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

Identificação de marcadores moleculares de diagnóstico e virulência em *Leishmania* por bioinformática

Gabriela Flávia Rodrigues Luiz

Belo Horizonte
2015

Gabriela Flávia Rodrigues Luiz

Identificação de marcadores moleculares de diagnóstico e virulência em *Leishmania* por bioinformática

Tese apresentada como requisito parcial para a obtenção do título de Doutor em Bioinformática pelo programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais.

Orientadora: Prof.ª Dr.ª Daniella Castanheira Bartholomeu

Belo Horizonte
2015

"Bear with us, while we think."

— http://slow-science.org/

“The problem with the world is that the intelligent people are full of doubts, while the stupid ones are full of confidence.”

— Charles Bukowski

AGRADECIMENTOS

À Profª. Dra. Daniella Castanheira Bartholomeu, pela oportunidade e disponibilidade desde o início da Iniciação Científica, e pela excelência de orientação, fatores imprescindíveis para a realização desse trabalho.

À minha família, pelo apoio, carinho e compreensão incondicionais durante todos os altos e baixos deste caminho.

À Thays, pelo carinho, paciência (nem sempre muito grande!) e por deixar minha vida mais divertida.

Aos membros da banca pela disponibilidade e contribuição crítica no aperfeiçoamento do trabalho.

Às pessoas que contribuíram diretamente no projeto, Mariana Cardoso, Robson, Eduardo Coelho, Daniel, Mariana Duarte e Thiago Souza.

Aos amigos do LIGP pelas sugestões que contribuíram para o desenvolvimento do trabalho, pela descontração e por compartilharem das mesmas dúvidas e angústias.

Às agências de fomento pelo apoio financeiro que tornou possível a realização desse projeto.

O meu sincero agradecimento a todos aqueles que tornaram possível a realização deste trabalho.

SUMÁRIO

RESUMO

Após 10 anos da publicação do primeiro genoma completo de uma espécie de *Leishmania*, os avanços na tecnologia de sequenciamento proporcionaram um grande acúmulo de informações genômicas, transcriptômicas e proteômicas do gênero. A convergência destas informações ômicas aliadas a ferramentas de bioinformática permite auxiliar na identificação de novos marcadores moleculares e determinação de características específicas do parasita. Sendo assim, o foco desse trabalho foi integrar esses dados com os recursos de bioinformática para contribuir em importantes questões biológicas: a identificação de novos biomarcadores para diferenciação das espécies e que contribuem para determinação de virulência.

Para tanto, foi desenvolvido uma ferramenta web capaz de rastrear genomas completos a fim de identificar conjuntos de primers táxon-específicos para genotipagem por PCR Multiplex. Esta ferramenta, nomeada TipMT, foi capaz de identificar 19.314 pares de primers utilizando os genomas de diferentes espécies de *Leishmania* como entrada para o programa. Na validação experimental verificou-se a eficiência dos pares de primers gerados pela ferramenta, capazes de diferenciar três espécies, *L. major*, *L. braziliensis* e *L. infantum*. TipMT oferece uma combinação de características que não está presente em outras aplicações web com a mesma finalidade: aceita múltiplas sequências como entrada, identifica regiões alvo automaticamente, testa a especificidade dos primers e como saída, gera um arquivo texto com as informações gerais dos primers e um gel de eletroforese virtual.

Em uma segunda parte do trabalho, foi avaliada a diferença do perfil de expressão de promastigotas com alta e baixa infectividade em *L. amazonensis* por RNA-seq. Nessa análise de transcriptômica, avaliou-se a expressão gênica de parasitos recém isolados de camundongos experimentalmente infectados e de parasitos que foram mantidos por 30 passagens de cultura *in vitro*. A avaliação de infectividade das amostras confirmou a diminuição da taxa de infecção após diversas passagens em

cultura axênica. O sequenciamento pela plataforma Illumina HiSeq 2000 gerou um total de 27,35 milhões de leituras e esses transcritos foram mapeados no genoma de referência de *L. amazonensis* com uma percentagem média de 86,57%. Identificou-se 626 genes com expressão diferencial significativa, sendo 66,13% com expressão diminuída após o cultivo *in vitro* seriado. Após 30 passagens *in vitro* foi detectada uma diminuição bastante significativa na expressão de genes, já descritos como envolvidos na infectividade, como a metalo-peptidase GP63, a triparedoxina peroxidase e a proteína de heatshock HSP70. A identificação das vias metabólicas mostrou que a perda de infectividade está relacionada a alterações no metabolismo, dados corroborados por estudo anterior utilizando técnicas de proteômica. Considerando esses resultados, a virulência em *Leishmania* pode estar associada com a eficiência de três processos: a interação parasito-hospedeiro mediada por proteínas de superfície do parasito; resposta ao estresse oxidativo; e metabolismo de aminoácidos e ácidos graxos. Os genes potencialmente associados à virulência de *Leishmania* apontados por esse estudo são, portanto, candidatos para posteriores estudos funcionais.

ABSTRACT

After 10 years of publication of the first complete genome of *Leishmania* genus, advances in sequencing technology have provided a large accumulation of genomic, transcriptomic and proteomic data of these taxa. The convergence of these -omics data combined with bioinformatics tools allows the identification of new molecular markers and determination of specific characteristics of the parasite. Thus, the focus of this work was to integrate these data with bioinformatics resources to contribute to a better understanding of important biological aspects of this parasite: identification of new biomarkers for taxa differentiation and associated with virulence.

For this purpose, we developed a web tool able to receive complete genomes to identify sets of taxon-specific primers for genotyping by multiplex PCR. This tool, named TipMT, was able to identify 19,314 pairs of primer using the genomic data from distinct *Leishmania* species as input. The experimental validation verified the efficiency of a set of primers designed by the tool, which were able to differentiate three species: *L. major, L. braziliensis* and *L. infantum*. TipMT offers a combination of features that are not present in other available web applications developed for this purpose. TipMT supports multiple sequences as input, identifies target regions automatically, tests the specificity of the primers and generates a text file with general information of the primers and virtual electrophoresis gel as output.

In a second part of this work, we evaluated the difference in expression profile of *L. amazonensis* promastigotes with high and low infectivity by RNA-seq. In this transcriptome study, we compared the expression profile of parasites freshly isolated from experimentally infected mice and parasites that were cultured after 30 *in vitro* passages. The evaluation of infectivity of the samples confirmed the decrease of infection rate after several passages in axenic culture. Sequencing by Illumina HiSeq 2000 platform generated 27.35 million of reads and these transcripts were mapped to the reference genome of *L. amazonensis* with an average percentage of 86.57%. We have identified 626 genes with significant differential expression, 66.13% with decreased

expression, after serial passages in *in vitro* culture. After 30 passages *in vitro,* we have identified a significant decrease in the expression of genes already described as involved in infectivity, as GP63 metallo-peptidase, tryparedoxin peroxidase and heatshock protein HSP70. The identification of metabolic pathways showed that the loss of infectivity is related to changes in metabolism, data corroborated by previous studies using proteomic techniques. Considering these results, the *Leishmania* virulence could be associated with the efficiency of three processes: parasite-host interaction mediated by parasite surface proteins; response to oxidative stress; and metabolism of amino acids and fatty acids. This study disclosed several other genes that are likely to be associated with *Leishmania* virulence and are good candidates for further functional studies.

Lista de figuras

Lista de tabelas

Lista de abreviações

BCV - Coeficiente de Variação Biológica

CPM - Contagens de transcritos por milhão

CSS - *Cascading Style Sheets*

DNA - ácido desoxirribonucleico

DNAc - DNA complementar

DNAk - DNA do cinetoplasto

EBI - European Bioinformatics Institute

EST - Expressed sequence tags

FDR - False Discovery Rate

GO - Gene Ontology

HTML - HyperText Markup Language

KEGG - Kyoto Encyclopedia of Genes and Genomes

logCPM - Logaritmo das Contagens de Transcritos por Milhão

logFC - Logaritmo de Fold-Change

MDS - Multidimensional Scaling

MLEE - Eletroforese de Enzimas Multilocus

NCBI - National Center for Biotechnology Information

OMS - Organização Mundial de Saúde

PCR - Reação em Cadeia Polimerase

PHP - Hypertext Preprocessor

RefSeq - The Reference Sequence database built by NCBI

RFLP - Restriction Fragment Length Polymorphism

RNA - Ácido ribonucleico

RNAm - RNA mensageiro

RNAr - RNA ribossômico

RNA-Seq - RNA sequencing

SNP - Single Nucleotide Polymorphism

SSR - Simple Sequence Repeat

TMM - Trimmed Mean of M values

INTRODUÇÃO

Considerações Gerais

As leishmanioses constituem um complexo de enfermidades, que podem ser categorizadas em três tipos: leishmaniose visceral, cutânea e mucocutânea. Essas manifestações clínicas são causadas por cerca de 20 espécies do gênero *Leishmania* encontradas em regiões tropicais e subtropicais do mundo (BAÑULS; HIDE; PRUGNOLLE, 2007). Essa doença afeta 88 países na África, Ásia e América Latina, com uma incidência estimada de 12 milhões de pessoas, com 2 milhões de novos casos por ano, e 350 milhões sob risco de infecção (DESJEUX, 2004).

As manifestações clínicas da leishmaniose são variadas e determinadas, principalmente, por fatores ambientais, do inseto vetor e da genética do parasita e do hospedeiro (SAKTHIANANDESWAREN; FOOTE; HANDMAN, 2009). As leishmanioses cutâneas são caracterizadas por lesões cutâneas papulares, nodulares ou ulcerosas e no Velho Mundo (Bacia Mediterrânea, Oriente Médio, África e Ásia), e as espécies comumente envolvidas nessa infecção são *L. major*, *L. tropica*, e *L. aethiopica*. Já no Novo Mundo (México, América Central e América do Sul) as principais espécies responsáveis são *L. mexicana*, *L. amazonensis*, *L. braziliensis*, *L. panamensis*, *L. peruviana*, e *L. guyanensis* (DESJEUX, 2004; HERWALDT, 1999).

A infecção por *Leishmania* se inicia com a introdução das formas infectivas, promastigotas metacíclicas, na pele do hospedeiro mamífero pela picada de insetos da subfamília Phlebotominae, conhecidos comumente como mosquito palha. Uma vez na pele, as promastigotas infectivas são fagocitadas por macrófagos e se diferenciam em amastigotas, forma replicativa no mamífero. Após uma multiplicação intensa, os macrófagos repletos de amastigotas rompem-se e liberam estas formas na corrente sanguínea, onde poderão ser fagocitadas por novas células fagocitárias. Essas células podem ser ingeridas por um flebotomíneo durante o repasto sanguíneo.

No intestino médio do inseto vetor, as amastigotas se diferenciam em promastigota e em seguida o parasito desenvolve uma série complexa de modificações

morfológicas e funcionais, originando as formas promastigotas metacíclicas (DE ALMEIDA et al., 2003) (Figura 1).

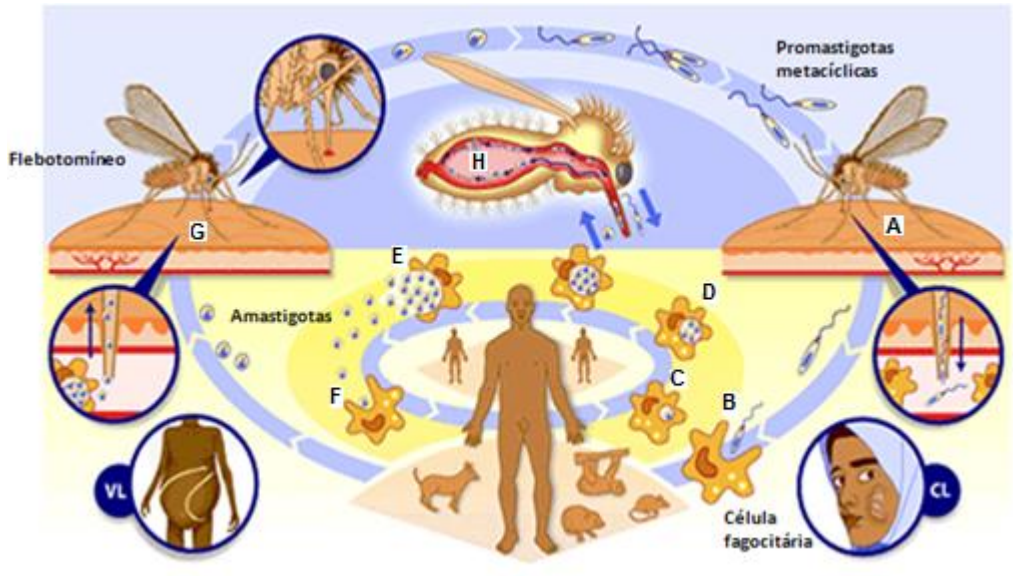


Figura 1: Ciclo de vida de Leishmania. Retirado de WHO (http://www.who.int/tdrold/diseases/leish/leish.htm). a: introdução das promastigotas metacíclicas na pele do hospedeiro mamífero pela picada do flebotomíneo; b: fagocitose das promastigotas pelas células fagocitárias; c: diferenciação das formas promastigotas em amastigotas; d: multiplicação intracelular das formas amastigotas; e: lise da célula e liberação das formas amastigotas; f: fagocitose de amastigotas por novas células fagocitárias do hospedeiro; g: ingestão de formas amastigotas pelo vetor; h: diferenciação das amastigotas em promastigotas, seguida das modificações morfológicas e funcionais para originar as formas promastigotas metacíclicas. VL, Leishmaniose visceral; CL, Leishmaniose cutânea.

A hipótese atual para a estruturação populacional é um reprodução clonal com episódios ocasionais de pseudorecombinação e recombinação intragênica, que simulam o processo de reprodução sexual (ROUGERON et al., 2009). O notável número de espécies de leishmanias é dividido em dois subgêneros: *(L.) Leishmania* e *(L.) Viannia*, que são separados em função da sua localização no intestino do vetor. No subgênero *Leishmania* os parasitos se desenvolvem no intestino médio e anterior do flebotomíneo e inclui o complexo *L. donovani, L. tropica, L. major, L. aethiopica* e *L. mexicana*. Já no subgênero *Viannia,* as leishmanias se desenvolvem no intestino posterior do flebotomíneo e inclui as espécies do complexo *L. braziliensis, L. peruviana, L. guyanensis, L. naifi* e *L. lainsoni*.

O primeiro genoma de *Leishmania* sequenciado foi o de *L. major*. Essa espécie apresenta 36 cromossomos, que totalizaram 32,8 Mb para o genoma haplóide, e foram preditos 911 genes de RNA, 39 pseudogenes, e 8272 genes codificadores de proteínas, dos quais 36% tiveram suas funções caracterizadas experimentalmente ou preditas por similaridade (IVENS et al., 2005). Em seguida, foram publicadas as sequências de *L. braziliensis* e *L. infantum*. As 2 espécies têm genomas de tamanhos parecidos, 32 Mb e 32,13 Mb em *L. braziliensis* e *L. infantum*, respectivamente. Além disso, o número de genes codificantes também apresenta semelhança, 8.314 e 8.195, entretanto, o número de pseudogenes apresenta maiores diferenças: 161 e 41, em *L. braziliensis* e *L. infantum*, respectivamente. A conservação entre as sequências codificadoras é alta: a média de identidade nucleotídica entre *L. major* e *L. infantum* é de 94%, *L. major* e *L. braziliensis* de 82%, e *L. infantum* e *L. braziliensis* de 81%. Comparações genômicas revelaram um alto nível de conservação e organização gênica no gênero, sendo que cerca de 99% dos genes dos três organismos mantêm a sintenia (EL-SAYED et al., 2005). Apesar dessa alta conservação, há cada vez mais evidências apontando que aneuploidias podem ocorrer no gênero, como a ploidia instável entre cepas, e células, de *L. infantum*. Isso sugere que o genoma em *Leishmania* é caracterizado por um "mosaico de aneuploidias" (CANTACESSI et al., 2015; ROGERS et al., 2011; STERKERS et al., 2012).

Apesar da variabilidade na patogenicidade e tropismo de tecidos em *Leishmania*, foram encontradas apenas cerca de 100 genes específicos para cada espécie (PEACOCK et al., 2007; ROGERS et al., 2011). A presença de poucos genes específicos sugere que as diferentes manifestações clínicas observadas tenham como explicação a expressão gênica diferencial, a diferença no número de cópias dos genes e a resposta imune diferencial do hospedeiro (KAYE; SCOTT, 2011; PEACOCK et al., 2007; ROGERS et al., 2011; TEIXEIRA et al., 2012).

Em relação à expressão gênica, a família Tripanosomatidae, a qual o gênero *Leishmania* pertence, apresenta características peculiares. Os genes estão em geral

organizados em clusters gênicos direcionais, os quais são co-transcritos, assemelhando-se às unidades de transcrição policistrônicas de procariotos. A transcrição policistrônica implica em um controle da expressão gênica primariamente pós-transcricional, já que genes presentes em uma mesma unidade de transcrição podem apresentar níveis de expressão distintos (MARTÍNEZ-CALVILLO et al., 2010; STEVENS et al., 2001). Um mecanismo importante nessa regulação pós-transcricional é o processamento do RNA policistrônico: a molécula é individualizada em RNAm monocistrônicos, em seguida ocorre o trans-splicing, com a adição do mini-éxon de 39 nucleotídeos, conhecido como “Spliced Leader”, na parte 5’ do mRNA e, finalmente, ocorre a poliadenilação nas extremidades 3’ dos mRNAs (ARAÚJO; TEIXEIRA, 2011).

Após 10 anos da publicação do primeiro genoma completo de uma espécie de Leishmania, os avanços na tecnologia de sequenciamento proporcionaram um grande acúmulo de informação genômica e transcriptômica. Uma profunda exploração desses dados contribuiria fortemente no entendimento da biologia do parasito, como na virulência, patogenicidade, imunobiologia e resistência às drogas. Esses estudos teriam aplicações potenciais no desenvolvimento de novas formas de diagnósticos, tratamentos e controle da leishmaniose (CANTACESSI et al., 2011, 2015; LEPROHON et al., 2015).

Diagnóstico molecular

A identificação precisa das espécies de parasitas é uma importante informação epidemiológica, que contribui para melhorar as estratégias de tratamento e controle de doenças. Para doenças parasitológicas, o diagnóstico é muitas vezes baseado nas manifestações clinicas, achados microscópicos, sorologia e/ou biopsia de tecidos infectados. Entretanto essas abordagens possuem diversas limitações, como sensitividade insuficiente, especialização do trabalhador e necessidade da cultivar o parasito (WONG et al., 2014). Mais recentemente, a técnica de PCR (Reação em Cadeia

Polimerase) tem sido utilizada como uma estratégia alternativa e/ou complementar aos testes parasitológicos clássicos usados em diagnóstico.

O nível de discriminação da tipagem baseada em PCR depende do marcador molecular utilizado. Dessa forma, identificar regiões alvo adequadas para o anelamento do primer é um passo crucial, pois essas regiões devem ser conservadas dentro do taxa alvo e diferir em taxa relacionados (LUCCHI et al., 2013; WONG et al., 2014). Os recentes avanços na tecnologia de sequenciamento permitem a realização de projetos genoma a um custo significativamente mais baixo, mesmo para organismos não-modelo. A exploração dos dados genômicos pode melhorar os métodos de detecção de infecções assintomáticas com baixas cargas parasitárias, uma vez que facilita a busca por genes multi-cópias, e a diferenciação de espécies ao identificar sequencias específicas (WONG et al., 2014).

Em *Leishmania*, a identificação de espécies tem sido feita pela Eletroforese de Enzimas Multilocus (MLEE), o padrão-ouro, que detecta diferentes alelos de genes *housekeeping* por meio da mobilidade eletroforética distinta das enzimas que eles codificam. Entretanto, esse método apresenta limitações, como a exigência da cultura do parasita, baixo poder de resolução e ser dispendioso. Diferentes métodos moleculares têm sido desenvolvidos e utilizados para contornar esse problema. Em laboratórios de diagnóstico molecular, o DNA do cinetoplasto (kDNA) é comumente utilizado, sendo que há duas abordagens. A primeira consiste na amplificação dos minicírculos de kDNA por primers específicos seguido da clivagem dos produtos de PCR por enzimas de restrição (RFLP) e a segunda, na amplificação dos minicírculos de kDNA por primers que geram um padrão polimórfico de bandas para todas as espécies, seguido de uma hibridização dos produtos de PCR por sondas especificas de kDNA (BAÑULS; HIDE; PRUGNOLLE, 2007). Estudos usando esse alvo apresentam altas taxas de detecção, variando entre 100% para leishmaniose cutânea e 97,1% para mucocutânea, e é capaz de diferenciar os subgêneros. Uma importante abordagem no diagnóstico da leishmaniose é a caracterização das espécies, que não ocorre para a

maioria das técnicas utilizadas. Para identificação de espécies, as técnicas podem ser basicamente divididas em dois grupos: a diferenciação baseada no tamanho do produto de amplificação ou na presença de sítios de enzimas de restrição. Na primeira, os alvos mais comumente usados são sequências repetitivas, como os microssatélites, e genes, como os genes da glicose-6-fosfato desidrogenase e manose fosfato isomerase (GOTO; LINDOSO, 2010). Na segunda, os alvos são os genes: GP63 (glycoprotein 63), ITS (gene do espaçador transcrito interno de DNA ribossomal), HSP70 (heatshock protein 70) e cisteino-peptidases (REITHINGER; DUJARDIN, 2007; SRIVIDYA et al., 2012).

O diagnóstico molecular pode, também, permitir a identificação de características específicas do parasita, como a virulência ou resistência às drogas (LUCCHI et al., 2013). Sendo assim, a quantidade de dados genômicos disponíveis combinada com ferramentas de bioinformática abrem a possibilidade de rastrear marcadores altamente informativos, como microssatélites e genes ortólogos, a fim de discriminar as diferentes espécies do gênero e características específicas (LI et al., 2009b; REITHINGER; DUJARDIN, 2007).

Microssatélites como marcadores moleculares para o diagnóstico

Os microssatélites, também conhecidos como *Simple Sequence Repeat* (SSR), são uma subdivisão da família de DNA satélites. Essas sequências nucleotídicas se repetem em série, sendo tradicionalmente definidas como arranjos formados pela combinação de 1 a 6 bases. Os microssatélites podem ser encontrados abundantemente em genomas de procariotos e eucariotos, em regiões codificadoras ou não-codificadoras e são classificados como perfeitos, imperfeitos, ou compostos. Sequências perfeitas são aquelas que não apresentam interrupções (e.g. $(GT)_{10}$), as imperfeitas são interrompidas por nucleotídeos não repetitivos (e.g. $(GT)_4AG(GT)_6$) e os compostos são formados por dois motivos adjacentes (e.g. $(GT)_{10}(CA)_7$). A herança de microssatélites segue o modelo mendeliano co-dominante e vários alelos podem ser encontrados para um único *locus* (OLIVEIRA; ZUCCHI; VENCOVSKY, 2006).

Embora amplamente utilizados em diversas áreas da genética, os mecanismos genéticos e evolutivos envolvidos no surgimento dos microssatélites ainda não são bem compreendidos. Vários mecanismos mutacionais foram propostos para explicar a variabilidade dessas repetições, como: o deslizamento da DNA polimerase durante a replicação ou reparo do DNA e o crossing-over desigual (RICHARD; KERREST; DUJON, 2008). Alguns estudos recentes têm sugerido que o tamanho dos microssatélites é o resultado do equilíbrio entre os eventos de deslizamento e mutações pontuais. Sendo que o primeiro favorece o crescimento e o segundo quebra as grandes sequências em outras menores (BHARGAVA; FUENTES, 2010). Durante a replicação ou o reparo do DNA, o deslizamento da DNA polimerase pode ocorrer em uma fita de DNA que temporariamente se dissocia da outra e religa rapidamente em uma posição diferente, um erro de pareamento. Esse processo se não corrigido pelo sistema de reparo de DNA leva ao aumento do número de repetições naquele alelo, se o erro ocorre na fita nova, ou uma diminuição, se o erro ocorre na fita molde. No crossing-over desigual, as unidades de repetição levam a um erro de pareamento dos cromossomos homólogos durante o quiasma, o que resultará em uma troca de fragmentos de tamanhos diferentes. Esse evento pode gerar mudanças drásticas no número de repetições de cada alelo envolvido na troca (OLIVEIRA; ZUCCHI; VENCOVSKY, 2006).

Estudos sobre a origem dos microssatélites no genoma indicam que a gênese não é aleatória e uma sequência mínima de 3 a 4 repetições é requerida para que os eventos de extensão possam ocorrer. Esses proto-microssatélites podem surgir a partir de mutações pontuais, como substituição de nucleotídeos e eventos de indel, ou trazidos de outros locais no genoma pelos elementos genéticos móveis (BHARGAVA; FUENTES, 2010).

O método convencional para descoberta de *loci* de microssatélites é laborioso e caro, e por isso tem sido substituído por análises *in silico* nos bancos de dados genômicos. De forma geral as ferramentas computacionais de mineração podem ser classificadas em três categorias com base na sua arquitetura. Um dos métodos detecta

as repetições em tandem seguindo regras específicas de construção e assegura uma busca exaustiva de todas as repetições, esse tipo de ferramenta encontra apenas microssatélites perfeitos. Outra técnica é a pesquisa em duas fases; na primeira, certas sequências são listadas como microssatélites com base nos parâmetros de busca, em seguida, as regiões são validadas de acordo com as regras estatísticas estipuladas. Nessa abordagem, a busca por repetições pode não ser exaustiva pois algumas sequências são selecionadas, mas não passam nos testes estatísticos. O terceiro tipo de algoritmo é o mais direto, um dado motivo, ou biblioteca de motivos, será alinhado ao longo das sequências genômicas e as regiões que apresentarem uma pontuação superior ao corte serão consideradas microssatélites (SHARMA; GROVER; KAHL, 2007).

A presença de alelos múltiplos, modo de herança co-dominante, abundância aliado à fácil detecção experimental e alta reprodutibilidade têm tornado os microssatélites os principais marcadores genéticos em diversos tipos de aplicações, como mapeamento genético, diagnóstico de doenças, investigação forense, análise populacional, estudos ecológicos, paternidade, biologia da conservação e discriminação de taxa muito próximos (DURAN et al., 2009; SHARMA; GROVER; KAHL, 2007).

Genes ortólogos como marcadores moleculares de diagnóstico

Sequências homólogas dividem ancestralidade comum e podem ser caracterizadas como, ortólogas, homologia por especiação, ou parálogas, por duplicação gênica. Esses conceitos são essenciais nas áreas de evolução, genômica comparativa, metagenômica e filogenômica. Ortólogos tipicamente mantêm a similaridade dos domínios protéicos e tem o mesmo nicho funcional, enquanto parálogos tendem a divergir com novas funções a partir de mutações pontuais e recombinação de domínios(POWELL et al., 2012).

A identificação de sequências homólogas tornou-se muito importante devido ao rápido crescimento do número de genomas sequenciados (LI; STOECKERT; ROOS,

2003). Anotação gênica, análises filogenéticas, reconstrução de redes metabólicas e identificação de elementos regulatórios são exemplos de aplicações dependentes dessa informação. A detecção automática de sequências ortólogas em larga escala é, portanto, um problema de extrema importância e de difícil implementação. Diversos métodos de predição foram desenvolvidos e podem ser divididos em duas categorias: baseados em árvores filogenéticas; e baseados em grafos e na similaridade entre pares de sequências(WHITESIDE et al., 2013).

O desenho de primers usando genes ortólogos como alvo permite utilizar essas sequências como marcadores moleculares, tornando possível sua utilização em estudos de diversidade, filogenética e genômica comparativa (FULTON et al., 2002; SAHU; GUPTA; DIXIT, 2011). Esses marcadores podem ser utilizados para amplificar os ortólogos correspondentes àquele táxon alvo e também uma ampla gama de taxa, incluindo os que não possuem um genoma sequenciado (WU et al., 2006). Diversos trabalhos já demonstraram a utilidade desses marcadores no estudo de taxa relacionados (FULTON et al., 2002; LI et al., 2008; SOLÍS-CALERO, 2008; WU et al., 2006). Entretanto, a busca por esses marcadores e o desenho dos iniciadores ocorre de forma não automatizada, necessitando que o pesquisador busque os alvos dentre uma série de parálogos ou ortólogos e desenhe os iniciadores específicos manualmente (GUERRERO et al., 2010).

Fatores de Virulência em *Leishmania*

A virulência no gênero *Leishmania* é manifestada por sua habilidade de produzir diferentes sintomas clínicos, variando de lesões cutâneas que se curam espontaneamente a uma doença visceral potencialmente fatal. Embora a virulência possa ser modulada por fatores ambientais e genéticos relacionados aos hospedeiros mamíferos e vetores, as características genéticas do parasito são elementos chave no estabelecimento da infecção. Portanto, genes e seus produtos que participem da capacidade do parasito de causar doença, mas não são necessários para a sua

sobrevivência por si só, são considerados fatores de virulência (BIFELD; CLOS, 2015; PEACOCK et al., 2007; RIVAS et al., 2004).

Um importante fator de virulência no gênero é a glicoproteína de superfície mais abundante nas formas promastigotas do parasito, a GP63. A proteína é uma metaloprotease que participa de várias etapas do início da infecção pelas promastigotas. As cisteino-proteinases são enzimas que exercem funções essenciais na interação parasito-hospedeiro. Outro fator é a proteína HSP23, relacionada à sobrevivência a temperaturas encontradas no hospedeiro humano. (BIFELD; CLOS, 2015; OLIVIER et al., 2012)

A manutenção *in vitro* de parasitas por um longo tempo é uma das primeiras abordagens para identificar genes de virulência de parasitas e desenvolvimento de cepas atenuadas (MITCHELL; HANDMAN; SPITHILL, 1984; RAFALUK et al., 2015). Estudos anteriores foram capazes de identificar algumas proteínas possivelmente envolvidas na perda de infectividade de espécies de *Leishmania* submetidas a várias passagens em cultura *in vitro*. Entretanto, muitos estudos foram feitos usando apenas uma abordagem proteômica, como por exemplo: a avaliação por proteômica de fatores de virulência em duas linhagens de *L. infantum* (DA FONSECA PIRES et al., 2014) e na mesma linhagem de *L. amazonensis* após várias passagens em cultura (MAGALHÃES et al., 2014). Novos estudos com uma abordagem mais ampla podem ser capazes de identificar um espectro maior de fatores de virulência e o real papel das proteínas nesse processo biológico (KAYE; SCOTT, 2011; MAGALHÃES et al., 2014).

Devido a expressão gênica não usual em *Leishmania*, estudos recentes têm combinado os dados da proteômica à transcriptômica para obter melhores resultados, como na identificação do papel do ferro na diferenciação em amastigotas (MITTRA et al., 2013) e na resposta do parasito a privação de purina (MARTIN et al., 2014). A junção desses dados proteômicos à analises de transcriptômica seriam de grande valia no entendimento de mecanismos de virulência e expressão gênica (LEPROHON et al., 2015).

O transcriptoma pode ser definido como o conjunto completo de transcritos em uma célula, e suas quantidades, em um estágio específico do desenvolvimento ou condição fisiológica, incluindo, RNA codificante (mRNA) e não codificante (rRNA, tRNA, iRNA e outros tipos de RNAs) (WANG; GERSTEIN; SNYDER, 2009a). Técnicas de microarranjos de DNA eram as mais utilizadas para a determinação de um amplo padrão de expressão gênica. Entretanto, a metodologia apresenta algumas limitações, como a especificidade do arranjo para cada tratamento, a saturação do fundo e a qualidade e densidade variáveis dos spots. Esses fatores têm dificultado a análise comparativa entre experimentos, levando a necessidade de desenvolver métodos normalizadores complexos (HINTON et al., 2004).

Devido às limitações das técnicas de sequenciamento convencional de ESTs e microarranjos, além do advento das novas tecnologias de sequenciamento de DNA em larga escala, foi proposto um novo método, denominado RNA-Seq (*Whole Transcriptome Shotgun Sequencing*), para o mapeamento e quantificação de transcriptomas (WANG; GERSTEIN; SNYDER, 2009a). Dentre as principais vantagens dessa técnica, podem ser citadas: detecção de transcritos não fica restrita somente àquelas sequências genômicas já conhecidas; determinação precisa dos limites de transcrição, com resolução de até uma única base; menor background, uma vez que as sequências podem ser mapeadas, sem ambiguidade, a regiões distintas do genoma; identificação de isoformas; quantificação de expressão alelo-específica (GARBER et al., 2011; WANG; GERSTEIN; SNYDER, 2009a).

Vários estudos têm demonstrado a eficiência desse método para analisar o transcriptoma de tripanosomatídeos. KOLEV et al. (2010) produziu um mapa genômico com resolução de um nucleotídeo do transcriptoma de *Trypanosoma brucei*; NILSSON et al. (2010) descobriu mais de 2500 eventos de *splicing* alternativos em *T. brucei*, resultado da adição da sequência do "*spliced leader*" em diferentes posições do mRNA,; SIEGEL et al. (2010) analisou a abundância de mRNA em dois estágios do ciclo de vida do *T. brucei* e mapeou as regiões não traduzidas dos transcritos; e RASTROJO et al.

(2013) determinou 1.884 novos genes e a expressão relativa de cada um dos 10.285 transcritos detectados em promastigotas de *L. major*.

Ferramentas computacionais na Bioinformática para a identificação de marcadores moleculares

Nos últimos anos, a computação tem se associado profundamente à biologia, tornando-se uma área de pesquisa interdisciplinar, conhecida como Bioinformática, que usa a tecnologia da informação para responder complexas questões biológicas. A solução desses problemas envolve o desenvolvimento de ferramentas computacionais que organizam, armazenam e recuperam informações de banco de dados (ABD-ELSALAM, 2003; ATTWOOD et al., 2011; FULLER et al., 2013).

Um protocolo computacional automatizado (também conhecido como *pipeline*) pode ser definido como uma execução sequencial de programas em uma base de dados. Linguagens de script, como o Perl e Phyton, fornecem uma forma fácil de automatizar a execução de programas e são especialmente úteis na definição dos parâmetros de entrada (CAVALCANTI et al., 2005). Uma notável característica dos pipelines é a flexibilidade de acrescentar e modificar componentes no processo de análise, visto que softwares são continuamente desenvolvidos e atualizados. Unir scripts em um pipeline tornou-se uma forma consagrada de criar sistemas de análise em bioinformática, sendo que dois grandes paradigmas se estabeleceram nas aplicações biológicas: a execução de várias tarefas independentes em paralelo ou múltiplas tarefas em série. Nesse último modelo, há uma interação entre as aplicações envolvidas, o resultado da primeira tarefa é a entrada da próxima. Portanto, o principal desafio dessa metodologia é a compatibilidade de formatos dos arquivos (CURCIN; GHANEM; GUO, 2005).

Essa metodologia apresenta algumas desvantagens, como a manutenção a longo-prazo do *pipeline*, devido ao acúmulo de múltiplas versões, a difícil interpretação dos *scripts* por terceiros e a forte dependência das configurações da máquina e do sistema operacional (BHOWMICK; SINGH; LAUD, 2003; CURCIN; GHANEM; GUO,

2005). A tecnologia de serviços web foi especialmente concebida para proporcionar interoperabilidade entre diferentes plataformas. A ideia de construir um serviço web ao invés de encapsular scripts na forma de um programa torna-se, portanto, uma opção mais vantajosa (CAVALCANTI et al., 2005).

Na bioinformática, esses serviços têm sido apontados como uma tecnologia capaz de contribuir significativamente na exploração plena dos dados biológicos e muitos projetos da área têm concentrado esforços nessa direção, viabilizando a execução, visualização e gestão dos dados através da internet (CAVALCANTI et al., 2005). Exemplos dessa tendência são as ferramentas e base de dados oferecidas pelas grandes instituições de pesquisa: European Bioinformatics Institute (EBI), National Center for Biotechnology Information (NCBI) (CURCIN; GHANEM; GUO, 2005).

Uma das tendências mais conhecidas na bioinformática é o crescimento exponencial da quantidade de dados genômicos disponíveis, devido principalmente aos crescentes avanços na tecnologia de sequenciamento (PAGANI et al., 2012). . Com o objetivo de aproveitar todo potencial destes dados, sequências genômicas devem ser convertidas em conhecimento biológico. Para isso diversos tipos de estudos pós-genômicos são realizados, identificando características nessas sequências nucleotídicas. Dessa forma, ferramentas de bioinformática têm sido desenvolvidas para realizar uma infinidade de funções, como por exemplo: predição de genes, domínios e homologia, e rastreamento de sequências repetitivas e polimorfismos (DURAN et al., 2009).

Vários programas de bioinformática já foram descritos na literatura para seleção de pares de primers a partir de uma sequência alvo, como um marcador molecular. A maioria destes programas está disponível gratuitamente, entretanto, a qualidade e manutenção das ferramentas são variáveis. Isso muitas vezes resulta na perda de links e incompatibilidades, portanto, o que anteriormente se mostrava útil pode não ser mais funcional (ABD-ELSALAM, 2003).

JUSTIFICATIVA

Os métodos de genotipagem baseados em PCR são uma técnica de diagnóstico molecular amplamente usada em estudos biológicos e biomédicos. Os marcadores baseados em PCR são passíveis de automação e, se bem desenhados, altamente específicos (FREDSLUND et al., 2006). PCR Multiplex é uma reação de amplificação utilizando mais de um conjunto de primers em uma única reação. Essa técnica tem sido usada amplamente em diversas áreas, incluindo: identificação de patógenos, genotipagem por SNP, análises de mutação, análises de deleção gênica, quantificação, análises de ligação, detecção de RNA e estudos forenses. A construção de primers para PCR multiplex de forma autônoma continua sendo um desafio computacional (SHEN et al., 2010).

O RNA-Seq já permitiu a geração de uma visão global sem precedentes a respeito do transcriptoma e sua organização para um número significativo de espécies e tipos celulares. Com sua alta resolução e sensibilidade, permitiu a identificação de novos genes e isoformas de genes conhecidos. O potencial, então, para determinar a estrutura e dinâmica de diversos transcriptomas é bastante amplo (WANG; GERSTEIN; SNYDER, 2009b).

Há um grande volume de dados na área de genômica, transcriptômica e proteômica no gênero *Leishmania*. A convergência destas informações ômicas aliadas a ferramentas de bioinformática permitiria auxiliar na identificação de novos marcadores moleculares tanto para identificação de espécies quanto características específicas de isolados do parasita, como a virulência ou resistência às drogas (CANTACESSI et al., 2011; REITHINGER; DUJARDIN, 2007). A integração destes dados podem contribuir em importantes questões na leishmaniose, como o diagnóstico molecular e a melhor compreensão do mecanismo de infecção e sobrevivência destes parasitos.

OBJETIVOS

Esse projeto tem como objetivo geral a busca de marcadores moleculares e de virulência em *Leishmania* usando abordagens de bioinformática. Para tanto, se propôs (i) o desenvolvimento de uma ferramenta web capaz de rastrear genomas completos a fim de identificar conjuntos de primers táxon-específicos para genotipagem por PCR Multiplex e validação desta ferramenta em genomas de diferentes espécies de *Leishmania* (ii) investigar a diferença do perfil de expressão de promastigotas com alta e baixa infectividade em *L. amazonensis* por RNA-seq.

Objetivos específicos

O capítulo 1 compreende a construção de uma ferramenta web para amplo uso da comunidade científica e geração de um conjunto de primers específicos para três espécies de *Leishmania*. Seus objetivos específicos foram:

- Desenvolver uma pipeline capaz de executar os diferentes desafios na geração de conjuntos de primers específicos: buscar alvos, desenhar primers e testar a especificidade dos mesmos;
- Disponibilizar a *pipeline* desenvolvida por meio da construção de uma ferramenta web;
- Verificar a especificidade dos primers encontrados pela pipeline *in vitro*.

No capítulo 2 é apresentado o sequenciamento em larga escala de bibliotecas de cDNA construídas a partir da extração do RNA total de promastigotas de *L. amazonensis* com alta e baixa infectividade. Seus objetivos específicos foram:

- Determinar a infectividade das formas promastigotas após sucessivas passagens em cultura axênica e obter as amostras de RNA.
- Avaliar a expressão gênica diferencial entre as passagens.

CAPÍTULO I

APRESENTAÇÃO

Uma técnica muito utilizada em análises de genômica comparativa, sistemática molecular e evolução são os marcadores moleculares (DURAN et al., 2009; SHARMA; GROVER; KAHL, 2007). Na ausência de sequências genômicas para o grupo de espécimes a ser analisado, os marcadores baseados em reação de PCR têm se mostrado bastante eficientes. Com a grande disponibilidade de dados genômicos surgiram métodos computacionais para ajudar na identificação dos marcadores moleculares e no desenho de pares de primers. Várias ferramentas foram desenvolvidas para ambas tarefas, entretanto, ainda existem muitas limitações que dificultam o uso por biólogos com pouca experiência computacional, como a construção dos primers, que é um passo crítico da PCR e pode ser bastante desafiadora dependendo da região alvo (ABD-ELSALAM, 2003; SOBHY; COLSON, 2012). Além disso, a concepção de um grande número de pares de primers de forma automatizada é extremamente desafiadora, tanto em termos de obtenção de boas regiões alvo, quanto no tempo necessário para gerar os resultados. Muitos métodos disponíveis atualmente são demorados e têm dificuldade em encontrar primers para múltiplas sequências alvo (CHUANG; CHENG; YANG, 2013).

Dessa forma, faz-se necessário o desenvolvimento de novas ferramentas capazes de preencher essa lacuna na prospecção de primers táxon-específicos. Como o gênero *Leishmania* possui espécies com genomas sequenciados de qualidade, os quais apresentam alto grau de similaridade, optou-se por realizar a validação da ferramenta desenvolvida neste projeto com estes organismos.

METODOLOGIA

Definição dos bancos de dados

As sequências dos genomas e genes preditos de *Leishmania infantum (JPCM5, versão 3)*, *Leishmania major (Friendlin, versão 5)* e *Leishmania braziliensis*

(MHOM/BR/75M2904, versão 2) foram obtidas do banco de dados TriTrypDB, http://tritrypdb.org/tritrypdb/ (ASLETT et al., 2010).

Identificação dos Microssatélites

Os microssatélites foram identificados usando o programa ProGeRF (LOPES et al., 2015). Repetições perfeitas compostas por um a seis nucleotídeos foram rastreadas nos três genomas com o seguinte critério de seleção dos microssatélites: pelo menos 12 unidades repetitivas para mono-nucleotídeos, 9 para di-, 6 para tri-, e 3 para tetra-, penta- e hexa-nucleotídeos. O programa foi desenvolvido utilizando a abordagem de dicionário para extrair regiões repetitivas com uma complexidade de tempo linear. O algoritmo armazena as informações de localização das “palavras” (repetições) em uma tabela *hash* baseada em listas circulares duplamente ligadas para uma identificação rápida e exaustiva de elementos repetitivos, perfeitos e imperfeitos. Essa técnica permite a busca de repetições de uma forma precisa e rápida mesmo em arquivos FASTA grandes (genomas completos).

Identificação dos Genes ortólogos

As sequências ortólogas foram identificadas utilizando um algoritmo amplamente utilizado pela comunidade científica para esse propósito, o OrthoMCL(LI; STOECKERT; ROOS, 2003), usando os parâmetros *default* do programa. Esse software fornece um método escalar para reconstruir grupos ortólogos entre múltiplos taxa eucarióticos, usando *best hit* bidirecional e o algoritmo de agrupamento de Markov (MCL) para agrupar prováveis pares de ortólogos e parálogos. Inicialmente prováveis ortólogos entre duas espécies são identificados a partir de pares de melhores hits recíprocos (BBH). Em seguida, parálogos recentes são identificados para cada espécie, adotando a premissa de serem mais similares ao gene homólogo identificado para a espécie em questão, do que às sequências de outros organismos.

Implementação da ferramenta de desenho de primers

Todos os scripts foram criados e executados em ambiente Linux, utilizando a linguagem Perl e programas de domínio público como BEDtools (QUINLAN; HALL, 2010), BioPerl (STAJICH et al., 2002), BLAST (ALTSCHUL et al., 1990), EMBOSS (RICE; LONGDEN; BLEASBY, 2000), e-PCR (ROTMISTROVSKY; JANG; SCHULER, 2004), MFEprimer (QU et al., 2012), MultiPLX (KAPLINSKI et al., 2005) e Primer3 (UNTERGASSER et al., 2012). Todas as execuções foram feitas em uma máquina com a seguinte configuração: Intel(R) Xeon(R) CPU E5640 2.67GHz, 16 processadores e 24 Gb de memória DDR3.

Construção da ferramenta web

A aplicação web, chamada "*Tool for Identification of Primers for Multiple Taxa*" (TipMT), foi criada utilizando a linguagem Perl e de desenvolvimento web. A ferramenta está hospedada em uma máquina com a seguinte configuração: Intel(R) Xeon(R) CPU X3430 2.40GHz, 4 processadores e 8 Gb de memória DDR3, e está disponível em http://200.131.37.155/tipMT/. As tecnologias do lado do cliente utilizadas foram o HTML (*HyperText Markup Language*) com as partes dinâmicas escritas em JavaScript para receber os parâmetros de entrada e processar os arquivos entre o usuário e a aplicação. Por sua vez, o PHP e MySQL foram as linguagens utilizadas do lado do servidor para armazenar os parâmetros de entrada e resultados.

Extração de DNA

Formas promastigotas das três espécies de *Leishmania*, *L. braziliensis* MHOM/BR/75/M2904, *L. infantum* MHOM/BR/74/PP75 e *L. major* Friendlin foram cultivadas em meio Schneider (Sigma) suplementado com 10% de soro fetal bovino inativado (Life Technologies) e 1% (v/v) penicilina/estreptomicina e mantido à 24 ºC. Amostras contendo $1x10^8$ parasitos em fase de crescimento logarítmico foram submetidas à centrifugação a 3.000 rpm por 15 minutos. Após a centrifugação, o sobrenadante foi descartado e o pellet lavado com 10 mL de PBS e centrifugado

novamente a 3.000 rpm por 15 minutos. O sobrenadante foi descartado mais uma vez e nova lavagem com PBS foi realizada.

A extração de DNA foi realizada usando o kit Wizard Genomic DNA Purification (Promega), seguindo-se as instruções do fabricante. O DNA extraído de cada espécie foi ressuspendido em água livre de DNase e quantificado no espectrofotômetro NanoDrop® Spectrophotometer ND-1000 e armazenado a -20°C até o momento do seu uso.

Reação em cadeia da polimerase

O DNA genômico obtido das três espécies de *Leishmania* foi utilizado como DNA molde nas reações de amplificação dos primers táxon específicos desenhados pelo TipMT (Tabela 2). Cada reação foi feita utilizando: 100ng de DNA genômico de cada espécie como molde, tampão IB da Phoneutria®, 200μM de dNTPs, 10ng de cada um dos primers forward e reverse selecionados e 1,25 μL de Taq DNA polimerase (Phoneutria®), em um volume final de 30 μL de reação. O programa utilizado para a amplificação consistiu em aquecimento à 94°C por cinco minutos, para desnaturação inicial do DNA, seguida por 30 ciclos de: 94°C por 30 segundos, 15 segundos à 60°C e 10 segundos à 72°C, seguido de uma extensão final de 72ºC por 7 minutos.

Análise dos produtos de amplificação

Os produtos de PCR obtidos foram fracionados por eletroforese em gel de agarose 2,5% em tampão TAE 1X (4,8 g/L Tris-base, 1,14 mL ácido acético glacial, 2 mL EDTA 0,5M, pH 8,0), contendo brometo de etídio (0,5 μg/μL).

RESULTADOS

Identificação dos microssatélites

A identificação dos microssatélites nos genomas é feita utilizando-se o programa ProGeRF (LOPES et al., 2015). A entrada para o programa é um arquivo no formato FASTA e as especificações de busca são definidas pelo usuário, utilizando cinco parâmetros. Os dois primeiros parâmetros definem a variação de n: tamanho inicial da subsequência

e tamanho final da subsequência. O terceiro e o quarto parâmetros permitem encontrar repetições imperfeitas, ao definirem a percentagem de degeneração e número máximo de gaps permitidos por repetição. Devido à janela deslizante percorrer toda a sequência, é comum encontrar resultados sobrepostos. O quinto parâmetro define, então, o grau máximo de sobreposição entre duas sequências.

Para identificar os microssatélites utilizou-se tamanho inicial da unidade repetitiva igual a 1 e final 6, 0% de degeneração, 0 gaps, 50% de porcentagem de sobreposição, o que significa que uma sobreposição maior que 50% entre duas sequências repetitivas implica que apenas o maior motivo é reportado.

A saída é um arquivo texto com as informações sobre as repetições encontradas. O arquivo foi o utilizado nas análises seguintes e contém: o identificador da sequência, o tamanho da sequência, a percentagem de degeneração, o número de gaps, o tamanho da subsequência, o número de vezes que a subsequência é repetida, a posição inicial, a posição final, o motivo repetitivo e a configuração da repetição.

Finalmente, no pipeline do TipMT é feito um filtro para microssatélites formados por mais de 60 nucleotídeos, tamanho interessante para utilização como marcador molecular.

Identificação dos genes ortólogos

Todos os passos para identificação de genes ortólogos pelo OrthoMCL foram automatizados em um *script*, que recebe as sequências dos genes preditos dos três genomas e tem como saída um arquivo contendo todos os grupos de ortólogos gerados pelo software. As sequências nucleotídicas são traduzidas pelo comando “transeq” do pacote EMBOSS, uma vez que o arquivo de entrada do OrthoMCL são sequências proteicas. Os genes ortólogos com apenas um gene de cada organismo, ortólogos 1:1, são identificados e usados nas etapas seguintes do pipeline. Esse procedimento exclui da análise genes parálogos e grupos sem uma das espécies, o que pode ocorrer no caso de genomas incompletos.

Construção do pipeline de desenho de primers

A construção da ferramenta para obtenção dos primers específicos foi feita de forma que o fluxo de processamento de dados fosse o seguinte (Figura 2).

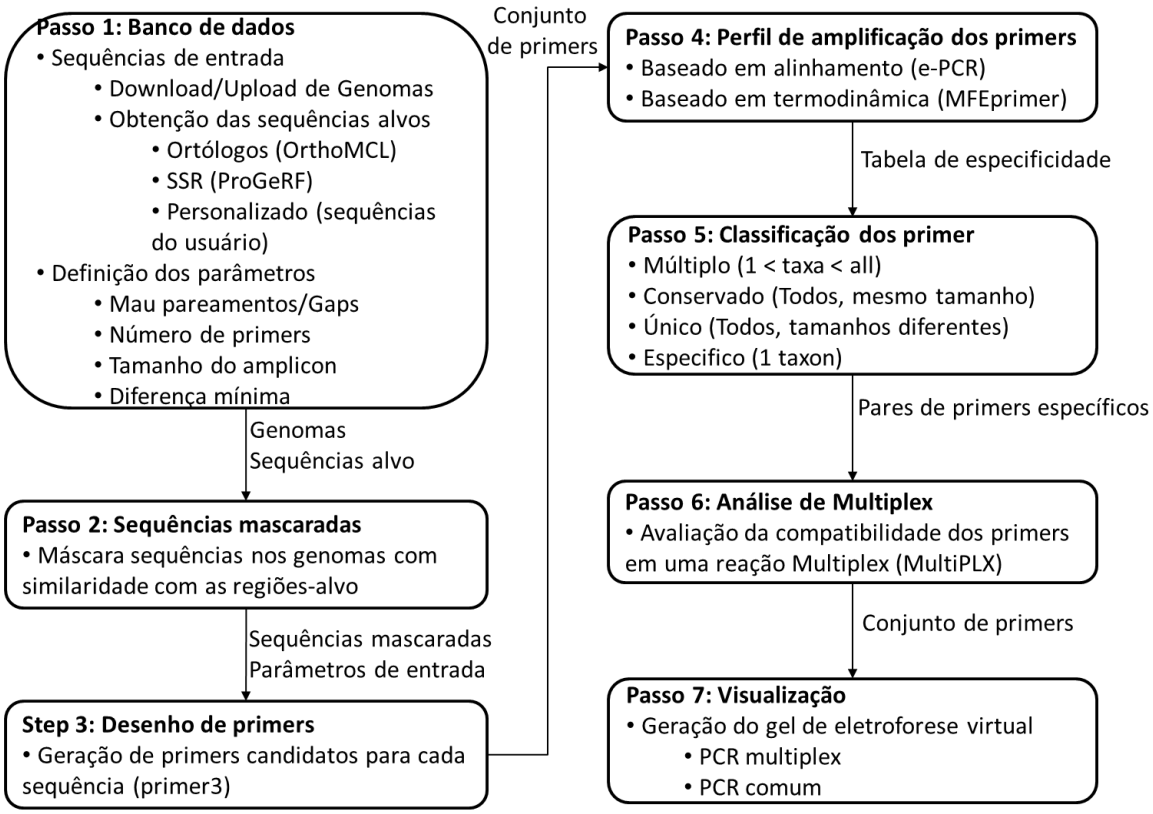


Figura 2: Fluxograma de funcionamento da ferramenta.

Os parâmetros de entrada básicos do pipeline são os genomas, o tipo de análise a ser feita e o tamanho inicial dos produtos de amplificação. A fim de otimizar os resultados, parâmetros mais específicos podem ser modificados, como: número de primers a serem desenhados por alvo, a diferença entre os produtos de amplificação em cada táxon e a estringência dos critérios de especificidade. O valor de estringência corresponde ao número de gaps e mau-pareamentos (*mismatches*) aceitos no anelamento do primer à sequência na execução do e-PCR. Portanto, valores mais altos garantem uma probabilidade menor de ocorrer amplificações inespecíficas, mas diminui a quantidade de primers encontrados no resultado final.

O primeiro passo é a obtenção dos parâmetros de entrada: genomas, sequências alvo e parâmetros da construção de primers. Os genomas devem estar no

formato FASTA e não há limite de quantos taxa poderão ser analisados, entretanto, o tempo de processamento aumenta substancialmente a cada táxon. Existem dois tipos de genomas na análise, os genomas alvo e os de reação cruzada. O primeiro será usado como DNA molde no desenho de primers e na verificação de especificidade. O segundo, são sequências de espécies em que podem ocorrer reações cruzadas na PCR, por exemplo, o genoma de um hospedeiro. Regiões nos genomas alvo em que há similaridades com os estes genomas de reação cruzada são evitadas no passo de desenho de primers.

Uma forma de melhorar a sensibilidade da PCR é encontrar regiões mais adequadas para construção do primer. A ferramenta utiliza três tipos de sequência alvo: microssatélites, genes preditos e sequências definidas pelo usuário, cada abordagem apresenta vantagens. Com exceção da abordagem de microssatélites, para as demais é necessário fornecer sequências adicionais. Nos genes ortólogos é preciso um arquivo no formato FASTA com os genes preditos de cada taxa e nas sequências do usuário, deverá ser fornecido um arquivo no formato FASTA com as sequências alvo pré-definidas pelo usuário para cada taxa.

Os microssatélites são muito polimórficos e tendem a ser menos conservados entre espécies correlatas quando comparado com sequências não repetitivas (GUICHOUX et al., 2011), o que aumenta as chances de encontrar primers específicos. Entretanto, regiões repetitivas tendem a apresentar mais erros durante o sequenciamento e montagem do genoma que as demais regiões. O erro mais comum é em relação ao tamanho exato da repetição, característica que afetaria a reprodução *in vitro* dos resultados. Dessa forma, optou-se por incluir como alvo alternativo sequências que tendem a apresentar uma menor chance de falhas na amplificação, os genes ortólogos 1:1. Por serem sequências de cópia única nos genomas, a montagem destas sequências são menos propensas a erros. Além disso, primers conservados entre taxa são mais facilmente encontrados nessa abordagem. Ao fornecer suas próprias sequências alvo, o usuário não utiliza o mecanismo de busca do TipMT, o que

torna a execução mais rápida. Ao oferecer essas três opções a ferramenta contempla organismos que possuem genomas bem anotados ou não.

Finalmente, os parâmetros para construção de primers são definidos. Como já mencionado anteriormente, o grau de estringência determina o número de maus pareamentos e gaps no anelamento do primer. Quanto maior esse número, menor será a chance de se encontrar amplificações inespecíficas (CAO et al., 2005), porém, muitos primers funcionais poderão ser descartados. Outro parâmetro é o número de primers por sequência alvo, um maior número de primers aumenta as chances de se obter mais primers funcionais, mas compromete o tempo de processamento e uso de recursos computacionais. A determinação dos tamanhos dos produtos de amplificação é feita pelo parâmetro de tamanho inicial e a diferença mínima entre taxa, o primeiro táxon terá o valor inicial e em cada táxon adicional será acrescido o valor da diferença mínima. Esses parâmetros possuem valores padrões de: um primer por alvo, grau de estringência igual à dois, tamanho do produto de amplificação inicial igual à 250 e diferença mínima, 50.

No segundo passo, regiões não ideais para construção de primer são mascaradas. A conservação nas regiões franqueadoras das sequências alvo é essencial, uma vez que primers com um alto número regiões de anelamentos podem causar falhas na PCR (ANDRESON; MÖLS; REMM, 2008). Regiões similares às sequências alvo são identificadas utilizando o MEGABLAST com parâmetros padrão contra todas as sequências genômicas. O MEGABLAST foi escolhido devido à velocidade de processamento e a habilidade de identificar diferenças sutis. Regiões com mais de 95% de identidade foram mascaradas com nucleotídeos minúsculos (inicialmente todas as sequências são alteradas para maiúsculo).

O terceiro passo consiste no desenho de primers propriamente dito. Os primers candidatos são gerados para cada sequência alvo usando o primer3 2.3.5 com parâmetros padrão, alterando-se apenas a opção para rejeitar a construção de primers onde há caracteres em minúsculo na região 3’. Mau-pareamentos na região 3’ dos

primers afetam muito mais a amplificação que aqueles localizados na 5', sendo que dois mau-pareamentos na região 3' em geral impedem a amplificação (YE et al., 2012). Baseado nessa premissa, esse procedimento diminui a chances de ocorrer amplificações inespecíficas ao evitar o desenho de primers em regiões em que há similaridade entre taxa (REMM; ANTS; METSPALU, 2004). As especificações padrão para a seleção dos primers no pipeline são: tamanho igual à 20 nucleotídeos e temperatura ideal de anelamento de 60°C. Nas análises em *Leishmania* foram utilizados os seguintes tamanhos de produto de amplificação: 350, 300 e 250 pares de base para *L. major*, *L. infantum* e *L. braziliensis*, respectivamente.

No quarto passo, todos os primers candidatos gerados são avaliados quanto à especificidade usando um software baseado em alinhamento, e-PCR, e outro baseado em termodinâmica, MFEprimer, ambos utilizando parâmetros padrão e grau de estringência definido no primeiro passo. Se o primer candidato apresentar o mesmo perfil de amplificação nas duas técnicas, ele continua na análise. Durante essa análise é avaliado também a não amplificação das combinações de primers "foward-foward" e "reverse-reverse".

No quinto passo a ferramenta indica todos os pares de primers potencialmente úteis para a diferenciação de taxa. Os primers selecionados são classificados em 4 categorias: especifico, múltiplo, conservado e único. Se o primer apresenta apenas um produto de amplificação no genoma alvo correspondente da sequência alvo, é definido como "específico". Se o primer tem um produto de amplificação em pelo menos mais um genoma, é classificado como múltiplo. Se ocorrem amplificações em todos os genomas com o mesmo tamanho, é um primer "conservado", e com tamanhos diferentes, um primer "único". Primers "únicos" são capazes de distinguir todos os taxa em uma única reação de PCR, pois tem diferentes tamanhos de produtos de amplificação para cada genoma.

No sexto passo, os primers específicos são agrupados por compatibilidade em uma reação de multiplex PCR usando o MultiPLX. Essa ferramenta testa a possibilidade

de ocorrer interações entre primers, como: formação de dímeros e diferenças incompatíveis na temperatura de anelamento.

Finalmente, no sétimo passo, são geradas as saídas visuais do TipMT: géis de eletroforese virtual. Após a escolha de um conjunto de primers pelo usuário, a ferramenta desenvolvida realiza um PCR eletrônica convencional ou multiplex, utilizando o MFEprimer. A função que realiza a PCR eletrônica convencional recebeu o nome de e-GEL e a multiplex, e-MPX. O resultado da execução do MFEprimer são as sequências dos produtos de amplificação encontradas, que são usadas para calcular a distância de migração eletroforética relativa para cada produto pelo pacote do BioPerl, “Bio::Tools::Gel”. Em seguida, uma imagem é construída mostrando o aspecto visual esperado para aquela reação de PCR. Dessa forma, é possível verificar a interação entre primers e genomas molde para a formação de produtos alternativos indesejados.

Ao final de todos esses passos, é esperado um resultado contendo 3 tipos de saída: uma lista de primers divididos pela sua categoria, uma lista de primers divididos em grupos compatíveis e figuras representando os géis de eletroforese virtual.

Desenvolvimento da ferramenta web

A interface gráfica com o usuário foi criada através da programação em HTML, linguagem padrão de criação de hipertexto, e o sistema de gerenciamento de layout foi desenvolvido em CSS. A linguagem PHP e JavaScript foram utilizadas para proporcionar uma dinamicidade na apresentação e recuperação de dados. Finalmente, o MySQL, sistema de gerenciamento de banco de dados relacional, foi utilizado para gerenciar todas as informações geradas pela ferramenta web. A figura 3 representa o fluxo de ações do sistema.

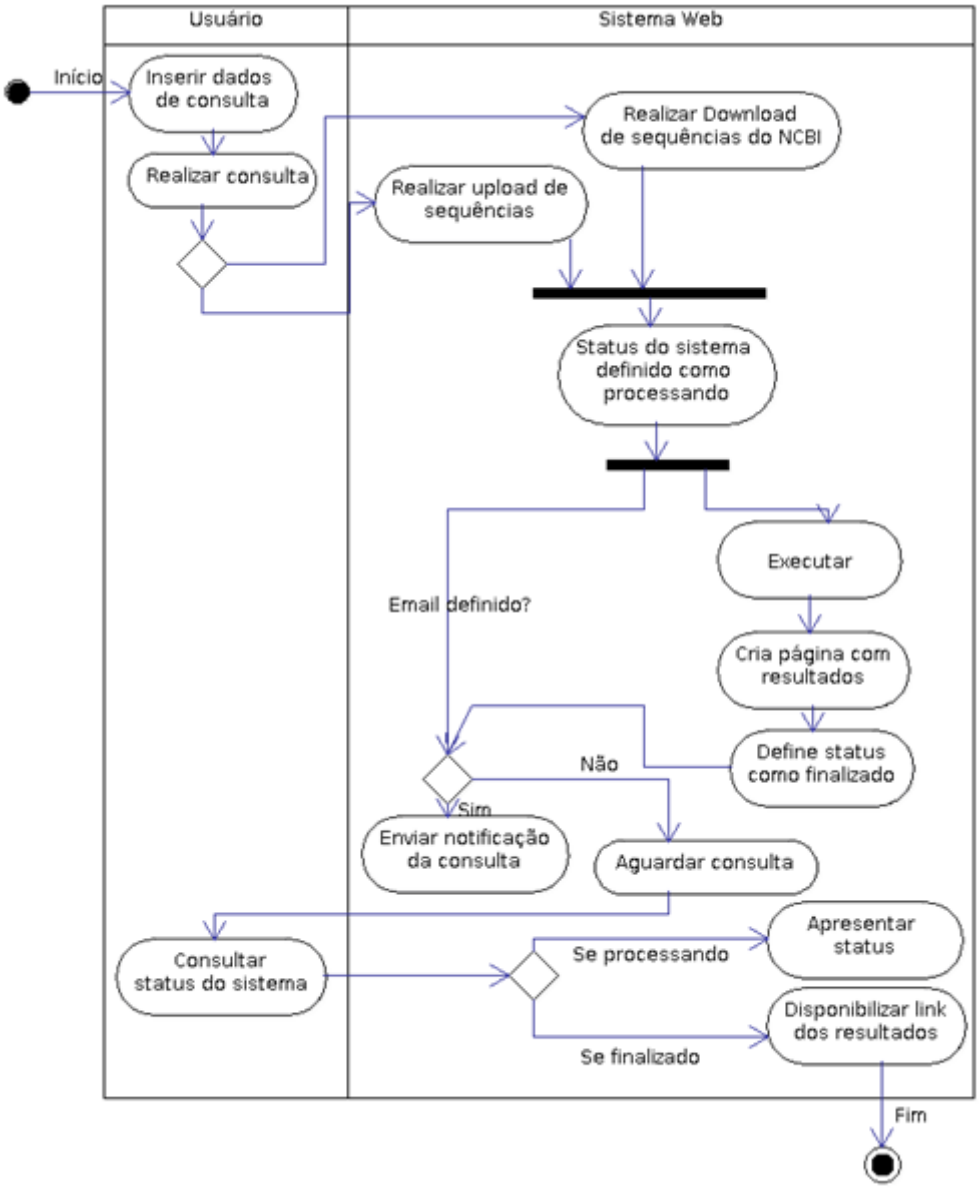


Figura 3: Fluxo de ações do sistema web. Ver detalhes no texto.

Na página principal da ferramenta web (Figura 4), o usuário inicia o processo selecionando o modo de execução: microssatélites, genes ortólogos ou sequências alvos do usuário. O próximo passo é a forma de obtenção das sequências de entrada no formato FASTA: o usuário pode enviar as sequências via upload de sua máquina ou download do banco de dados “RefSeq” do NCBI. Caso seja escolhida a segunda opção, o usuário deverá informar os números de acesso das sequências do NCBI. Se o código estiver incorreto, o sistema acusará um erro.

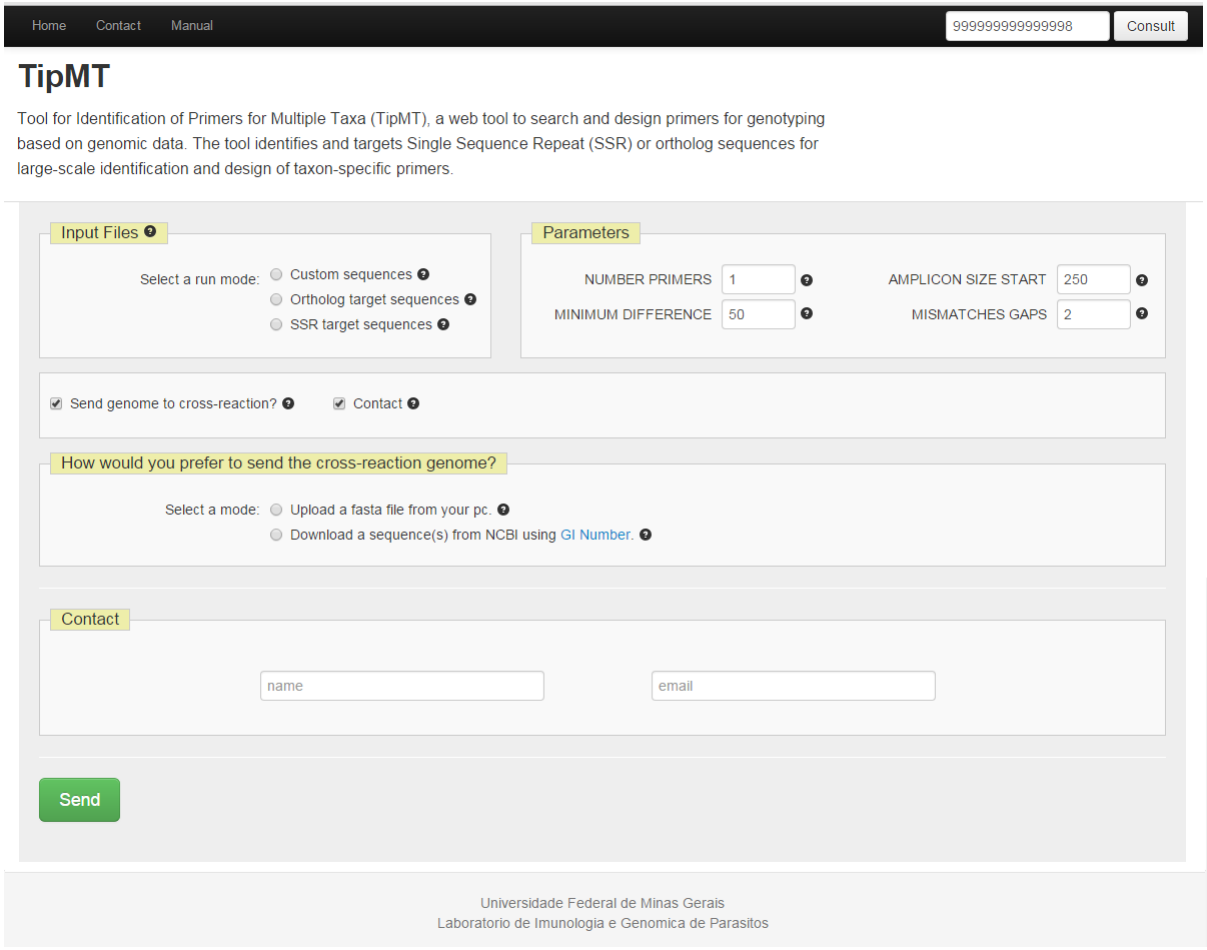


Figura 4: Página principal da ferramenta web, TipMT.

Em seguida, o usuário deve definir os demais parâmetros de entrada: tamanho dos produtos de amplificação, estringência, quantidade de primers por alvo, o tamanho inicial do produto de amplificação e a diferença mínima entre os produtos em cada táxon. Finalmente, o usuário pode fornecer os genomas em que ocorre reação cruzada (da mesma forma que os genomas de entrada) e seu e-mail caso deseje ser notificado após o término do processamento.

Ao enviar o formulário, clicando em "*Send*", o sistema iniciará o processamento e criará um código único, que identifica aquela consulta. Com esse identificador, o usuário pode realizar a consulta do status do seu processo. Se o processo ainda estiver em execução, o sistema apresentará uma página informando essa situação, entretanto, se estiver finalizado, será exibida a página com os resultados (Figura 5).

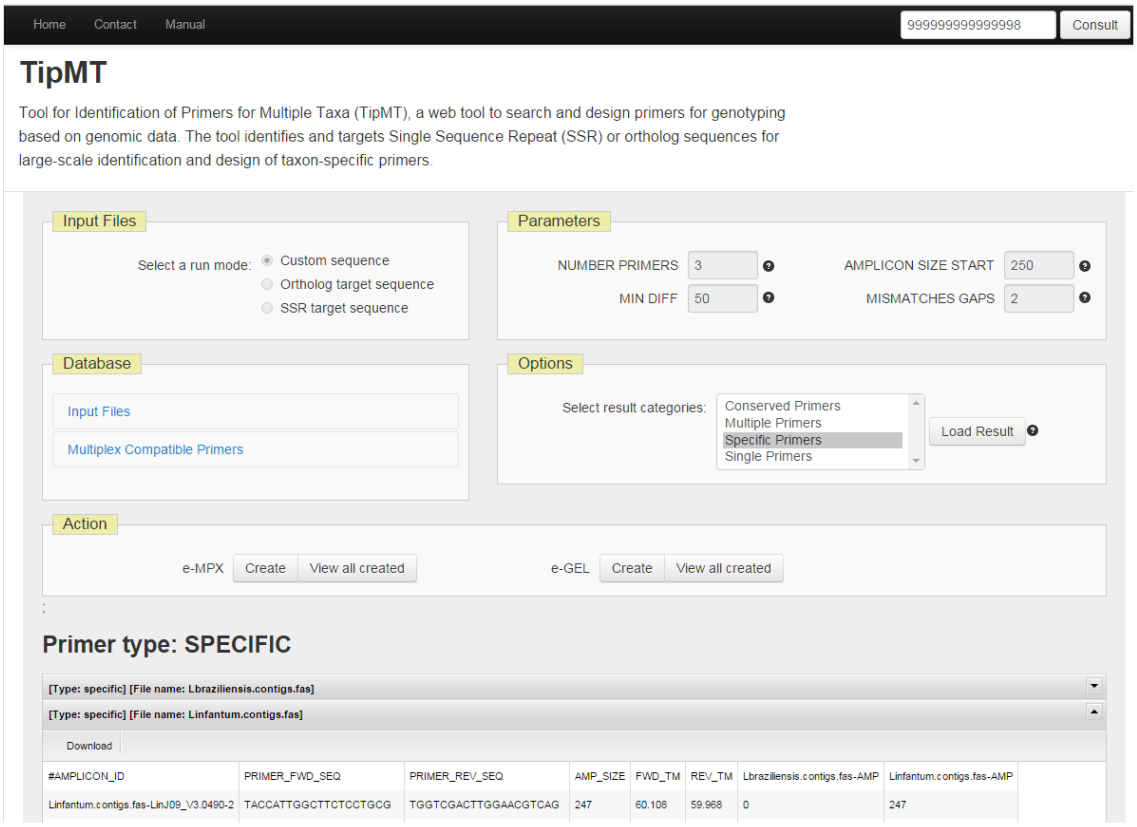


Figura 5: Página com os resultados de uma consulta no TipMT.

Todos os dados de cada execução, sequências e parâmetros de entrada, são apresentados na página de resultado, assim como, os primers para cada táxon e suas respectivas informações: tamanho esperado do produto de amplificação, temperatura de anelamento e classificação.

Os arquivos de saída são: a lista com os pares de primers divididos pela sua classificação, a lista com os pares de primers compatíveis em uma reação de multiplex PCR e o resultado visual, que consiste na imagem de um gel de eletroforese virtual, utilizando os pares de primers selecionados pelo usuário.

As saídas de texto são no formato tabular (Figura 6), onde cada coluna representa as seguintes informações: Nome do produto de amplificação, sequência do primer “forward”, sequência do primer “reverse”, temperatura de anelamento do primer “forward”, temperatura de anelamento do primer “reverse” e tamanho do produto de amplificação em cada genoma alvo.

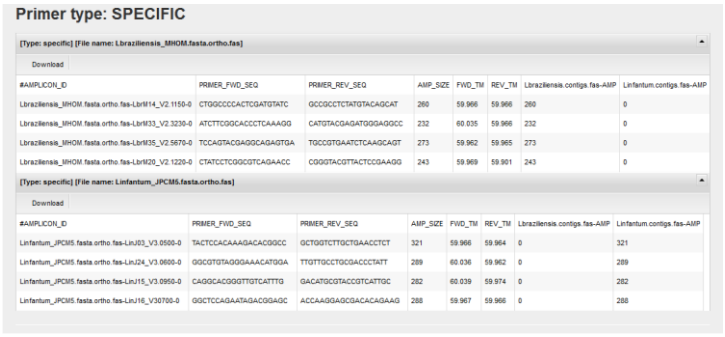


Figura 6: Tabela com as informações dos primers específicos gerados pelo TipMT.

A saída visual consiste na imagem de um gel de eletroforese virtual (Figura 7). Após selecionar um conjunto de pares de primers, o usuário pode gerar duas simulações de PCR: convencional (e-GEL) ou multiplex (e-MPX). Na função e-GEL, cada canaleta é uma reação entre um par de primer selecionado e a mistura de genomas alvo como DNA molde. A função e-MPX gera outro gel virtual com o perfil de amplificação de uma reação de PCR multiplex, onde cada canaleta contém a mistura de todos os pares de primers selecionados e um genoma alvo como molde.

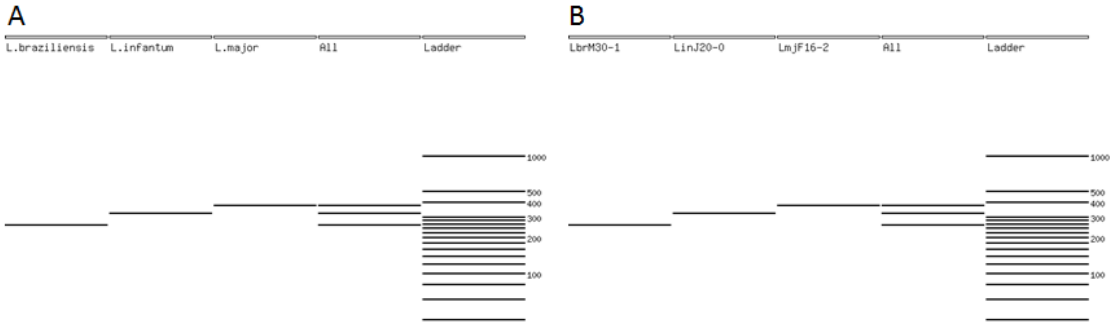


Figura 7: Gel de eletroforese virtual utilizando as funções e-MPX(A) e e-GEL(B) e os primers para genes ortólogos grupo G1.

A aplicação web conta com um manual de instruções onde são apresentados os passos para executar uma análise, descrição dos arquivos de saída e exemplos para serem executados ou, simplesmente, consultados. Devido à restrição de tamanho do arquivo de *Leishmania* de upload pelo servidor em que a ferramenta está hospedada (8 Mb), o exemplo para execução é para o gênero *Brucella* (*B. abortus*, 3Mb, e *B. ovis*,

3Mb) na opção genes ortólogos e o tempo de processamento para esse teste é de 15 minutos. O tempo de execução para *L. braziliensis* (31 Mb) e *L. infantum* (31 Mb) na abordagem de genes ortólogos é de 12 horas.

Validação dos primers específicos preditos

No total foram identificados: 19.246 pares de primers na abordagem para ortólogos e 68 para microssatélites, como mostrado mais detalhadamente na tabela 1.

Abordagem	Específico	Múltiplo	Único	Conservado
Microssatélites	65	3	0	0
Genes ortólogos	15.648	3.267	12	319

Tabela 1: Número de primers desenhados pela ferramenta de acordo com a abordagem e classificação.

Todos os primers apresentam as mesmas características para viabilizar a aplicação dos mesmos em PCR multiplex: 20 nucleotídeos de tamanho e temperatura ideal de anelamento em torno de 60°C. Além disso, os produtos de amplificação têm tamanhos diferentes para permitir a distinção dos mesmos em gel de agarose, a saber: aproximadamente 350, 300 e 250 pares de base para *L. major*, *L. infantum* e *L. braziliensis*, respectivamente. A fim de validar os resultados encontrados *in silico* foram sintetizados seis conjuntos de primers: um par para cada espécie de 3 grupos gerados para cada abordagem (Tabela 2).

Abordagem	Grupo	Espécie	Primer	Sequência Forward	Sequência Reverse	Amplicon (pb)
Microssatélites	G4	*L. braziliensis*	LbrM.20.1-1	CGTGAAGCTGCTTGGCAAAA	AGTGGTGGTGTGCGAAAAGA	225
		L. infantum	LinJ.14-2	GCTTCGAGGCTAACCCGATT	CACTCGCCTTTCCGCTATCT	289
		L. major	LmjF.33-1	TCTAAGTTTGCGCCAGGGTT	CGCGGGGTTTGTACTTGTTG	366
	G10	*L. braziliensis*	LbrM.03-1	AGCCATCGCTCACTAGAAGC	CCTTCCGTGATGCCAGGTAA	275
		L. infantum	LinJ.08-1	TCGATAACTGCACAGCTCGT	TGTGTGTGCTTGTGGCTCAT	301
		L. major	LmjF33-2	TTCTAAGTTTGCGCCAGGGT	CGCGGGGTTTGTACTTGTTG	367
	G0	*L. braziliensis*	LbrM15	TGTTTGGCTTTCTGGCTACA	CACCCACACAGTGACACACA	254
		L. infantum	LinJ09-12	AAGATGAAGCTCCTCCGTCA	CCGACTTCGTCCGTTATTCA	302
		L. major	LmjF07	CATCGTTTCCGTCTGTTGTG	CTCTCCTCTCCTACGCCTTG	351
Genes ortólogos	G718	*L. braziliensis*	LbrM31-2	CTCAGCGTCTCCTCATTGCA	TAGTTTCGCGCACCTCTGAG	229
		L. infantum	LinJ32-1	GCTTTTGCATGTCACCACGT	CGTCCATGCTACCCCTCAAG	305
		L. major	LmjF11-1	GACACCAGAGACAGCCCTTC	GGACAGCTTGGTCGGCTATT	362
	G1	*L. braziliensis*	LbrM30-1	ATCTCGGTGGAGGGAGACAA	AGATGCCAATGGTGGGTTGT	256
		L. infantum	LinJ20-0	GAAGACGGTGGTGAGAGTGG	CTCTTCAAGGGTGCCCAGAG	323
		L. major	LmjF16-2	TGTTGGAAGGACGACGGAAG	TCGAGGAGGAGAGGTGTCAG	375
	G0	*L. braziliensis*	LbrM34	CACCCCAAAAGAATCCAGAA	CTCTTTAGTGGATCAGCGCC	250
		L. infantum	LinJ26	CTTTGATAACATCACCGCCC	CCAAGTTTCTGCAGGTCCTC	306
		L. major	LmjF25	GCCTTTTTGGAACTCTGCTG	TCAACTTTCGAGCAATCACG	349
	Único	*L. major*	LmjF29-1	GCGGTGCTTGAATCACGTTT	GCGGTGTTTACATGACGACG	307-337

Tabela 2: Conjunto de primers sintetizados para validação *in vitro*.

Após a síntese dos conjuntos de primers, os mesmos foram usados em reações de PCR e, em seguida, foi feito a separação dos produtos de amplificação em géis de agarose e o perfil de amplificação obtido foi próximo do resultado predito nas análises *in silico* (Figura 8, 9 e 10).

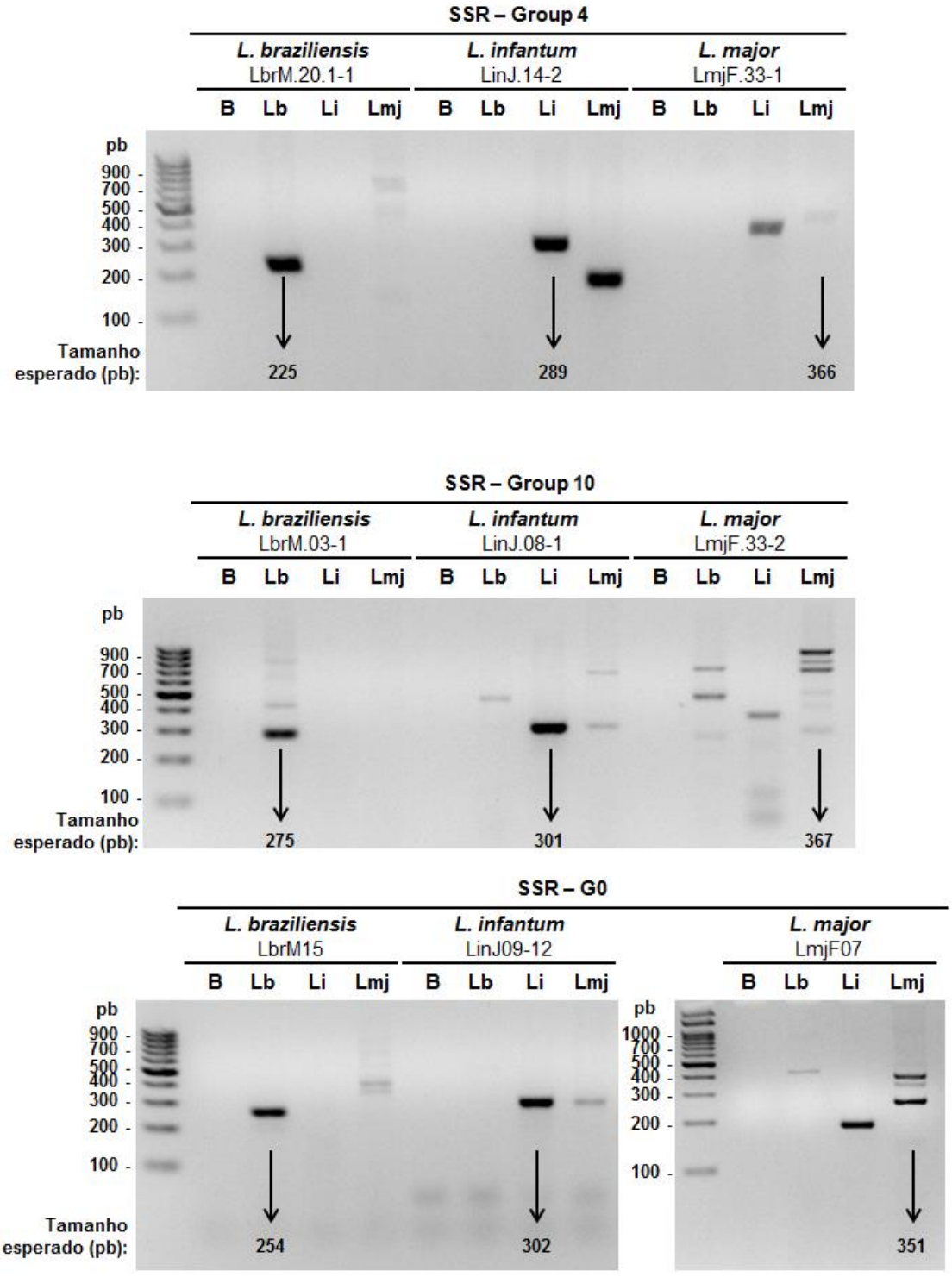


Figura 8: Perfil de amplificação dos conjuntos de primers sintetizados na abordagem de microssatélites. Cada canaleta corresponde à combinação do DNA genômico da espécie de *Leishmania* identificada acima e o primer indicado abaixo. As setas indicam o tamanho esperado para cada primer. Lb: *L. braziliensis*; Li: *L. infantum*; Lmj: *L.* major; B: branco; pb: pares de base;

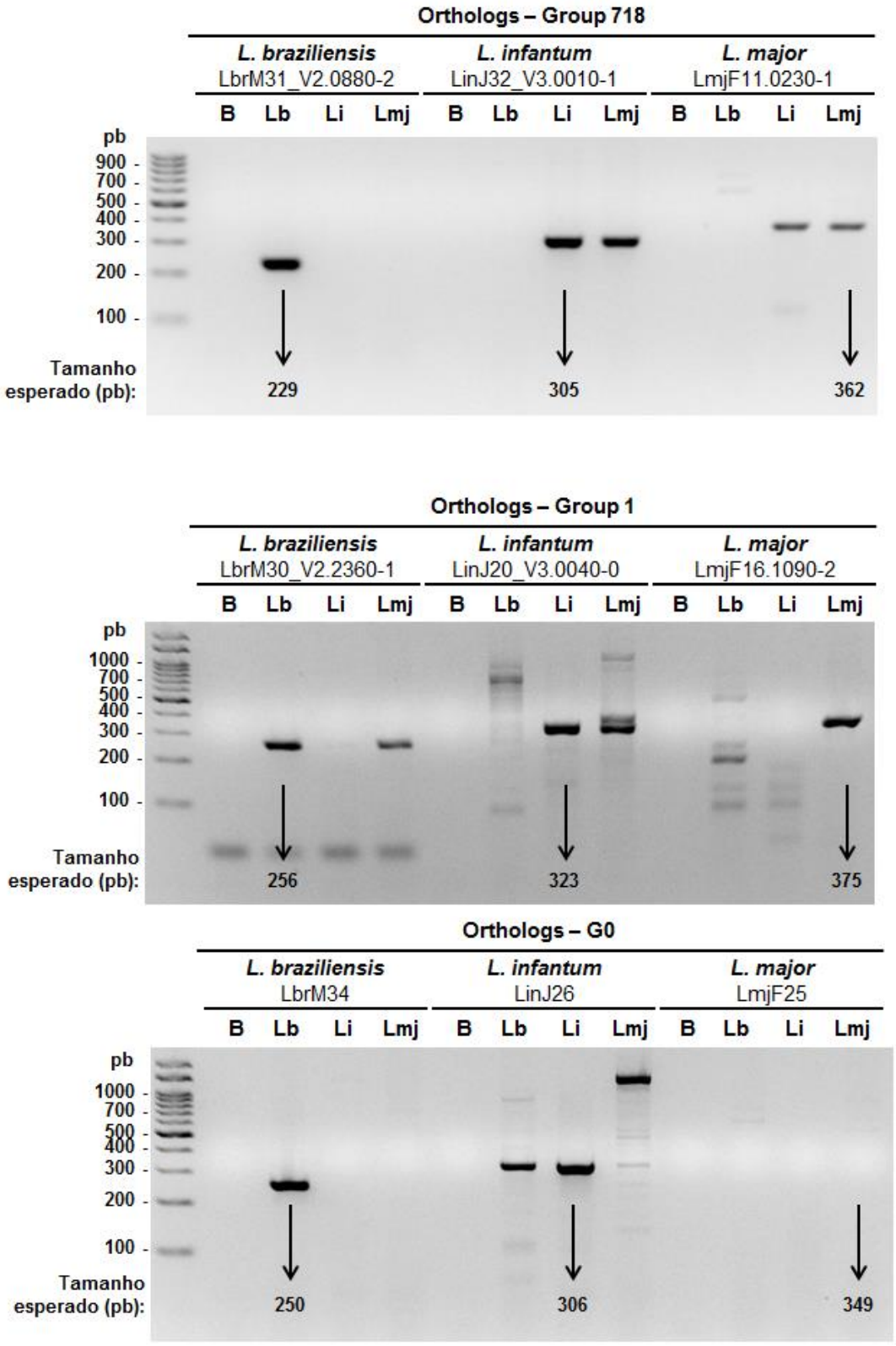


Figura 9: Perfil de amplificação dos conjuntos de primers sintetizados na abordagem de genes ortólogos. Cada canaleta corresponde à combinação do DNA genômico da espécie de *Leishmania* identificada acima e o primer indicado abaixo. As setas indicam o tamanho esperado para cada primer. Lb: *L. braziliensis*; Li: *L. infantum*; Lmj: *L.* major; pb: pares de base;

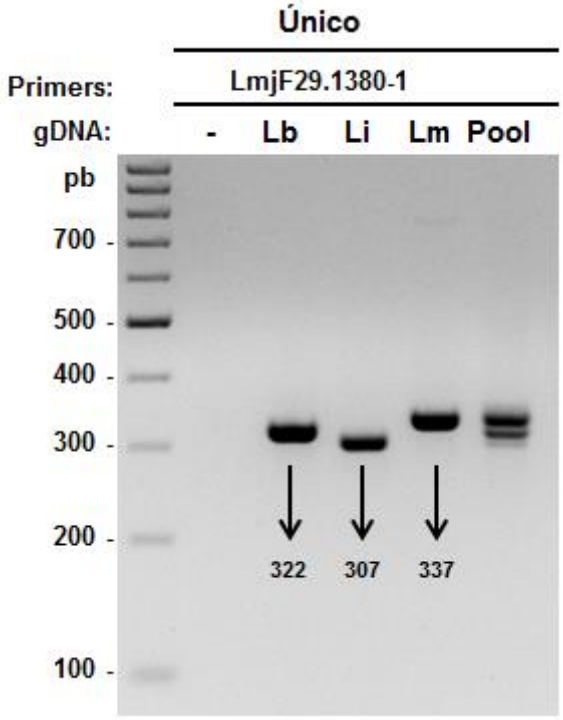


Figura 10: Perfil de amplificação do primer classificado como “único” na abordagem de genes ortólogos. Cada canaleta corresponde à combinação do DNA genômico da espécie de *Leishmania* identificada abaixo e o primer “único”. As setas indicam o tamanho esperado para cada primer. Lb: *L. braziliensis*; Li: *L. infantum*; Lmj: *L.* major; Pool: mistura do DNA genômico das três espécies; gDNA: DNA genômico; pb: pares de base;

Na abordagem dos microssatélites, todos os primers para *L. braziliensis* e *L. infantum* apresentaram a banda de amplificação específica para o seu genoma correspondente e no tamanho esperado. Os primers construídos para *L. major* apresentaram bandas inespecíficas, tanto para o seu genoma quanto para os demais. De qualquer forma, o perfil de amplificação obtido indica que os primers testados podem ser usados para diferenciação de *L. braziliensis* e *L. infantum*, visto que o perfil de amplificação é distinto para duas espécies. Além disso, os primers de *L. braziliensis* e *L. infantum* do SSR grupo 4, resultou em um perfil de amplificação distinto nas três espécies, apesar da amplificação inesperada do primer de *L. infantum* usando como molde o DNA genômico de *L. major*.

Na abordagem de genes ortólogos, o perfil de amplificação dos primers para as três espécies seguiu o padrão observado anteriormente obtido para os alvos de microsatélites, os primers para *L. braziliensis* e *L. infantum* apresentaram o resultado

esperado, enquanto as reações envolvendo *L. major* tiveram amplificações inespecíficas. Por sua vez, o perfil de amplificação do primer "único" foi o esperado contendo apenas as bandas esperadas para cada genoma, 322, 307 e 337 pares de base para *L. braziliensis*, *L. infantum* e *L. major,* respectivamente.

Devido ao bom resultado encontrado para *L. braziliensis* e *L. infantum*, o próximo passo foi verificar o comportamento de primers em uma reação de multiplex PCR, ou seja, utilizar os dois primers em uma mesma reação de amplificação, usando amostras de DNA genômico de *L. braziliensis* e *L. infantum* como molde (Figua 11).

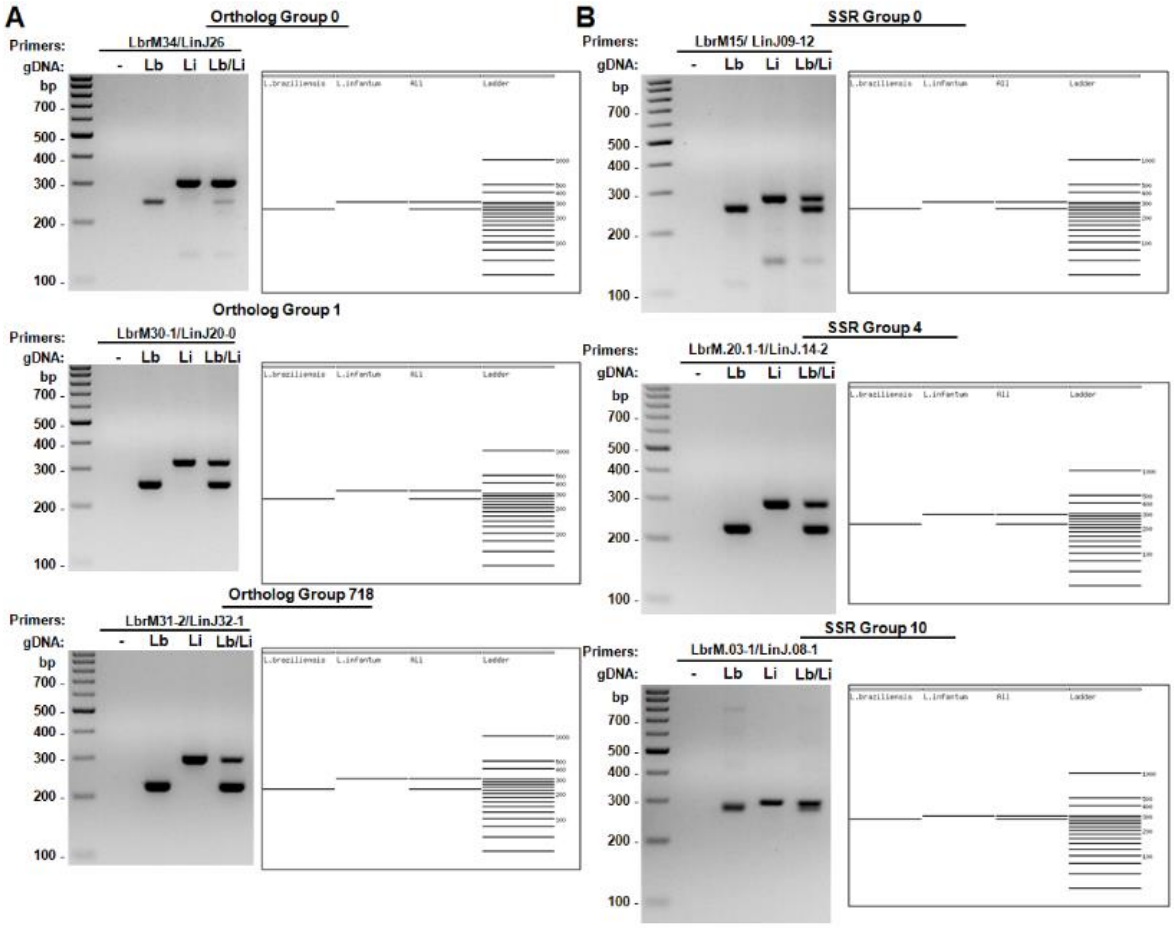


Figura 11: Gel de eletroforese real e virtual para os primers ortólogos (A) e microssatélites (B) em uma reação de multiplex PCR real e simulada (e-MPX). Cada canaleta corresponde a combinação do DNA genômico da espécie de *Leishmania* identificada acima e a mistura de primers. Lb: *L. braziliensis*; Li: *L. infantum*; gDNA: DNA genômico; bp: pares de base;

Comparação com ferramentas similares

Dentre as ferramentas para seleção de pares de primers a partir de uma sequência alvo disponíveis gratuitamente, foram analisadas apenas as que apresentavam características próximas às da ferramenta web desenvolvida nesse trabalho, construção de primers específicos, visto que muitos programas oferecem somente a função de selecionar primers a partir de uma sequência molde. As ferramentas escolhidas foram BatchPrimer3 (YOU et al., 2008), MPprimer (SHEN et al., 2010) e Primer-BLAST (YE et al., 2012). A comparação entre os pontos principais pode ser vista na tabela 3.

Ferramenta	Sequência de entrada	Tipo de alvo	Teste de especificidade	Saída	Compatibilidade multiplex
BatchPrimer 3	Até 500 sequências	SSR, SNP	Nenhum	Informação dos primers	Não
Primer-BLAST	Sequência única	Nenhum	Baseado em alinhamento	Alinhamento dos primers na sequência	Não
MPprimer	Múltiplas sequências	Nenhum	Baseado em termodinâmica	Informação dos primers	Sim
TipMT	Múltiplas sequências	SSR, genes ortólogos	Baseado em alinhamento e termodinâmica	Informação dos primers e gel virtual	Sim

Tabela 3 : Tabela comparativa entre o TipMT e outras aplicações web para desenho de primers especificos.

Um ponto em comum entre as ferramentas é o uso do Primer3 para construção dos primers. Entretanto, o teste de especificidade dos mesmos nem sempre era possível, o que torna os resultados menos confiáveis.

Observou-se também que as regiões alvos mais comuns dos programas são os microssatélites e SNPs. Entretanto, a ferramenta que determina alvo não verifica a possibilidade de amplificação inespecífica. Por sua vez, os serviços que testam especificidade exigem que o usuário forneça a sequência alvo, sendo que o Primer-BLAST só aceita uma por execução.

Cada ferramenta apresenta características peculiares. O BatchPrimer3 oferece diversas abordagens para construção de primers como: para sequenciamento, alelo específico, hibridização de oligos, microssatélites, genotipagem por SNP. Já o Primer-BLAST, fornece um teste de especificidade utilizando o BLAST contra qualquer genoma disponível no banco de dados do NCBI. Por sua vez, o MPprimer verifica a especificidade em apenas um número limitado e pré-definido de genomas, em sua maioria de organismos modelos, disponíveis em seu banco de dados. A determinação da especificidade é baseada na estabilidade termodinâmica entre os pares de primers e o DNA molde. Além disso, possui uma função que busca otimizar os pares de primers para uma PCR multiplex.

O TipMT aceita como entrada múltiplas sequências, identifica regiões alvo automaticamente e a especificidade dos primers é testada utilizando duas técnicas, baseada em alinhamento e termodinâmica. Como saída, a aplicação gera um arquivo texto com as informações gerais dos primers e possui uma função em que o usuário visualiza o resultado de uma reação de PCR, convencional e multiplex, o gel de eletroforese virtual.

DISCUSSÃO

O surgimento de novas tecnologias de sequenciamento de DNA proporciona uma produção de dados em larga escala, abrindo novas oportunidades para as análises pós-genômicas. O gênero *Leishmania* possui diversas espécies com genomas sequenciados, tendo sido o genoma de *L. major* o primeiro a ser publicado (IVENS et al., 2005). O genoma de *L. major* foi sequenciado usando a estratégia de sequenciamento *clone-by-clone,* ou seja o mapa físico de cada cromossomo já havia sido elucidado em estudos anteriores e, portanto, foi possível selecionar um conjunto mínimo de clones de tamanho de inserto grande (~150kb) que cobrisse toda a extensão de cada cromossomo. Estes clones individuais eram então submetidos à técnica de *shotgun* e a sequência original de cada clone reconstituída durante a montagem do

clone sequenciado (IVENS et al., 2005). Como a unidade de montagem nesta estratégia é o clone, esta etapa é extremamente menos complexa quando se compara com a estratégia de *whole genome shotgun*, em que todo o genoma é fragmentado e então montado. A estratégia de sequenciamento do genoma utilizada associada à natureza pouco repetitiva do genoma de *Leishmania* favoreceu a produção de um genoma de alta qualidade como o de *L. major*. Cabe salientar, entretanto, que os cromossomos montados de *L. major* representam pseudomoléculas, sendo cada cromossomo constituído por mosaicos de cada cromossomo homólogo. A disponibilidade de um genoma de referência (*L. major)* de alta qualidade e a alta similaridade e sintenia no gênero favoreceu a montagem e obtenção de genomas de outras espécies de *Leishmania* com qualidades bastante satisfatórias (PEACOCK et al., 2007; REAL et al., 2013; ROGERS et al., 2011).

A descoberta de marcadores moleculares não só ajuda no tratamento de questões biológicas, mas também facilita diversas aplicações, como mapeamento genético, diagnóstico de doenças, investigação forense, análise populacional, estudos ecológicos, paternidade e biologia da conservação (DURAN et al., 2009; SHARMA; GROVER; KAHL, 2007). Várias ferramentas foram desenvolvidas focando na construção de iniciadores, mas limitações dificultam o seu uso no desenho de primers para marcadores moleculares, seja pela pouca experiência computacional do usuário ou pela falta de funções na ferramenta (ABD-ELSALAM, 2003; SOBHY; COLSON, 2012). Como visto anteriormente na comparação entre ferramentas, a maioria das ferramentas disponíveis atualmente não possui a opção de busca de regiões alvo e/ou verifica a especificidade dos primers gerados. Outra lacuna, é o teste de compatibilidade entre pares de primers para uma reação de PCR Multiplex.

A aplicação que determina regiões alvos, microssatélites e SNPs, é o BatchPrimer3, que também oferece diversas abordagens para construção de primers (YOU et al., 2008). Entretanto, o teste de especificidade dos primers não é verificada, o que aumenta a possibilidade de ocorrer amplificações inespecíficas. O Primer-BLAST,

fornece um teste de especificidade utilizando o BLAST contra qualquer genoma disponível no banco de dados do NCBI, mas a ferramenta não determina alvo e só aceita uma sequência de entrada por execução (YE et al., 2012). O MPprimer também não busca regiões alvo, mas aceita múltiplas sequências e verifica a especificidade em um conjunto limitado e pré-definido de genomas (SHEN et al., 2010). Finalmente, essa aplicação possui uma função que busca otimizar os pares de primers para uma PCR multiplex. A ferramenta desenvolvida no presente trabalho oferece, portanto, uma combinação de características que não está presente em outras aplicações web. O TipMT aceita múltiplas sequências como entrada, identifica regiões alvo automaticamente e a especificidade dos primers é testada utilizando duas técnicas, baseada em alinhamento e termodinâmica. Como saída, a aplicação gera um arquivo texto com as informações gerais dos primers e possui uma função em que o usuário visualiza o resultado de uma reação de PCR, convencional e multiplex, o gel de eletroforese virtual.

Como o gênero *Leishmania* possui espécies com genomas sequenciados de qualidade, os quais apresentam alto grau de similaridade, optou-se por realizar a validação da ferramenta desenvolvida neste projeto com estes organismos. O TipMT foi capaz de identificar alvos apropriados e desenhar pares de primers, que na validação experimental diferenciaram as espécies corretamente. Os perfis encontrados para *L. braziliensis* e *L. infantum* foram bastante satisfatórios, pois se observou apenas as amplificações específicas. Entretanto, para *L. major* não foi possível distinguir um padrão de amplificação claro. A presença de amplificações inespecíficas pode ser devido ao sequenciamento e montagem dos genomas. A representação dos cromossomos montados de *L. major* como pseudomoléculas (IVENS et al., 2005; PEACOCK et al., 2007) torna a verificação de especificidade *in silico* falha, dificultando a identificação dessas amplificações inespecíficas. Os resultados encontrados para *L. major* podem também ser explicados, em parte, pela natureza muitas vezes errática dos experimentos de PCR e pela grande similaridade com *L. infantum* (ABD-ELSALAM,

2003; PEACOCK et al., 2007). A identidade nucleotídica média entre os genes ortólogos de *L. major* e *L. infantum* é de 94,78% (entre *L. infantum* e *L. braziliensis* é de 85,60%, para *L. major* e *L. braziliensis* é 85,42%), o que justifica a maior dificuldade em identificar primer específicos quando se analisa os genomas de *L. major* e *L. infantum*.

A ferramenta precisa ainda ser testada em outros genomas, mas a partir dos nossos resultados é possível dizer que a ferramenta é eficiente quando se analisa genomas com menos de 90% identidade, uma vez que desenhou primers capazes de distinguir *L. infantum* e *L. braziliensis* de forma correta e forneceu um gel virtual que corresponde ao resultado encontrado em um experimento real. Além disso, o comportamento dos primers para *L. braziliensis* e *L. infantum* em uma reação de multiplex PCR, seguiu o esperado, pois se observou um perfil bem próximo do criado pela função e-MPX. Finalmente, apesar da dificuldade encontrada para os primers específicos em *L. major*, com o primer "único" foi possível determinar a presença das três espécies em uma só reação de PCR.

Ao priorizar a construção de pares de primers específicos foi necessário excluir a chance de ocorrer amplificações cruzadas (REMM; ANTS; METSPALU, 2004). No entanto, durante a concepção dos primers percebeu-se que muitas sequências alvo eram perdidas, deixa-se de construir pares de primers, devido ao passo onde a sequência alvo é mascarada nas regiões em que possui similaridade com outros genomas. Dessa forma, percebeu-se a necessidade de criar uma nova opção para buscar primers. Para próximas versões, haverá uma opção em que o passo de exclusão de regiões similares não será aplicado, a fim de aumentar o número de iniciadores gerados e, consequentemente, o número de pares de primers taxa conservado.

CAPÍTULO II

APRESENTAÇÃO

O termo virulência engloba fatores microbiológicos, patológicos, ecológicos, e evolutivos, com uma complexidade adicional intrínseca ao organismo estudado. Informações sobre a evolução da virulência são fundamentais para vários campos da biologia, incluindo a investigação sobre o surgimento de novos patógenos, o desenvolvimento de vacinas e a compreensão da dinâmica e propagação de doenças infecciosas (RAFALUK et al., 2015; RIVAS et al., 2004).

A virulência ou o potencial patogênico de *Leishmania* é determinado por fatores ambientais e genéticos, relacionados aos hospedeiros mamíferos e invertebrados. Diversos genes já foram descritos como fatores de virulência em *Leishmania* e, em geral, participam de processos relacionados à tolerância a estresse, invasão celular e evasão do sistema imune. Quanto aos determinantes moleculares, pouco se sabe sobre os mecanismos envolvidos nesse processo (COELHO et al., 2012; RIVAS et al., 2004). Uma das primeiras abordagens para se estudar a dinâmica da evolução da virulência são experimentos de passagem em série. O cultivo seriado *in vitro* frequentemente leva a deleções no genoma de parasitas e ilhas de patogenicidade são frequentemente perdidas durante o crescimento microbiano nestas condições (EBERT, 1998).

Em *Leishmania*, a alteração na virulência por cultivo seriado já foi descrita em *L. major* (SILVA; SACKS, 1987), *L. donovani* (MUKHOPADHYAY et al., 1998), *L. amazonensis* (MAGALHÃES et al., 2014) e *L. braziliensis* (REBELLO et al., 2010). Esses estudos demonstraram que, em geral, os genes envolvidos na perda de virulência participam da tolerância à temperatura e estresse oxidativo, evasão e modulação do sistema imune (BIFELD; CLOS, 2015).

Neste contexto, entender como o parasito modula a expressão diferencial de genes ao longo do ciclo de vida e das passagens em cultura é uma importante questão a ser compreendida. A identificação dos genes envolvidos quer na infectividade dos parasitos nos hospedeiros mamíferos e invertebrados ou na sua manutenção em

culturas, devem ser, portanto, consideradas relevantes. Esta informação pode melhorar o entendimento sobre a interação parasito-hospedeiro e levar ao descobrimento de novos genes ou padrões associados com a infectividade e virulência de parasitos (TEIXEIRA et al., 2012).

METODOLOGIA

Cultivo dos parasitos e avaliação de infectividade

Os parasitos foram isolados a partir de lesões de camundongos BALB/c experimentalmente infectados com *L. amazonensis* (IFLA/BR/67/PH8) e, em seguida, lavados em meio Schneider (Gibco) com 1% de penicilina/ estreptomicina (Sigma) e 10% de soro fetal bovino (SFB) (Sigma) inativado pelo calor, e cultivados em meio Schneider completo. As passagens *in vitro* foram feitas a cada 5 dias até a 30ª passagem. Alíquotas dos parasitos cultivados foram recolhidas nos dias 0 (1ª passagem) e 150 dias de cultura (30ª passagem), como descrito por MAGALHÃES et al., 2014.

Alíquotas de cada passagem foram centrifugadas por 10 minutos e 2000 x *g*, à 4ºC. O sobrenadante foi removido e o *pellet* contendo os parasitos foi lavado três vezes com PBS estéril. Macrófagos murinos foram retirados de camundongos BALB/c e plaqueados em placas de cultura com 24 poços, à concentração de $5x10^5$ células por poço em meio RPMI suplementado com 10% de soro fetal bovino, 2 mM de L-glutamina, 200 U/mL de penicilina e 100 ug/mL de sulfato de estreptomicina, com o pH 7,4. Após 2 horas de incubação à 37ºC em 5% CO_2, promastigotas em fase estacionária foram quantificadas e adicionadas aos poços ($1x10^6$ e $5x10^6$, na proporção 1:2 e 1:10 macrófagos por parasitos, respectivamente). As culturas foram incubadas por 24 horas à 37ºC em 5% CO_2. Em seguida, as células foram lavadas e fixadas para determinar a percentagem de macrófagos infectados e o número de formas amastigotas intracelulares do parasito por macrófago, contando um total de 200 células. Esta etapa foi realizada em triplicata.

Obtenção das bibliotecas

As culturas foram lavadas e centrifugadas com PBS estéril e transferidas para tubos do 1,5 mL *Nuclease-free* após a coleta e imediatamente congeladas em nitrogênio líquido. Em seguida as amostras foram transferidas para um freezer a temperatura de -80ºC até o momento da extração do RNA. A extração foi feita utilizando o kit *Total RNA isolation* (Macherey-Nagel, Duren, Alemanha) de acordo com as recomendações do fabricante. As amostras de RNA de cada passagem foram quantificadas por um espectrofotômetro NanoDrop Spectrophotometer ND-1000 (Thermo Scientific, Wilmington, EUA) e a integridade das amostras de RNA total avaliada no equipamento Agilent 2100 Bioanalyzer (Agilent Technologies).

Construção das bibliotecas e sequenciamento

O enriquecimento para RNA mensageiro, síntese do cDNA e sequenciamento pela plataforma Illumina HiSeq 2000 foram realizados pela BGI Tech (China).

Para construção de bibliotecas de RNA-seq pair-end, o RNA mensageiro foi enriquecido usando oligos poli-T ligados a beads magnéticas. Utilizando um tampão de fragmentação, os RNAs foram clivados com tamanhos em torno de 200 pb, e então a primeira fita de cDNA foi sintetizada usando primers randômicos. Após a síntese da segunda fita, cDNAs de fita dupla foram purificados, lavados e as pontas foram convertidas em extremidades abruptas e adeniladas na extremidade 3'. Em seguida, os adaptadores de sequenciamento foram ligados nos fragmentos e enriquecidos por amplificação. Finalmente, as bibliotecas foram sequenciadas.

Processamento e mapeamento das reads

As sequências geradas foram analisadas utilizando o FastQC(FASTQC), um filtro de qualidade das reads. Após o filtro de qualidade, os transcritos foram mapeados no genoma de referência de *L. amazonensis* disponível em http://bioinfo08.ibi.unicamp.br/leishmania/, utilizando programas que alinham pequenas sequências a grandes genomas de forma veloz e eficiente, o BWA (LI; DURBIN, 2009)

e Bowtie2 (LANGMEAD; SALZBERG, 2012). Para todos os mapeamentos foram utilizados os parâmetros padrões de cada software. Em seguida, foram avaliados os parâmetros mínimos de qualidade, como os níveis de cobertura de cada região e o número de transcritos por sequência utilizando os pacotes SAMtools (LI et al., 2009a) e Bedtools (QUINLAN; HALL, 2010). Todas as execuções e análises foram feitas em ambiente Linux, utilizando a linguagem Perl e Bash Script, em uma máquina com a seguinte configuração: Intel(R) Xeon(R) CPU E5640 2.67GHz e 24 Gb de memória RAM DDR3.

Estimativa dos níveis de transcrição

A definição de um mapa preciso de todos os transcritos que são expressos em uma condição exige a montagem das reads alinhadas em unidades de transcrição. Antes de realizar as análises comparativas foi necessário, portanto, determinar os níveis de expressão dos genes ou transcritos de interesse e normalizar os dados, a fim de permitir uma estimativa acurada e sem tendenciosidades. O método utilizado para isso foi a ferramenta featureCount do pacote Subread (LIAO; SMYTH; SHI, 2014), uma ferramenta veloz, eficiente e precisa para quantificação do nível de expressão.

Análise da expressão gênica diferencial entre os tratamentos

Finalmente, após a quantificação dos transcritos, foi feita a análise estatística dos dados e análise diferencial entre as condições. Para isso foi usado o pacote EdgeR (*Empirical Digital Gene Expression analysis using* R) (ROBINSON; MCCARTHY; SMYTH, 2009) e scripts em R, software para análises computacionais estatísticas e gráficas. Foi aplicada a correção *False Discovery Rate* (FDR) visando corrigir os erros associados a múltiplos testes (BENJAMINI; HOCHBERG, 1995). Para a seleção dos genes diferencialmente expressos, foram aceitos valores com FDR ≤ 5%.

Avaliação funcional dos genes diferencialmente expressos

Um experimento bem sucedido de RNA-seq produz uma lista de genes diferentemente expressos entre os tratamentos. A lista em si não contribui muito para o

entendimento do fenômeno observado, sendo necessário inferir a função biológica daquele conjunto. Outra informação de interesse é identificar vias metabólicas pouco/bastante representadas. Para isso foi utilizado o Blast2GO (CONESA et al., 2005), uma ferramenta muito usada em anotação, visualização e análise funcional de dados genômicos. A avaliação de sublocalização celular foi feita utilizando a ferramenta online TargetP1.1 (EMANUELSSON et al., 2000) e a determinação de domínios proteicos foi por meio da ferramenta Batch Web CD-Search Tool (MARCHLER-BAUER et al., 2015). Ambas ferramentas foram executadas utilizando os parâmetros padrões de análise.

RESULTADOS

Avaliação das amostras

A avaliação de infectividade das amostras, R0 e R30, confirmou os achados dos trabalhos anteriores, ou seja, a diminuição da taxa de infecção após diversas passagens em cultura axênica (MAGALHÃES et al., 2014). Na tabela 4 é mostrado a média dos valores encontrados nas triplicadas.

MOI	**2:1**		**10:1**	
Amostra	**R0**	**R30**	**R0**	**R30**
Número de amastigotas por macrófago	1,78±0,174	0,9±0,028	7,1±0,360	2,87±0,687
Percentagem de macrófagos infectados	71,9±2,205	35,3±0,873	96,7±1,703	64,3±1,320

Tabela 4 : Percentagem de macrófagos infectados e relação entre o número de amastigotas por macrófago em infecções *in vitro* em cada amostra. MOI: *multiplicity of infection* (razão do número de promastigotas por número macrófagos). Valores representam a média de triplicatas e o desvio padrão

A integridade dos RNAs foi analisada pelo equipamento 2100 Bioanalyser e em todas as amostras pode-se visualizar os picos referentes aos rRNAs 24Sα 24Sβ e 18S, típicas de tripanosomatídeos (CAMPBELL et al., 1987). Os perfis dos eletroferogramas

(Figura 12) confirmam que todas as amostras apresentaram RNA íntegro, adequados à construção das bibliotecas.

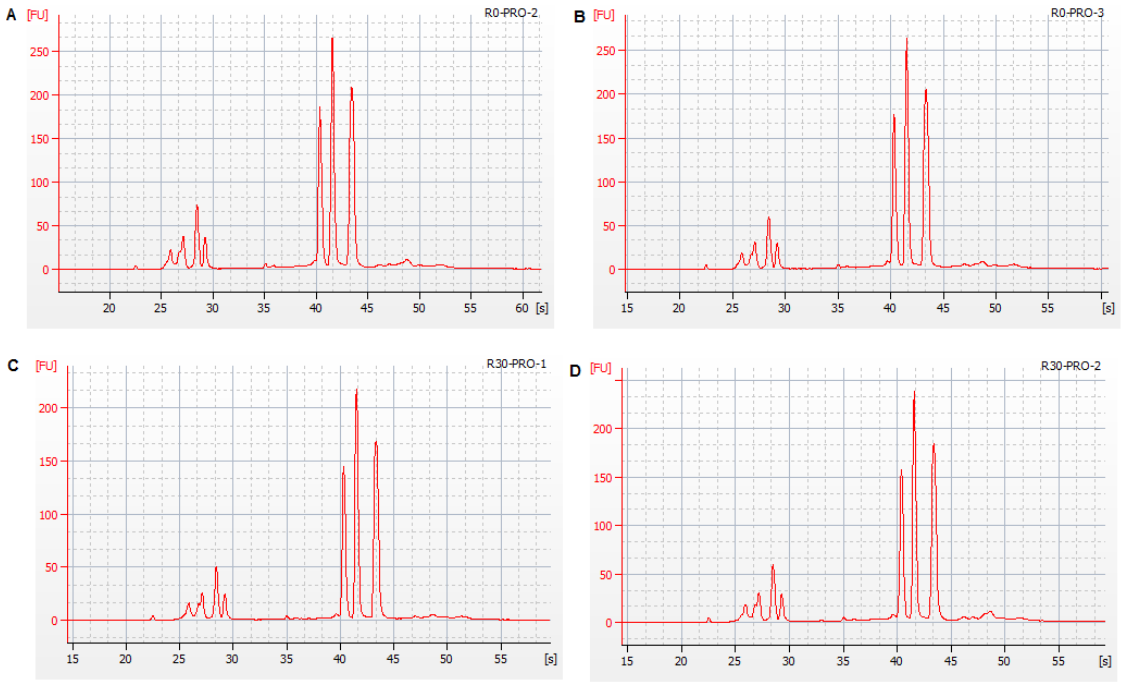


Figura 12: Eletroferogramas para cada amostra em duplicata sequenciada, R0 (A e B) e R30 (C e D).

Análise do sequenciamento

Quatro bibliotecas foram sequenciadas, correspondentes aos dois momentos de amostragem, R0 e R30, contendo em cada momento, duas repetições biológicas do tratamento.

Foi obtido um total de 27,35 milhões de leituras de 100 pb, com uma média de 6,4 milhões de leituras por biblioteca (Tabela 5). A quantidade de leituras por replicata em cada tratamento foi bastante homogênea, assim como a percentagem de alinhamento, 86,57%.

Biblioteca	RNA Concentração (ng/µL)	Número de leituras (milhões)	Percentagem de leituras alinhadas	Percentagem de leituras múltiplas
R0-1	813	5,62	86,17	2,81

R0-2	801	6,47	86,57	2,75
R30-1	534	7,44	85,46	2,70
R30-2	321	7,82	86,49	2,42

Tabela 5: Dados das quatro bibliotecas sequenciadas.

Os transcritos obtidos tiveram sua qualidade analisada pela ferramenta FastQC (FASTQC, [s.d.]), que fornece uma análise de qualidade para cada biblioteca. Após verificar a boa qualidade das sequencias (Figura 13) o próximo passo foi o mapeamento dos transcritos no genoma referência.

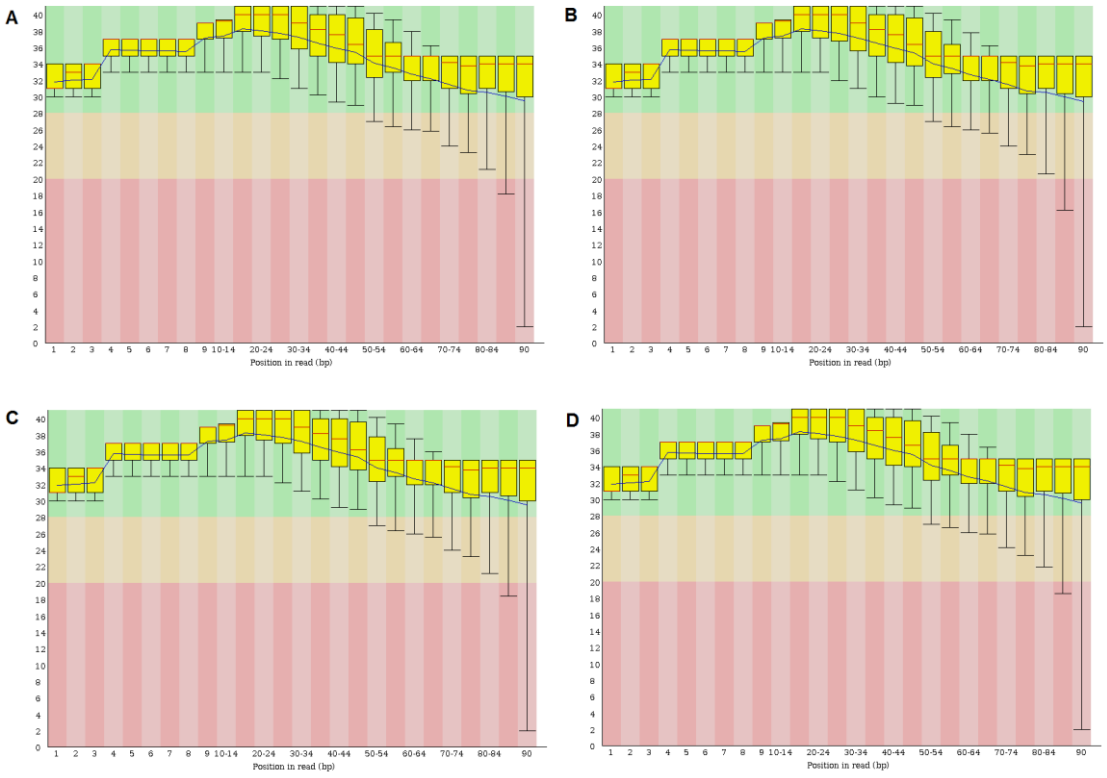


Figura 13: Gráficos com a qualidade média por base para cada amostra em duplicata R0 (A e B) e R30 (C e D). A cor de fundo dos gráficos representa o nível de qualidade das bases: muito boa (verde), razoável (laranja) e ruim (vermelho).

Mapeamento dos transcritos

Após o sequenciamento e avaliação de qualidade, os transcritos foram mapeados utilizando o genoma de referência. Conforme descrito na metodologia, foram utilizados os *software* Bowtie2 e BWA para o mapeamento baseado em referência. O

BWA apresentou uma maior percentagem média de transcritos mapeados, 86,2%, enquanto o Bowtie2, 81,7%. Preferiu-se, portanto, continuar as análises com os resultados obtidos com o BWA e apenas as leituras pareadas foram considerados nas análises seguintes.

Avaliação dos níveis de transcrição

Com os dados de mapeamento prontos, o número de leituras alocadas em cada gene é transformado em uma contagem digital da expressão gênica para todos os genes em cada condição de estudo. Para isso foi utilizado a função featureCount do pacote SubRead (LIAO; SMYTH; SHI, 2014) e o arquivo contendo as informações gênicas de *L. amazonensis.* Em seguida, a tabela com o número de transcritos por gene em cada biblioteca (Figura 14) foi analisada no edgeR (ROBINSON; MCCARTHY; SMYTH, 2009), pacote estatístico que apresenta uma boa performance na descoberta de verdadeiros positivos com o valor de FDR menor que 0,05 (ZHANG et al., 2014). Nesse trabalho foi utilizado o CPM (contagens de transcritos por milhão), onde número de contagens por gene é divido pelo tamanho da biblioteca (em milhões), para determinar com precisão a estimativa do nível de expressão uma vez que em trabalhos de análise diferencial estamos interessados em mudanças relativas de expressão entre as condições, ao invés de expressão absoluta (LAW et al., 2014).

GeneId	RO-1	RO-2	R3O-1	R3O-2
A22470	1582	1905	2543	2422
A22480	1443	1768	2183	2305
A22460	4994	5429	6974	6575
A22500	393	520	659	646
A22490	531	525	801	715
A22510	430	508	620	583
A22520	267	310	367	264
A22590	1053	1090	1984	1781
A22600	928	1086	1532	1446
A22610	814	1023	1416	1386
A22620	1959	2363	3656	3484
A22630	1250	1159	1765	1492
A22540	6	8	12	13

Figura 14: Exemplo da tabela contendo as contagens para cada gene. Na primeira coluna estão os identificadores dos genes e nas demais a contagem de transcritos em

cada biblioteca. R0-1 e R0-2, bibliotecas no tempo zero; R30-1 e R30-2, bibliotecas após 30 passagens em cultura.

A fim de minimizar artefatos devido ao baixo número de contagens, genes fracamente expressos foram retirados das análises (ANDERS et al., 2013).

As contagens foram corrigidas pelo método TMM do edgeR (*Trimmed Mean of M values*) (ROBINSON; OSHLACK, 2010), que corrige possíveis alterações no número de contagens para cada gene. A técnica calcula um fator de correção baseado na média ponderada da variação dos genes sem expressão diferenciada entre as amostras (RAPAPORT et al., 2013).

Em seguida, uma função é utilizada para estimar a dispersão entre as réplicas biológicas, o coeficiente de variação biológica (BCV). Para os dados desse trabalho foi encontrado o valor de 0,0715, que indica coerência entre as réplicas (HAYDOCK et al., 2015). Utilizando a técnica de "multidimensional scaling" (MDS) (ROBINSON; OSHLACK, 2010) e de análise de agrupamento hierárquico usando a distância euclidiana como métrica pode se verificar a relação entre as amostras (Figura 15). Ao analisarmos a dimensão 1 do MDS, ambas replicatas biológicas estão bastante próximas entre si e distantes entre tratamentos. As réplicas biológicas de R30 mostraram maior variabilidade biológica, demonstrada pela distância observada na dimensão 2, o que pode ser explicada pela própria complexidade biológica desse tipo de tratamento com longos tempos de cultura. Da mesma forma, quando a distância euclidiana entre as amostras foi calculada e usada para criar o *heatmap,* observou-se um padrão que descreve a separação entre os tratamentos e a proximidade entre as réplicas.

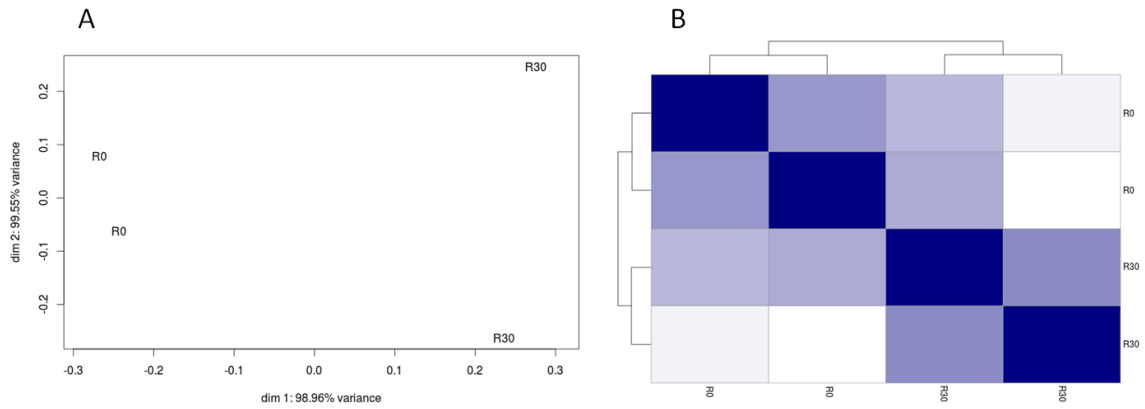


Figura 15: Relação entre as bibliotecas sequenciadas de *L. amazonensis* em um gráfico MDS (A) e em um heatmap de uma análise de agrupamento hierárquico usando a distância euclidiana como métrica (B). R0, bibliotecas no tempo zero; R30, bibliotecas após 30 passagens em cultura.

Uma vez verificada a consistência dos dados e qualidade técnica das amostras, o próximo passo foi determinar a expressão diferencial através do teste binomial negativo exato (ROBINSON; MCCARTHY; SMYTH, 2009). A partir do resultado do teste é possível identificar os genes diferencialmente expressos de acordo com o grau de significância.

Análise da expressão gênica diferencial entre os tratamentos

A análise de expressão diferencial dos genes e o número de genes significativos foi determinado pelo FDR, que corrige o valor *p* para múltiplas hipóteses pelo método de taxa de falsas descobertas. Obteve-se 626 genes significativos com FDR de 5% e se considerada alta estringência estatística (FDR 1%), foram identificados 297 genes significativos, sem redundância (Tabela 6).

Esse baixo número pode ser devido ao baixo poder estatístico inerente a um experimento com duas réplicas. Quando usado FDR igual a 5%, todos os genes selecionados apresentaram uma diferença de expressão de pelo menos 1,3 vezes (logFC maior que 0,3), o que aponta para uma significância biológica dos dados quando usado essa estringência estatística. O valor de 5% já foi descrito como o valor capaz de descobrir verdadeiros positivos e minimizar os falsos positivos (ZHANG et al., 2014).

Regulação/FDR	1%	5%	10%
Negativamente	216	414	629
Positivamente	81	212	354
Genes significativos	297	626	983

Tabela 6: Número de genes diferencialmente expressos em cada nível de significância (FDR).

Finalmente, do total de 8.100 genes preditos foram encontrados 626 diferencialmente expressos (FDR 5%), sendo 212 regulados positivamente e 414, negativamente. No gráfico de nível de expressão (logaritmo de “fold-change”, logFC) por abundância cada ponto representa um gene sendo que os vermelhos são aqueles significativamente expressos. Pelo gráfico podemos verificar que a maior parte dos genes com expressão diferencialmente são negativamente regulados e possuem um nível de expressão menor que duas vezes (linha azul, logFC = -1). Devido à utilização de apenas duas réplicas biológicas para cada tratamento e a variação biológica intrínseca da amostra após 30 passagens, o número de genes detectados como diferencialmente expressos não foi muito alto, principalmente considerando valores maiores de nível de expressão (logFC).

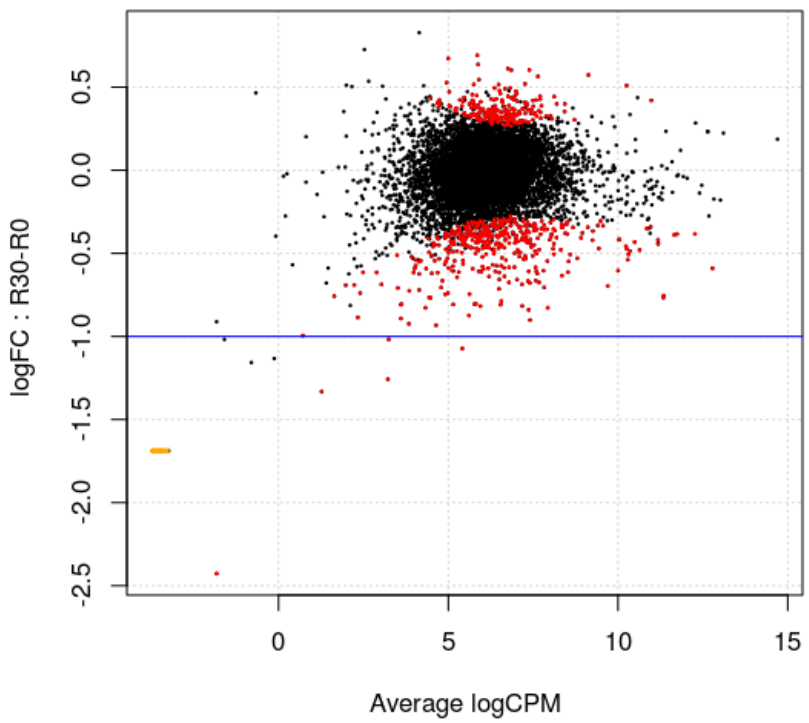


Figura 16: Gráfico com os genes diferencialmente expressos. Cada ponto representa um gene, sendo que os vermelhos são os significativamente expressos e a linha azul o ponto em que o nível de expressão é maior que duas vezes. logCPM, logaritmo das contagens de transcritos por milhão; logFC, nível de expressão (logaritmo de *fold-change*)

Para visualizar o perfil geral dos genes identificados como diferencialmente expressos foram feitos dois heatmaps, com agrupamentos hierárquicos dos genes, baseado nos níveis de expressão, um com todos os genes diferencialmente expressos (Figura 17) e outro com os 25 genes com maior diferença de expressão (Figura 18 e Tabela 7). Para realizar esses agrupamentos, foi utilizado o valor Z, ou valor de distribuição normal, de cada gene significativo para todas as réplicas biológicas, o que permite determinar quantos desvios padrão acima ou abaixo da média está a diferença entre a contagem observada e a contagem média em unidades de erro padrão. O agrupamento refletiu apenas o padrão de genes regulados positivamente ou negativamente, não apresentando subgrupos de genes regulados em conjunto.

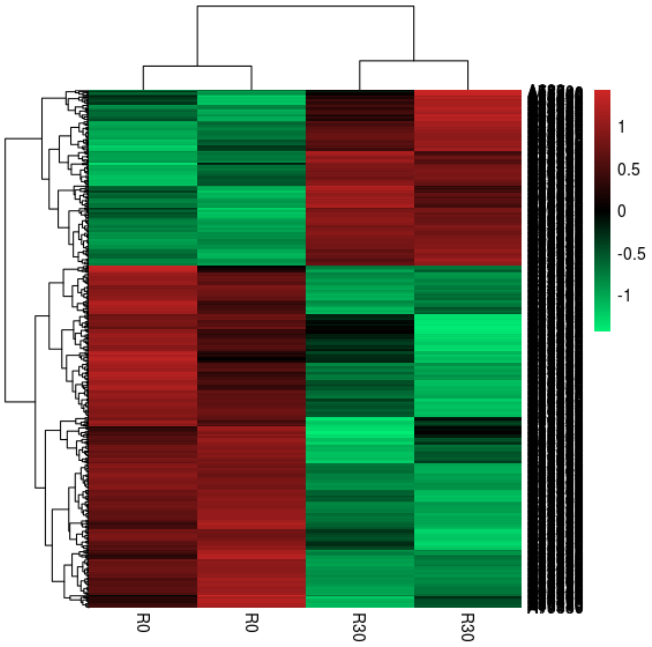


Figura 17: Heatmap dos 626 genes diferencialmente expressos entre os tratamentos agrupados pelo nível de expressão. As colorações verdes e vermelhas indicam, respectivamente, a diminuição e o aumento da expressão. A escala apresenta o valor Z.

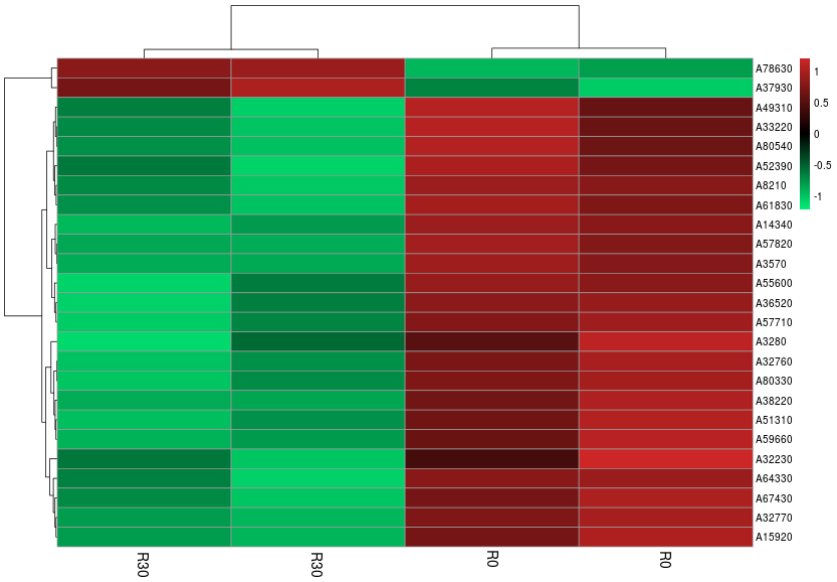


Figura 18: Heatmap dos 25 genes com maior diferença de expressão entre os tratamentos agrupados pelo nível de expressão. As colorações verdes e vermelhas indicam, respectivamente, a diminuição e o aumento da expressão. A escala apresenta o valor Z.

Identificador	logFC	Anotação
A37930	0.693261251462541	Pteridine transporter, putative
A78630	0.610785517210929	Glucosamine-fructose-6-phosphate aminotransferase, putative
A3280	-1.074946918212400	Nuclear lim interactor-interacting factor-like protein
A57710	-0.585219445158210	Delta-6 fatty acid desaturase, putative
A55600	-0.901550565546550	Conserved hypothetical protein
A59660	-0.934879531324690	Conserved hypothetical protein
A14340	-0.638750032910875	GrpE protein homolog
A33220	-0.656179033477561	Conserved hypothetical protein
A8210	-0.657802288204317	Fibrillarin
A80540	-0.663483992014052	Conserved hypothetical protein
A36520	-0.674416107299401	Conserved hypothetical protein
A3570	-0.674827247825761	Conserved hypothetical protein
A52390	-0.678124417568128	Ribosomal protein S6, putative
A64330	-0.686357284635476	RNA binding protein rbp16, putative
A61830	-0.696964800545597	DNA-binding protein HEXBP
A32770	-0.713839891086591	Polyadenylate-binding protein 1, putative
A51310	-0.789031583926092	Major surface protease gp63, putative
A80330	-0.806714104035247	Hypothetical protein
A49310	-0.807466698541907	Transcription factor, putative
A15920	-0.808432318704384	D-tyrosyl-tRNA(Tyr) deacylase
A32230	-0.817363222661118	Tuzin-like protein
A57820	-0.817399637016227	Hypothetical protein
A67430	-0.827512272451949	Amastin-like surface protein, putative
A32760	-0.840337036053002	Hypothetical protein
A38220	-0.875471873530039	Tagatose-6-phosphate kinase-like protein

Tabela 7: Os 25 genes com maior diferença de expressão entre os tratamentos. logFC, nível de expressão (logaritmo de *fold-change*).

Avaliação funcional dos genes diferencialmente expressos

Os genes foram analisados do ponto de vista funcional usando a ferramenta Blast2GO, que permite a identificação das funções gênicas por similaridade, pela associação com os termos oriundos do GO (Gene Ontology) e das vias metabólicas disponibilizadas pelo KEGG (Kyoto Encyclopedia of Genes and Genomes). As comparações foram feitas utilizando a amostra R0 como referência, portanto, genes diferencialmente expressos negativamente estão menos abundantes na amostra R30 que em R0.

Dos 626 genes significativos, 224 (35,58%) codificam para proteínas hipotéticas, que contribuem pouco nessa análise, mas muitas dessas proteínas podem ter um papel importante no processo de infectividade e ainda não tiveram a sua função determinada. Desse grupo de proteínas, apenas 24 (9,83%) são secretadas, apresentam peptídeo

sinal, e 59 (24,18%) apresentam peptídeo de direcionamento para a mitocôndria. Quanto a predição de domínios, 160 (65,57%) proteínas hipotéticas não apresentaram similaridade com algum domínio.

Dos genes regulados positivamente ou negativamente, 94 e 229 deles tiveram suas sequências associadas a termos GO, respectivamente. As análises de termos GO presentes, divididos em categorias ontológicas basais: processo biológico, função molecular e componente celular, foram realizadas tanto para todos os genes diferencialmente expressos (Figura 19), quanto para aqueles com nível de expressão aumentada ou diminuída (Figura 20).

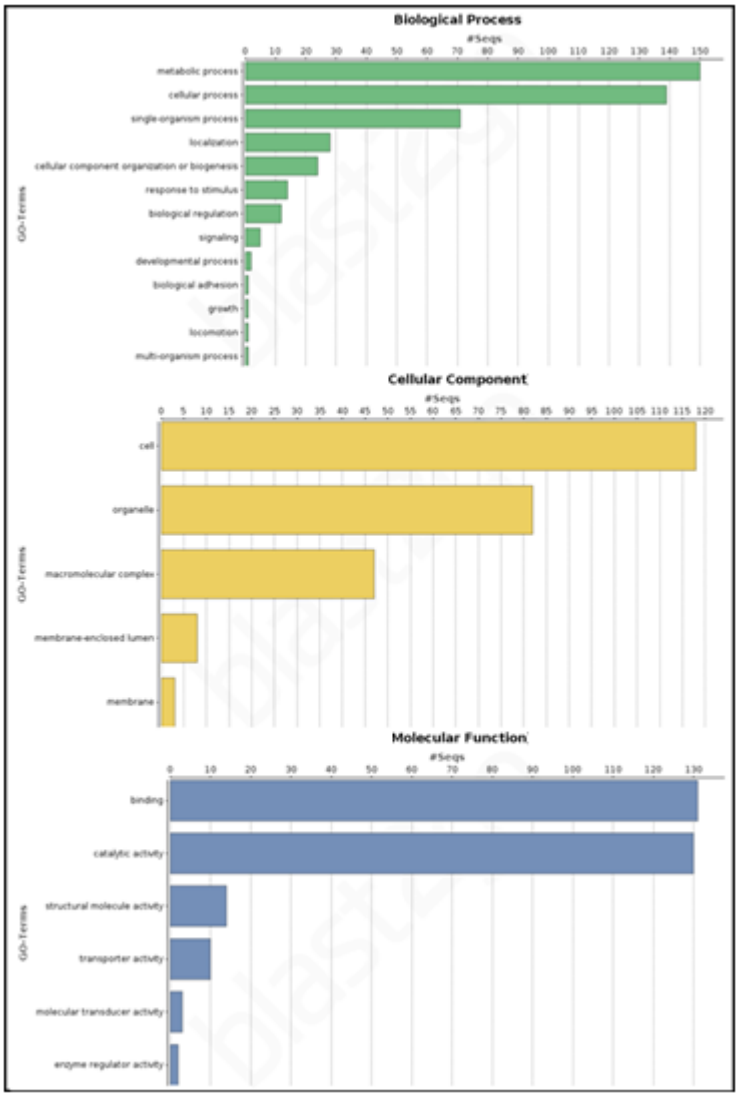


Figura 19: Termos GO encontrados em processos biológicos (verde), celulares (amarelo) e moleculares (azul) em todos os genes com expressão diferencial.

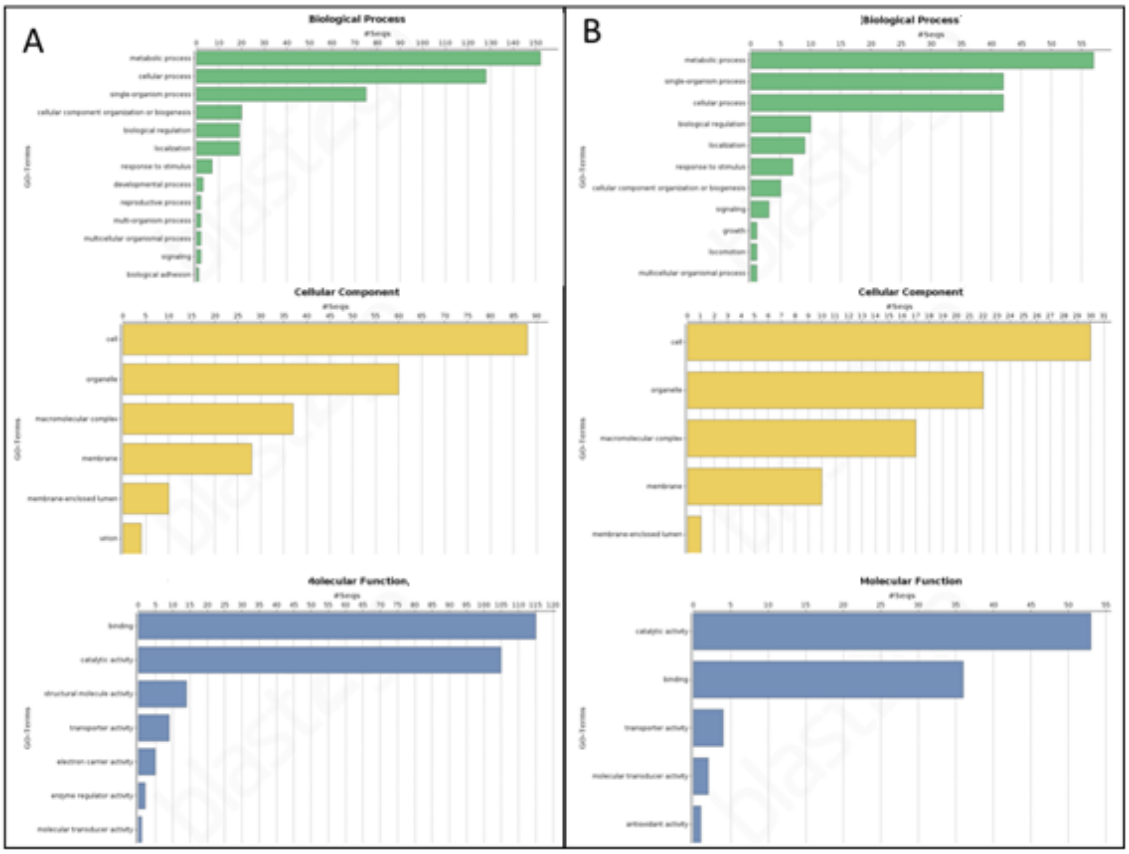


Figura 20: Termos GO encontrados em processos biológicos (verde), celulares (amarelo) e moleculares (azul) nos genes com expressão diminuída (A) e aumentada (B). As comparações foram feitas utilizando a amostra R0 como referência.

As principais categorias funcionais dos genes envolvidos na diminuição da infectividade em *L. amazonensis* foram associadas ao metabolismo (196 genes), processo celular (164 genes), atividade catalítica (154 genes), ligação (151 genes) e processos de organismos unicelulares (109 genes). Quando considerado separadamente os genes de expressão aumentada ou diminuída, percebe-se que os processos moleculares apresentam a maior diferença entre termos, a expressão aumentada possui a "atividade catalítica" como termo mais expressivo enquanto "ligação" é o termo mais abundante no grupo de genes com expressão diminuída. Além disso, há categorias exclusivas de cada expressão diferencial, como: "interação entre organismos" e "adesão celular", na regulação negativa, e "proliferação celular", na positiva.

O mapeamento nas vias metabólicas do KEGG permitiu verificar que os genes aqui identificados como diferencialmente expressos participam de 56 vias do metabolismo (Tabela 8).

Vias metabólicas	Número de sequências		
Vias	**Todos**	**Negativamente**	**Positivamente**
Purine metabolism	31	16	15
Thiamine metabolism	21	11	10
Pyrimidine metabolism	10	7	3
Alanine, aspartate and glutamate metabolism	5	4	1
Aminobenzoate degradation	5	4	1
Cysteine and methionine metabolism	5	5	0
Arginine and proline metabolism	4	4	0
Fatty acid biosynthesis	4	4	0
Glyoxylate and dicarboxylate metabolism	4	2	2
Valine, leucine and isoleucine degradation	4	1	3
Glycolysis / Gluconeogenesis	3	1	2
Inositol phosphate metabolism	3	2	1
Nicotinate and nicotinamide metabolism	3	1	2
Propanoate metabolism	3	1	2
Pyruvate metabolism	3	3	0
Amino sugar and nucleotide sugar metabolism	2	0	2
Ascorbate and aldarate metabolism	2	2	0
Citrate cycle (TCA cycle)	2	2	0
Fatty acid degradation	2	2	0
Galactose metabolism	2	1	1
Glycerolipid metabolism	2	2	0
Glycerophospholipid metabolism	2	1	1
Lysine degradation	2	2	0
Methane metabolism	2	1	1
Nitrogen metabolism	2	1	1
Oxidative phosphorylation	2	2	0
Pentose and glucuronate interconversions	2	2	0
Pentose phosphate pathway	2	2	0
Phosphatidylinositol signaling system	2	1	1
Steroid degradation	2	0	2
Steroid hormone biosynthesis	2	0	2
beta-Alanine metabolism	1	1	0
Biosynthesis of unsaturated fatty acids	1	1	0
Fructose and mannose metabolism	1	1	0
Glutathione metabolism	1	1	0
Glycine, serine and threonine metabolism	1	1	0
Histidine metabolism	1	1	0
Steroid biosynthesis	1	0	1
Tryptophan metabolism	1	1	0

Tabela 8: Vias metabólicas representadas nos genes diferencialmente expressos.

Vias metabólicas que parecem estar menos ativas em promastigotas com baixa infectividade: metabolismo de aminoácidos (alanina, aspartato, glutamato, cisteína,

metionina, arginina, prolina, glicina, serina, treonina, histidina, triptofano), síntese/degradação de ácidos graxos, metabolismo do piruvato, ciclo do ácido cítrico, fosforilação oxidativa, via das pentoses, metabolismo de glutationa e metabolismo de drogas. Por sua vez, a via síntese/degradação de esteróides parece ser a mais ativa.

DISCUSSÃO

A disponibilidade e exploração dos dados gerados nos projetos genomas das diferentes espécies de *Leishmania* é uma grande oportunidade de obter uma compreensão mais ampla da patogenia da leishmaniose. O presente estudo abrangeu dois tópicos considerados prioridade na pesquisa da leishmaniose: identificação de novos marcadores de diagnóstico e descoberta de fatores ligados à virulência do parasito (DESJEUX, 2004).

As tecnologias de sequenciamento de nova geração podem ser usadas também para acelerar a identificação de biomarcadores e foram aplicadas com sucesso em estudos de resistência a drogas em *Leishmania* (LEPROHON et al., 2015). Nesse trabalho foi proposto também a identificação de marcadores de virulência em *L. amazonensis* após sucessivas passagens do parasito *in vitro* e análise do perfil de expressão gênica pela técnica de RNA-seq.

A alteração na virulência em *Leishmania* devido ao cultivo a longo prazo *in vitro* e *in vivo* já foi descrito na literatura (LEI; ROMINE; BEETHAM, 2010; MAGALHÃES et al., 2014; MOREIRA et al., 2012; REBELLO et al., 2010; SILVA; SACKS, 1987), mas as causas ainda não foram totalmente elucidadas. Neste estudo, foi avaliado o perfil de expressão de *L. amazonensis* após passagens sucessivas de promastigotas em cultivo *in vitro*, investigando se há associação entre a sua expressão e a perda da virulência dos parasitos. De fato, o cultivo seriado diminuiu a virulência dos parasitos, modificação observada pela infecção *in vitro* dos macrófagos murinos. Esse experimento demonstrou que infecções utilizando parasitos de passagens iniciais (R0) apresentavam mais macrófagos infectados e por um número maior de amastigotas que as passagens tardias (R30). Essa diferença significativa de virulência entre a 1ª e 30ª passagens pode estar relacionada a mudanças que ocorrem nos antígenos de superfície, transcritos e até na estrutura dos cromossomos (BIFELD; CLOS, 2015; LEPROHON et al., 2015; ROGERS et al., 2011).

A análise dos genes diferencialmente expressos em parasitos R0 e R30 revelou 626 genes com expressão diferencial significativa (66,13% com expressão diminuída), com FDR 0,05 e uma diferença de pelo menos 1,3 vezes. A expressão diferencial de genes derivados de uma mesma unidade de transcrição policistrônica dos tripanosomatídeos implica em uma regulação gênica baseada nos mecanismos pós-transcricionais por meio do controle da estabilidade e da eficiência de tradução do mRNA (CAMPBELL; THOMAS; STURM, 2003; CLAYTON, 2002; MARTIN et al., 2014). Alterações mais sutis na expressão gênica são, portanto, comumente esperadas devido essa peculiaridade na expressão gênica dos tripanosomatídeos (DILLON et al., 2015; MARTIN et al., 2014). Do total de genes encontrados, 224 (35,58%) codificam para proteínas hipotéticas e dessa fração, 160 (65,57%) não apresentaram similaridade com algum domínio protéico conhecido. A caracterização funcional dessas proteínas pode revelar novas vias e estratégias destes parasitos associados à virulência e assim contribuir no entendimento do processo de infectividade do parasito.

Autores sugerem que passagens sucessivas *in vitro* favorecem a propagação de parasitos menos virulentos incapazes de se transformarem nas formas metacíclicas infectivas (LEI; ROMINE; BEETHAM, 2010; MOREIRA et al., 2012). Entretanto, no estudo aqui apresentado verificou-se que a expressão do gene marcador de diferenciação para a forma metacíclica, o META1 (BAÑULS; HIDE; PRUGNOLLE, 2007; GAMBOA et al., 2007), não sofreu alteração. MAGALHÃES et al., 2014 também observou em seu trabalho uma percentagem homogênea de promastigotas em fase estacionária em todas culturas, iniciais e após cultivo seriado. Pode se sugerir, portanto, que a perda de virulência não pode ser associada exclusivamente ao menor número de promastigotas infectivas nas culturas seriadas.

O processo de diferenciação para a forma metacíclica está ligada a uma redução de processos celulares, replicação de DNA e montagem de nucleossomos, processos ligados à tradução, metabolismo de proteínas e energia. Esse achado é consistente com a baixa demanda metabólica neste estágio de desenvolvimento do parasita (DILLON et

al., 2015). No estudo aqui apresentado, também verificamos uma diminuição desses processos nas amostras após 30ª passagens em cultura. Entretanto, os processos regulados positivamente como, sinalização celular e resposta ao estresse (DILLON et al., 2015), não foram detectados na amostra R30. A partir dessa observação pode se dizer, então, que ao longo das passagens *in vitro* os parasitos diminuem, gradualmente, a expressão de genes relacionados as vias necessárias para infectar e sobreviver ao meio intracelular, como a via de resposta ao estresse oxidativo e sinalização celular. A perda de virulência pode, portanto, não estar ligada somente ao processo de metaciclogênese propriamente dito, mas sim, à diminuição de moléculas envolvidas na interação parasita-hospedeiro e sobrevivência ao meio intracelular, pressões inexistentes no meio de cultura.

A diminuição da virulência em *Leishmania* também está associada a um acentuado decréscimo na produção da proteína de superfície GP63, uma metaloproteínase importante no processo de virulência. No presente trabalho, foi detectado uma diminuição bastante significativa na expressão desse gene após 30 passagens *in vitro*, estando este gene presente na lista dos 50 genes mais diferencialmente expressos. Resultados semelhantes foram obtidos para *L. chagasi* e *L. major* (BRITTINGHAM et al., 2001; LEI; ROMINE; BEETHAM, 2010; SÁDLOVÁ et al., 2006).

A identificação das vias metabólicas mostrou que a perda de infectividade em *L. amazonensis* está relacionada a mudanças no metabolismo. Dentre as 56 vias identificadas, encontram-se vias essenciais do metabolismo, tais como: ciclo do ácido cítrico, metabolismo do piruvato, via das pentoses, fosforilação oxidativa, síntese de substâncias orgânicas e macromoléculas.

Como verificado anteriormente, a diminuição de virulência está mais ligada à uma diminuição na expressão gênica. A partir dessa observação, pode se sugerir que parasitos mantidos em cultura axênica e que gradualmente reduzem a síntese de moléculas essenciais à sobrevivência nos hospedeiros, mas dispensáveis para a

sobrevivência em meio de cultura, são favorecidos (MUKHOPADHYAY et al., 1998). Um exemplo seria a diminuição observada nas vias de metabolismo de drogas, aminoácidos, ácidos graxos e glicerolipídeos.

Os dados encontrados no presente trabalho é corroborado pelo estudo realizado anteriormente utilizando técnicas de proteômica, eletroforese bidimensional e espectrometria de massa (MAGALHÃES et al., 2014). Nesse trabalho também foi analisado a virulência de *L. amazonensis* após passagens seriadas em cultura, 1ª e 30ª passagens. Foram identificadas 56 proteínas, sendo que 37 apresentaram uma diminuição significativa em sua abundância e 19, aumentaram. Do total de proteínas que apresentaram diminuição, 13 apresentaram resultado congruente com os dados encontrados na transcriptômica e estão presentes genes sabidamente alvos terapêuticos, a S-adenosilmetionina sintetase (DRUMMELSMITH et al., 2004), candidatos vacinais e de diagnóstico, o fator de iniciação eucariótico 4A, tuzina e a proteína ácida ribossomal P2 (BERBERICH et al., 2003; LAKSHMI; WANG; MADHUBALA, 2014). Há também genes já descritos como envolvidos na infectividade, como a metalo-peptidase GP63, a triparedoxina peroxidase e a proteína de heat shock HSP70. Como já dito anteriormente, a glicoproteína de superfície GP63 é uma metaloproteínase importante no processo de virulência. A proteína participa da inativação de processos de sinalização na célula hospedeira, evitando a ativação excessiva da resposta imunológica inata (OLIVIER et al., 2012). A triparedoxina peroxidase participa da resposta ao estresse oxidativo e tem um papel crucial na sobrevivência do parasita nestes ambientes (IYER et al., 2008). A proteína HSP70 faz parte da família de proteínas heat shock que estão envolvidas na proteção contra o estresse térmico, ácido e oxidativo no interior do hospedeiro mamífero (BIFELD; CLOS, 2015).

A comparação de dados da transcriptômica e proteômica mostrou também a existência de 8 proteínas discordantes, dentre elas a dissulfeto isomerase, glutamina sintetase e proteína paraflagelar. Entretanto, esses resultados discordantes podem ser

devido aos complexos mecanismos pós-transcricionais, que envolvem a eficiência de tradução e alterações na estabilidade pós-traducional de proteínas, como já foi descrito na literatura (MARTIN et al., 2014).

CONSIDERAÇÕES FINAIS

Como já dito anteriormente, dados genômicos e recursos bioinformáticos crescem exponencialmente em tamanho e complexidade, portanto, há uma necessidade de analisar, extrair e processar essas informações para aumentar nossa compreensão sobre os processos biológicos. Esses recentes avanços têm se mostrado uma ferramenta poderosa no estudo do gênero *Leishmania* (ROGERS et al., 2011). As leishmanioses encontram-se entre as doenças tropicais mais negligenciadas, atingindo principalmente às populações mais pobres, sendo uma das dez doenças tropicais que são alvos para pesquisa e financiamento pela Organização Mundial de Saúde (WHO, 2013). O uso dos dados produzidos pelas novas técnicas de sequenciamento e recursos bioinformáticos podem contribuir na identificação de novos biomarcadores tanto para identificação de espécies quanto características específicas do parasita, como demonstrado nesse trabalho.

O desenvolvimento de uma nova ferramenta, o TipMT, foi capaz de preencher uma lacuna na prospecção de primers táxon-específicos. Pares de primers ideais devem distinguir o táxon alvo sem reação cruzada com taxa relacionados. Para atingir esse objetivo o TipMT recebe sequências genômicas e integra todo o processo de desenho dos primers desde a busca de regiões alvo apropriadas até a verificação de especificidade. A utilização de duas técnicas para predição de amplificações, uma baseada em alinhamento e outra em termodinâmica, garantiu uma maior acurácia de predição. A validação experimental mostrou a efetividade da ferramenta proposta uma vez que a maioria dos primers foi capaz de identificar o táxon para qual foi desenhado.

A interface web construída permite a utilização do TipMT pela comunidade científica, mesmo por pesquisadores com pouca habilidade computacional. Nessa versão o usuário pode enviar genomas, determinar o tipo de sequência alvo e avaliar a performance do conjunto de primers selecionados com a ajuda do guia de utilização fornecido no site.

O TipMT pode ser aplicado em um amplo espectro de estudos, como diagnóstico molecular e análises evolutivas, representando uma contribuição para vários campos científicos. Pretendemos mantê-la sempre disponível e atualizada, agregando outras funcionalidades. Futuras versões do TipMT irão: considerar genes multi-cópias, que, atualmente, são excluídos por apresentarem mais de uma amplificação no mesmo genoma, mas que podem aumentar a sensibilidade da PCR; aceitar sequências curtas, geradas por sequenciadores de nova geração, como sequências de entrada; selecionar automaticamente um conjunto de primers ideais para PCR multiplex; e implementar uma opção de busca de primers conservados.

A busca por marcadores de virulência pela técnica de RNA-seq empregada no presente trabalho também mostrou resultados satisfatórios. O cultivo *in vitro* seriado de formas promastigotas de *L. amazonensis* confirmou a diminuição da virulência com o aumento no número de passagens. O sequenciamento do transcriptoma se mostrou um método eficiente para quantificação da expressão gênica e permitiu a identificação de um número considerável de genes diferencialmente expressos em resposta à diminuição da infectividade. A análise diferencial evidenciou, de forma geral, uma diminuição da expressão dos genes do parasito. As proteínas codificadas por esses genes muito possivelmente estão associadas com os mecanismos de infecção e virulência do parasito. A expressão diferencial de alguns genes verificada no presente trabalho foi corroborada por trabalhos prévios de proteômica (MAGALHÃES et al., 2014).

Do total de genes diferencialmente expressos, 35% codificam proteínas hipotéticas, que contribuíram pouco na análise funcional, mas que apresentam uma ótima oportunidade para estudos posteriores de caracterização molecular: analisar a expressão ao longo do ciclo de vida e localização celular, gerar parasitos superexpressores ou nocautes e comparar a infectividade *in vitro* e *in vivo* dos parasitos mutantes e selvagens. As respostas transcricionais frente ao cultivo seriado desencadeia respostas que envolvem uma mudança no metabolismo, com a diminuição

da expressão de vias essenciais do metabolismo, como vias do ciclo do ácido cítrico, metabolismo do piruvato, via das pentoses, fosforilação oxidativa, síntese de substâncias orgânicas e macromoléculas, metabolismo de drogas, aminoácidos, ácidos graxos e glicerolipídeos. Considerando esses resultados e o que foi descrito anteriormente, a virulência em *Leishmania* pode estar associada com a eficiência desses três processos: a interação parasito-hospedeiro mediada por proteínas de superfície do parasito; resposta ao estresse oxidativo; e metabolismo de aminoácidos e ácidos graxos.

A fim de aprofundar os estudos sobre os processos de infecção e transcrição em *L. amazonensis*, as seguintes abordagens adicionais podem ser realizadas: validar experimentalmente por PCR em tempo real a expressão de alguns genes regulados de forma distinta nos tratamentos, melhorar a anotação e as definições estruturais dos genes para *L. amazonensis* e realizar um estudo de RNA-seq utilizando transcritos polissomais, uma vez que há evidências de que a associação à maquinaria de tradução seja o mecanismo principal de controle da expressão diferencial em alguns tripanosomatídeos (FRAGOSO et al., 2003).

BIBLIOGRAFIA

ABD-ELSALAM, K. A. Bioinformatic tools and guideline for PCR primer design. **African Journal of Biotechnology**, v. 2, n. 5, p. 91–95, 2003.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410, 1990.

ANDERS, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. **Nature protocols**, v. 8, n. 9, p. 1765–86, set. 2013.

ANDRESON, R.; MÖLS, T.; REMM, M. Predicting failure rate of PCR in large genomes. **Nucleic acids research**, v. 36, n. 11, p. e66, jun. 2008.

ARAÚJO, P. R.; TEIXEIRA, S. M. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in Trypanosoma cruzi: a review. **Memórias do Instituto Oswaldo Cruz**, v. 106, n. 3, p. 257–66, maio 2011.

ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic acids research**, v. 38, n. Database issue, p. D457–62, jan. 2010.

ATTWOOD, T. K. et al. Concepts , Historical Milestones and the Central Place of Bioinformatics in Modern Biology : A European Perspective. In: **Bioinformatics - Trends and Methodologies**. [s.l: s.n.]. p. 3–39.

BAÑULS, A.-L.; HIDE, M.; PRUGNOLLE, F. Leishmania and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. **Advances in parasitology**, v. 64, p. 1–109, jan. 2007.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society: Series B**, v. 57, p. 289–300, 1995.

BERBERICH, C. et al. Dendritic cell (DC)-based protection against an intracellular pathogen is dependent upon DC-derived IL-12 and can be induced by molecularly defined antigens. **Journal of immunology (Baltimore, Md. : 1950)**, v. 170, n. 6, p. 3171–3179, 2003.

BHARGAVA, A.; FUENTES, F. F. Mutational dynamics of microsatellites. **Molecular biotechnology**, v. 44, n. 3, p. 250–66, mar. 2010.

BHOWMICK, S. S.; SINGH, D. T.; LAUD, A. Data Management in Metaboloinformatics: Issues and Challenges. p. 392–402, 2003.

BIFELD, E.; CLOS, J. The genetics of Leishmania virulence. **Medical Microbiology and Immunology**, 2015.

BRITTINGHAM, A. et al. Regulation of GP63 mRNA stability in promastigotes of virulent and attenuated Leishmania chagasi. **Molecular and Biochemical Parasitology**, v. 112, n. 1, p. 51–59, 2001.

CAMPBELL, D. A et al. Precise identification of cleavage sites involved in the unusual processing of trypanosome ribosomal RNA. **Journal of molecular biology**, v. 196, n. 1, p. 113–124, 1987.

CAMPBELL, D. A.; THOMAS, S.; STURM, N. R. Transcription in kinetoplastid protozoa: Why be normal? **Microbes and Infection**, v. 5, n. 13, p. 1231–1240, 2003.

CANTACESSI, C. et al. Bioinformatics Meets Parasitology. **Parasite Immunology**, v. 34, n. 5, p. 265–75, maio 2011.

CANTACESSI, C. et al. The past, present, and future of Leishmania genomics and transcriptomics. **Trends in Parasitology**, v. 31, n. 3, p. 100–108, 2015.

CAO, Y. et al. Information theory-based algorithm for in silico prediction of PCR products with whole genomic sequences as templates. **BMC bioinformatics**, v. 6, p.

190, jan. 2005.

CAVALCANTI, M. C. et al. Managing structural genomic workflows using Web services. **Data & Knowledge Engineering**, v. 53, n. 1, p. 45–74, abr. 2005.

CHUANG, L.-Y.; CHENG, Y.-H.; YANG, C.-H. Specific primer design for the polymerase chain reaction. **Biotechnology letters**, v. 35, n. 10, p. 1541–9, out. 2013.

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

COELHO, V. T. S. et al. Identification of proteins in promastigote and amastigote-like Leishmania using an immunoproteomic approach. **PLoS neglected tropical diseases**, v. 6, n. 1, p. e1430, jan. 2012.

CONESA, A. et al. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674–3676, 2005.

CURCIN, V.; GHANEM, M.; GUO, Y. Web services in the life sciences. **Drug discovery today**, v. 10, n. 12, p. 865–71, 15 jun. 2005.

DA FONSECA PIRES, S. et al. Identification of virulence factors in leishmania infantum strains by a proteomic approach. **Journal of Proteome Research**, v. 13, n. 4, p. 1860–1872, 2014.

DE ALMEIDA, M. C. et al. Leishmanial infection: analysis of its first steps. A review. **Memórias do Instituto Oswaldo Cruz**, v. 98, n. 7, p. 861–70, out. 2003.

DESJEUX, P. Leishmaniasis: current situation and new perspectives. **Comparative immunology, microbiology and infectious diseases**, v. 27, n. 5, p. 305–18, set. 2004.

DILLON, L. A. L. et al. Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation. **Nucleic Acids Research**, p. gkv656, 2015.

DRUMMELSMITH, J. et al. Differential protein expression analysis of Leishmania major reveals novel roles for methionine adenosyltransferase and S-adenosylmethionine in methotrexate resistance. **Journal of Biological Chemistry**, v. 279, n. 32, p. 33273–33280, 2004.

DURAN, C. et al. Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation. v. 61, n. 0, p. 16–27, 2009.

EBERT, D. Experimental evolution of parasites. **Science (New York, N.Y.)**, v. 282, n. 5393, p. 1432–1435, 1998.

EL-SAYED, N. M. et al. Comparative genomics of trypanosomatid parasitic protozoa. **Science (New York, N.Y.)**, v. 309, n. 5733, p. 404–9, 15 jul. 2005.

EMANUELSSON, O. et al. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. **J. Mol. Biol.**, n. 300, p. 1005–1016, 2000.

FASTQC. **FastQC**. Disponível em: <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>.

FRAGOSO, S. P. et al. Cloning and characterization of a gene encoding a putative protein associated with U3 small nucleolar ribonucleoprotein in Trypanosoma cruzi. **Molecular and biochemical parasitology**, v. 126, n. 1, p. 113–117, 2003.

FREDSLUND, J. et al. A general pipeline for the development of anchor markers for comparative genomics in plants. **BMC genomics**, v. 7, p. 207, jan. 2006.

FULLER, J. C. et al. Biggest challenges in bioinformatics. **EMBO reports**, v. 14, n. 4, p. 302–4, abr. 2013.

FULTON, T. M. et al. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. **The Plant cell**, v. 14, n. 7, p.

1457–1467, 2002.

GAMBOA, D. et al. Putative markers of infective life stages in Leishmania (Viannia) braziliensis. **Parasitology**, v. 134, n. Pt 12, p. 1689–1698, 2007.

GARBER, M. et al. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nature methods**, v. 8, n. 6, p. 469–77, jun. 2011.

GOTO, H.; LINDOSO, J. A. L. Current diagnosis and treatment of cutaneous and mucocutaneous leishmaniasis. **Expert review of anti-infective therapy**, v. 8, n. 4, p. 419–33, abr. 2010.

GUERRERO, D. et al. AlignMiner: a Web-based tool for detection of divergent regions in multiple sequence alignments of conserved sequences. **Algorithms for molecular biology AMB**, v. 5, n. 1, p. 24, 2010.

GUICHOUX, E. et al. Current trends in microsatellite genotyping. **Molecular ecology resources**, v. 11, n. 4, p. 591–611, jul. 2011.

HAYDOCK, A. et al. RNA-Seq Approaches for Determining mRNA Abundance in Leishmania. In: PEACOCK, C. (Ed.). . **Parasite Genomics Protocols**. Methods in Molecular Biology. [s.l.] Springer New York, 2015. v. 1201p. 207–219.

HERWALDT, B. L. Leishmania donovani. v. 354, p. 1191–1199, 1999.

HINTON, J. C. D. et al. Benefits and pitfalls of using microarrays to monitor bacterial gene expression during infection. **Current Opinion in Microbiology**, v. 7, n. 3, p. 277–282, 2004.

IVENS, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. **Science (New York, N.Y.)**, v. 309, n. 5733, p. 436–42, 15 jul. 2005.

IYER, J. P. et al. Crucial role of cytosolic tryparedoxin peroxidase in Leishmania donovani survival, drug response and virulence. **Molecular Microbiology**, v. 68, n. 2, p. 372–391, 2008.

KAPLINSKI, L. et al. MultiPLX: automatic grouping and evaluation of PCR primers. **Bioinformatics (Oxford, England)**, v. 21, n. 8, p. 1701–2, 15 abr. 2005.

KAYE, P.; SCOTT, P. Leishmaniasis: complexity at the host-pathogen interface. **Nature reviews. Microbiology**, v. 9, n. 8, p. 604–615, 2011.

KOLEV, N. G. et al. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. **PLoS pathogens**, v. 6, n. 9, p. e1001090, jan. 2010.

LAKSHMI, B. S.; WANG, R.; MADHUBALA, R. Leishmania genome analysis and high-throughput immunological screening identifies tuzin as a novel vaccine candidate against visceral leishmaniasis. **Vaccine**, v. 32, n. 30, p. 3816–3822, 2014.

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

LAW, C. W. et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. **Genome biology**, v. 15, n. R29, p. 1–17, 2014.

LEI, S. M.; ROMINE, N. M.; BEETHAM, J. K. Population changes in Leishmania chagasi promastigote developmental stages due to serial passage. **The Journal of parasitology**, v. 96, n. 6, p. 1134–1138, 2010.

LEPROHON, P. et al. Drug resistance analysis by next generation sequencing in Leishmania. **International Journal for Parasitology: Drugs and Drug Resistance**, v. 5, n. 1, p. 26–35, 2015.

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

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler

transform. **Bioinformatics (Oxford, England)**, v. 25, n. 14, p. 1754–60, 15 jul. 2009.

LI, J. et al. Hundreds of microsatellites for genotyping Plasmodium yoelii parasites. **Molecular and Biochemical Parasitology**, v. 166, p. 153–158, 2009b.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. **Genome research**, v. 13, n. 9, p. 2178–89, set. 2003.

LI, M. et al. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. **Cladistics**, v. 24, n. 5, p. 727–745, 2008.

LIAO, Y.; SMYTH, G. K.; SHI, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. **Bioinformatics**, v. 30, n. 7, p. 923–930, 2014.

LOPES, R. DA S. et al. ProGeRF: Proteome and Genome Repeat Finder Utilizing a Fast Parallel Hash Function. v. 2015, 2015.

LUCCHI, N. W. et al. Malaria diagnostics and surveillance in the post-genomic era. **Public health genomics**, v. 16, n. 1-2, p. 37–43, jan. 2013.

MAGALHÃES, R. D. M. et al. Identification of differentially expressed proteins from Leishmania amazonensis associated with the loss of virulence of the parasites. **PLoS neglected tropical diseases**, v. 8, n. 4, p. e2764, abr. 2014.

MARCHLER-BAUER, A. et al. CDD: NCBI's conserved domain database. **Nucleic Acids Research**, v. 43, n. Database issue, p. D222–D226, 28 jan. 2015.

MARTIN, J. L. et al. Metabolic Reprogramming during Purine Stress in the Protozoan Pathogen Leishmania donovani. **PLoS Pathogens**, v. 10, n. 2, 2014.

MARTÍNEZ-CALVILLO, S. et al. Gene expression in trypanosomatid parasites. **Journal of biomedicine & biotechnology**, v. 2010, p. 525241, jan. 2010.

MITCHELL, G. F.; HANDMAN, E.; SPITHILL, T. W. Vaccination against cutaneous leishmaniasis in mice using nonpathogenic cloned promastigotes of Leishmania major and importance of route of injection. **The Australian journal of experimental biology and medical science**, v. 62 (Pt 2), p. 145–153, 1984.

MITTRA, B. et al. Iron uptake controls the generation of Leishmania infective forms through regulation of ROS levels. **The Journal of experimental medicine**, v. 210, n. 2, p. 401–16, 11 fev. 2013.

MOREIRA, D. et al. Impact of continuous axenic cultivation in Leishmania infantum virulence. **PLoS Neglected Tropical Diseases**, v. 6, n. 1, 2012.

MUKHOPADHYAY, S. et al. Reduced expression of lipophosphoglycan (LPG) and kinetoplastid membrane protein (KMP)-11 in Leishmania donovani promastigotes in axenic culture. **The Journal of parasitology**, v. 84, n. 3, p. 644–647, 1998.

NILSSON, D. et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. **PLoS pathogens**, v. 6, n. 8, p. e1001037, jan. 2010.

OLIVEIRA, E. J.; ZUCCHI, M. I.; VENCOVSKY, R. Origin, evolution and genome distribution of microsatellites. v. 307, p. 294–307, 2006.

OLIVIER, M. et al. Leishmania virulence factors: Focus on the metalloprotease GP63. **Microbes and Infection**, v. 14, n. 15, p. 1377–1389, 2012.

PAGANI, I. et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. **Nucleic Acids Research**, v. 40, n. Database issue, p. D571–D579, jan. 2012.

PEACOCK, C. S. et al. Comparative genomic analysis of three Leishmania species

that cause diverse human disease. **Nature genetics**, v. 39, n. 7, p. 839–47, jul. 2007.

POWELL, S. et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. **Nucleic acids research**, v. 40, n. Database issue, p. D284–9, jan. 2012.

QU, W. et al. MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. **Nucleic Acids Research**, v. 40, n. Web Server issue, p. gks552–, jul. 2012.

QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics (Oxford, England)**, v. 26, n. 6, p. 841–2, 15 mar. 2010.

RAFALUK, C. et al. When experimental selection for virulence leads to loss of virulence. **Trends in Parasitology**, p. 1–9, 2015.

RAPAPORT, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. **Genome biology**, v. 14, n. 9, p. R95, 10 set. 2013.

RASTROJO, A. et al. The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. **BMC genomics**, v. 14, n. 1, p. 223, jan. 2013.

REAL, F. et al. The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. **DNA research : an international journal for rapid publication of reports on genes and genomes**, v. 20, n. 6, p. 567–81, dez. 2013.

REBELLO, K. M. et al. Leishmania (Viannia) braziliensis: Influence of successive in vitro cultivation on the expression of promastigote proteinases. **Experimental Parasitology**, v. 126, n. 4, p. 570–576, 2010.

REITHINGER, R.; DUJARDIN, J.-C. Molecular diagnosis of leishmaniasis: current status and future applications. **Journal of clinical microbiology**, v. 45, n. 1, p. 21–5, jan. 2007.

REMM;, M.; ANTS, K.; METSPALU, A. Primer Design for Large-Scale Multiplex PCR and Arrayed Primer Extension (APEX). In: **PCR technology: Current innovations**. 2nd. ed. ed. [s.l: s.n.]. p. 131–140.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends in genetics : TIG**, v. 16, n. 6, p. 276–7, jun. 2000.

RICHARD, G.-F.; KERREST, A.; DUJON, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. **Microbiology and molecular biology reviews : MMBR**, v. 72, n. 4, p. 686–727, dez. 2008.

RIVAS, L. et al. Virulence and disease in leishmaniasis: what is relevant for the patient? ˜. v. 20, n. 7, 2004.

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, n. 1, p. 139–140, 2009.

ROBINSON, M. D.; OSHLACK, A. A scaling normalization method for differential expression analysis of RNA-seq data. **Genome biology**, v. 11, n. 3, p. R25, jan. 2010.

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

ROTMISTROVSKY, K.; JANG, W.; SCHULER, G. D. A web server for performing electronic PCR. **Nucleic acids research**, v. 32, n. Web Server issue, p. W108–12, 1

jul. 2004.

ROUGERON, V. et al. Extreme inbreeding in Leishmania braziliensis. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 25, p. 10224–9, 23 jun. 2009.

SÁDLOVÁ, J. et al. Virulent and attenuated lines of Leishmania major: DNA karyotypes and differences in metalloproteinase GP63. **Folia Parasitologica**, v. 53, n. 2, p. 81–90, 2006.

SAHU, S.; GUPTA, D.; DIXIT, R. Data Mining and Analysis of COS Markers in Burma Agrimony. v. 4, n. 1, p. 10–13, 2011.

SAKTHIANANDESWAREN, A.; FOOTE, S. J.; HANDMAN, E. The role of host genetics in leishmaniasis. **Trends in parasitology**, v. 25, n. 8, p. 383–91, ago. 2009.

SHARMA, P. C.; GROVER, A.; KAHL, G. Mining microsatellites in eukaryotic genomes. **Trends in biotechnology**, v. 25, n. 11, p. 490–8, nov. 2007.

SHEN, Z. et al. MPprimer: a program for reliable multiplex PCR primer design. **BMC Bioinformatics**, v. 11, n. 1, p. 143, 2010.

SIEGEL, T. N. et al. 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, n. 15, p. 4946–57, ago. 2010.

SILVA, R. DA; SACKS, D. L. Metacyclogenesis Is a Major Determinant of Leishmania Promastigote Virulence and Attenuation. **Infection and Immunity**, v. 55, n. 11, p. 2802–2806, 1987.

SOBHY, H.; COLSON, P. Gemi: PCR primers prediction from multiple alignments. **Comparative and functional genomics**, v. 2012, p. 783138, jan. 2012.

SOLÍS-CALERO, C. Identificación in silico de un grupo de secuencias ortólogas conservadas (COS) de Ipomoea batatas. **Genome**, v. 15, n. 1, p. 79–84, 2008.

SRIVIDYA, G. et al. Diagnosis of visceral leishmaniasis: developments over the last decade. **Parasitology research**, v. 110, n. 3, p. 1065–78, mar. 2012.

STAJICH, J. E. et al. The Bioperl toolkit: Perl modules for the life sciences. **Genome research**, v. 12, n. 10, p. 1611–8, out. 2002.

STERKERS, Y. et al. Novel insights into genome plasticity in Eukaryotes: Mosaic aneuploidy in Leishmania. **Molecular Microbiology**, v. 86, n. 1, p. 15–23, 2012.

STEVENS, J. R. et al. The molecular evolution of Trypanosomatidae. **Advances in parasitology**, v. 48, p. 1–56, jan. 2001.

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

UNTERGASSER, A. et al. Primer3--new capabilities and interfaces. **Nucleic acids research**, v. 40, n. 15, p. e115, ago. 2012.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, 2009a.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature reviews. Genetics**, v. 10, n. 1, p. 57–63, jan. 2009b.

WHITESIDE, M. D. et al. OrthologueDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. **Nucleic acids research**, v. 41, n. Database issue, p. D366–76, jan. 2013.

WHO. **Leishmaniasis**. Disponível em: <http://www.who.int/topics/leishmaniasis/en/>.

WONG, S. S. et al. Molecular diagnosis in clinical parasitology: When and why?

Experimental biology and medicine (Maywood, N.J.), 25 mar. 2014.

WU, F. et al. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. **Genetics**, v. 174, n. 3, p. 1407–20, nov. 2006.

YE, J. et al. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. **BMC bioinformatics**, v. 13, p. 134, jan. 2012.

YOU, F. M. et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. **BMC bioinformatics**, v. 9, p. 253, jan. 2008.

ZHANG, Z. H. et al. A comparative study of techniques for differential expression analysis on RNA-seq data. **PLoS ONE**, v. 9, n. 8, 2014.

ANEXOS

Artigos não relacionados publicados durante o doutorado

1. HUGO O VALDIVIA; JOÃO L. REIS-CUNHA; **GABRIELA F. RODRIGUES-LUIZ**; RODRIGO P. BAPTISTA; G. CHRISTIAN BALDEVIANO; ROBERT V. GERBASI; DEBORAH E. DOBSON; FRANCINE PRATLONG; PATRICK BASTIEN; ANDRÉS G. LESCANO; STEPHEN M. BEVERLEY; DANIELLA C. BARTHOLOMEU. Comparative genomic analysis of *Leishmania* (Viannia) peruviana and *Leishmania* (Viannia) braziliensis. BMC Genomics, no prelo.

2. REIS-CUNHA , J. L. ; **RODRIGUES-LUIZ G.F.;** VALDIVIA, H. O. ; BAPTISTA, R. P. ; MENDES, T. A. O. ; MORAIS, G. L. ; GUEDES, R. ; MACEDO AM ; BERN, C. ; GILMAN, R. H. ; LOPEZ, C. T. ; ANDERSSON, B. ; VASCONCELLOS, A. T. R. ; BARTHOLOMEU, DC . Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct Trypanosoma cruzi strains. BMC Genomics, 2015.

3. STOCO, PATRÍCIA HERMES WAGNER, GLAUBER TALAVERA-LOPEZ, CARLOS GERBER, ALEXANDRA ZAHA, ARNALDO THOMPSON, CLAUDIA ELIZABETH BARTHOLOMEU, DANIELLA CASTANHEIRA LÜCKEMEYER, DÉBORA DENARDIN BAHIA, DIANA LORETO, ELGION PRESTES, ELISA BEATRIZ LIMA, FÁBIO MITSUO **RODRIGUES-LUIZ, GABRIELA** VALLEJO, GUSTAVO ADOLFO FILHO, JOSÉ FRANCO DA SILVEIRA SCHENKMAN, SÉRGIO MONTEIRO, KARINA MARIANTE TYLER, KEVIN MORRIS ALMEIDA, LUIZ GONZAGA PAULA DE ORTIZ, MAURO FREITAS CHIURILLO, MIGUEL ANGEL MORAES, MILENE HÖEHR DE CUNHA, OBERDAN DE LIMA MENDONÇA-NETO, RONDON SILVA, Rosane , et al. ; Genome of the Avirulent Human-Infective Trypanosome -Trypanosoma rangeli. PLoS Neglected Tropical Diseases (Online), v. 8, p. e3176, 2014.

4. MENDES, T. A. O. ; LOBO, F. P. ; RODRIGUES, T. S. ; **RODRIGUES-LUIZ, G. F.** ; DAROCHA, W. D. ; FUJIWARA, R. T. ; TEIXEIRA, S. M. R. ; BARTHOLOMEU, D. C. . Repeat-Enriched Proteins Are Related to Host Cell Invasion and Immune Evasion in Parasitic Protozoa. Molecular Biology and Evolution, v. 30, p. 951-963, 2013.

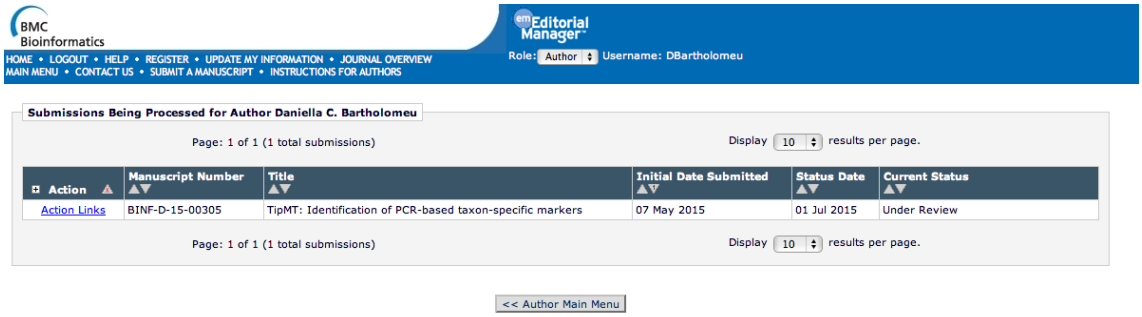
5. MOTTA, MARIA CRISTINA MACHADO SILVA, ROSANE MARTINS, ALLAN CEZAR DE AZEVEDO DE SOUZA, SILVANA SANT ANNA CATTA-PRETA, CAROLINA MOURA COSTA SILVA, Rosane KLEIN, CECILIA COIMBRA DE ALMEIDA, LUIZ GONZAGA PAULA DE LIMA CUNHA, OBERDAN CIAPINA, LUCIANE PRIOLI BROCCHI, MARCELO COLABARDINI, ANA CRISTINA DE ARAUJO LIMA, BRUNA MACHADO, CARLOS RENATO DE ALMEIDA SOARES, CÉLIA MARIA PROBST, CHRISTIAN MACAGNAN DE MENEZES, CLAUDIA BEATRIZ AFONSO THOMPSON, CLAUDIA ELIZABETH BARTHOLOMEU, DANIELLA CASTANHEIRA GRADIA, DANIELA FIORI PAVONI, DANIELA PARADA GRISARD, EDMUNDO C. FANTINATTI-GARBOGGINI, FABIANA MARCHINI, FABRICIO KLERYNTON **RODRIGUES-LUIZ, G. F.** , et al. ; Predicting the Proteins of Angomonas deanei, Strigomonas culicis and Their Respective Endosymbionts Reveals New Aspects of the Trypanosomatidae Family. Plos One, v. 8, p. e60209-e60229, 2013.

6. MENDES, TIAGO ANTÔNIO DE OLIVEIRA ; REIS CUNHA, JOÃO LUÍS ; DE ALMEIDA LOURDES, RODRIGO ; **RODRIGUES LUIZ, GABRIELA FLÁVIA** ; LEMOS, LUCAS DHOM ; DOS SANTOS, ANA RITA ROCHA ; DA CÂMARA, ANTÔNIA CLÁUDIA JÁCOME ; GALVÃO, LÚCIA MARIA DA CUNHA ; BERN, CARYN ; GILMAN, ROBERT H. ; FUJIWARA, RICARDO TOSHIO ; GAZZINELLI, RICARDO TOSTES ; BARTHOLOMEU, DANIELLA CASTANHEIRA . Identification of Strain-Specific B-cell Epitopes in Trypanosomacruzi Using Genome-Scale Epitope Prediction and High-Throughput Immunoscreening with Peptide Arrays. PLoS Neglected Tropical Diseases (Online), v. 7, p. e2524, 2013.

7. DOS SANTOS, SL ; FREITAS, LEANDRO M. ; LOBO, FP ; **RODRIGUES-LUIZ, G.F.** ; MENDES, T. A. O. ; OLIVEIRA, A. C. S. ; ANDRADE, L. O. ; CHIARI, E. ; GAZZINELLI, R. T. ; TEIXEIRA, S. M. ; FUJIWARA, R. T. ; BARTHOLOMEU, D. C. . The MASP Family of Trypanosomacruzi: Changes in Gene Expression and Antigenic Profile during the Acute Phase of Experimental Infection. PLoS Neglected Tropical Diseases (Online), v. 6, p. e1779, 2012.

8. FREITAS, LEANDRO M. ; DOS SANTOS, SARA LOPES ; **RODRIGUES-LUIZ, G. F.** ; MENDES, TIAGO A. O. ; RODRIGUES, THIAGO S. ; GAZZINELLI, RICARDO T. ; TEIXEIRA, SANTUZA M. R. ; FUJIWARA, RICARDO T. ; BARTHOLOMEU, DANIELLA C. . Genomic Analyses, Gene Expression and Antigenic Profile of the Trans-Sialidase Superfamily of Trypanosoma cruzi Reveal an Undetected Level of Complexity. Plos One, v. 6, p. e25914, 2011.

Manuscrito submetido à BMC Bioinformatics



BMC Bioinformatics

TipMT: Identification of PCR-based taxon-specific markers

--Manuscript Draft--

Manuscript Number:	
Full Title:	TipMT: Identification of PCR-based taxon-specific markers
Article Type:	Software
Section/Category:	Sequence analysis (Methods)
Funding Information:	Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (478570/2012-0) — Dr. Daniella C. Bartholomeu CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (51/2013) — Dr. Daniella C. Bartholomeu UNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE MINAS GERAIS (CBB - PPM-00219-13) — Dr. Daniella C. Bartholomeu Instituto Nacional de Ciência e Tecnologia de Vacinas (INCTV) (573547/2008-4) — Dr. Daniella C. Bartholomeu
Abstract:	Background: Molecular genetic markers are one of the most informative and widely used genome features in clinical and environmental diagnostic studies. A polymerase chain reaction (PCR)-based molecular marker is very attractive because it is suitable to high throughput automation and confers high specificity. However, the design of taxon-specific primers may be difficult and time consuming due to the need to identify appropriate genomic regions for annealing primers and to evaluate primer specificity. Results: Here, we report the development of a Tool for Identification of Primers for Multiple Taxa (TipMT), which is a web application to search and design primers for genotyping based on genomic data. The tool identifies and targets single sequence repeats (SSR) or ortholog sequences for genotyping using Multiplex PCR. This pipeline was applied to the genomes of two species of Leishmania (L. infantum and L. braziliensis) and validated by PCR using artificial genomic DNA mixtures of the two Leishmania species as templates. This experimental validation demonstrates the reliability of TipMT because amplification profiles showed discrimination of genomic DNA samples from the two species. Conclusions: The TipMT web tool allows for large-scale identification and design of taxon-specific primers and is freely available to the scientific community at http://200.131.37.155/tipMT/.
Corresponding Author:	Daniella C. Bartholomeu BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Gabriela Flávia Rodrigues-Luiz
First Author Secondary Information:	
Order of Authors:	Gabriela Flávia Rodrigues-Luiz Robson S. Lopes Mariana S. Cardoso

	Hugo O Valdivia
	Edward V. Ayala
	Thiago S. Rodrigues-Luiz
	Ricardo Toshio Fujiwara
	Daniella C. Bartholomeu
Order of Authors Secondary Information:	
Author Comments:	

TipMT: Identification of PCR-based taxon-specific markers

Gabriela F. Rodrigues-Luiz[1]; Robson S. Lopes[1,2]; Mariana S. Cardoso[1]; Hugo O. Valdivia[1]; Edward V. Ayala[1], Thiago de S. Rodrigues[3]; Ricardo T. Fujiwara[1]; Daniella C. Bartholomeu[1]*

[1]Laboratório de Imunologia e Genômica de Parasitos, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil;
[2]Departamento de Computação, Universidade Federal do Mato Grosso, Barra do Garças, Mato Grosso, Brazil;
[3]Departamento de Computação, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil;

GFRL: gab.luiz@gmail.com
RSL: robsonsilvalopes@ufmt.br
MSC: marianascardoso@yahoo.com.br
HOV: hugovalrod@gmail.com
EVA: edu_5_7@hotmail.com
TSR: tsouza@decom.cefetmg.br
RTF: fujiwara@icb.ufmg.br
DCB: daniella@icb.ufmg.br

*Corresponding author: daniella@icb.ufmg.br

ABSTRACT

Background: Molecular genetic markers are one of the most informative and widely used genome features in clinical and environmental diagnostic studies. A polymerase chain reaction (PCR)-based molecular marker is very attractive because it is suitable to high throughput automation and confers high specificity. However, the design of taxon-specific primers may be difficult and time consuming due to the need to identify appropriate genomic regions for annealing primers and to evaluate primer specificity.

Results: Here, we report the development of a **Tool for Identification of Primers for Multiple Taxa** (TipMT), which is a web application to search and design primers for genotyping based on genomic data. The tool identifies and targets single sequence repeats (SSR) or ortholog sequences for genotyping using Multiplex PCR. This pipeline was applied to the genomes of two species of *Leishmania* (*L. infantum and L. braziliensis*) and validated by PCR using artificial genomic DNA mixtures of the two *Leishmania* species as templates. This experimental validation demonstrates the reliability of TipMT because amplification profiles showed discrimination of genomic DNA samples from the two species.

Conclusions: The TipMT web tool allows for large-scale identification and design of taxon-specific primers and is freely available to the scientific community at http://200.131.37.155/tipMT/.

KEYWORDS: Molecular marker; specific primers; PCR; PCR Multiplex; web application

BACKGROUND

Polymerase chain reaction (PCR)-based typing methods are molecular diagnostic techniques widely used in biological and biomedical studies. The level of discriminatory power of PCR-based typing depends upon the molecular marker targeted. Therefore, identifying appropriate DNA target regions for primer annealing is a crucial step because these regions must be conserved within the target taxa but must vary among related taxa[1, 2].

Recent advances in next-generation sequencing technology are enabling genome sequencing projects at a significantly lower cost, even for non-model organisms. The resulting increase in the amount of genomic data available, combined with bioinformatics tools, have led to the identification of highly informative markers, such as microsatellites and ortholog genes[3].

Microsatellites or single sequence repeats (SSR) are tandem repeated stretches of short nucleotide motifs, usually ranging from 1 to 6 bp, ubiquitously distributed in the genomes of eukaryotic organisms. These regions are more prone to genetic variation and the differences in the length of individual SSR loci can be easily screened by PCR. In fact, this technique has been useful for several studies including strain typing and population genetics[4, 5].The conventional method of SSR discovery is time consuming and costly. Therefore, *in silico* mining analysis has been used to improve marker identification[6, 7].

Orthologs are homologous proteins that are related by speciation events and tend to show more functional similarity than other homologs. The identification of ortholog genes is useful in a wide range of contexts, such as inference of gene function, comparative genomics, evolutionary conservation and variability of molecular sequences[8]. Due to their importance, many tools have been developed to predict ortholog groups, including the widely used software OrthoMCL[9].

The demand is increasing for bioinformatic tools that automate analysis of genomic data generated by next-generation sequencing technology[10]. An example is the development of automated procedures to facilitate species-specific primer design for diagnostic methods[2]. Several web-based tools for facilitating primer design are available[11], but many of them are written mainly to assist in the primer design process and are not meant to search for targets and analyze primer specificity. The use of fully automated methods to search for molecular markers and the availability of genomic data for a growing number of taxa would increase the efficiency of PCR-based genotyping applications. Moreover, this strategy might save time and resources because the *in silico* evaluation of the candidate primers against the target genomic sequence are performed prior to testing them in the laboratory[12].Thus, there is a need for a tool to search for appropriate genomic target regions and then design specific primers towards the selected markers.

In this context, we have developed TipMT to meet the growing demand for easy-to-use software that facilitates the design of primers that target molecular markers to distinguish the genomic sequences of different taxa. This program only requires genomic sequences of a target species and offers, as an output, specific primers for a given taxa.

IMPLEMENTATION

Method Summary

The aim of the software pipeline is to provide a set of primer pairs flanking polymorphic sequences to identify taxa among related species, given their genomic sequences. By taking advantage of sequence data from related species, the pipeline identifies ortholog genes or SSR regions as target sequences that are likely to identify unique taxon or taxa. Because primer specificity is a key step in a PCR reaction, the pipeline identifies all potential annealing sites for the primers selected based on alignment and thermodynamics. Then, the TipMT evaluates compatibility among specific primers for designing multiplex PCR reactions. Finally, the program generates a virtual gel with the result of a simulated standard or multiplex PCR assay, where taxa can be identified by the size variation of the predicted PCR products.

The program workflow is shown in figure 1 and consists of the following steps: 1) the database with data required for the next steps is generated. In this step, the user provides the genomic sequences and defines the type of target sequences, ortholog genes or microsatellites, which are then extracted from each taxon; in this step, the user also defines the primer design constraints; 2) regions of target sequences with similarity to the other genomic sequences provided are masked; 3) candidate primers for each target sequence are designed; 4) the amplification profile of each taxon is obtained based on electronic PCR and the thermodynamic properties of the primers; 5) candidate primers are classified according to the number of taxa with amplicons; 6) the compatibility of specific primers in multiplex PCR reactions are checked; and 7) predicted PCR products for selected primers are visualized as virtual electrophoresis gels and analyzed.

User Input Data

TipMT is flexible because it accepts three different input data, genomes, predicted genes and previously defined target sequences. The user may provide sequences by uploading files in FASTA format or entering Nucleotide Accession numbers, and then, corresponding sequences will be downloaded from the NCBI RefSeq database. There are two types of genomic sequences: target taxa and cross-reaction taxa. The former sequences will be used

in the pipeline as templates for primer design and to check the specificity of the primers. The latter are sequences of species that should not cross-react during the PCR assays. Regions in the target taxa that have similarities to sequences in the cross-reaction taxa are not targeted during the primer-designing step. The number of taxa analyzed in TipMT is not limited, but the processing time is substantially increased for each taxon added.

The next step is the definition of the target sequences, and each approach offers advantages. SSRs are highly polymorphic and tend to be conserved between closely related species[5]. These features enhance the success rate in the search for taxon specific primers. Moreover, SSR targets only require genomic sequences. Ortholog sequences are less error-prone than repetitive regions during genome assembly, thus the amplification failure tends to be lower compared to SSRs. Additionally, conserved primers among closely related taxa are easily found using this approach. In case orthologs are selected as target sequences, the user should provide the sequences of the predicted genes in the genome. Also, users can have their custom sequences of interest be targets to design specific primers.

If the user chooses an SSR as a target, the pipeline will search SSR regions in the target genomes, and these regions will be the template sequences in the primer design step. On the other hand, if the user selects 'ortholog target', ortholog genes will be identified in the predicted coding regions provided by the user. Finally, the user may choose to provide target sequences instead of using the TipMT target search mechanism.

Lastly, the following primer design constraint parameters are defined: 1) mismatch and gap tolerance (both parameters are known to effect PCR specificity[12]); 2) PCR product sizes; 3) the number of primers per target (high values increase the chance to find specific primers but also increase the processing time); and 4) minimal difference (this value increases the PCR product size between taxa).

Pipeline mechanism

TipMT is a web-based application that was written in Perl language. The client side was built primarily in HTML markup language with dynamic parts written in JavaScript programming language, and Java applets are used to input data and process files between the user and the application. The server side runs on PHP, and MySQL database is used to store the input parameters and results.

The pipeline uses Primer3 core [13] as the primer design engine and is built around other public domain programs, such as BEDtools [14], BioPerl [15], BLAST [16], EMBOSS [17], e-PCR [18], MFEprimer [19], MultiPLX [20], OrthoMCL [9] and ProGeRF [21]. These programs

are used in the processing flow from raw sequences to the list of primers in the following list of procedures:

1. Target search. One way to improve sensitivity in PCR is to find the most appropriate template region for primer design. Ortholog sequences are identified in the predicted coding sequences provided by the user using OrthoMCL with default parameters values. ProGeRF searches for SSR regions, with a length of 1 to 6 bases, and without degeneration or gaps in the genomic sequences.

2. Mask similar regions. Conservation of the flanking regions of the target sequences is essential for a high quality PCR assay because a high number of primer annealing sites can cause failure of the PCR assay [22]. Similar regions between target sequences and genomes are identified using a MEGABLAST search with default parameters against all genomic sequences. MEGABLAST was chosen due to its speed and its ability to handle slight differences in genomic sequences. Next, regions with more than 95 percent of identity are masked with lowercase nucleotides (initially, all sequences are set as uppercase in the database).

3. Primer design. Candidate primers are generated for each of the DNA template sequences using primer3 2.3.5 with default parameters values and an option that rejects the primer candidates with lowercase letters in the first 3' end position. Because the high number of annealing sites influences primer specificity, this procedure decreases the rate of low-success specific PCR primers, avoiding primer design in regions with similarity between species[23].

4. Specificity check. All candidate pairs of primers generated are evaluated for specificity using the alignment based e-PCR algorithm and by thermodynamic properties using MFEprimer software, with mismatch and gap tolerance chosen by the user. If both tools predict PCR products with same length, the pair of primers is selected for the next step.

5. Primer classification. The program recovers all potentially useful pairs of primers for the differentiation of taxa. If a pair of primers has only one amplification product in the target genome sequence, it is defined as 'specific'. If a pair of primers has one amplification product in at least one other genome, it is defined as 'multiple'. If it amplifies the same size products in all genomes, it is named as 'conserved', but if the PCR products have different sizes in all genomes, then it is a 'single' primer. ‘Single’ pairs of primers are capable of distinguishing all taxa in a simple PCR reaction because each has different sizes of amplification products in each genome.

6. Compatibility check. Specific primers are clustered into compatible groups for multiplexing PCR using MultiPLX, which tests all primer pairs for interactions, including dimer formation and differences in their melting temperatures.

7. Gel visualization. After a set of primers is chosen, the relative electrophoretic migration distances are calculated based on the expected length of the amplification products. Then, a virtual electrophoresis gel is generated showing the expected amplification profile as a result of a standard PCR assay using a mixture of target genomes as the template. Another virtual gel is generated for the amplification profile of the multiplex PCR reaction, to check for interactions among primers that generate undesired alternative products.

Pipeline **validation**

Promastigote forms of *L. braziliensis* (M2904) and *L. infantum* (PP75) were cultured in Schneider's insect medium (Sigma) supplemented with 10% inactivated fetal bovine serum (Life Technologies) and 1% (v/v) penicillin/streptomycin and maintained at 24 ºC. Genomic DNA was extracted from 10^8 promastigotes in logarithmic growth phase using a Wizard Genomic DNA Purification Kit (Promega), resuspended in DNase-free water and quantified using a NanoDrop Spectrophotometer ND-1000. The genomic DNA obtained from both species of Leishmania was used as a template in PCR amplification reactions with selected taxon-specific primers designed by TipMT (Supplementary Table 1). Each PCR used 100 ng of DNA template, 1x Green GoTaq Reaction Buffer (Promega), 200 µM dNTPs mix, 10 pmol of each forward and reverse primers, and 1.25 U of Taq DNA polymerase (Phoneutria). The samples were incubated at 94 ºC for 5 min and submitted to 30 cycles of 94 ºC (30 seconds), 60 ºC (15 seconds) and 72 ºC (10 seconds), followed by a final extension of 72 ºC for 7 min. The PCR products were fractionated by electrophoresis in 2.5% agarose gels in TAE 1x buffer (4.8 g/L Tris-base, 1.14 mL acetic acid, 2 mL 0.5 M EDTA, pH 8.0) with 0.5 µg/mL ethidium bromide.

RESULTS AND DISCUSSIONS

TipMT Output

Primers set. Selected pairs of primers are classified in one of the four categories: specific, multiple, single and conserved. After the user chooses the categories, information regarding the primer characteristics is reported, such as primer sequence, melting temperature range (°C) and PCR product length (bp) (Figure 2). Additionally, users can choose a set of primers or save the primer information in a text file or visualize a virtual gel electrophoresis. The text file shows the following amplicon properties: 1) name; 2) forward primer sequence; 3) reverse primer sequence; and 4) primer melting temperature and amplicon in each target genome. A list of the compatible pairs of primers that are optimal for multiplex PCR is also available to download.

Virtual gel. After selecting a set of pair of primers, the user may choose to generate a virtual electrophoresis gel, a visual output with a simulated result of a conventional PCR e-GEL or multiplex PCR e-MPX (Figure 3). In both cases, TipMT takes the set of PCR products and calculates their respective migration distances based on the length of the amplicons.

In the e-GEL function, the visual output is a simulated conventional PCR assay, where each lane is a reaction with one selected primer and a mixture of target genomes as a template. However, the e-MPX function generates another virtual gel with the amplification profile of the multiplex PCR assay, where each lane is a mixture of all selected primers and one target genome or all target genomes as the template.

Multiplex PCR is a variant of PCR, which simultaneously amplifies many loci of interest in one single reaction by using more than one pair of primers. Setting up a multiplex PCR with consistent quality is not trivial; therefore, TipMT generates a file with groups of specific primers that are compatible in a multiplex assay, based on primer–primer interactions and differences in the melting temperatures.

Experimental Validation

Leishmania is a genus of flagellate protozoan that cause a broad spectrum of diseases, ranging from self-limiting localized cutaneous lesions to visceral leishmaniasis. More than 20 species of *Leishmania* cause infection in humans[24]. Despite the wide taxonomic complexity of this genus, the gold standard for diagnosing *Leishmania* infections, parasitological assays, only discriminates genus, not species. The reference method for species identification is multilocus enzyme electrophoresis (MLEE). However, this method has several limitations, including the relatively small number of characterized loci and the requirement of a parasite culture that potentially biases the results [25]. Therefore, the development of new molecular diagnostic methods could allow for a rapid and accurate diagnosis of *Leishmania* species in infection. Moreover, robust molecular markers might contribute to the characterization of parasite-specific features, such as virulence or drug resistance [26].

By applying this tool to the genomes of two species of *Leishmania*, we generated sets of primers for genotyping using Multiplex PCR. The PCR gel electrophoresis obtained experimentally using this protocol is shown in Figure 3. The result of the real experiment was nearly the same as that predicted by the virtual electrophoresis gel analysis (Figure 3).

Comparison to other primer design applications

We compared TipMT to other similar web-based primer design tools, such as BatchPrimer3 [27], Mprimer[28], and PrimerBlast [29]. A common feature to all tools is the use of Primer3 [13] as the primer design engine. A comparison of the main features of these tools is shown in Table 1.

Mprimer is a web-based tool that designs specific multiplex PCR primers, uses thermodynamic theories to estimate the stability of the primers, checks specificity against a limited and pre-defined list of genome databases provided by the tool and has a function for predicting a group of compatible multiplex primers. BatchPrimer3 allows users to design several types of primer, including generic primers, hybridization oligos, primers for SSR regions, SNP genotyping primers and DNA sequencing primers. Primer-BLAST provides a specificity check using BLAST to avoid nonspecific amplifications.

TipMT offers a combination of features that are not present in other available web applications. TipMT receives multiple sequences as input and identifies target regions automatically. Moreover, primer specificity is tested by both alignment- and thermodynamic-based properties. Furthermore, TipMT provides functions to generate a virtual electrophoresis gel for conventional or multiplex PCR assays. This output gives users a visual result before performing a real PCR reaction.

CONCLUSIONS

The emergence of large-scale DNA sequencing projects in recent years has produced large amounts of data, opening many opportunities for genomic analyses. Here, we focus our attention on identifying molecular markers and designing efficient primers for taxa differentiation. The ideal pair of primers should be capable of distinguishing the target taxon and should not cross-react with other closely related species. Toward this aim, TipMT receives genomic sequences as input and integrates the process of primer design, from the search for target sequences to the evaluation of primer specificity. As an output, the web-application generates a plain text file with general information on the pairs of primers, based on taxa-specificity. The output also includes an image showing the result of a simulated PCR assay with selected pairs of primers. Finally, experimental validation shows the effectiveness of the proposed tool in finding a taxon-specific pair of primers.

The pairs of primers generated by TipMT are suitable for use in conventional or multiplex PCR assays, as determined by the parameter settings during the primer design stage. Furthermore, primer design principles for conventional PCR and real-time quantitative PCR are quite similar. Thus, our tool could be used for designing primers for both methodologies by adjusting some parameters, such as PCR product size.

Future versions of TipMT will consider multi-copy genes as targets to improve PCR sensitivity and will also receive raw sequencing reads as input. Additional improvements may also be performed, for example, the automatic selection of ideal primers for multiplex PCR.

The application is platform independent, freely available and has a simple and user-friendly interface that allows for designing primers in a high-throughput manner, even for novice users. TipMT can be applied to a broad spectrum of research topics including both molecular diagnostic and evolutionary studies.

AVAILABILITY AND REQUIREMENTS

Project name: TipMT, Tool for Identification of Primers for Multiple Taxa

Project home page: http://200.131.37.155/tipMT/

Operating system(s): Platform independent

Programming language: PHP, JavaScript, PERL

Other requirements: Web browser (supported browsers: Firefox, Chrome)

Any restrictions to use by non-academics: no license needed

ABBREVIATIONS

PHP: PHP Hypertext Preprocessor; HTML: Hypertext Markup Language; BLAST: Basic Local Alignment Search Tool; RefSeq: Reference Sequence.

COMPETING INTERESTS

The authors declare no competing interests.

AUTHORS' CONTRIBUTIONS

GFRL carried out the code development, implementation, and drafted the manuscript. MCS, HOV and EVA validated SSR and Orthologs primers. RSL implemented the TipMT website construction. TSR and RTF provided scientific advice and the resources to develop the software. DCB conceived of the study, participated in its design and coordination, and helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This study was funded by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Instituto Nacional de Ciência e Tecnologia de Vacinas (INCTV)—Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). DCB, RTF are CNPq research

fellows. GFRL, HOV and EVA received scholarships from CAPES and MSC received a scholarship from CNPq.

REFERENCES

1. Wong SS, Fung KS, Chau S, Poon RW, Wong SC, Yuen K-Y: **Molecular diagnosis in clinical parasitology: When and why?** *Exp Biol Med (Maywood)* 2014.

2. Lucchi NW, Oberstaller J, Kissinger JC, Udhayakumar V: **Malaria diagnostics and surveillance in the post-genomic era.** *Public Health Genomics* 2013, **16**:37–43.

3. Li J, Zhang Y, Liu S, Hong L, Sullivan M, McCutchan TF, Carlton JM, Su XZ: **Hundreds of microsatellites for genotyping Plasmodium yoelii parasites**. *Mol Biochem Parasitol* 2009, **166**:153–158.

4. Ellegren H: **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5**:435–45.

5. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ: **Current trends in microsatellite genotyping.** *Mol Ecol Resour* 2011, **11**:591–611.

6. Sharma PC, Grover A, Kahl G: **Mining microsatellites in eukaryotic genomes.** *Trends Biotechnol* 2007, **25**:490–8.

7. Duran C, Appleby N, Edwards D, Batley J: **Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation**. 2009, **61**:16–27.

8. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P: **eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges.** *Nucleic Acids Res* 2012, **40**(Database issue):D284–9.

9. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178–89.

10. Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi ML: **Bioinformatics for next generation sequencing data.** *Genes (Basel)* 2010, **1**:294–307.

11. Abd-elsalam K a: **Bioinformatic tools and guideline for PCR primer design**. *African J Biotechnol* 2003, **2**:91–95.

12. Cao Y, Wang L, Xu K, Kou C, Zhang Y, Wei G, He J, Wang Y, Zhao L: **Information theory-based algorithm for in silico prediction of PCR products with whole genomic sequences as templates.** *BMC Bioinformatics* 2005, **6**:190.

13. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3--new capabilities and interfaces.** *Nucleic Acids Res* 2012, **40**:e115.

14. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841–2.

15. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz S a, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12**:1611–8.

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.

17. Rice P, Longden I, Bleasby a: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**:276–7.

18. Rotmistrovsky K, Jang W, Schuler GD: **A web server for performing electronic PCR.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W108–12.

19. Qu W, Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, Yang Y, Zhang C: **MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity.** *Nucleic Acids Res* 2012, **40**(Web Server issue):gks552–.

20. Kaplinski L, Andreson R, Puurand T, Remm M: **MultiPLX: automatic grouping and evaluation of PCR primers.** *Bioinformatics* 2005, **21**:1701–2.

21. Lopes R da S, Moraes JWL, Rodrigues TDS, Bartholomeu DC: **ProGeRF: Proteome and Genome Repeat Finder Utilizing a Fast Parallel Hash Function**. 2015, **2015**.

22. Andreson R, Möls T, Remm M: **Predicting failure rate of PCR in large genomes.** *Nucleic Acids Res* 2008, **36**:e66.

23. Remm; M, Ants K, Metspalu A: **Primer Design for Large-Scale Multiplex PCR and Arrayed Primer Extension (APEX)**. In *PCR technology: Current innovations*. 2nd. ed.; 2004:131–140.

24. **Leishmaniasis** [http://www.who.int/topics/leishmaniasis/en/]

25. Hernández C, Ramírez JD: **Molecular Diagnosis of Vector-Borne Parasitic Diseases**. 2013, **2**:1–10.

26. Reithinger R, Dujardin J-C: **Molecular diagnosis of leishmaniasis: current status and future applications.** *J Clin Microbiol* 2007, **45**:21–5.

27. You FM, Huo N, Gu YQ, Luo M-C, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD: **BatchPrimer3: a high throughput web application for PCR and sequencing primer design.** *BMC Bioinformatics* 2008, **9**:253.

28. Shen Z, Qu W, Wang W, Lu Y, Wu Y, Li Z, Hang X, Wang X, Zhao D, Zhang C: **MPprimer: a program for reliable multiplex PCR primer design**. *BMC Bioinformatics* 2010, **11**:143.

29. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL: **Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.** *BMC Bioinformatics* 2012, **13**:134.

FIGURE LEGENDS

Figure 1: Flowchart for TipMT analysis.

Figure 2: Primer report showing information about the primer characteristics.

Figure 3: Real and virtual gel electrophoresis for Orthologs (A) and SSR (B) primers, for a real and simulated (e-MPX) result of a multiplex PCR assay. Each lane corresponds to the combination of genomic DNA of *Leishmania* species identified at the top and a mixture of the two orthologs or SSR pair of primers. Lb: *L. braziliensis*; Li: *L. infantum*; gDNA: genomic DNA; bp: base pair;

TABLES

Table 1: Comparison of TipMT to other primer design applications

Tool	Sequence input	Target type	Specificity check method	Output	Multiplex check
BatchPrimer3	multiple sequences (up to 500)	SSR, SNP	none	Primers information	no
Primer-BLAST	single sequence	none	alignment	Primer alignment in template sequence	no
Mprimer	multiple sequences	none	thermodynamic	Primers information	yes
TipMD	multiple sequences	SSR, orthologs	alignment and thermodynamic	Primers information and virtual gel	yes

ADDITIONAL FILES

Supplementary Table 1: Primer list with selected taxon-specific primers designed by TipMT

Figure1
Click here to download Figure: Figure1.tiff

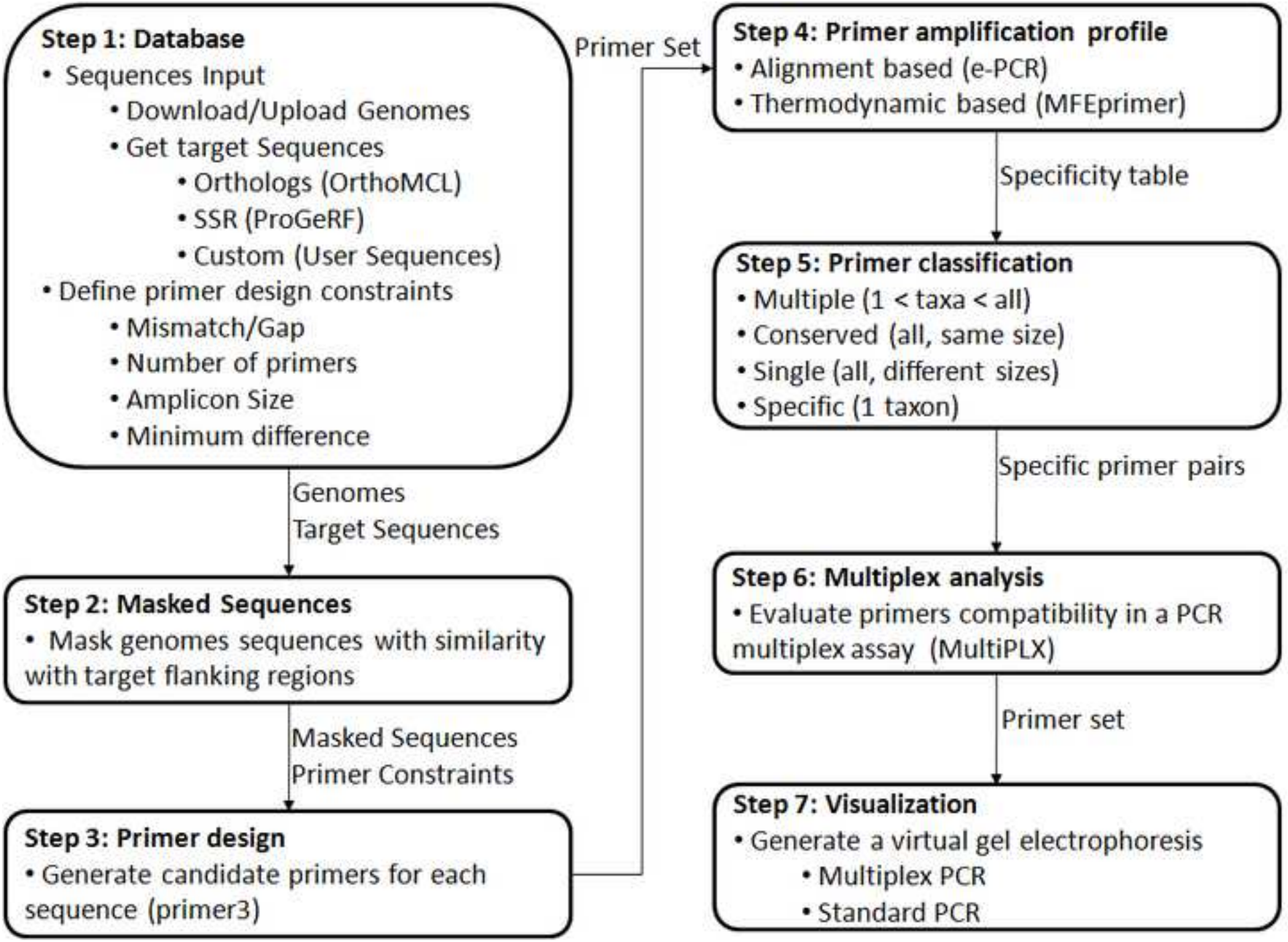
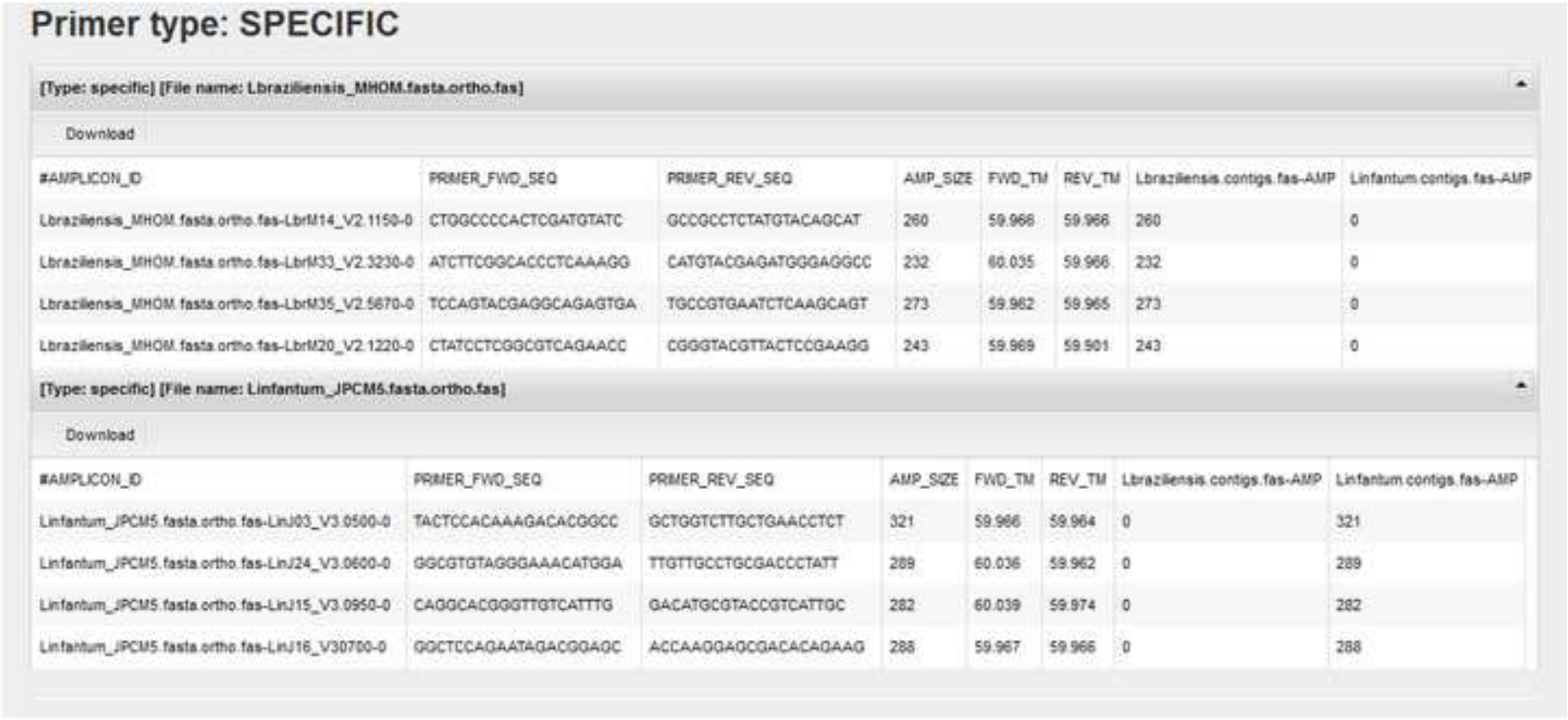


Figure2
Click here to download Figure: Figure2.tiff



Primer type: SPECIFIC

[Type: specific] [File name: Lbraziliensis_MHOM.fasta.ortho.fas]

Download

#AMPLICON_ID	PRIMER_FWD_SEQ	PRIMER_REV_SEQ	AMP_SIZE	FWD_TM	REV_TM	Lbraziliensis.contigs.fas-AMP	Linfantum.contigs.fas-AMP
Lbraziliensis_MHOM.fasta.ortho.fas-LbrM14_V2.1150-0	CTGGCCCCACTCGATGTATC	GCCGCCTCTATGTACAGCAT	260	59.968	59.966	260	0
Lbraziliensis_MHOM.fasta.ortho.fas-LbrM33_V2.3230-0	ATCTTCGGCACCCTCAAAGG	CATGTACGAGATGGGAGGCC	232	60.035	59.966	232	0
Lbraziliensis_MHOM.fasta.ortho.fas-LbrM35_V2.5670-0	TCCAGTACGAGGCAGAGTGA	TGCCGTGAATCTCAAGCAGT	273	59.962	59.965	273	0
Lbraziliensis_MHOM.fasta.ortho.fas-LbrM20_V2.1220-0	CTATCCTCGGCGTCAGAACC	CGGGTACGTTACTCCGAAGG	243	59.969	59.901	243	0

[Type: specific] [File name: Linfantum_JPCM5.fasta.ortho.fas]

Download

#AMPLICON_ID	PRIMER_FWD_SEQ	PRIMER_REV_SEQ	AMP_SIZE	FWD_TM	REV_TM	Lbraziliensis.contigs.fas-AMP	Linfantum.contigs.fas-AMP
Linfantum_JPCM5.fasta.ortho.fas-LinJ03_V3.0500-0	TACTCCACAAAGACACGGCC	GCTGGTCTTGCTGAACCTCT	321	59.966	59.964	0	321
Linfantum_JPCM5.fasta.ortho.fas-LinJ24_V3.0600-0	GGCGTGTAGGGAAACATGGA	TTGTTGCCTGCGACCCTATT	289	60.036	59.962	0	289
Linfantum_JPCM5.fasta.ortho.fas-LinJ15_V3.0950-0	CAGGCACGGGTTGTCATTTG	GACATGCGTACCGTCATTGC	282	60.039	59.974	0	282
Linfantum_JPCM5.fasta.ortho.fas-LinJ16_V30700-0	GGCTCCAGAATAGACGGAGC	ACCAAGGAGCGACACAGAAG	288	59.967	59.966	0	288

Figure3

Click here to download Figure: Figure3.tiff