

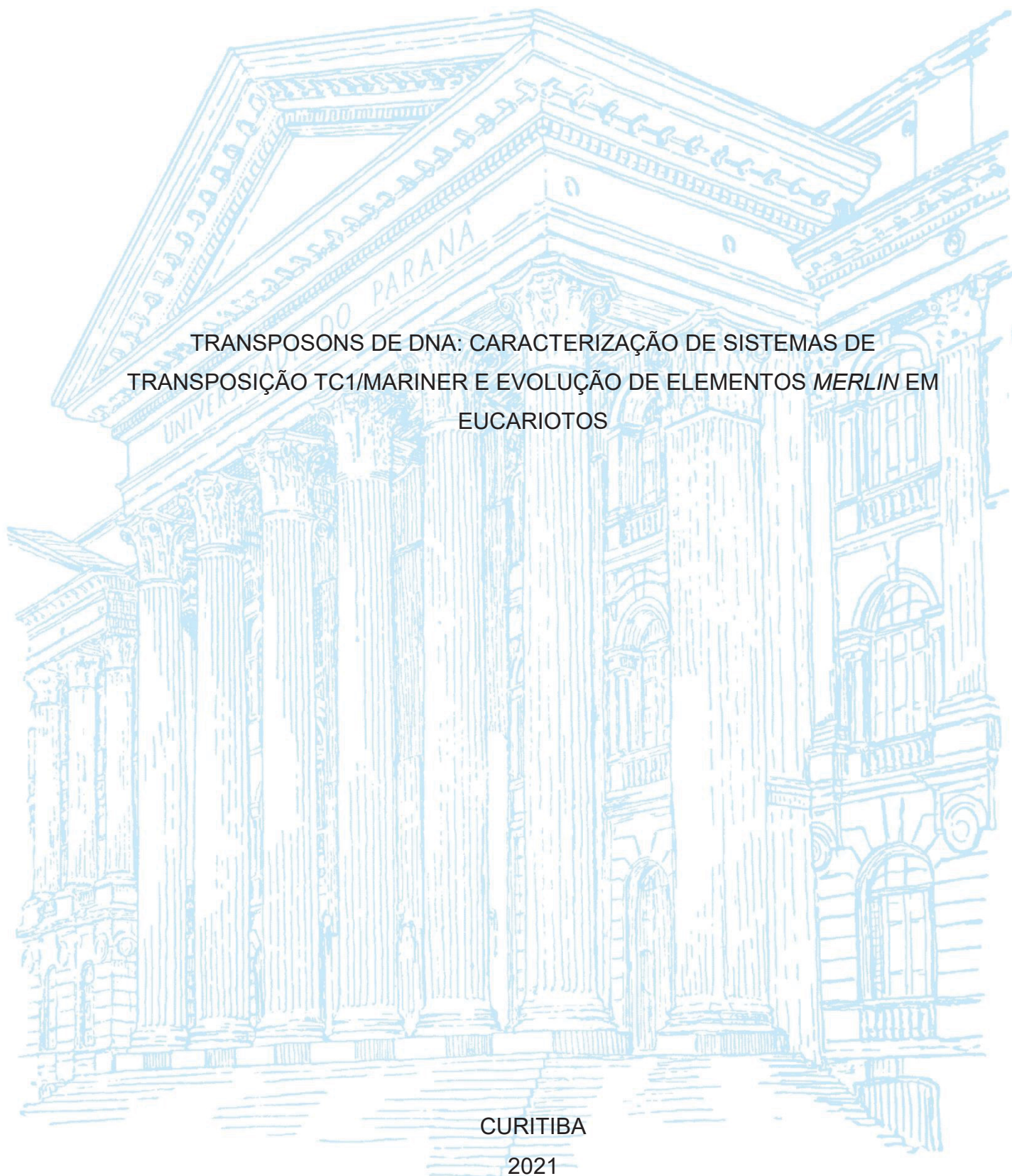
UNIVERSIDADE FEDERAL DO PARANÁ

ANA LUISA KALB LOPES

TRANSPOSONS DE DNA: CARACTERIZAÇÃO DE SISTEMAS DE
TRANSPOSIÇÃO TC1/MARINER E EVOLUÇÃO DE ELEMENTOS *MERLIN* EM
EUCARIOTOS

CURITIBA

2021



ANA LUISA KALB LOPES

TRANSPOSONS DE DNA: CARACTERIZAÇÃO DE SISTEMAS DE
TRANSPOSIÇÃO TC1/MARINER E EVOLUÇÃO DE ELEMENTOS *MERLIN* EM
EUCARIOTOS

Tese apresentada ao curso de Pós-Graduação em
Biologia Celular e Molecular, Setor de Ciências
Biológicas, Universidade Federal do Paraná, como
requisito parcial à obtenção do título de Doutor.

Orientador: Dr. Marco Aurelio Krieger

Coorientador(a): Dr^a. Adriana Ludwig

CURITIBA

2021

FICHA CATALOGRÁFICA

Universidade Federal do Paraná
Sistema de Bibliotecas
(Giana Mara Seniski Silva – CRB/9 1406)

Lopes, Ana Luisa Kalb

Transposons de DNA : caracterização de sistemas de transposição
TC1/mariner e evolução de elementos *Merlin* em eucariotos. / Ana Luisa
Kalb Lopes. – Curitiba, 2021.
163 p.: il.

Orientador: Marco Aurelio Krieger.
Coorientadora: Adriana Ludwig.

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

1. Elementos de DNA transponíveis. 2. Transposases. 3.
Célula eucariótica. I. Título. II. Krieger, Marco Aurelio. III. Ludwig, Adriana.
IV.

Universidade Federal do Paraná. Setor de Ciências Biológicas.
Programa de Pós-Graduação em Biologia Celular e Molecular.

CDD (22. ed.) 572.869

TERMO DE APROVAÇÃO



MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS BIOLÓGICAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO BIOLOGIA CELULAR E
MOLECULAR - 40001016007P8

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação BIOLOGIA CELULAR E MOLECULAR da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **ANA LUISA KALB LOPES** intitulada: **Transposons de DNA: Caracterização de Sistemas de Transposição *Tc1/Mariner* e evolução de elementos *Merlin* em eucariotos**, sob orientação do Prof. Dr. MARCO AURÉLIO KRIEGER, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora 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, 14 de Outubro de 2021.

Assinatura Eletrônica
12/11/2021 19:01:36.0
MARCO AURÉLIO KRIEGER
Presidente da Banca Examinadora

Assinatura Eletrônica
14/10/2021 16:52:05.0
PATRICIA SHIGUNOV
Avaliador Externo (INSTITUTO CARLOS CHAGAS)

Assinatura Eletrônica
17/11/2021 17:06:01.0
FABIANO PIMENTEL TORRES
Avaliador Externo (UNIVERSIDADE FEDERAL DO PAMPA)

Assinatura Eletrônica
28/10/2021 12:27:38.0
SILVIO MARQUES ZANATA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
14/10/2021 16:22:47.0
MARINDIA DEPREA
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL)

Ao grande amor da minha vida

Cristhiano

E nossos lindos filhos,

Larissa e Matheus

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus que me capacitou para realizar este trabalho e nunca me desamparou.

Ao meu orientador Dr Marco A. Krieger pela oportunidade e confiança depositada a mim.

À minha co-orientadora Adriana Ludwig pelo companheirismo, por ter me ensinado muito sobre bioinformática e esse mundo dos transposons (que ela tanto ama) que era novo para mim.

Ao pessoal do laboratório LACTAS do ICC e do Desenvolvimento do IBMP pelo companheirismo, troca de experiências e por compartilhar momentos agradáveis ao longo desta jornada. Ao Mateus que me ajudou muito com os experimentos de qPCR. À Mariana e ao Haruo que me ajudaram com os experimentos com célula CHO, com ensinamentos e sugestões. Ao Rafael Kessler pelo apoio e discussões ao longo dessa jornada.

Agradeço aos pesquisadores Eva Kriegová e Julius Lukeš que nos ajudaram com os ensaios de PCR do artigo e contribuições na escrita.

Aos pesquisadores Nilson Zanchin e Daniela Gradia pelo acompanhamento e sugestões ao longo do trabalho.

Agradeço imensamente a minha família que me deram muito apoio nesta jornada, principalmente minha mãe me ajudando com a Larissa desde o início deste trabalho quando ela tinha apenas 6 meses. À minha Irmã, pela companhia no laboratório e me entender como cientista.

Ao meu amor Cristhiano, pela compreensão, pelo apoio e pelos incentivos nos momentos difíceis. Obrigada por tudo que você tem feito pela nossa família.

Agradeço a Larissa que me acompanhou em toda a jornada sendo compreensiva e orgulhosa da mamãe que será doutora (rsrs) e o Matheus, ainda no forninho, que veio no final desta jornada para deixar ainda mais emocionante.

A Capes e IBMP pelo apoio financeiro.

Obrigada a todos

RESUMO

Os transposons de DNA podem ser vistos como veículos naturais de transferência devido a uma enzima chamada transposase, codificada pelo transposon, capaz de reconhecer repetições terminais invertidas (TIRs) presente nas suas extremidades. Quando expressa, a transposase se liga as TIRs e realizar sua excisão e reintegração em outra posição dentro do genoma. O Sistema Ingresso é uma plataforma desenvolvida para expressão heteróloga em células CHO (*Chinese hamster ovary*) que utiliza o transposon *Sleeping beauty* (SB) para integração genômica. Com o objetivo de entender o funcionamento do Sistema Ingresso como um todo, foram selecionadas 10 amostras (8 clones e 2 populações) de células CHO transfectadas com os vetores desse sistema contendo o gene de interesse Pró-trombina e selecionadas com o marcador GFP. Para a maioria das amostras, não foi observado uma correlação positiva entre a quantidade de cópias presente no genoma, o nível de mRNA e a quantidade de proteína expressa. Em relação ao local de inserção, foi possível verificar inserções em locais de atividade transcricional bem como regiões possivelmente silenciadas. O clone 28 que apresenta o maior nível de expressão da Pró-trombina mostrou o maior nível de mRNA e possui o maior número de cópias do transgene no genoma bem como maior número de locais de inserção exclusivos. Nessa amostra, foi confirmada a presença da SB no genoma, que pode levar a instabilidade genômica como a remobilização do cassete de expressão. Nossos dados indicam que o uso de uma transposase mais eficiente com capacidade de inserir um maior número de cópias do cassete por célula poderá levar a um ganho de expressão considerável da proteína de interesse. Como uma tentativa de aprimorar o sistema de transposição existente, uma nova Transposase também pertencente a superfamília *Tc1-Mariner* e suas respectivas TIRs foram adicionadas no lugar dessas regiões de SB. Apesar da atividade dessa nova Transposase não ter sido confirmada até o momento, uma população enriquecida com GFP, verificada por citometria de fluxo, sugere que houve atividade de mobilização do transgene da nova transposase. Nesse trabalho, também foi realizada uma análise da superfamília *Merlin* explorando a sua distribuição de maneira global nos eucariotos. Esses elementos foram identificados em animais, fungos, plantas e protistas e, pela primeira vez, sua presença foi relatada em Rhodophyceae, Metamonada, Discoba e Alveolata. Os resultados desse trabalho mostram que *Merlin* é mais amplamente distribuído do que inicialmente descrito. Além disso, este é o primeiro relato de um transposon de DNA em protozoários flagelados da ordem Kinetoplastida, com a Transposase *Merlin* conservada identificada em *Bodo saltans* e *Perkinsella* sp. As árvores evolutivas sustentam que *Merlin* é provavelmente uma superfamília antiga, com eventos iniciais de diversificação e perdas secundárias. Várias famílias potencialmente ativas foram caracterizadas, algumas, curiosamente, com TIRs altamente imperfeitas, e famílias contendo repetições *in tandem* internas. Ainda, vários casos potenciais de domesticação de *Merlin* foram encontrados. Estudos como esse, de caracterização de novos transposons, são importantes não somente para o entendimento do papel dessas sequências na evolução e funcionamento dos genomas, mas também fornecem informações relevantes no contexto de uso biotecnológico dessas sequências.

Palavras-chave: Transposon de DNA; *Tc1/Mariner*; *Merlin*; Sistema de transposição; Célula CHO, *Sleeping beauty*.

ABSTRACT

DNA transposons can be seen as natural transfer vehicles due to an enzyme called transposase, encoded by the transposon, capable of recognizing inverted terminal repeats (TIRs) present at their ends. When expressed, a transposase binds like TIRS and performs its excision and reintegration at another position within the genome. The Ingresso System is a platform developed for heterologous expression in CHO cells (Chinese hamster ovary) that uses the transposon *Sleeping Beauty* (SB) for genomic integration. In order to understand the functioning of the Ingresso System as a whole, 10 (8 clones and 2 units) of CHO cells transfected with the Ingresso System vectors containing the gene of interest Prothrombin were selected and selected with the GFP marker. For most sample, a positive correlation was not observed between the number of copies present in the genome, the level of mRNA and the amount of expression expressed. Regarding the insertion site, it was possible to verify the local insertions of transcriptional activity as well as possibly silenced regions. Clone 28 that has the highest level of Prothrombin expression exhibited the highest level of mRNA and has the highest number of copies of the transgene in the genome as well as the highest number of unique insertion sites. In this sample, the presence of SB in the genome was confirmed, which can lead to genomic instability such as the remobilization of the expression cassette. Our data indicate that the use of a more efficient transposase capable of inserting a greater number of copies of the cassette per cell can lead to a gain in the expression of the protein of interest. As an attempt to improve the existing transposition system, a new Transposase also belonging to the *Tc1-Mariner* superfamily and its latest TIRs were added in place of these SB regions. Although the activity of this novel transposase has not been confirmed so far, a population enriched with GFP, verified by flow cytometry, included the transgene mobilization activity of the novel transposase. In this work, an analysis of the superfamily *Merlin* was also carried out, exploring its global distribution in eukaryotes. These elements were identified in animals, fungi, plants and protists and, for the first time, their presence was reported in Rhodophyceae, Metamonada, Discoba and Alveolata. The results of this work show that Merlin is more widely distributed than initially described. Furthermore, this is the first report of a DNA transposon in flagellated protozoa of the order Kinetoplastida, with the conserved Merlin Transposase identified in *Bodo saltans* and *Perkinsella sp.* Evolutionary trees maintain that *Merlin* is likely an ancient superfamily, with early events of diversification and secondary losses. Several potentially active families have been characterized, some, interestingly, with highly imperfect TIRs, and families containing internal in tandem repeats. Also, several potential cases of Merlin domestication have been found. Studies like this one, for the characterization of new transposons, are important not only for understanding the role of these sequences in the evolution and functioning of genomes, but also provide relevant information in the context of biotechnological use of these sequences.

Keywords: DNA Transposon; Tc1/Mariner; Merlin; Transposition system; CHO cell, Sleeping beauty.

LISTA DE FIGURAS

Figura 1. Plataforma do Sistema Ingresso..	33
Figura 2. Desenho experimental da caracterização do Sistema Ingresso 1.0.....	38
Figura 3. Desenho experimental da caracterização de nova transposase para sistema de transposição.	38
Figura 4. Mapa do vetor <i>helper</i> do Sistema Ingresso.....	39
Figura 5. Mapa do vetor <i>donor</i> do Sistema Ingresso.....	39
Figura 6. Mapa do vetor <i>donor</i> do Sistema Ingresso 1.0.....	40
Figura 7. Esquema da etapa de clivagem da técnica de PCR splinkerette utilizada para identificar inserções gênicas realizadas pelo Sistema Ingresso.....	44
Figura 8: Desenho esquemático da análise das sequências geradas por Illumina da PCR splinkerette para investigar os locais de inserção genômica dos cassetes..	49
Figura 9. Esquema do protocolo da PCR Splinkerette.....	57
Figura 10. Eletroforese do resultado da PCR splinkerette para os clones 3, 5, 6, 7 e população 1.....	57
Figura 11. Análise do sequenciamento da PCR splinkerette para a amostra 7 (clone bacteriano 71).	58
Figura 12. Contexto genômico da inserção do cassete na célula CHO da amostra 6, clones bacterianos 61 e 62.....	59
Figura 13. Contexto genômico da inserção do cassete na célula CHO da amostra 7, clones bacterianos 71 e 74.....	60
Figura 14. Contexto genômico da inserção do cassete na célula CHO da amostra 7, clone bacteriano 72.....	60
Figura 15. Contexto genômico da inserção do cassete na célula CHO da amostra 7, clone bacteriano 73.....	60
Figura 16. Diagrama de Venn mostrando o número de inserções potenciais compartilhadas entre as amostras.	67
Figura 17. Inserção aleatória da SB no genoma de célula CHO para o clone 6.	71
Figura 18. Curva padrão do gene Glucagon.	72
Figura 19. Curva padrão do gene Pró-trombina.....	73
Figura 20. Valores obtidos para o número de cópias do cassete contendo o gene Pró-trombina nos clones e populações.	74
Figura 21. Curva padrão do gene Eif3i.....	76

Figura 22. Valores obtidos para a expressão relativa do gene Pró-trombina nos clones e populações.....	77
Figura 23. Gel de agarose da confirmação do mRNA bicistrônico contendo o gene da Pró-trombina e GFP nos clones e populações isolados de célula CHO inseridos pelo Sistema Ingresso 1.0.....	79
Figura 24. Confirmação da integração da sequência da transposase <i>SB</i> no genoma da célula CHO nos clones e população por qPCR.	Erro! Indicador não definido.
Figura 25. Confirmação da integração da sequência da transposase <i>SB</i> no genoma da célula CHO nos clones e populações por PCR.	81
Figura 26. Par 1171 de leituras (R1 e R2) oriundas do clone 28 que alinharam com o vetor <i>helper</i>	82
Figura 27. Esquema das novas Transposases e os domínios proteicos presentes em cada proteína.	84
Figura 28. Gel de agarose da clonagem dos genes das transposases 2, 5, 7 no vetor <i>helper</i> do Sistema Ingresso. c.....	84
Figura 29. Gel de agarose 1% dos fragmentos das TIRs 3' no vetor pUC57 e do vetor <i>donor</i> após a digestão com <i>EcoRV</i> e <i>NheI</i>	86
Figura 30. Visualização da eletroforese em gel de agarose 3% dos fragmentos das TIRs 5' amplificados por PCR.	87
Figura 31. Esquema do vetor <i>donor</i> e do gene sintético contendo as TIRs da <i>Tp2</i> (tamanho total 2407 pb) com as enzimas <i>EcoRV</i> e <i>BglII</i> nas extremidades para a clonagem.....	87
Figura 32. Gel de agarose da clonagem das TIRs 5' e 3' da <i>Tp2</i> no vetor <i>donor</i>	88
Figura 33. Gel de agarose da confirmação da clonagem do fragmento contendo as TIRs 5'e 3' da <i>Tp2</i> no vetor <i>donor</i>	89
Figura 34. Gel de agarose da confirmação da clonagem das TIRs 3' da <i>Tp5</i> e <i>Tp7</i> no vetor <i>donor</i>	90
Figura 35. Gel de agarose da confirmação da clonagem das TIRs 5'da <i>Tp5</i> no vetor <i>donor</i>	91
Figura 36. Gel de agarose da confirmação da retirada da transposase do vetor <i>helper</i>	91
Figura 37. Representação gráfica do resultado da citometria de fluxo 48 h após a transfecção em célula CHO dos vetores do Sistema ingresso contendo <i>SB</i> e <i>Tp2</i> ..	93

Figura 38. Dados das porcentagens de GFP obtidos na citometria de fluxo durante a seleção das células transfectantes.....	94
Figura 39. Dados das porcentagens de GFP obtidos na citometria de fluxo das populações 60 dias após a primeira seleção de células.. ..	95
Figura 40. Representação gráfica do resultado da citometria de fluxo de células CHO transfectadas com os vetores do Sistema ingresso e a <i>Tp2</i> , 60 dias após o segundo <i>cell sorting</i>	96
Figura 41. <i>Western Blot</i> das culturas 48 horas após a transfecção para verificar a expressão da <i>SB</i>	97
Figura 42. Gel de agarose da segunda PCR splinkerette das amostras transfectadas com <i>SB</i> e <i>Tp2</i>	98
Figura 43. Resultado do blastn de uma leitura R2 contra a TIR da TP2 (blast-2-sequence) e contra o genoma da CHO.....	99

LISTA DE TABELAS

Tabela 1. Análise quantitativa de Pró-trombina em sobrenadante de culturas de células provenientes de população e linhagens clonais do Sistema Ingresso 1.0 por ensaio de Elisa	41
Tabela 2. Região de inserção e contexto genômico na célula CHO das amostras inseridas pelo Sistema Ingresso 1.0.....	59
Tabela 3. Dados gerais do sequenciamento por Illumina dos fragmentos gerados por PCR splinkerette.	61
Tabela 4. Região de inserção e contexto genômico na célula CHO das amostras inseridas pelo Sistema Ingresso 1.0, clone 6.	62
Tabela 5. Região de inserção e contexto genômico na célula CHO das amostras inseridas pelo Sistema Ingresso 1.0, clone 7.	62
Tabela 6. Região de inserção e contexto genômico na célula CHO das amostras inseridas pelo Sistema Ingresso 1.0, clone 28..	62
Tabela 7. Região de inserção e contexto genômico na célula CHO das amostras inseridas pelo Sistema Ingresso 1.0, população.	64
Tabela 8. Inserções compartilhadas entre as amostras e as exclusivas de cada amostra.	67
Tabela 9. Dados gerais do sequenciamento por Illumina dos fragmentos gerados por PCR splinkerette para <i>Tp2</i>	98

LISTA DE QUADROS

Quadro 1. Sequência nucleotídica dos pares de primers utilizados na reação PCR em tempo real.....	41
Quadro 2. Sequência nucleotídica dos pares de primers utilizados na reação PCR Splinkerette.....	46
Quadro 3. Sequência nucleotídica dos pares de primers utilizados na reação PCR....	50
Quadro 4. Esquema dos vetores <i>helper</i> e <i>donor</i> utilizados na transfecção.....	89

LISTA DE ABREVIATURAS OU SIGLAS

TE - Elementos de transposição
RNA - Ácido ribonucleico
mRNA - RNA mensageiro
DNA - Ácido desoxirribonucleico
cDNA – DNA complementar
LTR - *long terminal repeats*
PLE - *Penelope-like elements*
LINE - *Long interspersed nuclear elements*
SINEs - *Short interspersed elements*
TIRs – *terminal inverted repeats*
TH - transferência horizontal
ORF – *open reading frame*
TSD – *target site duplication*
SB – *Sleeping beauty*
PB - *PiggyBac*
IS - *Insertion sequence*
CHO - *chinese hamster ovary*
HEK-293 - *human embryonic kidney*
Per.C6 - derivado de retina humana
BHK - *baby hamster kidney*
RMCE - *recombinase-mediated cassette exchange*
LCR - *locus control region*
MARs - *matrix attachment regions*
uCOEs - *ubiquitous chromatin-opening elements*
GFP - *green fluorescent protein*
ZFN - *Zinc-finger nucleases*
TALEN - *transcription activator-like effector nucleases*
FACS - *fluorescence-activated cell sorter*
IRES - *Internal ribosome entry sites*
CAG - *CMV early enhancer/chicken β actin*
CMV - Citomegalovirus
cHS4 - *chicken hypersensitive site 4*

TPE – *Tripartite element*

SV40 - *Símbion vacuolating virus 40*

UTR - *Untranslated Region*

qPCR – PCR quantitativa em tempo real

Ct - *Cycle Threshold*

LB - Luria Bertani

NCBI - *National Center for Biotechnology Information*

PBS - Tampão salino fosfatado

BSA - Albumina do soro bovino

EDTA - ácido etilenodiamino tetra-acético

CsCl – Cloreto de cézio

FISH - *Fluorescence In Situ Hybridization*

SUMÁRIO

1 INTRODUÇÃO	18
1.1 ELEMENTOS DE TRANSPOSIÇÃO	18
1.2 TRANSPOSONS DE DNA DA ORDEM TIR.....	22
1.3 TRANSPOSONS E SISTEMAS DE EXPRESSÃO DE PROTEÍNAS DE INTERESSE TERAPÊUTICO.....	25
1.4 SISTEMA DE TRANSPOSIÇÃO.....	28
1.5 SISTEMA INGRESSO	32
1.6 JUSTIFICATIVA	34
2 OBJETIVOS.....	35
2.1 OBJETIVO GERAL.....	35
2.2 OBJETIVOS ESPECÍFICOS	35
CAPÍTULO I.....	37
3 MATERIAL E MÉTODOS	38
3.1 DESENHO EXPERIMENTAL: CARACTERIZAÇÃO DO SISTEMA INGRESSO 1.0.....	38
3.2 DESENHO EXPERIMENTAL: CARACTERIZAÇÃO DE NOVA TRANSPOSASE PARA SISTEMA DE TRANSPOSIÇÃO.	38
3.3 VETORES	39
3.4 LINHAGEM CHO.....	39
3.5 CULTURAS DE CÉLULAS CHO EXPRESSANDO PROTEÍNA DE INTERESSE.....	40
3.6 EXTRAÇÃO DE ÁCIDOS NUCLEÍCOS	41
3.7 SÍNTESE DA PRIMEIRA FITA DE CDNA.....	41
3.8 PCR TEMPO REAL.....	42
3.9 PCR SPLINKERETTE	43
3.10 AVALIAÇÃO DA PERDA DO VETOR <i>HELPER</i>	50
3.11 DESENVOLVIMENTO DE NOVOS VETORES.....	50
3.12 TRANSFORMAÇÃO EM CÉLULA CÁLCIO COMPETENTE	52
3.13 TÉCNICA DE PCR DE COLÔNIA PARA CONFIRMAÇÃO DAS CLONAGENS.....	52
3.14 PREPARAÇÃO DE PLASMÍDEO EM PEQUENA ESCALA (MINIPREP)	53
3.15 TRANSFECCÃO DE PLASMÍDEO EM CÉLULA CHO POR POLIFECÇÃO.....	53
3.16 SELEÇÃO DOS TRANSFECTANTES	54

3.17 EXTRAÇÃO DE DNA EXTRACROMOSSOMAL DE CÉLULA CHO TRANSFECTADA	54
3.18 WESTERN-BLOT PARA DETECÇÃO DA PROTEÍNA TRANSPOSASE DO SB.....	55
3.19 AVALIAÇÃO DA PERDA DOS VETORES EPISSOMAIS	55
4 RESULTADOS E DISCUSSÃO	56
4.1 CARACTERIZAÇÃO DO SISTEMA INGRESSO	56
4.1.1 Local de integração genômica.....	56
4.1.2 Evidência de integração aleatória	70
4.1.3 Correlação entre o número de cópias do cassete e a expressão relativa do gene de interesse.....	72
4.1.4 Integração da transposase	79
4.2 NOVAS TRANSPOSASES	83
4.1.1 Clonagem nos vetores <i>helper</i> e <i>donor</i>	84
4.1.2 Transfecção em célula CHO para testar o novo sistema <i>Tp2</i>	92
4.1.3 Atividade da transposase <i>SB</i>	96
4.1.4 Local de integração genômica da <i>Tp2</i>	97
CAPÍTULO II	101
DISTRIBUTION OF <i>MERLIN</i> IN EUKARYOTES AND FIRST REPORT OF DNA TRANSPOSONS IN KINETOPLASTID PROTISTS	101
ABSTRACT.....	103
INTRODUCTION.....	103
MATERIALS AND METHODS.....	105
RESULTS AND DISCUSSION.....	107
ACKNOWLEDGMENTS	128
REFERENCES.....	129
SUPPORTING INFORMATION	134
5 CONSIDERAÇÕES FINAIS	137
6 PERSPECTIVAS.....	145
REFERÊNCIAS.....	146
ANEXOS	157

1 INTRODUÇÃO

1.1 Elementos de transposição

Os elementos de transposição (TEs) são definidos como sequências de DNA que têm a capacidade de se mover de um lugar para outro no genoma. Os TE estão presentes em quase todos os organismos desde procariotos a eucariotos (MUNHOZ-LÓPEZ e GARCIA-PÉREZ, 2010) e podem constituir uma considerável porção do genoma em humanos (~50%) (DE KONING, 2011) e plantas (>50%) (KARAKÜLAH, 2017), por exemplo. Esses elementos possuem características marcantes que os tornam importante na evolução e adaptação das populações. Eles são universalmente distribuídos, muito antigos, abundantes, ativamente móveis e, apesar de serem prejudiciais para alguns indivíduos, são benéficos para as linhagens (BOURQUE et al., 2018).

Os TEs são classificados num sistema hierárquico unificado, projetado com base no mecanismo de transposição, similaridade de sequência e relações estruturais. Esse sistema inclui os níveis de classe, subclasse, ordem, superfamília, família e subfamília (WICKER et al., 2007). Esses elementos são divididos em duas grandes classes, dependendo do mecanismo de transposição, que pode ser por meio de um intermediário de RNA (classe I) ou através de DNA (classe II). Essa divisão dos TEs em duas grandes classes foi proposta inicialmente por Finnegan (1989) e tem permanecido em todas as propostas alternativas de classificação de TEs posteriores.

Os elementos da Classe I, denominados retrotransposon, são aqueles que se transpõem por um intermediário de RNA, o qual é transcrito e utilizado como molde para síntese de cDNA via transcriptase reversa, codificada pelo próprio elemento.

De acordo com Wicker et al (2007), os retrotransposons podem ser divididos em cinco ordens com base em suas características mecânicas, organização e filogenia da transcriptase reversa: LTR, DIRS, PLE, LINE e SINEs.

Os elementos da Classe II, denominados transposons de DNA, realizam a excisão do DNA e o reinsere em outro local do genoma num mecanismo chamado corte e cola (FINNEGAN, 1989). Segundo a classificação de Wicker e colaboradores (2007), a Classe II pode ser dividida em duas subclasses, que são distinguidas pelo número de fitas de DNA cortadas durante a transposição. A subclasse 1 compreende

os transposons clássicos (ordem TIR) que realizam "corte e cola" da dupla fita de DNA pela enzima transposase e são caracterizados pela presença de repetições terminais invertidas (TIRs – *terminal inverted repeats*). Elementos dessa subclasse são o foco desse trabalho. Ainda, são classificadas nessa subclasse, um grupo de elementos pobremente conhecidos que codificam uma enzima chamada tirosina recombinase (ordem Crypton), pois possivelmente a transposição envolve a quebra de ambas as fitas de DNA mediada por essa enzima. A subclasse 2, realiza a clivagem de uma das fitas de DNA e são transpostos por um mecanismo que envolve o deslocamento de uma das fitas. São os elementos das ordens Helitron e Maverick.

Os TEs podem impactar os genomas de diferentes maneiras. Ao se inserir em um novo local do genoma eles podem interromper sequências como promotores, íntrons e regiões reguladoras, ou podem adicionar sequência regulatória ou marcas epigenéticas as quais podem afetar os genes próximos. A sua inserção pode, ainda, levar a perda total de função de um gene quando ocorre em uma sequência codificadora de um gene. Além disso, os TE promovem mudanças estruturais no genoma em grande magnitude. A natureza repetitiva dos elementos pode favorecer a recombinação entre sequências homólogas em regiões não homólogas, culminando em reestruturações cromossômicas (PAÇO, 2019). Isto pode levar a múltiplos benefícios para a linhagem hospedeira em termos de facilitar a adaptação a novos habitats bem como a sobrevivência ao meio ambiente (OLIVER e GREENE, 2009).

Os TEs também são capazes de contribuir com a expansão genômica pela geração de repetições em tandem. Essas repetições podem se originar nos próprios TEs através de locais predispostos para sua formação. Durante o processo de transposição, a transposase induz a ruptura do DNA que tem o potencial de originar as primeiras repetições em tandem durante o processo de reparo. Uma vez que uma repetição em tandem é gerada em um genoma, a expansão dessas novas unidades ocorre lentamente ao longo do tempo. Os TEs podem facilmente dispersá-las pelo processo de transposição por todo o genoma quando essas repetições estão dentro da sequência do elemento móvel (AHMED e LIANG, 2012; PAÇO, 2019). Além disso, os TEs estão presentes em regiões pericentroméricas de uma ampla gama de espécies, que são construídas principalmente por repetições em tandem o que facilita a dispersão destas sequências (PAÇO, 2019).

Uma das contribuições mais diretas dos TEs para a evolução do genoma é como fonte de matéria-prima para novos genes e funções. Como os TEs são ricos em

sequências codificantes e regulatórias, estas podem ser utilizadas pelo genoma hospedeiro para desenvolver novas funções celulares em um processo denominado domesticação ou exaptação (JANGAM et al, 2017). Muitos exemplos de genes derivados de diferentes tipos de TE foram identificados em diversos eucariotos, incluindo transposases, integrases, transcriptase reversa, Gag e envelope (VOLFF, 2006). Diferentes tipos de transposons de DNA e retrotransposon contribuíram com sequências codificadoras durante a evolução. Entre os casos de domesticação de transposon de DNA mais bem conhecidos e caracterizados estão os genes RAG1 e RAG2 (derivados do mesmo elemento ancestral *Transib*) envolvidos na diversificação de imunoglobulinas de célula T (JOLY-LOPEZ e BUREAU, 2018). Essa grande diversidade de genes codificadores de proteínas derivados de elementos transponíveis em genomas eucarióticos, incluindo animais, plantas e fungos, fornecem evidências do importante papel dos TE como reservatório para novas funções celulares (VOLFF, 2006).

A distribuição dos elementos nos genomas não ocorre de forma aleatória, mas exhibe vários níveis de preferência para sua inserção os quais são guiados por forças seletivas, num equilíbrio entre a sua propagação e os efeitos deletérios para o hospedeiro. A seleção natural e a deriva genética também são forças poderosas que moldam a distribuição e o acúmulo de TEs nos genomas, sendo que inserções fortemente deletérias são rapidamente removidas da população enquanto as inserções que têm pouco ou nenhum efeito sobre a função do genoma e a aptidão do hospedeiro podem atingir a fixação. Assim, a distribuição genômica, o sucesso e a diversidade de TEs em um genoma são moldados tanto por propriedades intrínsecas aos elementos quanto por forças evolutivas que atuam no nível da espécie hospedeira (BOURQUE et al., 2018). Alguns genomas contêm apenas poucas famílias, enquanto outros possuem uma grande diversidade de TEs, sendo que até agora, poucas espécies eucarióticas parecem ser desprovidas completamente de TEs como os protistas do filo Apicomplexa (*Plasmodium falciparum*, *Toxoplasma gondii*, e *Theileria parva*) (WELLS e FESCHOTTE, 2020).

A propagação dos TE ocorre tanto de maneira vertical pela linhagem germinativa, como de maneira horizontal entre os indivíduos e espécies. A transferência horizontal (TH) é um fenômeno comum que pode afetar todos os tipos de TEs, sendo que vírus e parasitas podem exercer um papel facilitador na dispersão desses elementos entre as espécies (DOTTO et al., 2018; WALLAU, VIEIRA e

LORETO, 2018). Uma das consequências diretas da TH é a amplificação dos TE resultando num aumento substancial no tamanho de genoma, além da possibilidade de inovação biológica (SCHAACK, GILBERT e FESCHOTTE, 2010). Além disso, a TH pode iniciar um novo ciclo de vida de um TE ao colonizar um novo hospedeiro. Nessa fase inicial ocorre a propagação e amplificação do número de cópias dessas sequências no hospedeiro. Durante esse processo, a interação entre o TE e o hospedeiro é muito importante para a sobrevivência de ambos (KIDWELL e LISCH, 2001). Ao longo do tempo, com os mecanismos epigenéticos de silenciamento desenvolvidos pelo hospedeiro ou pelo acúmulo de mutações, ocorre uma redução da mobilidade desses elementos. Em seguida, os TEs podem ser eliminados quando todas as suas cópias são inativadas ou perdidas (SCHAACK, GILBERT e FESCHOTTE, 2010). Esses processos que envolvem o ciclo de vida dos TE determinam a sua distribuição ao longo dos táxons.

Como já mencionado, o processo de transposição pode ser potencialmente deletério para o hospedeiro. Assim, estratégias são utilizadas no controle da transposição para diminuir o seu impacto de maneira a beneficiar tanto o hospedeiro quanto o elemento de transposição (MUNHOZ-LÓPEZ e GARCIA-PÉREZ, 2010). Dentre elas, a metilação do DNA desempenha um papel dominante na supressão dos TEs sendo a estratégia mais amplamente adotada por eucariotos superiores para impedir a expressão dos elementos (DENIZ, FROST e BRANCO, 2019). Outro mecanismo de silenciamento dos TE são os pequenos RNAs que podem atuar de maneira pós-transcricional por meio da degradação do RNA, bem como transcricionalmente, induzindo a deposição de modificações epigenéticas repressivas na sequência dos TE (DENIZ, FROST e BRANCO, 2019).

No caso de transposons de DNA, a própria transposase pode atuar como um inibidor da transposição num mecanismo chamado inibição por superprodução, no qual a sua atividade é diminuída quando há um excesso de concentração (CLAEYS BOUUAERT, 2013; HICKMAN e DYDA, 2016). Ainda sobre o controle de transposons de DNA, embora esses elementos sejam comuns na natureza, a grande maioria apresenta múltiplas mutações, apresentando transposases inativas as quais podem competir com transposases ativas pelo mesmo sítio de ligação. Como duas moléculas de transposase funcionais são necessárias para realizar a transposição, as transposase inativas atuam como inibidores negativos da transposição (MUNHOZ-LÓPEZ e GARCIA-PÉREZ, 2010).

Assim, para a persistência de um TE ao longo da evolução, é preciso atingir um equilíbrio delicado entre expressão e repressão para que seja possível promover a amplificação moderada desse TE, sem levar a desvantagem para o hospedeiro (BOURQUE, 2018).

À medida que novos genomas foram sendo sequenciados nos últimos anos, novos tipos e novas famílias de TEs foram identificados e a distribuição dos TEs conhecidos tem sido constantemente atualizada. A catalogação da diversidade de TEs e o entendimento da biologia dessas sequências é crucial para a compreensão global de como os TEs afetam a evolução e o funcionamento dos organismos hospedeiros.

1.2 Transposons de DNA da ordem TIR

Os transposons de DNA foram os primeiros TEs a serem descobertos e descritos por Barbara McClintock na década de 40 ao estudar mutações no milho como consequências do movimento dessas sequências, aos quais ela chamou de elementos controladores. Esses elementos são amplamente estudados e já foram identificados em essencialmente todos os genomas sequenciados, desde procariotos e eucariotos (HICKMAN e DYDA, 2016).

Os transposons mais simples contêm apenas uma fase de leitura aberta (ORF – *open reading frame*) que codifica a transposase. No entanto, em procariotos, os transposons podem conter várias ORFs sendo necessários para sua mobilidade ou, até mesmo, importantes na disseminação de genes de resistência a antibióticos entre as populações bacterianas. Em procariotos, elementos contendo apenas a transposase são chamados de Sequências de Inserção (IS) (SIGUIER, GOURBEYRE e CHANDLER, 2014; HICKMAN e DYDA, 2016).

Nesses TEs, a região codificadora da transposase é flanqueada pelas sequências TIRs, locais de ligação e reconhecimento da enzima transposase. Ao se ligar às TIRs, a transposase realiza a quebra do DNA, retirando o transposon e o reinserindo em outro lugar do genoma, num mecanismo denominado de corte e cola. As enzimas transposases são tipicamente proteínas de múltiplos domínios que incluem um domínio catalítico capaz de clivar ambas as fitas de DNA. Essa etapa de corte e cola gera uma duplicação do sítio alvo (TSD – *target site duplication*), devido a propriedade de corte desigual da dupla fita de DNA (criando extremidades coesivas)

pela transposase (WICKER et al., 2007). As TSDs representam uma característica única para cada grupo de transposon de DNA (FESCHOTTE e PRITHAM, 2007).

Segundo Wicker e colaboradores (2007), esses transposons são divididos em 9 superfamílias. Porém, uma divisão mais detalhada dessas sequências pode ser encontrada no Repbase, um banco de sequências repetitivas de eucariotos (JURKA, 2000 - <https://www.girinst.org/repbase/>), onde esses elementos são divididos em 23 superfamílias que se diferenciam pela sequência das TIRs e o tamanho da TSD. Dentre essas superfamílias, 18 codificam uma transposase com motivo catalítico conhecido como DDE/D, que consiste em dois resíduos de ácido aspártico (D) e um terceiro resíduo de ácido glutâmico (E) ou D, localizado em um núcleo conservado de α -hélices e folhas β (YUAN e WESSLER, 2011).

1.2.1 Superfamília *Tc1/Mariner*

A superfamília *Tc1/Mariner* representa o grupo de transposons de DNA mais amplamente distribuído na natureza e foi descoberto em genomas de invertebrados durante a década de 1980 (YUAN e WESSLER, 2011). Elementos dessa superfamília possuem em geral de 1300-2400 pb de comprimento e possuem sítio preferencial de inserção adjacente a TA, gerando TSDs de dois nucleotídeos (TA). A transposase dos elementos dessa superfamília possui dois domínios: domínio de ligação ao DNA na porção N-terminal e o domínio catalítico na porção C-terminal (BRILLET, BIGOT e AUGÉ-GOUILLOU, 2007).

Assim como para outras superfamílias, o fato de esses elementos poderem se transpor usando somente a sua transposase e não necessitarem de fatores específicos do hospedeiro para a sua transposição faz com que sejam uma ferramenta genética atrativa, podendo ser utilizados como vetores de transformação em procariotos e eucariotos (LAMPE, 2010; ROBERTSON et al., 2002). De fato, o primeiro sistema de transformação por transposição desenvolvido para vertebrados foi baseado no transposon dessa superfamília chamado *Sleeping Beauty (SB)*, um elemento inativo isolado de peixes salmonídeos, que foi reconstituído para esse fim (IVICS et al, 1997).

O *SB* apresenta uma região codificadora para uma transposase de 340 aminoácidos que é flanqueada por duas TIRs de aproximadamente 230 pb. A porção N-terminal apresenta o domínio de ligação ao DNA com dois subdomínios chamados

PAI e RED, seguido de um sinal de localização nuclear. Esse sinal se sobrepõe ao domínio de ligação ao DNA e a porção C-terminal do domínio catalítico (SANDOVAL-VILLEGAS et al, 2021). O domínio catalítico é responsável pela clivagem do DNA e pelas reações de transposição e contém a tríade já mencionada (IVICS Z, IZSVÁK, 2014).

O processo de transposição do *SB* se inicia com a ligação da transposase às TIRs do transposon. O subdomínio PAI tem o papel dominante no processo de ligação ao DNA. Após a ligação, ocorre a formação de um complexo sináptico no qual as extremidades TIRs são emparelhadas e mantidas juntas por quatro moléculas de transposases. Após a montagem do complexo sináptico, ocorre a excisão do transposon e o reparo da quebra de dupla fita de DNA. A etapa de excisão resulta na exposição de um grupo 3'OH livre nas extremidades do transposon o qual é essencial para a etapa de integração agindo como nucleófilo atacando a ligação fosfodiéster do DNA alvo. A integração do transposon contendo duas extremidades coesivas no DNA alvo resulta em lacunas de fita simples que são preenchidas pelo mecanismo de reparo de DNA, gerando TSDs (IVICS Z, IZSVÁK, 2014; OCHMANN e IVICS, 2021). Como para os demais membros dessa superfamília, a integração da *SB* ocorre preferencialmente nos dinucleotídeos TA, embora a integração de *SB* possa ocorrer de forma rara em um local não TA (GUO, ZHANG e HU, 2018).

Pela sua ampla distribuição, elementos dessa superfamília são muito estudados em eucariotos desde a sua distribuição (SCHEMBERGER et al., 2016; DA SILVA et al., 2017; CAVALCANTE et al., 2020), e história evolutiva em diferentes grupos filogenéticos (BERTOCCHI et al., 2017; DUPEYRON, et al., 2020; GAO et al., 2020; WANG et al., 2021; PALAZZO et al., 2021) até a descoberta de novas famílias (SANG et al., 2019; ZONG et al., 2020).

1.2.2 Superfamília *Merlin*

Ao contrário da superfamília *Tc1-Mariner*, há poucos relatos na literatura sobre a superfamília *Merlin*, descrita pela primeira vez em 2004 (FESCHOTTE, 2004) como um grupo de elementos detectados por análise computacional em diferentes genomas, que compartilham características estruturais e motivos de sequência comuns. Esses elementos possuem TIRs que variam em comprimento de 24 a 462 pb e são acompanhadas por 8-9 pb de TSDs. Os elementos *Merlin* têm proximidade

evolutiva ao grupo IS1016 de procariotos, compartilhando similaridade de sequência das proteínas e o tamanho da TSD. O grupo IS1016 junto com IS*Pna2*, IS*H4*, IS1595, IS*Sod11*, IS*Nwi* e IS*Nha5* constituem um grupo principal, denominado de família IS1595 (SIGUIER, GAGNEVIN e CHANDLE, 2009). Os membros desta família IS são geralmente flanqueados por TSDs de 8 pb, têm um único gene de transposase e, exceto para o grupo IS1016, todos os outros têm um domínio de dedo de zinco N-terminal (SIGUIER, GAGNEVIN e CHANDLE, 2009) denominado Zn_Tnp_IS1595. Esses elementos também têm um domínio C-terminal conservado contendo o motivo catalítico DDE compartilhado com os transposons *Merlin* de eucarióticos (FESCHOTTE, 2004; SIGUIER, GAGNEVIN e CHANDLE, 2009) denominado DDE_Tnp_IS1595. Todos os elementos *Merlin* e IS1016 descritos até agora não possuem o domínio Zn_Tnp_IS1595.

Esses elementos foram encontrados principalmente em animais, mas também em oomicetos, microsporídeos (FESCHOTTE, 2004), fungos e embriófitas (YUAN e WESSLER, 2011). Nenhum outro trabalho desde 2004 estudou a distribuição, características e evolução desses elementos.

1.3 Transposons e sistemas de expressão de proteínas de interesse terapêutico

A possibilidade de produzir proteínas de interesse em sistemas recombinantes como bactérias, fungos e animais teve um profundo impacto em várias áreas de pesquisa. As proteínas com interesse terapêutico formaram a base da biotecnologia farmacêutica nas últimas décadas e têm sido utilizadas para o tratamento de diversas doenças em humanos de maneira inovadora e eficiente. A indústria de biotecnologia movimentou, nos anos 2014 a 2017, 651 bilhões de dólares com 374 produtos biofarmacêuticos, incluindo fatores plasmáticos sanguíneos, anticoagulantes, vacinas, hormônios, fatores de crescimento, anticorpos monoclonais, entre outros. Esses produtos têm como alvo uma série de doenças incluindo câncer e doenças auto-imune (WALSH, 2018).

Organismos mais complexos como leveduras, células de inseto e células mamíferas têm sido utilizados para a expressão de proteínas recombinantes devido a presença de uma maquinaria mais complexa para expressão. O uso de leveduras proporciona alto rendimento de proteínas recombinantes secretadas a um custo, relativamente, baixo. A levedura metilotrófica *Pichia pastoris* é um dos sistemas mais

importantes para expressão heteróloga devido a presença de fortes promotores induzíveis por metanol e a sua alta capacidade de secreção (ZAHRL, 2017). No entanto, as leveduras expressam diversas proteases as quais levam a degradação da proteína e não são capazes de produzir padrões de glicosilação mais complexos como os encontrados em mamíferos (DALTON e BARTON, 2014).

Dentre os sistemas de expressão que utilizam células animais, as células de inseto têm sido bem utilizadas por apresentar um sistema de expressão de eucariotos superior além de vantagens como maior rendimento que as células mamíferas e alto nível de expressão quando infectado por baculovírus. Esse sistema é relativamente simples de usar, mas apresenta limitações no uso em larga escala (CHAA et al., 2005). Apesar do sistema célula de inseto/baculovírus apresentar a capacidade de processamento eucariótico de proteína sendo capaz de realizar modificação, dobramento, tráfego e montagem de polipeptídeos autênticos e solúveis, esse processamento apresenta limitações como é o caso da N-glicosilação, a qual é mais simples do que nos mamíferos (KOST, CONDREAY e JARVIS, 2005; SHI e JARVIS, 2007).

Os padrões de glicosilação de proteínas expressas em células mamíferas são mais consistentes com as observadas *in vivo* apresentando um papel importante na atividade biológica da proteína. Além disso, essas células têm a capacidade de realizar um sofisticado processamento pós-traducional bem como dobramento e montagem das proteínas de maneira autêntica (WALSH, 2014). Dessa maneira, o cultivo de células mamíferas tem se tornado um sistema dominante para a produção de proteínas com aplicações clínicas visto que, para o uso terapêutico, a glicosilação tem uma grande influência nas propriedades físicas e funcionais de glicoproteínas recombinantes, e é considerada um ponto importante para eficácia e segurança de um fármaco (HARTLEY, 2012; BUTLER e SPEARMAN, 2014; KUYSTERMANS e AL-RUBEAI, 2015).

Diversas linhagens celulares de mamíferos têm sido utilizadas para a produção de proteínas recombinantes, como por exemplo CHO (*chinese hamster ovary*), HEK-293 (*human embryonic kidney*), Per.C6 (derivado de retina humana), BHK (*baby hamster kidney*) e derivados de mieloma de camundongo, Sp2/0-Ag14 e NS0 (WURM, 2004; JAYAPAI et al. 2007; NOH, SHIN e LEE, 2019). Entre estas, a célula CHO é mais comumente usada, sendo atualmente responsável pela produção de cerca de 75% de todas as proteínas terapêuticas (WASH, 2018). De origem

epitelial, essa célula tem capacidade de se adaptar e crescer em culturas de suspensão com alta densidade celular que é ideal para produção em larga escala. Essas células têm a capacidade para realizar modificações pós-traducionais compatíveis e bioativos em humanos (KIM et al., 2012) e os padrões de glicosilação são semelhantes aos padrões encontrados em humanos (LAI, 2013; LALONDE e DUROCHER, 2017).

Apesar da possibilidade de expressão transiente de proteínas, a abordagem mais utilizada na indústria biofarmacêutica para a produção de proteínas recombinantes de interesse terapêutico é a expressão a partir de células estáveis. A obtenção de linhagens nas quais o gene de interesse é integrado de forma estável no genoma da célula hospedeira garante a produção por longos períodos e a reprodutibilidade. O primeiro passo no desenvolvimento de uma linhagem celular é a introdução do DNA recombinante dentro da célula hospedeira. Diversas metodologias são descritas na literatura para a transfecção do DNA, como precipitação com fosfato de cálcio, eletroporação, lipofecção, polifecção e transfecção retroviral (WURM, 2004; O'FLAHERTY et al, 2020). Uma vez no núcleo da célula, o DNA deve ser integrado no genoma para permitir a estabilidade da expressão da proteína. Diferentes técnicas têm sido utilizadas para a integração genômica como vetores virais, troca de cassete mediado por recombinação (RMCE), integrase (BÜSSOW, 2015), e ainda transposons de DNA que têm demonstrado aumentar significativamente a integração de genes no genoma da célula alvo de maneira estável (MATASCI et al., 2011; PETRAKIS et al., 2012).

Dependendo da metodologia utilizada, o local de integração do DNA será aleatório e o nível de expressão do gene pode ser afetado pela estrutura da cromatina ao redor do gene integrado (JAYAPAL, 2007). Dessa maneira, várias estratégias podem ser usadas para garantir a expressão gênica. Por exemplo, modificações no vetor de expressão como uso de promotores e elementos reguladores de DNA permitem atividade transcricional independentemente do local de inserção do gene no genoma (STERN, 2007). Dentre os elementos reguladores de DNA, podemos citar regiões controladoras de locus (LCR - *locus control region*), regiões de ligação à matriz nuclear (MARs - *matrix attachment regions*) e elementos de abertura de cromatina ubíquo (uCOEs - *ubiquitous chromatin-opening elements*) (LAI, 2013). Insuladores são elementos de DNA que possuem a habilidade de proteger o gene de interesse de sinais inapropriados e agem bloqueando a ação de um *enhancer* distante ou agem

como uma barreira contra o silenciamento do gene (WEST et al., 2002; VOGELMANN et al., 2011).

Após a integração do gene na célula hospedeira, um sistema de seleção é necessário para eliminar células não recombinantes. Agentes seletivos como antibióticos e proteínas fluorescente podem ser utilizadas nesta etapa. Células que integram estavelmente um vetor de expressão contendo GFP (*green fluorescent protein*), por exemplo, são diretamente identificadas por fluorescência intracelular e podem ser isoladas por citometria de fluxo (BÜSSOW, 2015).

Por fim, a seleção de células únicas é uma etapa crucial para o desenvolvimento de linhagens estáveis, pois permite a homogeneidade entre as células da mesma população (clones). O método convencional de diluição limitante é demorado e de baixa eficiência. Recentemente, a citometria de fluxo e *cell-sorting* permitem de forma automatizada a seleção de células de alta produção. Nessa metodologia é possível utilizar a co-expressão de um gene fluorescente que permite o monitoramento da expressão em tempo real e, ainda, a seleção das células únicas recombinantes através da intensidade de fluorescência reduzindo significativamente o tempo de seleção (LAI et al., 2013; BANDARANAYAKE e ALMO, 2014).

1.4 Sistema de transposição

Como já mencionado, uma etapa importante para a produção de proteínas recombinantes de interesse terapêutico é a introdução do DNA recombinante no genoma da célula hospedeira. A integração cromossômica pode ocorrer como um evento randômico numa taxa extremamente baixa (KOLACSEK et al., 2014a). Dessa forma, diferentes estratégias são utilizadas para aumentar a eficiência dessa etapa, como vetores virais, nucleases como ZFN e TALEN, integrases e os transposons de DNA.

Os transposons de DNA podem ser vistos como veículos naturais de transferência de DNA, seguros e com baixo custo de produção. A característica da enzima transposase que permite mobilizar sequências que contenham as suas respectivas TIRs, constitui a base de um sistema de transposição para inserção genômica. Nesse tipo de sistema, o DNA de interesse é clonado entre as TIRs de um transposon em um vetor (comumente chamado de vetor *donor* ou doador) que pode ser introduzido na célula juntamente com a transposase codificada como um

plasmídeo de expressão separado (chamado de vetor *helper* ou auxiliar), como RNA mensageiro ou como proteína recombinante. A transposase reconhece as TIRs e realiza a excisão do DNA presente no vetor e a sua integração no genoma da célula hospedeira, permitindo a sua expressão contínua (AMBERGER e IVICS, 2020). Essas características tornam os transposons uma ferramenta molecular para transferência de maneira eficiente e potencialmente menos tóxico e menos imunogênico, de fácil manipulação e produção em grande escala (HUDECEK et al, 2017; SANDOVAL-VILLEGAS et al, 2021).

Como já mencionado, o primeiro sistema de transposição desenvolvido para vertebrados foi baseado no transposon *SB*, elemento inativo isolado de peixes salmonídeos (IVICS,1997). Esse sistema foi criado a partir da síntese de uma sequência consenso eliminando mutações de códon de parada e de mudança de quadro de leitura. Esse sistema apresentou boa eficiência de transferência gênica mediada por transposição para diversos vertebrados, incluindo células humanas e murinas (IVICS,1997; LUO, 1998). A segunda geração desse sistema, denominada *SB11*, foi desenvolvida para aprimorar a atividade de transposição da *SB*. A substituição de 4 aminoácidos, incorporados na sequência da *SB11*, levou ao aumento de 3 vezes na atividade de transposição em relação a *SB* original (GEURTS et al., 2003). A versão mais hiperativa da transposase *SB* atualmente disponível, *SB100X*, exibe uma taxa de transposição 100 vezes maior que a proteína original (MATES et al., 2009). Essa versão demonstrou alta eficiência de transferência gênica tanto em linhagens germinativas em modelos mamíferos quanto em células-tronco (MATES et al., 2009; KATTER et al., 2013).

Vários outros transposons de diferentes espécies têm se mostrado ativos em vertebrados: *Frog Prince* de rã *Rana pipiens* (MISKEY, 2003), *PiggyBac* do lepidóptero *Trichoplusia ni* (DING, 2005), *Minus* de *Drosophila*, *Tol1* (KOGA, 2007), *Tol2* (KAWAKAMI, 2000) e *Passport* de peixes (CLARK, 2009), *Hsmar1* de humano (MISKEY, 2007), *Himar1* da mosca *Haematobia irritans* (ZHANG, 1998), *Ppmar* de bamboo (BING, 2017) entre outros (NI 2008).

Os sistemas de transposição são amplamente utilizados para uma variedade de aplicações como transferência gênica, mutagênese insercional, manipulações genômicas, terapia gênica e indução de células-tronco. Uma importante aplicação é a mutagênese insercional para a identificação de novos genes (HUDECEK et al., 2017;

TIPANEE et al, 2017, HUDECEK e IVICS, 2018). Os sistemas *SB*, *PiggyBac* e *Tol2* são os mais amplamente estudados e utilizados atualmente.

A funcionalidade da transposase *PiggyBac* (*PB*), pertencente a superfamília de mesmo nome, foi confirmada pela primeira vez em 1996 com a mobilização bem-sucedida em inseto (ELICK, BAUSER e FRASER, 1996). O *PB* possui TIRs com 13 pb e a transposase apresenta 594 aa e contém domínios de ligação ao DNA, um domínio catalítico incluindo a tríade catalítica DDD (posições 268, 346 e 447) interrompida por um domínio de inserção, um domínio N-terminal ácido e um domínio C-terminal rico em cisteína que contém um sinal de localização nuclear. Durante o mecanismo de transposição, o *PB* se insere em sequências TTAA, que são duplicadas e flanqueiam o transposon inserido após a integração (YUSA, 2015; SANDOVAL-VILLEGAS et al, 2021). O sistema *PB* já foi utilizado em uma ampla variedade de organismos desde levedura (MITRA, FAIN-THORNTON e CRAIG, 2008), plantas (JOHNSON e DOWD, 2014) e mamíferos (DING et al, 2005), incluindo humanos (WILSON, COATES e GEORGE, 2007) e atualmente o sistema *PB* é comercializado para uso em pesquisa (heraBioLabs, SBI SystemBioscience).

O elemento *Tol2*, pertencente a superfamília *hAT*, foi identificado a partir do genoma do peixe de água doce *Oryzias latipes* (KOGA et al, 1996). Este foi o primeiro relato de um transposon de DNA ativo em vertebrados. O *Tol2* possui 4700 pb de comprimento e contém uma região codificadora para proteína transposase composta de quatro éxons que leva a criação de diferentes isoformas de transposases onde a variante mais ativa é a de 649 aa (NI et al, 2016). Esse elemento contém TIRs imperfeitas de 17 bp e 19 bp, apresenta a tríade DDE catalítica (D192, D258 e E611), o motivo RW típico dos elementos da família *hAT*, domínio de ligação ao DNA e o domínio dedo de zinco BED. *Tol2* gera duplicação do sítio alvo de 8 pb e integração preferencial em regiões ricas em AT. O sistema *Tol2* tem alta capacidade de carga (> 11 Kb), baixa suscetibilidade à inibição de superprodução e têm se tornado agente popular de transferência de genes em uma variedade de sistemas animais, incluindo *zebrafish*, sapo africano, galinha, culturas de células de camundongos e humanos, incluindo células T primárias (KAWAKAMI, 2007; SANDOVAL-VILLEGAS et al, 2021).

Um estudo em 2006 (WU et al, 2006) comparou a atividade de transposição dos elementos *SB11* (primeira versão hiperativa da *SB*), *Tol2*, *PB*, e *Mos1*, em quatro linhagens de células mamíferas. *PB* foi o elemento que apresentou maior taxa de transposição em todas as células testadas (HeLa, CHO, HEK293 e H1299). O

elemento *SB11* apresentou moderada atividade nas células, com exceção da H1299, o qual foi inativo. Para as linhagens CHO e HeLa, o *SB11* teve cerca de 2,5 vezes mais eficiência que o *ToI2* (HUANG et al, 2010).

Como já mencionado, os transposons devem coexistir com seus hospedeiros e existe um balanço entre amplificação e repressão de TEs nos genomas. Assim, naturalmente, transposases são subótimas para evitar mutagênese insercional (MÁTÉS, 2009). Dessa forma, aumentar a atividade transposicional é um dos principais passos para o desenvolvimento de um sistema de transposição eficiente. Algumas estratégias já foram aplicadas para gerar mutantes hiperativos da *SB*, como importar aminoácidos e pequenos blocos de aminoácidos presentes em outras transposases relacionadas filogeneticamente com alto grau de conservação (GEURTS, 2003; ZAYED, 2004; BAUS, 2005), substituição racional de resíduos de aminoácidos selecionados (BAUS, 2005) e PCR *error prone* para *PB* (LI et al., 2013). Uma triagem genética de alto rendimento com aproximadamente 2 mil combinações possíveis de aminoácidos, produziu variantes da *SB* com atividades significativamente melhores. A versão mais hiperativa, denominada *SB100X*, apresentou taxa de eficiência 100 vezes maior que a *SB* original com as seguintes modificações: K14R; K33A; R115H; RKEN214-217DAVQ; M243H; T314N (MÁTÉS et al, 2009). Uma versão hiperativa do *PB* foi gerada combinando 7 mutações de aminoácidos, o qual apresentou um aumento de 17 vezes no processo de excisão e 9 vezes na integração (DOHERTY et al., 2012).

As versões hiperativas do elemento *PB* e *SB* foram comparadas em células humanas HEK293, HeLa e linfócitos T (DOHERTY et al., 2012). Em todas as concentrações testadas, o *PB* hiperativo demonstrou maior eficiência de transposição em ambas as células, sendo de 1.5 a 12 vezes mais ativo que o *SB100X*, dependendo da dosagem de DNA testado (DOHERTY et al., 2012). Quando comparada a habilidade de gerar linhagem celular para a produção de proteínas recombinantes em larga escala em células CHODG44, os elementos *PB* hiperativo e *SB100X* produziram resultados similares. Esses transposons apresentaram maior quantidade de células recombinantes estáveis em comparação ao *ToI2* (BALASUBRAMANIAN et al., 2016). Esta observação pode ter ocorrido pelo fato das transposases *PB* e *SB* serem versões aprimoradas e mais efetivas, enquanto a sequência codificadora da transposase de *ToI2* não foi modificada. Isso demonstra a importância desse aprimoramento das transposases para melhorar seu desempenho nos sistemas de transposição.

Uma questão importante dos sistemas de transposição é o local de inserção do transgene, pois a integração pode causar eventos adversos, como a ativação de proto-oncogenes ou a inativação de genes celulares essenciais ou ainda o transgene pode ser silenciado dependendo do estado da cromatina no local da inserção (YOSHIDA et al., 2017). Várias abordagens foram desenvolvidas para isolar os locais de inserção, incluindo PCR invertido, PCR vectorette, PCR splinkerette, entre outros. A PCR splinkerette é uma variante da PCR mediada por ligação e foi desenvolvida para superar algumas limitações presentes nas outras técnicas e se tornou uma das técnicas mais usadas atualmente, e aparentemente mais eficiente, para mapear os locais de inserção do transposon (UREN et al, 2009).

1.5 Sistema Ingresso

O Sistema Ingresso é uma plataforma de vetores para superexpressão heteróloga baseada em transposição, desenvolvida em nosso laboratório (Fig 1) (Pedido de Patente: WO2019010551A1). Esse sistema foi desenhado para garantir alto grau de integração do cassete de expressão nos cromossomos da célula hospedeira através do uso de sistema de transposição. Esse sistema foi desenvolvido para o uso em células CHO que são, atualmente, as mais utilizadas para expressão de proteínas terapêuticas.

Nesse sistema, a transposase do *SB* é expressa de maneira transiente pela transfecção do vetor *helper* que contém o gene da transposase. O vetor de expressão (*donor*) possui alguns componentes essenciais como a sequência da TIR específica do *SB* em ambas as extremidades que serão utilizadas no reconhecimento pela transposase para o processo de integração genômica; insuladores que garantem alto nível de transcrição independentemente do sítio de integração cromossômica; promotor CAG ou CMV; um gene repórter fluorescente que permite o monitoramento da expressão em tempo real e a seleção de células altamente produtoras da proteína de interesse por *fluorescence-activated cell sorter* (FACS); e um sítio interno de ligação ao ribossomo (IRES - *Internal ribosome entry sites*) para a expressão do gene alvo e marcador em um mesmo mRNA.

O insulador presente no sistema é o *CHS4* (*chicken hypersensitive site 4*) de galinha que está presente em ambas as extremidades do cassete de expressão e

apresenta tanto atividade de bloqueio de *enhancer* quanto atividade de barreira de silenciamento (BARKESS e WEST, 2012), impedindo, assim, o silenciamento do transgene independentemente do local de integração.

O promotor é um elemento crítico no sistema, pois pode influenciar no aumento da expressão e melhorar a estabilidade do gene de interesse. Os promotores de citomegalovírus (CMV) e CAG (uma combinação do CMV e o promotor de β -actina de galinha) são comumente utilizados para dirigir a expressão de proteínas recombinantes em células de mamíferos e são associados a altos níveis de expressão gênica (ZHU, 2012; ZHONG-JIE et al., 2019).

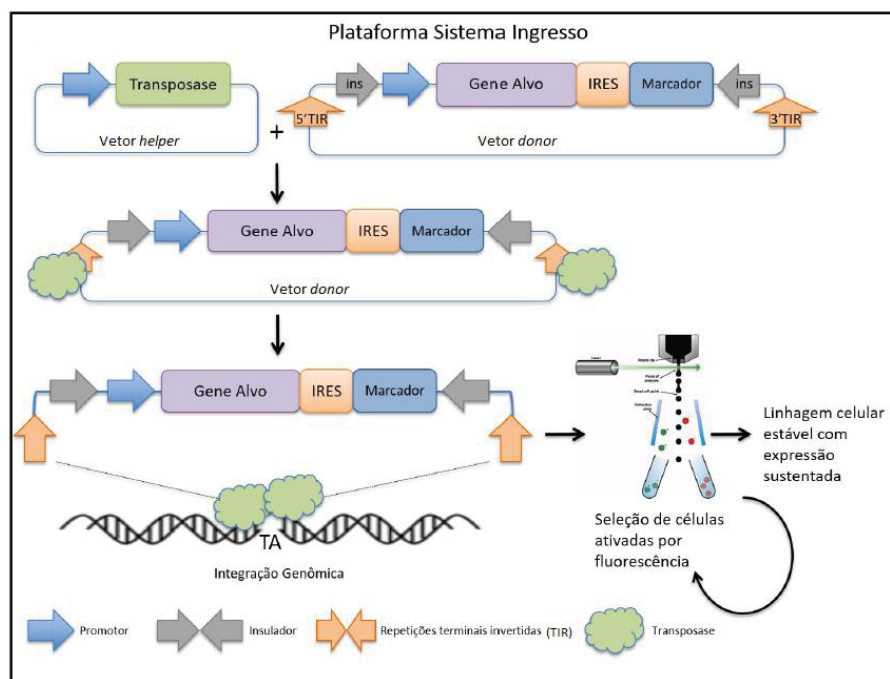


Figura 1. Plataforma do Sistema Ingresso. Este sistema possui dois vetores diferentes: um contendo o gene codificador da transposase *SB* (*vetor helper*) e um segundo contendo o gene de interesse a ser expresso juntamente com um marcador (*vetor donor*). O *vetor helper* tem como objetivo a expressão transiente da transposase, que será responsável por integrar a sequência alvo, presente no segundo vetor, no genoma. O *vetor donor* é composto pelas sequências 5' TIR e 3' TIR do transposon *SB*, insuladores (HS4 de *Gallus gallus*) em ambas as extremidades dos cassetes de expressão para regulação epigenética, um promotor forte (CAG ou CMV) *upstream* ao gene alvo de interesse, seguido da sequência IRES para ligação interna ao ribossomo e o gene marcador codificando uma proteína fluorescente (GFP, RFP, YFP) ou proteína de superfície (CD20, CD25, CD90) para seleção por FACS.

Recentemente, uma versão aprimorada da versão original, denominada Sistema Ingresso 1.0, foi desenvolvida acrescentando uma sequência 5'UTR, denominada de elemento tripartido (TPE – *Tripartite element*) ou elemento líder, na região *upstream* do gene de interesse. Essa sequência facilita a ligação ao ribossomo

aumentando a tradução da proteína. Além disso, em células mamíferas, a porção 5'UTR desempenha um papel importante na regulação da expressão gênica (BAIR et al., 2017; RAMKE et al., 2017). Essa nova construção foi testada e mostrou maior expressão da proteína de interesse em comparação com o vetor original (CAMPOS, 2021).

1.6 JUSTIFICATIVA

Com a finalidade de produção em larga escala de proteínas de interesse biotecnológico, o Sistema Ingresso previamente desenvolvido ainda está sendo aprimorado e testado pelo grupo com o objetivo de obter resultados melhores dessa produção. A adição do elemento TPE no vetor *donor* já mostrou uma melhora significativa na expressão do gene de interesse em nível de proteína. Entretanto, para entender o padrão de funcionamento desses vetores nas células, outros aspectos também precisam ser avaliados, como por exemplo, o número de insertos por células, padrão de expressão em nível de RNA e local de inserção. Assim, o primeiro objetivo desse trabalho consistiu nessa caracterização básica do Sistema Ingresso 1.0 baseado no transposon *SB*.

Outro aspecto passível de melhora do Sistema Ingresso é a transposase utilizada para promover a integração do cassete. Os sistemas mais eficientes como *piggyBac*TM e *SB100X* só podem ser usados para fins de pesquisa e seu uso para desenvolvimento de produto é restrito ou condicionado a pagamento de direitos autorais atrelados às patentes. Considerando a imensa diversidade de transposons de DNA que já foram encontrados e/ou descritos, nossa proposta foi escolher novas transposases para serem testadas em sistema de transposição com posterior tentativa de otimização por mutações na sua sequência. Com esse objetivo, anteriormente ao início deste projeto, foram analisadas centenas de transposases para selecionar candidatas para o trabalho. Essa análise consistiu no levantamento de diferentes transposons de diferentes superfamílias no Repbase e na literatura. Além disso, foram realizadas análises das sequências dos transposons observando a presença de TIRs bem definidas, presença de domínios proteicos característicos da superfamília da transposase em questão, tamanho das sequências da transposase, análises de similaridade e proximidade filogenética com outros elementos, etc. De um total de 640

elementos considerados nas análises, três foram selecionados como os mais promissores para iniciar este projeto, sendo codificados aqui (devido a possibilidade de patente) como *Tp2* (superfamília *Tc1-Mariner*), *Tp5* (superfamília *hAT*) e *Tp7* (superfamília *Piggybac*). Embora as clonagens iniciais foram realizadas para os três elementos, o projeto seguiu apenas com a *Tp2*.

Algumas superfamílias de transposons de DNA nunca foram testadas em sistemas de transposição e carecem de conhecimentos básicos, como a superfamília *Merlin* que foi descrita em 2004, tendo sido explorada apenas nesse trabalho original. Tendo em vista a importância do estudo dos TEs *per se* e também a relevância do entendimento dessas sequências para aplicações biotecnológicas, outra proposição dessa tese foi o estudo da superfamília *Merlin* em eucariotos avaliando a distribuição, conservação e evolução de grupo de TE.

2 OBJETIVOS

2.1 OBJETIVO GERAL

O objetivo principal deste trabalho é estudar transposons de DNA em diferentes contextos com uma abordagem biotecnológica através da caracterização e melhoramento de um sistema de transposição baseado em transposases da superfamília *Tc1/Mariner* e uma abordagem biológica/evolutiva pelo estudo da distribuição e evolução de elementos *Merlin* em eucariotos.

2.2 OBJETIVOS ESPECÍFICOS

2.2.1 CAPÍTULO I - Caracterização e melhoramento de um sistema de transposição baseado em transposases da superfamília *Tc1/Mariner*.

- a) Verificar o número de cópias do inserto em diferentes populações/clones expressando proteína de interesse a partir do Sistema Ingresso 1.0;
- b) Verificar a expressão relativa do gene de interesse nessas diferentes populações/clones;
- c) Determinação do local de inserção dos insertos nessas diferentes populações/clones;

- d) Verificar a perda esperada do vetor *helper* nessas diferentes populações/clones;
- e) Modificar os vetores do Sistema Ingresso original trocando as sequências da transposase e TIRs do *SB* por sequências dos novos transposons escolhidos.
- f) Caracterizar e comparar a eficiência de transposição do elemento *Tp2* em relação ao elemento *SB* no Sistema Ingresso.

2.2.2 CAPÍTULO II - Estudo da distribuição e evolução de elementos *Merlin* em eucariotos.

- a) Determinar a presença de sequências *Merlin* em diferentes grupos eucariotos;
- b) Caracterização básica de cópias de *Merlin* em grupos taxonômicos específicos;
- c) Inferência das relações evolutivas dos elementos *Merlin* encontrados.

CAPÍTULO I

Caracterização e melhoramento de um sistema de transposição baseado em transposases da superfamília *Tc1/Mariner*

3 MATERIAL E MÉTODOS

3.1 DESENHO EXPERIMENTAL: CARACTERIZAÇÃO DO SISTEMA INGRESSO

1.0.

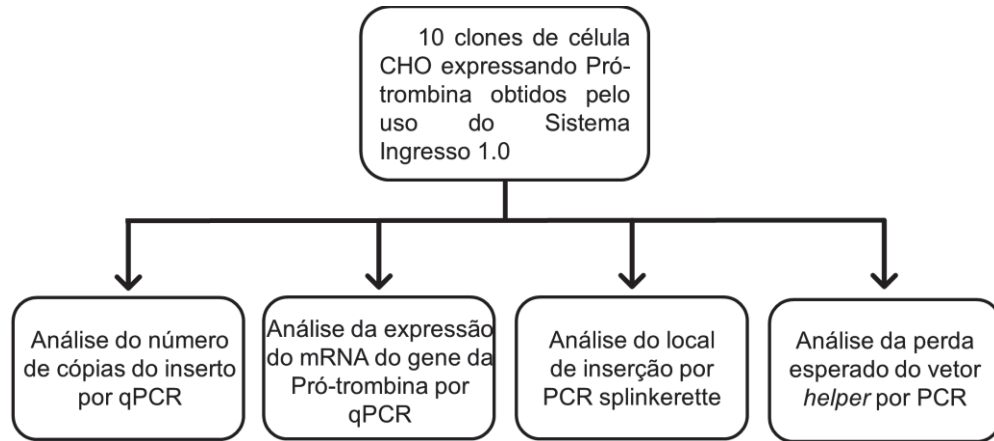


Figura 2. Desenho experimental da caracterização do Sistema Ingresso 1.0.

3.2 DESENHO EXPERIMENTAL: CARACTERIZAÇÃO DE NOVA TRANSPOSASE PARA SISTEMA DE TRANSPOSIÇÃO.

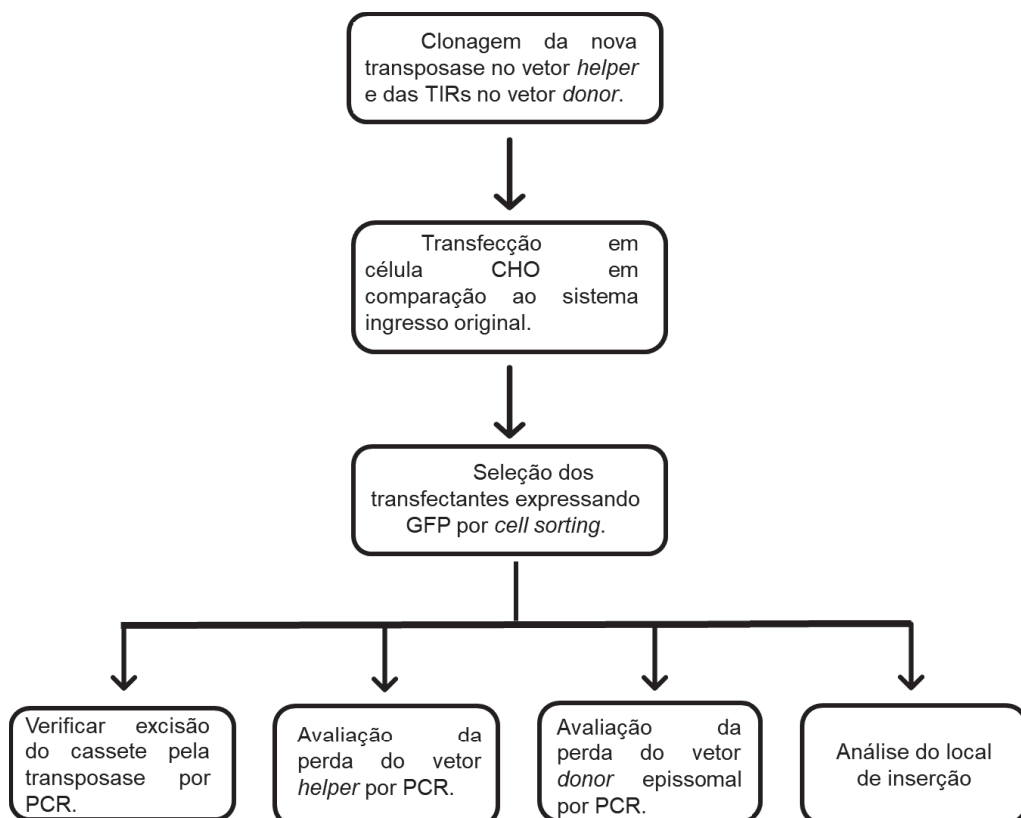


Figura 3. Desenho experimental da caracterização de nova transposase para sistema de transposição.

3.3 VETORES

Para a proposta de teste de novas transposases no Sistema Ingresso, foram utilizados os seguintes vetores da Plataforma Sistema Ingresso: 1) vetor *helper*, contendo o gene da transposase versão *SB11* sob controle do promotor CMV (Fig 4); 2) vetor *donor* do Sistema Ingresso original contendo o cassete de expressão e as TIRs 5' e 3' da *SB* (Fig 5). Nesse vetor inicial nenhum gene de interesse foi adicionado.

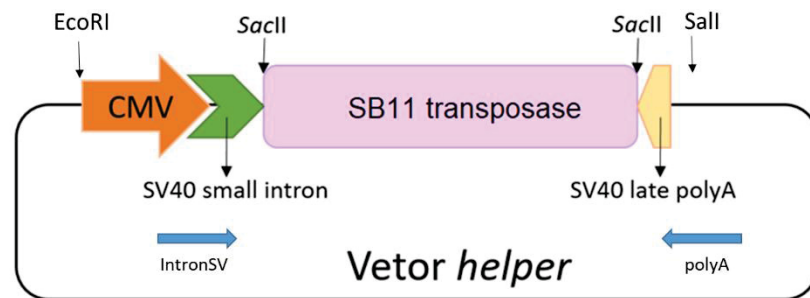


Figura 4. Mapa do vetor *helper* do Sistema Ingresso. Contém a região codificadora da transposase *SB11*, flanqueada por sítios da enzima *SacII*, o promotor CMV, e sinais de para processamento correto do mRNA, SV40 small intron (intron que deve ser processado por *splicing*) e sinal de poliadenilação SV40. Os primers IntronSV e polyA estão indicados na figura no local de anelamento.

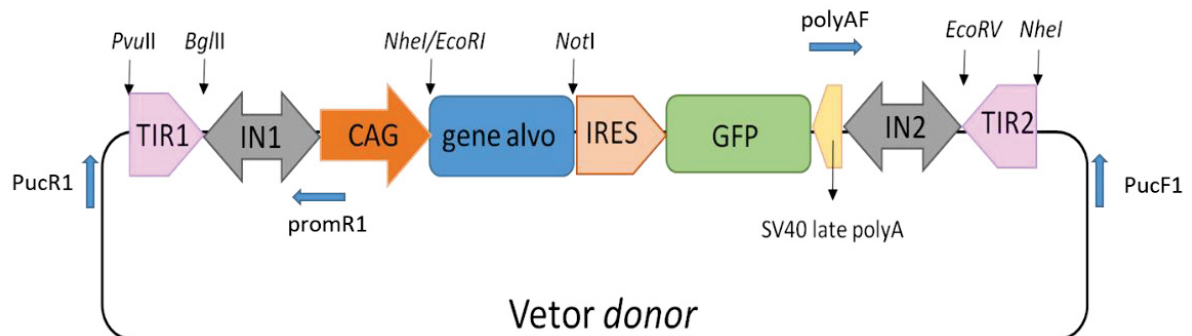


Figura 5. Mapa do vetor *donor* do Sistema Ingresso. Vetor de expressão contendo sequências de reconhecimento da transposase -TIR1 5' e TIR2 3', insuladores IN1 e IN2, promotor CAG, o gene de interesse, sequência IRES e o gene GFP. Os primers PucR1, promR1, polyAF e PucF1 estão indicados na figura no local de anelamento.

3.4 LINHAGEM CHO

As células ExpiCHO-S™ (Thermo Fisher Scientific) são derivadas de um subclone não modificado isolado a partir de células CHO-S. As células EXPICHO foram mantidas em meio Dynamis™ (Thermo Fisher Scientific) a 37°C, 8% de CO₂ e

120 rpm de agitação e se duplicam a cada 17 horas. As culturas foram mantidas com repiques a cada 4 dias de 5×10^5 células/ml.

3.5 CULTURAS DE CÉLULAS CHO EXPRESSANDO PROTEÍNA DE INTERESSE

Para a caracterização do Sistema Ingresso 1.0, foram selecionados clones isolados de células CHO expressando a proteína Pró-trombina humana e o marcador de seleção GFP clonados no vetor *donor*. Conforme já mencionado, o vetor *donor* do Sistema Ingresso 1.0 (Fig 6) contém uma região 5'UTR TPE e apresentou uma melhora significativa na expressão da proteína de interesse em relação ao vetor *donor* do Sistema Ingresso original.

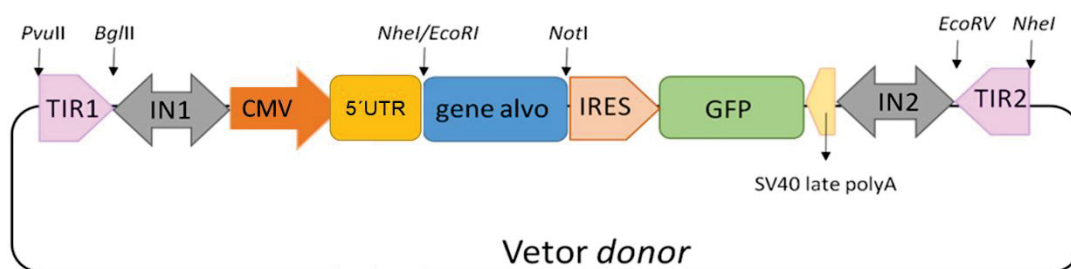


Figura 6. Mapa do vetor *donor* do Sistema Ingresso 1.0. Vetor de expressão contendo sequências de reconhecimento da transposase - TIR1 5' e TIR2 3', insuladores IN1 e IN2, promotor CMV e um fragmento do promotor CAG, sequência 5'UTR (TPE), gene alvo (Pró-trombina), sequência IRES e o gene GFP.

As culturas de células CHO expressando a proteína de interesse e o GFP foram obtidas pelo grupo de Proteínas Terapêuticas do IBMP. Células únicas foram selecionadas utilizando o equipamento Namocell Single Cell Dispenser (Namocell) e o marcador de seleção GFP, para a obtenção de clones. As células individualizadas denominadas de clones 5, 6 e 7 foram dispensados em meio condicionado e avaliados numa primeira etapa. Numa etapa posterior, uma única célula individualizada dos clones 4, 15, 28, 3 e 92 foi dispensada por poço em placa de 96 poços contendo meio semi-sólido Clonacell (ClonaCell™-CHO CD Medium - Stem Cell Technologies) e aditivos. Após o crescimento das células em erlenmeyer com (clones 4, 15, 28, 3 e 92) ou sem (clones 5, 6 e 7) retroalimentação, a quantificação da produção de Pró-trombina foi realizada por ensaio de Elisa nos 10 clones selecionados (Tabela 1). A população que deu origem as células individualizadas foi avaliada juntamente com os

clones, sendo que a população com crescimento sem retroalimentação foi denominada de 1 e a população com retroalimentação foi denominada 2. A população 2 foi avaliada posteriormente e possui um número maior de passagens.

Tabela 1. Análise quantitativa de Pró-trombina em sobrenadante de culturas de células provenientes de população e linhagens clonais do Sistema Ingresso 1.0 por ensaio de Elisa (Dissertação Bruna E. R. Campo, 2021 e dados obtidos pelo grupo Proteínas Terapêuticas do IBMP).

Amostra	Concentração de Pró-trombina (g/L)
28	0,40
4	0,35
15	0,35
5	0,36
7	0,26
6	0,13
92	0,05
3	0,02
População 1	0,35
População 2	0,25

3.6 EXTRAÇÃO DE ÁCIDOS NUCLEÍCOS

O DNA genômico das células CHO selvagens e das células CHO transfectadas foi extraído utilizando o High Pure PCR Template Preparation Kit® (Roche Diagnostics GmbH®), de acordo com o manual do fabricante. O RNA total das células CHO foi extraído de acordo com o kit RNAeasy (Qiagen) e tratado com DNaseI (Qiagen) conforme orientações do fabricante. A mensuração da concentração de DNA e RNA foi realizada através do uso do espectrofotômetro NanoDrop (Thermo Fisher Scientific).

3.7 SÍNTESE DA PRIMEIRA FITA DE CDNA

A síntese da primeira fita de cDNA foi realizada a partir de 2000 ng de RNA incubados com 1 µL de iniciador OligoDt (10 µM), 1 µL de dNTPs (10 mM) e água num volume final de 13 µL à 65°C por 5 min. Após a incubação, foram acrescentados, 4 µL de tampão 5X, 1 µL de RNase OUT, 1 µL de DTT (100mM) e 1 µL da enzima RT

SuperScript® IV. A reação permaneceu por 10 min a 50°C e, em seguida, a 80°C por 10 min.

3.8 PCR TEMPO REAL

Com o intuito de identificar o número de cópias do cassete integrado no genoma da célula CHO bem como a expressão relativa de gene de interesse, empregamos a técnica de PCR tempo real (qPCR).

A qPCR foi realizada com Eva Green (Biotium), usando o sistema QuantStudio5 (Thermo Fisher Scientific) e 7500 (Thermo Fisher Scientific) com volume de reação de 20 µl composto por 6,3 µl de mix de PCR do Kit Nat (IBMP), 0,8 µl de cada iniciador F e R (Quadro 1), 1 µl do DNA ou cDNA e 1 µl de Eva green 20X. O ciclo foi iniciado com 50°C por 2 min e 95°C por 10 min, seguido de 40 ciclos de 15 s a 95°C, 1 min anelamento a 60°C. A análise dos dados foi realizada no software do próprio equipamento.

Para a quantificação do número de cópias do cassete integrado no genoma da célula CHO foi realizada uma diluição seriada com fator de 1:10 do DNA plasmidial (vetor *donor*) contendo o gene de interesse a partir de 10⁸ cópias. O número de cópias do plasmídeo (quantidade inicial) foi calculado utilizando a seguinte fórmula: Número cópias = [DNA (g/µL) x 6,022*10²³ (pb/mol)] / [tamanho (pb) x 660 (g/mol)] (PREDIGER, 2017). A diluição seriada foi submetida a qPCR com primer específico da Pró-trombina (Quadro 1) para a determinação da curva padrão. O número de cópias de uma amostra foi obtido mediante a interpolação dos valores do Ct (*Cycle Threshold*) obtidos na curva padrão com DNA plasmidial e dividido pelo número de genomas das amostras obtidos através do normalizar de cópia única, gene Glucagon.

Para a expressão relativa do gene alvo (Pró-trombina), utilizamos o gene *Eif3i* como gene de referência, gene de expressão constitutiva envolvido na tradução (BAHR et al, 2009) e a análise dos dados foi realizada pelo método Ct comparativo (Delta-Delta-Ct). Para os normalizadores também foi realizada a curva padrão com diluições seriadas 1:10.

Para avaliar a presença da região codificadora da transposase e expressão em nível de RNA nos clones e população de célula CHO expressando Pró-trombina foi realizada qPCR com primers específicos para a *SB* (Quadro 1).

Quadro 1. Sequência nucleotídica dos pares de primers utilizados na reação PCR em tempo real.

Glucagon	F	ATTGCCAAACGCCACGAT
	R	CCAAGCAATGAATTCCTTTGC
Trombina	F	GAGGACGCCTCGAGATAAGC
	G	GACTCCTCCATAGCTGGCAC
Eif3i	F	CCACAACCTCCACCAGGATT
	R	ATGCGGACGTAACCATCTTC
SB11	F	TTAAAGCTTGGTCGCAAATGGG
	R	TGTCCTTAAGCCATTTCTCACG

3.9 PCR SPLINKERETTE

Para identificar as regiões de inserção dos transposon foi realizada a técnica de PCR splinkerette (UREN et al., 2009). Inicialmente o DNA genômico de clones/populações de célula CHO (transfectadas com os vetores do Sistema Ingresso 1.0 ou com os vetores da nova transposase *Tp2* que estão descritos no tópico 3.11), foi isolado de acordo com o tópico 3.6. Para essa técnica é preciso digerir o DNA com uma enzima que clive em uma das pontas do cassete, que possua corte relativamente frequente no genoma e que gere extremidades coesivas.

Devido à dificuldade de encontrar uma enzima que reconhece 4 pb apta a ser usada, optamos por usar duas enzimas que reconhecem 6 pb (*BglII* e *BamHI*) e que possuem as mesmas extremidades coesivas. O uso de duas enzimas aumenta o número de clivagens e diminui o tamanho dos amplicons. A enzima *BglII* corta o cassete no final da primeira TIR, então a região que será analisada é a região flangeadora 5' (Fig 7). Dessa maneira, o DNA genômico foi digerido com as enzimas *BglII* e *BamHI* nas seguintes condições: 2 ug de DNA, 10 U de cada enzima, 1X do tampão 3.1 em um volume de 30 ul, os quais foram incubadas a 37°C por 16 horas. Nessa etapa são gerados fragmentos contendo a porção da TIR 5' e uma região flangeadora de tamanho variável dependendo da localização do próximo sítio de clivagem.

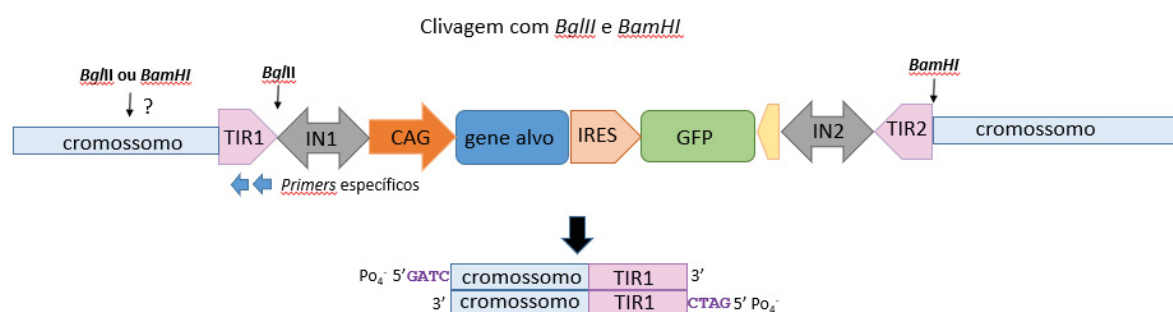


Figura 7. Esquema da etapa de clivagem da técnica de PCR splinkerette utilizada para identificar inserções gênicas realizadas pelo Sistema Ingresso. O DNA é clivado com *BglII* e *BamHI* gerando fragmentos coesivos, dentre eles os contendo a porção da TIR 5' e uma porção de tamanho variável da região de inserção. A esses fragmentos serão ligados adaptadores específicos da técnica e amplificação dos fragmentos contendo inserto se dá por PCR.

Em seguida, foi realizada a ligação dos adaptadores longo e curto específicos para essa técnica (Quadro 2) (UREN, 2009), os quais possuem complementariedade às regiões coesivas geradas na etapa de clivagem. Para isso, os adaptadores foram preparados previamente da seguinte maneira: os oligonucleotídeos liofilizados foram ressuspensos no tampão NEB 2 a uma concentração de 50 mM. Em um tubo, foram adicionados 50 µl de cada adaptador numa concentração final de 25 mM cada. Antes da reação de ligação, o mix de adaptadores foi desnaturado a 95°C por 5 min e resfriados a temperatura ambiente na taxa de 1°C a cada 15 segundos no termociclador. O DNA genômico digerido foi aquecido a 65°C por 20 min e colocado no gelo. A reação de ligação consistiu em 1 µl da enzima T4 ligase (Invitrogen), 1X de tampão, 2 µg de DNA genômico digerido, 1 µl do mix de adaptadores e água até o volume de 40 µl. A reação permaneceu a 4°C por 16 horas. Após a incubação, a ligase foi inativada a 65°C por 20 min. As amostras foram purificadas com o Kit QIAquick Gel extraction de acordo com o manual do fabricante, com exceção de que o DNA não foi extraído do gel de agarose.

Em seguida, duas etapas de reações de PCR são necessárias para a amplificação do inserto e região flanqueadora. Os primers estão no quadro 3. A primeira reação de PCR consistiu em 1 µl do DNA (10 ng) genômico com os adaptadores, 1 µl do primer Splink1 (0,1 µg/µl), 1 µl do primer R1 específico (0,1 µg/µl), 5 µl de mix de PCR do Kit Nat (IBMP) e água até o volume de 15 µl. As amostras foram incubadas a 95°C por 5 minutos e submetidas a 30 ciclos com as seguintes etapas: 95°C por 30 segundos, 59°C por 30 segundos e 72°C por 3 minutos, finalizando com uma etapa de extensão a 72°C por 10 minutos. Posteriormente, foi realizada a

segunda PCR com 1 µl do produto da primeira PCR e o mesmo protocolo descrito para a primeira PCR incluindo os primers específicos Splink2 e específico R2, dependendo do transposon e da metodologia de sequenciamento (quadro 2). Nós utilizamos dois métodos para obter o sequenciamento dos insertos, a partir da clonagem dos insertos, seguida do sequenciamento pela metodologia de Sanger e por sequenciamento direto dos insertos pela plataforma Illumina. O resultado da PCR splinkerette foi visualizado em gel de agarose 2% e corado com brometo de etídio.

Para o sequenciamento a partir de clonagem, os produtos de PCR foram clonados no vetor TOPO TA (Thermo Fisher Scientific) de acordo com as recomendações do fabricante. Os vetores contendo os produtos de PCR foram propagados em célula Top10 (tópico 3.12) e plaqueados em placas LB contendo ampicilina (100 µg/mL) e IPTG/X-gal (1 mmol/80 µg/mL). Cerca de 20 colônias foram selecionadas para extração de plasmídeo (como tópico 3.14) e enviados para sequenciamento pelo método de Sanger na empresa MacroGen. Por essa metodologia, foram analisados os clones 6 e 7 e a população 1.

A análise das sequências obtidas por clonagem foi feita pelo Staden Package (STADEN et al, 2000). A ferramenta Pregap permitiu a remoção da sequência do vetor e verificação da qualidade das bases nucleotídicas. A ferramenta Gap4 foi utilizada para alinhar e sobrepor as sequências originadas em ambos os sentidos e montar uma sequência consenso. A identificação da região TIR 5' nas sequências consenso foi obtida pelo alinhamento com a sequência esperada usando a opção blastn-2sequences do NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). Blastn online contra o genoma de CHO foi utilizado para identificar a região de inserção. O genoma que aqui chamamos de CHO, usado para o mapeamento das inserções, é o genoma de fêmea do hamster chinês *Cricetulus griseus* que foi submetido ao NCBI pela “CHO Genome Community” (acesso: GCA_003668045.3) em 06 de junho de 2020. Essa versão do genoma (CriGri-PICRH-1.0) tem uma cobertura de 31X e consiste na melhor montagem atualmente (HILLIARD, MACDONALD e LEE, 2020).

A ferramenta Graphics do NCBI foi usada para visualização gráfica das sequências e entender o contexto genômico das inserções (genes e sequências repetitivas nas proximidades e evidência de expressão de acordo com o mapeamento de leituras de dados de RNA-seq disponíveis para CHO).

As amostras sequenciadas pela plataforma Illumina foram purificadas, após a 2ª PCR, com o kit QIAquick PCR Purification e foram enviadas para o sequenciamento

pelo protocolo Nextera *paired-end* de 250 pb pela empresa GOGenetics. Foi realizado um primeiro sequenciamento dos clones 6, 7 e população 1 a partir da PCR splinkerette com o primer Spl_SB_R2_illumi e um segundo sequenciamento dos clones 6, 7, 28 e população 1 a partir da PCR splinkerette com o primer Spl_ilu_SB_R2_2.

Pela lógica da técnica de PCR splinkerette, na extremidade inicial dos amplicons está sempre presente o adaptador Splink2 e na extremidade final está a TIR5' (Fig 8A). Aos primers foram adicionadas as sequências para os adaptadores Nextera P5 e P7, sendo assim, a metodologia *paired-end* deve gerar, teoricamente, leituras polarizadas, sendo as leituras R1 correspondendo a sequência 5' do amplicon (similaridade com o adaptador Splink2 e CHO) e as leituras R2 correspondendo a região 3' do amplicon (similaridade com a fita complementar reversa da TIR 5' do transposon e contendo o sítio de inserção em CHO).

A análise das sequências geradas por Illumina consistiu em múltiplas etapas (Fig 8B): 1) A primeira etapa foi a limpeza das leituras pela ferramenta Trimmomatic (BOLGER, LOHSE e USADEL, 2014) na plataforma Galaxy (AFGAN et al., 2018), eliminando os adaptadores Nextera e regiões de baixa qualidade ($q < 20$) e também eliminando leituras menores que um determinado tamanho (menor que 100 pb no 1º sequenciamento e menor 220 pb no 2º sequenciamento). 2) A próxima etapa foi separar as leituras pareadas possivelmente corretas das leituras possivelmente espúrias. A partir da premissa apresentada na figura 8A, as leituras R1 foram alinhadas contra o genoma da CHO (GCA_003668045.3) e as leituras R2 contra a sequência da TIR5' do *SB* ou *Tp2* por blastn local (parâmetros: -outfmt 6 -evalue 0.0001 -max_target_seqs 1 -max_hsps 1). Apenas as leituras R1 e R2 positivas para os respectivos alinhamentos foram usadas e selecionadas (com código Python). 3) No final das leituras R2, teoricamente, está o sítio exato da inserção logo após a porção da TIR. Dessa forma, apenas essa região de cada leitura R2 foi recuperada usando um código em Python com as seguintes premissas: (i) a sequência precisa ter um tamanho mínimo (247 pb para *SB* no 1º sequenciamento e 220 pb para *SB* no 2º sequenciamento e para *Tp2*) para recuperar um tamanho suficiente de sequência a ser alinhada com a respectiva região de CHO; (ii) a sequência precisa ter uma etiqueta correspondendo aos primeiros 10 pb da TIR 5' (complementar reversa) do transposon em questão e precisa ser seguido da TSD (TA). Leituras que possuem a etiqueta, mas sem a TSD também foram recuperadas. 4) Para restringir o alinhamento às regiões

onde sabidamente estão essas inserções, as coordenadas de alinhamento das leituras R1 em CHO foram usadas para recuperação dessas regiões a partir da posição inicial do alinhamento de R1 até 7 kb após o final do alinhamento com um script em Perl. Em princípio, se o sítio de restrição *BamHI* ou *BglII* estiver localizado a mais de 7 kb do sítio de inserção, dificilmente será amplificado na PCR splinkerette.

5) O próximo passo é alinhar essas regiões isoladas no genoma da CHO para identificação precisa do sítio de inserção por blastn local (parâmetros: -outfmt 6 -max_target_seqs 1 -max_hsps 1). Para o primeiro sequenciamento do sistema *SB*, como a região recuperada de CHO nas leituras R2 é pequena (9-14 pb) foi necessário adicionar o parâmetro -task blastn-short para gerar resultados significativos nesse blast. Todos os scripts usados estão nos Anexos da tese.

Após a obtenção da tabela de inserções, foi realizada uma filtragem manual para eliminação de inserções que possam ser consideradas artefatos. Esses artefatos aconteceram principalmente pelo alinhamento das leituras em mais de uma posição no genoma com igual *score* de alinhamento. Isso aconteceu porque o resultado do blast foi limitado a um alinhamento (-max_target_seqs 1 -max_hsps 1) e ao acaso regiões distintas de mesmo *score* foram tabuladas para leituras de uma mesma inserção. Dessa forma, a filtragem foi realizada para eliminar possíveis inserções que produziram mais de um alinhamento de mesmo *score* no blastn do R1 contra CHO. Futuramente, os scripts podem ser atualizados para contemplar uma solução para esse problema.

Os sítios de inserção foram analisados individualmente pela ferramenta Genome Data Viewer do NCBI que permite a mesma visualização que a ferramenta Graphics já mencionada.

Para buscar possíveis inserções geradas de forma aleatória no genoma da CHO, ou seja, inserções independentes de transposição, as leituras obtidas foram mapeadas por blastn local (parâmetros: outfmt 6 -evalue 0.0001 -max_target_seqs 1 -max_hsps 1) contra a sequência esqueleto do vetor, sem o cassete de expressão. Leituras positivas foram isoladas e então alinhadas contra o genoma da CHO nos mesmos parâmetros do blastn

Quadro 2. Sequência nucleotídica dos pares de primers utilizados na reação PCR Splinkerette.

Adaptadores	Longo	CGAAGAGTAACCGTTGCTAGGAGAGACCGTGG CTGAATGAGACTGGTGTGCGACACTAGTGG
	Curto	GATCCCACTAGTGTGCGACACCAGTCTCTAATTT TTTTTTCAAAAAAA
1ª PCR	Splink1	CGAAGAGTAACCGTTGCTAGGAGAGACC
	SB_R1	CTTCTAAAGCCATGACATC
	Tp2_R1	TCGAGGTAAGCCAATGTTGC
2ª PCR Clonagem	Splink2	GTGGCTGAATGAGACTGGTGTGCGAC
	SB_R2	GGCACAGTCAACTTAGTG
2ª PCR Seq. Illumina	Splink2_illumi	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GGTGGCTGAATGAGACTGGTGTGCGAC
	Spl_SB_R2_illumi (1º sequenciamento)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGGGCACAGTCAACTTAGTG
	Spl_ilu_SB_R2_2 (2º sequenciamento)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGTAGATGTCCTAACTGACTTGCC
	Spl_ilu_TP2_R2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGTATTATATCGAACAAATCTGAATGGC

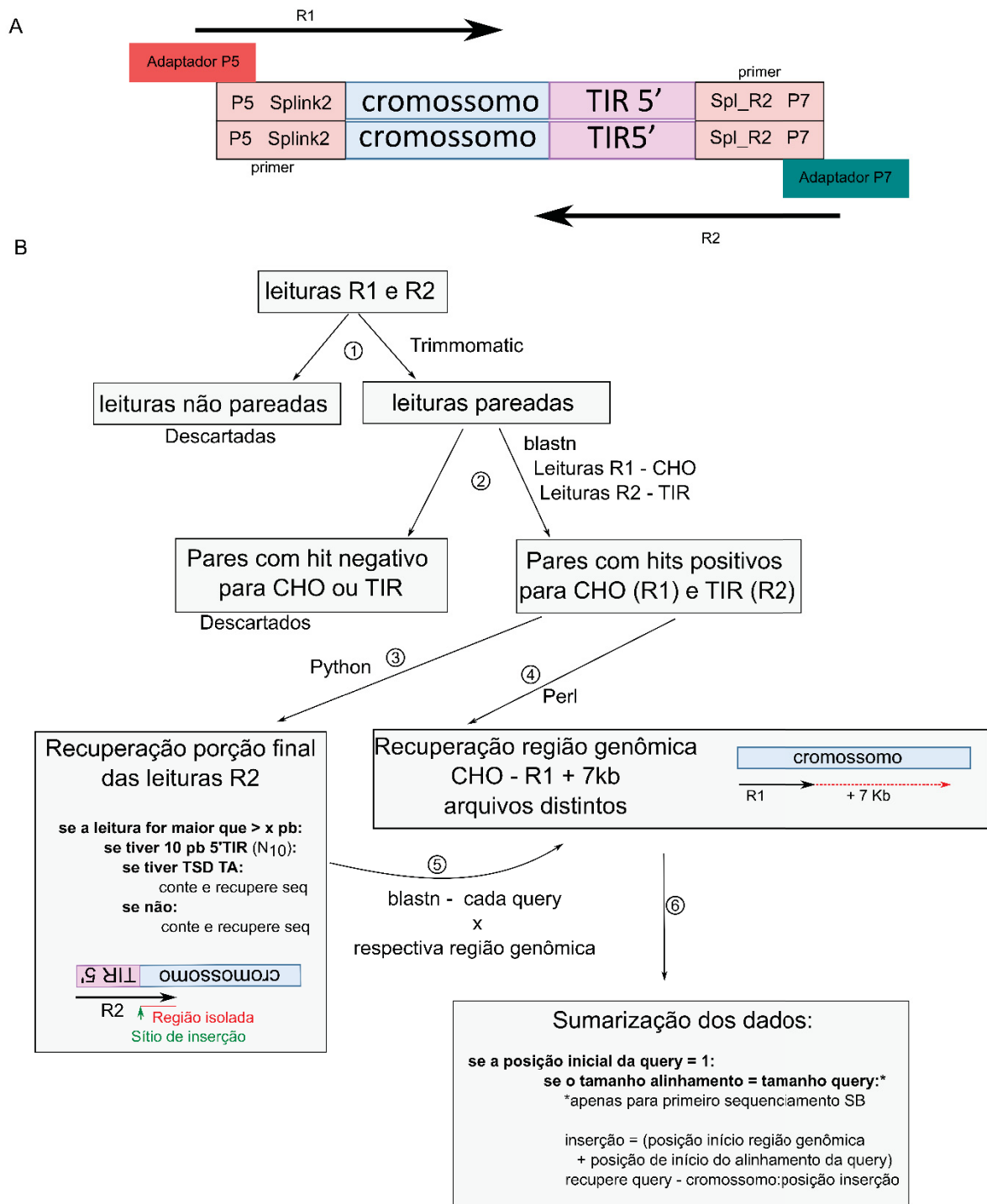


Figura 8: Desenho esquemático da análise das sequências geradas por Illumina da PCR splinkerette para investigar os locais de inserção genômica dos cassetes. As etapas estão detalhadas no texto.

3.10 AVALIAÇÃO DA PERDA DO VETOR *HELPER*

Para avaliar a perda do vetor *helper* foi realizada PCR convencional com os primers específicos internos (SB R e F, quadro 3), com primer que anela no vetor e na região da transposase (SB_F e polyA_R, quadro 3) e com primers desenhados para amplificar a transposase SB completa (SBpet28_F e SBpet28_R, quadro 3). Ainda, foi realizada qPCR com os primers específicos internos para o gene da transposase (SB R e F, quadro 3) a partir do DNA e cDNA dos clones/populações de célula CHO seguindo a metodologia descrita no tópico 3.8.

3.11 DESENVOLVIMENTO DE NOVOS VETORES

Como já mencionado, três transposons foram inicialmente escolhidos para serem testados no Sistema Ingresso no lugar do SB. As sequências codificadoras das transposases e das TIRs foram sintetizadas pela empresa GenScript incluindo os sítios de restrição apropriados para as clonagens. As sequências sintéticas, as quais foram inseridas no vetor pUC57, foram eluídas do papel filtro com água livre de nuclease e transformados em células cálcio competente (item 3.12) para a sua propagação.

As TIRs 5' e 3' de cada transposon foram amplificadas do vetor pUC57 com primers específicos (TIR2_5' F e R; TIR5_5' F e R; TIR7_5' F e R; TIR2_3' F e R; TIR5_3' F e R; TIR7_3' F e R) (quadro 3). Cada reação de PCR consistiu em 5 µl de mix de PCR do Kit Nat (IBMP), contendo todos os reagentes básicos comuns, 0,4 µl de cada iniciador (5 µM) F e R, 1 µl de *template* e água até o volume total de 15 µl. As amostras foram incubadas a 95°C por 5 minutos e submetidas a 30 ciclos de PCR com as seguintes etapas: 95°C por 30 segundos, 59°C por 30 segundos e 72°C por 1 minuto, finalizando com uma etapa de extensão a 72°C por 10 minutos. Os produtos amplificados foram analisados em gel de agarose 2%. As amostras amplificadas foram digeridas com as respectivas enzimas (TIR 5': *PvuII* e *BglII*; TIR 3': *EcoRV* e *BamHI*) em tampão 3.1 nas seguintes condições 3 µg de DNA, 10 U de cada enzima, 1X do tampão apropriado em um volume de 30 µl os quais foram incubadas a 37 °C por 16 horas. As reações de digestão foram purificadas com o Kit QIAquick PCR purification (Qiagen) para as TIRs 5'. O vetor *donor* foi digerido com as mesmas enzimas para a excisão das TIRs originais, nas condições citadas acima. A purificação do vetor *donor*

foi realizada por excisão do gel de agarose utilizando o kit QIAquick gel extraction (Qiagen).

Os genes das novas transposases foram digeridos do vetor pUC57 com a enzima *SacII*, bem como o gene da transposase SB do vetor *helper*. Nesta etapa, 3 µg de vetor contendo o gene das transposases foram digeridos com 10 U de *SacII*, tampão CutSmart (NEB) e água em volume final de 30 µl. A reação de clivagem foi incubada a 37°C por 16 horas.

A purificação dos genes das transposases e dos vetores *helper* sem a SB foi realizada por excisão do gel de agarose utilizando o kit QIAquick gel extraction (Qiagen). Após a purificação, o vetor *helper* sem o SB foi tratado com 1 µl de fosfatase alcalina (ThermoFischer Scientific - FastAP Thermosensitive Alkaline Phosphatase), 4 µl de tampão 10X e 5 µl de água, a mistura foi incubada a 37°C por 30 minutos. A reação foi interrompida por 15 min a 80°C.

Cada reação de ligação, do gene das transposases no vetor *helper* digerido e das TIRs no vetor *donor* digerido, consistiu em 1 µl da enzima T4 ligase (Invitrogen), 1X de tampão, 5 ng da PCR purificada e digerida da TIR ou o gene da transposase digerido, 50 ng de vetor e água até o volume de 20 µl. As reações permaneceram a 16°C por 16 horas. Após a incubação, 5 µl da reação de ligação foi transformada (item 3.12) em célula Top10. A seleção dos clones contendo plasmídeo com o inserto desejado foi realizada através da técnica de PCR de colônia (item 3.13). Com os clones positivos foi realizada a mini preparação de plasmídeo (item 3.14) e o material foi submetido ao sequenciamento de DNA, usando o serviço de sequenciamento da empresa Macrogen (Coréia do Sul).

Para o controle negativo das transfecções, utilizamos um vetor *helper* sem a região codificadora da transposase. Para isso, digerimos o vetor *helper* contendo a região do SB com a enzima *SacII* nas seguintes condições: 3 µg de DNA, 10 U da enzima, 1X do tampão CutSmart (NEB) em um volume de 30 µl, sendo que as reações foram incubadas a 37°C por 16 horas. A purificação do vetor *helper* aberto sem a transposase foi realizada por excisão do gel de agarose utilizando o kit QIAquick gel extraction (Qiagen). Após a purificação, o vetor foi religado como descrito acima e 5 µl da reação de ligação foi transformada (item 3.12) em célula Top10. A seleção dos clones foi realizada através da técnica de PCR de colônia (item 3.13) com os primers IntronSV e polyA (Fig 4 e quadro 3).

3.12 TRANSFORMAÇÃO EM CÉLULA CÁLCIO COMPETENTE

A transformação em células cálcio-competente foi realizada na linhagem de Top10 cálcio-competentes previamente preparadas. O plasmídeo foi adicionado a 100 µl de células. Estas células permaneceram no gelo por 30 minutos e, em seguida, foram incubadas a 42°C por 2 minutos, sendo colocadas novamente no gelo. Após a transformação, as células cresceram em 1 ml de meio LB por uma hora a 37°C sob agitação de 220 rpm. Em seguida, a cultura foi plaqueada em meio LB solidificado contendo ampicilina (100 µg/ml), onde permaneceram por 16 horas a 37°C.

3.13 TÉCNICA DE PCR DE COLÔNIA PARA CONFIRMAÇÃO DAS CLONAGENS

Após a transformação, as colônias que cresceram no meio LB seletivo foram transferidas para tubos de PCR com auxílio de um palito estéril, contendo 5 µl de mix de PCR do Kit Nat (IBMP), 0,4 µl de cada iniciador (5 µM) F e R (Quadro 3) e água até o volume total de 15 µl. As amostras foram incubadas a 95°C por 5 minutos e submetidas a 30 ciclos de PCR com as seguintes etapas: 95°C por 30 segundos, 59°C por 30 segundos e 72°C por 1 minuto, finalizando com uma etapa de extensão a 72°C por 10 minutos. Os produtos amplificados foram analisados em gel de agarose 2%. Posteriormente, o gel foi corado com brometo de etídio (0,5 µg/ml) por aproximadamente 10 minutos, lavado com água e analisado sob luz ultravioleta. O gel foi fotografado no sistema de foto-documentação L-Pix (Locus Biotecnologia, Brasil).

Quadro 3. Sequências nucleotídicas dos pares de primers utilizados nas reações de PCR.

TIR2_5'	F	GGCAGCTGAAGAAATCATTTTTTAT
	R	CCCAGATCTTAATCGAGGTAAGC
TIR5_5'	F	CAGCTGCAGAGACCTGC
	R	CAGATCTCACTCGAACATATTGATT
TIR7_5'	F	CAGCTGCCCTTTCCCGTC
	R	CAGATCTAACACGATTGTGAAATA
TIR2_3'	F	GATATCAAATTTAATTGAAAGTATACCAAG
	R	GCGGATCCAAGAAATCATTTTT
TIR5_3'	F	GATATCGTTAGCTTCGGGTG
	R	GCGGATCCCAGAGAGC
TIR7_3'	F	GATATCTATGCAAGTGTAAATGCTTG
	R	GCGGATCCCCCTTTCC

pucR1	F	CGAGGAAGCGGAAGAGC
pucF1	R	CTCATGAGCGGATACATATTTG
IntronSV	F	GATGTTGCCTTTACTTCTAGGC
polyA	R	AAGCAATAGCATCACAAATTTTAC
SB11	F	TTAAAGCTTGGTCGCAAATGGG
	R	TGTCCTTAAGCCATTTTCTCACG
Tp2	F	AGGGTGTGGAGAAAGCATGG
	R	ATGCCGTCGATCTTAGCCAG
Tp5	F	AGATAACGAGCAGAGCCAGC
	R	CTGCAGCTTGTTGAACACGG
Tp7	F	ATCGTGAAGAGCCGGCATAG
	R	TGGTGTAGCCATTGCTGGAG
SBpet28	F	GGGCCATGGGGCATCATCATCATCATATGGGAAAATCAAAGAAAT
	R	TCACTAGTATTTGGTAGCATTGCCTTTAAATTGTTAACTTGGG

3.14 PREPARAÇÃO DE PLASMÍDEO EM PEQUENA ESCALA (MINIPREP)

Os clones recombinantes foram cultivados em 5 ml de meio LB, contendo 100 µg/mL de ampicilina, a 37°C sob agitação de 220 rpm durante 18 h. As culturas foram então centrifugadas a 14.000 rpm por 1 minuto a temperatura ambiente e os plasmídeos recombinantes purificados com o sistema de mini preparação de plasmídeo QIAprep Spin Miniprep (QIAGEN) de acordo com as recomendações do fabricante. Após a purificação dos plasmídeos, fez-se a mensuração da concentração através do uso do espectrofotômetro NanoDrop (Thermo Fisher Scientific).

3.15 TRANSFEÇÃO DE PLASMÍDEO EM CÉLULA CHO POR POLIFECÇÃO.

As células EXPICHO (Thermo Fisher Scientific) foram mantidas em meio Dynamis™ (Thermo Fisher Scientific) a 37°C, 8% de CO₂ e 120 rpm de agitação. Para o experimento de transfecção, foram recuperados 1×10^7 células para cada teste. As células foram centrifugadas a 1700 rpm por 10 minutos a 20°C e, em seguida, ressuspendidas em 500 µl de meio Opti-Mem pH 7,1. Nos tubos contendo as células, foram adicionados 3,25 µg de DNA dos vetores *helper Tp2* ou *SB* (controle positivo) ou controle negativo (vetor sem transposase), 9 µg de DNA dos vetores *donor* (sem gene de interesse e contendo GFP como gene marcador) e 25 µl de polietilenoimina (PEI) 1 mg/ml. As amostras foram incubadas por 2 horas a 37°C, 8% de CO₂ e 120 rpm de agitação. Após o período de incubação, as amostras foram transferidas para

um recipiente contendo 20 ml de meio Dynamis e mantidas nas condições padrão já descritas no tópico 3.4.

3.16 SELEÇÃO DOS TRANSFECTANTES

Para a seleção das células transfectadas expressando o gene GFP, essas células foram centrifugadas a 1700 rpm por 10 minutos a 20°C e ressuspensas em PBS 1X contendo BSA 2% e antibióticos (50 U/ml de penicilina, 50 µg/ml de streptamicina e 1 µg/ml de ciprofloxacina). Em seguida, a fluorescência foi analisada por citometria de fluxo e essas células foram separadas por *cell sorting* utilizando o equipamento S3e Cell Sorter (BioRad). As células foram recuperadas em PBS 1X com BSA 2% e antibióticos listados acima. Após a separação, as células foram mantidas em meio Dynamis com meio condicionado e antibióticos até o próximo repique nas condições de cultivo padrão. As populações passaram por 2 a 5 eventos de enriquecimento.

3.17 EXTRAÇÃO DE DNA EXTRACROMOSSOMAL DE CÉLULA CHO TRANSFECTADA

Com as células transfectadas foi realizada a extração de DNA extracromossomal para isolamento do DNA plasmidial, segundo ARAD (1998). Um total de 1×10^6 de células foram recuperadas da cultura por centrifugação a 300 rpm por 10 minutos a 4°C. Em seguida, as células foram lavadas com PBS 1X e centrifugadas novamente. As células foram ressuspensas em 250 µl de tampão 1 (50 mM Tris pH 7,5, 10 mM EDTA, 50 µg/ml RNase A). Posteriormente, 250 µl do tampão 2 (1,2% SDS) foram adicionados e as células foram incubados a temperatura ambiente por 5 minutos. Após a incubação foram acrescentados 350 µl de tampão 3 (3 M CsCl, 1 M acetato de potássio, 0,67 M ácido acético) e incubados por 10 minutos. Em seguida, as células foram centrifugadas a 14,000 rpm por 10 minutos. Os tubos contendo as células foram invertidos algumas vezes e centrifugados novamente. As amostras foram inseridas em colunas de sílica e centrifugadas a 10.000 rpm por 1 minuto. As colunas foram lavadas com 750 µl de tampão de lavagem (60% etanol, 10 mM Tris pH 7,5, 50 µM EDTA, 80 mM acetato de potássio) e centrifugadas a 10,000 rpm por 1 minuto. As amostras foram centrifugadas mais uma vez para secar a coluna.

As amostras foram eluídas em 75 µl de tampão de eluição (Qiagen EB buffer e 0,5 M EDTA). A mensuração da concentração foi realizada através do uso do espectrofotômetro NanoDrop (Thermo Fisher Scientific).

3.18 WESTERN-BLOT PARA DETECÇÃO DA PROTEÍNA TRANSPOSASE DO SB

As culturas de células CHO 48 horas após a transfecção foram submetidas a ensaio de *Western blot* para verificar a expressão da proteína SB. Inicialmente, as amostras foram submetidas a técnica de SDS-PAGE utilizando 30 mA de corrente elétrica por gel. Em seguida, as proteínas foram transferidas do gel para uma membrana de nitrocelulose em tampão de transferência por 40 minutos a 50 - 60 V. Para a visualização das bandas, a membrana foi corada em solução de Ponceau S e o marcador de peso molecular foi marcado. A membrana foi lavada em água para remoção do Ponceau S e colocada em solução de bloqueio (PBS-Tween 0,05%, leite 5%) por 1 hora.

Após o bloqueio, a membrana foi lavada três vezes com PBS-Tween 0,05% por 5 min. A membrana foi incubada com o anticorpo primário (*Sleeping Beauty* anticorpo monoclonal, RD System) na concentração 1 µg/mL por 1 hora. A membrana foi lavada nas mesmas condições citada acima e a membrana foi incubada com anticorpo secundário conjugado à peroxidase (diluído 1:20.000). Após 40-60 minutos, a membrana foi novamente lavada. Para a revelação da membrana foi utilizado o kit SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo Fisher Scientific) e a visualização foi realizada no fotodocumentador L-PIX CHEMI (Loccus).

3.19 AVALIAÇÃO DA PERDA DOS VETORES EPISSOMAIS

Após a seleção das culturas transfectantes, expressando o gene GFP, separadas por pelo menos duas rodadas de *cell sorting* foi realizada o isolamento dos plasmídeos como item 3.17. Os plasmídeos isolados foram transformados em bactérias Top10 como item 3.12 e avaliados quanto a presença de vetores episomais.

4 RESULTADOS E DISCUSSÃO

4.1 CARACTERIZAÇÃO DO SISTEMA INGRESSO

Com o intuito de caracterizar o Sistema Ingresso 1.0 para entender melhor o desempenho do sistema como um todo, nós selecionamos clones isolados da célula CHO expressando Pró-Trombina como proteína de interesse e GFP como marcador de seleção para investigar o local de integração genômica por PCR splinkerette, número de cópias do cassete inserido e expressão relativa do mRNA de Pró-trombina por PCR em tempo real.

A seleção dos clones foi baseada no nível de expressão de Pró-Trombina, sendo os clones 4, 5, 15 e 28 de maior produção ($> 0,35$ g/L), os clones 3 e 92 ($< 0,1$ g/L) os menos produtores, e os clones 6 e 7 de produção intermediária ($> 0,1$ g/L). Ainda, os testes foram realizados com duas amostras provenientes de uma população de células a fim de comparação com os clones individualizados.

4.1.1 Local de integração genômica

Uma questão importante dos sistemas baseados em transposição é o local de inserção do transgene, pois a integração pode causar eventos adversos, como a ativação de proto-oncogenes ou a inativação de genes celulares essenciais. Uma das técnicas mais utilizadas atualmente, e aparentemente mais eficiente, para mapear os locais de inserção do transposon é a PCR splinkerette, como detalhado na figura 9 (UREN et al, 2009).

A PCR splinkerette foi padronizada com as amostras dos clones 3, 5, 6, 7 e população 1, como visualizada por eletroforese (Fig 10). Após a segunda PCR, os produtos devem aparecer como uma escada no gel de agarose.

Os produtos de PCR foram clonados em sistema TOPO e propagados em célula competente. Cerca de 20 clones bacterianos, 4 de cada amostra, foram enviados para sequenciamento e analisados quanto ao local de inserção.

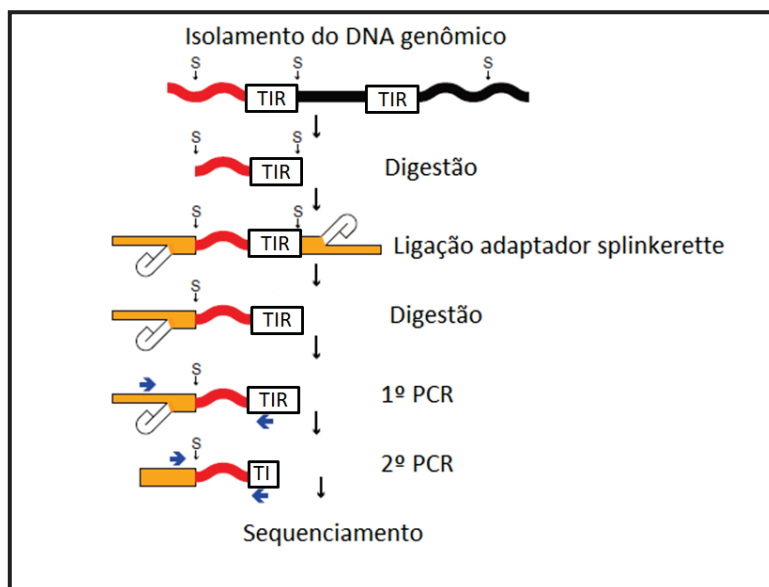


Figura 9. Esquema do protocolo da PCR Splinkerette. Modificado de Uren 2009. Nesta técnica, o DNA genômico é digerido com uma enzima de restrição (S) conhecida que clive a extremidade do cassete (região da TIR) e que produza extremidades coesivas. Nestas extremidades, é ligado um oligonucleotídeo (adaptador) que contém uma extremidade coesiva compatível, contém uma estrutura *hairpin* e não é fosforilado. Duas reações de PCR com primers específicos (flecha azul) são então realizadas para amplificar a sequência genômica entre a inserção do transposon (região da TIR) e a ligação do adaptador splinkerette. Os produtos de PCR podem ser sequenciados para a identificação do local de inserção no genoma.

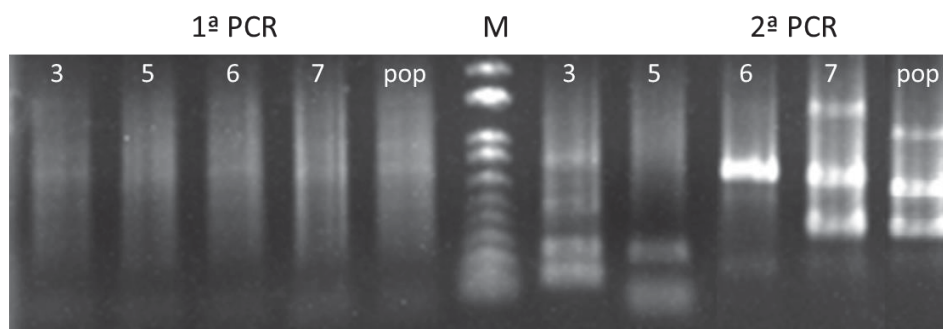


Figura 10. Eletroforese do resultado da PCR splinkerette para os clones 3, 5, 6, 7 e população 1. Resultados obtidos no gel de agarose da 1ª e 2ª PCR splinkerette. Amostras: clone 3, clone 5, clone 6, clone 7 e população (Pop); M: marcador de peso molecular 1Kb plus (Thermo Fisher Scientific).

Para o clone 3 e 5 de CHO, nenhum clone bacteriano mostrou sequência com qualidade adequada para a análise, por outro lado, foi possível identificar sítios de inserção do cassete para as amostras 6, 7 e o grupo população 1. O padrão esperado de sequência e obtido para essas amostras foi conforme o exemplificado na figura 11, onde foi possível identificar a sequência correspondente ao genoma da célula CHO, um TA (espera-se que na outra extremidade do cassete esteja presente outro TA,

formando assim as TSDs) e, em seguida, a porção da TIR 5' da SB no final da sequência.

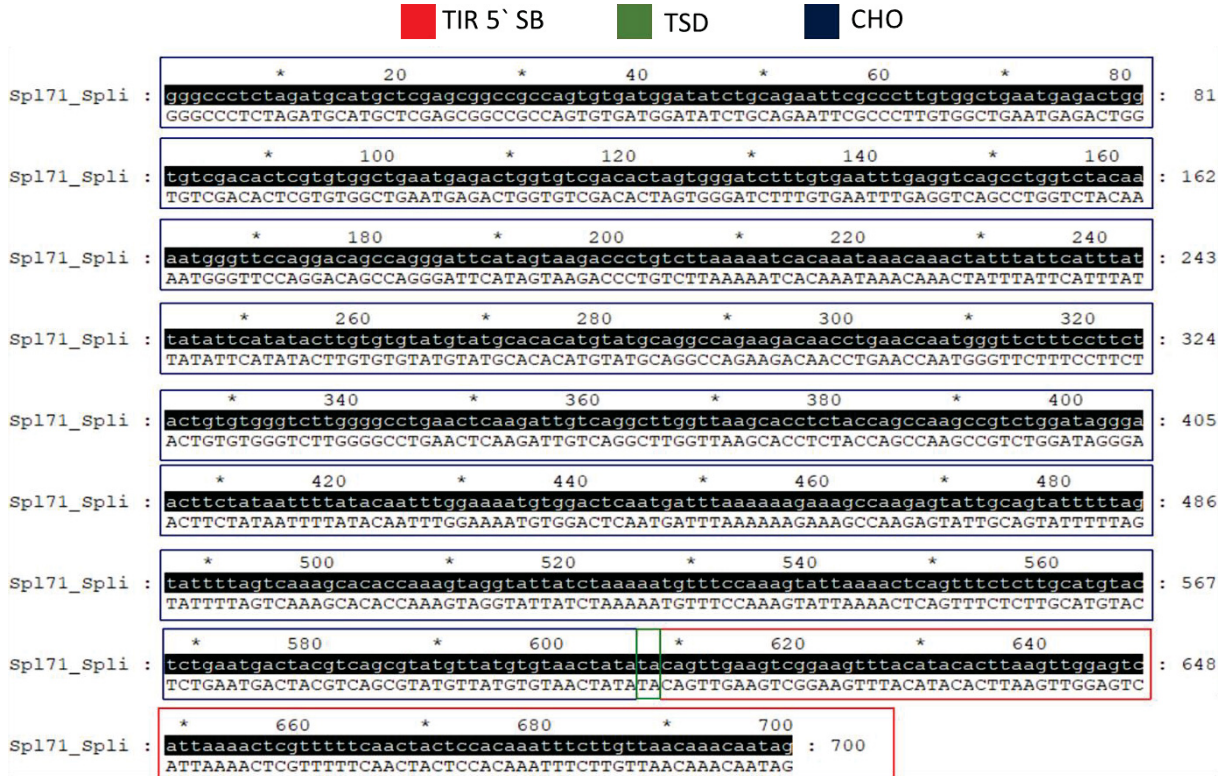


Figura 11. Análise do sequenciamento da PCR splinkerette para a amostra 7 (clone bacteriano 71). Caixas em vermelho indicam a sequência da TIR 5' do SB, em verde a sequência TA da TSD e em azul a região do genoma da célula CHO onde o cassete foi inserido (nesse a inserção aconteceu na posição 208112109 do *scaffold* NC_048596.1)

Através da ferramenta de visualização gráfica do NCBI foi possível observar o contexto genômico das inserções (tabela 2). Os clones 61 e 62 da amostra 6 representam a mesma inserção na região 5' UTR (780 pb *upstream*) do gene *Fkbp9* (peptidyl-prolyl cis-trans isomerase) (Fig 12). As sequências geradas para a amostra 7 indicaram inserção do cassete em 3 regiões diferentes da célula CHO. Os clones 71 e 74 foram inseridos em um íntron do gene *Epc1* (enhancer of polycomb homolog 1) (Fig 13), a inserção do clone 72 ocorreu *upstream* ao gene *Areg* (amphiregulin) (Fig 14) e o clone 73 foi inserido em um elemento SINE (Fig 15).

Três das quatro inserções ocorreram em regiões de atividade transcricional, como pode ser observado pela cobertura de leituras de RNA-seq disponibilizada no NCBI (Annotation Release 104) e mostradas nos gráficos (Figs 12, 13 e 14). Apenas uma inserção, que ocorreu dentro de um retrotransposon da ordem SINE, não mostrou

dados de mapeamento de RNA-seq. Regiões contendo TEs podem caracterizar locais de possível silenciamento, embora seja possível observar vários fragmentos de TEs em todas as regiões analisadas. É provável que todas essas inserções não tenham afetado de forma significativa o funcionamento da célula. Inserções em regiões não codificadoras são menos deletérias, embora, quando acontece nas proximidades de genes poderia levar a supressão ou desregulação da expressão desses genes e inserções em íntrons poderiam levar a *splicing* alternativo e exonização do cassette formando uma quimera proteica (DUBIN, SCHEID e BECKER, 2018).

Tabela 2. Região de inserção e contexto genômico na célula CHO das amostras inseridas pelo Sistema Ingresso 1.0.

Amostra	Região de inserção em CHO	Contexto genômico	Região ativa?
Clones bac 61 e 62 (clone CHO 6)	NC_048601.1:84464972	5' UTR (780 pb <i>upstream</i>) do gene <i>Fkbp9</i>	Sim
Clone bac 71, 74 (clone CHO 7)	NC_048596.1:208112109	intron do gene <i>Epc</i>	Sim
Clone bac 72 (clone CHO 7), Clones bac 1, 2 e 3 (população CHO)	NW_023276807.1:155906941	Região <i>upstream</i> do gene <i>Areg</i>	Sim
Clone bac 73 (clone CHO 7)	NC_048601.1: 43620480	Elemento SINE	Não

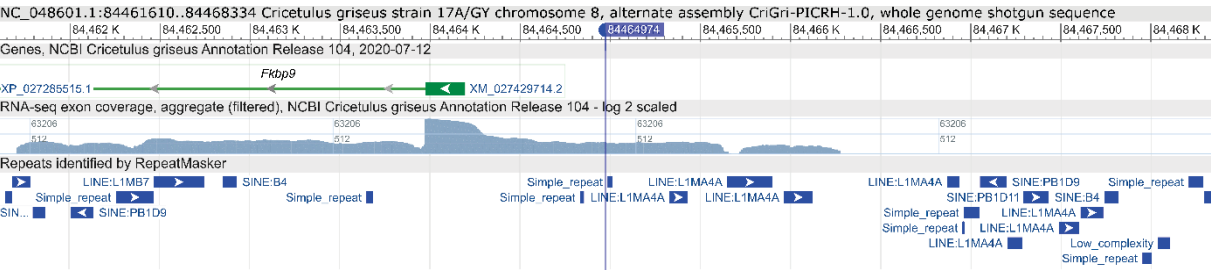


Figura 12. Contexto genômico da inserção do cassette na célula CHO da amostra 6, clones bacterianos 61 e 62. Dados da presença de genes, cobertura de leituras de RNA-seq e mapeamento de sequências repetitivas são mostrados. O cassette foi inserido na região 5' UTR (780 pb *upstream*) do gene *Fkbp9* (peptidyl-prolyl cis-trans isomerase).

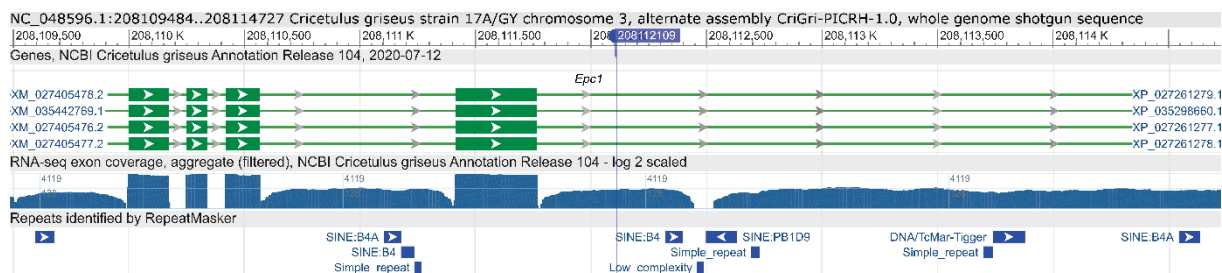


Figura 13. Contexto genômico da inserção do cassete na célula CHO da amostra 7, clones bacterianos 71 e 74. Dados da presença de genes, cobertura de leituras de RNA-seq e mapeamento de sequências repetitivas são mostrados. O cassete foi inserido em um íntron do gene *Epc1* (enhancer of polycomb homolog 1).

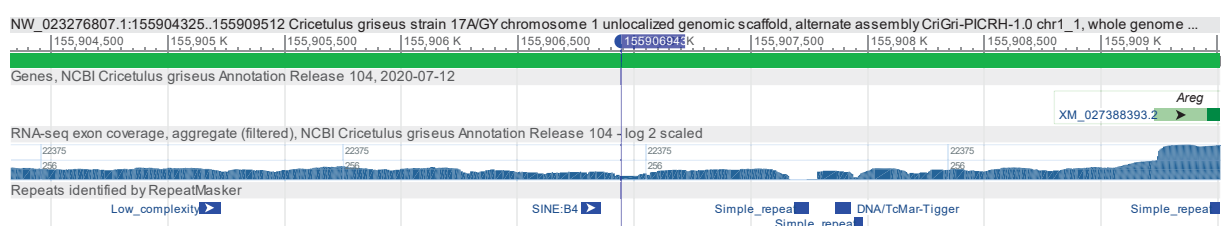


Figura 14. Contexto genômico da inserção do cassete na célula CHO da amostra 7, clone bacteriano 72. Dados da presença de genes, cobertura de leituras de RNA-seq e mapeamento de sequências repetitivas são mostrados. Inserção do cassete ocorreu região *upstream* ao gene *Areg* (amphiregulin).

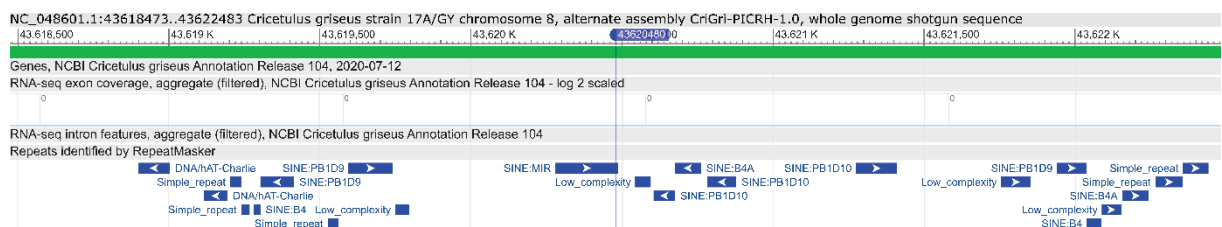


Figura 15. Contexto genômico da inserção do cassete na célula CHO da amostra 7, clone bacteriano 73. Dados da presença de genes, cobertura de leituras de RNA-seq e mapeamento de sequências repetitivas são mostrados. A inserção ocorreu em um elemento SINE.

Os mesmos clones foram submetidos a 2ª PCR do método splinkerette com primers específicos para o sequenciamento na plataforma Illumina. A análise dos resultados mostrou que essa técnica é muito útil e prática para a caracterização das inserções. O número de leituras foi satisfatório (mínimo de 21165 leituras pareadas para a amostra populacional). Após a limpeza das sequências para retirada de regiões de baixa qualidade e adaptadores do sequenciamento, foi aplicada a metodologia de bioinformática descrita no item 3.9 (materiais e métodos) para a obtenção da informação dos sítios de inserção. Foram consideradas para as análises apenas os pares de leituras cuja R1 teve identidade com CHO e R2 com a TIR 5' da SB. O primer desenhado para a 2ª PCR da SB não foi o mais adequado para o sequenciamento

Illumina, já que poucos nucleotídeos no final da leitura R2 corresponderiam a região genômica de CHO. Dessa forma, um novo primer foi desenhado em uma região mais central da TIR 5' permitindo uma leitura maior da região de inserção da CHO. Os clones 6, 7 e população foram re-sequenciados com esse novo primer bem como o clone 28, maior produtor de proteína.

Os dados das filtrações para as duas levas de sequenciamento estão mostrados na tabela 3. As leituras que não foram selecionadas pela análise no blastn provavelmente são leituras espúrias e não foram consideradas na análise de inserção. A maioria das leituras apresentaram TSD TA no local de inserção entre a TIR 5' e a célula CHO. Em todas as amostras houve uma diferença entre o número de possíveis inserções entre o 1º e 2º sequenciamento sendo que a população e o clone 28 apresentaram o maior número inserções. Conforme descrito nos materiais e métodos, os sítios de inserção obtidos foram filtrados para eliminação de inserções cujas leituras apresentaram alinhamento ambíguo (alinhamento em mais de uma posição no genoma de CHO com o mesmo *score*). Após essa filtração, teve uma grande diminuição do número de insertos que pode observado na última coluna tabela.

Tabela 3. Dados gerais do sequenciamento por Illumina dos fragmentos gerados por PCR splinkerette. Antes e após limpeza das sequências e filtrações de leituras corretas no 1º e 2º sequenciamento realizado com primers SB distintos.

Amostras	Total de leituras (R1 e R2)	Leituras pareados após Trimmomatic (R1 e R2) ¹	Leituras corretas (R1 – blastn CHO e R2 – blastn TIR)	Leituras R2 - <i>tag</i> do SB com TSD ²	Leituras R2 - <i>tag</i> do SB sem TSD ²	Número de possíveis inserções independentes (após filtração)
Clone 6 (SB) 1º	47700	38284	28453	8117	369	09 (4)
Clone 7 (SB) 1º	68974	32899	22589	12559	174	16 (5)
Pop (SB) 1º	21165	12205	9471	5470	83	85 (18)
Clone 6 (SB) 2º	103908	26057	11432	10721	107	08 (6)
Clone 7 (SB) 2º	79714	27599	23292	22240	185	04 (2)
Clone 28 (SB) 2º	218627	79435	71897	68277	520	32 (23)
Pop (SB) 2º	213377	81269	73699	70248	534	59 (34)

¹ – Sequências maiores que 100 bases para 1º sequenciamento e maiores que 220 bases para 2º sequenciamento.

² - Sequências maiores que 247 bases para o 1º sequenciamento e maiores que 220 bases para o 2º sequenciamento.

Os sítios de inserção (já filtrados) obtidos para as amostras estão sumarizados nas tabelas 4 (amostra 6), 5 (amostra 7), 6 (clone 28) e 7 (amostra populacional).

Tabela 4. Região de inserção do cassete inserido pelo Sistema Ingresso 1.0 e contexto genômico na célula CHO para o clone 6. O número de leituras se refere aquelas que possuem a etiqueta correspondendo aos primeiros 10 pb da TIR 5' (complementar reversa) seguido da TSD (TA). Todos os sítios foram checados pela ferramenta Genome Data Viewer do NCBI.

Leituras	Região de inserção	Compartilhada?	Região ativa?	Contexto genômico
6644 (1°) 9674 (2°)	NC_048601.1:84464972	pop	Sim	5' UTR (780 pb <i>upstream</i>) do gene <i>Fkbp9</i>
256 (1°) 231 (2°)	NW_023276807.1:155906939	todos	Sim	Região <i>upstream</i> do gene <i>Areg</i> (<i>amphiregulin</i>)
2 (1°)	NW_023276919.1:3928075	Clone 28	Não	Região não gênica
1 (1°)	NC_048596.1:170213246	Clone 28 e pop	Sim	1350 pb <i>downstream</i> do gene <i>Nudt21</i> (mRNA-nudix hydrolase 21)
1 (2°)	NC_048601.1:29288775	não	Sim	Intron do gene <i>Tsen2</i> (tRNA splicing endonuclease subunit 2)
1 (2°)	NC_048604.1:48106957	não	Não	Região não gênica
1 (2°)	NW_023276807.1:174122424	não	Não	Intron do gene <i>Rarb</i> (retinoic acid receptor beta)
1 (2°)	NW_023276807.1:27608663	não	Não	Região não gênica

Nota: 1° e 2° referem-se aos diferentes sequenciamentos desse clone

Tabela 5. Região de inserção do cassete inserido pelo Sistema Ingresso 1.0 e contexto genômico na célula CHO para o clone 7. O número de leituras se refere aquelas que possuem a etiqueta correspondendo aos primeiros 10 pb da TIR 5' (complementar reversa) seguido da TSD (TA). Todos os sítios foram checados pela ferramenta Genome Data Viewer do NCBI.

Leituras	Região de inserção	Compartilhada?	Região ativa?	Contexto genômico
11249 (1°) 21573 (2°)	NW_023276807.1:155906939	todos	Sim	Região <i>upstream</i> do gene <i>Areg</i> (<i>amphiregulin</i>)
1 (1°)	NC_048595.1:314402469	pop	Sim	Dentro de intron do gene <i>Iqgap2</i> (IQ motif containing GTPase activating)
1 (1°)	NC_048601.1:43627320	não	Não	Próximo de TEs
1 (1°)	NW_023276806.1:182870223	28 e pop	Sim	-
1 (1°)	NW_023276807.1:74974850	pop	Sim	Intron do gene <i>Rm3</i> (RRN3 homolog, RNA polymerase I transcription factor)
2 (2°)	NC_048596.1:208112109	não	Sim	Intron do gene <i>Epc1</i> (enhancer of polycomb homolog 1)

Nota: 1° e 2° referem-se aos diferentes sequenciamentos desse clone

Tabela 6. Região de inserção do cassete inserido pelo Sistema Ingresso 1.0 e contexto genômico na célula CHO para o clone 28. O número de leituras se refere aquelas que possuem a etiqueta

correspondendo aos primeiros 10 pb da TIR 5' (complementar reversa) seguido da TSD (TA). Todos os sítios foram checados pela ferramenta Genome Data Viewer do NCBI.

Leituras	Região de inserção	Compartilhada	Região ativa?	Contexto genômico
49721	NW_023276807.1:155906939	Todos	Sim	Região <i>upstream</i> do gene <i>Areg</i> (<i>amphiregulin</i>)
5300	NW_023276806.1:182870223	Clone 7 e pop	Sim	Região não gênica
5082	NW_023276919.1:3928075	Clone 6 e pop	Não	Região não gênica
2886	NC_048597.1:97137725	Pop	Sim	Intron do gene <i>Taok3</i> (TAO kinase 3)
1960	NC_048601.1:29541017	Pop	Sim	Região não gênica
793	NC_048596.1:170213246	Clone 6, pop	Sim	Região <i>downstream</i> do gene <i>Nudt21</i> (mRNA-nudix hydrolase 21)
283	NC_048597.1:115324104	Pop	Sim	Intron do gene <i>Cux1</i> (cut like homeobox 1)
219	NC_048597.1:109248388	Não	Sim	Intron do gene <i>Vkorc1/1</i> (vitamin K epoxide reductase complex subunit 1 like 1)
138	NC_048596.1:230963777	Pop	Não	Dentro de um elemento LTR
35	NC_048596.1:116895579	Não	Sim	3214 pb <i>downstream</i> do gene <i>lqgap1</i> (IQ motif containing GTPase activating protein 1)
26	NC_048595.1:151125255	Não	sim	Intron do gene <i>Asph</i> (aspartate beta-hydroxylase)
23	NC_048595.1:377625548	Pop	sim	Intron do gene <i>Enthd1</i> (ENTH domain containing 1) /Dentro de um LINE
17	NC_048595.1:385949881	Não	sim	Intron do gene <i>Ppp6r2</i> (protein phosphatase 6 regulatory subunit 2)
14	NC_048596.1:183489677	Pop	Sim	Intron do gene cytochrome b5 type B
6	NC_048595.1:177499390	Pop	Não	Região não gênica
3	NW_023276806.1:64262338	Não	sim	Intron do gene <i>Gatad2b</i> (GATA zinc finger domain containing 2B)
2	NC_048600.1:134353354	Pop	sim	Região <i>upstream</i> do gene Rn45s (45S ribosomal RNA) / Dentro de um SINE
1	NC_048595.1:262228516	Não	Sim	Dentro de um transposon hAT(Tip100) / <i>Upstream</i> do gene <i>Rev3l</i> (-REV3 like, DNA directed polymerase zeta catalytic subunit)
1	NC_048595.1:41233981	pop	sim	Intron do gene <i>Orc1</i> (origin recognition complex subunit 1)
1	NC_048599.1:101603605	Não	sim	Região <i>upstream</i> do gene <i>Wipf1</i> (WAS/WASL interacting protein family member 1) / Dentro de um LTR
1	NC_048600.1:44857609	Pop	Não	Intron do gene <i>Baiap3</i> (BAI1 associated protein 3)
1	NC_048603.1:28770725	Não	Não	Região não gênica
1	NW_023276806.1:87318060	Não	sim	Exón de um gene não codificante lncRNA (LOC113832892) / Dentro de um LTR

Tabela 7. Região de inserção do cassete inserido pelo Sistema Ingresso 1.0 e contexto genômico na célula CHO para a população. O número de leituras se refere aquelas que possuem a etiqueta correspondendo aos primeiros 10 pb da TIR 5' (complementar reversa) seguido da TSD (TA). Todos os sítios foram checados pela ferramenta Genome Data Viewer do NCBI.

Leituras	Região de inserção	Compartilhada?	Região ativa?	Contexto genômico
2790 (1°) 53558 (2°)	NW_023276807.1:155906939	Todos	Sim	Região <i>upstream</i> do gene <i>Areg</i> (<i>amphiregulin</i>)
2672 (2°)	NW_023276919.1:3928075	Clone 6 e 28	Não	Região não gênica
50 (1°) 822 (2°)	NW_023276806.1:182870223	Clone 7 e 28	Sim	Região não gênica
537 (2°)	NC_048596.1:39929115	Não	Sim	Região <i>downstream</i> do gene <i>Zfand5</i> (<i>zinc finger AN1-type containing 5</i>)
2 (1°) 487 (2°)	NC_048596.1:170213246	Clone 6 e 28	Sim	Região <i>dowstream</i> do gene <i>Nudt21</i> (mRNA-nudix hydrolase 21)
477 (2°)	NC_048601.1:29541017	Clone 28	Sim	Região não gênica
24 (1°) 429 (2°)	NC_048601.1:63165450	Não	Não	Região não gênica
20 (1°) 376 (2°)	NC_048597.1:97137725	Clone 28	Sim	Íntron do gene <i>Taok3</i> (TAO kinase 3)
213 (2°)	NC_048600.1:134353354	Clone 28	Sim	Região <i>upstream</i> do gene <i>Rn45s</i> (45S ribosomal RNA) / Dentro de um SINE
13 (1°) 176 (2°)	NW_023276807.1:74974850	Clone 7	Sim	Íntron do gene <i>Rrn3</i> (RRN3 homolog, RNA polymerase I transcription factor)
10 (1°) 168 (2°)	NC_048596.1:230963777	Clone 28	Não	Dentro de um elemento LTR
17 (1°) 167 (2°)	NC_048597.1:115324104	Clone 28	Sim	Intron do gene <i>Cux1</i> (cut like homeobox 1)
6 (1°) 133 (2°)	NC_048604.1:21665146	Não	Não	Região não gênica
1 (1°) 77 (2°)	NC_048600.1:44857609	Clone 28	Não	Intron do gene <i>Baiap3</i> (BAI1 associated protein 3)
5 (1°) 68 (2°)	NC_048595.1:41233981	Clone 28	Sim	Intron do gene <i>Orc1</i> (origin recognition complex subunit 1)
57 (1°)	NC_048601.1:29548155	Não	Não	Região não gênica
48 (2°)	NC_048601.1: 84464972	Clone 6	Sim	5' UTR (780 pb <i>upstream</i>) do gene <i>Fkbp9</i>
42 (2°)	NC_048595.1:314402469	Clone 7	Sim	Dentro de intron do gene <i>Iqgap2</i> (IQ motif containing GTPase activating)
5 (1°) 41 (2°)	NC_048596.1:168050306	não	Sim	Intron do gene <i>LOC103159710</i>
25(1°)	NC_048596.1:39936272	Não	Sim	Íntron do gene <i>Zfand5</i> (<i>zinc finger AN1-type containing 5</i>)
22 (2°)	NC_048597.1:199752729	Não	Não	Região não gênica
15 (2°)	NC_048597.1:72214419	Não	Não	Íntron do gene <i>LOC100760395</i>
12 (2°)	NW_023276806.1:12226116	não	Sim	Íntron do gene <i>Ttll7</i> (<i>tubulin tyrosine ligase like 7</i>)
9 (2°)	NC_048596.1:183489677	Não	Sim	Íntron do gene <i>LOC100760422</i> (cytochrome b5 type B)
8 (2°)	NC_048595.1:377625548	Clone 28	Sim	Intron do gene <i>Enthd1</i> (ENTH domain containing 1) /Dentro de um LINE
3 (1°) 8 (2°)	NC_048599.1:85721254	não	Sim	Intron do gene <i>Rapsn</i> (<i>receptor associated protein of the synapse</i>)
7 (2°)	NC_048601.1:89218177	Não	Não	Região não gênica
6 (2°)	NW_023276806.1:119718465	Não	Não	Região não gênica
4 (1°) 1 (2°)	NW_023276806.1:109258230	não	Sim	Intron do gene <i>Hspa4l</i> (<i>heat shock protein family A (Hsp70) member 4 like</i>)
3 (2°)	NC_048597.1:79485677	Não	Sim	Região não gênica

3 (1°)	NC_048597.1:94729468	não	Sim	Íntron do gene <i>Mlec (malectin)</i>
3 (2°)	NC_048599.1:105340420	não	Sim	Íntron do gene <i>Mettl5 (methyltransferase 5)</i>
2 (2°)	NW_023276807.1:1731919	Não	Sim	Íntron do gene <i>Mcu (mitochondrial calcium uniporter)</i>
1 (2°)	NC_048595.1:177499390	Clone 2 8	Não	Região não gênica
1 (2°)	NC_048595.1:41545531	Não	Não	Região não gênica
1 (2°)	NC_048602.1:25469999	Não	Sim	Região não gênica
1 (2°)	NW_023276806.1:98540139	Não	Não	Região não gênica
1 (1°)	NW_023276807.1:182650708	Não	Sim	Íntron do gene <i>Ccser2 (coiled-coil serine rich protein 2)</i>

Nota: 1° e 2° referem-se aos diferentes sequenciamentos dessa amostra

O número de leituras entre os sítios de inserção variou bastante, desde milhares de leituras a apenas uma. Esses sítios pouco representados poderiam caracterizar algum artefato da técnica. Porém, uma das inserções identificadas por meio de clonagem e sequenciamento de Sanger (NC_048596.1:208112109 do clone 7) foi representada por apenas 2 leituras no sequenciamento Illumina. Da mesma forma, é possível observar que alguns sítios com poucas leituras para um clone apresentam um maior número de leituras em outras amostras. Por exemplo, o sítio de inserção NC_048600.1:134353354 apresentou apenas 2 leituras para a clone 28, mas exibiu 213 leituras para a população e o sítio NW_023276806.1:182870223, o qual é compartilhado entre os clones 7, 28 e a população, apresentou 1, 5300 e 50 (1°) 822 (2°) leituras entre as amostras, respectivamente. Nesses casos, é muito provável que esses sítios com baixa leitura, para certas amostras, sejam de fato locais de inserção do transgene.

Para termos certeza de que as inserções com baixo número de leituras realmente caracterizam locais de inserção do transgene, a PCR splinkerette poderia ser repetida utilizando outras enzimas de restrição durante o preparo das amostras. Dessa forma, seriam gerados fragmentos de tamanhos distintos e assim poderíamos confirmar o local de inserção nos sítios que forem identificados em todas as PCRs, com as diferentes enzimas.

O número de inserções encontrados em cada amostra variou de 2 a 34 cópias. O clone 6 apresentou 4 (1°) e 6 (2°) inserções distintas, o clone 7, 5 (1°) e 2 (2°) inserções, o clone 28, 23 inserções e a população 18 (1°) e 34 (2°) inserções. É esperado, que as inserções dos clones sejam compartilhadas por todas as células desse clone, sendo assim, o número de inserções do clone 28 representa um valor acima do esperado para a *SB11* (KOLACSEK et al., 2014b; MATES et al., 2009). É

possível que esse alto número de inserções seja devido aos sítios com poucas leituras, que como discutido acima poderiam, em alguns casos, representar artefatos. Alternativamente, poderiam representar remobilização do cassete após a clonagem celular devido a permanência indesejável da transposase (ver discussão posterior). Nesse caso, a população clonal pode apresentar diferenças nas inserções em cada célula. Para a população, esse número maior de inserções já era esperado e pode ser facilmente explicado por que representa um *pool* heterogêneo de células.

A inserção do *SB* ocorre preferencialmente em locais ricos em dinucleotídeos TA. A maioria das sequências analisadas aqui apresentaram um TA no local de inserção, indicando o comportamento de inserção padrão dessa superfamília. As sequências com possível inserção não TA no primeiro sequenciamento não foram analisadas, pois a região era muito pequena. Já para o segundo sequenciamento, essas regiões foram alinhadas contra o genoma da célula CHO. Foi possível evidenciar pela inspeção visual do resultado na análise por blast que a maioria dessas leituras, na verdade, são resultado de inserções em regiões TA já identificadas e possuem substituições no sítio TA, indicando que houve erro durante o processo de amplificação ou no sequenciamento. Apenas 1 leitura (4393) do clone 6 apresentou característica de uma possível inserção não TA, onde os primeiros 118 nucleotídeos da leitura R2 alinharam perfeitamente ao início da TIR 5' da *SB* com posterior alinhamento com a célula CHO a partir do nucleotídeo 119 (posição NC_048596.1:254975977). Essa inserção, caso verdadeira, é em AA.

Apesar do número de clones analisado ser baixo, foi possível verificar inserções em locais de atividade transcricional bem como regiões de possível silenciamento. O clone 6 apresentou 4 inserções em local de atividade transcricional e 4 inserções em local sem atividade, o clone 7 apresentou 5 inserções em local de atividade e apenas uma inserção em local sem atividade. O clone 28, por sua vez, apresentou o maior número de inserções em local de atividade (18) entre os clones e 5 inserções sem atividade.

O conjunto das relações entre os sítios de inserção e as amostras, compartilhados e exclusivos, está representado graficamente no diagrama de Venn (Fig 16) e detalhado na tabela 8. É possível verificar sítios de inserção exclusivos de uma amostra bem como regiões que foram compartilhadas entre duas ou mais amostras.

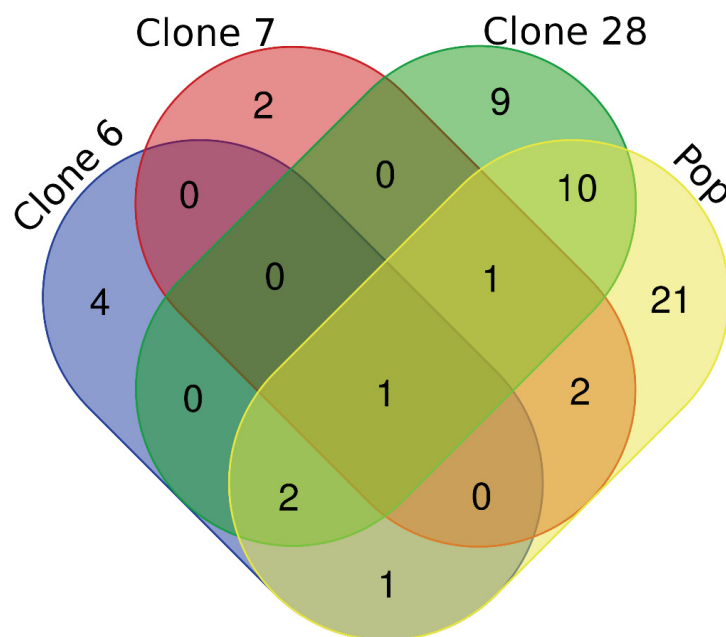


Figura 16. Diagrama de Venn mostrando o número de inserções potenciais compartilhadas entre as amostras. O diagrama foi gerado pela ferramenta disponível em: <http://bioinformatics.psb.ugent.be/webtools/Venn/>

Tabela 8. Inserções compartilhadas entre as amostras e as exclusivas de cada amostra.

Amostra	Total	Inserções
Clone 28 Clone 6 Clone 7 Pop	1	NW_023276807.1:155906939
Clone 28 Clone 6 Pop	2	NC_048596.1:170213246 NW_023276919.1:3928075
Clone 28 Clone 7 Pop	1	NW_023276806.1:182870223
Clone 6 Pop	1	NC_048601.1:84464972
Clone 7 Pop	2	NC_048595.1:314402469 NW_023276807.1:74974850
Clone 28 Pop	10	NC_048596.1:183489677 NC_048601.1:29541017 NC_048600.1:44857609 NC_048600.1:134353354 NC_048597.1:97137725 NC_048595.1:41233981 NC_048597.1:115324104 NC_048596.1:230963777 NC_048595.1:377625548 NC_048595.1:177499390

Clone 6	4	NC_048604.1:48106957 NC_048601.1:29288775 NW_023276807.1:174122424 NW_023276807.1:27608663
Clone 7	2	NC_048601.1:43627320 NC_048596.1:208112109
Clone 28	9	NC_048595.1:385949881 NC_048597.1:109248388 NC_048595.1:151125255 NC_048603.1:28770725 NW_023276806.1:64262338 NW_023276806.1:87318060 NC_048596.1:116895579 NC_048599.1:101603605 NC_048595.1:262228516
Pop	21	NW_023276807.1:182650708 NC_048597.1:199752729 NC_048601.1:89218177 NC_048596.1:168050306 NC_048596.1:39936272 NC_048599.1:105340420 NC_048602.1:25469999 NC_048597.1:72214419 NW_023276806.1:109258230 NC_048601.1:29548155 NW_023276806.1:98540139 NC_048601.1:63165450 NC_048597.1:94729468 NC_048599.1:85721254 NC_048604.1:21665146 NC_048596.1:39929115 NC_048597.1:79485677 NW_023276806.1:12226116 NW_023276806.1:119718465 NW_023276807.1:1731919 NC_048595.1:41545531

Todos os clones são derivados da mesma população de células CHO transfectadas com os vetores do Sistema Ingresso 1.0. No entanto, apenas um local de inserção foi comum entre todas as amostras analisadas (NW_023276807.1:155906939 - Região *upstream* do gene *Areg* - amphiregulin). Isso demonstra que os clones foram derivados de uma mesma célula inicial com essa inserção. Alguns sítios de inserção foram compartilhados entre a população e mais uma ou duas amostras e cada clone também apresentou inserções exclusivas como o clone 6 que apresentou 4 inserções específicas, o clone 7 duas inserções e o clone 28 nove inserções.

Se todas as inserções fossem igualmente identificadas na técnica, seria esperado que, a população tivesse todas as inserções presente nos diferentes clones,

pois a transposase deveria permanecer ativa apenas nas primeiras horas após a transfecção e ser eliminada durante as replicações celulares. O padrão de compartilhamento de inserções indica que a transposase pode ter ficado ativa pelo menos após algumas divisões celulares. As inserções exclusivas têm maior chance de ser algum artefato ou podem indicar que a transposase não foi eliminada e se manteve ativa após a clonagem celular gerando, dessa maneira, novas e exclusivas inserções em diferentes células das populações clonais.

O clone 28 é o que apresentou a maior expressão da proteína de interesse em relação as outras amostras e é também o que mostrou maior número de possível inserções. Essa expressão maior pode ser um efeito cumulativo da expressão de todos os cassetes inseridos ou ainda é possível que uma ou mais das inserções específicas presentes no seu genoma seja responsável por esse aumento de expressão. Dentre as 9 inserções exclusivas presentes no clone 28, apenas uma ocorreu em região sem atividade transcricional. As demais sequências foram inseridas em íntrons ou próximos a genes. Dentre esses genes estão o *Vkorc1l1* (*vitamin K epoxide reductase complex subunit 1 like 1*) o qual está envolvido no metabolismo da vitamina K, o *Ppp6r2* (*protein phosphatase 6 regulatory subunit 2*) que atua na regulação de uma quinase mitótica essencial, o *Asph* (*aspartate beta-hydroxylase*) que desempenha um papel importante na homeostase do cálcio, o *Gatad2b* (*GATA zinc finger domain containing 2B*) um repressor transcricional, o *Iqgap1* (*IQ motif containing GTPase activating protein 1*) que atua na regulação da morfologia e a motilidade celular, o *Wipf1* (*WAS/WASL interacting protein family member 1*) que apresenta um papel importante na organização do citoesqueleto de actina e o *Rev3l* (*REV3 like*) uma subunidade catalítica da DNA polimerase zeta. Como consequência da inserção próximo a um gene com expressão constitutiva, a expressão do cassete poderia ser amplificada em relação às outras inserções nos demais clones. Um bom exemplo é o gene *REV3l*, uma subunidade catalítica da DNA polimerase zeta que está envolvido com a estabilidade genômica e parece ser essencial para a sobrevivência das células em camundongo devido ao aumento da quebra cromossômica quando inativado (LANGE et al., 2016).

Ainda, é possível que as inserções importantes para o aumento da expressão da proteína de interesse no clone 28 sejam compartilhadas apenas com a população. Essas inserções também estão em íntrons e outras, assim como para algumas inserções exclusivas, estão localizadas dentro de TEs. É sabido que os TEs são ricos

fontes de sequências *cis*-regulatórias podendo funcionar como promotores ou *enhancers* (LORETO et al., 2018; SUNDARAM e WYSOCKA, 2020).

É preciso ressaltar, porém, que o Sistema Ingresso inclui a sequência de insuladores nas extremidades do cassete o que deveria, na teoria, impedir o efeito de ativação e repressão induzido pelas regiões adjacentes. No entanto, a atividade do insulador não foi testada nessa construção.

4.1.2 Evidência de integração aleatória

Em todas as amostras avaliadas, um pequeno número de leituras foi mapeado fora do cassete, principalmente na região *upstream* da TIR 5' (1ª leva de sequenciamento: 87 leituras R1 da população, 48 leituras R1 do clone 6 e 161 leituras R1 do clone 7; 2ª leva de sequenciamento: 15 pares de leituras R1 e R2 da população, 177 leituras R1 e 239 leituras R2 do clone 28, 1 par de leituras R1 e R2 dos clones 6 e 7). Numa inserção baseada em transposição, essa região mapeada não é integrada, uma vez que a transposase realiza a excisão do DNA na região das TIRs. A presença dessas leituras pode ter duas possíveis explicações: 1) são oriundos dos vetores originais que permaneceram na amostra. Essa hipótese é menos provável, pois é esperado que esses vetores não sejam replicados nas células CHO e sejam eliminados após algumas passagens. Embora tenha indicativos na literatura que vetores contendo o insulador HS4, presente no Sistema Ingresso, pode se replicar de maneira estável e eficiente em células mamíferas por pelo menos até 100 gerações (ZHANG et al., 2017) ou 2) são oriundos de uma integração não mediada pela transposase.

Todas essas leituras foram isoladas e mapeadas contra o genoma da célula CHO e apenas uma leitura (nº 16765) oriunda do clone 6 e algumas leituras do clone 28 mostraram também similaridade com o genoma da célula CHO. A análise manual da leitura do clone 6 mostrou que se trata de uma inserção aleatória, não mediada pela transposase (Fig 17). É possível observar que a parte inicial da leitura alinha com o genoma da célula CHO (dos nucleotídeos 33 ao 73), incluindo o sítio de *BamHI* onde o adaptador da PCR splinkerette se ligou, e a parte final da leitura alinha com o vetor *donor* a partir do nucleotídeo 63, indicando uma sobreposição de 11 nucleotídeos. Essa região de similaridade entre a célula CHO e o vetor (chamada por alguns autores de região de microhomologia) deve ter sido utilizada para a integração do vetor no

Figura 17. Inserção aleatória da *SB* no genoma de célula CHO para o clone 6. Resultado do blast do vetor *donor* da *SB* contra o genoma da célula CHO. A região destacada no quadrado vermelho é a região de 11 nucleotídeos do vetor *donor*, na região *ustream* da TIR 5', que gerou homologia com o genoma da célula CHO.

4.1.3 Correlação entre o número de cópias do cassete e a expressão relativa do gene de interesse.

Para a quantificação do número de cópias do cassete integrado no genoma da célula CHO dos clones selecionados, empregamos a técnica de PCR em tempo real a partir do DNA das amostras. O gene glucagon, gene endógeno de célula CHO, apresenta cópia única no genoma e foi selecionado como normalizador para essas análises (BAHR et al., 2009). Para identificar a quantidade de cassetes inseridos no genoma utilizamos primers específicos para o gene Pró-Trombina nas análises. A curva padrão de cada um desses genes foi construída com diluições seriadas 1:10 com DNA genômico da célula CHO para o gene glucagon (Fig 18) e com o DNA plasmidial do vetor *donor* contendo o gene da Pró-trombina (Fig 19).

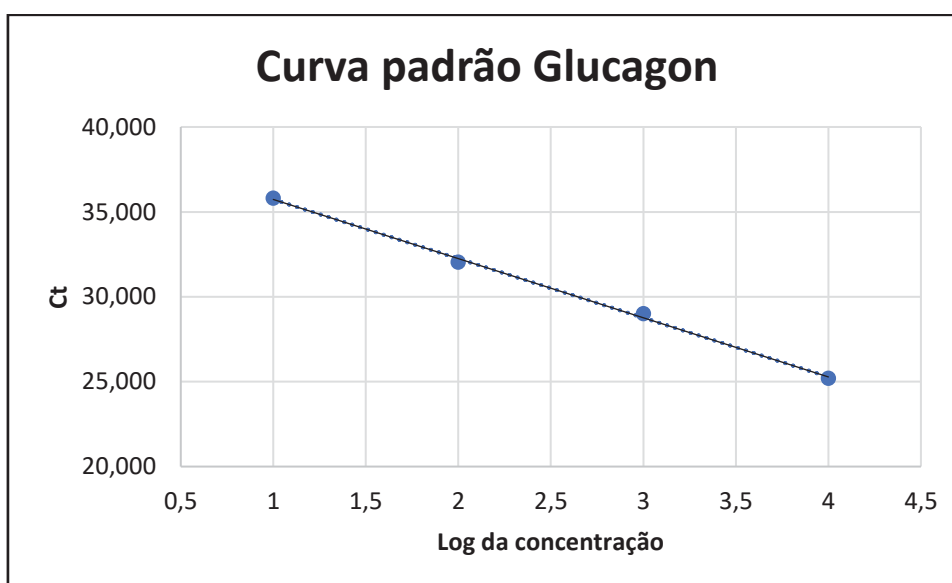


Figura 18. Curva padrão do gene Glucagon. Diluições seriadas do DNA de célula CHO de 10^4 cópias até 10 cópias.

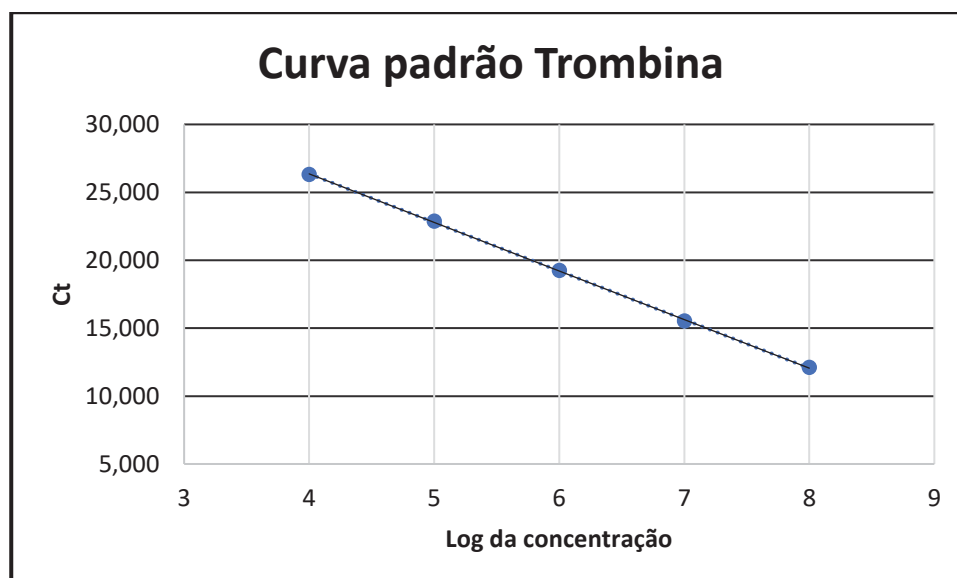


Figura 19. Curva padrão do gene Pró-trombina. Diluições seriadas do DNA plasmidial de 10^8 cópias até 10^4 cópias.

A curva padrão do gene glucagon foi construída com 4 concentrações da célula CHO em diluição seriada 1:10 (10^4 cópias - 10 cópias) e apresentou eficiência da reação de 93% com R^2 de 0,998 ($slope = -3,4846$). A curva-padrão do gene Pró-trombina foi construída com concentrações do DNA plasmidial (diluição seriada 1:10) iniciando em 10^8 cópias e apresentou eficiência da reação de 90% com R^2 de 0,999 ($slope = -3,576$). Uma eficiência de 100% indica que o número de moléculas dobrou a cada ciclo da reação, mas na prática é aceitável eficiências entre 90 e 115% ($slope$ entre -3,0 e -3,6) (OVERBERGH et al., 2017).

Para a obtenção do número de cópias dos clones selecionados, os resultados foram alcançados extrapolando os valores de Ct obtidos na equação da reta das curvas padrão para cada um dos genes. Os dados das análises de número de cópias foram compilados na figura 20. Para as amostras provenientes de clones, o esperado é que a maior parte das células tenham o mesmo número de cassetes enquanto os dados da população, o número de cópias representa uma média dos cassetes encontrados nas células.

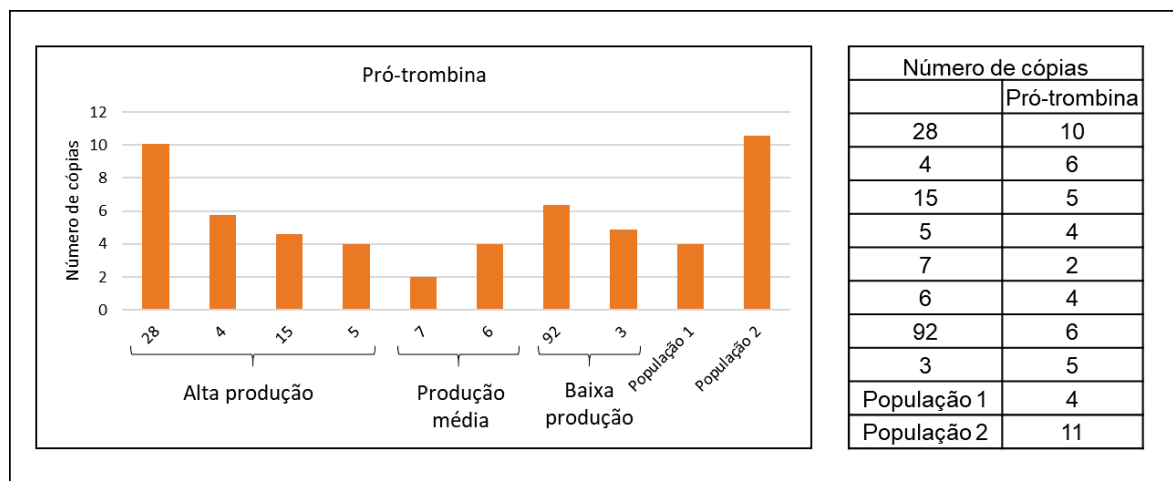


Figura 20. Valores obtidos para o número de cópias do cassete contendo o gene Pró-trombina nos clones e populações.

A amostra 7 apresentou o menor número de cópias inseridas por célula, apenas duas. Por outro lado, os dados obtidos para a amostra 28, que possui a maior taxa de expressão da proteína Pró-trombina, foi o maior entre os clones analisados, 10 cópias. As amostras 3 e 92 apresentaram as menores taxas de expressão da proteína de interesse, mas possuem o mesmo número de cópias do cassete (5 e 6 cópias, respectivamente) que amostras com alta taxa de expressão, clones 4 e 15. As amostras oriundas de população apresentaram em média de 4 a 11 cópias. Os dados obtidos para o Sistema Ingresso 1.0 demonstram uma variação no número de cópias do cassete inserido na célula CHO de 2 a 10 cópias entre as amostras derivadas de uma única célula.

Ao comparar esses dados com o número de inserções obtidos pelo sequenciamento, observa-se uma discrepância, pois um número muito maior de inserções foi obtido. Como já discutido, isso pode significar que existem artefatos na tabela de inserções ou é devido a remobilização do cassete. Nesse último caso, se a transposase retira o cassete de um lugar e insere em outro e o sítio original for reparado pelo mecanismo de junção de extremidades, o número de cassetes por células não vai aumentar, mas o número de sítios distintos na população do clone será maior.

Esses resultados demonstram que ao utilizar um sistema baseado em transposição para a transferência gênica é possível alcançar várias cópias do cassete de expressão por célula. A integração randômica, por outro lado, leva a integração de

apenas 1 cópia na maioria das vezes, por ser um processo muito ineficiente, menos de 1% das células são transfectadas (KOLACSEK et al., 2014a).

Uma comparação entre as diferentes formas hiperativas da *SB*, *SB11*, *SB32* e *SB100X* em célula HEK-293, utilizando diferentes quantidades de transposon e da transposase revelou um número de cópias elevado ao utilizar uma quantidade 10 vezes maior de vetor *donor* que do *helper* (10:1) variando de 19 cópias para *SB11* e *SB32* a 38 cópias para a *SB100X*. Por outro lado, ao utilizar 10 vezes menos quantidade de vetor *donor* em relação ao *helper* (1:10), o número de cópias encontrado foi reduzido drasticamente para 1,4 para a *SB11*, 3,5 para *SB32* e 5,8 para *SB100X*, muito provavelmente devido a inibição por superprodução da transposase (KOLACSEK et al., 2014b). Análise de diferentes clones derivados da transposição pela *SB100X* demonstrou a presença de uma a 4 cópias do cassete integradas em células HEK-293 (KOLACSEK et al, 2011), de duas a sete cópias em célula HeLa e de uma a quatro cópias em célula GABEB (TURCHIANO et al, 2014). Outro estudo utilizando células HeLa identificaram de 8 a 20 inserções (QUERQUES et al, 2019). A utilização direta da proteína *SB* hiperativa para integração genômica do cassete em célula HeLa revelou a inserção de 2 a 12 cópias por células dependendo da quantidade de proteína fornecida inicialmente (2,5 – 40 µg) (QUERQUES et al, 2019).

A alta atividade de excisão nem sempre é acompanhada por uma taxa elevada no número de cópias do cassete, como exemplificado pela versão hiperativa *SB100X*. Essa nomenclatura levou em conta a eficiência de remobilização de cópias cromossômica, mas o aumento do número de cópias por clones é de apenas 3 a 4 vezes. A eficiência da *SB100X* também é maior em termos de proporção de células transfectadas contendo a inserção (até 80X) (MÁTÉS et al, 2009).

Em geral, o que podemos observar é que o número de cópias inseridas mostra diferenças notáveis dependendo de qual transposase é utilizada, da construção do vetor *donor* e linhagens celulares analisadas e pode-se considerar que o número de insertos obtidos com a *SB11* nesse trabalho está dentro do esperado. Outro ponto importante a ser considerado é a quantidade de cassete utilizado durante o processo de transfecção. Alguns estudos demonstraram que o aumento do número de cópias é proporcional ao aumento da quantidade de vetor *donor* utilizado na transfecção (DOHERTY et al., 2012; CHEN et al., 2015).

A análise da expressão relativa é baseada na taxa de expressão de um gene alvo normalizado por um gene de referência. O gene *Eif3i* (gene endógeno da célula

CHO de expressão constitutiva) foi selecionado como referência para as análises de expressão gênica do RNA em células CHO por qPCR (BAHR et al, 2009). O gene *Eif3i* apresentou eficiência de reação de 96% com R^2 de 0,997 ($slope = -3,41$) (Fig 21).

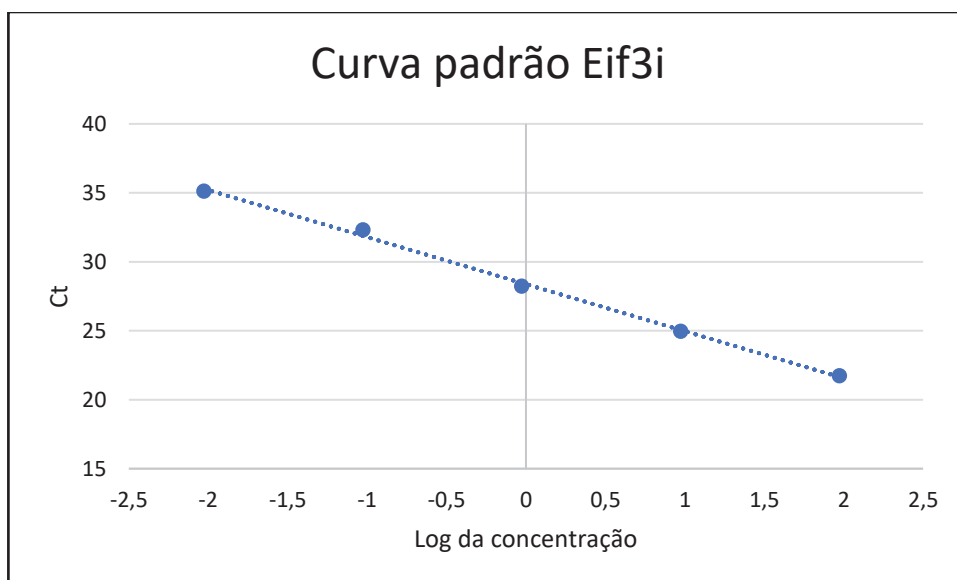


Figura 21. Curva padrão do gene Eif3i. Diluições seriadas do RNA de célula CHO de 94 ng até 0,0094 ng.

As amostras de cDNA foram amplificadas com primers específicos para os genes Pró-trombina e *Eif3i*. As análises de expressão relativa foram feitas através do método Ct comparativo, no qual os níveis de mRNA das amostras foram comparados com a amostra 3, que apresentou a menor expressão proteica no ensaio de Elisa.

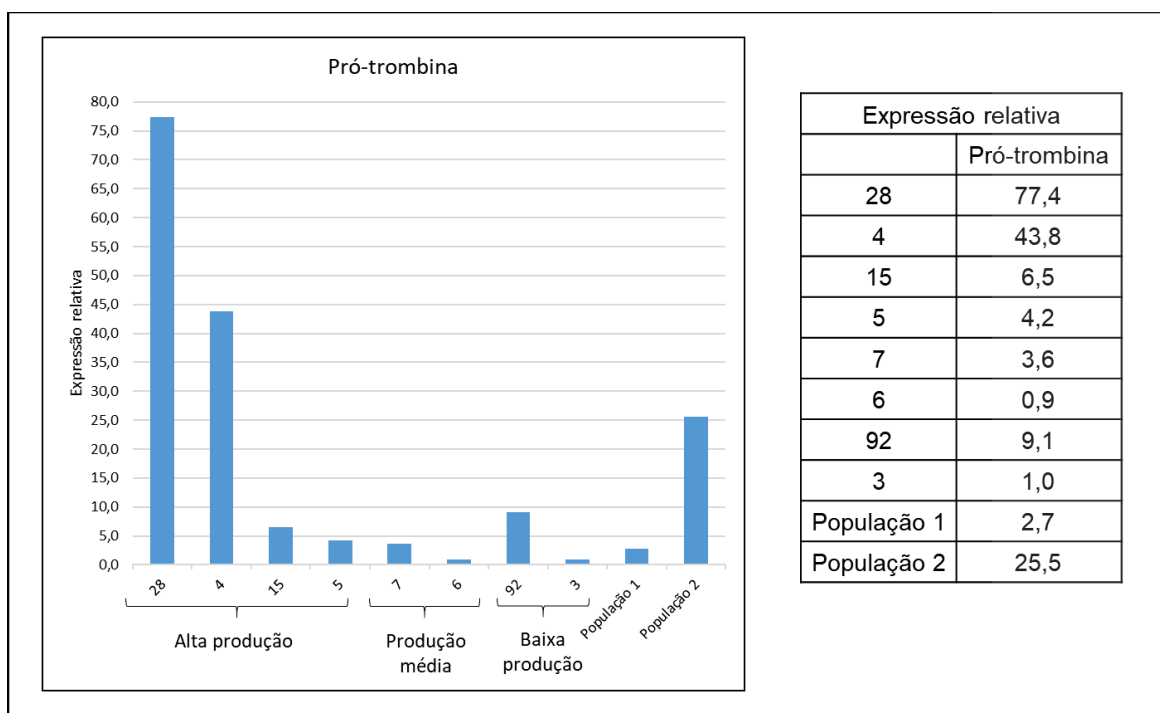


Figura 22. Valores obtidos para a expressão relativa do gene Pró-trombina nos clones e populações.

Os dados da expressão relativa (Fig 22) demonstraram que houve um aumento na expressão dos genes alvos em todas as amostras quando comparados a amostra 3, com exceção da amostra 6 que apresentou um nível semelhante de expressão a nível de RNA apesar de ter maior expressão proteica. A amostra 28 apresentou a maior taxa de expressão de Pró-trombina a nível de RNA sendo mais de 75 vezes maior que a amostra 3. A amostra 4 também apresentou um aumento significativo dos níveis de RNA para a Pró-trombina em relação a amostra 3, aproximadamente 44 vezes maior. Esses dados corroboram com as análises obtidas a nível de proteína para as amostras 28 e 4. Por outro lado, as amostras 15 e 5 que apresentaram alta taxa de expressão da proteína apresentaram um aumento sutil a nível de RNA, de 6,5 e 4,2 vezes, respectivamente. Enquanto a amostra 92 que apresentou baixa expressão proteica, semelhante a amostra 3, demonstrou aumento do RNA em 9 vezes, valor mais alto do que o obtido pelas amostras 15 e 5. Para as amostras derivadas de população a variação da taxa de RNA foi de 2,7 para 25,5 a mais que a amostra 3, sendo que ambas apresentaram nível de expressão proteica semelhante a amostra 5.

No geral, esses dados demonstram que houve um aumento no nível de expressão dos mRNAs em relação a amostra 3 como verificado para o nível de

proteína sugerindo uma correlação positiva entre mRNA e proteína. Os clones com maior expressão proteica também apresentaram um alto número de cópias (amostras 28 e 4). No entanto, essa relação não é perfeita uma vez que algumas amostras com baixa expressão proteica tiveram um grande aumento dos mRNAs e, por outro lado, algumas amostras com alta expressão apresentaram um aumento moderado no nível de mRNA. Essa relação também não foi perfeita para o número de cópias. Por exemplo, a amostra 7 apresentou o menor número de cópias, menos da metade que a amostra 3, mas teve um aumento de expressão de 3,6 vezes. Em alguns casos onde o número de cópias foi semelhante (clones 5, 15 e 3) houve uma diferença no nível de expressão do mRNA (4,2; 6,5 e 1, respectivamente). De fato, a relação entre o DNA, expressão do mRNA e proteína varia entre os genes (GUO et al., 2008). Essa discrepância pode ser atribuída aos mecanismos pós-transcricionais que podem moldar os níveis de proteína independentemente da abundância de mRNA (BUCCITELLI e SELBACH, 2020).

Dados da literatura mostram um aumento sutil do nível de mRNA do cassete integrado utilizando *SB100X*, além disso, os cassetes que apresentaram maior número de cópias inseridas apresentaram menor taxa de expressão a nível de mRNA (KOLACSEK e ORBÁN, 2018). Outro estudo, utilizando a *SB11* para inserção genômica em camundongo para a geração de animais transgênicos, demonstrou um aumento de 15 vezes na expressão do transcrito em relação ao animal não transfectado, o qual também foi verificado a nível de proteína por *Western Blot* (GAO et al., 2018).

Os dados obtidos para a expressão relativa do mRNA utilizando o Sistema Ingresso 1.0 foram bem elevados, alcançando um aumento máximo de 77 vezes em relação a amostra 3. Esses resultados sugerem que os elementos presentes no vetor *donor* podem favorecer a transcrição do mRNA da proteína de interesse. Os elementos *cis-acting* incluídos nos vetores de expressão, como promotores, *enhancers*, sinais de poliadenilação e outros elementos de expressão podem afetar os níveis de expressão do transgene (XU et al., 2019). Por exemplo, o insulador *chs4* muito utilizado em transferência de genes devido às suas duas funções iniciais de barreira e de bloqueio do intensificador, pode aumentar a expressão do cassete além de facilitar o estabelecimento do vetor levando ao aumento do número de células que expressam o transgene (RIVELLA et al., 2000; HAGEDORN, ANTONIOU e LIPPS, 2013). Além disso, esse elemento é capaz de diminuir o silenciamento e a metilação

do vetor de expressão (RIVELLA et al., 2000). Os promotores CMV e CAG estão entre os mais fortes em células de mamífero e apresentam alta taxa de expressão gênica em células CHO, além de melhorar a estabilidade do transgene (WANG et al., 2016; ZÚÑIGA et al., 2019). O vetor *donor* do Sistema Ingresso possui o insulador cHS4 nas extremidades 5' e 3' e o promotor CMV e um fragmento do promotor CAG. Esses elementos são importantes para o aumento dos mRNAs mas outros fatores podem ter influenciado nas amostras analisadas.

O vetor de expressão do Sistema Ingresso apresenta um sítio interno de ligação ao ribossomo (IRES) que permite a expressão do gene alvo e o GFP em um mesmo mRNA. Para verificar a se houve a expressão dos genes da Pró-trombina e GFP no mesmo mRNA, foi realizada PCR a partir do cDNA das amostras com primers anelando no início do gene da Pró-trombina e final do gene GFP, com tamanho esperado de 3250 pb. A maioria as amostras analisadas apresentaram amplificação no tamanho esperado indicando que o mRNA bicistrônico foi expresso corretamente (Fig 23). As amostras 15, 92 e 6 tiveram uma banda bem fraca no gel de agarose. Para a amostra 3 não foi possível visualizar uma banda correspondente, que pode sugerir uma baixa expressão do mRNA, como verificados nas análises de qPCR.

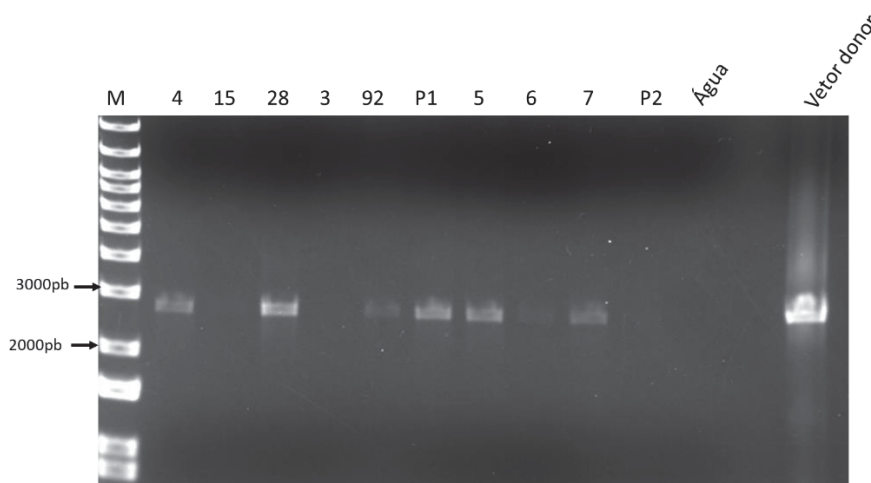


Figura 23. Gel de agarose da confirmação do mRNA bicistrônico contendo o gene da Pró-trombina e GFP nos clones e populações isolados de célula CHO inseridos pelo Sistema Ingresso 1.0. Legenda: M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific); O vetor donor utilizado nas transfecções foi usado com controle da PCR.

4.1.4 Integração da transposase

A transposase codificada pelo vetor *helper* não possui as TIRs propositalmente, para evitar a mobilização e integração dessa sequência no genoma. Porém, a integração aleatória desse vetor no genoma pode ocorrer levando à instabilidade genômica devido à expressão prolongada e incontrolável da transposase, resultando na remobilização contínua do cassete de expressão já integrado.

Para verificar se houve integração da transposase no genoma da célula CHO, a presença da *SB* foi investigada no DNA, RNA e proteína dos clones obtidos pelo Sistema Ingresso 1.0. Uma possível inserção da transposase no genoma da célula hospedeira foi confirmada por qPCR (dados não mostrados) e PCR convencional (Fig 24) com as amostras de DNA genômico. Todos os clones avaliados por PCR em tempo real apresentaram amplificação ao utilizar os primers internos da *SB*. Através da curva de dissociação (*melting curve*) foi possível verificar a especificidade do produto, confirmando a correta amplificação em todas as amostras.

Para a PCR convencional foram utilizados diferentes pares de primers: internos a sequência da *SB* (Fig 24 A, tamanho esperado de 85 pb), um primer interno da *SB* e outro do vetor *helper* na porção 3' (Fig 24 B, tamanho esperado de 440 pb) e pares que anelam no início e final da sequência da transposase do *SB* (Fig 24 C, tamanho esperado de 1040 pb).

Todas as amostras foram amplificadas tanto com os primers internos da *SB*, de maneira bem fraca, quanto com um primer do vetor *helper* na porção 3'. Porém, ao utilizar os primers para a amplificação da transposase completa, apenas as amostras 28 e a população apresentaram amplificação de tamanho esperado. Esse resultado pode indicar que apenas essas amostras, clone 28 e população, possuem a transposase completa a qual poderia levar a eventos de remobilização e causar uma instabilidade genômica. Infelizmente, as amostras dos clones 6 e 7 não foram avaliados nesses testes. A presença de apenas parte da transposase nos demais clones poderia ser explicada pela inserção aleatória do vetor *helper* usando uma região de microhomologia dentro da transposase. Assim, a transposase certamente não tem atividade funcional nos demais clones, por apresentar a sequência incompleta da proteína.

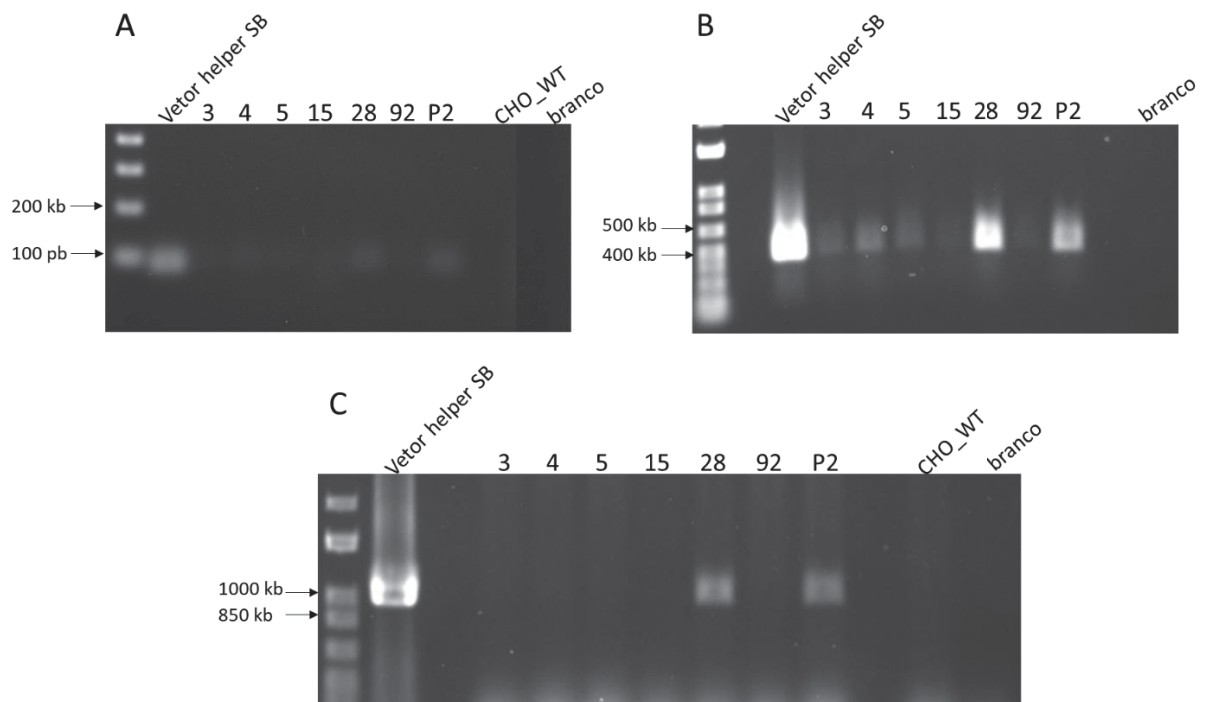


Figura 24. Confirmação da integração da sequência da transposase *SB* no genoma da célula CHO nos clones e populações por PCR. **A:** amplificação com primers internos da *SB* e tamanho esperado de 85 pb. **B:** amplificação com os primers *SB_F* e *polyA_R* (vetor *helper*) e tamanho esperado de 440 pb. **C:** amplificação com os primers externos *SBpet28_F* e *SBpet28_R* e tamanho esperado de 1040 pb.

Apesar da PCR splinkerette não ser desenhada para amplificação e sequenciamento de inserções do vetor *helper*, as leituras foram alinhadas por blastn contra a sequência desse vetor. Foi possível identificar leituras correspondentes ao vetor *helper* para o clone 28 (Fig 25). Esses dados indicam fortemente que a transposase *SB* está integrada no genoma da célula CHO para o clone 28. Essa observação corrobora a ideia de que, pelo menos para esse clone, o número maior de inserções identificadas no sequenciamento seja em parte devido a remobilização do cassete.



Figura 245. Par 1171 de leituras (R1 e R2) oriundas do clone 28 que alinharam com o vetor *helper*. Em vermelho está destacado a região codificadora da transposase *SB*.

Não houve amplificação das amostras a partir do cDNA por PCR convencional nem qPCR (dados não mostrados). Além disso, foi realizada ensaio de *Western Blot* utilizando anticorpos específicos para a *SB* nas culturas dos clones 5 e 28 para verificar a expressão da proteína nessas amostras. Não foi possível observar sinal de expressão nessas amostras no momento avaliado (dados não mostrados).

Casos de integração da *SB* já foram reportados anteriormente (HUANG et al, 2010b) a nível de DNA, mas não houve a detecção do transcrito de mRNA da transposase ou da expressão da proteína. Porém, é possível que a transposase seja expressa em pouca quantidade e que não tenha sido detectada. Ainda, a transposase pode ser ativada em condições/momentos específicos sendo indesejável a sua presença nas culturas.

Essa possível integração do vetor *helper* do Sistema Ingresso pode ter sido consequência da grande quantidade de vetor utilizado durante o processo de transfecção. A maioria dos trabalhos utiliza uma quantidade entre 100 a 1000 ng de

cada vetor (GEURTS et al., 2003; DING et al., 2005; LI et al., 2013) enquanto, neste trabalho, utilizamos 3,25 ug de dos vetores *helper* e 9 ug de vetor *donor*.

Uma alternativa que vem sendo descrita na literatura é a transfecção do RNA mensageiro que codifica para a transposase (WIEHE et al, 2007; WILBER et al, 2007). A utilização do mRNA aumenta a segurança dessa abordagem uma vez que não corre o risco de integração genômica espontânea (AMBERGER e IVICS, 2020). Além disso, a utilização direta da proteína recombinante permite um controle adicional sobre a expressão da transposase. A baixa meia vida em condições fisiológicas da proteína recombinante limita sua presença nas células reduzindo o risco de remobilização do transgene (SANDOVAL-VILLEGAS et al, 2021).

A proteína recombinante da SB100X foi usada para mediar a transgênese em diversas células de mamíferos, células-tronco e para gerar células CD19 CAR-T. Foi observado por análise de *Western Blot* uma perda substancial da proteína a partir de 24 horas após a transfecção, enquanto células transfectadas com o vetor plasmidial contendo a *SB* continuaram a expressar altos níveis de proteína por até 5 dias (QUERQUES et al, 2019).

Com o intuito de contornar essa possível integração da *SB* nas amostras de célula CHO, realizamos a clonagem do gene das transposases (*SB* e *Tp2*) no vetor de expressão bacteriano pET28a e como perspectivas iremos realizar a expressão em *E. coli* e testar a transfecção dessas proteínas em células CHO juntamente com o vetor *donor*.

4.2 NOVAS TRANSPOSASES

Como já mencionado, com o intuito de melhorar o Sistema Ingresso, escolhemos novas transposases para serem testadas. Iniciamos este estudo com os transposons *Tp2* (superfamília *Tc1-Mariner*), *Tp5* (superfamília *hAT*) e *Tp7* (superfamília *Piggybac*), os quais possuem diferentes características e domínios proteicos (Fig 26). O *Tp2* foi descrito numa espécie de formiga e apresenta similaridade com o elemento *SB* e nenhum artigo publicado. O *Tp5* pertence a uma espécie de díptera e apresenta os domínios característicos da superfamília *hAT* e proximidade filogenética a elementos sabidamente ativos. A *Tp7* é transcrionalmente ativa e foi isolada de uma espécie de coral.

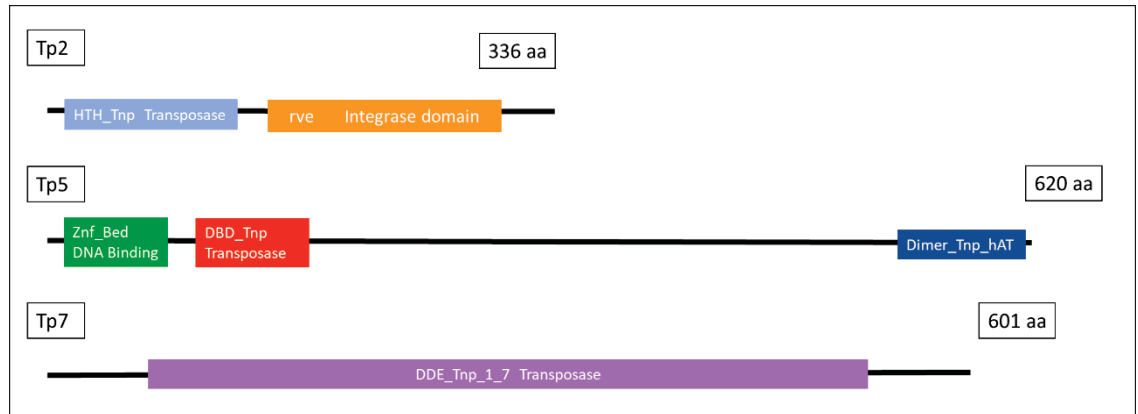


Figura 256. Esquema das novas Transposases e os domínios proteicos presentes em cada proteína.

4.1.1 Clonagem nos vetores *helper* e *donor*.

As sequências dos genes de *Tp2*, *Tp5* e *Tp7* foram otimizadas para a expressão em célula CHO, sintetizadas pela GenScript e inseridas no vetor pUC57. Para a inserção das novas transposases no vetor *helper* original, tanto o vetor quanto os genes foram digeridos com a enzima *SacII* (Fig 27A) e purificados (Fig 27B). Após a clonagem dos genes no vetor *helper* original sem a *SB*, foi realizada a confirmação da inserção por digestão. Nesta etapa verificamos a presença de diversos clones com perfil de bandas diferente do esperado (Fig 27C). Porém, para cada uma das transposases foi possível identificar ao menos um clone com perfil de bandas esperado (*Tp2* canaleta 1; *Tp5* canaletas 1 e 2; *Tp7* canaleta 3), os quais foram enviados para sequenciamento. Após análise, a correta inserção das transposases no vetor *helper* foi confirmada.

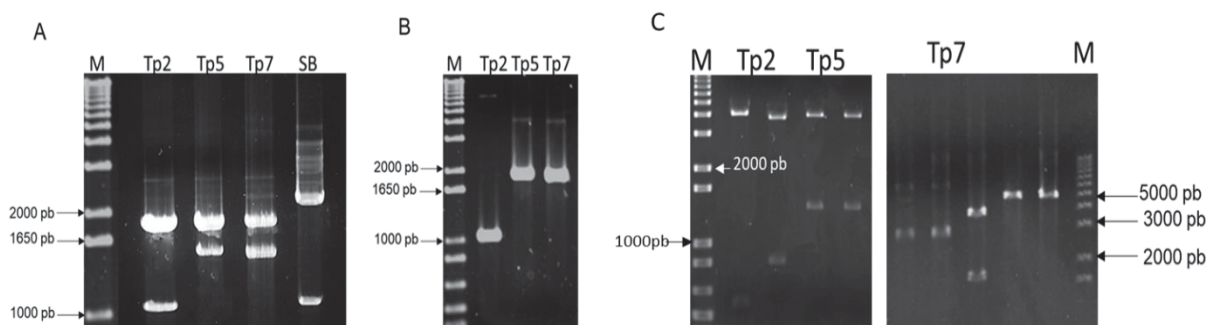


Figura 267. Gel de agarose da clonagem dos genes das transposases 2, 5, 7 no vetor *helper* do Sistema Ingresso. **A:** digestão dos vetores pUC57 liberando os genes das novas transposases e do vetor *helper*

liberando o gene da *SB* com a enzima *SacII*; **B**: produtos purificados após a digestão com *SacII*; **M**: marcador de DNA; *Tp2* (1028 pb); *Tp5* (1880 pb); *Tp7*: (1823 pb), *SB*: vetor *helper* (4732 pb) com *Sleeping Beauty* (1023 pb); **C**: Confirmação da inserção das novas transposases pela digestão do vetor *helper* com as seguintes enzimas: *Tp2*: *XhoI* e *EcoRV* libera fragmento de 845 pb; *Tp5*: *XhoI* e *EcoRV* libera fragmento de 1300 pb; *Tp7*: *XhoI* e *NheI* libera fragmento de 1735 pb.

As TIRs 3' e 5' de cada nova transposase foram sintetizadas e clonadas no vetor pUC57 pela GenScript. Aparentemente para alguns elementos de transposição são necessárias sequências internas, subterminais às TIRs, para a transposição. Por exemplo, para o elemento *To12* foram estabelecidos experimentalmente os tamanhos de 200 pb e 150 pb para as extremidades 5' e 3' respectivamente (URASAKI; MORVAN; KAWAKAMI, 2006). No caso do elemento *Ac* de milho esses valores são 238 pb e 209 pb (COUPLAND et al., 1989) e do elemento *Tag1* de *Arabidopsis*, 98 pb e 109 pb (LIU et al., 2001). Essa necessidade de regiões subterminais e sua extensão mínima para os transposons candidatos somente podem ser estabelecidas com testes envolvendo várias construções, o que não seria factível com o tempo de desenvolvimento desse projeto. Assim, foi escolhido utilizar um tamanho padrão de sequência de 200 pb do início e final dos elementos que incluem as TIR, e chamaremos por conveniência essas regiões terminais de TIRs

Iniciamos as clonagens com as TIRs 3' digerindo o vetor com as enzimas *EcoRV* e *NheI*. Como o vetor *donor* apresenta dois sítios de restrição para a enzima *NheI*, dois fragmentos, de 4592 pb e de 2331 pb referentes ao vetor, devem ser ligados a nova TIR 3' (Fig 28). Como essa etapa da clonagem apresentou vários problemas na ligação dos fragmentos, foram testadas diferentes etapas do processo como: quantidades e proporções de vetor e inserto, quantidade de ligase bem como diferentes fabricantes e diferentes cepas de bactérias no processo de transformação. Apesar dos esforços não obtivemos sucesso na ligação.

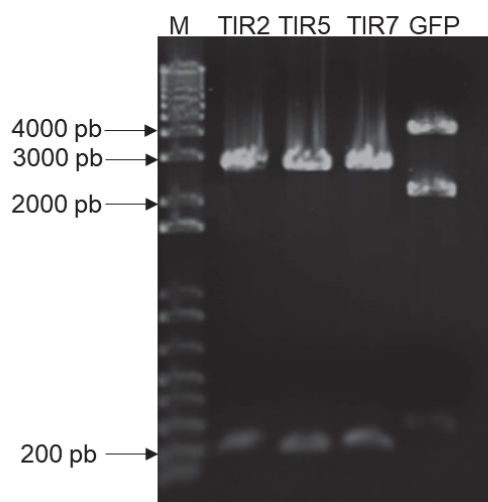


Figura 278. Gel de agarose 1% dos fragmentos das TIRs 3' no vetor pUC57 e do vetor *donor* após a digestão com *EcoRV* e *NheI*. Legenda: M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific); TIR2, TIR5 e TIR7= 200pb no vetor pUC57; GFP: Vetor *donor* (4592pb, 2331pb e 317pb).

A fim de contornar esse problema, desenhamos primers, contendo os sítios de restrição para as enzimas *EcoRI* (final da sequência promotora) e *SacII*, para o vetor *donor* com o objetivo de eliminar o segundo sítio de ligação da enzima *NheI* presente no final da sequência do promotor. A princípio, a região contendo o promotor seria amplificada eliminando o sítio da enzima *NheI* e, em seguida, clonada ao vetor *donor* com as enzimas *EcoRI* e *SacII*. Foram testados diferentes temperaturas e tempos de anelamento, diferentes enzimas Taq polimerase (Accuprime Pfx -Thermo Fisher Scientific; Kit Nat com DNA polimerase - IBMP). No entanto, não conseguimos obter uma banda específica para esta região.

Para a clonagem da região TIR 5' utilizamos primers específicos para amplificar essa região. Após a correta amplificação das TIRs 5' (Fig 29) e purificação, as mesmas foram digeridas com as enzimas *PvuII* e *BglII* e ligadas ao vetor *donor*. A confirmação da inserção da nova TIR 5' no vetor *donor* foi realizada por PCR de colônia com os mesmos primers utilizados na amplificação e as amostras foram enviadas para o sequenciamento. Após as análises do sequenciamento, observou-se que as TIR 5' não estavam presentes no local correto nos plasmídeos indicando que houve algum rearranjo no vetor, talvez facilitado pela presença de sequências repetidas no vetor.

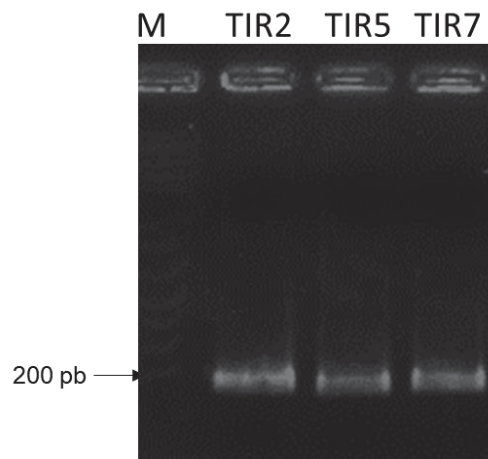


Figura 29. Visualização da eletroforese em gel de agarose 3% dos fragmentos das TIRs 5' amplificados por PCR. Legenda: M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific), TIR2, TIR5 e TIR7= 200pb.

Diante de todos os problemas encontrados nas clonagens das TIRs 3' e 5' das novas transposases, decidimos sintetizar toda a região do vetor *donor*, de um ponto *upstream* da TIR 5' até um ponto *downstream* da TIR 3', trocando as sequências das TIRs do SB pelas do *Tp2*. Além disso, foi retirado o segundo sítio de clivagem da *NheI* e inserido um novo sítio de clivagem para a enzima *BamHI* no local, para a clonagem da TIR 3' (Fig 30). A sequência foi sintetizada pela GenScript.

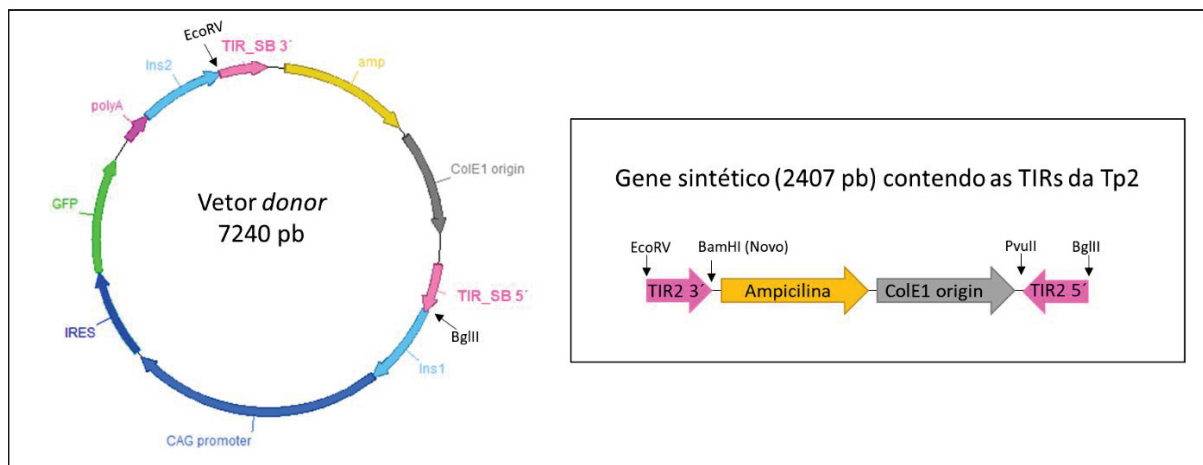
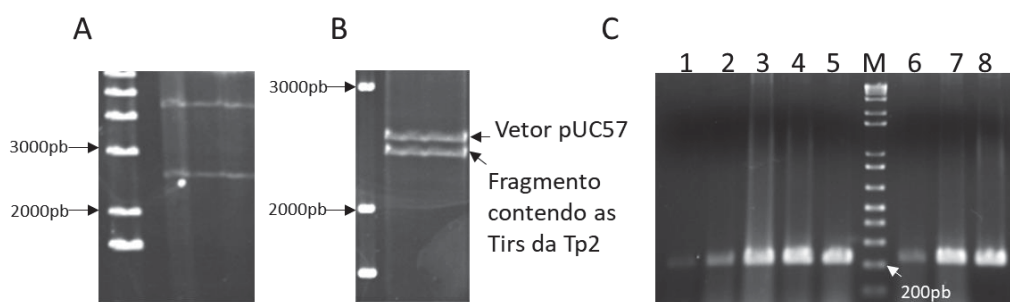


Figura 280. Esquema do vetor *donor* e do gene sintético contendo as TIRs da *Tp2* (tamanho total 2407 pb) com as enzimas *EcoRV* e *BglII* nas extremidades para a clonagem. No gene sintético foi inserido o sítio da enzima *BamHI* eliminando o sítio da enzima *NheI* para a clonagem da TIR 3'.

O vetor contendo a sequência sintética com as TIRs 5' e 3' do *Tp2* bem como o vetor *donor* SB foram digeridos com as enzimas *EcoRV* e *BglII* e purificados por

excisão do gel de agarose (Fig 31 A e B). Após a clonagem, a inserção do cassete contendo as TIRs do *Tp2* foi confirmada por PCR de colônia com os primers específicos da TIR5' 2, por serem os únicos primers específicos disponíveis no momento. A amplificação do fragmento apresentou o tamanho esperado de 200 pb para todos os clones testados, como mostra a figura 31 C.



Para confirmar a inserção correta do cassete no vetor *donor*, após a purificação dos 8 possíveis clones positivos, foi realizada a digestão com as enzimas *BamHI*, ausente no vetor *donor* original, *PvuII* e *NdeI*. O tamanho esperado das bandas foi 4024 pb, 1989 pb e 1016 pb. Diferentes perfis de bandas foram encontrados para os clones analisados indicando novamente algum tipo de rearranjo no vetor durante o processo de clonagem (Fig 32). Apenas os clones 2 e 5 apresentaram o perfil de bandas esperados, enquanto os demais (1, 3, 4, 6-8) tiveram diferentes tamanhos de bandas.

Com os clones 2 e 5 (Fig 32 A), foi realizada, ainda, PCR com primers do vetor *donor* e da TIR para confirmar a correta inserção no vetor (Fig 32 B). Fragmentos de tamanho esperado foram obtidos em todas as combinações de primers para os dois clones os quais foram enviados para sequenciamento e, após análise das sequências, a correta inserção das TIRs da *Tp2* no vetor *donor* foi confirmada.

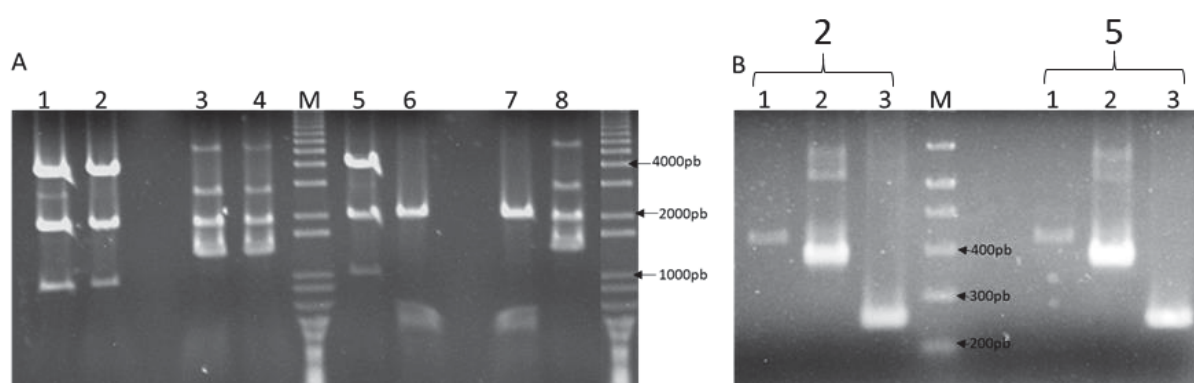


Figura 302. Gel de agarose da confirmação da clonagem do fragmento contendo as TIRs 5'e 3' da *Tp2* no vetor *donor*. **A**: digestão dos possíveis clones positivos (1-8) com as enzimas *BamHI*, *PvuII* e *NdeI* com tamanho esperado das bandas de 4024 pb, 1989 pb, 1016 pb; **B**: PCR dos clones 2 e 5 com diferentes pares de primers; 1: primers pucF1 e TIR5' 2_F (407 pb); 2: pucR1 e TIR3' 2_R (388 pb); 3: TIR5' 2 F e R (204 pb). M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific).

Em seguida, digerimos o vetor *donor* contendo as TIRs da *Tp2* para a clonagem das TIRs da *Tp5* e *Tp7* na região 3' com as enzimas *EcoRV* e *BamHI*. A confirmação da clonagem foi realizada por digestão do vetor (Fig 33) com a enzima *PvuII* para a *Tp5* e *SacI* para a *Tp7*. A TIR 3' da *Tp5* possui um único sítio para a enzima *PvuII* e o vetor *donor* também possui um sítio para essa enzima, dessa maneira a padrão de bandas esperado para a digestão dos possíveis clones foi de 6482 pb e 661 pb (Fig 33 A). A TIR 3' da *Tp7* possui um único sítio para a enzima *SacI* enquanto o vetor *donor* possui dois sítios, dessa maneira a padrão de bandas esperado para a digestão do vetor foi de 4047 pb, 2762 pb e 333 pb (Fig 33 B). Como controle, em ambos os testes, o vetor *donor* original da *SB* também foi digerido com as enzimas citadas acima. Somente a TIR 3' da *Tp7* apresentou clones positivos representado pelos números 5, 6 e 7 na figura 33 B, os quais foram confirmados por análise do sequenciamento.

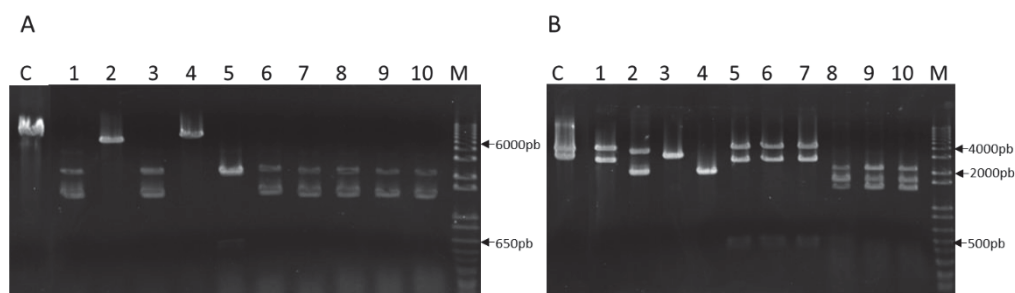


Figura 313. Gel de agarose da confirmação da clonagem das TIRs 3' da *Tp5* e *Tp7* no vetor donor. **A:** Digestão dos clones com *PvuII* para confirmar a inserção da TIR3' da *Tp5* com tamanho de bandas esperados de 6482 pb e 661 pb; C: vetor *donor* digerido com *PvuII*. **B:** Digestão dos clones com *SacI* para confirmar a inserção da TIR3' da *Tp7* com tamanho de bandas esperados de 4047 pb, 2762 pb e 333 pb. C: vetor *donor* digerido com *SacI*. M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific).

A seguir, digerimos o vetor *donor* para a clonagem da TIR 5' com as enzimas *BglII* e *PvuII*. A seleção das possíveis colônias contendo a TIR 5' das duas transposases foi realizada por PCR de colônia com um primer do vetor *donor* e um outro específico da seguinte maneira: PucR1 e TIR_5'R (Fig 34 A). Apenas uma colônia para a TIR 5' da *Tp5* foi amplificada e nenhum clone positivo foi encontrado para a *Tp7* (dados não mostrados). Como não havia nenhuma enzima de restrição com sítio presente apenas na sequência da TIR 5', optamos por realizar a confirmação da inserção no vetor *donor* por PCR utilizando diferentes primers juntamente com o vetor *donor* da *TP2* como controle (Fig 34 B). O primeiro par de primer anela na TIR 5' da *Tp2* e foi amplificado, como esperado, apenas no vetor *donor* da *Tp2*, indicando que essa região foi retirada corretamente do vetor. O segundo par de primers (TIR5_5' R e F) é específico da TIR 5' da *Tp5* mas também foi amplificado, inesperadamente, no vetor *donor* contendo a TIR da *Tp2*. Como estávamos tendo esse problema de especificidade utilizamos mais 2 pares de primers, um específico de TIR 5' da *Tp5* e um do vetor *donor* (linhas 3 e 4 da Fig 34 B). Além disso, realizamos a digestão do vetor com as enzimas *EcoRV*, *BamHI* e *PvuII* juntamente com o vetor *donor* da *Tp2*, e verificamos que aparentemente não houve nenhuma recombinação no vetor ao inserir a TIR 5' da *Tp5* (Fig 34 C). A correta inserção foi confirmada por análise de sequenciamento.

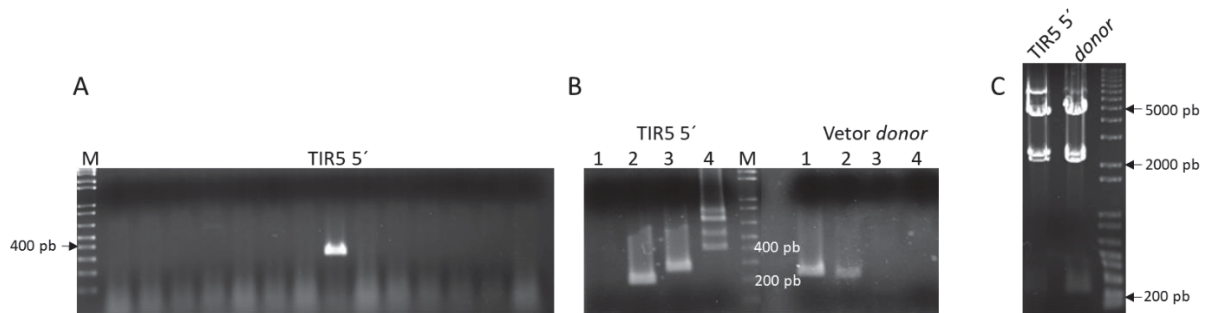


Figura 324. Gel de agarose da confirmação da clonagem das TIRs 5' da *Tp5* no vetor donor. **A**: PCR de colônia da TIR5 5' com primers TIR5_5' R e PucF1 (388 pb); **B**: PCR do clone positivo para a clonagem da TIR 5' do *Tp5* e vetor *donor* Tp2 como controle com os primers: 1. TIR2_5' F e R (200 pb); 2. TIR5_5' F e R (200 pb); 3. TIR5_5' R e PucF1 (388 pb); 4. TIR5_5' F e Prom (920 pb); **C**: digestão dos clones positivos com as enzimas *EcoRV*, *BamHI* e *PvuII* e com tamanhos esperados de 200 pb, 1989 pb e 5000 pb. M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific).

Apesar dos esforços, não conseguimos até o momento clonar a TIR5' da *Tp7*. Assim até o momento temos clonados as duas TIRs da *Tp2* no vetor *donor*, a TIR3' da *Tp7* e a TIR5' da *Tp5*. Para finalizar as clonagens e iniciar os testes de transposição ainda faltam clonar as sequências 5' da *Tp7* e a 3' da *Tp5*.

Para os ensaios de transfecção utilizamos um controle negativo caracterizado pelo vetor *helper* sem transposase que foi obtido pela digestão do vetor *helper* SB com uma enzima que libera a transposase, e religação do vetor. Essa construção foi confirmada por PCR com os iniciadores *IntronSV*, 86 pb antes da *SB*, e *polyA* que anela 113 pb após a transposase (Fig 35), cujo tamanho esperado de amplificação sem a *SB* foi de 199 pb. A correta eliminação da transposase também foi confirmada pela linearização do vetor após digestão com mesma enzima usada para liberação do fragmento durante a clonagem (dados não mostrados).

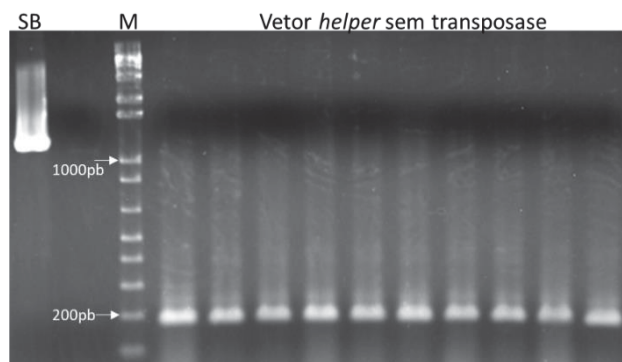


Figura 35. Gel de agarose da confirmação da retirada da transposase do vetor *helper*. PCR com iniciador *IntronSV* e *polyA*, tamanho esperado com a *SB*: 1253 pb, tamanho esperado sem transposase:

199 pb. SB: vetor *helper* contendo a SB, M: marcador de peso molecular (1 Kb plus – Thermo Fisher Scientific).

4.1.2 Transfecção em célula CHO para testar o novo sistema *Tp2*

Com os vetores *helper* e *donor* do transposon *Tp2* prontos, iniciamos o processo de transfecção.

As transfecções foram realizadas em duplicata com os vetores *helper* contendo a SB (controle positivo), o vetor *helper* sem a transposase (controle negativo) e vetor *helper* contendo a *Tp2*, juntamente com os vetores *donor* contendo suas respectivas TIRs (Quadro 4). Quarenta e oito horas após a transfecção, foi realizada a separação das células contendo a marcação verde do gene GFP por *cell sorting* utilizando o equipamento de citometria de fluxo S3e Cell Sorter (Bio-Rad).

Quadro 4. Esquema dos vetores *helper* e *donor* utilizados na transfecção.

Vetor <i>helper</i>		Vetor <i>donor</i>	
SB	4732 pb	TIRs da SB e GFP	7240 pb
Controle sem SB	3709 pb	TIRs da SB e GFP	7240 pb
<i>Tp2</i>	4737 pb	TIRs da <i>Tp2</i> e GFP	7008 pb
Controle sem <i>Tp2</i>	3709 pb	TIRs da <i>Tp2</i> e GFP	7008 pb

Para o primeiro *cell sorting* é esperado uma taxa de fluorescência mais elevada (40-60%) devido a presença de DNA plasmidial derivado da transfecção.

A porcentagem de células fluorescentes no primeiro *cell sorting* foi mais baixo do que o esperado, o que pode ter sido um problema pontual, pois as células ficaram inesperadamente sem agitação. A representação gráfica de uma réplica de cada experimento de citometria está exemplificada na figura 36. As amostras apresentaram as seguintes porcentagens de fluorescência: SB11.1: 1,18%; SB11.2: 0,75%; *Tp2*.1: 20%; *Tp2*.2: 16%; controle SB11.1: 6,92%; controle SB11.2: 8,78%; controle *Tp2*.1: 17,7%; controle *Tp2*.2: 20,3%.

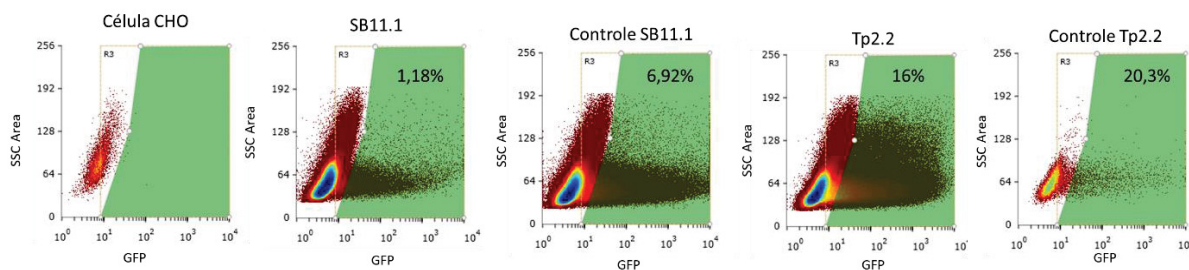


Figura 336. Representação gráfica do resultado da citometria de fluxo 48 h após a transfecção em célula CHO dos vetores do Sistema ingresso contendo *SB* e *Tp2*. Distribuição em dot plot das células CHO pela fluorescência de GFP e granulosidade (SSC), as células fluorescentes selecionadas para o *cell sorting* estão representadas no gate em verde. Para as células CHO e o controle *Tp2.2* foram utilizados menor quantidade de células para esta representação gráfica.

Após a primeira seleção das células fluorescentes, as células foram cultivadas para a perda dos plasmídeos episomais restando apenas o cassete integrado no genoma da célula hospedeira. Para a segunda separação das células é esperada uma porcentagem mais baixa de células fluorescentes em relação a primeira seleção devido justamente a essa perda do plasmídeo episomal. O objetivo de adicionar o controle negativo foi para comparar a fluorescência, pois na ausência da transposase, é esperado que os cassetes não sejam integrados e após algumas passagens, terá perda total da fluorescência devido a eliminação dos plasmídeos ao longo das divisões celulares da célula CHO. Como esperado, na segunda etapa de *cell sorting* (Fig 37) a porcentagem de células fluorescentes foi mais baixa que a primeira etapa para as amostras controle e *Tp2*. A média de células fluorescentes foi de 12,1% para as células transfectadas com *SB*, 2,1% para as células transfectadas com *Tp2* e 3,5% para as células controles transfectadas apenas com o vetor *donor*.

Após a seleção das células expressando GFP é possível verificar um elevado número de células contendo o marcador, como demonstrado para os dias 7 e 15 após o 1º e 2º *cell sorting*, respectivamente (Fig 37). O controle 1 da *Tp2* precisou ser descartado após o segundo *cell sorting* devido a contaminação da cultura.

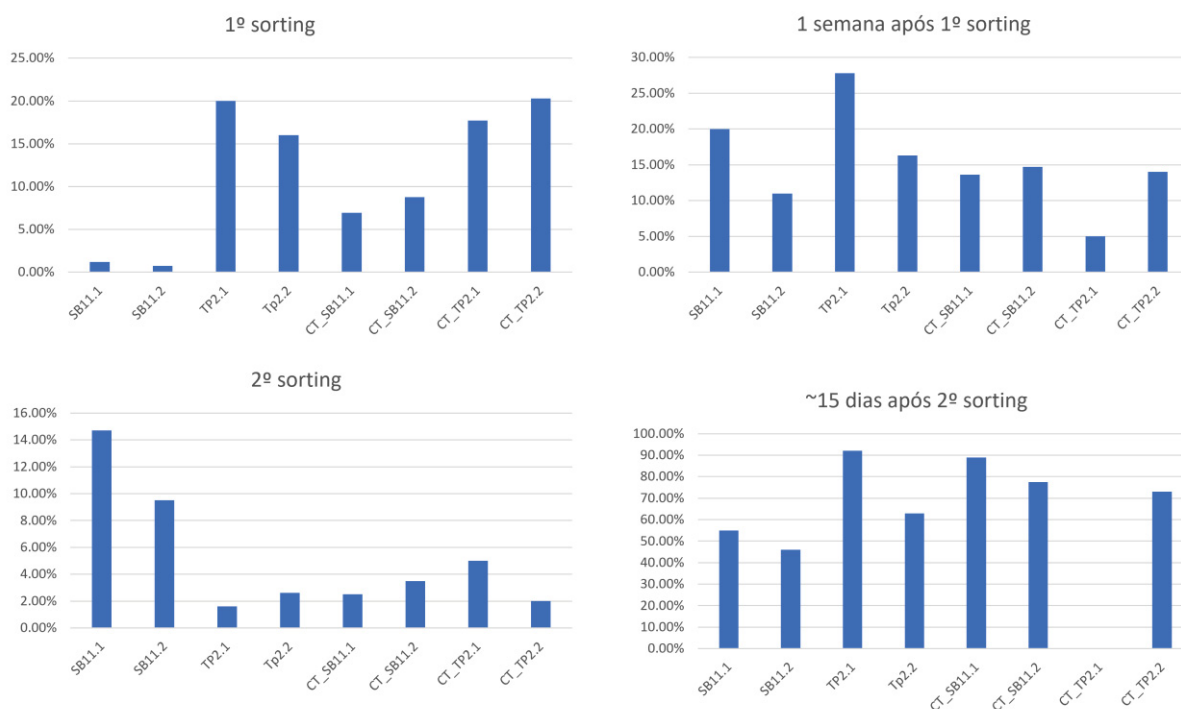


Figura 347. Dados das porcentagens de GFP obtidos na citometria de fluxo durante a seleção das células transfectantes. Amostras transfectadas com SB (SB11.1, SB11.2), amostras transfectadas com *Tp2* (TP2.1, TP2.2), controle SB (CT_SB11.1 e CT_SB11.2) e controles *Tp2*.

As culturas continuaram sendo mantidas com repiques semanais e eventuais seleções por *cell sorting* para o processo de enriquecimento das células contendo o cassete integrado nas amostras transfectadas com as transposases e esperando a eliminação do sinal de fluorescência dos controles. As seleções foram realizadas em dias distintos dependendo da situação de cada uma das culturas. Ao todo, cada cultura passou por aproximadamente 20 passagens desde a transfecção. A média de fluorescência de todas as culturas permaneceu alta (acima de 40%) até quase 60 dias após a primeira seleção das células.

A figura 38 mostra o nível de fluorescência das populações após cerca de 60 dias da transfecção e múltiplas seleções (SB11.1 – 4º *sorting*; SB11.2 - 3º *sorting*; TP2.2 – 3º *sorting*; CT_SB11.1 – 2º *sorting* e CT_TP2.2 – 2º *sorting*). Apenas neste momento foi possível verificar uma diminuição considerável da fluorescência do controle da SB, enquanto o controle da *Tp2* continuou apresentando um nível de fluorescência comparável ao teste (Fig 38).

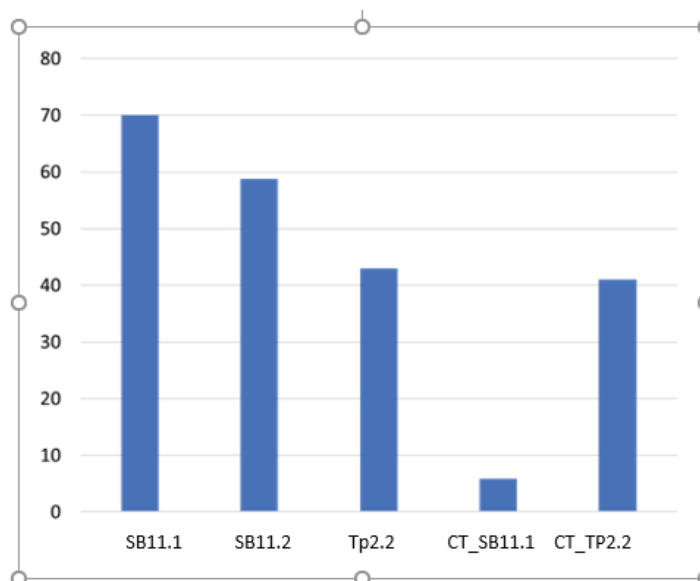


Figura 358. Dados das porcentagens de GFP obtidos na citometria de fluxo das populações 60 dias após a primeira seleção de células. Amostras transfectadas com SB (SB11.1, SB11.2), amostra transfectadas com *Tp2* (Tp2.2), controle SB (CT_SB11.1) e controle *Tp2* (CT_TP2.2).

Ao longo do enriquecimento das células fluorescentes, foi possível observar que as culturas controles apresentaram maior dificuldade para recuperação das células e, dessa maneira, passaram por um menor número de seleções até o momento. No entanto, seria esperado uma proporção mais baixa de células fluorescentes nessas culturas devido à perda dos vetores episomais. Essa persistência de células fluorescentes nas amostras controle pode indicar a presença de inserções aleatórias. É possível que um pequeno número de inserções aleatórias tenha sido selecionado no primeiro *round* de seleção e foram assim enriquecidas. Essa observação inviabiliza nosso objetivo inicial de evidenciar a atividade da nova transposase *Tp2* pelo nível de fluorescência em comparação com os controles. Por outro lado, na população dos testes realizados com a *Tp2* é possível observar uma subpopulação (deslocada para a direita) com nível de fluorescência bem maior do que o encontrado para o controle (Fig 39). Essa subpopulação de células fluorescentes pode indicar um maior número de cópias do cassete integrado no genoma das células, o que seria improvável de ser explicado apenas por inserções aleatórias. Alternativamente, nessa subpopulação, o cassete pode estar integrado de forma aleatória em uma região que propicia uma alta expressão.

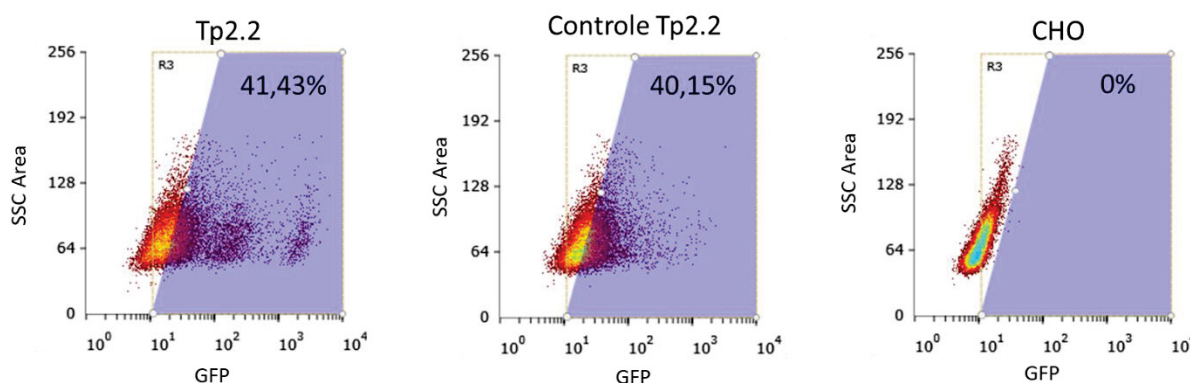


Figura 39. Representação gráfica do resultado da citometria de fluxo de células CHO transfectadas com os vetores do Sistema ingresso e a *Tp2*, 60 dias após o segundo *cell sorting*. Distribuição em dot plot das células CHO pela fluorescência de GFP e granulosidade (SSC), as células fluorescentes estão representadas no gate em azul.

Devido a uma contaminação geral das culturas do laboratório com levedura, todas as culturas foram descartadas, não tendo sido possível avaliar nesse momento a manutenção das fluorescências por mais tempo. Como perspectivas, as culturas estoques que foram guardadas ao longo dos experimentos serão descongeladas para a continuação das observações.

4.1.3 Atividade da transposase *SB*

Para verificar a atividade da transposase *SB*, amostras de proteína das culturas 48 horas após a transfecção foram submetidas a um ensaio de *Western Blot* utilizando anticorpos específicos para a *SB* (Fig 40). Dessa maneira, foi possível verificar a expressão da proteína *SB* apenas nas amostras que foram transfectadas com o vetor contendo a *SB* e ausência da proteína, conforme esperado, nos controles. Como não há disponível anticorpos contra a *Tp2* não foi possível verificar a expressão dessa proteína nas amostras transfectadas com o vetor contendo a *Tp2*.

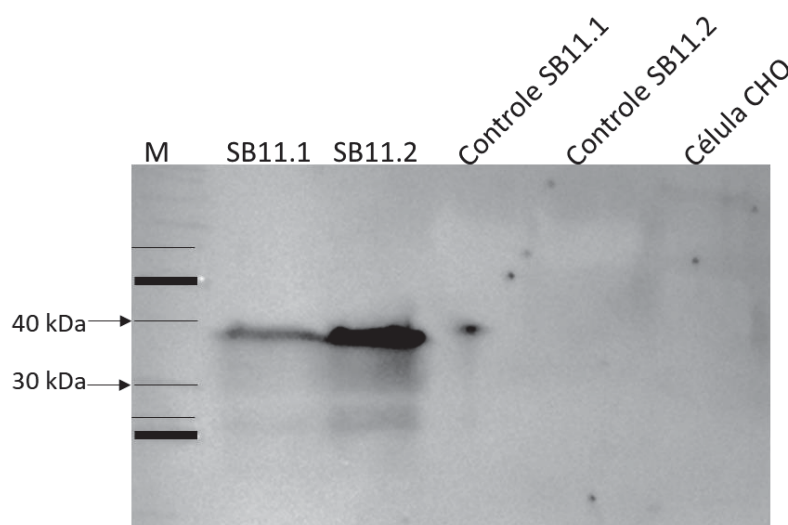


Figura 360. *Western Blot* das culturas 48 horas após a transfecção para verificar a expressão da *SB*. SB11.1, SB11.2 – duplicatas das culturas testes *SB*; Controle SB11.1, Controle SB11.2 – duplicatas dos controles *SB* (vetor *helper* sem transposase).

4.1.4 Local de integração genômica da *Tp2*

Com a população selecionada, após o segundo *cell sorting*, foi realizada a PCR splinkerette nos testes com a *Tp2* para verificar o local de inserção do transgene bem como a atividade da transposase. Infelizmente, as duplicatas *Tp2.1* e *CT_Tp2.1* (controle) foram perdidas ao longo dos experimentos e não foram analisadas. A segunda PCR da PCR splinkerette não mostrou um padrão de bandas em escada conforme o esperado (Fig 41), podendo indicar que alguma etapa na técnica não foi eficiente. Devido ao curto prazo para a finalização da tese, não foi possível repetir o experimento e assim as amostras *Tp2.2* e o controle *CT_Tp2.2* foram enviadas para o sequenciamento mesmo sem o padrão esperado.

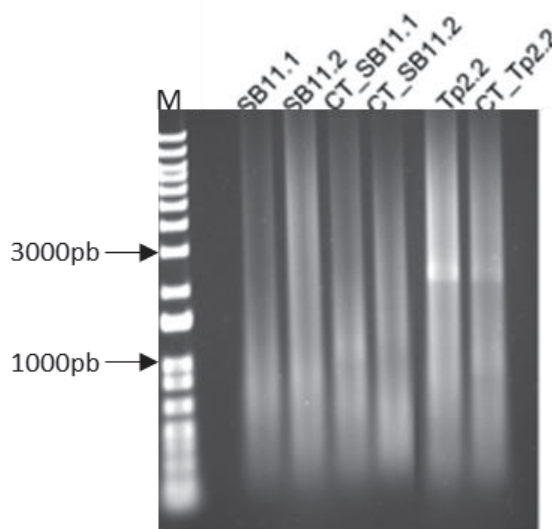


Figura 371. Gel de agarose da segunda PCR splinkerette das amostras transfectadas com *SB* e *Tp2*.

Os resultados do sequenciamento foram compilados na tabela 9 e indicam que a maior parte das leituras para as duas amostras devem ser espúrias e nenhuma leitura da população de *Tp2.2* teve o padrão correto com a etiqueta da TIR 5'. Para a amostra controle já era esperado que apenas leituras espúrias fossem geradas. Essa amostra foi sequenciada exatamente para servir de controle negativo, mostrando o padrão esperado para ausência de inserção mediada por transposição.

Tabela 9. Dados gerais do sequenciamento por Illumina dos fragmentos gerados por PCR splinkerette para *Tp2*. Antes e após limpeza das sequências e filtragens de leituras corretas.

Amostras	Total de leituras (R1 e R2)	Leituras pareados após Trimmomatic (R1 e R2) > 220 nucleotídeos	Leituras corretas (R1 – blastn CHO e R2 – blastn TIR)	Leituras R2 - tag do SB com TSD	Leituras R2 -tag do SB sem TSD	Número de possíveis inserções independentes (após filtragem)
TP2 Pop	96011	22123	101	0	0	0
CT TP2 Pop	210324	13844	204	0	0	0

Algumas leituras que passaram na filtragem de possíveis leituras corretas, por terem dado *hits* significativos com CHO (R1) e *Tp2* (R2), foram avaliadas em detalhes e parecem ser oriundas de inserções aleatórias, com inserções próximas ao final da

TIR 5' (Fig 42). Nesse exemplo, a leitura alinha com a TIR 5', da posição 3 a 86, mas não contém os nucleotídeos finais. Essa mesma leitura alinha com o genoma de CHO a partir do nucleotídeo 79, mostrando então uma região de sobreposição de 8 pb que pode ter sido usada para recombinação e inserção.

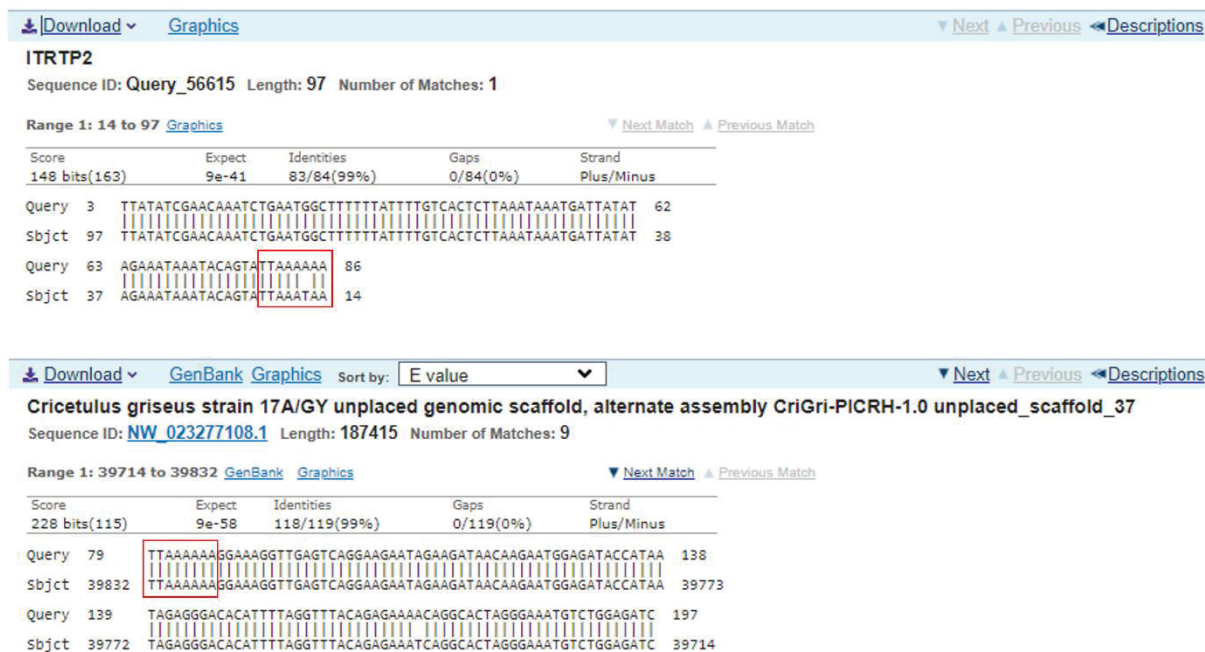


Figura 382. Resultado do blastn de uma leitura R2 contra a TIR da TP2 (blast-2-sequence) e contra o genoma da CHO. Em vermelho está destacada a região de microhomologia que deve ter sido usada para a inserção.

Para analisar a perda do vetor *helper* e *donor* episomais nas amostras após o segundo *sorting*, foi aplicado um protocolo de extração de DNA de baixo peso molecular, o qual recupera possíveis vetores que ainda estejam nas células. O DNA foi transformado em bactérias e 2 colônias foram recuperadas para uma das culturas teste de *SB* e 1 colônia para a cultura teste de *Tp2*. Isso indica a presença de plasmídeos contendo o gene de resistência a ampicilina. Os plasmídeos foram extraídos e mostraram um padrão de bandas após digestões distinto tanto dos vetores *helper* quanto dos vetores *donor* (dados não mostrados). Esses plasmídeos poderão ser sequenciados para entender a sua composição e os rearranjos que aconteceram. É possível que esses plasmídeos não tenham um grande impacto nas culturas, pois devem estar presentes em poucas células.

Da mesma forma como foi feito para os dados de sequenciamento dos clones de CHO transfectados com o Sistema Ingresso 1.0, as leituras oriundas da cultura

TP2 foram alinhadas por blastn contra o vetor *helper* de *Tp2* e mostrou hits significativos, sugerindo que pelo menos algumas células dessa população tenham o vetor *helper* inserido.

Diante do exposto, com os dados que foram obtidos até o momento, não podemos afirmar se a *Tp2* tem ou não atividade de transposição. É possível que a PCR splinkerette não tenha funcionado corretamente. Também, pode ter ocorrido inibição por superprodução, onde as transposases competem pelo mesmo sítio de ligação (TIRs) diminuindo sua atividade, já que uma grande quantidade de vetor *helper* foi utilizado para a transfecção. Como já mencionado, os dados do ensaio de citometria mostram uma subpopulação com expressão muito significativa de GFP na população teste de *Tp2* o que seria improvável de ser explicada apenas por inserções aleatórias. Dessa forma, como perspectivas, pretendemos selecionar as subpopulações para análise, realizar o isolamento de células com maior expressão para estabelecimento de linhagens clonais e repetir o experimento de PCR splinkerette e sequenciamento Illumina.

CAPÍTULO II

Distribution of *Merlin* in eukaryotes and first report of DNA transposons in kinetoplastid protists

Artigo publicado na revista PLoS One em 06 de maio de 2020.

Acesso ao artigo original:

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251133>

Citação: Lopes ALK, Kriegová E, Lukeš J, Krieger MA, Ludwig A. Distribution of *Merlin* in eukaryotes and first report of DNA transposons in kinetoplastid protists. PLoS One. 2021 May 6;16(5):e0251133. doi: 10.1371/journal.pone.0251133. PMID: 33956864; PMCID: PMC8101967.

Distribution of *Merlin* in eukaryotes and first report of DNA transposons in kinetoplastid protists

Ana Luisa Kalb Lopes^{1,2,3}, Eva Kriegová^{4,#}, Julius Lukeš^{4,5}, Marco Aurélio Krieger^{1,2}, Adriana Ludwig^{2,*}

¹Pós-Graduação em Biologia Celular e Molecular, Universidade Federal do Paraná, Curitiba, PR, Brazil

²Laboratório de Ciências e Tecnologias Aplicadas em Saúde (LaCTAS), Instituto Carlos Chagas, Fundação Oswaldo Cruz (Fiocruz), Curitiba, PR, Brazil

³Instituto de Biologia Molecular do Paraná, Curitiba, PR, Brazil

⁴Institute of Parasitology, Biology Center, Czech Academy of Sciences, České Budějovice (Budweis), Czech Republic

⁵Faculty of Sciences, University of South Bohemia, České Budějovice (Budweis), Czech Republic

#Present address: Institute of Entomology, Biology Centre, Czech Academy of Sciences, Ceske Budejovice, Czech Republic

* Corresponding author:

Email - adriludwig@gmail.com

ABSTRACT

DNA transposons are defined as repeated DNA sequences that can move within the host genome through the action of transposases. The transposon superfamily *Merlin* was originally found mainly in animal genomes. Here, we describe a global distribution of the *Merlin* in animals, fungi, plants and protists, reporting for the first time their presence in Rhodophyceae, Metamonada, Discoba and Alveolata. We identified a great variety of potentially active *Merlin* families, some containing highly imperfect terminal inverted repeats and internal tandem repeats. *Merlin*-related sequences with no evidence of mobilization capacity were also observed and may be products of domestication. The evolutionary trees support that *Merlin* is likely an ancient superfamily, with early events of diversification and secondary losses, although repeated re-invasions probably occurred in some groups, which would explain its diversity and discontinuous distribution. We cannot rule out the possibility that the *Merlin* superfamily is the product of multiple horizontal transfers of related prokaryotic insertion sequences. Moreover, this is the first account of a DNA transposon in kinetoplastid flagellates, with conserved *Merlin* transposase identified in *Bodo saltans* and *Perkinsella* sp., whereas it is absent in trypanosomatids. Based on the level of conservation of the transposase and overlaps of putative open reading frames with *Merlin*, we propose that in protists it may serve as a raw material for gene emergence.

INTRODUCTION

Transposable elements (TEs) are defined as repeated DNA sequences that can move within the host genome. TEs are not only present in both prokaryotes and eukaryotes, but they also constitute a significant fraction of numerous genomes, including those of humans [1] and plants [2,3]. TEs produce various genetic alterations that play a central role in the structural organization and plasticity of genomes [4]. Their insertions have the potential to inactivate or alter the expression of genes or gene regulatory elements. By ectopic recombination, TEs may trigger chromosomal rearrangements and contribute to mutagenesis [5–7]. Moreover, TEs are rich in coding

and regulatory sequences that can be co-opted by the host to develop novel cellular functions in a process called domestication or exaptation [8–10].

TEs exhibit a broad diversity in their structure and transposition mechanisms. A unified classification system for eukaryotic TEs was proposed, establishing two classes according to their transposition mechanisms, structures, and sequence similarity. Class I elements or retrotransposons move by a copy-paste mechanism that involves the reverse transcription of an RNA intermediate and insertion of its cDNA copy at a new site in the genome. Class II elements, or DNA transposons, move through a DNA intermediate [6,11]. The classical DNA transposons consist of a transposase gene that is flanked by two terminal inverted repeats (TIRs) forming the so-called TIR order [6]. During transposition, the transposase enzyme recognizes the TIRs and performs the excision of the transposon by double-strand DNA breaks, which is followed by the insertion into a new genomic location. Upon insertion, the target DNA site is duplicated, resulting in target site duplications (TSDs) [12]. The first 9 recognized TIR superfamilies [6] are distinguished by their sequences and the TSD size (*Tc1/mariner*, *PIF/Harbinger*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *piggyBac* and *CACTA*). Currently, in Repbase [13], a database of eukaryotic TEs and repetitive sequences, 18 superfamilies encoding a D-D-D/E-type transposase are distinguished.

The *Merlin* superfamily was first described by Feschotte in 2004 [14] as a group of elements detected by computational analysis in a wide range of genomes, which share common structural features and sequence motifs. These elements possess TIRs that range in length from 24 to 462 base pairs (bp) with conserved terminal 5'-GG-3', and are flanked by 8-bp or 9-bp TSDs. *Merlin* elements were found to be related to the IS1016 group of bacterial insertion sequence (IS), sharing sequence similarity in the C-terminal halves of the proteins and the TSD size [14]. IS1016 together with IS*Pna2*, IS*H4*, IS1595, IS*Sod11*, IS*Nwi* and IS*Nha5* constitute a major group so-called the IS1595 family [15]. Members of this IS family are usually flanked by 8-bp TSDs, have a single transposase gene, and, except for the IS1016 group, all others have an N-terminal zinc finger domain [15] named Zn_Tnp_IS1595. They also have a conserved C-terminal domain containing the DDE catalytic motif shared with the eukaryotic *Merlin* transposons [14,15] named DDE_Tnp_IS1595. All *Merlin* and IS1016 elements described so far lack the Zn_Tnp_IS1595 domain.

Elements from superfamily *Merlin* were identified in the genome of several animals, including nematodes, flatworms, mosquitos, ascidians, zebrafish, frogs, and

its relics are also present in humans [14]. Outside animals, the *Merlin* superfamily was described from the oomycete *Phytophthora sojae* (Stramenopila), the microsporidian *Nosema bombycis* (Microspora) [14], fungi and embryophytes [16]. Moreover, in the microsporidian *Anncaliia algerae*, *Merlin* represents the most abundant sequence element and seems to be involved in horizontal transfer events [17].

Within the last decade, the number of newly sequenced genomes is accelerating, providing data for an increased rate of identification of TEs. Here, we have performed bioinformatic analyses and updated our knowledge of the distribution of *Merlin* across eukaryotes, documenting for the first time its presence in Alveolata, Rhodophyceae, Metamonada and Discoba. This work also comprises the first record of DNA transposons in the well-studied kinetoplastid protists.

MATERIALS AND METHODS

Searches for *Merlin* sequences

It is possible to observe that canonical elements from IS1595 family (data not shown) and eukaryotic *Merlin* elements (S1 Table) have the DDE_Tnp_IS1595 domain (NCBI CDD accession: cl01481), which is on average 130 amino acids (aa) long, being a diagnostic feature.

Merlin sequences deposited in the Repbase database (<http://www.girinst.org/repbase/>) [13] were analyzed for the conservation of the DDE_Tnp_IS1595 domain using the NCBI CD-search tool [18]. Among them, *Merlin1_SM* from *Schmidtea mediterranea* was used in the initial blastp searches from NCBI server [19] against distinct eukaryotic groups. Sequences presenting a hit with an e-value lower than 1e-04 were retrieved and evaluated for the presence of the conserved domain. The region corresponding to the DDE_Tnp_IS1595 domain was extracted from all sequences retrieved from Repbase and blastp searches (around 300 sequences). Next, these sequences were used as query sequences in the second round of blastp searches (e-value cutoff of 1e-03) against each of the higher eukaryotic ranks, according to the classification of Adl et al. (2019). All retrieved sequences were analyzed in the CD-search for the presence of expected and other domains. To expand searches to taxa that did not produce significant hits in the blastp but have their genome sequence available, we performed online or local tblastn searches using the

same e-value cutoff. Possible significant hits were confirmed by extracting the sequences from the genomes (using an in-house Python code) and searching for open reading frames (ORFs) (using NCBI ORFfinder tool) and the DDE_Tnp_IS1595 domain by CD-search. Most blast searches were completed until November 2019.

To check for the possibility of bacterial sequence contamination, the retrieved proteins were used as a query on blastp against the nr NCBI database. Hits with more than 80% identity with a bacterial sequence were considered possible contaminations.

Structural analysis of Merlin in Rhodophyceae, Metamonada, Discoba and Alveolata

Since significant hits were found by blastp in several alveolates, all genomes available for this group were analyzed by tblastn to recover possible complete elements. For the analysis of significant tblastn hits from Rhodophyceae, Metamonada, Discoba and Alveolata, we prioritized the curated analysis to genomes where copies are more conserved and located in long contigs/scaffolds. Sequences retrieved from the same species were aligned with MAFFT [20], and the alignments were inspected to identify the limits of the copies. To verify the presence of the TIRs and TSDs, most conserved copies were analyzed by blastn with the parameters “align two or more sequences” and “somewhat similar sequences (blastn)” using the same sequence as a query and subject, and by visual inspection of alignments. Similarity with the eukaryotic TEs was checked using CENSOR tool from Repbase [21]. To evaluate divergence among copies in some species, protein and nucleotide sequences were aligned using MAFFT, and Mega X [22] was used to cluster the sequences using neighbor-joining (p-distance, 1000 replicate bootstrap) and to access the pairwise p-distance among them. The online Tandem Repeats Finder program [23] was used to help identify tandem repeats inside the *Merlin* copies. The secondary structure of *Merlin* from the kinetoplastids *Bodo saltans* and *Perkinsela* sp. was analyzed by Phyre2 [24]. All sequences analyzed are available in S1, S2 and S3 Appendices.

PCR confirmation of Merlin in Perkinsela sp.

The DNA of *Perkinsela* sp. (strain CCAP1560/4) was submitted to gradient PCR analysis. Primers were designed to anneal inside the transposon and the up- or downstream genes. Single-copy genes and those that are present across the

kinetoplastid flagellates were chosen. More information about the primers and the amplification conditions is available in S1 File.

Evolutionary analyses

The Transposase DDE_Tnp_IS1595 domain was isolated from identified sequences and filtered by size (>115 aa). Sequences were then filtered by identity (70%) using CD-HIT [25] to reduce the number of sequences and produce a less complex matrix. Three sequences from each group of IS1595 family, as classified in IS Finder (<https://www-is.biotoul.fr/index.php>), were added to the matrix (S2 Table). Sequences were aligned using PROMALS3D [26]. The alignment was trimmed using trimAl 1.4.1 [27] (0.4 of GAP threshold; 0.0 of similarity threshold). Additional filtering was done to eliminate some short sequences (<110 aa) from the final matrix. The evolutionary model LG+G was indicated by the test implemented in Mega X [22]. An additional alignment matrix was constructed including only the copies that have been curated by us and the ones from Repbase together with the IS1595 group sequences with no size filtering. For this matrix, the evolutionary model WAG+G was indicated. The trees were inferred by Bayesian Analysis (BA) in MrBayes 3.2.6 [28] with the indicated model and were run in the CIPRES gateway [29]. The Markov Chain Monte Carlo (MCMC) of the BA was run for at least 10,000,000 generations, sampling trees every 1,000 generations, and burning 25% of the initial results. Additionally, a maximum likelihood (ML) analysis with the indicated model was performed under the rapid bootstrap algorithm using RAxML-HPC BlackBox [30] implemented on CIPRES with automatic determination of bootstrap replicates.

RESULTS AND DISCUSSION

Merlin is widely distributed in eukaryotes

As new genomes from a broad range of taxa become available at an ever-increasing pace, it is possible to expand our knowledge about the distribution and evolutionary history of TEs. Here, we describe a broad analysis of the distribution of *Merlin* in eukaryotes. For this, we first performed a search for *Merlin* deposited in Repbase and identified 70 sequences, of which 32 displayed the expected DDE_Tnp_IS1595 domain and the DDE motif, an essential part of the catalytic site of

transposases [31]. In order to carry out a global search for these elements, we performed blastp searches in two rounds, using the sequences retrieved in the first search as multi-queries in a second round of blastp in specific taxonomic groups, increasing the chances to find similar sequences. The sequences were considered positive hits for *Merlin* only if the DDE_Tnp_IS1595 domain was found with no evidence of bacterial contamination.

Our results show that *Merlin* is present in a wide range of animals, fungi, plants and protists and for the first time, it was found in Alveolata, Rhodophyceae, Metamonada and Discoba species (Fig 1; S3 Table). For taxonomic assignments, we have used the recently revised classification of eukaryotes [32].

In Bilateria, few copies per species were found (Chordata: 90 sequences in 31 species; Echinodermata: 8 sequences in 2 species; Platyhelminthes: 58 sequences in 9 species; Protostomia: 1269 sequences in 126 species) (S1 Appendix). However, in a subset of non-closely related species, *Merlin* was more abundant, such as in the flatworm *Clonorchis sinensis* (24 sequences) and several protostomians, namely the octopid *Octopus bimaculoides* (191 sequences), the whipworm *Trichuris suis* (147 sequences), the spider *Araneus ventricosus* (112 sequences), the scorpion *Centruroides sculpturatus* (48 sequences) and few others.

In Cnidaria, 10 species showed significant hits varying from 5 to 23 sequences, except for the myxosporean *Thelohanellus kitauei*, where 185 significant hits were found (S1 Appendix). In fact, Yang et al. found 179 transposon sequences with the DDE_Tnp_IS1595 domain in the genome of this parasite, although they did not classify them as *Merlin* [33]. In Porifera, only 3 sequences were found in *Amphimedon queenslandica* (S1 Appendix).

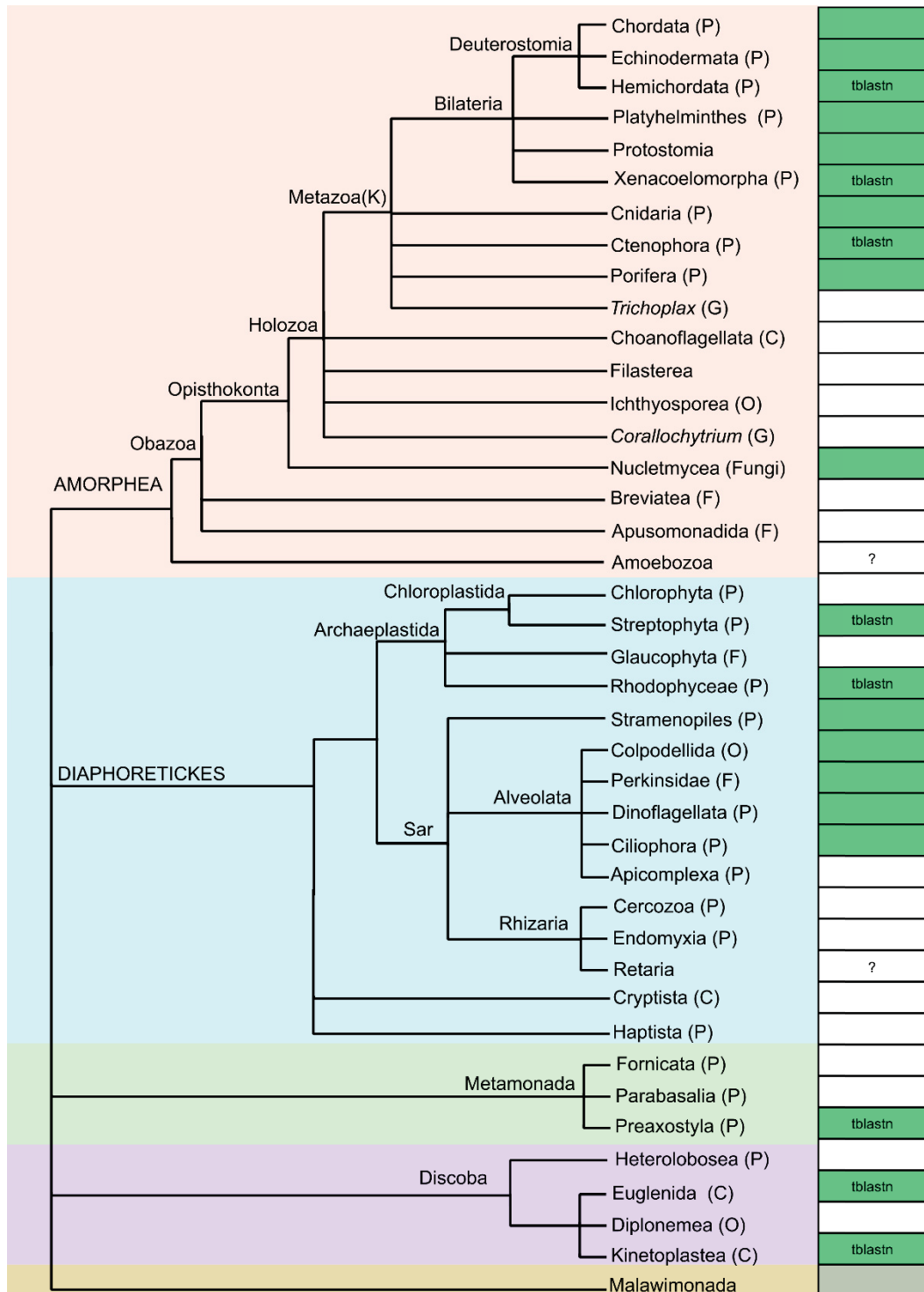


Figure 1: Phylogenetic distribution of the *Merlin* elements in eukaryotes. The cladogram was drawn based on [32], and the subdivisions of Bilateria followed the NCBI Taxonomy. Green boxes indicate the presence of *Merlin* in at least one species per each taxonomic group and those analyzed by tblastn were indicated; gray boxes indicate that the group has no genome sequence available; white boxes indicate that no *Merlin* sequence was found; ? indicates that the presence of *Merlin* remains unclear. G = genus; F = family; O = order; C = class; P = phylum; K = kingdom. Some taxa that have no genome available were omitted from the three (Chaetognatha, Gnathostomulida, Syssomonas from Holozoa; Colponemidia, Acavomonas from Alveolata; Jacobida, *Tsukubamonas* from Discoba; Symbiontida from Euglenozoa).

In Nucleotemycea, a group that includes fungi, 888 sequences were found in 28 genomes (S1 Appendix). A high number of sequences is present in two distinct strains of *A. algerae* (338 and 281 sequences) and other microsporidian species, namely *Nosema ceranae* (40 sequences), *Pseudoloma neurophilia* (38 sequences) and *Hepatospora eriocheir* (32 sequences). Evidence of a relatively recent spread of *Merlin* was reported in some microsporidian species [34]. Furthermore, *Merlin* was found in Mucoromycota (*Rhizopus delemar*: 57 sequences; *Lichtheimia corymbifera*: 14 sequences; *Phycomyces blakesleeana*: 11 sequences; *R. microspores*: 10 sequences, and others), a clade consisting of mycorrhizal fungi, root endophytes, and decomposers of plant material [35]. The queried elements were also identified in single representatives of the basal fungal phyla Cryptomycota (4 sequences in *Rozella allomyces*, also found by [36]) and Zoopagomycota (a single sequence in *Conidiobolus coronatus*).

The presence of *Merlin* in stramenopiles was previously reported [14] and here we found 128 sequences in 13 genomes from the genera *Phytophthora* and *Aphanomyces* (S1 Appendix). On the other hand, we provide the first report of *Merlin* in several alveolates mainly from Ciliophora, Dinophyceae and Perkinsozoa, and we explored its structure in more detail (see below).

By tblastn, possible *Merlin* elements were also found in 2 hemichordate genomes (S2 Appendix). *Saccoglossus kowalevskii* (GCA_000003605.1) showed around 30 hits, most of which represent remnants of *Merlin*, and 2 preserved transposase ORFs were found. In contrast, hundreds of hits were found in *Ptychodera flava* (GCA_001465055.1) along with at least 8 preserved transposase genes. Moreover, hundreds of significant hits were found in 2 available Xenacoelomorpha genomes, namely *Hofstenia miami* (GCA_900660155.1 and GCA_004352715.1) with at least 10 preserved transposase ORFs. Some hits were also found in 3 analyzed ctenophores genomes: *Mnemiopsis leidyi* (GCA_000226015.1; only remnants of ORFs), *Pleurobrachia bachei* (GCA_000695325.1; 4 preserved ORFs) and *Beroë ovata* (GCA_900239995.1; 6 preserved ORFs).

In Rhodophyceae, Metamonada and Discoba, we also found significant hits by tblastn, corresponding to the first report of *Merlin* superfamily in these taxa and we performed a curated analysis of copies (see below).

Some hits were also found in the genomes of 3 green algae (Chlorophyta) from 97 genomes analyzed, namely *Dunaliella* sp. (GCA_004335775.1), *Chloromonas* sp.

(GCA_004335635.1) and *Ulva prolifera* (GCA_004138255.1) (S2 Appendix). Since in *U. prolifera*, the transposase and flanking regions have high similarity to bacterial sequences (around 90%), contamination is strongly suggested. Although sequences from the other two species share only around 40% identity with bacterial sequences, they seem to be contaminations as well, since near genes also have these similarities and the contaminants contigs are nearly identical in both genomes.

From 45 analyzed amoebozoan genomes, hits were found in *Acanthamoeba mauritaniensis* (GCA_000826465.1), *Physarum polycephalum* (GCA_000413255.3) and *Synstelium polycarpum* (GCA_900092255.1) (S2 Appendix). Transposase and near genes identified in *A. mauritaniensis* and *P. polycephalum* are 90-100% identical with bacterial sequences. Hits for *S. polycarpum* are from truncated ORFs located in 4 very short contigs (488-941 bp) that do not show high similarity with bacterial sequences. However, since these contigs are not mapped on the genome, it is hard to judge whether these are remnants of *Merlin* or contaminations with eukaryotic reads. Thus, we do not have clear evidence for the presence of *Merlin* in this group.

In Rhizaria, significant hits were found in only 2 genomes from 60 analyzed ones of 11 species (S2 Appendix). In *Globobulimina* sp. (GCA_003354225.1) *Merlin* hits are from truncated ORFs located in 2 very short contigs with no evidence of bacterial contamination. Similar to *S. polycarpum*, it is not clear if these sequences are part of the genomes. In *Reticulomyxa filosa* (GCA_000512085.1) some significant hits containing truncated ORFs and 2 complete ORFs were found. A blastp of the transposases and neighboring genes revealed 40-60% identity to bacterial sequences. Despite a relatively low similarity, this might be contamination, since the contigs are not assembled in the genome. A hypothesis of horizontal transfer of bacterial fragments is unlikely because the flanking genes are distinct among contigs, and consequently, several independent events have to be implied. Glöckner et al (2014) have identified and removed some contigs in the *R. filosa* assembly that were derived from bacteria [37], but the contigs we identified are possibly from an unknown bacterial source that escaped this filtering. Thus, the presence of *Merlin* sequences in Rhizaria remains unclear.

Using blastp, we were unable to identify significant hits in plants (Streptophyta), although one *Merlin* element was previously identified in the spikemoss *Selaginella moellendorffi*, and is deposited in the Repbase database (*Merlin-1_Smo*) [13,16]. We have also analyzed almost 700 genomes using tblastn, yet most hits appear to be

contaminations. For example, one significant hit was found in a short contig of the *Lactuca sativa* genome (GCA_900243165.1) that has 99.91% identity to a genomic fragment of the bacterium *Proteiniphilum saccharofermentans*. Nevertheless, for a few species, there was no evidence of sequence contamination and *Merlin* is indeed present in their genomes, such as in *S. moellendorffii* (GCA_000143415.2), *S. kraussiana* (GCA_001021135.1) and *S. tamariscina* (GCA_003024785.1). However, functionally active copies are absent. The canonical *Merlin*-1_Smo and the copies found in the genomes have truncated DDE_Tnp_IS1595 domain. Since the genus *Selaginella* belongs to a basal lineage (class Lycopodiopsida) [38], it is plausible to assume that *Merlin* was lost in more derived plant lineages. This notion is supported by our finding of degenerate copies in *Rhodamnia argentea* (GCA_900635035.1) from the derived clade Euphyllophyta.

In some groups, for which whole genomes are available, namely Trichoplax (2 genomes), Choanoflagellata (2 genomes), Filasterea (1 genome), Ichthyosporea (9 genomes), Corallochytrium (1 genome), Breviatea (1 genome), Apusomonadidae (1 genome), Glaucophyta (1 genome), Cryptista (6 genomes), and Haptista (3 genomes), no significant hits for *Merlin* were found neither by blastp nor by tblastn.

It is important to notice that using blastp searches, we were looking for annotated proteins containing at least a partial DDE_Tnp_IS1595 domain and consequently, more degenerate copies were not identified. The analysis of a CD-search indicates that for some proteins the DDE_Tnp_IS1595 domain is incomplete (S1 and S2 Appendix) and these are possibly inactive transposases. Also, in several proteins, we observed additional domains that are not expected to be found in transposases such as the EGT51828.1 protein from *Caenorhabditis brenneri* that has a DDE_Tnp_IS1595 followed by a WD40 domain. These proteins could be a result of degeneration or yet artifact of genomic misassembly but are also good candidates for functional chimeras. Recently, Cosby et al. identified several examples of host-transposase fusion genes as a recurrent path for the emergence of transcription factors [39]. Interestingly, the Zn_Tnp_IS1595 domain was predicted in the N-terminal portion of a few proteins. For example, this domain is found in two proteins from *C. briggsae* (XP_002642078.1 and XP_002638395.1) with no indication of bacterial sequence contamination, questioning whether all eukaryotic *Merlin* are derived from IS1016 that lacks this domain.

In general, we observed a wider distribution of *Merlin* than previously reported. Although we did not explore the presence/absence in lower taxonomic ranks, *Merlin* is

clearly absent from a number of taxa, which may be a result of multiple independent losses and/or horizontal transfer. Extensive sequencing of the underrepresented taxa should help to distinguish between these two possibilities. The picture is further confused by contaminations that represent a common occurrence in large-scale sequencing and usually derive from microbiome present in the analyzed tissues or from the environment [40–43].

First report of Merlin in Rhodophyceae, Metamonada, Discoba and Alveolata

From 20 analyzed Rhodophyceae genomes, *Merlin*-related sequences were found only in *Porphyridium purpureum* (GCA_008690995.1) (S4 Table). Two copies were found sharing 99% nucleotide identity, and the alignment revealed imperfect TIRs of 39 bp, extendable with additional mismatches. One of these copies presents conserved TSDs of 9 bp that indicates recent insertion (Fig 2A, S3 Appendix). They encode proteins of 319 aa with conserved DDE_Tnp_IS1595 domain and DDE motif (S4 Table and Fig 3). CENSOR analysis indicates low similarity (33-39%) with known *Merlin* from Repbase. Only one additional remnant copy of *Merlin* was found in the genome. The presence of at least one potentially active and new copy and the absence in other red algae suggest a recent invasion of *Merlin* in this genome.

In Discoba, 95 genomes were searched (S4 Table), and significant hits were found in *Euglena gracilis* (GCA_900893395.1) and 2 kinetoplastids (see below). Most hits found in *E. gracilis* are in short contigs that hindered the recovery of complete copies (S3 Appendix). However, we were able to establish two different *Merlin* families based on the composition and size of TIRs and sequence divergence, both generating 8-bp TSDs (Fig 2B). They carry the expected conserved domain and present 47 to 56% similarity with *Merlin* from *S. mediterranea* (S4 Table); the DDE motif is conserved (Fig 3).

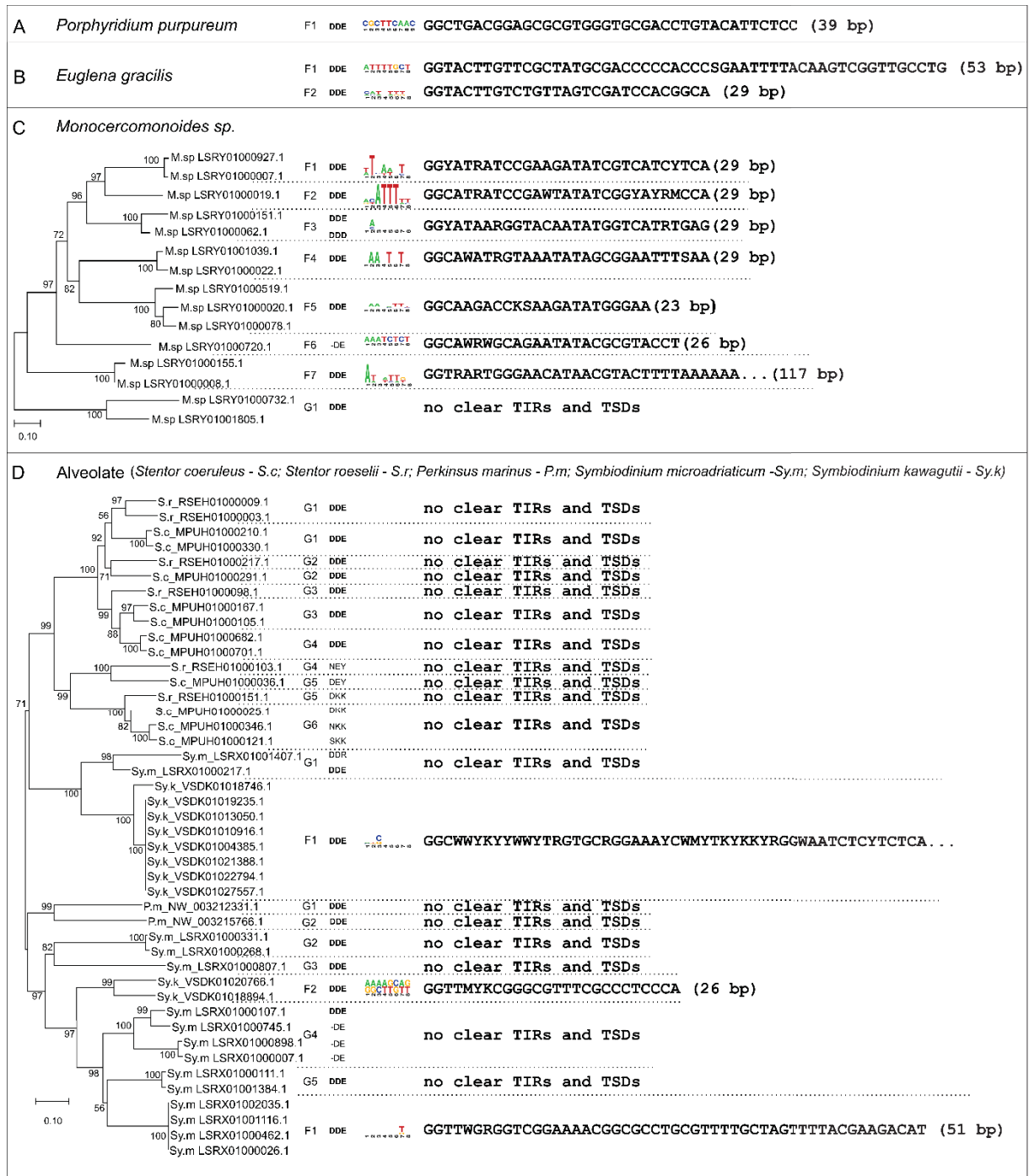


Fig 2. Main characteristics of *Merlin* sequences that were identified in this work. Similar *Merlin* copies within the same species sharing TIR sequences were grouped into families and identified by the letter F followed by a number. Groups of sequences within the same species with no TIRs and TSDs were divided according to the nucleotide divergence and identified by the letter G followed by a number. The residues aligned in the positions of the DDE motif are shown, and its conservation is highlighted in bold. TSD logos are shown and represent the nucleotide usage at each position and the y-axis ranges from a bit score of zero to two. TIR sequences are also shown and represented by both the 5' TIR and the reverse complement of the 3' TIR. Sequences are majority-rule consensus derived from the alignment of multiple copies of each family or individual copies in some cases and the mismatches between the two TIRs are shown as degenerate bases (R = A or G, Y = C or T, S = G or C, W = A or T, K = G or T, and M = A or C). A) *Merlin* family from *P. purpureum* has conserved DDE motif, TIRs of 39 bp and the 9-bp TSDs logo is a frequency plot based on one conserved copy. B) *Merlin* families (F1 and F2) from *E. gracilis* carry the conserved DDE motif, 8-bp TSDs and almost perfect TIRs. TSD logo from F1 is a frequency plot based on one preserved copy. C) Neighbor-joining tree of *Merlin* transposase proteins found in *Monocercomonoides* sp. showing at least 7 families (F1-F7) that present different TIRs and no

clear TSDs consensus. D) Neighbor-joining tree of *Merlin*-related proteins found in Alveolata (S.r – *Stentor roeselii*; S.c – *S.coeruleus*; P. p - *Porphyridium purpureum*; Sy.m – *Symbiodinium microadriaticum*; Sy.k – *Symbiodinium kawagutii*; P.m – *Perkinsus marinus*) based only on the conserved DDE_Tnp_IS1595 domain. However, the DDE motif is not conserved in all sequences and TIRs and TSDs were identified in only a few of them. TSD logo from F2 is a frequency plot based on two conserved copies. The limit of the TIRs from F1 is not clear.

In Metamonada, 24 genomes from 3 phyla were investigated (S4 Table), being found only in Preaxostyla. Few remnants of the *Merlin*-related sequences are found in *Streblomastix strix* (GCA_008636045.1) with no evidence of sequence contamination. It could be a result of an old invasion not followed by a successful amplification. On the other hand, the *Monocercomonoides* sp. PA203 genome (GCA_001643675.1) contains over 100 copies of *Merlin* with preserved or broken transposase ORFs (S3 Appendix). Within the most conserved sequences, we have found at least 7 *Merlin* families (Fig 2C) with different compositions of TIRs, all generating TSDs of 8 bp. Most families have TIRs ranging from 23 to 29 bp, but in some cases, the internal borders of those TIRs were not very clear, since they could be extended considering additional mismatches. The corresponding proteins have 28 to 42% similarity with other *Merlin* elements (S4 Table) and do not seem to be contaminations. As expected for functional transposases, copies of almost all families have the conserved DDE motif (Fig 3). We suggest that the high number of *Merlin* families in this species is the result of an ancient invasion followed by diversification. We found different levels of degeneration, with the transposase pseudogenization occurring in several copies that still contain nearly conserved TIRs and TSDs.

From 340 analyzed alveolate genomes, sequences homologous to *Merlin* were initially found in 7 species by blastp and subsequently in 33 genomes by tblastn. Significant hits were found in ciliates (24 from 33 genomes), dinoflagellates (5 from 7 genomes), Perkinsidae (3 from 8 genomes), and Coolpodellida (1 genome from 2 genomes). In 290 apicomplexans genomes tested, a taxon known to be devoid of TEs [44], only possible contaminants were found in two genomes (S4 Table; S3 Appendix). We restricted the curated analysis to 5 genomes with more conserved copies, where these are on long contigs/scaffolds.

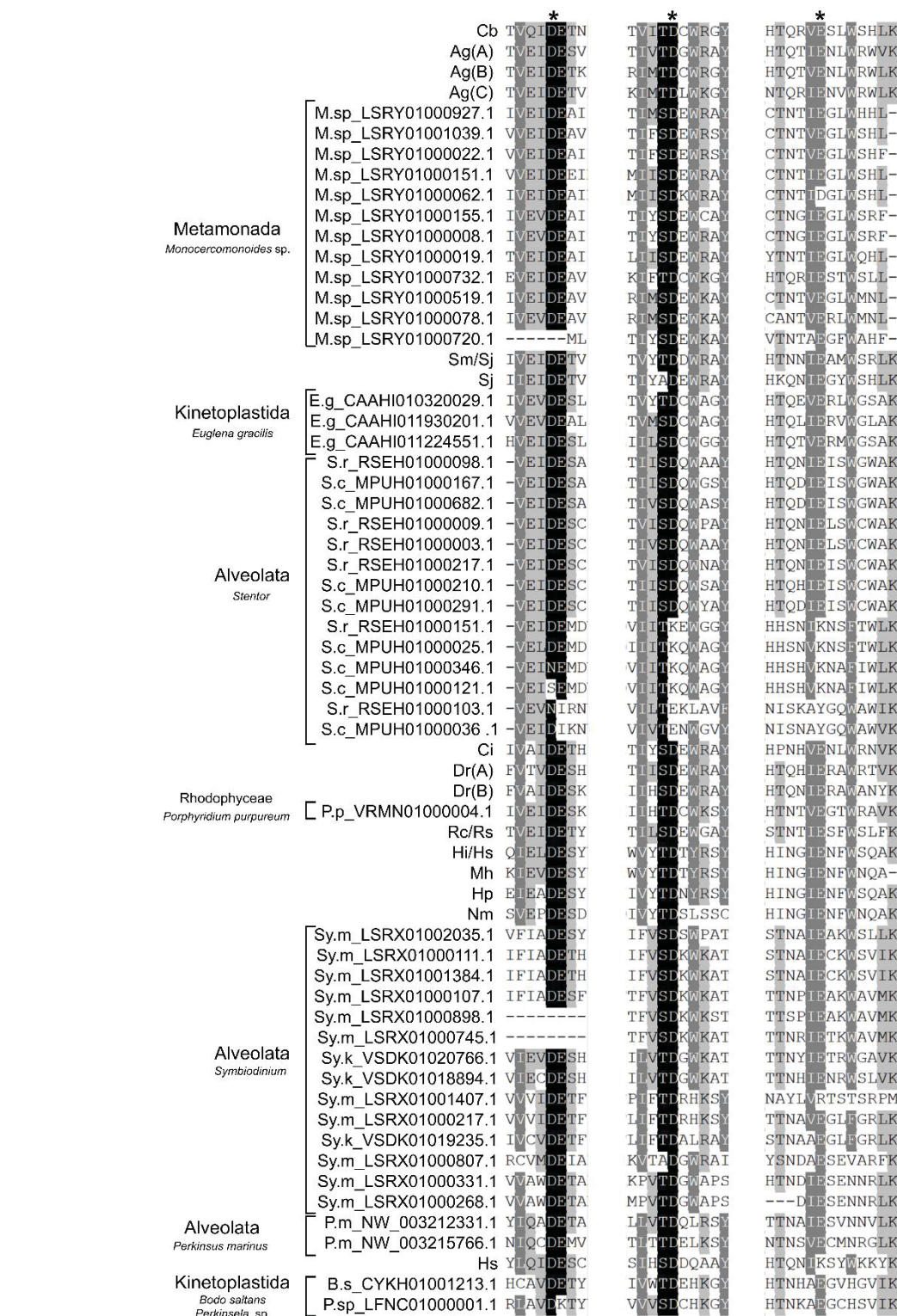


Fig 3. Alignment of the DDE catalytic motif region of *Merlin* families. The three conserved blocks of residues surrounding the DDE motif identified by Feschotte [14] are shown. The number of residues between blocks 1 and 2 varied from around 50 to 70 aa, except for the sequences Sy.m_LSRX01000331.1, Sy.m_LSRX01000268.1 and Sy.m_LSRX01000807.1 that present a larger region (around 110 aa). The DDE motif positions are highlighted with asterisks above the alignment. Colours on the sequences denote residues conservation: black > 90%; dark grey > 80%; light gray > 60%. All *Merlin* transposase proteins identified in this work were aligned with the *Merlin* sequence from *C. briggsae* (CAE74230). The consensus sequences for the three blocks of other *Merlin* transposases and IS1016 were obtained from [14] and added to the alignment (Cb – *C. briggsae*; Tm – *Trichuris muris*; Ag(A), Ag(B) and Ag(C) – *Anopheles gambiae*; Sm/Sj – *Schistosoma mansoni* and *S. japonicum*; Sj –

S. japonicum; Ci – *Ciona intestinalis*; Dr(A) and Dr(B) – *Danio rerio*; Hs – *Homo sapiens*; Rc/Rs – *Rickettsia conorii* and *R. sibirica*, Hi/Hs – *Haemophilus influenzae* and *H. somnus*, Hp – *H. paragallinarum*, Mh – *Mannheimia haemolytica*; Nm – *Neisseria meningitidis*). Sequences from this work are identified by initials (M.sp – *Monocercomonoides* sp.; E.g – *Euglena gracilis*; S.r – *Stentor roeselii*; S.c – *S.coeruleus*; P. p – *Porphyridium purpureum*; Sy.m – *Symbiodinium microadriaticum*; Sy.k – *Symbiodinium kawagutii*; P.m – *Perkinsus marinus*; B.s – *Bodo saltans*; P. sp – *Perkinsella* sp.) and the contig/scaffold ID. Some copies that are identical to others in these regions were omitted from the alignment (Sm_LSRX01000007.1 equal to Sm_LSRX01000898.1; M.sp_LSRX01000007.1 equal to M.sp_LSRX01000927.1; M.sp_LSRX01000020.1 equal to M.sp_LSRX01000078.1; M.sp_LSRX01001805.1 equal to M.sp_LSRX01000732.1; Sc_MPUH01000330.1 equal to Sc_MPUH01000210.1; Sc_MPUH01000105.1 and Sc_MPUH01000701.1 equal to Sc_MPUH01000682.1; Sy.k_VSDK01018746.1, Sy.k_VSDK01013050.1, Sy.k_VSDK01010916.1, Sy.k_VSDK01004385.1, Sy.k_VSDK01021388.1, Sy.k_VSDK01022794.1 and Sy.k_VSDK01027557.1 equal to Sy.k_VSDK01019235.1; Sy.m_LSRX01001116.1, Sy.m_LSRX01000462.1, Sy.m_LSRX01000026.1 equal to Sy.m_LSRX01002035.1).

In the ciliates *Stentor coeruleus* (GCA_001970955.1) and *S. roeselii* (GCA_006503475.1), several preserved ORFs were found encoding proteins varying from 381 to 425 aa containing the C-terminal DDE_Tnp_IS1595 domain. However, these proteins show considerable divergence among each other (Fig 2D; S4 Table). We were not able to find conserved TIRs and TSDs around the ORFs and very short intergenic regions separate them from the neighboring ORFs. CENSOR results for these proteins revealed similarity with known *Merlin* elements, although the DDE motif is not conserved in several ones (Figs 2 and 3). Some copies are probably the product of segment duplication rather than transposition since the nearest genes are shared (S.c_G1 and G6), and it is possible to observe syntenic conservation between both species (S.r_RSEH01000003.1 and S.c_G1 copies; S.r_RSEH01000098.1 and S.c_MPUH01000167.1). The absence of TIRs and TSDs indicates these copies are old *Merlin* insertions, and the conservation of transposase suggests they passed through domestication since rapid pseudogenization would be expected for ancestral copies.

In *Perkinsus marinus* (GCF_000006405.1), two *Merlin*-related preserved ORFs and several remnants of *Merlin* were found. The 350-aa proteins differ by 58% while both have conserved DDE_Tnp_IS1595 domain (S4 Table) and DDE motif (Fig 3). It was not possible to identify the TIRs and TSDs, which raises doubts about the activity of these sequences and the reason behind the conservation of transposase.

Symbiodinium microadriaticum (GCA_001939145.1) contains several preserved ORFs encoding to *Merlin*-related proteins and several remnants. The proteins vary in size from 181 to 445 aa and contain the DDE_Tnp_IS1595 domain, whereas in some of them the DDE motif is not conserved (Figs 2 and 3). The proteins are divided into

two major clusters (Fig 2) and additional subdivisions are possible due to sequence divergence. We could not identify conserved TIRs and TSDs, except for one group of sequences that represent a bona fide transposon family, which we call Sy.m_F1. In sequences from groups G4 and G5, we observed the expected 5'-CC-3' conservation in the 3' end of the alignment, while the 5' border could not be identified, hence representing copies in process of degeneration. Sequences from groups G1, G2 and G3 may be remnants of very old insertions and the conservation as ORFs suggests they could have been domesticated. The G2 copies seem to have been amplified by segment duplication rather than transposition.

The copies of Sy.m_F1 present 51-bp TIRs and 8-bp TSDs with no clear consensus sequence (Fig 2). The alignment of most conserved copies shows high similarity within 2.57 kb, indicating very recent insertions. There is a complex pattern of repeats in the 5' region of the element (Fig 4), including 2.3 units of a 52-bp tandem repeat that contains 5 units of a 6-bp tandem repeat, this region being duplicated in the opposite orientation. CENSOR does not indicate similarity with any known eukaryotic *Merlin* and the DDE_Tnp_IS1595 domain is poorly conserved as suggested by the high e-value (2e-01) (S4 table), yet the DDE motif is conserved (Fig 3). Blastp shows identity around 30% with *Merlin* from *A. algerae* (31% coverage, e-value 3e-04) and with an IS1595 family from *Taibaiella helva* (46% coverage, e-value 1.1e-02). Although the low similarity with both prokaryotic and eukaryotic transposases and the low conservation of the DDE_Tnp_IS1595 domain, the transposase is seemingly functional since the different composition of TSDs among copies show that they are products of transposition. Thus, this is a divergent but functional and active *Merlin* family.

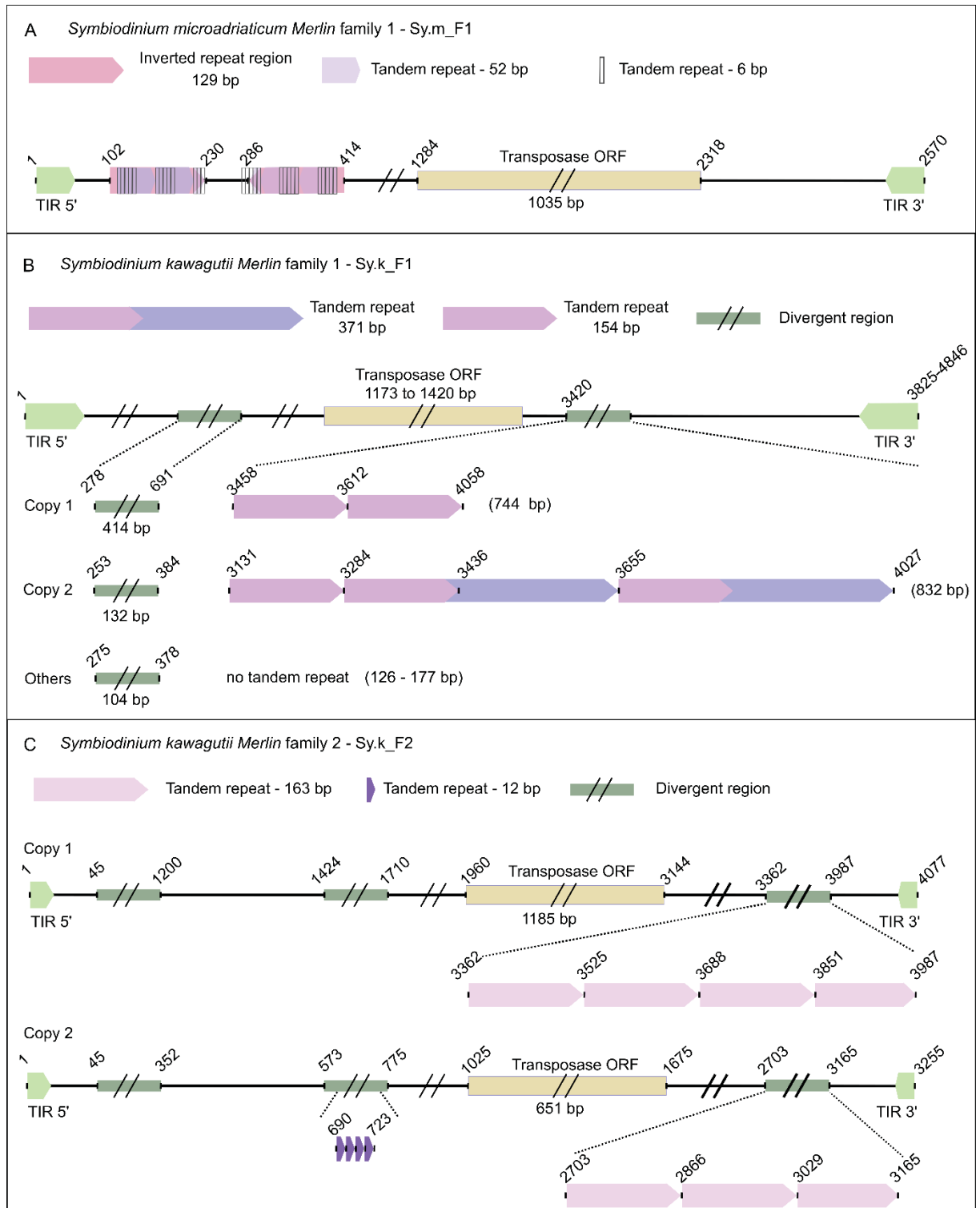


Fig 4. Schematic representation of *Merlin* families containing tandem repeats. A) Representation of Sy.m_F1 from *Sy. microadriaticum* containing a complex pattern of repeats in the 5' region of the element. B) Representation of Sy.k_F1 from *Sy. kawagutii* showing a 5' region that has high divergence among copies in sequence and size and a second divergent region in the 3' end that contains tandem repeats in two copies. Copy 1 - VSDK01027557.1, Copy 2 - VSDK01013050.1. C) Sy.k_F2 from *Sy. kawagutii* with three indicated divergent regions. Due to missing data, we cannot estimate the size of the first region. The second divergent region contains 4 units of a 12-bp repeat in copy 2, while the third divergent region contains a 163-bp tandem repeat. Copy 1 - VSDK01020766.1, Copy 2 - VSDK01003368.1.

In *Sy. kawagutii* (GCA_009767595.1), we found sequences homologous to *Merlin* that fell into two groups (Fig 2) corresponding to different families. Sy.k_F1 has several copies with high conservation in the coding sequence while the 5' and 3' regions are divergent, and some copies contain tandem repeats (Fig 4). The DDE_Tnp_IS1595 domain is predicted with high confidence (S4 Table) and the DDE motif is conserved (Fig 3). Some copies also have a C-terminal zinc finger domain (ZZ superfamily; cl00295). The most intriguing feature of this family is that the 5' and 3' TIRs are very divergent, presenting 12 mismatches in the first 26 positions (Fig 2; S1 Fig). We were able to determine the limits of the element and the TSDs via the alignment of several copies, but the internal limits of TIRs remained unclear. Few mismatches between 5' and 3' TIRs were found for some *Merlin* families [14], nevertheless, this is the first report of highly imperfect TIRs. The identified copies contain one additional 1-kb ORF overlapping with the transposase in the opposite orientation. The predicted protein has no similarity with sequences available in the NCBI and no domain was predicted, being probably generated by chance and kept in all copies due to the conservation in the transposase gene. The conservation of TSDs indicates that despite the great divergence between 5 and 3' TIRs, these copies were amplified by transposition and most insertions are relatively recent, this being a seemingly active family. Sy.k_F2 is composed of only 2 copies that have TIRs of 26 bp (Fig 2), and also have conserved and divergent regions (Fig 4). Copy 1 contains 4 units of a 163-bp tandem repeat that is present as 3 units in copy 2, which carries a 12 bp-long tandem repeat. The DDE_Tnp_IS1595 domain and DDE motif (Fig 3) are conserved.

All the new TE families described here have the 5' TIR that initiates with the nucleotides GG (Fig 2), as shown for all other *Merlin* families [14] and all contain the expected DDE_Tnp_IS1595 domain. Hence, we are confident to consider them as members of the *Merlin* superfamily. Imperfect TIRs were common in the families that we described. This feature is found from several active TEs from different superfamilies and in some cases, TIRs are completely absent [45–47]. It has been shown that for *Tc1/mariner*, the transposase binding has different affinities to imperfect TIRs diminishing the transposition rate [47,48]. Although we have no evidence that *Merlin* behaves like *Tc1/mariner*, we know that TIR sequences are expected to coevolve with the transposase sequence [14,49]; thus, a suboptimal arrangement for transposition

could be positively selected since high rates of transposition could be detrimental to the host.

First report of DNA transposon in kinetoplastid protists

From 89 kinetoplastid genomes analyzed (S4 Table), we found *Merlin*-related sequences in the free-living *B. saltans* (GCA_001460835.1) [50,51] and the parasitic *Perkinsela* sp. CCAP 1560/4 (GCA_001235845.1) [52]. Only retrotransposons were previously reported for these [51,52] and other kinetoplastids, such as members of the human-pathogenic genera *Trypanosoma* and *Leishmania* for which high-quality genomes are available [53–55].

The best tblastn hit for *B. saltans* and *Perkinsela* sp. produced a low score (57.8 and 70.9, respectively), showing around 40% identity with the query over a region of 100 aa, but with a significant e-value ($1.56\text{e-}10$ and $8.32\text{e-}16$, respectively). The presence of the DDE_Tnp_IS1595 domain confirms that this is not a spurious result (S4 Table). However, since we found only a few divergent copies of *Merlin* in both species, the first step was to exclude sequence contamination, the most probable source of which in the case of *B. saltans* is the feeder bacteria and for *Perkinsela* sp., it is its host *Paramoeba pemaquidensis* [52]. Still, bacterial contamination could happen at any point along the sequencing process. In this case, blastp searches against the NCBI nr database would possibly show high similarity with some bacterial sequences, as was the case for some other species, yet the blast results revealed only around 30% identity with IS1595 (S5 Table). Moreover, we did not find *Merlin* in the *P. pemaquidensis* genome (GCA_002151225.1). This was expected since the above-discussed genome assemblies were well treated to eliminate possible sequence contaminations [51,52].

To further verify the presence of *Merlin* in *Perkinsela* sp., PCR was performed using 5 pairs of primers that anneal to the transposon copies and neighboring genes. All 5 combinations of primers presented amplification of expected size, suggesting that the genome assembly is correct and indeed contains *Merlin* (S1 File). Thus, there is strong evidence for *Merlin* being present in these kinetoplastid flagellates. Since this sequence element is clearly absent from related trypanosomatid genomes, it is reasonable to speculate that *Merlin* was present in the last common ancestor of Kinetoplastea predicted to exist about 1 billion years ago [56], and is maintained until present in the basal lineages, while it was lost in the more derived and obligatory

parasitic trypanosomatids. The reasons behind the maintenance of this TE in the extremely reduced genome of *Perkinsela* sp. is an intriguing question to be addressed in the future.

Merlin transposase is preserved yet inactive in Perkinsela sp. and B. saltans

The *Merlin* copies LFNC01000001.1(A) from *Perkinsela* sp. and CYK01001213.1 from *B. saltans* are the highest conserved ones in each genome and were used as reference copies (Fig 5A). They are 762 bp and 1.29 kb long in *Perkinsela* sp. and *B. saltans*, respectively (although there is an alternative start codon in the latter species). The predicted proteins have 43% identity and 64% similarity in a 219-aa conserved region. The structure of *Merlin* from *Perkinsela* sp., as predicted by Phyre2, indicates similarity to c3hosA, a *Tc1-Mariner* Mos1 element from *Drosophila mauritiana* (94.2% confidence, 12% identity) (S2 Fig.), while the confidence for *Merlin* from *B. saltans* is lower (57.72% confidence). Both proteins retain the characteristic DDE motif (Fig 3), indicating that they still can be functional transposases.

The comparisons of copies in both species revealed low conservation as can be seen by the long branches in the nucleotide sequence trees (Figs 5B and 5C). The *Perkinsela* sp. copies LFNC01000001.1(B) and (C) are possibly derived from segment duplication or assembly artifact.

It was not possible to establish the limits of copies and no conserved TIRs and TSDs were found in *Perkinsela* sp., suggesting *Merlin* is likely very ancient and no longer mobilized. The *Merlin* reference copy (Fig 5B; box 5) is annotated as hypothetical protein XU18_0102 (KNH09417.1) and is located in a strand switch region (SSR), same as some other copies. The *Merlin* ORF found in LFNC01000542.1 (Fig 5B; box 10) is not annotated and is located in the opposite orientation as compared to the neighboring genes, with the predicted protein carrying the DDE_Tnp_IS1595 domain. These copies of *Merlin* in *Perkinsela* sp. possess only 85% identity on the nucleotide level for the entire ORFs alignment and 81% identity on the amino acid level in the conserved core alignment of 199 aa.

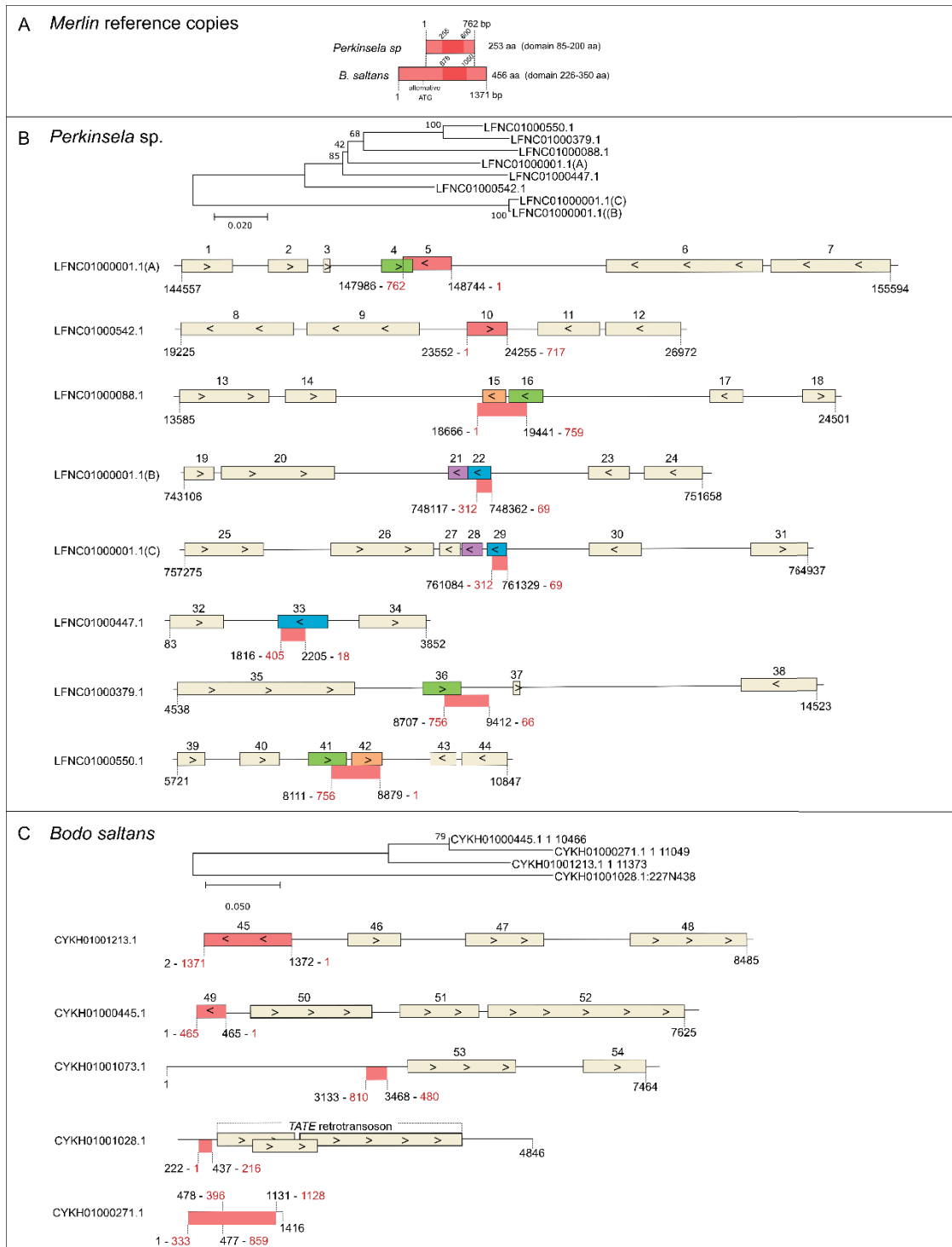


Fig 5. Merlin copies from *B. saltans* and *Perkinsela* sp. A) Representation of the most conserved *Merlin* copy from each species. An alternative internal ATG is shown. Both proteins possess the DDE_Tnp_IS1595 domain whose coding region is represented by dark color. B) Neighbor-joining tree of *Merlin* copies based on nucleotide sequences and the representation of their genomic context in the *Perkinsela* sp. genome. C) Neighbor-joining tree of *Merlin* copies based on nucleotide sequences and the representation of their genomic context in the *B. saltans* genome. ORFs are represented with boxes and numbers and the arrows indicate their direction. Additional information on genes is available in S6 Table. Colored boxes are related ORFs. Red boxes are *Merlin* copies and red boxes without an outline are the non-coding regions with similarity to *Merlin* protein in the tblastn. The relative position of the alignment with *Merlin* reference copy is written in red.

Interestingly, we noticed some overlapping ORFs, in both the same and the opposite orientations to *Merlin* copies, some of which are annotated genes (Fig 5B; S6 Table). Boxes 4, 16, 36 and 41 comprise related ORFs that overlap with *Merlin* in the opposite orientation. The ORF 33 encodes a 258-aa protein with no predicted domain which, however, shares similarity in the C-terminal portion with *Merlin*. Considering the similarity on the nucleotide level throughout the entire ORF, this is likely a copy of *Merlin* that underwent frameshift mutations. Moreover, two other short ORFs also seem to have originated from *Merlin* (boxes 15 and 42). These findings are exciting, since these ORFs may represent new genes derived from *Merlin*, although there is currently no evidence regarding their expression.

The only conserved encoding copy in *B. saltans* (Fig 5C; box 45) is located at the beginning of contig CYKH01001213.1. It was not possible to analyze the presence of TIRs as no sequence for the 3' region of the element is available. The same applies to the copy located in contig CYKH01000445.1 and CYKH01000271.1. The other hits correspond to very degenerate copies, generally located in the opposite orientation concerning the neighboring genes. Thus, the current genome assembly of *B. saltans* does not allow drawing conclusions regarding its activity, however, we would expect more similar copies in the case of active elements.

The orientation of *Merlin* insertions in both kinetoplastid genomes is an interesting point. Unlike other eukaryotes, genes in kinetoplastids are organized in polycistronic units and the transcription initiation by RNA polymerase II preferentially occurs at divergent SSRs [57,58]. The TE copies inserted in the same orientation as near genes may be eliminated faster from the genomes since they would be always expressed in the sense strand of the polycistronic transcript. Correspondingly, copies inserted in the opposite orientation or in the SSR region (as is the case for *Merlin* in both species) could be in “safe havens”, reducing their deleterious effects on the host.

We cannot discard the possibility that genome assembly issues impaired us from finding complete copies of *Merlin* in *B. saltans* and *Perkinsela* sp. Recently, we have detected complete copies of *VIPER* retrotransposons in two *Trypanosoma cruzi* Dm28c assemblies sequenced by PacBio that were missed in 454-based assemblies [59]. The former technology generates long reads and thus allows better quality assembly of the repetitive sequences, which is not yet available for *B. saltans* and *Perkinsela* sp.

In case the *Merlin* copies of *B. saltans* and *Perkinsela* sp. are indeed inactive, we wondered why is the transposase gene conserved. Even for inactive copies, one would expect to find TIRs with or without mismatches, yet they are completely absent, indicating an ancient origin of these insertions. It follows that in such case, the transposase would be subject to pseudogenization. Hence, the conservation of *Merlin* transposase could be the result of domestication, a well-documented process in diverse eukaryotes, including kinetoplastids [60,61].

Merlin and IS1595 family evolutionary tree

An evolutionary tree of curated *Merlin* families was constructed based on the conserved transposase domain DDE_Tnp_IS1595 (Fig 6) and was presented as unrooted, given the absence of a well-supported outgroup choice, since we assume that eukaryotic *Merlin* could have independent origins. Members from the seven groups of IS1595 family were included in the tree and formed a monophyletic clade while the internal monophyly of IS*Nwi* and IS*Sod11* were not recovered. We can see that all eukaryotic elements also form a monophyletic group that could suggest a common ancient origin of these sequences.

Sequences from the same taxon such as Protostomia, Alveolata and Chordata, are grouped in distinct clades and most eukaryotic clades are branching from the same node. These observations reinforce the idea of a great divergence time of these sequences with several independent ancient diversification events. No clear cases of horizontal transfer were identified, although it could explain some inconsistencies in the relationships, such as involving *E. gracilis* and *P. purpureum*. As for other TEs, the stochastic loss is certainly part of the complex evolutionary pattern.

Merlin from *B. saltans* and *Perkinsela* sp. were grouped in the tree, corroborating the idea that *Merlin* is ancient in kinetoplastids. An acquisition via independent horizontal transfer is a less plausible scenario since it should be from a related source and both species have very distinct ecological niches.

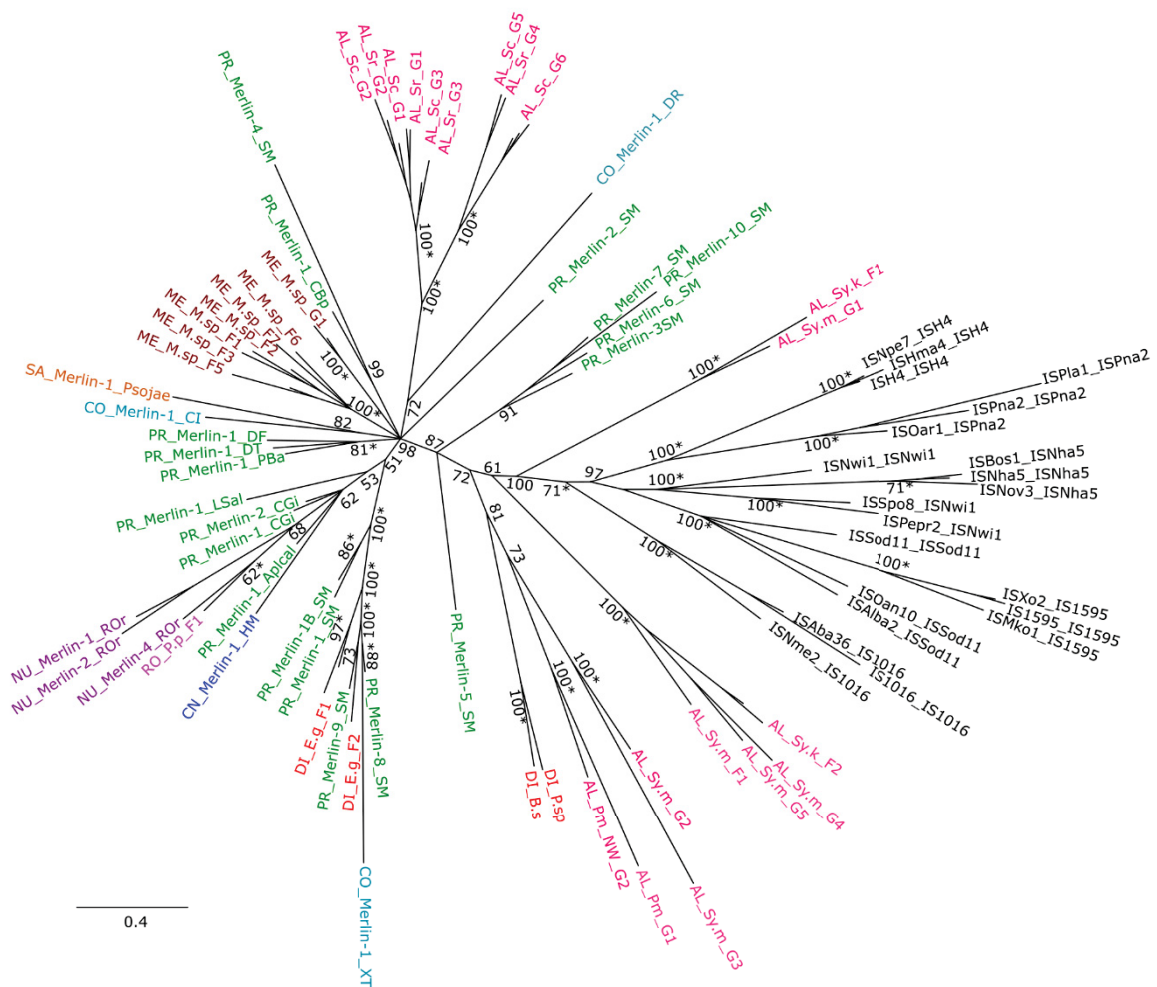


Fig 6. Unrooted 50% majority rule consensus Bayesian tree (WAG + G) of *Merlin* and IS1595 group sequences based on the amino acid sequence of the conserved transposase domain DDE_Tnp_IS1595 (168 positions). Posterior probability values (PP) are indicated near the nodes and some of the values from derived clades were omitted. The * sign near PP values indicates the clade was supported with bootstrap higher than 50 in the ML tree. *Merlin* sequences from different taxonomic groups are highlighted in different colors and identified with initials: AL – Alveolata, CN – Cnidaria, CO – Chordata, NU – Nucleotmycea, PR – Protostomia, SA – Stramenopiles, RO – Rhodophyceae, ME – Metamonada, DI – Discoba. Information of Repbase *Merlin* and IS1595 sequences are found in S1 and S2 Tables, respectively. *Merlin* sequences characterized in this work are identified by the initial of taxon followed by the abbreviation of species name (S.r – *Stentor roeselii*; S.c – *S.coeruleus*; P.p – *Porphyridium purpureum*; Sy.m – *Symbiodinium microadriaticum*; Sy.k – *Symbiodinium kawagutii*; P.m – *Perkinsus marinus*; E.g – *Euglena gracilis*; B.s – *Bodo saltans*; P.sp – *Perkinsella* sp.) and the group or family.

We can observe that the IS1016 is the closest related IS group to the eukaryotic *Merlin*, corroborating the initial hypothesis [14]. However, considering the greater diversity of *Merlin* sequences that were not included in this tree, we cannot rule out the possibility that other *Merlin* elements would have a closer similarity to other IS1595 groups. In fact, including a higher number of sequences, leads to a largely unsolved tree (S3 Fig) where most clades are branching from the base, possibly due to insufficient phylogenetic signal reflecting very ancient separation events. The problem

of character insufficiency is well known for modest size D-D-D/E-type transposases resulting in poor resolution of most phylogenies [62].

Final remarks

DNA transposons are important components of eukaryotic genomes with great diversity at the superfamily and family levels. The relatively recently described *Merlin* superfamily seems to be less widely distributed as compared to other superfamilies such as *Tc1-Mariner* and *hAT* [16]. Here, we have described *Merlin*-related sequences for the first time in Alveolata, Rhodophyceae, Metamonada and Discoba. The sequences identified from all taxa, either complete copies, remnants, or derived genes, contain the characteristic domain DDE_Tnp_IS1595, and there is no evidence for recent horizontal transfer from bacterial ISs, thus all sequences were classified as *Merlin* superfamily regardless of having a common origin.

Our results indicate that *Merlin* is more widespread than thought before and its presence in all major eukaryotic supergroups for which complete genomes are available (Amorphea, Diaphoretickes, Metamonada and Discoba) is compatible with the notion that *Merlin* is a truly ancestral genetic element. Feschotte [14] suggested that *Merlin* and the IS1016 proteins belong to a distinct monophyletic group of transposases that have differentiated from other transposases before the divergence of eukaryotes and prokaryotes. The evolutionary tree obtained with the curated sequences supports this idea. However, due to the diversity of *Merlin* found, and the lack of resolution in the more embracing tree, we cannot rule out the possibility that this superfamily is a result of independent invasions of related prokaryotic IS1595 family members into the eukaryote genomes. Moreover, the patchy distribution and diversity that we see in some groups are better explained by multiple horizontal transfers of *Merlin*.

The presence of conserved *Merlin*-related sequences with no evidence of mobilization capacity preservation was observed in some species, suggesting that *Merlin* may have been domesticated during evolution. Several studies indicate that transposases were co-opted to act in DNA binding, modulation of chromatin structure and TE repression [9,63]. They are also associated with programmed genome rearrangement in the ciliates *Paramecium*, *Tetrahymena* and *Oxytricha* [64], and in some cases, the transposase domain and the D-D-D/E motif remain conserved [63].

Thus, we describe several potential *Merlin* domestication cases that worth further investigation.

Concerning the putatively active *Merlin* families, we can highlight that several of them exhibit signs of recent transposition activity. *Merlin* was successful in colonizing and diverging in the *Monocercomonoides* sp. genome where 7 potentially active families were identified. This co-existence of divergent *Merlin* families in the same genome was already reported [14,17]. We also observed that *Merlin* can have highly imperfect TIRs as seen for Sy.k_F1, a feature not described before for this superfamily. Finally, another interesting finding is the observation of tandem repeats inside *Merlin* sequences. Although this was not described before, we can observe that several of Repbase *Merlin* families also have internal tandem repeats (S1 Table). The close relationship of tandem repeats and TEs has been recently well documented, with several micro, mini and satellite DNAs found embedded within TEs [65–70]. The existence of tandem repeats in multiple families and copies of *Merlin* indicates this TE could help to spread tandem repeats by transposition as proto-satellites that could be next amplified and homogenized such as the model suggested by Paço and colleagues [67].

Importantly, we document for the first time DNA transposons in the kinetoplastid genomes that were thought to be devoid of these genetic elements. The absence of TIRs and TSDs in the *Merlin* copies found in *Perkinsella* sp. contrasts with the conservation of the transposase coding region. Thus, it is possible that complete copies were not assembled, or the transposase is being maintained for a currently unknown cellular function. Our findings represent a starting point for understanding the impact of these sequences on protists and reveal greater diversity of TEs than thought previously.

ACKNOWLEDGMENTS

Authors are grateful to the Instituto Carlos Chagas and Instituto de Biologia Molecular do Paraná for general support, to the two anonymous reviewers for constructive comments and suggestions and to the research community who are providing an increasing number of genome sequences in public databases allowing research works such as the one presented here.

REFERENCES

1. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 2011;7. doi:10.1371/journal.pgen.1002384
2. Karakülah G, Suner A. PlanTEEnrichment: A tool for enrichment analysis of transposable elements in plants. *Genomics.* 2017;109: 336–340. doi:10.1016/j.ygeno.2017.05.008
3. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet.* 2002;3: 329–41. doi:10.1038/nrg793
4. Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution.* 2001;55: 1–24. doi:10.1111/j.0014-3820.2001.tb01268.x
5. Oliver KR, McComb JA, Greene WK. Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biol Evol.* 2013;5: 1886–1901. doi:10.1093/gbe/evt141
6. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007;8: 973–82. doi:10.1038/nrg2165
7. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer MA. L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A.* 2008;105: 19366–71. doi:10.1073/pnas.0807866105
8. Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays.* 2006;28: 913–22. doi:10.1002/bies.20452
9. Jangam D, Feschotte C, Betrán E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* 2017;33: 817–831. doi:10.1016/j.tig.2017.07.011
10. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. *Mol Ecol.* 2019;28: 1537–1549. doi:10.1111/mec.14794
11. Finnegan DJ. Eukaryotic transposable elements and genome evolution. *Trends Genet.* 1989;5: 103–7. doi:10.1016/0168-9525(89)90039-5
12. Munoz-Lopez M, Garcia-Perez J. DNA transposons: Nature and applications in genomics. *Curr Genomics.* 2010;11: 115–128. doi:10.2174/138920210790886871
13. Jurka J, Kapitonov VV V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 2005;110: 462–7. doi:10.1159/000084979
14. Feschotte C. *Merlin*, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. *Mol Biol Evol.* 2004;21: 1769–80. doi:10.1093/molbev/msh188
15. Siguier P, Gagnevin L, Chandler M. The new IS1595 family, its relation to IS1 and the frontier between insertion sequences and transposons. *Res Microbiol.* 2009;160: 232–41. doi:10.1016/j.resmic.2009.02.003

16. Yuan Y-W, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci U S A*. 2011;108: 7884–9. doi:10.1073/pnas.1104208108
17. Parisot N, Pelin A, Gasc C, Polonais VV, Belkorchia A, Panek J, et al. Microsporidian genomes harbor a diverse array of transposable elements that demonstrate an ancestry of horizontal exchange with metazoans. *Genome Biol Evol*. 2014;6: 2289–2300. doi:10.1093/gbe/evu178
18. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*. 2004;32: W327–31. doi:10.1093/nar/gkh454
19. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res*. 2013;41. doi:10.1093/nar/gkt282
20. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res*. 2019. doi:10.1093/nar/gkz342
21. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*. 2006;7: 474. doi:10.1186/1471-2105-7-474
22. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Battistuzzi FU, editor. *Mol Biol Evol*. 2018;35: 1547–1549. doi:10.1093/molbev/msy096
23. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27: 573–80. doi:10.1093/nar/27.2.573
24. Kelley LA, Sternberg MJE. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc*. 2009;4: 363–71. doi:10.1038/nprot.2009.2
25. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26: 680–2. doi:10.1093/bioinformatics/btq003
26. Pei J, Grishin N V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol*. 2014;1079: 263–71. doi:10.1007/978-1-62703-646-7_17
27. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25: 1972–3. doi:10.1093/bioinformatics/btp348
28. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61: 539–42. doi:10.1093/sysbio/sys029
29. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES science gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop, GCE*. New Orleans, LA; 2010. pp. 1–8. doi:10.1109/GCE.2010.5676129
30. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30: 1312–3. doi:10.1093/bioinformatics/btu033

31. Shao H, Tu Z. Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics*. 2001;159: 1103–15. Available: <http://www.ncbi.nlm.nih.gov/pubmed/11729156>
32. Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, et al. Revisions to the classification, nomenclature, and diversity of eukaryotes. *J Eukaryot Microbiol*. 2019;66: 4–119. doi:10.1111/jeu.12691
33. Yang Y, Xiong J, Zhou Z, Huo F, Miao W, Ran C, et al. The genome of the myxosporean *Thelohanellus kitauei* shows adaptations to nutrient acquisition within its fish host. *Genome Biol Evol*. 2014;6: 3182–3198. doi:10.1093/gbe/evu247
34. de Albuquerque NRM, Ebert D, Haag KL. Transposable element abundance correlates with mode of transmission in microsporidian parasites. *Mob DNA*. 2020;11: 19. doi:10.1186/s13100-020-00218-8
35. Spatafora JW, Chang Y, Benny GL, Lazarus K, Smith ME, Berbee ML, et al. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia*. 2016;108: 1028–1046. doi:10.3852/16-042
36. Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K. Cut-and-paste transposons in fungi with diverse lifestyles. *Genome Biol Evol*. 2017;9: 3463–3477. doi:10.1093/gbe/evx261
37. Glöckner G, Hülsmann N, Schleicher M, Noegel AA, Eichinger L, Gallinger C, et al. The genome of the foraminiferan *Reticulomyxa filosa*. *Curr Biol*. 2014;24: 11–18. doi:10.1016/j.cub.2013.11.027
38. PPG I. A community-derived classification for extant lycophytes and ferns. *J Syst Evol*. 2016;54: 563–603. doi:10.1111/jse.12229
39. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science*. 2021;371. doi:10.1126/science.abc6405
40. Sangiovanni M, Granata I, Thind AS, Guarracino MR. From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics*. 2019;20: 168. doi:10.1186/s12859-019-2684-x
41. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, et al. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog*. 2014;10: e1004437. doi:10.1371/journal.ppat.1004437
42. Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One*. 2014;9: e110808. doi:10.1371/journal.pone.0110808
43. Francois CM, Durand F, Figuet E, Galtier N. Prevalence and implications of contamination in public genomic resources: a case study of 43 reference arthropod assemblies. *G3 (Bethesda)*. 2020;10: 721–730. doi:10.1534/g3.119.400758
44. Kissinger JC, DeBarry J. Genome cartography: charting the apicomplexan genome. *Trends Parasitol*. 2011;27: 345–54. doi:10.1016/j.pt.2011.03.006

45. Jurka J, Kapitonov V V, Kohany O, Jurka M V. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet.* 2007;8: 241–59. doi:10.1146/annurev.genom.8.080706.092416
46. Yan F, Di S, Takahashi R. CACTA-superfamily transposable element is inserted in MYB transcription factor gene of soybean line producing variegated seeds. *Genome.* 2015;58: 365–74. doi:10.1139/gen-2015-0054
47. Trubitsyna M, Grey H, Houston DR, Finnegan DJ, Richardson JM. Structural basis for the inverted repeat preferences of mariner transposases. *J Biol Chem.* 2015;290: 13531–40. doi:10.1074/jbc.M115.636704
48. Augé-Gouillou C, Hamelin MH, Demattei M V, Periquet M, Bigot Y. The wild-type conformation of the Mos-1 inverted terminal repeats is suboptimal for transposition in bacteria. *Mol Genet Genomics.* 2001;265: 51–7. doi:10.1007/s004380000385
49. Naumann TA, Reznikoff WS. Tn5 transposase with an altered specificity for transposon ends. *J Bacteriol.* 2002;184: 233–40. doi:10.1128/jb.184.1.233-240.2002
50. d'Avila-Levy CM, Boucinha C, Kostygov A, Santos HLC, Morelli KA, Grybchukleremenko A, et al. Exploring the environmental diversity of kinetoplastid flagellates in the high-throughput DNA sequencing era. *Mem Inst Oswaldo Cruz.* 2015;110: 956–65. doi:10.1590/0074-02760150253
51. Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Curr Biol.* 2016;26: 161–172. doi:10.1016/j.cub.2015.11.055
52. Tanifuji G, Cenci U, Moog D, Dean S, Nakayama T, David V, et al. Genome sequencing reveals metabolic and cellular interdependence in an amoeba-kinetoplastid symbiosis. *Sci Rep.* 2017;7: 11688. doi:10.1038/s41598-017-11866-x
53. Bringaud F, Ghedin E, El-Sayed NM a, Papadopoulou B. Role of transposable elements in trypanosomatids. *Microbes Infect.* 2008;10: 575–81. doi:10.1016/j.micinf.2008.02.009
54. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science.* 2005;309: 409–15. doi:10.1126/science.1112631
55. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet.* 2007;39: 839–47. doi:10.1038/ng2053
56. Butenko A, Hammond M, Field MC, Ginger ML, Yurchenko V, Lukeš J. Reductionist pathways for parasitism in euglenozoans? Expanded datasets provide new insights. *Trends Parasitol.* 2021;37: 100–116. doi:10.1016/j.pt.2020.10.001
57. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell.* 2003;11: 1291–9. doi:10.1016/s1097-2765(03)00143-6
58. Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.* 2003;5: 1231–1240. doi:10.1016/j.micinf.2003.09.005

59. Ribeiro YC, Robe LJ, Veluza DS, dos Santos CMB, Lopes ALK, Krieger MA, et al. Study of VIPER and TATE in kinetoplastids and the evolution of tyrosine recombinase retrotransposons. *Mob DNA*. 2019;10: 34. doi:10.1186/s13100-019-0175-2
60. Ludwig A, Krieger MA. Genomic and phylogenetic evidence of VIPER retrotransposon domestication in trypanosomatids. *Mem Inst Oswaldo Cruz*. 2016;111: 765–769. doi:10.1590/0074-02760160224
61. Smith M, Bringaud F, Papadopoulou B. Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome. *BMC Genomics*. 2009;10: 240. doi:10.1186/1471-2164-10-240
62. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA*. 2017;8: 19. doi:10.1186/s13100-017-0103-2
63. Gao B, Wang Y, Diaby M, Zong W, Shen D, Wang S, et al. Evolution of pogo, a separate superfamily of IS630-Tc1-mariner transposons, revealing recurrent domestication events in vertebrates. *Mob DNA*. 2020;11: 25. doi:10.1186/s13100-020-00220-0
64. Vogt A, Goldman AD, Mochizuki K, Landweber LF. Transposon domestication versus mutualism in ciliate genome rearrangements. *PLoS Genet*. 2013;9: e1003659. doi:10.1371/journal.pgen.1003659
65. Inukai T. Role of transposable elements in the propagation of minisatellites in the rice genome. *Mol Genet Genomics*. 2004;271: 220–7. doi:10.1007/s00438-003-0973-5
66. Smýkal P, Kalendar R, Ford R, Macas J, Griga M. Evolutionary conserved lineage of Angela-family retrotransposons as a genome-wide microsatellite repeat dispersal agent. *Heredity (Edinb)*. 2009;103: 157–67. doi:10.1038/hdy.2009.45
67. Paço A, Freitas R, Vieira-da-Silva A. Conversion of DNA Sequences: from a transposable element to a tandem repeat or to a gene. *Genes (Basel)*. 2019;10. doi:10.3390/genes10121014
68. Belyayev A, Josefiová J, Jandová M, Mahelka V, Krak K, Mandák B. Transposons and satellite DNA: on the origin of the major satellite DNA family in the *Chenopodium* genome. *Mob DNA*. 2020;11: 20. doi:10.1186/s13100-020-00219-7
69. Wilder J, Hollocher H. Mobile elements and the genesis of microsatellites in dipterans. *Mol Biol Evol*. 2001;18: 384–392. doi:10.1093/oxfordjournals.molbev.a003814
70. Kapitonov V V, Holmquist GP, Jurka J. L1 repeat is a basic unit of heterochromatin satellites in cetaceans. *Mol Biol Evol*. 1998;15: 611–2. doi:10.1093/oxfordjournals.molbev.a025963

SUPPORTING INFORMATION

S1 Fig. Alignment of flanking regions of most conserved copies of *Merlin* family 1 from *Sy. kawagutii* (Sy.k_F1). The first and last (reverse complement) 130 nucleotides of copies were aligned and it is possible to observe highly imperfect TIRs with no clear limit. TSDs for each copy are shown highlighted in different colors.

<https://doi.org/10.1371/journal.pone.0251133.s001> (TIFF)

S2 Fig. Secondary structure prediction of *Merlin* reference protein from *Perkinsela* sp. modeled used Phyre2. The secondary structure was predicted using c3hosA template, a Tc1-Mariner Mos1 element from *Drosophila mauritiana*. The structure was predicted with 94.2% confidence.

<https://doi.org/10.1371/journal.pone.0251133.s002> (TIF)

S3 Fig. Bayesian tree (LG + G) of *Merlin* and IS1595 group sequences. The tree is based on the amino acid sequence of the conserved transposase domain DDE_Tnp_IS1595 (142 positions) and was rooted by the midpoint.

<https://doi.org/10.1371/journal.pone.0251133.s003> (PNG)

S1 File. Information on PCR for confirming the presence of *Merlin* elements in *Perkinsela* sp.

<https://doi.org/10.1371/journal.pone.0251133.s004> (PDF)

S2 File. Alignment of the conserved transposase domain DDE_Tnp_IS1595 of curated *Merlin* sequences and IS1595 used to generate the tree shown in Fig 6.

<https://doi.org/10.1371/journal.pone.0251133.s005> (ALN)

S3 File. Alignment of the conserved transposase domain DDE_Tnp_IS1595 of *Merlin*/IS1595 used to generate the tree shown in S3 Fig.

<https://doi.org/10.1371/journal.pone.0251133.s006> (ALN)

S1 Appendix. *Merlin*-related proteins identified by blastp. Sequences are provided as fasta files for each major taxonomic group and the results of the CD-search are also provided.

<https://doi.org/10.1371/journal.pone.0251133.s007>

(ZIP)

S2 Appendix. *Merlin*-related sequences identified by tblastn from Rhizaria, Streptophyta, Amoebozoa, Chlorophyta, Hemichordata, Ctenophora and Xenacoelomorpha. Contigs containing positive hits and protein sequences are provided as fasta files. The results of the CD-search are also provided. A README file is provided with additional information about files.

<https://doi.org/10.1371/journal.pone.0251133.s008> (ZIP)

S3 Appendix. Sequences analyzed in this work from Discoba, Metamonada, Rodophyceae and Alveolata. Contigs containing positive hits for each species are provided as fasta files. Nucleotide sequences from each *Merlin* group or family are also provided. All protein sequences were provided in a single file. A README file is provided with additional information about files.

<https://doi.org/10.1371/journal.pone.0251133.s009> (ZIP)

S1 Table. Summary of main features of *Merlin* canonical families available in the Repbase23.11.

<https://doi.org/10.1371/journal.pone.0251133.s010> (XLSX)

S2 Table. Information of IS1595 group sequences used in the phylogenetic tree.

<https://doi.org/10.1371/journal.pone.0251133.s011> (XLSX)

S3 Table. Summary of blastp and tblastn against the major taxonomic groups.

<https://doi.org/10.1371/journal.pone.0251133.s012> (XLSX)

S4 Table. General results obtained by Rhodophyceae, Rhizaria, Metamonada, Discoba and Alveolata groups. The data is provided in different tabs, including the genomes analyzed, CD-search and CENSOR results and protein divergence. The description of each tab is available in the first tab.

<https://doi.org/10.1371/journal.pone.0251133.s013> (XLSX)

S5 Table. Summary of blastp result against protein nr database using *Merlin* protein sequence from Bodo saltans and *Perkinsela* sp. as a query.

<https://doi.org/10.1371/journal.pone.0251133.s014> (XLSX)

S6 Table. Information about the ORFs overlapping and near *Merlin* copies that are represented in Fig 5.

<https://doi.org/10.1371/journal.pone.0251133.s015> (XLSX)

5 CONSIDERAÇÕES FINAIS

Os elementos de transposição se caracterizam pela sua capacidade de se mover dentro do genoma e constituem uma relevante porção nos genomas de diferentes espécies. Esses elementos são forças poderosas de mudança genética e têm desempenhado um papel significativo na evolução de muitos genomas. Dessa maneira, o entendimento da biologia dessas sequências bem como da sua diversidade é crucial para a compreensão global de como os TEs afetam a evolução e o funcionamento dos organismos hospedeiros. Esse estudo se torna ainda mais relevante quando novos tipos e novas famílias de TEs são identificados, o que vem acontecendo à medida que novos genomas são sequenciados.

Além disso, os transposons de DNA podem ser usados como ferramenta genética através da transferência gênica e mutagênese insercional podendo ser utilizados para análise de regulação dos genomas, estudar o desenvolvimento embrionário, identificar genes e vias implicadas em doenças ou patogênese de patógenos e, até mesmo, contribuir para a terapia gênica. Dentre as ferramentas de engenharia genética destacamos os sistemas baseados em transposição, no qual um gene de interesse é clonado entre as TIRs de um transposon em um vetor o qual é usado para inserção genômica mobilizado por uma transposase que deve ser suplementada. Esses sistemas representam uma ferramenta muito promissora para uma ampla variedade de estudos genômicos e para produção de proteína recombinante.

Diante da importância dos transposon de DNA para diferentes áreas da biologia, neste trabalho esses elementos foram estudados em diferentes contextos. Primeiramente, por uma abordagem biotecnológica realizamos a caracterização e melhoramento de um sistema de transposição baseado em transposases da superfamília *Tc1/Mariner* e, segundo, empregamos uma abordagem biológica e evolutiva dos transposons de DNA através do estudo da distribuição e evolução dos elementos *Merlin* em eucariotos.

Iniciamos uma caracterização básica do Sistema Ingresso 1.0 previamente desenvolvido e modificado em nosso laboratório, que consiste em uma plataforma de vetores para superexpressão heteróloga baseada em transposição. Essa etapa é essencial para o melhor entendimento dos diversos componentes presentes neste sistema e do desempenho do sistema como um todo. Neste contexto, avaliamos o

número de cópias do cassete de expressão inseridas bem como o nível de expressão de mRNA do gene de interesse em diferentes clones isolados de células CHO expressando diferentes níveis da proteína de interesse. O número de cópias obtidos para o cassete está dentro do esperado (de 2 a 10 cópias) considerando a baixa eficiência da *SB11*, em relação a outras versões aprimoradas como a *SB100X*. Possivelmente, a metodologia de seleção dos transfectantes expressando GFP por FACS, que permite a seleção de células únicas através da intensidade de fluorescência, contorna a baixa eficiência da transposase *SB11* em termos de número de células contendo inserções.

Para alguns clones, como o 28 e o 4, a relação entre o número de cópias inseridas no genoma e o nível de expressão do mRNA e de proteína foi condizente. Esses clones apresentaram altas taxa de expressão proteica (0,4 e 0,35 g/L), elevado número de cópias de DNA (10 e 6 cópias) e alta taxa de expressão a nível de mRNA (> 40 vezes em relação a amostra de menor expressão). No entanto, esse padrão não foi observado para todos os clones sendo verificado uma correlação imperfeita entre o nível de mRNA e expressão proteica indicando que outros fatores devem influenciar nessa relação. De fato, a expressão gênica é um processo complexo e é regulado em muitos níveis como transcrição, processamento do mRNA, tradução, modificações pós traducionais entre outros (BUCCITELLI e SELBACH, 2020).

O número de inserções sequenciadas e identificadas para cada amostra variou de 2 a 34 cópias, o que representa um valor acima do esperado para a *SB11* e maior do que o verificado pela análise de qPCR. Como já mencionado, uma possível explicação para essa discrepância é a remobilização do cassete de expressão pela transposase, caso ela esteja inserida no genoma da célula CHO. Dessa maneira, o número de cópias entre as células do mesmo clone não varia, mas existem diferentes locais de inserção no genoma devido a atividade de remobilização do cassete pela transposase e, dessa maneira, o número de diferentes inserções observadas pela análise do sequenciamento é maior. O clone 28, por exemplo, apresentou 10 cópias do cassete na análise de qPCR e 23 inserções diferentes pelo sequenciamento. A presença da *SB* no genoma das células foi confirmada para o clone 28 e na população, sendo assim essa hipótese plausível.

A maioria dos elementos transponíveis não se transpõem de forma completamente aleatória, mas parecem ter locais preferenciais para integração, os quais podem ser chamados de *hotspots*. A superfamília *Tc1/Mariner* possui sítio

preferencial de inserção adjacente a região TA (LIU et al., 2005). Ao verificar o local de inserção do transgene nos clones selecionados, praticamente todos apresentaram o dinucleotídeo TA no local de inserção, indicando o comportamento de inserção padrão da superfamília *Tc1/Mariner*. A preferência por uma região rica em TA pode ser o primeiro nível de seleção do local da inserção da *SB* uma vez que esses locais podem facilitar o relaxamento dos pares de bases durante a reação de integração e demonstram que a estrutura do DNA pode desempenhar um papel na seleção do sítio alvo de *SB* (YANT et al, 2005). Mais recentemente, foi demonstrado que inserções não-TA podem ocorrer também, embora de forma mais rara, aproximadamente 1,4% das inserções (GUO, ZHANG e HU, 2018). Apenas um sítio de inserção no clone 6 foi avaliado como uma possível inserção não TA e sim AA.

As inserções do Sistema Ingresso não parecem mostraram relação com quaisquer características genômicas e epigenômicas e ocorreram tanto em locais de atividade transcricional bem como regiões de possível silenciamento. Embora precisaríamos avaliar um maior número de clones ou populações experimentais para poder avaliar estatisticamente. Esse resultado corrobora com outros estudo os quais sugerem que a *SB* pode acessar regiões de cromatina compactadas e excisar eficientemente o transposon em regiões de heterocromatina (IKEDA et al., 2007; YOSHIDA et al., 2017). Em princípio, esperamos que o local de inserção não deva interferir quanto a expressão do gene de interesse, uma vez que o vetor do Sistema Ingresso apresenta insuladores em ambas as extremidades do cassete que agem como uma barreira contra o silenciamento do gene e ação de *enhancers*. Porém, conforme já mencionado, a atividade desses insuladores não foi testada nessa construção.

De todos os clones expressando Pró-trombina, o clone 28 é o mais interessante para o grupo em termos de produção de proteína em escala industrial. Assim, entender as razões pelas quais esse clone tem maior expressão pode ajudar a estabelecer adequações desejáveis ao Sistema Ingresso. No entanto, estabelecer essas razões não é uma tarefa fácil. Não é possível afirmar se a maior expressão do gene de interesse no clone 28, em nível de RNA e proteína, é o resultado do número maior de inserções ou do local dessas inserções no genoma. Talvez esses dois fatores estejam interferindo. De qualquer forma, é possível pensar que a chance de ter uma maior expressão em células com 10 cópias por genoma ou mais seria maior do que com apenas 2 a 6 cópias que foram encontradas para os demais clones. Então, essa

característica de gerar um grande número de inserções por célula seria desejável para os Sistema Ingresso, o que com a *SB* não hiperativa talvez não seja possível obter de forma eficiente. É possível que o número de cópias maior para o clone 28 tenha sido adquirido devido a persistência da *SB* por essa população. Além da remobilização, a presença da *SB* ativa pode levar a um aumento do número de cópias. Esse aumento do número de cópias pode acontecer, por exemplo, quando cópias do cassete mobilizadas durante a replicação do DNA, saindo da fita já replicada, são reparadas por recombinação de fitas homólogas usando a fita irmã que ainda retém a cópia original. Assim, essa persistência da *SB* no clone 28 parece ter mimetizado uma situação de transposase mais ativa. Num primeiro momento pode parecer interessante, porém, a presença da *SB* pode ser um problema para a expressão estável da proteína de interesse ao longo do tempo, uma vez que a transposase pode levar a instabilidade genômica e a remobilização do cassete pode levar a saída de um local de alta atividade transcricional e ser inserido em um local de silenciamento diminuindo a expressão da proteína de interesse. Assim, essa descoberta é muito importante para planejar os ensaios futuros do grupo, levando ao desenho de alternativas para a entrega da transposase, como na forma de proteína (expressa em bactérias). Os dados também indicam, como já esperado, que o uso de transposases mais eficientes com capacidade de inserir um maior número de cópias do cassete por célula poderá levar a um ganho de expressão considerável da proteína de interesse.

Com o intuito de melhorar o Sistema Ingresso, novas transposases foram selecionadas para verificar a atividade de integração em células mamíferas. Desde o início do projeto, essa foi uma proposta arriscada visto que não havia comprovação de que as transposases escolhidas apresentassem atividade em células de mamíferos. As dificuldades nas clonagens para obter os vetores finais acabaram postergando muito o início dos testes de transfecção. Dentre os três transposons escolhidos, consideramos o *Tp2* como o mais promissor, e por isso suas TIRs foram sintetizadas junto com o fragmento do vetor. Além disso, após a finalização dos vetores *Tp2*, os equipamentos de citometria de fluxo disponíveis no Instituto estavam com problemas e a situação foi agravada pela pandemia da Covid-19. Com os ensaios realizados até o momento, não podemos afirmar que a *Tp2* apresentou atividade de mobilização do transgene. No entanto, as evidências de uma população enriquecida com GFP nas amostras transfectadas com *Tp2* sugerem que várias cópias do cassete possam estar presentes nessa subpopulação, o que seria explicado mais facilmente

por atividade de transposição do que apenas por inserções aleatórias. Como já mencionado, não seria esperado que a *Tp2*, como uma transposase coevoluindo no genoma hospedeiro, tivesse uma atividade tão elevada. Dessa forma, o objetivo inicial desse trabalho era tentar aprimorá-la. A escolha da *Tp2* foi baseada na sua similaridade em nível de proteína com a *SB* (aproximadamente 60%) que permitiria realizar modificações posteriores na proteína com base nas modificações da SB100X como uma tentativa de aprimoramento. Assim, algumas perguntas permanecem sem respostas até a presente data, o que esperamos que sejam respondidas num futuro próximo.

Concomitantemente aos trabalhos de bancada, um trabalho de bioinformática foi realizado. Através de uma abordagem biológica/evolutiva aprofundamos o conhecimento dos elementos *Merlin* em eucariotos. Foi realizada uma abordagem para estudar toda uma superfamília de transposons de DNA, que carecia de estudos desde sua primeira identificação. As buscas foram feitas pela ferramenta blast em dois *rounds* de buscas, onde todas as sequências identificadas no primeiro *round* foram usadas como *query* no segundo round. Isso aumentou as chances de encontrar sequências mais divergentes da *query* inicial. Como a contaminação de amostras de DNA derivados da microbioma de tecidos normais ou contaminação ambiental é uma ocorrência comum em estudos de sequenciamento em grande escala, as proteínas recuperadas foram avaliadas quanto a presença de bactérias pela ferramenta blastp e hits com mais de 80% de identidade com uma sequência bacteriana foram consideradas possíveis contaminações.

Neste trabalho, descrevemos uma distribuição global de elementos *Merlin* em animais, fungos, plantas e protistas, relatando pela primeira vez sua presença em Rhodophyceae, Metamonada, Discoba e Alveolata. Nesses quatro grupos, procedemos análises detalhadas das cópias para identificação das características básicas das famílias como identificação das TIRs e TSDs e avaliação da conservação do motivo DDE. Identificamos uma grande variedade de famílias de *Merlin* potencialmente ativas, com sinais de mobilização recentes (TIRs e TSDs conservadas), algumas coabitando o mesmo genoma.

Interessantemente, algumas famílias descritas possuem TIRs altamente imperfeitas, uma característica nunca relatada para essa superfamília. A presença de TSDs conservadas nesses casos certifica que as inserções são recentes e que os *mismatches* encontrados entre as TIRs 5' e 3' foram gerados anteriormente à

inserção, ou seja, essas TIRs são reconhecidas pela transposase. Isso pode representar um arranjo subótimo para transposição, assim como é sugerido para elementos *Tc1/Mariner* (NAUMANN e REZNIKOFF, 2002), diminuindo efeitos deletérios no hospedeiro. Transposases que apresentam TIRs imperfeitas são boas candidatas para serem usadas em sistemas de transposição uma vez que modificando as sequências das TIRs, tornando-as iguais, há um aumento da atividade de mobilização (PLEDGER et al., 2004; PLEDGER e COATES, 2005).

Também foram identificadas repetições *in tandem* internas em algumas famílias, uma associação que, embora descrita para outros elementos, não havia sido descrita para elementos *Merlin*. A existência de repetições em várias famílias e cópias de *Merlin* indica que este TE poderia ajudar a espalhar repetições *in tandem* por transposição como protossatélites que podem ser amplificados e homogeneizados em seguida, como o modelo sugerido por Paço et al (2019). O fato de os elementos *Merlin* apresentarem repetições *in tandem* em suas sequências em cópias conservadas pode ser um indicativo da alta capacidade de carga desses elementos o que os torna promissores para o uso em sistemas de transposição. O sistema *SB*, por exemplo, apresenta uma limitação na sua capacidade de carga e a sua eficiência de transposição é inversamente proporcional ao tamanho da carga (GEURTS et al., 2003; SANDOVAL-VILLEGAS et al., 2021).

Também foram observadas sequências relacionadas aos elementos *Merlin* com capacidade codificadora, mas sem evidência de capacidade de mobilização (sem qualquer sinal da presença de TIRs e TSDs) e podem ser produtos de domesticação. O número de genes derivados de TEs identificados tem aumentado consideravelmente com o avanço dos estudos dos TEs em novos genomas (JANGAM et al, 2017). A comprovação, porém, de que essas cópias de *Merlin* representam genes domesticados necessitaria de estudos específicos nesses organismos.

As árvores evolutivas sustentam que *Merlin* é provavelmente uma superfamília antiga, com eventos iniciais de diversificação e perdas secundárias, embora reinvasões repetidas provavelmente tenham ocorrido em alguns grupos, o que também explicaria sua diversidade e distribuição descontínua. Embora na árvore com apenas as sequências caracterizadas nesse artigo, *Merlin* tenha formado um grupo monofilético relacionado ao grupo *IS1016*, a árvore completa de sequências *Merlin* apresentou uma baixa resolução filogenética. Assim, não podemos descartar a

possibilidade de que essa superfamília seja o produto de múltiplas transferências horizontais ancestrais de sequências de inserção procarióticas relacionadas.

Além disso, este trabalho consistiu no primeiro relato de um transposon de DNA em flagelados da ordem Kinetoplastida. Até então, apenas retrotransposons foram descritos para essa ordem de protozoários, o que representou um resultado muito importante dessa tese. A transposase de *Merlin* foi encontrada conservada em *Bodo saltans* e *Perkinsela* sp., embora esteja ausente em tripanossomatídeos, o grupo mais estudado dentro de Kinetoplastida que incluem as espécies causadoras de doenças, *T. cruzi*, *T. brucei* e o gênero *Leishmania*. A presença de *Merlin* em *Perkinsela* sp foi confirmada por PCR usando primers que anelam no transposon e em genes próximos (sabidamente genes de Kinetoplastida). Devido ao baixo número de cópias encontrado nesses dois genomas e a possibilidade de contaminação essa etapa de confirmação foi essencial para a confiabilidade dos resultados. Não conseguimos colaboração para comprovar a presença de *Merlin* em *B. saltans*, mas a confirmação em *Perkinsela* sp. já foi suficiente para termos confiança nos dados encontrados. Embora a região codificadora da transposase tenha sido encontrada preservada nas duas espécies, nenhum indício de TIRs e TSDs foram observados. Durante o ciclo de vida de um transposon, as mutações devem acontecer tanto na região codificadora quanto nas TIRs e TSDs e o esperado para um transposon em processo de degeneração é que a região codificadora acumule mutações e perca sua capacidade codificadora antes mesmo de perder completamente suas TIRs e TSDs. O melhoramento da montagem desses dois genomas talvez revele a presença de cópias completas. Porém, essa conservação da transposase de *Merlin* pode ser resultado de domesticação dessa proteína para alguma função nesses organismos. Também foi observada a presença de ORFs potenciais com alguma sobreposição às cópias de *Merlin* na fita senso ou antisenso. Embora não haja evidência da expressão dessas ORFs, esses dados podem indicar novos genes derivados de regiões de *Merlin*. Nos cinetoplastídeos os genes são organizados em Unidades Policistrônicas, e de forma geral não tem promotores e íntrons (CAMPBELL, THOMAS e STURM, 2003). Nesse contexto, o processo de expressão gênica dos cinetoplastídeos em RNAs policistrônicos poderia facilitar o surgimento de novos genes pela inserção de TEs dentro de Unidades Policistrônicas, que seriam transcritos juntamente com os genes adjacentes. Assim, propomos que *Merlin*, pode servir como matéria-prima para a emergência de genes em protistas.

Em resumo, nesta tese, os transposons de DNA foram explorados dentro de um sistema de transposição buscando o entendimento do seu funcionamento dentro desse sistema e sua real aplicabilidade para a transformação genética de células CHO visando a produção em larga escala de proteínas de interesse biotecnológico. Os dados obtidos sugerem que uma transposase mais eficiente que a *SB11* deveria ser usada nesse sistema, bem como é necessário desenvolver alternativas de entrega dessa transposase (na forma de RNA ou proteína) para o estabelecimento de linhagens celulares produtoras estáveis. Embora os dados obtidos até o momento não tenham comprovado a atividade da *Tp2*, ela ainda é uma boa candidata e em breve teremos essa resposta. Ainda, com a experiência adquirida pelo grupo, outras transposases podem ser testadas. Neste trabalho, também foi possível aumentar o conhecimento dos elementos da superfamília *Merlin* com uma metodologia que englobou buscas em todos os eucariotos, tendo em vista a importância do estudo de uma superfamília pouco explorada e também a relevância do entendimento dessas sequências para desenvolvimento de ferramentas. Até o momento, por exemplo, não foi descrito nenhum sistema de transposição utilizando elementos da família *Merlin* e nossos dados indicam que algumas famílias podem ser boas candidatas. O estudo das características básicas de uma determinada superfamília/família, como o realizado nesse trabalho, é fundamental não somente para o entendimento do papel dessas sequências nos genomas, mas também é útil para entender seu potencial para aplicação biotecnológica.

6 PERSPECTIVAS

- Realizar ensaios de FISH (*Fluorescence In Situ Hybridization*) para verificar o local de inserção da *SB* e ajudar a estabelecer o número de cópias por células.
- Avaliar o cariótipo da célula CHO selvagem e das populações de interesse, como o clone 28, para verificar se a conservação da estrutura cromossômica .
- Repetir a PCR splinkerette do clone 28 usando outras enzimas para ligação dos adaptadores para confirmar os sítios de inserção.
- Realizar a expressão das proteínas *SB* e *Tp2* em bactérias a partir dos vetores pET28a para testar a transfecção em célula CHO como uma alternativa a transfecção do vetor *helper*, evitando assim a atividade indesejada da transposase a longo prazo.
- Repetir a PCR splinkerette na população de células transfectadas com *Tp2* para verificar a atividade da transposase.
- Selecionar as subpopulações de *Tp2* para avaliação.

REFERÊNCIAS

AHMED, M., e LIANG, P. Transposable elements are a significant contributor to tandem repeats in the human genome. **Comp and funct genomics**, v. 2012, 947089, 2012.

AMBERGER, M.; IVICS, Z. Latest Advances for the Sleeping Beauty Transposon System: 23 Years of Insomnia but Prettier than Ever: Refinement and Recent Innovations of the Sleeping Beauty Transposon System Enabling Novel, Nonviral Genetic Engineering Applications. **Bioessays**, v.42, n. 11, 2020.

ATES, I.; RATHBONE, T.; STUART, C.; BRIDGES, P. H.; COTTLE, R. N. Delivery Approaches for Therapeutic Genome Editing and Challenges. **Genes** (Basel), v. 11, n. 10, p. 1113, 2020.

BAIR, C. R.; NARAYAN, P. K. L.; KAJON, A. E. The tripartite leader sequence is required for ectopic expression of HAdV-B and HAdV-E E3 CR1 genes. **Virology**, v. 505, p. 139-147, 2017.

BAHR, S.M.; BORGSCHULTE, T.; KAYSER, K. J.; LIN, N. Using microarray technology to select housekeeping genes in Chinese hamster ovary cells. **Biotechnol. Bioeng**, v.104, p. 1041–1046, 2009.

BALASUBRAMANIAN, S.; RAJENDRA, Y.; BALDI, L.; HACKER, D. L.; WURM, F. M. Comparison of three transposons for the generation of highly productive recombinant CHO cell pools and cell lines. **Biotechnol Bioeng**, v. 113, n. 6, p. 1234-43, 2016.

BANDARANAYAKE, A. D.; ALMO, S.C. Recent advances in mammalian protein production. **FEBS Lett**, v. 588, n.2, p. 253-60, 2014.

BARKESS, G.; WEST, A. G. Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. **Epigenomics**, v. 4, n. 1, p. 67-80, 2012.

BAUS, J.; LIU, L.; HEGGESTAD, A. D.; SANZ, S.; FLETCHER, B. S. Hyperactive transposase mutants of the Sleeping Beauty transposon. **Mol. Ther**, v. 12, p. 1148–1156, 2005.

BERTOCCI, N. A.; TORRES, F. P.; GARNERO, A.; GUNSKI, R. J.; WALLAU, G. L. Evolutionary history of the mariner element galluhop in avian genomes. **Mobile DNA**, v. 8, n. 11, 2017

BOURQUE, G.; BURNS, K. H.; GEHRING, M.; GORBUNOVA, V.; SELUANOV, A.; HAMMELL, M.; IMBEAULT, M.; IZSVÁK, Z.; LEVIN, H. L.; MACFARLAN, T. S.; MAGER, D. L.; FESCHOTTE, C. Ten things you should know about transposable elements. **Genome Biol**, v. 19, n. 1, p. 199, 2018.

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

BUCCITELLI, C.; SELBACH, M. mRNAs, proteins and the emerging principles of gene expression control. **Nature Reviews Genetics**, v. 21, n.10, p. 630–644, 2020.

BÜSSOW, K. Stable mammalian producer cell lines for structural biology. **Curr Opin Struct Biol**, v. 32, p. 81-90, 2015.

BRILLET, B.; BIGOT, Y.; AUGÉ-GOUILLOU, C. Assembly of the Tc1 and mariner transposition initiation complexes depends on the origins of their transposase DNA binding domains. **Genetica**, v. 130, n. 2, p. 105–120, 1997.

BUTLER, M.; SPEARMAN, M. The choice of mammalian cell host and possibilities for glycosylation engineering. **Curr Opin Biotechnol**, v. 30, p. 107–112, 2014.

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

CAVALCANTE, M. G.; NAGAMACHI, C. Y.; PIECZARKA, J. C.; NORONHA, R. C. R. Evolutionary insights in Amazonian turtles (Testudines, Podocnemididae): co-location of 5S rDNA and U2 snRNA and wide distribution of Tc1/Mariner. **Biol Open**, v. 9, n. 4, p. 049817, 2020.

CHAVEZ, C. L.; CALOS, M. P. Therapeutic applications of the Φ C31 integrase system. **Curr Gene Ther**. V. 11, n. 5, p. 375–81, 2011.

CHEN, X.; CUI, J.; YAN, Z.; ZHANG, H.; CHEN, X.; WANG, N.; SHAH, P.; DENG, F.; ZHAO, C.; GENG, N.; LI, M.; DENDULURI, S. K.; HAYDON, R. C.; LUU H. H.; REID, R. R.; HE T. Sustained high level transgene expression in mammalian cells mediated by the optimized piggyBac transposon system. **Genes Dis**, v. 2, n. 1, p. 96–105, 2015.

CLAEYS BOUUAERT, C.; LIPKOW, K.; ANDREWS, S. S.; LIU, D.; CHALMERS, R. The autoregulation of a eukaryotic DNA transposon. **Elife**, v. 2, p. e00668, 2013.

CLARK, K. J.; CARLSON, D. F.; LEAVER, M. J.; FOSTER, L. K.; FAHRENKRUG, S. C. Passport, a native Tc1 transposon from flatfish, is functionally active in vertebrate cells. **Nucleic Acids Res**, v. 37, n. 4, p. 1239–47, 2009.

COLLIER, L. S.; CARLSON, C. M.; RAVIMOHAN, S.; DUPUY, A. J.; LARGAESPADA, D. A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. **Nature**, v. 436, n. 7048, p. 272–276, 2005.

COUPLAND, G. et al. Sequences near the termini are required for transposition of the maize transposon Ac in transgenic tobacco plants. **PNAS**, v. 86, n. 23, p. 9385–8, 1989.

DA SILVA, K. R.; MARIOTTO, S.; CENTOFANTE, L.; PARISE-MALTEMPI, P. P. Chromosome mapping of a Tc1-like transposon in species of the catfish *Ancistrus*. **Comp Cytogenet**, v. 11, n. 1, p. 65–79, 2017.

DENİZ, Ö.; FROST, J. M.; BRANCO, M. R. Regulation of transposable elements by DNA modifications. **Nature reviews Genetics**, v. 20, n. 7, p. 417–431, 2019.

DE KONING, A. P. J.; GU, W.; CASTOE, T. A.; BATZER, M. A.; POLLOCK, D. D. Repetitive elements may comprise over two-thirds of the human genome. **PLoS Genet**. v. 7, n. 12, 2011.

DE JONG, J.; AKHTAR, W.; BADHAI, J.; RUST, A. G.; RAD, R. Chromatin Landscapes of Retroviral and Transposon Integration Profiles. **PLoS Genet**, v. 10, n. 4, 2014.

DING, S.; WU, X.; LI, G.; HAN, M.; ZHUANG, Y.; XU, T. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. **Cell**, v. 122, n. 3, p. 473–83, 2005.

DOHERTY, J. E.; HUYE, L. E.; YUSA, K.; ZHOU, L.; CRAIG, N. L.; WILSON, M. H. Hyperactive piggyBac gene transfer in human cells and in vivo. **Hum Gene Ther**, v. 23, n. 3, p. 311-20, 2012.

DOTTO, B. R., CARVALHO, E. L., DA SILVA, A. F., DEZORDI, F. Z., PINTO, P. M., CAMPOS, T. L., REZENDE, A. M., & WALLAU, G. (2018). HTT-DB: new features and updates. **Database (Oxford)**, bax102, 2018.

DUPEYRON, M.; BARIL, T.; BASS, C.; HAYWARD, A. Phylogenetic analysis of the Tc1/mariner superfamily reveals the unexplored diversity of pogo-like elements. **Mob DNA**, v. 11, n. 21, 2020.

DUBIN, M. J.; MITTELSTEN SCHEID, O.; BECKER, C. Transposons: a blessing curse. *Current opinion in plant biology*, 42, 23–29, 2018.

ELICK, T.A.; BAUSER, C. A.; FRASER, M. J. Excision of the piggyBac transposable element in vitro is a precise event that is enhanced by the expression of its encoded transposase. **Genetica**, v. 98, p. 33–41, 1996.

FESCHOTTE, C. *Merlin*, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. **Mol biol evol**, v. 21, n. 9, p. 1769–80, 2004.

FESCHOTTE, C.; PRITHAM, E. J. DNA transposons and the evolution of eukaryotic genomes. **Annua rev genet**, v. 41, p. 331–68, 2007.

FINNEGAN, D. J. Eukaryotic transposable elements and genome evolution. **Trends in genetics**, v. 5, p.103–7, 1989.

GAO, B.; WANG, W.; WU, H.; CHEN, C.; SHEN, D.; WANG, S.; CHEN, W.; ZHANG, L.; CHAN, S.; SONG, C. Changes in Skeletal Muscle and Body Weight on Sleeping Beauty Transposon-Mediated Transgenic Mice Overexpressing Pig mIGF-1. **Biochem Genet**, v. 56, p. 341–355, 2018.

GAO, B.; SANG, Y.; ZONG, W.; DIABY, M.; SHEN, D.; WANG, S.; WANG, Y.; CHEN, C.; SONG C. Evolution and domestication of Tc1/mariner transposons in the genome of African coelacanth (*Latimeria chalumnae*). **Genome**, v. 63, n. 8, p. 375-386, 2020.

GEURTS, A. M.; YANG, Y.; CLARK, K. J.; LIU, G.; CUI, Z.; DUPUY, A. J.; BELL, J. B.; LARGAESPADA, D. A.; HACKETT, P. B. Gene transfer into genomes of human cells by the Sleeping Beauty transposon system. **Mol. Ther**, v. 8, p. 108–117, 2003.

GUO, Y.; XIAO, P.; LEI, S.; DENG, F.; XIAO, G. G.; LIU, Y.; CHEN, X.; LI, L.; WU, S.; CHEN, Y.; JIANG, H.; TAN, L.; XIE, J.; ZHU, X.; LIANG, S.; DENG, H. How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. **Acta biochim biophys Sin**, v. 40, n. 5, p. 426–436, 2008.

GUO, Y.; ZHANG, Y.; HU, K. Sleeping Beauty transposon integrates into non-TA dinucleotides. **Mobile DNA**, v. 9, n. 8, 2018.

HARTLEY, J. L. Why Proteins in Mammalian Cells? In: Hartley J. (eds) Protein Expression in Mammalian Cells. *Methods in Molecular Biology (Methods and Protocols)*, vol 801. Humana Press, 2012.

HAGEDORN, C.; ANTONIOU, M. N.; LIPPS, H. J. Genomic cis-acting Sequences Improve Expression and Establishment of a Nonviral Vector. *Mol ther Nucleic acids*, v. 2, n. 8, p. e118, 2013.

HICKMAN, A. B.; DYDA, F. DNA Transposition at Work. **Chem Rev.**, v. 116, n. 20, p. 12758-12784, 2016.

HILLIARD, W.; MACDONALD, M. L.; LEE, K. H. Chromosome-scale scaffolds for the Chinese hamster reference genome assembly to facilitate the study of the CHO epigenome. **Biotechnol bioeng**, v. 117, n. 8, p. 2331–2339, 2020.

HUANG, X.; GUO, H.; TAMMANA, S.; JUNG, Y. C.; MELLGREN, E.; BASSIt, P.; CAO, Q.; TU, Z. J.; KIM, Y. C.; EKKER, S. C.; WU, X.; WANG, S. M.; ZHOUp, X. Gene transfer efficiency and genome-wide integration profiling of Sleeping Beauty, Tol2, and piggyBac transposons in human primary T cells. **Mol Therapy**, v. 18 n. 10, p. 1803–1813, 2010a.

HUANG, X., HALEY, K., WONG, M., GUO, H., LU, C., WILBER, A., ZHOU, X. Unexpectedly high copy number of random integration but low frequency of persistent expression of the Sleeping Beauty transposase after trans delivery in primary human T cells. **Human gene therapy**, v. 21, n. 11, p. 1577–1590, 2010b.

HUDECEK, M.; IZSVÁK, Z.; JOHNNEN, S.; RENNER, M.; THUMANN, G.; IVICS, Z. Going non-viral: the SleepingBeauty transposon system breaks on through to the clinical side. **Crit Rev Biochem Mol Biol**, v. 52, n. 4, p. 355-380, 2017.

HUDECEK, M.; IVICS, Z. Non-viral therapeutic cell engineering with the Sleeping Beauty transposon system. **Curr Opin Genet Dev**, v. 52, p. 100–108, 2018.

IKEDA, R., KOKUBU, C., YUSA, K., KENG, V. W., HORIE, K., TAKEDA, J. Sleeping beauty transposase has an affinity for heterochromatin conformation. **Molecular and cellular biology**, v. 27, n. 5, p. 1665–1676, 2007.

IVICS, Z.; HACKETT, P. B.; PLASTERK, R. H.; IZSVÁK, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. **Cell**, v. 91, p. 501-10, 1997.

IVICS, Z.; IZSVÁK, Z. Sleeping beauty transposition. **Microbiol spectrum** v. 3, n. 2, p. 3-0042, 2014.

JANGAM, D.; FESCHOTTE, C.; BETRÁN E. Transposable element domestication as an adaptation to evolutionary conflicts. **Trends Genet**, v. 33, p. 817–831, 2017.

JAYAPAL, K. P.; WLASCHIN, K. F.; HU, W. S.; YAP, M. G. Recombinant protein therapeutics from CHO cells—20 years and counting. **Chem Eng Prog**, v. 103, p. 40–47, 2007.

JOHNSON, E. T.; DOWD, P. F. A non-autonomous insect piggyBac transposable element is mobile in tobacco. **Mol. Genet. Genom**, v. 289, p. 895–902, 2014.

JOLY-LOPEZ, Z.; BUREAU, T. E. Exaptation of transposable element coding sequences. **Curr Opin Genet Dev**, v. 49, p. 34-42, 2018.

JURKA, J. Repbase update: a database and an electronic journal of repetitive elements. **Trends Genet**, v. 16, n. 9, p. 418-20, 2000.

KARAKÜLAH, G.; SUNER, A. PlanTEEnrichment: A tool for enrichment analysis of transposable elements in plants. **Genomics**. v. 109, p. 336–340, 2017.

KATTER, K.; GEURTS, A. M.; HOFFMANN, O.; MÁTÉS, L.; LANDA, V.; HIRIPI, L.; MORENO, C.; LAZAR, J.; BASHIR, S.; ZIDEK, V.; POPOVA, E.; JERCHOW, B.; BECKER, K.; DEVARAJ, A.; WALTER, I.; GRZYBOWSKI, M.; CORBETT, M.; FILHO, A. R.; HODGES, M. R.; BADER, M.; IVICS, Z.; JACOB, H. J.; PRAVENEK, M.;

BOSZE, Z.; RÜLICHE, T.; IZSVÁK, Z. Transposon-mediated transgenesis, transgenic rescue, and tissue-specific gene expression in rodents and rabbits. **FASEB J**, v. 27, n. 3, p. 930-41, 2013.

KAWAKAMI, K.; SHIMA, A.; KAWAKAMI, N. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. **Proc Natl Acad Sci U S A**, v. 97, n. 21, p. 11403-8, 2000.

KAWAKAMI, K. Tol2: a versatile gene transfer vector in vertebrates. **Genome Biol**, v. 8 Suppl 1(Suppl 1):S7, 2007.

KIM, J.Y.; KIM, Y.G.; LEE, G.M. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. **Appl Microbiol Biotechnol**, 93, p. 917-30, 2012

KOLACSEK, O.; KRÍZSIK, V.; SCHAMBERGER, A.; ERDEI, Z.; APÁTI, A.; VÁRADY, G.; MÁTÉS, L.; IZSVÁK, Z.; IVICS, Z.; SARKADI, B.; ORBÁN, T. I. Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. **Mobile DNA**, v. 2, n. 5, 2011.

KOLACSEK, O.; IZSVÁK, Z.; IVICS, Z.; SARKADI, B.; ORBÁN, T. I. Quantitative analysis of DNA transposon-mediated gene delivery: the Sleeping Beauty system as an example. In: Genomics III - Methods, Techniques and Applications. iConcept Press Ltd., Hong Kong, 2014a.

KOLACSEK, O.; ERDEI, Z.; APÁTI, A.; SÁNDOR, S.; IZSVÁK, Z.; IVICS, Z.; SARKADI, B.; ORBÁN, T. I. Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. **Hum Gene Ther Methods**, v. 25, n. 4, p. 241–252, 2014b.

KOLACSEK, O.; ORBÁN, T. I. Transcription activity of transposon sequence limits Sleeping Beauty transposition. **Gene**, v. 15, n. 676, p 184-188, 2018.

KOGA, A.; SHIMADA, A.; KUROKI, T.; HORI, H.; KUSUMI, J.; KYONO-HAMAGUCHI, Y.; HAMAGUCHI, S. The Tol1 transposable element of the medaka fish moves in human and mouse cells. **J Hum Genet**, v. 52, n. 7, p. 628-635, 2007.

KUYSTERMANS, D.; AL-RUBEAI, M. Biopharmaceutical Products from Animal Cell Culture. In: Al-Rubeai M. (eds) Animal Cell Culture. Cell Engineering, vol 9. Springer, Cham, 2015.

LAI, T.; YANG, Y.; NG, S.K. Advances in Mammalian Cell Line Development Technologies for Recombinant Protein Production. **Pharmaceuticals**, v. 6, p. 579-603, 2013.

LALONDE, M. E.; DUROCHER, Y. Therapeutic glycoprotein production in mammalian cells. **J Biotechnol**, v. 251, p. 128–140, 2017.

LANGE, S.S.; TOMIDA, J.; BOULWARE, K.S.; BHETAWAL, S.; WOOD, R.D. The Polymerase Activity of Mammalian DNA Pol ζ Is Specifically Required for Cell and Embryonic Viability. **PLoS Genet**, v. 12, n. 1, p. e1005759, 2016.

LAMPE, D. J. Bacterial genetic methods to explore the biology of mariner transposons. **Genetic**. v. 138, n. 5, p. 499-508, 2010.

LI, X.; EWIS, H.; HICE, R. H.; MALANI, N.; PARKER, N.; ZHOU, L.; FESCHOTTE, C.; BUSHMAN, F. D.; ATKINSON, P. W.; CRAIG, N. L. A resurrected mammalian hAT

transposable element and a closely related insect element are highly active in human cell culture. **Proc Natl Acad Sci U S A**, v. 110, n. 6, p. 478–487, 2013.

LI, X.; BURNIGHT, E. R.; COONEY, A. L.; MALANI, N.; BRADY, T.; SANDER, J. D.; STABER, J.; WHEELAN, S. J.; JOUNG, J. K.; MCCRAY, P. B.; BUSHMAN, F. D.; SINN, P. L.; CRAIG, N. L. piggyBac transposase tools for genome engineering. **Proc Natl Acad Sci U S A**, v. 110, n. 25, p. E2279-87, 2013.

Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. **Bioinformatics**, 25:1754-60, 2009.

LIU, G., GEURTS, A. M., YAE, K., SRINIVASAN, A. R., FAHRENKRUG, S. C., LARGAESPADA, D. A., TAKEDA, J., HORIE, K., OLSON, W. K., & HACKETT, P. B. Target-site preferences of Sleeping Beauty transposons. **Journal of molecular biology**, v. 346, n. 1, p. 161–173, 2005.

LIU, D. et al. Functional dissection of the cis-acting sequences of the Arabidopsis transposable element Tag1 reveals dissimilar subterminal sequence and minimal spacing requirements for transposition. **Genetics**, v. 157, n. 2, p. 817–30, 2001.

LORETO, E.; DEPRÁ, M.; DIESEL, J. F. PANZERA, Y.; VALENTE-GAIESKY, V. Drosophila relics hobo and hobo-MITEs transposons as raw material for new regulatory networks. **Genetics and molecular biology**, v. 41 (1 suppl 1), p. 198-205, 2018.

LUO, G.; IVICS, Z.; IZSVÁK, Z.; BRADLEY, A. Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. **Proc Natl Acad Sci U S A**, v. 95, n. 18, p. 10769–10773, 1998.

MALEKI, A. F.; SEKHAVATI, M. H. Application of phiC31 integrase system in stem cells biology and technology: a review, **Frontiers in Life Science**, v. 11, n.1, p. 1-10, 2018.

MATASCI, M.; HACKER, D.L.; BALDI, L.; WURM, F.M. Recombinant therapeutic protein production in cultivated mammalian cells: current status and future prospects. **Drug Discov Today: Technol**, v. 5, n 2-3, p. 37-42, 2008.

MATASCI, M.; BALDI, L.; HACKER, D.L.; WURM, F.M. The PiggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability **Biotechnol Bioeng**, v. 108, n. 9, p. 2141–2150, 2011.

MÁTÉS, L.; CHUAH, M.K.; BELAY, E.; JERCHOW, B.; MANOJ, N.; ACOSTA-SANCHEZ, A.; GRZELA, D.P.; SCHMITT, A.; BECKER, K.; MATRAI, J.; MA, L.; SAMARA-KUKO, E.; GYSEMANS, C.; PRYPUTNIEWICZ, D.; MISKEY, C.; FLETCHER, B.; VANDENDRIESSCHE, T.; IVICS, Z.; IZSVÁK, Z. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. **Nat Genet**, v. 41, n. 6, p. 753-61, 2009.

MING-BING, Z.; HUI, H.; MISKEY, C.; LAZAROW, K.; IVICS, Z.; KUNZE, R.; GUOJUN, Y.; IZSVÁK, Z.; DING-QIN, T. Transposition of the bamboo Mariner-like element Ppmar1 in yeast, **Mole Phylogenet Evol**, v. 109, p. 367-374, 2017.

MISKEY, C.; IZSVÁK, Z.; PLASTERK, R. H.; IVICS, Z. The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. **Nucleic Acids Res**, v. 31, n. 23, p 6873-81, 2003.

MISKEY, C.; PAPP, B.; MÁTÉS, L.; SINZELLE, L.; KELLER, H.; IZSVÁK, Z.; IVICS, Z. The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. **Molecular and cellular biology**, v. 27, p. 12, p. 4589–4600, 2007.

MITRA, R.; FAIN-THORNTON, J.; CRAIG, N.L. piggyBac can bypass DNA synthesis during cut and paste transposition. **EMBO J**, v. 27, p. 1097–1109, 2008.

MUÑOZ-LÓPEZ, M.; GARCÍA-PÉREZ, J.L. DNA Transposons: Nature and Applications in Genomics. **Current Genomics**, v. 11, n. 2, p. 115-128, 2010.

NAUMANN, T. A.; REZNIKOFF, W. S. Tn5 transposase with an altered specificity for transposon ends. **J Bacteriol**, v. 184, p. 233–40, 2002.

NI, J.; CLARK, K.J.; FAHRENKRUG, S.C.; EKKER, S.C. Transposon tools hopping in vertebrates. **Briefings in Functional Genomics and Proteomics**, v. 7, n. 6, p. 444–453, 2008.

NOH, S. M.; SHIN, S.; LEE, G. M. Cell Line Development for Therapeutic Protein Production. In *Cell Culture Engineering* (Lee, G. M., and Kildegaard, H. F., Ed.), pp 23–47, WileyVCH Verlag GmbH & Co. KGaA, Weinheim, 2019.

OHBAYASHI, F.; BALAMOTIS, M. A.; KISHIMOTO, A.; AIZAWA, E.; DIAZ, A.; HASTY, P.; GRAHAM, F. L.; CASKEY, C. T., & MITANI, K. Correction of chromosomal mutation and random integration in embryonic stem cells with helper-dependent adenoviral vectors. **Proc Natl Acad Sci U S A**, v. 102, n. 38, p. 13628–13633, 2005.

O'FLAHERTY, R.; BERGIN, A.; FLAMPOURI, E.; MOTA, L. M.; OBAIDI, I.; QUIGLEY, A.; XIE, Y.; BUTLER, M. Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing. **Biotechnol Adv**, v. 43, p. 107552, 2020.

OLIVER, K. R.; GREENE, W. K. Transposable elements: powerful facilitators of evolution. **BioEssays**, v. 31, n. 7, p. 703–14, 2009.

OCHMANN, M.T.; IVICS, Z. Jumping ahead with Sleeping Beauty: Mechanistic Insights into Cut-and-Paste Transposition. **Viruses**, v. 13, n. 76, 2021.

OVERBERGH, L.; VIG, S.; COUN, F.; MATHIEU, C. Chapter 4 - Quantitative Polymerase Chain Reaction. In: Patrinos, G. P., editor. *Molecular Diagnostics*. Academic Press; San Diego, CA pp. 41-58, 2017.

PAÇO, A.; FREITAS, R.; VIEIRA-DA-SILVA, A. Conversion of DNA Sequences: from a transposable element to a tandem repeat or to a gene. **Genes (Basel)**, v. 10, 2019.

PALAZZO, A.; ESCUDER, E.; D'ADDABBO, P.; LOVERO, D.; MARSANO, R. M. A genomic survey of Tc1-mariner transposons in nematodes suggests extensive horizontal transposon transfer events. **Mol Phylogenet Evol**, v. 158, p. 107090, 2021.

PETRAKIS, S.; RASKO, T.; MÁTÉS, L.; IVICS, Z.; IZSVÁK, Z.; KOUZI-KOLIAKOU, Z.; KOLIAKOS, G. Gateway-compatible transposon vector to genetically modify human embryonic kidney and adipose-derived stromal cells. **Biotechnol J**, v. 7, n. 7, p. 891–897, 2012.

PLEDGER, D. W.; FU, Y. Q.; COATES, C. J. Analyses of cis -acting elements that affect the transposition of Mos1 mariner transposons in vivo. **Molecular genetics and genomics**, v. 272, n. 1, p. 67–75, 2004.

PLEDGER, D. W.; COATES, C. J. Mutant Mos1 mariner transposons are hyperactive in *Aedes aegypti*. **Insect biochemistry and molecular biology**, v. 35, n. 10, p. 1199–1207, 2005.

QU, L.; WANG, L.; ZHU, X. Global mapping of binding sites for phic31 integrase in transgenic maden-darby bovine kidney cells using ChIP-seq. **Hereditas**, v. 156, n. 3, 2019.

QUERQUES, I.; MADES, A.; ZULIANI, C.; MISKEY, C.; ALB, M.; GRUESO, E.; MACHWIRTH, M.; RAUSCH, T.; EINSELE, H.; IVICS, Z.; HUDECEK, M.; BARABAS, O. A highly soluble Sleeping Beauty transposase improves control of gene insertion. **Nat Biotechnol**, v. 37, n. 12, p. 1502–1512, 2019.

RAUSCHHUBER, C.; EHRHARDT, A. RNA interference is responsible for reduction of transgene expression after Sleeping Beauty transposase mediated somatic integration. **PloS one**, v. 7, n. 5, p. e35389, 2012.

RAMKE, M.; LEE, J. Y.; DYER, D. W.; SETO, D.; RAJAIYA, J.; CHODOSH, J. The 5'UTR in human adenoviruses: leader diversity in late gene expression. **Sci Rep**, v. 7, n. 1, p. 618, 2017.

RIBEIRO, Y. C.; ROBE, L. J.; VELUZA, D. S.; DOS SANTOS, C. M. B.; LOPES, A. L. K.; KRIEGER, M. A.; LUDWIG, A. Study of VIPER and TATE in kinetoplastids and the evolution of tyrosine recombinase retrotransposons. **Mob DNA**, v. 5, p.10:34, 2019.

RIVELLA, S.; CALLEGARI, J. A.; MAY, C.; TAN, C. W.; SADELAIN, M. The cHS4 insulator increases the probability of retroviral expression at random chromosomal integration sites. **Journal of virology**, v. 74, n. 10, p. 4679–4687, 2000.

ROBERTSON, H.M.; SOTO-ADAMES, F.N.; WALDEN, K.K.; AVANCINI, R.M.; LAMPE, D.J. The *mariner* transposons of animals: horizontally jumping genes. In: Kado CI, editor. **Horizontal gene transfer**. Academic Press; San Diego, CA pp. 173–185, 2002.

RUTHERFORD, K.; PARKHILL, J.; CROOK, J. Artemis: sequence visualization and annotation. **Bioinformatics**, v. 16, p. 944–945, 2000.

SANDOVAL-VILLEGAS, N.; NURIEVA, W.; AMBERGER, M.; IVICS, Z. Contemporary Transposon Tools: A Review and Guide through Mechanisms and Applications of Sleeping Beauty, piggyBac and Tol2 for Genome Engineering. **Int. J. Mol. Sci**, v. 22, n. 5084, 2021.

SANG, Y.; GAO, B.; DIABY, M. Incomer, a DD36E family of Tc1/mariner transposons newly discovered in animals. **Mobile DNA**, v. 10, n. 45, 2019.

SCHAACK, S.; GILBERT, C.; FESCHOTTE, C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology and Evolution*, v. 30, n. 10, p. 1–10, 2010.

STADEN, R.K.; BEAL, K.F.; BONFIELD, J.K. The Staden Package, 1998. **Methods in molecular biology** (Clifton, N.J.), v. 132, p. 115–30, 2000.

SIGUIER, P.; GOURBEYRE, E.; CHANDLER M. Bacterial insertion sequences: their genomic impact and diversity. **FEMS Microbiol Rev**, v. 38, n. 5, p. 865–91, 2014.

STERN, B.; OLSEN, L. C.; TRÖßE, C.; RAVNEBERG, H.; PRYM, I. F. Improving mammalian cell factories: The selection of signal peptide has a major impact on

recombinant protein synthesis and secretion in mammalian cells. **Trends Cell Mol Biol**, v. 2, p. 1–17, 2007.

SUNDARAM, V., WYSOCKA, J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 375(1795), 20190347, 2020.

SCHEMBERGER, M. O.; NOGAROTO, V.; ALMEIDA, M. C.; ARTONI, R. F.; VALENTE, G. T.; MARTINS, C.; MOREIRA-FILHO, O.; CESTARI, M. M.; VICARI, M. R. Sequence analyses and chromosomal distribution of the Tc1/Mariner element in Parodontidae fish (Teleostei: Characiformes). **Gene**, v. 593, n. 2, p. 308-14, 2016.

TIPANEE, J.; CHAI, Y. C.; VANDENDRIESSCHE, T.; CHUAH, M. K. Preclinical and clinical advances in transposon-based gene therapy. **Biosci Rep**, v. 37, n. 6, 2017.

TURCHIANO, G.; LATELLA, M. C.; GOGOL-DÖRING, A.; CATTOGLIO, C.; MAVILIO, F.; IZSVÁK, Z.; IVICS, Z.; RECCHIA, A. Genomic Analysis of Sleeping Beauty Transposon Integration in Human Somatic Cells. **PLOS ONE**, v. 9, n. 11, 2014.

UREN, A. G.; MIKKERS, H.; KOOL, J.; VAN DER WEYDEN, L.; LUND, A. H.; WILSON, C. H.; RANCE, R.; JONKERS, J.; VAN LOHUIZEN, M.; BERNIS, A.; ADAMS, D. J. A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. **Nat. Protoc**, v. 4, p. 789–798, 2009.

URASAKI, A.; MORVAN, G.; KAWAKAMI, K. Functional Dissection of the Tol2 Transposable Element Identified the Minimal cis-Sequence and a Highly Repetitive Sequence in the Subterminal Region Essential for Transposition. **Genetics**, v. 174, n. 2, p. 639–649, 2006.

VOGELMANN, CUVIER, O.; NOLLMANN, M. Roles of chromatin insulator proteins in higher-order chromatin organization and transcription regulation. **Nucleus**, v. 2, p. 358-69, 2011.

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

WALSH, G. Biopharmaceutical benchmarks. **Nat Biotechnol**, v. 32, p. 992-1000, 2014.

WALSH G. Biopharmaceutical benchmarks. **Nat Biotechnol**, v. 36, p. 1136-1145, 2018.

WANG, X. Y.; ZHANG, J. H.; ZHANG, X. Impact of Different Promoters on Episomal Vectors Harboring Characteristic Motifs of Matrix Attachment Regions. **Sci Rep**, v. 6, p. 26446, 2016.

WANG, S.; DIABY, M.; PUZAKOV, M.; ULLAH, N.; WANG, Y.; DANLEY, P.; CHEN, C.; WANG, X.; GAO, B.; SONG, C. Divergent evolution profiles of DD37D and DD39D families of Tc1/mariner transposons in eukaryotes. **Mol Phylogenet Evol**, v. 161, p. 107143, 2021.

WEST, A.G.; GASZNER, M.; FELSENFELD, G. Insulators: Many functions, many mechanisms. **Genes & Dev**, v. 1, p. 271– 288, 2002.

WICKER, T., SABOT, F., HUA-VAN, A., BENNETZEN, J. L., CAPY, P., CHALHOUB, B., FLAVELL, A., LEROY, P., MORGANTE, M., PANAUD, O., PAUX, E.,

SANMIGUEL, P., SCHULMAN, A. H. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, p. 973–82, 2007.

WILSON, M.H.; COATES, C.J.; GEORGE, A.L. PiggyBac transposon-mediated gene transfer in human cells. **Mol. Ther**, v. 15, p. 139–145, 2007.

WU, S.C.; MEIR, Y.J.; COATES, C.J.; HANDLER, A.M.; PELCZAR, P.; MOISYADI, S.; KAMINSKI, J.M. piggyBac is a flexible and highly active transposon as compared to Sleeping Beauty, Tol2, and Mos1 in mammalian cells. **PNAS**, v. 103, n. 41, p. 15008–15013, 2006.

WURM, F.M. Production of recombinant protein therapeutics in cultivated mammalian cells. **Nat Biotechnol**, v. 22, p. 1393–8, 2004.

XU, Z. J.; JIA, Y. L.; WANG, M.; YI, D. D.; ZHANG, W. L.; WANG, X. Y.; ZHANG, J. H. Effect of promoter, promoter mutation and enhancer on transgene expression mediated by episomal vectors in transfected HEK293, Chang liver and primary cells. **Bioengineered**, v. 10, n. 1, p. 548–560, 2019.

YAN, B. W., ZHAO, Y. F., CAO, W. G., LI, N., & GOU, K. M. Mechanism of random integration of foreign DNA in transgenic mice.

Transgenic research, v. 22, n. 5, p. 983–992, 2013.

YANT, S. R.; WU, X.; HUANG, Y.; GARRISON, B.; BURGESS, SM.; KAY, M. A. High-resolution genome-wide mapping of Transposon integration in mammals. **Mol Cell Biol**, v. 25, p. 2085–94, 2005.

YOSHIDA, J.; AKAGI, K.; MISAWA, R.; KOKUBU, C.; TAKEDA, J.; HORIE, K. Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. **Sci Rep**, v. 2, n. 7, p. 43613, 2017.

YUSA, K. piggyBac Transposon. **Microbiology spectrum**, v. 3, p. 3–28, 2015.

YUAN, Y. W.; WESSLER, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. **Proc Natl Acad Sci**, v. 108, p. 7884–7889, 2011.

ZHANG, L.; SANKAR, U.; LAMPE, D. J.; ROBERTSON, H. M.; GRAHAM, F. L. The Himar1 mariner transposase cloned in a recombinant adenovirus vector is functional in mammalian cells. **Nucleic Acids Res**, v. 26, n. 16, p. 3687–93, 1998.

ZHANG, X., WANG, X. Y., JIA, Y. L., GUO, X., WANG, Y. F., & WANG, T. Y. A Vector Based on the Chicken Hypersensitive Site 4 Insulator Element Replicates Episomally in Mammalian Cells. **Current gene therapy**, v. 16, n. 6, p. 410–418, 2017.

ZAYED, H.; IZSVAK, Z.; WALISKO, O.; IVICS, Z. Development of hyperactive Sleeping Beauty transposon vectors by mutational analysis. **Mol. Ther**, v. 9, p. 292–304, 2004.

ZONG, W.; GAO, B.; DIABY, M.; SHEN, D.; WANG, S.; WANG, Y.; SANG, Y.; CHEN, C.; WANG, X.; SONG, C. Traveler, a New DD35E Family of Tc1/Mariner Transposons, Invaded Vertebrates Very Recently. **Genome Biology and Evolution**, v. 12, n. 3, p. 66–76, 2020.

ZHU J. Mammalian cell protein expression for biopharmaceutical production. **Biotechnol Adv**, v. 30, n. 5, p. 1158–70, 2012.

ZÚÑIGA, R. A.; GUTIÉRREZ-GONZÁLEZ, M.; COLLAZO, N.; SOTELO, P. H.; RIBEIRO, C. H.; ALTAMIRANO, C.; LORENZO, C.; AGUILLÓN, J. C.; MOLINA, M. C.

Development of a new promoter to avoid the silencing of genes in the production of recombinant antibodies in chinese hamster ovary cells. Journal of biological engineering, v. 13, n. 59, 2019.

ANEXOS

Scripts utilizados

Os scripts foram escritos para analisar o resultado do sequenciamento NGS gerado a partir de PCR-splinkerette para identificar inserções de cassete mediadas por transposons.

Pela lógica da técnica PCR-splinkerette, o adaptador Splink2 estará na extremidade esquerda dos amplicons e a TIR 5' do transposon estará na extremidade direita. As sequências para os adaptadores Nextera P5 e P7 foram adicionadas aos primers. As leituras obtidas pela metodologia *paired-end* devem teoricamente gerar leituras polarizadas: Leituras R1 correspondentes à sequência 5' do amplicon (identidade com o adaptador Splink2 e CHO) e as leituras de R2 correspondentes à região 3' do amplicon (identidade com o TIR do transposon). A partir dessa premissa, as análises seguiram várias etapas:

1) Corte de leituras; O tamanho mínimo de leitura pode ser escolhido nesta etapa;

2) Com as leituras pareadas: as leituras R1 foram alinhadas por blastn local contra o genoma CHO (GCA_003668045.3) e as leituras R2 contra a sequência SB ou TP2 TIR5 ' (parâmetros: outfmt 6 -evaluate 0,0001 -max_target_seqs 1 -max_hsp 1).

3) Apenas os pares de leituras com *hits* positivo para ambos foram selecionados usando o **Python script 1** que gerou uma lista de IDs das leituras. Estas são consideradas as possíveis leituras corretas.

4) As leituras corretas de R2 foram então selecionadas usando a lista de IDs no **Python script 2** para gerar um arquivo multifasta dessas sequências.

5) O **Python script 3** foi usado para verificar se as leituras R2 têm a etiqueta do transposon (complementar reversa dos primeiros 10 bp da TIR 5' do transposon). Este script também verifica a presença da TSD TA esperado. A *string* para a etiqueta de cada transposon é diferente. O tamanho mínimo das leituras também pode ser definido nesse script. Esse script salvará um arquivo multifasta contendo a região de cada leitura logo após a TIR, começando com a TSD TA. O script também gera um outro arquivo com possível sequências de inserções em não-TA.

6) O **Python script 4** é usado para separar cada sequência fasta gerada no script anterior. Em teoria, cada uma dessas sequências corresponde à sequência

CHO e a primeira posição é o local de inserção. Eles são usados como query na próxima busca de blastn.

7) Para aumentar a confiabilidade da análise, essas sequências não foram alinhadas contra todo o genoma, mas apenas contra a região genômica que eles deveriam estar usando as informações do blastn do seu par de leituras R1. Para separar esta região, o **Perl script 1** foi usado para ler a tabela do blastn de R1 contra CHO e cada região genômica mais 7 kb a partir da última posição de alinhamento do blast foi salva em arquivos separados.

8) O blastn local foi realizada usando programação em Linux para executá-lo em vários arquivos:

```
Makeblastdb: for file in *_data; do makeblastdb -dbtype nucl -in $file -out $file
; done &
```

```
Blast: for file in *_query; do blastn -query $file -db ${file%_query}_data -out
${file%_query}_result -outfmt 6 -max_target_seqs 1 -max_hsps 1 -num_threads 10;
done
```

Para queries curtas (como aquelas do primeiro experimento de SB), um parâmetro deve ser adicionado para permitir a obtenção de resultados significativos [-task blastn-short]

9) Apenas alinhamentos que comecem na primeira posição da query devem ser considerados como corretos. Para o primeiro sequenciamento de SB, como as queries eram muito curtas, foram consideradas apenas sequências com alinhamento em toda extensão (tamanho do alinhamento = tamanho da query). Para separar essas sequências corretas os IDs e o tamanho das sequências foram isolados em um arquivo usando o **Python script 5** e a informação usada na próxima etapa.

10) O **Python script 6** foi usado para identificar as posições de inserção, usando a informação da posição cromossômica da região isolada de CHO acrescentada da posição inicial do alinhamento no blast da sequência final de R2. Esse script gera uma tabela com o ID de cada leitura e a posição cromossômica da inserção correspondente

11) Finalmente, para sumarizar o número de leituras para cada inserção distinta, a tabela gerada anteriormente foi analisada no linux: cut -f 2 | sort | uniq -c > sumario

#Python script 1

Python script to recover IDs whose sequences generated significant hits in two different blast searches
 # Reads R1 against CHO and reads R2 against the TIR of the transposon

```
import sys
id1 = []
count1 = 0
id2 = []
count2 = 0
file3 = open(sys.argv[1], "r") # file containing the tabular result of blast for reads R1
file4 = open(sys.argv[2], "r") # file containing the tabular result of blast for reads R2
file5 = open(sys.argv[3], "w") # output file
result = ""
for line in file3:
    line = line.rstrip()
    temp1 = line.split("\t")
    id1.append(temp1[0])
    count1 += 1
for line in file4:
    line = line.rstrip()
    temp = line.split("\t")
    id2.append(temp[0])
    count2 += 1
for i in range(count1):
    for ii in range(count2):
        if (str(id1[i]) in str(id2[ii])) & (len(id1[i]) == len(id2[ii])):
            result = id1[i] + "\n"
            print(id1[i])
            file5.write(str(result))
            break
```

#Python script 2

Python script to recover sequences from a fasta file using a list of IDs.

```
import sys
header = []
sequence = []
count = -1
file_open = open(sys.argv[1], "r") # fasta file to be analyzed
for line in file_open:
    line = line.rstrip()
    if ">" in line:
        temp = line.split(" ")
        header.append(temp[0])

        count += 1
        sequence.insert(count, "")
    else:
        sequence[count] += line
con = count + 1

print("The number of sequences in this file is:") # for checking
print(con)
result = ""
file3 = open(sys.argv[2], "r") # file containing the IDs of sequences, one per line
file4 = open(sys.argv[3], "w") # output file

for line in file3:
    line = line.rstrip()

    for i in range(con):
        if (line in header[i]) & (len(line) == len(header[i]) - 1):
```

```

result=str(header[i])+"\n"+ str(sequence[i]) + "\n"
file4.write(str(result))

```

#Python script 3

```

# Python script to recover sequences that match some criteria
#It was written to evaluate the reads R2 from NGS sequencing of PCR Splinkerette
#It was written to evaluate reads from SB insertions (first primers)
# Criteria: >247 bp, have a specific string (5' transposon tag + TSD)
import sys
header = []
sequence = []
count = -1
file1 = open(sys.argv[1], "r") #fasta file to be analyzed
for line in file1:
    line = line.rstrip()
    if ">" in line:
        temp=line.split(" ")
        header.append(temp[0])
        count+=1
        sequence.insert(count,"")
    else:
        sequence[count]+=line
con=count+1
print ("The number of sequences in this file is:") #for checking
print (con)
result=""
file2 = open(sys.argv[2], "w") # output file with TSD
file3 = open(sys.argv[3], "w") # output file without TSD
count2=0
count3=0
count4=0
string=""
for i in range (con):
    if len(sequence[i]) >247: # select reads > 247 bp (for the first primer SB designed, this size is
        important to be able to analyze
        # 9-14 bp of CHO. For the second primers the size could be smaller; i.e. >220 bp)
        string=sequence[i]
        #The read R2 is from antisense strands and starts with the reverse complement of the TIR
        region, after we should see the TA
        #and the CHO region
        if "ACTTCAACTG" in (string): # select reads that contain this string - reverse complement of first
            10 bp of the TIR- 5' transposon tag
            ini= string.find("ACTTCAACTG")
            ini=ini+10
            fin= len(sequence[i])
            count2+=1
            nucleotide = sequence[i]
            region = nucleotide[int(ini):int(fin)]
            if (region[0]=="T" and region[1] == "A"):
                count3+=1
                result = str(header[i]) + "\n"+str(region)+"\n"
                file2.write(str(result))
            else:
                count4+=1
                result = str(header[i]) + "\n"+str(region)+"\n"
                file3.write(str(result))

print ("Number of sequences with the required criteria (>247 and 10bp TIR 5' SB:") #for checking
print (count2) #for checking
print ("Number of sequences with TSD TA:") #for checking

```

```

print (count3) #for checking
print ("Number of sequences without TSD TA:") #for checking
print (count4) #for checking

```

#Python script 4

Python script to separate sequences from a multifasta into independent files named as the ID

```

import sys
header = []
sequence = []
count = -1
file1 = open(sys.argv[1],"r") #genome file
for line in file1:
    line = line.rstrip()
    if ">" in line:
        header.append(line)
        count+=1
        sequence.insert(count,"")
    else:
        sequence[count]+=line
con=count+1
result=""
for i in range (con):
    name=header[i]
    name=name[1:]
    file2 = open("%s_query" % name,"w")
    result = header[i]+"\\n"+ sequence[i]
    file2.write(str(result))

```

#Python script 5

#Python script to recover a table with the IDs and size of the sequences used as query (the end of R2 sequences with TSD)

```

import sys
header = []
sequence = []
count = -1
file1 = open(sys.argv[1],"r") #fasta file to be analyzed
for line in file1:
    line = line.rstrip()
    if ">" in line:
        temp=line.split(" ")
        header.append(temp[0])
        count+=1
        sequence.insert(count,"")
    else:
        sequence[count]+=line
con=count+1
result=""
file2 = open(sys.argv[2],"w")
for i in range (con):
    result = str(header[i]) + "\\t"+str(len(sequence[i]))+"\\n"
    file2.write(str(result))

```

#Python script 6

Python script to recover the exact position of insertion site

It is based on the blast result of the portion of R2 just after the end of the TIR against to the CHO genomic region

#It will generate a table with the ID of read and the corresponding insertion site position

```

import sys
file1 = open(sys.argv[1],"r") #table with the ids \\t size of queries
file2 = open(sys.argv[2],"a") #out

```

```

result=""
for line in file1:
    line = line.rstrip()
    temp0 = line.split("\t")
    id=temp0[0]
    size=temp0[1]
    print(line) #for checking
    file3 = open("%s_result" %id,"r") # for each ID in the table (file 1), open the respective blast results
    for line2 in file3:
        temp = line2.split("\t")
        temp2 = temp[1].split("_")
        print (temp[6]) #for checking
        if int(temp[6])==1: # the blast alignment should start from the first position
            if int(temp[3]==size): # the blast alignment should encompass the entire query size. This could
                or should be deleted if the query is not short
                    if (temp2[2]<temp2[4]):

                        position = (int(temp2[2]) + int(temp[8])-1)
                        result= str(id) + "\t" + temp2[0] + "_" + temp2[1]+":"+ str(position) + "\n"
                        file2.write(str(result))
        else:
            print (temp2[2] + "\t" +temp2[4] + "\n")
            position = (int(temp2[2]) - int(temp[8]))
            result= str(id) + "\t" + temp2[0] + "_" + temp2[1]+":"+ str(position) + "\n"
            file2.write(str(result))

```

Perl script 1

```

# Perl script to recover sequences from a genome from blast search results
# It was written to isolate genomic regions of CHO that contain the insertions sequenced after PCR
splinkerette
# It will recover the sequence from initial match position to 7 kb of end match position
# For each read R1 with positive hit in the blast (limited to 1 hit), a separate file will be created containing
the genomic region
open (IN1, $ARGV[0]) || die "problem to open genome";#genome
@entrada=<IN1>;
close IN1;
$counter=-1;
foreach $line(@entrada){
    chomp$line;
    if ($line=~>/){
        $counter++;
        $header[$counter]=$line;
    }
    else {
        $sequence[$counter]=$line;
    }
}
print "The genome was read";

open (IN2,$ARGV[1]) || die "problem to open blast table"; #blast result
while ($line=<IN2>) {
    chomp $line;
    print($line);
    @temp=split(/\t/,$line);
    print ($temp[1]);
    if ($temp[8]<=$temp[9]) {#8=subject start; 9 = subject end
        $initial=$temp[8];
        $final=$temp[9];
    } elsif ($temp[8]>=$temp[9]){
        $final=$temp[8];
    }
}

```

```

        $initial=$temp[9];
    }
    $final=$final+7000;# 7kb was chosen since a larger distance of restriction site and the
insertion would have difficulties
    #to be amplified. It can be changed

    @trim=($temp[1],$initial,$final);
    for($i=0;$i<=$counter;$i++){

        if($header[$i]=~/ $temp[1] /i){
            open OUT,">$temp[0]_data";#output files
            if ($temp[8]>$temp[9]) {
                $nucleotide=substr($sequence[$i],$trim[1]-1,$trim[2]-$trim[1]);
                $nucleotide =~ tr/ACGTacgt/TGCAtgca/;
                $seq{$temp[1]}= reverse $nucleotide;
                print OUT">$trim[0]_ $trim[2]_to_ $trim[1]\n$seq{$temp[1]}\n";
            } else{
                $nucleotide=substr($sequence[$i],$trim[1]-1,$trim[2]-$trim[1]);
                $seq{$temp[1]}= $nucleotide;
                print OUT">$trim[0]_ $trim[1]_to_ $trim[2]\n$seq{$temp[1]}\n";
            }
        }
    }
}

```