

UNIVERSIDADE FEDERAL DO PARANÁ

YASMIN CARLA RIBEIRO

ANÁLISE EVOLUTIVA DOS RETROTRANSPÓSONS *VIPER* E *TATE* NOS
GENOMAS DE TRIPANOSOMATÍDEOS

CURITIBA

2018

YASMIN CARLA RIBEIRO

ANÁLISE EVOLUTIVA DOS RETROTRANSPOSONS *VIPER* E *TATE* NOS
GENOMAS DE TRIPANOSOMATÍDEOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Biologia Celular e Molecular, no Curso de Pós-Graduação em Biologia Celular e Molecular, Setor de Ciências Biológicas, da Universidade Federal do Paraná.

Orientador: Dr. Marco Aurélio Krieger
Co-orientadora: Dra. Adriana Ludwig

CURITIBA

2018

Universidade Federal do Paraná. Sistema de Bibliotecas.
Biblioteca de Ciências Biológicas
(Telma Terezinha Stresser de Assis –CRB/9-944)

Ribeiro, Yasmin Carla

Análise evolutiva dos retrotransposons *VIPER* e *TATE* nos genomas de
Tripanosomatídeos. / Yasmin Carla Ribeiro. – Curitiba, 2018.
128 p. : il. ; 30cm.

Orientador: Marco Aurélio Krieger

Co-orientadora: Adriana Ludwig

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de
Ciências Biológicas. Programa de Pós-Graduação em Biologia Celular e
Molecular.

1. Protozoário. 2. Evolução. 3. Retroelementos. I. Título II. Krieger,
Marco Aurélio. III. Ludwig, Adriana. IV. Universidade Federal do Paraná.
Setor de Ciências Biológicas. Programa de Pós-Graduação em Biologia
Celular e Molecular.

CDD (20. ed.) 593.1



**PROGRAMA DE PÓS-GRADUAÇÃO EM
BIOLOGIA CELULAR E MOLECULAR**

Departamento de Biologia Celular
Setor de Ciências Biológicas
Universidade Federal do Paraná

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOLOGIA CELULAR E MOLECULAR da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **YASMIN CARLA RIBEIRO** intitulada “**Análise evolutiva dos retrotransposons VIPER e TATE nos genomas de Tripanosomatídeos**”, após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVADA no rito de defesa.

A outorga do título de Mestre está sujeita à homologação pelo Colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 28 de março de 2018.

Prof. Dr. Marco Krieger
Orientador e presidente da banca
ICC-Fiocruz/PR (avaliador interno)

Prof. Dr. Daniel Pacheco Bruschi
UFPR (avaliador externo)

Prof. Dr. Leonardo Magalhães
UFPR (avaliador externo)

Dedico as conquistas deste trabalho aos meus avós, os quais sempre me motivaram a ser a mulher forte e independente que hoje consigo reconhecer em mim.

AGRADECIMENTOS

Agradeço, primeiramente, aos meus orientadores que me auxiliaram na execução do trabalho. Ao Dr. Marco Krieger pela confiança, acolhida e apoio e, em especial a Dra. Adriana Ludwig que não apenas compartilhou seu conhecimento comigo, mas também foi o meu principal apoio no mestrado, me incentivando, motivando, inspirando e me ensinando a amar esse mundo maravilhoso dos transposons.

Meu segundo agradecimento não poderia deixar de ser à minha família. Agradeço ao apoio deles, a preocupação com a minha saúde emocional e o incentivo para seguir nessa carreira que me completa.

Agradeço ao acolhimento, apoio e motivação do grupo de pesquisa, que intitulamos de TEs&Tryps (Adri, Danila, Cyndia, Ana Luisa). Também a todos os colegas do laboratório de genômica funcional e aos bons amigos da pós.

À Dra. Lizandra Jaqueline Robe, da FURG de Rio Grande, pela forte colaboração estabelecida, nos auxiliando na resolução dos complexos alinhamentos e filogenias dos retroelementos e da família de tirosina recombinases, compartilhando seu conhecimento conosco e estando sempre disponível para nos auxiliar.

Ao programa de pós-graduação em Biologia Celular e Molecular e a UFPR pelo suporte dado aos alunos. Ao ICC-Fiocruz/PR pela disponibilização da estrutura. A todos os professores da pós-graduação, em especial àqueles que eu tive o prazer de frequentar suas disciplinas, pela inspiração e influência que exerceram para mim.

Às agências de fomento CAPES e CNPq pelo apoio financeiro que possibilitou a realização deste trabalho.

Por fim, o mestrado foi acompanhado de momentos difíceis e conturbados. Por muitas vezes, a vontade de desistir e o pensamento de que não era merecedora quase me derrubaram, mas um dos grandes ensinamentos que a ciência me trouxe foi que, assim como os sistemas biológicos, não conseguimos chegar a lugar nenhum sozinhos, necessitamos de uma rede que auxilia a engrenar o melhor de nós. Por isso, um dos meus agradecimentos mais especiais vai para todos os meus amigos. Amigos que eu trouxe da vida (Camylla, Jéssica), da faculdade (Melina, Marina, Gabriel) do laboratório (Adri, Ueriton, Cyndia, Denise, Beth, Juliana, Bruno, Danila, Matheus, Claudemir, Fernanda, Yiris, Aline). Tê-los comigo, ao meu lado,

vivenciando e ajudando meus momentos de fraqueza, foi o alívio, o remédio que eu precisava para me manter firme e prosseguir nos meus sonhos. Divido a conquista desse mestrado com vocês, que cansaram de que ouvir desabafar durante os períodos de crises, durante momentos que eu não sabia mais o que fazer. Deixo aqui o meu mais sincero obrigado.

“Todos vamos morrer, e é isso que nos torna pessoas de sorte. Existe uma imensa maioria que nunca vai morrer, pois nunca vai nascer. As pessoas que em potencial poderiam estar no meu lugar, mas que nunca verão a luz do dia, superam em número os grãos de areia do Saara. Entre esses fantasmas com certeza há poetas superiores a Keats. E cientistas superiores a Newton. Disso sabemos porque o conjunto de pessoas possíveis que nosso DNA permite supera maciçamente o número de pessoas que existem. A despeito dessa probabilidade chocante, somos você e eu, em toda nossa banalidade, que aqui estamos”

Richard Dawkins

RESUMO

Tripanosomatídeos são protozoários flagelados e algumas espécies são conhecidos patógenos humanos. Pertencem a essa família o *Trypanosoma cruzi*, agente etiológico da doença de Chagas; *T. brucei* agente da doença do sono; e *Leishmania* sp., que desencadeiam os quadros de leishmaniose. Pelo sequenciamento dos primeiros genomas de tripanosomatídeos, foi observado que cerca de 2 a 5% do genoma haploide desses protozoários é composto por elementos de transposição, sendo encontrado apenas retrotransposons que se transpõem através de intermediário de RNA. Elementos de transposição são componentes móveis presentes em diversos genomas, sendo capazes de se multiplicar e se integrar no genoma auxiliando na variabilidade genética o que contribui para a evolução e adaptação das espécies. Poucos elementos foram descritos em tripanosomatídeos, podendo ser agrupados em três famílias: CRE-like, Ingi-like e DIRS. Dentro dos elementos classificados na ordem DIRS, *T. cruzi* possui um elemento denominado *VIPER*, enquanto *L. braziliensis* possui um elemento *TATE*. É de conhecimento que os elementos DIRS codificam uma provável Gag-like, uma transcriptase reversa para realizar a transcrição reversa da sequência de RNA a ser transposta, e uma tirosina recombinase (principal característica dos elementos DIRS) que está associada com a integração do elemento no sítio-alvo. O estudo escasso e a ausência de uma caracterização clara desses elementos, levantam dúvidas sobre a conservação, distribuição e evolução desses elementos nos genomas dos tripanosomatídeos. Tendo isso em vista, o objetivo deste trabalho foi estudar os elementos *VIPER* e *TATE* em um contexto evolutivo, tanto em associação com a evolução dos organismos hospedeiros quanto na história evolutiva dos retrotransposons. 40 espécies de tripanosomatídeos e *Bodo saltans* foram utilizadas para realizar as análises, que se concentraram em análises *in silico* das cópias dos elementos presentes nos genomas, reconstruções filogenéticas, análise de divergência e de transferência horizontal. Em suma, os elementos apresentaram uma distribuição descontínua ao longo da família dos tripanosomatídeos e cópias completas, possivelmente codificadoras, foram encontradas para os dois elementos em diferentes espécies. A coexistência dos dois elementos completos no genoma de *B. saltans* nos auxiliou a propor um cenário evolutivo onde esses elementos estariam presentes no ancestral dos tripanosomatídeos, tendo sido mantidos por mais de 400 milhões de anos e/ou podendo ter sofrido eventos de reativação ao longo da evolução. Propomos também um outro cenário evolutivo, sobre a origem dos elementos da ordem DIRS no contexto da história evolutiva dos retrotransposons. Concluimos que a evolução desses elementos ocorre de forma modular, tendo origens independentes. Conseguir elucidar essas questões faz com que esse trabalho represente uma importante contribuição na área dos retrotransposons e um primeiro passo para análises funcionais que serão necessárias para melhor compreensão do funcionamento e modo de mobilização desses elementos.

Palavras-chave: Retrotransposons, Tripanosomatídeos, DIRS, Evolução.

ABSTRACT

Trypanosomatids are flagellated protozoa and some species are human pathogens, like as *Trypanosoma cruzi*, the etiological agent of Chagas disease; *T. brucei*, the etiological agent of sleeping sickness, and *Leishmania* sp that unleash a several cases of leishmaniasis. After sequencing of first trypanosomatids genomes it was observed that about 2 – 5% of the haploid genome is composed by transposable elements and only retrotransposons were identified. Transposable elements are mobile sequences that are present in essentially all organisms. They are able to self-replicate and integrate in new positions at the genome, contributing to the genome adaptation and evolution. Few elements were described in trypanosomatids, and we can separate them into three group: CRE-like, Ingi-like and DIRS. Inside the DIRS group, there are two elements, *VIPER* from *T. cruzi* and *TATE* from *L. braziliensis*. We know that the DIRS group elements encode a Gag-like protein, a reverse transcriptase to perform the reverse transcription of RNA molecule and a tyrosine recombinase (major characteristic of DIRS group) to do the integration of element in the target site. The scarce study and the absence of the characterization of these elements, raise some doubts about their conservation, distribution and evolution in trypanosomatids genomes. Thus, the purpose of this work was study *VIPER* and *TATE* in an evolutionary view, both in the context of host evolution and in the context of retroelements evolution. 40 trypanosomatids species and *Bodo saltans* were used for the analysis that concentrated in *in silico* searches and analysis of copies, phylogenetic reconstructions, divergence and horizontal transfer analysis. In our analysis, the elements *VIPER* and *TATE* show a discontinuous distribution over the family Trypanosomatidae and full copies, potentially encoding all genes, were found from both elements in some different species. The *VIPER* and *TATE* coexistence in *B. saltans* suggest an evolutionary scenario where both elements were present in ancestor of trypanosomatids, surviving for more than 400 million years and/or suffering possible reactivation events along the trypanosomatids evolution. We also proposed the evolutionary scenario about the origin of order DIRS elements inside retrotranspon evolution history. We concluded that the evolution of these elements occurs in a modular form, presenting independent origins. By elucidating these issues, this work has an important contribution to the mobile elements science and it is a first step to further functional analysis about the function of the genes and the mobilization mechanism of these elements.

Key-words: Retrotransposons, Trypanosomatids, DIRS, Evolution.

LISTA DE FIGURAS

FIGURA 1: VISÃO GERAL DA FILOGENIA DOS CINETOPLASTÍDEOS	17
FIGURA 2: HISTÓRIA EVOLUTIVA DOS TRIPANOSOMATÍDEOS	18
FIGURA 3: ESTÁGIOS MORFOLÓGICOS APRESENTADOS PELOS TRIPANOSOMATÍDEOS.....	20
FIGURA 4: ESQUEMA DO PROCESSO DE TRANSCRIÇÃO GÊNICA EM TRIPANOSOMATÍDEOS.....	22
FIGURA 5: MECANISMO BÁSICO DE TRANSPOSIÇÃO DOS ELEMENTOS DE CLASSE II (TRANSPOSONS)	25
FIGURA 6: ESQUEMA DO PROCESSO DE MOBILIZAÇÃO DOS LTR- RETROTRANSPOSONS	27
FIGURA 7: ESTRUTURA DOS DIFERENTES GRUPOS DE RETROTRANSPOSONS CODIFICADORES DE YR.....	30
FIGURA 8: ESQUEMA DA ORIGEM DOS RETROELEMENTOS DE XIONG E EICKBUSH (1990).....	33
FIGURA 9: ESTRUTURA TEÓRICA INICIALMENTE DEFINIDA PARA O ELEMENTO <i>TATE</i>	50
FIGURA 10: ESQUEMA DAS RELAÇÕES FILOGENÉTICAS DAS ESPÉCIES DE TRIPANOSOMATÍDEOS USADAS NESSE TRABALHO E A DISTRIBUIÇÃO DOS ELEMENTOS <i>VIPER</i> E <i>TATE</i>	52
FIGURA 11: ESQUEMA DA ESTRUTURA DOS ELEMENTOS <i>VIPER</i> E <i>TATE</i> COMPLETOS DE TRIPANOSOMATÍDEOS	58
FIGURA 12: REPRESENTAÇÃO ESQUEMÁTICA DE <i>VIPER</i> COMPLETO E <i>SIRE</i> DE <i>T. CRUZI</i>	61
FIGURA 13: MOTIVO CH-BOX ENCONTRADO EM SEQUÊNCIAS DE GAG-LIKE DE <i>VIPER</i>	63
FIGURA 14: REPRESENTAÇÃO ESQUEMÁTICA DO <i>TATE</i> COMPLETO DE <i>L.</i> <i>BRAZILIENSIS</i> E DA CÓPIA CONSENSO DISPONÍVEL NO REPBASE	69
FIGURA 15: FILOGENIA DE <i>VIPER</i> E <i>TATE</i> DE CINETOPLASTÍDEOS COM BASE NA SEQUÊNCIA DA TRANSCRIPTASE REVERSA	71
FIGURA 16: COMPARAÇÃO DA IDENTIDADE DAS SEQUÊNCIAS DOS ELEMENTOS <i>TATE</i> E <i>VIPER</i> COM PROTEÍNAS DO HOSPEDEIROS ENTRE <i>LEP.</i> <i>PYRROCORIS</i> E <i>T. THEILERI</i>	73

FIGURA 17: RELAÇÕES FILOGENÉTICAS DOS DIFERENTES GRUPOS DE RETROELEMENTOS.....	74
FIGURA 18: RELAÇÕES EVOLUTIVAS ENTRE PROTEÍNAS DA FAMÍLIA DE TIROSINA RECOMBINASES.....	76
FIGURA 19: REPRESENTAÇÃO DE UM POSSÍVEL CENÁRIO EVOLUTIVO PARA A ORIGEM DOS RETROTRANSPOSONS.....	84

LISTA DE TABELAS

TABELA 1: GÊNEROS CORRESPONDENTES A FAMÍLIA TRYPANOSOMATIDAE	17
TABELA 2: RELAÇÃO DE GENOMAS SELECIONADOS PARA ANÁLISES <i>IN SILICO</i>	40
TABELA 3: GENES ESCOLHIDOS PARA VERIFICAR POSSÍVEL CASO DE TH DE <i>TATE</i> E <i>VIPER</i> . O NOME E O ID DOS GENES DE <i>LEP. PYRRHOCORIS</i> SÃO INDICADOS.	45
TABELA 4: DADOS DAS SEQUÊNCIAS DE RT USADAS PARA RECONSTRUÇÃO FILOGENÉTICA	46
TABELA 5: DADOS DAS SEQUÊNCIAS DE YR USADAS PARA RECONSTRUÇÃO FILOGENÉTICA	47
TABELA 6: NÚMERO TOTAL DE <i>HITS</i> ENCONTRADOS NA ANÁLISE DE TBLASTN	51
TABELA 7: RESUMO DAS CARACTERÍSTICAS DAS CÓPIAS DE <i>VIPER</i> POTENCIALMENTE CODIFICADORAS	55
TABELA 8: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA GAG-LIKE DE <i>VIPER</i>	59
TABELA 9: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA YR DE <i>VIPER</i>	59
TABELA 10: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA RT/RH DE <i>VIPER</i>	60
TABELA 11: RESUMO DAS CARACTERÍSTICAS DAS CÓPIAS DE <i>TATE</i> POTENCIALMENTE CODIFICADORAS	64
TABELA 12: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA GAG-LIKE DE <i>TATE</i>	66
TABELA 13: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA YR DE <i>TATE</i>	66
TABELA 14: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA RT/RH DE <i>TATE</i>	67

LISTA DE SIGLAS E ABREVIATURAS

BA	- Análise bayesiana
CA	- Capsídeo
DNA	- Ácido desoxirribonucleico
dS	- Taxa de substituições sinônimas
EN	- Endonuclease
HTT	- Transferência horizontal de elementos de transposição
HTT-DB	- Banco de dados de transferência horizontal de elementos de transposição
ICR	- Região interna de complementariedade
IN	- Integrase
IR	- Região intergênica
kDNA	- DNA mitocondrial
LTR	- Longas repetições terminais
MA	- Matriz
MHR	- Região de maior homologia
ML	- Máxima verossimilhança
mRNA	- RNA mensageiro
NC	- Nucleocapsídeo
NCBI	- Centro nacional de informação biotecnológicas
NJ	- Neighbor joining
ORF	- Fase aberta de leitura
Pol II	- RNA polimerase II
RH	- RNaseH
RNA	- Ácido ribonucleico
RT	- Transcriptase reversa
SDR	- Repetição direta dividida
SL	- <i>Splice-leader</i> (mini-éxon)
TE	- Elemento de transposição
TH	- Transferência horizontal

TIR	- Repetição terminal invertida
TSD	- Sítio de duplicação alvo
TV	- Transferência vertical
UTR	- Região não traduzida
VLP	- Partícula viral- <i>like</i>
YR	- Tirosina recombinase

SUMÁRIO

1	INTRODUÇÃO.....	16
1.1	TRIPANOSOMATÍDEOS.....	16
1.1.1	Posicionamento filogenético e evolução.....	16
1.1.2	Características básicas dos tripanosomatídeos	19
1.2	ELEMENTOS DE TRANSPOSIÇÃO.....	23
1.2.1	Classe II – Transposons.....	24
1.2.2	Classe I – Retrotransposons	25
1.2.4	Elemento <i>VIPER</i>	30
1.2.5	Elemento <i>TATE</i>	31
1.2.6	Origem dos retrotransposons	32
1.2.7	Evolução dos elementos de transposição nos genomas.....	34
1.2.8	Impacto dos elementos de transposição nos genomas.....	35
2	OBJETIVOS	39
3	MATERIAIS E MÉTODOS.....	40
3.1	Genomas.....	40
3.2	Análise estrutural do <i>TATE</i> de <i>L. braziliensis</i>	41
3.3	Busca dos elementos <i>VIPER</i> e <i>TATE</i> no genoma dos tripanosomatídeos	42
3.4	Análises evolutivas de <i>TATE</i> e <i>VIPER</i>	44
3.5	Evolução dos elementos da ordem DIRS.....	46
4	RESULTADOS	49
4.1	Análise estrutural do elemento <i>TATE</i> em <i>L. braziliensis</i>	49
4.2	Análise da distribuição de <i>VIPER</i> e <i>TATE</i> nos tripanosomatídeos.....	50
4.3	<i>VIPER</i>	54
4.4	<i>TATE</i>	63
4.5	Evolução dos elementos <i>VIPER</i> e <i>TATE</i> nos genomas dos cinetoplastídeos.....	70
4.6	História evolutiva dos elementos DIRS	73
5	DISCUSSÃO.....	78
5.1	Busca dos elementos nos genomas dos cinetoplastídeos – A influência da montagem dos genomas na análise genômica de TEs	78
5.2	<i>VIPER</i> e <i>TATE</i> : sobreviventes nos genomas dos cinetoplastídeos	80
5.3	Origem independente dos elementos contendo YR	82
5.4	O que ainda precisamos saber sobre os elementos DIRS?	84

6	CONCLUSÕES	87
7	PERSPECTIVAS.....	88
	REFERÊNCIAS.....	89
	ANEXO A – SCRIPT <i>PERL</i> UTILIZADO PARA RECUPERAÇÃO DAS SEQUÊNCIAS.....	99
	ANEXO B – DESCRIÇÃO DAS CÓPIAS DE <i>VIPER</i> ANALISADAS	100
	ANEXO C – DESCRIÇÃO DAS CÓPIAS DE <i>TATE</i> ANALISADAS.....	107
	ANEXO D – ALINHAMENTO DO GENE DOMESTICADO DE <i>VIPER</i> COM GENE PREDITO PARA <i>A. deanei</i>	115
	ANEXO E – RECONSTRUÇÕES FILOGENÉTICAS OBTIDAS PARA OS ELEMENTOS <i>VIPER</i> E <i>TATE</i>	117
	ANEXO F – EXPRESSÃO DE PROTEÍNAS PARA A PRODUÇÃO DE ANTICORPOS CONTRA AS PROTEÍNAS DE <i>VIPER</i>	124

1 INTRODUÇÃO

1.1 TRIPANOSOMATÍDEOS

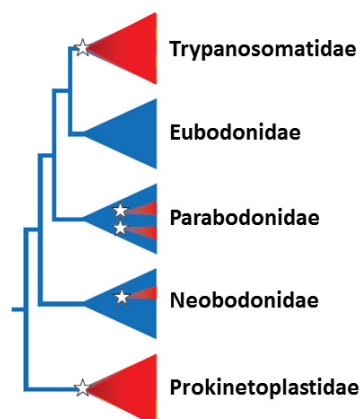
1.1.1 Posicionamento filogenético e evolução

Os tripanosomatídeos pertencem à ordem Kinetoplastida que corresponde a protozoários da linhagem Euglenozoa, podendo ser parasitas ou de vida livre. Esses protozoários são caracterizados dentro dessa ordem por possuírem uma organela específica e altamente especializada, o cinetoplasto (SHLOMAI, 1994).

Inicialmente a divisão taxonômica da ordem Kinetoplastida se baseava nas características morfológicas e nos ciclos de vida dos organismos, estabelecendo assim a divisão em duas subordens Bodonina e Trypanosomatina (VICKERMAN, 1976 apud STEVENS, 2008). Em novos estudos de classificação, baseados em análises de genes conservados, os cinetoplastídeos (membros da ordem Kinetoplastida) foram reagrupados em cinco famílias (Figura 1), Eubodonidae, Parabodoidae, Neobodonidae, Prokinetoplastidae e Trypanosomatidae, sendo que estas duas últimas possuem membros parasitas, enquanto as outras famílias correspondem a membros de vida livre (YAZAKI et al., 2017). Esses autores propõem que esses cinco grupos sejam ranqueados como ordens que compõe a classe Kinetoplastea. Nesse trabalho, preferimos usar a nomenclatura antiga, ainda mais utilizada.

A família Eubodonidae é a mais próxima dos tripanosomatídeos, os quais se separaram do último ancestral comum há mais de 400 milhões de anos, antes da transição evolutiva para o parasitismo (LUKEŠ et al., 2014). Enquanto os tripanosomatídeos se tornaram organismos parasitas, os eubodonídeos mantiveram o modo de vida livre e são encontrados em ambientes aquáticos (OPPERDOES et al., 2016). Dentro dos eubodonídeos, o *Bodo saltans*, que possui o genoma sequenciado e disponibilizado (JACKSON; QUAIL; BERRIMAN, 2008), pode ser utilizado como grupo externo para análises evolutivas e pode ajudar a compreender os padrões de perda e ganho de genes e também auxiliar no entendimento da evolução para o parasitismo.

FIGURA 1: VISÃO GERAL DA FILOGENIA DOS CINETOPLASTÍDEOS



LEGENDA: Filogenia dos cinetoplastídeos inferida a partir de 43 genes que codificam proteínas altamente conservadas. Os ramos azuis indicam espécies de vida livre, enquanto os clados em vermelho indicam espécies parasitas. Os prováveis eventos de mudança de estilo de vida são indicados por estrelas.

FONTE: Modificada de YAZAKI et al. (2017)

Os tripanosomatídeos são parasitas obrigatórios e correspondem atualmente a 19 gêneros, podendo ser classificados ainda em dois grupos: monoxênicos e dixênicos (Tabela 1). Os protozoários monoxênicos representam a maioria das espécies e têm a característica de possuírem apenas um hospedeiro (invertebrados), enquanto os dixênicos possuem mais de um hospedeiro, tendo capacidade de parasitar espécies vegetais e animais (KAUFER et al., 2017).

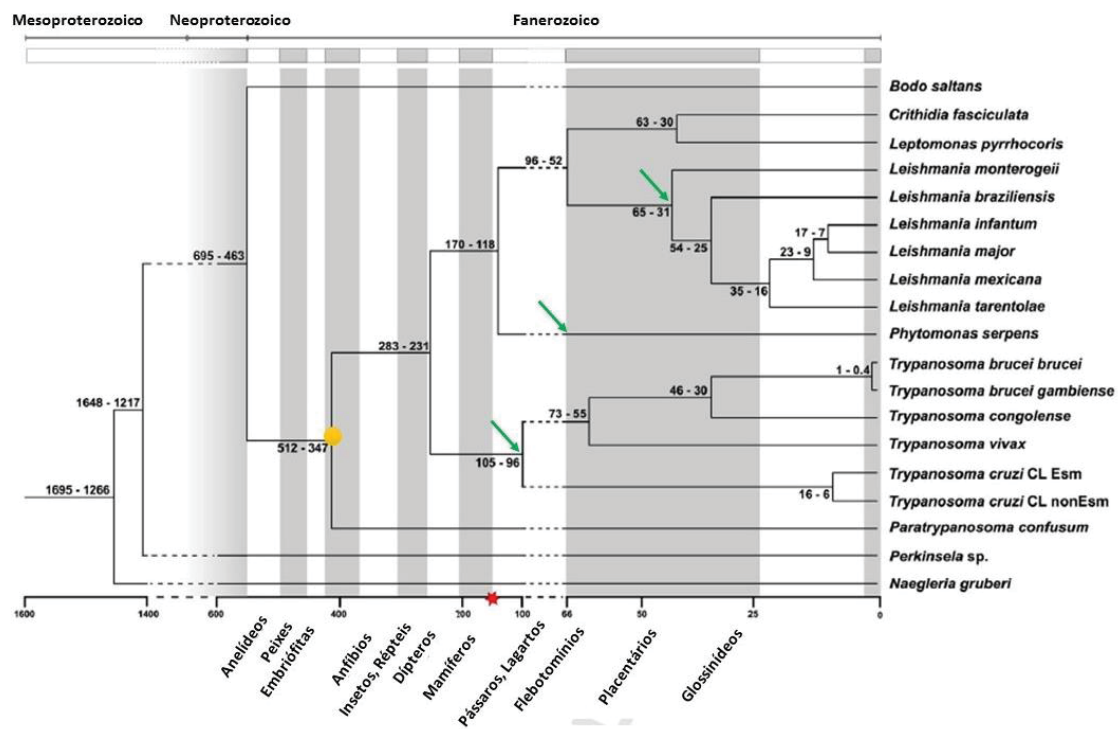
TABELA 1: GÊNEROS CORRESPONDENTES A FAMÍLIA TRYPANOSOMATIDAE

	Gêneros
Monoxênicos	<i>Anqomonas</i>
	<i>Blastocrithidia</i>
	<i>Blechomonas</i>
	<i>Crithidia</i>
	<i>Herptomonas</i>
	<i>Kentomonas</i>
	<i>Leptomonas</i>
	<i>Lotmaria</i>
	<i>Novymonas</i>
	<i>Paratrypanosoma</i>
	<i>Sergeia</i>
	<i>Strigomonas</i>
	<i>Wallaceina</i>
	<i>Zelonia</i>
Dixênicos	<i>Endotrypanum</i>
	<i>Leishmania</i>
	<i>Phytomonas</i>
	<i>Porcisia</i>
	<i>Trypanosoma</i>

FONTE: Modificado de KAUFER, et al., 2017.

Um cenário mais parcimonioso para elucidar o início da atividade parasítica em vertebrados seria a adaptação de uma linhagem de tripanosomatídeo ancestral, que parasitava insetos e acabou se tornando dependente do parasitismo em vertebrados através de eventos de adaptação ocorridos ao longo da evolução (STEVENSON, 2008). Porém, sugere-se que, ao longo da evolução dos tripanosomatídeos, aconteceram três possíveis eventos de transição para a vida dixênica (Figura 2), que ocorreram nos gêneros *Trypanosoma* e *Leishmania*, parasitando vertebrados e em *Phytomonas*, parasitando plantas (LUKEŠ et al., 2014).

FIGURA 2: HISTÓRIA EVOLUTIVA DOS TRIPANOSOMATÍDEOS



O desenvolvimento da vida dixênica é compatível com o tempo estimado de surgimento dos seus hospedeiros. A capacidade de parasitar mamíferos permitiu à algumas espécies o desenvolvimento de especializações que colaboraram com a capacidade patogênica desses parasitas, fazendo com que hoje possuam uma

importância no desenvolvimento de algumas doenças, consideradas de grande impacto para a saúde humana (LUKEŠ et al., 2014; STEVENS, 2008).

Alguns tripanosomatídeos dixênicos são conhecidos no cenário clínico por serem agentes etiológicos de doenças consideradas negligenciadas e que afetam diretamente a saúde pública. A existência de espécies patogênicas na família Trypanosomatidae torna o estudo da biologia básica desses organismos ainda mais importante. Espécies de *Trypanosoma* e *Leishmania* são responsáveis por desencadear diferentes doenças, sendo a tripanossomíase americana conhecida como doença de Chagas, ocasionada pelo *T. cruzi* (CHAGAS, 1909), tripanossomíase africana conhecida como doença do sono, ocasionada pelo *T. brucei*, e leishmaniose ocasionada por diferentes espécies do gênero *Leishmania* (JACKSON, 2015).

1.1.2 Características básicas dos tripanosomatídeos

Os tripanosomatídeos são protozoários que possuem um flagelo que varia de tamanho conforme seu estágio de desenvolvimento e apresentam algumas estruturas especializadas, incomuns a outros eucariotos. A principal estrutura dos tripanosomatídeos e que define os cinetoplastídeos, é o cinetoplasto, uma porção da mitocôndria única que mantém um arranjo complexo e organizado do DNA mitocondrial (kDNA) (DE SOUZA, 2002), sendo este constituído por moléculas circulares denominadas de maxicírculos (22 a 28 kb) e minicírculos (1,5 kb) (SHLOMAI, 1994).

Outra estrutura presente nos tripanosomatídeos são os acidocalciossomos. Essa organela é constituída de vacúolos armazenadores de cálcio, ricos em ortofosfato, pirofosfato e polifosfato, e parece estar associada com a sinalização por meio de cálcio e a osmoregulação da célula (DOCAMPO, 2016). Os tripanosomatídeos também possuem o chamado glicosomo, um peroxissomo especializado que recebe esse nome devido à presença de enzimas da via glicolítica. Algumas espécies também possuem reservossomos, onde são armazenadas as macromoléculas endocitadas.

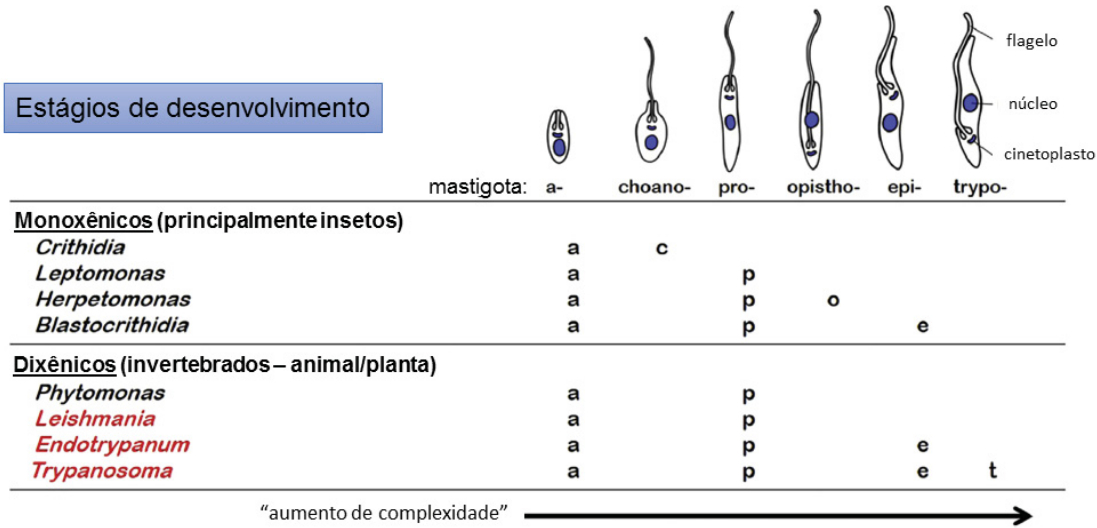
Uma outra importante característica dos tripanosomatídeos é a intensa realização de endocitose. Na maioria das espécies a atividade endocítica e

exocítica ocorre através da bolsa flagelar, que é uma estrutura especializada originada a partir da invaginação da membrana plasmática. Contudo, algumas espécies apresentam o cistóstoma, uma região especializada que atua na ingestão de macromoléculas (SOUZA, 2008).

Os tripanosomatídeos apresentam diferentes estágios celulares ao longo de seu ciclo de vida, alterando assim sua morfologia. A figura 3 traz em resumo os estágios que podem ser apresentados ao longo do ciclo de vida. As formas amastigotas e promastigotas são compartilhadas pela maioria dos gêneros, e é possível notar também que o fato do gênero ser dixênico ou monoxênico não influencia na complexidade da morfologia (O'DONOGHUE, 2017).

Estes protozoários apresentam seis classes morfológicas. As formas epimastigota e tripanomastigota representam a classe *juxtaforma* (do latim, *juxta*: ao lado) e apresentam o flagelo emergindo lateralmente do corpo do celular. As formas coanomastigota, promastigota e opistomastigota representam a classe de *liberforma* (do latim, *liber*: livre) que apresentam o flagelo emergindo centralmente do corpo celular (KAUFER et al., 2017; O'DONOGHUE, 2017).

FIGURA 3: ESTÁGIOS MORFOLÓGICOS APRESENTADOS PELOS TRIPANOSOMATÍDEOS



LEGENDA: Diferentes estágios de desenvolvimento apresentados durante o ciclo de vida dos tripanosomatídeos, tanto para monoxênicos quanto para dixênicos. As letras representam as formas existentes no ciclo de vida da espécie referida, a: amastigota; c: coanomastigota; p: promastigota; o: opistomastigota; e: epimastigota; t: tripanomastigota. A posição de flagelo, núcleo e cinetoplasto também se alteram conforme a mudança do estágio de vida.

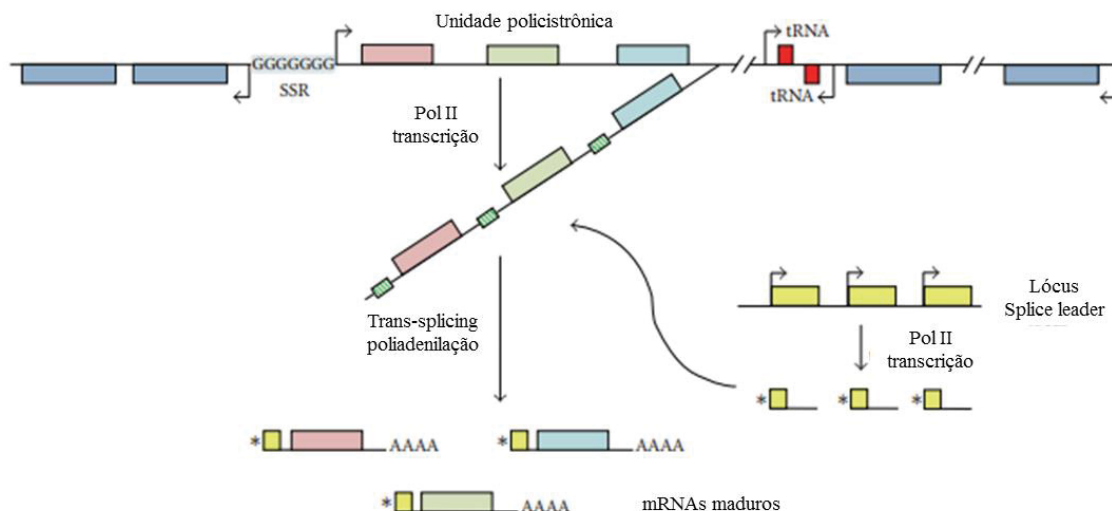
FONTE: Traduzido de O'DONOGHUE, 2017.

Ao longo da diferenciação das formas morfológicas ocorre um reposicionamento do núcleo e do cinetoplasto. Além do reposicionamento destas organelas, também ocorrem modificações no flagelo, como tamanho e posição na célula (DOCAMPO et al., 2005; WHEELER; GLUENZ; GULL, 2013).

Os tripanosomatídeos também compartilham algumas peculiaridades genômicas e funcionais que diferem dos demais eucariotos. Os genomas desses parasitas são organizados em longas unidades policistrônicas, onde vários genes, não necessariamente relacionados, estão organizados sequencialmente na mesma fita de DNA. Devido a esta organização, a transcrição ocorre de forma policistrônica onde todos os genes sequenciais são transcritos de uma única vez (TEIXEIRA, 1998).

A maioria dos genes dos tripanosomatídeos não possuem promotores para a RNA polimerase II (Pol II), ficando assim a regulação da expressão gênica majoritariamente em um nível pós-transcricional. Os genes em geral não possuem íntrons, e são separados por regiões intergênicas (IRs) que são fundamentais para o processamento do policistron (MART et al., 2010). Após a transcrição das unidades policistrônicas irão ocorrer dois eventos necessários para a maturação dos RNAs mensageiros (mRNAs), o *trans-splicing*, quando ocorre a adição de uma sequência de RNA de 39 nucleotídeos conhecida como *splice leader* (SL) ou mini-exon à extremidade 5' dos mRNAs e a adição da cauda poli-A à extremidade 3' (VANHAMME; PAYS, 1995). O gene que codifica a sequência do SL RNA é um dos únicos no genoma dos tripanosomatídeos que apresenta promotor, e devido a isso, sua expressão ocorre de forma independente pela Pol II (MART et al., 2010) (Figura 4).

FIGURA 4: ESQUEMA DO PROCESSO DE TRANSCRIÇÃO GÊNICA EM TRIPANOSOMATÍDEOS



LEGENDA: Representação dos processos de transcrição gênica e processamento dos RNAs em tripanosomatídeos. Os blocos correspondem aos genes dispostos em sequência, formando a unidade policistrônica; a região SSR (*strand-switch region*) é a região de troca de fitas entre duas unidades policistrônicas e onde ocorre o reconhecimento da Pol II para o início da transcrição. Em outro locus, o SL-RNA será transcrito e processado e então cada unidade será adicionada aos mRNAs da unidade policistrônica através de *trans-splicing*. Os mRNAs também sofrem o processo de poliadenilação (representado na figura por quatro As). O asterisco indica a presença do Cap na porção 5' do SL RNA.

FONTE: Modificado de MARTÍNEZ-CALVILLO et al. (2010).

Em 2005, os genomas de três importantes espécies patogênicas da família Trypanosomatidae foram sequenciados: *Trypanosoma cruzi*, *T. brucei* e *L. major* (BERRIMAN et al., 2005; EL-SAYED et al., 2005a; IVENS et al., 2005). Com o sequenciamento simultâneo dos genomas foi possível realizar diversas análises de genômica comparativa, onde foi visto que os três genomas compartilham um proteoma básico de 6200 proteínas (EL-SAYED et al., 2005b). Apesar dos altos níveis de divergência no nível da sequência, já esperados para espécies que divergiram há pelo menos 100 milhões de anos (ver figura 2), essas três espécies apresentam uma grande conservação na ordem dos genes (sintenia). Os poucos locais de rearranjo do genoma entre essas espécies são marcados pela presença de retrotransposons (que serão discutidos posteriormente), sugerindo que essas sequências podem ter desempenhado um papel importante na organização dos genomas dos tripanosomatídeos (GHEDIN et al., 2004).

Após o sequenciamento dos primeiros tripanosomatídeos e com o advento do sequenciamento de nova geração, atualmente encontramos cerca 40 sequências genômicas de tripanosomatídeos disponíveis no banco de dados TriTrypDB, banco de dados de tripanosomatídeos (ASLETT et al., 2009) e NCBI (National Center for

Biotechnology Information), incluindo espécies mononômicas e dixêmicas (BERRIMAN et al., 2005; BUTLER; JASKOWSKA; KELLY, 2017; EL-SAYED et al., 2000, 2005a; FLEGONTOV et al., 2016; IVENS et al., 2005; JACKSON; QUAIL; BERRIMAN, 2008; KELLY et al., 2014; PORCEL et al., 2014; RAYMOND et al., 2012; RUNCKEL; DERISI; FLENNIKEN, 2014; STOCO et al., 2014). Essa grande quantidade de genomas da família Trypanosomatidae sequenciados e em andamento é uma valiosa fonte de dados para diversos tipos de estudos, como o entendimento da variabilidade genética, evolução de famílias gênicas e suas funções, evolução do parasitismo e dos mecanismos da patogenicidade. Nesse contexto, também pode ser destacado a grande oportunidade para a realização de estudos dos elementos de transposição presentes nesses genomas, como o apresentado aqui.

1.2 ELEMENTOS DE TRANSPOSIÇÃO

Elementos de transposição (TEs) são componentes móveis dos genomas de procariotos e eucariotos. Possuem a capacidade de se replicar e se integrar em novas posições no genoma por diferentes mecanismos de transposição, gerando assim um grande impacto sobre a variabilidade dos genomas. Essas transposições são responsáveis por grande parte das modificações na estrutura dos genomas, o que influencia na evolução e adaptação das espécies (BIÉMONT; VIEIRA, 2005; FESCHOTTE; PRITHAM, 2007; KIDWELL; LISCH, 1997; VOLFF, 2006; WESSLER, 2006).

Os TEs utilizam da maquinaria molecular do organismo hospedeiro para codificar suas próprias enzimas, necessárias para a sobrevivência e atividade do elemento e são observados nos genomas como sequências repetitivas que se estendem ao longo das sequências cromossômicas dos organismos (FESCHOTTE, 2008).

Estes elementos podem ser agrupados em duas classes, conforme a molécula intermediária usada na transposição. São divididos em classe I, que se transpõem por intermédio de moléculas de RNA e em classe II, transponíveis por intermédios de DNA (FINNEGAN, 1989). Posteriormente, novos trabalhos aprofundaram ainda mais a classificação dos TEs, propondo divisões dentro das classes baseadas na estrutura e mecanismo de transposição dos elementos (CAPY; BAZIN; LANGIN, 1998; WICKER et al., 2007). No sistema de classificação proposto

por WICKER et al. (2007), pode-se classificar os TEs em: classe, subclasse, ordem, superfamília, família e subfamília.

Poucos estudos têm abordado os TEs de tripanosomatídeos e apenas poucas famílias foram descritas. Estima-se que eles perfazem 2 a 5% do genoma haploide (BERRIMAN et al., 2005; EL-SAYED et al., 2005a). Essa estimativa é muito menor do que a proporção encontrada para outros organismos, como o próprio genoma humano, que é composto por aproximadamente 50% de sequências derivadas de TEs (LANDER et al., 2001). Nenhum TE de classe II foi detectado até o momento nesses genomas (BRINGAUD et al., 2008a). Os retrotransposons descritos para os tripanosomatídeos são poucos e as análises do nosso grupo confirmam essa baixa diversidade (Adriana Ludwig, dados não publicados) e podem ser agrupados em três grandes grupos:

1 - *CRE-like*: elementos inseridos sempre na mesma posição em genes de SL-RNA. Composto pelos elementos *CRE1* e *CRE2* de *Crithidia fasciculata* (GABRIEL et al., 1990; TENG; WANG; GABRIEL, 1995), *SLACS* de *T. brucei* (AKSOY et al., 1990) e *CZAR* de *T. cruzi* (VILLANUEVA et al., 1991).

2 - *Ingi-like*: elementos amplamente distribuídos nos tripanosomatídeos, e são os mais estudados até o momento (BRINGAUD; ROGERS; GHEDIN, 2015). São os elementos chamados: *Ingi*, *RIME* e *TbSIDER* de *T. brucei* (BRINGAUD et al., 2004, 2007), *L1Tc* e *NARTc* de *T. cruzi* (BRINGAUD et al., 2002; MARTÍN et al., 1995), *LmSIDER1* e *LmSIDER2* de *L. major* (SMITH; BRINGAUD; PAPADOPOULOU, 2009).

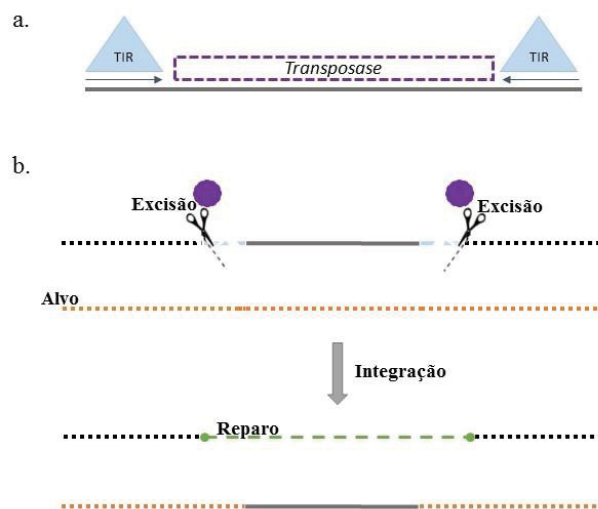
3 - *DIRS*: são os elementos alvo do nosso estudo, que serão descritos a seguir. Compartilham a característica de codificar uma proteína tirosina recombinase (YR), responsável pela inserção no genoma. São os elementos *VIPER* de *T. cruzi* (LORENZI; ROBLEDO; LEVIN, 2006a) e *TATE* de *L. braziliensis* (PEACOCK et al., 2007).

1.2.1 Classe II – Transposons

Os transposons de DNA (figura 5) se integram nos genomas através de um mecanismo de “cópia e cola”, através de uma proteína Transposase codificada pelo próprio elemento. Para que ocorra o processo de transposição, os elementos clássicos desta classe também apresentam sequências de repetições terminais

invertidas, as chamadas TIRs (do inglês: *Terminal Inverted Repets*). Essas sequências serão reconhecidas pela enzima para que ocorra o processo de transposição. Durante o processo de transposição acontece uma duplicação do sítio alvo, formando as TSDs (do inglês: *Target Site Duplications*) (KAWAKAMI; LARGAESPADA; IVICS, 2017; MUÑOZ-LOPEZ et al., 2016). Após a transposição, ocorre o reparo do DNA na região onde ocorreu a excisão por junção das extremidades (*Non-homologous end joining-NHEJ*) ou por recombinação homóloga, aumentando o número de cópias nesse último caso (YU et al., 2004). Estes são os transposons de DNA clássicos, denominados como subclasse I, ordem TIR. Os transposons de DNA não clássicos compreendem os grupos de elementos Crypton, Helitron e Maverick, os quais apresentam processo de transposição diferente dos elementos TIR e enzimas distintas da Transposase (WICKER et al., 2007).

FIGURA 5: MECANISMO BÁSICO DE TRANSPOSIÇÃO DOS ELEMENTOS DE CLASSE II (TRANSPOSONS)



LEGENDA: Em “a” está representado a estrutura básica do transposon, que consiste em uma sequência codificadora de Transposase e duas TIRs. Essa sequência está inserida em sequências cromossômicas do genoma hospedeiro. Em “b” está representado o mecanismo de transposição do elemento, quando ocorre a excisão do transposons, através do reconhecimento das TIRs pela transposase. A sequência de DNA é então inserida em uma nova posição no genoma. A região de onde a sequência foi retirada deve acontecer o reparo, podendo aumentar ou não o número de cópias.

FONTE: A autora, 2018.

1.2.2 Classe I – Retrotransposons

A classe I dos TEs corresponde aos elementos que se transpõem por intermédio de RNA, os chamados retrotransposons. Ao contrário do mecanismo realizado pelos elementos de classe II, os retrotransposons necessitam realizar o

processo de transcrição reversa do seu RNA para que a transposição possa ocorrer (EICKBUSH; JAMBURUTHUGODA, 2008; MUÑOZ-LOPEZ et al., 2016).

Os retrotransposons codificam as próprias enzimas necessárias para a sua mobilização e transposição, sendo ao menos uma transcriptase reversa (RT), para realizar a transcrição reversa do RNA em cDNA, e uma enzima para fazer a inserção das novas cópias no genoma, sendo em geral ou uma integrase (IN) ou uma endonuclease (EN) (BOEKE; STOYE, 1997).

Apesar das diferentes estruturas, divergência entre as sequências e diferentes modos de transposição, todos os retrotransposons compartilham uma proteína RT, que possui regiões catalíticas mais conservadas entre os elementos e é comumente empregada para o estudo filogenético dos retrotransposons, sendo a base da classificação e separação dos grupos reportados até o momento (EICKBUSH; JAMBURUTHUGODA, 2008).

Classicamente, os retrotransposons foram agrupados em dois grupos, conforme a ausência ou presença de longas repetições terminais (LTRs- do inglês, *Long Terminal Repeats*), os retrotransposons com LTRs e os retrotransposons sem LTR (CAPY; BAZIN; LANGIN, 1998). Atualmente, essa classificação tornou-se ineficiente, pois retrotransposons com diferentes estruturas foram sendo descobertos, o que levou a proposição de diferentes ordens e superfamílias.

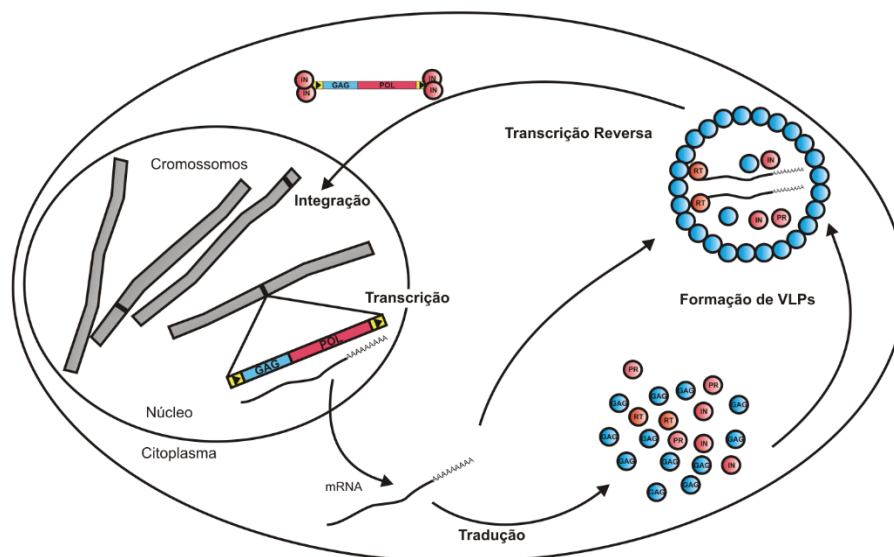
A divisão hoje conhecida dos retrotransposons consiste em 5 ordens: (1) LTR, que possuem as LTRs e estrutura similar a retrovírus, codificando as proteínas Gag e Pol (RT, IN/RH, Protease) (Superfamílias: *Copia*, *Gypsy*, *Bel-Pao*, *Retrovirus*, *ERV*); (2) DIRS, que possuem repetições terminais distintas das LTRs e codificam pelo menos uma RT e uma tirosina recombinase (YR) para a transposição (Superfamílias: DIRS, Ngaro, *VIPER*); (3) PLE, que possuem uma EN, também não possuem LTRs e sua RT é mais próxima filogeneticamente às telomerasas do que às RTs dos retrotransposons com LTR ou LINEs (Superfamília: *Penelope*); (4) LINE do inglês, *Long Interspersed Nuclear Element*), que consistem em elementos que codificam uma RT, uma EN e sua região 3' pode apresentar uma cauda poli-A (Superfamílias: *R2*, *RTE*, *Jockey*, *L1*, *I*); e SINE (do inglês, *Short Interspersed Nuclear Element*, são elementos não autônomos mobilizados pelos LINEs (Superfamílias: *tRNA*, *7SL*, *5S*) (WICKER et al., 2007).

Para este trabalho, os grupos mais importantes a serem discutidos são os LTRs e DIRs. WICKER et al. (2007) não agrupou DIRS juntamente com os LTR,

pois menciona que, provavelmente, o mecanismo de mobilização dos elementos dessa classe seja diferente dos LTR. Já, outro trabalho em resposta ao trabalho de WICKER et al. (2007) continua agrupando os elementos DIRS juntamente com os LTR-retrotransposons (KAPITONOV; JURKA, 2008), o que gera questionamentos sobre a real classificação desses retrotransposons. Apesar de relacionados, notamos que os elementos desses grupos são claramente distintos, e acreditamos que a classificação de WICKER et al. (2007) seja a mais adequada. Os elementos DIRS serão discutidos no tópico 1.2.3.

Por serem evolutivamente relacionados, sugere-se que os LTR-retrotransposons possuem um ciclo de mobilização semelhante aos retrovírus, com a formação de uma partícula tipo viral (VLP - do inglês: *virus like particle*) (Figura 6). Esses elementos apresentam um gene *gag*, que codifica uma proteína Gag (processada em matriz, capsídeo e nucleocapsídeo) requerida para a produção de VLPs. A mobilização dos LTR-retrotransposons consiste na transcrição do elemento dentro do núcleo, no citoplasma as proteínas são traduzidas e a proteína Gag é então recrutada para formar a partícula-viral (VLP) onde ocorrerá então a transcrição reversa do elemento pela RT que, ao ser transcrito em cDNA voltará para o núcleo para se inserir em uma nova posição no genoma com ajuda da proteína IN (HAVECKER; GAO; VOYTAS, 2004; JOHN M; SH; VARMUS, 1997).

FIGURA 6: ESQUEMA DO PROCESSO DE MOBILIZAÇÃO DOS LTR-RETROTRANSPONONS



LEGENDA: Inicialmente ocorre a transcrição do elemento em mRNAs que seguem para o citoplasma onde são traduzidos e ocorre a formação de VLPs onde irá acontecer a transcrição reversa de cópias do RNA. O cDNA gerado é então integrado novamente no genoma. GAG codifica Matriz, Capsídeo e Nucleocapsídeo; POL codifica PR-Protease; RT/RH-Transcriptase Reversa/RNaseH; IN-Integrase. FONTE: LUDWIG (2010).

1.2.3 Grupo DIRS

Devido à escassez de trabalhos na literatura a respeito dos elementos pertencentes ao grupo DIRS, eles ainda são pouco compreendidos. O nome do grupo faz referência ao primeiro elemento descrito deste grupo, o elemento DIRS1 de *Dictyostelium* (CAPPELLO; HANDELSMAN; LODISH, 1985a).

Sabe-se que elementos DIRS apresentam uma estrutura diferente dos LTR-retrotransposons, como já mencionado, composta por um gene que codifica para uma YR no lugar de uma IN, uma proteína com domínios RT e RH, e não codifica nenhuma protease (WICKER et al., 2007). A YR é uma recombinase sítio-específica que utiliza um sítio ativo nucleofílico de tirosina para a quebra da fita de DNA e está presente também nos transposons de DNA chamados *Crypton*. Elas formam uma família de YR sítio-específicas juntamente com outras sequências que incluem recombinases de bacteriófagos, resolvases e transposases procarióticas e flipases de fungos (JAYARAM et al., 2015). Essas sequências têm sido utilizadas em filogenias para tentar entender a evolução dos elementos da ordem DIRS (GOODWIN; POULTER, 2001; LORENZI; ROBLEDO; LEVIN, 2006a; POULTER; BUTLER, 2015a).

Os elementos DIRS também possuem uma ORF (ou fase aberta de leitura – do inglês: *open reading frame*) adicional na região 5', que codifica um produto que tem sido nomeado *gag-like* pela comparação com LTR-retrotransposons, pois está na mesma ordem e apresenta um motivo típico CH-box de proteínas Gag retrovirais (POULTER; BUTLER, 2015a; WICKER et al., 2007). A Gag é caracterizada por conter matriz (MA), capsídeo (CA) e nucleocapsídeo (NC) e é importante para a formação da partícula viral em retrovírus. A Gag apresenta um motivo na região do CA chamado de “região de maior homologia”, o MHR (do inglês: *major homology region*) que apresenta uma alta conservação entre os retrovírus (FREED, 2015). Outro motivo conservado na Gag de retrovírus é o CH-box (CX₂CX₄HX₄C) presente na região do NC que configura a capacidade de ligação à ácidos nucleicos (JOHN M; SH; VARMUS, 1997). O produto da ORF1 dos DIRS, porém, não possui qualquer similaridade de sequência com proteínas Gag e nenhum trabalho experimental tem abordado a função dessa proteína.

Esses elementos também podem codificar uma Metiltransferase (MT) ou uma Hidrolase, cujas funções ainda são desconhecidas. Outra característica

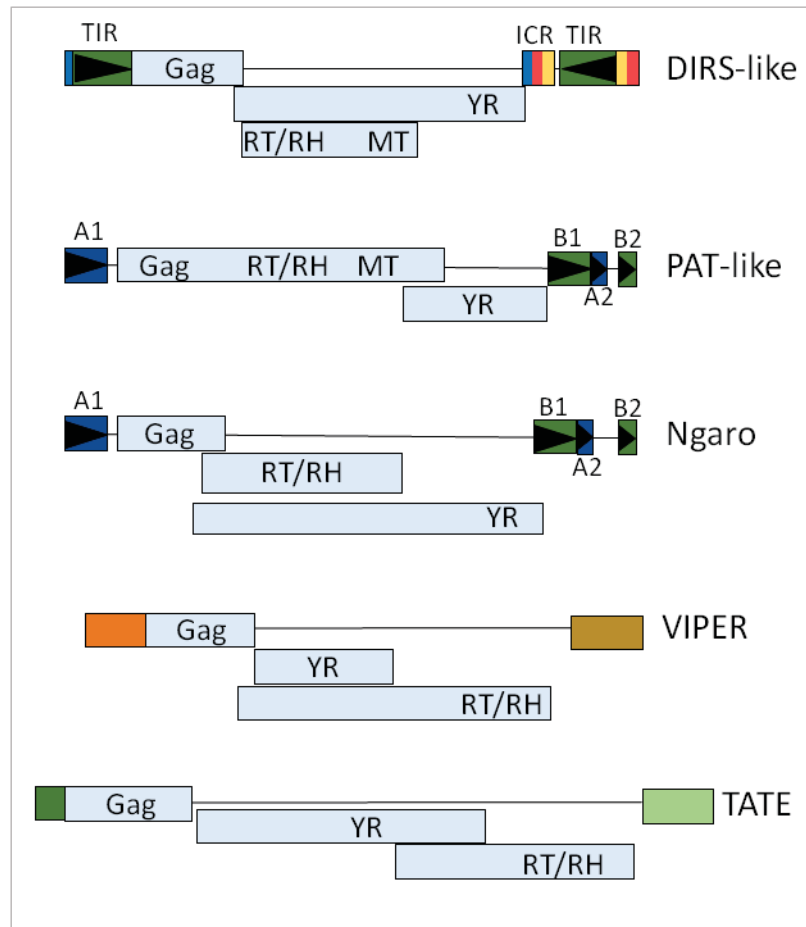
interessante de vários elementos DIRS é a grande sobreposição das ORFs que codificam as proteínas YR e RT/RH (PIEDNOËL et al., 2011a).

Vários trabalhos envolvendo análises genômicas têm relatado a presença de elementos da ordem DIRS em variados genomas, como peixes, anfíbios, répteis, fungos, nematoides e plantas, porém poucos elementos foram caracterizados em detalhes (MUSZEWSKA; STECZKIEWICZ; GINALSKI, 2013; POULTER; BUTLER, 2015a; SZITENBERG et al., 2014).

A figura 7 mostra a estrutura de diferentes elementos que podem ser considerados da ordem DIRS. Eles diferem em algumas características. Os elementos que chamamos de DIRS-like são os que apresentam características similares a do elemento *DIRS1* de *Dictyostelium* apresentando regiões terminais invertidas (TIRs) e uma região interna complementar (ICR – do inglês *internal complementary region*). Elementos PAT-like e Ngaro possuem regiões repetidas diretas divididas (SDRs – *split direct repeats*). Os elementos *VIPER* e *TATE*, alvos do nosso estudo, que são os elementos DIRS de tripanosomatídeos até o momento não possuem padrão de repetições determinadas.

Essas repetições têm importância para o processo de transcrição reversa, pois através delas seriam geradas regiões de anelamento necessárias para o reconhecimento e ação da RT. Esse processo de transcrição reversa não é bem detalhado, mas com certeza difere do processo conhecido para os LTR-retrotransposons (mesmo mecanismo utilizado pelos retrovírus), e é sugerido que ao final gere uma molécula de DNA circular que será inserida no genoma pela YR (CAPPELLO; HANDELSMAN; LODISH, 1985b; GOODWIN; POULTER, 2004; POULTER; GOODWIN, 2005). Ao contrário do que é visto para outros elementos, os elementos da ordem DIRS não geram TSDs quando integrados pela YR (PIEDNOËL et al., 2011b; POULTER; BUTLER, 2015b).

FIGURA 7: ESTRUTURA DOS DIFERENTES GRUPOS DE RETROTRANSPOSONS CODIFICADORES DE YR.



LEGENDA: Regiões repetidas estão representadas por triângulos. Elementos DIRS-like apresentam regiões terminais invertidas (TIR) e uma região interna complementar (ICR). Elementos PAT-like e Ngaro possuem regiões repetidas diretas divididas (SDRs) (A1, B1, A2, B2). *VIPER* e *TATE* não possuem repetições. As ORFs estão representadas por caixas em azul com a indicação dos domínios codificados. Gag, primeira ORF dos elementos da ordem DIRS; RT/RH - Transcriptase Reversa/RNaseH; MT – Metiltransferase; YR – Tirosina Recombinase. A figura não foi desenhada em escala. FONTE: Modificada de POULTER e BUTLER (2015).

Apesar de todos os elementos dessa ordem codificarem uma YR, diferenças estruturais e análises filogenéticas sugerem que os elementos DIRS não formam um grupo monofilético, ou seja, esses elementos devem ter tido mais de uma origem na história evolutiva dos retrotransposons. Esse tópico, sobre a origem desses elementos, é apresentado no item 1.2.6.

1.2.4 Elemento *VIPER*

VAZQUEZ et al., (2000) descreveu pela primeira vez o elemento *VIPER* em *T. cruzi* como um elemento incompleto, degenerado de 2,326 pb. Esses autores também descrevem a existência de uma versão truncada de *VIPER*, chamada *SIRE*

(*Short Interspersed Repetitive Element*). O elemento SIRE é composto pelos primeiros 185 pb da porção 5' do *VIPER*, uma região de aproximadamente 30 pb exclusiva e mais 220 pb da porção 3' do *VIPER*. *SIRE* não é autônomo, pois não codifica nenhuma proteína e é possivelmente mobilizado pelas enzimas do *VIPER*.

VIPER foi pouco estudado e inicialmente classificado como um LTR-retrotransposon (VAZQUEZ et al., 2000), mas assim como os outros elementos DIRS, *VIPER* possui um gene que codifica YR (LORENZI; ROBLEDO; LEVIN, 2006a), tendo sido classificado posteriormente, como já mencionado, como uma superfamília da ordem DIRS (WICKER et al. 2007).

Esse elemento foi considerado inativo, pois nenhuma cópia íntegra foi encontrada no genoma de *T. cruzi* CL Brener, a primeira linhagem sequenciada (BRINGAUD et al., 2008a; LORENZI; ROBLEDO; LEVIN, 2006a). Uma sequência consenso de *VIPER* foi construída a partir do alinhamento várias cópias e está disponível no Repbase, um banco de dados de sequências repetitivas (JURKA et al., 2005) (http://www.girinst.org/protected/repbase_extract.php?access=TcVIPER). A estrutura da sequência consenso de 4390 pb de *VIPER* pode ser observada na figura 7 na qual foram identificadas três ORFs, codificando Gag-like, YR e RT/RH. Não foi reportado até o momento a existência de repetições de nenhum tipo nas extremidades desse elemento, assim como são observadas em outros elementos da ordem DIRS.

A presença do elemento também foi relatada em *T. brucei* e *T. vivax* (LORENZI; ROBLEDO; LEVIN, 2006) e mais recentemente LUDWIG e KRIEGER, (2016) relatam sequências remanescentes do elemento em várias espécies de tripanosomatídeos e em *B. saltans*, sugerindo que o elemento já estava presente no ancestral dessas espécies.

Nesse trabalho apresentamos uma análise detalhada da distribuição, conservação e evolução desse elemento em tripanosomatídeos.

1.2.5 Elemento *TATE*

O elemento *TATE* (*telomere-associated transposable elements*) foi encontrado como cópias em *tandem* nos telômeros de *L. braziliensis*, possuindo um padrão de inserção em uma sequência telomérica hexamérica GGGTTA, e foi mencionado rapidamente no artigo do sequenciamento deste genoma (PEACOCK et

al., 2007). Por apresentar uma YR no lugar de uma IN, *TATE* foi mencionado como sendo um elemento relacionado a IN de fago (SMITH; PEACOCK; CRUZ, 2007). Recentemente, esse elemento também foi reportado em *L. panamensis* e *B. saltans* (JACKSON et al., 2016).

A estrutura básica do elemento descrita no material suplementar de PEACOCK et al., (2007) não delimita as ORFs e não mostra a presença de um gene *gag-like*. Já a sequência consenso presente no Repbase (http://www.girinst.org/protected/repbase_extract.php?access=TATE) mostra um elemento de 6986 pb, codificando três ORFs (*Gag-like*, YR, RT/RH), como pode ser visto na figura 7. Assim como para o elemento *VIPER*, nenhum tipo de repetição foi reportado nas extremidades desse elemento.

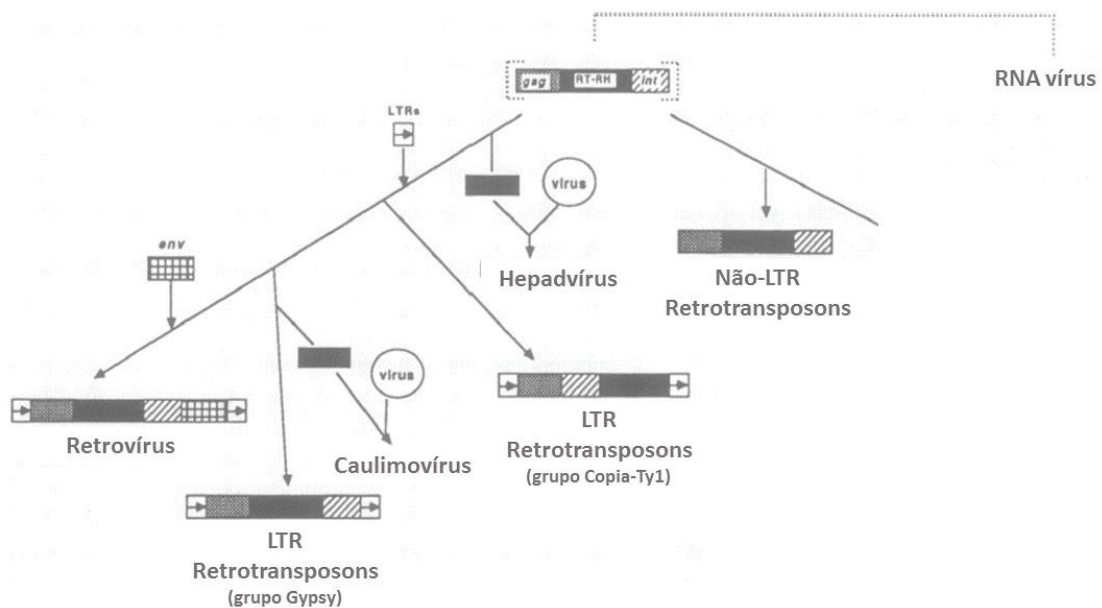
Ainda não há trabalhos específicos sobre esse elemento, mas foi reportada a produção de pequenos RNAs endógenos em *L. braziliensis* contra *TATE* (representado cerca de 75% do repertório de pequenos RNAs encontrado) indicando um mecanismo de regulação do elemento mediado por RNA de interferência (RNAi) (ATAYDE et al., 2013).

Nesse trabalho apresentamos uma análise detalhada da distribuição, conservação e evolução de *TATE* nos genomas dos tripanosomatídeos.

1.2.6 Origem dos retrotransposons

XIONG e EICKBUSH (1990) elucidam hipóteses sobre a origem dos retroelementos através das relações entre sete regiões conservadas da proteína RT de 82 retroelementos de animais, plantas, protozoários e bactérias (totalizando 178 resíduos). Segundo esses autores, um retrotransposon ancestral sem LTRs possuía os genes *gag-like* e *pol-like* e divergiram nas linhagens que originaram os LTR-retrotransposons e os retrotransposons sem LTR muito primitivamente (Figura 8). Posteriormente, em uma das linhagens, a aquisição de LTRs deu origem aos grupos de retrotransposons *Ty1-copia* e *Ty3-gypsy* e em seguida, por meio da aquisição de um gene envelope (*env*), deu origem aos retrovírus. A partir desta mesma linhagem originaram-se os grupos *Hepadnavirus* (exemplo, *Hepatite B virus*) e *Caulimovirus* (vírus de plantas como *Cauliflower mosaic virus*) pela combinação com algum vírus pré-existente (XIONG; EICKBUSH, 1990).

FIGURA 8: ESQUEMA DA ORIGEM DOS RETROELEMENTOS DE XIONG E EICKBUSH (1990)



LEGENDA: A figura mostra um esquema da origem dos retroelementos. As diferentes ORFs, gag, RT/RH, Int (IN) e env são mostradas em diferentes tonalidades e padrões. LTRs são representadas por uma caixa contendo uma seta. As aquisições das LTRs e do gene env são indicadas na linhagem que deu origem aos LTR-retrotransposons e retrovírus. Os grupos Ty1/copia e Ty3/gypsy, além de formarem clados distintos na filogenia da RT apresentam uma ordem diferente dos domínios de pol (IN-RT/RH e RT/RH-IN, respectivamente). Para *Hepadnavirus* e *Caulimovirus* é assumido que somente parte do gene pol, contendo o domínio RT se associou com algum vírus já existente.

FONTE: Modificada de XIONG E EICKBUSH (1990).

Na filogenia apresentada por XIONG e EICKBUSH (1990) é possível observar dois elementos que não formavam um grupo monofilético com os outros LTR-retrotransposons, os elementos *Pao* de *Bombyx mori* e *DIRS1* (já mencionado)(XIONG; EICKBUSH, 1990). Estudos posteriores indicam a separação dos elementos dessa linhagem em seis clados: grupo Ty1/copia, BEL (grupo do elemento *Pao*), *DIRS1* (clado que inicialmente apresentava apenas o elemento *DIRS1*), Ty3/Gypsy, Caulimovírus e Retrovírus, sendo que elementos BEL contém LTRs (MALIK; HENIKOFF; EICKBUSH, 2000).

Dessa forma, fica claro que DIRS estava separado dos demais retrotransposons. Ao longo dos anos, uma série de novos retrotransposons codificadores de YR foram sendo descobertos e WICKER et al. (2007) os classificaram, como já mencionado, como elementos da ordem DIRS. Esses autores separaram a ordem DIRS em três superfamílias, DIRS, Ngaro e VIPER, mas não mencionaram os critérios adotados para tal classificação. Outro trabalho mais

recente, que considerou a filogenia de RT e YR, separou o grupo DIRS em DIRS-*like*, Ngaro-*like* e PAT-*like*, porém neste trabalho o elemento *VIPER* não foi mencionado (POULTER; BUTLER, 2015b). Nas análises filogenéticas da RT apresentada por LORENZI et al. (2004) é possível observar a separação em três clados: *VIPER*, Ngaro e DIRS (incluindo PAT no mesmo clado). O elemento *TATE* nunca foi incluído em análises filogenéticas e não sabemos se forma um grupo juntamente com *VIPER* ou se seria um grupo independente. Para resolver tal discordância é necessário investigar a história evolutiva dos retrotransposons incluindo novos elementos e acrescentando na mesma análise os elementos *VIPER* e *TATE*. Esse também foi um dos pontos abordados neste trabalho.

1.2.7 Evolução dos elementos de transposição nos genomas

Os TEs, em geral, como fazem parte do genoma, são transmitidos entre gerações por transferência vertical (TV) e conservados ou não ao longo dos processos de especiação durante a evolução. Nos últimos anos, porém, outra forma de transmissão de TEs tem chamado a atenção, a transferência Horizontal (TH). A TH consiste na transferência de material genético independente da relação de descendência, entre espécies distantemente relacionadas, por meio não sexual, ou para organismos não sexuais, de uma célula para outra não descendente (SCHAACK; GILBERT; FESCHOTTE, 2010).

Esse processo é frequente em procariotos tendo uma grande importância na evolução das populações e aquisição de resistência a antibióticos (SOUCY; HUANG; GOGARTEN, 2015). Já em eucariotos a TH foi, por muito tempo, considerada rara, mas vários trabalhos têm reportado esse fenômeno cada vez mais nos genomas eucariotos, particularmente quando consideramos a evolução dos TEs (GILBERT; CORDAUX, 2013; LUDWIG; LORETO, 2007; LUDWIG; VALENTE; LORETO, 2008; ROSSATO et al., 2014; SCHAACK; GILBERT; FESCHOTTE, 2010; WALLAU; ORTIZ; LORETO, 2012; WALLAU; VIEIRA; LORETO, 2018).

Até o momento, mais de 2800 eventos de HT de TEs (HTT) foram reportados na literatura (WALLAU; VIEIRA; LORETO, 2018). O número crescente de casos reportados e a sua importância para a história evolutiva dos elementos, levaram a criação de um banco de dados (HTT-DB) a fim de manter as informações a respeito de novos eventos (DOTTO et al., 2015, 2018).

A maior parte dos eventos de HT ocorre com transposons de DNAs, porém também há relatos de eventos relacionados com retrotransposons, sendo em maior parte relatados com LINEs e LTR-retrotransposons (PECCOUD et al., 2017). Até o presente momento nenhum trabalho relatou eventos de HT para elementos do tipo DIRS como observamos no HTT-DB. Também, nunca foram evidenciados casos de HTT de tripanosomatídeos.

Algumas primícias são necessárias serem observadas para considerarmos a possibilidade de HTT: incongruências filogenéticas que mostram discrepâncias na distribuição dos TEs, quando associados com a filogenia da espécie hospedeira; a alta similaridade dos genes entre espécies filogeneticamente distantes; e a presença dessas espécies em um mesmo nicho ecológico (PANAUD, 2016). Diferentes abordagens podem ser utilizadas para testar se a possibilidade de HTT. Podemos verificar HTT, por exemplo, através da comparação da taxa de substituições sinônimas (dS), entre o TE e os genes da espécie hospedeira, neste caso a taxa dS do TE deve ser menor que a taxa dos genes do hospedeiro (LUDWIG; VALENTE; LORETO, 2008; WALLAU; ORTIZ; LORETO, 2012).

A transferência horizontal e ativação de TEs podem ser uma vantagem evolutiva para as espécies, uma vez que os elementos podem carregar consigo genes das espécies hospedeiras enquanto se transpõem em diferentes células (FROST et al., 2005), o que pode, através da integração das sequências por intermédio dos mecanismos de transposição dos TEs, colaborar com a aquisição de sequências regulatórias importantes para as espécies (FESCHOTTE, 2008).

1.2.8 Impacto dos elementos de transposição nos genomas

Os TEs já foram considerados DNA lixo ou DNA egoísta, no entanto, é demonstrado que a ativação dos elementos em um genoma pode gerar grandes impactos para o organismo hospedeiro, tanto positivamente gerando variabilidade genética podendo ser importante para a evolução da espécie, quanto negativamente, influenciando no surgimento de mutações e doenças (DUBIN; MITTELSTEN SCHEID; BECKER, 2018).

Durante o processo de transposição, os TEs podem se inserir em outros locais ao longo das sequências, e essas inserções, dependendo da localização, se próximo ou dentro de um gene, podem afetar diretamente na expressão gênica. Os eventos de inserção podem ocorrer dentro de um éxon ou íntron gerando transcritos truncados, *splicing* alternativo ou exonização ou ainda uma combinação destes eventos. Outra forma de afetar a expressão é pela inserção em regiões de *enhancers* ou promotores fazendo com que a transcrição seja reduzida ou aumentada, ou também alterando o perfil de metilação do DNA gerando silenciamento da transcrição (DUBIN; MITTELSTEN SCHEID; BECKER, 2018).

Para melhor compreender o impacto e a dinâmica os TEs nos genomas, muitos estudos utilizam *Drosophila* como modelo experimental e foi observado que nessas espécies cerca de 80% das mutações espontâneas ocorrem pela atividade dos TEs (GREEN, 1988 *apud* GUERREIRO, 2012) e que situações de estresse estão intrinsecamente relacionadas com a ativação dos elementos (GUERREIRO, 2012). Mutações espontâneas por intermédio de TEs também são reportadas no genoma humano, sendo 0,17% dessas mutações ocasionadas pelo retrotransposon LINE-1 (LISCH; KIDWELL, 2002).

A alteração no perfil de expressão dos genes ocasionada pelos eventos de inserção dos TEs pode levar ao desencadeamento de processos patológicos nos organismos hospedeiros. Estudos de mobiloma têm tentado encontrar e relatar a predominância dos TEs em processos patológicos e, estudos já realizados mostram a associação dos elementos em cerca de 75 doenças humanas como hemofilia, esclerose lateral amiotrófica, degeneração lombar frontotemporal e diversos tipos de câncer (MCCULLERS; STEINIGER, 2017). Com relação ao câncer, estudos têm caracterizado a superexpressão de alguns TEs no prognóstico dessas doenças. Um elemento bem estudado nessa questão é o retrotransposon LINE-1 nos cânceres de mama, ovário, pâncreas, gastrointestinal, orofaringe e retal, atuando na heterogeneidade da população de células malignas devido aos seus eventos de inserção, podendo ser utilizado inclusive como marcador de prognóstico destas doenças (BURNS, 2017).

Ao longo da evolução, os TEs podem gerar vantagens para os organismos hospedeiros. A ativação de um TE em certo ponto da evolução de uma espécie, também pode realizar modificações importante nos genes ou até mesmo colaborar com a formação de novos genes, como ocorre nos casos de domesticação. A

domesticação de um TE acontece quando uma sequência do elemento (região codificadora total, parcial ou região regulatória) é cooptada para desempenhar alguma função específica no organismo hospedeiro (JANGAM; FESCHOTTE; BETRÁN, 2017). Alguns exemplos de domesticação são: gene *cas9*, importante para a integração de DNA viral no locus CRISPR, que derivou do domínio *tnpB* de um transposon de DNA (KAPITONOV; MAKAROVA; KOONIN, 2015); os genes *MAIL1* e *MAIN*, importantes para o silenciamento de TEs, derivados dos LTR-retrotransposons *Gypsy* e *Copia* (IKEDA et al., 2017); e os genes que atuam na formação placentária em humanos (*Syncytin 1* e *Syncytin 2*) que foram derivados da proteína Env do retrovírus endógeno HERV-W (LAVIALLE et al., 2013). Como fonte de regiões regulatórias, pode-se destacar a importância dos TEs no genoma humano, onde é observado que 25% das sequências promotoras de genes contêm regiões de TEs (JORDAN et al., 2003).

Diversos são os exemplos conhecidos de domesticação de TEs em vários organismos, e um caso recentemente elucidado é o gene *Arc* que derivou da proteína Gag dos retrotransposons do tipo Ty3/*gypsy* em tetrápodes e *Drosophila*. É expresso na região neuronal de mamíferos e tem importante influência na memória de longo prazo e plasticidade sináptica, sendo relatado também em distúrbios neurológicos. Assim como a Gag de retrotransposons e retrovírus, a proteína Arc forma uma VLP que são liberadas pelos neurônios em vesículas extracelulares, podendo assim se transferir de um neurônio para o outro (PASTUZYN et al., 2018).

Casos de domesticação também estão presentes em tripanosomatídeos. Foi reportado um caso da ação regulatória que os TEs podem desempenhar. É visto que em *L. major* um elemento chamado *LmSIDER2* localiza-se principalmente na região 3'UTRs (*untranslated regions*) de diversos genes e leva a instabilidade dos mRNAs devido a característica de ribozima do elemento, atuando assim diretamente sobre a regulação da expressão gênica (BRINGAUD et al., 2007, 2008a; SÁNCHEZ-LUQUE et al., 2012; SMITH; BRINGAUD; PAPADOPOULOU, 2009).

Recentemente, foi reportado um possível caso de domesticação do elemento *VIPER* em algumas espécies de tripanosomatídeos. Um gene de aproximadamente 3 kb foi encontrado em posição sintênica em *L. braziliensis*, *L. panamensis*, *Leptomonas pyrrhocoris*, e *C. fasciculata*. O produto deste gene tem similaridade na região carboxi com a região carboxi da proteína RT de *VIPER*. Embora não é conhecida a função desse gene, a conservação de uma ORF tão

grande por mais de 50 milhões de anos indica a atuação de seleção para remoção de códons de parada e a proteína deve estar desempenhando algum papel nesses organismos (LUDWIG; KRIEGER, 2016).

É possível observar uma enorme variedade de efeitos dos TEs nos mais diversos organismos, que são difíceis de prever de antemão. Com certeza, muitos outros exemplos de domesticação e de efeitos dos TEs nos genomas permanecem desconhecidos, justificando o estudo dessas sequências em diferentes espécies. Apesar da aparente baixa quantidade e diversidade dos TEs em tripanosomatídeos, os casos de domesticação relatados indicam que eles podem servir como importante fonte de novidades evolutivas e a caracterização básica dos TEs que compõem esses genomas é um primeiro passo para o melhor entendimento do impacto dessas sequências nesses organismos (LUDWIG; KRIEGER, 2016). Assim, nesse trabalho apresentamos uma análise da distribuição, conservação e evolução de dois elementos dos tripanosomatídeos como um ponto de partida para estudos de caracterização molecular e funcional.

2 OBJETIVOS

2.1 Objetivo geral

Estudar os retrotransposons *VIPER* e *TATE* em uma perspectiva evolutiva nos genomas dos tripanosomatídeos.

2.2 Objetivos específicos

- Caracterizar a estrutura geral do elemento *TATE*.
- Entender a distribuição, conservação e evolução dos elementos *VIPER* e *TATE* em tripanosomatídeos por meio de buscas *in silico* e análises filogenéticas.
- Entender a origem dos elementos *VIPER* e *TATE* por meio de estudos evolutivos em comparação com os outros elementos DIRS e outros retrotransposons.

3 MATERIAIS E MÉTODOS

3.1 Genomas

Para as análises, utilizamos um total de 44 genomas de tripanosomatídeos disponíveis, junto com o genoma de *B. saltans*, que foi utilizado como grupo externo para as análises. Os genomas foram obtidos a partir dos bancos de dados TritrypDB (ASLETT et al., 2009) e NCBI, conforme destacado na Tabela 2.

TABELA 2: RELAÇÃO DE GENOMAS SELECIONADOS PARA ANÁLISES *IN SILICO*

Fonte: TritrypDB			
Organismo	Versão	Nível de montagem/Cobertura	Fonte do dado/Contato
<i>Blechnomonas ayalai</i> B08-376	Versão: 2017-01-18	Não informado	Vyacheslav Yurchenko
<i>Crithidia fasciculata</i> strain Cf-CI	Versão: 2015-06-01	Cromossomo/ 130x	Stephen M. Beverley
<i>Endotrypanum monterogeii</i> strain LV88	Versão: 2015-01-16	Cromossomo/ 50x	Stephen M. Beverley
<i>Leishmania aethiopica</i> L147	Versão: 2014-12-16	Cromossomo/ 110x	Stephen M. Beverley
<i>Leishmania amazonensis</i> MHOM/BR/71973/M2269	Versão: 2013-07-25	Scaffold/ 96x	Diana Bahia
<i>Leishmania arabica</i> LEM1108	Versão: 2014-12-16	Scaffold/ 187x	Stephen M. Beverley
<i>Leishmania braziliensis</i> MHOM/BR/75/M2903	Versão: 2014-12-16	Cromossomo/ 72x	Stephen M. Beverley
<i>Leishmania braziliensis</i> MHOM/BR/75/M2904	Versão: 2016-05-28	Cromossomo	Pathogen Sequencing Unit
<i>Leishmania donovani</i> BPK282A1	Versão: 2016-05-28	Cromossomo/ 52x	Matthew Berriman
<i>Leishmania enriettii</i> LEM3045	Versão: 2014-12-16	Scaffold/ 124x	Stephen M. Beverley
<i>Leishmania gerbilli</i> LEM452	Versão: 2014-12-16	Scaffold/ 140x	Stephen M. Beverley
<i>Leishmania infantum</i> JPCM5	Versão: 2015-12-07	Cromossomo	GeneDB
<i>Leishmania major</i> Friedlin	Versão: 2016-05-28	Cromossomo	GeneDB
<i>Leishmania mexicana</i> MHOM/GT/2001/U1103	Versão: 2016-05-28	Cromossomo	GeneDB
<i>Leishmania panamensis</i> MHOM/COL/81/L13	Versão: 2014-12-16	Cromossomo	Stephen M. Beverley
<i>Leptomonas pyrrhocoris</i> H10	Versão: 2015-05-05	Cromossomo/ 37x	Julius Lukes
<i>Leptomonas seymouri</i> ATCC 30220	Versão: 2015-05-06	<i>Leptomonas seymouri</i> ATCC 30220 Sequence and Annotation	Vyacheslav Yurchenko
<i>Leishmania</i> sp. MAR LEM2494	Versão: 2014-12-16	<i>Leishmania</i> sp. MAR LEM2494 Genome sequence and annotation	Stephen M. Beverley
<i>Leishmania tarentolae</i> Parrot-Tarll	Versão: 2011-06-22	16x	Jacques Corbeil
<i>Leishmania tropica</i> L590	Versão: 2014-12-16	Cromossomo/ 253x	Stephen M. Beverley
<i>Leishmania turanica</i> LEM423	Versão: 2014-12-16	Scaffold/ 108x	Stephen M. Beverley

<i>Trypanosoma brucei brucei</i> TREU927	Versão: 2016-05-19	Cromossomo	Pathogen Sequencing Unit
<i>Trypanosoma brucei gambiense</i> DAL972	Versão: 2016-05-28	Cromossomo	Pathogen Sequencing Unit
<i>Trypanosoma brucei</i> Lister strain 427	Versão: 2010-10-20	Cromossomo	Pathogen Sequencing Unit
<i>Trypanosoma congolense</i> IL3000	Versão: 2016-06-06	Não informado	Matthew Berriman
<i>Trypanosoma cruzi</i> CL Brener Esmeraldo-like	Versão: 2015-12-07	Cromossomo	Pathogen Sequencing Unit
<i>Trypanosoma cruzi</i> CL Brener Non-Esmeraldo-like	Versão: 2015-12-07	Cromossomo	GeneDB
<i>Trypanosoma cruzi marinkellei</i> strain B7	Versão 1.0	Contig/ 30 x	Oscar Frazen
<i>Trypanosoma evansi</i> STIB 805	Versão: 2014-06-03	Não informado	Achim Schnauffer
<i>Trypanosoma grayi</i> ANR4	Versão: 2014-06-17	Scaffold/ 10 x	Mark Field
<i>Trypanosoma rangeli</i> SC58	Versão: 2013-10-30	Contigs	Edmundo Carlos Grisard
<i>Trypanosoma vivax</i> Y486	Versão: 2016-06-06	Não informado	Pathogen Sequencing Unit
Fonte: NCBI			
Organismo	GenBank acesso	Cobertura/nível de montagem	Remetente
<i>Angomonas deanei</i> ATCC PRA-265	GCA_001659865.1	344x/ Scaffold	Heinrich Heine University Duesseldorf
<i>Bodo saltans</i> Lake Konstanz	GCA_001460835.1	170x/ Contig	Wellcome Trust Sanger Institute
<i>Crithidia bombi</i> IL132	GCA_002216585.1	100x/ Scaffold	University of Massachusetts
<i>Crithidia mellificae</i> ATCC 30862	GCA_002216565.1	100x/ Scaffold	University of Massachusetts
<i>Leishmania peruviana</i> PAB-4377_V1	GCA_001403675.1	59x/ Cromosomo	UFMG
<i>Lotmaria passim</i> ATCC PRA-422	GCA_002216525.1	100x/ Scaffold	University of Massachusetts
<i>Phytomonas francai</i>	GCA_001766655.1	200x/ Scaffold	University of Oxford
<i>Phytomonas</i> sp. isolate EM1	GCA_000582765.1	20x/ Scaffold	Genoscope CEA
<i>Phytomonas</i> sp. isolate Hart1	GCA_000982615.1	20x/ Scaffold	Genoscope CEA
<i>Strigomonas culicis</i> TCC012E	GCA_000482145.1	23x/ Contig	Virginia Commonwealth University
<i>Trypanosoma cruzi</i> Dm28c PB1	GCA_002219105.2	150x/ Contig	Instituto Carlos Chagas - ICC/Fiocruz-PR
<i>Trypanosoma equiperdum</i> OVI V2	GCA_001457755.2	104x/ Scaffold	ANSES
<i>Trypanosoma theileri</i> Edinburgh	GCA_002087225.1	100x/ Scaffold	University of Edinburgh

3.2 Análise estrutural do *TATE* de *L. braziliensis*

Para analisar o elemento *TATE* de *L. braziliensis* foi realizado BLASTn online no TritypDB (ASLETT et al., 2009) contra o genoma de *L. braziliensis* MHOM/BR/75/M2904. Utilizamos como *query* a sequência de nucleotídeos canônica (consenso) de *TATE* disponível no Repbase (JURKA et al., 2005), e consideramos para as análises as cópias encontradas no BLAST dentro do *cut-off* de *e-value* de

10^{-10} . As cópias foram obtidas manualmente do genoma utilizando a ferramenta “*genome browser*” e baixando uma região de aproximadamente 20 Kb em arquivos FASTA.

As 15 cópias mais conservadas foram alinhadas no programa no MUSCLE (EDGAR, 2004) e visualizadas no GeneDoc 2.7 (NICHOLAS; NICHOLAS; DEERFIELD, 1997) e o limite dos elementos foram definidos. Para verificar a existência de repetições terminais, foi realizado BLASTn (NCBI), opção “*Align two or more sequences*” e “*Somewhat similar sequences (blastn)*”, utilizando a sequência de nucleotídeos a ser analisada como *query* e como *subject*. As sequências nucleotídicas contendo *TATE* foram submetidas à ferramenta ORFfinder do NCBI para verificar as ORFs presentes nas sequências e as sequências aminoacídicas codificadas por cada uma das ORFs foram usadas para busca de domínios conservados utilizando a ferramenta CD- *search* do NCBI (MARCHLER-BAUER; BRYANT, 2004).

3.3 Busca dos elementos *VIPER* e *TATE* no genoma dos tripanosomatídeos

Realizamos uma busca pela presença de *VIPER* e *TATE* em todos os genomas apresentados na tabela 2. Utilizamos a opção tBLASTn por meio de BLAST local (Versão 2.2.26) contra os genomas baixados. Os *hits* dentro do *cut-off* de *e-value* de 10^{-10} foram considerados significativos. Utilizamos como *query* as sequências das três proteínas (*Gag-like*, YR e RT) das cópias canônicas de *VIPER* e *TATE* presentes no Repbase. Todas as cópias de todas as espécies foram recuperadas com auxílio de um código de programação Perl (anexo A) para as análises posteriores. Esse código foi escrito para recuperar uma região flangeadora *upstream* (do inglês: a montante) e *downstream* (do inglês: a jusante) de 8 kb, tamanho suficiente para recuperar as possíveis cópias completas dos elementos. Sequências redundantes foram excluídas por inspeção visual durante as análises. Para confirmar a identidade das sequências, foi utilizada a ferramenta online Repeat Masker presente no software CENSOR (KOHANY et al., 2006), que busca similaridade de sequências contra as sequências repetitivas presentes no Repbase.

As quinze cópias mais conservadas (menor *e-value* ou maior valor de *score* quando muitos *hits* apresentaram *e-value*=0.0) de cada espécie foram individualmente analisadas, para maior acurácia. Para analisar a presença das

ORFs foi utilizada a ferramenta ORFfinder do NCBI. Para as ORFs encontradas foram buscados os domínios, utilizando a sequência de aminoácidos predita, através da ferramenta CD-search (MARCHLER-BAUER; BRYANT, 2004). As ORFs codificadoras de YR e RT/RH foram identificadas pela similaridade com as cópias canônicas das proteínas de *VIPER* e *TATE* ou pelos domínios presentes. A ORF codificadora de *Gag-like* foi predita pela proximidade ou sobreposição com a ORF2, e também pela ausência de proteínas similares em outras espécies por análise de BLASTp no NCBI. Essa abordagem de identificação da proteína *Gag-like* foi estabelecida durante as primeiras análises quando foi observada uma grande divergência nessa proteína. A existência de proteínas similares à proteína da ORF predita em várias outras espécies de tripanosomatídeos indicaria que a proteína em questão não deve ser parte do elemento. A ausência de proteínas similares em outras espécies é um indicativo maior de que a proteína predita é uma *Gag-like*.

As cópias foram consideradas potencialmente codificadoras quando foram encontradas as três ORFs codificando proteínas com tamanhos significativos (> 300 aminoácidos (aa) para *Gag-like*, > 250 aa para YR, > 600 aa para RT/RH). As cópias foram consideradas degeneradas quando ao menos uma ORF apresentou tamanho menor do que o padronizado para as análises.

Com as sequências aminoacídicas codificadas por cada uma das ORFs preditas foi realizada uma busca de domínios conservados utilizando a ferramenta CD- search do NCBI (MARCHLER-BAUER; BRYANT, 2004).

As ORFs em potencial tiveram suas características anotadas (posição na sequência, tamanho, domínios) e suas aminoacídicas recuperadas em arquivos FASTA.

Para tentar achar os limites dos elementos presentes em cada espécie, as cópias mais conservadas foram alinhadas no programa no MUSCLE (EDGAR, 2004) disponível online ou disponível no software MEGA7 (KUMAR; STECHER; TAMURA, 2016) e visualizadas no GeneDoc 2.7 (NICHOLAS; NICHOLAS; DEERFIELD, 1997). Para verificar a existência de repetições terminais foi realizado BLASTn (NCBI), opção “Align two or more sequences” e “Somewhat similar sequences (blastn)”, utilizando a sequência de nucleotídeos a ser analisada como *query* e como *subject*.

3.4 Análises evolutivas de *TATE* e *VIPER*

Para entender a evolução dos elementos *VIPER* e *TATE* nos genomas dos cinetoplastídeos, inferimos a filogenia das proteínas YR e RT/RH separadamente. Utilizamos as sequências das proteínas do elemento DrNgaro (AAN71721.1) de *Danio rerio* como grupo externo.

Todas as ORFs de RT/RH e YR de tamanho significativo foram incluídas. As sequências de aminoácidos foram alinhadas no PSI-coffee, disponível no pacote online T-coffee (DI TOMMASO et al., 2011; FLODEN et al., 2016) e visualizadas no GeneDoc 2.7 (NICHOLAS; NICHOLAS; DEERFIELD, 1997) para edição manual e seleção de regiões de alinhamento confiáveis (verificadas com base no código de cores apresentado pelo programa de alinhamento).

Diferentes matrizes de alinhamento foram testadas para cada proteína (1 – apenas sequências de *VIPER*; 2 - apenas sequências de *TATE*; 3 – apenas sequências de *VIPER* e grupo externo; 4 – apenas sequências de *TATE* e grupo externo; 5 – apenas sequências de *VIPER* e *TATE*; 6 – sequências de *VIPER* e *TATE* e grupo externo). Esses diferentes alinhamentos foram importantes para verificar a capacidade de cada matriz em resolver os agrupamentos, já que, a análise conjunta de *VIPER* e *TATE* reduz a quantidade de sítios, pois um número maior de regiões fica com alinhamento não confiável.

As filogenias foram construídas pelo método de máxima verossimilhança (ML) pela ferramenta RAXML-HPC BlackBox (STAMATAKIS, 2014) implementada no portal CIPRES (MILLER; PFEIFFER; SCHWARTZ, 2010), usando o modelo de substituição de aminoácidos indicado pela opção “*Find Best DNA/Protein Models*” do software MEGA7 (KUMAR; STECHER; TAMURA, 2016). As árvores em formato *newick* obtidas foram visualizadas pela ferramenta online Evolview (HE et al., 2016). Também foram obtidas árvores pelo método de Neighbor-joining (NJ) com *bootstrap* de 1000 réplicas usando o modelo mais indicado, segundo o teste mencionado acima, que estava disponível para NJ (já que existe um número menor de modelos disponíveis para esse método). As árvores de NJ foram obtidas pelo Mega7 (KUMAR; STECHER; TAMURA, 2016).

Para análise de divergência intraespecífica das proteínas preditas de *VIPER* e *TATE*, as sequências foram alinhadas pela ferramenta Expresso, disponível no pacote online T-coffee (DI TOMMASO et al., 2011). Regiões de alinhamento com

baixa confiabilidade (com base no código de cores apresentado pelo programa) foram excluídas com uso do programa GeneDoc 2.7. Foram usadas as sequências das proteínas preditas para cada uma das ORFs identificadas, independentemente se faz parte de uma cópia do elemento completa ou degenerada.

As divergências entre sequências (nucleotídeos ou aminoácidos) foram obtidas pela estimativa de distância-p, com a opção “*pairwise deletion*” pelo software MEGA7 (KUMAR; STECHER; TAMURA, 2016). Essa distância é a proporção (p) dos sítios diferentes entre duas sequências, obtida pela divisão do número de diferenças de nucleotídeos ou aminoácidos pelo número total de sítios comparados. A divergência média foi obtida pela opção “*Compute overall mean distance*” e a divergência par a par foi obtida pela opção “*Compute pairwise distance*”.

Análises de sintenia foram realizadas em alguns casos para evidenciar se determinadas cópias de uma espécie estavam em um mesmo contexto cromossômico de cópias de outras espécies. Para isso, identificamos os genes localizados *upstream* e *downstream* da cópia analisada, e as sequências de aminoácidos desses genes foram usadas como *query* em BLAST local, opção tBLASTn, contra o genoma desejado. A sintenia é comprovada se os genes são encontrados no mesmo *contig/scaffold* que a cópia do elemento e em posição *upstream* e *downstream* conforme esperado.

Dois possíveis casos de TH foram testados dos elementos *TATE* e *VIPER* entre *Lep. pyrrhocoris* e *T. theileri*. Para isso, comparamos a identidade encontrada para as sequências de aminoácidos de *TATE* e *VIPER* entre as duas espécies com a identidade encontrada para 30 proteínas ortólogas desses organismos escolhidas aleatoriamente (Tabela 3).

TABELA 3: GENES ESCOLHIDOS PARA VERIFICAR POSSÍVEL CASO DE TH DE *TATE* E *VIPER*. O NOME E O ID DOS GENES DE *LEP. PYRRHOCORIS* SÃO INDICADOS.

Gene	ID
<i>G-actin binding protein</i>	LpyrH10_05_2460
<i>piwi</i>	LpyrH10_14_0470
<i>Vps35</i>	LpyrH10_14_0530
<i>endomembrane protein</i>	LpyrH10_20_0480
<i>metacaspase 5</i>	LpyrH10_01_3410
<i>Arp3</i>	LpyrH10_26_0310
<i>adenylate kinase</i>	LpyrH10_14_1580
<i>SLA/LP</i>	LpyrH10_16_1550
<i>polyamine transporter</i>	LpyrH10_01_6810
<i>j-binding protein</i>	LpyrH10_17_0040
<i>tRNA nucleotidyltransferase</i>	LpyrH10_24_1320

<i>alpha-glucosidase</i>	LpyrH10_13_0100
<i>SIR2RP1</i>	LpyrH10_01_8080
<i>AP1g</i>	LpyrH10_05_4290
<i>Nob1</i>	LpyrH10_06_2900
<i>COP1b</i>	LpyrH10_25_0440
<i>aquaporin</i>	LpyrH10_01_1410
<i>katanin</i>	LpyrH10_02_7750
<i>translation initiation factor IF-2</i>	LpyrH10_01_1240
<i>myosin 13</i>	LpyrH10_02_4330
<i>phosphoglycerate mutase</i>	LpyrH10_28_1120
<i>amino acid permease</i>	LpyrH10_01_6810
<i>mitochondrial DNA topoisomerase II</i>	LpyrH10_26_0420
<i>RBP38</i>	LpyrH10_03_2720
<i>Pno1</i>	LpyrH10_01_5870
<i>6-phosphogluconate dehydrogenase</i>	LpyrH10_01_5690
<i>Rab1</i>	LpyrH10_07_0990
<i>enolase</i>	LpyrH10_17_1430
<i>GTP-binding protein</i>	LpyrH10_03_4140
<i>eIF-6</i>	LpyrH10_09_1030

3.5 Evolução dos elementos da ordem DIRS

Para tentar entender a origem dos retrotransposons *VIPER* e *TATE*, inferimos a filogenia das proteínas RT/RH e YR separadamente, incluindo sequências dos elementos ou proteínas relacionadas.

Para a RT/RH utilizamos as regiões conservadas, denominadas de box 1 a 7, do domínio RT estabelecidas por XIONG e EICKBUSH (1990) e incluímos diversos retrotransposons dos diferentes grupos e sequências de *Hepadnavirus* e *Caulimovirus* (Tabela 4). Essas regiões foram cortadas das sequências originais e submetidas a um alinhamento progressivo múltiplo com ClustalW com matriz de proteína BLOSUM, com penalização de abertura de *gap* de 10 e de extensão de 0,2. Este alinhamento foi verificado em relação às posições homólogas atribuídas anteriormente dentro de cada um dos sete domínios RT e regiões variáveis entre domínios foram retiradas.

TABELA 4: DADOS DAS SEQUÊNCIAS DE RT USADAS PARA RECONSTRUÇÃO FILOGENÉTICA

Nome	Organismo	ID /NCBI ou Uniprot	Classificação
HBV	<i>Hepatitis B virus</i>	AAK19542.1	<i>Hepadnavirus</i>
WHV	<i>Woodchuck hepatitis virus 59</i>	P12899.1	<i>Hepadnavirus</i>
DHBV	<i>Duck hepatitis B virus</i>	P30028.1	<i>Hepadnavirus</i>
LvNgaro	<i>Lytechinus variegatus</i>	AC131494.1	Ngaro
DrNgaro	<i>Danio rerio</i>	AAN71721.1	Ngaro
MuLV	<i>Murine leukemia virus</i>	P03356.3	Retrovírus
HTLV1	<i>Human T-cell lymphotropic virus type 1</i>	P03362.3	Retrovírus
MMTV	<i>Mouse mammary tumor virus</i>	P03365.3	Retrovírus

MPMV	<i>Mason-Pfizer monkey virus</i>	P07572.1	Retrovírus
SIV	<i>Simian immunodeficiency virus</i>	P22382.2	Retrovírus
HIV2	<i>Human immunodeficiency virus type 2</i>	Q74120.3	Retrovírus
CAEV	<i>Caprine arthritis encephalitis virus</i>	P33459.1	Retrovírus
FMV	<i>Figwort mosaic virus</i>	NP_619548.1	<i>Caulimovirus</i>
CerV	<i>Carnation etched ring virus</i>	P05400.1	<i>Caulimovirus</i>
CaMV	<i>Cauliflower mosaic virus</i>	Q02964.1	<i>Caulimovirus</i>
17.6	<i>Drosophila melanogaster</i>	P04323.1	Ty3/gypsy
297	<i>Drosophila melanogaster</i>	P20825.1	Ty3/gypsy
Gypsy	<i>Drosophila melanogaster</i>	P10401.1	Ty3/gypsy
Ty3	<i>Saccharomyces cerevisiae</i>	CAA97115.1	Ty3/gypsy
Copia	<i>Drosophila melanogaster</i>	P04146.3	Ty1/copia
1731	<i>Drosophila melanogaster</i>	S00954	Ty1/copia
Ty1	<i>Saccharomyces cerevisiae</i>	P47100.3	Ty1/copia
TcPAT	<i>Crassostrea gigas</i>	Repbase	PAT
Kangaroo	<i>Volvox carteri f. nagariensis</i>	AAM94957.1	PAT
PAT	<i>Panagrellus redivivus</i>	Q26106	PAT
DIRS2	<i>Gasterosteus aculeatus</i>	Repbase	DIRS
DrDIRS1	<i>Danio rerio</i>	Repbase	DIRS
DIRS1	<i>Dictyostelium discoideum</i>	M11339	DIRS
PAO	<i>Bombyx mori</i>	Repbase	BEL
BEL	<i>Drosophila melanogaster</i>	U23420.1	BEL
Ingi	<i>Trypanosoma brucei</i>	M16068.1	Não-LTR
Jockey	<i>Drosophila melanogaster</i>	P21328.1	Não-LTR
SLACS	<i>Trypanosoma brucei</i>	CAA34931.1	Não-LTR

Para a filogenia da YR incluímos sequências de IN de fagos, Transposases e Resolvases de bactérias e Flipases de fungos (Tabela 5) e utilizamos duas regiões conservadas (box 1 e 2) entre essas proteínas indicada por alguns autores (LORENZI; ROBLEDO; LEVIN, 2006a; NUNES-DÜBY et al., 1998). O alinhamento foi realizado com os mesmos critérios descritos acima, exceto que a região intermediária foi mantida sempre que algum nível de conservação pudesse ser detectado.

TABELA 5: DADOS DAS SEQUÊNCIAS DE YR USADAS PARA RECONSTRUÇÃO FILOGENÉTICA

Nome	Organismo	ID/ NCBI ou Uniprot	Classificação
FLP_KLULA	<i>Kluyveromyces lactis</i>	P13783.1	Flipases
FLP_1_ZYGBI	<i>Zygosaccharomyces bisporus</i>	P13784.1	Flipases
FLP_2_YEAST	<i>Saccharomyces cerevisiae</i>	P03870.1	Flipases
YF72_HAEIN	<i>Haemophilus influenzae</i>	P46495.2	Transposase
RCI2_ECOLI	<i>Escherichia coli</i>	P16470.1	Transposase
XERC_ECOLI	<i>Escherichia coli</i>	P22885	Resolvase
XERD_ECOLI	<i>Escherichia coli</i>	P21891	Resolvase
XERD_HAEIN	<i>Haemophilus influenzae</i>	L42023.1	Resolvase

BP186	<i>Escherichia virus 186</i>	P06723.1	Fago
BPHP1	<i>Haemophilus phage</i>	P21422.1	Fago
BPP2	<i>Escherichia virus P2</i>	P36932.2	Fago
BP434	<i>Phage 434</i>	P27078.1	Fago
BPHK0	<i>Escherichia virus HK022</i>	P16407.1	Fago
BPMFR	<i>Mycobacterium phage FRAT1</i>	P25426.1	Fago
DIRS1	<i>Dictyostelium discoideum</i>	M11339.1	DIRS
PAT	<i>Panagrellus redivivus</i>	X60774.1	PAT
DrNgaro1	<i>Danio rerio</i>	AY152729.1	Ngaro
LvNgaro	<i>Lytechinus variegatus</i>	AC131494.1	Ngaro
DrDIRS1	<i>Danio rerio</i>	Repbase	DIRS
Cn1	<i>Cryptococcus neoformans</i>	AY248893	Crypton

As análises filogenéticas foram então realizadas para cada um desses conjuntos de dados utilizando a análise bayesiana (BA) implementada no MrBayes 3.2.6 (RONQUIST et al., 2012), sob um modelo misto. A cadeia de Markov Monte Carlo (MCMC) do BA foi realizada por 10.000.000 de gerações, amostrando árvores a cada mil gerações. Os 25% dos resultados iniciais, considerados “*burn-in*” foram descartados.

Além disso, uma análise de ML foi realizada pelo RaxML sob o algoritmo Bootstrap rápido (STAMATAKIS, 2014), seguindo o modelo evolutivo sugerido pela ferramenta “*Find Best DNA/Protein Models*”, conforme medido em MEGA7 (KUMAR; STECHER; TAMURA, 2016). Bootstrap de 1.000 réplicas foi realizado para acessar a confiabilidade dos ramos. As árvores filogenéticas resultantes da BA e ML foram visualizadas e editadas em FigTree 1.4.2 (RAMBAUT, 2008) e Dendroscope (HUSON; SCORNAVACCA, 2012), respectivamente. Em ambos os casos, as análises da RT foram enraizadas com o uso dos retrotransposons sem LTR *ingi*, *jockey* e *SLACS* como grupo externo. Alternativamente, dada a ausência de um bom grupo externo para as sequências YR, as árvores filogenéticas reconstruídas com este conjunto de dados foram enraizadas pelo ponto médio.

4 RESULTADOS

4.1 Análise estrutural do elemento *TATE* em *L. braziliensis*

Como ainda não existe uma caracterização estrutural do elemento *TATE* e por ele ter sido descrito inicialmente em *L. braziliensis*, escolhemos analisar esta espécie primeiramente por ferramentas online com a finalidade de entendermos melhor a estrutura do elemento antes das demais análises. O BLASTn realizado no TritypDB mostrou 37 *hits* dentro do *cut-off* definido para a linhagem escolhida.

Das 15 cópias analisadas, a maioria apresentou as três ORFs em uma disposição padrão, que está representada na figura 9 com os respectivos domínios encontrados. Para a ORF1 não foi encontrado nenhum domínio conservado, porém a chamamos *Gag-like* conforme convenção definida por outros autores. Já a ORF2 apresentou o domínio DNA_BRE_C (cd00397) correspondente à enzima YR. A ORF3 apresentou o domínio RT_ *super family* (cl02808) correspondente à enzima RT e o domínio Peptidase_A17, que inicialmente pensamos tratar-se de uma PR, enzima codificada por retrovírus e LTR-retrotransposons, responsável pelo processamento das poliproteínas, mas que nunca foi reportada para elementos DIRS. Porém, esse último domínio, apesar de descrito (pfam05380) como homólogo das proteinases aspárticas, é na verdade um domínio RNaseH e foi nomeado de forma imprecisa (MAJOREK et al., 2014). A ORF2 para algumas cópias apresentou também outros 2 domínios adicionais: SPOR (cl27087) e KAR9 (pfam08580) que são preditos para a mesma região na porção amino-terminal da proteína. SPOR é um domínio de ligação a peptidoglicanos presente em muitas proteínas bacterianas e contribuem para uma variedade de processos que envolvem a remodelação do peptidoglicano, especialmente durante a divisão celular (YAHASHIRI; JORGENSON; WEISS, 2017). KAR9 é um domínio presente na proteína KAR9 de *Saccharomyces cerevisiae*, requerida para a correta montagem do fuso mitótico durante a divisão celular (MILLER; ROSE, 1998). A presença desses dois domínios em retrotransposons nunca foi evidenciada e não é possível afirmar se essa região realmente apresenta alguma função relacionada.

Com base nos alinhamentos das cópias, observamos que os limites do elemento estão de acordo com o apresentado no Repbase. As cópias também foram submetidas a BLASTn, alinhando-as contra si mesmas, para verificar repetições

terminais invertidas ou diretas, porém nenhum tipo de repetição terminal foi encontrado em nenhuma das cópias analisadas do genoma de *L. braziliensis*. As cópias variaram de 6844 a 7053 pb, com uma identidade média de 90,7% (menor identidade: 84%; maior identidade 98%). Posteriormente, durante nossas análises, evidenciamos que *TATE* pode apresentar um padrão complexo de repetições longas (apresentadas no item 4.3) e, como havíamos cortado as sequências de acordo com a sequência consenso do Repbase, não encontramos nas primeiras análises.

FIGURA 9: ESTRUTURA TEÓRICA INICIALMENTE DEFINIDA PARA O ELEMENTO *TATE*



LEGENDA: Estrutura desenhada com base no consenso das análises das cópias. Em verde está representado os limites dos elementos sem repetições. As ORFs estão representadas em azul e dentro de cada ORF estão representados os domínios presentes: laranja - domínio DNA_BRE_C; azul escuro - domínio RT_like super Family; cinza - domínio peptidase_A17, que corresponde a um domínio do tipo RNaseH. Em preto estão representados os dois domínios adicionais. A estrutura básica está representada em escala e na mesma disposição encontrada nas análises *in silico*.

FONTE: A autora, 2018.

4.2 Análise da distribuição de *VIPER* e *TATE* nos tripanosomatídeos

Os resultados das análises realizadas no BLAST local, que indicam a presença das proteínas de cada elemento, estão sumarizados na tabela 6, onde consta o número de *hits* encontrados dentro do *cut-off* definido. A abordagem de tBLASTn foi empregada, pois permite a descoberta de sequências mais divergentes que não seriam recuperadas por BLASTn e também permite achar cópias degeneradas, o que seria dificultado pelo uso de BLASTp.

Quase todas as espécies analisadas apresentaram *hits* significativos para ao menos uma proteína de *VIPER* ou *TATE*, sendo que algumas espécies apresentaram *hits* para os dois elementos. De 45 espécies, apenas *L. tarenrolae*, *Lep. seymouri*, *Lotmaria passim*, *Phytomonas* sp. isolados EM1 e HART1 e *T. rangeli* não apresentaram *hits* significativos, o que pode indicar ausência desses elementos nestas espécies. Para cada uma das proteínas encontramos resultados distintos, sendo que para Gag-like notamos um menor número de *hits* do que para as proteínas YR e RT/RH, para ambos elementos.

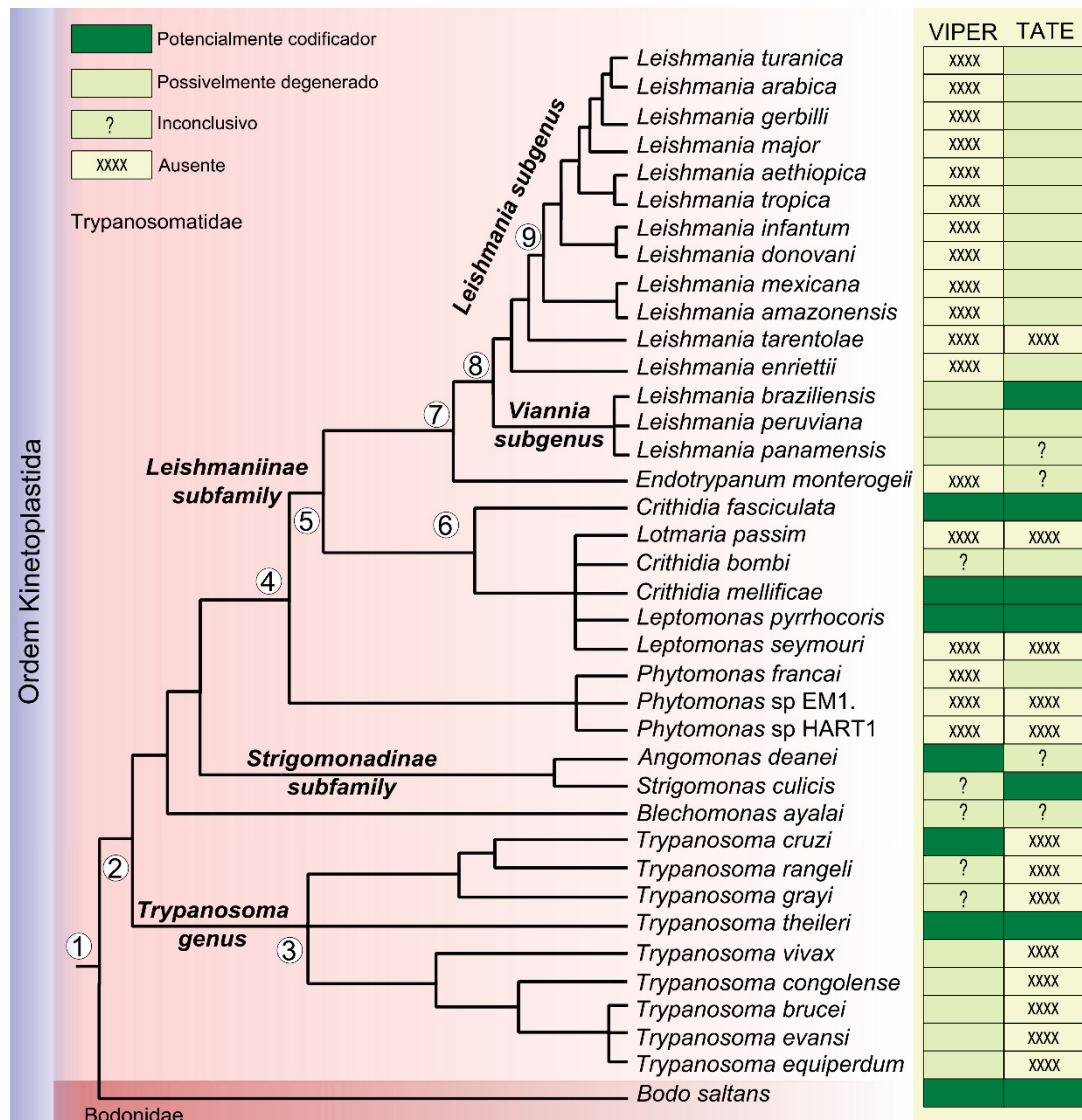
TABELA 6: NÚMERO TOTAL DE *HITS* ENCONTRADOS NA ANÁLISE DE TBLASTN

	VIPER			TATE		
	Gag-like	YR	RT/RH	Gag-like	YR	RT/RH
<i>Bodo saltans</i> Lake Konstanz	0	0	10	0	0	5
<i>Angomonas deanei</i> ATCC PRA-265	0	3	3	0	12	5
<i>Blechomonas ayalai</i> B08-376	0	0	1	0	77	36
<i>Crithidia fasciculata</i> strain Cf-CI	0	25	41	0	107	151
<i>C. bombi</i> IL132	0	2	12	0	36	19
<i>C. mellificae</i> ATCC 30862	0	13	33	0	9	8
<i>Endotrypanum monterogeii</i> strain LV88	0	0	0	12	25	45
<i>Leishmania aethiopica</i> L147	0	0	0	0	6	0
<i>L. amazonensis</i> MHOM/BR/71973/M2269	0	0	0	0	1	3
<i>L. arabica</i> strain LEM1108	0	0	0	0	1	3
<i>L. braziliensis</i> MHOM/BR/75/M2903	0	0	7	31	89	77
<i>L. braziliensis</i> MHOM/BR/75/M2904	0	0	11	34	60	50
<i>L. donovani</i> BPK282A1	0	0	0	0	4	2
<i>L. enriettii</i> strain LEM3045	0	0	0	0	67	45
<i>L. gerbilli</i> strain LEM452	0	0	0	0	4	6
<i>L. infantum</i> JPCM5	0	0	0	0	4	4
<i>L. major</i> strain Friedlin	0	0	0	0	3	1
<i>L. mexicana</i> MHOM/GT/2001/U1103	0	0	0	0	4	3
<i>L. panamensis</i> MHOM/COL/81/L13	0	0	3	12	21	17
<i>L. peruviana</i> PAB-4377_V1	0	0	10	27	109	94
<i>L. tarentolae</i> Parrot-Tarll	0	0	0	0	0	0
<i>L. tropica</i> L590	0	0	0	0	4	6
<i>L. turanica</i> strain LEM423	0	0	0	0	0	3
<i>Leishmania</i> sp. MAR LEM2494	0	0	1	0	17	18
<i>Leptomonas pyrrhocoris</i> H10	0	29	69	2	43	53
<i>Lep. seymouri</i> ATCC 30220	0	0	0	0	0	0
<i>Lotmaria passim</i> ATCC PRA-422	0	0	0	0	0	0
<i>Phytomonas francai</i>	0	0	0	0	7	3
<i>Phytomonas</i> sp. isolate EM1	0	0	0	0	0	0
<i>Phytomonas</i> sp. isolate Hart1	0	0	0	0	0	0
<i>Strigomonas culicis</i> TCC012E	0	0	1	0	31	10
<i>Trypanosoma brucei</i> brucei TREU927	0	36	58	0	0	0
<i>T. brucei</i> gambiense DAL972	0	17	24	0	0	0
<i>T. brucei</i> Lister strain 427	0	26	30	0	0	0
<i>T. congolense</i> IL3000	0	44	80	0	0	0
<i>T. cruzi</i> Dm28c PB1	752	617	1425	0	0	0
<i>T. cruzi</i> CL Brener Esmeraldo	272	254	578	0	0	0
<i>T. cruzi</i> CL Brener Non-Esmeraldo	321	264	663	0	0	0
<i>T. cruzi</i> marinkellei strain B7	369	767	671	0	0	0
<i>T. equiperdum</i> OVI V2	0	15	25	0	0	0
<i>T. evansi</i> strain STIB 805	0	17	28	0	0	0
<i>T. grayi</i> ANR4	0	72	159	0	0	0
<i>T. rangeli</i> SC58	0	0	0	0	0	0
<i>T. theileri</i> Edinburgh	7	39	76	0	19	17
<i>T. vivax</i> Y486	0	101	256	0	0	0

O primeiro passo para tentar entender essa distribuição e a evolução do *VIPER* e *TATE* consistiu em analisar a distribuição dos elementos em uma perspectiva filogenética. A figura 10 mostra as relações evolutivas entre as espécies de tripanosomatídeos utilizada neste trabalho. Essa filogenia foi desenhada por nós com base em diversos estudos considerando apenas as relações com alta confiabilidade (CORTEZ et al., 2006; DU; MASLOV; CHANG, 1994; FLEGONTOV et

al., 2013; HARKINS et al., 2016; KATO; CÁCERES; HASHIGUCHI, 2016; KELLY et al., 2014; LUKEŠ et al., 2014; RUNCKEL; DERISI; FLENNIKEN, 2014; SCHWARZ et al., 2015; STOCO et al., 2014; TEIXEIRA et al., 2011).

FIGURA 10: RELAÇÕES FILOGENÉTICAS CONSENSO DAS ESPÉCIES DE TRIPANSOMATÍDEOS USADAS NESSE TRABALHO E A DISTRIBUIÇÃO DOS ELEMENTOS VIPER E TATE.



LEGENDA: A filogenia dos tripanosomatídeos foi desenhada com base em diversos estudos. Espécies cujo posicionamento não é claro foram adicionadas como ramos politômicos. Números nos círculos representam o tempo estimado de divergência em milhões de anos atrás (mya) conforme inferido LUKES et al. (2014): 1 – 463-695 mya; 2 – 231-283 mya; 3 – 96-105 mya; 4 – 118-170 mya; 5 – 52-96 mya; 6 – 30-63 mya; 7 – 31-65 mya; 8 – 25-54 mya; 9 – 9-23 mya. As ramificações não estão desenhadas em escala. A presença de pelo menos uma cópia potencialmente codificadora do elemento na espécie é indicada como um retângulo verde escuro, a presença apenas de cópias degeneradas é representada por um retângulo verde claro, espécies onde não foi possível concluir sobre a conservação do elemento apresentam um ?, e a provável ausência completa do elemento é representada por XXXX.

FONTE: A autora, 2018.

Como a presença de *hits* significativos no BLAST não nos mostra se as cópias presentes são completas ou degeneradas, para um entendimento da conservação dos elementos, analisamos as 15 cópias consideradas mais conservadas de todas as espécies. Os anexos B e C mostram em detalhes as análises das ORFs de cada uma dessas cópias de *VIPER* e *TATE*, respectivamente. Para a maioria das espécies, os dois elementos parecem estar degenerados, porém não foi possível analisar a conservação de algumas cópias para algumas espécies, pois estas cópias estavam localizadas no final de *contigs/scaffolds* ou em regiões de dados faltantes. Assim, não é possível concluir sobre a conservação dessas cópias. Ao lado de cada espécie na figura 10 é indicada a ausência completa de cópias ou a presença de cópias potencialmente codificadoras, somente degeneradas ou cópias inconclusivas

Como já mencionado, poucas espécies apresentaram *hits* significativos para a ORF codificadora de *Gag-like*, tanto de *VIPER* quanto de *TATE*, o que poderia indicar que a maioria das cópias estão degeneradas. No entanto, algumas cópias dos elementos apresentaram uma ORF com tamanho significativo justaposta ou com sobreposição à ORF da YR, sugerindo que essa ORF estava fazendo parte do elemento, embora não tendo similaridade com a proteína *Gag-like* usada como *query*. A ausência de sequências similares em outras espécies corroborou com a sugestão de que essas sequências codificam uma proteína *Gag-like*. Isso mostra que essa proteína é bem menos conservada que a YR e RT/RH. Essa maior divergência da *Gag-like* poderia ser explicada por uma seleção relaxada ou por seleção positiva atuando sobre a proteína.

Em geral, notamos uma distribuição descontínua dos elementos entre as espécies. Essa descontinuidade pode ser resultado de perdas dos elementos ao longo da evolução, porém também poderiam ser explicadas por TH dos elementos entre as espécies se conectadas a outras evidências.

Nos tópicos a seguir descrevemos separadamente as cópias encontradas de *VIPER* e *TATE*, com ênfase para as cópias potencialmente codificadoras.

4.3 VIPER

VIPER parece ter sido completamente perdido na maioria das espécies de *Leishmania* analisadas, em *Endotrypanum montegerogei*, *Lo. passim*, *Lep. seymouri*, *Phytomonas* sp. Isolate EM1 e Hart1, *P. francai*.

Foram encontradas cópias degeneradas de *VIPER* em *L. braziliensis*, *L. panamensis*, *L. peruviana*, *T. brucei*, *T. congolense*, *T. equiperdum*, *T. evansi* e *T. vivax*. Apesar destes organismos mostrarem regiões de similaridade com *VIPER*, não foram encontradas ORFs potencialmente codificadoras e a localização dos *hits* em grandes *scaffolds* sugere que o elemento possa estar realmente degenerado nessas espécies. Em *L. braziliensis*, *L. panamensis* e *L. peruviana*, espécies proximamente relacionadas, um dos *hits* apresentados por cada corresponde ao gene domesticado derivado de *VIPER* já citado (LUDWIG; KRIEGER, 2016). A presença desse gene em *L. peruviana*, embora esperada, não havia sido reportada anteriormente.

Para alguns outros genomas que possuem cópias de *VIPER*, não foi possível concluir sobre a conservação do elemento. Por exemplo, o genoma de *T. grayi* apresenta diversos *hits* para *VIPER*, mas a maioria das cópias está localizada no final dos *contigs* ou em *contigs* curtos, perdendo a região da primeira ou primeira e segunda ORF. *Blechnomonas ayalai* apresentou apenas um *hit* significativo para a RT, e a sequência está em um *contig* curto. Embora, aparentemente as ORFs para Gag-like e YR estejam degeneradas, não podemos descartar a possibilidade de que outras cópias estejam presentes e foram excluídas na montagem do genoma. Este genoma ainda não está publicado e não temos informação sobre a cobertura da montagem utilizada.

Também não é possível concluir sobre conservação de *VIPER* em *Strigomonas culicis*, pois apesar de apresentar apenas uma cópia aparentemente degenerada, o genoma está montado em nível de *contig* (cobertura 23X) e é possível que cópias completas possam ter sido perdidas na montagem. Já para *C. bombi*, que apresentou vários *hits*, a maior parte está presente em *contigs* muito curtos. Assim, não é possível afirmar se essas espécies possuem sequências codificadoras de *VIPER*.

T. rangeli não apresentou *hits* significativos para *VIPER*, sugerindo perda do elemento. Porém, havia sido reportado a presença de resquícios de *VIPER* usando

um *cutoff* de e-value de 10^{-5} , menos estrigente que o que usamos (STOCO et al., 2014). Quando analisando os *hits* encontrados com esse *cutoff*, não foi possível confirmar a identidade das sequências como sendo *VIPER*, usando Repeat Masker. Dessa forma, é possível que esse elemento tenha sido completamente perdido, contudo é impossível concluir apenas com esses dados, pois a montagem do genoma está em nível de pequenos contigs.

Por outro lado, pelo menos uma cópia potencialmente codificadora de *VIPER* foi encontrada em *B. saltans*, *T. cruzi*, *T. theileri*, *Angomonas deanei*, *C. fasciculata*, *C. mellificae* e *Lep. pyrrhocoris*.

B. saltans apresenta 10 cópias de *VIPER* sendo que, quatro delas podem ser consideradas potencialmente codificadoras (Tabela 7), porém apenas uma apresenta todos os domínios esperados conservados. Surpreendentemente, as sequências apresentam uma alta divergência, impossibilitando o alinhamento das sequências nucleotídicas e o estabelecimento dos limites das cópias por alinhamento. As sequências de aminoácidos também apresentam alta divergência (Tabelas 8, 9 e 10, respectivamente para Gag-like, YR e RT/RH), sendo que as proteínas Gag-like preditas não apresentaram alinhamento confiável. Essa grande divergência sugere que essas sequências separaram uma da outra há muito tempo. A avaliação das repetições ficou comprometida para algumas das cópias, pois estão localizadas principalmente nas extremidades de *contigs/scaffolds*. Porém, em uma das cópias foi possível encontrar duas repetições diretas distintas, e essa cópia pode representar um elemento *VIPER* completo nessa espécie (Figura 11, Tabela 7). Esse padrão é semelhante ao encontrado para alguns elementos Ngara e PAT-like (Figura 7). Um fato intrigante é que não foi possível verificar a conservação do domínio da RT nessa sequência.

TABELA 7: RESUMO DAS CARACTERÍSTICAS DAS CÓPIAS DE VIPER POTENCIALMENTE CODIFICADORAS

<i>Scaffold/contig_posição</i>	ORF1	ORF2	ORF3	Repetições
	Posição Domínios	Posição Domínios	Posição Domínios	Estrutura básica, tamanho e identidade
<i>B. saltans</i>				
CYKH01000025.1_4008-9037	1-1335	1136-2419	1615-4173	5' – ausente
	N	DNA_BRE_C	RT_like	3' → A1 → B1 → A2 → B2 A: 285 pb, 96%;

			RNase	B: 198 pb, 97%
CYKH01002013.1_423-4623	1-1206	1281-2189	1031-3778	5' → A1 → A2
	N	DNA_BRE_C	RNase	3' – dados faltantes A: 83 pb, 93%
CYKH01001028.1_613-5045	1-1203	1079-2095	1288-3819	5' – ausente
	N	DNA_BRE_C	N	3' – ausente
CYKH01000216.1_959-6047	1-1236	1600-2472	1503-4130	5' → A1
	N	DNA_BRE_C	RNase	3' → B1 → B2 → A2 A: 83 pb, 100%; B: 57 pb, 100%
<i>T. cruzi</i>*				
MBSY01000735.1_110090-114144	1-1341	1342-2359	1218-2163	5' → A1
	CwIO1	DNA_BRE_C	RT_like RNase	3' → B1 → A2 → B2 A: 173 pb, 100%; B: 219 pb, 100%
<i>T. theileri</i>				
NBCO01000001.1_190527_200565	1-1266	1399-2359	1577-3551	5' – ausente
	N	N	RT-like	3' → A1 → A2 A: 95 pb, 100%
<i>A. deanei</i>				
KV452601.1_66_3952	1-1281	1308-2220	1898-3821	5' – dados faltantes
	SMC_N	N	RVT_1 RNase	3' – dados faltantes
<i>C. fasciculata</i>				
CfaC1_21_76128-86740	1-1182	2057-2888	1120-4573	5' → A1
	SMC_N	DNA_BRE_C	RT-like RNase	3' → B1 → A2 → B2 A: 136 pb, 99%; B: 288 pb, 100%
CfaC1_27_1306980-1317400	1-1251	1356-2184	1247-3782	5' → A1
	N	N	Herpes_ICP4 RT-like RNase	3' → B1 → A2 → B2 A: 144 pb, 100%; B: 308 pb, 100%
CfaC1_29_2163302-2173637	1-1227	939-1860	1238-3641	5' – ausente
	N	N	RT-like RNase	3' – ausente
CfaC1_21_237095-247550	1-1416	2387-3302	1917-4911	5' → A1
	GRS1	N	RT-like RNase	3' → B1 → A2 → B2 A: 414 pb, 93 %; B: 294 pb, 99%
<i>Lep. pyrrhocoris</i>				

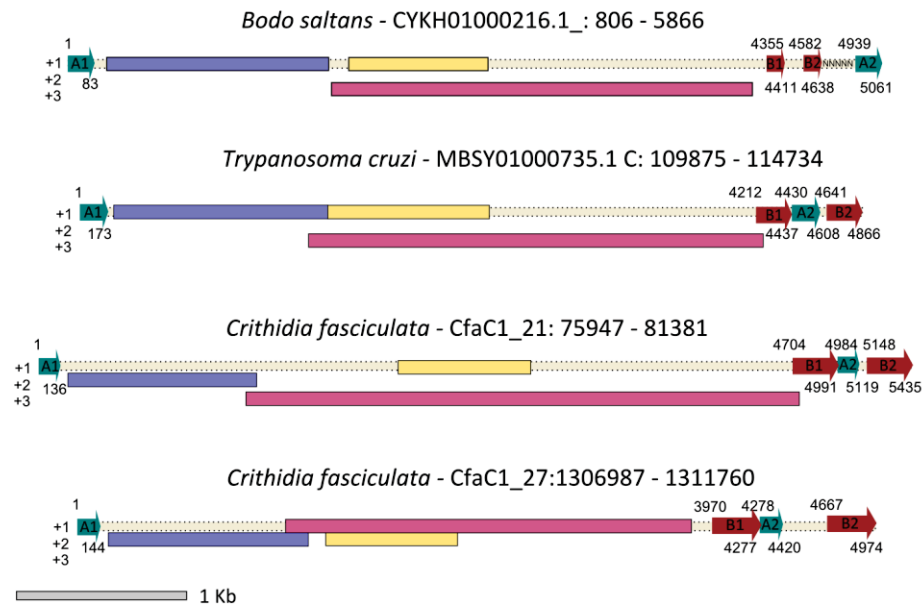
LpyrH10_23_286594-293282	1-1191 N	1187-2447 N	1105-4090 RT- <i>like</i> RNase	5' – ausente 3' – ausente
LpyrH10_04_536700-547229	1-1689 N	1685-2945 N	1603-4588 RVT_1 RNase	5' – ausente 3' – ausente
<i>C. mellificae</i>				
MDUD01008257.1_1-4861	1-1695 SMC_N Atrophin_1	1442-2966 N	3112-4621 RT- <i>like</i> RNase	5' – dados faltantes 3' – ausente

Nota: A posição no *scaffold/contig* refere-se à posição da cópia no *scaffold* (primeira posição da ORF 1 até a última posição da ORF3). A posição das ORFs refere-se à posição destas dentro da cópia, iniciando pela primeira posição da ORF1. Repetições distintas estão representadas por letras diferentes (A e B). As posições (5' ou 3') dos fragmentos repetitivos estão indicadas e os fragmentos estão representados em ordem). O tamanho e identidade de cada repetição são indicados. Os domínios presentes são indicados: N – nenhum domínio predito;

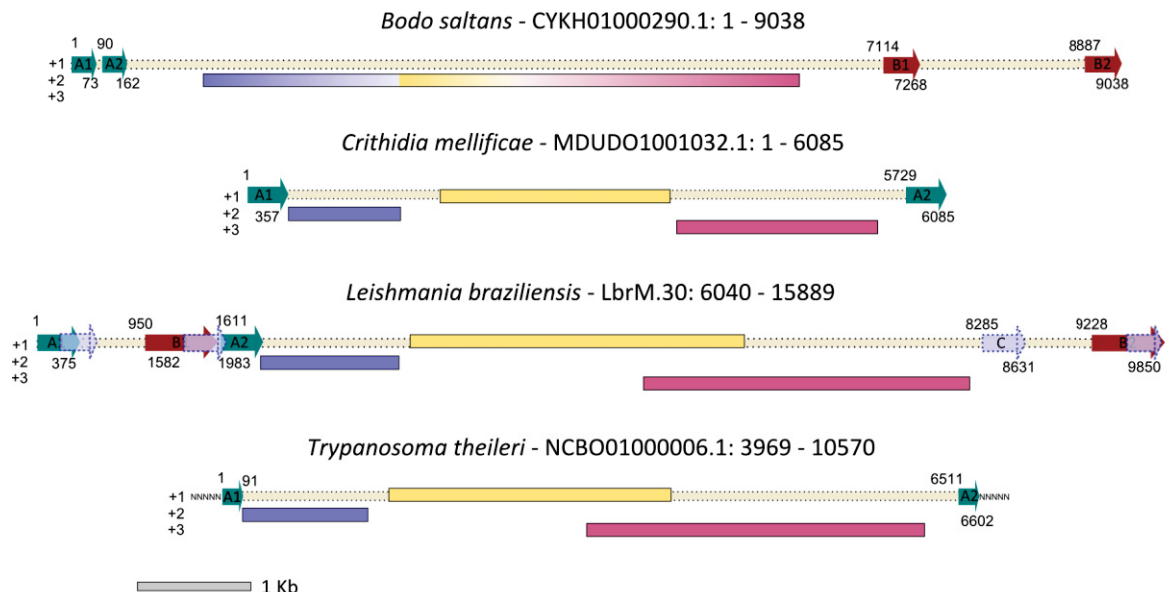
* somente uma cópia de *T. cruzi* foi representada, já que existe uma grande conservação entre as cópias. Informações adicionais das cópias são encontradas no anexo B.

FIGURA 11: ESQUEMA DA ESTRUTURA DOS ELEMENTOS *VIPER* E *TATE* COMPLETOS DE TRIPANOSOMATÍDEOS

VIPER



TATE



LEGENDA: Representação esquemática dos elementos *VIPER* e *TATE* possivelmente completos e potencialmente codificadores encontrados no nosso trabalho. Retângulos de diferentes cores representam as ORFs gag-like (azul), YR (amarelo), e RT/RH (rosa) e foram desenhadas conforme a fase de leitura (+1, +2 ou +3). As setas indicam as repetições encontradas. O scaffold e posição da cópia são indicados ao lado do nome da espécie. Somente duas das três cópias de *C. fasciculata* foram desenhadas. A figura está desenhada em escala, sendo distinta para *VIPER* e *TATE* (repetição A de *VIPER* de *B. saltans*, tamanho foi duplicado; repetição A *TATE* de *B. saltans*, tamanho foi triplicado; repetição de *TATE* de *T. theileri*, tamanho foi duplicado). Regiões de dados faltantes foram representadas por NNNN.

FONTE: A autora, 2018.

TABELA 8: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA GAG-LIKE DE VIPER

Espécie (cópias, aa)	Gag-like		
	% divergência		
	Média	Mínima	Máxima
<i>B. saltans</i> (5, 292 aa)		na	
<i>T. cruzi</i> (13, 492 aa)	2,1	0	4,7
<i>T. theileri</i> (3, 240 aa)	67,9	49,2	78,7
<i>A. deanei</i> (2, 210 aa)	76,7	76,7	76,7
<i>C. fasciculata</i> (5, 420 aa)	69	1,5	84,8
<i>Lep. pyrrhocoris</i> (2, 396 aa)	0,3	0,3	0,3
<i>C. mellificae</i> (3, 700 aa)		na	

Nota: Análise de divergência realizada utilizando distância-p. A similaridade foi estimada apenas para as regiões que apresentaram alinhamento confiável. O número de cópias da ORF para cada espécie e o tamanho do alinhamento considerado, em aminoácidos (aa), são mostrados entre parênteses. na - não analisada devido à ausência de alinhamento confiável.

TABELA 9: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA YR DE VIPER

Espécie (cópias, aa)	YR		
	% divergência		
	Média	Mínima	Máxima
<i>B. saltans</i> (5, 150 aa)	67,3	41,3	73,2
<i>T. cruzi</i> (15, 338 aa)	1,8	0	3,6
<i>T. theileri</i> (2, 272 aa)	52,9	52,9	52,9
<i>A. deanei</i> (2, 283 aa)	51,6	51,6	51,6
<i>C. fasciculata</i> (4, 398 aa)	59,2	40,6	69,9
<i>Lep. pyrrhocoris</i> (13, 430 aa)	38,4	0	61,7
<i>C. mellificae</i> (3, 507 aa)	0	0	0

Nota: Análise de divergência realizada utilizando distância-p. A similaridade foi estimada apenas para as regiões que apresentaram alinhamento confiável. O número de cópias da ORF para cada espécie e o tamanho do alinhamento considerado, em aminoácidos (aa), são mostrados entre parênteses.

TABELA 10: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA RT/RH DE VIPER

Espécie (cópias, aa)	RT/RH		
	% divergência		
	Média	Mínima	Máxima
<i>B. saltans</i> (7, 294 aa)	67	28	77
<i>T. cruzi</i> (14, 948 aa)	3,2	0	7
<i>T. theileri</i> (5, 701 aa)	42,4	21	56,6
<i>A. deanei</i> (3, 611 aa)	73,3	68,6	78,3
<i>C. fasciculata</i> (5, 621 aa)	59,2	40,6	69,9
<i>Lep. pyrrhocoris</i> (11, 1091 aa)	27,9	0	63,8
<i>C. mellificae</i> (2, 393 aa)	57	57	57

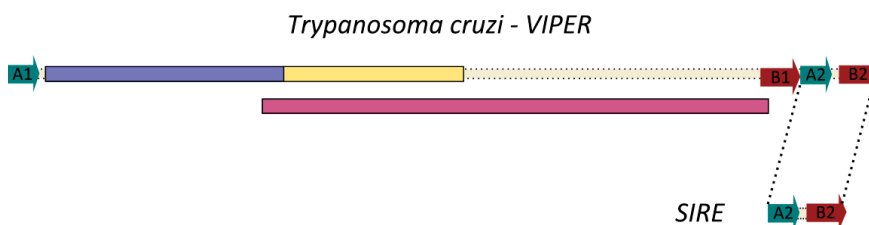
Nota: A similaridade foi estimada apenas para as regiões que apresentaram alinhamento confiável. O número de cópias da ORF para cada espécie e o tamanho do alinhamento em aminoácidos (aa) considerado são mostrados entre parênteses.

Cópias codificadoras de *VIPER* também foram encontradas em duas espécies de *Trypanosoma*. Avaliamos quatro diferentes linhagens de *T. cruzi* e apenas a linhagem Dm28c apresentou possíveis cópias codificadoras. Na linhagem CL Brener já havia sido reportada a ausência de cópias completas (LORENZI; ROBLEDO; LEVIN, 2006a). O alinhamento de nucleotídeos de cópias conservadas permitiu isolar claramente os limites do elemento. Observamos algumas cópias mais divergentes e as cópias mais conservadas apresentaram 4859 pb. As buscas por sequências repetitivas mostraram a presença de repetições no *VIPER* dessa espécie (Figura 11). Até o momento, essas repetições de *VIPER* não haviam sido identificadas. A divergência média entre 13 cópias mais conservadas de *VIPER* é de 5%, indicando grande conservação. A divergência encontrada para as sequências de aminoácidos preditas para cada ORF também é baixa (Tabelas 8, 9 e 10). Interessantemente, encontramos um domínio CwIO1 super family em quase todas as cópias da Gag-like.

Com a identificação das repetições de *VIPER* de *T. cruzi* é possível observar que o elemento *SIRE*, que apresenta centenas de cópias no genoma dessa espécie, corresponde às repetições A2B2 de *VIPER* (Figura 12). Esse achado abre a discussão sobre a mobilidade do elemento *SIRE*. Uma das hipóteses era que esse elemento curto e não autônomo tivesse sido gerado por deleção interna do elemento e fosse mobilizado pelas enzimas fornecidas pelo elemento *VIPER* (LORENZI; ROBLEDO; LEVIN, 2006a), já que eles possuíam a mesma região flanqueadora.

Como esse novo cenário, não podemos afirmar se as cópias de *VIPER* podem ser mobilizadas, ou se são apenas resquícios da presença do elemento completo.

FIGURA 12: REPRESENTAÇÃO ESQUEMÁTICA DE *VIPER* COMPLETO E *SIRE* DE *T. CRUZI*



LEGENDA: Representação esquemática do elemento *VIPER* completo de *T. cruzi* e comparação com o elemento *SIRE* da mesma espécie. Retângulos de diferentes cores representam as ORFs gag-like (azul), YR (amarelo), e RT/RH (rosa). As setas representam as repetições.

FONTE: A autora, 2018.

Em *T. theileri*, três cópias são mais conservadas, porém apenas uma está completa, segundo nosso critério. Nessa cópia, encontramos apenas uma repetição (A) na região 3' de 95 pb (100% identidade). A outra repetição esperada (B) não foi encontrada. As divergências das proteínas de *VIPER* nessa espécie são altas (Tabelas 9, 10 e 11) indicando que as cópias divergiram há bastante tempo. De acordo com essa divergência, o alinhamento das sequências de nucleotídeos revelou pouca similaridade entre as cópias, não sendo possível estabelecer os limites do elemento.

A. deanei apresenta quatro cópias e pelo menos uma das cópias apresenta as três ORFs potencialmente codificadoras (Tabela 7), porém, não detectamos o domínio característico da YR na proteína predita para a ORF 2 e não foi possível avaliar a presença de repetições, pois a região codificadora dessa cópia perfaz praticamente todo o *contig*. A avaliação da divergência das proteínas codificadas mostrou uma alta divergência entre as cópias (Tabelas 8, 9 e 10).

Um dos *hits* obtidos para a ORF3 de *A. deanei* representa um gene (*scaffold* KV452459.1: 737275- 739732) de 2457 pb codificando uma proteína de 818 aminoácidos com similaridade com a região de 260-788 da proteína RT/RH de *VIPER*. Os domínios da RT e RH não estão conservados e é encontrado um domínio N-terminal Atrophin-1 em uma região sem similaridade com *VIPER*. Ainda, curtas regiões intergênicas separam essa ORF dos genes *upstream* e *downstream*, não apresentando qualquer remanescente das ORFs de Gag-like e YR. Essa estrutura gênica é semelhante a que foi encontrada por nosso grupo como sendo um gene derivado de *VIPER* em algumas espécies, evento de domesticação do

elemento mencionado em 1.2.8 (LUDWIG; KRIEGER, 2016). A presença do gene domesticado não foi reportada em *A. deanei*, porém, como nesse trabalho estamos analisando uma montagem de genoma de *A. deanei* distinta da usada no trabalho anterior, existe a possibilidade de que esse gene seja ortólogo do reportado e represente o mesmo caso de domesticação. Para testar essa hipótese alinhamos a sequência de aminoácidos desse gene de *A. deanei* com as sequências de proteínas do gene domesticado das outras espécies e observamos uma baixa similaridade (25%), praticamente restrita a região carboxi (Anexo D). Além disso, esse gene de *A. deanei* não se encontra em região sintênica em relação ao gene domesticado de *VIPER* das outras espécies. Assim, não temos indícios de se tratar do mesmo evento de domesticação. Esse gene pode representar um evento de domesticação independente (caso o gene seja ativo e tenha alguma função em *A. deanei*) ou é apenas uma cópia de *VIPER* degenerada que ainda mantém essa ORF.

C. fasciculata parece apresentar várias cópias de *VIPER* (aproximadamente 30) e possui pelo menos quatro cópias potencialmente codificadoras, e em três são encontrados padrões de repetição terminais (Tabela 7, Figura 11). A análise da divergência das proteínas codificadas pelas cópias e da sequência de nucleotídeos dessas cópias indicam que as elas são bastante divergentes.

C. mellificae apresentou vários *hits* correspondendo a aproximadamente 30 cópias e pelo menos uma cópia é codificadora (Tabela 7). Não foi possível identificar regiões de repetição, porém a região codificadora perfaz praticamente todo o *contig* e não é possível saber se possíveis repetições estariam na parte dos dados faltantes.

Lep. phyrocoris tem duas cópias potencialmente codificadoras de *VIPER* (Tabela 7). Outras várias cópias apresentaram padrão de ORF2 e ORF3 semelhantes, mas com ORF1 aparentemente degenerada. Alinhamos as sequências de nucleotídeos de 13 cópias para encontrar limites para o elemento nessa espécie. Interessantemente, pode-se observar a formação de dois grupos de sequências claramente distintos de *VIPER*. Nenhum dos tipos de *VIPER*, porém parece apresentar repetições terminais.

Entre os motivos MHR e CH-box (CX₂CX₄HX₄C) reportados na sequência de *VIPER* de *T. cruzi* por LORENZI et al (2016), encontramos apenas o CH-box (figura 13), nas proteínas Gag-like de *T. cruzi*, *T. theileri* e *L. pyrrhocoris*, porém em

nenhuma foi possível encontrar o motivo MHR semelhante ao reportado. Nas demais espécies não foi possível encontrar nenhum dos motivos.

FIGURA 13: MOTIVO CH-BOX ENCONTRADO EM SEQUÊNCIAS DE GAG-LIKE DE VIPER

MBSY010001	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010007	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010006	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010002	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010005	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010006	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010005	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010003	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010007	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010000	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	426															
MBSY010006	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	492															
MBSY010005	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
MBSY010006	:	--AQLFVDRRETQ--R----	TVFCQ	Q	R	C	H	I	F	G	H	E	A	S	T	C	N	V	R	P	R	-----	GNYSQ--YR	R	P	S	K	N	G	R	G	A	A	R	R	--S	:	446															
LpyrH10_23	:	LAQRPRVRSRTPS--RTAGPGI	F	C	Y	Y	G	I	A	G	H	V	S	T	E	C	R	K	K	L	S	-----	DQK--KTEK	G	H	A	--S	E	A	A	--T	---	R	:	396																		
NBCO010000	:	--AKPLVEKEKGT--K----	S	I	F	C	L	R	C	K	T	E	G	R	T	A	A	T	C	N	V	R	L	P	A	F	N	K	P	R	F	N	T	H	N	--H	P	R	Y	S	K	N	V	R	G	A	V	A	L	Q	---	:	301
NBCO010000	:	--AFKLQKKKSAPPPS----	K	A	T	C	Y	R	C	G	R	I	G	H	Y	S	T	C	R	V	Q	L	Q	---	K	E	A	Y	K	P	K	N	--G	F	R	E	R	K	-----	Q	---	:	421										
NBCO010000	:	--AAKIKRKHSAAPPPS----	K	A	F	C	E	R	G	R	V	G	H	Y	S	T	T	C	R	V	N	L	Q	---	K	E	T	Y	K	P	K	N	--D	Y	R	E	R	K	-----	Q	---	:	403										

LEGENDA: Alinhamento da região da proteína Gag-like contendo o motivo CH-box (quadro vermelho) com indicação dos aminoácidos chave (CX₂CX₄HX₄C) por *. MBSY: *T. cruzi* Dm28c; LpyrH10: *Lep. pyrrocoris*; NBCO: *T. theileri*.

FONTE: A autora, 2018.

4.4 TATE

Como evidente na tabela 6, pelas análises conseguimos perceber que *TATE* parece ter sido completamente perdido em *Phytomonas* sp. isolados EM1 e Hart1, nas espécies de *Trypanosoma* (exceto *T. theileri*), *Lo. passim*, *Lep. seymouri* e *L. tarentolae*.

Foram encontradas cópias degeneradas de *TATE* nas espécies *L. aethiopica*, *L. amazonensis*, *L. arabica*, *L. donavani*, *L. enrietti*, *L. gerbilli*, *L. infantum*, *L. major*, *L. mexicana*, *L. peruviana*, *C. bombi*, *L. tropica*, *L. turanica* e *P. francai*.

Resultados inconclusivos foram encontrados para *A. deanei*, *Bl. ayalai* e *monterogeii*. Aparentemente não há cópias completas potencialmente codificadoras nessas espécies, mas o nível de montagem desses genomas dificultou as análises, uma vez que elas estavam presentes em *contigs* curtos, em regiões de dados faltantes ou no final dos *contigs*, desta forma não foi possível concluir sobre a conservação do elemento nessas espécies.

Cópias potencialmente completas de *TATE*, apresentando as três ORFs potencialmente codificadoras, foram encontradas em *B. saltans*, *C. fasciculata*, *C. mellificae*, *S. culicis*, *L. braziliensis*, *Lep. pyrrhocoris* e *T. theileri*.

B. saltans apresenta duas cópias potencialmente codificadoras para *TATE*. Contrariamente ao que é encontrado em *L. braziliensis*, para as duas cópias, uma única ORF foi encontrada. O domínio DNA_BRE_C foi encontrado em ambas as cópias, o domínio RVT_1_super family parece estar incompleto e o domínio RH ausente (Tabela 11). A divergência das proteínas completas codificadas por essas cópias é de 44%. O alinhamento de nucleotídeos entre as cópias não permitiu verificar os limites do elemento, mas para uma das cópias foi encontrado um padrão de repetição (Figura 11). Apesar de possivelmente ser uma repetição do *TATE*, a distância encontrada entre o final da ORF e uma das repetições é bastante grande e inesperada. Dessa forma, não é possível afirmar se essa repetição realmente faz parte do elemento.

Diferentemente das outras espécies do gênero, *T. theileri* apresentou *hits* significativos para *TATE*. *T. theileri* possui uma cópia com as três ORFs potencialmente codificadoras, porém o domínio RH é ausente. Problemas na montagem do genoma podem ter impedido de encontrarmos adicionais cópias completas, uma vez que a maioria das cópias analisadas está em regiões de dados faltantes. Inclusive, encontramos uma repetição (A1/A2) cujas sequências estão exatamente flanqueadas por dados faltantes (Tabela 11, figura 11). Ou seja, a existência de outra repetição não foi possível ser avaliada. A análise de divergência de aminoácidos mostrou uma maior conservação da proteína RT/RH entre as cópias (Tabelas 12, 13 e 14).

TABELA 11: RESUMO DAS CARACTERÍSTICAS DAS CÓPIAS DE TATE POTENCIALMENTE CODIFICADORAS

Scaffold/contig_posição	ORF1	ORF2	ORF3	Repetições
	Posição Domínio	Posição Domínio	Posição Domínio	Estrutura básica, tamanho e identidade
<i>B. saltans</i>				
CYKH01000290.1_1157-6382		1-5226		5' → A1 → A2
				3' → B1 → B2
		DNA_BRE_C RVT_1		A: 73 pb, 99%; B: 161 pb, 81%

CYKH01000240.1_371-4765		1-4395		5' – ausente 3' – ausente
		DNA_BRE_C		
		RVT_1		
<i>C. fasciculata</i>				
CfaC1_16_593910-598902	1-1225	1269-3471	2960- 4994	5' – ausente 3' – ausente
		DNA_BRE_C		
			RT_like	
			DNA_pol3	
<i>C. mellificae</i>				
MDUD01001032.1_377-5156	1-987	1338-3360	3389- 5156	5' →A1 3' →A2
		DNA_BRE_C		A: 357 pb, 100%
			RT_like	B: dados ausentes
			RNase	
<i>S. culicis</i>				
AUXH01000761.1_1187-6266	1-840	933-3987	3566- 6266	5' →A1 3' →A2
		Atrophin-1		A: 346 pb, 99%
			RT_like	
<i>Lep. pyrrhocoris</i>				
LpyrH10_03_995010-1005122	1-1095	674-3938	3151- 5869	5' – ausente 3' – ausente
		DNA_BRE_C		
LpyrH10_45_4163-13642	1-1200	1623-3486	2732- 5318	5' – ausente 3' – ausente
		DNA_BRE_C		
			RT_like	
			dnaA	
LpyrH10_10_418649-428385	1-1014	969-3291	3317- 5138	5' – ausente 3' – ausente
		DNA_BRE_C		
			RT_like	
<i>L. braziliensis</i>				
LbrM.30_1128659-1142875	1-1218	1311-4245	3356- 6218	5' →A1→B1→A2 3' →C→B2
		DNA_BRE_C		A: 375 pb, 98%
			RT_like	B: 637 pb, 94%
			RNase	C: 293 pb, 98%

LbrM.20.2_676198-690033	1-1218	1311-4260	3371-5837	5' ←A1 3' ←A2 A: 71 pb, 100%
		DNA_BRE_C SPOR	RT_like RNase GcrA	
LbrM.34_19966225-2010719	1-3642		2754-5616	5' →A1 3' →A2 A: 382 pb, 89%
		DNA_BRE_C SPOR	RT_like RNase	
<i>T. theileri</i>				
NBCO01000006.1_199922-210051	1-1104	1281-3762	3014-5981	5' →A1 3' →A2 A: 91 pb, 100%
		DNA_BRE_C	RT_like	

Nota: A posição no *scaffold/contig* refere-se à posição da cópia (primeira posição da ORF 1 até a última posição da ORF3). A posição das ORFs refere-se à posição destas dentro da cópia, iniciando pela primeira posição da ORF1. Repetições distintas estão representadas por letras diferentes (A e B). As posições (5' ou 3') dos fragmentos repetitivos estão indicadas e os fragmentos estão representados em ordem). O tamanho e identidade de cada repetição são indicados.

TABELA 12: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA GAG-LIKE DE TATE

Espécie (cópias, aa)	Gag-like		
	% divergência		
	Média	Mínima	Máxima
<i>L. braziliensis</i> M2903 (2, 397 aa)	17,4	17,4	17,4
<i>L. braziliensis</i> M2904 (8, 450 aa)	26,7	0	50,4
<i>L. panamensis</i> (2, 404 aa)	44,9	44,9	44,9
<i>L. peruviana</i> (2, 257 aa)	45,9	45,9	45,9
<i>C. mellificae</i> (2, 254 aa)	74,4	74,4	74,4
<i>Lep. Pyrrhocoris</i> (5, 194 aa)	70,3	53,6	77,1
<i>T. theileri</i> (3, 378 aa)	61,8	35,7	75,4

Nota: A similaridade foi estimada apenas para as regiões que apresentaram alinhamento confiável. O número de cópias da ORF para cada espécie e o tamanho do alinhamento em aminoácidos (aa) considerado são mostrados entre parênteses.

TABELA 13: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA YR DE TATE

Espécie (cópias, aa)	YR		
	% divergência		
	Média	Mínima	Máxima

<i>C. fasciculata</i> (7, 727 aa)	55,2	2,1	74
<i>C. mellificae</i> (3, 434 aa)	63,7	53,6	69,5
<i>L. braziliensis</i> M2903 (5, 432 aa)	3,3	0	4,6
<i>L. braziliensis</i> M2904 (15, 982 aa)	4,4	0	12,2
<i>L. enrieti</i> (3, 250 aa)	35	27,4	39,1
<i>Lep. pyrrhocoris</i> (10, 468 aa)	51,1	0	64,2
<i>T. theileri</i> (3, 584 aa)	54,7	35,9	65,8

Nota: A similaridade foi estimada apenas para as regiões que apresentaram alinhamento confiável. O número de cópias da ORF para cada espécie e o tamanho do alinhamento em aminoácidos (aa) considerado são mostrados entre parênteses.

TABELA 14: DIVERGÊNCIA INTRAESPECÍFICA MÉDIA, MÍNIMA E MÁXIMA APRESENTADA PELA PROTEÍNA RT/RH DE TATE

Espécie (cópias, aa)	RT/RH		
	% divergência		
	Média	Mínima	Máxima
<i>C. fasciculata</i> (3, 473 aa)	46,4	45,6	46,8
<i>L. braziliensis</i> M2904 (14, 954 aa)	4	0	10,6
<i>Lep. Pyrrhocoris</i> (8, 882 aa)	44,4	0	55,9
<i>T. theileri</i> (4, 698 aa)	13,1	0	17,5

Nota: A similaridade foi estimada apenas para as regiões que apresentaram alinhamento confiável. O número de cópias da ORF para cada espécie e o tamanho do alinhamento em aminoácidos (aa) considerado são mostrados entre parênteses.

Em *S. culicis*, a maioria dos *hits* está em regiões de dados faltantes e fica impossível analisar a conservação do elemento. Porém, uma das cópias apresentou todas as ORFs codificadoras e um padrão de repetição (Tabela 11). O domínio DNA_BRE_C não está presente, mas foi verificado na ORF 2 a presença do domínio Atrophin-1. O domínio correspondente a RH também não está presente.

C. fasciculata apresenta pelo menos uma cópia (Tabela 11) que possui as três ORFs potencialmente codificadoras com os domínios DNA_BRE_C e RT_like. O domínio da RH não foi encontrado, porém foi encontrado o domínio DNA_pol3_gamma3 super family mas com valor *e-value* menos significativo na região amino da proteína. A cópia completa de *C. fasciculata* não apresentou padrão de repetições terminais. Várias cópias apresentando apenas uma ou duas ORFs conservadas foram encontradas, sendo que a menor divergência proteica foi vista para YR.

Entre as cópias encontradas em *C. mellificae* verificamos que ao menos uma delas está completa e é potencialmente codificadora (Tabela 11, figura 11),

possuindo também os domínios conservados correspondentes a YR e RT/RH, e apresentando um padrão de repetição direta. Apesar dessa cópia completa e de uma das repetições ter sido encontrada, dados faltantes na região 3' da cópia impossibilitaram a avaliação da presença ou ausência de mais repetições. Dados faltantes na região da terceira ORF também deixaram inconclusiva a análise para outras cópias, que apresentam duas ORFs codificadoras, comprometendo assim a avaliação de outras possíveis cópias completas.

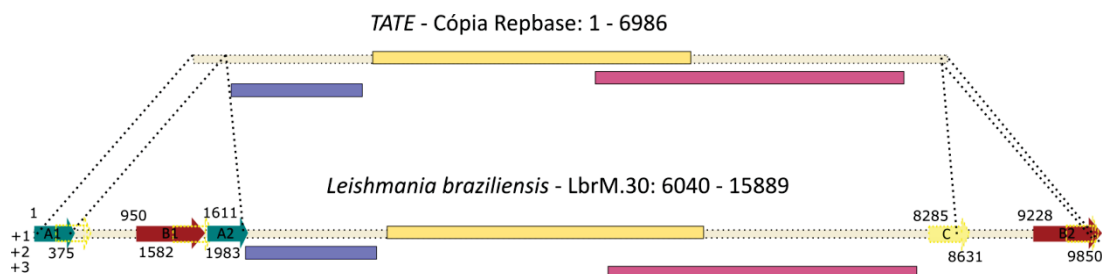
Em *Lep. pyrrhocris* encontramos três cópias completas e potencialmente codificadoras (Tabela 11). As três cópias apresentam as ORFs codificando os domínios correspondentes a YR e RT, mas o domínio da RH não foi encontrado em nenhuma das ORFs. Assim como *C. fasciculata*, as cópias de *Lep. pyrrhocris* não apresentaram padrão de repetição. Os resultados do BLAST contra a ORF1, correspondente a *Gag-like*, mostraram uma similaridade da *Gag-like* de *Lep. pyrrhocris* com o elemento *TATE* de *T. theileri*, e ao alinhar as sequências de *Gag-like* dessas duas espécies obtivemos regiões de alinhamento confiáveis que mostram semelhanças entre as sequências. Esse fato chamou a atenção, pois a proteína *Gag-like* apresenta divergência tão significativa entre as sequências que normalmente não se encontram ortólogos (na busca de BLASTp estabelecida), e as cópias entre si não apresentam um bom alinhamento. Essa similaridade poderia ser um indicativo de TH de *TATE* entre essas espécies.

O genoma da linhagem *L. braziliensis* MHOM/BR/75/M2903 apresenta um grande número de regiões com dados faltantes e foi possível recuperar apenas uma cópia codificadora. Já para a linhagem MHOM/BR/75/M2904, a maioria das 15 cópias mais conservadas (ordenadas pelo *score*) apresentam o elemento completo. Quatro cópias apresentam a ORF para *Gag-like* degenerada e, em apenas duas não foi possível analisar devido a sua posição em região de dados faltantes. A análise de divergência das proteínas mostrou uma maior conservação do que o observado em outras espécies, indicando que algumas cópias podem ter divergido uma da outra há pouco tempo, dentro da linhagem. Notamos um padrão de repetições nas cópias, porém a avaliação completa das repetições foi dificultada devido a presença de mais de uma cópia no mesmo *contig*, criando quimeras. A cópia mais conservada de *L. braziliensis* (Figura 11) mostrou um tipo de repetição diferente. Além de apresentar duas repetições (A e B), como mostrado na tabela 11, também apresenta uma repetição C com sobreposição às repetições A e B, tanto na extremidade 5' quanto

3'. Nenhum caso de repetição com esse padrão foi descrito até o momento e embora lembre uma ICR presente nos elementos *DIRS-like*, as repetições são todas diretas. Como essa repetição não é esperada, sua real existência deve ser corroborada com a análise de cópias adicionais de *TATE* dessa espécie.

Assim, encontramos uma diferença entre a cópia consenso de *TATE* (depositada no banco Repbase) e as cópias analisadas. A cópia consenso do elemento apresenta um tamanho menor, pois nela não estão representadas as repetições que encontramos na cópia de *L. braziliensis* e aparentemente possuem deleções nas extremidades. As diferenças e semelhanças entre as cópias estão representadas na figura 14. É possível observar que os 300 pb iniciais da cópia consenso são semelhantes a repetição A1, mas a região restante que contém as outras repetições está ausente.

FIGURA 14: REPRESENTAÇÃO ESQUEMÁTICA DO *TATE* COMPLETO DE *L. BRAZILIENSIS* E DA CÓPIA CONSENSO DISPONÍVEL NO REPBASE



LEGENDA: Representação esquemática de uma possível cópia completa do elemento *TATE* de *L. braziliensis* e uma comparação com a sequência consenso do elemento *TATE* da mesma espécie disponível no Repbase. Retângulos de diferentes cores representam as ORFs *gag-like* (azul), YR (amarelo), e RT/RH (rosa). As setas representam as repetições.

FONTE: A autora, 2018.

Pela proximidade de *L. panamensis* e *L. peruviana* com *L. braziliensis*, imaginávamos encontrar também cópias codificadoras nessas espécies. Porém, aparentemente as cópias de *L. peruviana* estão degeneradas e em *L. panamensis*, a maioria das cópias apresentam ORFs em regiões de dados faltantes, dificultando assim as análises.

As cópias que apresentam ORFs potencialmente codificadoras para *Gag-like* de *TATE* de todas as espécies não possuem o motivo de ligação a DNA CH-box. Buscamos também por variações do CH-box clássico, porém não conseguimos identifica-lo. O motivo MHR também não foi encontrado.

4.5 Evolução dos elementos *VIPER* e *TATE* nos genomas dos cinetoplastídeos

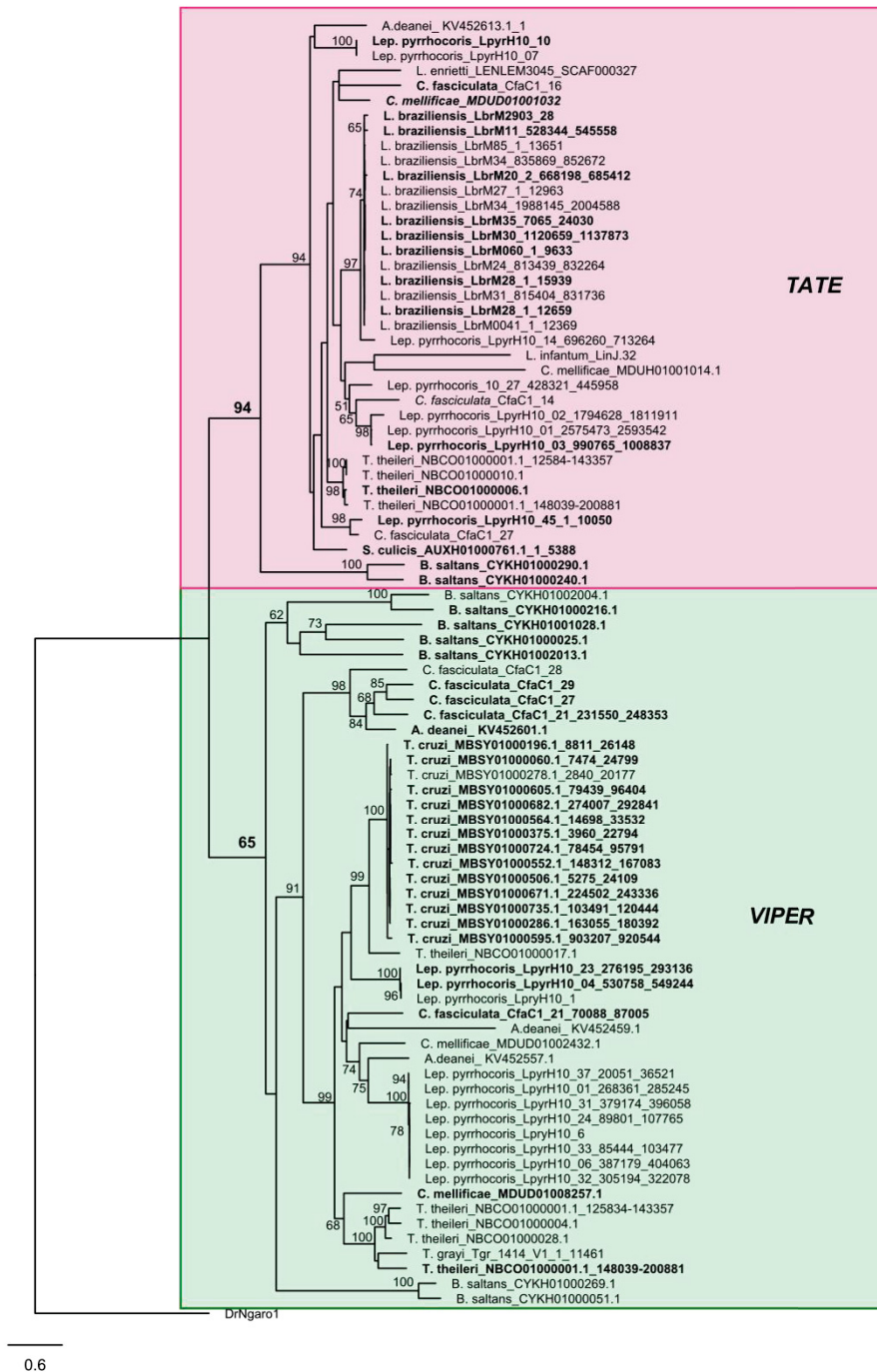
Realizamos construções filogenéticas dos elementos *VIPER* e *TATE* com base nas proteínas RT/RH e YR. As diferentes matrizes analisadas indicam que a adição de ambos os elementos na mesma árvore e o uso de grupo externo interferiu pouco na resolução das filogenias, pois as filogenias obtidas das matrizes de alinhamento individuais não mostraram valores de bootstrap mais confiáveis para a maioria dos clados. A construção da filogenia incluindo os dois elementos faz sentido, uma vez que observamos que esses elementos formam um grupo monofilético, como será apresentado no tópico a seguir. Ainda, cabe salientar que a análise desses elementos de forma separada dos demais foi importante, pois permitiu o uso de regiões maiores de sequências.

As filogenias, tanto pelo método de ML quanto por NJ, apresentaram topologia similar. ML apresentou um suporte melhor nos ramos, porém, mesmo com isso não foi possível resolver alguns grupamentos. Por isso, optamos por apresentar apenas as filogenias obtidas por ML que estão disponíveis no anexo E. A proteína YR apresentou baixa resolução filogenética para todas as matrizes e poucas relações podem ser estabelecidas.

A figura 15 mostra a árvore de ML de *VIPER* e *TATE*, incluindo o grupo externo, mostrando a separação em dois grandes clados que correspondem às sequências dos elementos nas diferentes espécies.

Para tentar entender o padrão de evolução dos elementos, comparamos a filogenia obtida com as relações filogenéticas das espécies hospedeiras (Figura 10). Para ambos os elementos, *B. saltans* apresenta uma posição basal, o que está de acordo com o esperado. Porém, é possível evidenciar inúmeras incongruências nos grupamentos derivados, como proximidade inesperada de cópias (por exemplo uma das cópias de *VIPER* de *C. mellificae* agrupa com *T. theileri* e *T. grayi*) e cópias da mesma espécie estão localizadas em clados distintos (*TATE* de *Lep. pyrrhocoris*, *C. fasciculata*, *C. mellificae*; *VIPER* de *Lep. pyrrhocoris*, *T. theileri*, *C. fasciculata*, *C. mellificae*, *A. deanei*). A separação de sequências da mesma espécie em clados distintos era esperada uma vez que grande divergência foi observada para as proteínas (discutido no tópico anterior).

FIGURA 15: FILOGENIA DE *VIPER* E *TATE* DE CINETOPLASTÍDEOS COM BASE NA SEQUÊNCIA DA TRANSCRIPTASE REVERSA



LEGENDA: Árvore de ML (modelo LG + G, 341 aa) de cópias de *VIPER* e *TATE* de cinetoplastídeos com base na sequência de aminoácidos da RT. A sequência da RT de *DrNgaro1* foi usada como grupo externo para enraizar a árvore. Valores de bootstrap de 1000 réplicas são mostrados próximos aos ramos (valores menores que 50 foram omitidos). Sequências que fazem parte de possíveis cópias codificadoras estão destacadas em negrito.

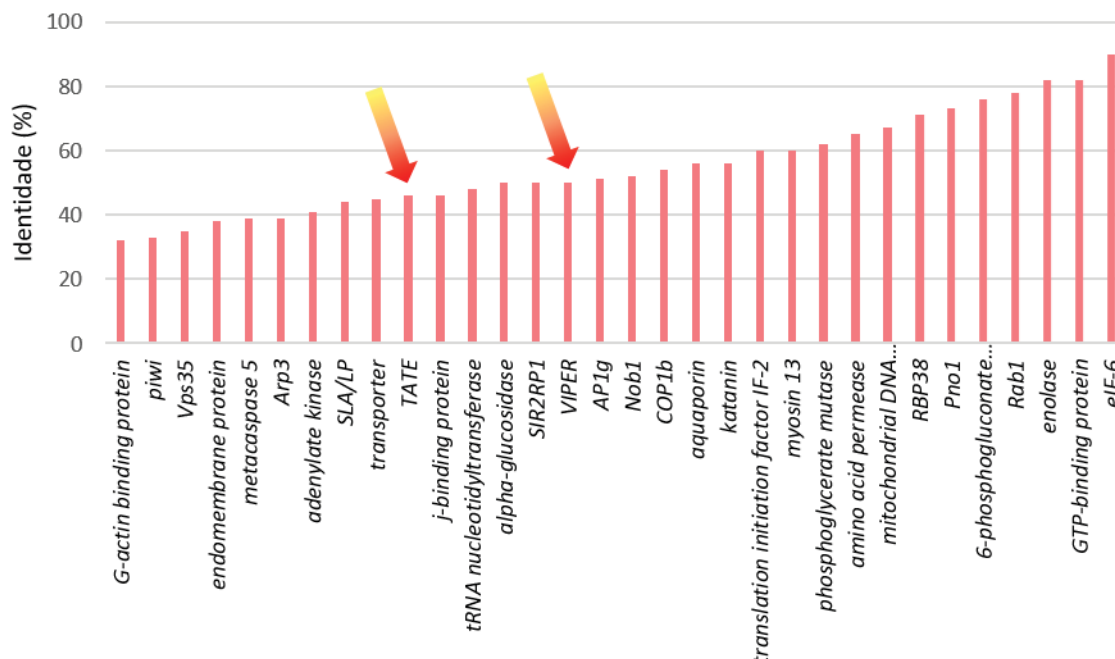
FONTE: A autora, 2018.

As cópias potencialmente codificadoras estão presentes em clados diversos, inclusive para cópias da mesma espécie, como por exemplo as sequências codificadoras de *TATE* de *Lep. pyrrhocris* estão presentes em três clados.

Duas incongruências chamaram a atenção e são sugestivas de TH: *TATE* que, apesar de parecer completamente perdido em todas as espécies de *Trypanosoma*, permanece completo em *T. theileri*, e, como já mencionado, a proteína Gag-like de *TATE* dessa espécie apresenta similaridade considerável com a proteína de *TATE* de *Lep. pyrrhocris*, o que não era esperado; *VIPER* de *Lep. pyrrhocris* divide-se em dois grupos, e em um agrupa com *T. theileri* e com *T. cruzi* (apesar de baixo bootstrap), sendo tanto as cópias de *T. cruzi* quanto as cópias de *Lep. pyrrhocris* potencialmente codificadoras. Além disso, as espécies se encontram em um mesmo contexto ambiental e vetorial, o que perfaz as primícias para TH, nos sugerindo ainda mais essa hipótese (FLEGONTOV et al., 2016; NGOMTCHO et al., 2017).

Para testar essa hipótese, comparamos a divergência observada para *VIPER* e *TATE* com a divergência das proteínas das espécies hospedeiras. Escolhemos comparar *T. theileri* e *Lep. pyrrhocris*. A figura 16 mostra um histograma da porcentagem de identidade das proteínas entre as duas espécies. Para aceitar TH do elemento, a sequência da RT de *TATE* ou *VIPER* deve ser mais conservada que as sequências das proteínas analisadas. Em caso de TV, a sequência da RT deve apresentar uma divergência média ou maior que as outras proteínas, pois em princípio não sofrem seleção purificadora. Podemos observar que para ambos os elementos, as divergências estão de acordo com TV dos elementos.

FIGURA 16: COMPARAÇÃO DA IDENTIDADE DAS SEQUÊNCIAS DOS ELEMENTOS TATE E VIPER COM PROTEÍNAS DO HOSPEDEIROS ENTRE *LEP. PYRROCORIS* E *T. THEILERI*



LEGENDA: Histograma mostrando a porcentagem de identidade de 30 proteínas ortólogas e a RT de TATE e VIPER entre *Lep. pyrrhcoris* e *T. theileri* ordenadas da esquerda para direita das mais divergentes as mais similares. As setas indicam as posições dos elementos TATE e VIPER.

FONTE: A autora, 2018.

4.6 História evolutiva dos elementos DIRS

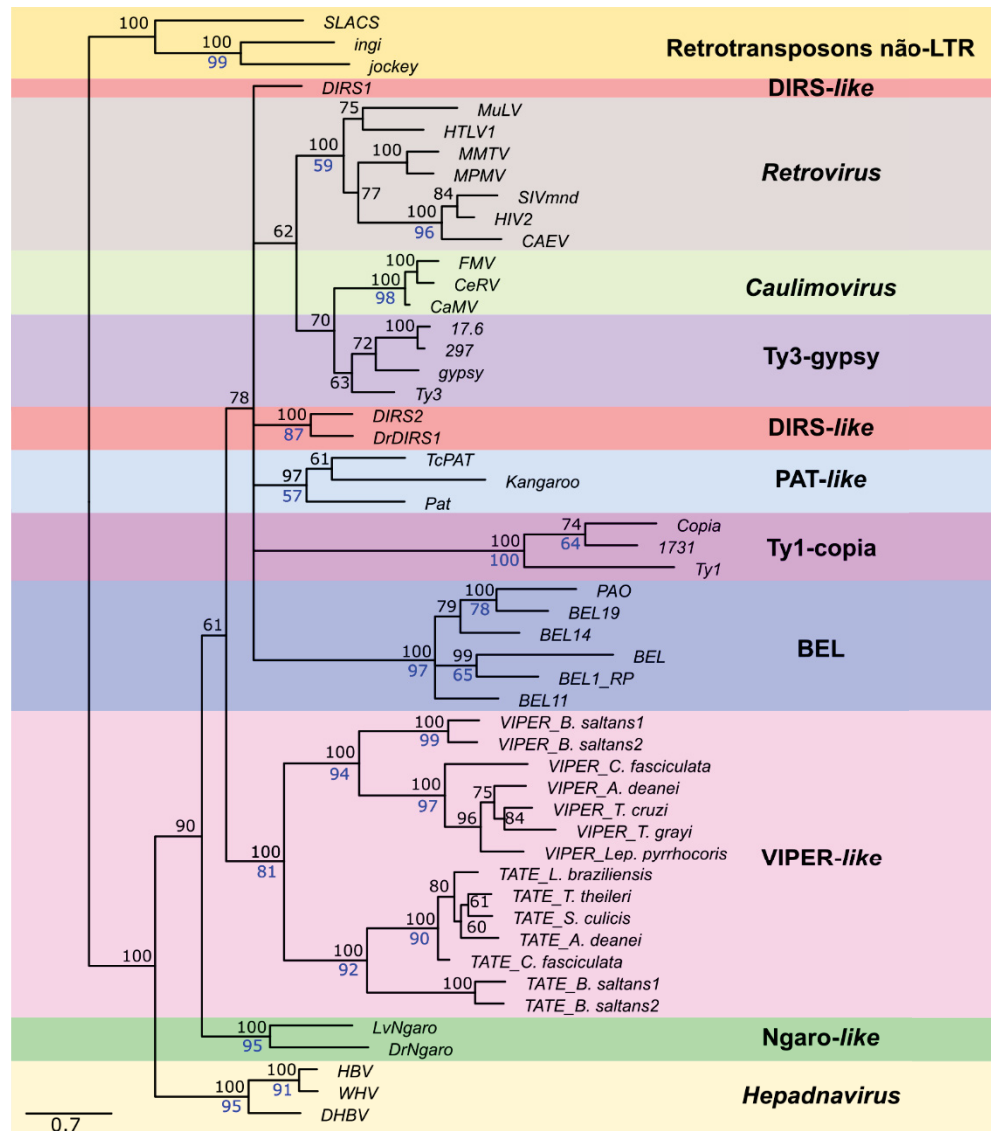
Para avaliar as relações filogenéticas entre os elementos contendo YR e outros retrotransposons, árvores filogenéticas foram construídas para YR e RT/RH, contendo sequências representativas dos diferentes grupos.

As filogenias por ML e BA apresentaram topologia similar, porém as MLs apresentaram menor suporte de clados e resolução. Optamos por apresentar as filogenias por BA e os valores de bootstrap obtidos para os clados pelas MLs são mostrados na mesma figura (quando maiores que 50%).

A figura 17 mostra a árvore por BA da RT com topologia geral similar à encontrada em outros estudos (MALIK; EICKBUSH, 2001; XIONG; EICKBUSH, 1990), com a separação mais primitiva dos *Hepadnavirus* e a confiável relação entre Ty3-gypsy, *Caulimovirus* e retrovírus. O posicionamento dos grupos Ty1-copia e BEL, como esperado, formam um clado juntamente com Ty3-gypsy, *Caulimovirus* e retrovírus. Porém, não foi possível obter a monofilia para os elementos do grupo DIRS-like. Os mesmos elementos DIRS-like (*DIRS1*, *DIRS2* e *DrDIRS1*) foram usados por LORENZI et al. (2006) e formaram um grupo monofilético juntamente

com os elementos *PAT-like* (PAT, Kangaroo, TcPAT). Elementos *DIRS-like* e *PAT-like* também formam um grupo monofilético em MALIK e EICKBUSH (2001). Em nossas análises, os elementos *PAT-like* formaram um grupo monofilético, porém não agruparam com os elementos *DIRS*.

FIGURA 17: RELAÇÕES FILOGENÉTICAS DOS DIFERENTES GRUPOS DE RETROELEMENTOS



LEGENDA: Árvore filogenética por BA (modelo LG + G + I, 181 aa) com base na sequência da proteína RT de diferentes grupos de retroelementos. Valores de probabilidade posterior são indicados acima dos nós e clados com suporte menor do que 50% foram colapsados. Valores de bootstrap apresentados para cada clado na análise de ML estão apresentados em azul abaixo dos nós (somente valores acima de 50). A árvore foi enraizada pelo ponto médio.

FONTE: A autora, 2018.

As sequências dos elementos *Ngaro-like* parecem ter divergido da linhagem dos demais retrotransposons primitivamente, conforme podemos observar pelo alto suporte da ramificação (90%).

As sequências de *VIPER* e *TATE* também formam um grupo monofilético (suporte 100% BA, 81% ML) que aparentemente ramificou antes dos demais retroelementos. Essa monofilia comprova a relação já esperada desses dois elementos, já que ambos estão presentes em tripanosomatídeos e possuem uma estrutura de similar.

De acordo com esses resultados, podemos sugerir a divisão dos retroelementos contendo YR em quatro grupos, que nomeamos de acordo com o primeiro elemento descrito seguido da indicação “like”: *DIRS-like* (grupamento suportado por outros autores, conforme já mencionado), *PAT-like*, *Ngaro-like* e *VIPER-like*. Assim, essa proposição difere, ou complementa, os demais trabalhos que mencionam a classificação dos elementos da ordem DIRS, pois WICKER et al. (2007) os classificaram em três superfamílias, DIRS, *Ngaro* e *VIPER*, e POULTER e BUTLER (2015) os separaram em três grupos distintos, *DIRS-like*, *Ngaro-like* e *PAT-like*.

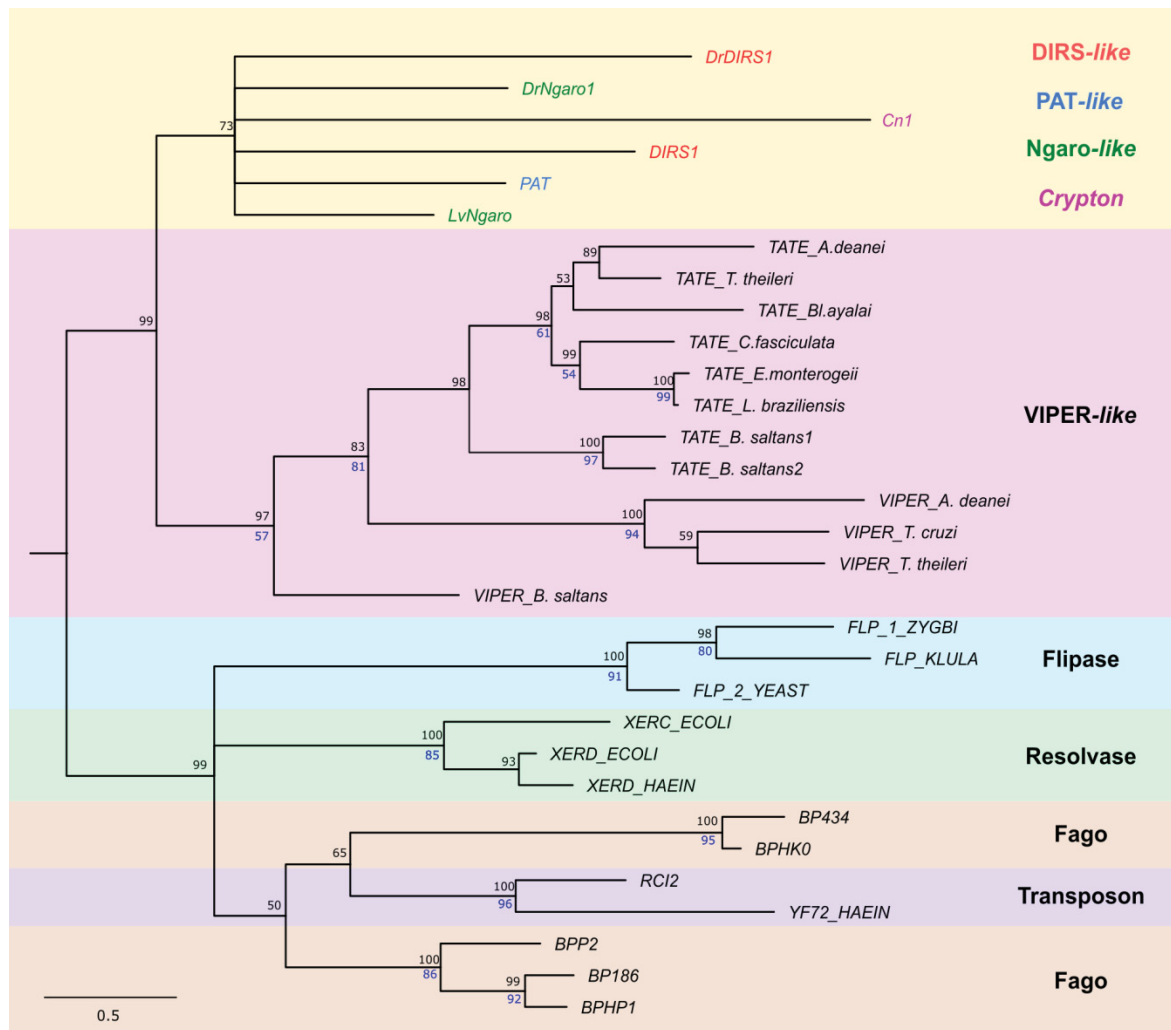
A separação dos elementos *PAT-like* e *DIRS-like* deve-se principalmente pela diferença estrutural das repetições terminais, já que filogeneticamente eles são proximamente relacionados.

Em relação a filogenia da YR, a resolução dos clados foi menor devido a maior divergência entre as proteínas usadas quando comparada com a RT. Apesar disso, foi possível obter algumas relações bem suportadas (Figura 18). Observamos inicialmente dois grandes clados, um contendo os retrotransposons e o transposon *Crypton* e outro contendo as demais YRs.

Os elementos *VIPER* e *TATE* formam um grupo monofilético indicando uma origem em comum também dessa proteína. Diferentemente do esperado, os elementos *Ngaro-like* agruparam com os elementos *Pat-like* e *DIRS-like*, indicando uma origem em comum dessa proteína para os três grupos, o que não é observado para a RT. Por isso, a análise dos domínios RT e YR separadamente foi importante, pois já tínhamos indícios dessas relações distintas entre as proteínas (LORENZI et al., 2006) e também, porque é conhecido que os retrotransposons e retrovírus possuem um perfil de evolução modular com repetidos eventos de rearranjos de domínios ou ainda origem reticulada de seus componentes, ou seja, cada domínio

pode ter sido originado de diferentes fontes (MALIK; EICKBUSH, 1999; MALIK; HENIKOFF; EICKBUSH, 2000; SMYSHLYAEV et al., 2013; TERZIAN; PÉLISSON; BUCHETON, 2001).

FIGURA 18: RELAÇÕES EVOLUTIVAS ENTRE PROTEÍNAS DA FAMÍLIA DE TIROSINA RECOMBINASES



LEGENDA: Árvore filogenética por BA (modelo LG + G, 132 aa) dos retrotransposons contendo YR (DIRS-like, PAT-like, Ngaro-like e VIPER-like) juntamente com outras proteínas da família de tirosina recombinases (transposons Crypton, transposons de procariotos, flipases, resolvases e IN de fagos). Valores de probabilidade posterior são indicados perto dos nós e clados com suporte menor do que 50% foram colapsados. Valores de bootstrap apresentados para cada clado na análise de ML estão apresentados em azul abaixo dos nós (somente valores acima de 50). A árvore foi enraizada pelo ponto médio.

FONTE: A autora, 2018.

A relação do transposon do grupo Crypton com os elementos contendo YR já havia sido descrita e também encontramos na filogenia obtida em nosso trabalho, porém não conseguimos recuperar a monofilia dos diferentes grupos, sendo todos apresentados como um clado politômico. Essa falta de resolução na verdade

também é observada nas filogenias de YR mostrada por outros autores (GOODWIN e POULTER, 2001; LORENZI et. al, 2006; POULTER e BUTLER, 2015), porém, na maioria das vezes as árvores não possuem os ramos de baixo suporte colapsados, como apresentamos no nosso trabalho, e a falta de suporte para os clados também não é mencionada, assim, observamos conclusões que consideramos precipitadas, pela falta de confiabilidade dos dados apresentados. Por exemplo, LORENZI et al. (2006) afirma haver relação dos elementos *DIRS-like*, *PAT-like* e *Ngaro-like* com as flipases e do elemento *VIPER* com transposases procarióticas, porém, essas relações apresentaram suporte abaixo de 50%. Já em nossos dados, obtivemos uma clara separação dos retroelementos e essas sequências.

5 DISCUSSÃO

5.1 Busca dos elementos nos genomas dos cinetoplastídeos – A influência da montagem dos genomas na análise genômica de TEs

Os TEs são sequências de DNA de natureza repetitiva e esse tipo de sequência causa problemas na hora da montagem dos genomas. O problema principal com sequências repetitivas é que os programas de montagem não conseguem distinguir de qual cópia é determinada leitura do sequenciamento e assim, as regiões flanqueadoras podem ser montadas de forma errônea. O grau de dificuldade que as repetições causam depende do tamanho das leituras, quanto menores, mais difícil ficará a montagem. Dessa forma, as sequências repetitivas podem muitas vezes ser mascaradas durante a montagem dos genomas, sendo então tratadas como região faltante e as informações sobre esse tipo de sequência muitas vezes acabam sendo incompletas. Esse problema pode ocorrer, principalmente com o sequenciamento de nova geração devido as curtas leituras geradas (TREANGEN; SALZBERG, 2013).

Os genomas de cinetoplastídeos usados nesse trabalho apresentam diferentes graus de montagem, como pode ser observado na tabela 2. Assim, precisamos tomar cuidado com as conclusões que tiramos a partir desses dados. Apesar da necessária cautela, dados genômicos continuam sendo uma fonte valiosa de estudo dos TEs, sendo amplamente utilizados para o melhor entendimento dos elementos nos contextos genômicos de diferentes organismos (DE LA CHAUX; WAGNER, 2011; DUFRESNE et al., 2011; GUYOT et al., 2016; KEIDAR; DORON; KASHKUSH, 2018).

Por exemplo, *Lep. pyrrhocoris* é uma das espécies que apresenta melhor montagem do genoma, quando comparado com os genomas de outras espécies analisadas, com montagem a nível cromossômico e apresenta cobertura de 37 vezes (FLEGONTOV et al., 2016). Mesmo assim, encontramos regiões de dados faltantes que dificultaram as análises.

Embora não tenhamos conseguido concluir sobre a presença e conservação dos elementos em alguns genomas, nós obtivemos importantes resultados, como a descoberta de cópias completas e potencialmente ativas de *VIPER* de *T. cruzi* na linhagem Dm28c, o que não foi encontrado no genoma de *T. cruzi* CL Brener

(BRINGAUD et al., 2008b; LORENZI; ROBLEDO; LEVIN, 2006b). Não podemos concluir se as diferenças encontradas entre as linhagens de *T. cruzi* são devido a problemas de montagem do genoma ou se são diferenças genéticas reais entre as linhagens. Porém, sabemos que o tipo de sequenciamento empregado para a linhagem Dm28c, o PacBio, gera montagens de sequências repetitivas com melhor qualidade. Essa tecnologia emprega o sequenciamento de uma única molécula de DNA em tempo real sem amplificação e é indicada para sequências repetitivas pois gera longos *reads* durante o sequenciamento (NAKANO et al., 2017).

VIPER parece estar degenerado em *T. brucei* e espécies próximas, *T. equiperdum*, *T. evansi*, *T. congolense* e *T. vivax*. Porém, assim como para *T. cruzi* que foi considerado inicialmente degenerado, a análise de diferentes linhagens ou montagens genômicas poderiam revelar cópias funcionais.

Ainda, encontramos cópias potencialmente codificadoras de *VIPER* em várias outras espécies, levantando a hipótese de que esse elemento ainda esteja ativo nessas espécies, o que é corroborada pela presença de repetições terminais em algumas espécies.

Em relação ao elemento *TATE* também obtivemos importantes achados, como o complexo padrão de repetições terminais em *L. braziliensis* e a presença de cópias potencialmente codificadoras em algumas espécies com distintos padrões de repetição. O genoma de *L. braziliensis* apresenta boa cobertura e montagem (PEACOCK et al., 2007), porém a natureza repetitiva de *TATE* e a presença de cópias em tandem dificultaram a análise, principalmente das repetições terminais.

Como pode ser observado nas tabelas 8 e 12, algumas repetições terminais apresentadas pelas cópias de *VIPER* e *TATE* apresentam 100% de identidade. Se essas regiões são importantes para o processo de transcrição reversa, assim como são as LTRs de LTR-retrotransposons, é provável que no momento da inserção as repetições sejam idênticas. Assim, é possível sugerir que essas cópias representam inserções recentes no genoma, podendo representar cópias ativas dos elementos. Cópias ativas em geral estariam representadas em maior número no genoma, porém, a qualidade da montagem dos genomas e os problemas envolvendo a montagem de sequências repetitivas poderiam explicar a identificação de apenas uma cópia.

Apesar das questões de montagem de genoma e dúvidas em relação a algumas espécies, é possível observar que *VIPER* e *TATE* apresentam uma

distribuição descontínua ao longo da filogenia das espécies hospedeiras, o que foi melhor entendida com as análises filogenéticas. Todas as inferências evolutivas foram feitas com base no cenário mais parcimonioso de acordo com os dados atuais. Isso implica, que a adição futura de novas sequências ou novas espécies, pode levar a mudanças das conclusões.

5.2 *VIPER* e *TATE*: sobreviventes nos genomas dos cinetoplastídeos

Inicialmente, LORENZI et al. (2006) havia proposto que *VIPER* teria colonizado os genomas de espécies de *Trypanosoma* após a separação deste gênero e *Leishmania*, devido à ausência de *VIPER* nas espécies de *Leishmania* previamente analisadas. Nossos dados mudam completamente esse cenário, pois encontramos cópias potencialmente codificadoras e resquícios de *VIPER* em *B. saltans* e em espécies da subfamília *Leishmaniinae*.

A presença de *VIPER* e *TATE* em *B. saltans* já havia sido mencionada anteriormente (JACKSON et al., 2016). Apesar do padrão de distribuição descontínuo dos elementos poder indicar que houve TH envolvendo *B. saltans*, temos observado que esse padrão de evolução é comum para vários genes. *B. saltans* parece manter inúmeros genes que estavam presentes no ancestral antes da separação dos tripanosomatídeos e esses genes acabaram sendo perdidos de forma descontínua em alguns grupos específicos ao longo da evolução dos tripanosomatídeos (JACKSON et al., 2016) como é o exemplo da família das miosinas recentemente reportado pelo grupo (DE SOUZA et al., 2018).

Devido a divergência encontrada entre as cópias de *VIPER* e *TATE* dentro da mesma espécie, e o padrão de separação dessas cópias em clados distintos podemos sugerir a existência de polimorfismo ancestral, pois seria improvável o acúmulo de divergência em tal nível em pouco tempo. A presença de cópias de *VIPER* distintas em *B. saltans* indica a possível existência de polimorfismo no ancestral comum de tripanosomatídeos e essa espécie. Porém, a topologia da árvore (ausência de sequências adicionais de *B. saltans* nos clados derivados) indica que a maior parte do polimorfismo de *VIPER* e *TATE* existente nos tripanosomatídeos surgiu após a separação dessa família com ancestral de *B. saltans*.

Apesar das inconclusões em alguns pontos da filogenia podemos sugerir vários eventos de perdas independentes dos elementos ao longa da evolução dos tripanosomatídeos. *VIPER* parece ter sido perdido em todas linhagens de *Phytomonas* e *TATE* nas linhagens de *Phytomonas* sp. EM1 e HART1. *Phytomonas* sp isolados EM1 e Hart1 estão com cobertura menor que *P. francai* (BUTLER; JASKOWSKA; KELLY, 2017), mas não encontramos nem resquícios dessas sequências, embora apresentem cópias degeneradas de elementos do tipo Ingi-like (PORCEL et al., 2014). A degeneração completa de *TATE* nas espécies de *T. cruzi* e *VIPER* na maioria das espécies de *Leishmania* também é observada e suportada pela ausência de *hits*, sendo vários dos genomas montados em nível razoável, que seria pelo menos esperado encontrar alguns *hits*. A degeneração provavelmente começou a acontecer no ancestral comum entre as espécies ou até mesmo a perda completa dos elementos. As espécies de *Lep. seymouri* e *Lo. passim* também perderam *VIPER* e *TATE*, *E. monterogeii* perdeu *VIPER* e *L. tarentolae* não apresenta mais resquícios de *TATE*, o que já foi reportado anteriormente (RAYMOND et al., 2012).

Além do padrão de polimorfismos ancestral seguido de perdas e degeneração dos elementos *VIPER* e *TATE* ao longo da evolução, os casos mais interessantes são os de manutenção da atividade. Podemos observar pelo menos três possíveis eventos recentes de “explosão” (do inglês: *burst*) de transposição, quando ocorre uma massiva mobilização dos elementos no genoma. Esses eventos podem ser observados por grandes clados formados por sequências da mesma espécie com tamanho de ramos extremamente curtos. Aparentemente, essa grande atividade pode ter ocorrido para o elemento *VIPER* em *T. cruzi* e em *Lep. pyrrhocoris* e para o elemento *TATE* em *L. braziliensis*.

Concluindo, a presença do *VIPER* e do *TATE* em *B. saltans*, sugere que os dois elementos já estavam presentes no último ancestral comum dessas espécies de cinetoplastídeos há mais de 460 milhões de anos atrás e ainda parecem estar ativos (ou foram reativados) em algumas espécies. Essa “sobrevivência” por tantos milhões de anos levanta questionamentos sobre a co-evolução dessas sequências com o genoma dos seus hospedeiros, tais como, o modo que as sequências têm escapado da inativação e como esses genomas controlam essas sequências. Ainda, tantos milhões de anos de co-evolução, é impossível não pensar em qual tem sido o

impacto dessas sequências na evolução das espécies e como eles podem ter contribuído para organização e funcionalidade dos genomas dos cinetoplastídeos.

5.3 Origem independente dos elementos contendo YR

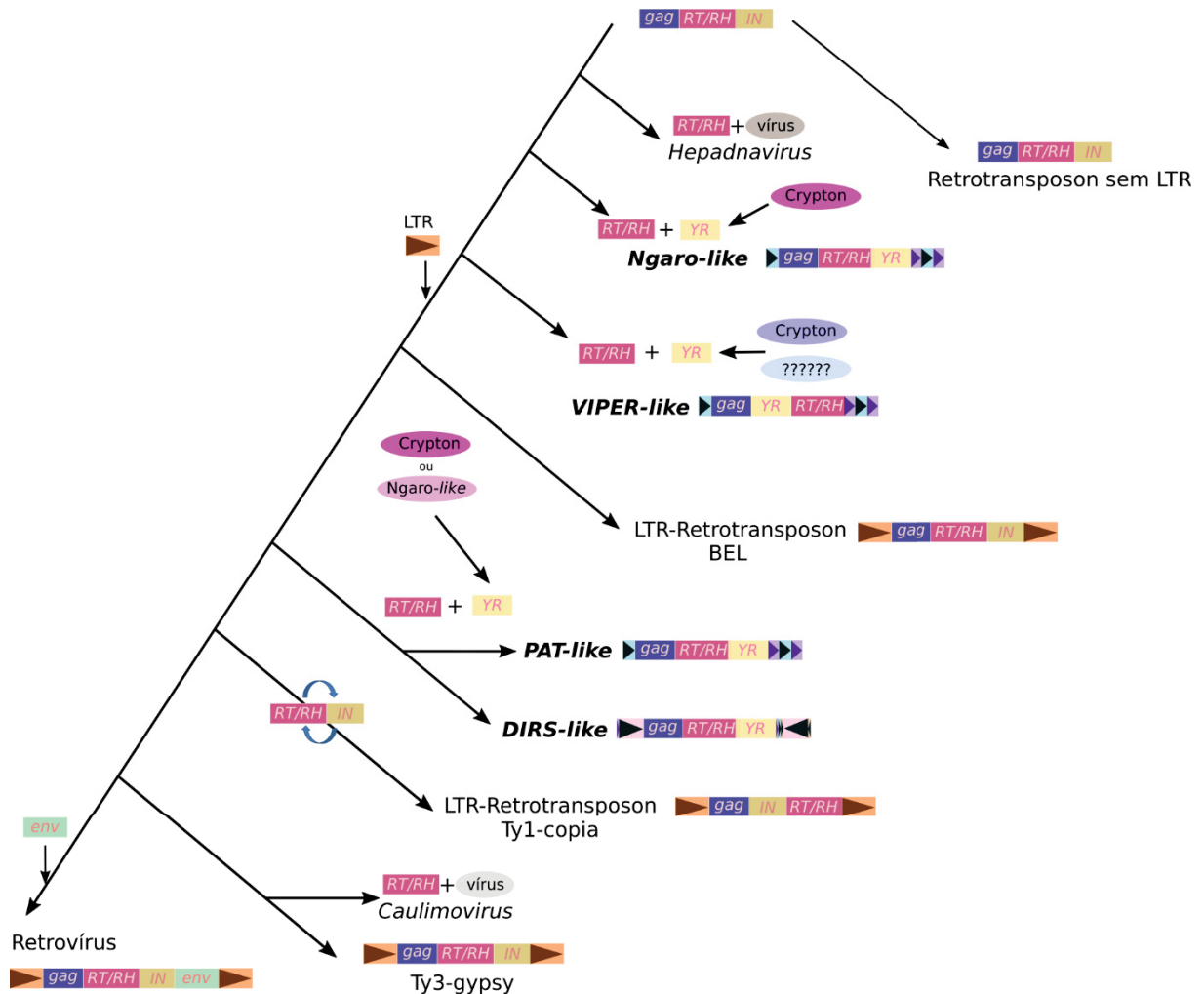
POULTER e BUTLER (2015) sugerem que retrotransposons do tipo YR tenham surgido pela combinação de elementos de DNA do tipo Crypton e RT/RH de um retrotransposon Ty3-gypsy pré-existente. E pelas diferentes características sugerem que essa transição de um transposon de DNA para um retrotransposon possa ter acontecido diversas vezes formando os três grupos, que seriam *DIRS-like*, *PAT-like* e Ngaro. No nosso trabalho, mostramos evidência filogenética para a origem independente dos elementos YR e separação em quatro grupos, incluindo os elementos *VIPER-like*.

A habilidade de poder recuperar a história evolutiva dessas sequências depende do tempo em que esses elementos surgiram e das pressões seletivas a que essas sequências foram impostas. Conforme já esperávamos, observamos a não resolução de algumas relações, possivelmente devido à falta de sinal filogenético que pode refletir eventos de separação muito antigos, talvez antes da divergência dos principais grupos eucariotos. Mesmo assim, nossos dados juntamente com os dados da literatura permitiram a sugestão de um cenário evolutivo para essas sequências (Figura 19).

Segundo esse cenário, um progenitor sem LTRs deu origem a duas linhagens, dos retrotransposons sem LTR e dos demais retrotransposons. A aquisição das LTRs é indicada na figura, dando origem aos LTR-retrotransposons dos grupos BEL, Ty1/-copia (com inversão na ordem dos domínios RT/RH-IN) e Ty3-gypsy. E a aquisição do gene *env* deu origem aos Retrovírus. Para *Hepadnavirus* e *Caulimovirus* é assumido que somente parte do gene *pol*, contendo o domínio RT se associou com algum vírus já existente. Os elementos do grupo Ngaro-like podem ter se formado pela combinação de um elemento Crypton (fonte da YR) com uma RT/RH dos retrotransposons, muito primitivamente. Da mesma forma, os elementos *VIPER-like* podem ter surgido primitivamente. A origem da YR dos elementos *VIPER-like* pode ser de algum elemento Crypton ou ainda de alguma origem desconhecida no momento. O possível surgimento desse elemento antes mesmo do

surgimento dos LTR-retrotransposons indica que elementos desse grupo poderiam ter uma distribuição mais abrangente do que é vista (restrita aos cinetoplastídeos). Isso indica que o sequenciamento de novos genomas pode revelar uma distribuição mais ampla desse grupo ou ainda, que sequências desse grupo representam uma adaptação para “parasitar” esses genomas tão peculiares. Devido à proximidade da YR de *PAT-like* e *DIRS-like* com a YR de elementos *Ngaro-like*, pode-se sugerir que eles tiveram uma mesma fonte da YR, ou ainda, como os *Ngaro-like* devem ter surgido anteriormente, os elementos *PAT-like* e *DIRS-like* podem ter sido originados pela recombinação de elementos *Ngaro-like* que trocaram suas RT/RH com LTR-retrotransposons. Elementos *PAT-like* e *Ngaro-like* apresentam padrão semelhante de repetições terminais, já os elementos *DIRS-like* sofreram rearranjos importantes na estrutura das repetições e consequentemente no mecanismo de transcrição reversa.

FIGURA 19: REPRESENTAÇÃO DE UM POSSÍVEL CENÁRIO EVOLUTIVO PARA A ORIGEM DOS RETROTRANSPOSONS



LEGENDA: Esquema representando um possível cenário para a origem dos retroelementos com base na proposição inicial de XIONG e EICKBUSH (1990) e atualizada de acordo como os nossos dados. Os genes estão representados por caixas de diferentes cores, gag, IN (integrasse), RT/RH (Transcriptase Reversa/RNaseH), YR (tirosina recombinase), envelope (env). LTRs e repetições são representadas por caixas contendo setas. O surgimento dos grandes grupos de retroelementos é indicado começando com *Hepadnavirus*, *Ngaro-like*, *VIPER-like*, *BEL*, *PAT-like* e *DIRS-like*, *Ty1-copia*, *Ty3-gypsy*, *Caulimovirus* e *Retrovírus*. Nesse cenário não consideramos a origem do gene *Gag-like* dos retrotransposons da ordem *DIRS*, que pode ter sido originada do próprio retrotransposon de quem foi adquirida a RT/RH ou a partir de uma origem distinta. Também não estamos considerando possível origem alternativa ou recombinação do domínio RH. Origem das ORFs extras dos *DIRS* (hidrolases e metiltransferases) não foi considerada. Classes de sequências possivelmente derivadas da linhagem de retrotransposons sem LTR foram omitidas.

FONTE: A autora, 2018.

5.4 O que ainda precisamos saber sobre os elementos DIRS?

Apesar do estudo evolutivo realizado neste trabalho ter norteado o entendimento de algumas questões sobre o contexto dos elementos DIRS, alguns pontos ainda permanecem em dúvida.

Uma das grandes questões a respeito dos DIRS é sobre o modo de transposição desses elementos. Sabemos que a proteína Gag de retrovírus é requerida para a formação das VLPs, o que nos faz pensar se nos elementos DIRS a *Gag-like* também desempenharia esse papel de formação de VLP, como ocorreria esse processo e qual seria o endereçamento desta partícula. Verificamos que as proteínas *Gag-like* de *VIPER* de algumas espécies possuem o CH-box, enquanto as de *TATE* não possuem. Ainda, observamos uma grande divergência nessas sequências que dificulta o estabelecimento de relações com outras proteínas. Essa grande divergência poderia ser explicada por uma seleção relaxada ou por seleção positiva atuando sobre a proteína. Estudos funcionais dessa proteína podem ajudar a resolver essa questão, como o que o nosso grupo está fazendo com o elemento *VIPER* de *T. cruzi*.

Pouco sabemos também a respeito da expressão e processamento do RNA destes elementos. Existe um promotor específico para que a transcrição possa ocorrer? É relatado que os elementos *Ingi-like* de tripanosomatídeos tem um promotor próprio (HERAS et al., 2007), o que nos leva a pensar que talvez os elementos *VIPER* e *TATE* possam também possuir um promotor. Alternativamente, eles seriam transcritos em unidades policistronicas juntamente com os demais genes. Além disso, como o mRNA desses elementos é processado? Seria uma molécula única ou cada ORF formaria um mRNA? Como não existem regiões intergênicas entre as ORFs, é provável que sejam processadas como um mRNA único.

Outro questionamento que levantamos é sobre a tradução do elemento, qual seria a sinalização para que ocorra a tradução da próxima ORF? A suposição é as diferentes proteínas sejam traduzidas por *ribosomal frameshift*, através do qual sinais específicos no mRNA sinalizam para uma mudança de fase (DINMAN, 2012). Essa estratégia de tradução é usada por muitos vírus, principalmente retrovírus (BOLINGER; BORIS-LAWRIE, 2009; DINMAN, 2012). Um dos pontos, visto durante as análises, que nos ajuda a iniciar uma sustentação para a hipótese é a presença de cada uma das ORFs em *frames* diferentes e uma grande sobreposição de algumas ORFs. As ORFs 1 e 2 de *VIPER* de *T. cruzi*, porém, estão na mesma fase.

Ainda sobre a tradução dos elementos, observamos que *TATE* de *B. saltans* apresenta uma única ORF onde possivelmente os domínios funcionais são expressos como uma poliproteína. Diferentemente dos LTR-retrotransposons, os

elementos da ordem DIRS não codificam uma PR para realizar o processamento do polipeptídeo. Assim, seria necessário a atuação de alguma protease do hospedeiro.

Um dos pontos mais intrigantes quando pensamos em TEs de tripanosomatídeos é em relação a regulação dessas sequências. Quando os TEs apresentam alta taxa de transposição, eles podem ser prejudiciais ao hospedeiro. Assim, os genomas apresentam mecanismos que silenciam TEs para estabilizar os genomas (GEBERT; ROSENKRANZ, 2015). Um dos processos mais importantes envolve o mecanismo de interferência de RNA (RNAi), onde pequenos RNAs não codificadores funcionam como guias para proteínas efetoras da família Argonauta que exercem atividades enzimáticas para inibir a expressão gênica, principalmente por clivagem de transcritos ou bloqueando a síntese proteica (GEBERT; ROSENKRANZ, 2015). Esse mecanismo foi visto aparentemente controlando *TATE* em *L. braziliensis* (ATAYDE et al., 2013) e *SLACS* e *Ingi* em *T. brucei* (DIKENG et al., 2001). Porém, a maquinaria clássica de RNAi não está presente em todas as espécies de tripanosomatídeos, incluindo *T. cruzi* (ULLU; TSCHUDI; CHAKRABORTY, 2004). Dessa forma, deve existir algum mecanismo alternativo de controle dos TEs nessa espécie, já que observamos a presença de cópias potencialmente ativas de *VIPER* nessa espécie.

Acreditamos que o elemento *VIPER* de *T. cruzi* Dm28c pode ser um ótimo modelo para o estudo do mecanismo de transposição e funcionamento dos elementos contendo YR e nosso trabalho representa um importante ponto de partida para futuros estudos funcionais.

6 CONCLUSÕES

Um dos primeiros pontos abordados no trabalho foi a respeito da ausência de uma descrição do elemento *TATE*, que ainda é pouco estudo e não possuía a descrição de uma estrutura bem definida na literatura. A análise de várias cópias de *TATE* conservadas no genoma de *L. braziliensis* nos permitiu avaliar o padrão estrutural do elemento, atingindo assim nosso primeiro objetivo proposto: descrever a estrutura geral do elemento *TATE*.

Com a busca acurada de cópias completas, degeneradas e ausentes de *VIPER* e *TATE*, conseguimos verificar os padrões de disposição das ORFs e repetições terminais e distribuição dos elementos entre as espécies. Tendo caracterizado as cópias completas e remanescentes nos genomas, foi possível começar a elucidar cenários para a evolução desses elementos ao longo da evolução dos tripanosomatídeos, conseguindo assim atingir o objetivo proposto de entender a distribuição, conservação e evolução dos elementos no contexto evolutivo das espécies hospedeiras.

Por fim, elucidar um cenário de origem para os elementos *VIPER* e *TATE* dentro da história evolutiva dos retrotransposons compunha um dos objetivos mais abrangentes e desafiador do trabalho. Com nossas análises conseguimos chegar ao desenvolvimento de um cenário mais parcimonioso sobre a evolução dos retrotransposons, incluindo outros elementos da ordem DIRS além dos elementos *VIPER* e *TATE*, contribuindo assim para uma maior compreensão das relações existentes entre os retrotransposons.

Com o desfecho desse trabalho, conseguimos preencher lacunas pré-existentes sobre os retrotransposons *VIPER* e *TATE* e também sobre os DIRS em um contexto mais geral. Um dos fatos importantes, e que devemos salientar, é que os estudos evolutivos são realizados pensando no cenário mais parcimonioso para o momento presente, podendo ser confirmado ou refutado posteriormente. A consistência dos resultados não só ajudou na compreensão de pontos duvidosos sobre os elementos, mas também colaborou com a abertura de novos pensamentos e questionamentos, que motivam o crescimento de estudos sobre esses elementos.

7 PERSPECTIVAS

- Avaliar o contexto genômico das cópias dos elementos, incluindo distância de genes, posicionamento e orientação nas unidades poliestrônicas;
- Analisar um número maior de cópias de algumas espécies;
- Avaliar a presença de pequenos elementos derivados de *VIPER* e *TATE* em diferentes genomas;

REFERÊNCIAS

- AKSOY, S. et al. SLACS retrotransposon from *Trypanosoma brucei gambiense* is similar to mammalian LINEs. **Nucleic acids research**, v. 18, n. 4, p. 785–92, fev. 1990.
- ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research**, v. 38, n. Database issue, p. D457-62, jan. 2009.
- ATAYDE, V. D. et al. The structure and repertoire of small interfering RNAs in *Leishmania (Viannia) braziliensis* reveal diversification in the trypanosomatid RNAi pathway. **Molecular microbiology**, v. 87, n. 3, p. 580–93, fev. 2013.
- BERRIMAN, M. et al. The genome of the African trypanosome *Trypanosoma brucei*. **Science (New York, N.Y.)**, v. 309, n. 5733, p. 416–22, jul. 2005.
- BIÉMONT, C.; VIEIRA, C. What transposable elements tell us about genome organization and evolution: the case of *Drosophila*. **Cytogenetic and genome research**, v. 110, n. 1–4, p. 25–34, jan. 2005.
- BOEKE, J. D.; STOYE, J. P. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: COFFIN, J. M.; HUGHES, S.; VARMUS, H. (Eds.). **Retroviruses**. Cold Spring Harbor: Cold Spring Harbor Press, 1997. p. 343–435.
- BOLINGER, C.; BORIS-LAWRIE, K. Mechanisms employed by retroviruses to exploit host factors for translational control of a complicated proteome. **Retrovirology**, v. 6, p. 8, 24 jan. 2009.
- BRINGAUD, F. et al. Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. **Molecular and biochemical parasitology**, v. 124, n. 1–2, p. 73–8, 2002.
- BRINGAUD, F. et al. The ingi and RIME non-LTR Retrotransposons are Not Randomly Distributed in the Genome of *Trypanosoma brucei*. **Molecular Biology and Evolution**, v. 21, n. 3, p. 520–528, 2004.
- BRINGAUD, F. et al. Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. **PLoS pathogens**, v. 3, n. 9, p. 1291–307, set. 2007.
- BRINGAUD, F. et al. Role of transposable elements in trypanosomatids. **Microbes and Infection**, v. 10, n. 6, p. 575–581, 2008a.
- BRINGAUD, F. et al. Role of transposable elements in trypanosomatids. **Microbes and infection / Institut Pasteur**, v. 10, n. 6, p. 575–81, maio 2008b.
- BRINGAUD, F.; ROGERS, M.; GHEDIN, E. Identification and analysis of ingi-related retroposons in the trypanosomatid genomes. **Methods in molecular biology (Clifton, N.J.)**, v. 1201, p. 109–22, 2015.
- BURNS, K. H. Transposable elements in cancer. **Nature reviews. Cancer**, v. 17, n. 7, p. 415–424, 2017.
- BUTLER, C. E.; JASKOWSKA, E.; KELLY, S. Genome Sequence of *Phytomonas françai*, a Cassava (*Manihot esculenta*) Latex Parasite. **Genome announcements**, v. 5, n. 2, jan. 2017.

- CAPPELLO, J.; HANDELSMAN, K.; LODISH, H. F. Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. **Cell**, v. 43, n. 1, p. 105–15, nov. 1985a.
- CAPPELLO, J.; HANDELSMAN, K.; LODISH, H. F. Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. **Cell**, v. 43, n. 1, p. 105–15, nov. 1985b.
- CAPY, P.; BAZIN, C.; LANGIN, T. **Dynamics and evolution of transposable elements**. Austin: Landes Bioscience, 1998.
- CHAGAS, C. Nova tripanozomíaze humana: estudos sobre a morfologia e o ciclo evolutivo do *Schizotrypanum cruzi* n. gen., n. sp., agente etiológico de nova entidade morbida do homem. **Memórias do Instituto Oswaldo Cruz**, v. 1, n. 2, p. 159–218, ago. 1909.
- CORTEZ, A. P. et al. The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa. **Parasitology**, v. 133, n. Pt 2, p. 159–69, ago. 2006.
- DE LA CHAUX, N.; WAGNER, A. BEL/Pao retrotransposons in metazoan genomes. **BMC evolutionary biology**, v. 11, p. 154, 4 jun. 2011.
- DE SOUZA, D. A. S. et al. Evolutionary analyses of myosin genes in trypanosomatids show a history of expansion, secondary losses and neofunctionalization. **Scientific reports**, v. 8, n. 1, p. 1376, jan. 2018.
- DE SOUZA, W. Basic cell biology of *Trypanosoma cruzi*. **Current pharmaceutical design**, v. 8, n. 4, p. 269–85, 2002.
- DI TOMMASO, P. et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. **Nucleic acids research**, v. 39, n. Web Server issue, p. W13-7, jul. 2011.
- DINMAN, J. D. Mechanisms and implications of programmed translational frameshifting. **Wiley interdisciplinary reviews. RNA**, v. 3, n. 5, p. 661–73, 2012.
- DJIKENG, A. et al. RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24-26-nucleotide RNAs. **RNA (New York, N.Y.)**, v. 7, n. 11, p. 1522–30, nov. 2001.
- DOCAMPO, R. et al. Acidocalcisomes - Conserved from bacteria to man. **Nature Reviews Microbiology**, v. 3, n. 3, p. 251–261, 2005.
- DOCAMPO, R. The origin and evolution of the acidocalcisome and its interactions with other organelles. **Molecular and biochemical parasitology**, v. 209, n. 1–2, p. 3–9, 2016.
- DOTTO, B. R. et al. HTT-DB: horizontally transferred transposable elements database. **Bioinformatics (Oxford, England)**, v. 31, n. 17, p. 2915–7, 1 set. 2015.
- DOTTO, B. R. et al. HTT-DB: new features and updates. **Database : the journal of biological databases and curation**, v. 2018, 1 jan. 2018.
- DU, Y.; MASLOV, D. A.; CHANG, K. P. Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. **Proceedings of the National Academy of Sciences of the United States of America**, v. 91, n. 18, p. 8437–41, ago. 1994.

- DUBIN, M. J.; MITTELSTEN SCHEID, O.; BECKER, C. Transposons: a blessing curse. **Current Opinion in Plant Biology**, v. 42, p. 23–29, 2018.
- DUFRESNE, M. et al. Genome-wide comparative analysis of pogo-like transposable elements in different *Fusarium* species. **Journal of molecular evolution**, v. 73, n. 3–4, p. 230–43, out. 2011.
- EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic acids research**, v. 32, n. 5, p. 1792–7, jan. 2004.
- EICKBUSH, T. H.; JAMBURUTHUGODA, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. **Virus research**, v. 134, n. 1–2, p. 221–34, jun. 2008.
- EL-SAYED, N. M. et al. The African trypanosome genome. **International Journal for Parasitology**, v. 30, p. 329–345, 2000.
- EL-SAYED, N. M. et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science (New York, N.Y.)**, v. 309, n. 5733, p. 409–15, jul. 2005a.
- EL-SAYED, N. M. et al. Comparative genomics of trypanosomatid parasitic protozoa. **Science (New York, N.Y.)**, v. 309, n. 5733, p. 404–9, jul. 2005b.
- FESCHOTTE, C. The contribution of transposable elements to the evolution of regulatory networks. **Nature Reviews Genetics**, v. 9, n. 5, p. 397–405, 2008.
- FESCHOTTE, C.; PRITHAM, E. J. DNA transposons and the evolution of eukaryotic genomes. **Annual review of genetics**, v. 41, p. 331–68, jan. 2007.
- FINNEGAN, D. J. Eukaryotic transposable elements and genome evolution. **Trends in genetics : TIG**, v. 5, n. 4, p. 103–7, abr. 1989.
- FLEGONTOV, P. et al. *Paratrypanosoma* is a novel early-branching trypanosomatid. **Current Biology**, v. 23, n. 18, p. 1787–1793, set. 2013.
- FLEGONTOV, P. et al. Genome of *Leptomonas pyrrhocoris*: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of *Leishmania*. **Scientific Reports**, v. 6, n. October 2015, p. 23704, 2016.
- FLODEN, E. W. et al. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. **Nucleic Acids Research**, p. gkw300, abr. 2016.
- FREED, E. O. HIV-1 assembly, release and maturation. **Nature reviews. Microbiology**, v. 13, n. 8, p. 484–96, ago. 2015.
- FROST, L. S. et al. Mobile genetic elements: The agents of open source evolution. **Nature Reviews Microbiology**, v. 3, n. 9, p. 722–732, 2005.
- GABRIEL, A. et al. A rapidly rearranging retrotransposon within the minixon gene locus of *Crithidia fasciculata*. **Molecular and cellular biology**, v. 10, n. 2, p. 615–624, 1990.
- GEBERT, D.; ROSENKRANZ, D. RNA-based regulation of transposon expression. **Wiley interdisciplinary reviews. RNA**, v. 6, n. 6, p. 687–708, 2015.
- GHEDIN, E. et al. Gene synteny and evolution of genome architecture in trypanosomatids. **Molecular and biochemical parasitology**, v. 134, n. 2, p. 183–91, abr. 2004.
- GILBERT, C.; CORDAUX, R. Horizontal transfer and evolution of prokaryote

transposable elements in eukaryotes. **Genome Biology and Evolution**, v. 5, n. 5, p. 822–832, 2013.

GOODWIN, T. J. D.; POULTER, R. T. M. A new group of tyrosine recombinase-encoding retrotransposons. **Molecular biology and evolution**, v. 21, n. 4, p. 746–59, abr. 2004.

GOODWIN, T. J.; POULTER, R. T. The DIRS1 group of retrotransposons. **Molecular biology and evolution**, v. 18, n. 11, p. 2067–82, nov. 2001.

GUERREIRO, M. P. G. What makes transposable elements move in the Drosophila genome? **Heredity**, v. 108, n. 5, p. 461–468, 2012.

GUYOT, R. et al. Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories. **Molecular genetics and genomics : MGG**, v. 291, n. 5, p. 1979–90, out. 2016.

HARKINS, K. M. et al. Phylogenomic reconstruction supports supercontinent origins for Leishmania. **Infection, Genetics and Evolution**, v. 38, p. 101–109, mar. 2016.

HAVECKER, E. R.; GAO, X.; VOYTAS, D. F. The diversity of LTR retrotransposons. **Genome biology**, v. 5, n. 6, p. 225, 2004.

HE, Z. et al. Evolvview v2: an online visualization and management tool for customized and annotated phylogenetic trees. **Nucleic acids research**, v. 44, n. W1, p. W236–41, jul. 2016.

HERAS, S. R. et al. The L1Tc non-LTR retrotransposon of Trypanosoma cruzi contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. **Nucleic acids research**, v. 35, n. 7, p. 2199–214, 2007.

HUSON, D. H.; SCORNAVACCA, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. **Systematic biology**, v. 61, n. 6, p. 1061–7, dez. 2012.

IKEDA, Y. et al. Arabidopsis proteins with a transposon-related domain act in gene silencing. **Nature communications**, v. 8, p. 15122, 3 maio 2017.

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, jul. 2005.

JACKSON, A. P. Genome evolution in trypanosomatid parasites. **Parasitology**, v. 142, n. S1, p. S40–S56, 2015.

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

JACKSON, A. P.; QUAIL, M. A.; BERRIMAN, M. Insights into the genome sequence of a free-living kinetoplastid: Bodo saltans (kinetoplastida: Euglenozoa). **BMC Genomics**, v. 9, p. 1–11, 2008.

JANGAM, D.; FESCHOTTE, C.; BETRÁN, E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. **Trends in Genetics**, v. 33, n. 11, p. 817–831, 2017.

JAYARAM, M. et al. An Overview of Tyrosine Site-specific Recombination: From an F1p Perspective. **Microbiology spectrum**, v. 3, n. 4, ago. 2015.

JOHN M, H. C.; SH, H.; VARMUS. **Retroviruses**. Cold Spring Harbor: Cold Spring

Harbor Laboratory Press, 1997.

JORDAN, I. K. et al. Origin of a substantial fraction of human regulatory sequences from transposable elements. **Trends in Genetics**, v. 19, n. 2, p. 68–72, 2003.

JURKA, J. et al. Repbase Update, a database of eukaryotic repetitive elements. **Cytogenetic and genome research**, v. 110, n. 1–4, p. 462–7, jan. 2005.

KAPITONOV, V. V.; JURKA, J. A universal classification of eukaryotic transposable elements implemented in Repbase. **Nature reviews. Genetics**, v. 9, n. 5, p. 411–2; author reply 414, maio 2008.

KAPITONOV, V. V.; MAKAROVA, K. S.; KOONIN, E. V. ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. **Journal of bacteriology**, v. 198, n. 5, p. 797–807, 28 dez. 2015.

KATO, H.; CÁCERES, A. G.; HASHIGUCHI, Y. First Evidence of a Hybrid of *Leishmania* (*Viannia*) *braziliensis*/L. (*V.*) *peruviana* DNA Detected from the Phlebotomine Sand Fly *Lutzomyia tejadai* in Peru. **PLoS neglected tropical diseases**, v. 10, n. 1, p. e0004336, jan. 2016.

KAUFER, A. et al. The evolution of trypanosomatid taxonomy. **Parasites & Vectors**, v. 10, n. 1, p. 287, 2017.

KAWAKAMI, K.; LARGAESPADA, D. A.; IVICS, Z. Transposons As Tools for Functional Genomics in Vertebrate Models. **Trends in genetics : TIG**, v. 33, n. 11, p. 784–801, nov. 2017.

KEIDAR, D.; DORON, C.; KASHKUSH, K. Genome-wide analysis of a recently active retrotransposon, Au SINE, in wheat: content, distribution within subgenomes and chromosomes, and gene associations. **Plant cell reports**, v. 37, n. 2, p. 193–208, fev. 2018.

KELLY, S. et al. A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. **Scientific Data**, v. 1, ago. 2014.

KIDWELL, M. G.; LISCH, D. Transposable elements as sources of variation in animals and plants. **Proceedings of the National Academy of Sciences of the United States of America**, v. 94, n. 15, p. 7704–11, jul. 1997.

KOHANY, O. et al. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. **BMC bioinformatics**, v. 7, p. 474, jan. 2006.

KUMAR, S.; STECHER, G.; TAMURA, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Molecular Biology and Evolution**, v. 33, n. 7, p. 1870–1874, jul. 2016.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 2001.

LAVIALLE, C. et al. Paleovirology of “syncytins”, retroviral env genes exapted for a role in placentation. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 368, n. 1626, p. 20120507, 19 set. 2013.

LISCH, D. R.; KIDWELL, M. G. Transposable Elements as Sources of Genomic Variation. In: **Mobile DNA II**. [s.l.] American Society of Microbiology, 2002. p. 59–90.

LORENZI, H. A.; ROBLEDO, G.; LEVIN, M. J. The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-encoding retrotransposons.

Molecular and Biochemical Parasitology, v. 145, n. 2, p. 184–194, 2006a.

LORENZI, H. A.; ROBLEDO, G.; LEVIN, M. J. The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-encoding retrotransposons.

Molecular and biochemical parasitology, v. 145, n. 2, p. 184–94, fev. 2006b.

LUDWIG, A.; KRIEGER, M. A. Genomic and phylogenetic evidence of VIPER retrotransposon domestication in trypanosomatids. **Memorias do Instituto Oswaldo Cruz**, v. 111, n. 12, p. 765–769, 2016.

LUDWIG, A.; LORETO, E. L. S. Evolutionary pattern of the gtwin retrotransposon in the *Drosophila melanogaster* subgroup. **Genetica**, v. 130, n. 2, p. 161–168, 2007.

LUDWIG, A.; VALENTE, V. L. S.; LORETO, E. L. S. Multiple invasions of Errantivirus in the genus *Drosophila*. **Insect molecular biology**, v. 17, n. 2, p. 113–24, abr. 2008.

LUKEŠ, J. et al. Evolution of parasitism in kinetoplastid flagellates. **Molecular and biochemical parasitology**, v. 195, n. 2, p. 115–22, jul. 2014.

MAJOREK, K. A. et al. The RNase H-like superfamily: New members, comparative structural analysis and evolutionary classification. **Nucleic Acids Research**, v. 42, n. 7, p. 4160–4179, 2014.

MALIK, H. S.; EICKBUSH, T. H. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. **Journal of virology**, v. 73, n. 6, p. 5186–5190, 1999.

MALIK, H. S.; EICKBUSH, T. H. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. **Genome research**, v. 11, n. 7, p. 1187–97, jul. 2001.

MALIK, H. S.; HENIKOFF, S.; EICKBUSH, T. H. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. **Genome Research**, v. 10, n. 9, p. 1307–1318, 2000.

MARCHLER-BAUER, A.; BRYANT, S. H. CD-Search: protein domain annotations on the fly. **Nucleic acids research**, v. 32, n. Web Server issue, p. W327–31, jul. 2004.

MART, S. et al. Gene Expression in Trypanosomatid Parasites. v. 2010, 2010.

MARTÍN, F. et al. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes. **Journal of molecular biology**, v. 247, n. 1, p. 49–59, mar. 1995.

MCCULLERS, T. J.; STEINIGER, M. Transposable elements in *Drosophila*. **Mobile genetic elements**, v. 7, n. 3, p. 1–18, 2017.

MILLER, M. A.; PFEIFFER, W.; SCHWARTZ, T. **Creating the CIPRES Science Gateway for inference of large phylogenetic trees** 2010 Gateway Computing Environments Workshop, GCE 2010. **Anais...IEEE**, nov. 2010

MILLER, R. K.; ROSE, M. D. Kar9p is a novel cortical protein required for cytoplasmic microtubule orientation in yeast. **The Journal of cell biology**, v. 140, n. 2, p. 377–90, jan. 1998.

MUÑOZ-LOPEZ, M. et al. Study of Transposable Elements and Their Genomic Impact. **Methods in molecular biology (Clifton, N.J.)**, v. 1400, p. 1–19, 2016.

MUSZEWSKA, A.; STECZKIEWICZ, K.; GINALSKI, K. DIRS and Ngaro

- Retrotransposons in Fungi. **PLoS ONE**, v. 8, n. 9, p. 1–10, 2013.
- NAKANO, K. et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. **Human cell**, v. 30, n. 3, p. 149–161, jul. 2017.
- NGOMTCHO, S. C. H. et al. Molecular screening of tsetse flies and cattle reveal different *Trypanosoma* species including *T. grayi* and *T. theileri* in northern Cameroon. **Parasites & vectors**, v. 10, n. 1, p. 631, 2017.
- NICHOLAS, K. B.; NICHOLAS, H. B. J.; DEERFIELD, D. W. I. Genedoc: a tool for editing and annotating multiple sequence alignments. **EMBNEW.NEWS**, v. 4, n. 14, 1997.
- NUNES-DÜBY, S. E. et al. Similarities and differences among 105 members of the Int family of site-specific recombinases. **Nucleic acids research**, v. 26, n. 2, p. 391–406, jan. 1998.
- O'DONOGHUE, P. Haemoprotozoa: Making biological sense of molecular phylogenies. **International Journal for Parasitology: Parasites and Wildlife**, v. 6, n. 3, p. 241–256, 2017.
- OPPERDOES, F. R. et al. Comparative Metabolism of Free-living Bodo saltans and Parasitic Trypanosomatids. **Journal of Eukaryotic Microbiology**, v. 63, n. 5, p. 657–678, 2016.
- PANAUD, O. Transferts horizontaux d'éléments transposables chez les eucaryotes : les gènes volants. **Comptes Rendus - Biologies**, v. 339, n. 7–8, p. 296–299, 2016.
- PASTUZYN, E. D. et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. **Cell**, v. 172, n. 1–2, p. 275–288.e18, 2018.
- 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.
- PECCOUD, J. et al. Massive horizontal transfer of transposable elements in insects. **Proceedings of the National Academy of Sciences**, v. 114, n. 18, p. 4721–4726, 2017.
- PIEDNOËL, M. et al. Eukaryote DIRS1-like retrotransposons: an overview. **BMC Genomics**, v. 12, n. 1, p. 621, 2011a.
- PIEDNOËL, M. et al. Eukaryote DIRS1-like retrotransposons: an overview. **BMC Genomics**, v. 12, n. 1, p. 621, 2011b.
- PORCEL, B. M. et al. The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. **PLoS genetics**, v. 10, n. 2, p. e1004007, fev. 2014.
- POULTER, R. T. M.; BUTLER, M. I. Tyrosine Recombinase Retrotransposons and Transposons. **Microbiology spectrum**, v. 3, n. 2, p. MDNA3-0036-2014, 2015a.
- POULTER, R. T. M.; BUTLER, M. I. Tyrosine Recombinase Retrotransposons and Transposons. **Microbiology spectrum**, v. 3, n. 2, p. MDNA3-0036-2014, abr. 2015b.
- POULTER, R. T. M.; GOODWIN, T. J. D. DIRS-1 and the other tyrosine recombinase retrotransposons. **Cytogenetic and genome research**, v. 110, n. 1–4, p. 575–88, 2005.
- RAMBAUT, A. **FigTree**, 2008.

- RAYMOND, F. et al. Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. **Nucleic Acids Research**, v. 40, n. 3, p. 1131–1147, fev. 2012.
- RONQUIST, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. **Systematic biology**, v. 61, n. 3, p. 539–42, maio 2012.
- ROSSATO, D. O. et al. *But2* Is a Member of the Third Major Group of hAT Transposons and Is Involved in Horizontal Transfer Events in the Genus *Drosophila*. **Genome Biology and Evolution**, v. 6, n. 2, p. 352–365, fev. 2014.
- RUNCKEL, C.; DERISI, J.; FLENNIKEN, M. L. A draft genome of the honey bee trypanosomatid parasite *Crithidia mellificae*. **PLoS one**, v. 9, n. 4, p. e95057, 2014.
- SÁNCHEZ-LUQUE, F. et al. Pr77 and L1TcRz: A dual system within the 5'-end of L1Tc retrotransposon, internal promoter and HDV-like ribozyme. **Mobile genetic elements**, v. 2, n. 1, p. 1–7, 1 jan. 2012.
- SCHAACK, S.; GILBERT, C.; FESCHOTTE, C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. **Trends in ecology & evolution**, v. 25, n. 9, p. 537–46, set. 2010.
- SCHWARZ, R. S. et al. Characterization of Two Species of Trypanosomatidae from the Honey Bee *Apis mellifera*: *Crithidia mellificae* Langridge and McGhee, and *Lotmaria passim* n. gen., n. sp. **Journal of Eukaryotic Microbiology**, v. 62, n. 5, p. 567–583, set. 2015.
- SHLOMAI, J. The assembly of kinetoplast DNA. **Parasitology Today**, v. 10, n. 9, p. 341–346, 1994.
- SMITH, D. F.; PEACOCK, C. S.; CRUZ, A. K. Comparative genomics: from genotype to disease phenotype in the leishmaniasis. **International journal for parasitology**, v. 37, n. 11, p. 1173–86, set. 2007.
- SMITH, M.; BRINGAUD, F.; PAPADOPOULOU, B. Organization and evolution of two *SIDER* retroposon subfamilies and their impact on the *Leishmania* genome. **BMC genomics**, v. 10, p. 240, 2009.
- SMYSHLYAEV, G. et al. Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. **Proceedings of the National Academy of Sciences of the United States of America**, v. 110, n. 50, p. 20140–5, 10 dez. 2013.
- SOUKY, S. M.; HUANG, J.; GOGARTEN, J. P. Horizontal gene transfer: building the web of life. **Nature reviews. Genetics**, v. 16, n. 8, p. 472–82, ago. 2015.
- SOUZA, W. DE. An Introduction to the Structural Organization of Parasitic Protozoa. p. 822–838, 2008.
- STAMATAKIS, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics (Oxford, England)**, v. 30, n. 9, p. 1312–3, maio 2014.
- STEVENS, J. R. Kinetoplastid phylogenetics, with special reference to the evolution of parasitic trypanosomes. **Parasite (Paris, France)**, v. 15, n. August, p. 226–232, 2008.
- STOCO, P. H. et al. Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. **PLoS Neglected Tropical Diseases**, v. 8, p. e3176, 2014.

SZITENBERG, A. et al. The evolution of tyrosine-recombinase elements in. **PLoS ONE**, v. 9, n. 9, 2014.

TEIXEIRA, M. M. G. et al. Phylogenetic Validation of the Genera *Angomonas* and *Strigomonas* of Trypanosomatids Harboring Bacterial Endosymbionts with the Description of New Species of Trypanosomatids and of Proteobacterial Symbionts. **Protist**, v. 162, n. 3, p. 503–524, jul. 2011.

TEIXEIRA, S. M. Control of gene expression in Trypanosomatidae. **Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica ... [et al.]**, v. 31, p. 1503–1516, 1998.

TENG, S. C.; WANG, S. X.; GABRIEL, A. A new non-LTR retrotransposon provides evidence for multiple distinct site-specific elements in *Crithidia fasciculata* minixon arrays. **Nucleic acids research**, v. 23, n. 15, p. 2929–36, ago. 1995.

TERZIAN, C.; PÉLISSON, A.; BUCHETON, A. Evolution and phylogeny of insect endogenous retroviruses. **BMC evolutionary biology**, v. 1, p. 3, 2001.

TREANGEN, T. J.; SALZBERG, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. **Nat Rev Genet.**, v. 13, n. 1, p. 36–46, 2013.

ULLU, E.; TSCHUDI, C.; CHAKRABORTY, T. RNA interference in protozoan parasites. **Cellular microbiology**, v. 6, n. 6, p. 509–19, jun. 2004.

VANHAMME, L.; PAYS, E. Control of gene expression in trypanosomes. **Microbiological reviews**, v. 59, n. 2, p. 223–240, 1995.

VAZQUEZ, M. et al. The short interspersed repetitive element of *Trypanosoma cruzi*, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons. **Proceedings of the National Academy of Sciences of the United States of America**, v. 97, p. 2128–2133, 2000.

VILLANUEVA, M. S. et al. A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of *Trypanosoma cruzi*. **Molecular and cellular biology**, v. 11, n. 12, p. 6139–48, dez. 1991.

VOLFF, J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. **BioEssays : news and reviews in molecular, cellular and developmental biology**, v. 28, n. 9, p. 913–22, set. 2006.

WALLAU, G. L.; ORTIZ, M. F.; LORETO, E. L. S. Horizontal transposon transfer in eukarya: Detection, bias, and perspectives. **Genome Biology and Evolution**, v. 4, n. 8, p. 689–699, 2012.

WALLAU, G. L.; VIEIRA, C.; LORETO, É. L. S. Genetic exchange in eukaryotes through horizontal transfer: connected by the mobilome. **Mobile DNA**, v. 9, n. 1, p. 6, 2018.

WESSLER, S. R. Transposable elements and the evolution of eukaryotic genomes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 47, p. 17600–1, nov. 2006.

WHEELER, R. J.; GLUENZ, E.; GULL, K. The limits on trypanosomatid morphological diversity. **PLoS ONE**, v. 8, n. 11, 2013.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, 2007.

- XIONG, Y.; EICKBUSH, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. **The EMBO journal**, v. 9, n. 10, p. 3353–62, 1990.
- YAHASHIRI, A.; JORGENSEN, M. A.; WEISS, D. S. The SPOR Domain, a Widely Conserved Peptidoglycan Binding Domain That Targets Proteins to the Site of Cell Division. **Journal of bacteriology**, v. 199, n. 14, jul. 2017.
- YAZAKI, E. et al. Global Kinetoplastea phylogeny inferred from a large-scale multigene alignment including parasitic species for better understanding transitions from a free-living to a parasitic lifestyle. **Advance Publication by J-STAGE Genes & Genetic Systems**, v. xx, n. xx, p. 1–25, 2017.
- YU, J. et al. Microhomology-dependent end joining and repair of transposon-induced DNA hairpins by host factors in *Saccharomyces cerevisiae*. **Molecular and cellular biology**, v. 24, n. 3, p. 1351–64, fev. 2004.

ANEXO A – SCRIPT *PERL* UTILIZADO PARA RECUPERAÇÃO DAS SEQUÊNCIAS

Código também disponível em CD

```

1  #Esse código recupera todas as sequencias de Viper e Tate de um genoma em dois arquivos.
2  #Previamente o blast local deverá ser realizado colocando as sequências query dos dois elementos em um mesmo arquivo e gerando resultado tabular.
3  #As sequências de hits plus/minus serão obtidas na fita correta.
4  open (IN1, "Genoma.fasta") || die "nao abre genoma"; $nome do arquivo do genoma+extensão
5  $entrada=<IN1>;
6  close IN1;
7  $contador=-1;
8  foreach $linha($entrada){
9      chomp $linha;
10     if ($linha=~>/>){
11         $contador++;
12         $cabecalho[$contador]=$linha;
13     }
14     else {
15         $sequencia[$contador]=$linha;
16     }
17 }
18 open SAIDA, ">viper.fasta";
19 open SAIDA2, ">tate.fasta";
20 open (IN2, "tabelablast") || die "nao abre 2"; $nome do arquivo da tabela de resultado do blast com 12 colunas
21 while ($linha=<IN2>){
22     chomp $linha;
23     @temp=split(/\t/, $linha); #splitar as colunas da tabela
24     if ($temp[0] =~ /viper/i) { #para salvar VIPER
25         if ($temp[8] <= $temp[9]) { #8=subject start; 9 = subject end
26             $inicial=$temp[8];
27             $final=$temp[9];
28         } elsif ($temp[8] >= $temp[9]) {
29             $final=$temp[8];
30             $inicial=$temp[9];
31         }
32         if ($inicial <= 8000) {
33             $inicial=1;
34             $final=$final+8000;
35         } else {
36             $inicial=$inicial-8000;
37             $final=$final+8000;
38         }
39         print "$inicial\t$final\t";
40         $trim=($temp[1], $inicial, $final); #trim=("nome do cromossomo", "posição inicial", "posição final";
41         for ($i=0; $i <= $contador; $i++){
42             if ($cabecalho[$i] =~ />/i) {
43                 if ($temp[8] > $temp[9]) {
44                     $nucleotideos=substr($sequencia[$i], $trim[1]-1, $trim[2]-$trim[1]);
45                     $nucleotideos =~ tr/ACGTacgt/TGCAtgca/;
46                     $seq{$temp[1]}=reverse $nucleotideos;
47                 } else {
48                     $nucleotideos=substr($sequencia[$i], $trim[1]-1, $trim[2]-$trim[1]);
49                     $seq{$temp[1]}= $nucleotideos;
50                 }
51                 print SAIDA">stableinfo[1]_stableinfo[2]_trim[0]_ position $trim[1] to $trim[2]\n$seq{$temp[1]}\n";
52             }
53         }
54     } else { #para salvar TATE
55         if ($temp[8] <= $temp[9]) { #8=subject start; 9 = subject end
56             $inicial=$temp[8];
57             $final=$temp[9];
58         } elsif ($temp[8] >= $temp[9]) {
59             $final=$temp[8];
60             $inicial=$temp[9];
61         }
62         if ($inicial <= 8000) {
63             $inicial=1;
64             $final=$final+8000;
65         } else {
66             $inicial=$inicial-8000;
67             $final=$final+8000;
68         }
69         print "$inicial\t$final\t";
70         $trim=($temp[1], $inicial, $final); #trim=("nome do cromossomo", "posição inicial", "posição final";
71         for ($i=0; $i <= $contador; $i++){
72             if ($cabecalho[$i] =~ />/i) {
73                 if ($temp[8] > $temp[9]) {
74                     $nucleotideos=substr($sequencia[$i], $trim[1]-1, $trim[2]-$trim[1]);
75                     $nucleotideos =~ tr/ACGTacgt/TGCAtgca/;
76                     $seq{$temp[1]}=reverse $nucleotideos;
77                 } else {
78                     $nucleotideos=substr($sequencia[$i], $trim[1]-1, $trim[2]-$trim[1]);
79                     $seq{$temp[1]}= $nucleotideos;
80                 }
81                 print SAIDA2">stableinfo[1]_stableinfo[2]_trim[0]_ position $trim[1] to $trim[2]\n$seq{$temp[1]}\n";
82             }
83         }
84     }
85 }
86

```


CiaC1_28_2171120-2187941	degenerada										1251 416		7631	8881	sem domínio conservado				2994 997		7462	10455	RT_ like super family RNase_H_ like super family	642-685 877-993	9.02e-03 1.08e-10									
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	GRS1 super family										1416 471		6494	7909	157-241		4.15e-03																	
	degenerada																										degenerada							
	degenerada																										degenerada							
	degenerada																										degenerada							
C. bombi	pblaa										pblaa	início	final					pblaa	início	final			domínio	posição	e-value									
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	região faltante										região faltante																região faltante							
	região faltante										região faltante																região faltante							
	degenerada										degenerada																degenerada							
	região faltante										região faltante																região faltante							
	degenerada										degenerada																degenerada							
	região faltante										região faltante																região faltante							
	região faltante										região faltante																região faltante							
C. mellificae	pblaa										pblaa	início	final					pblaa	início	final			domínio	posição	e-value									
	SMC_N super family										1695 564	241	1935	3-191	5.23e-07	sem domínio conservado														RT_ like super family		120-166	6.06e-03	
	Atrophin-1 super family										região faltante																RNase_H_ RT_ DIRS1		353-469	1.58e-13				
	região faltante										região faltante																RT_ like super family		20-73	2.53e-04				
	região faltante										região faltante																RNase_H_ like super family		261-377	3.55e-09				
	região faltante										região faltante																degenerada							
	região faltante										região faltante																degenerada							
	região faltante										região faltante																degenerada							
	região faltante										região faltante																degenerada							
	região faltante										região faltante																degenerada							
L. braziliensis M2903	pblaa										pblaa	início	final					pblaa	início	final			domínio	posição	e-value									
	degenerada										degenerada																domesticated gene (Ludwig e Krieger 2016)				degenerada			
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
L. braziliensis M2904	pblaa										pblaa	início	final					pblaa	início	final			domínio	posição	e-value									
	degenerada										degenerada																domesticated gene (Ludwig e Krieger 2016)				degenerada			
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							
	degenerada										degenerada																degenerada							

NBCO01000004.1_941135_957797	1212 403	6356	7567	ToIC HemX	42+58 106-178	4.38e-04 3.29e-04	degenerada				960 319	7886	8845	sem domínio conservado	degenerada	2514 837	7868	10381	RT_ like super family RNase H_ like super family	377-489 678-797	1.33e-05 8.41e-08
NBCO01000001.1_184039_200881	1266 421	6488	7753				região faltante									1974 657	8065	10038	RT_ like super family	370-557	5.16e-06
NBCO01000008.1_892435_899754							região faltante												região faltante		
NBCO01000035.1_148891_165733							degenerada												degenerada		
NBCO01000021.1_1_3195							degenerada												degenerada		
NBCO01000040.1_266531_282172							região faltante												região faltante		
NBCO01000018.1_233889_50971							degenerada												degenerada		
NBCO01000038.1_237362_253664							degenerada												degenerada		
NBCO01000037.1_111595_128803							degenerada												degenerada		
NBCO01000036.1_4776_21153							degenerada												degenerada		
NBCO01000035.1_232086_248910							degenerada												degenerada		
T. vivax																					
Ty486_11_2839738 to 2856622	pblaa	início	final	domínio	posição	e-value					pblaa	início	final	domínio	degenerada				domínio	posição	e-value
Ty486_11_4546517 to 4563137				degenerada											degenerada				degenerada		
Ty486_11_550849 to 567142				degenerada											degenerada				degenerada		
Ty486_11_551138 to 567608				degenerada											degenerada				degenerada		
Ty486_11_4339338 to 4356285				degenerada											degenerada				degenerada		
Ty486_11_3128323 to 3145228				degenerada											degenerada				degenerada		
Ty486_11_159615 to 176364				degenerada											degenerada				degenerada		
Ty486_11_3725402 to 3741731				degenerada											degenerada				degenerada		
Ty486_09_1957548 to 1974051				degenerada											degenerada				degenerada		
Ty486_09_314916 to 331272				degenerada											degenerada				degenerada		
Ty486_10_430434 to 447264				degenerada											degenerada				degenerada		
Ty486_03_1 to 16403				degenerada											degenerada				degenerada		
Ty486_03_512004 to 528744				degenerada											degenerada				degenerada		
Ty486_08_426680 to 443532				degenerada											degenerada				degenerada		
Ty486_02_519592 to 536476				degenerada											degenerada				degenerada		

ANEXO C – DESCRIÇÃO DAS CÓPIAS DE TATE ANALISADAS

Tabela que sumariza as análises das 15 cópias mais conservadas de cada espécie para o elemento *TATE*. As informações indicadas são: *Contigs/scaffolds* de cada cópia com posição inicial e final da região isolada, tamanho de cada ORF predita em pares de base (pb) e aminoácidos (aa) e a posição inicial e final da ORF na sequência isolada, domínio encontrados na proteína com sua posição na sequência e *e-value* da busca. Tabela disponível também em CD.

Genoma/Posição cromossômica	ORF1				ORF2				ORF3			
	ORF1	início	final	domínio	posição	e-value	início	final	domínio	posição	e-value	domínio
<i>B. subtilis</i> CYKH01000290.1	5226 1741	1157	6382	DNA_BRE_C	894-1032	2.55e-07	Gag-like, YR and RT/RH in a single ORF					
				RVT_1 super family	1294-1377	1.87e-03						
	4395 1464	371	4765	DNA_BRE_C	640-783	4.91e-05	Gag-like, YR and RT/RH in a single ORF					
CYKH01000240.1				RVT_1 super family	1052-1122	1.05e-04						
CYKH01000124.1				degenerada			degenerada					
CYKH01001237.1				degenerada			degenerada					
CYKH01001135.1				degenerada			degenerada					
<i>A. deanei</i>	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value
KV452613.1				degenerada			1527 508	6676	8202	sem domínio conservado		
KV452575.1				região faltante						região faltante		
KV452489.1				região faltante						região faltante		
LXWQ01001884.1				região faltante						região faltante		
KV452523.1				degenerada						degenerada		
KV452539.1				degenerada						degenerada		
KV452469.1				degenerada						degenerada		
KV452484.1				degenerada						degenerada		
KV452489.1				região faltante			1152 383	8065	9216	DNA BRE C super family	381-500	7.92e-03
KV452604.1				degenerada			1068 355	7518	8595	sem domínio conservado		
<i>B. ayala</i>	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value
BayA_534				região faltante						degenerada		
BayA_481				degenerada						degenerada		
BayA_396				degenerada						degenerada		
BayA_076				degenerada						degenerada		
BayA_525				degenerada						degenerada		
BayA_265				degenerada						degenerada		
BayA_052				degenerada						degenerada		
BayA_381				degenerada						degenerada		
BayA_016				degenerada						degenerada		
BayA_354				degenerada						degenerada		
BayA_535				degenerada						degenerada		
BayA_432				degenerada						degenerada		
BayA_045				degenerada						degenerada		

BayA_130	degenerada				degenerada				degenerada			
BayA_459	região faltante				degenerada				degenerada			
C. fasciculata	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value
CfaC1_30_2950106to2965213	degenerada				degenerada				degenerada			
CfaC1_30_2490278to2508005	degenerada				DNA_BRE_C				2.34e-03			
CfaC1_30_944521to960832	degenerada				degenerada							
CfaC1_30_511957to529090	degenerada				ULP1 super family				2.50e-04			
CfaC1_28	degenerada				degenerada							
CfaC1_05	degenerada				degenerada							
CfaC1_27	degenerada				DNA_BRE_C				1.04e-04			
CfaC1_14	degenerada				DNA_BRE_C super family				2.74e-07			
CfaC1_23	degenerada				degenerada							
CfaC1_22	degenerada				degenerada							
CfaC1_26	1385 454	4901	626	5	sem domínio conservado							
CfaC1_11	degenerada				degenerada							
CfaC1_18	degenerada				ULP1 super family				3.86e-04			
CfaC1_16	1125 374	4451	557	5	sem domínio conservado				3.86e-03			
CfaC1_19	degenerada				DNA_BRE_C				2.62e-09			
E. montrogei	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value
	1101 366	8001	910	1	sem domínio conservado							
KB312623.1	região faltante				região faltante							
AOF501002192.1	degenerada				degenerada							
KB312616.1	região faltante				região faltante							
AOF501001613.1	região faltante				DNA_BRE_C				2.04e-06			
KB312643.1	degenerada				degenerada							
AOF501002410.1	região faltante				região faltante							
AOF501001643.1	região faltante				degenerada							
AOF501002159.1	1098 365	100	119	7	sem domínio conservado							
L. aethiopica	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value
KE598534.1	degenerada				degenerada							
KE598637.1	degenerada				degenerada							
KE598529.1	degenerada				degenerada							
AUMB01001235.1	degenerada				degenerada							
L. amazonensis	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value
KE391216.1	degenerada				degenerada							
KE391474.1	degenerada				degenerada							
KE390981.1	degenerada				degenerada							

	1143 380	8000	914 2	PUFD_like super family	183-278	3.76e-03	2919 972	9250	12168	DNA_BRE_C super family	807-951	1.65e-04	2862 963	11280	14141	RT_like super family	Peptidase_A17 super family	591	5.71e-04
LbrM28_1 to 15939										DNA_BRE_C super family									
										DNA_BRE_C super family	783-946	1.15e-04	2862 963	0	14141				
										DNA_BRE_C	1029-1188	1.57e-06							
LbrM34_1988145 to 2004588				Sobrepostas com ORF2			3642 1213	8480	12121	SPOR super family	401-539	5.05e-05	2862 963	11233	14094				
										DNA_BRE_C	788-947	2.30e-06							
LbrM00_60_575438-582815_1 to 9633	1194 397	467	1660	sem domínio conservado			2919 972	1768	4686	SPOR super family	160-298	1.26e-05	2862 963	3798	6659				
										DNA_BRE_C	165-303	2.74e-03							
LbrM24_813439 to 832264				degenerada			2514 837	5941	8454	SPOR super family			2862 963	8001	10862				
										DNA_BRE_C super family	783-946	1.15e-04	2862 963	8000	10861				
LbrM28_1 to 12659	1143 380	4720	5862	PUFD_like super family	183-278	3.76e-03	2919 972	5970	8888	DNA_BRE_C super family									
										DNA_BRE_C	263-318	8.08e-03							
LbrM30_147253 to 164377	1332 443	7797	9128	sem domínio conservado			2622 873	9221	11842	tRNA-synt_1g super family									
										DNA_BRE_C super family	287-431	8.33e-06	2652 883	8000	10651				
LbrM00_41_423585-436033_1 to 12369				degenerada			1374 457	7515	8888	DNA_BRE_C super family									
				missing						DNA_BRE_C	689-849	5.95e-07							
LbrM00_85_724234-731913_1 to 13651							2625 874	293	2917	SPOR super family	62-200	6.75e-06	3129 1042	1762	4890				
L. donavani	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value	ORF3	início	final	domínio	posição	e-value	
Ld32				degenerada						degenerada									
				degenerada						degenerada									
Ld30 (ver posição)				degenerada						degenerada									
Ld30 (ver posição)				degenerada						degenerada									
Ld20 (ver posição)				degenerada						degenerada									
Ld20 (ver posição)				degenerada						degenerada									
L. entretiti	ORF1	início	final	domínio	posição	e-value	ORF2	início	final	domínio	posição	e-value	ORF3	início	final	domínio	posição	e-value	
KE48264.1				degenerada						degenerada									
KE48263.1				degenerada						degenerada									
KE48258.1				degenerada						degenerada									
KE48217.1				degenerada						degenerada									
KE48204.1				degenerada						degenerada									
KE48275.1				degenerada						degenerada									
KE48290.1				degenerada			1170 389	7844	9013	DNA_pol3_gamma3 super family	33-170	5.09e-05					região faltante		
				degenerada						EBV-NA3 super family	19-302	8.82e-05							
KE48233.1				degenerada						degenerada									
KE48288.1				degenerada						degenerada									
KE48222.1				degenerada						degenerada									
KE48304.1				região faltante						degenerada									
KE48280.1				degenerada			1323 440	6745	8067	sem domínio conservado									

KE148238.1	degenerada				degenerada				degenerada			
KE148287.1	degenerada				1086 361	6798	7883	sem domínio conservado				degenerada
KE148306.1	região faltante				região faltante				degenerada			
L. gerbilli	ORF1	início	final	domínio	ORF2	início	final	domínio	ORF3	início	final	domínio
KE561409.1	degenerada				degenerada				degenerada			
KE561408.1	degenerada				degenerada				degenerada			
KE561439.1	degenerada				degenerada				degenerada			
KE561431.1	degenerada				degenerada				degenerada			
KE561412.1	degenerada				degenerada				degenerada			
ATBK01001157.1	degenerada				degenerada				degenerada			
L. infantum	ORF1	início	final	domínio	ORF2	início	final	domínio	ORF3	início	final	domínio
LinJ_32	degenerada				degenerada				degenerada			
LinJ_30_p1272412to1289083	degenerada				degenerada				degenerada			
LinJ_30_p1036645to1063277	degenerada				degenerada				degenerada			
LinJ_20_p64936to81340	degenerada				degenerada				degenerada			
LinJ_20_p320708to337028	degenerada				degenerada				degenerada			
LinJ_24	degenerada				degenerada				degenerada			
L. major	ORF1	início	final	domínio	ORF2	início	final	domínio	ORF3	início	final	domínio
LinJF_32	degenerada				degenerada				degenerada			
LinJF_30_p1315453to1331845	degenerada				degenerada				degenerada			
LinJF_30_p1075153to1091545	degenerada				degenerada				degenerada			
LinJF_13	degenerada				degenerada				degenerada			
L. mexicana	ORF1	início	final	domínio	ORF2	início	final	domínio	ORF3	início	final	domínio
LinMxM_20_p2701039to2717500	degenerada				degenerada				degenerada			
LinMxM_24	degenerada				degenerada				degenerada			
LinMxM_12	degenerada				degenerada				degenerada			
LinMxM_29	degenerada				degenerada				degenerada			
LinMxM_20_p2947223to2963704	degenerada				degenerada				degenerada			
L. panamensis	ORF1	início	final	domínio	ORF2	início	final	domínio	ORF3	início	final	domínio
	1218 405	330	1547	sem domínio conservado	região faltante				região faltante			
	AOND01002634.1_1.9805				região faltante				degenerada			
	KB453324.1				região faltante				região faltante			
	AOND01002656.1				região faltante				região faltante			
	KB453475.1				região faltante				região faltante			
	KB453565.1				degenerada				região faltante			
	KB453290.1	8000	9472	Lon super family	região faltante				região faltante			
	KB453322.1				região faltante				região faltante			
	AOND01002915.1				região faltante				região faltante			
AOND01003046.1	região				degenerada				região			

MDUD01002261.1_1 to 8713										degenerada										degenerada									
MDUD01008473.1_1 to 8419										degenerada										degenerada									
MDUD01004842.1_1 to 8805										degenerada										degenerada									
MDUD01001414.1_1 to 9037										degenerada										degenerada									
L. tropica										degenerada										degenerada									
XE147471.1_46268 to 63608										degenerada										degenerada									
XE147314.1_234387 to 251076										degenerada										degenerada									
XE147331.1_198221 to 214763										degenerada										degenerada									
XE147387.1_42881 to 59004										degenerada										degenerada									
XE147405.1_70117 to 86239										degenerada										degenerada									
XE147318.1_478985 to 495101										degenerada										degenerada									
XE147336.1_191854 to 208120										degenerada										degenerada									
L. turanica										degenerada										degenerada									
XE537394.1_330719 to 347141										degenerada										degenerada									
XE537429.1_12042 to 28308										degenerada										degenerada									
Leishmania sp. MARLEM2494										degenerada										degenerada									
LMARLEM2494_09_259835 to 277412										degenerada										degenerada									
LMARLEM2494_28_375015 to 392166										degenerada										degenerada									
LMARLEM2494_30_1118759 to 1138850										degenerada										degenerada									
LMARLEM2494_01_235741 to 252601										degenerada										degenerada									
LMARLEM2494_11_159126 to 175839										degenerada										degenerada									
LMARLEM2494_15_396525 to 413214										degenerada										degenerada									
LMARLEM2494_24_1 to 9208										degenerada										degenerada									
LMARLEM2494_13_104217 to 120759										degenerada										degenerada									
LMARLEM2494_35_418182 to 434724										degenerada										degenerada									
LMARLEM2494_08_363895 to 380425										degenerada										degenerada									
LMARLEM2494_36_2610406 to 2626831										degenerada										degenerada									
LMARLEM2494_17_368696 to 385085										degenerada										degenerada									
LMARLEM2494_27_361628 to 378032										degenerada										degenerada									
LMARLEM2494_19_565386 to 581688										degenerada										degenerada									
LMARLEM2494_05_344880 to 361482										degenerada										degenerada									
Lep. pyrthocoris										degenerada										degenerada									
LpyrH10_14_696280 to 713264										sem domínio conservado										domínio									
LpyrH10_03_990765 to 1008837										sem domínio conservado										sem domínio conservado									
LpyrH10_16_467696 to 485768										sem domínio conservado										sem domínio conservado									
LpyrH10_01_2575473 to 2593542										sem domínio conservado										sem domínio conservado									
LpyrH10_37_13614 to 31398										sem domínio conservado										sem domínio conservado									
LpyrH10_45_1 to 10050										sem domínio conservado										sem domínio conservado									
LpyrH10_44_1 to 16569										sem domínio conservado										sem domínio conservado									

[illegible]

ANEXO D – ALINHAMENTO DO GENE DOMESTICADO DE VIPER COM GENE PREDITO PARA *A. deanei*

Visualização do alinhamento, em duas partes, das sequências de aminoácidos do gene domesticado reportado por LUDWIG e KRIEGER (2016) das diferentes espécies (seis primeiras sequências) juntamente a proteína predita para o gene de *Angomonas deanei* e a porção C-terminal da proteína RT/RH de *VIPER* de *T. cruzi*.

Parte 1

```

LpyrH10_03 : MAE-----LHMTTPFNADVCTSSSIDDDPRMISYDHEVELEPRAEYETYPYSACRCRAVIRSVVLWAMNSTOTOFVIAWNNMNNRLGQAL : 87
Chritidia_ : MAEPVVNAPGCTAAPYTNKPPDPFPHVITWSPFIDDDPRMISYDHEVELEPRAEYETYPYSACRCRAVIRSVVLWAMNSTOTOFVIAWNNMNNRLGQAL : 98
LMARLEM249 : MTTI-----SARRRDPFPHVATWYRFIEQDDPRMITCELRGELESARDVHPYSACRCROSVVKSLLWAAASACAAFPWNGETNNLQGLQNL : 87
LBRM2903_3 : MNH-----SSSIPFIADHVVWVSCFIEQDDPRMITCELRGELESAREYYPYSACRCROSLISSLLWAGIARACQHVHQWNSTDNINLGGGL : 86
LbrM.33.34 : MNH-----SSSIPFIADHVVWVSCFIEQDDPRMITCELRGELESAREYYPYSACRCROSLISSLLWAGIARACQHVHQWNSTDNINLGGGL : 86
LPAL13_330 : MNH-----SSSIPFIADHVVWVSCFIEQDDPRMITCELRGELESAREYYPYSACRCROSLISSLLWAGIARACQHVHQWNSTDNINLGGGL : 86
Angomonas : MSL-----DEE-----DDYEDPSOIGMFLC----- : 21
VIPER : ----- : -

m s f f v t w i q d d r m i l r e l e a v s a c r q s l w a r r w n l g q l

LpyrH10_03 : VOHMLDLTAEELGVTNTEIYAAALCAECHPFLRMHEIEQCSKTHEGWSRHRDRAAAAIETATSTTIYALSPLLGVSQCRMPORAEITVAPL-- : 182
Chritidia_ : VOHMLDLTAEELGVTNTEIYAAALCAECHPFLRMHEIEQCSKTHEGWSRHRDRAAAAIETATSTTIYALSPLLGVSQCRMPORAEITVAPL-- : 189
LMARLEM249 : VOHMLDLTAEELGVTNTEIYAAALCAECHPFLRMHEIEQCSKTHEGWSRHRDRAAAAIETATSTTIYALSPLLGVSQCRMPORAEITVAPL-- : 178
LBRM2903_3 : IOHIFDLTAEMCGINTTIIYAAALCAECHPFLRMHEIEQCSKTHEGWSRHRDRAAAAIETATSTTIYALSPLLGVSQCRMPORAEITVAPL-- : 177
LbrM.33.34 : IOHIFDLTAEMCGINTTIIYAAALCAECHPFLRMHEIEQCSKTHEGWSRHRDRAAAAIETATSTTIYALSPLLGVSQCRMPORAEITVAPL-- : 177
LPAL13_330 : IOHIFDLTAEMCGINTTIIYAAALCAECHPFLRMHEIEQCSKTHEGWSRHRDRAAAAIETATSTTIYALSPLLGVSQCRMPORAEITVAPL-- : 177
Angomonas : -----CLOEPEKSKIFA-----DS-SV-----RIANPKR-PRFPSILGRNEGNSAPSSSVLQTEGKQN : 75
VIPER : ----- : -

q h d l t a e g d i a a l a e h p f l k g w g e r p p e

LpyrH10_03 : -ARESLAAAASE--FAI---LQTVIAETPPP-----ATSPGAENGRF-----S---KAVTK--AMGSKFA--QP-KFA-A-AATA----- : 245
Chritidia_ : -ARESLAPOP-----NGCAEAPIH-----V-AVAVDIAPOEGAAEESR--WSALPPPPVALKPEPPT--FAVKAQP-FLPAETKAV : 258
LMARLEM249 : LTSEPTVKSEPALTHDAQALPAKMAIAPQTPQGTTPPSPFPFGVAFVSVKYLKPAALGNTAGYGTAP--STLMQGSVVPVNAIE-RAFAHRMP-- : 268
LBRM2903_3 : -SREQLPRESAEIYPCVETPFTNKASPTTM-----LYPVGYCNETLQNAADIR-----GGVLS--AKLLKDTA--CGETK-RP-EVSLPS-- : 254
LbrM.33.34 : -SREQLPRESAEIYPCVETPFTNKASPTTM-----LYPVGYCNETLQNAADIR-----GGVLS--AKLLKDTA--CGETK-RP-EVSLPS-- : 254
LPAL13_330 : -SREQLPRESAEIYPCVETPFTNKASPTTM-----LYPVGYCNETLQNAADIR-----GGVLS--AKLLKDTA--CGETK-RP-EVSLPS-- : 254
Angomonas : NVGSEKDKPPQFAVLSSSTTTP-----RQ---NGVLA--DP--TD--TM--NSQTE-MK-KIRHEE-- : 124
VIPER : ----- : -

e p

LpyrH10_03 : -----EPTTRPVKVVAAA--FA--V-----ESKIAAARFPFP--FFVAAVNATKANGHAAPPPPSKKKEESERTP----- : 305
Chritidia_ : TAQPPTPNAPAFSAARF--P--PSEMATQKSKSNGHAAPPPPHIPFAHAVTRETLVRKVVPFPQ-----PAAPAIAAA-----KAA : 336
LMARLEM249 : -----ASRSVSATVNE--KPLRS-----KGQAGAAAAFAA--PSSTEASSLRRAAQPVVLRQESYTVRPHQVSGQEPFEGLATQSMCV : 344
LBRM2903_3 : -----ASKIVANHAND--KPLRS-----KGDIAKPADVFAA--SLIRDSLPPESVPSAAIYRLIHAVDTAPSVST-----KGV : 320
LbrM.33.34 : -----ASKIVANHAND--KPLRS-----KGDIAKPADVFAA--SLIRDSLPPESVPSAAIYRLIHAVDTAPSVST-----KGV : 320
LPAL13_330 : -----ASKIVANHAND--KPLRS-----KGDIAKPADVFAA--SLIRDSLPPESVPSAAIYRLIHAVDTAPSVST-----KGV : 320
Angomonas : -----APSPHPNQNNSGPFVFM-----AGSSNVFLA--E-----RQYRMPGSGGIPYMTMPSNTPHVTHTFPPTYPGNIFFPH : 192
VIPER : ----- : -

a p

LpyrH10_03 : -----GV-----SLRPRPPP--PPFA--QP-----AAFAAPKNEALVQAEKEVYSDSEEFLLSS : 351
Chritidia_ : PVPEFAVKNLPQ-----KKVALAAP--FSPPVAKKAV-----AAPTKPQPCVHVHPKKQNTYSDSEES-- : 394
LMARLEM249 : AAKDVAGSVAAD-----LA--VKK--AVDSSDRTSAPSTIALKSPTVEVLQSSSPFAESVRP--EFSKKT--TYSDSEGT-- : 415
LBRM2903_3 : -----TA-----SSK--FATRLFAKSDVNFTSATKPFAAAAAPLLPAFAAAKEPPYIPKKVTIVYTDSDS-- : 393
LbrM.33.34 : VAKPIARNGEAR-----TA-----SSK--FATRLFAKSDVNFTSATKPFAAAAAPLLPAFAAAKEPPYIPKKVTIVYTDSDS-- : 393
LPAL13_330 : VAKPIARNGEAR-----TA-----SSK--FATRLFAKSDVNFTSATKPFAAAAAPLLPAFAAAKEPPYIPKKVTIVYTDSDS-- : 393
Angomonas : YGGPLPRYAWNFPQTVHCPYQANTPSPFQ-----TRNLQVNVKPPSQTSVPFT-----PVSTRYSPQTNHT : 257
VIPER : ----- : -

p p p p y d s

LpyrH10_03 : RSSSAAAPAPKTAANVAAKMAEPFVSFPR---GATALATEAELLKYGHGDTENLLQAC---L-----DEMADPDGVSAAITELKLAVERKRRK : 437
Chritidia_ : -SVSSSEDPAGSPPAASLPQSMMPTEFATNGSVKAAPAVVLGEAELLRYGADGPDLLQAC---W-----SDAADPDGLSAIFDRKLAVTESRRK : 482
LMARLEM249 : -TESDTSAPAPSTPVRTASAPSTTEIAATTPS-----TV-PAALLKMRDGGTDLQAC---W-----SDLSVDPDGISAISERLAITEDGQOI : 496
LBRM2903_3 : -DDVTSSAVVHDINASASASGLALETIV-----TEALLKYGDGKSADLMQAF-----W-----NDLPADPDGVSIVFGRLAITEDGROV : 467
LbrM.33.34 : -DDVTSSAVVHDINASASASGLALETIV-----TEALLKYGDGKSADLMQAF-----W-----NDLPADPDGVSIVFGRLAITEDGROV : 467
LPAL13_330 : -DDVTSSAVVHDINASASASGLALETIV-----TEALLKYGDGKSADLMQAF-----W-----NDLPADPDGVSIVFGRLAITEDGROV : 467
Angomonas : -----PFA-----GTADG-----RAALLRLAMSPGFAEDEKCEAMTAPRTAPPGVRFPTAEEQTV : 310
VIPER : ----- : -

a p p l l y l q a w a d p d g w l a t 2 d

```

Parte 2

```

*      600      *      620      *      640      *      660      *      680
LpyrH10_03 : RLIAEFPFLLNINRSRMEVRYVKRLDOLFQYLLADAPLELLEH-----NDAARFLAADIAADAVGLIRESRRAGPDT--VIAFSVMDGTEQET : 528
Chritidia_ : PLCVEVPFSAISVEKVRARMEKRVAKRIEHLFOYLSADAPVPELRLH-----EAAARFLAADVDRISTARLIRESRQTEGPDT--VIAFSVVDKAEKDP : 573
LMARLEM249 : ELEVAFLAFLSLDLKLRMRMEKRVKRVLDOLFQYLLHTEFAAALPEH-----QEVSLILSGDQTECLRAAEIRESRRKSAGPDT--VLAFTVVEKSKKET : 587
LBRM2903_3 : FLMAAASPPILCLERVKORMQSKVAKRLDOLFQYLLHABAPTPLPEER-----NEVAALLSARDIECLEAAKLIREFCQSAGPDT--VLAFTVVEKATNA : 558
LbrM.33.34 : FLMAAASPPILCLERVKORMQSKVAKRLDOLFQYLLHABAPTPLPEER-----NEVAALLSARDIECLEAAKLIREFCQSAGPDT--VLAFTVVEKATNA : 558
LPALL3_330 : FLMAAASPPILCLERVKORMQSKVAKRLDOLFQYLLHABAPTPLPEER-----NEVAALLSARDIECLEAAKLIREFCQSAGPDT--VLAFTVVEKATNA : 558
Angomonas : FLRHVTVPPPLPLEQLRSRFENKKTNRERLWGC-----ILVRSNLS-----HAFATLRCAAEVIVVACMPAVPSSAETRGWVAEFTILMEAT--G : 399
VIPER : ELQOVNVFVNLNLEWIRSRINFATLERIMOVWGLVG---RSPFFSSSGKARIGSRRIPTAARLIRKAGLIEDASSITGGW---IFSVVVEEKITGL : 104
pL a ppL 6 64 R v kR q65qy ap l p e aa 6s D L a 6Irp k agpdt 6 aF366

```

```

*      700      *      720      *      740      *      760      *      780
LpyrH10_03 : TRRFSLWPC-GSNRNAMIDA-DPRVSLHHISTYLDVAVAFAGCAESIAACFPDQILHDMQERFVFTDTSRRWCLTRLPMTHVAEIVQIILCLC : 625
Chritidia_ : SRRFLSWPC-GSNRSALDGYEPHVLHHISSYLDVAVAAAGCAETLMSSFDQVLPNPVQERFVFTDAKRWCLTRLPMTHVAEIVMILLSTLC : 670
LMARLEM249 : RRRFLSWPR-GSNEQVILSSGYEPHVLHHISTYLDVAVNAPGCVSELENCFDQVLPDQLERFLTDARERWCLTRLPLGHIVSMEIVQIILSTLC : 684
LBRM2903_3 : RRRFLTWPR-GSNQHTILSGYESHPLHHISTYLDVAVNESGCFVQSIFFGCFDQVLPDQLERFVFTDARDHRWCLTRLPLGHIVSMEIVQIIVSTLS : 655
LbrM.33.34 : RRRFLTWPR-GSNQHTILSGYESHPLHHISTYLDVAVNESGCFVQSIFFGCFDQVLPDQLERFVFTDARDHRWCLTRLPLGHIVSMEIVQIIVSTLS : 655
LPALL3_330 : RRRFLTWPR-GSNQHTILSGYESHPLHHISTYLDVAVNESGCFVQSIFFGCFDQVLPDQLERFVFTDARDHRWCLTRLPLGHIVSMEIVQIIVSTLS : 655
Angomonas : RRRFMFWPLGFFREEVKDAPFKIOWEHIKIGQS--EASAAI--VVEDACQRLDLPFGCRARHRLTDEAGTLTDLTRLPTPCPPSEMIHLLLSALV : 493
VIPER : RRRWFWPR-DKRR---DPEANVRLIHSHVLPFVMAEASCLDTRASEFQVSLHRETRHLRCRVEDGLVLVLTRLPMGYKASHEILQIITIAHA : 198
RR56 WP gsn d ye 6p6hHis yl av g 6 fdq6 LPp qer5 ftda r cLTRLP h vs E66q66 st6

```

```

*      800      *      820      *      840      *      860      *      880
LpyrH10_03 : GHFSYTLPRFAADSAVLVVDVVRHVRYFYAQKGVCQSRDKFRACKECSVAQOLNKGK-GGGVVFIFGVHFDHTAHVVCVEPELLRRIA--AFRRLLN : 720
Chritidia_ : GHFSYTAQHSFAISPEVSDVVRVTRVPYCTQAAACARNKFRACKECCGTLQPCATD-TSGVYHFGVHFDHDAHSVVEPEMVRRIA--AFRCSTH : 765
LMARLEM249 : GHFSYTLQSHATANAVHSDVVRVTRVLYGPNVDVCKARKKWPACIDCSATASSGGGNATHEFLGVVFDHGAHEVSVADDLRVE--SKRRMS : 780
LBRM2903_3 : GHFSYTLQSYAANAVRIDVNVYVYTABRCGPAKDVQSASAAARACDECCGIVTTSWRDVPTSDNFFLGVHFDHSAHEVSVASDILSVD--LVRVVIS : 751
LbrM.33.34 : GHFSYTLQSYAANAVRIDVNVYVYTABRCGPAKDVQSASAAARACDECCGIVTTSWRDVPTSDNFFLGVHFDHSAHEVSVASDILSVD--LVRVVIS : 751
LPALL3_330 : GHFSYTLQSYAANAVRIDVNVYVYTABRCGPAKDVQAASAAARACDECCGIVTTSWRDVLTSDNFFLGVHFDHSAHEVSVASDILSVD--LVRVVIS : 751
Angomonas : KTEPTVSLPHLVSKGRSKVVVVDVRLFYGETQRVQKGKDECVRQAQAGITFH-HACE-VKLESEFLGIHEDHIQRSVSLSDSIRIHAHSLFNQLT : 589
VIPER : GVTGVVHPPLAAAPPLVRVDVVIDNIRIAGSKSATLWEAQLNRNADSRHTMGEDRES-GATQTFELGVCFDHTHRAVSLSDKIVRSVRAMPALNSLT : 295
ghp3y l a v dVw6 r G v aac c a y F G6hFDH ah Vs6 6 6 p

```

```

*      900      *      920      *      940      *      960      *      980
LpyrH10_03 : TRDLERQYTLQLLSAAVTHIAVSKEVTHARTVYWRLSQTFAEVELLKEELPQDSFAHLEKWKHDAHHAARSAKVALNMT--LFWELSLSWFA : 816
Chritidia_ : TRDLERHYTLQLLSASVLGVAVSREKRLRTVWRLSAALENEDTLKVEPLPADSYAQLEKNHKSVLHAAPRQVKTESLENMT--LYWELSSSWF : 861
LMARLEM249 : TRDLERYTLQLLFCSAVLQVAVSEYVYVLYVYVRLSTLQEHPEQLHTESEPLECTCTQVQRWHEALLKTAARRVGVVSSSCMS--LFWELTAKNQV : 876
LBRM2903_3 : TRDLERQYTLQLLCSSTLRVIAAGYVTVLTEVCHRLGLVQLHPQQLHTESELLPGVVMQLQKREAVLKNTRHREAVTSSMS--LFWVLNKANCOA : 847
LbrM.33.34 : TRDLERQYTLQLLCSSTLRVIAAGYVTVLTEVCHRLGLVQLHPQQLHTESELLPGVVMQLQKREAVLKNTRHREAVTSSMS--LFWVLNKANCOA : 847
LPALL3_330 : TRDLERQYTLQLLCSSTLRVIAAGYVTVLTEVCHRLGLVQLHPQQLHTESELLPGVVMQLQKREAVLKNTRHREAVTSSMS--LFWVLNKANCOA : 847
Angomonas : LSPLEVFSLRVLYADPVMDPOLYQYVPLFKRMRRCSFVFKKETTMGATIQLSFSQMKLLKLLYCAVLRNTRHREAVTSSMS--LFWVLNKANCOA : 687
VIPER : IAEVEVMASRFLAAALIGTRLCDHFFPIHAWRRRLSANRGIVLETSSANPEKAVGLGERLRHIENRRRIIRKPEKASAA--HITDASLHWGA : 391
trd6Er y3q L5 6 a s Y 6 rl p L p w l n pR 6 p 6 v l a

```

```

*      1000      *      1020      *      1040      *      1060      *      10
LpyrH10_03 : VLIQEETOETW-TSYGS--LEDMLG-----TRACASITFAIECFQPCIEKGAHWLILIGGKSLGRN-----D : 877
Chritidia_ : VLIQEESQETW-NTGGS--LEDMLG-----TRVACAISFAVECFQPCIEKGSVRIQDLGNNVMSRN-----D : 922
LMARLEM249 : VLVNEETOETW-NAEGS--LEDMLG-----TAVAAAITSFVGFQPCIEKGAHVRIKIMGGKPTGGS-----S : 937
LBRM2903_3 : VLVSQETOETW-NASME--LEDTLG-----TRVASVAVTFAVVGFPQPCIEKGAHVRIKIVIGGEAKGPN-----G : 908
LbrM.33.34 : VLVSQETOETW-NASME--LEDTLG-----TRVASVAVTFAVVGFPQPCIEKGAHVRIKIVIGGEAKGPN-----G : 908
LPALL3_330 : VLVSQETOETW-NASME--LEDTLG-----TRVASVAVTFAVVGFPQPCIEKGAHVRIKIVIGGEAKGPN-----G : 908
Angomonas : LILFFNRRRLGG-M-GGKFFRRPRSGVEMEVQAVRIALESILNANSFGEMTILIDSSESIPEGKRSDDQATFSSEYVW-SEVQKIHAETKHKLTVTFR : 782
VIPER : VFPDSDGDVKIAGGKWE--RKPFPI-----MQAEARAVRIALSASAI--PSTMDWVDN--TSLOEAA-----N : 451
61 e qetw l d lg t a a Fq c ekg h gg g

```

```

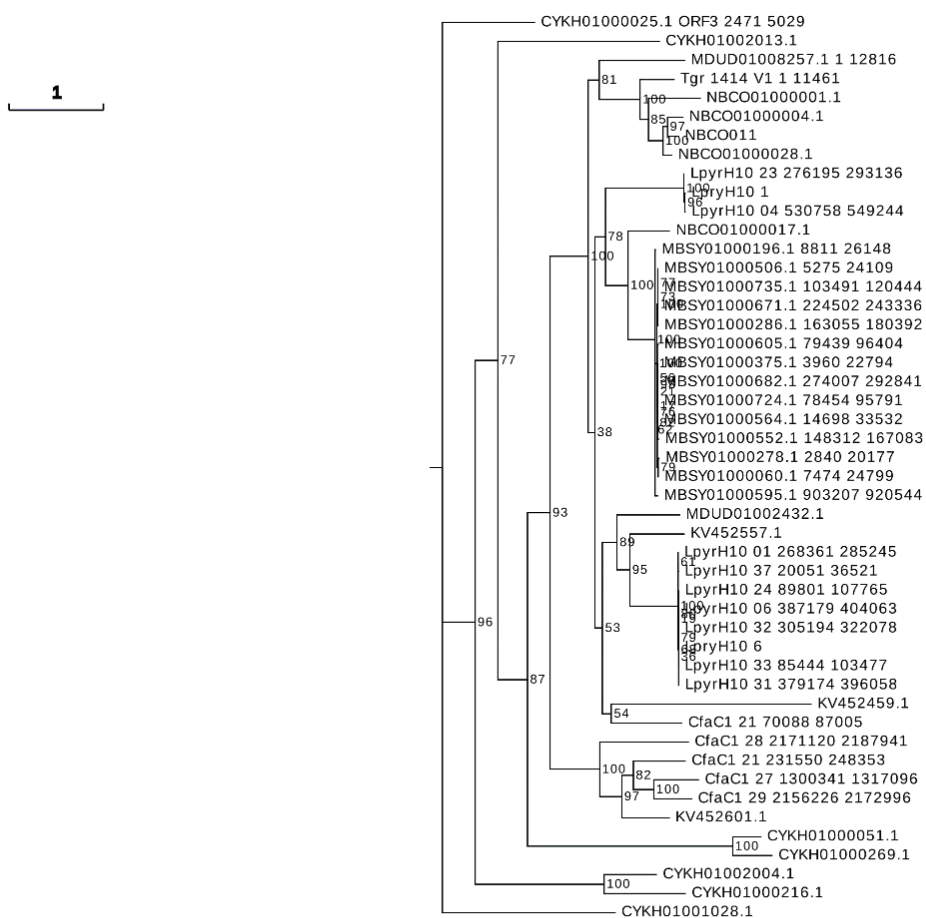
*      80      *      1100      *      1120      *      1140      *
LpyrH10_03 : TPIFKNESKDVADCIIRHLVTR-TFSFEVRYLGEASKANGSTAA-G--QKFAAKEQRKEYA----- : 937
Chritidia_ : KPIYRNEPEKVEDDCIRHLVTR-TFSFEVENCGNEGKNGTAA-H--SKFAYKCKW-----A : 978
LMARLEM249 : -----TPEEASVHIRSHLVAR-TFSFEVAYLDVEA-SSNGTADA-AAITLKLAHNGH-----E : 989
LBRM2903_3 : KPIYSEPEDSDAVSIRHLVAR-IFSFEVAYLNLPG-SAGGVAGD-TAACNDAQMCANRRLIAPPS--SPPPITYPS : 981
LbrM.33.34 : KPIYSEPEDSDAVSIRHLVAR-IFSFEVAYLNLPG-SAGGVAGD-TAACNDAQMCANRRLIAPPS--SPPPITYPS : 981
LPALL3_330 : KPIYSEPEDSDAVSIRHLVAR-IFSFEVAYLNLPG-SAGGVAGD-TAACNDAQMCANRRLIAPPS--SPPPITYPS : 981
Angomonas : -----SIKNEEHSFIRTEAGVPLDPEEWAMASREL-LQR-----N : 818
VIPER : KGSSEKSHAMTWELQRIYEPDSR-GIQATHAYVRSANFADGISRGVFTILOLAKENWNLRRGAGSCGWRTPKSAT : 528
s i Ir hlv r fsefayv g a a g

```

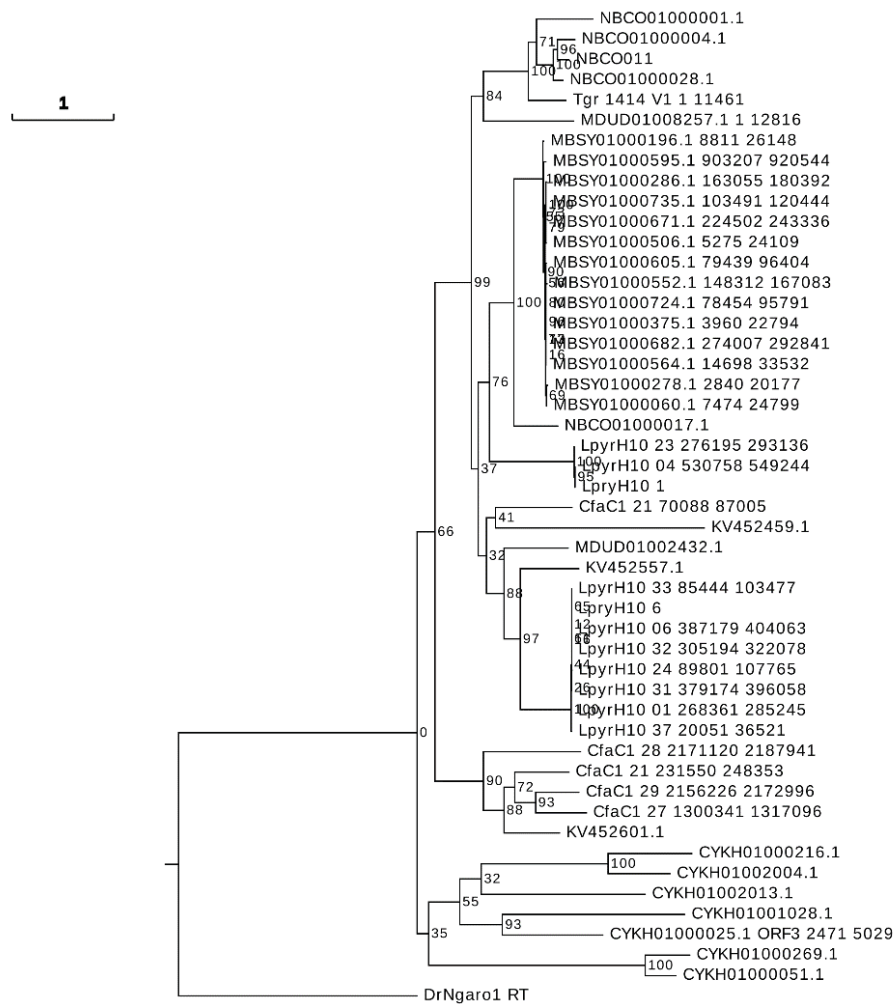
ANEXO E – RECONSTRUÇÕES FILOGENÉTICAS OBTIDAS PARA OS ELEMENTOS *VIPER* E *TATE*

Reconstruções filogenéticas dos elementos *VIPER* e *TATE* utilizando cada uma das proteínas (YR ou RT) através do método de ML utilizando diferentes matrizes de alinhamento.

Árvore de ML (modelo evolutivo LG+G, 625 aminoácidos) da proteína RT de *VIPER* de cinetoplastídeos. Nenhum grupo externo foi utilizado.



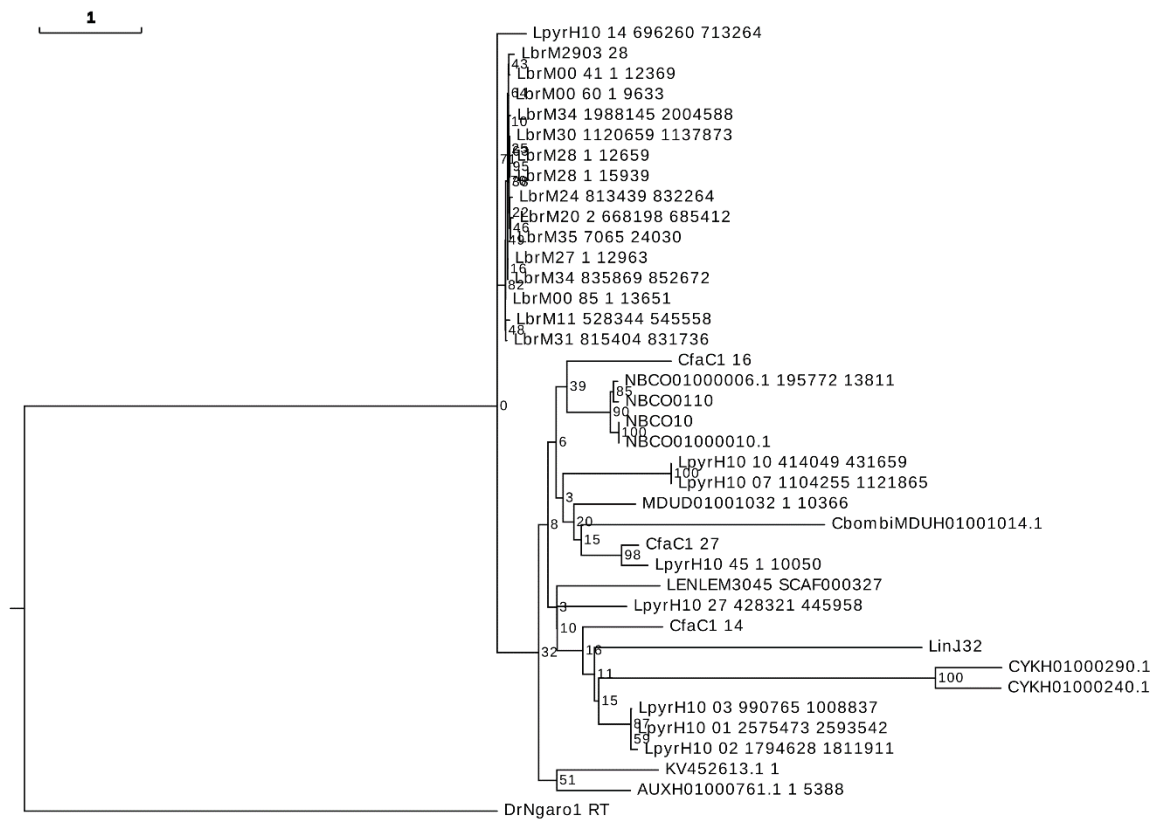
Árvore de ML (modelo evolutivo LG+G, 647 aminoácidos) da proteína RT de *VIPER* de cinetoplastídeos. A proteína RT do elemento DrNgaro1 foi utilizada como grupo externo.



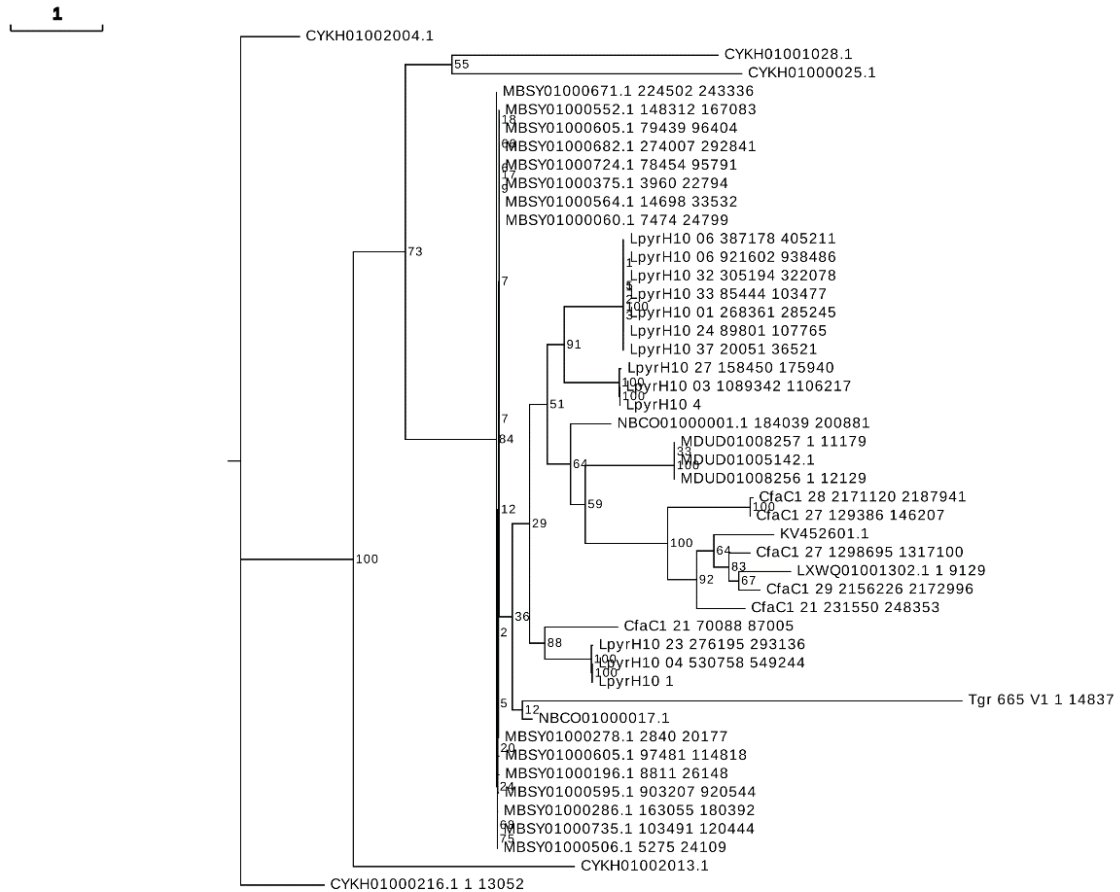
Árvore de ML (modelo evolutivo LG+G, 841 aminoácidos) da proteína RT de *TATE* de cinetoplastídeos. Nenhum grupo externo foi utilizado.



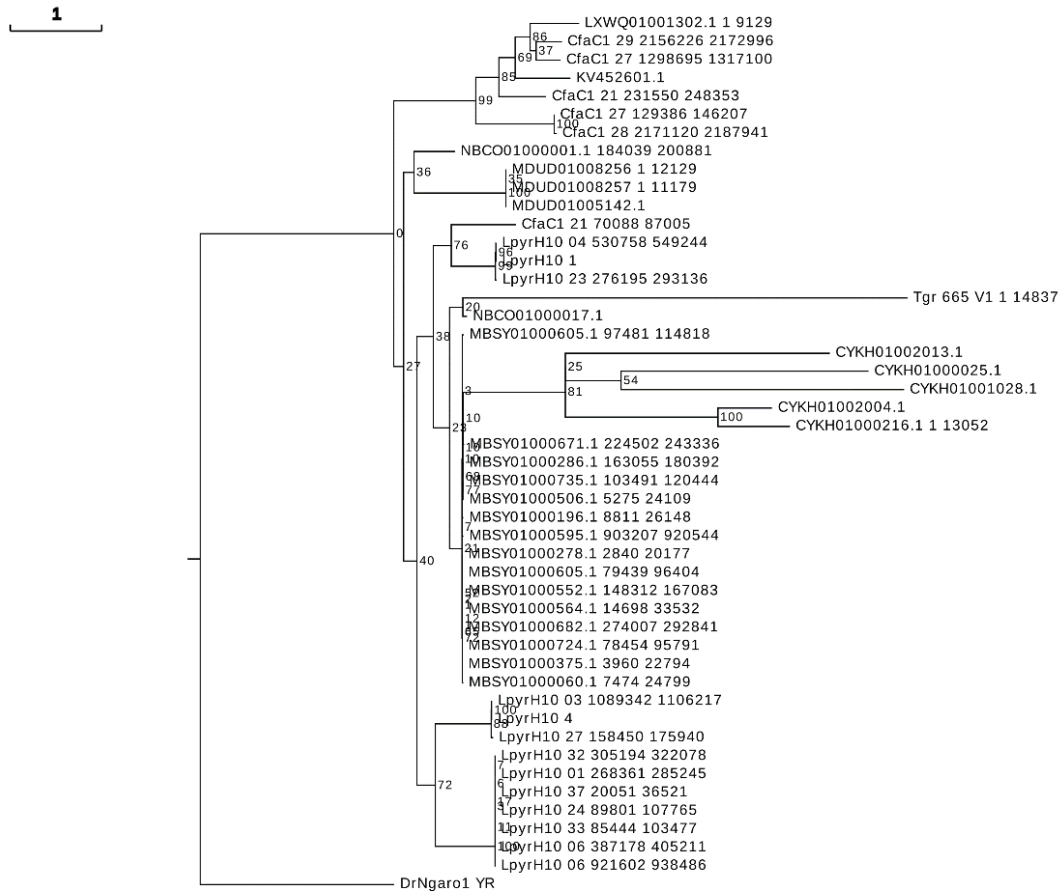
Árvore de ML (modelo evolutivo LG+G, 831 aminoácidos) da proteína RT de *TATE* de cinetoplastídeos. A proteína RT do elemento DrNgaro1 foi utilizada como grupo externo.



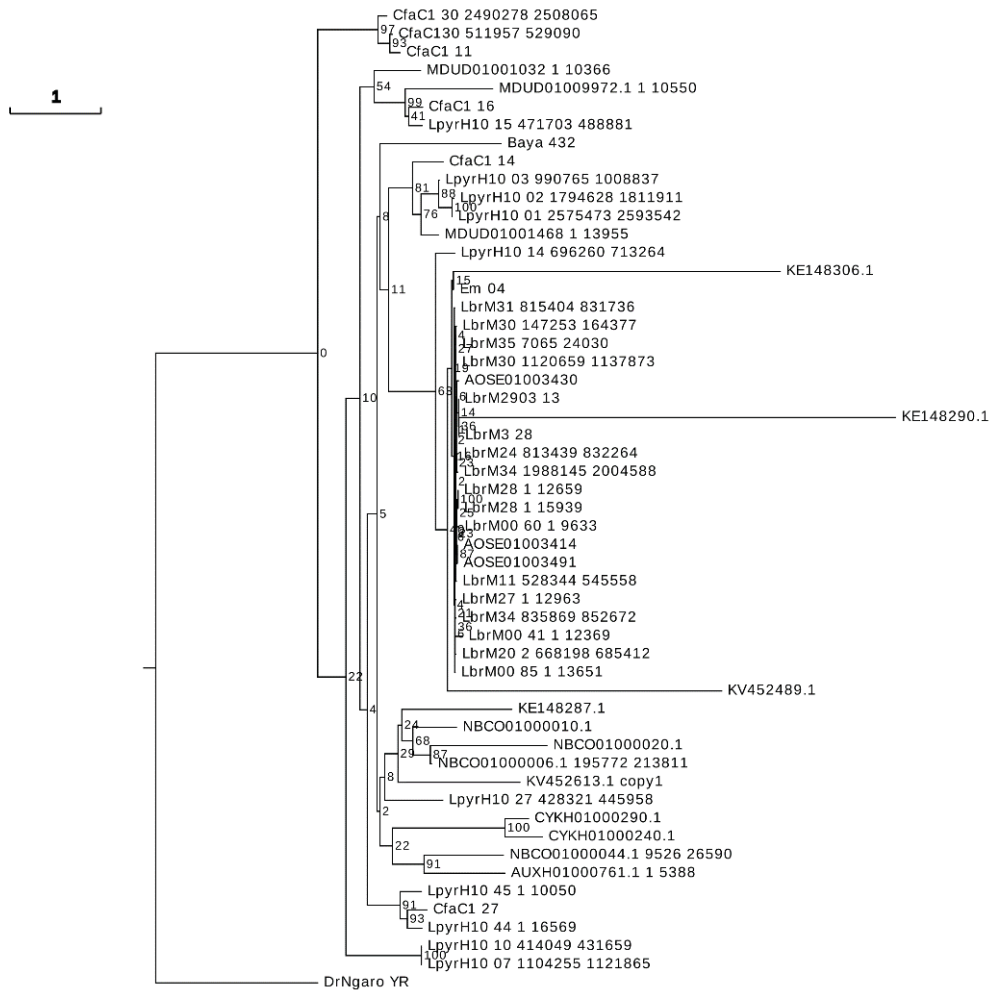
Árvore de ML (modelo evolutivo LG+G, 275 aminoácidos) da proteína YR de *VIPER* de cinetoplastídeos. Nenhum grupo externo foi utilizado.



Árvore de ML (modelo evolutivo LG+G, 249 aminoácidos) da proteína YR de *VIPER* de cinetoplastídeos. A proteína YR do elemento DrNgaro1 foi utilizada como grupo externo.



Árvore de ML (modelo evolutivo LG+G, 543 aminoácidos) da proteína YR de *TATE* de cinetoplastídeos. A proteína YR do elemento DrNgaro1 foi utilizada como grupo externo.



ANEXO F – EXPRESSÃO DE PROTEÍNAS PARA A PRODUÇÃO DE ANTICORPOS CONTRA AS PROTEÍNAS DE *VIPER*

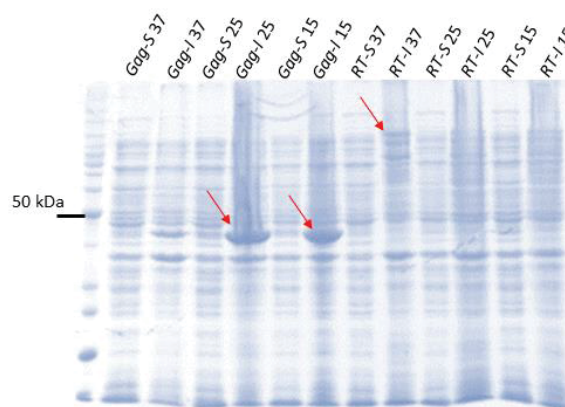
Um dos objetivos iniciais da dissertação também foi a produção de anticorpos contra as proteínas de *VIPER* para a geração de ferramentas para estudos funcionais. A primeira etapa de clonagem e testes de expressão das proteínas foi realizada. Sumarizamos as principais etapas desenvolvidas como anexo, pois devido ao curto período de desenvolvimento do trabalho (apenas 10 meses, devido a troca de orientação/projeto) e as dificuldades iniciais apresentadas, decidimos seguir apenas com as análises evolutivas.

- As ORFs 1 e 3 (*Gag-like* e RT/RH) já estavam clonadas em vetor pdonr221 (vetor de entrada do Sistema Gateway® - Invitrogen, usado para as clonagens). A ORF 2 (YR) teve de ser clonada e tivemos dificuldades em obter clones positivos, o que atrasou o processo de expressão dessa proteína.

- As três ORFs de *VIPER* foram clonadas no vetor pDEST17 pela recombinação com os vetores de entrada. Esse vetor possui a etiqueta 6xHis N-terminal (Figura 1).

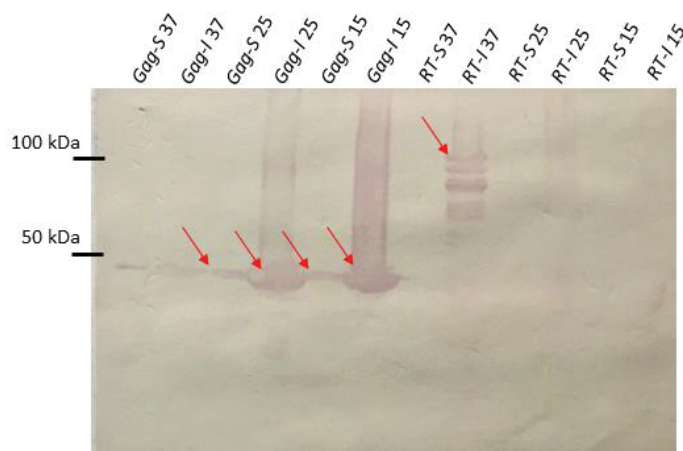
- Os testes de expressão da *Gag-like* e RT/RH foram realizados na bactéria *E. coli* cepa BL21 DE3 Star pela indução com 0,5 mM de IPTG e com as seguintes condições de cultivo: 37°C por 4 horas, 25°C por 16 horas, 15°C por 16 horas. Os extratos foram analisados em gel SDS-page e *Western Blot* (Figuras 1 e 2). As técnicas confirmaram a expressão das proteínas *Gag-like* (49,6 kDa) e RT/RH (104 kDa) conjugadas com 6xHis, porém para a proteína RT a única condição que apresentou um resultado positivo, mostrou um padrão de possível degradação.

- A expressão da proteína YR (38,7 kDa) foi testada utilizando a cepa BL21 DE3 Star, com a condição de 0,5 mM de IPTG para indução em uma temperatura de 37°C durante 4 horas. Analisamos o resultado em gel de SDS-page e confirmação da expressão *Western blot* (Figuras 3 e 4). É possível observar uma banda superexpressa principalmente na fração insolúvel, porém com tamanho um pouco maior do que o esperado de aproximadamente 40 kDa.

FIGURA 1 - TESTE DE EXPRESSÃO DAS PROTEÍNAS GAG-LIKE E RT DE *VIPER*

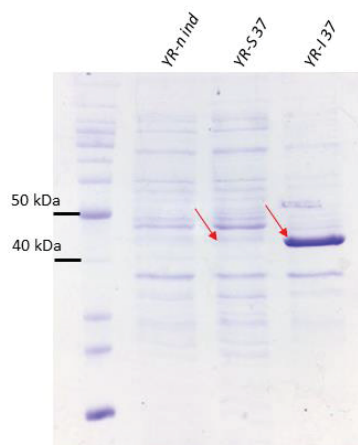
LEGENDA: Gel SDS-*page* 13% mostrando os resultados dos testes de expressão em diferentes condições. Gag-S: expressão da ORF1 (gag-like) fração solúvel; Gag-I: expressão da ORF1 (gag-like) fração insolúvel; RT-S: expressão da ORF3 (RT/RH) fração insolúvel; RT-I: expressão da ORF3 (RT/RH) fração insolúvel. 37, 25 e 15 representam as temperaturas de expressão: 37 °C – 4 h; 25 °C – 16 h; 15 °C – 16 h. As setas vermelhas apontam para as bandas correspondentes com o peso molecular de cada proteína.

FONTE: A autora, 2018.

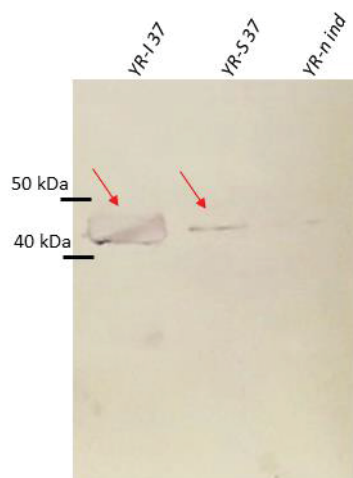
FIGURA 2 - *WESTERN BLOT* CONFIRMATÓRIO DA EXPRESSÃO DE GAG-LIKE E RT.

LEGENDA: *Western blot* correspondente ao gel representado na acima utilizando anticorpo anti-his. Gag-S: expressão da ORF1 (Gag-like) fração solúvel; Gag-I: expressão da ORF1 (Gag-like) fração insolúvel; RT-S: expressão da ORF3 (RT) fração insolúvel; RT-I: expressão da ORF3 (RT) fração insolúvel. 37, 25 e 15 representam as temperaturas de expressão: 37°C – 4 h; 25°C – 16 h; 15°C – 16 h. As setas vermelhas correspondem ao sinal obtido para cada proteína.

FONTE: A autora, 2018.

FIGURA 3 - TESTE DE EXPRESSÃO DA PROTEÍNA YR DE *VIPER*

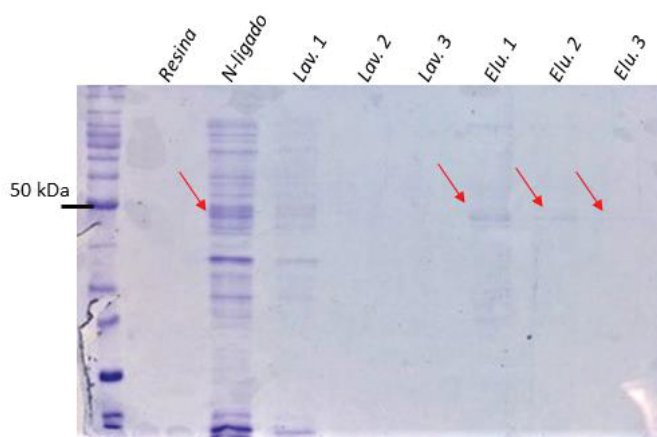
LEGENDA: Gel SDS-page 13% mostrando o resultado do teste de expressão. YR n-ind: extrato sem a indução da proteína – plasmídeo com o vetor contendo a proteína, porém sem acréscimo de IPTG; YR-S 37: expressão da YR fração solúvel; YR-I 37: Expressão da YR fração insolúvel. As setas vermelhas apontam para as bandas que poderiam estar correspondentes a expressão de proteínas.
 FONTE: A autora, 2018.

FIGURA 4 - *WESTERN BLOT* CONFIRMATÓRIO DA EXPRESSÃO DA YR.

LEGENDA: *Western blot* correspondente ao gel da expressão mostrando na figura acima anticorpo anti-his. YR-I 37: expressão da YR fração insolúvel; YR-S: expressão da YR fração solúvel; YR-n ind: amostra não induzida, contendo o plasmídeo com a YR, porém sem acréscimo de IPTG. As setas vermelhas correspondem ao sinal pela revelação.
 FONTE: A autora, 2018.

- Elaboramos hipóteses sobre a possível atividade funcional das proteínas RT/RH e YR gerarem toxicidade para a célula, já que poderiam talvez interferir em processos celulares. Testes de crescimento das bactérias expressando a RT/TH foram iniciados, porém para entender o padrão de expressão das proteínas e elaborar um protocolo satisfatório de expressão, teríamos que investir bastante tempo em novos testes. Assim, prosseguimos para os testes de purificação da proteína *Gag-like* por resina de níquel. A purificação foi avaliada em gel SDS-*page* (Figura 5). Percebemos que houve uma ligação não tão forte da proteína na resina pois visualizamos uma banda correspondente ao tamanho da proteína na fração não ligada, e a segunda eluição (250 mM de imidazol) apresentou o melhor resultado, pois observamos a proteína quase completamente purificada (ausência de bandas inespecíficas).

FIGURA 5 - RESULTADO DO TESTE DE PURIFICAÇÃO EM RESINA DE NÍQUEL PARA PROTEÍNA GAG-LIKE.



LEGENDA: Gel SDS-*page* 13% mostrando o resultado do primeiro teste de purificação em resina de níquel utilizando a fração solúvel obtida no teste de expressão. Resina: resina utilizada ressuspensa em tampão de lavagem; N-ligado: fração não ligada; Lav. 1, Lav. 2, Lav. 3: frações recolhidas a cada etapa de lavagem; Elu. 1: fração eluída com 100 mM de imidazol; Elu. 2: fração eluída com 250 mM de imidazol; Elu. 3: fração eluída com 500 mM de imidazol. As setas vermelhas indicam para bandas correspondentes ao tamanho da proteína.

FONTE: A autora, 2017.

- Alterações no protocolo de purificação foram realizadas para tentarmos obter um melhor resultado. Aparentemente havíamos definido um protocolo de expressão e purificação satisfatório, porém ao aumentar a escala de expressão da proteína *Gag-like* para proceder a purificação das proteínas em maior quantidade para inoculação nos camundongos, começamos a ter dificuldade na expressão da

proteína e, quando expressa, o protocolo de purificação previamente definido passou a não ser satisfatório.

- Dessa forma, apesar da importância da produção de anticorpos contra as proteínas de *VIPER*, decidimos focar apenas nas análises *in silico*. Porém, o desenvolvimento desta parte do trabalho por outro aluno será facilitado pelos primeiros testes realizados mencionados aqui.