C00117	D-Ribose 5-phosphate	6
C00020	AMP	6
C00068	Thiamin diphosphate	6
C00332	Acetoacetyl-CoA	6
C00065	L-Serine	6
C15972	Enzyme N6-(lipoyl)lysine	5
C00029	UDP-glucose	5
C00111	Glycerone phosphate	5
C00006	NADP+	5
G10610	UDP-N-acetyl-D-glucosamine	5
C00074	Phosphoenolpyruvate	5
C00267	alpha-D-Glucose	5
C01172	beta-D-Glucose 6-phosphate	5
C15973	Enzyme N6-(dihydrolipoyl)lysine	5
C00036	Oxaloacetate	5
C00008	ADP	5
C02090	Trypanothione	5
C00033	Acetate	5
C00206	dADP	4
C00365	dUMP	4
C00002	ATP	4
C06250	Holo-[carboxylase]	4
C00275	D-Mannose 6-phosphate	4
C00199	D-Ribulose 5-phosphate	4
C00144	GMP	4
C00143	5,10-Methylenetetrahydrofolate	4
C00044	GTP	4
C00242	Guanine	4
C00035	GDP	4
C00877	Crotonoyl-CoA	4
C00361	dGDP	4
C00352	D-Glucosamine 6-phosphate	4
C00350	Phosphatidylethanolamine	4
C00051	Glutathione	4
C00154	Palmitoyl-CoA	4
C00641	1,2-Diacyl-sn-glycerol	4
C01222	GDP-4-dehydro-6-deoxy-D-mannose	4
C00026	2-Oxoglutarate	4
C00110	Dolichyl phosphate	4
C00005	NADPH	4
C00082	L-Tyrosine	4
C00042	Succinate	4
C01144	(S)-3-Hydroxybutanoyl-CoA	4
C00315	Spermidine	4
C00073	L-Methionine	4
C03069	3-Methylcrotonoyl-CoA	4
C00130	IMP	4
C00064	L-Glutamine	4
C00197	3-Phospho-D-glycerate	4
C00258	D-Glycerate	4

O tamanho médio de caminho entre dois nós é alto, pois a rede é grande. Essa propriedade indica quantos passos, em média, é necessário realizar para ir de um nó a outro da rede; isto é, quantas reações enzimáticas, em média, separam um gene do outro da rede. Finalmente, o número médio de vizinhos representa quantos genes estão conectados a um determinado gene específico. Esse valor é relativamente alto na rede metabólica de *T. cruzi.*

TABELA 3 – Valores dos parâmetros das redes 1 e 2, não direcional e direcional.

	Rede 1		Rede 2	
Parâmetro	**Não Direcional**	**Direcional**	**Não Direcional**	**Direcional**
Coeficiente de clusterização	0,396	0,307	0,393	0,303
Componentes conectados	15	15	17	17
Diâmetro da rede	17	17	17	20
Centralização da rede	0,126	-	0,127	-
Caminhos mais curtos	78%	50%	74%	48%
Tamanho de caminho caracteristico	5,295	5,232	5,584	6,002
Número médio de vizinhos	6,356	6,356	5,951	5,951

Da FIGURA 13 a FIGURA 15, vemos a distribuição do grau de conexões que cada gene apresenta. De maneira geral, as redes, tanto direcionada quanto não direcionada, apresentam uma mesma distribuição.

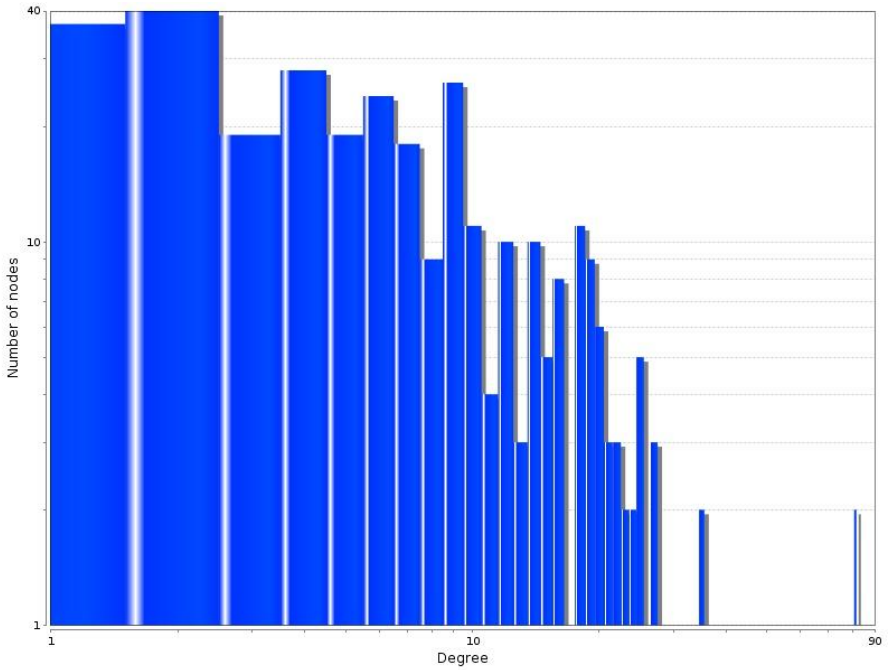


FIGURA 13 – Grau de distribuição dos nós da rede não direcional.

Uma rede direcionada apresenta dois tipos de conexões em relação a um determinado nó, de entrada e de saída. Consequentemente, esse tipo de rede apresenta dois tipos de graus de conexão: de entrada e de saída (*in-degree* e *out-degree*). Podemos observar que em relação aos genes que apresentam uma conexão, essa é mais frequente como saída (OUT, n=48) do que entrada (IN, n=30); já para os genes que apresentam duas conexões, tanto a entrada (n=73) quanto a saída (n=60) são bem frequentes.

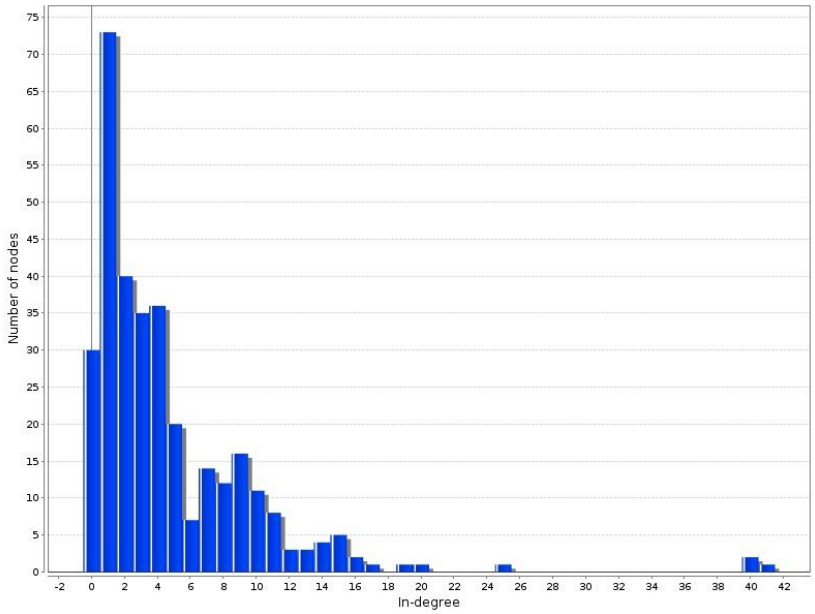


FIGURA 14 – Grau de distribuição dos nós da rede direcional IN.

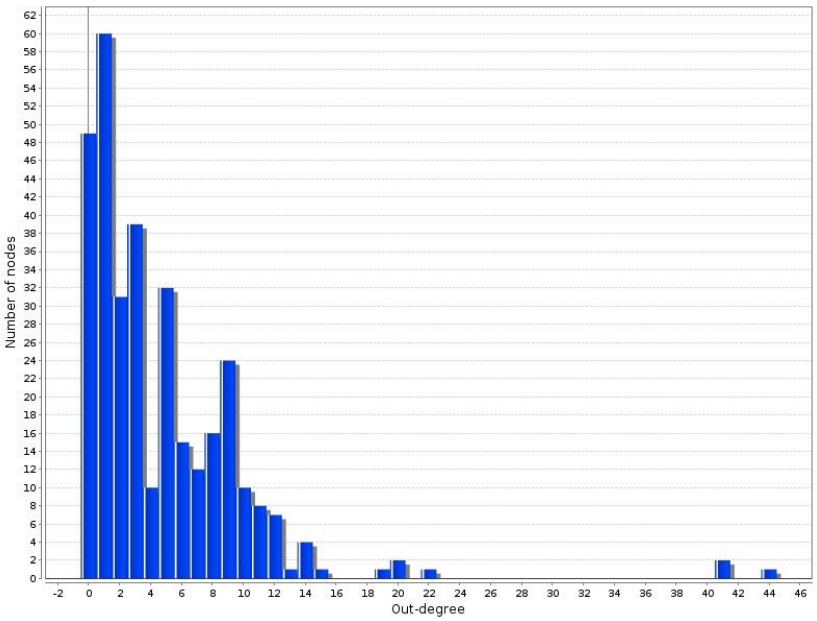


FIGURA 15 – Grau de distribuição dos nós da rede direcional OUT.

Como a rede é muito conectada, removemos dois dos compostos que mais possuem conexões, acetil-CoA (C00024) com 13 arestas e amônia (C00014) com 9 arestas, para avaliar se seria possível, com essa exclusão, seria possível modularizar melhor a rede, isto é, observar um menor grau de clusterização global. Esses dois compostos são muito genéricos, envolvidos em muitas reações e, ao afetarem fortemente a conectividade da rede, sem agregarem especificidade funcional, poderiam estar limitando nossas análises. A nova rede (rede 2) formada com a remoção desses dois compostos possui 325 nós (genes) e 1.422 arestas (relação de adjacência entre os genes), demonstrando o alto grau de conectividade da rede.

Ao retirar esses dois compostos, conseguimos diminuir a conectividade da rede (TABELA 3). O tamanho médio aumentou de 5,2 para 6,0 e o número médio de vizinhos caiu de 6,4 para 6,0. No entanto, essas modificações foram de pequena monta e não foi possível aprimorar a utilização da rede para identificar módulos de co-regulação (ver abaixo).

Na FIGURA 18, temos a representação final da rede analizada no presente trabalho.

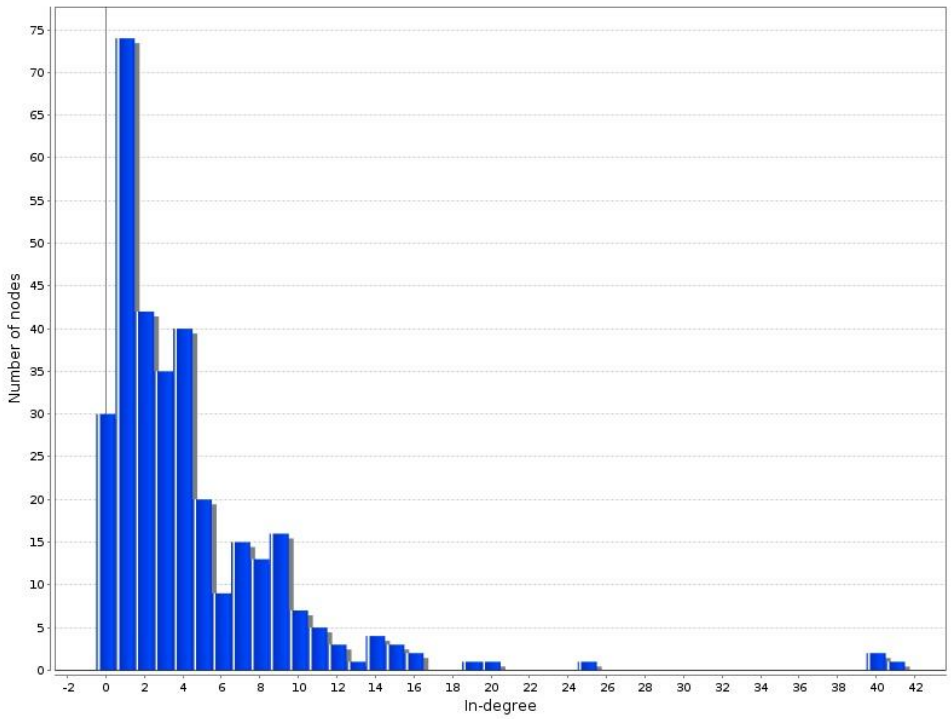


FIGURA 16 – Grau de distribuição dos nós da rede direcional (com dois compostos deletados) IN.

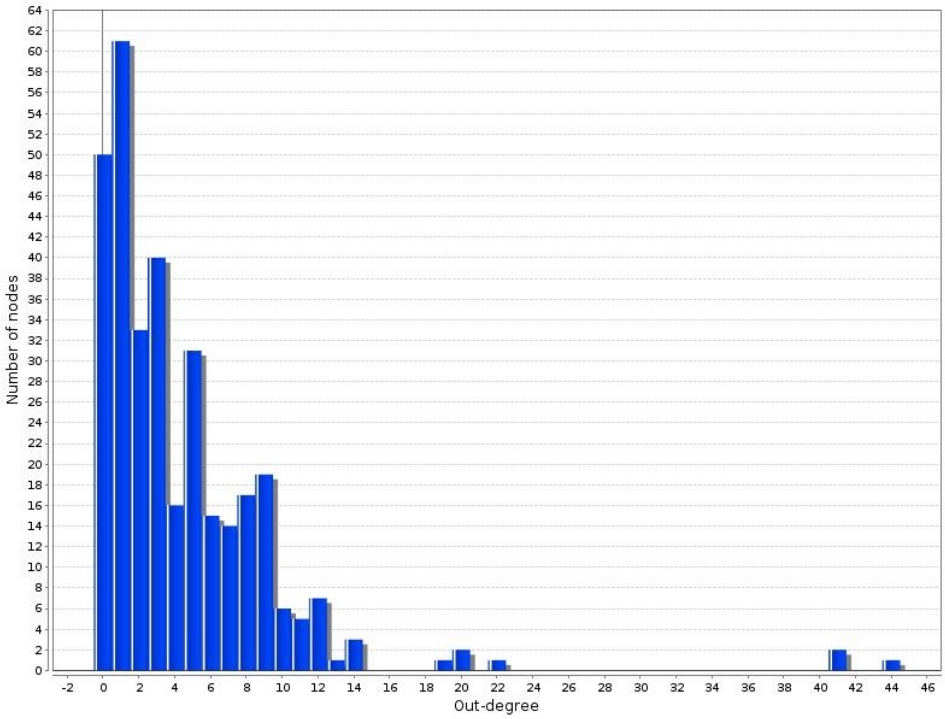


FIGURA 17 – Grau de distribuição dos nós da rede direcional (com dois compostos deletados) OUT.

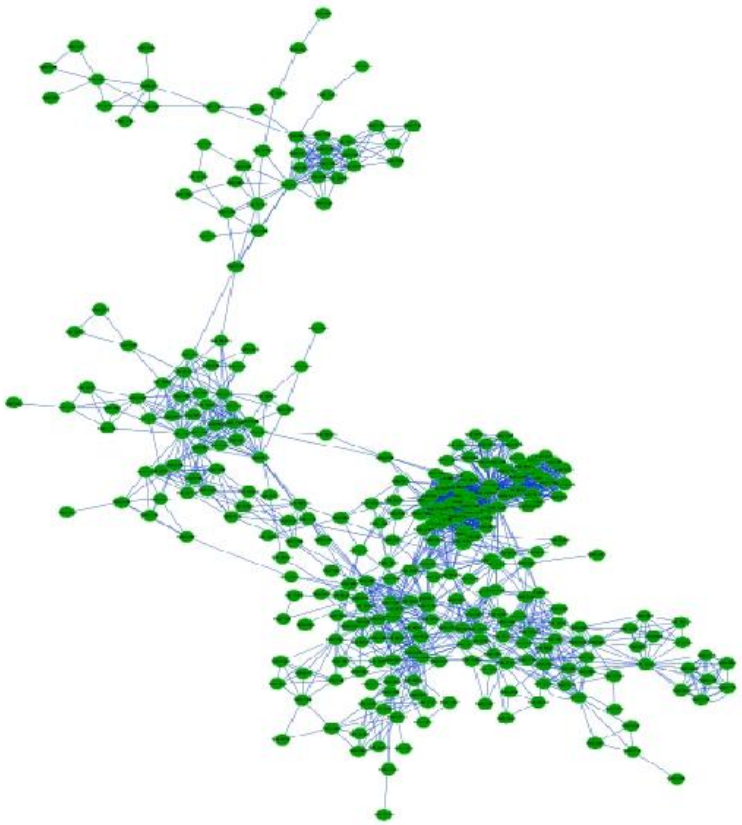


FIGURA 18 – Rede metabólica de *T. cruzi* mostrando a relação entre genes adjacentes nas vias metabólicas.

Com essa informação de “vizinhança” da rede pudemos calcular os número de genes adjacentes em uma via metabólica que são diferencialmente expressos nas 6 comparações entre as formas do ciclo de vida. A TABELA 4 mostra o número de DEGs encontrados e o tipo de associação entre os genes, para cada uma das 6 comparações.

TABELA 4 – Número de DEGs para cada comparação do ciclo de vida para cada uma das 3 situações.

Comparação	+/- ou -/+	+/+	-/-
Ama x Trp	9	1	37
Epi x Ama	48	8	123
Epi x Met	58	13	63
Epi x Trp	59	13	63
Met x Ama	17	5	19
Met x Trp	21	3	59

Para saber se esses númerso encontrados eram estatisticamente significativos, fizemos uma simulação com 10.000 réplicas, utilizando a mesma estrutura da rede, o mesmo número de DEGs da comparação em questão, apenas mudando aleatoriamente os genes na rede que eram diferencialmente expressos.

Para todas as comparações, a probabilidade de encontrar o mesmo número de DEGs nas diferentes situações foi maior do que 5%. Na FIGURA 19 podemos observar a distribuição do número pares de DEGs de acordo com as 10.000 réplicas da simulação para a comparação Epi versus Trp. A moda da simulação foi 12, enquanto o número de pares DEGs para a comparação Epi versus Trp foi 9.

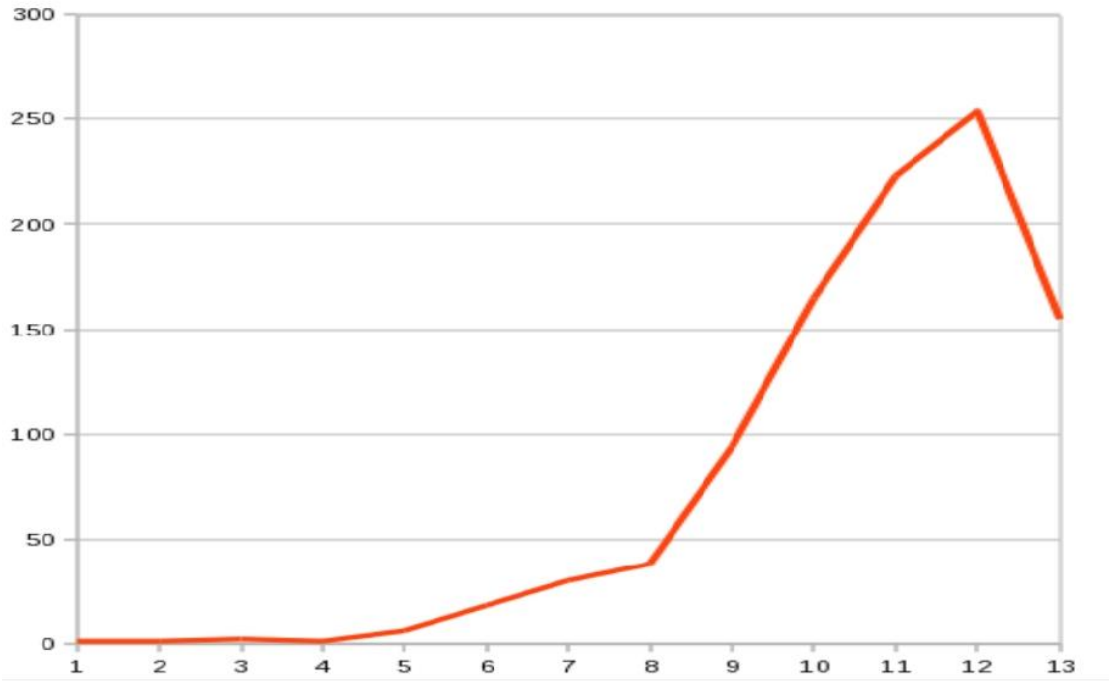


FIGURA 19 – Distribuição do número de DEGs (eixo x) de acordo com as 10.000 réplicas da simulação (eixo y).

Para tentar entender o porquê disso fomos analisar o número de conexões de cada gene na rede. A FIGURA 20 mostra o número de conexões na rede (eixo y) para cada um dos genes (eixo x). A rede metabólica não direcional de *T. cruzi* é altamente conectada, e é devido a essa alta conectividade que a probabilidade de encontrar genes vizinhos que sejam diferencialmente expressos é alta, pois os padrões de modulações tem maior poder de propagação na rede.

Isso não quer dizer que os padrões que foram identificados não sejam reais e biologicamente relevantes. Somente indica que, devido à alta conectividade das redes metabólicas em geral, a sua análise integrada tem pouco poder estatístico para se identificar padrões biológicos reais.

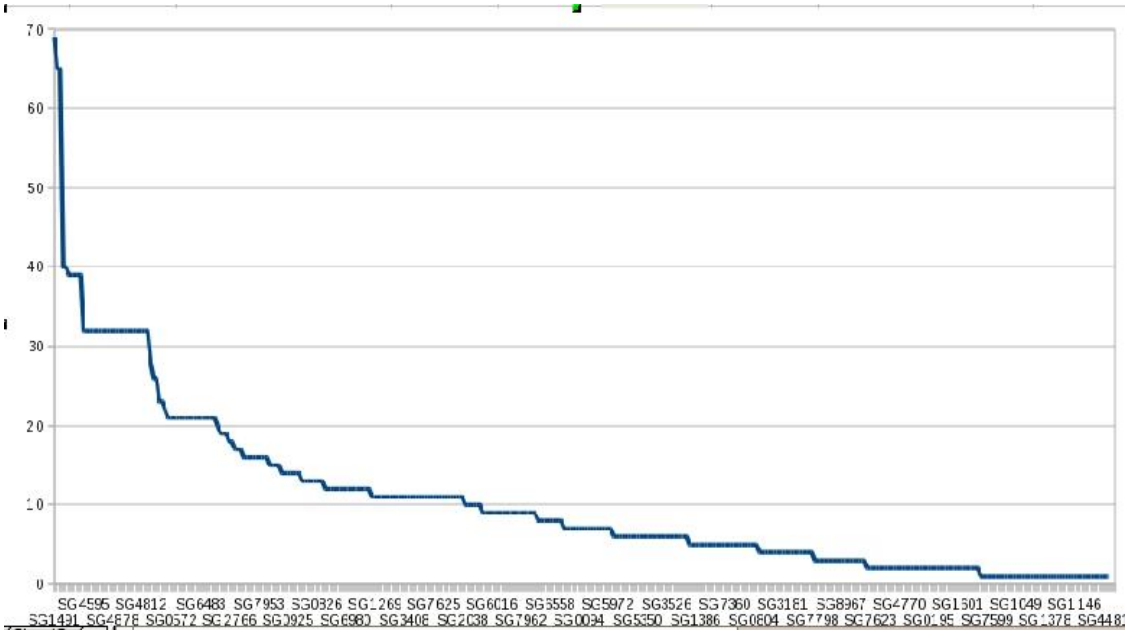


FIGURA 20 – Número de conexões na rede (eixo y) para cada um dos genes (eixo x).

5.3 GENES MARCADORES E ELEMENTOS REGULATÓRIOS

5.3.1 Genes marcadores identificados no ciclo de vida de *T. cruzi*

Com o intuito de encontrar genes que fossem marcadores de determinada fase do ciclo de vida, ou seja, genes diferencialmente expressos de uma fase do ciclo com relação às demais. Esse conceito é diferente de genes estágio-específicos, que são expressos somente em uma fase e não nas demais.

Foram feitas todas as combinações possíveis entre as quatro formas do parasita: formas individuais, e combinações de duas ou três. Foram, portanto, estabelecidas 15 categorias diferentes de genes marcadores, que ao longo do texto são referidas com as seguintes abreviações: Epi, Met, Ama, Trp, EpiMeta, EpiAma, EpiTrp, MetaAma, MetTrp, AmaTrp, EpiMetAma, EpiMetTrp, EpiAmaTrp, MetAmaTrp, além da categoria *housekeeping*, com os genes que não modularam em nenhuma das quatro fases do ciclo.

Na TABELA 5 estão os genes marcadores (supergenes diferencialmente expressos) de cada categoria, utlizando o critério de razão de mudança de 1,5 vezes, conforme descrito em material e métodos. Como nem todos os genes identificados possuíam 3’UTR que passaram nos critérios de seleção, o número de

seqüências utilizadas para a busca de motivos é menor, mas a proporção de perda é muito pequena (TABELA 5).

Da FIGURA 21 a FIGURA 26 é mostrada a clusterização hierárquica de cada categoria, onde as colunas representam as cinco réplicas biológicas de cada uma das formas do ciclo de vida, indicadas na parte superior da figura, e as linhas correspondem a cada um dos genes marcadores da respectiva categoria. Estas figuras são para inspeção visual das categorias e também mostram as diferenças entre as réplicas.

A lista com o identificador e a anotação dos supergenes que compõem cada uma das categorias encontra-se no apêndice.

TABELA 5 – Número de genes marcadores (supergenes diferencialmente expressos) em cada categoria.

Categoria	supergenes	supergenes com 3'UTR	Proporção com 3'UTR	Proporção do total (n=2728)
Ama	63	61	96,8%	2,3%
AmaTrp	179	171	95,5%	6,6%
Epi	272	252	92.6%	10,0%
EpiAma	57	55	96,5%	2,1%
EpiAmaTrp	144	141	97,9%	5,3%
EpiMet	175	171	97,7%	6,4%
EpiMetAma	70	64	91,4%	2,6%
EpiMetTrp	46	46	100,0%	1,7%
EpiTrp	1	1	100,0%	0,03%
Met	240	233	97,1%	8,8%
MetAma	4	4	100,0%	0,1%
MetAmaTrp	334	324	97,0%	12,2%
MetTrp	33	30	90,9%	1,2%
Housekeeping	547	525	96,0%	20,1%
Trp	563	528	93,8%	20,6%

Dos 63 genes marcadores encontrados para categoria Ama, 35 são proteínas hipotéticas ou hipotéticas conservadas (55,6%). Como esperado, o gene de amastina foi identificado nessa categoria. Nessa categoria, foi encontrado enriquecimento para proteínas com o domínio ABC_tran (PF0005), um domínio de ligação de ATP em transportadores ABC. (valor p= $8{,}8 \times 10^{-5}$).

Dos 179 genes encontrados em AmaTrp, foram encontrados um grande

número de genes de superfície, como trans-sialidades, mucinas, GP63, MASPs. Mais de um terço das proteínas encontradas foram hipotéticas ou hipotéticas conservadas. Como esperado pelo grande número de trans-sialidades encontradas, o domínio Tr-sialidase_C (PF11052) esteve enriquecido nessa categoria (valor p= $2,5x10^{-21}$), assim como o domínio mucin (PF01456, valor p= $7,5x10^{-8}$) dos genes de mucina. Também é possível observar que, em geral, o padrão de expressão é mais forte em Trp do que em Ama para a grande maioria dos genes marcadores dessa categoria.

A categoria Epi teve 272 genes marcadores identificados. Uma série de proteínas ribossomais, quinases, e novamente mais de 50% de proteínas hipotéticas ou hipotéticas conservadas. Um domínio de amino-transferase, Aminotran_1_2 (PF00155), está enriquecido nessa categoria (valor p= $6,4x10^{-5}$).

A categoria EpiAma corresponde às duas formas replicativas do parasita e teve 57 genes marcadores. Foram encontrados genes com os domínios de DNA polimerase, DNA_pol_A (PF00476), e PfkB (PF00294), uma família de quinases de carboidratos, com valores de significância de $4,5x10^{-4}$ e $9,3x10^{-4}$, respectivamente. Embora o número de genes identificados como específicos dessas categorias seja pequeno, é importante ressaltar que as categorias PFAM enriquecidas são compatíveis com seu estado replicativo. A proporção de proteínas hipotéticas, conservadas ou não, foi conforme esperado, 54,4%.

O mesmo foi visto para EpiAmaTrp, onde cerca de 50% dos seus 144 genes marcadores anotados como proteínas hipotéticas ou hipotéticas conservadas. Foram encontrados domínios enriquecidos de proteína ribossomal Ribosomal_L24e (PF01246, valor p=$2,3x10^{-3}$); e dois domínios de ligação a RNAs nucleolares: Nop (PF01798) e NOSIC (PF08060), com valores de significância p= $5,6x10^{-3}$ e p= $3,8x10^{-3}$, respectivamente.

Nas formas encontradas no inseto vetor, categoria EpiMet, foram identificados 175 genes marcadores, os quais se apresentaram enriquecidos para um domínio sem função conhecida DUF1935 (PF09149, valor p= $1x10^{-6}$). Também estão presentes genes com domínio Rhodanese (PF00581, valor p= $2,1x10^{-3}$), envolvido em destoxificação; domínio Guanylate_cyc (PF00211, valor p= $8,2x10^{-3}$), domínio catalítico de adenilato e guanilato ciclase; e domínio Peptidase_C2 (PF00648, valor p= $5,4x10^{-3}$), uma calpaína peptidase.

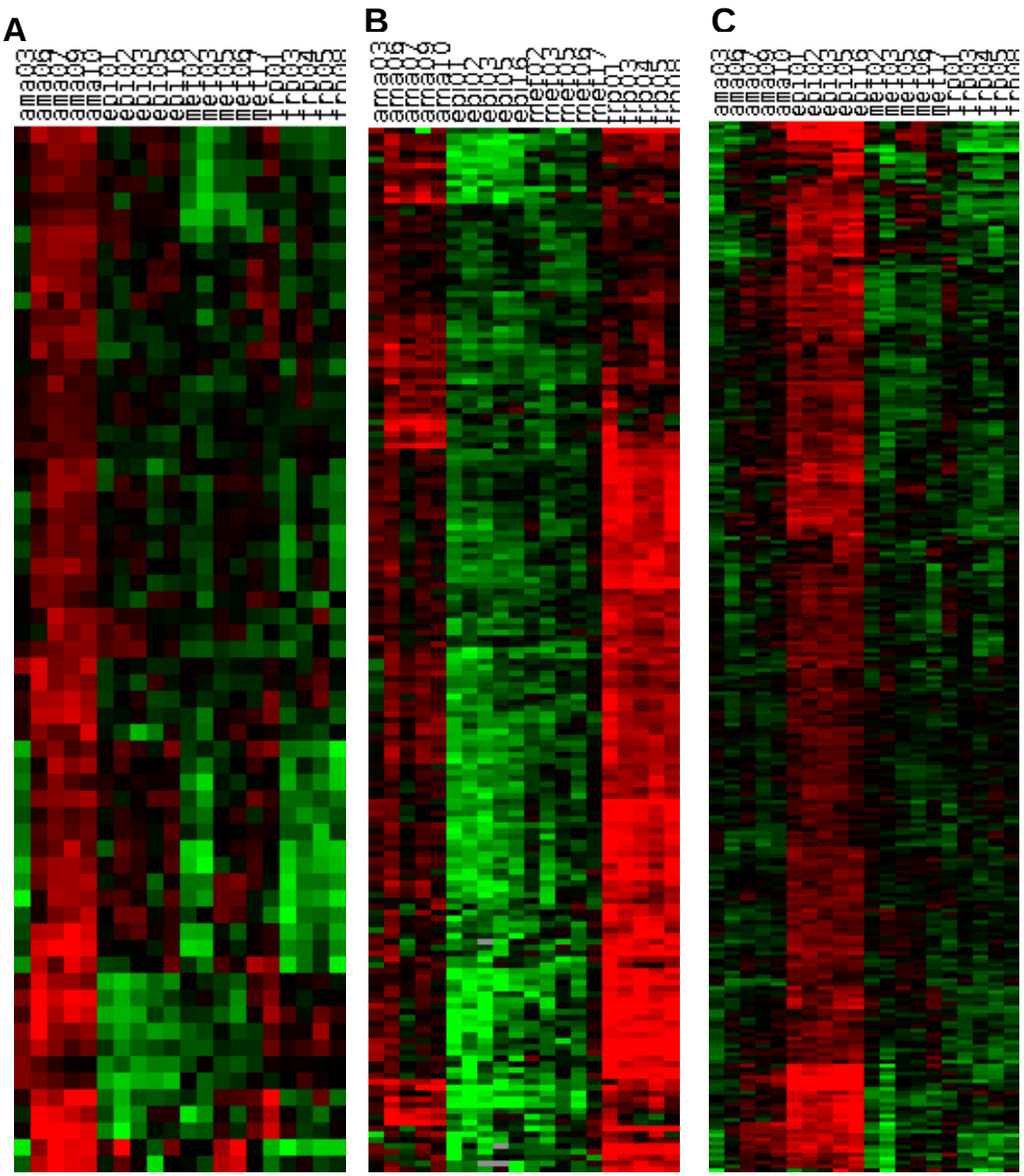


FIGURA 21 – Clusterização hierárquica dos genes marcadores em Ama (A), AmaTrp (B) e Epi (C).

Foram identificados 63 genes marcadores em Ama, 179 em AmaTrp e 272 em Epi

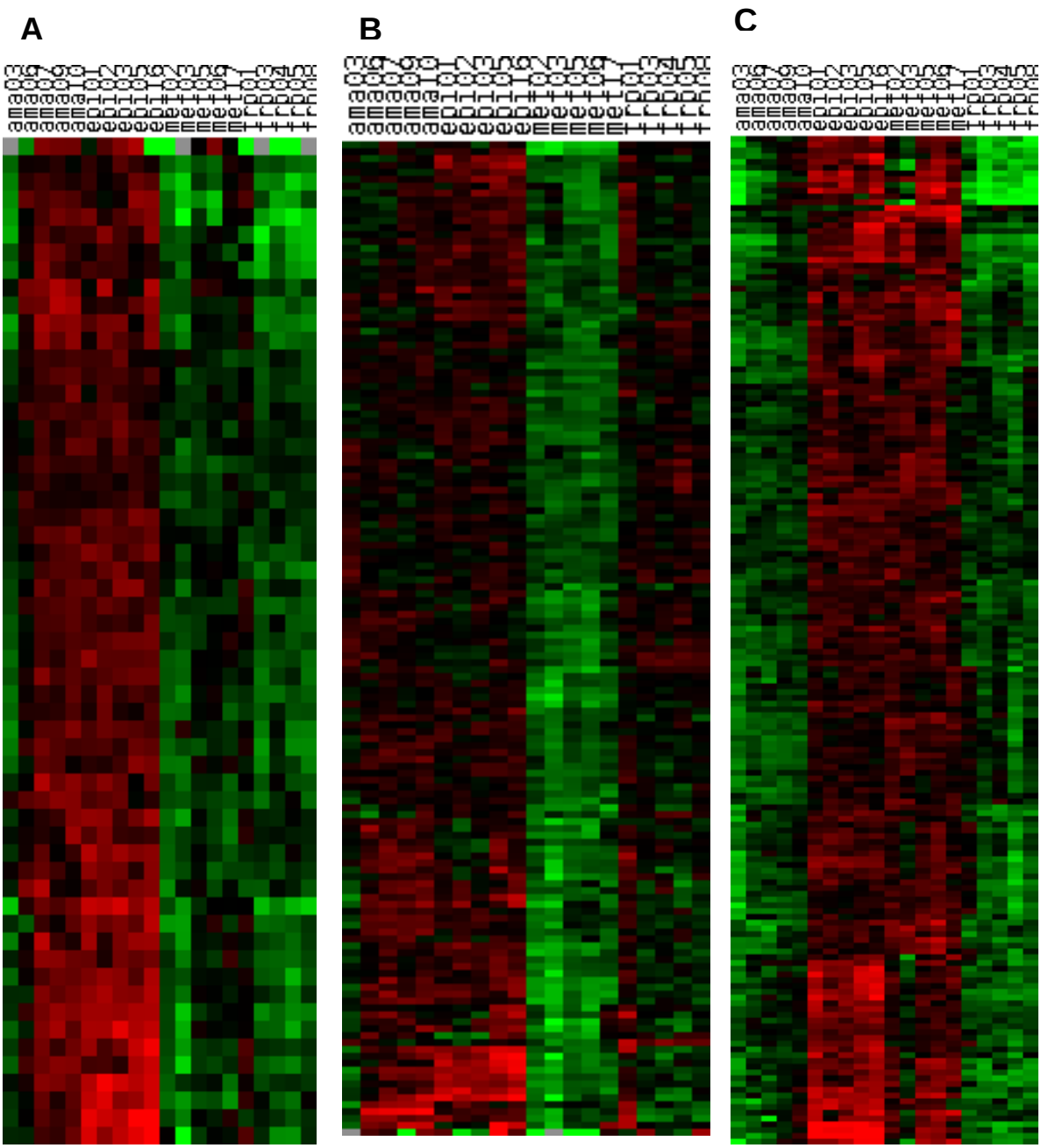


FIGURA 22 – Clusterização hierárquica dos genes marcadores em EpiAma (A), EpiAmaTrp (B) e EpiMet (C).

Foram identificados 57 genes marcadores em EpiAma, 144 em EpiAmaTrp e 175 em EpiMet.

Em EpiMetAma foram identificados 70 genes marcadores e a proporção de proteínas hipotéticas ou hipotéticas conservadas foi de acordo com o esperado, ~50%. Entre os domínios enriquecidos estão do gene da subunidade maior da calpaína Calpain_III (PF01067, valor p= $1{,}8 \times 10^{-3}$) e um domínio de ligação a oxi-redutase NAD_binding_1 (PF00175, valor p= $6{,}4 \times 10^{-3}$).

Em EpiMetTrp foram identificados 46 genes marcadores, os quais tiveram

enriquecimento para genes com domínio de repetições ricas em leucina LRR_1 (PF00560, valor p= $6,1x10^{-6}$). Interessantemente, nessa categoria encontramos uma proporção significativamente maior de proteínas hipotéticas, conservadas ou não (n=30, 65,2% do total).

Em Meta, foram identificados 240 genes candidatos, e três domínios PFAM se encontraram enriquecidos: fosfodiesterase (PDEase_I, PF00233, valor p=$1,5x10^{-3}$), super-óxido dismutase (Sod_Fe_N, PF00081, valor p=$8x10^{-3}$) e domínio de ligação a AMP cíclico (cNMP_binding, PF00027, valor p=$5x10^{-3}$). Da mesma forma que para EpiMetTrp, a proporção de proteínas hipotéticas nessa categoria é significativamente maior (67,1%).

A categoria EpiTrp apresenta somente um gene candidato, uma proteína hipotética conservada sem identificação de domínios. apenas uma proteína hipotética foi identificada. Essa proteína apresenta similaridade distante, identificada pelo uso da ferramenta PSI-BLAST, com a cadeia pesada de miosina de insetos e nematódeos. É interessante ressaltar que um padrão EpiTrp é biologicamente intrigante, pois a distância entre as formas é muito grande, em relação à sua biologia, tanto que somente foi identificado um gene marcador. Esse dado foi corroborado por experimentos mais antigos de nosso grupo utilizando microarranjo de DNA, no qual o padrão EpiTrp também foi observado, mas em menor grau, o que deve se dever primariamente à menor sensibilidade, especificidade e intervalo dinâmico do microarranjo.

Em MetAma apenas quatro genes foram identificados: uma proteína de membrana gp63, um precursor de acetilase de GPI-inositol, uma permease de aminoácidos e uma proteína hipotética. Essa proteína hipotética é só identificada em *T. cruzi* (CL Brener, Dm28c e Sylvio, os genomas que temos disponível) e em *T. vivax.* Essa última identificação já a torna uma proteína hipotética conservada. Ela tem uma similaridade baixa com uma proteína hipotética identificada em *Leishmania* sp., identificada pelo PSI-BLAST. Na análise do PSI-BLAST, essa proteína não foi identificada em outras espécies de *Trypanosoma sp.* É importante ressaltar que essa proteína de *Leishmania* tem como melhor similaridade recíproca (*Best reciprocal hit*) a proteína de *T. cruzi.* Portanto, é possível que essa proteína identificada como mais expressa em tripomastigota metacíclico e amastigota seja um ortólogo distante de uma proteína hipotética de *Leishmania,* a qual não foi encontrada em outras espécies de *Trypanosoma,* os quais não tem, geralmente, um estágio intracelular.

Geralmente, há um padrão de compartilhamento entre tripomastigota metacíclico, amastigota e tripomastigota sanguíneo; conforme veremos logo abaixo; é intrigante imaginar porque essa proteína não tem uma expressão maior em tripomastigota sanguíneo.

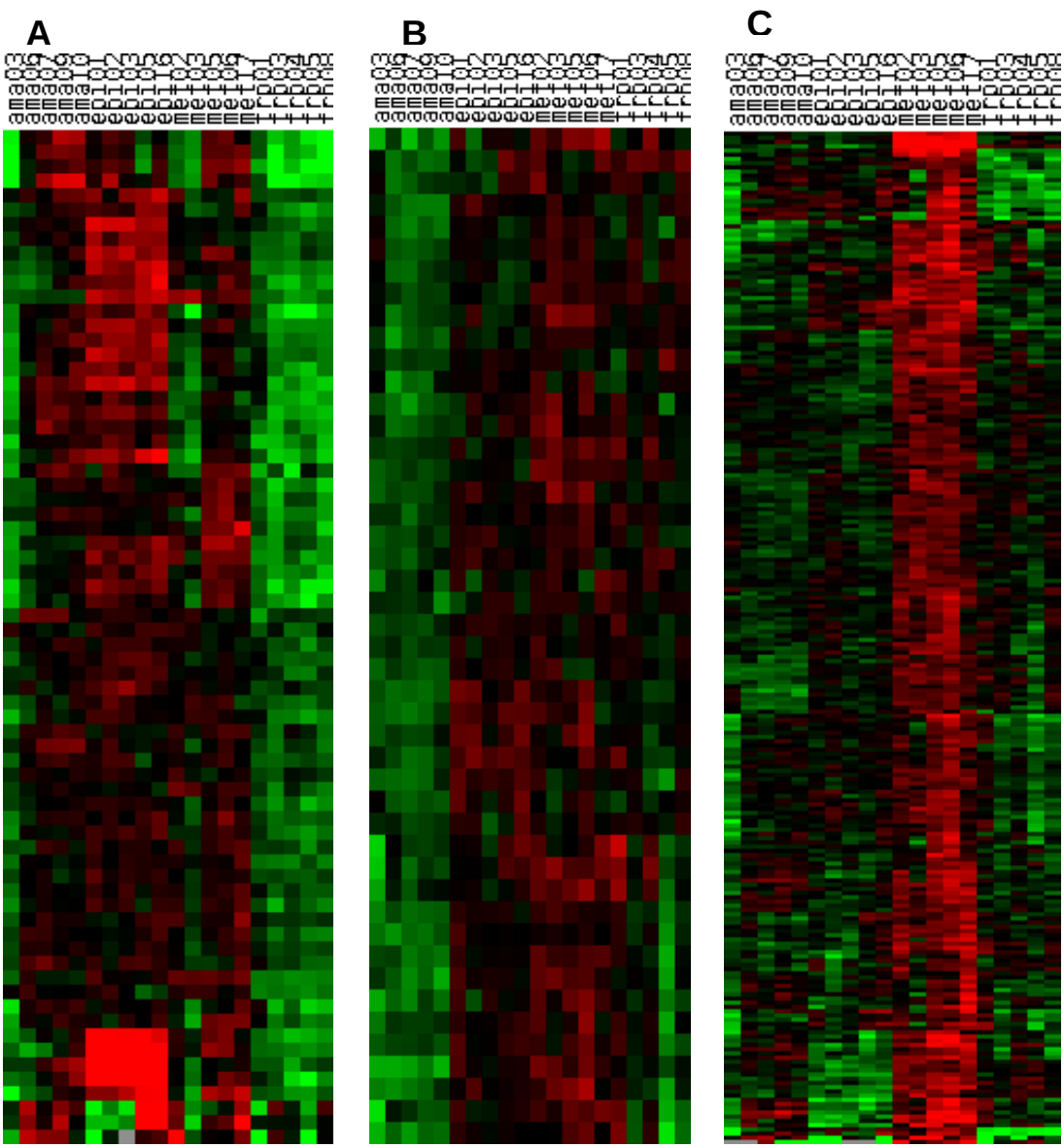


FIGURA 23 – Clusterização hierárquica dos genes marcadores em EpiMetAma (A), EpiMetTrp (B) e Met (C).

Foram identificados 70 genes marcadores em EpiMetAma, 46 em EpiMetTrp e 240 em Met.

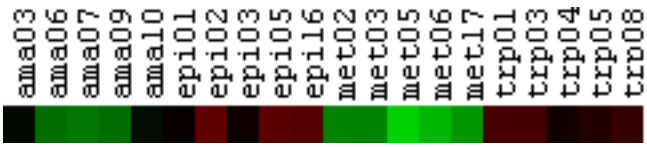


FIGURA 24 – Clusterização hierárquica do único gene marcador de EpiTrp.

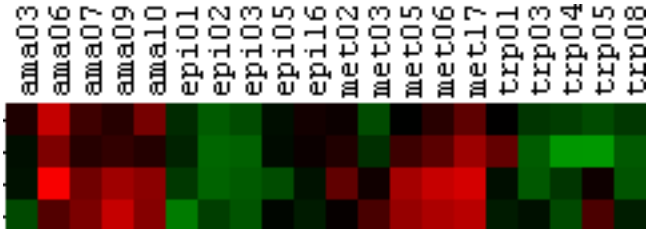


FIGURA 25 – Clusterização hierárquica dos quatro genes marcadores de MetAma.

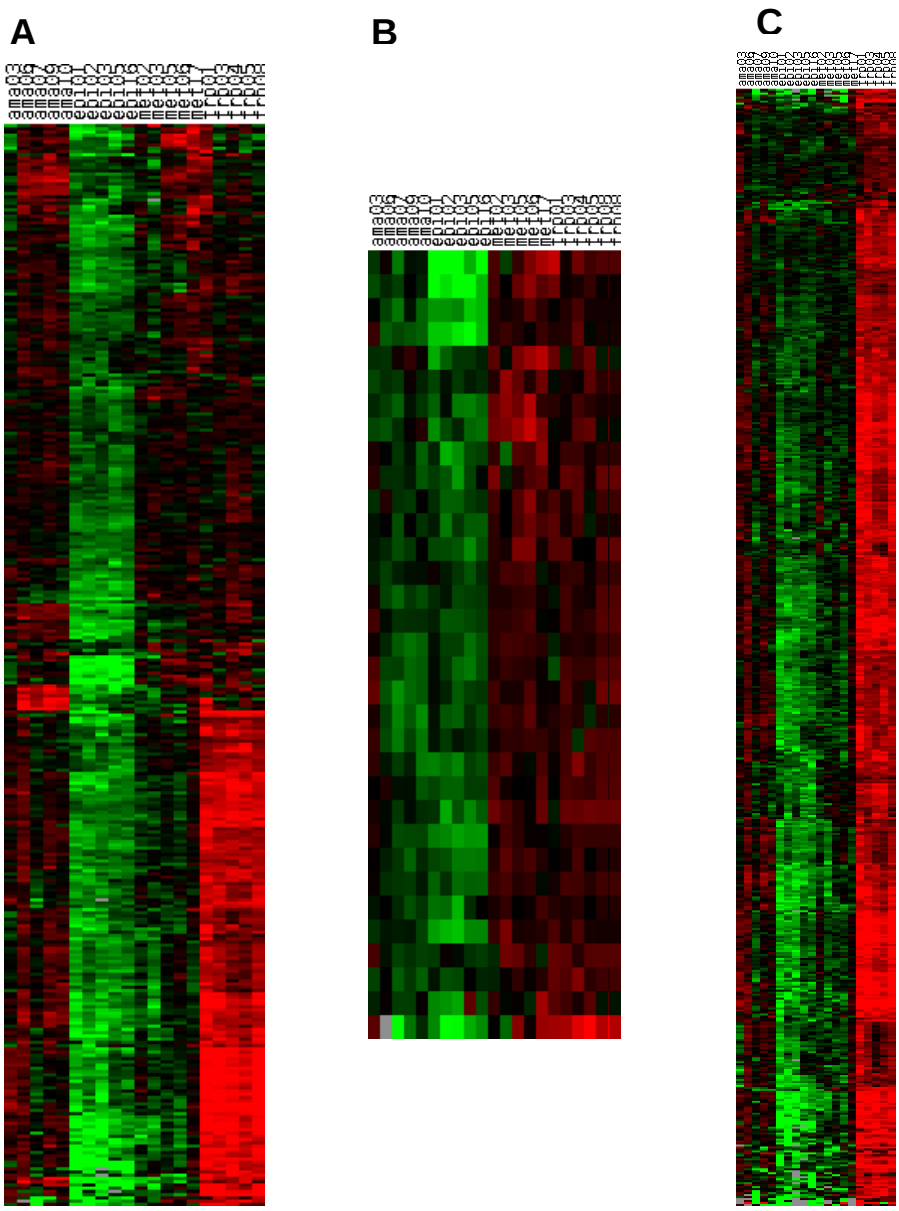


FIGURA 26 – Clusterização hierárquica dos genes marcadores de MetAmaTrp (A), MetTrp (B) e Trp(C).

Foram identificados 334 genes marcadores em MetAmaTrp, 33 em MetTrp e 563 em Trp.

Em MetAmaTrp, as formas infectivas gerais, foram identificados 334 genes candidatos. Desses, a grande maioria são trans-sialidades, gp63, mucinas e MASPs, e cerca de 50% de proteínas hipotéticas ou hipotéticas conservadas. Entre os

domínios enriquecidos estão os de mucina (Mucin, PF01456, valor p= $4,6x10^{-18}$) e trans-sialidade (Tr-sialidase_C, PF11052, valor p= $1,1x10^{-8}$).

Em MetTrp, as formas infectivas de transição entre os hospedeiros, 33 genes candidatos, sendo que 25 correspondiam a proteínas hipotéticas conservadas (75,8%). Foram indentificados como enriquecidos os seguintes domínios: Sof1 (PF04158), p25-alpha (PF05517), RAMP4 (PF06624) e DUF1930 (PF09122), todos com valor de p= $7,3x10^{-3}$, pois somente uma proteína contendo esse domínio foi identificada nesse grupo, sendo que em todo o proteoma de *T. cruzi,* essa é a única ocorrência do domínio.

Dos 563 genes identificados em Trp, a grande maioria foi de proteínas de superfície: 280 MASPs (49,7%), 115 mucinas (20,4%), 85 trans-sialidases (15,1%), entre outras proteínas de superfície. E ainda cerca de 50 proteínas hipotéticas (8,9%). Como esperado os domínios enriquecidos são mucina (Mucin - PF01456) e trans-sialidase (Tr-sialidase_C - PF11052), com valores de p= $3,8x10^{-59}$ e p= $5,2 x10^{-88}$, respectivamente.

Foram encontrados 547 genes *housekeeping*, supostamente expressos constitutivamente ao longo de todo o ciclo de vida do parasita. Interessantemente, 400 desses genes são de proteínas hipotéticas ou hipotéticas conservadas (73,2%). Isso é contra-intuitivo, pois genes *housekeeping* são geralmente muito conservados evolutivamente e, portanto seria esperado que eles tivessem uma anotação informativa. Diversas hipóteses podem explicar esse padrão:

- Erro de predição do gene: essas regiões consideradas codificadoras na realidade não o são e os mapeamentos das leituras de RNA-Seq são errôneos. Geralmente, nessa situação os mapeamentos espúrios tem uma distribuição uniforme, o que causaria a aparente ausência de modulação de um transcrito real. Contra essa hipótese, observamos que a grande maioria das proteínas hipotéticas presentes nesse grupo são conservadas, isto é, provavelmente apresentam uma região codificadora real.
- Genes *house-keeping* não conservados evolutivamente: a maioria dos organismos modelos são distante de *T. cruzi* e portanto é possível que existam genes de manutenção da homeostasia celular de tripanossomatídeos que são bastante conservados. Essa é uma hipótese que merece ser avaliada em mais detalhes futuramente.

- Genes restritos primariamente aos tripanossomatídeos, com funções pontuais e não essenciais, cuja potencialidade de regulação não é alta, isto é, eles não são selecionados evolutivamente para serem modulados. Por exemplo, não estariam relacionados a funções essenciais de *T. cruzi* que necessitam modulação, como resposta ao ambiente, infecção etc.

5.3.2 Busca por motivos no 3'UTR dos genes marcadores do ciclo de vida

A fim de identificar possíveis elementos regulatórios no 3'UTR que estejam relacionados à regulação estágio-específica de genes durante o ciclo de vida, foram feitas buscas por motivos nas seqüências 3'UTR de cada uma das 15 categorias.

As categorias EpiTrp e MetAma não tiveram nenhum motivo encontrado nas UTRs dos genes que as compõem. Isso é possivelmente devido ao fato do pequeno número de seqüencias, apenas uma para EpiTrp, o que inviabiliza a análise, e 4 para MetAma, o que dificulta a identificação de motivos nesses genes.

Para as demais 13 categorias, foram identificados 232 motivos, na análise com janela entre 6 e 15 nucleotídeos, e 211 motivos com janela entre 16 e 26. A significância de cada motivo em dada categoria foi avaliada utilizando como controles: todo o genoma de *T. cruzi* (geral), os genes não modulados (*housekeeping*), e a categoria inversa (ver correspondência em material e métodos).

5.3.2.1 Motivos de Epi

5.3.2.1.1 Motivos de Epi com janela 6-15

Como pode ser visto na TABELA 6, para os 11 motivos encontrados em Epi apenas os motivos 10 e 11 foram enriquecidos no 3'UTR dos genes dessa categoria em geral, sendo que para o motivo 10 o enriquecimento em relação ao controle inverso é mais significativo. Para os motivos 2 e 4, o enriquecimento só foi estatisticamente significativo para o controle inverso. É importante reforçar que a análise de enriquecimento significativo permite eliminar motivos frequentes nos

genes da categoria Epi, mas que são muito comuns em outros genes, como por exemplo o motivo 1, que foi o mais frequente em Epi, mas que é também muito comum nos controles. Isso é mais evidenciado ao verificamos que a composição desse motivo é extremamente simples, sendo um trecho de poli-T.

TABELA 6 – Motivos identificados em Epi com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	TTTTTTTTTTTTTTT	1.27E-01	3.94E-01	6.13E-02
2	[AG][AC]A[AC][AG][AG][AG][AC]AAAAA[AC]A	2.35E-02	1.00E+00	3.73E-08
3	G[AG][GA][GA]G[AG][GA][AG][GA][AG][GA][GA][AG][GA]	8.49E-03	6.61E-02	2.02E-04
4	GTGTGTGTGTGT[GT]TG	4.80E-03	4.58E-02	1.33E-07
5	[TA]TT[TA]TT[AT]TT[AT]TT[TA]TT	2.07E-01	1.30E-01	6.86E-01
6	A[TC]ATA[TC]A[TC][AG][TC]ATATA	1.00E+00	1.99E-01	9.07E-03
7	[CT][TC][CT][CT]C[TC][CT][TC][CT][TC][CT][TC][CT][TCA]C	7.49E-01	7.57E-01	7.12E-03
8	TTT[GT]TT[TG][TC]T[GC]T[TG][TG][TC]T	9.39E-01	9.27E-01	3.48E-01
9	AAAAAA[AT][AG][AC]AA[TCA][GAT]A[AC]	7.93E-01	4.37E-01	2.27E-03
10	[GAC][AG][GC]G[CA][GCA]G[GA]G[AG][GA]G[AG][GA][GA]	1.59E-15	2.04E-07	1.91E-27
11	[CA]GG[GC][ATC][GA]CA[GC]C[AG]GC[AC]	6.67E-20	7.74E-06	4.25E-08

O motivo 10 de Epi foi identificado com um e-value de $1x10^{-12}$, e esteve presente em 59 das 252 sequências utilizadas para busca por motivos (23,4%). A visualização do motivo no formato LogoPlot está representado na FIGURA 29.

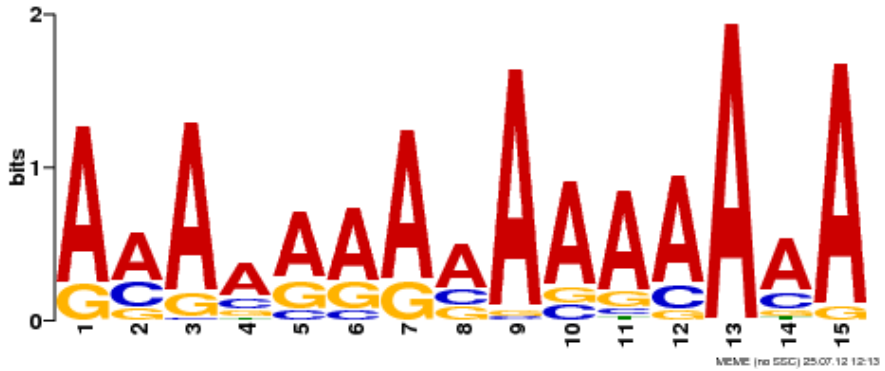


FIGURA 27 – Logo do motivo 2 de Epi.

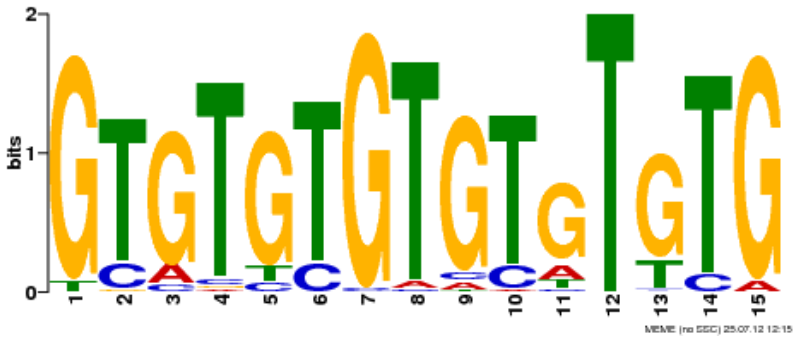


FIGURA 28 – Logo do motivo 4 de Epi.

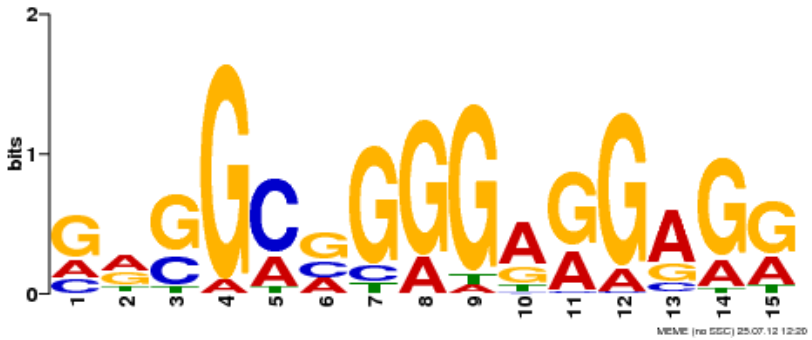


FIGURA 29 – Logo do motivo 10 de Epi.

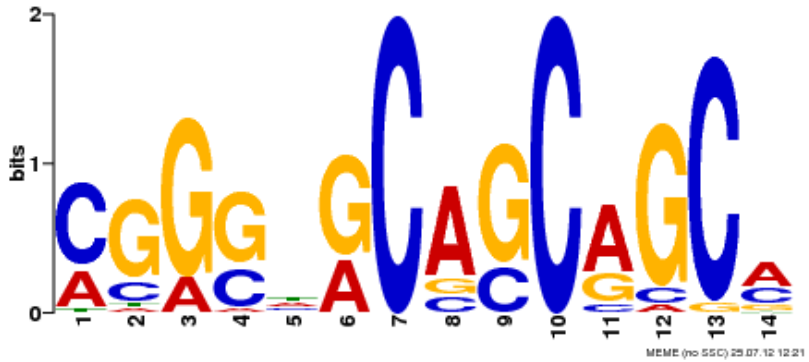


FIGURA 30 – Logo do motivo 11 de Epi.

O motivo 11 de Epi foi identificado com um *e-value* de $6x10^{-7}$, e esteve presente em 21 das 252 sequências utilizadas para busca por motivos (8,3%). A visualização do motivo no formato LogoPlot está representado na FIGURA 30.

Na busca desses dois motivos contra os 3 controles (geral, *housekeeping* e inverso) os valores de p foram altamente significativos, indicando que esses motivos podem estar relacionados a regulação da expressão gênica desse conjunto de genes nessa fase do ciclo de vida.

A fim de evidenciar a distribuição dos motivos significativos identificados em

Epi nas diferentes seqüências, fizemos uma análise de compartilhamento, isto é, com qual frequência um determinado motivo co-ocorre em uma seqüência (TABELA 7).

TABELA 7 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em Epi 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	2	4	10
4	57			
10	59			
11	21			

É possível evidenciar que os motivos mais enriquecidos (10 e 11) apresntam um padrão de co-ocorrência mediano para fraco (38%), enquanto que os motivos 10 e 11 co-ocorrem frequentemente com o motivo 2 (72%). A co-ocorrência entre os motivos 2 e 4 é media (56%). É interessante notar que a ocorrência do motivo 4 com os motivos 10 e 11 é fraca (19%). Esses dados reforçam a possibilidade que os domínios sejam reais por sua co-ocorrência em genes modulados fortalece a hipótese de sítios distintos que atuam em conjunto, interagindo com diferentes fatores regulatórios em trans.

Na FIGURA 31, a via de metabolismo de cisteína e metionina, que apresentou uma maior quantidade de genes modulados em Epi está ilustrada, mostrando a ocorrência dos genes com os motivos mais enriquecidos na presente análise.

5.3.2.1.2 Motivos de Epi com janela 16-25

Os motivos maiores identificados em Epi tendem a ser menos complexos, mas mesmo assim são significativos. Embora geralmente as RBPs atuem em motivos lineares pequenos, é possível que elas interajam com motivos maiores ou que esses representem estruturas secundárias conservadas cuja projeção na representação primária ainda apresenta alto grau de conservação. Obviamente, a ocorrência aleatória de motivos maiores é bem menos provável do que a de motivos maiores, o que justifica o enriquecimento estatisticamente significativo de motivos

não tão complexos. No entanto, o único motivo que apresentou enriquecimento significativo nas três comparações é o motivo 10, o motivo mais complexo dentre os enriquecidos.

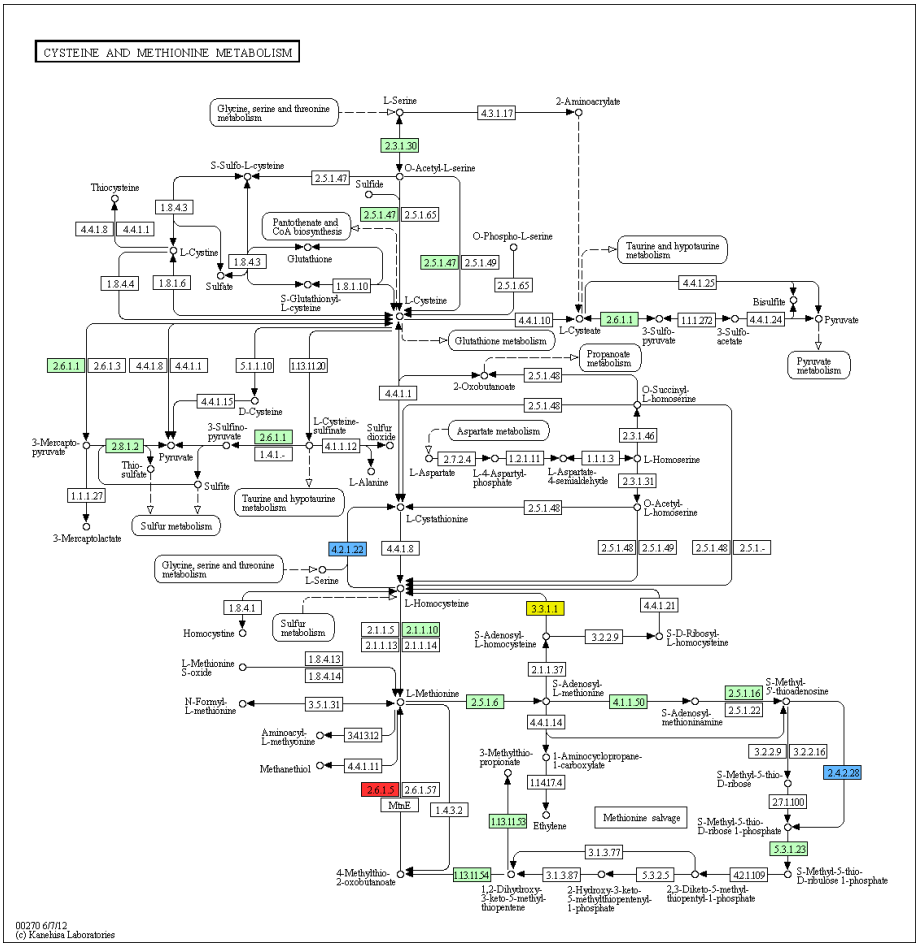


FIGURA 31 - Representação do mapa do KEGG da via metabólica de cisteína e metionina com genes modulados em Epi contendo os motivos mais enriquecidos

Os genes com o motivo 10 estão marcado em vermelho, os genes com o motivo 11 estão marcado em azul e os genes nos quais os dois motivos ocorrem estão marcados em amarelo. Os demais genes de *T. cruzi* que estão nessa via estão marcados em verde.

TABELA 8 – Motivos identificados em Epi com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	3.98E-17	6.97E-01	2.97E-03
2	2.53E-03	7.59E-01	3.20E-10
3	6.84E-01	2.48E-01	2.81E-03
4	3.15E-03	7.85E-02	1.32E-06
5	7.20E-02	3.04E-01	4.40E-01
6	8.09E-05	1.42E-02	3.50E-15
7	3.28E-01	7.42E-01	4.70E-04
8	3.88E-01	3.93E-01	9.91E-06
9	1.12E-02	2.38E-02	4.15E-01
10	5.69E-25	1.58E-08	2.70E-10
11	7.15E-01	2.30E-01	6.24E-02
12	4.26E-05	1.09E-02	3.61E-02

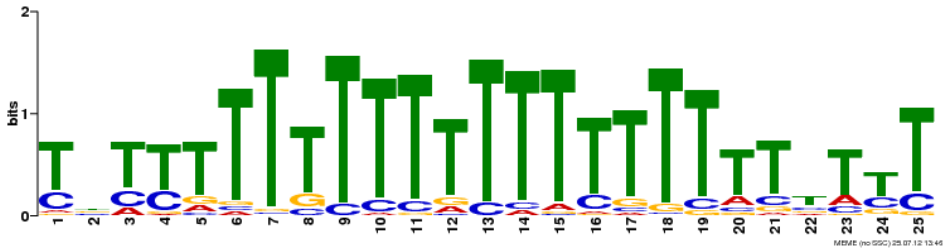


FIGURA 32 – Logo do motivo 1 de Epi.

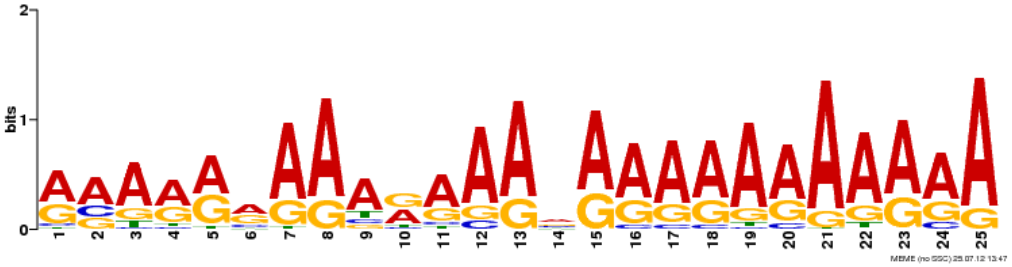


FIGURA 33 – Logo do motivo 2 de Epi.

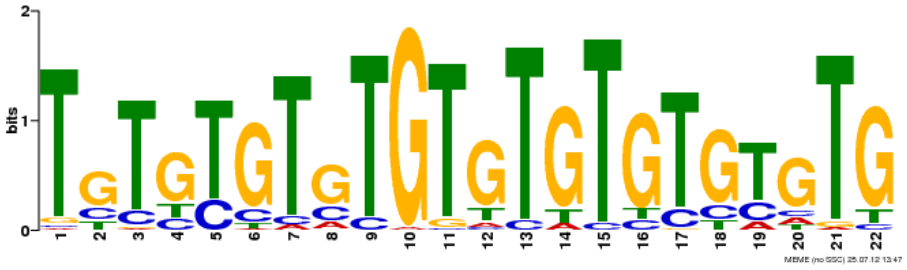


FIGURA 34 – Logo do motivo 4 de Epi.

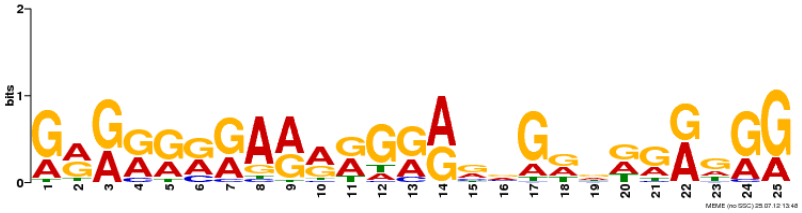


FIGURA 35 – Logo do motivo 6 de Epi.

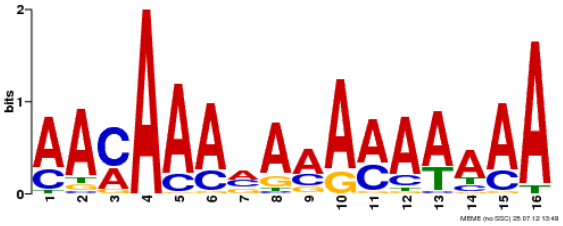


FIGURA 36 – Logo do motivo 8 de Epi.

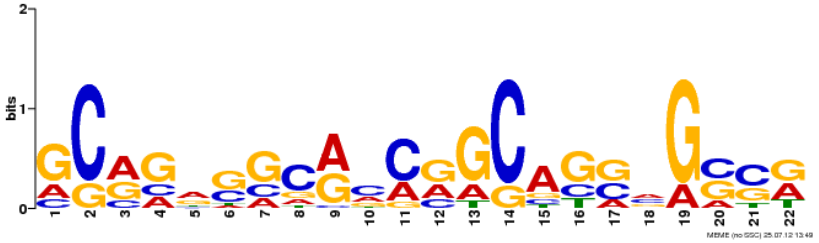


FIGURA 37 – Logo do motivo 10 de Epi.

A análise de compartilhamento está representada na TABELA 9. De maneira geral, o grau de co-ocorrência é alto: o motivo 1 co-ocorre em cerca de 72% das sequencias com os outros motivos (58% para o motivo 8); o motivo 2 co-ocorre em 89% das vezes com o motivo 8, e 57% com o motivo 4; já o motivo 4 co-ocorre raramente com os motivos 8 (8%) e 10 (13%); os motivos 6, 8 e 10 co-ocorrem em 59% das vezes.

TABELA 9 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em Epi janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	1	2	4	6	8
2	144	-				
4	49	-	-			
6	112	-	-	-		
8	48	-	-	-	-	
10	37	-	-	-	-	-

Esses resultados podem reforçar a noção de que esses elementos podem interagir para criar o padrão identificado, o de aumento em Epi.

5.3.2.2 Motivos de Met

5.3.2.2.1 Motivos de Met janela 6-15

Dos 13 motivos encontrados em Met, três deles apresentaram significância estatística, sendo que somente um para os três controles utilizados (TABELA 10).

TABELA 10 – Motivos identificados em Met com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	TTTTTTTTTTTTTTT	8.73E-01	9.24E-01	1.00E+00
2	AAAAAAAAAAA[AG]AAA	2.94E-03	3.07E-01	2.01E-01
3	ATATATA[TC]ATATATA	6.11E-07	1.17E-08	3.25E-06
4	[AC]A[ACG][AG][AC]A[ACG][AG][ACG]AAA[CA]A[AC]	7.81E-05	2.73E-02	1.10E-01
5	TGT[GC]TGT[GT]T[GC]T[GT]T[GT]T	4.25E-01	5.82E-01	3.95E-01
6	[TA]TT[TA]TT[TA]TT[AT]TT[TA]TT	4.41E-01	2.77E-01	1.00E+00
7	G[AG][AG][AG][GA][AG][GA][GA][AG][AG][GA][GA][GA][AG][AG]	1.33E-02	2.08E-01	7.49E-01
8	[TC][TC][CT][CTG][CT][CT][TC][CT][CT][TA][CT][CT][CT][CT]C	3.01E-01	1.00E+00	1.00E+00
9	AT[AG]TATATATAT[ACG][TC]	4.00E-08	1.41E-08	1.02E-05
10	[AG][AG][AG]A[AG][AG]AA[AG][AG][GAT][GA]A[AG]G	3.02E-03	3.06E-01	5.23E-01
11	T[TG]TT[GT]TT[TG]TT[GT][TC]TT[TC]	3.50E-01	5.87E-01	4.61E-01
12	[ACT]AA[TAC]AA[TCA]AA[TAC]AA[AT][AG]A	8.89E-04	3.37E-01	3.88E-01
13	[CGA][CG][GC]C[ACG][CG][ACG]C[GAC][GA][CA][CGA][AC][CG][CG]	4.91E-07	8.77E-03	3.02E-01

O motivo 3 (FIGURA 38) foi identificado com e-value de $1x10^{-184}$ em 98 das 233 sequências utilizadas para a busca do motivo (42%). Já o motivo 9 foi identificado em 39 sequências (16,7%), com e-value de $7x10^{-13}$ (FIGURA 39). É importante ressaltar que ambos motivos são muito semelhantes em sua composição, sendo que o motivo 9 apresenta algumas posições degeneradas, enquanto que o motivo 3 é mais conservado. O motivo 13 (FIGURA 40) tem uma composição bem distinta dos outros dois motivos.

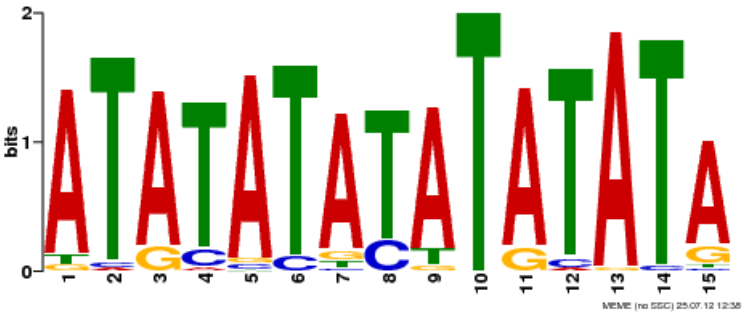


FIGURA 38 – Logo do motivo 3 de Met.

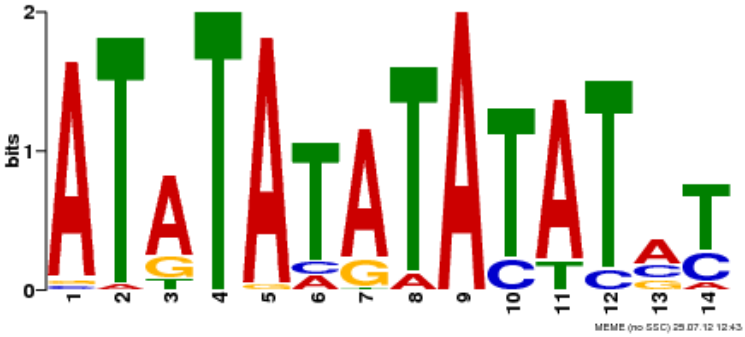


FIGURA 39 – Logo do motivo 9 de Met.

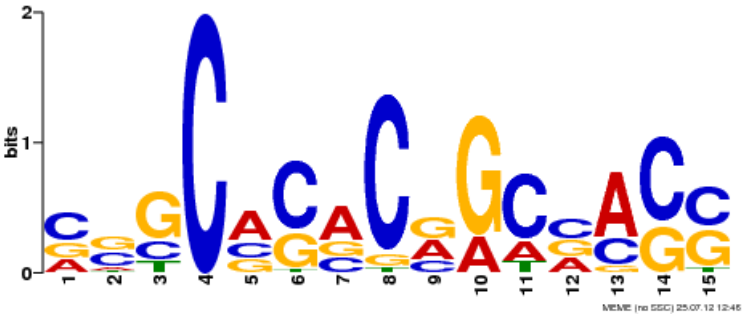


FIGURA 40 – Logo do motivo 13 de Met.

Na TABELA 11, é possível identificar o grau de co-ocorrência dos motivos enriquecidos em Met. Conforme esperado, os motivos 3 e 9 apresentam um grau de co-ocorrência alto (80%), enquanto que a ocorrência dos mesmos com o motivo 9 é menor (36% e 16%). No entanto, conforme mencionado anteriormente, os motivos 3 e 9 são muito semelhantes e provavelmente essa medida indicaria mais uma identificação conjunta por similaridade e não por co-ocorrência, isto é, o motivo 3 pode ser considerado como presente em uma sequência na mesma região do motivo 9, mas a similaridade com esse é maior. Esse tipo de análise discriminatória é mais complexa, e planejamos incorporá-la no sistema no desenvolvimento futuro.

TABELA 11 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em Met Janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	3	9
9	39	-	
13	25	.	.

Na FIGURA 41, a via da pentose-fosfato, que apresentou uma maior quantidade de genes modulados em Epi está ilustrada, mostrando a ocorrência dos genes com os motivo 3. É interessante salientar que ambos genes estão no mesmo nível da via metabólica.

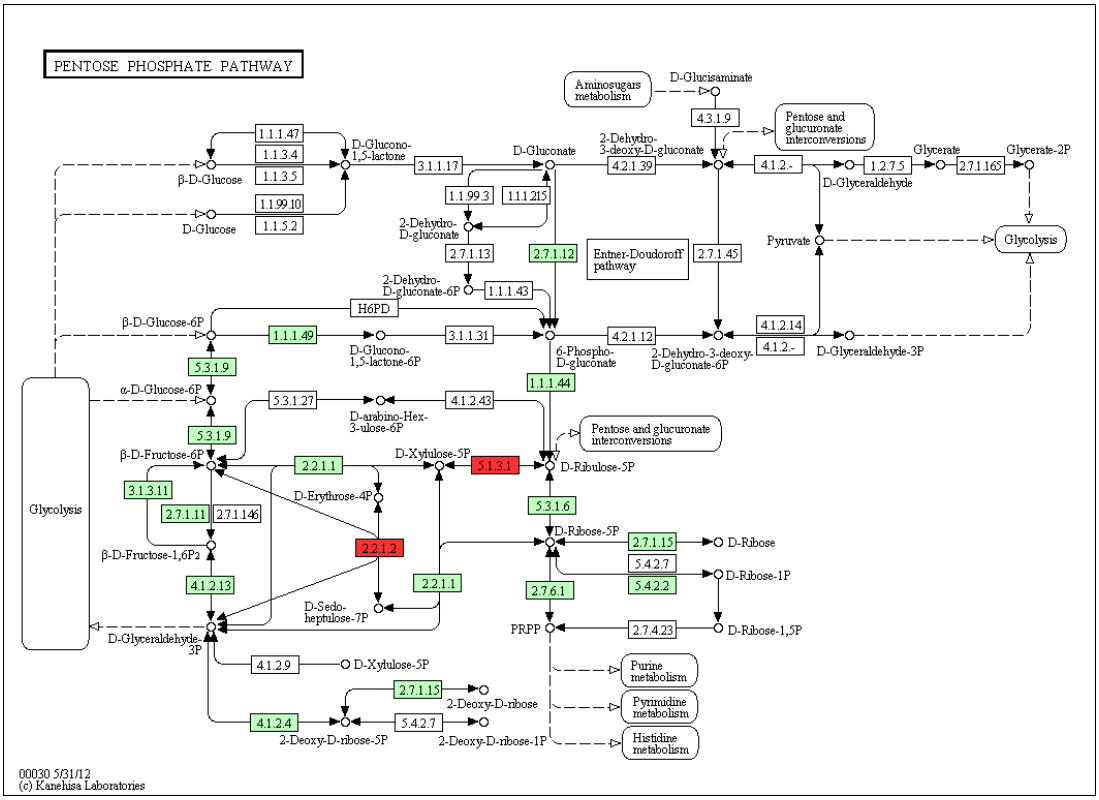


FIGURA 41 - Representação do mapa do KEGG da via metabólica de cisteína e metionina com genes modulados em Met contendo os motivos mais enriquecidos

Os genes com o motivo 3 estão marcado em vermelho. Os demais genes de *T. cruzi* que estão nessa via estão marcados em verde.

5.3.2.2.2 Motivos de Met janela 16-25

Na TABELA 12, é possível evidenciar que em Met, janela de 16 a 25, foi selecionado somente um motivo enriquecido. Da mesma forma que para Epi, esse motivo apresenta uma complexidade relativamente baixa, muito semelhante aos identificados em Met, janela 6-15 (motivos 3 e 9). É interessante ressaltar que, embora os motivos com janela entre 16 e 25 sejam menos complexos, os identificados em Epi e Met são bem distintos.

TABELA 12 – Motivos identificados em Met com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	7.19E-02	8.75E-01	6.69E-01
2	1.78E-01	6.08E-01	8.91E-01
3	1.93E-08	3.56E-07	7.39E-06
4	8.61E-04	1.33E-01	2.83E-01
5	1.02E-01	9.20E-01	6.90E-01
6	3.18E-01	6.36E-01	1.00E+00
7	2.71E-01	1.49E-01	2.06E-01
8	1.52E-02	1.80E-01	3.37E-01
9	4.05E-03	3.43E-01	5.22E-01
10	5.94E-01	7.88E-01	3.25E-01

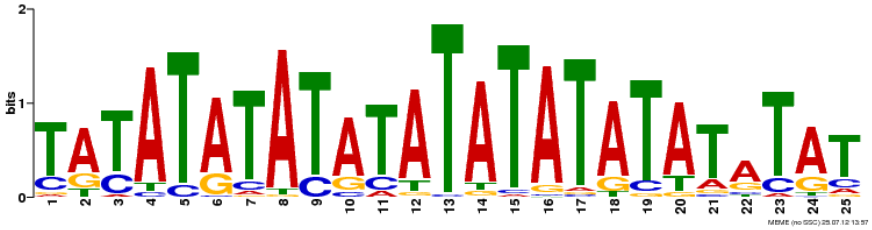


FIGURA 42 – Logo do motivo 3 de Met.

5.3.2.3 Motivos de Ama

5.3.2.3.1 Motivos de Ama width 6-15

Os motivos 1 e 6 de Ama tiveram um enriquecimento com grau fraco em relação a somente um dos controles, sendo que o motivo 1 é pouco complexo (FIGURA 43); o motivo 6 apresenta um padrão melhor definido (FIGURA 44). Já o motivo 8 (FIGURA 45) apresentou valores significativos para dois dos controles, e é

um motivo bem definido.

TABELA 13 – Motivos identificados em Ama com janela 6-15 com seleção dos enriquecidos..

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	AA[AG]AA[AG]AA[AG]AA[AG]AAA	4.32E-06	1.14E-04	1.84E-02
2	[TC][TC]TTTTTTT[TC]TTT[TC]T	3.48E-01	4.13E-01	8.29E-01
3	[TC][TGA]TTT[AT]TTT[AT]TT[TA]TT	6.43E-01	1.00E+00	5.18E-01
4	G[TG]G[TG]G[TA]G[TG]GTG[GT][GCT][TG]G	5.83E-02	4.46E-02	5.56E-01
5	AAA[GC]AAA[AG]A[AG]AA[AGC]AA	1.62E-05	6.50E-04	4.94E-02
6	[CT][ATC][CT][GCT]C[AT]C[TC]C[TG]C	3.00E-03	3.64E-06	2.24E-04
7	[TG][TG]T[GT]TT[GT]TT[TG]TT[TG]TT	8.32E-02	1.07E-01	6.68E-01
8	G[GA]G[GA]GAGG[GA]A[AG]	1.20E-07	4.08E-06	2.09E-01

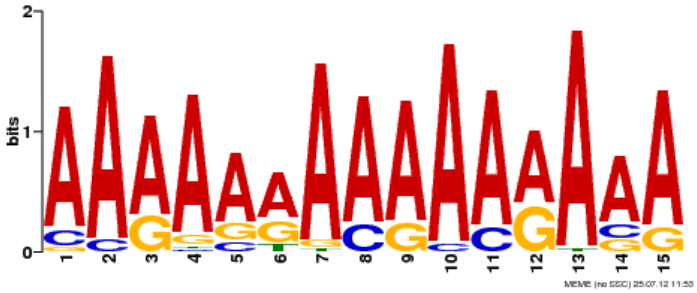


FIGURA 43 – Logo do motivo 1 de Ama.

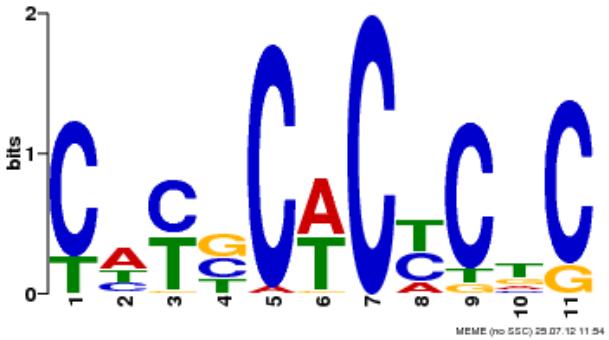


FIGURA 44 – Logo do motivo 6 de Ama.

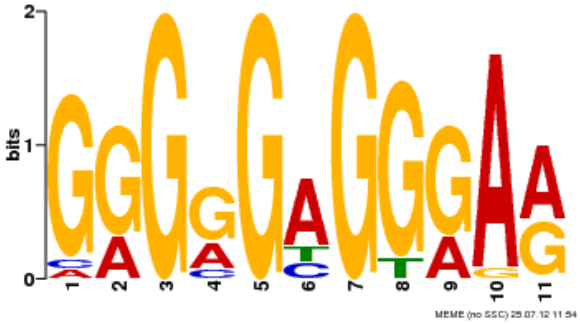


FIGURA 45 – Logo do motivo 8 de Ama.

Na TABELA 14, vemos o padrão de co-ocorrência dos motivos enriquecidos em Ama. Embora os três padrões sejam bem distintos, eles co-ocorrem com frequência alta: os motivos 1 e 8 co-ocorrem em 89% das sequencias; os motivos 1 e 6 em 59% e os motivos 6 e 8 em 39%. Pelo fato de que os motivos são distintos, isso indica que esses elementos podem estar atuando em sinergia para a determinação da expressão diferencial observada nessa categoria.

TABELA 14 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em Ama janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	1	6
6	32	-	
8	18	-	-

5.3.2.3.2 Motivos de Ama janela 16-25

Na TABELA 15, é possível evidenciar que em Ama, janela de 16 a 25, não foi selecionado nenhum motivo enriquecido.

TABELA 15 – Motivos identificados em Ama com janela 16-25 com seleção dos enriquecidos.

Motivo	Geral	Housekeeping	Inverso
1	3.86E-04	1.50E-02	4.94E-02
2	6.51E-01	3.16E-01	1.00E+00
3	2.65E-01	2.43E-01	6.54E-01
4	4.73E-04	2.84E-04	5.28E-02
5	3.92E-04	4.19E-03	4.94E-02
6	6.81E-01	2.25E-01	2.89E-01

5.3.2.4 Motivos de Trp

5.3.2.4.1 Motivos de Trp janela 6-15

Na categoria Trp foi identificado o limite de 50 domínios (TABELA 16). Podemos observar uma grande quantidade de motivos enriquecidos

significativamente nessa categoria. No entanto, esse resultado deve ser considerado com bastante cuidado, pois a composição dessa categoria é predominantemente de famílias multigênicas de superfície, e mesmo com a redução de redundância realizada por nós, com a construção dos super-genes de *T. cruzi,* os representantes de cada super-gene de uma mesma família podem compartilhar regiões 3'-UTR mais por origem comum do que por conservação funcional de elementos reguladores.
Conforme esperado, o compartilhamento desses domínios entre os genes constituintes é muito alto (TABELA 17). Os motivos 1 a 20 são extremamente compartilhados entre si; os motivos 21 e 23 a 47 são menos compartilhados, com exceção em relação aos motivos 1, 2, 13 e 22, que são extremamente compartilhados entre todos os genes dessa categoria. Esses resultados reforçam a necessidade de análises adicionais que necessitam ser realizadas para avaliar a existência de motivos regulatórios nas regiões UTRs de genes pertencentes à família multi-gênicas.

5.3.2.4.2 Motivos de Trp janela 16-25

Da mesma forma que na análise com a janela menor, a categoria Trp alcançou o limite de 50 domínios (TABELA 18) com a janela maior. As mesmas considerações feitas para a análise anterior são válidas. O padrão de compartilhamento entre os motivos mais frequentes é mantido, pois os motivos de 1 a 12 são compartilhados entre si. O motivo 11 é encontrado frequentemente com todos os outros motivos, o motivo 16 com os motivos 11 a 50 e o motivo 19 com os motivos 1 a 11. No entanto, em relação ao compartilhamento de motivos entre as seqüências, os motivos são menos compartilhados.

TABELA 16 - Motivos identificados em Trp com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	[CT]GCC[CA]GCAC[AT]CAC[AC]C	3.06E-258	1.67E-130	9.47E-31
2	[CG]CCACG[AT][GAT]G[CA]AC[AG]CA	0.00E+00	7.08E-192	2.06E-39
3	A[TC][GA]CAC[AT]C[CA][CT]ATGCA	5.47E-189	1.24E-93	4.13E-17
4	AGT[GA]AG[TG][GA]AGAGA[GA]A	3.38E-69	4.32E-40	3.00E-09
5	GGCGGAGCACCGTAA	1.73E-195	2.86E-97	4.81E-18
6	[AT]CCGCTGCCGT[CG]CTG	1.20E-316	1.56E-161	5.55E-33
7	GCATGGACTCTCGCC	1.90E-202	1.35E-100	1.06E-18
8	AATGTGCCCATTGTA	9.32E-187	6.29E-95	1.53E-17
9	GCATTGAGTGGGCGA	3.08E-169	8.22E-86	8.03E-16
10	TTTTGCAC[GC]ACACGC	3.83E-243	1.07E-120	7.70E-23
11	GCCCGTTGGTGCTCT	6.04E-232	1.75E-115	9.28E-22
12	TGTGTG[TC]GTC[CT][CA]CT[TG]	2.75E-106	4.92E-68	1.68E-20
13	T[GT][TA]TT[CGT]T[GT]TTTTG[TC]T	4.69E-31	7.26E-20	7.69E-08
14	CTCTCTCTCTCCCT[GC]	3.11E-39	2.02E-25	1.44E-17
15	CTGGCTGCATG	2.19E-239	3.62E-120	9.92E-23
16	TC[GT]TC[CT]TTC[TA]CTC[TA]C	2.62E-82	9.53E-59	2.45E-19
17	TGTG[TC]G[TG]G[CT]T[TG]TGC[GT]	2.65E-105	3.13E-62	6.21E-10
18	TCACCTCA	0.00E+00	1.09E-98	3.30E-18
19	GCC[GA]TGCA	0.00E+00	3.25E-98	3.62E-18
20	TGTGCTCCG[CT]GTTG[CT]	4.61E-99	3.61E-47	2.10E-08
21	GGCAATCACTATGGA	7.04E-47	6.85E-22	5.47E-04
22	[TG]TTTTT[AGT]TT[TG][TC]T[TA]TT	4.05E-23	1.99E-14	7.57E-09
23	[CG]A[CG][TG]GG[AT]GCC[AG]T[GT][TG]G	4.30E-102	3.61E-47	2.10E-08
24	TGGTCCGATGT	7.50E-62	1.19E-29	2.42E-05
25	[CT]GACACGC[AT][CT]AC[AG]AC	8.26E-59	1.39E-26	6.78E-05
26	GGAGTGACTGG	2.79E-45	6.73E-21	5.30E-04
27	TGTG[TC]GTGT[TG]T	2.98E-55	2.50E-36	9.15E-09
28	TTTTCTTTT[CT]T[AT][CT]T[CT]	2.38E-14	1.79E-07	7.91E-02
29	CT[CT]CCC[TA]GCAC[AC]GA[AC]	7.88E-49	1.48E-22	3.16E-04
30	TT[CT]CCGCCCAACTGC	8.01E-24	1.58E-11	2.43E-02
31	[GT]CTGCACTCTTCCTG	7.25E-44	3.07E-20	8.65E-04
32	GTTTGCGTGGAC[GT]CA	0.00E+00	1.63E-116	5.69E-22
33	[CT][CT][TC][TC][TC]TTTAATT[GA]TT	3.68E-15	3.87E-08	2.31E-02
34	TATATG	1.00E+00	1.00E+00	1.00E+00
35	[CT][GA][AG][GTA]AA[TC][GTA]AA[TA][TA]T[ATC]T	7.89E-45	4.90E-15	1.48E-01
36	[TG][CT]CCC[AG]CATCC	2.11E-147	3.62E-50	5.51E-09
37	[AG]CGGGGA[CT]	1.00E+00	1.00E+00	1.00E+00
38	[TC][TC][CT][TC]T[GT]TTTGCTTT[TG]	3.63E-15	1.60E-12	1.48E-04
39	TGGTTTGC[GA]G[GT]GCGG	3.19E-15	1.87E-07	1.59E-01
40	CACACGCGGTGCCGG	2.82E-10	5.52E-05	3.90E-01
41	AA[ACT]A[AG][GA]A[AGT]AAA[AG][TA][AG]A	5.63E-35	2.57E-26	1.14E-13
42	CCACGGACTGCACGA	0.00E+00	4.49E-101	9.10E-19
43	[AT][AT][AT][TA][AT][AT][CA][TA]TT[TA][TA]TT[TA]	1.07E-21	3.76E-08	1.83E-07
44	AGGA[GT][TG]GGA[GA]GA[AG][ACG]A	5.37E-33	7.61E-33	4.69E-07
45	TGTGGG[CA]T[GA]GTGGAA	5.29E-08	7.00E-03	3.79E-01
46	TGTGCT[GT]GTG[CT]ATGC	8.04E-02	5.47E-01	3.89E-01
47	[TC][CT][CT]GTGG[AC][CA]CAACAA	5.90E-10	2.72E-05	2.40E-01
48	GCTCCCTGTAT	4.00E-21	2.18E-08	9.86E-02
49	TATATTGC	1.00E+00	1.00E+00	1.00E+00
50	TTTGTGCAGGTAATA	3.33E-06	7.61E-03	1.00E+00

TABELA 17 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em Trp janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

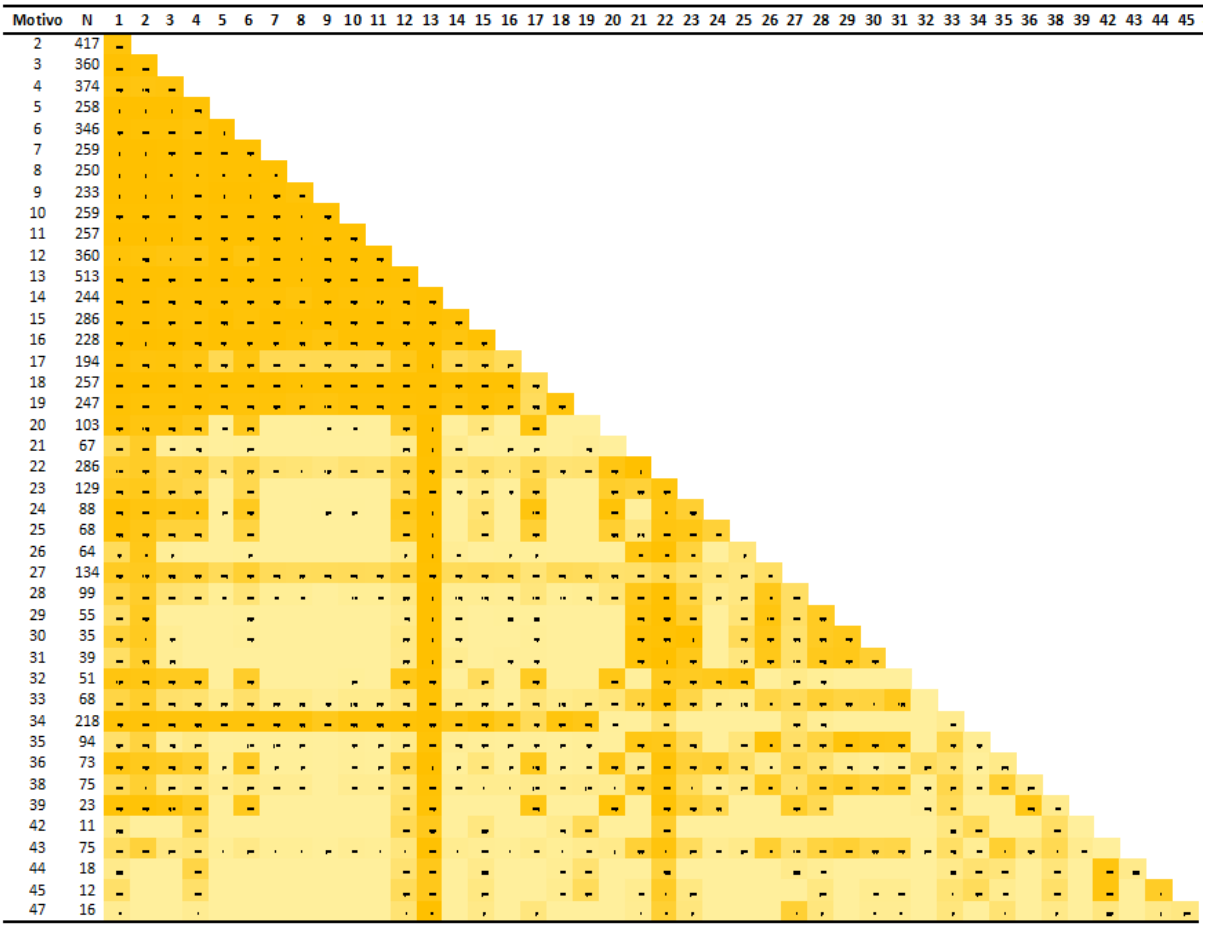


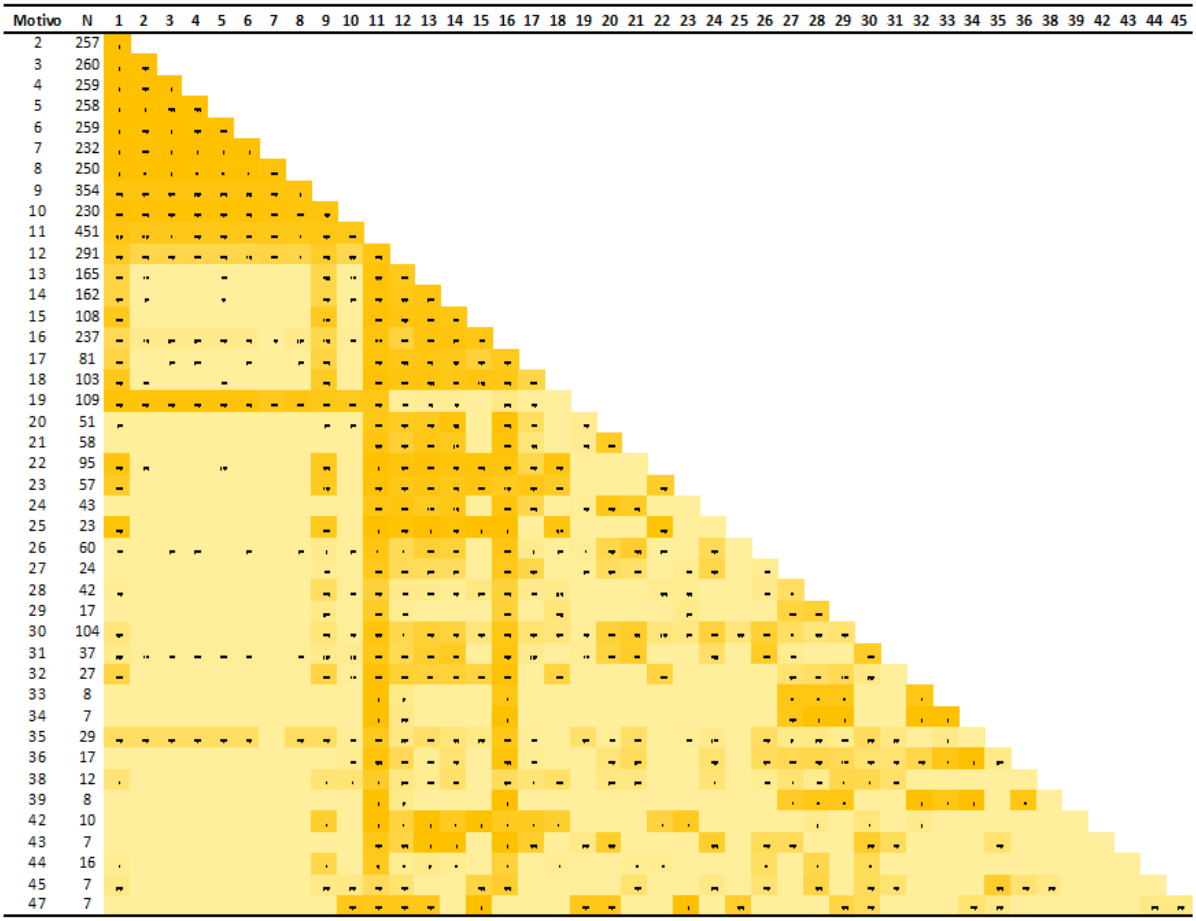
TABELA 18 – Motivos identificados em Trp com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	1.48E-305	3.31E-164	1.65E-32
2	1.88E-244	1.50E-117	3.47E-22
3	6.07E-211	1.94E-105	1.46E-19
4	2.01E-243	1.07E-120	7.70E-23
5	1.86E-219	6.73E-109	1.53E-18
6	5.98E-233	4.94E-117	4.44E-22
7	1.76E-179	1.66E-88	2.45E-16
8	5.25E-188	2.86E-97	4.81E-18
9	1.41E-59	2.19E-30	2.08E-08
10	2.88E-21	1.19E-11	1.88E-07
11	5.13E-25	1.74E-17	1.08E-09
12	6.67E-200	1.52E-126	7.40E-39
13	7.83E-140	1.11E-68	1.55E-12
14	2.37E-105	2.33E-58	3.65E-12
15	4.16E-77	3.95E-39	4.51E-07
16	3.71E-09	4.78E-03	2.07E-02
17	2.14E-114	2.74E-48	1.06E-08
18	1.05E-187	2.12E-110	1.03E-20
19	7.73E-28	1.13E-12	4.82E-09
20	8.87E-12	4.50E-06	6.86E-01
21	3.18E-48	6.85E-22	3.53E-03
22	9.38E-73	8.18E-34	4.68E-06
23	4.01E-38	1.44E-20	8.93E-04
24	3.10E-79	1.10E-30	1.37E-05
25	2.18E-129	2.05E-49	5.62E-09
26	6.09E-28	5.71E-21	1.00E+00
27	8.26E-17	2.18E-08	9.86E-02
28	5.99E-73	1.15E-48	5.20E-22
29	2.47E-233	5.34E-72	5.64E-13
30	1.07E-06	5.43E-03	3.02E-02
31	3.23E-14	1.03E-09	4.30E-02
32	2.55E-25	3.64E-09	8.08E-01
33	1.85E-290	3.75E-83	4.60E-15
34	5.33E-06	1.53E-02	1.00E+00
35	2.80E-47	1.13E-31	2.02E-15
36	1.07E-53	1.44E-25	1.16E-04
37	4.79E-01	7.70E-01	3.36E-02
38	5.73E-07	1.72E-03	1.00E+00
39	0.00E+00	2.49E-96	1.26E-17
40	3.52E-05	1.13E-02	6.11E-01
41	2.51E-05	4.20E-04	1.39E-04
42	7.65E-71	1.63E-34	2.58E-06
43	1.80E-21	3.82E-07	1.55E-01
44	9.67E-24	1.17E-16	1.21E-02
45	8.52E-51	1.47E-24	2.02E-04
46	1.86E-04	3.08E-02	1.00E+00
47	3.03E-22	1.39E-10	3.90E-02
48	6.05E-05	1.53E-02	1.00E+00
49	5.79E-04	6.19E-02	1.00E+00
50	8.47E-02	5.00E-01	3.84E-02

TABELA 19 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em Trp janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.



5.3.2.5 Motivos de EpiMet

5.3.2.5.1 Motivos de EpiMet janela 6-15

Em EpiMet foram identificados 17 motivos (TABELA 20). Desses, os motivos 2, 6, 7, 8, 10, 11, 12, 13, 14, 15 e 16 foram considerados significantes para essa categoria. O motivo 14 é o único significativo para os três controles. O motivo 2 (FIGURA 46) apresenta um padrão pouco complexo, no entanto se mostrou com significância relativamente alta contra o controle inverso. Esse resultado é interessante por reforçar a abordagem diferente que é a análise pelo controle inverso, pois esse motivo é relativamente frequente no conjunto total de 3'-UTRs de *T. cruzi,* mas esse motivo é pouco frequente no grupo AmaTrp.

Os demais motivos estão representados da FIGURA 47 à FIGURA 56. É muito interessante salientar que a maioria dos motivos identificados nessa categoria são bastante complexos e com baixo grau de entropia, isto é, com diversas posições extremamente conservadas. Uma hipótese que será avaliada em mais detalhes futuramente é a relevância desses motivos na determinação da expressão de genes de grande importância para o estágio de *T. cruzi* no hospedeiro inseto. Os motivos 10, 11 e 12 possuem genes com a presença do domínio mucina (Mucin - PF01456) e portanto devem ser múltiplos elementos regulatórios de mucinas.

TABELA 20 – Motivos identificados em EpiMet com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	TTTTTTTTTTTTTTT	2.35E-01	3.95E-01	7.22E-02
2	AAA[AC]AAA[AG]AAAA[AG]AA	4.58E-03	2.53E-01	1.60E-13
3	T[TA]T[TA]T[TA]T[TA]T[AT][TA][TA]T[AT]T	1.56E-01	3.66E-02	8.92E-01
4	[GA][ACG][AG]G[GA][AG][AG]G[AG][GA][AG][GA][GA][ACG][AG]	3.41E-02	3.20E-02	1.59E-02
5	C[CT][CT][CT][TCG]T[CTG][TC][CT]T[CT][TC][CT][TCG][CT]	5.28E-01	2.82E-01	7.37E-04
6	TGT[GC]TGTGTG[TC]GTG[TC]	1.40E-01	2.86E-01	9.51E-07
7	[CG][AC][GC][ACG]C[GAC][CA][GAC]C[CA][TCG]CC[AG]C	2.56E-06	2.79E-02	1.35E-13
8	[AG]AAA[AG]A[AC][AG][AG]CAA[AC]AA	2.03E-02	2.87E-01	1.41E-09
9	TT[TG]TT[TG]T[TG][GT][TC]TG[TC]T[GT]	9.27E-01	1.00E+00	5.03E-02
10	[CG]ACCC[TG][GC][CG][AT]GCGG[AT]G	1.05E-12	6.54E-07	1.71E-03
11	CC[CG][GT]TTGT[GT]T[TG][GC][GC]GC	1.04E-11	6.54E-07	1.71E-03
12	G[CT]GACGG[CAT][CA]G[CG][GC]C[GC]C	1.86E-14	6.54E-07	1.71E-03
13	[GC][GC]TGCGCCGCGTATA	4.47E-08	2.06E-04	2.99E-02
14	G[GT][GT]GG[AT][GA]G[CT][GA][AGTCC][TCA]GG	4.14E-13	6.71E-12	2.24E-09
15	GC[AT]C[GT][TC]GGCGTG[GA][AC]C	5.72E-09	4.92E-05	1.47E-02
16	[CG]GGCCGA[AG]TGAC[GA]TG	1.17E-06	8.56E-04	6.07E-02
17	[ATG]TTATT[AG]T[TC]ATT[AT]TT	6.67E-01	6.95E-01	1.49E-02

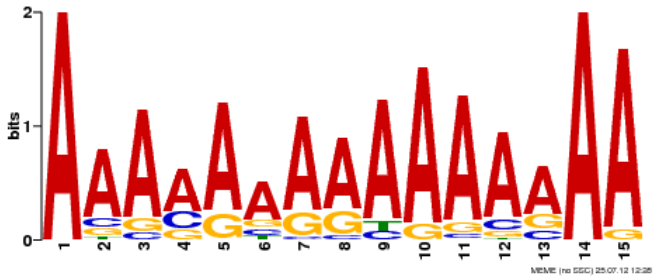


FIGURA 46 – Logo do motivo 2 de EpiMet.

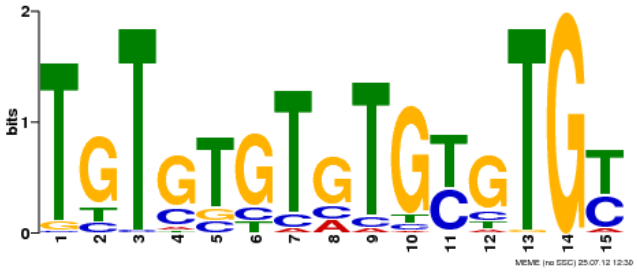


FIGURA 47 – Logo do motivo 6 de EpiMet.

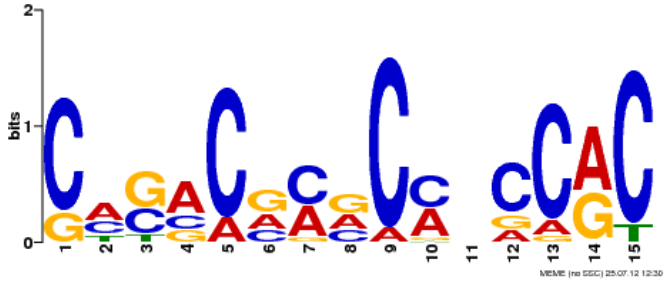


FIGURA 48 – Logo do motivo 7 de EpiMet.

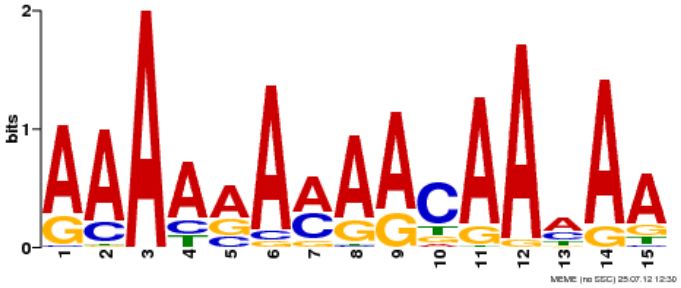


FIGURA 49 – Logo do motivo 8 de EpiMet.

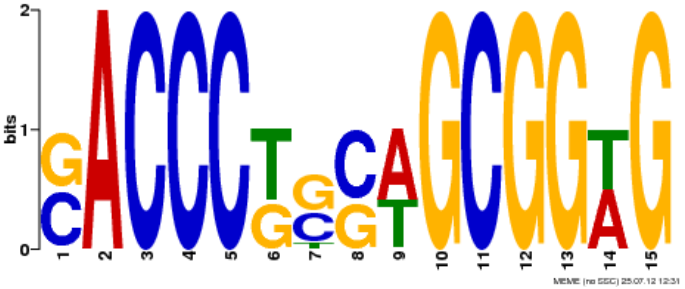


FIGURA 50 – Logo do motivo 10 de EpiMet.

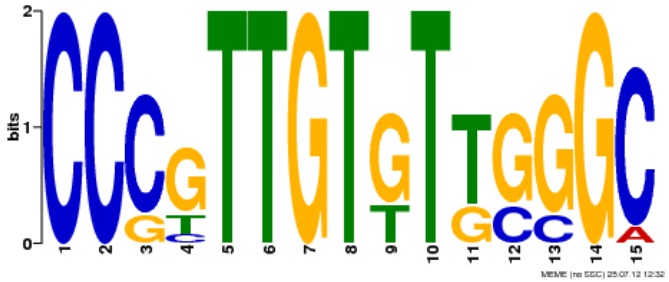


FIGURA 51 – Logo do motivo 11 de EpiMet.

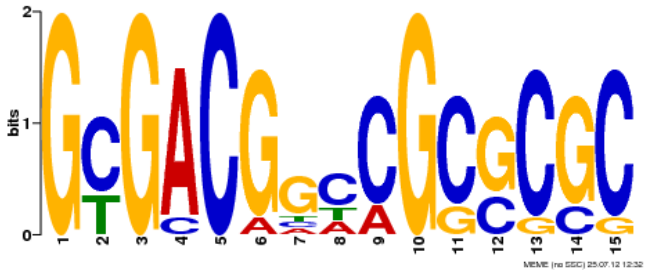


FIGURA 52 – Logo do motivo 12 de EpiMet.

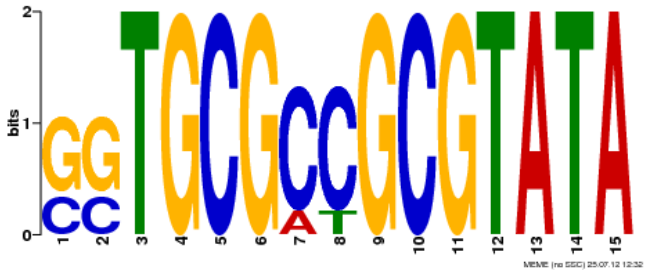


FIGURA 53 – Logo do motivo 13 de EpiMet.

FIGURA 54 – Logo do motivo 14 de EpiMet.

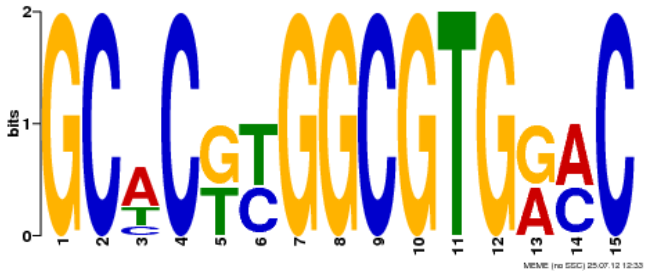


FIGURA 55 – Logo do motivo 15 de EpiMet.

FIGURA 56 – Logo do motivo 16 de EpiMet.

Na TABELA 21, vemos o padrão de co-ocorrência dos motivos enriquecidos em EpiMet. De maneira geral, a co-ocorrência desses motivos apresenta um padrão dicotômico. Mais precisamente, certos motivos são encontrados preferencialmente com um ou poucos motivos, e até mesmo não identificado em conjunto com outros motivos.

Por exemplo, os motivos 7, 10 e 11 co-ocorrem com o motivo 15 em 100% das sete seqüencias que contém o motivo 15; o motivo 7 ocorre em 90% das 10 seqüências que contém os motivos 10 e 11; esses últimos motivos ocorrem em 90% de suas sequencias. O motivo 8 ocorre conjuntamente com o motivo 2 em 95,5% de suas 44 seqüências, mas nunca é visto com os motivos 11, 12, 13, 15 e 16. O motivo 16 é visto em 80% de suas cinco seqüencias com os motivos 7, 10, 11, 14 e 15, mas nunca é visto conjuntamente com os motivos 6, 8, 12 e 13.

Esse padrão é muito interessante, ainda mais considerando a conservação dos mesmos. Uma hipótese é que a co-ocorrência desses motivos pode distinguir os 171 genes candidatos selecionados nessa categoria em sub-especializações que necessitam de regulação diferencial. Para isso, iremos avaliar a expressão desses genes nas demais condições incluídas em nosso banco de dados do projeto Reguloma, incluir dados de conservação evolutiva, presença de domínios proteicos, interatoma, entre outros. Essa análise está fora do escopo do presente trabalho, cujo objetivo primário é ilustrar a utilidade da plataforma bioinformática desenvolvida, mas demonstra a sua potencialidade em identificar padrões interessantes para futuros estudos.

TABELA 21 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiMet janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	2	6	7	8	10	11	12	13	14	15
6	51										
7	36										
8	44										
10	10										
11	10										
12	9										
13	6										
14	29										
15	7										
16	5										

5.3.2.5.2 Motivos de EpiMet janela 16-25

Os motivos 2, 7, 8, 9, 10, 11, 14 e 17 foram enriquecidos significativamente (TABELA 22). Nenhum motivo foi encontrado enriquecido nas três análises realizadas. Da mesma forma que na análise com a janela menor, um dos motivos (motivo 2, FIGURA 57) apresentou significância relativamente alta somente na comparação contra o controle inverso; esse motivo é muito similar ao menor que apresentou o mesmo padrão e foi encontrado nas mesmas sequencias. Outra característica comum aos motivos encontrados nessa análise é a sua maior complexidade e conservação das posições.

TABELA 22 – Motivos identificados em EpiMet com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	5.86E-02	3.86E-01	1.28E-01
2	6.54E-03	1.87E-01	3.89E-15
3	7.53E-01	7.17E-01	5.10E-01
4	2.09E-05	7.78E-04	3.43E-04
5	8.58E-01	3.92E-01	5.59E-02
6	1.67E-01	7.88E-02	2.98E-02
7	1.47E-13	3.59E-08	1.12E-02
8	4.25E-11	6.54E-07	1.71E-03
9	8.08E-02	1.00E+00	1.17E-08
10	6.17E-12	6.54E-07	1.66E-01
11	1.88E-01	5.93E-01	1.13E-08
12	4.15E-01	1.00E+00	3.75E-01
13	2.83E-04	3.55E-03	1.23E-01
14	3.86E-17	3.59E-08	3.13E-02
15	3.63E-02	1.33E-02	4.17E-02
16	6.35E-05	3.55E-03	1.23E-01
17	5.01E-01	7.69E-01	5.81E-09

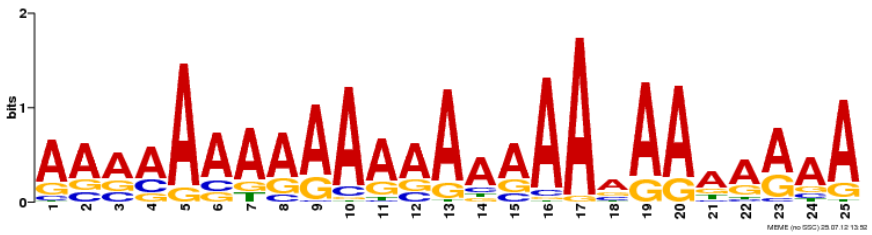


FIGURA 57 – Logo do motivo 2 de EpiMet.

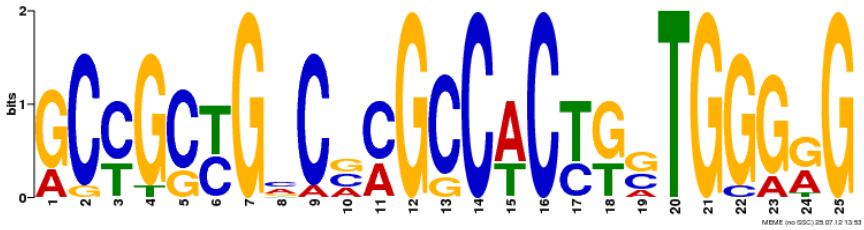


FIGURA 58 – Logo do motivo 7 de EpiMet.

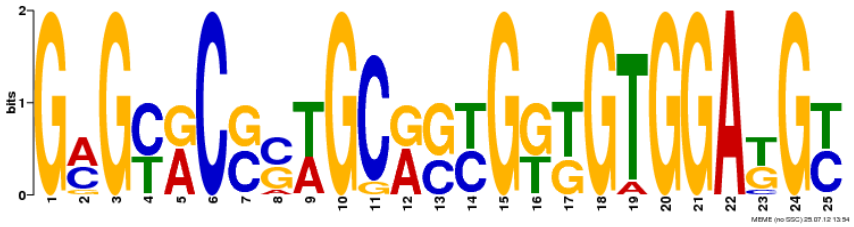


FIGURA 59 – Logo do motivo 8 de EpiMet.

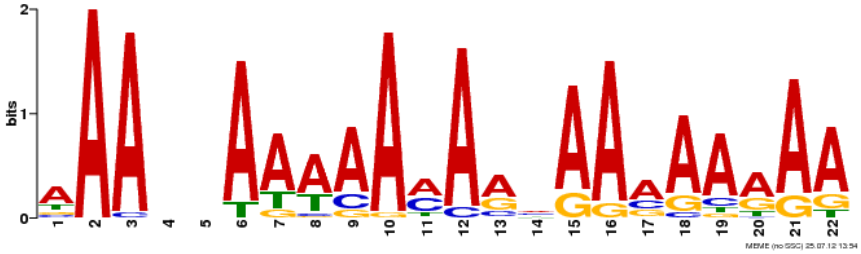


FIGURA 60 – Logo do motivo 9 de EpiMet.

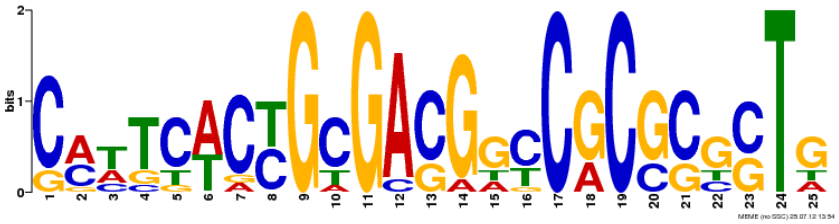


FIGURA 61 – Logo do motivo 10 de EpiMet.

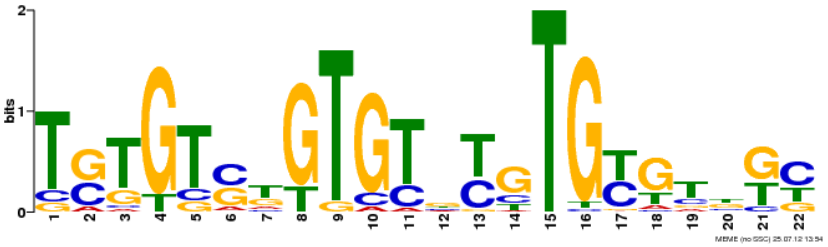


FIGURA 62 – Logo do motivo 11 de EpiMet.

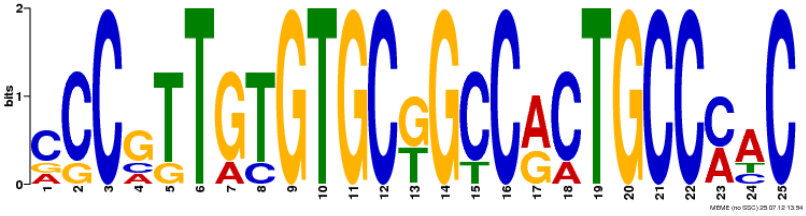


FIGURA 63 – Logo do motivo 14 de EpiMet.

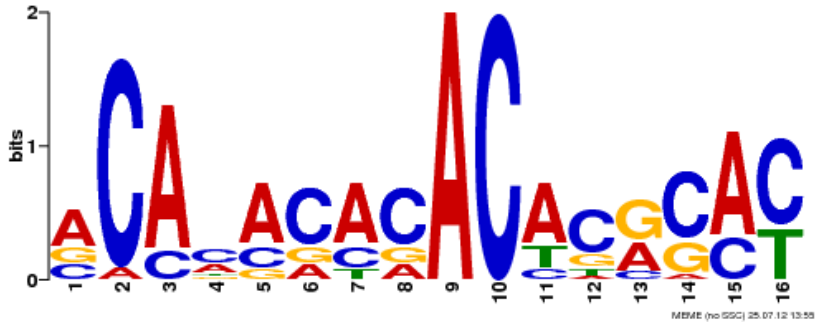


FIGURA 64 – Logo do motivo 17 de EpiMet.

Na TABELA 23, vemos o padrão de co-ocorrência dos motivos enriquecidos em EpiMet janela de 16 a 25. O padrão dicotômico ainda é visto, no entanto em menor grau. Como exemplos, o motivo 2 é encontrado em 86% das 29 seqüências que contém o motivo 9, que não está presente praticamente com os outros motivos; o motivo 7 ocorre em 80% das 10 seqüências que contém o motivo 8; esse motivo não ocorre conjuntamente com os motivos 9 e 17, e raramente com o motivo 2. O motivo 17 ocorre com o motivo 2 em 59% de suas 17 seqüências; esse motivo não ocorre praticamente com as outras seqüências, então em grande proporção ocorre sozinho.

TABELA 23 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiMet janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	2	7	8	9	10	14
7	11	-					
8	10		.				
9	29	-					
10	10	.	.	.	.		
14	5	.	.	.	.	.	
17	17	-		.	-	.	

5.3.2.6 Motivos de EpiAma

5.3.2.6.1 Motivos de EpiAma janela 6-15

Na categoria EpiAma foram identificados 3 motivos enriquecidos, somente

para as comparações com os controles geral e *housekeeping* (TABELA 24). O motivo 2 (FIGURA 65) foi identificado com e-value de $1x10^{-62}$ em 40 das 55 sequências dessa categoria (73%). Já o motivo 3 (FIGURA 66) foi identificado em 42 sequências, com e-value de $5x10^{-33}$ (76%).

TABELA 24 – Motivos identificados em EpiAma com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	TTTTTT[TC]TTTTTTTT	5.30E-01	8.64E-01	4.90E-02
2	A[AG]AA[ACG][AG][AG][AC]AA[AG]AA[AC]A	1.94E-07	5.20E-06	2.51E-05
3	G[AG]G[GA]G[AT]G[AG][GAT][AG][GA][AG][GA][AG][GA]	8.48E-07	2.15E-06	4.91E-04
4	[GT]TGTGTGTGTG[TC]GTG	8.92E-01	1.00E+00	6.48E-01
5	TT[GCT]TT[TG]TT[GTC]TT[GT]TT[TA]	8.74E-01	1.00E+00	7.78E-01
6	[TC]ATA[TC]AT[AT]TA[TC][AT][TC][AT]T	1.00E-06	2.87E-09	1.00E+00
7	A[ACG]A[CA][GA][GA][AG][AG][AG][GA][AG]AA[GC][AG]	3.67E-05	1.57E-03	1.21E-04
8	[TC]T[TC]T[TC][TC][CT][TC][TC][TC]C[TA][TC][TC]C	2.52E-01	1.79E-01	3.14E-02

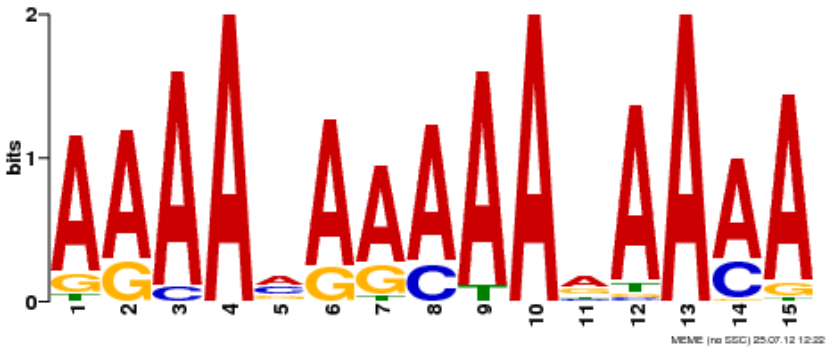


FIGURA 65 – Logo do motivo 2 de EpiAma.

FIGURA 66 – Logo do motive 3 de EpiAma

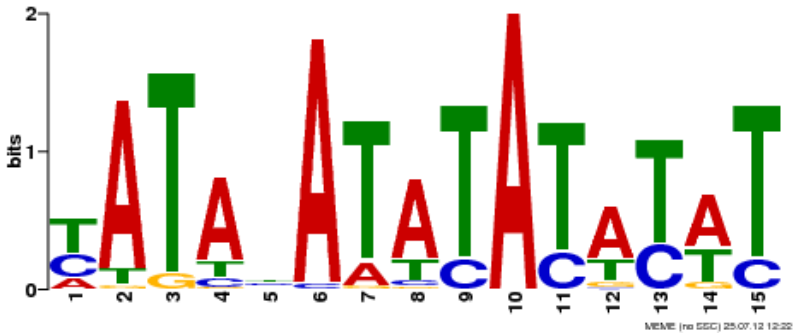


FIGURA 67 – Logo do motivo 6 de EpiAma.

Em relação à co-ocorrência dos motivos, os motivos 2 e 3 ocorrem em 88% das 40 seqüências que contém o motivo 2. O motivo 6 co-ocorre em média em 62% das suas 35 seqüências com os motivos 2 e 3. Portanto, a co-ocorrência dos motivos nessa categoria é alta (TABELA 25).

TABELA 25 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiAma janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	2	3
3	42	.	
6	35	.	.

5.3.2.6.2 Motivos de EpiAma janela 16-25

Utilizando a janela maior na comparação EpiAma, são identificados dois motivos, significativos somente para as comparações contra os controles geral e *housekeeping,* de forma idêntica ao encontrado para a janela menor (TABELA 26).

No entanto, a composição desses motivos é mais complexa do que a dos motivos menores (FIGURA 68, FIGURA 69), o que é raro: geralmente, os motivos maiores, por possuírem maior número de posições, tendem a ser significativos mesmo quando a composição é menos complexa.

TABELA 26 – Motivos identificados em EpiAma com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	7.54E-01	8.62E-01	1.24E-01
2	3.31E-03	2.19E-02	5.96E-03
3	5.41E-04	6.63E-03	5.48E-04
4	2.64E-01	3.10E-01	1.00E+00
5	1.78E-01	5.06E-02	3.66E-01
6	1.00E+00	6.70E-01	3.07E-02
7	9.02E-05	2.79E-03	1.37E-03
8	7.06E-14	4.02E-09	4.61E-02
9	3.55E-21	1.60E-13	3.15E-03

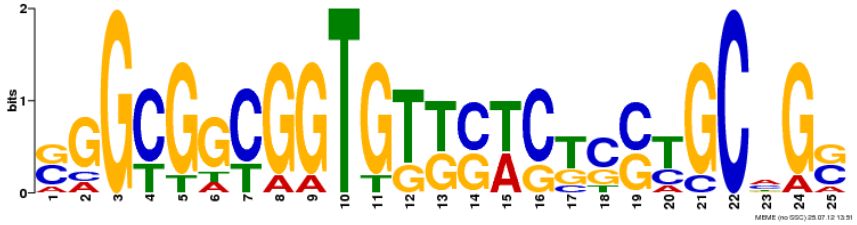


FIGURA 68 – Logo do motivo 8 de EpiAma.

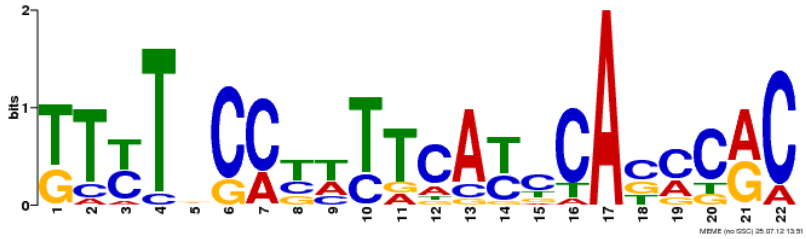


FIGURA 69 – Logo do motivo 9 de EpiAma.

A co-ocorrência entre esses dois motivos é intermediária, isto é, 43% das sete sequências que contém o motivo 8 também contém o motivo 9 (TABELA 27).

TABELA 27 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiAma janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

Motivo	N	8
9	13	-

5.3.2.7 Motivos de MetTrp

5.3.2.7.1 Motivos de MetTrp janela 6-15

Em MetTrp, foram identificados 4 motivos (TABELA 28), mas nenhum deles é significativo o suficiente em relação aos controles para ser considerado um elemento regulatório específico confiável.

TABELA 28 – Motivos identificados em MetTrp com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	A[TC][AG][TC][AC][TC]ATA[CT]A[CT]A[TC]A	5.82E-05	3.71E-05	1.21E-03
2	AA[AC]AA[GA][AG]AAAA[AG]AA[AC]	5.79E-01	3.45E-01	2.40E-02
3	G[GT]G[AG][GA]G[GA]AG[AG][GA][AG][GA][AG]G	1.34E-01	1.27E-01	6.54E-01
4	TTTTTTT[CTG]T[TG]TTTTT	2.16E-03	2.97E-03	6.55E-03

5.3.2.7.2 Motivos de MetTrp janela 16-25

Em MetTrp foi identificado somente um motivo enriquecido, somente contra a categoria Housekeeping e com um valor de p muito próximo ao limite do critério de significância mínima. Esse motivo é relativamente simples (FIGURA 70).

A incapacidade de se identificar elementos regulatórios específicos em MetTrp, as duas formas infectivas principais de *T. cruzi,* pode parecer a princípio algo contra-intuitivo. No entanto, o número de genes candidatos nessa categoria é pequeno (n=30), o que dificulta a identificação de elementos regulatórios significativos. Na verdade, a maioria dos genes que são aumentados em MetTrp também o são em Ama (categoria MetAmaTrp, n=334, a qual será analisada subsequentemente).

5.3.2.8 Motivos de AmaTrp

5.3.2.8.1 Motivos de AmaTrp janela 6-15

Em AmaTrp foram encontrados 39 motivos (TABELA 30). A maioria desses motivos é significativamente enriquecida em relação aos 3 controles, da mesma forma do que foi observado para a categoria Trp. Essa categoria também está enriquecida para trans-sialidases e mucinas.

TABELA 29 – Motivos identificados em MetTrp com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	1.46E-05	8.53E-06	1.46E-03
2	3.51E-01	2.53E-01	2.86E-03
3	2.02E-02	2.77E-02	2.50E-01
4	8.72E-02	8.00E-02	8.21E-01
5	1.00E+00	8.49E-01	3.70E-01

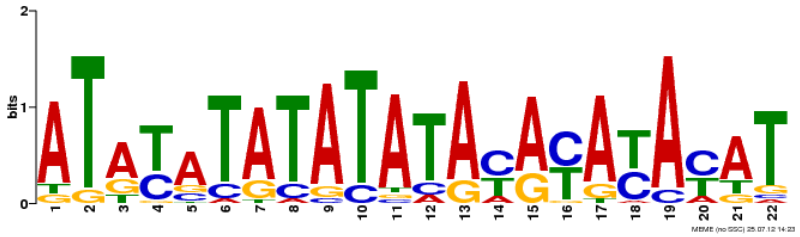


FIGURA 70 – Logo do motivo 1 de MetTrp.

Em relação à co-ocorrência desses motivos (TABELA 31), o padrão também é muito semelhante ao de Trp. Os motivos 1 a 15 co-ocorrem praticamente totalmente entre si; o motivo 12 praticamente co-ocorre com todos os motivos significativos (n=35). Os demais motivos co-ocorrem com padrões bem mais específicos, com exceção do motivo 25.

5.3.2.8.2 Motivos de AmaTrp janela 16-25

Da mesma forma que para a janela menor, a grande maioria dos 32 motivos identificados em AmaTrp com a janela maior são significativos (TABELA 32). A

análise de sua co-ocorrência apresenta padrões similares (TABELA 33).

TABELA 30 – Motivos identificados em AmaTrp com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	GCACCGCCC[GA]CACAC	4.85E-73	2.40E-72	6.36E-43
2	GGCGGAGCACCGTAA	3.47E-52	3.06E-48	5.82E-25
3	ATGCACACCCATGCA	1.28E-49	8.02E-50	8.07E-26
4	ACACGCACCGCTGCC	1.60E-72	5.06E-63	4.22E-33
5	GAG[GT]G[AT]GAGAGAGCC	3.16E-18	7.84E-21	1.47E-10
6	CA[CA]GA[AG]GCA[CA]ACA[CA]A	2.81E-08	9.66E-06	4.83E-04
7	GCA[TA][GT][GC]ACTCT[CG]G[CA]C	8.62E-93	4.28E-83	1.28E-44
8	CTGCCTGGCTGCATG	3.97E-81	3.68E-81	1.75E-43
9	AATGTGCCCATTGTA	1.89E-48	6.76E-46	4.51E-22
10	GCCCGTTGGTGCTCT	9.90E-55	2.04E-51	1.09E-26
11	[TG][CG]TGTGTGTGTCC[CG]C	5.28E-21	5.86E-20	3.92E-16
12	TTT[CT]T[GT]TTTT[GT]TTTT	1.72E-05	1.35E-05	2.16E-03
13	[AG]CACTC[GT]TCCT[TG]C[TGA][CA]	3.50E-81	8.36E-71	1.31E-38
14	TGAGTGGGCGACCTC	1.11E-38	2.69E-38	1.24E-19
15	CTCTCCCT[GC]TGTGTG	2.51E-25	1.25E-18	8.78E-34
16	CTCACCTCA[CT]C	2.76E-55	7.90E-49	2.98E-26
17	TTTTT[TC]TTT[TG][CT]T[TA][TC]T	1.87E-04	1.52E-04	6.74E-03
18	A[AT]AAA[AG][AT]AAGA[AG][AT]AA	1.15E-06	1.29E-06	1.94E-05
19	CTTTTGCA	2.38E-123	1.86E-47	1.55E-24
20	GGGACA[CA]TTGCGACC	2.31E-17	5.95E-14	4.96E-07
21	[TC][TC][TGC][TC]TTTAATTGTTT	9.35E-06	2.07E-05	1.62E-04
22	ATCCG[TC]CACGTTGAA	3.91E-09	1.02E-10	2.11E-05
23	T[GT]TTTTGCTTT	4.50E-07	1.10E-07	1.00E-05
24	AA[ATG][AG][AC][AG]A[GC]AA[AC]AA[GCA]A	5.09E-08	7.99E-09	1.32E-05
25	[TC]ATTGCAT	1.44E-87	3.65E-33	6.34E-17
26	AGCCACGGGGACATG	3.01E-05	6.54E-07	1.71E-03
27	CTCCGCGTTG[CT][TG][TC]CC	9.45E-11	1.02E-10	2.11E-05
28	C[CG]C[TA]GCA[CG][AC]GA[AC][TA][TC]A	1.24E-46	3.60E-28	2.57E-14
29	TTT[TG][GCT]TG[TC]G[CT]GT[TG]T[TG]	4.76E-09	4.46E-07	1.13E-11
30	CCACC[GA][AC]CACGCTCA	2.81E-11	3.59E-08	4.00E-04
31	[AT]G[GT][AG]GG[GA][GA][AG][AG]GG[GA]A[ACG]	1.05E-04	2.21E-05	3.79E-02
32	AGAGGTGTGTG	6.86E-09	1.43E-07	2.67E-02
33	ATATA[TA]ATAT[AT][TA]A[TA]A	6.48E-09	9.22E-08	1.68E-05
34	A[CA]T[GC]TG[GA]TCCGATGT	1.17E-05	1.54E-07	8.27E-04
35	AAT[GAT]AATTT[CA]T[CT]T[CT][CA]	9.06E-24	1.32E-14	2.32E-07
36	ATATGC	1.00E+00	1.00E+00	1.00E+00
37	AAAA[AG]AA[GAT][AG]A	4.04E-09	1.33E-08	1.78E-06
38	A[AC][CT]T[GA]TTTGC[GA]TGGA	2.59E-06	2.77E-06	3.51E-03
39	C[TA][AC]ATCCC[CT]CA[TC]CCC	3.85E-63	1.01E-31	3.61E-16

TABELA 31 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em AmaTrp janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocorrência.

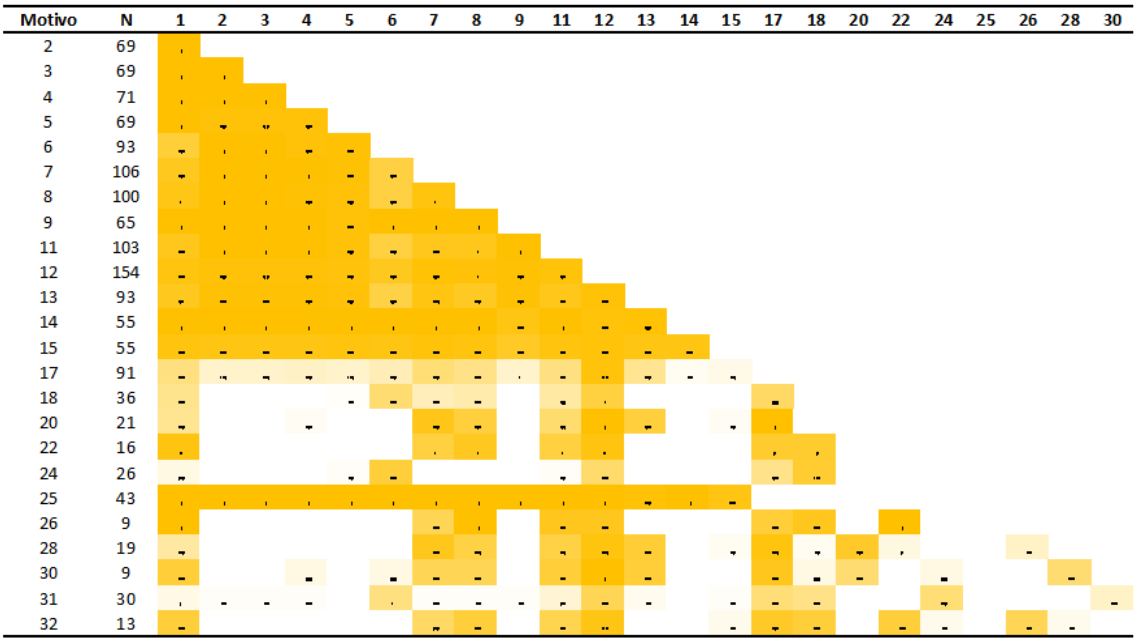


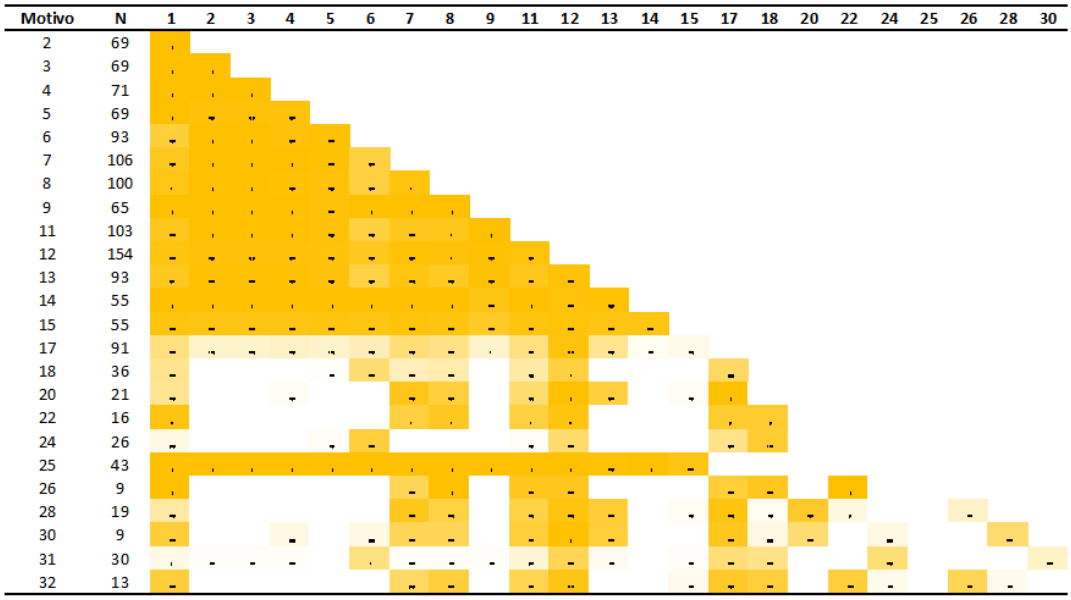
TABELA 32 – Motivos identificados em AmaTrp com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	1.72E-64	5.96E-58	2.81E-30
2	1.99E-54	8.02E-50	8.07E-26
3	3.68E-61	1.80E-55	6.56E-29
4	5.21E-62	1.80E-55	6.56E-29
5	1.52E-59	1.18E-54	1.85E-28
6	4.30E-60	1.18E-54	1.85E-28
7	1.35E-45	4.03E-45	2.81E-23
8	2.77E-37	1.48E-37	3.07E-19
9	7.73E-14	4.84E-13	2.79E-10
10	8.19E-05	5.72E-05	5.88E-03
11	2.14E-10	1.49E-08	2.69E-11
12	1.22E-21	2.23E-25	7.19E-13
13	8.12E-06	3.34E-06	3.33E-04
14	3.11E-21	3.06E-17	1.61E-07
15	3.29E-76	1.21E-73	6.36E-43
16	1.13E-03	1.57E-03	3.70E-02
17	2.94E-09	5.26E-12	4.76E-06
18	3.76E-17	5.95E-14	4.96E-07
19	1.43E-02	2.98E-02	2.72E-03
20	9.31E-22	1.37E-20	2.09E-10
21	1.39E-03	1.13E-02	9.34E-03
22	1.11E-72	2.08E-62	9.64E-34
23	2.44E-01	1.57E-01	6.65E-01
24	2.72E-07	3.59E-08	4.00E-04
25	8.35E-13	3.32E-14	2.77E-08
26	3.87E-25	5.46E-13	1.08E-07
27	5.66E-05	6.41E-03	1.47E-02
28	2.18E-09	1.04E-10	4.85E-08
29	3.93E-01	4.87E-01	3.79E-01
30	7.38E-10	5.70E-08	5.11E-08
31	3.69E-09	4.44E-10	4.43E-05
32	3.27E-05	6.54E-07	1.71E-03

TABELA 33 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em AmaTrp janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocurrência.



5.3.2.9 Motivos de EpiMetAma

5.3.2.9.1 Motivos de EpiMetAma janela 6-15

Nessa categoria foram identificados 8 motivos (TABELA 34), sendo que 5 são enriquecidos significativamente. O motivo 7 (FIGURA 74) apresentou significância contra os três controles; pelo fato de que os demais motivos significativos foram identificados somente na comparação contra o controle inverso, é necessário ter cautela nas comparações dessa categoria, mesmo com valores de p extremamente significativos, pois a categoria inversa é justamente a Trp, que contém um grande número de famílias multi-gênicas e, portanto, a sua heterogeneidade da composição das regiões 3'-UTR é baixa.

Na TABELA 35, vemos a análise de co-ocorrência de motivos; nela podemos observar que o motivo 7 tem uma co-ocorrência significativa com os outros motivos, que foram selecionados somente na comparação contra Trp. Isso não é de todo inesperado, pois os outros motivos são relativamente comuns no genoma de *T. cruzi,* e portanto também devem ocorrer nas sequencias que contém o motivo 7.

TABELA 34 – Motivos identificados em EpiMetAma com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	TTTTTTTTTTTTTTT	2.02E-02	9.86E-03	1.00E+00
2	AAA[AG][AG]AAAAAA[AC]AAA	2.57E-04	4.91E-03	1.75E-30
3	T[GT]T[GT][TC][GT]TGTG[TC]GTG	8.02E-01	6.93E-01	5.86E-05
4	A[TC]A[TC]ATA[TC]A[TC]A[TC]A[TC]A	9.61E-03	1.09E-02	1.19E-18
5	G[AG][GA]G[GA][GA][GA]A[GA][GA]G[GA][GA][AG]	4.96E-04	1.26E-03	3.91E-06
6	[TC]TTTTTT[TC]T[TG]T[CT]TT[TC]	2.28E-01	3.44E-01	1.15E-02
7	C[TC]CC[CA][CA][CTG]C[CA][CT][TC]	4.97E-09	9.57E-07	9.50E-13
8	[AC]A[CAG]AA[AG][CA][AG][AC]A[AC][AT][AG][AC]A	9.75E-05	7.09E-03	4.23E-34

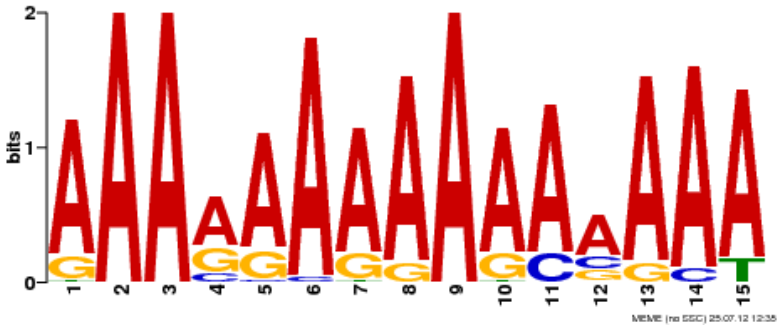


FIGURA 71 – Logo do motivo 2 de EpiMetAma.

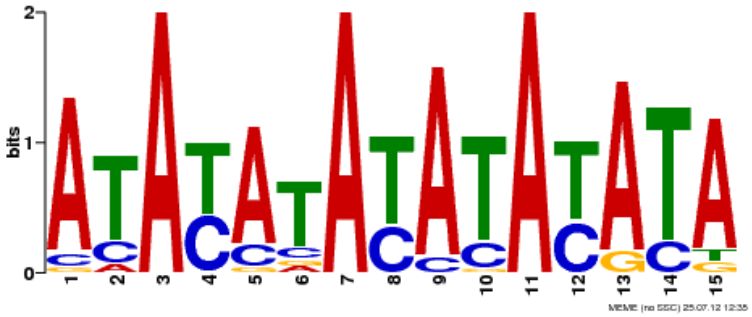


FIGURA 72 – Logo do motivo 4 de EpiMetAma.

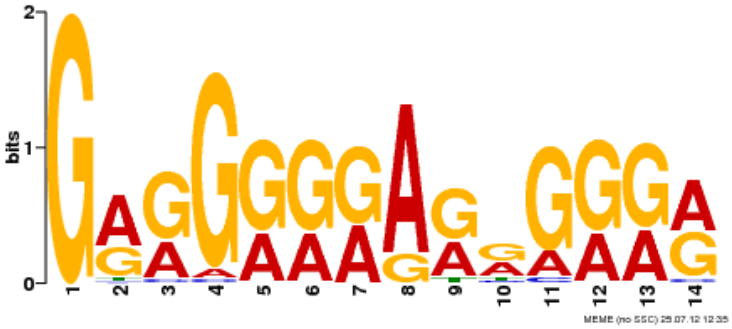


FIGURA 73 – Logo do motivo 5 de EpiMetAma.

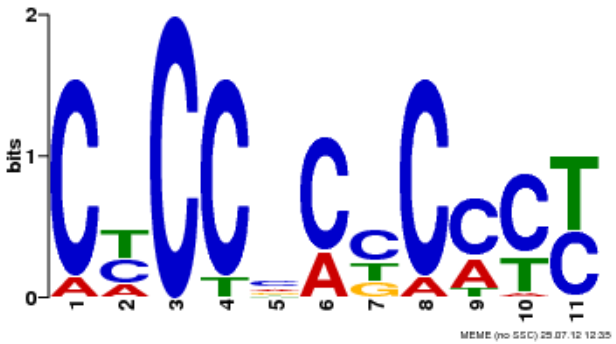


FIGURA 74 – Logo do motivo 7 de EpiMetAma.

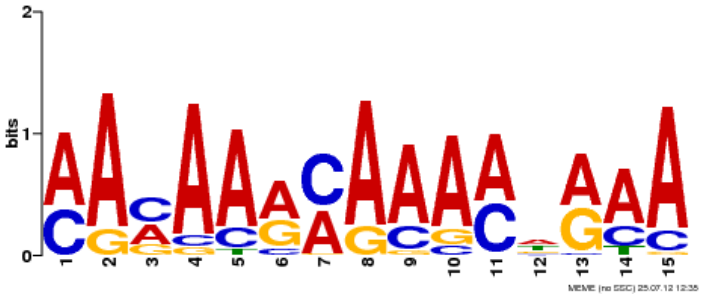


FIGURA 75 – Logo do motivo 8 de EpiMetAma.

TABELA 35 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiMetAma janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocurrência.

Motivo	N	2	4	5	7
4	25	.			
5	41	.	.		
7	21	-	-	-	
8	35	-	.	-	-

Na FIGURA 76, a via da glicólise e gliconeogênese, que apresentou uma maior quantidade de genes modulados em Epi está ilustrada, mostrando a ocorrência dos genes com os motivo 7.

5.3.2.9.2 Motivos de EpiMetAma janela 16-25

Da mesma forma que para a janela menor, foram identificados 8 motivos (TABELA 36), sendo que 4 são enriquecidos significativamente. O motivo 8 (FIGURA 80) foi o único que apresentou significância contra os três controles; a mesma

observação realizada para os motivos com a janela menor é válida para essa análise. O motivo 8 é o mais complexo dos 4 motivos enriquecidos, o que justifica a sua identificação significativa na comparação com os três controles.

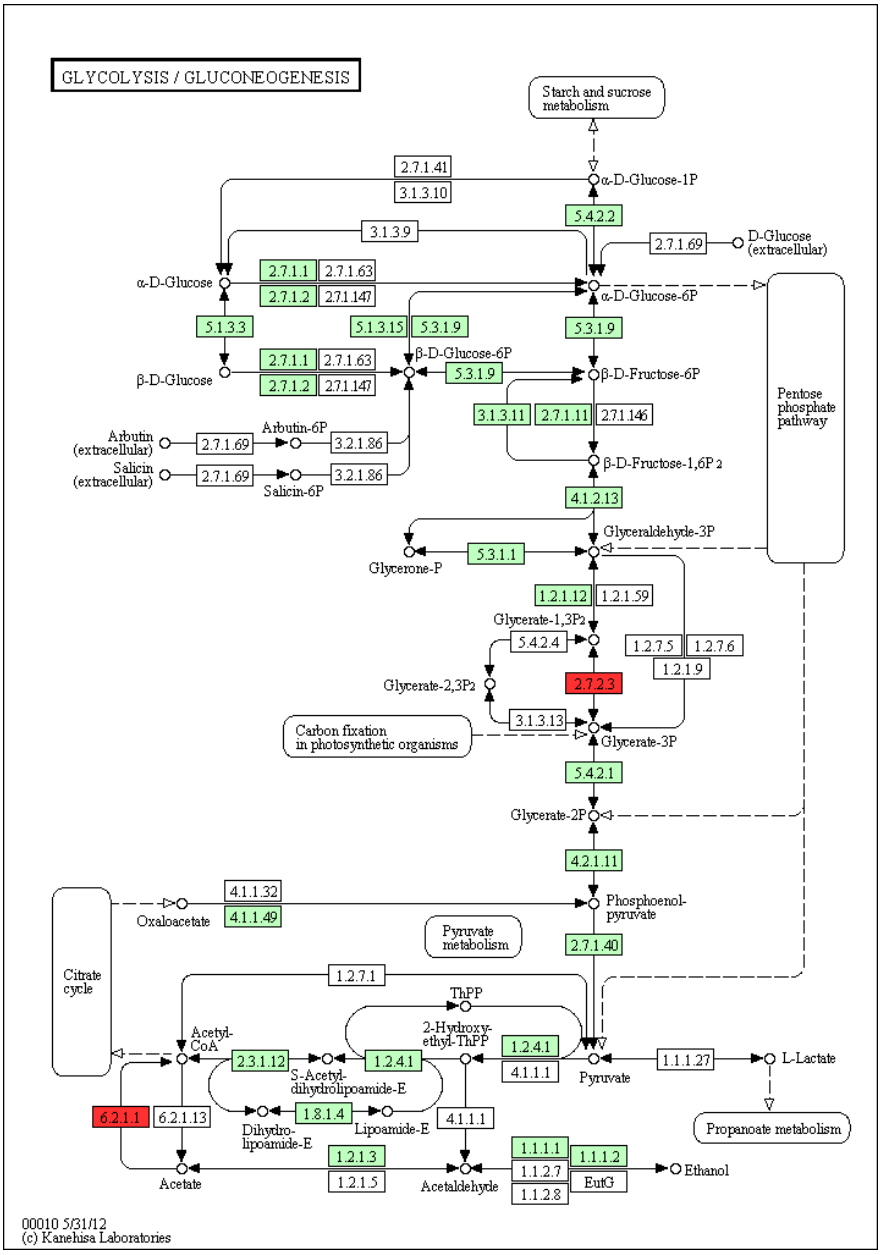


FIGURA 76 - Representação do mapa do KEGG da via metabólica da glicólise e gliconeogênese, com genes modulados em EpiMetAma contendo os motivos mais enriquecidos

Os genes com o motivo 7 estão marcado em vermelho. Os demais genes de *T. cruzi* que estão nessa via estão marcados em verde.

Na comparação de co-ocorrência (TABELA 37), podemos verificar que os motivo 8 co-ocorre frequentemente com os motivos 2 e 5 (63%), mas não com o

motivo 7 (33%).

TABELA 36 – Motivos identificados em EpiMetAma com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	3.27E-05	8.66E-01	2.18E-01
2	8.81E-04	1.16E-02	2.15E-32
3	4.08E-01	1.78E-01	1.29E-03
4	8.98E-01	8.94E-01	3.12E-03
5	2.09E-05	2.05E-05	4.87E-41
6	6.58E-01	6.40E-01	3.96E-02
7	7.60E-05	1.19E-03	1.23E-34
8	6.94E-15	3.48E-10	9.25E-07

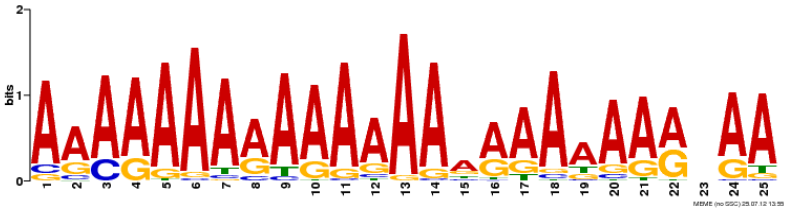


FIGURA 77 – Logo do motivo 2 de EpiMetAma.

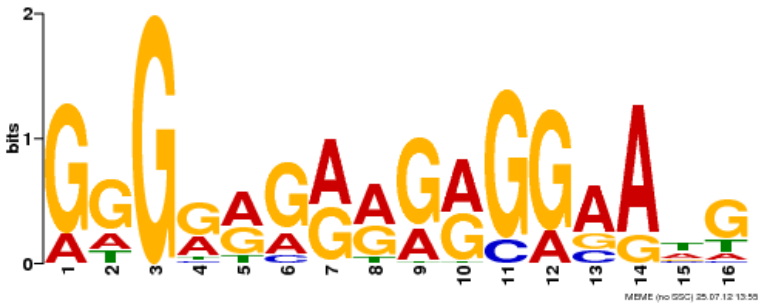


FIGURA 78 – Logo do motivo 5 de EpiMetAma.

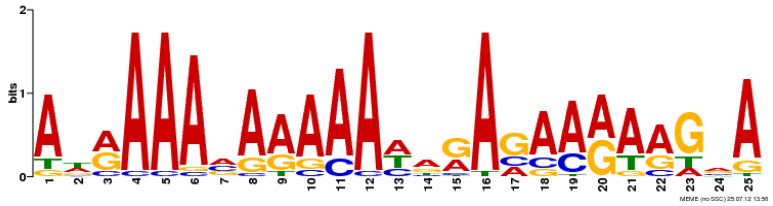


FIGURA 79 – Logo do motivo 7 de EpiMetAma.

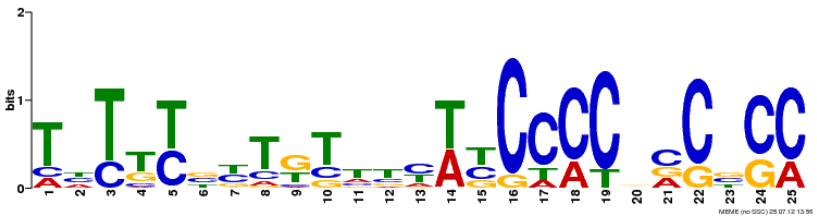


FIGURA 80 – Logo do motivo 8 de EpiMetAma.

TABELA 37 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiMetAma janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocurrência.

Motivo	N	2	5	7
5	41			
7	22			
8	18			

5.3.2.10 Motivos de EpiMetTrp

5.3.2.10.1 Motivos de EpiMetTrp janela 6-15

EpiMetTrp apresentaram 5 motivos identificados (TABELA 38). O motivo 5 (FIGURA 81) foi identificado em 14 das 46 sequências analisadas (30%) para busca pelo motivo, com e-value de $2x10^{-5}$.

TABELA 38 – Valor Motivos identificados em EpiMetTrp com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	TT[TC]TT[TC]TTTTT[TC]TTT	1.00E+00	1.00E+00	4.93E-01
2	[AG]AA[AG]AA[AG]AA[ACG]AAAAA	6.65E-03	4.61E-02	1.00E+00
3	G[AT][GA][AG][GA][AG]GA[AG][GA][GA][AG][GA][AG][GA]	4.25E-04	1.76E-03	3.20E-02
4	G[TC][GA]TGTG[TC]GT[GCA][TG]G[TC]G	3.48E-01	2.48E-01	4.18E-01
5	[CG][CT][GC][CT][TCA][CT][CG]C[TA]C[CT][CG]C[CT]	4.44E-14	2.78E-09	8.50E-03

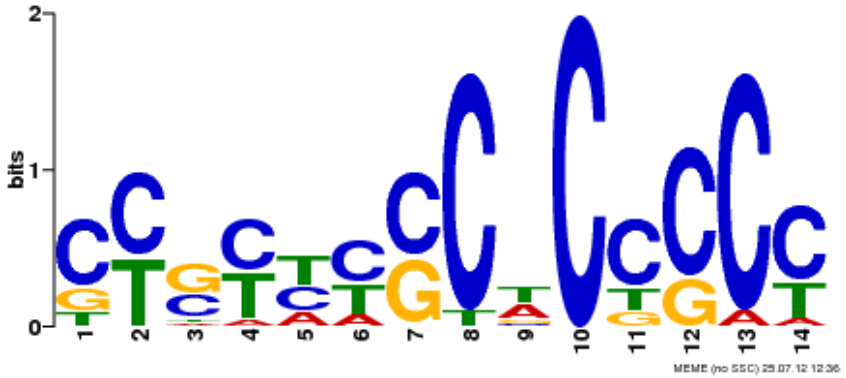


FIGURA 81 – Logo do motivo 5 de EpiMetTrp.

5.3.2.10.2 Motivos de EpiMetTrp janela 16-25

Nessa comparação, também só foi identificado um motivo significativo (TABELA 39), o qual é relativamente semelhante ao motivo obtido com janela menor (FIGURA 82).

TABELA 39 –Motivos identificados em EpiMetTrp com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	2.34E-03	1.00E+00	2.69E-01
2	5.05E-04	1.29E-02	3.05E-01
3	1.63E-02	9.01E-02	3.26E-01
4	2.07E-01	1.90E-01	6.05E-02
5	1.02E-15	2.71E-14	1.23E-05

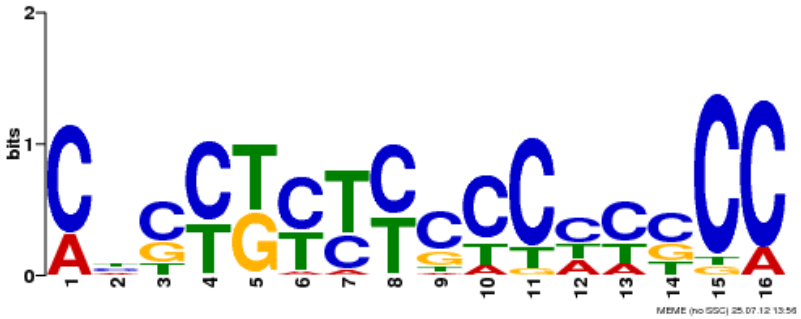


FIGURA 82 – Logo do motivo 5 de EpiMetTrp.

5.3.2.11 Motivos de EpiAmaTrp

5.3.2.11.1 Motivos de EpiAmaTrp janela 6-15

Na categoria EpiAmaTrp, foram identificados 12 motivos (TABELA 40). Desses, apenas o motivo 12 (FIGURA 84) foi identificado como enriquecido em duas comparações contra os controles. Ele apresenta um padrão mais complexo do que o motivo 10 (FIGURA 83).

A co-ocorrência desses motivos é baixa (20%, TABELA 41).

TABELA 40 –Motivos identificados em EpiAmaTrp com janela 6-15 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	[TC]TTTTTTTTTTTTTT	3.08E-01	3.57E-01	3.00E-01
2	AAAAA[AG]AAAAAA[AG]AA	2.36E-02	4.46E-01	9.15E-01
3	[TC]TTTT[TC]T[TC][TC][CT][TC][TC]T[TC]T	1.32E-01	2.07E-01	7.55E-02
4	[GA]AG[AG][AG][GA][GA][AG][GA][AG][GA][GA]GA[AG]	2.14E-02	7.14E-02	1.35E-01
5	[GC]TGTGTG[TC][GA]TG[TA]GTG	3.83E-01	1.00E+00	4.21E-01
6	[AC][CA]A[CAT]A[CA]A[CTA][AG][CA]A[CA][AG][CA]A	1.40E-03	2.09E-01	2.80E-01
7	TT[GA]TT[AGT]TT[TGA]TT[AGT]TT	7.66E-01	1.00E+00	8.02E-01
8	GA[AG][GA][GA]A[AG][AG]A[AGC]A	6.11E-04	1.51E-02	1.64E-01
9	CC[CTG][CTA]C[CA]C[TC]C[CT]C	1.48E-04	1.93E-02	8.22E-03
10	TAT[AT]TATATA[TC]ATAT	1.15E-02	2.88E-01	2.12E-09
11	[AG]AAAA[ATC][AC][AG]A[AC]AA[CGTAA]AA	1.34E-01	4.98E-01	3.36E-01
12	GG[TCG]GG[AGT]G[GA][GT]G[GA][CG]A[CG][GA]	8.22E-07	4.27E-03	5.08E-06

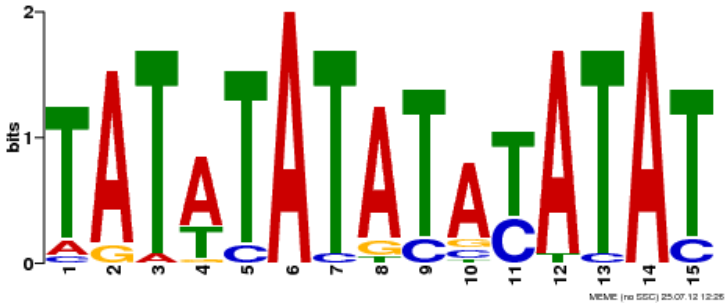


FIGURA 83 – Logo do motivo 10 de EpiAmaTrp.

FIGURA 84 – Logo do motivo 12 de EpiAmaTrp.

TABELA 41 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em EpiAmaTrp janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocurrência.

Motivo	N	10
12	34	.

5.3.2.11.2 Motivos de EpiAmaTrp janela 16-25

Nessa comparação, somente um motivo foi identificado como significativo; no entanto, ele apresentou significância contra os três controles, o que não aconteceu com os motivos identificados com a janela menor; além disso, sua composição é distinta dos dois motivos encontrados anteriormente.

TABELA 42 – Motivos identificados em EpiAmaTrp com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre $1x10^{-5}$ e $1x10^{-10}$; em azul numa tonalidade intermediária valor p entre $1x10^{-10}$ e $1x10^{-20}$; e em azul mais escuro valor p < $1x10^{-20}$.

Motivo	Geral	Housekeeping	Inverso
1	3.21E-03	1.06E-01	9.03E-02
2	6.32E-03	1.55E-01	5.91E-01
3	9.76E-04	1.56E-01	3.38E-01
4	9.38E-02	7.22E-01	5.10E-01
5	7.91E-01	7.68E-01	1.67E-01
6	5.85E-03	1.29E-01	8.93E-02
7	5.17E-15	7.47E-11	1.08E-06
8	7.37E-02	8.33E-01	2.13E-05

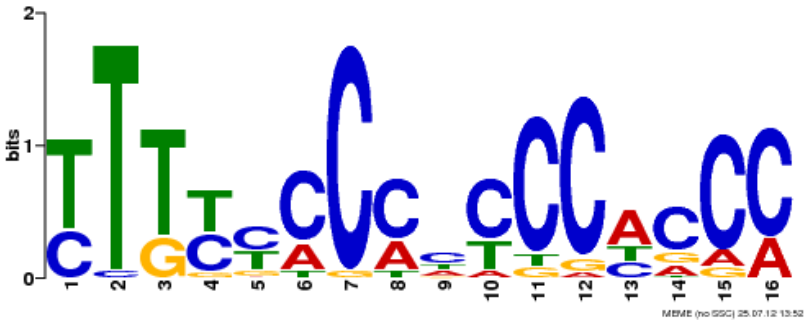


FIGURA 85 – Logo do motivo 7 de EpiAmaTrp.

5.3.2.12 Motivos de MetAmaTrp

5.3.2.12.1 Motivos de MetAmaTrp janela 6-15

Nessa categoria foram identificados 42 motivos (TABELA 43), sendo que 30 foram considerados enriquecidos (71%). Existe uma grande quantidade de proteínas de superfície, especialmente MASPs, o que pode explicar a identificação dessa

grande proporção de motivos. No entanto, diferentemente das outras duas análises com Trp que também apresentavam essas categorias de proteínas (Trp e TrpAma), nessa análise há uma maior proporção de proteínas com outras funções (n=184, de um total de 334, 55%), o que reforça a importância de se estabelecer formas diferentes de análise em relação às famílias multigênicas.

TABELA 43 – Motivos identificados em MetAmaTrp com janela 6-15 com seleção dos enriquecidos.

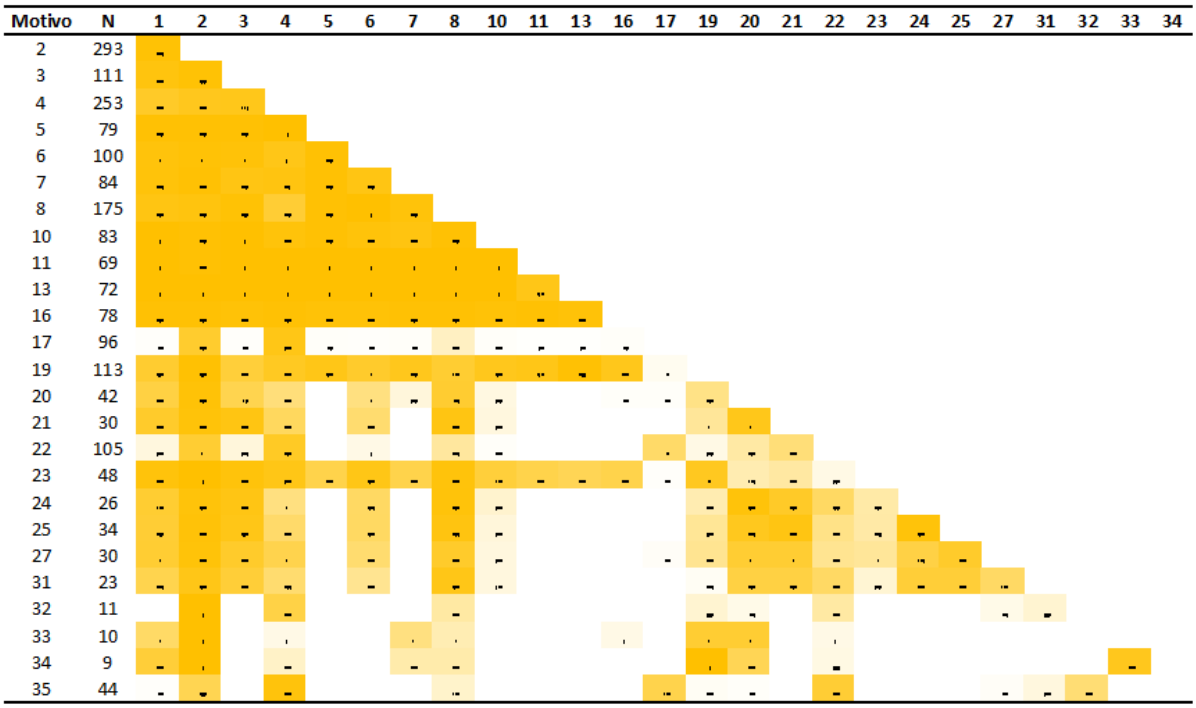
Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre 1x10-5 e 1x10-10; em azul numa tonalidade intermediária valor p entre 1x10-10 e 1x10-20; e em azul mais escuro valor p < 1x10-20

Motivo	Expressão Regular	Geral	Housekeeping	Inverso
1	GCACCGCCCGCACAC	6.83E-64	5.02E-62	3.72E-38
2	TTT[TC]T[TG]TTTT[TG]TTTT	2.53E-07	8.28E-06	2.75E-03
3	CCGCTGCCGT[CG]CTG[CT]	9.71E-65	2.80E-57	1.38E-36
4	[GA]A[AG][GA][AC]A[CA]A[CA]A[AC]A[AT]AA	7.84E-06	3.70E-03	9.99E-05
5	CACACCCATGCAGGC	1.54E-49	5.82E-44	3.79E-29
6	CATG[GC]GCG[GC][AT]G[CG]ACC	4.20E-60	9.45E-54	3.78E-33
7	ACTCTCGCCCTCACC	1.53E-46	2.99E-44	4.47E-27
8	CT[CG][TG][CG]TGTG[TC]GTG[TC][CG]	5.97E-17	4.10E-13	4.43E-18
9	GA[GA][TAG]GA[GA][GT][GA]A[GA]AGA[GA]	2.56E-02	3.22E-01	2.39E-01
10	TTTGCAC[GC]ACACGCA	8.76E-53	7.60E-47	8.25E-29
11	TGAGTGGGCGACCTC	5.21E-39	1.10E-34	1.38E-19
12	CCGTTGGTGCTCTCT	3.08E-51	2.28E-47	3.79E-29
13	GTGCCCATTGTATAT	2.66E-36	5.73E-33	2.82E-21
14	CTCTCTC[CT]CT[GC]TGTG	1.53E-10	1.45E-08	7.18E-09
15	CCCTTGCATGG	2.92E-38	1.14E-35	6.68E-22
16	ACACTC[GT]TCCT	9.04E-46	1.04E-42	5.54E-26
17	ATA[TA]A[TA]A[TA]A[TA]A[TA]A[TA]A	3.60E-02	3.41E-02	8.58E-01
18	TTTTTT[TA]TT[TGA][TC]TTTT	1.31E-01	1.91E-01	8.29E-01
19	CACCC[CA]AC	2.04E-95	2.42E-29	5.22E-18
20	[TG]C[CT]G[TC]C[AG]CGT[TG]G[AG]AC	2.94E-19	1.32E-17	1.36E-11
21	GTGCTCCGCGTTG[TC][GT]	2.95E-18	6.89E-16	8.56E-10
22	A[AT][AG]AA[GA][ATG]AA[GA][AG][AG][ATC]AA	4.89E-03	1.99E-03	9.89E-02
23	TGTGTTTTGCT	2.10E-13	3.64E-11	4.24E-08
24	CGTC[TC]ACTGTGGTCC	3.23E-13	6.97E-12	3.52E-07
25	[AG]CGGGGA[CT][TA]TGTG[TC][GA]	3.48E-18	2.16E-15	2.35E-10
26	C[AC]C[AGC][GC]ACAC[AG]C[TA]CA[TC]	7.04E-36	9.32E-25	4.33E-21
27	AGA[GA]GTGT[GA]TG	2.87E-21	1.52E-14	6.09E-09
28	TTT[TC]T[TC]TTT[TC]T[TA][TC][TC]T	5.94E-04	1.29E-02	4.17E-03
29	[TC]ATTGCAT	6.04E-97	8.02E-30	2.61E-18
30	[AG]ACTGTTTGCGTGGA	3.20E-09	5.92E-08	6.08E-05
31	AGCACTCA[AG][CA][CG]	1.28E-22	6.89E-16	8.56E-10
32	CACGCGGTGCCGGCC	1.82E-08	2.25E-05	3.21E-03
33	GGGCAATCACT[AG]TGG	9.47E-04	6.01E-05	3.21E-03
34	GGGTGCCGTGTGTTT	4.23E-05	2.25E-05	3.21E-03
35	A[ACG]G[GA][AG][AC][TAG][GA][ACG]A[AC][GA][AG]A[GA]	1.07E-03	3.35E-04	3.22E-02
36	TGCATG	1.00E+00	1.00E+00	1.00E+00
37	[TC]AT[ATG]TAT[TA]TAT[ATG][TA]AT	6.70E-02	7.58E-01	3.25E-01
38	[CA][AC][TA][AT]C[TG]GTGCAGCT[CT]	1.85E-11	4.32E-07	2.24E-04
39	CC[CT]T[CT]TTTAATTGTT	6.09E-01	6.20E-01	7.36E-01
40	TGGCTG	1.00E+00	1.00E+00	1.00E+00
41	CCTTGCGGGGACGGC	8.77E-06	3.00E-03	3.78E-02
42	CAC[TA][GT]TTGAG[GA]AT[GA]G	1.21E-06	6.01E-05	3.21E-03

Da mesma forma que para as demais análises de identificação de possíveis motivos regulatórios que continham uma grande proporção de genes de famílias multigênicas, existe um grande grau de co-ocorrência de motivos, especialmente para aqueles mais frequentes (TABELA 44). O motivo 2 co-ocorre com todos os outros motivos com uma frequência significativa (em sua grande maioria, frequência maior do que 80%). Alguns motivos mostram co-ocorrência bem mais restritua, como por exemplo o motivo 17, que praticamente só co-ocorre com os motivos 2 e 4.

TABELA 44 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em MetAmaTrp janela 6-15

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocurrência.



5.3.2.12.2 Motivos de MetAmaTrp janela 16-25.

Nessa categoria foram identificados 35 motivos (TABELA 45), sendo que 26 foram considerados enriquecidos (74%), proporção semelhante à da análise com janela menor.

O padrão de co-ocorrência é bastante similar ao observado com a janela menor (TABELA 46). Os motivos 1 a 11 co-ocorrem frequentemente entre si (geralmente, frequência maior do que 80%); o motivo 7 co-ocorre frequentemente

com todos os outros motivos, com a menor frequência com o motivo 22 (59%); o motivo 1 também ocorre frequentemente com a maioria dos outros motivos (freqüência menor com o motivo 33, 45%), com exceção dos motivos 24, 32, 22, 31 e 35, sendo que para os três últimos sua co-ocorrência é nula. Para diversos outros motivos, o padrão de co-ocorrência é muito mais limitado e, para o motivo 22, a sua ocorrência com outro motivo é inexistente, com exceção do motivo 7, já citado, cuja ocorrência é 59%; em outras palavras, esse motivo ocorre muitas vezes sozinho.

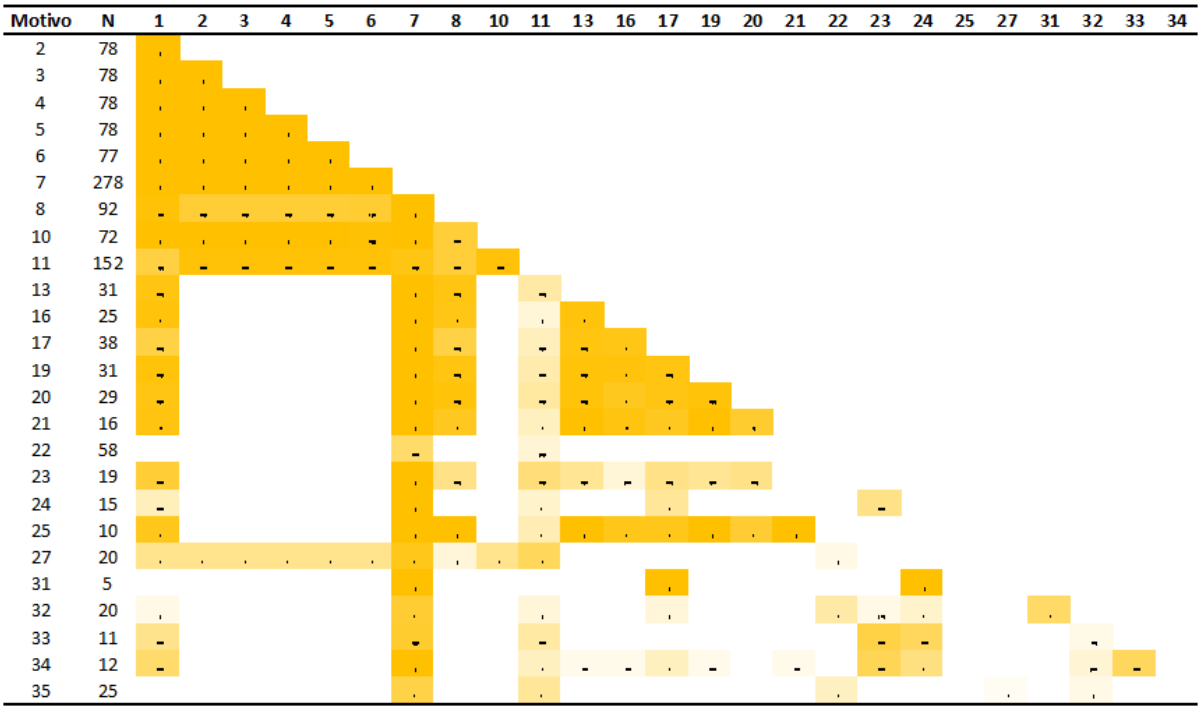
TABELA 45 – Motivos identificados em MetAmaTrp com janela 16-25 com seleção dos enriquecidos.

Os valores de p seguem uma escala na cor azul. Em azul mais claro valor p entre 1x10-5 e 1x10-10; em azul numa tonalidade intermediária valor p entre 1x10-10 e 1x10-20; e em azul mais escuro valor p < 1x10-20

Motivo	Geral	Housekeeping	Inverso
1	8.16E-66	1.70E-59	4.89E-38
2	2.93E-47	6.80E-40	1.27E-25
3	1.52E-40	6.62E-36	4.90E-20
4	2.96E-46	2.13E-40	6.02E-26
5	1.60E-42	3.69E-39	5.52E-24
6	4.90E-46	1.13E-40	5.66E-25
7	2.06E-30	1.15E-03	1.33E-02
8	3.87E-54	3.89E-52	4.33E-32
9	5.44E-03	5.63E-02	3.56E-02
10	1.79E-36	1.10E-34	1.38E-19
11	1.45E-05	2.47E-03	8.92E-08
12	8.80E-03	7.25E-05	1.47E-01
13	3.83E-20	4.67E-19	1.36E-11
14	6.59E-05	1.59E-03	6.02E-03
15	4.21E-02	9.38E-01	1.92E-03
16	1.62E-19	1.33E-18	4.19E-10
17	4.25E-21	3.80E-18	5.65E-11
18	1.40E-02	7.28E-01	8.33E-01
19	6.34E-110	5.09E-57	4.31E-35
20	2.15E-16	7.84E-16	1.23E-10
21	6.22E-10	8.04E-09	1.69E-05
22	1.48E-07	1.14E-08	3.94E-04
23	1.51E-08	4.29E-05	1.56E-07
24	5.75E-08	5.92E-08	6.08E-05
25	3.88E-06	6.01E-05	3.21E-03
26	6.17E-02	2.65E-03	8.40E-01
27	1.03E-07	1.31E-07	4.96E-07
28	1.14E-03	4.36E-04	3.21E-03
29	9.78E-05	7.94E-03	7.12E-02
30	6.86E-04	3.41E-04	6.20E-05
31	3.20E-06	3.00E-03	3.78E-02
32	3.85E-29	1.32E-19	1.03E-16
33	2.08E-04	8.40E-06	1.64E-03
34	1.86E-06	1.16E-06	2.73E-03
35	1.88E-07	2.07E-07	2.30E-04

TABELA 46 – *Heatmap* representando a proporção de co-ocorrência dos motivos significativos identificados em MetAmaTrp janela 16-25

A intensidade da cor das células representa o grau de co-ocorrência dos motivos nas sequencias, sendo que quanto mais forte a intensidade, maior a proporção de co-ocurrência.



5.3.2.13 Motivos da categoria *housekeeping*

5.3.2.13.1 Motivos de *housekeeping* janela 6-15

Foram identificados 15 motivos nessa categoria (TABELA 47), sendo que 6 motivos foram considerados enriquecidos (40%). O motivo 10 (FIGURA 88) foi o único considerado enriquecido tanto na comparação contra a totalidade do 3'-UTRoma quanto contra a categoria inversa. Os motivos 2 (FIGURA 86) e 3 (FIGURA 87) são pouco complexos; os demais são mais complexos, em geral com algumas posições bastante conservadas. Nessa comparação, a categoria inversa se refere a todos os genes marcadores, isto é, todo o conjunto de genes que mostraram modulação significativa no ciclo de vida. Portanto, esses motivos, em especial o 10, podem ser considerados como elementos mais raros em genes modulados, sendo que a sua importância na manutenção de níveis estáveis de mRNA é uma hipótese a

ser testada.

TABELA 47 – Motivos identificados como *housekeeping* com janela 6-15 com seleção dos enriquecidos.

Motivo	Expressão Regular	Geral	Inverso
1	TTTTTTTTTTTTTTT	2.57E-01	1.18E-01
2	AAAAA[AG]AAAAAAAAA	2.26E-03	7.23E-09
3	GT[GT]TGTGTGTGTGTG	8.21E-01	7.71E-02
4	[GA][AG][GA][AG][GA][AG][GA][AG][AG][GA][GA][AG][GA][AG][GA]	2.15E-02	5.14E-03
5	[TC]T[TC]T[TC]T[TC][TC][CT][TC][CT]T[TC][TC][CT]	3.89E-01	1.24E-01
6	A[TC]A[TC]A[TC]A[TC]A[TC]A[TC]A[TC]A	7.01E-01	5.18E-03
7	TTT[TAG]TT[TGA]TT[GAT]TT[TAG]TT	9.57E-01	5.85E-01
8	A[AC]AA[CA]AA[AC]AA[CA]AAAA	2.14E-03	1.90E-08
9	A[TC]AA[TA]AATAA[TA][AT]A[AT]A	7.02E-03	1.13E-05
10	[GTC]C[AC][CG][CA][AC][CG]CAC[CA][AC][CA][CG][AC]	1.89E-14	3.04E-13
11	[GAC]A[GA]A[GA][GA][AG][AG]GA[AC][AG][AG][AG]A	2.82E-03	2.06E-10
12	G[TG][GT][TG]GTG[TG]GTG[TG]G[TG][GC]	4.30E-01	5.37E-03
13	TT[TC][TA]TTT[AT]TTT[TA]TT[TC]	5.37E-01	4.39E-01
14	[CA][AG]CAC[GA]C[AG]	6.81E-01	2.75E-10
15	TG[GT][TC]G[TG]TG[GC]TG[CT][TC]G[CG]	2.30E-04	1.45E-13

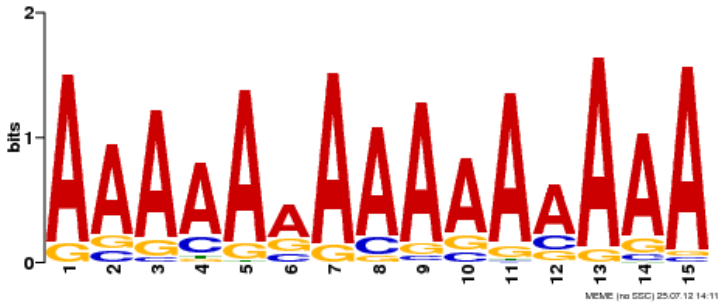


FIGURA 86 – Logo do motivo 2 da categoria housekeeping.

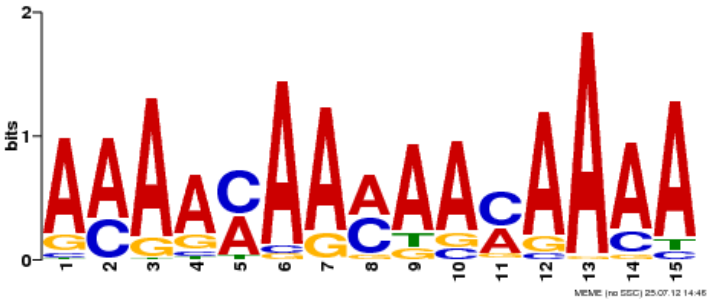


FIGURA 87 – Logo do motivo 8 da categoria housekeeping.

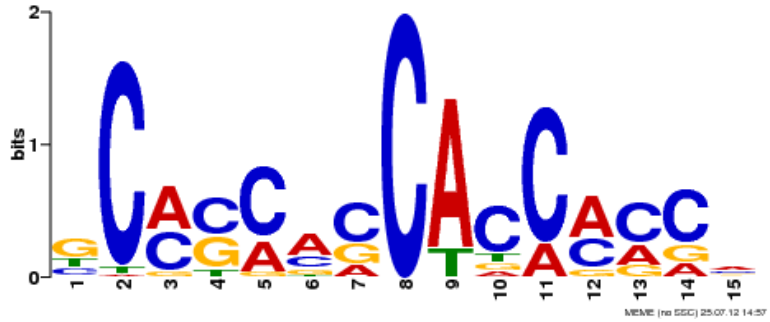


FIGURA 88 – Logo do motivo 10 da categoria housekeeping.

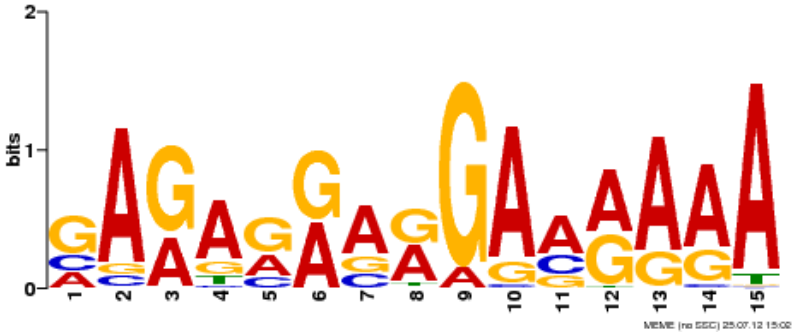


FIGURA 89 – Logo do motivo 11 da categoria housekeeping.

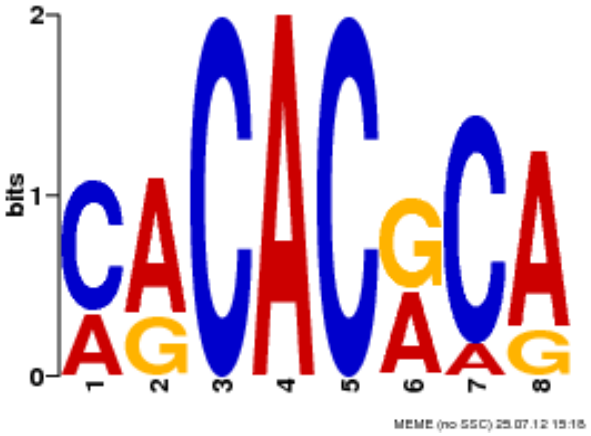


FIGURA 90 – Logo do motivo 14 da categoria housekeeping.

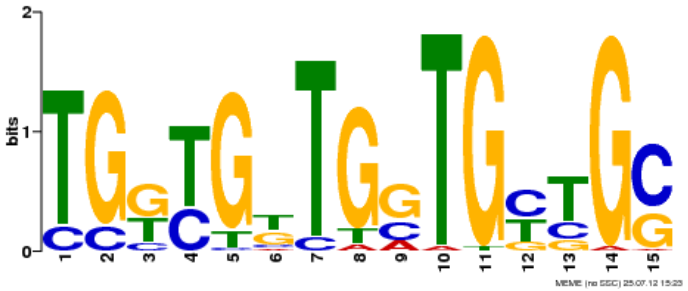


FIGURA 91 – Logo do motivo 15 da categoria housekeeping.

5.3.2.13.2 Motivos de *Housekeeping* janela 16-25

Foram identificados 14 motivos nessa categoria (TABELA 48), sendo que 7 motivos foram considerados enriquecidos (50%). O motivo 11 **(**FIGURA 97**)** foi o único considerado enriquecido tanto na comparação contra a totalidade do 3'-UTRoma quanto contra a categoria inversa. As considerações realizadas para a categoria anterior também são válidas para a comparação atual. De maneira geral, os motivos são pouco complexos.

TABELA 48 – Motivos identificados como housekeeping com janela 16-25 com seleção dos enriquecidos.

Motivo	Geral	Inverso
1	1.02E-04	1.03E-14
2	5.28E-37	8.07E-02
3	3.16E-01	1.42E-02
4	1.63E-03	1.20E-13
5	1.93E-03	2.32E-01
6	4.83E-01	1.94E-03
7	2.39E-03	1.16E-08
8	7.65E-01	6.07E-02
9	1.15E-03	8.26E-12
10	7.43E-01	5.85E-02
11	1.05E-20	8.82E-14
12	5.25E-01	2.09E-01
13	1.25E-02	6.86E-03
14	6.07E-02	9.28E-08

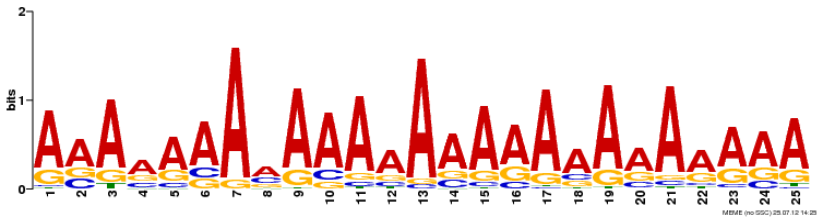


FIGURA 92 – Logo do motivo 1 da categoria housekeeping.

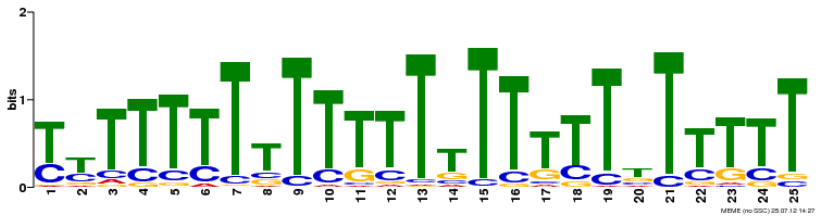


FIGURA 93 – Logo do motivo 2 da categoria housekeeping.

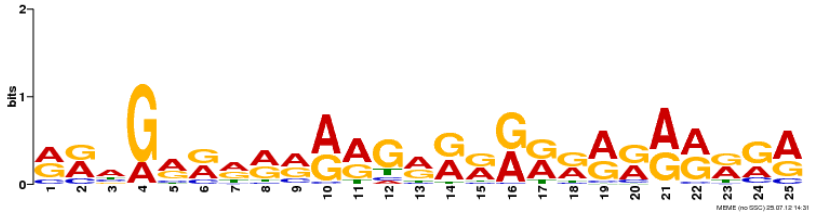


FIGURA 94 – Logo do motivo 4 da categoria housekeeping.

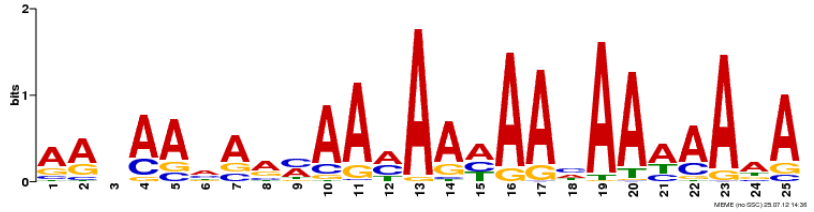


FIGURA 95 – Logo do motivo 7 da categoria housekeeping.

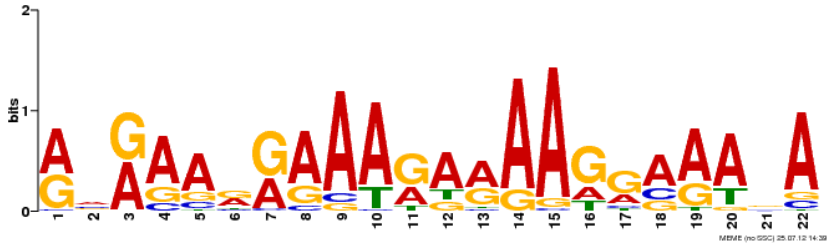


FIGURA 96 – Logo do motivo 9 da categoria housekeeping.

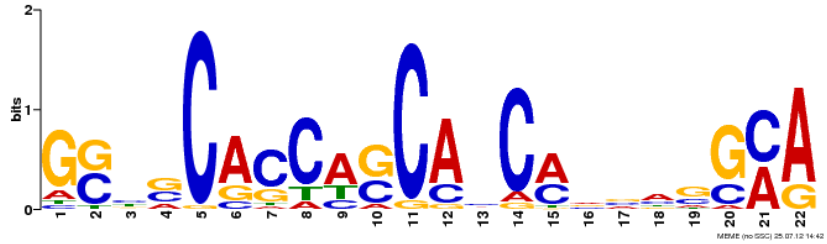


FIGURA 97 – Logo do motivo 11 da categoria housekeeping.

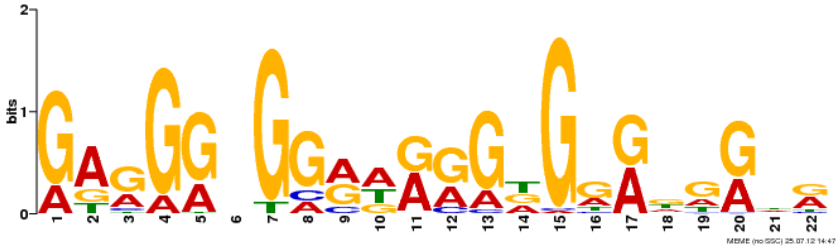


FIGURA 98 – Logo do motivo 14 da categoria housekeeping.

Na comparação com as vias do KEGG, somente um gene contendo o motivo 11 foi identificado, a proteína ligação ao motivo TATA (TBP), na via de fatores de transcrição basal (FIGURA 99).

5.3.3 CONSIDERAÇÕES GERAIS SOBRE A PREDIÇÃO DE ELEMENTOS REGULATÓRIOS NOS GENES CANDIDATOS

A análise realizada no tópico anterior foi relativamente exaustiva. Optamos por apresentá-la de maneira completa, sem selecionar alguns exemplos característicos, para reforçar os padrões que foram identificados em cada uma das análises, bem como as suas nuances, principalmente em relação a aspectos biológicos e técnicos que influenciariam a análise.

Sabemos que a representação completa dos padrões identificados, sua significância estatística e a análise co-ocorrência são repetitivas. Isso é reforçado pelo fato de que os padrões identificados são comentados de forma somente descritiva, sem análises subsequentes que comprovem sua participação em regulação. No entanto, é importante reforçar que o objetivo primário do presente trabalho é a construção de uma base bioinformática que possibilite a análise integrada, dentro de uma perspectiva ômica, de diferentes aspectos relacionados à regulação da expressão gênica. Nesse sentido, a identificação de elementos conservados em região 3'-UTR tem uma importância cabal e devido a isso optamos por essa descrição detalhada.

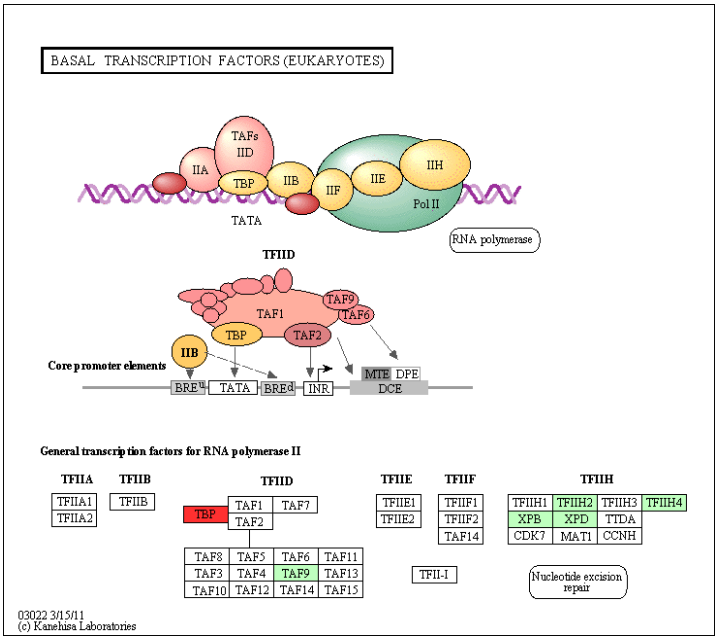


FIGURA 99 - Representação do mapa do KEGG da via metabólica da glicólise e gliconeogênese, com genes modulados em EpiMetAma contendo os motivos mais enriquecidos

Os genes com o motivo 11 estão marcado em vermelho. Os demais genes de *T. cruzi* que estão nessa via estão marcados em verde.

Com os diferentes resultados, foi possível identificar diversas alterações que devem ser realizadas posteriormente para aumentar o grau de confiabilidade na presunção de que elementos conservados identificados sejam realmente regulatórios. Dentre diversas modificações, mencionamos o tratamento das famílias multi-gênicas, a análise de co-ocorrência em dimensões maiores do que 2 a 2, a incorporação de análises de conservação evolutiva (posições menos conservadas nos motivos identificados tendem a variar mais frequentemente em outros genomas similares a *T. cruzi*) e a incorporação de uma maior variedade de análises transcriptômicas à seleção dos conjuntos de genes co-expresso, abrangendo as outras análises transcriptômicas que estão sendo realizadas por nosso grupo.

Independentemente às objeções descritas acima, foi possível evidenciar a relevância da abordagem empregada, reforçando a questão do controle inverso, que permite identificar candidatos com maior potencial de estarem diretamente relacionados ao fenômeno específico em questão (presente em um grupo, ausente em seu inversO), a análise de co-ocorrência para identificar elementos compartilhados (indicativo de regulação combinatória) ou isolados (maior potencial de determinar diretamente o padrão de co-expressão analisado).

5.4 ASSOCIAÇÃO ENTRE ESTRUTURA DE REDE METABÓLICA E EXPRESSÃO DIFERENCIAL

Além das análises realizadas anteriormente utilizando a estrutura do KEGG para as redes metabólicas de *T. cruzi*, podemos utilizá-la para avaliar se o tipo de reação enzimática pode estar associada ao padrão de expressão diferencial. Conforme pode ser visto na FIGURA 100, no qual vemos os supergenes de *T. cruzi* representado em círculos (nós) e o tipo de reação enzimática como uma seta (arestas), pode-se perceber que o nó SG1641 só tem reações direcionais; caso sua expressão seja modulada, por exemplo como repressão, o prosseguimento da reação enzimática fica bloqueado. Já o supergene SG41189 tem uma reação reversível e caso seja reprimido, irá direcionar a reação para outro lado.

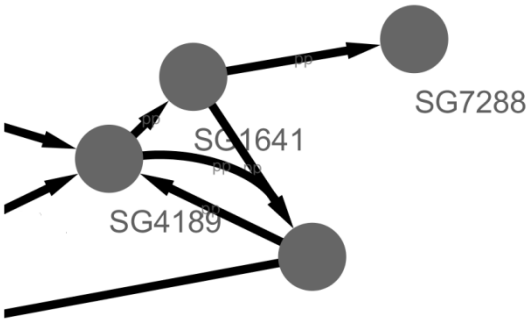


FIGURA 100 – Reações enzimáticas avaliadas na análise de expressão diferencial.

Quando analisamos a rede inteira do KEGG de *T. cruzi,* encontramos 141 genes envolvidos em reações irreversíveis (como o SG1641) e 185 genes envolvidos em reações reversíveis (como o SG4189). A pergunta que foi feita, para ilustrar a utilização de forma global dos dados do KEGG em relação à regulação da expressão gênica foi a seguinte: há algum padrão de modulação preferencial nos genes diferencialmente expressos (DEG), em geral ou em cada uma das 15 categorias identificadas para a análise de motivos regulatórios, se ele catalisa uma reação reversível ou irreversível? Na , podemos verificar as proporções identificadas e o resultado do teste estatístico aplicado (teste exato de Fisher), o qual demonstra que não foi identificado uma associação entre DEGs e o tipo de reação que eles codificam.

TABELA 49 – Análise estatística da associação de genes diferencialmente expressos com o tipo de reação enzimática catalizado.

	Contagem Total		Contagem Categoria		
Categoria	**Rev.**	**Irrev.**	**Rev.**	**Irrev.**	**p**
Todos	183	135	55	39	90.6%
Ama	183	135	2	0	51.1%
AmaTrp	183	135	0	1	42.6%
Epi	183	135	23	12	37.4%
EpiAma	183	135	3	3	70.2%
EpiAmaTrp	183	135	5	4	100.0%
EpiMet	183	135	10	6	79.9%
EpiMetAma	183	135	6	2	47.5%
EpiMetTrp	183	135	1	2	57.8%
EpiTrp	183	135	0	0	100.0%
Met	183	135	9	8	80.3%
MetAma	183	135	0	0	100.0%
MetAmaTrp	183	135	4	4	72.7%
MetTrp	183	135	0	2	18.3%
Trp	183	135	1	0	100.0%
Housekeeping	183	135	2	4	40.8%

5.5 IDENTIFICAÇÃO DAS EXTREMIDADES DOS mRNAs

Para identificar as extremidades dos mRNAs de *T. cruzi*, utilizamos uma grande base de dados inicial, a qual representa o esforço continuado do nosso grupo realizado nos últimos dois anos em avaliar o transcriptoma de *T. cruzi* por sequenciamento de nova geração. Devido ao processamento não enviesado das bibliotecas de fragmentos de mRNAs utilizados para o sequenciamento, a probabilidade de que um fragmento contenha uma seqüência de mini-exon ou poli-A, mesmo parcial, é muito baixa. Isso é intensificado pela degradação inicial que possa já estar ocorrendo no mRNA, mesmo com a manipulação cuidadosa do mesmo, bem como heterogeneidade no processo de fragmentação, no qual as extremidades do mRNA seriam perdidas preferencialmente. Todos esses fatores resultam em uma pequena proporção de elementos indicadores das extremidades do mRNA (mini-exon e cauda poli-A) nas leituras processadas.

De fato, obtivemos uma proporção relativamente baixa de leituras com mini-éxon ou cauda poli-A. De um total de 2,7 bilhões de seqüências, mesmo utilizando um critério de seleção pouco estringente, somente 75.325.282 leituras (3,0%) apresentaram seqüência indicativa de mini-éxon e 105.009.114 leituras(4,0%) apresentaram seqüência indicativa de cauda poli-A.

Ao realizar o mapeamento dessas seqüências no haplótipo Não-Esmeraldo de CL Brener, obtivemos mapeamento positivo para 27.954.690 seqüências contendo um mini-éxon putativo (37,1% do total) e 42.832.093 seqüências contendo uma cauda poli-A putativa (40,8% do total). A baixa quantidade de leituras mapeadas provavelmente se deve principalmente a seqüências de má qualidade e, minoritariamente, às divergências entre as cepas CL Brener, utilizada como referência, e a Dm28c, da qual foram gerados os dados. O total de mapeamentos obtidos foi 46.981.948 para o mini-éxon (1,7 mapeamento por leitura) e 72.971.396 para a cauda poli-A (1,7 mapeamento por leitura).

A distribuição do número de mapeamentos por leitura pode ser visto na FIGURA 101. A grande maioria das leituras contendo mini-éxon e cauda poli-A mapearam em somente uma posição (24.162.933 para mini éxon (86,4%); 37.521.110 para poli-A (87,6%)); em seguida, o maior número de mapeamentos identificado foi 2 (1.835.714 para mini-éxon (6,6%); 2.734.438 para poli-A (6,3%)).

O número máximo de posições permitida no *software* de mapeamento foi 500. Para mini-éxon, 5 leituras mapearam em 500 posições, 24.415 leituras mapearam em mais de 100 posições (0,09%), 308.794 leituras mapearam em mais de 10 posições (1,1%), e 827.155 mapearam em mais de 5 posições (3,0%); para poli-A, 317 leituras mapearam em 500 posições, 59.060 leituras mapearam em mais de 100 posições (0,14%), 439.464 leituras mapearam em mais de 10 posições (1,0%), 1.091.051 mapearam em mais de 5 posições (2,5%).

Há uma pequena tendência de que as leituras contendo poli-A mapeiem em mais posições; no entanto, essa tendência é revertida quando comparamos mapeamentos múltiplos em menor quantidade (>5 mapeamentos). E na comparação dos dois histogramas de mapeamento, o padrão geral é de grande similaridade.

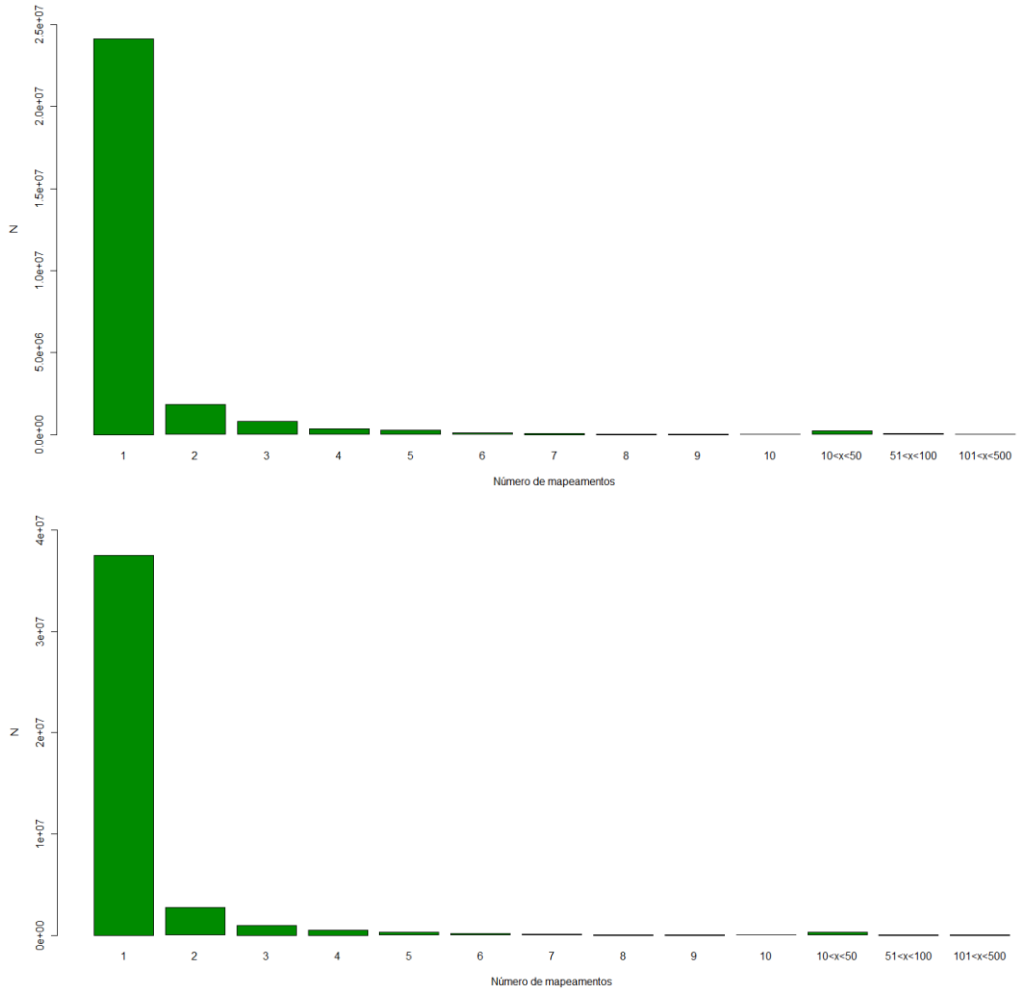


FIGURA 101 – Histograma do número de mapeamentos obtidos por leitura

Acima, mapeamentos com leituras contendo mini-éxon. Abaixo, mapeamentos com leituras contendo poli-A

Da FIGURA 102 a FIGURA 104 temos algumas representações esquemáticas

do resultado da identificação inicial das seqüências de mini-éxon, para ilustrar os resultados obtidos de forma mais quantitativa global, utilizando como tamanho mínimo 10 posições.

Na FIGURA 102, avaliamos a densidade de leituras identificadas quando comparamos a posição inicial da seqüência de mini-éxon (eixo Y) e o grau de similaridade obtido (eixo X). Podemos observar que a grande maioria das leituras foram identificadas com alta similaridade (>99%) e predominantemente como posição inicial 1. As demais leituras preferencialmente estão localizadas na parte inferior do gráfico, ou seja, identificação da seqüência do mini-éxon no começo da leitura, com diferentes graus de similaridade, e na região superior esquerda, isto é, identificação da seqüência de mini-éxon no final da leitura e baixa similaridade (entre 70% e 80%, sendo que 70% foi nosso limiar mínimo de similaridade para que a leitura seja selecionada.

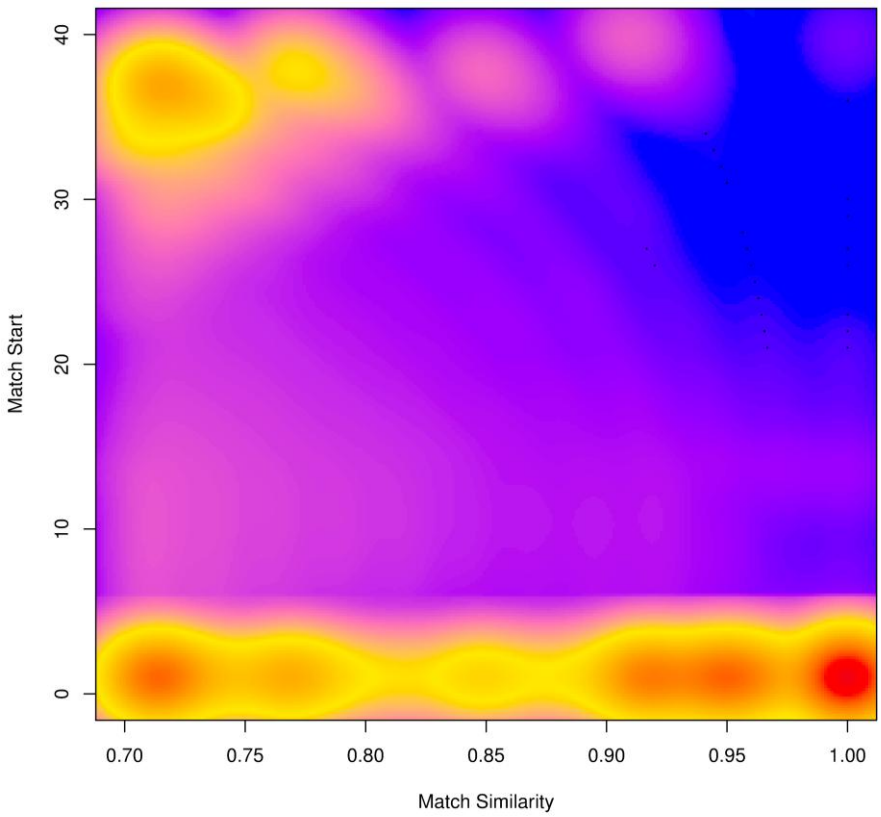


FIGURA 102 – Gráfico de densidade comparando a posição do início da identificação (eixo Y) com o grau de similaridade da região identificada com a seqüência do mini-éxon (eixo X).
As cores representam a densidade de leituras plotadas na região, indo do azul até o vermelho, passando por rosa, amarelo e laranja.

A interpretação desse gráfico é a seguinte: identificações de mini-éxon em posições além da primeira são indicativas de mapeamento espúrio. Podemos ver que a grande maioria das identificações foi realizada no primeiro nucleotídeo, o que é esperado para uma identificação correta. Somente quando baixamos o grau de similaridade com a seqüência de mini-éxon, é que uma quantidade pequena, mas importante, de identificações é feitas com o final da seqüência (posição inicial maior do que 35). Isso se deve ao fato de que, quanto menor a seqüência que será utilizada para a identificação (e o limite inferior é 10, ou seja, posição inicial de 40 na leitura de 50 nucleotídeos), maior é a probabilidade de que um pareamento espúrio seja feito com a seqüência de mini-éxon cuja similaridade seja maior do que o limite selecionado (70%). Podemos observar isso claramente na figura, pois identificações no final da seqüência praticamente não existem quando utilizamos critérios de similaridade mais estringentes (>90%). Portanto, provavelmente a totalidade das identificações no final da seqüência são espúrias, isto é, similaridade aleatória com a seqüência de mini-éxon.

Aparentemente, a quantidade de mapeamento com baixa similaridade no início (extremidade inferior esquerda) e no final da sequencia (extremidade superior esquerda) é muito parecida. Esperamos também identificações espúrios no começo da sequencia, o que seria indicado por um menor grau de similaridade com a seqüência. Como o número de identificações com baixa similaridade no começo da sequencia é o dobro do que foi identificado no final da sequencia, estimamos que cerca de 50% dos mapeamentos com baixa similaridade no começo da seqüência sejam verdadeiros.

Na FIGURA 103, avaliamos a densidade de leituras identificadas quando comparamos o tamanho da região identificada como similar ao mini-éxon (eixo Y) e o grau de similaridade obtido (eixo X). O tamanho da região identificada representa quantos nucleotídeos foram utilizados para avaliar a similaridade do mini-éxon com a seqüência. Esse tamanho pode variar entre 10 (critério de tamanho mínimo utilizado para essa análise) e 39 (tamanho total do mini-éxon).

A interpretação desse gráfico é a seguinte: podemos observar que a grande maioria das leituras foram identificadas contendo uma região de pareamento com o mini-éxon de tamanho pequeno, tanto com alta similaridade (extremidade inferior direita) quanto com baixa similaridade (extremidade inferior esquerda) (>99%). As identificações com tamanhos maiores estão presentes somente com alta

similaridade. e predominantemente como posição inicial 1. Esses resultados indicam que a grande maioria dos pareamentos pequenos ocorre com seqüências pequenas (esperado tanto pela distribuição aleatória quanto pelo provável tamanho da sequencia do mini-éxon presente no fragmento de mRNA que foi utilizado para a análise); e que os pareamentos de grande tamanho e, portanto muito mais prováveis de serem realmente o mini-éxon, praticamente só ocorrem com grau de similaridade alto, o que reforça a acurácia do sequenciador SOLiD em obter a leitura correta.

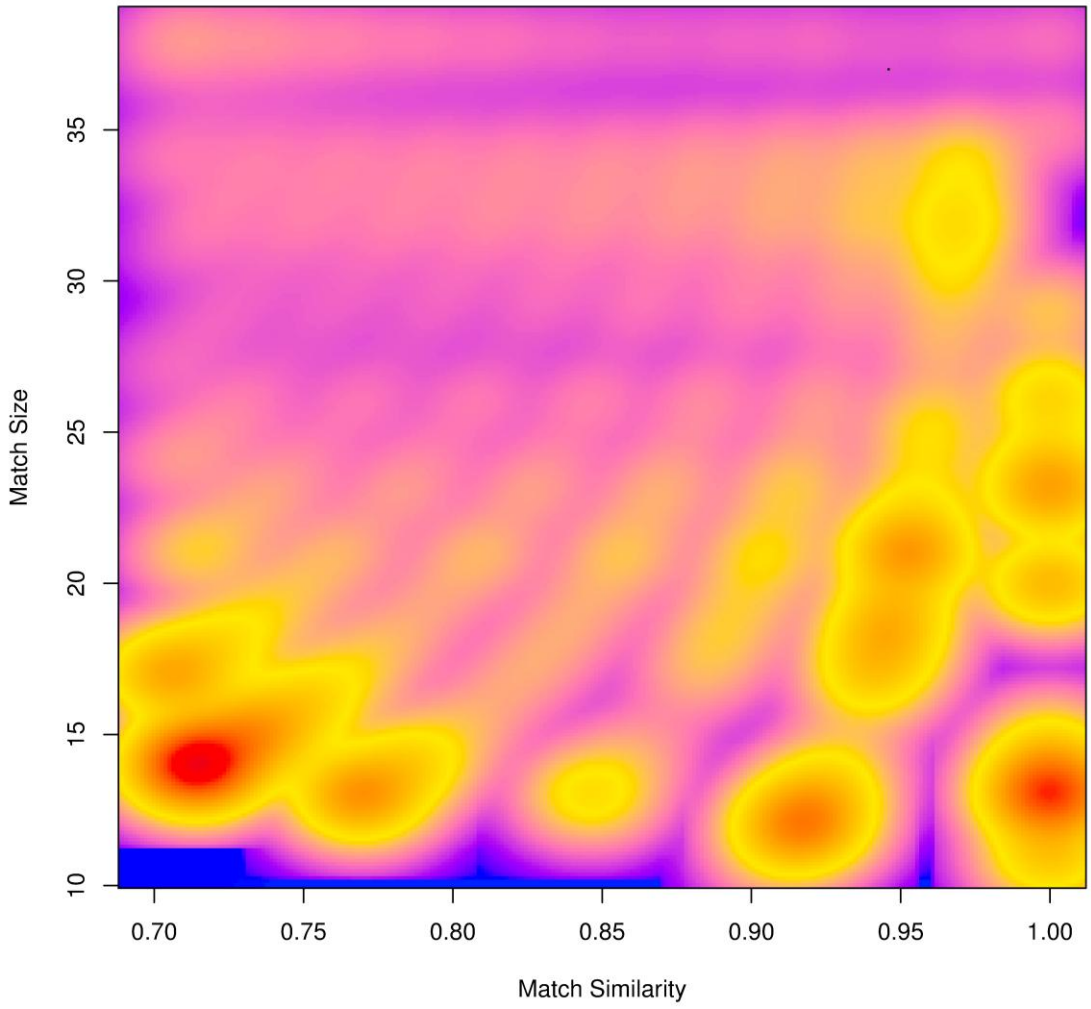


FIGURA 103 – Gráfico de densidade comparando o tamanho da região de pareamento da identificação (eixo Y) com o grau de similaridade da região identificada com a seqüência do mini-éxon (eixo X).

As cores representam a densidade de leituras plotadas na região, indo do azul até o vermelho, passando por rosa, amarelo e laranja.

Na FIGURA 104, avaliamos a densidade de leituras identificadas quando comparamos o tamanho da região identificada como similar ao mini-éxon (eixo Y) e

a posição de início da região de pareamento na leitura (eixo X), para três critérios distintos de similaridade. O tamanho da região identificada representa quantos nucleotídeos foram utilizados para avaliar a similaridade do mini-éxon com a seqüência. Esse tamanho pode variar entre 10 (critério de tamanho mínimo utilizado para essa análise) e 39 (tamanho total do mini-éxon). Os pareamentos cuja posição inicial é o início da leitura são projetados como contendo pareamento corretos.

A interpretação desse gráfico é a seguinte: quando a similaridade da região de pareamento é muito alta (>90%, gráfico superior), praticamente todos os mapeamentos foram feitos no começo da seqüência (pareamentos corretos) e com tamanho da região pequena. Tamanhos maiores de região estão associados ao pareamento no começo da leitura, o que também é indicativo que esses pareamentos são corretos, mas a probabilidade de obtermos segmentos grandes do mini-éxon nos fragmentos de mRNA decresce progressivamente com o aumento do tamanho do pareamento. A diagonal observada contém poucas leituras e representa pareamentos espúiros que podem ocorrer com alta similaridade em posições dentro da leitura (considerados como identificações espúrias). Há poucos pareamentos com tamanho pequeno, posição no final da seqüência e similaridade alta (extremidade inferior direita). Quando a similaridade da região de pareamento é alta (>80%, gráfico intermediário), as observações feitas para o gráfico superior são apropriadas; só observamos que a quantidade de mapeamentos espúrios (diagonal e canto inferior direito) é maior, embora em uma proporção bem pequena. Quando a similaridade da região de pareamento é média (>70%, gráfico inferior), podemos observar que: a) ainda temos uma quantidade maior de pareamentos no começo da sequencia e com tamanho pequeno (canto inferior esquerdo quando comparado ao canto inferior direito), b) o número de identificações corretas com região de pareamento grande é muito pouco (praticamente não temos pareamentos corretos grandes com similaridade média) e c) a proporção de pareamentos espúrios (diagonal e canto inferior direito) são significativamente maiores, especialmente para os pareamentos pequenos no final da seqüência com similaridade média (canto inferior direito).

Resumindo, esses três gráficos nos sugerem que a maioria das identificações maiores do que 10 nucleotídeos é correta (nosso limite inferior para a análise geral é 5), que pareamentos feitos no começo da seqüência são corretos (o que se espera pela lógica do *trans-splicing*) e que a proporção entre verdadeiro-positivos e falso-

positivos nos pareamentos pequenos e com similaridade média é cerca de 2 vezes.

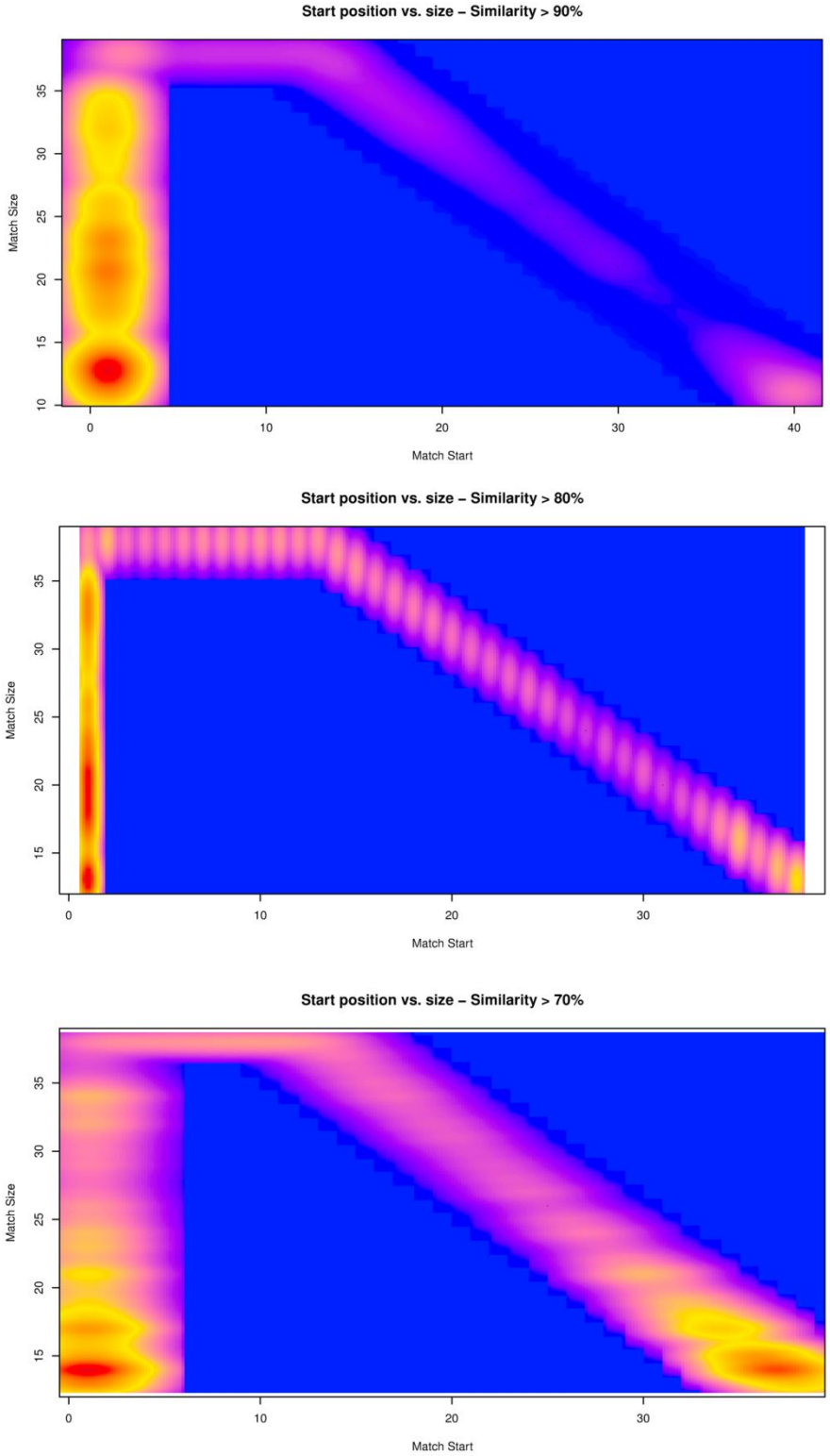


FIGURA 104 – Gráfico de densidade comparando o tamanho da região de pareamento da identificação (eixo Y) com a posição de início (eixo X).

As cores representam a densidade de leituras plotadas na região, indo do azul até o vermelho, passando por rosa, amarelo e laranja. Os três gráficos se referem a limites diferentes de similaridade.

Para a eliminação dos falso-positivos, a última etapa é a comparação da região de pareamento com o genoma. Caso a similaridade dessa região com o genoma seja alta, é provável que a leitura originalmente não continha o elemento avaliado (mini-éxon ou cauda poli-A) e que a identificação foi espúria.

Após as etapas de identificação permissiva de seqüências e mapeamento no genoma, filtramos as leituras para um conjunto altamente provável de conter os elementos através da comparação com o genoma. Caso a região da seqüência que apresentou similaridade com os elementos também apresentasse similaridade com o genoma, essa leitura era descartada. Utilizamos quatro critérios diferentes de estringência (TABELA 50).

TABELA 50 – Número de mapeados que passaram em quatro critérios de estringência diferentes.

Condição	A		B		C		D	
Elemento	N	%	N	%	N	%	N	%
Mini-éxon	18.267.829	38.9%	15.147.831	32.2%	13.425.510	28.6%	11.862.968	25.3%
poli-A	45.083.823	61.8%	35.133.680	48.1%	23.650.714	32.4%	11.916.521	16.3%

Os diferentes critérios são referentes à diferença da similaridade da região da leitura identificada como contendo o elemento quando comparada com o elemento e com o genoma. A) 20% de diferença, B) 30% de diferença, C) 40% de diferença, D%) 50% de diferença. Por exemplo, se uma determinada região da leitura tivesse uma similaridade com o mini-éxon de 80%, essa leitura seria selecionada para análises subsequentes; a mesma região que se assemelhou ao mini-éxon é comparada contra o genoma, e o grau de semelhança foi de 40%. A diferença entre as similaridades, de 40%, é considerada para a seleção final e essa leitura seria contada nas análises A a C.

Para a determinação das leituras que provavelmente continham o mini-éxon, utilizamos o critério D (diferença da similaridade entre mini-éxon e genoma da região identificada maior do 50%). Com esses critérios, obtivemos um total de 6.641.038 leituras mapeadas que muito provavelmente contem o mini-éxon, que correspondem a 0,3% do total de leituras utilizadas para a análise. Essas leituras foram mapeadas em 11.862.968 sítios, o que corresponde a 1,8 mapeamento por leitura.

O mesmo critério foi utilizado para a seleção das leituras que provavelmente continham a cauda poli-A, obtendo-se um total de 11.887.684 leituras mapeadas,

que muito provavelmente contém a cauda poli-A, as quais correspondem a 0,5% do total de leituras utilizadas para a presente análise. Essas leituras foram mapeadas em 11.916.521 sítios, o que corresponde a 1,0 mapeamento por leitura, uma redução significativa em relação à população de mapeamento total.

Na figura FIGURA 105 a FIGURA 107, podemos verificar o padrão de mapeamento dessas leituras positivas em trechos específicos do genoma de *T. cruzi,* demonstrando a eficiência do processo de identificação de leituras marcadoras dos sítios de trans-splicing, bem como algumas situações particulares de trans-splicing.

Nessas figuras é possível observar a identificação correta dos sítios de adição de mini-exon, os quais são representados por uma grande densidade de mapeamento em posições especïficas do genoma, geralmente a 5’ das regiões codificadoras conforme esperado. Há algumas evidëncias de que alguns picos de inserção do mini-exon que são localizadas no interior da região codificadora sejam na realidade indicativos de erro na predição da mesma, e não do sitio de trans-splicing.

Outros padrões interessantes são a existência de múltiplos sítios de *trans-splicing*, os quais podem representar mecanismos reguladores da expressão diferencial desses genes, o que demonstra o desdobramento dos resultados desse trabalho. De acordo com diferentes 5’ UTR e 3’ UTR, é possível que um mesmo gene possa produzir proteínas com diferenças no seu padrão de expressão temporal, localização celular e até mesmo função.

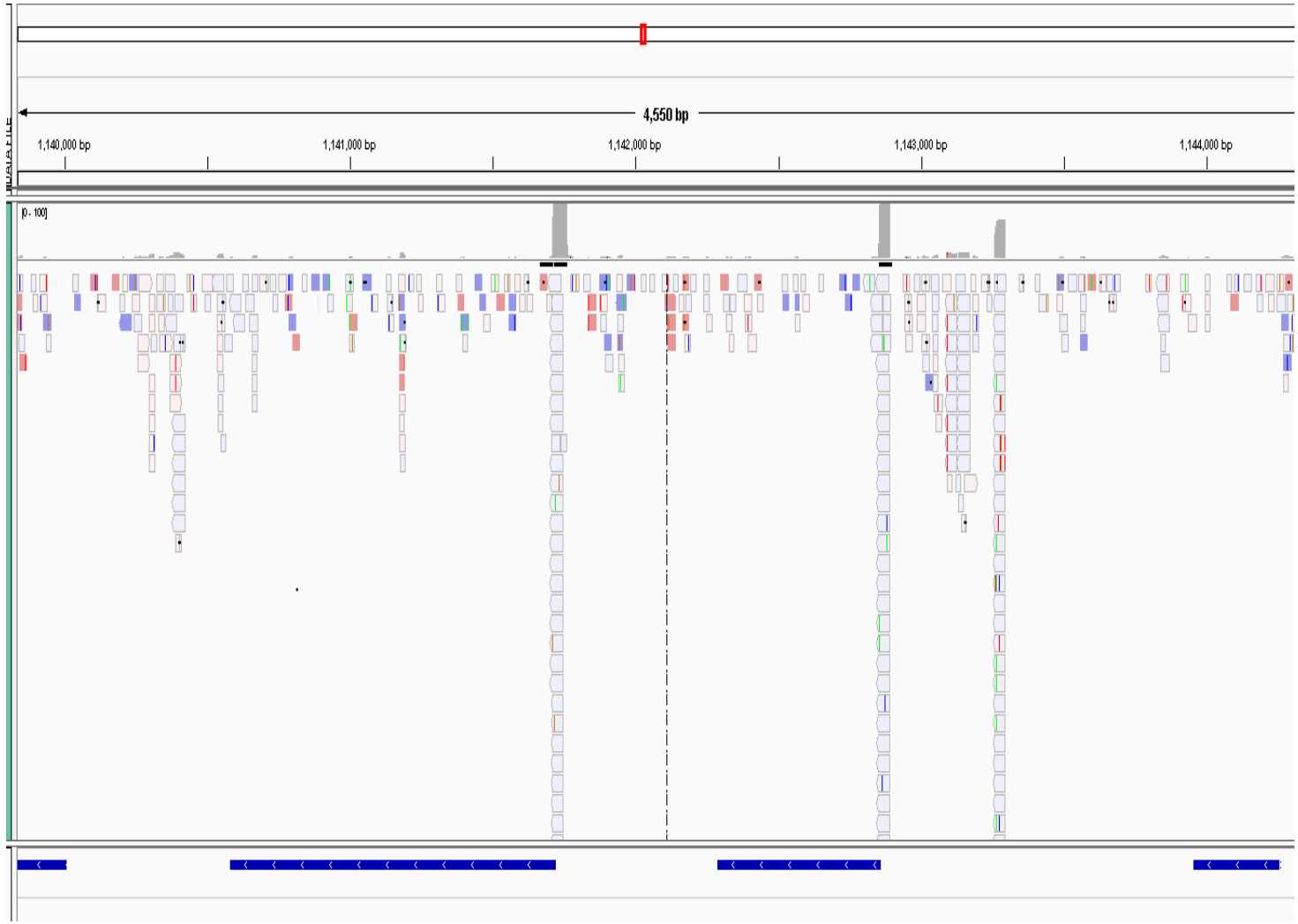


FIGURA 105 – Mapeamento das porções dos reads que foram identificados como contendo mini-éxon no genoma de *T. cruzi*.

Na porção inferior, os retângulos azuis indicam CDS de *T. cruzi,* as setas internas indicam o sentido da codificação; no painel central, a representação das leituras mapeadas como retângulos, e na parte superior um mapa de densidade. Nessa figura, é possível identificar 3 sítios claros de adição do mini-éxon (picos evidentes na parte superior), sendo que um dos genes apresenta dois sítios (trans-splicing alternativo), sendo que o sítio mais próximo da extremidade 5'da região codificadora é o mais frequente.

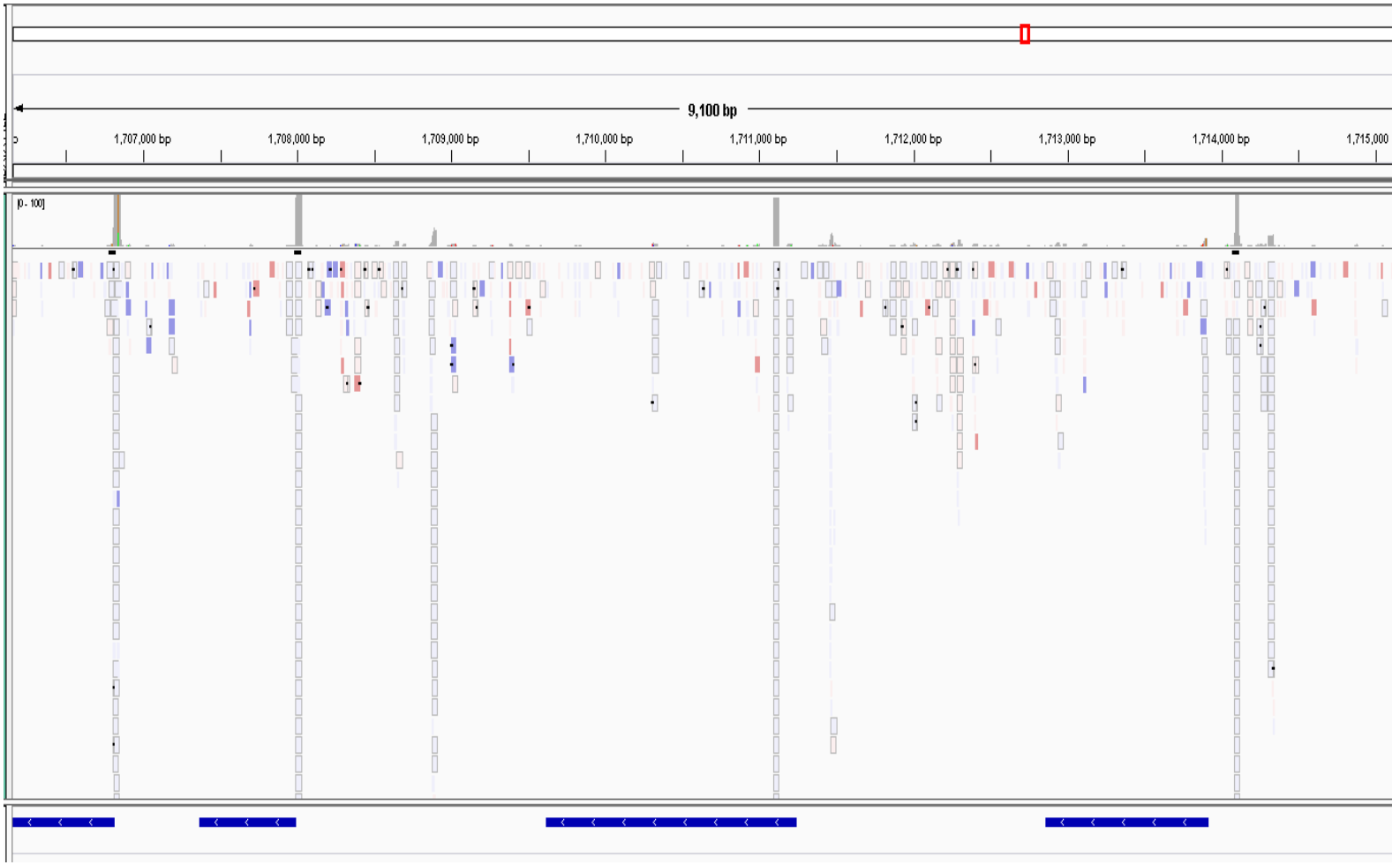


FIGURA 106 – Mapeamento das porções dos reads que foram identificados como contendo mini-éxon no genoma de *T. cruzi*.

Legenda idêntica à da figura 105. Nessa porção, podemos observar claramente que todos os genes visualizados apresentam um mini-éxon identificado um pouco antes do início da sua região codificadora. A exceção é o terceiro gene, cuja região de mini-éxon está interna à CDS, o que indica um erro de predição do códon de inicio da tradução para essa CDS.

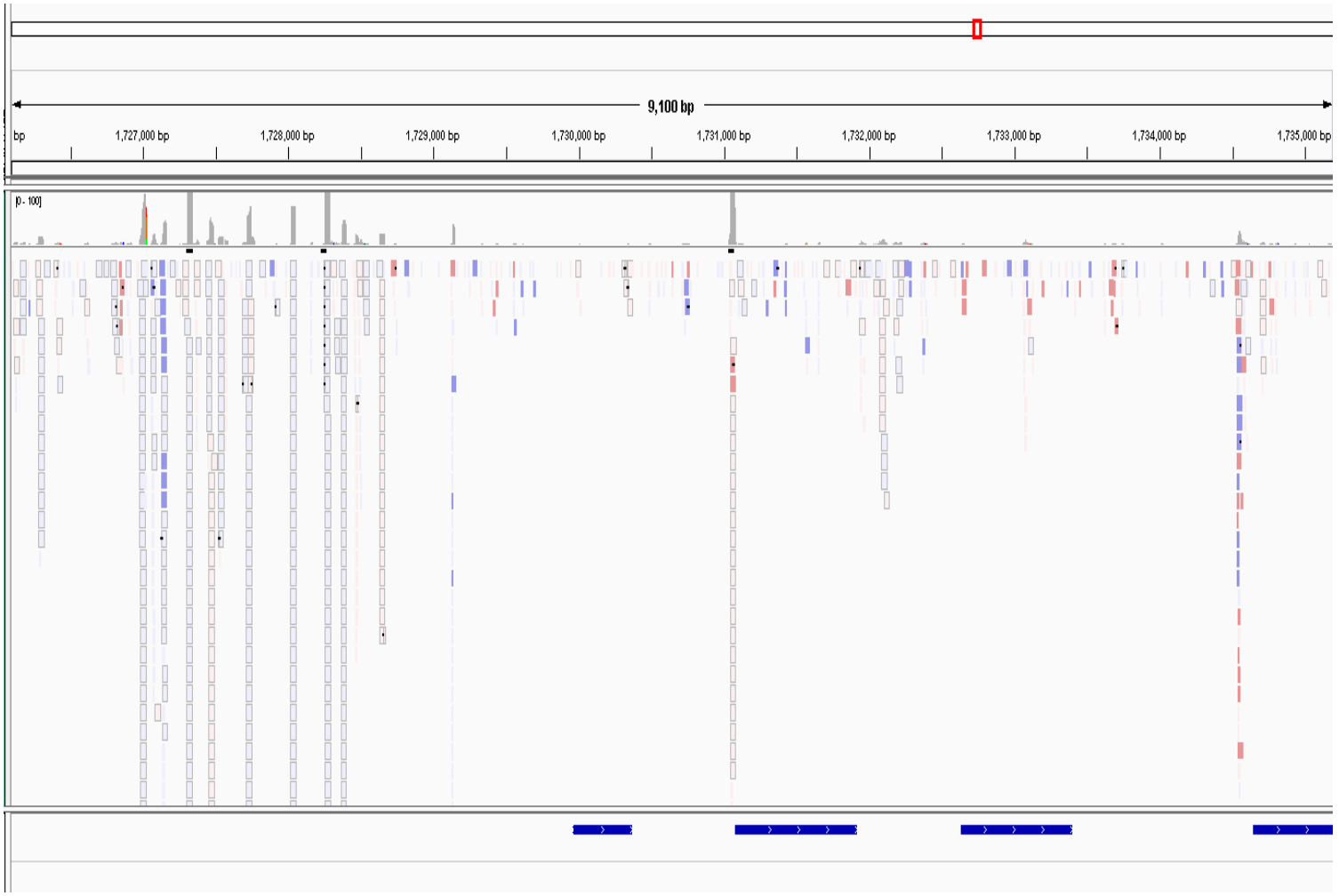


FIGURA 107 – Mapeamento das porções dos reads que foram identificados como contendo mini-éxon no genoma de *T. cruzi*

Legenda idêntica à da figura 105. Nessa porção, evidenciamos padrões estranhos de mapeamento, onde uma ilha de picos com mini-éxon é evidenciada. Uma explicação para esse fenômeno seria um loco de codificação para o gene de mini-éxon o qual não teria sido anotado..

6 DISCUSSÃO

Trypanosoma cruzi é o agente causador da Doença de Chagas, e junto com outras doenças causadas por tripanossomatídeos, como a Doença do Sono e as leishmanioses, afetam milhões de pessoas em todo mundo. Essas doenças são endêmicas dos Trópicos e geralmente acometem populações pobres da América Latina, África e Ásia. Por isso, não chamam a atenção das empresas farmacêuticas e são consideradas doenças negligenciadas (HOTEZ & PECOUL, 2010). São doenças que necessitam de pesquisa para o melhor entendimento da biologia desses parasitas, que possam levar ao desenvolvimento de vacinas e fármacos para combater essas doenças.

Os tripanossomatídeos divergiram muito cedo da linhagem evolutiva do eucariotos e por isso possuem características biológicas muito peculiares, bastante diferentes das dos outros eucariotos. Em relação às diferenças relacionadas à expressão gênica, eles possuem grupos de genes direcionais, dispostos em fitas separadas nos cromossomos, que formam longas unidades policistrônicas e são processadas por trans-splicing para gerar os mRNAs que serão traduzidos. Nesse processo, é adicionada a seqüência do mini-éxon na extremidade 5'UTR dos transcritos e a cauda poli-A na extremidade 3' (CLAYTON, 2002).

Como a regulação da expressão gênica é majoritariamente pós-transcricional, o *trans-splicing* é um importante passo de regulação da expressão gênica. Uma vez que o transcrito primário possua múltiplos sítios com sinal para inserção do mini-éxon ou da cauda poli-A, o processamento pode gerar diferentes variantes do transcrito processado, que embora tenham a mesma seqüência da região codificadora possuem diferentes extremidades 5' e 3'. E os motivos presentes nas regiões 5' e 3' dos transcritos são outro importante ponto para a regulação pós-transcricional. Estes elementos podem controlar a meia-vida deste mRNA, direcioná-lo para um outro compartimento celular, amarzená-lo ou direcioná-lo para a degradação. Isso depende da proteína ou proteínas e complexos que se ligam a esses elementos regulatórios.

Nesse ponto da regulação da expressão gênica, dois capítulos deste

trabalho vêm a contribuir para a elucidação desses mecanismos de regulação. Primeiro é a identificação de possíveis elementos regulatórios na 3' UTR dos transcritos de *T. cruzi*. Normalmente são feitas buscas por motivos nas regiões intergênicas e esses prováveis elementos regulatórios são comparados com os presentes nas regiões codificadoras dos genes, utilizando esse parâmetro como controle da significância desses motivos como elementos regulatórios. Essa procura baseia-se na similaridade de seqüência entre diferentes regiões do genoma. Essa busca de motivos por similaridade de seqüência pode produzir melhores resultados quando se analisa comparativamente outras espécies.

Nesse trabalho, nós utilizamos uma abordagem, que agrega uma outra informação à busca por similaridade de seqüência, a co-expressão gênica. Nós partimos de dados de RNA-seq do ciclo de vida do parasita e com base nesses dados classificamos genes marcadores de 15 categorias: Epi, Met, Ama, Trp, EpiMeta, EpiAma, EpiTrp, MetaAma, MetTrp, AmaTrp, EpiMetAma, EpiMetTrp, EpiAmaTrp, MetAmaTrp, e *housekeeping* (genes de expressão constitutiva). Partindo do pressuposto que genes co-expressos possam estar sendo co-regulados, e que a co-regulação seja causada pelos mesmos motivos em suas 3'UTRs, fizemos então a busca por motivos em cada uma destas categorias. Ao associar essas duas informações aumentamos a nossa chance de encontrar elementos regulatórios específicos das fases do ciclo de vida do parasita.

Outra parte importante do trabalho foi delimitar as extremidades 5' e 3' dos transcritos. Para a identificação dos elementos reguladores na região 3'UTR, utilizamos um critério empírico de 300 nucleotídeos como sendo o tamanho máximo desse elemento. Ao identificar os limites do transcrito, é possível realizar a predição dos elementos reguladores com maior confiabilidade. Como demonstramos nesse trabalho, um gene pode ter diferentes sítios de inserção do mini-éxon e, principalmente, de cauda poli-A, ou trans-splicing alternativo, podendo gerar variantes com a presença ou ausência de um determinado motivo, ou ainda diferentes combinações de motivos que podem se ligar a complexos protéicos de regulação. A delimitação correta das extremidades dos transcritos, ou das possíveis combinações, nos permite identificar os diferentes elementos regulatórios presentes num transcrito. Tendo em mãos esses limites e identificando os genes que apresentam sítios alternativos de trans-splicing, iremos realizar futuramente uma análise de qual seriam os possíveis sítios reguladores presentes nas regiões

alternativas dos mRNAs.

Outras abordagens em larga escala que estão sendo utilizadas no Instituto Carlos Chagas podem vir a contribuir para o refinamento na determinação de genes co-regulados, e consequentemente na identificação de motivos. O mesmo que foi feito neste trabalho utilizando dados de RNA-seq do ciclo de vida, também pode ser feito com RNA-seq de diferentes situações fisiológicas do parasita. Como por exemplo, analisar o transcriptoma do *T. cruzi* submetido a diferentes tipos de estresse (nutricional, oxidativo, populacional, etc), em resposta a diferentes drogas, parasitas que tenham genes nocauteados, etc. Essas mesmas condições podem ser avaliadas em nível de proteína através da avaliação do proteoma por espectrometria de massas.

Além disso, experimentos de ribonômica, determinação da população de mRNAs que estão associados a uma determinada proteína ligadora a RNA, são de extrema importância para a identificação de elementos regulatórios nos transcritos. Uma vez que se sabe que o conjunto de transcritos que se associa a uma mesma proteína, pode-se pressupor que a proteína ligadora a RNA se liga a esses transcritos pelo mesmo motivo. Fazendo uma busca por motivos por similaridade de seqüência associada a essa tipo de informação aumenta a confiabilidade e a probabilidade de se encontrar elementos regulatórios, pois temos um forte indício de que esses mRNAs são co-regulados pela mesma proteína. Infelizmente a busca por motivos por similaridade de seqüência não é de todo efetivo, pois essas seqüências formam estruturas secundárias, e é a conformação da estrutura secundária que vai dar a especificidade do sítio à ligação com as proteínas (CHEN *et al*., 2006; PARKER *et al*., 2011). Por isso também é importante utilizar abordagens *in silico* que busquem por estruturas secundárias, mas ferramentas para determinação e a busca por conservação da estrutura secundária ainda não são satisfatórias porque a variação na seqüência primária pode levar a uma mesma estrutura secundária e tais programas ainda precisam de melhorias.

O que ajuda na elaboração dos modelos de predição dos programas são as estruturas secundárias já conhecidas a partir de dados experimentais. Esses dados de estrutura são escassos, dado o trabalho de estudo individual de cada transcrito. Mas recentemente foi desenvolvida uma técnica para determinação da estrutura secundária de transcritos utilizando clivagem com enzimas de clivam RNA apenas em fita simples, e emzimas que clivam apenas em fita dupla, combinado a RNA-seq.

Essa técnica é chamada de PARS (do inglês *Parallel Analysis of RNA Structure*) e resumidamente utiliza um score para predizer se determinado nucleotídeo do transcrito é fita simples ou fita dupla. Esse score é a razão entre os número de tags ,mapeado em determinada base, de uma biblioteca com relação a outra, ou seja, a biblioteca que foi feita utilizando a enzima que cliva fita simples (nuclease S1) e a biblioteca feita utilizando a enzima que cliva fita dupla (RNase V1) (KERTESZ *et al*., 2010). Nesse trabalho foram identificadas as estruturas secundárias de ~3.000 transcritos, correspondendo a 50% do genoma de *Saccharomyces cerevisiae.*

Uma outra abordagem em larga escala que complementaria os outros dados é a análise do interatoma (conjunto de interações proteína-proteína) de *T. cruzi*. Ao saber a rede de interação de proteínas podemos entender melhor como a regulação está ocorrendo; além disso, podemos usar os dados de interatoma para associar função às proteínas hipotéticas (que perfazem 50% do genoma) de *T. cruzi*. Estamos produzindo atualmente o interatoma completo de *T. cruzi,* focando inicialmente nas RBPs (PRETI *et al.,* em andamento).

Um outro nível de análise que seria bastante interessante fazer é criar a rede combinada de interação proteína-proteína e de interação proteína ligadora a RNA e seus alvos. Tal análise nos daria uma visão ainda mais integrada de como a regulação da expressão gênica está ocorrendo em *T. cruzi*.

Essas abordagens de obtenção de dados em larga escala por transcriptômica, proteômica, ribonômica e interatoma fazem parte do Projeto Reguloma do Instituto Carlos Chagas, que visa integrar todos esses dados, a fim de montar as redes de regulação da expressão gênica em *T. cruzi*. Para trabalhar com o conjunto de todos os genes de *T. cruzi* e facilitar a transferência de genes de uma plataforma de caracterização funcional para outra (por exemplo, para vetores de expressão com diferentes tags ou para os vetores para ensaio de duplo-híbrido em levedura) foi construído no Instituto Carlos Chagas o ORFeoma de *T. cruzi*, biblioteca de clones de todos os genes codificadores de proteína.

Esse conjunto de dados em larga escala de diferentes fontes pode nos levar a um entendimento global do sistema molecular do parasita. Mas para isso é necessário armazenar e organizar toda essa informação, o que não é uma tarefa trivial. Neste ponto entra a 3ª parte deste trabalho, a implementação de um banco de dados capaz de armazenar de maneira integrada dados biológicos complexos. Nesse sentido, foi implementado um banco de dados, KinetoDB, com as

informações genômicas de 5 espécies de tripanossomatídeos e um visualizador genômico. Embora já existam bancos de dados públicos de tripanossomatídeos com esse tipo de informação, existe a necessidade de termos um banco deste tipo no Instituto não só para o armazenamento mas também para a mineração dos dados, e aproveitamento dos dados ainda não publicados. Esse banco de dados seria direcionado para análises intensas e não para a disponibilização dos dados para a comunidade científica interessada, o que justifica a construção local de um sistema de armazenamento de informações.

A genômica comparativa é uma poderosa ferramenta para o entendimento de diversos aspectos da biologia, se aplicando também no contexto de regulação da expressão gênica. Por isso, embora o foco do Instituto seja o estudo da regulação da expressão gênica em *T. cruzi*, a incorporação dos dados genômicos e de expressão gênica de outras espécies de tripanossomatídeos também é importante. Desse modo, fica mais claro de encontrar os padrões, motivos, e mecanismos de regulação gênica que são mais conservados entre as espécies. Estamos atualmente iniciando diversos projetos de sequenciamento genômico de tripanossomatídeos, com o foco principal de incorporar ao sistema de dados as posições variantes existentes nos diferentes genomas em suas regiões reguladoras da expressão gênica.

Finalmente, a quarta parte de contribuição desse trabalho foi a abordagem da análise da expressão gênica de *T. cruzi* projetada em redes metabólicas do parasita. Isso permite que vejamos as reações metabólicas como um todo, como os genes e diferentes vias interagem entre si, e não como vias isoladas. Seguindo nessa filosofia de integração de dados, foram associados dados de expressão gênica do ciclo de vida do parasita para tentar entender o seu impacto nas redes metabólicas de *T. cruzi*.

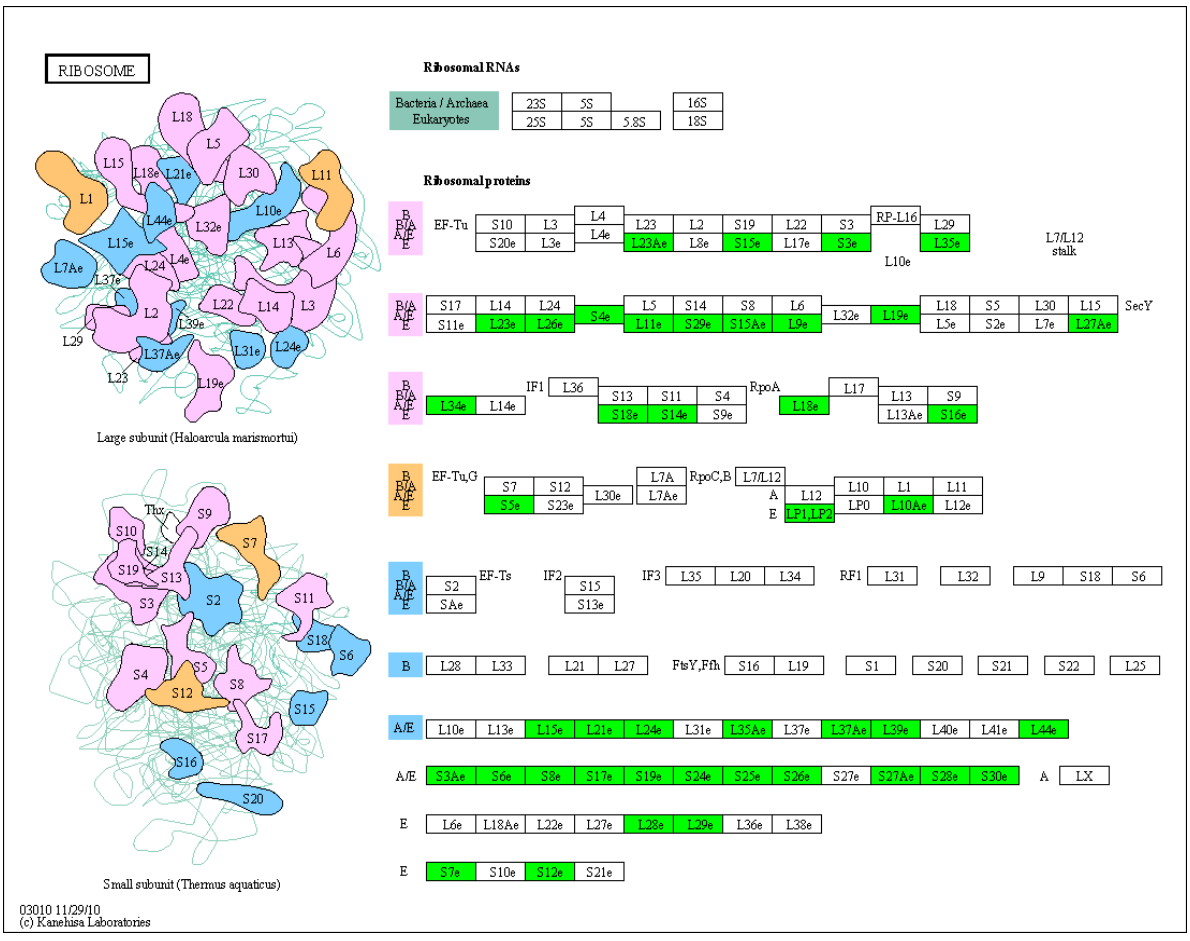
As diferentes análises que realizamos em relação à rede metabólica não conseguiram identificar padrões interessantes relacionados à regulação da expressão gênica. Isso se deve à diversos fatores, dentre os quais a alta conectividade da rede metabólica, o número relativamente grande de genes dentro de cada função metabólica (o que aumenta o número de nós conectados e arestas), o pequeno número de genes selecionados dentro de cada grupo que tinham anotação no KEGG (o que impede a realização de análises com poder estatístico) e ao estado de desconhecimento sobre a função dos genes de *T. cruzi,* visto que a

maioria das categorias analisadas tinham ~55% de proteínas hipotéticas, sendo que para algumas categorias esse porcentual chegava a 70%.

No entanto, embora as análises realizadas no presente trabalho não identificaram padrões interessantes sobre a regulação da expressão gênica associada ao metabolismo, a implementação da estrutura bioinformática para a realiazação dessas análises foi muito importante. Como exemplos, selecionamos dois casos no qual a integração entre KEGG e expressão gênica foi interessante.

Na FIGURA 108, vemos a ilustração dos dados de RNA-Seq comparando epimastigotas e tripomastigotas metacíclicos, no qual é possível evidenciar uma forte e clara diminuição na expressão de diversas proteínas ribossomais em metacíclicos. Esse tipo de análise está sendo utilizado na escrita da descrição das modulações evidenciadas no ciclo de vida de *T. cruzi,* o que está fora do escopo do presente trabalho, mas ilustra a utilização da plataforma construída.

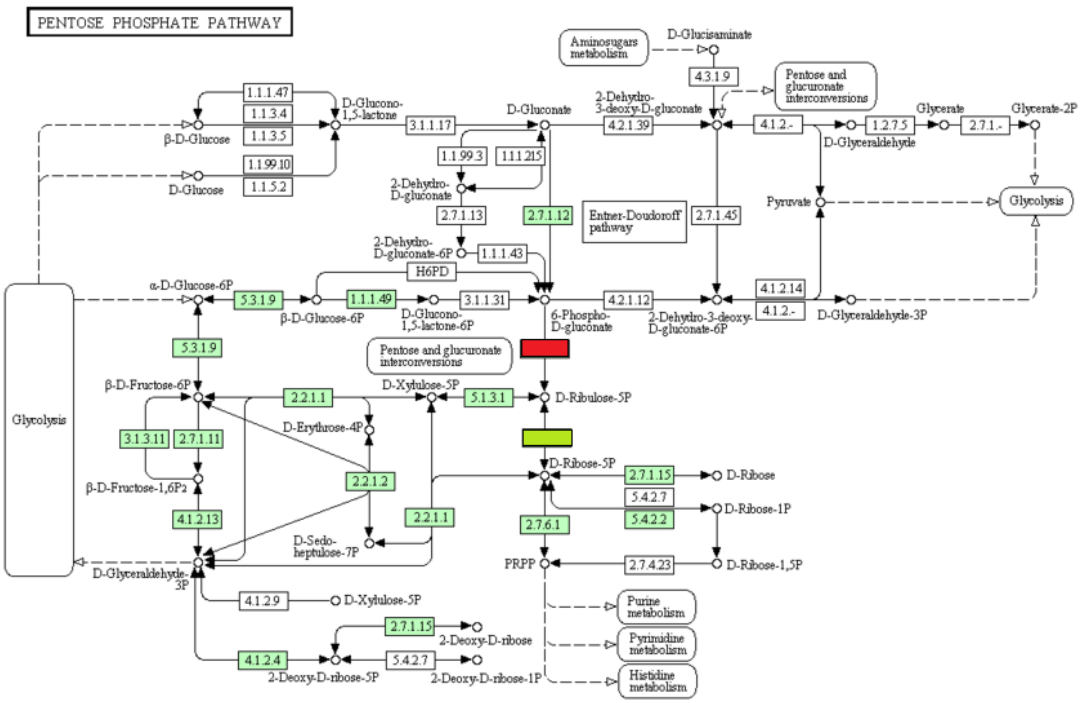
FIGURA 108 – Visualização do sistema Ribossomo (KEGG) na comparação entre Epi e Met.



Genes diminuídos em meta estão marcados em verde. Podemos obsevar a grande quantidade de genes modulados nessa comparação.

Na FIGURA 109, vemos a ilustração dos dados de RNA-Seq analisando a resposta de epimastigotas ao tratamento com inibidores da biossíntese do ergosterol. Na via da pentose-fosfato, só foi observado dois genes modulados, mas o padrão de modulação é bem interessante: os genes estão adjacentes e apresentam regulação inversa. Isso pode ser um interessante mecanismo de controle do fluxo metabólico. Esse tipo de análise é mais próxima ao que foi realizado no presente trabalho, no entanto não evidenciamos esse padrão com confiança estatística. Isso pode se dever a duas coisas: a análise da rede metabólica como um todo, e não em vias separadas, pode ter um grau de conectividade tão alto que as modulações evidenciadas acabam tendo uma propagação muito maior. Em segundo lugar, a análise do ciclo de vida pode evidenciar menos frequentemente modulações desse tipo, sendo que o uso de um inibidor farmacológico possa modular de forma mais específica esses fenômenos. Ou seja, o grau de modulação no ciclo de vida de *T. cruzi* é muito grande, e esse grau de modulação é utilizado na análise estatística. Portanto, esse é outro elemento que pode aumentar a propagação dos padrões aleatórios na rede, dificultando a identificação de padrões específicos com confiança estatística.

FIGURA 109 – Visualização da via pentose-fosfato na resposta à inibidores da biossíntese de ergosterol.



Dentro do contexto do Projeto Reguloma de *T. cruzi* do Instituto Carlos Chagas, esse trabalho contribuiu para a implementação inicial de uma base de dados que suporte essa quantidade e complexidade de informação; integrou dados de co-expressão para refinar a identificação de elementos regulatórios na 3'UTR dos transcritos; utilizou dados de RNA-seq para delimitar os sítios de inserção do mini-éxon e cauda poli-A nos transcritos; e incorporou de maneira mais eficaz a análise de redes metabólicas.

Portanto, consideramos que alcançamos o objetivo primário do trabalho que seria criar uma base computadorizada de armazenamento e análise de dados ômicos dentro do contexto de regulação pós-transcricional da expressão gênica em *T. cruzi,* principalmente. Além disso, a determinação dos limites dos mRNAs de *T. cruzi,* objetivo secundário mas essencial, representará um grande avanço na capacidade de realizarmos as análises referentes às redes regulatórias da expressão gênica.

7 CONCLUSÃO

Os principais desenvolvimentos tecnológicos e conclusões científicas do presente trabalho foram:

- Implementação de uma base de dados que comporte a quantidade e a complexidade dos dados que estão sendo gerados pelo Projeto Reguloma do Instituto Carlos Chagas;
- A análise combinada de redes metabólicas associada a dados de expressão gênica para identificar pontos importantes da regulação da expressão gênica;
- A identificação de genes marcadores para as fases do ciclo de vida de *T. cruzi*.
- A combinação da informação de genes co-expressos em cada fase do ciclo de vida para a identificação de elementos regulatórios na 3'UTR dos transcritos;
- A identificação de diversos possíveis elementos regulatórios associados a genes co-expressos e provavelmente co-regulados nas diferentes fases do ciclo de vida de *T. cruzi*;
- Estabelecimento de controles para uma melhor avaliação de possíveis elementos regulatórios nas 3'UTRs: presença do motivo em todo o genoma, nos genes não modulados, e na outra fase do ciclo de vida que não a que está sendo analisada;
- A delimitação das extremidades 5' e 3' dos transcritos de *T. cruzi* utilizando dados de RNA-seq de transcriptoma total, não enriquecidos para as extremidades
- Identificação de diversos padrões de trans-splicing alternativo em *T. cruzi*.

8 PERSPECTIVAS

Dado a filosofia prioritariamente de desenvolvimento tecnológico do presente trabalho, as realizações realizadas até o momento abrem um amplo conjunto de perspectivas associadas:

- Inserção dos dados ômicos gerados pelo Instituto no sistema implementado (KinetoDB) ;
- Criação de ferramentas web para visualização e mineração de dados do KinetoDB;
- Utilização das UTRs delimitadas nesse trabalho para fazer nova busca por motivos;
- Incorporar uma anotação *in house* de Gene Ontology (GO) e realizar a análise de enriquecimento de GO no grupo de genes que compartilham um mesmo motivo no 3’ UTR;
- Fazer busca por motivos no 5’UTR;
- Fazer busca por motivos nos genes diferencialmente expressos que são vizinhos na rede metabólica.
- Aprimorar o algoritmo de utilização do MEME para a predição de motivos primários.
- Incorporar análises relacionadas à evolução para a predição dos elementos reguladores, pela incorporação de uma grande variedade de genomas que serão sequenciados no futuro.
- Incorporar análises mais elaboradas, porém complexas, de estrutura secundária do RNA para a identificação dos elementos reguladores.

REFERÊNCIAS

ALCOLEA, P.J.; ALONSO, A.; GÓMEZ, M.J.; MORENO, I.; DOMÍNGUEZ, M.; PARRO, V.; LARRAGA, V. Transcriptomics throughout the life cycle of *Leishmania infantum*: high down-regulation rate in the amastigote stage. **International Journal for Parasitology**, v. 40, p. 1497-1516, 2010.

ANDRADE, L.O.; ANDREWS, N.W. The *Trypanosoma cruzi* – host-cell interplay: location, invasion, retention. **Nature Reviews Microbiology**, v. 3, p. 819-823, 2005.
ARCHER, S.K.; INCHAUSTEGUI, D.; QUEIROZ, R.; CLAYTON, C. The cell cycle regulated transcriptome of *Trypanosoma brucei*. **PLoS One**, v. 31, 6(3):e18425, 2011.

ASLETT, M.; AURRECOECHEA, C.; BERRIMAN, M.; BRESTELLI, J.; BRUNK, B.P.; CARRINGTON, M.; DEPLEDGE, D.P.; FISCHER, S.; GAJRIA, B.; GAO, X.; GARDNER, M.J.; GINGLE, A.; GRANT, G.; HARB, O.S.; HEIGES, M.; HERTZ-FOWLER, C.; HOUSTON, R.; INNAMORATO, F.; IODICE, J. *et al*. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research**, v. 38, (Database issue):D457-D462, 2010.

AZEVEDO, H.P.; ROITMAN, I. Cultivation of *Trypanosoma cruzi* in defined media. In: **Genes and Antigens of Parasites** – A laboratory manual, 2nd ed., Fundação Oswaldo Cruz, Rio de Janeiro, pg. 29-36, 1984.

BAILEY, T.; AND CHARLES, E. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. **Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology**, v. 2, p. 28-36, 1994.

BAPTISTA, C.S.; VÊNCIO, R.Z.; ABDALA, S.; VALADARES, M.P.; MARTINS, C.; DE BRAGANÇA PEREIRA, C.A.; ZINGALES, B. DNA microarrays for comparative genomics and analysis of gene expression in *Trypanosoma cruzi*. **Molecular and Biochemistry Parasitology**, v. 138, p. 183-94, 2004.

BELLI, S.I. Chromatin remodelling during the life cycle of trypanosomatids. **International Journal of Parasitology**, v. 30, p. 679-687, 2000.

BERRIMAN, M.; GHEDIN, E.; HERTZ-FOWLER, C.; BLANDIN, G.; RENAULD, H.; BARTHOLOMEU, D.C.; LENNARD, N.J.; CALER, E.; HAMLIN, N.E.; HAAS, B.; BÖHME, U.; HANNICK, L.; ASLETT, M.A.; SHALLOM, J.; MARCELLO, L.; HOU, L.; WICKSTEAD, B.; ALSMARK, U.C.; ARROWSMITH, C.; *et al*. The genome of the African trypanosome *Trypanosoma brucei.* **Science**, v. 15, p. 416-422, 2005.

BREMS, S.; GUILBRIDE, D.L.; GUNDLESDODJIR-PLANCK, D, BUSOLD C, LUU VD, SCHANNE M, HOHEISEL J, CLAYTON C. The transcriptomes of *Trypanosoma brucei* Lister 427 and TREU927 bloodstream and procyclic trypomastigotes. **Molecular and Biochemistry Parasitology**, v. 139, p. 163-172, 2005.

BRENER, Z. Terapêutica experimental na doença de Chagas. In: BRENER, Z., ANDRADE, Z., BARRAL-NETTO, M. ***Trypanosoma cruzi* e Doença de Chagas**, 2 ed., Guanabara Koogan, Rio de Janeiro, p. 379-388, 2000.

BRENER, Z.; CANÇADO, J.R.; GALVÃO, L.M.; DA LUZ, Z.M.P.; FILARDI, L.S.; PEREIRA, M.E.S.; SANTOS, L.M.T.; CANÇADO, C.B. An experimental and clinical assay with ketoconazole in the treatment of Chagas disease. **Memórias do Instituto Oswaldo Cruz**, v. 88, p. 149–53, 1993.

BRISSE, S.; BARNABE, C.; TIBAYRENC, M. Identification of six *Trypanosoma cruzi* phylogenetic lineages by random amplified polymorphic DNA and multilocus enzyme electrophoresis. **Internation Journal for Parasitology**, v. 30, p. 35-44, 2000.

BRITTO, C.; RAVEL, C.; BASTIEN, P.; BLAINEAU, C.; PAGÈS, M.; DEDET, J.P.; WINCKER, P. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes. **Gene**, v. 222, p. 107-117, 1998.

CAMPBELL, D.A.; THOMAS, S.E.; STURM, N.R. Transcription in kinetoplastid protozoa: why be normal? **Microbes Infection**, v. 5, p. 1231-1240, 2003.

CANO, M.I.; GRUBER, A.; VAZQUEZ, M.; CORTÉS, A.; LEVIN, M.J.; GONZÁLEZ, A.; DEGRAVE, W.; RONDINELLI, E.; ZINGALES, B.; RAMIREZ, J.L.; ALONSO, C.; REQUENA, J.M.; DA SILVEIRA, J.F. Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* genome project. **Molecular Biochemistry and Parasitology,** v. 71, p. 273-278, 1995.

CAVALIER-SMITH, T. Kingdom protozoa and its 18 phyla. **Microbiological Reviews**, v. 57, p. 953-994, 1993.

CAVALIER-SMITH, T. Kingdoms protozoa and chromista and the eozoan root of the eukaryotic tree. **Biology Letters**, v. 6, p. 342-345, 2010.

Center for Disease Control. In: Laboratory Identification of Parasites of Public Health Concern. <http://www.dpd.cdc.gov/dpdx/HTML/TrypanosomiasisAmerican.htm>

CHAGAS, C. Nova tripanosomiase humana: Estudos sobre a morfologia e o ciclo evolutivo do *Schizotrypanum cruzi* n. gen., n. sp., agente etiológico de nova entidade mórbida do homem. **Memórias do Instituto Oswaldo Cruz**, v.1, p.159-218, 1909.

CHEN, J.M.; FÉREC, C.; COOPER, D.N. A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3' UTR variants. **Human Genetics**, v. 120(3), p. 301-33, 2006.

CLAYTON, C.; SHAPIRA, M. Post-transcriptional regulation of gene expression in trypanosomes and *Leishmania*s. **Molecular Biochemistry and Parasitology**, v. 156, p. 93-101, 2007.

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

COURA, J.R. Chagas disease: what is known and what is needed – A background article. **Memórias do Instituto Oswaldo Cruz**, v. 102, p. 113-122, 2007.

COURA, J.R. Transmission of chagasic infection by oral route in the natural history of

Chagas' disease. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 39, p. 113-117, 2006.

COURA, J.R.; DE CASTRO, S.L. A Critical Review on Chagas Disease Chemoterapy. **Memórias do Instituto Oswaldo Cruz**, v. 97, p. 3-24, 2002.

DA SILVA, C.V.; KAWASHITA, S.Y.; PROBST, C.M.; DALLAGIOVANNA, B.; CRUZ, M.C.; DA SILVA, E.A.; SOUTO-PADRÓN, T.C.; KRIEGER, M.A.; GOLDENBERG, S.; BRIONES, M.R.; ANDREWS, N.W.; MORTARA, R.A. Characterization of a 21kDa protein from *Trypanosoma cruzi* associated with mammalian cell invasion. **Microbes and Infection**, v. 11, p. 563-570, 2009.

DALLAGIOVANNA ,B.; CORREA, A.; PROBST, C.M.; HOLETZ, F.; SMIRCICH, P.; DE AGUIAR, A.M.; MANSUR, F.; DA SILVA, C.V.; MORTARA, R.A.; GARAT, B.; BUCK, G.A.; GOLDENBERG, S.; KRIEGER, M.A. Functional genomic characterization of mRNAs associated with TcPUF6, a pumilio-like protein from *Trypanosoma cruzi*. **The Journal of Biological Chemistry**, v. 28, p. 8266-8273, 2008.

DE GAUDENZI, J.; FRASCH, A.C.; CLAYTON, C. RNA-binding domain proteins in Kinetoplastids: a comparative analysis. **Eukaryotic Cell**, v. 4, p. 2106-2114, 2005.

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

DIAS, J.C.P. Epidemiologia. In: BRENNER, Z.; ANDRADE, A.Z.; BARRAL-NETTO, M. (eds). ***Trypanosoma cruzi* e Doença de Chagas**. 2.ª Ed. Guanabara Koogan SA. Rio de Janeiro, 2000.

D'ORSO, I.; DE GAUDENZI, J.G.; FRASCH, A.C. RNA-binding proteins and mRNA turnover in trypanosomes. **Trends in Parasitology**, v. 19, p. 151-155, 2003.

DIEHL S, DIEHL F, EL-SAYED NM, CLAYTON C, HOHEISEL JD. Analysis of stage-specific gene expression in the bloodstream and the procyclic form of *Trypanosoma brucei* using a genomic DNA-microarray. **Molecular and Biochemistry Parasitology**, v. 28, p. 115-123, 2002.

DO MONTE-NETO, R.L.; COELHO, A.C.; RAYMOND, F.; LÉGARÉ, D.; CORBEIL, J.; MELO, M.N.; FRÉZARD, F.; OUELLETTE, M. Gene expression profiling and molecular characterization of antimony resistance in *Leishmania amazonensis*. **PLoS Neglected Tropical Disease**, v. 5(5):e1167, 2011.

D'ORSO, I.; FRASCH, A.C. Functionally different AU- and G-rich cis-elements confer developmentally regulated mRNA stability in *Trypanosoma cruzi* by interaction with specific RNA-binding proteins. **Journal of Biological Chemistry**, v. 276, p. 15783-15793, 2001.

DOWNING, T.; IMAMURA, H.; DECUYPERE, S.; CLARK, T.G.; COOMBS, G.H.; COTTON, J.A.; HILLEY, J.D.; DE DONCKER, S.; MAES, I.; MOTTRAM, J.C.; QUAIL, M.A.; RIJAL, S.; SANDERS, M.; SCHÖNIAN, G.; STARK, O.; SUNDAR, S.; VANAERSCHOT, M.; HERTZ-FOWLER, C.; DUJARDIN, J.C.; BERRIMAN, M. Whole

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

DUMONTEIL, E. Vaccine development against *Trypanosoma cruzi* and *Leishmania* species in the post-genomic era. **Infection, Genetics and Evolution**, v. 9, p. 1075-1082, 2009.

ELEMENTO, O.; SLONIM, N., TAVAZOIE, S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. **Molecular Cell**, v. 28, p. 337-350, 2007.

EL-SAYED, N.M.; MYLER, P.J.; BARTHOLOMEU, D.C.; NILSSON, D.; AGGARWAL, G.; TRAN, A.N.; GHEDIN, E.; WORTHEY, E.A.; DELCHER, A.L.; BLANDIN, G.; WESTENBERGER, S.J.; CALER, E.; CERQUEIRA, G.C.; BRANCHE, C.; HAAS, B.; ANUPAMA, A.; ARNER, E.; ASLUND, L.; ATTIPOE, P.; BONTEMPI, E. *et al*. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science**, v. 309, p. 409-415, 2005a.

EL-SAYED, N.M.; MYLER, P.J.; BLANDIN, G.; BERRIMAN, M.; CRABTREE, J.; AGGARWAL, G.; CALER, E.; RENAULD, H.; WORTHEY, E.A.; HERTZ-FOWLER, C.; GHEDIN, E.; PEACOCK, C.; BARTHOLOMEU, D.C.; HAAS, B.J.; TRAN, A.N.; WORTMAN, J.R.; ALSMARK, U.C.; ANGIUOLI, S.; ANUPAMA, A. *et al*. Comparative genomics of trypanosomatid parasitic protozoa. **Science**, v. 309, p. 404-409, 2005b.

ENGSTLER, M.; BOSHART, M. Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in *Trypanosoma brucei*. **Genes and Development,** v. 18, p. 2798-2811, 2004.

FENG, X.; FEISTEL, T.; BUFFALO, C.; MCCORMACK, A.; KRUVAND, E.; RODRIGUEZ-CONTRERAS, D.; AKOPYANTS, N.S.; UMASANKAR, P.K.; DAVID, L.; JARDIM, A.; BEVERLEY, S.M.; LANDFEAR, S.M. Remodeling of protein and mRNA expression in *Leishmania mexicana* induced by deletion of glucose transporter genes. **Molecular and Biochemistry Parasitology,** v. 175, p. 39-48, 2011.

FILARDI, L.S.; BRENER, Z. Susceptibility and natural resistance of *Trypanosoma cruzi* strains to drugs used clinically in Chagas disease. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 81, p. 755–759, 1987.

FRANCO DE GODOY, L.M.; MARCHINI, F.K.; PAVONI, D.P.; DE CÁSSIA PONTELLO RAMPAZZO, R.; PROBST, C.M..; GOLDENBERG, S.; KRIEGER, M.A. Quantitative proteomics of *Trypanosoma cruzi* during metacyclogenesis. **Proteomics**, Jul 4. doi: 10.1002/pmic.201200078. [Epub ahead of print], 2012.

FRANZÉN, O.; OCHAYA, S.; SHERWOOD, E.; LEWIS, M.D.; LLEWELLYN, M.S.; MILES, M.A.; ANDERSSON, B. Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio X10/1 and comparison with *T. cruzi* VI CL Brener. **PLoS Neglected Tropical Diseases**, v. 8, 5(3):e984, 2011.

GARCIA, A.; COURTIN, D.; SOLANO, P.; KOFFI, M.; JAMONNEAU, V. Human African trypanosomiasis: connecting parasite and host genetics. **Trends in Parasitology**, v. 22, p. 405-409, 2006.

GENERIC MODEL ORGANISM DATABASE PROJECT. http://gmod.org/wiki/Main_Page

GOODARZI, H.; NAJAFABADI, H.S.; OIKONOMOU, P.; GRECO, T.M.; FISH, L.; SALAVATI, R.; CRISTEA, I.M.; TAVAZOIE, S. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. **Nature**, v. 8, p. 264-268, 2012.

GRANT, C.E.; BAILEY, T.L.; NOBLE, W.S. FIMO: Scanning for occurrences of a given motif. **Bioinformatics**, v. 27, p. 1017–1018, 2011.

GRYNBERG, P.; PASSOS-SILVA, D.G.; MOURÃO, M.D.E., HIRATA, J.R.; MACEDO, A.M.; MACHADO, C.R.; BARTHOLOMEU, D.C.; FRANCO, G.R. *Trypanosoma cruzi* gene expression in response to gamma radiation. **PLoS One**, v. 7(1):e29596, 2012.

HAFNER, M.; LANDTHALER, M.; BURGER, L.; KHORSHID, M.; HAUSSER, J.; BERNINGER, P.; ROTHBALLER, A.; ASCANO, J.R.; JUNGKAMP, A.C.; MUNSCHAUER, M.; ULRICH, A.; WARDLE, G.S.; DEWELL, S.; ZAVOLAN, M.; TUSCHL, T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. **Cell**, v. 141, p. 129-141, 2010.

HAILE, S.; PAPADOPOULOU, B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. **Current Opinion in Microbiology**, v. 10, p. 569-577, 2007.

HALL, N.; BERRIMAN, M.; LENNARD, N.J.; HARRIS, B.R.; HERTZ-FOWLER, C.; BART-DELABESSE, E.N.; GERRARD, C.S.; ATKIN, R.J.; BARRON, A.J.; BOWMAN, S.; BRAY-ALLEN, S.P.; BRINGAUD, F.; CLARK, L.N.; CORTON, C.H.; CRONIN, A.; DAVIES, R.; DOGGETT, J.; FRASER, A. *et al*. The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organization, recombination and polymorphism. **Nucleic Acids Research**, v. 31, p. 4864-4873, 2003.

HARTMANN, C.; BENZ, C.; BREMS, S.; ELLIS, L.; LUU, V.D.; STEWART, M.; D'ORSO, I.; BUSOLD, C.; FELLENBERG, K.; FRASCH, A.C.; CARRINGTON, M.; HOHEISEL, J.; CLAYTON, C.E. Small trypanosome RNA-binding proteins TbUBP1 and TbUBP2 influence expression of F-box protein mRNAs in bloodstream trypanosomes. **Eukaryotic Cell**, v. 6, p. 1964-1978, 2007.

HIDE, G. History of sleeping sickness in East Africa. **Clinical Microbiology Reviews**, v.12, p. 112-25, 1999.

HOLETZ, F.B.; ALVES, L.R.; PROBST, C.M.; DALLAGIOVANNA, B.; MARCHINI, F.K.; MANQUE, P.; BUCK, G.; KRIEGER, M.A.; CORREA, A.; GOLDENBERG, S. Protein and mRNA content of TcDHH1-containing mRNPs in *Trypanosoma cruzi.* **FEBS Journal**, v. 277, p. 3415-3426, 2010.

HOLZER, T.R.; MCMASTER, W.R.; FORNEY, J.D. Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana*. **Molecular and Biochemistry Parasitology,** v. 146, p. 198-218, 2006.

HOTEZ, P.J.; PECOUL, B. "Manifesto" for advancing the control and elimination of neglected tropical diseases. **PLoS Neglected Tropical Diseases**, v. 4(5):e718, 2010.

IRMER, H.; CLAYTON, C. Degradation of the unstable EP1 mRNA in *Trypanosoma brucei* involves initial destruction of the 3'-untranslated region. **Nucleic Acids Research**, v. 29, p. 4707-4715, 2001.

IVENS, A.C.; PEACOCK, C.S.; WORTHEY, E.A.; MURPHY, L.; AGGARWAL, G.; BERRIMAN, M.; SISK, E.; RAJANDREAM, M.A.; ADLEM, E.; AERT, R.; ANUPAMA, A.; APOSTOLOU, Z.; ATTIPOE, P.; BASON, N.; BAUSER, C.; BECK, A.; BEVERLEY, S.M.; BIANCHETTIN, G.; BORZYM, K.; BOTHE, G. *et al*. The genome of the kinetoplastid parasite *Leishmania major*. **Science**, v. 309, p. 436-442, 2005.

JACKSON, A.P.; SANDERS, M.; BERRY, A.; McQUILLAN, J.; ASLETT, M.A.; CHUKUALIM, B.; CAPEWELL, P.; MacLEOD, A.; MELVILLE, S.E.; GIBSON, W.; BARRY, J.D.; BERRIMAN, M.; HERTZ-FOWLER, C. The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human african trypanosomiasis. **PLoS Neglected Tropical Diseases**, 4: e6, 2010.

JÄGER, A.V.; De GAUDENZI, J.G.; CASSOLA, A.; D'ORSO, I.; FRASCH. A.C. mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes. **Proceedings of the National Academy of Science of the USA,** v. 104, p. 2035-2042, 2007.

JUNQUEIRA, A. C.; DEGRAVE, W.; BRANDÃO, A. Minicircle organization and diversity in *Trypanosoma cruzi* populations. **Trends in Parasitology**, v. 21, p. 270-272, 2005.
KABANI, S.; FENN, K.; ROSS, A.; IVENS, A.; SMITH, T.K.; GHAZAL, P.; MATTHEWS, K. Genome-wide expression profiling of in vivo-derived bloodstream parasite stages and dynamic analysis of mRNA alterations during synchronous differentiation in *Trypanosoma brucei*. **BMC Genomics**, 10:427, 2009.

KERTESZ, M.; WAN, Y.; MAZOR, E.; RINN, J.L.; NUTTER, R.C.; CHANG, H.Y.; SEGAL, E. Genome-wide measurement of RNA secondary structure in yeast. **Nature**, v. 467(7311), p. 103-107, 2010.

KOLEV, N.G.; FRANKLIN, J.B.; CARMI, S.; SHI, H.; MICHAEL, S.; TSCHUDI, C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. **PLoS Pathogens**, v. 6(9):e1001090, 2010.

KOLEV, N.G.; FRANKLIN, J.B.; CARMI, S.; SHI, H.; MICHAELI, S.; TSCHUDI, C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. **PLoS Pathogens**, v. 9(9):e1001090, 2010.

KOUMANDOU, V.L.; NATESAN, S.K.; SERGEENKO, T.; FIELD, M.C. The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages. **BMC Genomics**, 9:298, 2008.

LAI, D.H.; HASHIMI, H.; LUN, Z.R.; AYALA, F.J.; LUKES, J. Adaptations of *Trypanosoma brucei* to gradual loss of kinetoplast DNA: *Trypanosoma equiperdum* and *Trypanosoma evansi* are petite mutants of *T. brucei*. **Proceedings of the National Academy of Sciences of the USA**, v.105, p.1999-2004, 2008.

LUU, V.D.; BREMS, S.; HOHEISEL, J.D.; BURCHMORE, R.; GUILBRIDE, D.L.; CLAYTON, C. Functional analysis of *Trypanosoma brucei* PUF1. **Molecular and Biochemistry Parasitology**, v. 150, p. 340-349, 2006. Erratum in: **Molecular and Biochemistry Parasitology,** v. 169, p. 70, 2010.

MAIR, G.; SHI, H.; LI, H.; DJIKENG, A.; AVILES, H.O.; BISHOP, J.R.; FALCONE, F.H.; GAVRILESCU, C.; MONTGOMERY, J.L.; SANTORI, M.I.; STERN, L.S.; WANG, Z.; ULLU, E.E.; TSCHUDI, C. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. **RNA**, v. 6, p.163-169, 2000.

MANFUL, T.; FADDA, A.; CLAYTON, C. The role of the 5'-3' exoribonuclease XRNA in transcriptome-wide mRNA degradation. **RNA**, v. 17, p. 2039-2047, 2011.

MARCHINI, F.K.; DE GODOY, L.M.; RAMPAZZO, R.C.; PAVONI, D.P.; PROBST, C.M.; GNAD. F.; MANN, M.; KRIEGER, M.A. Profiling the *Trypanosoma cruzi* phosphoproteome. **PLoS One**, 6(9):e25381, 2011.

MAYA, J.D.; CASSELS, B.K.; ITURRIAGA-VÁSQUEZ, P.; FERREIRA, J.; FAÚNDEZ, M. *et al*. Mode of action of natural and synthetic drugs against *Trypanosoma cruzi* and their interaction with the mammalian host. **Comparative Biochemistry and Physiology**, v. 146, p. 601-620, 2007.

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**, v. 6, p. 3567-3581, 2006.

MICHAELI, S.; DONIGER, T.; GUPTA, S.K.; WURTZEL, O.; ROMANO, M.; VISNOVEZKY, D.; SOREK, R.; UNGER, R.; ULLU, E. RNA-seq analysis of small RNPs in *Trypanosoma brucei* reveals a rich repertoire of non-coding RNAs. **Nucleic Acids Research**, v. 40, p.1282-1298, 2012.

MINNING, T.A.; BUA, J.; GARCIA, G.A.; MCGRAW, R.A.; TARLETON, R.L. Microarray profiling of gene expression during trypomastigote to amastigote transition in *Trypanosoma cruzi*. **Molecular and Biochemistry Parasitology**, v. 131, p. 55-64, 2003.

MINNING, T.A.; WEATHERLY, D.B.; ATWOOD, J.; ORLANDO, R.; TARLETON, R.L.The steady-state transcriptome of the four major life-cycle stages of *Trypanosoma cruzi*. **BMC Genomics**, 10:370, 2009.

MOREIRA, D.; VON DER HEYDEN, S.; BASS, D.; LÓPEZ-GARCIA, P.; CHAO, E.; CAVALIER-SMITH, T. Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. **Molecular Phylogenetics and Evolution**, v. 44, p. 255-266, 2007.

MUNGALL, C.J.; EMMERT, D.B.; AND THE FLYBASE CONSORTIUM. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. **Bioinformatics,** v. 23, p. 337-346, 2007.

MURTA, S.M.; KRIEGER, M.A.; MONTENEGRO, L.R.; CAMPOS, F.F.; PROBST, C.M.; AVILA, A.R.; MUTO, N.H.; DE OLIVEIRA, R.C.; NUNES, L.R.; NIRDÉ, P.; BRUNA-

ROMERO, O.; GOLDENBERG, S.; ROMANHA, A.J. Deletion of copies of the gene encoding old yellow enzyme (TcOYE), a NAD(P)H flavin oxidoreductase, associates with in vitro-induced benznidazole resistance in *Trypanosoma cruzi*. **Molecular and Biochemistry Parasitology,** v. 146, p. 151-162, 2006.

NEVES, D.P. **Parasitologia Humana**. 8ª Ed. Atheneu. São Paulo, 1991.

OPEN BIOLOGICAL AND BIOMEDICAL ONTOLOGY (OBO). <http://www.obofoundry.org>

OPPERDOES, F.R.; MICHELS, P.A. Horizontal gene transfer in trypanosomatids. **Trends in Parasitology**, v. 23, p. 470-476, 2007.

OVERATH, P.; HAAG, J.; LISCHKE, A.E.; O'HUIGIN, C. The surface structure of trypanosomes in relation to their molecular phylogeny. **International Journal of Parasitology**, v. 31, p. 468-471, 2001.

PARKER, B.J.; MOLTKE, I.; ROTH, A.; WASHIETL, S.; WEN, J.; KELLIS, M.; BREAKER, R.; PEDERSEN, J.S.. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. **Genome Research**, v. 21(11), p. 1929-1943, 2011.

PEACOCK, C.S.; SEEGER, K.; HARRIS, D.; MURPHY, L.; RUIZ, J.C.; QUAIL, M.A.; PETERS, N.; ADLEM, E.; TIVEY, A.; ASLETT, M.; KERHORNOU, A.; IVENS, A.; FRASER, A.; RAJANDREM, M.A.; CARVER, T.; NORBERTCZAK, H.; CHILLINGWORTH, T.; HANCE, Z.; JAGELS, K.; MOULE, S.; ORMOND, D.; RUTTER, S.; SQUARES, R.; WHITEHEAD, S.; RABBINOWITSCH, E.; ARROWSMITH, C.; WHITE, B.; THURSTON, S.; BRINGAUD, F.; BALDAUF, S.L.; FAULCONBRIDGE, A.; JEFFARES, D.; DEPLEDGE, D.P.; OVOLA, S.O.; HILLEY, J.D.; BRITO.; L.O.; TOSI, L.R.; BARREL, B.; CRUZ, A.K.; MOTTRAM, J.C.; SMITH, D.F.; BERRIMAN, M. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. **Nature Genetics**, v. 39, p. 839-847, 2007.

PostgreSQL GLOBAL DEVELOPMENT GROUP. <http://www.postgresql.org>

QUEIROZ, R.; BENZ, C.; FELLENBERG, K.; HOHEISEL, J.D.; CLAYTON, C. Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. **BMC Genomics,** 10:495, 2009.

RAYMOND, F.; BOISVERT, S.; ROY, G.; RITT, J.F.; LÉGARÉ, D.; ISNARD, A.; STANKE, M.; OLIVIER, M.; TREMBLAY, M.J.; PAPADOPOULOU, B.; OUELLETTE, M.; CORBEIL, J. Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. **Nucleic Acids Research**, v. 40, p. 1131-1147, 2012.

ROBINSON, M.D.; McCARTHY, D.J.; SMYTH, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, p. 139-140, 2010.

ROCHETTE, A.; RAYMOND, F.; UBEDA, J.M.; SMITH, M.; MESSIER, N.; BOISVERT, S.; RIGAULT, P.; CORBEIL, J.; OUELLETTE, M.; PAPADOPOULOU, B. Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between

the two species. **BMC Genomics**, 9:255, 2008.

RUMBLE, S.M.; LACROUTE, P.; DALCA, A.V.; FIUME, M.; SIDOW, A.; *et al*. SHRiMP: Accurate Mapping of Short Color-space Reads. **PLoS Computational Biology,** 5(5): e1000386, 2009.

SAXENA, A.; LAHAV, T.; HOLLAND, N.; AGGARWAL, G.; ANUPAMA, A.; HUANG, Y.; VOLPIN, H.; MYLER, P.J.; ZILBERSTEIN, D. Analysis of the *Leishmania donovani* transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. **Molecular and Biochemistry Parasitology**, v. 152, p. 53-65, 2007.

SAXENA, A.; WORTHEY, E.A.; YAN, S.; LELAND, A.; STUART, K.D.; MYLER, P.J. Evaluation of differential gene expression in *Leishmania major* Friedlin procyclics and metacyclics using DNA microarray analysis. **Molecular and Biochemistry Parasitology,** v. 129, p. 103-114, 2003.

SERPELONI, M.; MORAES, C.B.; MUNIZ, J.R.; MOTTA, M.C.; RAMOS, A.S.; KESSLER, R.L.; INOUE, A.H.; daROCHA, W.D.; YAMADA-OGATTA, S,F,; FRAGOSO, S.P.; GOLDENBERG, S.; FREITAS-JUNIOR, L.H.; AVILA, A.R. An essential nuclear protein in trypanosomes is a component of mRNA transcription/export pathway. **PLoS One**, v. 6(6):e20730, 2011a.

SERPELONI, M.; VIDAL, N.M.; GOLDENBERG, S.; AVILA, A.R.; HOFFMANN, F.G. Comparative genomics of proteins involved in RNA nucleocytoplasmic export. **BMC Evolutionary Biology**, 11:7, 2011b.

SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N.S.; WANG, J.T.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome Research**, v. 13, p. 2498-504, 2003.

SHAPIRO, T. A.; ENGLUND, P. T. The structure and replication of kinetoplast DNA. **Annual Review of Microbiology,** v. 49, p. 117-43, 1995.
SIEGEL, T.N.; HEKSTRA, D.R.; WANG, X.; DEWELL, S.; CROSS, G.A. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. **Nucleic Acids Research**, v. 38, p. 4946-4957, 2010.

SIEGEL, T.N.; HEKSTRA, D.R.; WANG, X.; DEWELL, S.; CROSS, G.A. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. **Nucleic Acids Research**, v. 38(15), p. 4946-4957, 2010.

SINGH, N.; ALMEIDA, R.; KOTHARI, H.; KUMAR, P.; MANDAL, G.; CHATTERJEE, M.; VENKATACHALAM, S.; GOVIND, M.K.; MANDAL, S.K.; SUNDAR, S. Differential gene expression analysis in antimony-unresponsive Indian kala azar (visceral *Leishmania*sis) clinical isolates by DNAmicroarray. **Parasitology**, v. 134, p. 777-787, 2007.

SONENBERG, N.; DEVER, T.E. Eukaryotic translation initiation factors and regulators. **Current Opinion in Structural Biology**, v. 13, p. 56-63, 2003.

SOUZA, R.T.; LIMA, F.M.; BARROS, R.M.; CORTEZ, D.R.; SANTOS, M.F.; CORDERO, E.M.; RUIZ, J.C.; GOLDENBERG, S.; TEIXEIRA, M.M.; DA SILVEIRA, J.F. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. **PLoS One**. V. 6, 2011, (e23042).

SRIVIDYA, G.; DUNCAN, R.; SHARMA, P.; RAJU, B.V.; NAKHASI, H.L.; SALOTRA, P. Transcriptome analysis during the process of in vitro differentiation of *Leishmania donovani* using genomic microarrays. **Parasitology**, v. 134, p. 1527-1539, 2007.

STEIN, L.D.; MUNGALL, C.; SHU, S.; CAUDY, M.; MANGONE, M.; DAY, A.; NICKERSON, E.; STAJICH, J.E.; HARRIS, T.W.; ARVA, A.; LEWIS, S. The generic genome browser: a building block for a model organism system database. **Genome Research**, v. 12, p. 1599-1610, 2002.

STURM, N. R.; VARGAS, N. S.; WESTENBERGER, S. J.; ZINGALES, B.; CAMPBELL, D. A. Evidence for multiple hybrid groups in *Trypanosoma cruzi*. **International Journal for Parasitology**, v. 33, p. 269-279, 2003.
STURM, N.R.; CAMPBELL, D.A. Alternative lifestyles: The population structure of *Trypanosoma cruzi*. **Acta Tropica**, v. 115, p. 35-43, 2009.

TEIXEIRA, A.R.; NASCIMENTO, R.J.; STURM, N.R. Evolution and pathology in Chagas disease – A Review. **Memórias do Instituto Oswaldo Cruz**, v. 101, p. 463-491, 2006.
TELLERIA, J.; LAFAY, B.; VIRREIRA, M.; BARNABÉ, C.; TIBAYRENC, M.; SVOBODA, M. *Trypanosoma cruzi*: Sequence analysis of the variable region of kinetoplast minicircles. **Experimental Parasitology,** v. 114, p. 279–88, 2006.

THE APACHE SOFTWARE FOUNDATION. Apache. http://www.apache.org

THE PERL PROGRAMMING LANGUAGE. Perl5. <http://www.perl.org>

THE R PROJECT FOR STATISTICAL COMPUTING. R. <www.r-project.org>

TIBAYRENC, M. Genetic subdivisions within *Trypanosoma cruzi* (discrete typing units) and their relevance for molecular epidemiology and experimental evaluation. **Kinetoplastid Biology and Disease**, v. 1, 2-12, 2003.
ULE, J.; JENSEN, K.B.; RUGGIU, M.; MELE, A.; ULE, A.; DARNELL, R.B. CLIP identifies Nova-regulated RNA networks in the brain. **Science**, v. 14, p. 1212-1215, 2003.

VAN DONGEN, S. A cluster algorithm for graphs. **Technical Report INS-R0010,** National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 1998.

VASCONCELOS, E.J.; TERRÃO, M.C.; RUIZ, J.C.; VÊNCIO, R.Z.; CRUZ, A.K. *In silico* identification of conserved intercoding sequences in *Leishmania* genomes: unraveling putative *cis*-regulatory elements. **Molecular Biochemical Parasitology**, v. 183(2), p. 140-150, 2012.

VEITCH, N.J.; JOHNSON, P.C.; TRIVEDI, U.; TERRY S.; WILDRIDGE, D.; MACLEOD, A. Digital gene expression analysis of two life cycle stages of the human-infective parasite, *Trypanosoma brucei gambiense* reveals differentially expressed clusters of

co-regulated genes. **BMC Genomics**, 11-124, 2010.

WINCKER, P.; RAVEL, C.; BLAINEAU, C.; PAGES, M.; JAUFFRET, Y.; DEDET, J.P. The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. **Nucleic Acids Research**, v. 24, p. 1688-1694, 1996.

WRAY, G.A.; HAHN, M.W.; ABOUHEIF, E.; BALHOFF, J.P.; PIZER, M.; ROCKMAN, M.V.; ROMANO, L.A. The evolution of transcriptional regulation in eukaryotes. **Molecular Biology and Evolution**, v. 20, p.1377-1419, 2003.

YEO, M.; ACOSTA, N.; LLEWELLYN, M.; SÁNCHEZ, ADAMSON, S. *et al*. Origins of Chagas disease: Didelphis species are natural hosts of *Trypanosoma cruzi* I and armadillos hosts of *Trypanosoma cruzi* II, including hybrids. **International Journal for Parasitology**, v. 35, p. 225–233, 2005.

YOON, H.S.; GRANT, J.; TEKLE, Y.I.; WU, M.; CHAON, B.C.; COLE, J.C.; LOGSDON, J.M.; PATTERSON, D.J.; BHATTACHARYA, D.; KATZ, L.A. Broadly sampled multigene trees of eukaryotes. **BMC Evolutionary Biology,** v. 18, p. 8-14, 2008.

YOSHIDA, N. Molecular mechanism of *Trypanosoma cruzi* infection by oral route. **Memórias do Instituto Oswaldo Cruz**, v. 104, p. 101-107, 2009.

ZINGALES, B.; ANDRADE, S. G.; BRIONES, M. R.; CAMPBELL, D. A.; CHIARI, E.; FERNANDES, O.; GUHL, F.; LAGES-SILVA, E.; MACEDO, A. M.; MACHADO, C. R.; MILES, M. A.; ROMANHA, A. J.; STURM, N. R.; TIBAYRENC, M.; SCHIJMAN, A. G. *et al*. New consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. **Memórias do Instituto Oswaldo Cruz**, v. 104, p. 1051-1054, 2009.

ZINGALES, B.; MILES, M.A.; CAMPBELL, D.A.; TIBAYRENC, M.; MACEDO, A.M.; TEIXEIRA, M.M.; SCHIJMAN, A.G.; LLEWELLYN, M.S.; LAGES-SILVA, E.; MACHADO, C.R.; ANDRADE, S.G.; STURM, N.R. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. **Infection, Genetics and Evolution**, v. 12, p. 240-53, 2012.

ZINGALES, B.; STOLF, B. S.; SOUTO, R. P.; FERNANDES, O.; BRIONES, M. R. S. Epidemiology, biochemistry and evolution of *Trypanosoma cruzi* lineages based on ribosomal RNA sequences. **Memórias do Instituto Oswaldo Cruz**, v. 94, p. 159-164, 1999.