



Ministério da Saúde  
FIOCRUZ  
Fundação Oswaldo Cruz  
Instituto Aggeu Magalhães

FUNDAÇÃO OSWALDO CRUZ  
INSTITUTO AGGEU MAGALHÃES  
PROGRAMA DE PÓS GRADUAÇÃO EM BIOCÊNCIAS E BIOTECNOLOGIA EM  
SAÚDE

Elverson Soares de Melo

**Caracterização genômica e funcional dos elementos de transposição presentes em  
culicídeos vetores de patógenos**

Recife

2022

Elverson Soares de Melo

**Caracterização genômica e funcional dos elementos de transposição presentes em  
culicídeos vetores de patógenos**

Tese apresentada ao Doutorado em Biociências  
e Biotecnologia em Saúde, do Instituto Aggeu  
Magalhães, Fundação Oswaldo Cruz, para  
obtenção do grau de Doutor em Ciências

**Orientador:** Dr. Gabriel da Luz Wallau

**Coorientador:** Dr. Antonio Mauro Rezende

Recife

2022

**Catálogo na fonte: Biblioteca Luciana Borges Abrantes dos Santos  
Instituto Aggeu Magalhães / Fundação Oswaldo Cruz**

---

M528c Melo, Elverson Soares de.  
Caracterização genômica e funcional dos elementos de transposição presentes em culicídeos vetores de patógenos/Elverson Soares de Melo. — Recife: [s. n.], 2022.  
260 p.: il.; 30 cm.

Tese (doutorado) – Instituto Aggeu Magalhães, Programa de Pós-graduação em Biociências e Biotecnologia em Saúde, Recife, 2022.

Orientador: Gabriel da Luz Wallau.  
coorientador: Antonio Mauro Rezende.

1. Elementos de DNA transponíveis. 2. Retroelementos. 3. Transferência genética horizontal. 4. Culicídeos. 5. Biologia Computacional. 6. Expressão gênica. 7. RNA longo não codificante. I. Wallau, Gabriel da Luz. II. Rezende, Antonio Mauro. III. Título.

Elverson Soares de Melo

**Caracterização genômica e funcional dos elementos de transposição presentes em  
culicídeos vetores de patógenos**

Tese apresentada ao Doutorado em Biociências  
e Biotecnologia em Saúde, do Instituto Aggeu  
Magalhães, Fundação Oswaldo Cruz, para  
obtenção do grau de Doutor em Ciências

Aprovado em: 06 de junho de 2022.

**BANCA EXAMINADORA**

---

Dr. Gabriel da Luz Wallau  
Instituto Aggeu Magalhães/Fundação Oswaldo Cruz

---

Dra. Constância Flávia Junqueira Ayres  
Instituto Aggeu Magalhães/Fundação Oswaldo Cruz

---

Dra. Tatiany Patrícia Romão Pompílio de Melo  
Instituto Aggeu Magalhães/Fundação Oswaldo Cruz

---

Dra. Ana Maria Benko Iseppon  
Universidade Federal de Pernambuco

---

Dr. Elgion Lucio Silva Loreto  
Universidade Federal de Santa Maria

## AGRADECIMENTOS

Ao meu orientador Dr. Gabriel da Luz Wallau pela amizade, compreensão pelas discussões científicas enriquecedoras, incentivo a seguir na carreira da ciência e pela oportunidade de participar de um grupo pesquisa tão dedicado.

Ao meu coorientador Dr. Antonio Mauro Rezende pelas discussões científicas em temas sobre os quais eu não tinha tanto domínio.

Aos meus amados pais Samuel e Lucicleide, pelo esforço que fizeram para me educar, pelo estímulo infindável na minha carreira acadêmica, pelo enorme apoio e conselhos nos momentos em que me senti confuso em meio aos problemas decorrentes dessa carreira.

À minha querida noiva Sophia, pelo carinho, amor e companheirismo no dia a dia. Por me ajudar a ser mais forte na vida e no trabalho, por me dar suporte quando enfrentei as dificuldades inerentes da carreira na pesquisa científica e pela compreensão quando necessitei me dedicar integralmente a algumas etapas do doutorado.

Aos meus colegas do Wallaulab e do núcleo de Bioinformática pelos momentos de companheirismo, de discussões científicas enriquecedoras, pelo auxílio em alguns problemas em pipelines e no servidor, e pelos momentos de churrasco gaúcho (quando eu conseguia estar presente).

À Fiocruz Pernambuco por ter me dado à oportunidade de estudar e construir minha vida de pesquisa em uma instituição de qualidade, desde o início do primeiro estágio de iniciação científica em 2010.

À banca examinadora, por ter aceitado o convite e pela contribuição oferecida a este trabalho.

À CAPES pelo incentivo financeiro durante todo o período de doutorado, sem o qual seria mais complicado manter 4 anos de pesquisa dedicada.

A todos os pesquisadores, espalhados por todos os continentes do mundo, que indiretamente contribuíram para este trabalho ao gerar e depositar uma série de dados primários em bancos de dados públicos.

Por fim, gostaria de agradecer aos meus amigos e familiares pelo carinho e pela compreensão nos momentos em que a dedicação aos estudos foi exclusiva, e muitas vezes precisei estar ausente.

“A vida sem ciência é uma espécie de morte.” (Sócrates)

## RESUMO

MELO, Elverson Soares de. Caracterização genômica e funcional dos elementos de transposição presentes em culicídeos vetores de patógenos. 2022. Tese (Doutorado Acadêmico em Biociências e Biotecnologia em Saúde) - Instituto Aggeu Magalhães, Fundação Oswaldo Cruz, Recife, 2022.

Elementos de transposição (TEs) são sequências repetitivas inseridas nos genomas de quase todos os organismos e que conseguem sintetizar proteínas que auxiliam sua mobilidade pelo genoma do hospedeiro. Em diversos casos essa movimentação pode trazer consequências negativas para o indivíduo ou espécie, variando desde modificação da expressão de um gene até a perda funcional de uma ORF. Entretanto, em alguns casos a transposição traz vantagem à espécie. Por exemplo, em mosquitos, alguns TEs são responsáveis por variações genéticas que conferem resistência a inseticidas. Apesar da sua importância, devido ao fato de muitos mosquitos estarem implicadas na transmissão de patógenos, muitos deles só tiveram seu genoma conhecido na última década. Entretanto, o conjunto de TEs dessas espécies ainda é pouco explorado, e quase nada se sabe sobre o modo de herança e impactos dos TEs nesses genomas. Dessa forma, o objetivo deste trabalho foi caracterizar os TEs em mosquitos desde o genoma até o proteoma, avaliando seu modo de herança e o potencial impacto no hospedeiro. Para isso foram analisados genomas de 24 espécies de mosquitos, juntamente a 1035 corridas de RNA-Seq de 12 espécies, além de 704 corridas de espectrometria de massas de três espécies. As análises foram realizadas por métodos computacionais a partir de dados públicos de experimentos depositados em bases de dados. Os resultados mostraram que os TEs foram amplamente trocados por transferência horizontal (HTT) (172 famílias de TEs) entre os mosquitos nos últimos 30 milhões de anos. Pesquisando em mais de 32.000 genomas, também descobrimos HTTs entre mosquitos e dois diferentes filos, Cnidaria e Nematoda, e dois subfilos, Chelicerata e Crustacea, identificando o verme *Wuchereria bancrofti* como provável vetor da propagação horizontal de um elemento *Tc1-Mariner* entre várias espécies de *Anopheles*. A avaliação do transcriptoma dos mosquitos mostrou que muitas das famílias de TEs são expressas. No entanto, apenas uma minoria dos transcritos tem potencial para codificar proteínas completas de TEs, a maioria deles parece formar RNAs longos não codificantes (lncRNAs). De fato, mais de um terço dos lncRNAs de *Ae. albopictus* e *Cx. quinquefasciatus* tem alguma relação com TEs, incluindo cinco lncRNAs capazes de influenciar genes relacionados ao sistema de defesa antiviral. Além disso, outros poucos transcritos de TEs também são coexpressos com genes próximos, sugerindo potenciais efeitos de ação em *cis*. Os transcritos de TEs também apresentam expressão diferencial em algumas situações, sendo fortemente reprimidos nos ovários dos mosquitos. Entretanto, sob estresse, a ativação de sua expressão não segue uma regra geral como mostrada em estudos anteriores, sendo dependente do tipo, intensidade e intervalo pós-estresse analisado. Por fim, poucos TEs são expressos ao nível de proteína e sua expressão é altamente tecido específica. Os resultados da pesquisa mostram que os TEs não são estruturas transcricionalmente estáticas nos mosquitos, evadindo-se dos mecanismos de silenciamento, e que as generalizações de seu comportamento, baseadas em estudos de poucas espécies modelo, são inadequadas.

**Palavra-chave:** elementos de DNA transponíveis; retroelementos; transferência genética horizontal; bioinformática; expressão gênica; RNA longo não codificante.

## ABSTRACT

MELO, Elverson Soares de. Genomic and functional characterization of transposable elements present in Culicidae pathogen vectors. 2022. Thesis (Doctorate in Biosciences and Biotechnology in Health) - Instituto Aggeu Magalhães, Fundação Oswaldo Cruz, Recife, 2022.

Transposable elements (TEs) are repetitive sequences inserted in the genomes of almost all organisms. They also synthesize proteins that help their mobility through the host genome. In many cases, this movement can have negative consequences for the individual or species, ranging from modification of the expression of a gene to the functional loss of an ORF. However, in some cases, the transposition brings an advantage to the species. For example, in mosquitoes, some TEs are responsible for genetic variations that confer insecticide resistance. Despite its importance, since many mosquitoes are involved in the transmission of pathogens, many of them only had their genome known in the last decade. However, the set of TEs of these species is still poorly explored, and almost nothing is known about the mode of inheritance and impacts of TEs on these genomes. Thus, the objective of this work was to characterize TEs in mosquitoes from the genome to the proteome, evaluating the mode of inheritance of these TEs and their potential impact on the host. For this, genomes of 24 species of mosquitoes were analyzed, together with 1035 RNA-Seq samples from 12 species, in addition to 704 mass spectrometry samples from 3 species. The analyzes were performed using computational methods using public data from experiments deposited in databases. The results showed that TEs were largely exchanged by horizontal transfer (172 families of TEs) between mosquitoes over the last 30 million years. Searching over 32,000 genomes, we also discovered transfers between mosquitoes and two different phyla—Cnidaria and Nematoda, and two subphyla—Chelicerata and Crustacea, identifying the worm *Wuchereria bancrofti*, which acted as a vector for the horizontal transfer of a *Tc1-Mariner* element between several *Anopheles* species. Evaluation of the mosquito transcriptome showed that many of the TE families are expressed. However, only a minority of transcripts have the potential to encode complete TE proteins, most of which appear to form long non-coding RNAs (lncRNAs). In fact, more than a third of the lncRNAs of *Ae. albopictus* and *Cx. quinquefasciatus* has some association with TEs, including five lncRNAs capable of influencing genes related to the antiviral defense system. In addition, a few other TE transcripts are also co-expressed with nearby genes, suggesting potential cis-acting effects. TE-derived transcripts also show differential expression in some situations, being strongly repressed in the ovaries. However, under stress, the activation of its expression does not follow a general rule as shown in previous studies, being dependent on the type, intensity, and post-stress interval analyzed. Finally, few TEs are expressed at the protein level and their expression is highly tissue-specific. Research results show that TEs are not transcriptionally static structures in mosquitoes, evading silencing mechanisms, and that generalizations of their behavior, based on studies of a few model species, are inadequate.

**Keywords:** DNA transposable elements; retroelements; gene transfer, horizontal; bioinformatics; gene expression; RNA, long noncoding.



## LISTA DE FIGURAS

<b>Figura 1</b> – Árvore esquemática englobando elementos que possuem o domínio de transcriptase reversa .....	23
<b>Figura 2</b> – Classificação dos elementos não-LTRs (LINEs) .....	34
<b>Figura 3</b> – Estrutura geral de um SINE de tRNA .....	36
<b>Figura 4</b> – Filogenia dos elementos TIR .....	44
<b>Figura 5</b> – Origem e evolução proposta para os <i>Mavericks</i> .....	49
<b>Figura 6</b> – Distribuição das espécies <i>Ae. aegypti</i> e <i>Ae. albopictus</i> .....	53
<b>Figura 7</b> – Predição da distribuição de <i>Culex quinquefasciatus</i> no globo .....	54
<b>Figura 8</b> – Predição da distribuição de 34 espécies de <i>Anopheles</i> transmissores de <i>Plasmodium</i> spp. ....	55
<b>Figura 9</b> – Tipos de influência exercidas pelos TEs nos genes do hospedeiro .....	66
<b>Figura 10</b> – Visão geral da proteogenômica .....	75
<b>Figura 11</b> – Procedimento utilizado na caracterização dos TEs e das HTTs ocorridas .....	85
<b>Figura 12</b> – Procedimento utilizado na caracterização dos TEs expressos .....	87
<b>Figura 13</b> – Função modificada no código do programa TEffectR para lidar com <i>batch effects</i> .....	95
<b>Figura 14</b> – Passo a passo da avaliação do envolvimento dos TEs e lncRNAs .....	99
<b>Figura 15</b> – Fluxograma da análise de dados proteômicos .....	102
<b>Figura 16</b> – Elementos transponíveis dos mosquitos .....	106
<b>Figura 17</b> – Fração do genoma ocupada por TEs, considerando sequências caracterizadas por ambos os métodos, <i>de novo</i> e baseado em homologia .....	108
<b>Figura 18</b> – Proporção de cada superfamília TE nos genomas de mosquitos .....	109
<b>Figura 19</b> – Proporção do mobiloma com as ordens de acordo com o sistema de classificação de Wicker <i>et al.</i> (2007) .....	111
<b>Figura 20</b> – Datação intragenômica das quatro ordens TE mais abundantes de genomas de mosquitos .....	112
<b>Figura 21</b> – HTT entre mosquitos e superfamílias representativas de TEs .....	113
<b>Figura 22</b> – Distribuição de valores de P de todas as comparações de HTT positivas par a par .....	116
<b>Figura 23</b> – Perfil de idade das famílias TE que sofreram transferências horizontais .....	116
<b>Figura 24</b> – Influência das HTTs no genoma do hospedeiro e suas datações .....	117

<b>Figura 25</b> – Transferência horizontal de um TE do clado R4 da superfamília R2 .....	118
<b>Figura 26</b> – Falta de correlação entre sobreposição geográfica e número de casos de transferência horizontal em mosquitos.....	119
<b>Figura 27</b> – Rede de transferências horizontais a partir de uma perspectiva geográfica .....	120
<b>Figura 28</b> – Transposons transmitidos horizontalmente para outras espécies.....	122
<b>Figura 29</b> – Distribuição dos elementos de transposição em relação aos genes .....	124
<b>Figura 30</b> – Sobreposição de lncRNAs e elementos transponíveis em <i>Aedes albopictus</i> e <i>Culex quinquefasciatus</i> .....	126
<b>Figura 31</b> – Diferentes distribuições de superfamílias TE no genoma e em éxons de lncRNAs.....	128
<b>Figura 32</b> – Análise de enriquecimento de superfamílias TE em éxons de lncRNAs.....	129
<b>Figura 33</b> – Idade relativa para cópias de TE dentro e fora de éxons de lncRNAs.....	130
<b>Figura 34</b> – Distribuição do número de <i>reads</i> mapeadas por genes e TEs dos mosquitos...	131
<b>Figura 35</b> – Frações das famílias de TEs que são expressas em nível de RNA .....	132
<b>Figura 36</b> – <i>Bloxplot</i> da expressão dos TE nas 12 espécies de acordo com as diferentes Ordens. ....	133
<b>Figura 37</b> – <i>Heatmap</i> da expressão das diferentes superfamílias de TEs nos mosquitos.....	134
<b>Figura 38</b> – Expressão diferencial dos TEs nas gônadas e carcaças de machos e fêmeas ...	135
<b>Figura 39</b> – Expressão diferencial dos TEs em mosquitos sob diferentes situações de estresse. ....	137
<b>Figura 40</b> – Expressão diferencial dos TEs em <i>Aedes aegypti</i> infectado com diferentes patógenos.....	139
<b>Figura 41</b> – Correlação entre a expressão de genes e TEs nos mosquitos. ....	140
<b>Figura 42</b> – Rede de associações físicas e funcionais com a única proteína de <i>An. minimus</i> cuja expressão é fortemente associada à expressão de TEs. ....	149
<b>Figura 43</b> – Rede de associações físicas e funcionais com a única proteína de <i>An. dirus</i> cuja expressão é fortemente associada à expressão de um TEs.....	150
<b>Figura 44</b> – Rede de associações físicas e funcionais com a única proteína de <i>An. gambiae</i> cuja expressão é moderadamente associada à expressão de TEs. ....	150
<b>Figura 45</b> – Rede de associações físicas e funcionais das proteínas de <i>An. albimanus</i> cuja expressão é forte ou moderadamente associada à expressão de TEs. ....	152
<b>Figura 46</b> – Rede de associações físicas e funcionais das proteínas de <i>An. stephensi</i> cuja expressão é forte ou moderadamente associada à expressão de TEs. ....	153

<b>Figura 47</b> – Rede de associações físicas e funcionais das proteínas de <i>An. funestus</i> cuja expressão é fortemente associada à expressão de TEs. ....	154
<b>Figura 48</b> – Rede de associações físicas e funcionais das proteínas de <i>An. coluzzii</i> cuja expressão é fortemente associada à expressão de TEs. ....	155
<b>Figura 49</b> – Rede de associações físicas e funcionais das proteínas de <i>Cx. quinquefasciatus</i> cuja expressão é fortemente associada à expressão de TEs. ....	157
<b>Figura 50</b> – Completude dos transcriptomas montados e de referência verificada pelo BUSCO.....	160
<b>Figura 51</b> – Fração do transcriptoma composta por transcritos relacionados a TEs.....	160
<b>Figura 52</b> – Proporção de transcritos de TEs capazes de gerar ORFs significativas .....	161
<b>Figura 53</b> – Número de potenciais isoformas completas de TEs dentre as ORFs preditas..	162
<b>Figura 54</b> – Famílias de potenciais isoformas completas de TEs dentre as ORFs preditas .	163
<b>Figura 55</b> – Proteínas de TEs identificadas por espectrometria de massas .....	165

## LISTA DE QUADROS

<b>Quadro 1</b> – Divisão taxonômica da superfamília <i>Bel-Pao</i> .....	26
<b>Quadro 2</b> – Diferentes transposases da superfamília Tc1-mariner .....	40
<b>Quadro 3</b> – Lista de montagens de genomas representativos, depositados no NCBI, das espécies de mosquito. ....	56
<b>Quadro 4</b> – Experimentos de RNA-Seq de exposição a estresse abióticos e bióticos analisados .....	90
<b>Quadro 5</b> – Experimentos de RNA-Seq de em linhagens somáticas e germinativas dos mosquitos .....	94
<b>Quadro 6</b> – Dados dos experimentos utilizados para análise proteômica de elementos de transposição nesse trabalho. ....	101
<b>Quadro 7</b> – Elementos de transposição que potencialmente agem em <i>cis</i> modelando a expressão de genes adjacentes.....	141

## LISTA DE ABREVIATURAS E SIGLAS

AP	<i>Apurinic Endonuclease</i>
CPM	<i>Counts per million</i>
CUB	<i>Codon usage bias</i>
dc-megablast	<i>Discontiguous megablast</i>
dS	<i>Synonymous substitution rate</i>
EN	Endonuclease
Env	Envelope
ESI	<i>Electrospray</i>
FDR	<i>False discovery rate</i>
GTF	<i>General Transfer Format</i>
HPLC	<i>High-performance liquid chromatography</i>
HSPs	<i>High scoring pairs</i>
HTH	<i>Helix-turn-helix</i>
HTT	<i>Horizontal transposon transfer</i>
ICTV	<i>International Committee on Taxonomy of Viruses</i>
INT	Integrase
LARD	<i>Large retrotransposon derivative</i>
LINE	<i>Long interspersed nuclear element</i>
lncRNAs	<i>Long non-coding RNAs</i>
LTR	<i>Long terminal repeats</i>
MITE	<i>Miniature inverted-repeat transposable elements</i>
mtDNA	DNA mitochondrial
Non-LTR	<i>Non-Long terminal repeats</i>
PEP	<i>Posterior error probability</i>
piRNAs	<i>Piwi-interacting RNAs</i>
PIT	<i>Proteomics informed by transcriptomics</i>
RH	RNase H
RT	<i>Reverse transcriptase</i>
SINE	<i>Short interspersed nuclear element</i>
siRNAs	<i>Small interfering RNA</i>
snRNA	<i>Small nuclear RNA</i>

SNV	<i>Single nucleotide variations</i>
Tase	Transposase
TE	<i>Transposable elements</i>
TFs	<i>Transcription factors</i>
TIR	<i>Terminal inverted repeats</i>
TPRT	<i>Target-primed reverse transcription</i>
TRIM	<i>Terminal-repeat retrotransposons in miniature</i>
TSD	<i>Target site duplications</i>
DE	Diferencialmente expressos
dpi	Dias pós-infecção
hpi	Horas pós-infecção
PBS	<i>Primer Binding Site</i>
PPT	<i>Polypurine Tract</i>
PPI	<i>Protein-protein interaction</i>

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b> .....	18
1.1	DELIMITAÇÃO DO TEMA .....	18
1.2	PERGUNTA CONDUTORA.....	20
1.3	HIPÓTESE .....	20
<b>2</b>	<b>REFERENCIAL TEÓRICO</b> .....	21
2.1	ELEMENTOS DE TRANSPOSIÇÃO.....	21
2.2	OS ELEMENTOS DE CLASSE I.....	23
<b>2.2.1</b>	<b>Os elementos da Ordem LTR</b> .....	23
2.2.1.1	Superfamília <i>Copia</i> .....	24
2.2.1.2	Superfamília <i>Gypsy</i> .....	25
2.2.1.3	Superfamília <i>Bel-Pao</i> .....	26
2.2.1.4	TRIMs .....	27
2.2.1.5	LARDs.....	28
<b>2.2.2</b>	<b>Os elementos da Ordem DIRS</b> .....	29
<b>2.2.3</b>	<b>Os elementos da Ordem PLE</b> .....	30
<b>2.2.4</b>	<b>Os elementos não-LTR com terminações indefinidas</b> .....	31
2.2.4.1	Elementos Autônomos – LINEs (Long Interspersed Nuclear Element) .....	31
2.2.4.2	Elementos não-autônomos – SINEs .....	36
2.3	OS ELEMENTOS DE CLASSE II .....	37
<b>2.3.1</b>	<b>Elementos da Ordem TIR</b> .....	38
2.3.1.1	Superfamília <i>CACTA</i> .....	38
2.3.1.2	Superfamília <i>PIF-Harbinger</i> .....	38
2.3.1.3	Superfamília <i>Tc1-Mariner</i> .....	39
2.3.1.4	Superfamília <i>hAT</i> .....	40
2.3.1.5	Superfamília <i>PiggyBac</i> .....	42
2.3.1.6	Superfamília <i>Transib</i> .....	43
2.3.1.7	Superfamília <i>Mutator</i> .....	44
2.3.1.8	Superfamília <i>Merlin</i> .....	44
<b>2.3.2</b>	<b>Elementos TIR não inclusos no esquema de Wicker <i>et al.</i> (2007)</b> .....	44
<b>2.3.3</b>	<b>MITEs</b> .....	46
<b>2.3.4</b>	<b>Elementos da Ordem Crypton</b> .....	48

2.3.5	<b>Elementos da Ordem Maverick</b> .....	48
2.3.6	<b>Elementos da Ordem Helitron</b> .....	49
2.4	<b>OS CULICÍDEOS</b> .....	52
2.5	<b>BANCO DE DADOS GENÔMICOS E OS GENOMAS DOS CULICÍDEOS</b> .....	56
2.6	<b>TRANSFERÊNCIA HORIZONTAL DE TRANSPOSONS (HTT)</b> .....	59
2.7	<b>ELEMENTOS DE TRANSPOSIÇÃO, TRANSCRIÇÃO E EFEITOS NA EXPRESSÃO DE GENES DO HOSPEDEIRO</b> .....	62
2.7.1	<b>Mecanismos do hospedeiro para regulação da expressão de TEs</b> .....	62
2.7.2	<b>Elementos de transposição que escapam do controle transcricional</b> .....	64
2.7.3	<b>Impacto de TEs no organismo hospedeiro</b> .....	65
2.7.4	<b>Expressão de elementos de transposição sob diferentes condições</b> .....	68
2.7.5	<b>Domesticação ou cooptação de TEs</b> .....	69
2.7.6	<b>Formas de estudo do mobiloma expresso</b> .....	70
2.8	<b>PROTEÔMICA E ELEMENTOS DE TRANSPOSIÇÃO</b> .....	71
2.8.1	<b>Proteômica clássica e proteômica de <i>shotgun</i></b> .....	71
2.8.2	<b>Proteogenômica</b> .....	74
2.8.3	<b>Proteômica informada pela transcriptômica - PIT</b> .....	77
2.8.4	<b>Proteômica <i>de novo</i></b> .....	78
3	<b>OBJETIVOS</b> .....	80
3.1	<b>OBJETIVO GERAL</b> .....	80
3.2	<b>OBJETIVOS ESPECÍFICOS</b> .....	80
4	<b>METODOLOGIA</b> .....	81
4.1	<b>CARACTERIZAÇÃO DO MOBILOMA</b> .....	81
4.1.1	<b>Definição das espécies e obtenção do genoma</b> .....	81
4.1.2	<b>Identificação dos Elementos de Transposição através de uma metodologia <i>de novo</i></b> .....	82
4.1.3	<b>Identificação de elementos transponíveis usando uma abordagem baseada em homologia</b> .....	83
4.1.4	<b>Mapeamento de consensos e cálculo da divergência das famílias de TEs</b> .....	83
4.1.5	<b>Análise de Transferência Horizontal</b> .....	84
4.1.6	<b>Datando as transferências horizontais</b> .....	86
4.1.7	<b>Análise da distribuição geográfica de mosquitos vetores e relação com as HTTs</b> .....	87



4.2	ABORDAGEM TRANSCRIPTÔMICA .....	88
4.2.1	<b>Anotação dos genomas .....</b>	88
4.2.2	<b>Caracterização dos Transcritos derivados de TEs.....</b>	89
4.2.3	<b>Avaliação da expressão diferencial.....</b>	90
4.2.4	<b>Verificação da coexpressão entre TEs e genes.....</b>	94
4.2.5	<b>Reconstrução de transcritos derivados de TEs.....</b>	98
4.2.6	<b>Avaliação da sobreposição de TEs e RNAs mensageiros.....</b>	99
4.2.7	<b>Avaliação da sobreposição de TEs e RNAs longos não-codificantes .....</b>	99
4.3	ABORDAGEM PROTEÔMICA .....	101
4.3.1	<b>Definição das espécies e obtenção dos espectros de massas.....</b>	101
4.3.2	<b>Análise dos dados proteômicos.....</b>	102
5	<b>CONSIDERAÇÕES ÉTICAS .....</b>	105
6	<b>RESULTADOS.....</b>	106
6.1	RESULTADOS DA ANÁLISE GENÔMICA.....	106
6.1.1	<b>O conteúdo de TE é altamente variável entre as diferentes espécies de mosquitos.....</b>	106
6.1.2	<b>Diversas superfamílias TE são encontradas nos mosquitos .....</b>	109
6.1.3	<b>Espécies do complexo <i>An. gambiae</i> possuem atividades mais recentes de TEs</b>	111
6.1.4	<b>Centenas de eventos HTT ocorreram nas espécies de mosquitos .....</b>	113
6.1.5	<b>As transferências horizontais de transposon impactam diretamente no tamanho do genoma dos mosquitos .....</b>	118
6.1.6	<b>As transferências horizontais de elementos transponíveis ocorreram nos últimos 30 milhões de anos .....</b>	119
6.1.7	<b>A localização geográfica não afeta o número de transferências entre as espécies de mosquitos .....</b>	120
6.1.8	<b>Os elementos <i>Tc1-Mariner</i> foram transferidos horizontalmente para espécies eucarióticas distantemente relacionadas .....</b>	120
6.2	RESULTADOS DAS ANÁLISES FUNCIONAIS .....	123
6.2.1	<b>Centenas de TEs estão sobrepostos a genes nos mosquitos .....</b>	124
6.2.2	<b>Os lncRNAs de algumas espécies de mosquitos são enriquecidos de TEs.....</b>	126
6.2.3	<b>Cópias de TEs nos lncRNAs são mais jovens.....</b>	130
6.2.4	<b>Diversas famílias de TEs são expressas nos mosquitos .....</b>	131
6.2.5	<b>Os TEs são fortemente reprimidos nos ovários .....</b>	135

6.2.6	<b>Expressão dos TEs em situação de estresse e de infecção</b> .....	137
6.2.7	<b>Coexpressão entre TEs e genes</b> .....	140
6.2.8	<b>Remontagem dos transcritos de TEs</b> .....	160
6.2.9	<b>Poucos TE chegam ao nível de expressão de proteínas</b> .....	163
7	<b>DISCUSSÃO</b> .....	167
7.1	<b>COMPOSIÇÃO DO MOBILOMA E TRANSFERÊNCIA HORIZONTAL</b> .....	167
7.2	<b>EXPRESSÃO DOS TEs</b> .....	173
7.3	<b>ELEMENTOS DE TRANSPOSIÇÃO E RNAs LONGOS NÃO CODIFICANTES</b> .....	187
7.4	<b>LIMITAÇÕES E PERSPECTIVAS FUTURAS</b> .....	189
8	<b>CONCLUSÕES</b> .....	192
	<b>REFERÊNCIAS</b> .....	194
	<b>APÊNDICE A – ARTIGO PUBLICADO NA REVISTA PLOS GENETICS</b> .	228
	<b>APÊNDICE B – ARTIGO PUBLICADO NA REVISTA GENETICS AND MOLECULAR BIOLOGY</b> .....	229
	<b>APÊNDICE C – MONTAGENS DE GENOMA DE MOSQUITO USADOS NESTE ESTUDO</b> .....	230
	<b>APÊNDICE D – LISTA DE CORRIDAS DE RNA-SEQ UTILIZADAS NESTE ESTUDO</b> .....	233
	<b>APÊNDICE E – GARANTIA DA QUALIDADE DAS CORRIDAS DE RNA- SEQ UTILIZADAS</b> .....	259

# 1 INTRODUÇÃO

Esta seção introduz o objeto de estudo da tese ao leitor, assim como apresenta as questões e hipóteses investigadas no trabalho.

## 1.1 DELIMITAÇÃO DO TEMA

O estudo e caracterização dos genomas revelaram padrões inesperados, como a presença massiva de “parasitas genômicos”, conhecidos como elementos de transposição (TEs, do inglês *Transposable Elements*), na grande maioria das espécies eucarióticas. Esses elementos utilizam a maquinaria molecular do organismo hospedeiro (organismo que possui TEs em seu genoma) para promover sua própria propagação pelo genoma do indivíduo (KIDWELL; LISCH, 2001). Dessa forma, os TEs são sequências de DNA que podem se mover de um *locus* genômico para outro. Eles foram descobertos por Barbara McClintock no final da década de 1940, no genoma do milho (MCCLINTOCK, 1950, 1954), mas não foram muito destacados até a década de 70, quando foram redescobertos em outras espécies e sua genética e importância evolutiva reveladas em todos os ramos da vida (RAVINDRAN, 2012). Estes parasitas são naturalmente “confinados” ao genoma dos organismos e conseqüentemente são geralmente transmitidos de uma espécie para outra através de transferência vertical, isto é, de espécies ancestrais para espécies descendentes. No entanto, há evidências de que os mesmos podem sair de um genoma e invadir novos genomas através do mecanismo de transferência horizontal (CHALOPIN *et al.*, 2015; GILBERT; FESCHOTTE, 2018).

Os TEs são responsáveis por grande parte do conteúdo do genoma de muitos eucariontes como consequência da sua mobilidade e replicação. Devido a essa mobilidade, eles são capazes de gerar diferenças estruturais e influenciar diretamente o tamanho dos genomas podendo alterar a estrutura e regulação gênica (HIRSCH; SPRINGER, 2016). Sabe-se agora que os TEs remodelam profundamente os genomas de todas as espécies de inúmeras maneiras. A natureza repetitiva dos TEs gera sítios abundantes para rearranjos cromossômicos. Além disso, sua transposição pode ter sérias consequências para os genomas do hospedeiro por meio da inserção em regiões codificadoras de genes hospedeiros ou nos íntrons. Isso pode dar origem a novas caudas poli(A), causando pulo de éxon ou criando promotores alternativos (BIÉMONT; VIEIRA, 2006; COWLEY; OAKLEY, 2013). Os efeitos desses diversos impactos no hospedeiro variam entre positivos e negativos.

As transposições de TEs e as potenciais recombinações geradas (BENNETZEN; WANG, 2014) podem provocar consequências desastrosas para o genoma, como disfunção de genes codificantes. Devido a isso, os organismos desenvolveram um arsenal completo de mecanismos moleculares que podem silenciar esses elementos (HIRANO; SIOMI, 2015; ITO; KAKUTANI, 2014; SAITO; SIOMI, 2010). Apesar desses mecanismos de silenciamento, algumas famílias de TEs ainda produzem proteínas funcionais que podem, em alguns casos, ser recrutadas para exercer uma nova função na fisiologia de um organismo (CHUONG; ELDE; FESCHOTTE, 2017). Tal processo é amplamente conhecido como cooptação ou domesticação molecular (JANGAM; FESCHOTTE; BETRÁN, 2017). Entre 50 e 100 proteínas de plantas e mamíferos foram descritas como derivadas de TEs, como por exemplo, as proteínas RAG1 e 2, que possuem papéis essenciais na recombinação para geração de variabilidade na resposta imunológica a novos patógenos do sistema imunológico de mamíferos (KAPITONOV; JURKA, 2005). Considerando todas as possibilidades, os TEs podem contribuir significativamente para a formação e reestruturação de redes gênicas regulatórias. Isto ocorre principalmente pela criação de novos sítios de ligação de fatores de transcrição (TFs) trazidos pelos TEs, que em camundongos, representam 40% do total de sítios de ligação da espécie (SUNDARAM *et al.*, 2014).

Mesmo que estudos envolvendo análise genômica, e mais recentemente transcriptômica de TEs, venham se tornando mais recorrentes, poucos estudos abordam um enfoque capaz de identificar proteínas derivadas de TEs, havendo apenas um realizado até hoje utilizando metodologia de larga escala a fim de identificar todo o proteoma derivado de TEs em eucariotos. Neste estudo, realizado apenas em uma linhagem de cultura de células da espécie *Aedes aegypti*, identificou-se 136 proteínas, principalmente de retrotransposons LTR, que podem estar envolvidos com a defesa antiviral nessa espécie (MARINGER *et al.*, 2017).

Uma parcela significativa do genoma de algumas espécies é composta por TEs. Por exemplo, estes elementos podem chegar a compor 80% do organismo vegetal modelo *Zea mays* (SCHNABLE *et al.*, 2009) e quase metade do genoma do mosquito *Aedes aegypti* (NENE *et al.*, 2007). Este último pertence à família Culicidae, que são organismos de extrema importância para os humanos uma vez que algumas de suas espécies são vetores de arbovírus, tais como: vírus da Dengue, Chikungunya, Zika, febre amarela; e vetores de outros patógenos, como os protozoários causadores da malária e nematódeos causadores da filariose linfática (BENELLI; MEHLHORN, 2016). Os culicídeos de maior importância médica pertencem aos gêneros *Anopheles*, *Culex* e *Aedes*. Embora os TEs correspondam a quase metade do genoma de *Aedes*

*aegypti*, eles compõem uma proporção diferente na composição do genoma de *Anopheles gambiae* (aproximadamente 16%) (ARENSBURGER *et al.*, 2011b).

Nos últimos 10 anos ocorreu um aumento abrupto no número de anofelinos com montagens de genoma disponíveis em bancos de dados públicos, de apenas duas espécies antes de 2013 para 22 espécies em 2018. Contudo, esses são pobremente caracterizados quanto a seu conteúdo e ao impacto dos elementos de transposição. Informações mais detalhadas só existem para os primeiros genomas de Culicídeos sequenciados: *An. gambiae* e *Ae. aegypti*, porém, até o momento nenhum estudo avaliou o impacto dos TEs na expressão gênica dessas espécies.

De fato, avanços científicos para entender o papel funcional dos TEs somente existem para pouquíssimas espécies de animais, como o homem e o camundongo. Para os genomas de Culicídeos muitas questões ainda precisam de averiguação como: I) avaliar a proximidade dos TEs a genes do hospedeiro; II) avaliar a transcrição dos TEs; III) avaliar o potencial codificante dos TEs nestas espécies; IV) avaliar a produção de proteínas derivadas dos TEs no hospedeiro.

## 1.2 PERGUNTA CONDUTORA

Quais os elementos de transposição (TEs) presentes nos genomas de 24 espécies de culicídeos dos gêneros *Anopheles*, *Aedes* e *Culex* e dentre esses, quais são expressos ou ainda podem atuar modulando a expressão gênica?

## 1.3 HIPÓTESE

Diferentes famílias de TEs ocorrem nos mosquitos e parte delas é expressa, impactando o fenótipo e a expressão gênica ao se inserir nas adjacências de genes ou a partir da interação de seus produtos com mRNAs e proteínas dessas espécies.

## 2 REFERENCIAL TEÓRICO

Esta revisão está dividida em 8 partes que abordam desde os conceitos básicos sobre as maiores superfamílias de elementos de transposição até métodos computacionais para realizar inferências sobre esses elementos. Informações sobre seu ciclo de vida, expressão e modo de transmissão (herança vertical ou transferência horizontal) também são apresentadas.

### 2.1 ELEMENTOS DE TRANSPOSIÇÃO

Os elementos de transposição (*transposable elements* – TEs) são sequências de DNA integradas no genoma, que têm a capacidade de se movimentar por esse genoma através de inserções em diferentes regiões. Eles estão presentes no genoma de boa parte dos organismos vivos e, em alguns destes, constituem porção significativa do genoma (BIÉMONT; VIEIRA, 2006; CHÉNAIS *et al.*, 2012).

Os TEs foram descobertos na década de 1940 pela pesquisadora Barbara McClintock ao realizar trabalhos com quebra de cromossomos no milho. Ela observou que os padrões de coloração do milho mudavam a cada geração e a cada cruzamento, atribuindo esse comportamento a presença de elementos genéticos que estavam regulando de alguma forma a expressão de alguns genes. Ela então considerou esses elementos como constituintes normais do genoma do organismo, que eram capazes de controlar a expressão de genes através de sua movimentação pelo genoma durante a proliferação celular e os denominou de *Ac* e *Ds* (BIÉMONT; VIEIRA, 2006; CHUONG; ELDE; FESCHOTTE, 2017; MCCLINTOCK, 1950; RAVINDRAN, 2012).

Esses elementos permaneceram sem muito destaque no campo da genética, até sua redescoberta, na década de 70, em outras espécies bastante distintas do milho, como em bactérias e vírus bacteriófagos. Um grande foco na pesquisa de elementos de transposição surgiu com o conhecimento do genoma total de diversas espécies a partir do novo milênio. Hoje sabe-se que alguns organismos possuem mais de 80% do seu genoma constituído por TEs (BIÉMONT; VIEIRA, 2006; RAVINDRAN, 2012).

Os TEs apresentam uma ampla diversidade estrutural, diferindo principalmente em relação a seus mecanismos de transposição. Essa diferença permite a divisão dos TEs em dois grupos iniciais: os transposons de DNA e os retrotransposons de RNA. Os transposons de DNA se transpõem através de um intermediário de DNA, ou seja, o próprio elemento é transposto

para outra região do genoma. Esse intermediário é transposto através do mecanismo conhecido como “*cut and paste*”, onde o TE é recortado (do inglês *cut*) de um local do genoma e depois “colado” (do inglês *paste*) em outra região do genoma, realizado por uma transposase, ou através de uma replicação “*rolling circle*” como ocorre nos elementos *Helitrons*. Os retrotransposons de RNA, por sua vez, se transpõem através de um intermediário de RNA, ou seja, o elemento inserido no genoma serve de molde para a transcrição de um RNA. Esse RNA é então reversamente transcrito em DNA por meio de uma DNA polimerase RNA-dependente e posteriormente inserido no genoma, sendo o processo conhecido como “*copy and paste*”, pois o TE é primeiramente copiado (do inglês *copy*) em RNA e transcrito reversamente para DNA e depois “colado” (do inglês *paste*) em outra região do genoma. Entretanto os mecanismos de transposição variam bastante entre os diferentes grupos de retrotransposons de RNA (BENNETZEN; WANG, 2014; LEVIN; MORAN, 2011).

Devido à imensa gama de tipos de elementos de transposição existentes, há certos problemas quanto à sua classificação. Muitas vezes a classificação dos TEs aparenta confusão, não existindo apenas um esquema de classificação para esses elementos. Cada banco de dados desses elementos adapta um esquema de classificação havendo sobreposições e diferenças. Exemplos de bancos desse tipo são o Repbase (BAO; KOJIMA; KOHANY, 2015; JURKA *et al.*, 2005), o TEFam (<http://tefam.biochem.vt.edu/tefam/index.php>), assim como bancos específicos para TEs de determinadas espécies como o SoyTEdb, no caso da soja (DU *et al.*, 2010).

Os principais esquemas de classificação de TEs são: o esquema criado por Kapitonov e Jurka (2008) para o banco de dados de TEs Repbase, e o esquema proposto por Wicker *et al.* (2007). Embora haja diferenças em relação a alguns conceitos e ao agrupamento de TEs com base em algumas características similares, há diversos pontos de concordância entre os esquemas. O esquema de classificação de Wicker *et al.* (2007) divide os TEs em categorias similares às usadas na classificação taxonômica de espécies. De acordo com ele, os TEs são divididos em duas Classes (Classe I – TEs de RNA e Classe II – TEs de DNA), estas podem ter subclasses, que por sua vez são divididas em Ordens. Cada Ordem então é dividida em uma ou mais superfamílias, que divergem entre si quanto à presença de duplicações criadas nas bordas do local de inserção do TE (*target site duplications* - TSD) e quanto a sua sequência de DNA. A junção da classe, ordem e superfamília gera um código de trinca para cada TE (WICKER *et al.*, 2007).

Nos próximos tópicos dessa revisão sobre elementos de transposição serão descritos vários grupos de TEs que foram classificados por Wicker *et al.* (2007) e que merecem destaque na análise do mobiloma de animais. Boa parte desses grupos ainda possui elementos autônomos, enquanto alguns grupos são compostos completamente por elementos não autônomos, que perderam toda a maquinaria enzimática necessária para a transposição. O esquema de classificação de Wicker *et al.* (2007) foi escolhido pois apresenta melhor estrutura e maior grau de granularidade, principalmente em relação a classificação dos LINEs, o que facilita a descrição e o entendimento dos grupos de TEs.

## 2.2 OS ELEMENTOS DE CLASSE I

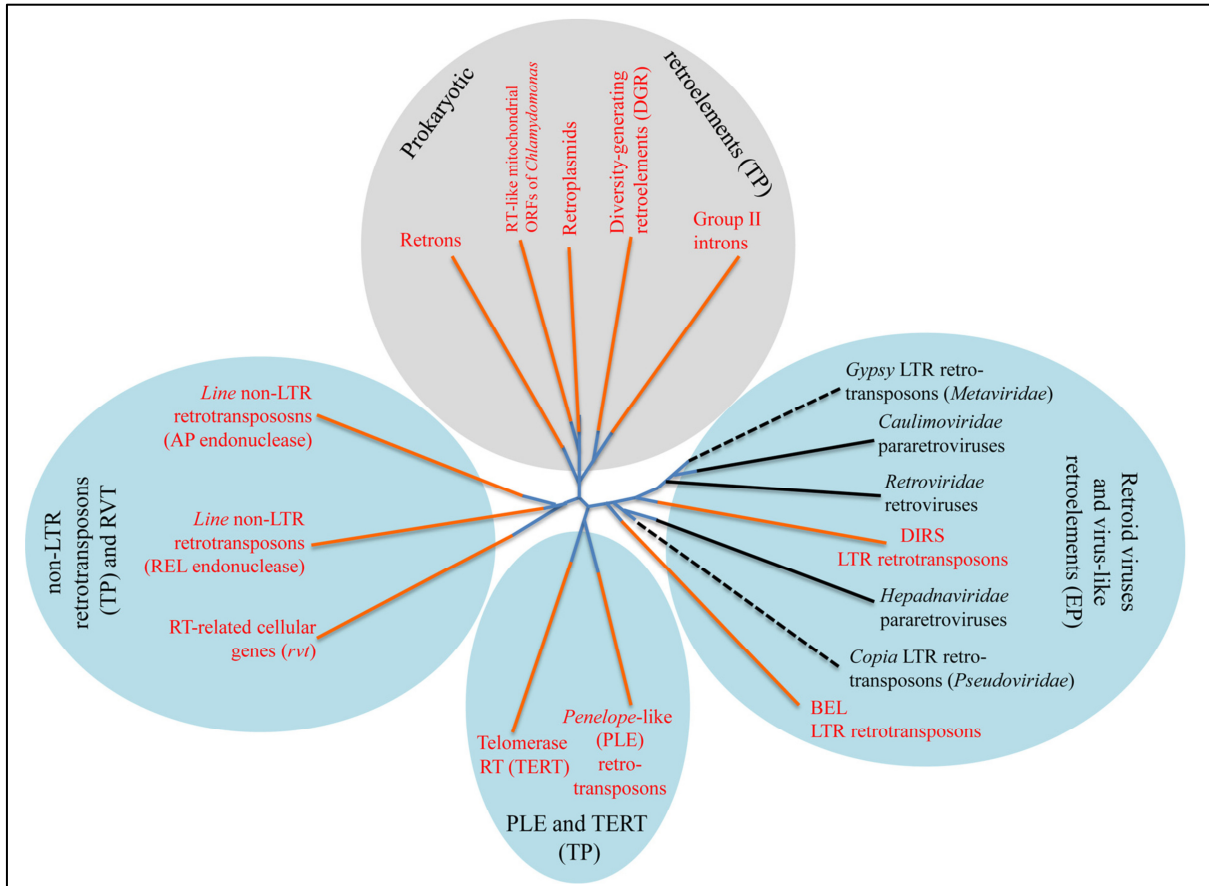
De acordo com o código proposto por Wicker *et al.* (2007), os elementos de Classe I compõem os retrotransposons e se dividem em cinco ordens: LTR, DIRS, PLE, LINE e SINE. Essas ordens divergem muito entre si quanto à estrutura dos seus elementos de transposição e quanto ao mecanismo de inserção de suas novas cópias. A seguir são descritas todas essas ordens.

### 2.2.1 Os elementos da Ordem LTR

De acordo com o esquema de classificação de Wicker *et al.* (2007), os elementos LTRs formam uma Ordem de TEs de Classe I que é muito abundante nos genomas de animais e principalmente de plantas. Eles possuem longas repetições terminais (LTR), e variam desde elementos com centenas a dezenas de milhares de nucleotídeos. Flanqueando a região interna existem duas regiões, uma em cada LTR, elas são chamadas de PBS (*Primer Binding Site*) e sítio PPT (*Polypurine Tract*). Esses TEs geralmente possuem duas ORFs (*open reading frames*), podendo variar em alguns casos. A ORF GAG, produz uma proteína de capsídeo, e a ORF POL produz uma integrase, uma proteinase, uma transcriptase reversa e um domínio de RNase H. Os elementos LTR são muito similares a retrovírus, sua sequência de DNA no genoma é transcrita pela RNA polimerase II. Ao fim da cadeia de expressão, esses TEs produzem proteínas muito similares a proteínas virais. Alguns autores sugerem que esses elementos deram origem aos atuais retrovírus (PACHULSKA-WIECZOREK; LE GRICE; PURZYCKA, 2016). O posicionamento filogenético de alguns TEs LTRs, assim como de outros TEs de Classe I pode ser visto na Figura 1.



Figura 1 – Árvore esquemática englobando elementos que possuem o domínio de transcriptase reversa.



**Fonte:** Koonin e Dolja (2014)

**Legenda:** O círculo cinza representa elementos de procariotos, enquanto os azuis representam elementos eucarióticos. Os ramos pretos representam elementos que ainda podem portar capsídeo viral (os ramos pretos tracejados representam os TEs), já os ramos laranjas indicam elementos que não possuem mais capsídeos.

### 2.2.1.1 Superfamília *Copia*

Os elementos *Copia* foram primeiramente caracterizados e bem estudados em *Drosophila melanogaster* (FINNEGAN *et al.*, 1978). Inicialmente os pesquisadores descreveram elementos que possuíam cerca de 5 kb de tamanho e possuíam cerca de 300 pb em cada LTR. Essas LTRs eram bastante similares a sequências do elemento *Ty1* de *Saccharomyces cerevisiae*, e do vírus da necrose esplênica aviária (LEVIS; DUNSMUIR; RUBIN, 1980). Anos mais tarde, elementos da superfamília *Copia* foram descritos pela primeira vez em plantas, na espécie *Arabidopsis thaliana* (VOYTAS; AUSUBEL, 1988). Hoje sabe-se que esses elementos estão presentes em plantas, animais, fungos e em protistas.

De acordo com Wicker *et al.* (2007), os TEs da superfamília *Copia* diferem das outras superfamílias de LTRs na ordem dos domínios proteicos dentro da ORF POL. Os TEs *Copia*

possuem a sequência de domínios: AP (endonuclease apurínica), INT (integrase), RT (transcriptase reversa) e RH (RNase H). Enquanto isso, os TEs das outras superfamílias possuem a sequência AP, RT, RH, INT.

Os TEs *Copia* se replicam através de uma *virus-like particle* (VLP), que é composta majoritariamente pela proteína Gag (capsídeo e nucleocapsídeo) e pelos genes de integrase e transcriptase reversa imaturos. A proteinase do TE cliva o RNA imaturo preparando-o para a transcrição reversa (VOYTAS, 2008). Ainda que produzam partículas similares às partículas virais, as VLPs desses TEs não são infectantes principalmente pelo fato desses TEs não produzirem proteínas de envelope viral (KING *et al.*, 2011). Por serem capazes de produzir muitas das proteínas virais, autores argumentam que eles apenas não são considerados vírus porque são incapazes de infectar novas células como os vírus fazem. Apesar disso, diferindo da classificação de *Copia* como uma superfamília de TE proposta por Wicker *et al.* (2007), TEs desse grupo também são classificados como membros da família *Pseudoviridae*, de acordo com o Comitê Internacional de Taxonomia de Vírus (*International Committee on Taxonomy of Viruses* - ICTV) (KOONIN; DOLJA, 2014), que a divide em três gêneros (KING *et al.*, 2011):

- a) Gênero *Pseudovirus*: composto de TEs das famílias *Ty1*, *Ty2*, *Ty4* de leveduras, onde geralmente são encontrados em um número variando entre 3 e 40 cópias completas e vários fragmentos, como solo-LTRs (uma única extremidade LTR remanescente do TE completo) ou cópias bastante truncadas. Também compõe vários TEs encontrados em plantas.
- b) Gênero *Hemivirus*: compõe os TEs *Copia stricto sensu*. Esses elementos são encontrados em fungos, algas, e principalmente em insetos, sendo comum em mosquitos, como no genoma de *Ae. aegypti*.
- c) Gênero *Sirevirus*: compõe TEs encontrados apenas em plantas. Alguns desses TEs possuem adicionalmente um gene de envelope viral (*env-like gene*).

#### 2.2.1.2 Superfamília *Gypsy*

Os TEs dessa superfamília possuem uma organização e sequência genéticas mais próxima a retrovírus, que os elementos *Copia*. Entretanto, diferentemente deste último, os *Gypsy* possuem em sua sequência o domínio INT logo após os domínios de RT e RH. Assim como os TE da superfamília *Copia*, os *Gypsy* também são classificados em uma família viral pelo ICTV, sendo conhecidos como membros da família viral *Metaviridae*, estando dispersos pelo genoma de todos os eucariotos estudados (KOONIN; DOLJA, 2014; WICKER *et al.*, 2007).

Esses TEs são elementos bastante grandes, variando entre 4 e 10 kb. Alguns deles possuem LTRs tão pequenas quanto 77 pb, enquanto em outros cada LTR pode ultrapassar 2 kb. Seguindo o código de classificação viral do ICTV, essa família (*Metaviridae*) possui dois gêneros (KING *et al.*, 2011):

- a) Gênero *Metavirus*: compõe o TE *Ty3* e TEs com estruturas similares. Está presente em plantas, fungos, animais vertebrados e invertebrados;
- b) Gênero *Errantivirus*: compõe os *Gypsy stricto sensu* e TEs similares. Muitos dos elementos desse grupo possuem uma ORF que codifica uma proteína similar à proteína do envelope (*env*) dos retrovírus, porém com baixa similaridade em nível de sequência de DNA. Está presente em muitos invertebrados.

Para vários pesquisadores de elementos de transposição, no entanto, essa classificação possui diversas inconsistências. De acordo com o *Gypsy database* (LLORENS *et al.*, 2011), os TEs da superfamília *Gypsy*, podem ser divididos em 34 clados baseados em estudos filogenéticos. Desse total, quatro clados ocorrem entre os Artrópodes: *Gypsy (stricto sensu)*, *17.6*, *412/mdg1*, *Micropia/mdg3*.

### 2.2.1.3 Superfamília *Bel-Pao*

Esses TEs são mais ancestrais que os *Gypsy* e que os retrovírus, formando um clado basal entre os elementos LTR (Figura 1) (KOONIN; DOLJA, 2014). O elemento *Pao* foi inicialmente caracterizado no genoma do Lepidoptera *Bombyx mori*, contendo repetições em tandem dentro de suas LTRs e uma ORF com 1158 aminoácidos (XIONG; BURKE; EICKBUSH, 1993). Já o elemento *BEL* foi inicialmente caracterizado no genoma de *D. melanogaster* durante um estudo de duplicações de sequências e outras deficiências produzidas em fêmeas heterozigóticas (GOLDBERG *et al.*, 1983).

Além dessas duas linhagens de TEs, essa superfamília pode ser dividida em mais três linhagens de acordo com alguns esquemas de classificação: *Suzu*, *Sinbad* e *Tas*. Essas cinco linhagens podem ser alocadas em três ramos filogenéticos (LLORENS *et al.*, 2011). O Quadro 1 traz mais informações sobre esses TEs e sua divisão.

De acordo com a divisão de TEs da ordem LTR em famílias virais, os TEs dessa superclasse são inseridos na família *Belpaoviridae*. Essa família inclui apenas um gênero, o

*Semotivirus*, que inicialmente foi colocado dentro da Família Metaviridae (*Gypsy*) devido a uma organização genética similar (KRUPOVIC *et al.*, 2018; LLORENS *et al.*, 2011).

**Quadro 1** – Divisão taxonômica da superfamília *Bel-Pao*

<b>Linhagem</b>	<b>Filos hospedeiros</b>	<b>Clado</b>	<b>Envelope</b>
Linhagem 1	<i>Metazoa</i> (Nematoda e Cnidaria)	<i>Tas</i>	Presente em alguns elementos
	<i>Metazoa</i> (Insecta)	<i>Bel</i>	Presente em um elemento
Linhagem 2	<i>Metazoa</i> (Insecta e Vertebrata)	<i>Pao</i>	Presente em um elemento
	<i>Metazoa</i> (Invertebrata)	<i>Sinbad</i>	Ausente
Linhagem 3	<i>Metazoa</i> (Echinodermata e Vertebrata)	<i>Suzu</i>	Ausente

**Fonte:** Adaptado do banco de dados GyDB (LLORENS *et al.*, 2011).

É provável que os elementos *Bel-Pao* tenham surgido no início da evolução dos metazoários, sendo o segundo grupo de LTR mais abundante entre esses, atrás apenas da superfamília *Gypsy*. TEs *Bel-Pao* são abundantes em dípteros, ocorrendo em várias espécies do gênero *Drosophila*. Também é relatado um alto número de cópias desses TEs no genoma bem estudado do mosquito *Ae. aegypti* (DE LA CHAUX; WAGNER, 2011).

#### 2.2.1.4 TRIMs

Esses retrotransposons são elementos pequenos e não autônomos que possuem LTRs em suas extremidades. Seu nome quer dizer retrotransposons de repetição terminal em miniatura (*Terminal-repeat retrotransposons in miniature*). Com LTRs variando de 100 a 250 nucleotídeos e região interna variando de 100 a 300 nucleotídeos, os TRIMs foram primeiramente descritos em angiospermas (WITTE *et al.*, 2001), sendo considerados elementos fósseis no genoma do arroz (BARUCH; KASHKUSH, 2012).

Assim como os outros elementos LTRs, os TRIMs possuem em sua sequência os motivos PBS e PPT, assim como uma duplicação (*Target Site Duplication* – TSD) na região flangeadora (WITTE *et al.*, 2001). Por outro lado, o domínio interno dos TRIMs, além de ser pequeno, não codifica quaisquer domínios encontrados comumente em outros TEs LTRs

(WICKER *et al.*, 2007) e diverge muito entre os outros TRIMs de um genoma, ao mesmo tempo em que as LTRs dos TRIMs são muito parecidas (BARUCH; KASHKUSH, 2012). Embora o comum seja presença de LTRs nas extremidades dos retrotransposons LTR, alguns TRIMs possuem de três até sete LTRs numa mesma cópia do elemento, provavelmente derivadas de recombinação (LIU *et al.*, 2015)

Há indícios que os TRIMs podem ter se originado de deleções das regiões internas de grandes retrotransposons LTR (GAO *et al.*, 2016). Apesar disso, os TEs que lhes deram origem, TEs parentais, não são conhecidos nos genomas, o que faz com que essas sequências não possam ser ligadas a nenhuma superfamília de TEs, sendo o nome TRIM apenas criado a partir de suas características estruturais e sem significado taxonômico (WICKER *et al.*, 2007).

Os TRIMs são capazes de se inserir em íntrons ou em promotores, sendo sua ocorrência bastante enriquecida em regiões próximas a genes, onde podem fazer a captura de um éxon de um gene (GAO *et al.*, 2016). Eles também são encontrados em regiões codificantes, como nas famílias de plantas Solanaceae e Fabaceae com sobreposição a genes conhecidos como P450 em *Medicago truncatula* e glicolato oxidase no tomate (WITTE *et al.*, 2001).

Embora a grande maioria dos estudos sobre os TRIMs tenham sido realizados em plantas, eles também já foram descritos em animais. Por exemplo, no cestoda *Echinococcus multilocularis*, onde os TRIMs possuíam LTRs e região interna de aproximadamente 190 pb e 590 pb, respectivamente, estando presente em alto número de cópias, embora a maior parte delas não possuísse sítios PBS e ou PPT. Esses mesmos TRIMs também foram identificados em outros cestodas como as *Taenia solium*, *Taenia asiatica* e *Taenia taeniaeformis* (KOZIOL *et al.*, 2015). Além desses platelmintos, os TRIMs também já foram descritos em espécies de formigas, onde muitas das cópias desse grupo de TEs se encontravam próximas a genes (ZHOU; CAHAN, 2012). Recentemente foi caracterizado um elemento TRIM de 807 pb em *D. melanogaster* (RECH *et al.*, 2022), entretanto até o ano de 2020 esse tipo de elemento não tinha sido descrito em espécies dessa ordem (um dos artigos desta tese relata a ocorrência dessas estruturas de TEs em diversas espécies de mosquitos).

#### 2.2.1.5 LARDs

O nome LARD significa *large retrotransposons derivatives* (grandes derivados de retrotransposons). Esses elementos são similares aos TRIM, porém seu tamanho é bem maior, em geral eles ultrapassam os 4 kb de comprimento. No genoma do trigo, por exemplo, o

tamanho médio desses elementos é de cerca de 13 kb (SABOT *et al.*, 2005). Suas terminações são compostas por LTRs. Na região adjacente interna de cada LTR existem as mesmas estruturas presentes nos outros LTRs: o sítio PBS na região 5' e o sítio PPT na região 3', que são semelhantes em todos os LARDs de plantas. As semelhanças com os outros LTRs terminam neste ponto. Diferentemente de LTR autônomos, os LARDs não possuem quaisquer sequências internas codificadoras, nem mesmo fragmentos de domínios clássicos de elementos de transposição. Como exemplo, as ORFs presentes em LARDs de gramíneas são muito curtas e não codificam para nada conhecido. Apesar da característica de ausência de ORFs clássicas dos TEs LTR, alguns LARDs são transcritos em diferentes variedades de cevada (KALENDAR *et al.*, 2004).

Assim como os TRIMs, os LARDs não representam um grupo monofilético. De fato, eles nem mesmo representam um grupo taxonômico. Esses elementos apenas são agrupados juntos por suas características estruturais, uma vez que não se tem conhecimento dos transposons que os originaram. Além disso, a falta de genes codificantes impede que qualquer análise filogenética seja feita (WICKER *et al.*, 2007).

Embora bastante comuns no genoma de plantas, esses TEs já foram identificados em outros reinos. LARDs foram relatados no genoma do fungo *Laccaria bicolor*, onde representavam 0,08% do genoma e ocorriam preferencialmente em regiões próximas a genes (LABBÉ *et al.*, 2012). Em insetos, dois artigos sobre o genoma de *Cx. quinquefasciatus* reportaram a presença de grandes elementos com LTRs, mas sem região codificadora (ARENSBURGER *et al.*, 2010; MARSANO *et al.*, 2012), o que estruturalmente os classifica como LARDs.

### 2.2.2 Os elementos da Ordem DIRS

De acordo com o sistema de classificação de Wicker *et al.* (2007) essa ordem de elementos se divide em três superfamílias: *DIRS*, *Ngaro* e *VIPER*. A principal diferença desses elementos para os elementos LTR, é que em todas as superfamílias da ordem DIRS, o domínio de integrase está ausente, estando presente o domínio de *tirosina recombinase*. Isso distancia essa ordem de todos os outros elementos de Classe I, já que o modo de integração desses TEs no genoma do hospedeiro é bem distinto dos elementos que codificam a enzima integrase.

Os elementos dessa Ordem não possuem padrões nas terminações da sua sequência de DNA. Elementos da superfamília *DIRS* possuem regiões terminais invertidas (TIR), discutidas

mais adiante, já elementos *Ngaro* e *Viper* possuem outras estruturas conhecidas como *split direct repeats*, um tipo especial de terminação onde mais de um tipo de repetição direta ocorre em uma das extremidades do TE (WICKER *et al.*, 2007). As regiões codificantes são similares entre as superfamílias dessa ordem, porém diferem quanto à sua disposição no TE. No geral, esses elementos possuem 3 ORFs, que podem se encontrar sobrepostas umas às outras em muitos casos (POULTER; GOODWIN, 2005).

O elemento que dá nome à ordem foi descoberto na espécie *Dictyostelium discoideum*, um organismo unicelular da classe Eumycetozoa, que compreende os *slime molds*. Ele tem cerca de 4,5 kb e mais de 40 cópias estão presentes no genoma desta espécie (CHUNG; ZUKER; LODISH, 1983). Hoje sabe-se que TEs da superfamília DIRS e Ngaro estão presentes em muitas linhagens de eucariotos e são relativamente comuns em espécies de insetos (PETERSEN *et al.*, 2019; POULTER; BUTLER, 2015).

### 2.2.3 Os elementos da Ordem PLE

De acordo com o sistema de classificação de Wicker *et al.* (2007), a ordem PLE possui apenas uma superfamília, que é denominada *Penelope*. Estes elementos possuem uma estrutura bem distinta dos outros elementos. Esses TEs foram primeiramente descritos em *Drosophila virilis*, entretanto estão presentes em algumas linhagens dessa espécie e ausentes em outras, mostrando um padrão irregular de distribuição. Além disso, a presença desse elemento propicia a ocorrência da disgenesia híbrida (doença que causa esterilidade, altas taxas de mutação, entre outros problemas) e a transposição de outros elementos. Esses primeiros elementos descritos possuíam uma ORF com cerca de 2,5 kb de comprimento, e uma estrutura diversa, principalmente nas suas terminações, que variavam entre sequências diretas (pseudo-LTRs) ou invertidas (TIRs) (EVGEN'EV *et al.*, 1997).

Embora os primeiros elementos *Penelope* descritos tivessem uma região 5' bastante variável, sua região central possuía um padrão conservado. Parte desses elementos possui uma única ORF que codifica uma transcriptase reversa, mais similar à RT da telomerase que dos outros TEs de Classe I, e uma endonuclease de restrição, que também é diferente de todas as outras endonucleases de TEs. A outra parte dos elementos apresenta uma grande variabilidade até mesmo na região codificante, com diversas inserções, deleções e variações de nucleotídeo único (SNVs). Alguns elementos da ordem PLE contêm também trechos de repetições teloméricas próximos da sua região 3'. Outra característica que os torna bastante diferente de

vários outros elementos de transposição é a capacidade que os PLE têm de possuir diversos íntrons, em geral de pequeno tamanho (EVGEN'EV; ARKHIPOVA, 2005; GLADYSHEV; ARKHIPOVA, 2007).

Esses elementos são extremamente dispersos entre organismos, estando presentes em diversos filos de animais, protistas, fungos e plantas (em maior raridade). Seu padrão irregular de distribuição entre as espécies faz com que eles estejam presentes em uma espécie, mas não em outras espécies do mesmo gênero. Isso indica altas taxas de eliminação no genoma, além de que levanta a suspeita da existência um padrão de transmissão que foge do padrão tradicional herança vertical, onde a geração parental passa o gene aos seus descendentes (ARKHIPOVA, 2006; EVGEN'EV; ARKHIPOVA, 2005). Esse padrão de transmissão de fato já foi comprovado em alguns casos como TEs que se transferiram horizontalmente<sup>1</sup> entre artrópodes e coníferas (LIN; FARIDI; CASOLA, 2016).

Embora bastante dispersos entre os insetos, há algumas discordâncias sobre a ocorrência dos PLEs em algumas espécies de mosquito. Enquanto Arensburger *et al.* (2010) afirmaram não haver PLEs nos genomas de *An. gambiae* e *Cx. quinquefasciatus*, Peterson *et al.* (2019) relataram a presença desses TEs nestes dois genomas e também no genoma de *Ae. aegypti*.

## 2.2.4 Os elementos não-LTR com terminações indefinidas

Esse conjunto abrange alguns pseudogenes, alguns pseudogenes de snRNA (*small nuclear RNA*) e SINEs (KAPITONOV; JURKA, 2003a). De acordo com o sistema de classificação de Wicker *et al.* (2007) eles são divididos em duas Ordens: LINEs e SINEs. Enquanto os primeiros representam elementos que são capazes de codificar as proteínas responsáveis pela sua transposição, embora haja elementos defectivos, os segundos representam elementos não autônomos formados através de um processo específico de fusão entre estruturas.

### 2.2.4.1 Elementos Autônomos – LINEs (Long Interspersed Nuclear Element)

Esses TEs são muito abundantes em linhagens de eucariotos, e amplamente distribuídos entre os reinos. Entretanto, inicialmente eles não eram reconhecidos como elementos

---

<sup>1</sup> O tópico de transferências de elementos de transposição será visto na seção 2.6



autopropagantes e foram primordialmente agrupados como pseudogenes. Apenas posteriormente, foi descoberto que esses elementos possuíam ORFs com similaridade com transcriptases reversas e que provavelmente eram autônomos. Além de poder se inserir dentro de genes, passando a ser cotranscritos (WICKER *et al.*, 2007).

Eles têm um modo de transposição distinto dos outros TEs vistos até agora. As principais diferenças são: I) não possuem integrase; II) a região 5' é geralmente truncada e parece não ter efeito na transposição; III) as TSD de diferentes cópias do mesmo elemento podem ser bem diferentes. Após a transcrição, a endonuclease dos LINEs faz uma quebra no DNA alvo (todas as superfamílias de LINEs produzem endonucleases). A região 3' do DNA clivado em uma fita então serve como *primer* para transcrição reversa do TE, fenômeno que é conhecido como *target-primed reverse transcription* (TPRT). Após a extensão do elemento, a segunda fita do DNA é clivada, e o RNA do híbrido precisa ser degradado, no entanto apenas poucos elementos produzem RNase H. Para os elementos que não possuem essa enzima, a degradação do híbrido deve ser fruto da atividade da maquinaria celular (EICKBUSH; MALIK, 2002). De acordo com o sistema proposto por Wicker *et al.* (2007) existem cinco superfamílias de LINEs: *RTE*, *R2*, *Jockey*, *L1* e *I*.

A flexibilidade do mecanismo de transposição parece permitir uma diversidade de estruturas entre as superfamílias. No geral, superfamílias mais novas possuem duas ORFs, sendo que pelo menos uma delas, chamada de *pol*, codifica para a transcriptase reversa (RT). A outra ORF, por sua vez, não tem função bem estabelecida em todos TEs, acreditando-se ter função de sítio de ligação e atividade de chaperona (EICKBUSH; MALIK, 2002; HAN, 2010). Enquanto isso, elementos mais ancestrais, na base da filogenia da ordem, possuem apenas a ORF *pol* (WICKER *et al.*, 2007). A ORF *pol* das diferentes superfamílias têm diferentes capacidades de codificação, a única característica comum é que todas elas possuem um domínio de endonuclease e um domínio de RT, sendo esse último bem conservado. As endonucleases, por outro lado, são de dois tipos, EN (endonuclease) e APE (endonuclease apurínica).

A classificação dos LINEs é deveras complicada porque existem diversas linhagens e diferentes autores as tratam de diferentes formas. É bastante comum a divisão dos LINEs em *clades* (clados). De acordo com dados atuais, o Repbase classifica esses elementos em 31 clados distintos (GENETIC INFORMATION RESEARCH INSTITUTE, 2022). É interessante notar que há um mapeamento entre os esquemas que dividem os LINEs em clados e o esquema de Wicker *et al.* (2007). Boa parte dos clados podem ser associados às cinco superfamílias de LINEs desse último esquema. No entanto, associar novas famílias de TEs a uma superfamília

ou a um clado não é uma tarefa trivial, pois análises baseadas apenas em estrutura podem levar a erros. Para a tarefa de associação, deve-se conduzir uma análise filogenética a partir do domínio RT, que é ubíquo entre os transposons não-LTRs. Em geral, são usados 440 resíduos de aminoácidos de domínio RT com essa finalidade (KAPITONOV; TEMPEL; JURKA, 2009; WICKER *et al.*, 2007). Um exemplo de filogenia dos LINEs é mostrado na Figura 2a. Nota-se que alguns clados dão nome à superfamília ao qual pertencem. Uma classificação com maior número de clados e com nomenclatura utilizada no Repbase pode ser vista na Figura 2b.

Os elementos da superfamília **R2** compõem os TEs mais antigos. As relações filogenéticas dos clados que o compõe não são totalmente claras, mas sabe-se que esse grupo é parafilético formado por um grande grupo *R2* e um grupo menor *CRE*. Os elementos dessa superfamília se inserem em posições específicas. No que diz respeito aos clados, os *R2* e *R4* se inserem em regiões 28S, já os *CRE* e *NeSL* se inserem em *tandem repeats splice leader*, enquanto os elementos Dong parecem se inserir em regiões de *simple repeats* TAA. Esses TEs parecem não causar potenciais impactos ao hospedeiro porque eles se inserem nas vizinhanças de genes que possuem diversas cópias, caso alguma seja desconfigurada, o impacto negativo não seria tão grande quanto em genes de cópia única. Elementos da superfamília R2 realizam a transcrição reversa mesmo quando não há anelamento entre RNA e DNA. A transcriptase, nesse caso, aproxima o DNA genômico do RNA do TE e realiza várias iniciações abortivas antes de estender o elemento. Todos os elementos R2 possuem uma EN C-terminal que geralmente tem dedos de zinco ou c-myb na região N-terminal (EICKBUSH; MALIK, 2002; KAPITONOV; TEMPEL; JURKA, 2009; KOJIMA, 2019). Existe um clado chamado *Randi/Dualen*, que não se adequa ao esquema de Wicker *et al.* (2007), sendo um clado ancestral como *R2*, que, assim como esse, possui apenas uma ORF. Essa ORF, no entanto, é muito grande, ultrapassando os 9kb. Além disso, ela codifica para duas endonucleases, diferindo totalmente dos outros TEs (KAPITONOV; TEMPEL; JURKA, 2009; KOJIMA; FUJIWARA, 2005).

Os **RTE** são os menores TEs pertencentes aos LINEs, possuindo a estrutura de codificação mais simples. Apresentam regiões 3' UTR extremamente pequenas, que terminam em repetições em tandem. Eles não são sítio-específicos como os *R2*, e assim como essa superfamília, a maior parte dos *RTE* contém apenas uma ORF. Essa codifica um domínio RT e uma endonuclease apurínica N-terminal. Apesar disso, alguns dos seus elementos são tão truncados que foram inicialmente confundidos com SINEs. Alguns *RTEs* são também conhecidos por realizarem transferência horizontal entre vertebrados (EICKBUSH; MALIK, 2002; IVANCEVIC *et al.*, 2018; SUH *et al.*, 2016)

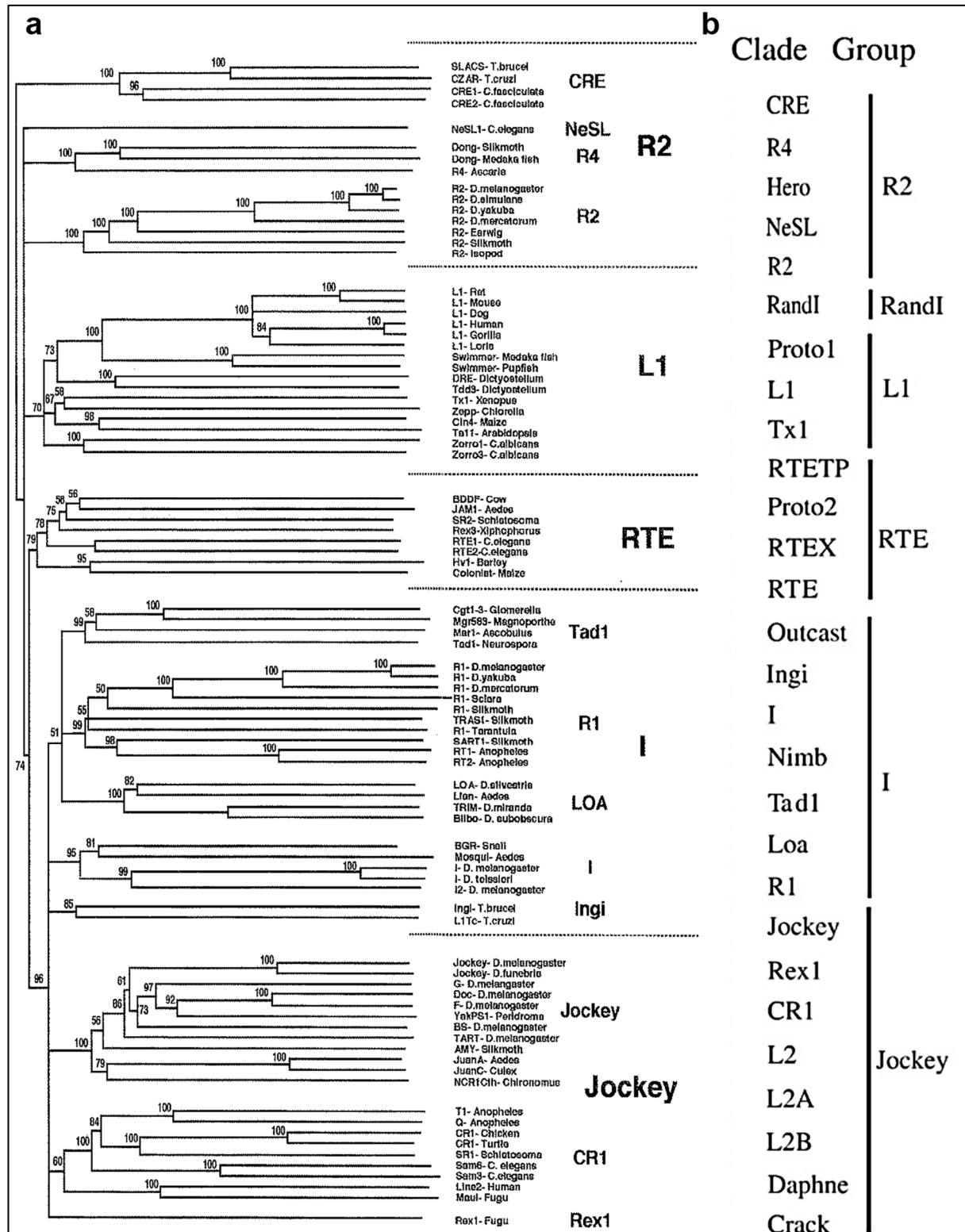
Os elementos das três superfamílias restantes, em geral, codificam duas ORFs. Os elementos da superfamília *LI* são divididos em três clados, sendo esta, a superfamília com o menor número de clados: *LI*, *Tx1*, *Proto1*. A maior parte dos membros tem baixa especificidade em relação aos sítios alvo e são espalhados no genoma de diversas espécies. Entretanto, alguns elementos possuem um padrão de inserção preferencial. O *Tdd-3* se insere comumente próximo a genes de RNA transportadores, enquanto os elementos *Tx1* se inserem em regiões que causam pouco dano ao hospedeiro (EICKBUSH; MALIK, 2002; KOJIMA, 2019). O mais estudado de seus elementos, o *LI*, é muito comum em mamíferos, onde seu número de cópias é geralmente alto, compondo parte significativa do genoma humano. Ele também está presente em outros vertebrados, em insetos, em plantas, onde muitas vezes são potencialmente ativos (IVANCEVIC *et al.*, 2016). Enquanto isso, os elementos *Proto1*, descritos inicialmente em protistas, não são tão comuns em animais, sendo conhecidos entre os artrópodes apenas no genoma do *Pediculus humanus* (KAPITONOV; TEMPEL; JURKA, 2009; PETERSEN *et al.*, 2019).

A superfamília *I* forma um grupo que não é bem distinto da superfamília *Jockey*. Entretanto, é mais simples distinguir-se quais elementos formam a superfamília *I* do que aqueles que formam a superfamília *Jockey*, uma vez que uma das características da *I* é a presença de um domínio de RNase H (Ribonuclease H). Contudo, nem todos os elementos desse grupo possuem esse domínio, a exemplo de *R1* e *Tad1*, onde possivelmente houve perda da RNase H. Os *I* formam o grupo com o segundo maior número de clados. A classificação dentro dessa superfamília é mais complexa, uma vez que alguns clados são parafiléticos (EICKBUSH; MALIK, 2002; KOJIMA, 2019). Além dos clados mostrados na Figura 2b, essa superfamília ainda abriga o clado *Vingi* (KOJIMA; KAPITONOV; JURKA, 2011). Os elementos *I* são comuns em artrópodes, sendo encontrados em todas as espécies investigadas (PETERSEN *et al.*, 2019). *Loner* e *Outcast* são dois clados de *I* muito comuns em mosquitos (BOULESTEIX; BIÉMONT, 2005).

A última superfamília compreende os elementos *Jockey*. Essa é a mais diversa superfamília entre todos os elementos de transposição, com nove clados distintos. Todos os elementos dessa família ocorrem em animais, sendo extremamente abundantes em insetos (EICKBUSH; MALIK, 2002; KOJIMA, 2019). Enquanto a ORF2 desses elementos é mais conservada e codifica a transcriptase reversa e uma endonuclease, a ORF1 é muito diversa estruturalmente, dividindo-se em cinco tipos distintos (METCALFE; CASANE, 2014).

Evidências mostram que esses elementos são capazes de se transferir horizontalmente entre artrópodos, mesmo que em baixa frequência (REISS *et al.*, 2019; SIMÃO *et al.*, 2018).

Figura 2 – Classificação dos elementos não-LTRs (LINEs)



Fonte: Adaptado de Eickbush e Malik (2002) (a) e Kapitonov, Tempel e Jurka, (2009) (b).

Legenda: A árvore mostrada em a foi construída a partir do domínio RT, comum a todos os LINEs. Os nomes das superfamílias estão em letras maiores à direita dos nomes dos clados (letras menores). Em b está mostrada uma divisão atual de cada superfamília (Group) de LINE e seus respectivos clados.

#### 2.2.4.2 Elementos não-autônomos – SINEs (*Short INterspersed Elements*)

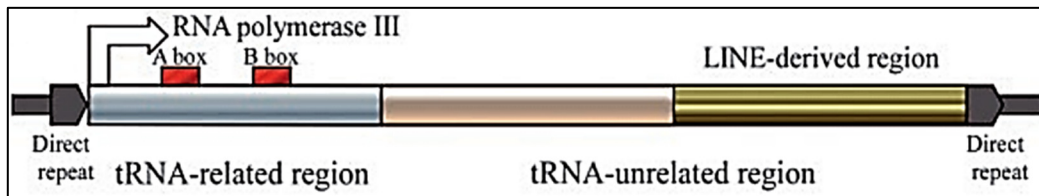
Os SINEs ou elementos pequenos interespçados são transposons não-LTRs com até 600 pb, embora dois terços das famílias conhecidas de SINEs possuam entre 150 e 300 pb. Eles são sequências abundantes nos genomas de eucariotos, ainda que não estejam presentes em eucariotos unicelulares, podendo atingir mais de um milhão de cópias em um genoma. Apesar dessa grande quantidade de cópias, esses TEs não possuem sequência codificante e dessa forma são classificados como elementos não autônomos. Para sua transposição eles usam a RNA polimerase III e necessitam das RDDP (DNA polimerases dependentes de RNA ou transcriptases reversas) pertencentes a outros elementos não-LTRs autônomos e de sequência similar (KRAMEROV; VASSETZKY, 2011).

Grande parte dos SINEs descritos derivam de tRNAs, enquanto frações menores são derivadas do 7SL RNA e do 5S rRNA (OHSHIMA; OKADA, 2005). Alguns autores impõem restrições adicionais para classificação de uma sequência como um SINE, como pelo menos 100 cópias no genoma e 60% de identidade com algum tRNA, 5S rRNA ou 7SL RNA em pelo menos 60 nucleotídeos de alinhamento (VASSETZKY; KRAMEROV, 2013). Devido ao seu tamanho reduzido sua estrutura é simples: regiões 5' e 3' terminais separadas por uma sequência não codificante única de cada família e de origem incerta. A região 5' possui alta similaridade com outras estruturas que são transcritas pela RNA polimerase III, incluindo seus promotores. Já a região 3' frequentemente possui similaridade com a região 3' de um TE não-LTR do genoma do hospedeiro, dando indícios que essa região possibilita os SINEs usarem a maquinaria enzimática de LINEs similares, num mecanismo de *trans mobilização* (KAJIKAWA; OKADA, 2002; KRAMEROV; VASSETZKY, 2011; OHSHIMA; OKADA, 2005). A estrutura geral de um SINE é similar à mostrada na Figura 3, onde se vê a estrutura de um SINE de tRNA. Embora a maior parte dos SINEs tenha seu evento de transposição iniciado com a transcrição da cópia do SINE por uma RNA pol III, alguns SINEs também podem ser transcritos por uma RNA pol II em pré-mRNAs (KRAMEROV; VASSETZKY, 2011).

Os SINEs foram descritos inicialmente no início da década de 1980 em primatas (*Alu*) e em camundongos (B1) onde as cópias estavam frequentemente flanqueadas por repetições diretas (SINGER, 1982). Alguns anos depois foi mostrado que os SINEs *Alu* possuíam muita semelhança com os RNAs 7SL. As regiões 5' e 3' desse RNA possuíam homologia com os *Alu*,

embora sua região interna não mostrasse qualquer identidade, sendo então proposto que o SINE *Alu* fosse derivado de deleção da região central do 7SL RNA (ULLU; TSCHUDI, 1984).

Figura 3 – Estrutura geral de um SINE de tRNA.



Fonte: Ohshima e Okada (2005).

Em 2003 foi então descrito um tipo de SINE que não era derivado de tRNAs ou de 7SL RNA, mas sim do 5S rRNA. Esse elemento, o SINE3, está presente em pelo menos  $10^4$  cópias no genoma do peixe zebra, e sua região 3', de aproximadamente 60 pb, se assemelha a um TE do clado *CRI*, da superfamília *Jockey*. Foi então criada a hipótese de que esse SINE3 teria surgido da inserção de um *CRI* truncado na região 5' próximo de onde se retrotranspôs um mRNA de 5S rRNA. Um evento de transcrição pôde então ter gerado um mRNA quimérico contendo o 5S rRNA, uma região do genoma entre o 5S rRNA e o *CRI* truncado e por último o *CRI* truncado. Finalmente, a transcriptase reversa produzida pelo *CRI* autônomo reconheceria o fragmento de *CRI* e então amplificaria essa nova estrutura pelo genoma (KAPITONOV; JURKA, 2003a). Mais recentemente foi descrito um novo TE cuja região 5' é derivada de um *small nuclear RNA* (snRNA) sendo classificado pelos descritores como um tipo alternativo de SINE (KOJIMA, 2015).

Os SINEs podem estar integrados com a região codificadora de genes, podendo ser expressos como proteínas quiméricas. Mais de 200 SINEs já foram encontrados em mRNAs de humanos, ou como acentuadores e silenciadores de genes (OHSHIMA; OKADA, 2005). É importante notar que devido ao seu modo de nascimento, novas famílias de SINEs parecem estar surgindo *de novo* durante a evolução das espécies (KRAMEROV; VASSETZKY, 2011).

### 2.3 OS ELEMENTOS DE CLASSE II

De acordo com o código proposto por Wicker *et al.* (2007), os elementos de Classe II se dividem em duas subclasses. A primeira abrange os chamados “*cut and paste*” clássicos, contendo as ordens TIR, que possui uma variedade imensa de superfamílias de TE, e a ordem Crypton. A subclasse II abrange duas ordens: Helitron e Maverick.

### 2.3.1 Elementos da Ordem TIR

Elementos dessa ordem são vastos. Novos elementos vêm sendo descobertos ao longo dos anos e alguns deles não se enquadram nas superfamílias formuladas por Wicker *et al.* (2007). De acordo com esse esquema existem nove ordens. Enquanto algumas são comuns em insetos, outras são extremamente raras nesses organismos. A característica comum aos elementos dessa ordem é a presença de terminações repetidas invertidas (TIRs).

#### 2.3.1.1 Superfamília *CACTA*

O nome *CACTA* deriva do motivo mais frequente no início das TIRs da sequência de DNA deste grupo de TEs em plantas (WICKER *et al.*, 2007). Os TEs dessa superfamília também são conhecidos pelo nome *EnSpm*. Esse nome é uma abreviação de *Enhancer/Suppressor-Mutator*, dois elementos que foram descobertos na década de 1950 (MCCLINTOCK, 1954; PETERSON, 1960). O último, o *Spm*, é um elemento que possui duas ORFs: uma delas codifica a transposase e a outra codifica uma proteína regulatória da atividade do TE (FEDOROFF, 2013). Entretanto, para algumas famílias de *CACTA*, a função da segunda ORF não está comprovada. As proteínas desse elemento podem ser codificadas por diversos éxons, variando de um a onze, em plantas. Essa diversidade de número de éxons reflete diretamente no tamanho de cada ORF. Enquanto algumas possuem cerca de 500 aminoácidos, em outras esse valor pode chegar a surpreendentes 4785 aminoácidos (BUCHMANN *et al.*, 2014). Evidências mostram a presença desses múltiplos éxons faz com que possa ocorrer *splicing* alternativo para produção de múltiplas isoformas das proteínas destes TEs (DEMARCO; VENANCIO; VERJOVSKI-ALMEIDA, 2006; MASSON *et al.*, 1989).

Esses elementos são um dos mais abundantes em plantas, onde a maior parte dos estudos sobre esses TEs está disponível. Entretanto, eles também ocorrem em menor quantidade em fungos e em diversos filos de animais como platelmintos, equinodermos, cnidários, cordados e artrópodes, entre eles, os mosquitos *Ae. aegypti* e *An. gambiae* (DEMARCO; VENANCIO; VERJOVSKI-ALMEIDA, 2006; TANG *et al.*, 2015).

#### 2.3.1.2 Superfamília *PIF-Harbinger*

O nome PIF deriva de *P instability factor* (Fator de instabilidade do elemento *P*) no genoma do milho (WALKER *et al.*, 1997) que, juntamente com *PIFa* (PIF autônomo),

formaram os primeiros elementos *PIF*. Foi visto inicialmente que os elementos *PIFa* codificavam uma proteína com homólogos em *Arabidopsis thaliana*, arroz, sorgo, nematódeos e fungos. O compartilhamento de características em espécies distintas e a similaridade dos *PIF* com o TE *Harbinger* em *A. thaliana* (KAPITONOV; JURKA, 1999) levou à criação dessa superfamília (ZHANG *et al.*, 2001). Os elementos autônomos dessa família codificam uma transposase com cinco motivos conservados, além de uma proteína de ligação ao DNA com um domínio SANT/myb/trihelix conservado que é necessário para a expressão da transposase. Além disso, esses TEs possuem TIRs que variam de 10 a 700 nucleotídeos (KAPITONOV; JURKA, 2004). Outra característica dos *Harbinger* é que eles estão envolvidos com a transposição de um MITE (*Miniature Inverted-repeat Transposable Element*, discutidos mais adiante), o elemento *Tourist-like* no milho e em outras plantas (ZHANG *et al.*, 2001). O primeiro consenso de *Harbinger* identificado em *An. gambiae* possuía TIRs imperfeitas e codificava uma transposase de 471 aminoácidos e uma proteína de ligação ao DNA de 245 aminoácidos (KAPITONOV; JURKA, 2004). Em outra espécie do gênero *Anopheles*, *An. funestus*, também já foi descrita a presença de *Harbinger*, inclusive de um *Harbinger* transcricionalmente ativo (FERNÁNDEZ-MEDINA *et al.*, 2017).

Essa superfamília de TEs está presente em pelo menos 75 espécies de plantas, animais e fungos, com 88 famílias potencialmente ativas, estando ausente nos mamíferos, aves, esponjas, tubarões e celacantos (HAN *et al.*, 2015).

### 2.3.1.3 Superfamília *Tc1-Mariner*

Essa superfamília de elementos de transposição é composta pelos TEs mais amplamente dispersos nas formas vivas, desde fungos a plantas, protozoários e diversos tipos de animais (PLASTERK; IZSVÁK; IVICS, 1999). Ela é composta por famílias como *Tc1*, um elemento primeiramente descrito em diferentes linhagens de *Caenorhabditis elegans*, com um tamanho de 1,7 kb, alta variação do número de cópias e estrutura similar aos elementos IS de bactérias (EMMONS *et al.*, 1983); *mariner* um elemento descrito inicialmente em *Drosophila mauritiana* com 1286 nucleotídeos codificando uma ORF de 356 aminoácidos, e que causava mutação na cor dos olhos dessa espécie (JACOBSON; MEDHORA; HARTL, 1986); *pogo* um elemento primeiramente caracterizado em *Drosophila melanogaster* com tamanho variando de cerca de 1 kb a cerca de 2 kb, com múltiplas cópias truncadas de cerca de 200 nucleotídeos e que foi inicialmente colocado em um grupo de TEs que incluía o elemento *P* e o elemento *hobo*



(TUDOR *et al.*, 1992); **tigger** um elemento de 2,5 kb de tamanho e com duas ORFs cujo produto é muito similar ao do *pogo*, sendo descrito inicialmente no genoma humano (SMIT; RIGGS, 1996); **gambol**, uma família descoberta em *An. gambiae*, e que é similar aos Tc1 (COY; TU, 2005).

Outra denominação para essa superfamília é dada pela sigla ITm, onde o **T** representa a família *Tc1*, o **m** a família *mariner* e o **I** o elemento IS630 de bactérias (SHAO; TU, 2001). Os elementos dentro dessa superfamília divergem bastante em relação à sequência primária da sua transposase, com até cerca de 85% de divergência, embora exista uma provável origem monofilética. A região N-terminal das transposases dessa superfamília possui motivos de hélice-volta-hélice, que estão relacionados à ligação com DNA e que variam em quantidade de acordo com a família considerada (PLASTERK; IZSVÁK; IVICS, 1999).

O domínio catalítico das transposases dessa superfamília possui o motivo DDE ou DDD. O D representa o aminoácido aspartato, enquanto o E representa o aminoácido glutamato. Esses aminoácidos não se seguem um após o outro na sequência primária da proteína, eles são intercalados com outros aminoácidos. A distância entre os dois primeiros Ds é variável, já entre o segundo D e o último D ou E é frequentemente invariável dentro de uma família de ITms (SHAO; TU, 2001). De acordo com essas características e com análises filogenéticas, foram identificadas famílias com oito motivos diferentes como mostrado no Quadro 2.

#### 2.3.1.4 Superfamília *hAT*

Os primeiros elementos de transposição foram identificados no genoma do milho na década de 1940, eles foram chamados de *Ac/Ds*. Os elementos *Ac* (*Activator*) são membros da superfamília *hAT*. Desta forma, esta é a superfamília que conhecemos há mais tempo. Não apenas sua descoberta é antiga, mas a própria superfamília possui uma origem ancestral. Esses elementos possuem TIRs pequenas, variando no geral entre 5 e 17 nucleotídeos, e seu tamanho total é de cerca de 4 kb. Os *hAT* também possuem uma característica de gerar duplicações de oito nucleotídeos no seu sítio de inserção (KEMPKEN; WINDHOFER, 2001). A ORF dos *hAT* possuem quatro domínios funcionais: o domínio BED, um domínio de ligação de DNA presentes em todas as transposases dos *hAT*; um segundo domínio de ligação; um domínio catalítico; um domínio de inserção entre o segundo D e o E da tríade DDE, uma das regiões menos conservadas dos *hAT* (ATKINSON, 2015).

Quadro 2 – Diferentes transposases da superfamília Tc1-mariner

Motivo	Famílias e subfamílias	Referências
DD34D	Mariner (Hsmar1, Himar1, Mos1, Hsmar2)	(JACOBSON; MEDHORA; HARTL, 1986; TELLIER; BOUUAERT, 2015)
DD37D	maT	(CLAUDIANOS <i>et al.</i> , 2002)
DD39D	Soymar1 (ancestral dos mariner), OSMAR1	(JARVIK; LARK, 1998; TARCHINI <i>et al.</i> , 2000)
DD41D	<i>rosa</i> , <i>leidy</i>	(PUZAKOV; PUZAKOVA, 2019; ZHANG <i>et al.</i> , 2016)
DD34E	Tc1, Gambol	(COY; TU, 2005; EMMONS <i>et al.</i> , 1983)
DD35E	IS630	(MATSUTANI <i>et al.</i> , 1987; TENZEN; MATSUTANI; OHTSUBO, 1990)
DD37E	ITmD37E	(BIEDLER; SHAO; TU, 2007; SHAO; TU, 2001)
DDDxD	pogo, Tigger	(SMIT; RIGGS, 1996; TUDOR <i>et al.</i> , 1992)

**Fonte:** Elaborado pelo autor

Apesar de ser uma superfamília conhecida há bastante tempo, poucos TEs completos e com TIRs e ORFs intactas são conhecidos. Um deles é o elemento *Buster* em *Ae. aegypti*. O uso de ferramentas biotecnológicas pode auxiliar na ressurreição de *hAT* inativos, transformando-os em elementos capazes de se transpor e que assim podem ser utilizados como vetores biotecnológicos, como foi realizado com o transposon *SPIN* (LI *et al.*, 2013b). Por ter uma capacidade de se inserirem e permanecerem ativos no genoma de diferentes espécies, os *hAT* vem sendo constantemente usados como ferramentas de transferência de genes (ATKINSON, 2015).

Os elementos da superfamília *hAT* são dispersos entre as mais variadas espécies de eucariotos, abrangendo fungos, plantas, animais, e alguns protistas, como as amebas. Há diferenças significativas entre as ORFs dos *hAT*, sugerindo uma divisão em dois grupos bem definidos (ambos com diferenças claras de especificidade de sítio alvo): os *Buster*, mais comuns em animais e os *Ac*, mais dispersos em diversos reinos de eucariotos (ARENSBURGER *et al.*, 2011a; ATKINSON, 2015). Um terceiro grupo de *hAT*, chamado de *Tip*, também foi proposto. Ele é mais próximo aos *Buster*, que ao *Ac* e é encontrado em plantas, cnidários e insetos das ordens Lepidoptera, Hemiptera, Hymenoptera, Coleoptera e Diptera (ROSSATO *et al.*, 2014).

### 2.3.1.5 Superfamília *PiggyBac*

Essa superfamília teve uma peculiaridade em sua descoberta, 20 anos se passaram entre a descoberta do primeiro elemento, numa mariposa, e a descoberta do segundo elemento em outra espécie. O período de 20 anos entre a descoberta do primeiro e o segundo elemento ocorreu numa época em que muitos elementos de transposição passaram a ser descobertos e estudados, entre os anos 1980 e 2000. Esse primeiro elemento foi identificado graças a uma transferência do próprio elemento realizada a partir do genoma da mariposa para o genoma de um vírus que a infectava. Ele tinha TIRs muito pequenas com 13 pb e seu tamanho médio era de 2475 pb. Sua ORF com 1800 pb codificava uma transposase de aproximadamente 600 aminoácidos (YUSA, 2015).

Sabe-se hoje que os elementos dessa superfamília não possuem transposases com o motivo catalítico DDE conservado que ocorre na maior parte das transposases dos elementos da ordem TIR. Ao invés disso, eles possuem um motivo DDD-D/G (quatro aspartatos (D) com tolerância para uma glicina (G) ou um glutamato (E) na última posição) (BOUALLÈGUE *et al.*, 2017; KOJIMA, 2019).

Boa parte dos elementos dessa superfamília mostra uma inserção preferencial, transpondo-se para as adjacências de sequências TTAA. Entretanto, transposições raras para outros 5 sítios também são encontradas. Uma análise envolvendo dezenas de elementos dessa superfamília mostrou que as TIRs desses elementos estão entre as menores dos transposons, possuindo cerca de 14 pb, além disso esses TEs podem chegar a cerca de 8500 pb de tamanho. Uma característica interessante é que esses TEs podem ter estruturas como repetições diretas ou TIRs subterminais associadas as terminações TIRs clássicas. Devido a essas variações foi proposta a divisão dessa superfamília em 4 grupos estruturais. Enquanto algumas espécies possuem apenas um dos grupos, outras possuem todos eles (BOUALLÈGUE *et al.*, 2017).

Os TEs dessa superfamília são bastante dispersos em animais, entre eles, muitos insetos (PETERSEN *et al.*, 2019). Entretanto, eles também ocorrem em uma diversidade de eucariotos como protozoários, algas vermelhas, mamíferos, chegando a colonizar até mesmo alguns vírus. A única exceção é a sua ausência em genomas de plantas (BOUALLÈGUE *et al.*, 2017). Sua ampla dispersão e estudo possibilitou o amplo uso desse TE como ferramenta biotecnológica em todos os filões de eucariotos, principalmente entre animais, com destaque para 5 ordens de insetos (YUSA, 2015). Entre eles, vetores de relevância na saúde pública, como mosquitos dos

gêneros *Anopheles* e *Aedes*, onde os TEs *PiggyBac* foram usados como *promoter-traps* (LOBO *et al.*, 2001; REID *et al.*, 2018).

#### 2.3.1.6 Superfamília *Transib*

Esses elementos foram inicialmente caracterizados no genoma de *D. melanogaster*. Eles eram TEs pequenos e degenerados, com menos de 2,2 kb cujas TIRs possuíam cerca de 45 pb e ORFs que codificavam transposases com tamanho variando entre 400 e 600 pb. Sabe-se hoje que formas autônomas desses TEs possuem entre 3 e 4 kb, ORFs com cerca de 2,1 kb e TIRs que podem ultrapassar os 500 pb (CHEN; LI, 2008; KAPITONOV; JURKA, 2003b). Esses elementos também já foram encontrados no genoma de várias outras espécies do gênero *Drosophila*, em vespas *Nasonia*, no besouro *Tribolium castaneum*, em *An. gambiae*, *Ae. aegypti*, alguns vermes nematódeos, lepidópteros, fungos e hidrozoários (CHEN; LI, 2008; KAPITONOV; JURKA, 2003b, 2005).

Todos os elementos dessa superfamília possuem apenas uma ORF, com uma peculiaridade que os colocou em bastante destaque no estudo de TEs. Isso deriva do fato da transposase codificada por elementos dessa superfamília possuir características bastante distintas das transposases dos outros TEs. Ao invés da similaridade com TEs, a transposase dos *Transib* possui alta similaridade com cerca de 600 aminoácidos da proteína RAG1 do sistema imune de gnatostomados (KAPITONOV; JURKA, 2005).

Além da similaridade de sequência de aminoácidos, a transposase dos *Transib* realiza uma função molecular igual a função exercida pela RAG1 no que diz respeito à clivagem do DNA. Além disso, a transposase do *Transib* ainda possui estruturas adicionais que realizam um papel similar à função da proteína RAG2 (ambas RAG1 e RAG2 são proteínas necessárias para a recombinação V(D)J). Essas estruturas adicionais, similares a RAG2, dão à transposase desse TE, uma capacidade de realizar um papel complexo em sua transposição, uma vez que essa única proteína do TE é capaz de realizar funções moleculares que são realizadas por um par de proteínas (RAG1 e RAG2) dos vertebrados (LIU; YANG; SCHATZ, 2019). Apesar dessa diferença de números, a alta similaridade entre esse TE e as proteínas RAG permitiram a proposição de que tanto as proteínas RAG1 quanto as RAG2 evoluíram a partir de um único *Transib* ancestral, que possuiria duas regiões codificantes em sua sequência nucleotídica (KAPITONOV; KOONIN, 2015).

### 2.3.1.7 Superfamília *Mutator*

Os elementos dessa superfamília foram primeiramente descobertos no genoma do milho. Seu nome possui uma relação direta com seu efeito, eles são altamente mutagênicos. Nessa espécie, suas terminações possuem um tamanho longo, cerca de 220 nucleotídeos, já sua região interna é variável (LISCH, 2002). Considerando todas as famílias de *Mutator* conhecidas, suas TIRs variam de 14 a 500 nucleotídeos, havendo algumas famílias que nem mesmo possuem TIRs. Embora eles sejam extremamente dispersos em angiospermas, também existem em fungos, protistas e animais. Recentemente foram descobertas duas famílias de *Mutator* em insetos: *Ghost* e *Spectre*, entretanto sua ocorrência é mais comum em besouros e afídeos. Os autores do estudo, todavia, sugeriram que muitos elementos *Mutator* ainda estão aguardando sua descoberta em artrópodes (DUPEYRON *et al.*, 2019). Essa superfamília também já foi encontrada em mosquitos. No ano de 2017, foi descoberto um *Mutator*, o *Muta1*, que potencialmente se encontra ativo no genoma de *Ae. aegypti* (LIU; WESSLER, 2017).

Devido ao seu caráter mutacional, esses elementos tiveram suas propriedades bem estudadas e se tornaram uma importante ferramenta genética no estudo de elementos de transposição e seus efeitos (LISCH; JIANG, 2009).

### 2.3.1.8 Superfamília *Merlin*

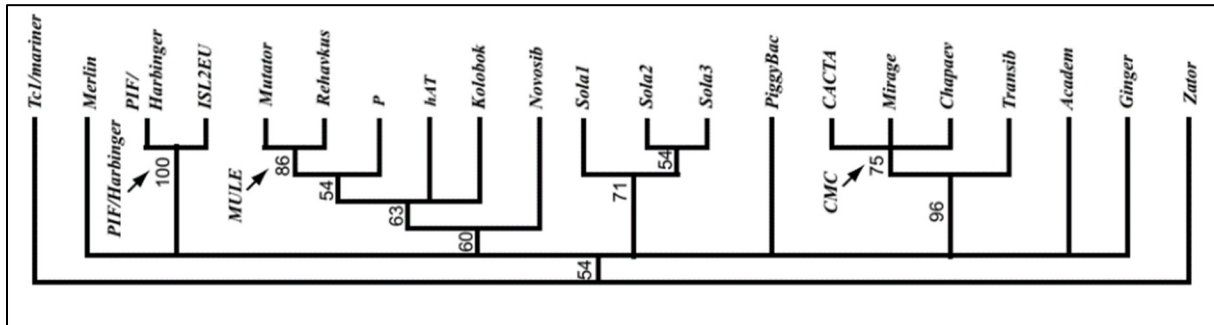
Essa é uma superfamília pouco estudada, estando seus elementos majoritariamente presentes em animais, com bem menos ocorrência em fungos Zygomycota e plantas licófitas (YUAN; WESSLER, 2011). Um grupo muito similar de TEs, os elementos *IS1016*, ocorrem em bactérias. Com os elementos *IS1016*, os *Merlin* compartilham TIRs similares e duplicações de inserção de 8 a 9 nucleotídeos. Além disso, possuem cerca de 50% de similaridade entre suas sequências proteicas, com uma região conservada na região C-terminal da proteína transposase (FESCHOTTE, 2004).

## 2.3.2 Elementos TIR não inclusos no esquema de Wicker *et al.* (2007)

Desde a criação do esquema de classificação por Wicker *et al.* (2007) alguns transposons que possuem TIRs foram descobertos. Parte deles não se encaixam neste sistema e formam grupos separados, enquanto outros são agrupados como elementos das superfamílias existentes.

Uma filogenia consenso inferida a partir da presença ou ausência dos resíduos de aminoácidos altamente conservados das transposases e mostrando diversos grupos de TIRs pode ser vista na Figura 4. Alguns novos grupos de TEs que possuem TIRs são *Novosib*, *Kolobok*, *ISL2EU*, *Sola*, *Zator*, *Ginger1*, *Ginger2*, *Academ*, *Zisupton*, *IS3EU*, *Chapaev* e *Mirage* (YUAN; WESSLER, 2011).

Figura 4: Árvore consenso dos elementos TIRs



**Fonte:** Adaptado de Yuan e Wessler (2011).

**Legenda:** Os valores numéricos próximos aos nós representam os valores de *bootstrap*.

Os *ISL2EU* são TEs muito similares aos *PIF-Harbinger* e, dessa forma, inseridos nesta superfamília. A descoberta de outros elementos como o *Spy*, que estaria evolutivamente relacionado com esses dois grupos, levou a proposição de uma grande superfamília PHIS que englobaria todos esses TEs (HAN *et al.*, 2014). Os elementos *Rehavkus* são muito similares à superfamília *Mutator*, e dessa forma foram inseridos nela. Outro caso de similaridade de novos grupos a superfamílias bem estabelecidas é o agrupamento de *CACTA*, *Chapaev* e *Mirage* no mesmo clado (Figura 4), o que levou à proposição da superfamília CMC (*CACTA*, *Mirage*, *Chapaev*) (YUAN; WESSLER, 2011).

Sabe-se muito pouco de alguns elementos. Alguns deles foram descritos apenas no banco de dados Repbase, como *Novosib*, *Academ* e *IS3EU*. Outros grupos também não são muito estudados, como *Kolobok*, que existem em algumas plantas, animais e protozoários; *Zisupton*, que são encontrados em cordados, hemicordados, urocordados, alguns fungos e algas; *Ginger*, um grupo de TIRs estritamente relacionados com o LTR *Gypsy* que está presente em animais e plantas que não produzem sementes (BÖHNE *et al.*, 2012; KOJIMA, 2019; YUAN; WESSLER, 2011)

Enquanto *Novosib* e *Kolobok* parecem agrupar com *hAT*, *P* e *Mutator* (Figura 4), os grupos *Academ* e *Ginger* parecem não ser relacionados com nenhuma outra superfamília descrita e não possuem características estruturais similares a nenhum outro grupo (YUAN;

WESSLER, 2011). Outro grupo que não possui qualquer similaridade com qualquer outra superfamília descrita por Wicker *et al.* (2007) são os *Sola*, que são formados por 3 linhagens: *Sola1*, *Sola2* e *Sola3*. O grupo interessante está presente tanto em procariotos quanto em eucariotos. Nesses últimos eles ocorrem em protistas, plantas e animais, dentre eles os mosquitos (BAO *et al.*, 2009; GOUBERT *et al.*, 2015). Outros TEs de ocorrência em mosquitos são os *Zator*, um grupo que ocorre em protistas e animais e que provavelmente evoluiu de um TE bacteriano. Esse grupo de TEs aparentemente possui alguma relação com a superfamília *Tc1-Mariner* (BAO *et al.*, 2009)

### 2.3.3 MITEs

Alguns dos TEs de Classe II são muitos pequenos, possuem TIRs, mas não codificam proteínas que auxiliam na sua transposição, esses elementos são difíceis de classificar uma vez que pouco se sabe sobre as peculiaridades com que cada família se multiplica dentro de um genoma. A esses elementos se deu o nome de MITEs, ou elementos de transposição de repetições invertidas em miniatura (*Miniature Inverted-repeat Transposable Elements*). Esses elementos estão presentes em grande número de cópias dentro do genoma das espécies, uma parte deles muito provavelmente surgiu de deleções de elementos TIRs completos, como o MITE *Emigrant* no genoma de *A. thaliana* (FESCHOTTE; MOUCHÈS, 2000) e o elemento *mPing* que deriva de um elemento *Ping* autônomo (KIKUCHI *et al.*, 2003). Entretanto, o grande número de cópias das famílias de MITEs presentes nos genomas de muitas espécies não se deriva de deleções diretas de TEs autônomos (FATTASH *et al.*, 2013). Deve-se separar o momento em que um MITE surge, de sua amplificação. É possível que TEs não autônomos precursores de MITEs surjam a partir de deleções de elementos autônomos que foram eliminados do genoma, e que o grande número de cópias derivado de diversas mobilizações seja fruto do uso da maquinaria enzimática de outros elementos (JIANG *et al.*, 2004).

Um dos primeiros MITEs estudados foi o elemento *Tourist*, que foi caracterizado no genoma do milho, esse elemento tem 128 nucleotídeos de extensão e elementos de sua família tem tamanhos próximos, ele também possui TIRs conservadas flanqueadas por repetições diretas, e são frequentemente encontrados em associação com genes no milho (BUREAU; WESSLER, 1992). MITEs também são frequentemente encontrados dentro de genes em outras espécies além do milho e alguns se sobrepõe diretamente com regiões codificantes (0,3% dos

MITEs de *Oryza sativa*), por esse motivo eles foram implicados com impacto direto em mecanismos de regulação gênica e na estrutura gênica (OKI *et al.*, 2008).

Um MITE bastante estudado é o *Stowaway*. Esse elemento, extremamente abundante no genoma de plantas, se multiplica a partir do uso de uma transposase de um TE relativamente distante, um *Mariner* denominado de *Osmar*, através de um mecanismo de *cross-mobilization* (mobilização por um elemento semelhante, porém de origem diferente), que pode ser o mesmo mecanismo usado pelos MITEs para se amplificar dentro dos genomas de diversos outros hospedeiros (YANG *et al.*, 2009).

Poucos MITEs ativos já foram identificados, *dTstul* pertencente à família *Stowaway* é capaz de se mobilizar no genoma do Tomate, se inserindo dentro de um éxon de um gene (MOMOSE; ABE; OZEKI, 2010). Entretanto, o MITE ativo mais estudado é o mPing (*Ping* em miniatura) que é associado ao elemento *Ping*, um *PIF-Harbinger*, que se move pelo genoma do arroz, *O. sativa*. O mPing é derivado de deleções do *Ping* (KIKUCHI *et al.*, 2003) e para sua transposição é necessária a expressão das ORF1 e ORF2 desse elemento *Ping* ou do elemento *Pong* através de um mecanismo conhecido como *cross-mobilization*, onde um elemento não autônomo é transposto utilizando-se das enzimas de um elemento funcional de outra linhagem (YANG *et al.*, 2007).

Os MITEs estão presentes em vários dos grandes grupos de organismos como plantas, animais, fungos e bactérias, entretanto, não existe correlação entre o número de MITEs e o percentual de TEs de um organismo (FATTASH *et al.*, 2013). Mais recentemente foi demonstrado que é comum a ocorrência de MITEs em algumas famílias de vírus, como *Ascoviridae*, *Poly-dnaviridae* (PDV) and *Pandoraviridae*. Em um dos vírus da última, os MITEs estão presentes em até 251 cópias, 1,5% do genoma viral. Esses MITEs provavelmente colonizam vírus através de transferência gênica entre vírus e hospedeiros (ZHANG *et al.*, 2018).

MITEs já foram encontrados no genoma de *Ae. aegypti* e *An. gambiae* (TU, 2001). Nesta última espécie, alguns MITEs derivam de elementos da superfamília P que invadiram o genoma de *A. gambiae*, porém a região interna desses MITEs, região entre as TIRs, não apresenta similaridade com os elementos parentais, provavelmente tendo origem numa sequência ectópica (QUESNEVILLE; NOUAUD; ANXOLABÉHÈRE, 2006). Em *A. aegypti* já foram reportados 143 famílias de MITEs correspondendo a 16% do genoma da espécie (NENE *et al.*, 2007), algumas delas possivelmente ativas como as famílias *Aton-I* e *Aton-II* que possuem 112



pb e 144 pb respectivamente, e 294 e 317 cópias, muitas delas idênticas entre as cópias das famílias, indicando uma atividade recente (YANG; WONG; ROOKE, 2012).

### 2.3.4 Elementos da Ordem Crypton

Esta ordem de TE possui apenas uma superfamília, de mesmo nome. Esta superfamília de TEs é bastante distinta das outras famílias de TE da Classe II. Os TEs *Crypton* não possuem domínios de transposase, domínios de RT ou até mesmo regiões terminais marcadas por estruturas bem definidas como LTRs ou TIRs. Em sua estrutura principal, eles possuem apenas um domínio de Tirosina recombinase, que é frequentemente encontrado em bactérias, porém raro em eucariotos, sendo descritos apenas compondo TEs (KOJIMA; JURKA, 2011; WICKER *et al.*, 2007). Esse tipo de elemento de transposição foi originalmente descrito em vários fungos patogênicos, tendo tamanho aproximadamente de 4 kb e presente em baixo número de cópias. O elemento originalmente descrito também possuía uma longa ORF interrompida por pequenos íntrons (GOODWIN; BUTLER; POULTER, 2003).

Os Cryptons foram posteriormente identificados em diversas outras espécies. Dentre elas, estão presentes espécies de fungos, stramenopiles, diatomáceas e animais, incluindo artrópodes, cordados, equinodermas, moluscos e cnidários, onde geralmente representam cópias não autônomas (KOJIMA; JURKA, 2011). Eles inclusive estão presentes de forma inativada e em baixo número de cópias no genoma humano (KOJIMA, 2018). Em insetos, os Cryptons possuem uma estrutura um pouco distinta da estrutura encontrada em outros animais, provavelmente derivada de uma evolução mais rápida (KOJIMA; JURKA, 2011).

### 2.3.5 Elementos da Ordem Maverick

Assim como a ordem Crypton, a ordem Maverick também possui apenas uma superfamília, de mesmo nome. Embora possuam terminações repetitivas invertidas (TIR) eles são isolados numa ordem separada, porque possuem características próprias e bem distintas de quaisquer TEs de classe II (WICKER *et al.*, 2007). Esses elementos também são chamados de *Polintons* e são conhecidos desde 2005. Por primeiramente se encontrar apenas uma de suas ORFs em diversas espécies e pela pressão de seleção negativa observada, inicialmente esse TE foi classificado erroneamente como mais um gene do organismo. Entretanto, características

como a presença de TSD e TIRs indicaram que se tratava de um elemento de transposição (FESCHOTTE; PRITHAM, 2005).

Os *Mavericks* são elementos muito grandes, um dos maiores entre todos os TEs, sendo conhecido como um dos transposons gigantes (ARKHIPOVA; YUSHENOVA; ANGERT, 2019). Seu tamanho também reflete na sua complexidade. Um Maverick pode chegar a 40 kb de tamanho e codificar mais de uma dezena de ORFs. Suas proteínas mais comuns são uma DNA polimerase B e uma C-integrase, contudo ainda possuem uma série de proteínas não bem caracterizadas (PRITHAM; PUTLIWALA; FESCHOTTE, 2007). Alguns elementos dessa superfamília são capazes de expressar proteínas com a mesma estrutura de um capsídeo viral e dessa forma potencialmente geram virions. Adicionalmente eles codificam proteínas homólogas a proteínas virais: FtsK-ATPase e *cysteine protease*. Assim como alguns elementos LTRs são considerados um misto de vírus de RNA com elementos de transposição, os *Mavericks* passaram a ser considerados por alguns autores como um tipo especial que alterna entre “estilo de vida” como TE ou como vírus de DNA, chamados de Polintonvírus (KRUPOVIC; BAMFORD; KOONIN, 2014). Foi proposto então que os *Mavericks* evoluíram a partir de bacteriófagos e originaram os atuais vírus gigantes da ordem Megavirales assim como Adenovírus e vírus de invertebrados da família *Bidnaniridae*, como mostrado na Figura 5 (KOONIN; KRUPOVIC; YUTIN, 2015; KRUPOVIC; KOONIN, 2016).

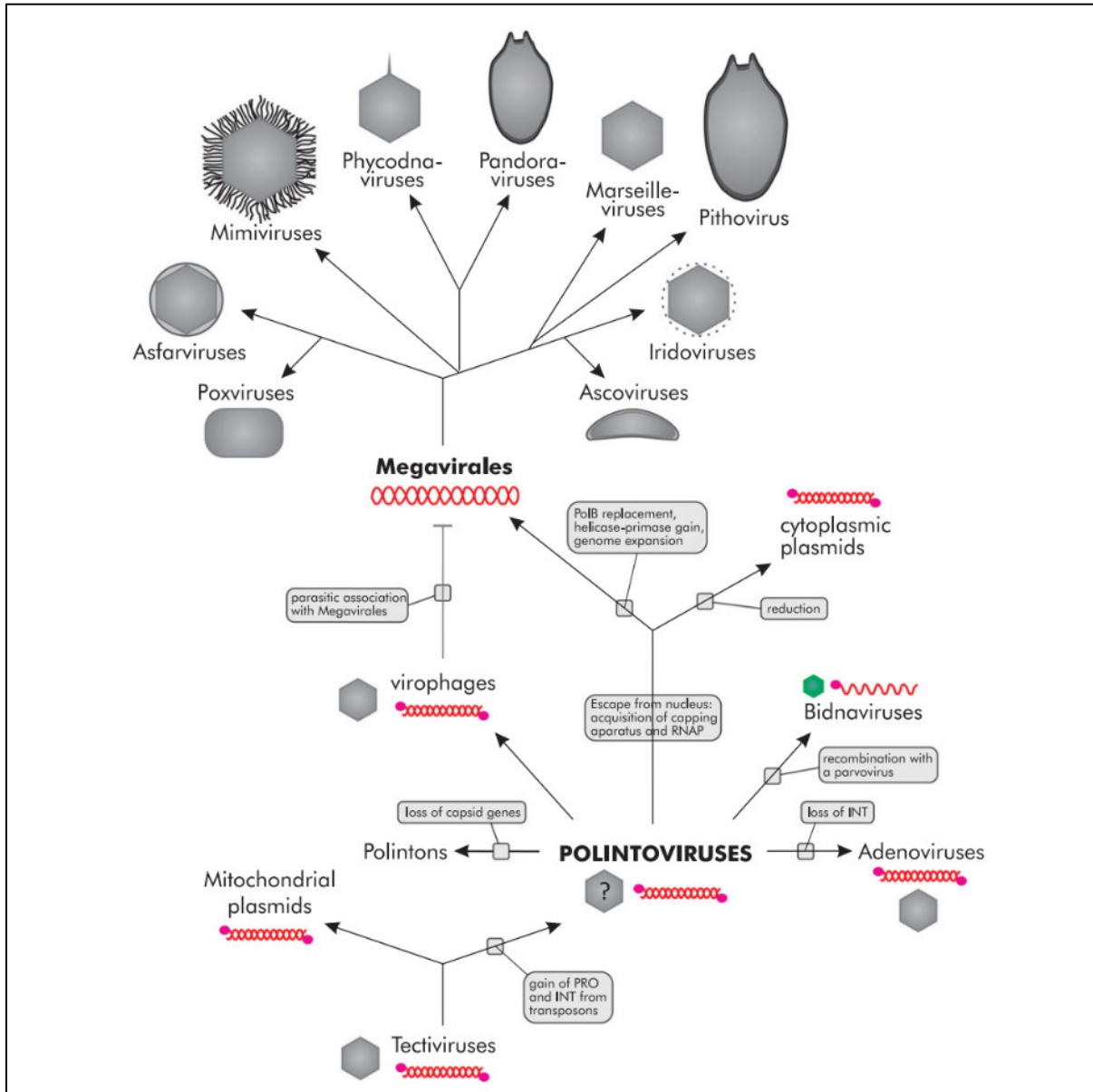
Esses elementos são comuns em animais não mamíferos, mas são raros em outros reinos e não são encontrados em plantas. Em algumas espécies eles podem ocupar grande parte do genoma, como em *Trichomonas vaginalis*, onde cerca de 30% do genoma é formado por *Mavericks* (PRITHAM; PUTLIWALA; FESCHOTTE, 2007). Esses elementos são comuns no gênero *Drosophila*, ocorrendo em nove espécies. Uma análise filogenética com as proteínas DNA pol-B e C-integrase mostrou que duas linhagens diferentes de *Mavericks* existem em insetos, essas linhagens são tão antigas que provavelmente divergiram antes dos ancestrais dos Cnidários (HAAPA-PAANANEN; WAHLBERG; SAVILAHTI, 2014).

### 2.3.6 Elementos da Ordem Helitron

Esses elementos, assim como *Maverick*, representam uma superfamília que é o único componente de uma ordem que leva o mesmo nome. Diferentemente da maior parte dos outros TEs, os *Helitron* não produzem duplicações nos flancos do elemento. Eles também não possuem limites muito bem definidos, uma vez que não possuem LTRs nem TIRs. Esses

elementos foram definidos no ano de 2001 no genoma de plantas e no verme *C. elegans*. No geral são elementos muito grandes, podendo alcançar cerca de 15 kb de comprimento, alguns com íntrons e éxons. No entanto, a maior parte dos *Helitrons* se encontra na forma de elementos não autônomos de tamanho bem menor, mas que conservam características que permitem que sejam transpostos por helicases de *Helitrons* autônomos (KAPITONOV; JURKA, 2001).

Figura 5 – Origem e evolução proposta para os *Mavericks* (*Polintons*)



**Fonte:** Koonin, Krupovic e Yutin (2015).

**Legenda:** Segundo o esquema proposto, os Mavericks (Polintons) são formas de Polintovírus que perderam a capacidade de produzir o capsídeo viral. Estes possivelmente evoluíram a partir de bacteriófagos da família *Tectiviridae*, ao adquirir genes de cisteína protease e integrase de transposons de Classe II da superfamília Ginger. É provável que no passado, Mavericks que permaneceram com o capsídeo tenham gerado os atuais Adenovírus, através da perda do gene de integrase adquirido, ou vírus de fita de simples de DNA como os Bidnavírus. Através do escape do ambiente nuclear da célula e da aquisição de genes de transcriptase reversa e genes da maquinaria de adição do Cap, os Polintovírus evoluíram até os atuais vírus gigantes da ordem Megavirales.

Existem 4 variantes estruturais (a exemplo de Helentrons) e diversas famílias de *Helitrons*. Estas diferem em suas terminações e na quantidade de proteínas e tipos de domínio que possui. Algumas famílias possuem em sua sequência até mesmo múltiplas repetições em *tandem*. Entretanto, sua característica principal, e que define todos esses TEs, é a presença de um gene que codifica para uma proteína com característica de helicase, ao invés de sequências para transcriptases reversas ou transposases características dos TIR. Além do domínio de helicase, que desenrola o DNA, a proteína do *Helitron* também possui um domínio de endonuclease, que está envolvido na quebra do DNA e um motivo de dedo de zinco que é responsável pela ligação ao DNA. Diferentes variantes de *Helitrons* podem carregar diferentes ORFs, algumas com função desconhecida, em adição a ORF principal que codifica a Helicase (THOMAS; PRITHAM, 2015).

Em seu modo de replicação próprio, o *Helitron* cliva uma das fitas de DNA de sua sequência, a excisa e a insere em outro local do genoma. Esse modo de replicação só foi elucidado em 2016 a partir da reconstrução de um *Helitron* ativo. Devido ao seu modo de transposição, esses TEs não raramente capturam genes do hospedeiro e os integram em sua sequência. Isso gera uma disseminação de genes e elementos regulatórios por todo o genoma, assim como facilita a construção de transcritos quiméricos. Essa captura também se deve ao fato de algumas famílias se inserirem preferencialmente em regiões de promotores gênicos, embora não haja preferência por regiões com maior adensamento gênico (GRABUNDZIJA *et al.*, 2016).

Os elementos dessa ordem são dispersos entre os diversos reinos de eucariotos, embora evidências indiquem que os *Helitrons* tenham se originado de elementos similares a plasmídeos que colonizavam genomas procarióticos, tendo invadidos os genomas dos eucariotos logo no início da evolução desse grupo (HERINGER; KUHN, 2022). Interessantemente os *Helitrons* são extremamente raros em mamíferos, mas muitos comuns em genomas de invertebrados. Petersen *et al.* (2019) encontraram evidências da presença de Helitrons em 62 genomas de artrópodes (100% das espécies do estudo). Além de eucariotos, estes TEs também já foram encontrados em vírus da família Polydnviridae, que é composta por vírus que infectam vespas, ao mesmo tempo em que foram encontrados nos seus hospedeiros (COATES, 2015; THOMAS; SCHAACK; PRITHAM, 2010).

## 2.4 OS CULICÍDEOS

Os mosquitos compõem os organismos da família Culicidae. Eles representam mais de 3500 espécies conhecidas e mais uma provável grande quantidade ainda desconhecida. Essa família faz parte da ordem Diptera, que provavelmente se originou há 260 milhões de anos e parte da infraordem Culicomorpha, que forma um clados basais dos dípteros (WILKERSON; LINTON; STRICKMAN, 2021).

Eles possuem um ciclo de vida que passa por quatro etapas, ovo, larva, pupa e adulto, ou seja, são holometábolos. As três primeiras fases são aquáticas, diferentemente do adulto. As larvas passam por quatro fases L1, L2, L3 e L4, e possuem muitas características morfológicas que auxiliam na taxonomia de cada espécie. A maior parte das pupas possui forma similar, um corpo com fusão de cabeça e tórax, formando um cefalotórax. Já os adultos são alados, com um par de asas alongadas, e corpo dividido em três grandes partes, cabeça, tórax e abdômen. Na cabeça estão os olhos, palpos, antenas e o aparelho bucal em forma de uma longa probóscide, que é uma das peças bucais mais especializadas entre organismos picadores. Já o tórax é uma unidade rígida que liga a cabeça e o abdômen e onde estão ligadas as patas e as asas. O abdômen, por sua vez, é uma região flexível, com capacidade de expansão, formada por segmentos visíveis (FOSTER; WALKER, 2019).

Uma fração dessas espécies de mosquito tem o sangue de organismos vertebrados como parte de sua alimentação. Por conta disso há mais facilidade no carregamento de patógenos, sendo os mosquitos, os primeiros artrópodes descritos como hospedeiros de parasitos de vertebrados (FOSTER; WALKER, 2019). Dessa forma, várias espécies mostram competência vetorial, ou seja, são capazes de se infectar ou permitir que agentes infecciosos se repliquem dentro delas. Parte destes mosquitos é capaz de levar diversos patógenos causadores de doenças ao homem, agindo de fato como vetores no campo por mostrar capacidade vetorial (CLEMENTS, 2012).

Os mosquitos habitam quase todos os continentes do planeta, exceto a Antártida, no entanto o número de espécies é maior em regiões de clima tropical, sendo o número de espécies inversamente proporcional à latitude do planeta. A grande dispersão dos mosquitos com capacidade vetorial está intimamente ligada à dispersão de doenças causadas pelos patógenos transmitidos por eles. Exemplos de doenças causadas por esses patógenos são Dengue, Chikungunya, Zika, febre amarela, malária e filariose linfática. Essas são doenças mais comuns,

entretanto os mosquitos também estão envolvidos na transmissão de diversos outros patógenos menos conhecidos (BENELLI; MEHLHORN, 2016).

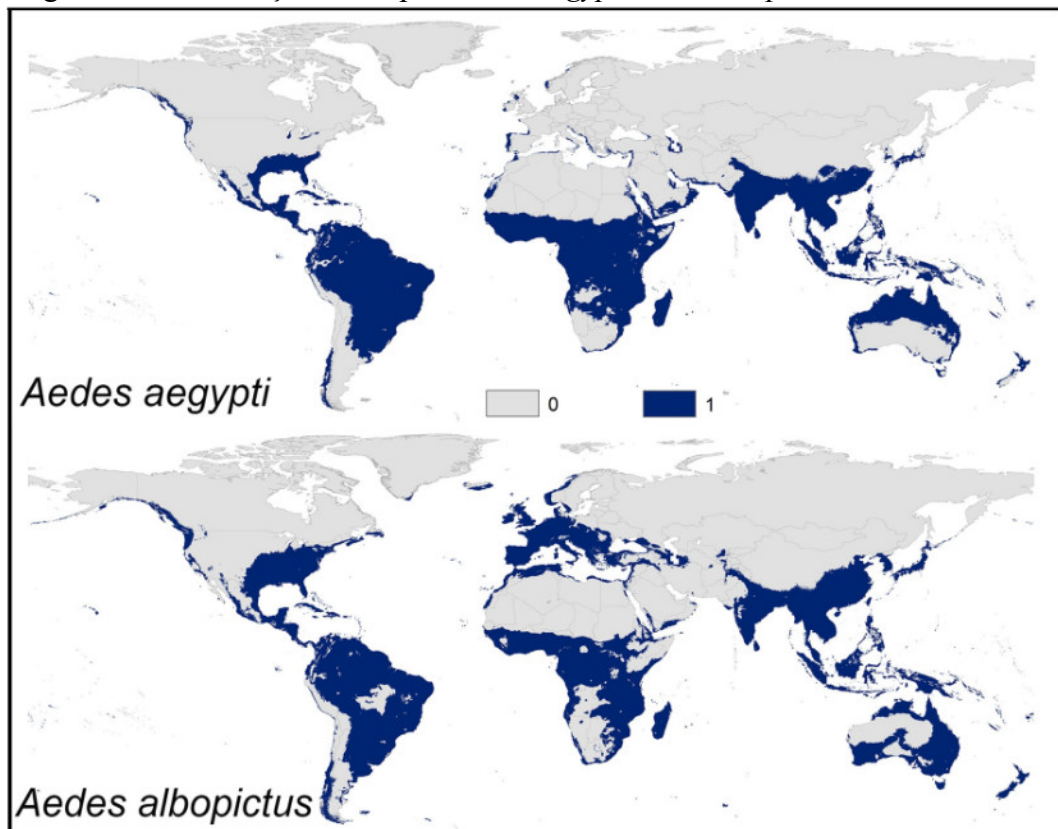
A Família Culicidae está dividida em duas grandes subfamílias. A maior e com maior diversidade de espécies é a subfamília Culicinae, que é dividida em 11 tribos, as quais por sua vez são divididas em vários gêneros (diferentes esquemas de classificação divergem em relação à quantidade de gêneros existentes). Dentro dessa subfamília está uma das espécies de mosquito mais conhecidas, em grande parte por sua incriminação na transmissão de diversos patógenos, o *Ae. aegypti*. Junto com outra espécie do mesmo gênero, o *Ae. albopictus*, são capazes de transmitir diversos patógenos (FOSTER; WALKER, 2019; HARBACH, 2013).

Até poucos anos atrás se acreditava amplamente que o *Ae. aegypti* tivesse se originado em regiões distribuídas pela África subsaariana, onde a forma ancestral chamada *Ae. aegypti formosus* (considerada uma subespécie) ainda poderia ser encontrada (POWELL; TABACHNICK, 2013). No entanto, com o aumento dos dados de sequenciamento de populações dessa espécie ficou claro que *Ae. aegypti* provavelmente originou-se nas ilhas do sudoeste do Oceano Índico, nas proximidades da ilha de Madagascar, mas adaptou-se a convivência doméstica e se dispersou através das regiões tropicais do globo, chegando às Américas através de embarcações (SOGHIGIAN *et al.*, 2020). Já o *Ae. albopictus* surgiu no extremo oriente asiático, no entanto nos últimos 40 anos ele se espalhou rapidamente por todas as massas territoriais, exceto a Antártida, invadindo o Brasil no ano de 1986, logo se espalhando por diversos estados (BONIZZONI *et al.*, 2013). A dispersão de *Ae. albopictus* por diversas regiões temperadas do globo foi facilitada pela expressão de genes de diapausa, que permitem que o mosquito sobreviva às estações mais frias do ano. Um estudo já relatou que essa característica foi fortemente influenciada pela presença de elementos de transposição próximos aos genes relacionados à diapausa (GOUBERT *et al.*, 2017). Ambas as espécies estão se dispersando cada vez mais no globo causando um aumento nos surtos de arboviroses (KAMAL *et al.*, 2018). A distribuição dessas espécies no globo pode ser vista na Figura 6.

Outra espécie da subfamília Culicinae e que é bastante conhecida é o *Cx. quinquefasciatus*. Ela faz parte de um gênero que contém 769 espécies conhecidas e pertence ao complexo de espécies *Cx. pipiens*. As espécies desse complexo possuem hábitos de alimentação generalista, alimentando-se de sangue de uma gama de mamíferos. No entanto, esse hábito diverge dentro das espécies do complexo. Enquanto *Cx. pipiens*, que não é distribuído pelo Brasil, possui hábitos ornitofílicos, o *Cx. quinquefasciatus*, que está presente em nosso país, possui hábitos de predominância antropofílica, ocasionalmente alimentando-se

de sangue de aves ou outros mamíferos. Sua presença também está bastante relacionada a atividade humana, visto que se desenvolve em dentro ou ao redor de ambientes domésticos, em água limpa ou mais comumente em suja, rica em materiais orgânicos derivados das mais diversas fontes resultantes de atividades humanas (WILKERSON; LINTON; STRICKMAN, 2021).

Figura 6 – Distribuição das espécies *Ae. aegypti* e *Ae. albopictus*

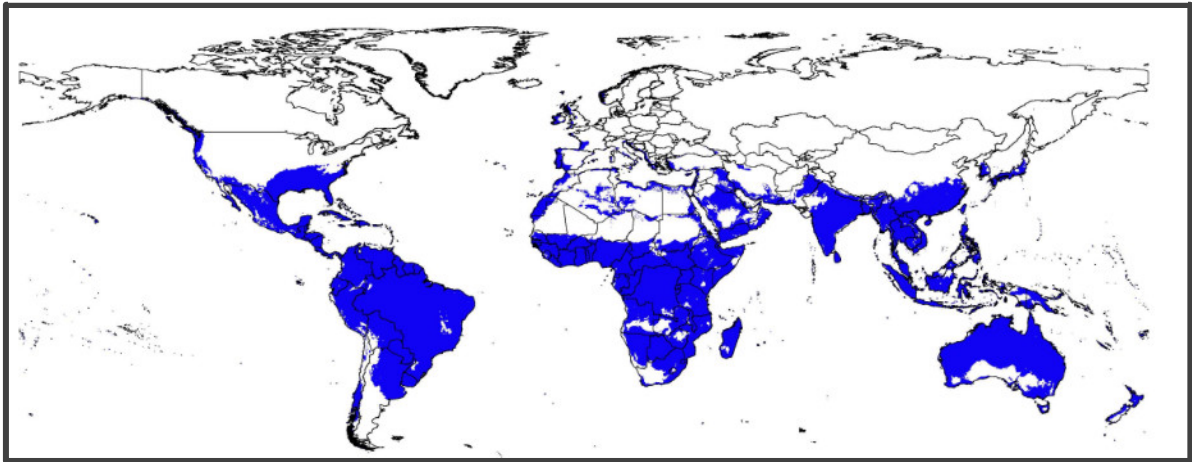


Fonte: Kamal *et al.* (2018).

Legenda: A região marcada em azul corresponde a área de distribuição das espécies.

No hemisfério Sul, esta espécie, que provoca bastante incômodo por suas picadas, geralmente se alimenta de sangue humano com o crepúsculo vespertino. Por esse hábito hematofágico ele é capaz de se infectar e transmitir a *Wuchereria bancrofti* no Brasil, sendo o vetor da filariose linfática. Em outras áreas das Américas, o *Cx. quinquefasciatus* também transmite uma série de outros patógenos, como o arbovírus causador da encefalite Saint Louis, e outros flavivírus que também causam encefalites, além da Febre do Nilo Ocidental (BRASIL, 2011). Sua distribuição potencial abrange áreas tropicais e subtropicais como pode ser visto na Figura 7 (SAMY *et al.*, 2016).

Figura 7 – Predição da distribuição de *Culex quinquefasciatus* no globo



**Fonte:** Samy et al. (2016)

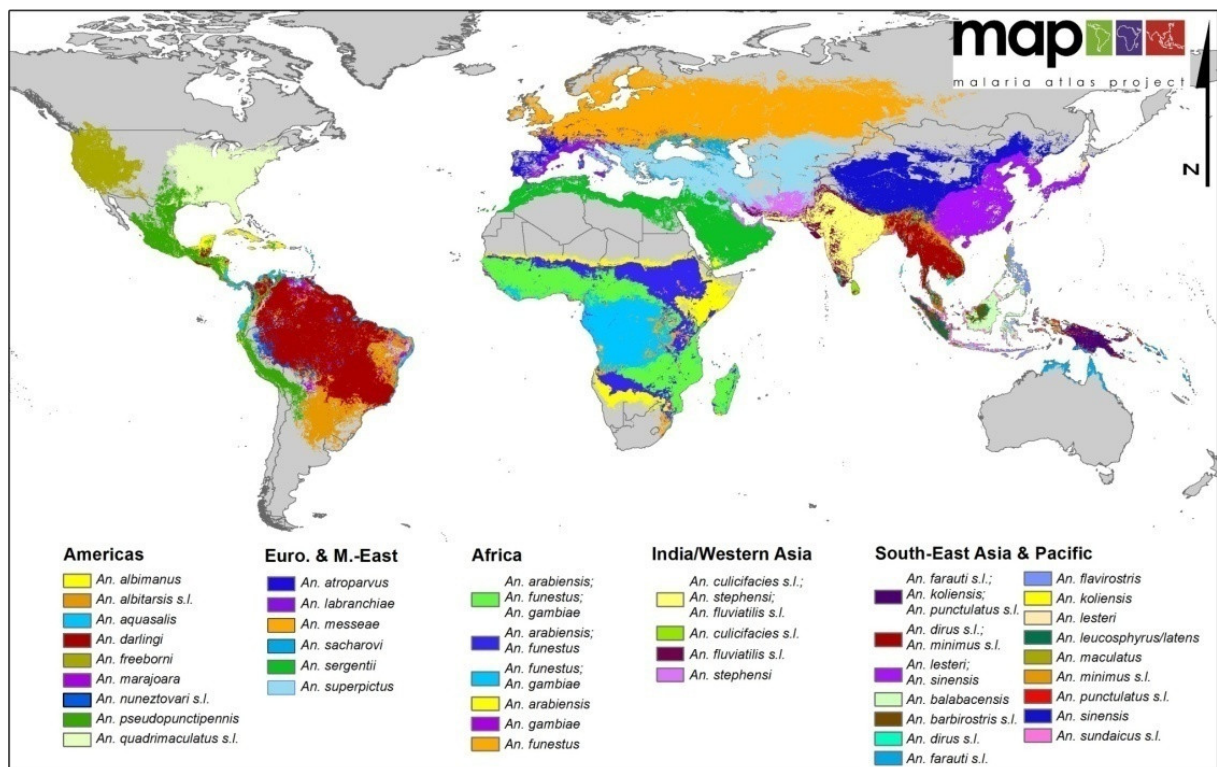
**Legenda:** A região marcada em azul corresponde a área de distribuição da espécie.

A outra subfamília da família Culicidae, é a subfamília Anophelinae. Ela possui três gêneros e é considerada o grupo primitivo dos culicídeos, tendo divergido dos Culicinae há cerca de 195 milhões de anos (FOSTER; WALKER, 2019; TANG *et al.*, 2018). O maior dos gêneros dessa subfamília é o *Anopheles*. Com mais de 465 espécies, esse grupo engloba diversos hospedeiros intermediários de patógenos, sendo os mais comuns, parasitas do gênero *Plasmodium*, causadores da malária. Cerca de 70 espécies de *Anopheles* são capazes de transmitir os patógenos causadores da malária aos seres humanos, com destaque para 41 espécies que se apresentam como vetores dominantes. Entre elas há destaque para o complexo *An. gambiae*, que é formado por oito espécies geneticamente distintas, sendo o *An. gambiae* o mais antropofílico e mais estudado de todos.

Os mosquitos desse gênero estão dispersos em todo o globo, atingindo até mesmo áreas temperadas e bastante frias como partes da Rússia, Finlândia, Suécia e partes do Canadá. Por ter ampla distribuição geográfica (Figura 8) e por transmitir a parasitose de maior impacto na Terra, a Malária, muita atenção foi dada a mosquitos do gênero *Anopheles* (SINKA, 2013).



Figura 8 – Predição da distribuição de 34 espécies de *Anopheles* transmissores de *Plasmodium* spp.



Fonte: Sinka (2013)

## 2.5 BANCO DE DADOS GENÔMICOS E OS GENOMAS DOS CULICÍDEOS

O principal banco de dados para montagens de genomas é o NCBI Assembly (<https://www.ncbi.nlm.nih.gov/assembly>). Ele fornece informações sobre a estrutura e outros metadados de genomas montados, relatórios estatísticos das montagens (apresentam métricas como N50, quantidade de contigs, entre outras) e links para os dados de sequência genômica. Dependendo da qualidade do genoma também é disponibilizado ao usuário a predição gênica, assim como suas sequências nucleotídicas e proteicas. As montagens nesse banco podem possuir as mais diversas qualidades, desde montagens extremamente fragmentadas até genomas completos. Estes últimos são representados pela junção de: I) montagem primária, conta com as sequências de todos os cromossomos nucleares da espécie; II) montagem de DNA mitocondrial; III) grupo de unidades de montagem de *locus* alternativo (KITTS *et al.*, 2016). Atualmente (junho de 2022) há por volta de 24 mil montagens de genomas de eucariotos no banco NCBI Assembly, destas 8.819 pertencem a espécies de animais, sendo apenas 2 totalmente completas: *Caenorhabditis elegans* e *Homo sapiens*. Há, no entanto, redundâncias de espécies no quantitativo total de montagens, uma vez que uma espécie pode ter várias

montagens de genoma. Existem 98 montagens de genomas de culicídeos atualmente disponíveis no NCBI, sendo 36 em nível de cromossomo, 45 em nível de *scaffold* e 17 em nível de *contig*. Estas são representativas de 34 espécies (Quadro 3), 8 das quais só tiveram as montagens de seus genomas disponibilizadas nos últimos 5 anos e embora muitas dessas possuam genoma montado em nível de cromossomo, seu score de N50 é muito baixo, indicativo de montagens muito fragmentadas (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2022).

**Quadro 3** – Lista de montagens de genomas representativos, depositados no NCBI, das espécies de mosquito (Continua).

Código de acesso	Nome	Tamanho da montagem	Contig N50 (Nº de bases)	Nível da montagem	Data da submissão
GCF_000005575.2	<i>An. gambiae</i>	265.011.681	85.548	Cromossomo	16/10/2006
GCA_000349145.1	<i>An. dirus</i>	216.307.690	76.300	Scaffold	28/03/2013
GCA_000349105.1	<i>An. epiroticus</i>	223.486.714	88.996	Scaffold	28/03/2013
GCA_000349025.1	<i>An. minimus</i>	201.793.324	123.292	Scaffold	28/03/2013
GCA_000349065.1	<i>An. quadriannulatus</i>	283.828.998	39.415	Scaffold	28/03/2013
GCA_000349165.1	<i>An. christyi</i>	172.658.580	8.461	Scaffold	02/04/2013
GCA_000439205.1	<i>An. nili</i>	98.320.046	2.278	Scaffold	22/07/2013
GCA_000473375.1	<i>An. culicifacies</i>	202.998.806	13.414	Scaffold	30/09/2013
GCA_000211455.3	<i>An. darlingi</i>	136.935.538	115.072	Contig	20/12/2013
GCA_000473445.2	<i>An. farauti</i>	183.103.254	161.110	Scaffold	17/01/2014
GCA_000473525.2	<i>An. melas</i>	224.162.116	11.312	Scaffold	17/01/2014
GCA_000441895.2	<i>An. sinensis</i>	220.777.669	30.137	Scaffold	14/07/2014
GCA_000956275.1	<i>An. koliensis</i>	151.110.088	4.659	Contig	19/03/2015
GCA_000956255.1	<i>An. punctulatus</i>	146.157.495	10.256	Contig	19/03/2015
GCA_002091845.1	<i>An. cracens</i>	326.465.827	9.003	Scaffold	14/04/2017
GCA_002091835.1	<i>An. maculatus</i>	462.688.933	38.030	Scaffold	14/04/2017
GCF_002204515.2	<i>Ae. aegypti</i>	1.278.715.314	11.758.062	Cromossomo	22/06/2017
GCA_002846955.1	<i>An. aquasalis</i> <sup>1</sup>	162.944.031	14.431	Scaffold	21/12/2017
GCA_003951495.1	<i>An. funestus</i>	444.530.063	107.489	Cromossomo	20/12/2018
GCF_006496715.1	<i>Ae. albopictus</i>	2.538.371.206	1.184.735	Scaffold	28/06/2019
GCA_007018985.1	<i>An. bwambae</i> <sup>1</sup>	288.964.812	20.985	Scaffold	16/07/2019
GCA_008817895.1	<i>An. fontenillei</i> <sup>1</sup>	232.626.592	18.769	Scaffold	03/10/2019
GCF_013141755.1	<i>An. stephensi</i>	243.460.342	38.117.870	Cromossomo	09/06/2020
GCF_013758885.1	<i>An. albimanus</i>	172.602.732	25.000.043	Cromossomo	06/08/2020
GCF_015732765.1	<i>Cx. quinquefasciatus</i>	573.214.445	2.875.282	Cromossomo	04/12/2020
GCA_016170015.1	<i>An. longipalpis</i> <sup>1</sup>	347.073.236	27.950	Cromossomo	22/12/2020
GCA_016254315.1	<i>An. parensis</i> <sup>1</sup>	278.978.218	44.720	Cromossomo	22/12/2020
GCA_016170025.1	<i>An. vaneedeni</i> <sup>1</sup>	308.699.844	40.521	Cromossomo	22/12/2020

**Quadro 3** – Lista de montagens de genomas representativos, depositados no NCBI, das espécies de mosquito (Conclusão).

Código de acesso	Nome	Tamanho da montagem	Contig N50 (Nº de bases)	Nível da montagem	Data da submissão
GCF_016801865.1	<i>Cx. pipiens pallens</i> <sup>1</sup>	566.659.442	838.087	Contig	02/02/2021
GCF_016920715.1	<i>An. arabiensis</i>	256.807.969	23.935.993	Cromossomo	10/03/2021
GCF_016920705.1	<i>An. coluzzii</i>	273.408.546	8.269.537	Cromossomo	10/03/2021
GCF_017562075.2	<i>An. merus</i>	294.364.394	2.729.089	Cromossomo	24/03/2021
GCA_018908115.1	<i>An. rivulorum</i> <sup>1</sup>	321.482.242	25.213	Cromossomo	17/06/2021
GCA_914969975.1	<i>An. atroparvus</i>	277.747.802	44.960.226	Contig	28/09/2021

Fonte: Elaborado pelo autor

Legenda: <sup>1</sup> Espécies cujo genoma só se tornou conhecido nos últimos 5 anos.

Do total de montagens, encontram-se disponíveis genomas sequenciados de 30 espécies do gênero *Anopheles*, 21 dos quais foram analisados nesta tese<sup>2</sup> (o restante apenas se tornou disponível para acesso após o início das análises ou então possuíam qualidade muito baixa). Além dos *Anopheles* spp., outros dois gêneros de mosquito possuem espécies com seus genomas sequenciados: o gênero *Aedes*, representado por *Ae. aegypti*, *Ae. albopictus* e o gênero *Culex*, representado por *Cx. quinquefasciatus* e *Cx. pipiens pallens*. Este último não foi investigado neste trabalho, uma vez que seu genoma só se tornou disponível apenas no ano de 2021.

Uma outra fonte de dados de genoma de mosquitos é o banco de dados VectorBase ([www.vectorbase.org](http://www.vectorbase.org)). Esse é um banco de dados específico de informações sobre animais invertebrados vetores de patógenos. Além da sequência genômica, vários outros recursos são disponibilizados, como por exemplo dados de expressão gênica, dados de genômica comparativa, dados de vias metabólicas, entre outros (GIRALDO-CALDERÓN *et al.*, 2015). No geral, o VectorBase possui dados mais acurados e informativos sobre os genomas dos mosquitos que o NCBI Assembly e por isso é mais utilizado pelos estudiosos de vetores. Recentemente o VectorBase teve sua estrutura alterada e passou a fazer parte do EuPathDB, formando assim o VEuPathDB, com novas funcionalidades (GIRALDO-CALDERÓN *et al.*, 2022). Diferentemente do NCBI, o VectorBase possui genomas com processamento adicional, como remoção de sequências redundantes devido aos processos de montagem do genoma. Devido a essas etapas de processamentos adicionais esse banco contém o genoma de apenas 22 espécies de mosquito: *Ae. aegypti*, *Ae. albopictus*, *An. albimanus*, *An. arabiensis*, *An.*

<sup>2</sup> Mais detalhes sobre as espécies analisadas nesta tese estão disponibilizados na seção de Materiais e Métodos. A tabela contendo as versões dos genomas investigados neste trabalho está disponibilizada no Apêndice C.

*atroparvus*, *An. christyi*, *An. coluzzii*, *An. culicifacies*, *An. darlingi*, *An. dirus*, *An. epiroticus*, *An. farauti*, *An. funestus*, *An. gambiae*, *An. maculatus*, *An. melas*, *An. merus*, *An. minimus*, *An. quadriannulatus*, *An. sinensis*, *An. stephensi* e *Cx. quinquefasciatus* (THE VEUPATHDB PROJECT TEAM, 2022).

A diferença do número de espécies sequenciadas entre Anophelinae e Culicinae também se deve em parte à diferença de tamanho do genoma dos indivíduos desses dois grupos. No geral, os genomas dos Culicinae são bem maiores que os dos Anophelinae (por volta de 3 a 8 vezes maiores). O aumento do tamanho do genoma se dá em parte pelo aumento do repertório de genes ligados à captação de feromônios, a receptores gustativos e a genes de resposta imune e resistência a inseticidas (RUZZANTE; REIJNDERS; WATERHOUSE, 2019). Esses dois últimos grupos têm ganhado bastante atenção, principalmente na tentativa de encontrar genes chave na relação parasita/hospedeiro que possam gerar novos métodos de controle da transmissão de doenças por esses vetores. Esse objetivo, juntamente com a diminuição dos custos das tecnologias de sequenciamento de larga escala, permitiu que fossem realizados diversos estudos de transcriptômica através da técnica de RNA-Seq. Hoje, cada vez mais se tornam comuns os estudos de proteômica nessas espécies. Todavia, esses estudos de expressão focam, em sua maioria, os conjuntos de genes codificantes anotados, com um menor número visando o estudo de pequenos RNAs e RNAs longos não codificantes (lncRNAs), e pouquíssimos sobre a fração móvel do genoma (mobiloma), composta por elementos de transposição (RUZZANTE; REIJNDERS; WATERHOUSE, 2019).

O mobiloma é justamente um outro fator importante para a diferença no tamanho do genoma entre Anophelinae e Culicinae. Embora não tão bem caracterizados a nível de superfamília para muitas espécies, o total de TEs em muitos genomas foi estimado. Enquanto em Culicinae o conteúdo de TEs representa mais de um quarto do genoma (ARENSBURGER *et al.*, 2010; NENE *et al.*, 2007), nos Anophelinae a proporção de TEs raramente ultrapassa os 10% de todo o genoma (HOLT *et al.*, 2002; NEAFSEY *et al.*, 2015).

## 2.6 TRANSFERÊNCIA HORIZONTAL DE TRANSPOSONS (HTT)

Os elementos de transposição são geralmente passados através das espécies por intermédio da herança parental. Assim, estima-se que os TEs da maioria das espécies atuais de eucariotos tenham sido herdados de seus ancestrais. No entanto, alguns TEs podem passar de um organismo para outro por um meio que não envolve troca sexual ou herança parental. Esse

método é conhecido como transferência horizontal de transposons (HTT) (WALLAU; VIEIRA; LORETO, 2018). Como apresentado anteriormente, existe uma diversidade enorme de elementos de transposição. Da mesma forma, há também uma diferença na quantidade e propensão funcional para se transferir horizontalmente para outra espécie. Entretanto, todos os tipos de TE podem vir a realizar esse processo (SILVA; LORETO; CLARK, 2004).

Existem algumas características que contribuem na identificação de casos de transferência horizontal. A primeira delas diz respeito à alta similaridade entre duas cópias de TEs em genomas de espécies distintas. Para exemplificar, quando essas espécies são distantes filogeneticamente, assumindo-se que um determinado gene não sofre seleção negativa, não é esperado que cópias deste gene sejam altamente similares em nível de nucleotídeo a outro gene na espécie mais distante. Quando essa alta similaridade é vista entre TEs, que na maioria dos casos não colabora para o *fitness* do hospedeiro e evolui de forma neutra, temos um indicativo de transferência horizontal. Esse é o pré-requisito mais comum de detecção de HTTs (SILVA; LORETO; CLARK, 2004).

A segunda característica é de fácil visualização, a HTT pode ser detectada através de incongruências entre a filogenia das espécies construída a partir de genes nucleares e uma filogenia construída a partir de TEs. Quando as árvores resultantes possuem topologias distintas, é possível se indagar sobre um caso de transferência horizontal. Essa visualização pode ser feita com o uso de ferramentas como o tanglegram, implementado em muitos programas como o Dencroscope (HUSON; SCORNAVACCA, 2012) e Dendextend (GALILI, 2015). Essas ferramentas permitem verificar com facilidade diferenças existentes na topologia real da filogenia das espécies (árvores bem suportadas e construídas a partir de um conjunto de genes) em relação a uma topologia de árvore construída a partir das sequências de TEs.

A terceira característica se dá através da percepção de um comportamento chamado de “*patchy distribution*” de um TE. Neste caso, um elemento está presente em uma espécie, mas ausente em outras espécies muito próximas. Isso pode ser um indicativo que ele não foi herdado verticalmente, ou então que ele foi perdido nessas espécies irmãs. Entretanto, uma vez que muitos TEs não são estritamente essenciais para a sobrevivência de uma espécie, eles podem ser naturalmente perdidos durante o tempo evolutivo, enfraquecendo o poder de discriminação dessa característica (SILVA; LORETO; CLARK, 2004).

Quando um TE possui todas essas três características (a presença de apenas algumas delas não dá poder de discriminação): I) similaridade maior que a esperada entre as cópias de duas

espécies; II) incongruência filogenética; III) distribuição totalmente desigual entre o maior número de espécies próximas possíveis (do mesmo gênero) ou entre linhagens da mesma espécie, é muito provável que ele tenha sido transferido horizontalmente entre as espécies. Entretanto, é muito difícil avaliar a presença dessas características em espécies muito próximas. Alguns métodos foram desenvolvidos para avaliar quantitativamente todos os casos de transferência horizontal, até mesmo entre espécies bastante próximas. Um método bastante robusto é implementado pelo software VHICA. Ele compara o sinal de transferência vertical que é mostrado por genes ortólogos herdados verticalmente e o sinal gerado por um elemento de transposição entre um par de espécies. Esses sinais são derivados de medidas como o desvio do uso de códons dos genes e sua dS (taxa de substituição sinônima), podendo quantificar com precisão a significância de um evento de HTT (WALLAU *et al.*, 2016).

Para que um elemento de transposição seja transferido horizontalmente e fixado numa espécie não é necessária apenas a troca de material genético entre duas espécies, o elemento também deve conseguir se integrar ao genoma. Na maioria dos animais, essa troca deve ocorrer especificamente na linhagem celular germinativa da nova espécie, de outra forma a aquisição do novo TE seria eliminada da população assim que o indivíduo receptor da transferência fosse eliminado da população. Para que a troca de DNA aconteça, também é necessário contato íntimo entre duas espécies, assim como ocorre em casos de transferência horizontal de genes não ligados a TEs. Esse contato pode ocorrer devido a algumas relações ecológicas como a simbiose, a infecções virais de vírus generalistas, entre outras (WALLAU; VIEIRA; LORETO, 2018).

É possível que haja também relação da interação parasita-hospedeiro com eventos de HTTs. Recentemente foi reportado um potencial caso de transferência de um retrotransposon não LTR pertencente à superfamília RTE, denominado AviRTE, entre vermes filariais e algumas espécies de aves. Esse TE está presente em cópias autônomas em sete clados de pássaros e dois clados de nematódeos. Evidências como *patch distribution*, centenas de cópias nos genomas, alta similaridade entre os consensos das cópias desse TEs presentes nos genomas reforçam a ideia da ocorrência de transferência horizontal. Espécies de mosquitos e insetos do gênero *Chrysops* (conhecidos como mutucas) foram levantados como os potenciais vetores dessas transferências (SUH *et al.*, 2016).

## 2.7 ELEMENTOS DE TRANSPOSIÇÃO, TRANSCRIÇÃO E EFEITOS NA EXPRESSÃO DE GENES DO HOSPEDEIRO

Além da capacidade de se replicar dentro de um genoma e de ser capaz de se mover de uma espécie para outra, os elementos de transposição podem causar diversos impactos em seus hospedeiros. Na maioria das vezes esses impactos geram consequências negativas, já em outros casos, os TEs podem contribuir trazendo uma vantagem adaptativa ao hospedeiro e podem influenciar a evolução das espécies. Assim como qualquer outro gene, os elementos de transposição devem transcrever seus genes em RNAs mensageiros. Estes são então traduzidos em proteínas essenciais para a transposição do TE.

### 2.7.1 Mecanismos do hospedeiro para regulação da expressão de TEs

Ao expressar proteínas íntegras, os TEs são capazes de mover-se para outras regiões do genoma, o que pode levar a um dano (derivado da mutagênese causada pela inserção do TE em uma nova região) no espécime hospedeiro. Essa inserção, juntamente com rearranjos cromossômicos causados pelos TEs, pode gerar várias consequências negativas para um organismo, como exemplo, diversos tipos de câncer em seres humanos (BELANCIO; ROY-ENGEL; DEININGER, 2010; LEVIN; MORAN, 2011; MARASCA *et al.*, 2020). Devido ao alto potencial deletério da transposição para os espécimes (se a transposição ocorrer em tecidos somáticos) ou para a espécie (se a transposição ocorrer em células germinativas, e assim ser transmitida aos descendentes), os organismos desenvolveram uma série de mecanismos moleculares de defesa contra os TEs. Esses mecanismos agem tanto em nível de regulação transcricional como pós-transcricional (CHUONG; ELDE; FESCHOTTE, 2017).

Uma das formas de prevenir a expressão de elementos de transposição é aumentar a compactação do DNA nas regiões em que ocorrem, formando heterocromatina, e consequentemente dificultando a ação da maquinaria de transcrição. Essa compactação muitas vezes é um fenômeno que está ligado a metilação das histonas. Além disso, nessas regiões, o DNA geralmente também se encontra metilado. Isso se torna claro quando mutações nos genes das metiltransferases de *A. thaliana*, onde muitos TEs encontram-se metilados, são responsáveis por permitir a transposição de alguns TEs (ITO; KAKUTANI, 2014; LIPPMAN *et al.*, 2004). Hipóteses indicam que é possível que a metilação de DNA tenha, inclusive, surgido como uma forma de defesa contra transposição (CHUONG; ELDE; FESCHOTTE, 2017).

Um dos mecanismos de regulação de TEs mais bem conhecidos é o controle pós-transcricional de elementos de transposição através de uma classe de pequenos RNAs não codificantes, os PIWI-*interacting* RNAs (piRNAs). A inativação desses piRNAs levam ao descontrole e ativação dos TEs. Alterações nos genes que são envolvidos nas vias desses piRNAs também levam ao descontrole dos TEs. Esses RNAs, cujo tamanho geralmente varia entre 26 e 32 nt, são carregados dentro das proteínas PIWI, que são proteínas da família das Argonautas, proteínas ligadas à clivagem de sequências de mRNA. Dessa forma os piRNAs formam o complexo piRISC, e silenciam os TEs através do pareamento de bases e clivagem dos mRNAs dos elementos de transposição. Em vertebrados as proteínas PIWI são expressas apenas nas células germinativas. A perda dessas proteínas, por exemplo, leva a célula à apoptose no peixe zebra, onde há piRNAs com similaridade apenas a TEs de Classe I. Esse alinhamento em uma única classe de TEs não é uma constante para os piRNAs já que em *Xenopus* existe uma grande proporção desses RNAs mapeando em TEs de Classe II. Embora haja menos piRNAs alinhando em TEs em mamíferos, já foi mostrado que essa via também tem algum papel no silenciamento de TEs (HOUWING *et al.*, 2007; SAITO; SIOMI, 2010). Sendo mostrado que a via do piRNAs tem papel importante no controle do TE *L1* em células germinativas de seres humanos (FRIEDLI; TRONO, 2015).

O mecanismo de controle via piRNAs é particularmente bem conhecido em *Drosophila melanogaster*, onde foi descrito o mecanismo de “*ping-pong*”. A ação dessa via é essencial para proteção dos gametas e da fertilidade (SIOMI; MIYOSHI; SIOMI, 2010), mas além disso os piRNAs também realizam a regulação de TEs em células somáticas. Nesta espécie, cerca de 50% dos piRNAs possuem similaridade com regiões de cópias de TEs. Entretanto, isso não é uma realidade para todos os organismos. Em *Ae. aegypti*, também da ordem Diptera, apenas 19% de todos os piRNAs possuem alguma similaridade a transposons. A maior parte desses RNAs alinham com vírus e genes do hospedeiro. O que pode indicar diferenças entre o sistema de piRNAs de mosquitos e das moscas do gênero *Drosophila*. De fato, em *Ae. aegypti* o pareamento de outro tipo de pequenos RNAs, os *short interfering RNAs* (siRNAs) são mais comuns em TEs do que o pareamento de piRNA em TEs (ARENSBURGER *et al.*, 2011b).

Há hipóteses que algumas regiões do genoma (regiões de até 200kb, conhecidas como clusters de piRNA) funcionem como uma armadilha de TEs, se um TE se inserir nessas ilhas, fica imobilizado na região. Caso esse evento ocorra em células germinativas, haverá transmissão dessa nova estrutura para as próximas gerações. Então, esse TE acabará funcionando como fonte de novos piRNAs para silenciar outros TEs. Segundo o modelo, se um TE for muito ativo ele terá



mais chance de se inserir nestes clusters, gerando, dessa forma, mais piRNAs que vão atuar regulando o próprio TE (HIRANO; SIOMI, 2015; SAITO; SIOMI, 2010).

### 2.7.2 Elementos de transposição que escapam do controle transcricional

O mecanismo de silenciamento de TEs não é 100% eficaz, pois alguns TEs conseguem expressar RNA e proteínas e realizar sua transposição. Geralmente os TEs são mais expressos em tecidos embrionários, onde vias metabólicas alternativas são ativadas, reduzindo o controle sobre a expressão desses elementos. No entanto, alguns elementos também conseguem expressar seus produtos proteicos e se transpor em tecidos somáticos. Em humanos, onde os TEs são amplamente estudados, sabe-se que há expressão de ERV1 em diversos tecidos. Um estudo realizado em seres humanos indicou que cerca de 2% dos TEs são expressos. Mesmo em um número tão baixo, essa fração possui representantes de cerca de 95% de todas as subfamílias de TEs presentes no genoma humano. Nesta espécie também foi observado que muitos LINEs e SINEs são expressos e possivelmente têm papel regulatório sobre diversos genes, além disso evidenciou-se uma relação entre a idade de um indivíduo e o número de TEs expressos (BOGU *et al.*, 2019).

A expressão de TEs parece ocorrer nos mais diversos organismos de diversos reinos. No milho, onde cerca de 85% do genoma é composto por TEs, diversos elementos são expressos: LINEs, MITEs, TRIMs, vários elementos das superfamílias *Copia*, *CACTA* e principalmente *Gypsy* (VICIENT, 2010). Em culicídeos, alguns estudos identificaram TEs expressos em algumas espécies dos gêneros *Anopheles* e *Aedes*. Em *Ae. fluviatilis*, algumas ORFs de elementos de transposição (codificando para transposase e transcriptase reversa) foram encontradas diferencialmente expressas entre mosquitos saudáveis e infectados com bactérias do gênero *Wolbachia* (CARAGATA *et al.*, 2017). Cerca de 15 TEs também foram encontrados expressos no sialotranscriptoma (transcriptoma da glândula salivar) de *An. sinensis* (FANG; YAN; CHEN, 2018). Entretanto, poucos estudos focam exclusivamente na investigação do mobiloma expresso. Um estudo com esse escopo revelou resultado semelhante ao visto em humanos, a maior parte das superfamílias de TEs presentes em *An. funestus*, possui ao menos uma sequência expressa, sendo a grande maioria (85%) pertencente a TEs de Classe I, principalmente retrotransposons com LTRs (FERNÁNDEZ-MEDINA *et al.*, 2017).

Além dos RNA mensageiros tradicionais da expressão de um TE como: transposases, transcriptases reversas e outras proteínas específicas para cada superfamília; e dos pequenos

RNAs, os TEs também estão envolvidos na transcrição de RNAs longos não codificantes (lncRNAs), compondo parte do transcrito. Em humanos, cerca de 83% dos RNAs longos intergênicos não codificantes possuem sobreposição com TEs e cerca de 42% são derivados de elementos de transposição. Os lncRNAs possuem papel na regulação do processamento celular, e lncRNAs ligados a TEs possuem uma maior capacidade de diversificação que os que não estão ligados (KELLEY; RINN, 2012). Além disso, estudos em vertebrados mostram que inserções de TEs podem estar envolvidas com a criação de novos lncRNAs, bem como estabilizam a estrutura secundária dos lncRNAs, possivelmente contribuindo para a obtenção de sua correta conformação. Dessa forma, os TEs parecem ser capazes de influenciar o transcriptoma não codificante de um organismo (KAPUSTA *et al.*, 2013).

Alguns autores argumentam ainda que os TEs são diretamente responsáveis pelos domínios de atividade dos éxons dos lncRNAs. Essa contribuição pode ser dada através do fornecimento de sítios de ligação a fatores de transcrição, fatores de reparo do DNA, proteínas de ligação ao RNA, RNA Polimerase II, todos originalmente presentes na estrutura básica do TE. Outra possibilidade é a ação por complementaridade de bases com regiões específicas do genoma onde estão as outras cópias do mesmo TE que está inserido no lncRNA (JOHNSON; GUIGÓ, 2014). Essa complementaridade entre as sequências de lncRNA e o DNA genômico pode resultar na formação de estruturas chamadas de R-loops, onde se forma um híbrido DNA-RNA, ficando uma das fitas do DNA sem anelamento, ou na formação de triplex RNA-DNA. Ambas as estruturas podem atuar na regulação da expressão gênica (MARCHESE; RAIMONDI; HUARTE, 2017).

### 2.7.3 Impacto de TEs no organismo hospedeiro

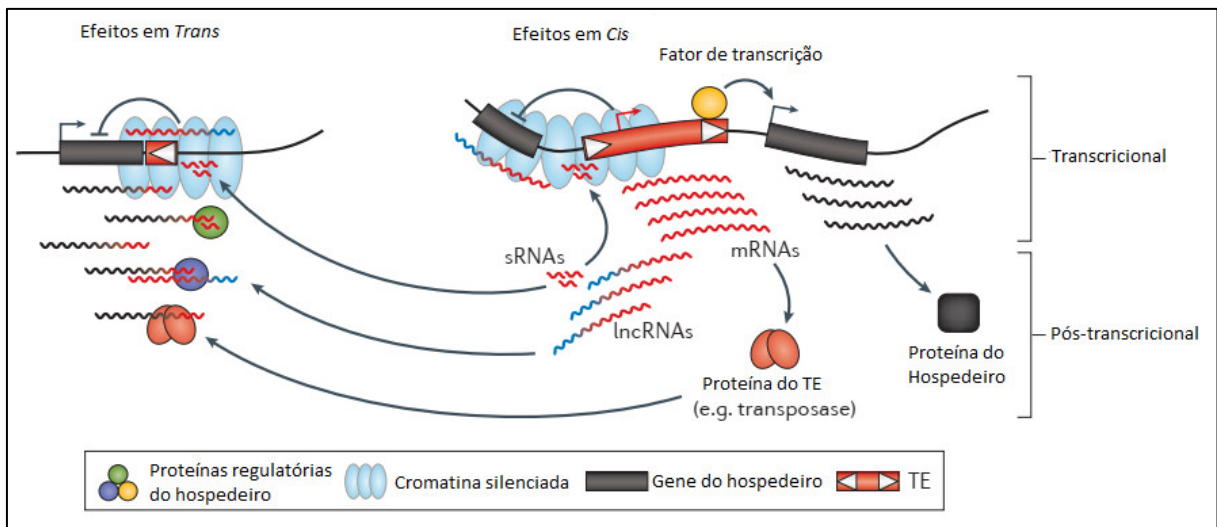
Os TEs podem ser expressos fusionados a mRNAs codificantes da espécie. Em alguns casos, parte de éxons de genes do hospedeiro são englobados por uma família de TE, essa nova estrutura se multiplica e se transforma em uma nova família, como o transposon *CpG-SVA* que possui 76 cópias no genoma humano (BANTYSH; BUZDIN, 2009). Em outros casos é o éxon do TE que é capturado pelo gene do hospedeiro, criando um novo gene, como ocorreu com o gene SETMAR, quando o gene SET se fundiu com a transposase do TE *Hsmar1* em primatas antropoides (CORDAUX *et al.*, 2006). O transcriptoma de um organismo pode conter diversos transcritos fusionados a elementos de transposição. Por exemplo, em *A. thaliana*, cerca de 7,8% dos transcritos (2373 genes) mostram alguma região com alta similaridade com elementos de

transposição que variam entre Helitrons, TIRs, LINEs, SINEs, e LTRs (LOCKTON; GAUT, 2009). Entretanto, em outras espécies o número de transcritos fusionados pode ser muito baixo, apenas 63 transcritos de *D. melanogaster* parecem estar fusionados com TEs, sendo observado que as fusões variam entre diferentes linhagens dessa mosca. A frequência de ocorrência de TE quiméricos mais baixa que o esperado para *D. melanogaster* evidencia o potencial deletério de algumas inserções de TEs, que por causar efeitos negativos para a *fitness* de um espécime, estão presentes em apenas algumas linhagens, visto que a alteração de um produto gênico pode causar a morte de um organismo (LIPATOV *et al.*, 2005).

Ao formar um transcrito quimérico, os transposons podem afetar diretamente o fenótipo de um organismo. Esses transcritos podem ser originados de diversas formas pelos TEs. Eles podem se inserir diretamente na ORF de um gene, alterando a estrutura final da proteína expressa ou gerando uma proteína truncada pela inserção de um ou mais códons de parada. Em outras situações os TEs podem se inserir em um íntron do gene, podendo causar variação nos mecanismos de *splicing* desse gene ou criar um novo éxon para o mesmo. Além disso, essa inserção em íntrons também pode alterar o fim da transcrição, onde o TE inserido pode passar a se comportar como uma nova cauda poli(A) causando o encurtamento do mRNA. Alternativamente, ao se inserir a montante (*upstream*) de um gene, o TE também pode causar alterações no promotor do mesmo. Como o TE precisa recrutar fatores de transcrição para expressão de suas proteínas, quando próximo a um gene, ele pode doar seus sítios criando vários novos sítios de ligação para fatores de transcrição. Além dessa alteração na região promotora de genes, os TEs podem até mesmo criar um promotor alternativo derivado de sua sequência (COWLEY; OAKLEY, 2013).

Além dessas implicações, os TEs também podem impactar a expressão gênica ao agir como acentuadores, isoladores ou repressores da expressão. Esses efeitos descritos, são conhecidos como efeitos em *cis*, porém eles não são os únicos que um TE pode causar. Há uma outra classe, denominada de efeitos em *trans*, onde os produtos de um TE (RNAs e proteínas) agem impactando um gene que está bem distante dele (Figura 9). Por exemplo, a transposase de um TE ativo pode, além de atuar na sua própria transposição, atuar transpondo genes fusionados a TEs, onde o sítio de reconhecimento da transposase ainda se encontra íntegro. Evidências concretas dos efeitos de TEs na expressão de genes no hospedeiro podem ser verificados através de diversas técnicas, por exemplo com o uso de genes repórter e *knockout* gênico. Essas técnicas, entretanto, não se aplicam em análises de larga escala onde há uma grande quantidade de TEs a ser testada (CHUONG; ELDE; FESCHOTTE, 2017).

Figura 9 – Tipos de influência exercidas pelos TEs nos genes do hospedeiro.



**Fonte:** Adaptado de Chuong, Elde e Feschotte (2017).

**Legenda:** O painel da direita indica potenciais efeitos em *cis* dos TEs, como inserção na região promotora e alteração do controle de expressão de um gene, silenciamento do gene pela formação de heterocromatina gerada com a finalidade de silenciar o TE. Já o painel à esquerda mostra os potenciais efeitos em *trans* dos TEs sobre os genes, como por exemplo realizar controle pós-transcricional de genes fusionados a TEs através de pequenos RNAs, ou transpor regiões de genes que estão fusionados a regiões terminais de um TE, uma vez que estas podem ser reconhecidas por proteínas de TEs produzidas por cópias íntegras em *trans*.

Há casos descritos sobre o impacto de TEs sobre alguns genes em insetos, principalmente em organismos bem estudados como as moscas do gênero *Drosophila*. A inserção do transposon *Accord* da ordem LTR *upstream* no gene *Cyp6g1*, um gene citocromo P450, em *D. melanogaster* está relacionada com o aumento da resistência ao inseticida Dicloro-Difenil-Tricloroetano (DDT) (CHUNG *et al.*, 2007). Em outro caso, um transposon de DNA da família *Tc1*, inserido na região intergênica dos genes *Jheh*, está associado ao aumento da expressão desses genes e ao aumento da resistência ao estresse oxidativo em populações de *D. melanogaster* (GUIO; BARRÓN; GONZÁLEZ, 2014). A influência de TEs nos genes também pode ser vista na mariposa, onde a inserção de um grande TE no íntron do gene *cortex* gerou um aumento do melanismo, auxiliando na adaptação ao ambiente urbano das mariposas que possuíam essa inserção (HOF *et al.*, 2016). Em mosquitos *Ae. albopictus*, a presença de TEs influencia a expressão de genes de diapausa, o que trouxe vantagens adaptativas à espécie, facilitando a invasão de ambientes temperados (GOUBERT *et al.*, 2017).

Por diversos efeitos no hospedeiro, as inserções de TEs podem afetar diretamente as redes regulatórias e de transcrição de um organismo. A inserção do mesmo TE próximo a diversos genes pode ocasionar a ativação em conjunto desses genes, já que o TE irá requerer os mesmos fatores de transcrição. Um exemplo desse impacto é a capacidade de engravidar que os mamíferos apresentam. Essa adaptação é provavelmente originária de efeitos da transcrição de

TEs. Isso se deu uma vez que o início da via de sinalização responsável por formar a decídua está relacionado ao hormônio progesterona, que apresenta moléculas sinalizadoras cujos sítios de ligação se originaram a partir de inserções de TEs, que dessa forma foram capazes de alterar a rede gênica regulatória pré-existente (COWLEY; OAKEY, 2013).

#### **2.7.4 Expressão de elementos de transposição sob diferentes condições**

Assim como os genes nucleares que compõem o organismo, os elementos de transposição também sofrem variações de expressão sob diferentes condições a que um organismo pode ser exposto. Uma das mais clássicas condições reportadas na literatura é o estresse. Num cenário de estresse, é possível que haja risco de morte para o organismo, logo a expressão de TEs poderia gerar variabilidade genética, potencialmente trazendo algum benefício para a espécie. Em vários casos já reportados, uma condição de estresse é capaz de elevar a expressão de elementos de transposição. Em muitos deles, a elevação da expressão de alguns TEs também gera elevação na expressão de genes em regiões flangeadoras desses TEs. O estresse, no entanto, pode ser gerado por uma série de fatores, como calor, pouco alimento, baixa oxigenação e exposição à radiação (HORVÁTH; MERENCIANO; GONZÁLEZ, 2017).

Em *Drosophila melanogaster*, a exposição a altas doses de radiação é capaz de elevar a expressão de diversas famílias de TEs. Da mesma forma, outros tipos de estresse como estresse oxidativo (exposição ao formaldeído e dioxina) também elevam as taxas de expressão dos elementos de transposição, apesar disso, no geral, a fração de TEs com diferença de expressão é bem menor que o total de TEs existentes nessa espécie (OLIVEIRA *et al.*, 2021). A exposição ao calor também parece favorecer a superexpressão de elementos de transposição, em 7 de 10 ecótipos de *A. thaliana* expostos a crescimento sob temperatura de 38° C houve aumento considerável da expressão de TEs (BARAH *et al.*, 2013).

Um outro estímulo que é capaz de alterar as taxas de expressão de elementos de transposição de um organismo é a sua infecção por vírus. Seis dias após a infecção das moscas *D. melanogaster* e *D. simulans* com o vírus SINV (Sindbis vírus), a expressão de diversas famílias de TEs se torna reduzida no corpo, ao mesmo tempo que a expressão de pequenos RNAs derivados de elementos de transposição aumenta. Isso ocorre uma vez que a resposta antiviral por meio de pequenos RNAs também atua silenciando a expressão de elementos de transposição nessas espécies. Embora uma infecção viral possa afetar a expressão de elementos de transposição em tecidos somáticos, a expressão de TEs não é afetada em tecido germinativo

(ovários) de *D. melanogaster* e *D. simulans*, o que provavelmente é resultado de uma já presente repressão da expressão de TEs nos ovários (ROY *et al.*, 2020). Assim como os ovários de moscas do gênero *Drosophila*, diversos tecidos em diferentes organismos vertebrados também apresentam um padrão diferenciado de expressão de elementos de transposição (PASQUESI *et al.*, 2020).

### 2.7.5 Domesticação ou cooptação de TEs

Os TEs podem ser parasitas genômicos, podem ser fontes de sítios de ligação a fatores de transcrição, podem influenciar a variedade de lncRNAs, assim como podem agir em *cis* e em *trans* de diversas formas sobre as regiões codificantes do organismo. Mas além disso, em nível proteico, os TEs também podem contribuir para o *fitness* das espécies hospedeiras. Quando há uma adaptação favorável à espécie, o TE ou suas proteínas podem ser cooptados (domesticados) para uma função diferente da original (JANGAM; FESCHOTTE; BETRÁN, 2017). Em camundongos, por exemplo, vários TEs foram domesticados e atualmente funcionam como acentuadores da expressão gênica tecido específica (TODD *et al.*, 2019).

Diversos elementos de transposição podem ser domesticados para funções diferentes em diferentes organismos. Em humanos, por exemplo, há sete proteínas que são similares a Cryptons e seis delas codificam o domínio DUF3504, que não tem função conhecida, mas está bastante disperso entre os gnatostomados (vertebrados com mandíbula). Esse domínio derivado de um *Crypton* ancestral, de fato não constitui um pseudogene, uma vez que há indícios da produção de produtos funcionais (KOJIMA; JURKA, 2011). Outro exemplo de elemento de transposon domesticado (talvez o mais famoso caso de domesticação de TEs) também em gnatostomados são elementos da superfamília *Transib* que foram, há cerca de 500 milhões de anos, cooptados para novas funções, gerando a proteína RAG1 e RAG2, que são essenciais para a recombinação V(D)J, responsável pelo sistema imune adaptativo dos vertebrados (AGRAWAL; EASTMAN; SCHATZ, 1998; BROECKER; MOELLING, 2019; KAPITONOV; JURKA, 2005; KAPITONOV; KOONIN, 2015).

Em humanos, pelo menos cinco genes são derivados de elementos *PiggyBac*, entretanto, estudos mostram que cada uma dessas sequências derivadas de *PiggyBac* foi originada por um mesmo *PiggyBac* ancestral domesticado (BOUALLÈGUE *et al.*, 2017). Embora esses eventos de domesticação de TEs não sejam de comum ocorrência, eles podem acontecer diversas vezes durante a evolução de algumas espécies. Em duas espécies de rãs do gênero *Xenopus*, pelo

menos 10 genes codificadores de transposases de TEs da família Kolobok foram domesticados e perderam sua capacidade de transposição (HIKOSAKA; KONISHI, 2018).

Até a domesticação, uma linhagem de TE pode passar por seis estágios. Inicialmente o TE está ativo no genoma, gerando muitas cópias, o que pode levar a consequências deletérias para o espécime. Devido a isso, este monta respostas para silenciar o TE. Num segundo momento, este TE passa a influenciar o fenótipo do organismo. Como agora seu produto influencia o espécime, o TE passa a sofrer uma pressão de seleção, e em períodos de mudança ambiental, o fenótipo alterado pelo TE pode gerar vantagens adaptativas ocasionando uma seleção negativa no TE. O próximo passo é a imobilização desse TE numa região do genoma, por exemplo por mutação nas suas extremidades, e aumento da sua estabilidade de expressão. Após a estabilização, esse TE é fixado na população, agindo mais comumente como elementos moduladores da expressão, que como proteínas efetoras. Por último, mudanças no ambiente da espécie podem eliminar as pressões de seleção sobre o gene derivado do TE, a partir desse momento ele está apto a passar por uma seleção positiva, onde há diversificação de sua sequência e consequentes funções (JOLY-LOPEZ; BUREAU, 2018).

### **2.7.6 Formas de estudo do mobiloma expresso**

Talvez a principal forma de estudo de transcritos contendo sequências de TEs seja o sequenciamento em massa do transcriptoma do organismo por RNA-Seq, havendo outras formas de análise como o CAGE-Seq (*Cap Analysis of Gene Expression*), que permite mapear as regiões promotoras (CHUONG; ELDE; FESCHOTTE, 2017).

Depois de sequenciado, o transcriptoma total pode ser montado por métodos *de novo*, onde não se utiliza o genoma como referência. Os milhares de transcritos resultantes podem ser mapeados contra bases de dados de TEs para verificação da similaridade, ao mesmo tempo em que é realizada uma busca por domínios conservados de TEs (FERNÁNDEZ-MEDINA *et al.*, 2017). Entretanto, essa abordagem necessita de muito recurso computacional e bastante tempo, ambos variando proporcionalmente com o número de *reads* de RNA-Seq sequenciados. Além disso, o uso de dados de RNA-Seq de diversas linhagens de uma espécie pode resultar em um alto número de transcritos quiméricos após a montagem *de novo*. Uma abordagem utilizando o mapeamento de *reads* contra o consenso dos TEs é bem menos viável, uma vez que as cópias individuais podem desviar bastante dos consensos, ocasionando seu não mapeamento. Outra dificuldade é quantificar a expressão de uma família de TEs, uma vez que ela possui múltiplas

cópias. Uma abordagem alternativa e comumente utilizada é aplicada no Tetranscripts, um pipeline desenhado para o estudo da expressão de elementos de transposição. O software utiliza um arquivo de anotação das cópias de TEs caracterizadas na espécie juntamente com o genoma como fonte de mapeamento de *reads* derivadas de RNA-Seq. Ele é capaz de lidar com *reads* que mapeiam em mais um local e aplica um algoritmo para o cálculo do número de *reads* mapeados por família de TE, além disso ele também suporta análises de expressão diferencial de TEs (JIN *et al.*, 2015).

Esse software permite catalogar quais famílias de elementos de transposição estão expressas no organismo, mas não permite verificar se há expressão de transcritos ligados a genes do hospedeiro, uma vez que ele processa apenas as cópias de TEs. Uma abordagem complementar utilizada para estudar o mobiloma expresso que potencialmente causa efeitos em *cis* na expressão de genes do hospedeiro é verificar se genes e TEs se coexpressam em diversas amostras e condições. Essa abordagem, implementada no software TEffectR (KARAKÜLAH *et al.*, 2019) permite realizar uma análise de regressão linear entre a expressão de TEs na região 5' o 3' de um determinado gene em diversas amostras.

## 2.8 PROTEÔMICA E ELEMENTOS DE TRANSPOSIÇÃO

Mesmo com todos os mecanismos regulatórios que os organismos possuem para frear a expressão dos elementos de transposição, alguns elementos ainda conseguem evadir esse controle e produzir proteínas funcionais. A atividade de TEs pode estar relacionada com alguns tipos de câncer em seres humanos, não se sabendo se ela seria uma das causas ou uma das consequências dessa doença. A metodologia da proteômica pode auxiliar na detecção da expressão de um determinado TE com algum fenótipo de um espécime, também pode verificar se o nível de expressão desse TE está relacionado com a característica fenotípica de um espécime (RUGGLES *et al.*, 2017).

### 2.8.1 Proteômica clássica e proteômica de *shotgun*

O termo proteômica refere-se ao estudo do proteoma, que é o conjunto de todas as proteínas expressas pelos genes de um organismo num determinado tempo. A primeira etapa numa análise proteômica é a obtenção da mistura proteica de um indivíduo ou tecido. Após extraídas, as proteínas devem ser purificadas e posteriormente clivadas com uma protease,



geralmente tripsina, dependendo da metodologia usada (POMASTOWSKI; BUSZEWSKI, 2014). Esse conjunto de proteínas ou peptídeos pode ser então submetido a diferentes métodos que podem ser mais ou menos adequados dependendo do objetivo do projeto. No entanto, os métodos de identificação de um proteoma não avançaram com a mesma velocidade que os métodos de estudos de genomas e transcriptomas. Dessa forma, muitas das proteínas preditas sobre o genoma de uma espécie, com a utilização de ferramentas de bioinformática, não tem quaisquer evidências experimentais. Ao contrário do que ocorre com as isoformas de transcritos de RNA, que são confirmadas por sequenciamento de cDNA ou por RNA-Seq (O'NEILL, 2019).

No campo da proteômica experimental, umas das técnicas clássicas mais usadas para avaliar o proteoma de um indivíduo e suas variações qualitativas e quantitativas é a técnica de eletroforese bidimensional (eletroforese 2-D), que vem sendo utilizada desde a década de 1970. Nessa técnica, o conjunto de proteínas purificadas é levado para focalização isoeletrica, uma etapa conhecida como a primeira dimensão da eletroforese. Após as proteínas atingirem um pH onde sua carga elétrica se anula, é realizada uma reação para equilibrar as proteínas e se segue para a segunda dimensão da eletroforese, uma eletroforese SDS-PAGE. Ao fim do experimento as proteínas estão dispersas em duas dimensões no gel de eletroforese, a primeira é o pH onde a proteína atinge o ponto isoeletrico e a segunda é peso molecular da proteína (POMASTOWSKI; BUSZEWSKI, 2014).

A eletroforese bidimensional é uma técnica excelente para separar milhares de proteínas numa mistura. Entretanto, é bastante direcionada para avaliar proteínas que estão diferencialmente expressas em alguma condição. Um software de análise de imagens, compara os géis resultantes de diversas condições experimentais. Os *spots* (pontos no gel que representam uma determinada proteína) podem então ser excisados do gel, tripsinizados e levados para a análise por espectrometria de massas. Geralmente a análise é realizada num aparelho do tipo MALDI-TOF, onde o espectro da proteína é capturado. Esse espectro é então comparado com um banco de dados de espectros e a proteína é identificada (MAGDELDIN *et al.*, 2014; POMASTOWSKI; BUSZEWSKI, 2014). Todo esse *workflow* torna difícil a análise de milhares de proteínas de um indivíduo. Análises mais abrangentes vêm sendo realizadas a partir de uma metodologia de *shotgun*.

Métodos de *shotgun* são métodos de análises de proteínas em larga escala. Neles, as proteínas extraídas de uma amostra são submetidas a uma clivagem proteolítica gerando uma mistura de peptídeos. Essa mistura é levada geralmente a uma cromatografia líquida de alta

performance (HPLC) acoplado a um espectrômetro de massas que utiliza o *electrospray*, como fonte ionizadora, acoplado a dois analisadores de massas (ESI-MS/MS). O *electrospray* (ESI) ioniza os peptídeos e lança aqueles que adquirem carga no primeiro analisador de carga (MS1), nele os íons (fragmentos de peptídeos eletricamente carregados) são separados pela sua razão massa-carga ( $m/z$ ). Os íons que passam por um determinado filtro de  $m/z$  são então fragmentados, por exemplo por colisão com um gás, e analisados no segundo analisador de massas (MS2). Esse procedimento resulta em espectros, a partir dos quais podemos identificar a sequência de um peptídeo, seja a partir de métodos *de novo*, ou mais comumente através de buscas em um banco de dados (AEBERSOLD; MANN, 2003; O'NEILL, 2019).

Um banco de dados para uma busca de proteínas deve conter todas as possíveis proteínas contidas na mistura inicial que foi levada ao espectrômetro. Por esse motivo, geralmente é utilizado o proteoma predito de uma espécie, que se encontra depositado em bancos como o NCBI ou UniProt. Existem diversas ferramentas de busca para espectros, algumas delas são ferramentas comerciais e amplamente utilizadas como o MASCOT (PERKINS *et al.*, 1999), outras são ferramentas públicas como X!Tandem (CRAIG; BEAVIS, 2004), OMSSA (GEER *et al.*, 2004), Andromeda (COX *et al.*, 2011), Comet (versão acadêmica do software comercial SEQUEST®) (ENG; JAHAN; HOOPMANN, 2013), MS-GF + (KIM; PEVZNER, 2014) e MSFragger (KONG *et al.*, 2017). Essas ferramentas constroem espectros teóricos baseados no banco de dados passado, e compara esses espectros contra os espectros experimentais. Uma proteína é identificada quando há um *peptide-spectrum match* (PSM) que é mantido após o cálculo do score baseado em um método estatístico que varia entre as ferramentas. Uma opção para aumentar o número de proteínas capturadas na etapa de análise de dados é o uso desses diversos softwares no mesmo *dataset*. Essa abordagem foi implementada no SearchGui, um software que permite automatizar a análise contra uma base de dados utilizando muitos dos diferentes motores de busca citados anteriormente (BARSNES; VAUDEL, 2018).

Embora o uso de diversos algoritmos de busca possa aumentar o número de proteínas identificadas, deve-se levar em consideração que a grande maioria deles usa a mesma abordagem de busca: uma abordagem *Target-Decoy*. De acordo com essa estratégia, sequências inexistentes de proteínas (*decoys*) são adicionadas ao banco de dados. Esses *decoys* podem ser construídos por diversas metodologias, que impactam o resultado final da análise. O objetivo é estimar os scores dos matches na base *decoy* e com isso estimar a taxa de falsos positivos da busca (*false discovery rate* - FDR), consequentemente eliminando os matches com baixo suporte (ELIAS; GYGI, 2007, 2010).

A escolha do banco de dados é crucial para a análise de proteínas de acordo com a metodologia de busca em bases. Uma base com menos de 1000 sequências proteicas pode levar a uma estimativa errônea da FDR, outro fator de erro é restringir a base a apenas alguns componentes, excluindo proteínas que provavelmente estão na mistura (KNUDSEN; CHALKLEY, 2011). As bases escolhidas podem ser derivadas: I) do proteoma de referência da espécie; II) de todas as proteínas conhecidas para um determinado nível taxonômico que abrange a espécie; III) de todo SwissProt; IV) de bancos customizados. Cada banco possivelmente resultará em um resultado distinto. Bancos muito grandes geram um número de *decoys* muito grande, isso aumentará o número de *matches* aleatórios nos *decoys*. O alto número de *matches* irá causar impacto na análise via FDR, uma vez que o número de *matches* com a base de proteínas reais não aumentará na mesma proporção. O resultado disso é que a aplicação de um *threshold* para a FDR irá descartar muitos casos de proteínas identificadas corretamente, uma vez que estarão presentes *matches* aleatórios com *score* mais alto (HEDGES *et al.*, 2015).

Os resultados de estudos proteômicos, assim como os dados dos espectros de massa geralmente são depositados em bancos de dados proteômicos, com objetivo de livre acesso e reuso por parte da comunidade científica. Esses no geral seguem as recomendações do *ProteomeXchange Consortium* (DEUTSCH *et al.*, 2020), cujos membros são: PRIDE (EMBL-EBI, Cambridge, Reino Unido), PeptideAtlas (ISB, Seattle, WA, EUA), MassIVE (UCSD, San Diego, CA, EUA), jPOST (várias instituições, Japão) , iProx (National Center for Protein Sciences, Pequim, China) e Panorama Public (Universidade de Washington, Seattle, WA, EUA). Apesar da grande quantidade de membros, em alguns desses repositórios, os dados de espectrometria não estão muito bem estruturados e em boa parte deles não há muitos experimentos de espécies não modelos, como é o caso dos mosquitos. O mais completo e bem estruturado dos banco de dados de espectrometria de massa é o PRIDE (*Proteomics Identifications Database* - [www.ebi.ac.uk/pride/](http://www.ebi.ac.uk/pride/)) (PEREZ-RIVEROL *et al.*, 2019), e é nele que se encontram a maior parte dos dados experimentais proteômicos de mosquitos.

### 2.8.2 Proteogenômica

Embora a abordagem proteômica de *shotgun* LC-MS/MS seguida pela busca no proteoma de referência da espécie seja útil em grande parte dos casos, algumas análises requerem algo um pouco mais complexo. Um desses campos de análise é a proteogenômica, cujo resumo pode ser visto na Figura 10.

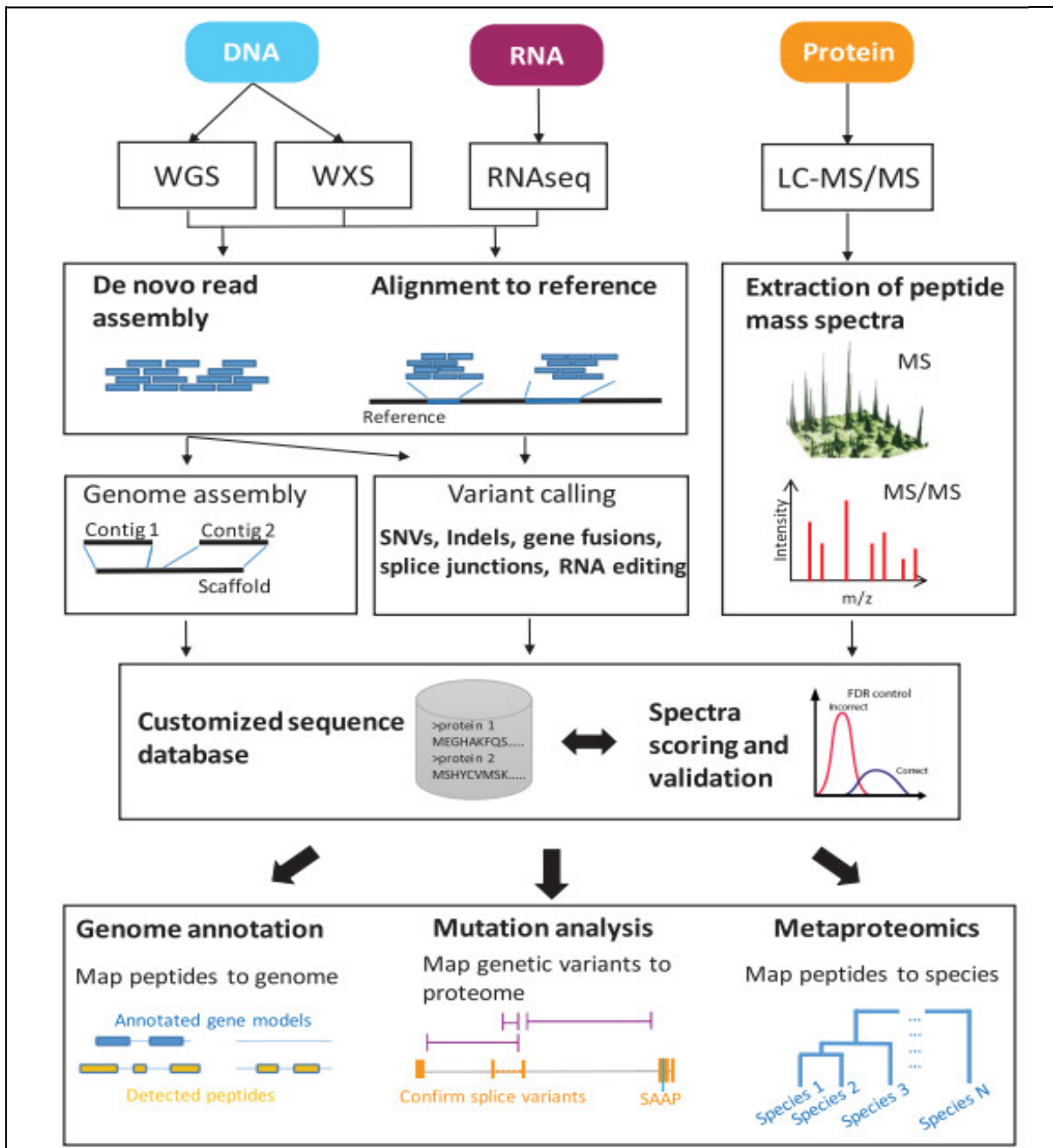
Dentro dessa área é possível integrar dados genômicos e proteômicos mais diretamente. Um dos usos mais comuns dessa técnica é a anotação de áreas do genoma que ainda não são bem estudadas e modeladas como regiões codificantes. Nesse sentido, o maior número de regiões codificantes (com o mínimo de regiões irrelevantes) é utilizado como base para ferramentas de busca de proteínas. Verifica-se então se houve *match* com alguma região ainda não modelada como proteína, auxiliando assim na anotação do genoma (RUGGLES *et al.*, 2017).

Visto isso, uma área que está em ascensão é o estudo da expressão da parte do genoma de uma espécie que é formada por inserções virais antigas e elementos de transposição, e quais os efeitos dessa expressão (RUGGLES *et al.*, 2017). Uma análise proteogenômica em *A. thaliana* revelou que uma fração significativa dos espectros de peptídeos dessa espécie representava proteínas produzidas por elementos de transposição, que costumeiramente não são incluídos no proteoma de referência de uma espécie (CASTELLANA *et al.*, 2008). A expressão proteica de TEs também foi mostrada no parasita *Coccidioides posadasii*, onde os autores do estudo identificaram um TE da ordem LTR potencialmente ativo (MITCHELL *et al.*, 2018).

Apesar da ascensão da proteômica nos últimos anos, ela não é tão comum como o sequenciamento de DNA. Desta forma não há tantos dados proteômicos de mosquitos disponíveis em bancos de dados, como há de espécies modelo como *Homo sapiens*, *Mus musculus* e *Drosophila melanogaster* (PEREZ-RIVEROL *et al.*, 2019). Estudos de proteínas de mosquitos focando em análises proteogenômicas são ainda mais raros. Há apenas um estudo com esse enfoque em culicídeos, realizado em na espécie *Anopheles stephensi*, onde foram estudados 15 diferentes tecidos/órgãos de mosquitos adultos e os estágios de larva e pupa (PRASAD *et al.*, 2017).

A desvantagem dos métodos de proteogenômica é o alto número de possíveis peptídeos no banco de dados, já que muitas vezes é realizada a predição de ORFs nos seis quadros abertos de leitura (ORFs) do genoma. Esse aumento no espaço de busca leva ao aumento do tempo das análises, além de poder reduzir o número de peptídeos identificados, levando à necessidade de métodos mais laboriosos para a análise de dados (LI *et al.*, 2016).

Figura 10: Visão geral da proteogenômica.



Fonte: Ruggles *et al.* (2017)

**Legenda:** Tecnologia de larga escala de sequenciamento de DNA pode ser utilizada para montagens *de novo* ou alinhamento a uma referência. Os dados gerados podem ser utilizados para criação de um banco de dados próprio que será utilizado para realizar o mapeamento dos espectros obtidos por meio de LC-MS/MS. Os resultados da análise podem ser utilizados para auxiliar a anotação do genoma em regiões anteriormente não anotadas, identificar mutações específicas bem como novas variantes de splicing e também pode ser utilizado no campo da metaproteômica.

### 2.8.3 Proteômica informada pela transcriptômica - PIT

Embora estudos de proteogenômica utilizando as ORFs possíveis de um organismo, preditas a partir de sua sequência genômica, tenham permitido a identificação de TEs expressos em algumas espécies, o que não era comum ao utilizar bases do proteoma padrão, alguns autores argumentam que muitos dados sobre TE ainda podem ser perdidos durante a análise. Essa perda deve-se à presença de mutações que podem causar diferenças significativas entre as ORFs preditas de um genoma de referência de uma espécie e as ORFs reais existentes na linhagem que será utilizada para extração de proteínas. Essa diferença de sequências de DNA entre duas linhagens é mais comum em regiões não codificantes do genoma assim como nos elementos de transposição, e é causada pela baixa pressão seletiva nestas regiões, que permite substituições nucleotídicas sem grandes danos ao hospedeiro (JOLY-LOPEZ; BUREAU, 2018; MARINGER *et al.*, 2017).

Uma vez que sequenciar e montar o genoma de várias linhagens da mesma espécie muitas vezes não é possível, devido ao alto custo do experimento, alguns pesquisadores iniciaram o uso de RNA-Seq seguido de montagem do transcriptoma da mesma linhagem em que será realizado o estudo proteômico. Isso elimina problemas relacionados à variação das sequências de TEs entre diferentes linhagens de uma mesma espécie. Além disso, o uso do transcriptoma é muito útil para espécies não modelo, que não têm genoma disponível ou em que o genoma montado ainda é muito fragmentado. Este é o principal fato que distancia a PIT da proteogenômica, a falta de necessidade de um genoma bem montado. Outra vantagem da utilização do transcriptoma como base é a possibilidade de criação de um banco de dados tecido específico, uma vez que se deseje estudar o transcriptoma e o proteoma de algum tecido de um determinado organismo (SAHA; MATTHEWS; BESSANT, 2016).

A PIT vem identificando transcritos que inicialmente não foram anotados, ao mesmo tempo em que é capaz de mostrar que RNAs do organismo, que eram considerados não codificantes, são capazes de gerar proteínas funcionais. Isso é uma realidade até mesmo em organismos bem estudados e com genomas bem montados como os dos camundongos (CHOCU *et al.*, 2014). Uma integração da proteogenômica com informações passadas pelo sequenciamento do transcriptoma levou ao aumento do número de genes descobertos na espécie *Anopheles stephensi*, assim como a revisão de 87 transcritos anteriormente classificados como não codificantes, mas cujas proteínas foram detectadas por análise de LC-MS/MS (PRASAD *et al.*, 2017).

Esse método tem se mostrado eficiente para estudos de expressão de elementos de transposição. Davidson, Matthews e Maringer (2017) argumentam que a técnica de PIT abre novas fronteiras no estudo da expressão de transposons. Isso se dá principalmente pelo fato do genoma de referência muitas vezes não refletir o mobiloma da amostra estudada. Outro fator muito importante é que, uma vez que se trata de um método de *high throughput*, o número de proteínas identificadas é bem maior que a proteômica de TEs baseada em gel bidimensional.

#### 2.8.4 Proteômica *de novo*

Um método alternativo, para a busca de similaridade entre os espectros de massa e o DNA codificante do organismo, é o sequenciamento *de novo* de peptídeos. Nesse método, análises manuais ou computacionais são realizadas nos espectros derivados da LC-MS/MS. Essa metodologia não era classicamente utilizada como a busca em bases de dados, pois sofreu bastante com algumas falhas na predição da sequência do peptídeo, o que muitas vezes impedia a obtenção do real peptídeo. Entretanto, atualmente os espectrômetros de massa possuem cada vez mais resolução, principalmente após a democratização dos analisadores de massa do tipo Orbitrap. Seguindo a mesma evolução, os programas de análise *de novo* se tornaram mais robustos, melhorando bastante a eficácia da predição da sequência peptídica. Atualmente, grandes expectativas vêm sendo criadas sobre essa metodologia, como uma futura “Era do Ouro” da proteômica *de novo* (MUTH *et al.*, 2018).

Mais de 25 algoritmos de sequenciamento *de novo* de peptídeos foram publicados nos últimos 25 anos. Uma grande quantidade deles é adequada apenas para finalidades específicas, enquanto outros foram descontinuados. Entretanto, alguns desses softwares geram sequências peptídicas com relativa confiança (MUTH; RENARD, 2018). Um exemplo é o Novor, um algoritmo muito rápido e que usa árvores de decisão (um tipo de metodologia de aprendizagem de máquina) treinadas para predizer cada aminoácido corretamente num peptídeo baseado em seu espectro de MS/MS (MA, 2015).

Um estudo de *benchmarking* realizado com diferentes algoritmos de sequenciamento *de novo* sobre diferentes *datasets* concluiu que esse método ainda não pode ser, sozinho, uma alternativa para os métodos clássicos de análise em bancos de dados. Entretanto, ele pode ser de grande auxílio em estudos proteogenômicos ou de metaproteômica, uma vez que esses métodos dependem de banco de dados bastante extensos, o que causa um grande aumento no

tempo da análise aliado a um potencial descarte de *matches* positivos (KUMAR; YADAV; DASH, 2017; MUTH; RENARD, 2018).

Após o sequenciamento *de novo* dos peptídeos, estes podem ser mapeados no banco de dados através de algoritmos de alinhamento como o BLAST. Esse alinhamento permite a identificação das proteínas que estão expressas num determinado organismo. Alguns softwares como o PepExplorer (LEPREVOST *et al.*, 2014) e o SearchGui (BARSNES; VAUDEL, 2018) são capazes de realizar esse procedimento. O sequenciamento *de novo* adicionado ao mapeamento dos peptídeos em um banco de dados parece ser uma metodologia complementar a proteogenômica/PIT para estudar a expressão de elementos de transposição. Isso pode ocorrer uma vez que esses elementos não possuem uma pressão de seleção tão grande quanto os genes do hospedeiro e podem sofrer mutações com mais facilidade entre diversas linhagens. Conseqüentemente o uso de uma base de dados construída com uma linhagem distinta de um organismo (quando não é possível ter dados de RNA-Seq da mesma linhagem) pode trazer diversas variações de aminoácidos únicos, prejudicando a identificação por algoritmos que realizam os *matches* entre espectros predito e observado.



### 3 OBJETIVOS

#### 3.1 OBJETIVO GERAL

Caracterizar genômica e funcionalmente os elementos de transposição (TEs) presentes em 24 espécies de Culicídeos vetores de patógenos, dos gêneros *Anopheles*, *Aedes* e *Culex* e entender o papel desses elementos na expressão gênica.

#### 3.2 OBJETIVOS ESPECÍFICOS

##### 3.2.1 Abordagem genômica

3.2.1.1 Caracterizar a diversidade de elementos de transposição presentes nos genomas de mosquitos vetores;

3.2.1.2 Avaliar o modo de herança dos elementos de transposição nos genomas das espécies estudadas.

##### 3.2.2 Abordagem transcriptômica

3.2.2.1 Identificar quais TEs são expressos em níveis de RNA a partir de dados públicos de RNA-Seq.

3.2.2.2 Verificar a coexpressão entre genes e TEs localizados na região *upstream* do gene a fim de identificar potenciais ações em *cis* sobre a expressão gênica;

3.2.2.3 Verificar a associação de TEs e RNAs longos não codificantes;

3.2.2.4 Analisar o padrão de expressão diferencial de TEs em mosquitos expostos a diferentes condições;

3.2.2.5 Identificar os transcritos TE potencialmente capazes de gerar proteínas de TE intactas.

##### 3.2.3 Abordagem proteômica

3.2.3.1 Identificar quais TEs estão sendo expressos no nível de proteína usando conjuntos de dados de espectrometria de massas disponíveis ao público (espécies *Ae. aegypti*, *An. gambiae* e *An. stephensi*).

## 4 METODOLOGIA

Esta metodologia está dividida em três grandes partes. A primeira parte diz respeito à caracterização dos elementos de transposição no nível do genoma. A segunda parte aborda métodos utilizados na análise funcional de elementos de transposição em nível de RNA. A última parte foca nos métodos de análise proteômica utilizados na descoberta de TEs que estão expressos em nível de proteína.

### 4.1 CARACTERIZAÇÃO DO MOBILOMA

#### 4.1.1 Definição das espécies e obtenção do genoma

Foram caracterizados os TEs presentes nos genomas das espécies *Anopheles gambiae*, *An. albimanus*, *An. koliensis*, *An. dirus*, *An. atroparvus*, *An. arabiensis*, *An. christyi*, *An. coluzzii*, *An. culicifacies*, *An. darlingi*, *An. epiroticus*, *An. farauti*, *An. funestus*, *An. maculatus*, *An. melas*, *An. merus*, *An. minimus*, *An. punctulatus*, *An. quadriannulatus*, *An. sinensis*, *An. stephensi*, *Culex quinquefasciatus*, *Aedes aegypti* e *Ae. albopictus*. O genoma dessas espécies foi obtido das montagens presentes nos bancos de dados do VectorBase (<https://www.vectorbase.org>) e NCBI (<https://www.ncbi.nlm.nih.gov/assembly>) até dezembro de 2017. Em etapas posteriores do estudo, montagens mais recentes dessas 24 espécies foram utilizadas, à medida em que montagens de melhor qualidade foram sendo disponibilizadas. Dados sobre os genomas das espécies estudadas, como códigos de acesso, cobertura de sequenciamento e origem do espécime sequenciado estão presentes no Apêndice C.

O DNA mitocondrial (mtDNA) foi montado usando as *raw reads* de algumas espécies e a versão 1.9.1 do software Mitobin (HAHN; BACHMANN; CHEVREUX, 2013). Essas sequências, juntamente com os genomas de mtDNA já disponíveis no NCBI, foram usados para construir uma árvore filogenética das espécies estudadas. A árvore foi construída com o método de máxima verossimilhança usando PhyML (GUINDON *et al.*, 2010) com 1000 réplicas de *bootstrap*, de acordo com o modelo GTR + G + I, cuja escolha foi baseada no critério de informação de Akaike, no software SMS.

#### 4.1.2 Identificação dos Elementos de Transposição através de uma metodologia *de novo*

Para a identificação dos TEs, inicialmente foi utilizada uma abordagem *de novo*, para isso foi utilizado o TEdenovo, um *pipeline*<sup>3</sup> integrante do pacote REPET (FLUTRE *et al.*, 2011), juntamente com dados capturados pelo RepeatScout (PRICE; JONES; PEVZNER, 2005). Os elementos repetitivos identificados por essa última ferramenta foram passados para processamento junto ao TEdenovo.

As montagens do genoma de todas as gênero *Anopheles* foram utilizadas como dado de entrada do pipeline. No entanto, devido à alta densidade de TEs e ao grande tamanho do genoma de *Cx. quinquefasciatus*, *Ae. aegypti* e *Ae. albopictus*, foi necessário realizar uma subamostragem aleatória dos contigs/scaffolds desses genomas, de forma que apenas um terço do genoma dessas 3 espécies foi passado como entrada ao TEdenovo. Isso não interfere na identificação da maioria dos TEs, uma vez que eles são bastante distribuídos pelo genoma. Através do uso deste pipeline, realizamos o autoalinhamento do genoma de cada espécie com o BLASTER, para encontrar regiões repetitivas. Uma pesquisa estrutural de TEs da superfamília LTR também foi realizada com o LTRharvest (ELLINGHAUS; KURTZ; WILLHOEFT, 2008) (opção --struct) para aumentar o número de TEs identificados (que seriam perdidos devido ao baixo número de cópias). Uma vez que essa busca por estruturas de LTR, realizada com o LTRharvest, gera uma série de sequências (artefatos) que podem não estar relacionadas a TEs, como duplicações segmentares e regiões de baixa complexidade, foi executado um segundo algoritmo sobre as sequências preditas pelo LTRharvest, a fim de filtrar sequências compostas de mais de 50% de regiões de baixa complexidade. Nas sequências restantes após a aplicação desse primeiro filtro foi executado o LTRfinder (XU; WANG, 2007), para excluir sequências de LTR sem a presença dos sítios PBS e PPT.

Na segunda etapa do pipeline, usamos os algoritmos Recon (BAO; EDDY, 2002), Piler (EDGAR; MYERS, 2005), Grouper (QUESNEVILLE; NOUAUD; ANXOLABÉHÈRE, 2003) e MCL. Em seguida, os clusters foram alinhados pelo MAP (HUANG, 1994). Após a obtenção dos consensos, a ferramenta PASTEC foi usada para classificar os TEs. Por meio dessa ferramenta, executamos blastx e tblastx contra a versão Repbase REPET (RepBase20.05\_REPET) e sequências ribossomais de mosquitos. Também realizamos uma pesquisa em domínios usando Hmmer (EDDY, 2011) contra o banco de dados Pfam. Na última

---

<sup>3</sup> Conjuntos de programas executados em cadeia, compondo diferentes análises.

etapa do pipeline, foi realizada uma etapa de filtragem que eliminou potenciais famílias multigênicas dos mosquitos e regiões de microssatélites.

Tendo em vista a ausência de elementos caracterizados pelo TEdenovo ou RepeatScout em alguns genomas, examinamos adicionalmente as *raw reads* de sequenciamentos genômicos dessas espécies usando o TEdna (ZYTNIKI; AKHUNOV; QUESNEVILLE, 2014) e o dnaPipeTE (GOUBERT *et al.*, 2015). Isso foi realizado para verificar se os TEs poderiam ter sido removidos antes ou durante a etapa de montagem do genoma.

#### 4.1.3 Identificação de elementos transponíveis usando uma abordagem baseada em homologia

Nesse estágio, os TEs foram identificados com base na homologia com os TEs descritos anteriormente nos bancos de dados Repbase e TEFam. Inicialmente, foi executada uma busca utilizando blastn e tblastn nestes dois bancos de dados, de forma individual, contra o genoma de cada espécie. Somente *HSPs*<sup>4</sup> (*High-scoring Segment Pairs*) resultantes com *bit score* acima de 200 foram mantidos, de modo a excluir *matches*<sup>5</sup> aleatórios. Posteriormente, os *HSPs* derivados do blastn e tblastn foram mesclados em um único arquivo para reconstruir as cópias derivadas de *HSPs* que correspondiam a elementos da mesma família, espaçados não mais que 1 kb entre si.

Todas as cópias de elementos de transposição em um determinado genoma foram agrupadas pelo algoritmo CD-HIT-est (LI; GODZIK, 2006) considerando um *cutoff* de 80% de identidade e 80% de cobertura, e utilizando-se de uma estratégia de alinhamento global. As cópias de cada cluster foram então extraídas e alinhadas usando o algoritmo MAP, que reconstruiu o consenso representativo de cada variante estrutural. Finalmente, um arquivo contendo todos os consensos de TE para cada espécie foi gerado para cada um dos bancos de dados utilizados.

#### 4.1.4 Mapeamento de consensos e cálculo da divergência das famílias de TEs

As sequências consenso derivadas do TEdenovo e da busca por homologia foram classificadas usando três programas diferentes: PASTEC (HOEDE *et al.*, 2014), um

---

<sup>4</sup> Um alinhamento local, sem *gaps*, que atinge uma das pontuações de alinhamento mais altas em uma determinada busca.

<sup>5</sup> Um mapeamento perfeito entre duas cadeias de caracteres que são iguais.

componente do pacote REPET; TEsorter, um programa recentemente descrito que classifica os TEs de acordo com seus domínios proteicos conservados (ZHANG *et al.*, 2019); e o RepeatClassifier, um componente do RepeatModeler (FLYNN *et al.*, 2020). Os consensos foram então alinhados no genoma com o RepeatMasker, que decifrou o número de cópias e pares de bases cobertos por cada TE de cada genoma. Também foi extraída a distância Kimura 2 parâmetros (K2P) de cada família TE usando os *scripts* auxiliares do RepeatMasker. Essa distância foi utilizada para estimar a datação intragenômica<sup>6</sup> desses elementos.

Adicionalmente, foi realizada uma análise da correlação entre a fração de transposons e o tamanho da montagem do genoma de cada espécie estudada. Isso foi feito usando a função *cor()* (método = "spearman") no software R. O coeficiente de correlação foi usado para testar a força da correlação.

#### 4.1.5 Análise de Transferência Horizontal

Para identificar possíveis ocorrências de transferência horizontal de TEs entre as espécies estudadas foi utilizado o pacote VHICA, no ambiente R. Este pacote compara a relação entre a regressão linear derivada da taxa de substituição sinônima (dS) e do viés de uso de códons (*codon usage bias* - CUB) de genes ortólogos, herdados verticalmente e de cópia única, das espécies contra a dS e CUB de elementos transponíveis. Uma transferência vertical é o cenário mais provável se o dS-CUB de um TE não for significativamente diferente daquele dos genes hospedeiros. Por outro lado, um desvio significativo nos valores de dS-CUB dos genes hospedeiros indica transferência horizontal.

Para o uso do VHICA foram necessárias, além das sequências das famílias de TEs de cada espécie, 50 sequências de DNA de genes ortólogos de cópia única e presentes em todos os genomas. Estes genes foram identificados na base de dados do OrthoDB (KRIVENTSEVA *et al.*, 2007), que armazena sequências de aminoácidos de genes ortólogos. O ID de cada um desses genes foi utilizado para recuperar a sequência de codificação de nucleotídeos para cada gene no VectorBase.

Após a obtenção das sequências dos TEs pelos métodos *de novo* e baseados em homologia, foi realizada uma clusterização de todos os consensos dos TEs das espécies

---

<sup>6</sup> Datação relativa baseada na divergência de cada cópia de TE em relação ao consenso da família da qual ela deriva.

estudadas. Para isso foi utilizado o algoritmo CD-HIT-EST, com um *cutoff* de 80% de identidade e 80% de cobertura, usando o parâmetro de refinamento (-g 1) e um alinhamento global. Como método de validação, a cópia que mais semelhante a cada consenso (com base no melhor *hit* de pontuação) foi extraída a partir de uma busca por Blastn de cada consenso de TE contra o genoma das suas respectivas espécies. Isso criou um segundo conjunto de *clusters* com a mesma estrutura dos *clusters* de consenso, usando as cópias de TEs em vez do consenso de um TE. Os *clusters* remanescentes na análise foram os seguintes: I) aqueles que possuíam sequências em pelo menos duas espécies; II) aqueles em que todas as sequências tinham pelo menos 600 nucleotídeos; e III) aqueles que possuíam pelo menos uma sequência com ORF que codifica um polipeptídeo com mais de 200 aminoácidos de tamanho.

As sequências de cada *cluster*, e também de cada conjunto de ortólogos, foram alinhadas dentro de seus respectivos conjuntos através de alinhamento de códons realizado pelo programa MACSE (RANWEZ *et al.*, 2011), tomando como referência (parâmetro -seq), a sequência nucleotídica da maior ORF encontrada entre as sequências agrupadas. As demais sequências foram então passadas para um arquivo FASTA separado no parâmetro -seq\_lr. Dessa forma, os alinhamentos dos TEs assim como os alinhamentos dos genes ortólogos foram oferecidos como *input* ao pacote VHICA para a verificação da presença ou ausência de transferências horizontais.

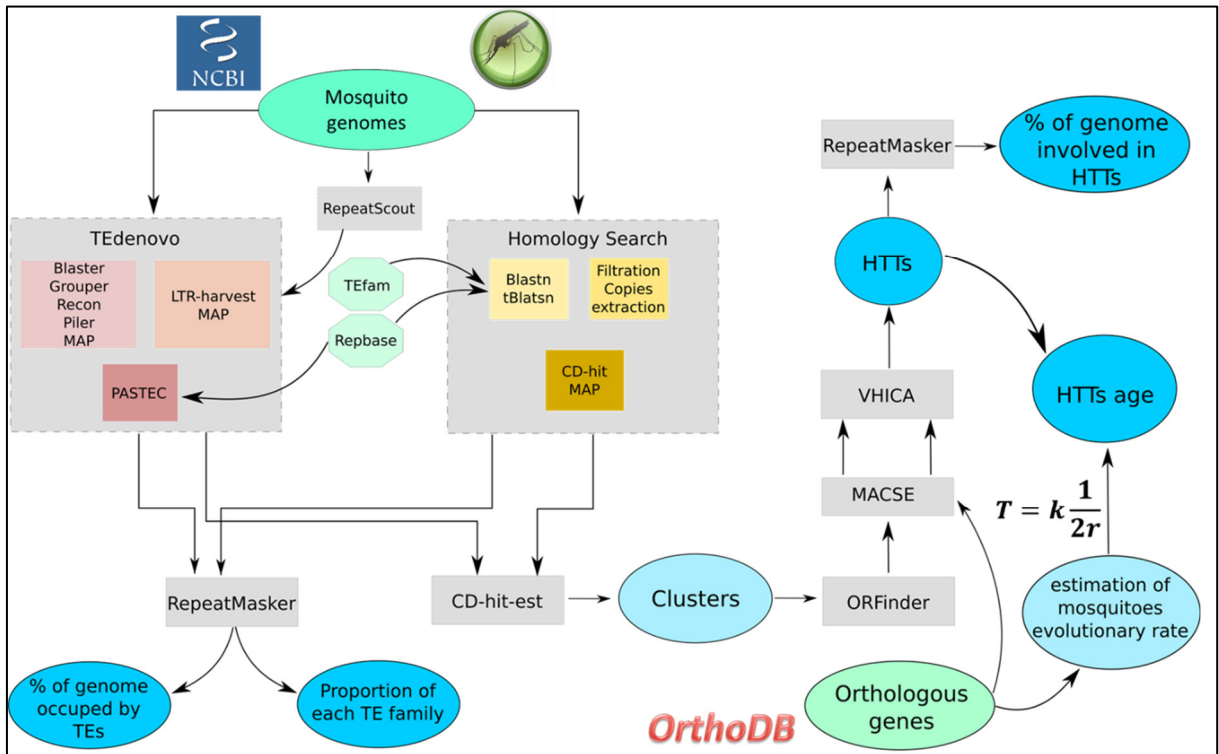
Aqueles *clusters* de TE cujos *p-values* foram menores que 0,01 em um teste estatístico unilateral foram considerados eventos de transferência horizontal. Além disso, a porcentagem do genoma envolvido em eventos de transferência horizontal, bem como o parâmetro K2P (*Kimura two-parameter*) para cada família TE em cada espécie, foi calculada usando o software RepeatMasker e seus scripts auxiliares.

Para investigar se algumas das famílias de TE envolvidas em eventos de transferência horizontal entre as 24 espécies de mosquitos estudadas também poderiam estar envolvidas em transferências horizontais para outras espécies, realizamos um blastn (-task dc-megablast) das sequências desses TEs contra: Banco de dados *nt* do NCBI; todos os genomas de protostomados, plantas, fungos, protistas, platelmintos, vírus, Echinodermata, Hemichordata e Chordata presentes no NCBI até de janeiro de 2019. Aqueles *matches* que tinham um alto grau de identidade no nível de nucleotídeos, com mais de 80% de cobertura da sequência TE do mosquito e aquelas que tinham um número de cópias superior a cinco foram consideradas prováveis HTTs. Quando uma espécie tinha mais de uma montagem de genoma derivada de

amostras diferentes, dispensamos a necessidade de um número de cópias se as correspondências estivessem presentes na maioria dos diferentes conjuntos de genomas dessa espécie.

O resumo da metodologia utilizada na abordagem genômica de estudo dos elementos de transposição está mostrado na Figura 11.

Figura 11 – Procedimento utilizado na caracterização dos TEs e das HTTs ocorridas.



Fonte: Elaborado pelo autor

**Legenda:** Blaster, LTR-harvest, RepeatScout, Grouper, Recon e Piler são programas para identificação de TEs. TEfam e Rebase são banco de dados de TEs. MAP é um algoritmo de alinhamento. PASTEC é um programa de classificação dos consensos de TEs. RepeatMasker é um programa que alinha sequências de TEs contra o genoma das espécies a partir do uso do algoritmo RMBlast. CD-hit-est é um programa utilizado aqui para remoção de redundância e para formar grupos compostos pelos consensos de uma mesma família de TE que está presente em múltiplas espécies. ORFinder é um algoritmo para predição de quadro abertos de leitura e foi utilizado para predição de ORFs nas sequências de TEs. MACSE é um programa de alinhamento de códons e utilizado para alinhar os *clusters* de TEs.

#### 4.1.6 Datando as transferências horizontais

A datação dos eventos HTT foi realizada aplicando a fórmula  $T = k / 2r$  (GRAUR; LI 2000), onde  $T$  é o tempo,  $k$  é a taxa de substituição sinônima (dS) entre cópias de TE de duas espécies, e  $r$  é a taxa evolutiva dos grupos de espécies.

O primeiro passo foi então encontrar a taxa evolutiva dos culicídeos. O cálculo dessa taxa foi realizado utilizando-se da mesma fórmula usada para datar as transferências. Para isso

usamos como tempo T a datação de fósseis de alguns ancestrais de cada grupo: *Anopheles dominicanus* (ancestral da subfamília Anophelinae, usado para calcular a taxa evolutiva de transferências ocorridas dentro do gênero *Anopheles*, idade = 34 milhões de anos) (ZAVORTINK; POINAR, 2000), *Paleoculicis minutus* (ancestral da subfamília Culicinae, idade = 79 milhões de anos, usado para calcular a taxa evolutiva de transferências ocorridas dentro dos Culicinae) (POINAR *et al.*, 2000) e *Burmaculex antiquus* (ancestral da família Culicidae, usado para calcular a taxa evolutiva de transferências que ocorrem entre o gênero *Anopheles* e os gêneros *Aedes* ou *Culex*, idade = 99 milhões de anos) (BORKENT; GRIMALDI, 2006). O valor k foi considerado o valor central entre todos os valores de todas as dS obtidas de cada par de espécies. Essas informações foram extraídas dos arquivos gerados pelo pacote VHICA usando comparações par-a-par entre as espécies. Além disso, calculamos o intervalo interquartil como a taxa de erro das estimativas da taxa de evolução.

Usando os valores de dS para cada táxon de mosquito a partir dos 50 genes ortólogos de cópia única e os dados fósseis, foram identificadas as taxas evolutivas  $r$ :  $17,567 \times 10^{-3}$  mutações por milhão de anos para transferências entre mosquitos da subfamília Anophelinae,  $9,205 \times 10^{-3}$  para transferências entre mosquitos dentro da subfamília Culicinae e  $10,006 \times 10^{-3}$  para transferências entre as subfamílias Anophelinae e Culicinae.

O passo final foi utilizar o valor de dS (calculado pelo VHICA) entre os TEs horizontalmente transferidos entre cada par de espécies e a taxa evolutiva descoberta, para calcular a idade das transferências horizontais em milhões de anos.

#### 4.1.7 Análise da distribuição geográfica de mosquitos vetores e relação com as HTTs

A região de distribuição de cada espécie de mosquito do gênero *Anopheles* foi extraída das distribuições apresentadas por Sinka *et al.* (SINKA, 2013; SINKA *et al.*, 2010a, 2010b, 2011) e a distribuição prevista usando o *Malaria Atlas Project*. Essas regiões foram consideradas habitats ancestrais desses mosquitos, uma vez que a maioria dos mosquitos *Anopheles* não são espécies invasoras e, portanto, se dispersam muito pouco. Consideramos as ilhas do sudoeste do Oceano Índico como o habitat ancestral do *Ae. aegypti* (SOGHIGIAN *et al.*, 2020) e a região leste e sudeste da Ásia que se estende à Índia como o habitat nativo de *Ae. albopictus* (BONIZZONI *et al.*, 2013). Para avaliar se a sobreposição e/ou proximidade da distribuição das espécies de mosquitos tem algum impacto sobre a probabilidade da transferência horizontal, foi realizada uma análise de correlação ponto-bisserial usando a função

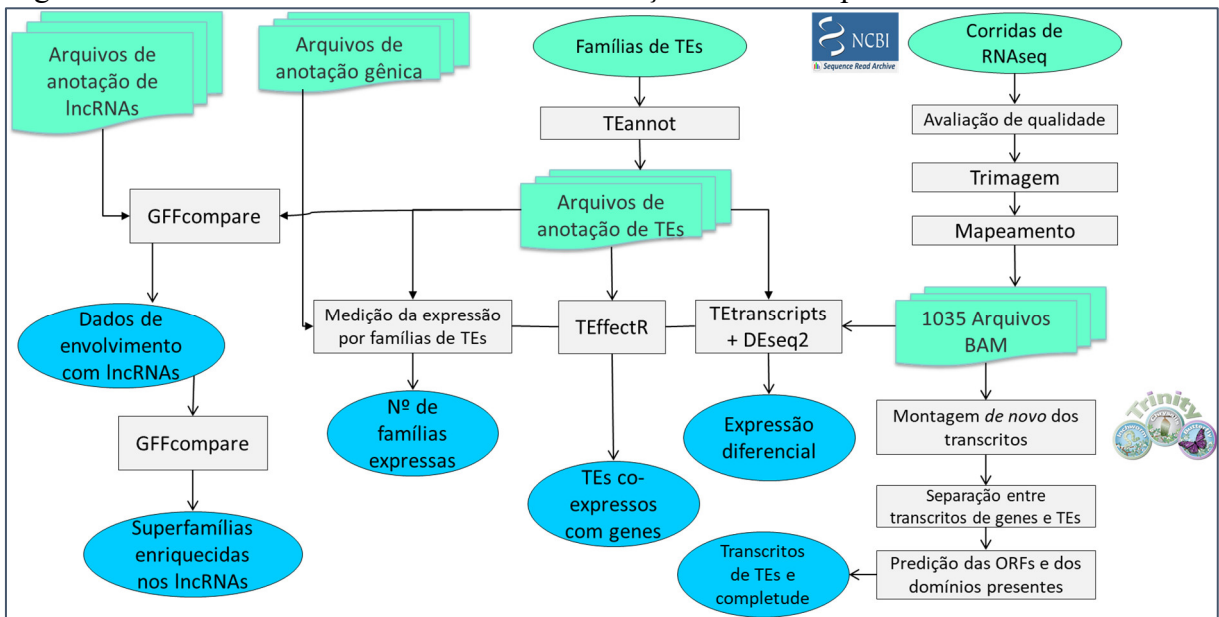


biserial.cor() do software R. Essa função gera um coeficiente de correlação biserial (rpb), que foi usado para medir a relação entre uma variável binária (presença ou ausência de sobreposição geográfica) e uma variável contínua (número de transferências que ocorreram entre duas espécies de mosquitos).

## 4.2 ABORDAGEM TRANSCRIPTÔMICA

Nessa etapa se inicia o afunilamento do trabalho, embora haja dados de genoma para as 24 espécies, o mesmo não ocorre para dados de transcriptoma. Logo, essa etapa foi realizada apenas para 12 espécies: *Ae. aegypti*, *Ae. albopictus*, *Cx. quinquefasciatus*, *An. gambiae*, *An. coluzzii*, *An. arabiensis*, *An. merus*, *An. minimus*, *An. albimanus*, *An. funestus*, *An. dirus* e *An. stephensi*. Outras 12 espécies não possuem dados públicos de RNA-Seq ou os dados presentes não são suficientes para suportar conclusões. O resumo geral da abordagem transcriptômica está mostrado na Figura 12.

Figura 12 – Procedimento utilizado na caracterização dos TEs expressos.



Fonte: Elaborado pelo autor.

**Legenda:** A figura mostra um resumo da metodologia utilizada para avaliação da expressão de TEs em nível transcriptômico. As formas coloridas em verde mostram as fontes de dados utilizadas. Os arquivos de anotação gênica e de IncRNAs, assim como as corridas de RNAseq, foram obtidos de bancos de dados públicos. Já os arquivos BAM e famílias de TEs e suas anotações foram gerados durante a tese. Os quadrados em cinza representam os métodos ou programas de processamento utilizados. Os círculos azuis, por outro lado, representam os resultados finais dos processamentos realizados.

### 4.2.1 Anotação dos genomas

O passo inicial para a verificação da expressão dos elementos de transposição é a criação de uma boa anotação de todas as suas cópias no genoma das espécies. Neste trabalho apenas consensos de TEs caracterizados pela metodologia *de novo* foram usados inicialmente para anotar genomas de mosquitos. O primeiro passo foi remover todas as redundâncias possíveis na biblioteca TE de cada espécie. Primeiro, aplicamos a regra 80-80 para remover a redundância na etapa 6 do TEdenovo e removemos os elementos repetitivos sem quaisquer características de TE e elementos cujo consenso era quimérico. Depois disso, mapeamos cada consenso TE contra o genoma do mosquito usando blastn (BLAST 2.2.30+) usando os seguintes parâmetros: -task rmbblastn, -dust no, -word\_size 11, -soft\_masking false. Apenas o consenso da TE que tinha pelo menos uma correspondência nos genomas foi mantido para a próxima etapa. Posteriormente, usamos o script filter-stage-1.prl do software RepeatScout (PRICE; JONES; PEVZNER, 2005) a fim de remover consensos de TEs formados por mais de 50% de regiões de baixa complexidade (esses consensos são provavelmente TEs malformados) e criar uma biblioteca final de TE que foi usada para anotar genomas de mosquitos.

O pipeline TEannot (FLUTRE *et al.*, 2011) foi usado para criar os arquivos de anotação (formato GFF) de elementos transponíveis nos genomas das 12 espécies de mosquitos. Na etapa 2 do pipeline, apenas algoritmos Censor e BLR foram usados para mapear TEs contra o genoma do mosquito, isso foi seguido pelo uso de um filtro estatístico na etapa 3. Também foi executado um tblastx para anotar possíveis TEs por homologia usando sequências de TEs presentes no banco de dados Repbase (etapa 6 do TEannot). No final, o "procedimento de junção longa" (*long join procedure*) foi realizado e apenas as correspondências no genoma com comprimento superior a 300 nt foram mantidas no arquivo de anotação gff3 (como o uso do blast contra o Repbase poderia resultar em um maior índice de falsos positivos, apenas correspondências maiores que 800 nt foram mantidas, para reduzir a possibilidade de inserção de *matches* aleatórios). Um script personalizado foi usado para remover qualquer possível anotação sobreposta de elementos de transposição (mais de um TE na mesma região) e para criar arquivos GTF (.gft) usados para estimar a expressão TEs no pacote TEtranscripts (JIN *et al.*, 2015).

#### 4.2.2 Caracterização dos Transcritos derivados de TEs

Para avaliar quais dos TEs caracterizados nas etapas anteriores são expressos em nível de RNA, dados públicos de RNA-Seq foram obtidos da base de dados *Sequence Read Archive* (SRA) do NCBI, que armazena dados brutos de sequenciamentos de larga escala. Foram

analisados apenas dados sequenciados na plataforma Illumina<sup>®</sup>, que foram sequenciados a partir de *paired-end reads* e cuja biblioteca tenha sido construída a partir de seleção por cauda poli-A do mRNA ou por depleção do RNA ribossomal (Apêndice D).

Os programas FastQC (ANDREWS, 2010) e MultiQC (EWELS *et al.*, 2016) foram utilizados para avaliar a qualidade das corridas de RNA-Seq antes e depois do procedimento de trimagem. Os arquivos fastq foram então trimados para remoção de adaptadores e *reads* de baixa qualidade com o software Trimmomatic (BOLGER; LOHSE; USADEL, 2014), com os seguintes parâmetros usados para a maioria dos experimentos: LEADING:3, TRAILING:3, AVGQUAL:20, MAXINFO:75:0.5, MINLEN:35. O alinhador STAR (DOBIN *et al.*, 2013), versão 2.7.3a, foi escolhido para mapear leituras de RNA-Seq contra genomas de mosquitos, uma vez que permite facilmente multimapeamento durante o alinhamento. Para cada corrida de RNA-Seq foi mantido um máximo de 100 alinhamentos por *reads* (--outFilterMultimapNmax 100) para permitir a captura de *reads* de diferentes cópias do mesmo consenso TE. O software MultiQC foi então usado para avaliar a qualidade do mapeamento. Qualquer experimento em que mais de 50% das leituras não se alinharam ao genoma de referência foi excluído do estudo.

O programa TEcount (parte do pacote Tetranscripts) foi usado para estimar o número de *reads* que mapeiam para cada gene ou TE. Este programa recebeu os arquivos .bam, um arquivo GTF dos genes dos mosquitos e um arquivo GTF dos TEs dos mosquitos e, como resultado, gerou uma tabela de contagem de *reads* mapeadas para cada estrutura (gene ou TE) em cada experimento. As tabelas de contagem de todos os experimentos da mesma espécie de mosquito foram concatenadas. Genes e TEs com valores de contagem muito baixos (quando a estrutura possuía *reads* mapeando em menos de 10% das corridas de RNA-Seq) em todos os experimentos foram removidos. Este procedimento foi realizado para evitar que o ruído (mapeamento errado das *reads*) superestimasse o número de TEs expressos.

### 4.2.3 Avaliação da expressão diferencial

O programa Tetranscripts foi usado para realizar a análise de expressão diferencial de experimentos de domínio público de várias espécies diferentes. Foram analisados apenas experimentos envolvendo: exposição a algum tipo de estresse abiótico e exposição à infecção (estresse biótico) (Quadro 4), expressão em linhagens somáticas versus linhagens germinativas e mosquitos masculinos versus femininos (Quadro 5). Além disso foram comparados experimentos de mesma origem, a fim de evitar a inserção de viés na análise. Para evitar

também a inserção de ruído, definimos o parâmetro --minread igual a 10 e consideramos TEs expressos diferencialmente apenas aqueles com  $\log$  de *fold change*  $> |1|$ .

Quadro 4 – Experimentos de RNA-Seq de exposição a estresse abióticos e bióticos analisados (Continua)

Espécie	Artigo de origem	Condições	Código do SRA	Descrição do grupo
<i>Aedes albopictus</i>	(HUANG; POELCHAU; ARMBRUSTER, 2015)	Controle	SRR1663754 SRR1663769 SRR1663843 SRR1663911	Fêmeas adultas submetidas a dias longos (16 horas de luz) por 11 dias
		Criados sob condições indutoras de diapausa	SRR1663685 SRR1663687 SRR1663689 SRR1663697	Fêmeas adultas submetidas a dias curtos (8 horas de luz) por 11 dias
	(POELCHAU <i>et al.</i> , 2013)	Controle	SRR652101 SRR652102 SRR652103	Ovos com 11 dias após a oviposição e derivados de fêmeas parentais submetidas a dias longos (16 horas de luz)
		Criados sob condições indutoras de diapausa	SRR652072 SRR652088 SRR652090	Ovos com 11 dias após a oviposição e derivados de fêmeas parentais submetidas a dias curtos (8 horas de luz)
		Controle	SRR652116 SRR652118 SRR672801	Ovos com 21 dias após a oviposição e derivados de fêmeas parentais submetidas a dias longos (16 horas de luz)
		Criados sob condições indutoras de diapausa	SRR652091 SRR652092 SRR652097	Ovos com 21 dias após a oviposição e derivados de fêmeas parentais submetidas a dias curtos (8 horas de luz)
		Controle	SRR652119 SRR652120	Ovos com 40 dias após a oviposição e derivados de fêmeas parentais submetidas a dias longos (16 horas de luz)
		Criados sob condições indutoras de diapausa	SRR652098 SRR652100	Ovos com 40 dias após a oviposição e derivados de fêmeas parentais submetidas a dias curtos (8 horas de luz)
<i>Anopheles albimanus</i>	(MACKENZIE-IMPOINVIL <i>et al.</i> , 2019)	Controle	SRR8128633 SRR8128632 SRR8128635	Mosquitos adultos sadios
		Tratado com deltametrina	SRR8128646 SRR8128631 SRR8128630	Mosquitos adultos que sobreviveram a exposição à deltametrina, numa concentração de 12,5 µg/garrafa num bioensaio de garrafa do CDC
<i>Anopheles gambiae</i>	(SEAMAN <i>et al.</i> , 2015)	Controle	SRR2094645 SRR2094649	Mosquitos fêmeas com dois dias pós emergência alimentadas com sangue
		Tratado com ivermectina	SRR2094651 SRR2094652	Mosquitos fêmeas com dois dias pós emergência alimentadas com sangue contendo 11.75 ng/ml de ivermectina

Quadro 4 – Experimentos de RNA-Seq de exposição a estresse abióticos e bióticos analisados (Continuação)

Espécie	Artigo de origem	Condições	Código do SRA	Descrição do grupo
<i>Anopheles gambiae</i>	(SEAMAN <i>et al.</i> , 2015)	Controle	SRR2094655 SRR2094656	Mosquitos fêmeas com seis dias pós emergência alimentadas com sangue
		Tratado com ivermectina	SRR2094658 SRR2094659	Mosquitos fêmeas com seis dias pós emergência alimentadas com sangue contendo 11.75 ng/ml de ivermectina
<i>Anopheles merus</i>	(UYHELJI; CHENG; BESANSKY, 2016)	Controle	SRR2932565 SRR2932577 SRR2932589	Larvas com 24 horas de idade criadas em água doce
		Expostos a salinidade (20%)	SRR2932567 SRR2932579 SRR2932591	Larvas com 24 horas de idade expostas a solução salina de NaCl (15.85 g/L) por 6 horas.
		Expostos a salinidade (50%)	SRR2932570 SRR2932582 SRR2932594	Larvas com 24 horas de idade expostas a solução salina de NaCl (6.34 g/L) por 6 horas.
<i>Anopheles coluzzii</i>	(UYHELJI; CHENG; BESANSKY, 2016)	Controle	SRR2932571 SRR2932583 SRR2932595	Larvas com 24 horas de idade criadas em água doce
		Expostos a salinidade (20%)	SRR2932573 SRR2932585 SRR2932597	Larvas com 24 horas de idade expostas a solução salina de NaCl (15.85 g/L) por 6 horas.
		Expostos a salinidade (50%)	SRR2932576 SRR2932588	Larvas com 24 horas de idade expostas a solução salina de NaCl (6.34 g/L) por 6 horas.
<i>Aedes aegypti</i>	(KANG <i>et al.</i> , 2018)	Controle	SRR7356171 SRR7356172 SRR7356173	Fêmeas adultas derivadas de larvas crescidas sob condições ótimas, e cujo RNA foi 3h após a alimentação sanguínea
		Derivado de estresse populacional	SRR7356183 SRR7356184 SRR7356185	Fêmeas adultas derivadas de larvas crescidas sob densidade populacional 10X maior que as condições ótimas, e cujo RNA foi 3h após a alimentação sanguínea
		Controle	SRR7356177 SRR7356178 SRR7356179	Fêmeas adultas derivadas de larvas crescidas sob condições ótimas, e cujo RNA foi 18h após a alimentação sanguínea
		Derivado de estresse populacional	SRR7356189 SRR7356190 SRR7356191	Fêmeas adultas derivadas de larvas crescidas sob densidade populacional 10X maior que as condições ótimas, e cujo RNA foi 18h após a alimentação sanguínea
<i>Aedes aegypti</i>	(DONG; BEHURA; FRANZ, 2017)	Controle 1 dia	SRR5288091 SRR5288094 SRR5288075	Intestino médio de fêmeas com uma semana de idade
		Infectados 1 dia	SRR5288084 SRR5288099 SRR5288086	Intestino médio de fêmeas com uma semana de idade e expostas a infecção com CHIKV (1 dpi)

Quadro 4 – Experimentos de RNA-Seq de exposição a estresse abióticos e bióticos analisados (Continuação)

Espécie	Artigo de origem	Condições	Código do SRA	Descrição do grupo
<i>Aedes aegypti</i>	(DONG; BEHURA; FRANZ, 2017)	Controle 2 dias	SRR5288079 SRR5288078 SRR5288095	Intestino médio de fêmeas com uma semana de idade
		Infectados 2 dias	SRR5288098 SRR5288074 SRR5288076	Intestino médio de fêmeas com uma semana de idade e expostas a infecção com CHIKV (2 dpi)
	(CHOWDHURY <i>et al.</i> , 2020)	Controle 7 dias	SRR8921124 SRR8921123	Glândulas salivares de fêmeas sadias
		Infectados CHIKV 7 dias	SRR8921132 SRR8921131	Glândulas salivares de fêmeas expostas a infecção com CHIKV (7 dpi)
		Controle 14 dias	SRR8921126 SRR8921125	Glândulas salivares de fêmeas sadias
		Infectados DENV 14 dias	SRR8921130 SRR8921129	Glândulas salivares de fêmeas expostas a infecção com DENV2 (14 dpi)
		Controle 14 dias	SRR8921126 SRR8921125	Glândulas salivares de fêmeas sadias
		Infectados ZIKV 14 dias	SRR8921128 SRR8921127	Glândulas salivares de fêmeas expostas a infecção com ZIKV (14 dpi)
	(ETEBARI <i>et al.</i> , 2017)	Controle 2 dias	SRR5955010 SRR5955011 SRR5955012	Corpo total de fêmeas sadias
		Infectados ZIKV 2 dias	SRR5955013 SRR5955015 SRR5955014	Corpo total de fêmeas expostas a infecção com ZIKV (2 dpi)
		Controle 7 dias	SRR5955017 SRR5955016 SRR5955018	Corpo total de fêmeas sadias
		Infectados ZIKV 7 dias	SRR5955019 SRR5955021 SRR5955020	Corpo total de fêmeas expostas a infecção com ZIKV (7 dpi)
		Controle 14 dias	SRR5955023 SRR5955022 SRR5955024	Corpo total de fêmeas sadias
		Infectados ZIKV 14 dias	SRR5955026 SRR5955025 SRR5955027	Corpo total de fêmeas expostas a infecção com ZIKV (14 dpi)
	(KANG <i>et al.</i> , 2018)	Controle 3 horas	SRR7356171 SRR7356172 SRR7356173	Corpo total de fêmeas sadias
		Infectados DENV 3 horas	SRR7356168 SRR7356169 SRR7356170	Corpo total de fêmeas expostas a infecção com DENV2 (3 hpi)
		Controle 18 horas	SRR7356177 SRR7356178 SRR7356179	Corpo total de fêmeas sadias

Quadro 4 – Experimentos de RNA-Seq de exposição a estresse abióticos e bióticos analisados (Conclusão)

Espécie	Artigo de origem	Condições	Código do SRA	Descrição do grupo
<i>Aedes aegypti</i>	(KANG <i>et al.</i> , 2018)	Infectados DENV 3 horas	SRR7356174 SRR7356175 SRR7356176	Corpo total de fêmeas expostas a infecção com DENV2 (18 hpi)
	(CHOI <i>et al.</i> , 2014)	Controle 1 dia	SRR1061170 SRR1061168 SRR1061169	Tórax de fêmeas (Liverpool) sadias
		Infectados <i>B. malayi</i> 1 dia	SRR1060807 SRR1060808 SRR1060996	Tórax de fêmeas (Liverpool) infectadas com <i>Brugia malayi</i> (1 dpi)
		Controle 3 dias	SRR1061196 SRR1061197 SRR1061198	Tórax de fêmeas (Liverpool) sadias
		Infectados <i>B. malayi</i> 3 dias	SRR1060839 SRR1060848 SRR1061020	Tórax de fêmeas (Liverpool) infectadas com <i>Brugia malayi</i> (3 dpi)
		Controle 5 dias	SRR1061224 SRR1061225 SRR1061226	Tórax de fêmeas (Liverpool) sadias
		Infectados <i>B. malayi</i> 5 dias	SRR1060869 SRR1060867 SRR1061049	Tórax de fêmeas (Liverpool) infectadas com <i>Brugia malayi</i> (5 dpi)
		Controle 7 dias	SRR1061252 SRR1061254 SRR1061253	Tórax de fêmeas (Liverpool) sadias
		Infectados <i>B. malayi</i> 7 dias	SRR1060895 SRR1060893 SRR1061075	Tórax de fêmeas (Liverpool) infectadas com <i>Brugia malayi</i> (7 dpi)

Fonte: Elaborado pelo autor

#### 4.2.4 Verificação da coexpressão entre TEs e genes

Para verificar possíveis efeitos *cis* regulatórios dos elementos de transposição sobre genes próximos foi utilizado o pacote TEffectR (KARAKÜLAH *et al.*, 2019), que é executado no ambiente R. Para isso, os arquivos GTF de anotação dos genes e TEs foram transformados nos formatos de arquivo exigidos pelo pipeline. Os arquivos .bam de mapeamento das *reads* de RNA-Seq gerados nas etapas anteriores também foram passados ao pipeline, que aplica uma série de métricas para inferir, normalizar e transformar os dados de expressão gênica, e em seguida aplicar uma análise de regressão para verificar se duplas de genes e TEs se coexpressam significativamente ao longo de diversas corridas de RNA-Seq.

Quadro 5 – Experimentos de RNA-Seq de em linhagens somáticas e germinativas dos mosquitos (Continua)

Espécie	Artigo de origem	Tipo de amostra	Código do SRA
<i>Anopheles albimanus</i>	(PAPA <i>et al.</i> , 2017)	Carçaça de fêmeas	SRR4104312; SRR4104313; SRR4104314
		Ovários	SRR4104319; SRR4104340; SRR4104341
		Carçaça de machos	SRR4104309; SRR4104310; SRR4104311
		Testículos e glândulas acessórias	SRR4104316; SRR4104317; SRR4104318
<i>Anopheles gambiae</i>		Carçaça de fêmeas	SRR1805328; SRR1805349; SRR1805330
		Ovários	SRR1805332; SRR1805341; SRR1805334
		Carçaça de machos	SRR1805322; SRR1805324; SRR1805343
		Testículos e glândulas acessórias	SRR1805326; SRR1805345; SRR1805347
<i>Anopheles arabiensis</i>		Carçaça de fêmeas	SRR4104315; SRR4104342; SRR4104343
		Ovários	SRR4104218; SRR4104219; SRR4104347
		Carçaça de machos	SRR4104216; SRR4104217; SRR4104276
		Testículos e glândulas acessórias	SRR4104344; SRR4104345; SRR4104346
<i>Anopheles minimus</i>		Carçaça de fêmeas	SRR4104225; SRR4104223; SRR4104224
		Ovários	SRR4104279; SRR4104278; SRR4104308
		Carçaça de machos	SRR4104221; SRR4104222; SRR4104220
		Testículos e glândulas acessórias	SRR4104247; SRR4104246; SRR4104277
<i>Anopheles coluzzii</i>	(CASSONE <i>et al.</i> , 2017)	Carçaça de machos	SRR1587094; SRR1587095; SRR1587096; SRR1587097
		Testículos	SRR1587089; SRR1586022; SRR1587087; SRR1587088
<i>Anopheles merus</i>		Carçaça de machos	SRR1587098; SRR1587099; SRR1587100; SRR1587101
		Testículos	SRR1587090; SRR1587091; SRR1587092; SRR1587093
<i>Aedes aegypti</i>	(MATTHEWS <i>et al.</i> , 2018)	Carçaça de fêmeas	SRR6311360; SRR6311357; SRR6311358
		Ovários	SRR6311330; SRR6311331; SRR6311393



Quadro 5 – Experimentos de RNA-Seq de em linhagens somáticas e germinativas dos mosquitos (Conclusão)

Espécie	Artigo de origem	Tipo de amostra	Código do SRA
<i>Aedes aegypti</i>	(MATTHEWS <i>et al.</i> , 2018)	Carcaça de machos	SRR6311355; SRR6311356; SRR6311353
		Testículos	SRR6311396; SRR6311395; SRR6311391

Fonte: Elaborado pelo Autor

A fim de aumentar a quantidade de informação de expressão gênica de cada espécie, e dessa forma potencializar a análise de regressão linear, foram utilizados dados de diversos experimentos para cada espécie. Como o uso de dois ou mais experimentos conduzidos por pessoas diferentes, em locais diferentes ou utilizando equipamentos diferentes pode inserir ruído na análise de expressão, foi necessária a utilização de uma abordagem para correção desses ruídos. Esses efeitos indesejados, conhecidos como *batch effects*, já são conhecidos e tratados na meta-análise de dados de *Microarrays* e RNA-Seq através de algumas metodologias estatísticas de correção como o ComBat (JOHNSON; LI; RABINOVIC, 2007).

Uma vez que a construção das matrizes de expressão das cópias dos TEs próximas a genes é realizada dentro de umas das funções do pacote TEffectR, a correção para remoção dos *Batch effects* deve ser feita enquanto a função está em execução. Dessa forma foi necessária a alteração do código em R do programa original para inserir a opção de correção dos *Batch effects*. Como isso deve ser realizado no R, foi utilizada a função ComBat do pacote estatístico SVA (*Surrogate Variable Analysis* - versão 3.32.1) (LEEK *et al.*, 2012, 2019; LEEK; STOREY, 2007) disponibilizado no Bioconductor (GENTLEMAN *et al.*, 2004). Os trechos alterados/adicionados na função *apply\_lm* estão descritos na Figura 13.

Figura 13 – Função modificada no código do programa TEffectR para lidar com *batch effects*.

```

1 apply_lm_modified<-function(gene.annotation, gene.counts, repeat.counts, batch_effects=NULL,
2                             covariates=NULL, prefix){
3     ##### bloco não modificado
23     ##### Ajuste dos batch effects com ComBat #####
24     if(!is.null(batch_effects)){
25         batch = batch_effects$Study
26         v$E = ComBat(dat=v$E, batch=batch)
27     }
28     ##### bloco não modificado
88 }

```

Fonte – Elaborado pelo autor.

Neste trecho de código, uma matriz da expressão dos genes e TEs, já normalizada através de CPM e cujos dados encontram-se em escala logarítmica, é passada para a função ComBat,

que também recebe como argumento, um arquivo contendo os possíveis *batch effects*. Neste estudo os grupos de *batch* (lote) foram definidos como os códigos identificadores de cada experimento de RNA-Seq obtidos do SRA, ou seja, cada amostra de um estudo foi colocada num mesmo lote, amostras de estudos diferentes estão em diferentes lotes, como já realizado em estudos anteriores (VANDENBON, 2022).

Para a análise foram avaliados os potenciais efeitos *cis*-regulatórios dos TEs. Dessa forma foram avaliadas inserções de TEs localizados até 10kb *upstream* da região de início de um gene e que se localizavam na mesma fita deste, ou seja, são transcritos no mesmo sentido do gene. Para reduzir ao máximo o número de inferências inadequadas, foram consideradas relações significativas apenas aquelas cujo *p-value* foi menor que 0,01 e cujo coeficiente de correlação (R) foi maior que 0,5. O R pode variar entre -1 e +1, onde os extremos indicam forte correlação e o valor central (0) nenhuma relação linear entre as variáveis. Com base em padrões já estabelecidos sobre análise de correlação, relações lineares com  $R > 0,9$  foram consideradas muito fortes, com  $0,9 \leq R \leq 0,7$  foram consideradas fortes, com  $0,7 \leq R \leq 0,5$  foram consideradas moderadas e com  $R < 0,5$  foram consideradas fracas ou muito fracas (MUKAKA, 2012). Este último grupo não foi reportado neste trabalho.

A fim de descobrir vias potencialmente sobre influência dos TEs, as proteínas dos genes que apresentaram coexpressões significativas com TEs foram utilizadas para construir redes de interação usando o banco de dados STRING (SZKLARCZYK *et al.*, 2021). Para isso, os primeiros interagentes de cada proteína foram extraídos. A lista de proteínas e seus interagentes foi então passada como entrada para construção de uma rede única no STRING. A rede construída levou em consideração interações físicas e funcionais entre as proteínas. Foram consideradas significativas as interações com *score* maior que 0,5. Nestas redes, foram realizadas análises de agrupamentos utilizando o algoritmo de *k-means* com o número de clusters igual ao número de genes influenciados por TEs numa determinada espécie. Para cada cluster foi realizada uma análise de enriquecimento de GO e vias KEGG.

Redes de interação física proteína-proteína, contendo todo o proteoma de uma espécie, também foram extraídas do STRING para cada uma das espécies de mosquito onde a detecção de genes e TEs coexpressos foi positiva. Para isso, os códigos de identificação das proteínas que compõem o proteoma das espécies foram extraídos do UniProt proteomes (BATEMAN *et al.*, 2021) e passados ao STRING, sendo consideradas significativas interações com *score* maior que 0,7. A rede resultante foi passada ao Cytoscape (SHANNON *et al.*, 2003) para o cálculo

das proteínas que se comportam como *hubs* (nós com maior influência e relevância nas redes). Para isso, foram utilizadas duas métricas de centralidade dos nós, com o objetivo de verificar se os TEs influenciam a expressão de algum gene que se comporta como *hub*. Uma das métricas é classicamente utilizada nos estudos das redes e a outra foi escolhida por mostrar o melhor resultado na definição dos *hubs* em redes de interação proteína-proteína (ASHTIANI *et al.*, 2018) :

- a) *Degree centrality*: Mede a quantidade de conexões que é realizada por cada nó;
- b) *Closeness centrality*: mede a rapidez com que as informações podem fluir de/para um nó para/de todos os outros nós na rede, ou seja, é uma forma de identificar os nós que propagam informações de forma mais eficiente na rede. Calculada como o inverso da soma do comprimento dos caminhos mais curtos entre o nó e todos os outros nós no grafo;

#### 4.2.5 Reconstrução de transcritos derivados de TEs

Para recuperar a sequência nucleotídica de quaisquer transcrições potenciais derivadas de elementos transponíveis, os arquivos BAM de cada espécie foram submetidos ao software de montagem Trinity (versão: Trinity-v2.12.0) (GRABHERR *et al.*, 2011). O Trinity guiado pelo genoma (*genome-guided Trinity Assembly*) foi escolhido para reduzir a chance de montagem artefatos e para evitar que o transcrito de TE obtido fosse na verdade um consenso de diversos transcritos derivados de cópias similares, como ocorreu em artigos que utilizaram uma montagem de Trinity sem nenhuma referência (VENDRELL-MIR *et al.*, 2020). Outra forma de obter construções mais confiáveis foi usar corridas de RNA-Seq da mesma linhagem utilizada para construção do genoma de referência de cada espécie. Para homogeneizar a descoberta de transcritos derivados de TEs, apenas experimentos realizados com mosquitos adultos foram usados na montagem. Para avaliar a integridade da montagem, executamos o BUSCO (SIMÃO *et al.*, 2015) (versão 3.1.0) para cada transcriptoma montado, usando o banco de dados Diptera. Transcritos derivados de TEs foram descobertos mapeando-se os transcritos montados contra os consensos de TE de cada espécie de mosquito (utilizando blastn) e contra a base de proteínas de TEs do Repbase (utilizando blastx).

A fim de recuperar TEs expressos com potencial codificante para proteínas, foi usado o TransDecoder (-m 200) a fim de predizer os quadros de leitura aberta (Open Reading Frames - ORFs) para cada transcrição derivada de TE. Para evitar a predição de ORFs espúrias em

detrimento das ORFs corretas de um TE, ou seu local de início incorreto, incluímos pesquisas de homologia como critérios de retenção de ORF. Um mapeamento de blastn foi realizado contra um banco de dados que consiste em sequências de proteínas de elementos transponíveis disponíveis em Repbase (RepBase20.05 - 24195 proteínas), além de sequências de proteínas TE (3145 proteínas) disponíveis em UniProt SwissProt (BATEMAN *et al.*, 2021). Além disso, realizamos uma pesquisa de perfis usando HMMER 3.1 (EDDY, 2011) para identificar domínios de proteínas dentro das ORFs, usando o banco de dados ProfilesBankForREPET\_Pfam27.0\_GypsyDB.hmm (disponível em: <https://urgi.versailles.inra.fr/Tools/REPET>), que inclui os modelos ocultos de Markov depositados no Pfam (MISTRY *et al.*, 2021) e GypsyDB (LLORENS *et al.*, 2011).

#### 4.2.6 Avaliação da sobreposição de TEs e RNAs mensageiros

Alguns TEs podem se sobrepor aos genes dos hospedeiros e assim podem influenciar sua expressão. O software GffCompare, um programa do *GFF Utilities* (PERTEA; PERTEA, 2020), foi usado para comparar a anotação de referência da região de codificação de cada mosquito estudado (como opção de referência) e o arquivo GTF de anotação TE (como *query*). Isso estimou a proporção de TEs em regiões intergênicas e intrônicas assim como com sobreposição a éxons.

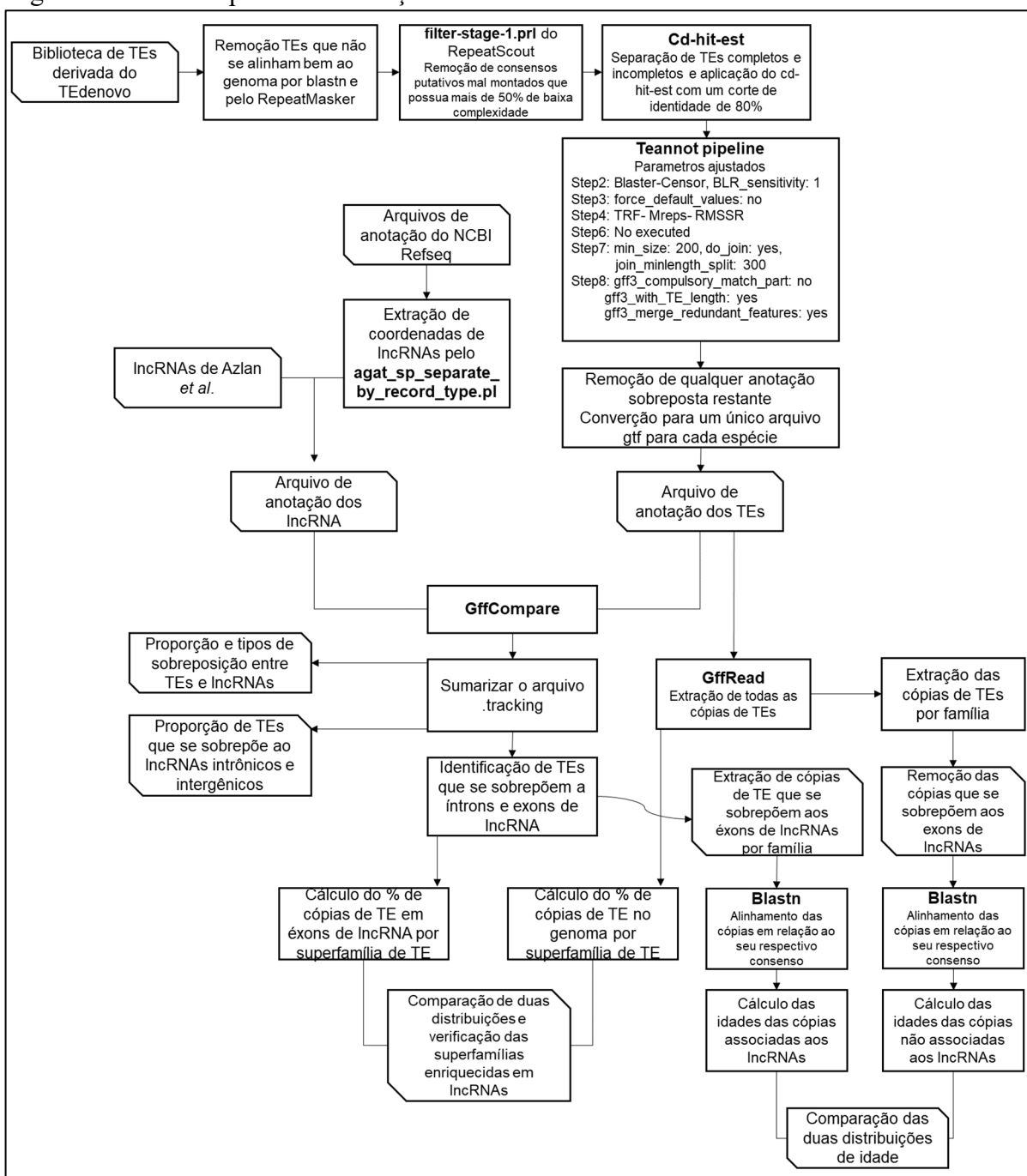
#### 4.2.7 Avaliação da sobreposição de TEs e RNAs longos não-codificantes

A fim de investigar o papel dos TEs na biogênese do lncRNAs das espécies foram utilizadas as espécies *Ae. albopictus* e *Cx. quinquefasciatus*, uma vez que apenas nelas há caracterizações sistemáticas dos lncRNAs em montagens de genoma de alta qualidade (GCF\_006496715.1 e GCF\_015732765.1, respectivamente). A Figura 14 mostra todo o fluxo de trabalho de análise. Os arquivos de anotação TE foram gerados usando o pipeline TEannot (FLUTRE *et al.*, 2011) com parâmetros personalizados. Apenas cópias TE anotadas maiores que 200 nucleotídeos foram mantidas, uma vez que esse é o tamanho mínimo de um lncRNA. Além disso, diminuir o tamanho da cópia TE aumentaria a chance de correspondências espúrias ou atribuições incorretas de superfamília.

Em seguida, o conjunto de anotações de lncRNAs disponíveis do *Ae. albopictus* RefSeq GCF\_006496715.1 foram obtidas e combinadas com o conjunto de dados de lncRNAs anotado

por Azlan *et al.* (2021) para gerar as coordenadas de lncRNAs desta espécie. O conjunto de dados de anotação de *Cx. quinquefasciatus* RefSeq VPISU\_Cqui (GCF\_015732765.1) foi usado como fonte de lncRNA. Também foi utilizado o conjunto de dados de lncRNA anotado por Azlan *et al.* (2020) para comparar entre a versão de montagem atual e anterior (Cpip2). Por fim, o software GffCompare (PERTEA; PERTEA, 2020) foi utilizado para encontrar o tipo de sobreposição entre os TEs e os lncRNAs em ambas as espécies.

Figura 14 – Passo a passo da avaliação do envolvimento dos TEs e lncRNAs.



Fonte: Elaborado pelo autor.

Posteriormente foi avaliado se a sobreposição entre genes de lncRNAs e TEs ocorre com maior frequência em íntrons ou éxons de lncRNAs. Também foi executada uma análise de enriquecimento usando o pacote clusterProfiler (YU *et al.*, 2012), versão 3.10.1, para verificar se os TEs envolvidos com lncRNAs são igualmente distribuídos entre os diversos tipos de TEs ou se há a seleção de algumas superfamílias específicas.

Para avaliar se existem padrões de idade distintos de cópias TE coocorrendo ou não com éxons de lncRNAs, a divergência de cada cópia TE que se sobrepõe com os lncRNAs e a divergência de cada cópia TE que não se sobrepõe com lncRNAs foram estimadas em relação ao seu respectivo consenso de TE (família), usando blastn. A divergência de cada cópia TE para a respectiva sequência de consenso da família TE representa a idade relativa de cada cópia desde sua separação do elemento ancestral, a sequência de consenso da família TE. Ou seja, cópias TE mais jovens tiveram menos tempo para acumular mutações desde sua separação do TE ancestral e apresentam menor divergência ao consenso, enquanto cópias TE mais antigas tiveram mais tempo para acumular mutações e são caracterizadas por maior divergência ao consenso. Um teste de Wilcoxon Rank Sum (wilcox.test em R) foi usado para comparar a divergência das cópias entre os dois conjuntos de dados.

### 4.3 ABORDAGEM PROTEÔMICA

#### 4.3.1 Definição das espécies e obtenção dos espectros de massas

Para a identificação de elementos de transposição que ainda são capazes de gerar proteínas no hospedeiro, experimentos de espectrometria de massa foram analisados. Uma busca foi realizada no banco de dados de identificações proteômicas (PEREZ-RIVEROL *et al.*, 2019) (PRIDE - <https://www.ebi.ac.uk/pride/>) para recuperar experimentos de espectrometria de massa para todas as 12 espécies estudadas. Destas, apenas três espécies (Quadro 6), *An. gambiae*, *An. stephensi* e *Ae. aegypti*, apresentaram dados satisfatórios (dados de experimentos de espectrometria de massa *in tandem* – LC-MS/MS, experimentos bem descritos e com mosquitos adultos como amostra).

Os arquivos .raw de cada experimento foram baixados do PRIDE e convertidos localmente para o formato .mzML usando ThermoRawFileParser (opção *Peak Picking* definida) (HULSTAERT *et al.*, 2020) integrado na interface SearchGUI (BARSNES; VAUDEL, 2018).

Quadro 6 – Dados dos experimentos utilizados para análise proteômica de elementos de transposição neste trabalho.

Espécies	Código do PRIDE	Referência	Tecido	Modificações
<i>Ae. aegypti</i>	PXD002468	(DHAWAN <i>et al.</i> , 2017)	Glândula salivar	Carbamidometilação de cisteína (fixa); oxidação da metionina (variável)
<i>Ae. aegypti</i>	PXD010293	(DEGNER <i>et al.</i> , 2019)	Sêmen	Carbamidometilação de cisteína (fixa); oxidação de metionina e desamidação de glutamina e asparagina (variável)
<i>Ae. aegypti</i>	PXD022665	(MARTINS <i>et al.</i> , 2021)	Cabeça	Carbamidometilação de cisteína (fixa); iTRAQ 4-plex (Tyr, Lys e peptídeo N-terminal), fosfato (Ser, Thr e Tyr) e oxidação de metionina (variável)
<i>Ae. aegypti</i>	PXD003799	(MARINGER <i>et al.</i> , 2017)	Linhagem celular Aag2	Carbamidometilação de cisteína (fixa); oxidação da metionina e acetilação proteica N-terminal (variável)
<i>An. gambiae</i>	PXD000630	(DWIVEDI <i>et al.</i> , 2014)	Cérebro	Carbamidometilação de cisteína (fixa); oxidação da metionina e acetilação proteica N-terminal (variável)
<i>An. gambiae</i>	PXD016300	(ZOURÉ <i>et al.</i> , 2020)	Corpo inteiro	Carbamidometilação de cisteína (fixa); oxidação da metionina (variável)
<i>An. stephensi</i>	PXD001128	(PRASAD <i>et al.</i> , 2017)	Múltiplos tecidos	Carbamidometilação de cisteína (fixa); oxidação da metionina e acetilação proteica N-terminal (variável)

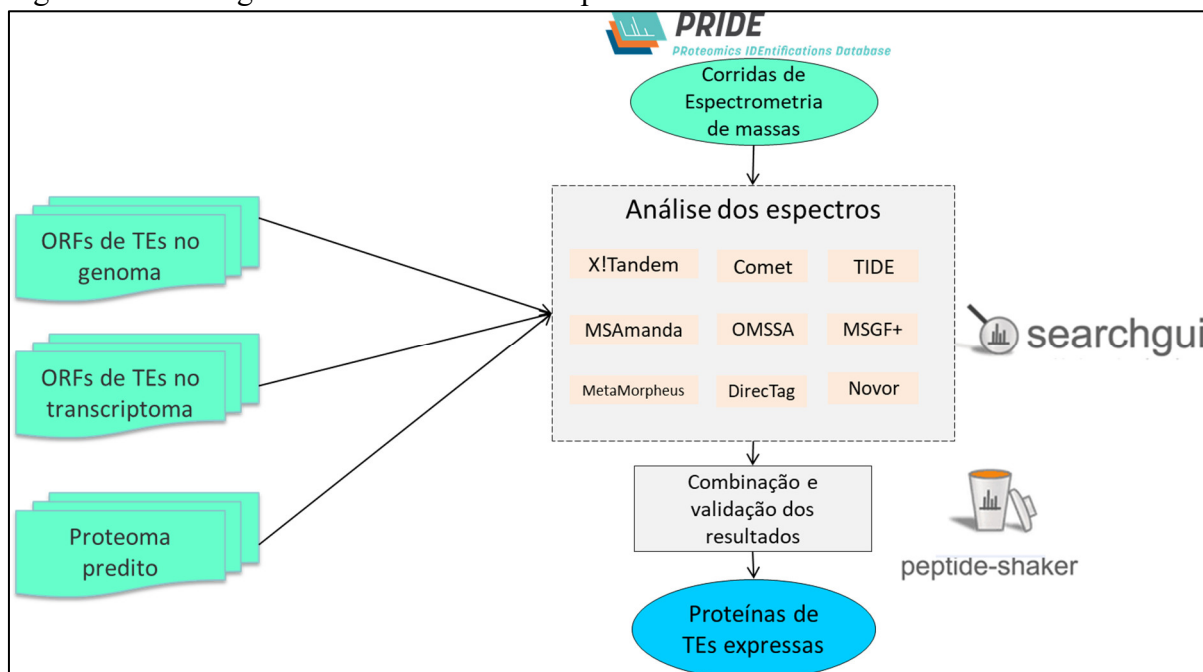
Fonte: Elaborado pelo autor

#### 4.3.2 Análise dos dados proteômicos

Os resultados das corridas de espectrometria de massa foram analisados utilizando-se a técnica de proteômica informada pela transcriptômica (*Proteomics informed by transcriptomics* – PIT), que utiliza sequências proteicas preditas diretamente do transcriptoma montado, e utilizando-se também de conjunto das proteínas de TEs preditas diretamente do genoma. Dessa forma, um banco de dados diferente para cada espécie foi construído abrangendo: 243 proteínas contaminantes comuns, o conjunto de dados de proteínas de TE previsto pelo Transdecoder tanto no transcriptoma montado pelo Trinity quanto no genoma, e o proteoma de referência daquela espécie (derivado do UniProt). Para a construção do conjunto de dados de proteínas TEs uma série de passos foi realizada: em primeiro lugar, proteínas de TE completas foram separadas das proteínas TE truncadas. Depois, a redundância em ambos os subconjuntos foi removida usando o programa cd-hit (parâmetros: -c 0.98 -G 1) (LI; GODZIK, 2006). Posteriormente, as proteínas do subconjunto de proteínas truncadas com alta similaridade (>

99%) com as proteínas do subconjunto completo foram removidas usando cd-hit-2d, isso mantém apenas uma isoforma para cada TE, de preferência a isoforma mais completa e que contém mais informação para a etapa de mapeamento. Finalmente, os dois subconjuntos foram unidos para construir o conjunto de dados de proteínas de TEs de cada espécie. O resumo das etapas da análise proteômica está mostrado na figura 15.

Figura 15 – Fluxograma da análise de dados proteômicos.



Fonte: Elaborado pelo autor.

As análises proteômicas foram conduzidas no SearchGUI utilizando a abordagem *target-decoy* e testando diferentes ferramentas de busca para cada experimento: X! TANDEM (CRAIG; BEAVIS, 2004), MS Amanda (DORFER *et al.*, 2014), MS-GF + (KIM; PEVZNER, 2014), Comet (ENG; JAHAN; HOOPMANN, 2013) e Tide (DIAMENT; NOBLE, 2011). A base *decoy* utilizada foi criada a partir das sequências reversas das sequências de aminoácidos das proteínas reais dos mosquitos. Esse software também nos permitiu executar uma análise *de novo* dos peptídeos a partir da utilização do programa Novor (MA, 2015).

As análises dos espectros foram realizadas com especificidade de clivagem de enzimática; Enzima: Tripsina; Carga do precursor: 2-4; e *peptide mass tolerance, fragment ion mass tolerance*, o número máximo de clivagens perdidas, modificações fixas e variáveis de acordo com os métodos descritos nos manuscritos originais dos experimentos. Foi utilizada uma FDR (*false discovery rate*) em nível de proteína de 0,01, definida como o ponto de corte



para as identificações. Os resultados de diferentes mecanismos de pesquisa foram combinados usando PeptideShaker (Vaudel *et al.* 2015). Uma combinação específica de motores de busca foi escolhida entre aquelas que aumentam o número de grupos de proteínas e proteínas TE identificadas e minimizam a PEP *resolution* (resolução de probabilidade de erro posterior).

## **5 CONSIDERAÇÕES ÉTICAS**

Esta tese teve como objetivo principal caracterizar os elementos de transposição presentes nos Culicídeos, assim como verificar seu padrão de expressão e possíveis implicações funcionais. Toda a metodologia foi realizada através de métodos de análise computacional de dados públicos referentes aos mosquitos presentes em diversas bases de dados. Portanto, não foram utilizados animais de experimentação nem material biológico oriundo de seres humanos. Desta forma este trabalho não necessita de parecer da Comissão de Ética em Pesquisas, nem da Comissão de Ética no Uso de Animais do IAM/FIOCRUZ.

## 6 RESULTADOS

Nesta seção são mostrados primeiramente os resultados da análise genômica dos elementos de transposição com a recharacterização do conteúdo de TEs dos mosquitos e avaliação da ocorrência de transferência horizontal desses TEs dentro e fora da família Culicidae. Como resultado, centenas de transferências horizontais de TEs que ocorreram nos últimos 30 milhões de anos e que contribuíram significativamente para a expansão genômica dos genomas de mosquitos foram encontradas. Alguns desses TEs transferidos horizontalmente estão envolvidos em HTTs com espécies distantes de outros Filos. Também foi encontrada uma espécie de vetor, o nematoide *Wuchereria bancrofti*, que facilitou a disseminação horizontal de um TE específico entre os mosquitos. Esses dados trouxeram uma contribuição substancial para a compreensão do fenômeno HTT e seu impacto nos genomas dos hospedeiros.

Posteriormente são mostrados os resultados da análise funcional dos TEs, onde foi visto que grande parte das famílias de TEs são expressas, com presença de alguns transcritos de TEs potencialmente íntegros e completos, com presença de proteínas de TEs expressas nos mosquitos. Foi observado, no entanto, que muitos dos transcritos não possuem muito potencial codificante e parecem ser lncRNAs. A contribuição dos TEs para a biogênese do lncRNA em *Aedes albopictus* e *Culex quinquefasciatus* foi então investigada. Observou-se que uma grande fração (cerca de 40-52%) dos loci de lncRNA coocorre com loci de TEs em ambas as espécies. A maioria das associações lncRNA/TE são representadas por sequências derivadas de TE que são expressas como um ou todos os éxons de lncRNAs, incluindo cinco lncRNAs que parecem influenciar genes relacionados ao sistema imunológico envolvidos na resposta antiviral. Em suma, os resultados indicam que elementos transponíveis contribuem substancialmente para a biogênese de lncRNAs e em escala bem menor para produção de proteínas nos mosquitos.

### 6.1 RESULTADOS DA ANÁLISE GENÔMICA

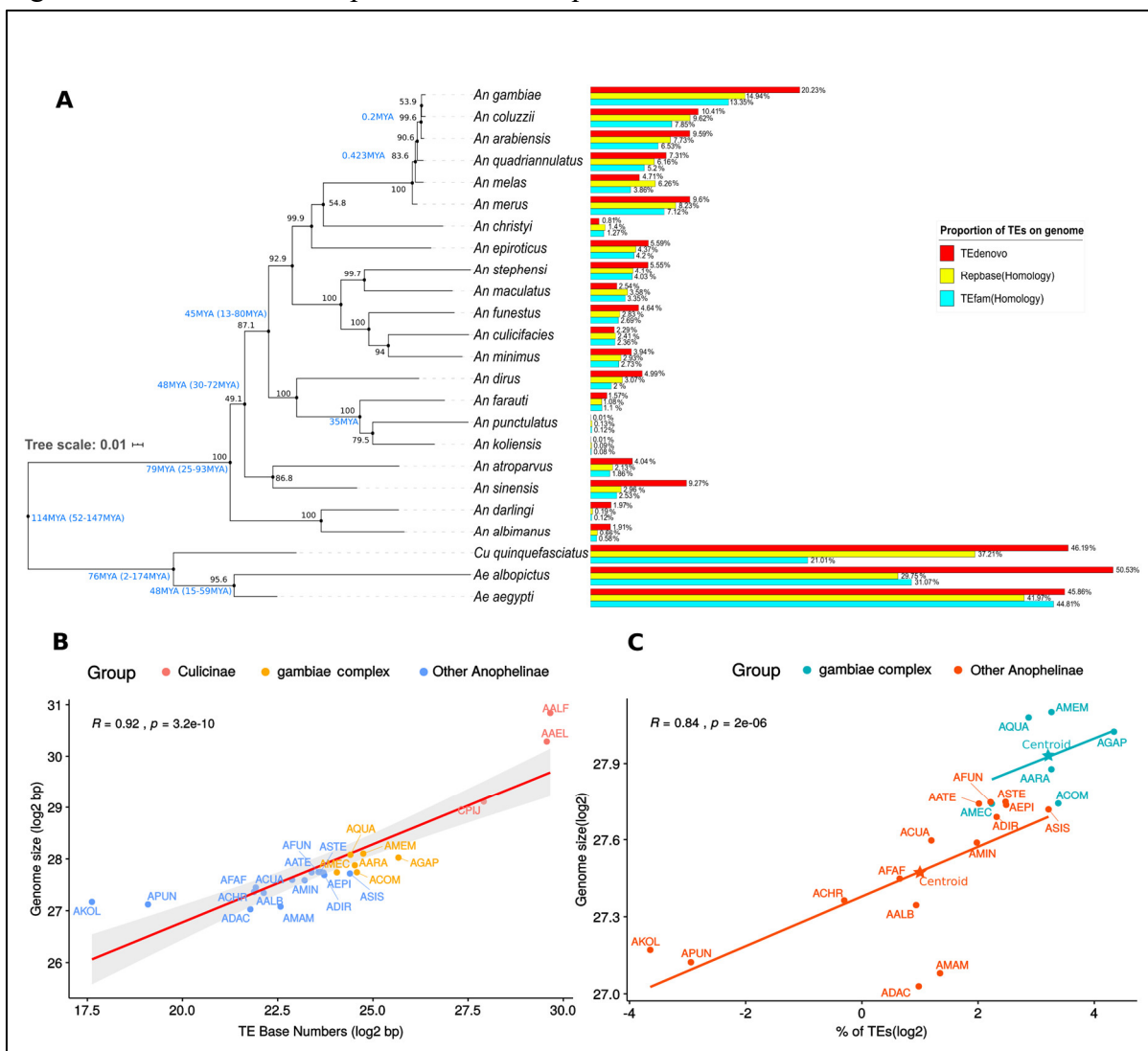
Nesta seção estão descritos os resultados dos diversos experimentos realizados utilizando como base a sequência genômica dos mosquitos.

#### 6.1.1 O conteúdo de TE é altamente variável entre as diferentes espécies de mosquitos

Foi realizada uma recharacterização de TEs em 24 genomas de mosquito disponíveis usando uma abordagem *de novo* e baseada em homologia. Para a primeira, foi utilizado o

pipeline TEdenovo (parte do pacote REPET), enquanto para o último foi aplicado o método baseado em homologia com dois bancos de dados TE diferentes, Repbase e TEfam. Todas as sequências resultantes foram classificadas por três programas diferentes, PASTEC, RepeatClassifier e TEsorter. Uma variação substancial foi detectada no conteúdo de TE de várias espécies, por exemplo, *An. gambiae* - a espécie de mosquito mais extensivamente estudada e que tem o maior número de TEs caracterizados armazenados em bancos de dados apresentou um conteúdo de TE de 20,23% pelo TEdenovo, 14,94% por pesquisa de homologia contra Repbase e 13,35% contra TEfam (Fig. 16A). A discrepância entre os três métodos é mais evidente em uma das espécies menos estudadas, *An. darlingi*. Uma diferença de 10 vezes foi observada entre TEdenovo (1,97%) e pesquisa baseada em homologia contra Repbase (0,19%).

Figura 16 – Elementos transponíveis dos mosquitos.



Fonte: Elaborado pelo autor.

Legenda: (A) Uma inferência filogenética baseada na máxima verossimilhança, com suporte de bootstrap sobre os nós, relatando a proporção do tamanho do genoma ocupado por elementos transponíveis; os valores das letras

azuis perto do suporte de *bootstrap* representam o tempo de divergência médio obtidas no TimeTree (KUMAR *et al.*, 2017) entre as espécies, as barras horizontais coloridas representam a proporção de TEs caracterizados por métodos diferentes. (B) Análise de correlação dos tamanhos do genoma do mosquito e fração TE. No complexo de espécies *An. gambiae* elementos não-TE contribuem mais para o tamanho do genoma do que em outros Anophelinae (C). As abreviações das espécies de quatro letras seguem a Tabela 1.

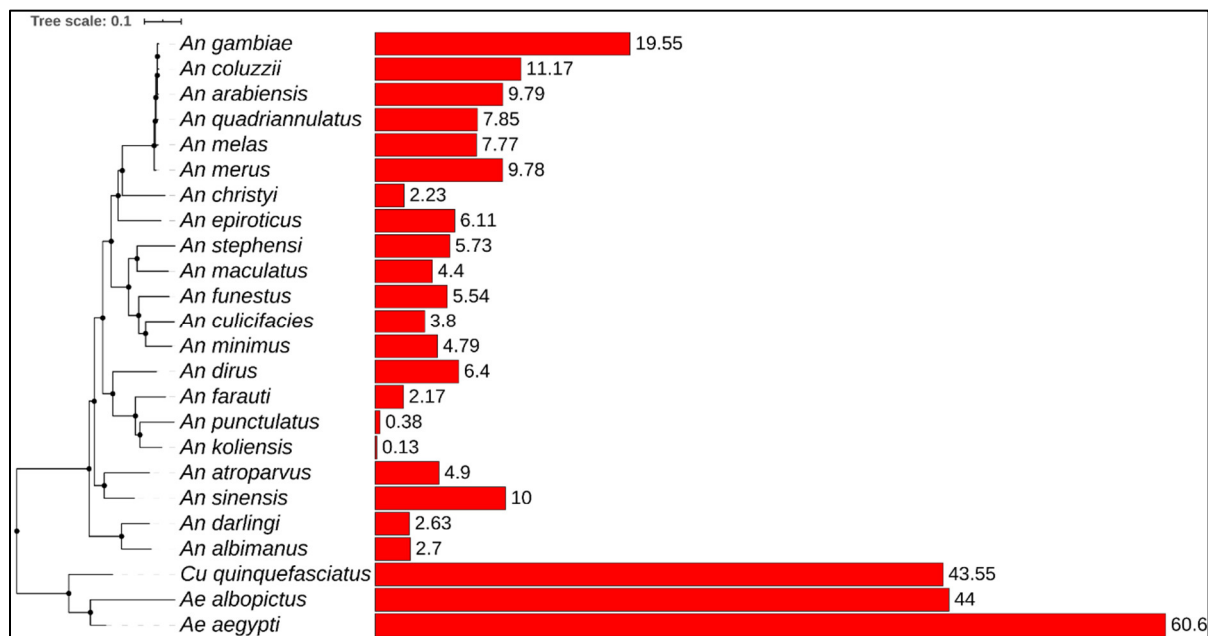
O complexo *An. gambiae*, que contém oito espécies africanas (BARRÓN *et al.*, 2019), seis das quais foram analisadas aqui — *An. gambiae*, *An. coluzzii*, *An. arabiensis*, *An. quadriannulatus*, *An. melas* e *An. merus* — mostrou a maior proporção de genoma coberto por elementos móveis entre todas as espécies de *Anopheles* investigadas. No outro extremo, os mosquitos da Nova Guiné, *An. farauti*, *An. punctulatus* e *An. koliensis* apresentaram um teor de TE muito baixo, entre 0,13% e 1,57%, considerando as estimativas mais altas para todas as metodologias. Uma vez que, nem a pesquisa baseada em homologia, nem a abordagem *de novo* identificou um conteúdo de TE maior do que 1% nas duas últimas espécies, tentamos outros métodos de caracterização de TE usando diretamente as *reads* de sequenciamento para verificar se essas estimativas podem ter sido influenciadas pelas etapas de remoção de TE durante a montagem original do genoma. O software Tedna encontrou muitas sequências repetitivas, porém poucas delas são elementos transponíveis. Por outro lado, as três espécies de mosquitos da subfamília Culicinae (*Ae. aegypti*, *Ae. albopictus* e *Culex quinquefasciatus*) têm um conteúdo de TE muito maior do que as espécies de Anophelinae, variando em cerca de 44% a mais de 50% em *Ae. albopictus* e *Ae. aegypti* respectivamente.

Para gerar um conjunto de dados final dos TEs recuperados usando cada metodologia, foram utilizados todos os TEs recuperados pelas três abordagens diferentes. Esse conjunto foi utilizado como base para alinhamentos usando o RepeatMasker, que estimou a proporção final de TEs em cada genoma (Fig. 17).

O conteúdo final de TE obtido a partir do mascaramento de todos os TEs recuperados foi usado para avaliar se o tamanho do genoma do mosquito se correlaciona com o conteúdo do mobiloma. A análise indicou uma forte correlação positiva ( $R = 0,92$ ) entre o conteúdo de TE e o tamanho do genoma dos mosquitos (Fig. 16B). Olhando apenas para as espécies de *Anopheles*, existem dois padrões, espécies do complexo *An. gambiae* tem um tamanho de genoma maior para um conteúdo semelhante de TE em comparação com outros mosquitos Anophelinae (Fig. 16C), sugerindo que TEs e sequências não-TE desempenham um papel importante no tamanho do genoma no complexo *An. gambiae*. Outros fenômenos na escala genômica, como duplicação segmentar e/ou expansão do conteúdo repetitivo não relacionado a TEs, como satélites e expansão da família de genes de múltiplas cópias, também podem ter

contribuído para a expansão do genoma, enquanto a expansão conduzida por TE tem um papel mais significativo no tamanho do genoma das espécies restantes de *Anopheles*.

Figura 17 – Fração do genoma ocupada por TEs, considerando sequências caracterizadas por ambos os métodos, *de novo* e baseado em homologia

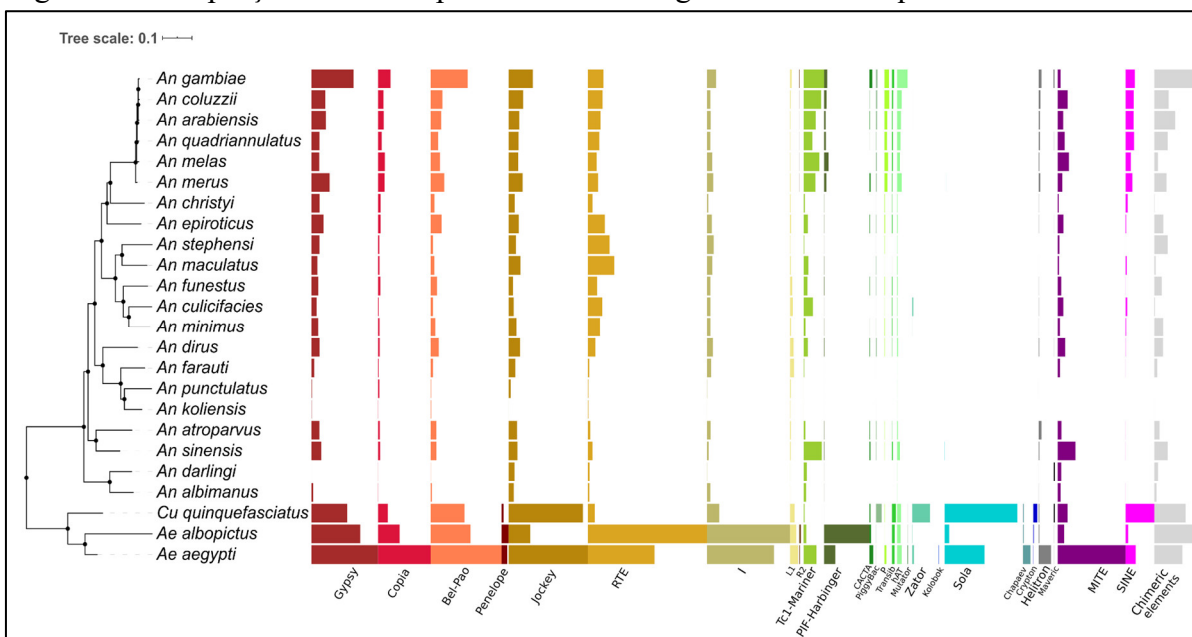


Fonte: Elaborado pelo autor.

### 6.1.2 Diversas superfamílias TE são encontradas nos mosquitos

Pelo menos 24 superfamílias de TEs foram encontradas nos genomas dos mosquitos (Fig. 18). Esse mobiloma diverso compreende a maioria das ordens de elementos de transposição estabelecidas por Wicker *et al.* (2007) e algumas superfamílias que não haviam sido descritas na época. Superfamílias LTR, como *Copia*, *Gypsy* e *Bel-Pao*, são onipresentes nos genomas de mosquitos, respondendo por pelo menos 6% do genoma de *Ae. aegypti* e 4,3% de *An. gambiae*. Os elementos LINE são abundantes na maioria das espécies, representando 14,5% do genoma de *Ae. aegypti*, 15,7% de *Ae. albopictus* e 6,2% de *Cx. quinquefasciatus*. *RTE* e *Jockey* são as superfamílias mais abundantes nos genomas dos mosquitos e ambas pertencem a ordem LINE. Doze diferentes superfamílias TIR estão bem representadas em mosquitos. Como esperado, *Tcl-Mariner*, a superfamília de TEs mais difundida entre os artrópodes, está presente em todas as espécies de mosquitos, mesmo naquelas com baixo teor de TE. Elementos TIR não autônomos, como prováveis MITEs, e SINEs, também estão presentes na maioria das espécies.

Figura 18 – Proporção de cada superfamília TE nos genomas de mosquitos.



**Fonte:** Elaborado pelo autor.

**Legenda:** Cada barra colorida representa a proporção de cada superfamília TE (nomes no eixo X) seguindo Wicker *et al.* (2007) esquema hierárquico. TEs não caracterizados ou caracterizados somente em nível de ordem são mostrados apenas na figura 19.

As superfamílias das Classes I e II são distribuídas de forma heterogênea entre os mosquitos. A superfamília Penelope foi encontrada apenas nos três mosquitos Culicinae estudados e cobriu apenas uma pequena fração desses genomas (Fig. 18). O mesmo padrão pode ser visto com elementos *Mutator*, *Chapaev* e *Crypton-like*. A diferença mais notável na distribuição da superfamília foi encontrada no caso dos elementos *Sola*. Este elemento é um dos TEs mais abundantes no genoma de *Cx. quinquefasciatus* (2,86%). Também está presente em mosquitos *Aedes* e em apenas uma espécie de Anophelinae – *An. sinensis* (representando quase 0,01% de seu genoma). Outra diferença clara é a ausência de elementos das superfamílias *CACTA*, *P* e *Transib* entre os anofelinos neotropicais e da Nova Guiné. *R2* é outra superfamília que possui uma intrigante distribuição irregular entre os mosquitos. Ela está presente apenas em mosquitos *Aedes* e em espécies do complexo *An. gambiae* (Fig. 18).

Do ponto de vista dos genomas dos mosquitos, o complexo de espécies *An. gambiae* tem uma distribuição de superfamílias de TEs semelhante, enquanto há uma variação substancial na distribuição e abundância entre os mosquitos *Aedes* e *Anopheles*. No entanto, é claro que *Cx. quinquefasciatus* possui o perfil de superfamílias de TEs mais distinto da família Culicidae, com a maior abundância de elementos *Zator*, *Sola* e *Crypton-like* em comparação com todas as outras espécies de mosquitos estudadas. Outra característica distintiva do mobiloma de *Cx. quinquefasciatus* é a baixa proporção genômica de superfamílias ubíquas entre mosquitos como

*Tc1-Mariner*, *RTE* e *I*. A diferença gigantesca do mobiloma de *Cx. quinquefasciatus* para o dos outros mosquitos possivelmente se deve à falta de dados genômicos de outras espécies do gênero *Culex* e também de gêneros próximos como *Lutzia*, *Deinocerites* e *Coquillettidia*.

Não foi possível classificar alguns elementos ao nível da superfamília, a maioria destes foram classificados apenas como Classe I e Classe II (Fig. 19). Esses elementos representam cerca de 15 a 20% de todos os TEs em mosquitos do complexo de espécies *An. gambiae* e 18% de *Cx. quinquefasciatus*, mas representam apenas uma pequena proporção nos outros genomas de mosquitos (2E-4 a 3%). Elementos quiméricos, formados por *nested elements* de duas ou mais superfamílias distintas, estão presentes em quase todas as espécies, mas uma maior proporção de TEs quiméricos foi encontrada nos genomas de *Ae. aegypti*, *An. arabiensis*, *An. gambiae* e *An. quadriannulatus* (Fig. 19).

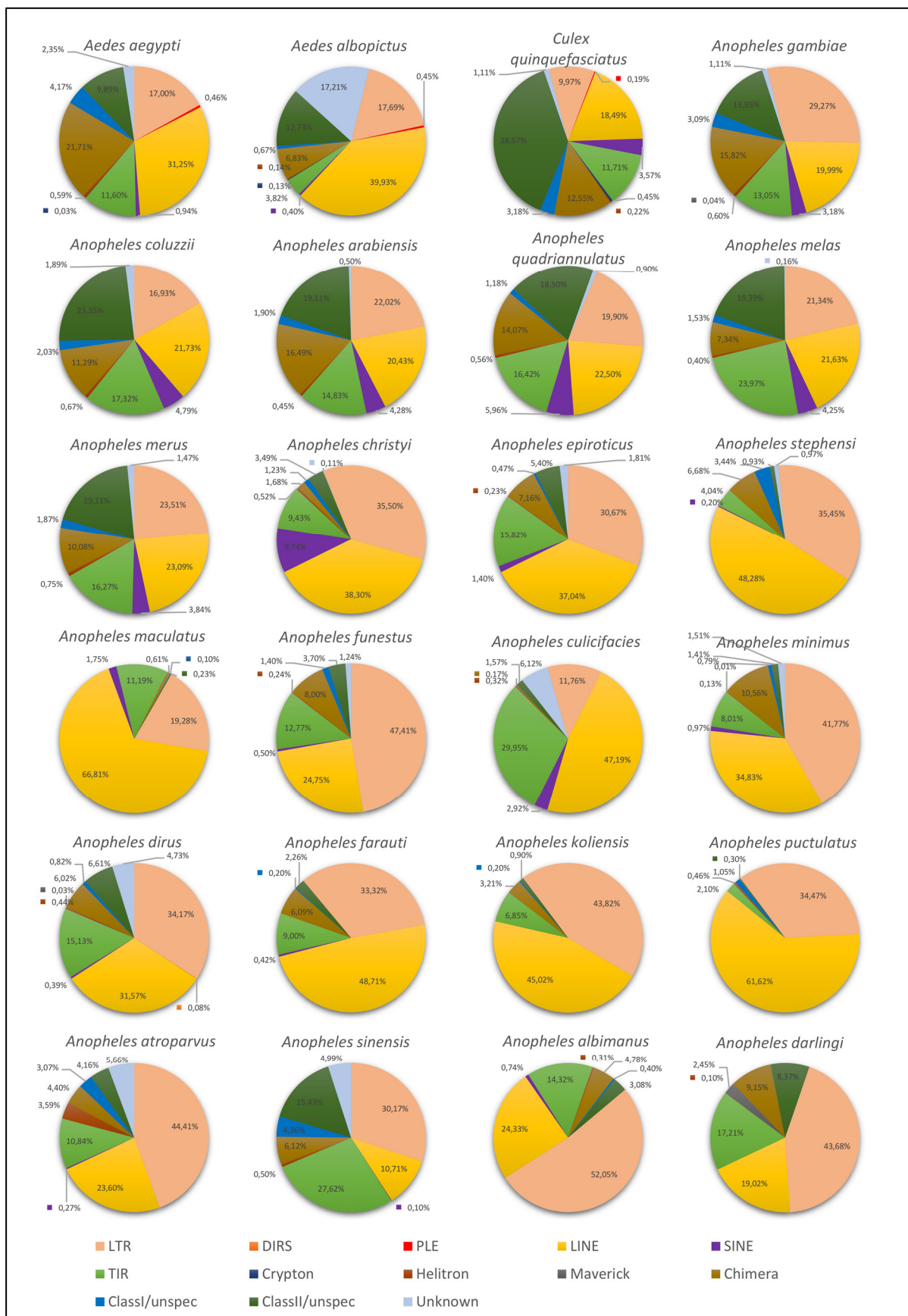
### 6.1.3 Espécies do complexo *An. gambiae* possuem atividades mais recentes de TEs

Para examinar a distribuição de TEs dentro de genomas de espécies ao longo do tempo, estimamos a distância Kimura-2-parâmetros (K2P) de todos os TEs. Agrupamos os TEs nas quatro ordens mais abundantes: LTR, LINE, TIR e SINE. Os *boxplots* na Fig. 20 revelam padrões interessantes: mosquitos do complexo *An. gambiae* tendem a ter famílias de TEs das ordens TIR, LINE e LTR mais recentes do que outros Anophelinae. Valores baixos de K2P nesses seis mosquitos indicam que muitas famílias TE estão atualmente ativas ou estiveram ativas recentemente. Em contraste, a grande maioria das famílias LINE em *An. darlingi*, *An. albimanus*, *An. punctulatus* e *An. koliensis* — que são espécies com teor de TE muito baixo — são antigas e provavelmente não estão mais ativas.

A maioria dos elementos SINE no genoma de mosquitos *Anopheles* parece ter sofrido rajadas de transposição antigas e não parecem mais estar ativos nesses mosquitos (Fig. 20D). Um padrão distinto pode ser observado nas espécies de Culicinae, onde detectamos expansão recente de SINEs. Em geral, o padrão K2P fornece pistas de que os TEs são mais jovens e mais recentemente ativos em Culicinae do que em espécies de Anophelinae fora do complexo *An. gambiae*.

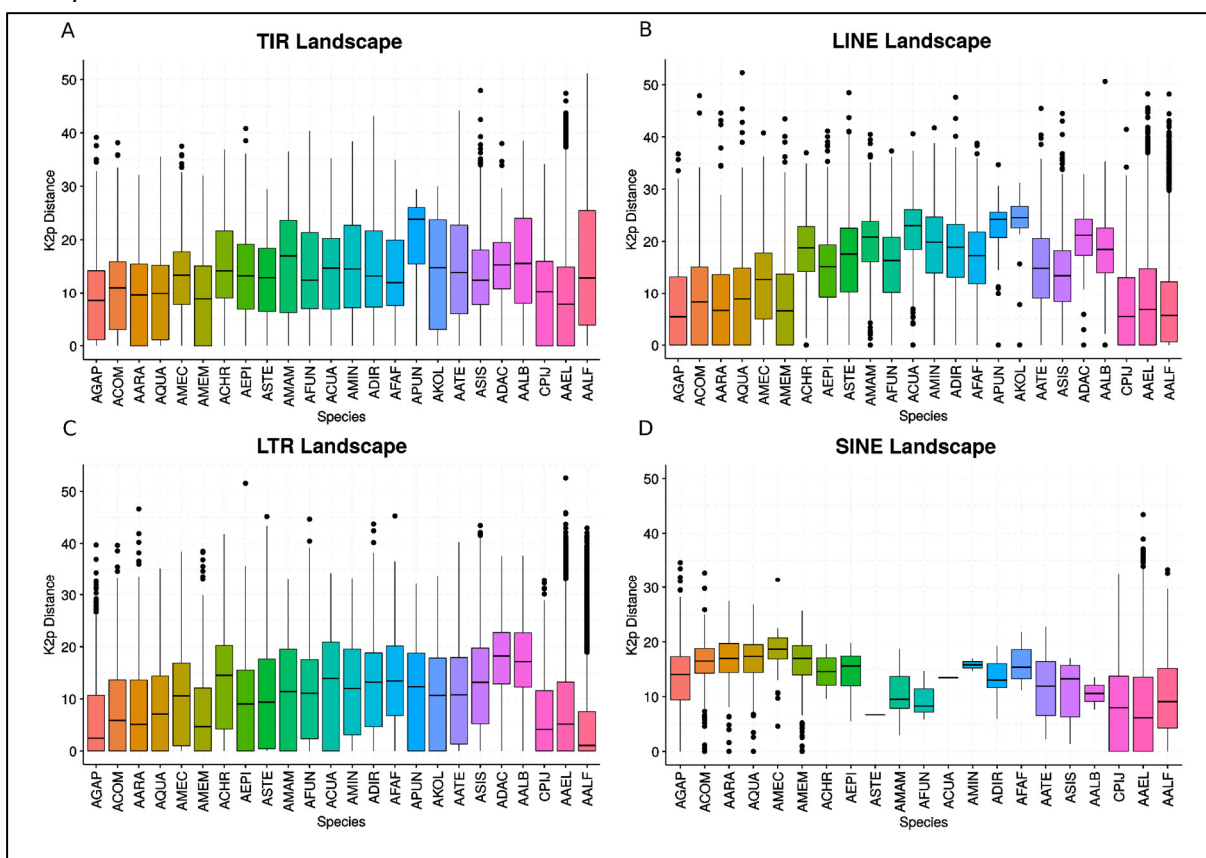


Figura 19 – Proporção do mobiloma com as ordens de acordo com o sistema de classificação de Wicker *et al.* (2007).



Fonte: Elaborado pelo autor.

Figura 20 – Datação intragenômica das quatro ordens TE mais abundantes de genomas de mosquitos.



Fonte: Elaborado pelo autor

Legenda: As colorações variam de acordo com a espécie avaliada. Os nomes das espécies foram transformados em códigos de quatro letras para simplificar o gráfico, o mapeamento de código para nome pode ser encontrado na Tabela 1.

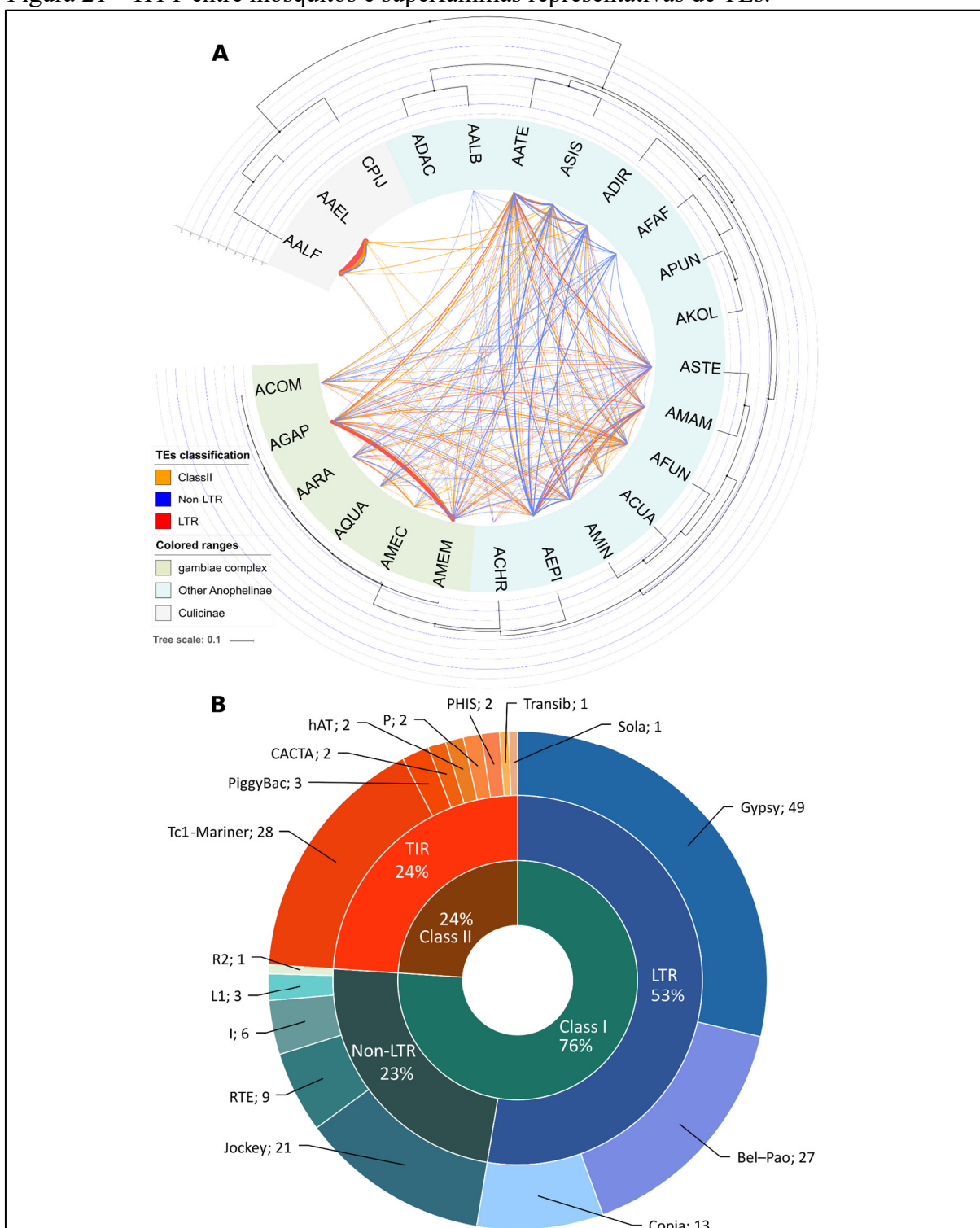
### 6.1.4 Centenas de eventos HTT ocorreram nas espécies de mosquitos

Encontramos 172 famílias TE com sinais significativos de transferência horizontal (HT) entre as espécies estudadas, usando o pacote VHICA. No entanto, é importante notar que esta é certamente uma subestimativa <sup>7</sup> do real número de eventos HTT, uma vez que várias famílias de TEs têm mais de um sinal HTT par a par significativo (Fig. 21 A). Dessa forma, as HTT então descritas aqui em termos mais gerais, mantendo pelo menos um evento HT para cada família de TE, tendo em vista a complexidade dos sinais encontrados e a falta de algoritmos atualmente disponíveis para determinar o cenário HTT mais provável e o número mínimo de eventos para explicar esses sinais. Quatorze famílias adicionais mostraram um sinal HTT positivo. Entretanto, essas famílias não estão presentes em todas as versões de montagem do

<sup>7</sup> Também é importante relatar que esse número não representa uma superestimativa dos eventos de HTT, pois pelo menos um evento deve ocorrer entre duas espécies para que uma HTT seja detectada. E para cada família, as HTTs foram contabilizadas apenas uma vez

genoma de uma espécie (não é possível confirmar esse polimorfismo) ou não foi possível diferenciar o evento HTT de um evento de introgressão genética, uma vez que há formação de híbridos em entre algumas espécies (HTTs entre espécies com histórico de introgressão não foram consideradas).

Figura 21 – HTT entre mosquitos e superfamílias representativas de TEs.



Fonte: Elaborado pelo autor.

**Legenda:** (A) Uma árvore filogenética cujas arestas representam eventos HTT entre duas espécies, a espessura das arestas é proporcional ao número de eventos. Os nomes das espécies foram transformados em códigos de quatro letras para simplificar o gráfico, o mapeamento de código para nome pode ser encontrado na Tabela 1. (B) Classificação das 172 famílias TE envolvidas em HTT entre os mosquitos.

Mais de três quartos dos transposons envolvidos em eventos de HT pertencem a TEs de Classe I, abrangendo oito superfamílias diferentes (Fig. 21B), enquanto 8 superfamílias de TEs de Classe II foram detectadas sofrendo HT. A superfamília com maior número de famílias que realizam HTT é a *Gypsy*, que responde por mais de 28% de todas as transferências, seguida pelos elementos *Tc1-Mariner*. Essa superfamília merece menção especial, pois possui 9 famílias distintas envolvidas em eventos de HT entre quatro ou mais espécies de mosquitos.

Quase 20% dos sinais de HTT detectados (34 famílias) ocorreram apenas entre *Ae. aegypti* e *Ae. albopictus*. Cento e trinta e três famílias de TE envolvidas em eventos de HT foram detectadas entre as espécies do gênero *Anopheles*, representando 18 das 21 espécies. Apenas três espécies de mosquitos deste gênero, *An. darlingi*, *An. punctulatus* e *An. koliensis*, não estiveram envolvidas em eventos HTT. Entre as espécies de *Anopheles*, *An. merus*, uma espécie divergente inicial do complexo *An. gambiae*, esteve envolvida no maior número de HTTs (Fig. 3A). A maioria das espécies do gênero *Anopheles* está envolvida em HTTs de TEs Classe I e Classe II (Tabela 1). Entretanto, em três espécies detectamos apenas HTTs de TEs de Classe I: elementos não-LTR em *An. albimanus* e *An. farauti*, e um não-LTR e um LTR em *An. christyi*.

Com relação às HTTs entre diferentes gêneros de mosquitos, seis eventos foram encontrados. Cinco desses pertencem à superfamília *Tc1-Mariner*, enquanto apenas um R4, um retrotransposon não-LTR, foi transferido de *Ae. albopictus* ao ancestral dos mosquitos do complexo *An. gambiae*, como já descrito em um estudo anterior (BIEDLER; CHEN; TU, 2015). Nenhum evento de transferência horizontal envolvendo *Cx. quinquefasciatus* foi encontrado, possivelmente devido ao seu mobiloma tão distinto das outras espécies avaliadas e uma ausência de outros genomas conhecidos da tribo Culicini. É importante notar que todos os eventos de HTT possuem alta significância mostrando *p-values* muito baixos (Fig. 22). Além disso, foram associados a várias evidências de HTT adicionais como distribuição irregular e incongruências filogenéticas entre TE e a filogenia do hospedeiro, conforme discutido nas próximas seções.

Tabela 1 – Número de famílias TE em cada espécie que está envolvida em eventos de HTT

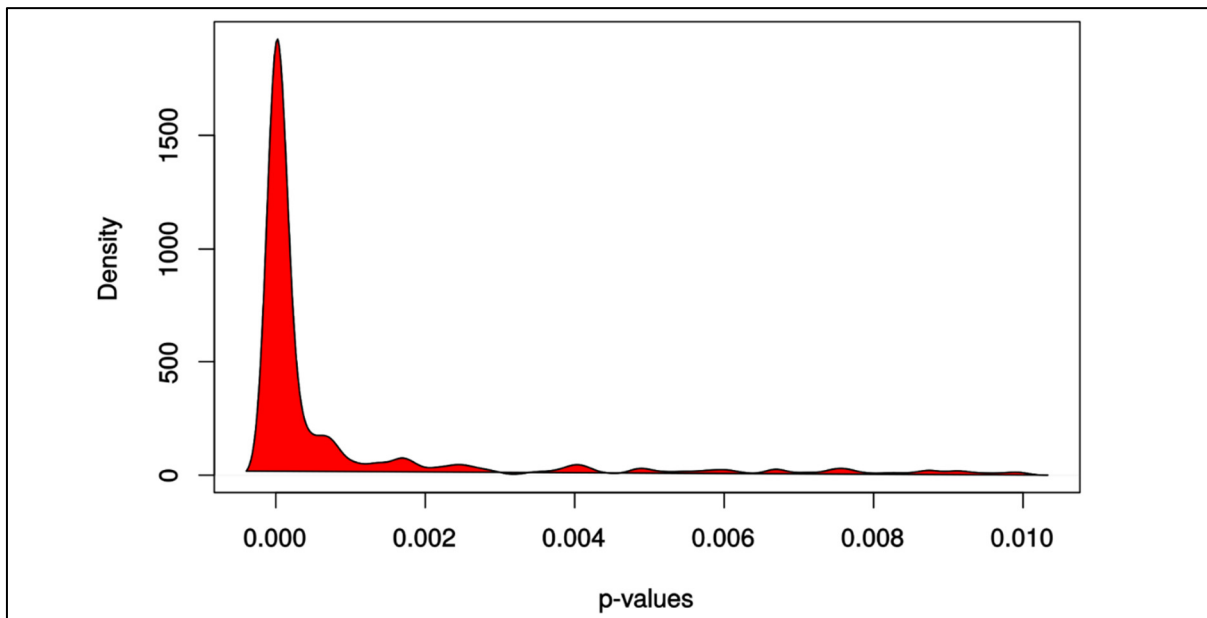
Espécies	Código da espécie	Classe II	Classe I		Total
			Não-LTR	LTR	
<i>Anopheles merus</i>	AMEM	9	13	21	43
<i>Anopheles gambiae</i>	AGAP	10	6	26	42
<i>Aedes aegypti</i>	AAEL	12	6	22	40
<i>Aedes albopictus</i>	AALF	11	7	22	40
<i>Anopheles epiroticus</i>	AEPI	5	15	15	35
<i>Anopheles atroparvus</i>	AATE	9	7	17	33
<i>Anopheles maculatus</i>	AMAM	5	9	12	26
<i>Anopheles sinensis</i>	ASIS	10	7	9	26
<i>Anopheles dirus</i>	ADIR	8	9	6	23
<i>Anopheles minimus</i>	AMIN	7	9	7	23
<i>Anopheles stephensi</i>	ASTE	4	7	12	23
<i>Anopheles arabiensis</i>	AARA	6	5	7	18
<i>Anopheles funestus</i>	AFUN	8	3	7	18
<i>Anopheles coluzzii</i>	ACOM	4	2	5	11
<i>Anopheles melas</i>	AMEC	6	1	3	10
<i>Anopheles culicifacies</i>	ACUA	5	3	1	9
<i>Anopheles quadriannulatus</i>	AQUA	4	3	1	8
<i>Anopheles farauti</i>	AFAF	0	6	0	6
<i>Anopheles albimanus</i>	AALB	0	2	0	2
<i>Anopheles christyi</i>	ACHR	0	1	1	2
<i>Anopheles darlingi</i>	ADAC	0	0	0	0
<i>Anopheles koliensis</i>	AKOL	0	0	0	0
<i>Anopheles punctulatus</i>	APUN	0	0	0	0
<i>Culex quinquefasciatus</i>	CPIJ	0	0	0	0

Fonte: Elaborado pelo autor

Cerca de 10% das redes HTT envolvem pelo menos 4 espécies. Essa proporção é pequena, principalmente entre as famílias LTR, onde quatro ou mais espécies de mosquitos foram detectadas em menos de 10% do total de HTTs envolvendo famílias LTR. De todas as superfamílias estudadas, *Tc1-Mariner* é aquela cujos elementos são mais frequentemente transferidos horizontalmente entre muitas espécies, 9 famílias de TE dessa superfamília se transferiram entre mais de 2 espécies de mosquitos. Detalhes sobre a participação das espécies em cada rede de transferência horizontal da família TE podem ser vistos no arquivo online <https://doi.org/10.1371/journal.pgen.1008946.s002>, que lista todas as imagens de saída do VHICA para casos de HTT positivos. A idade intragenômica relativa de cada família TE submetida a HT por genoma de mosquito é dada na Fig. 23. Percebe-se um perfil bem

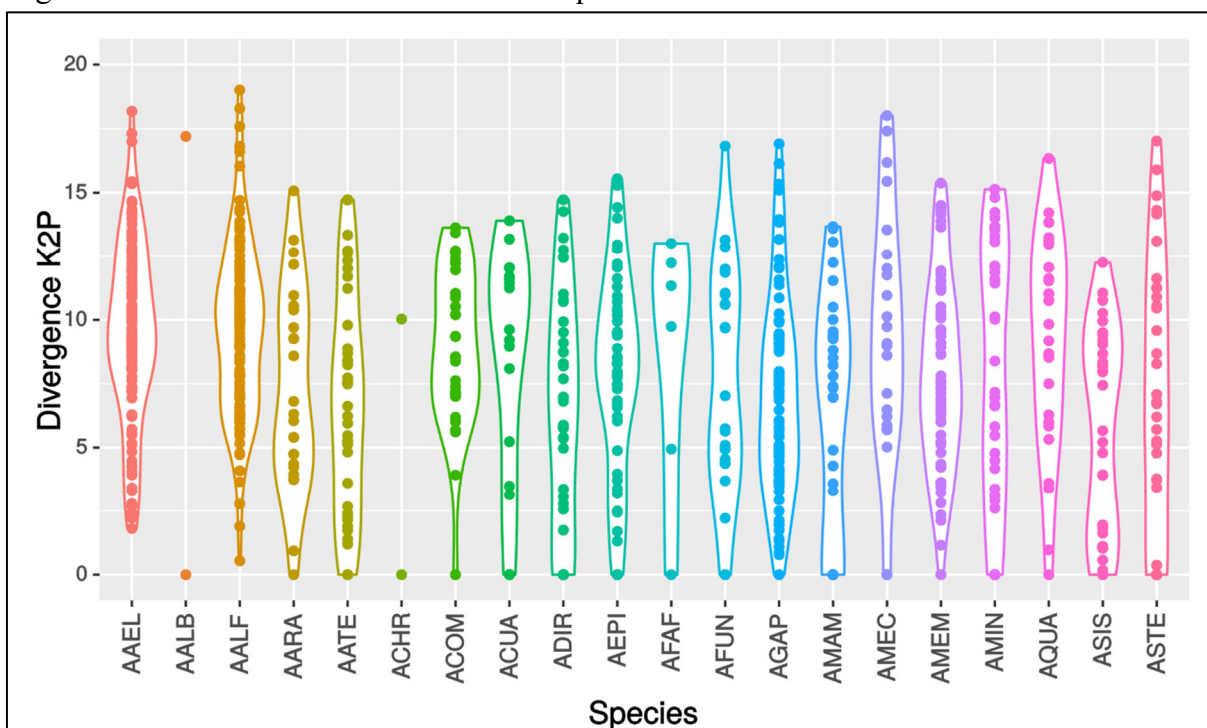
distribuído de idade das famílias, onde tanto famílias bastante antigas quanto famílias recentes estão envolvidas em eventos de HTT.

Figura 22 – Distribuição de valores de P de todas as comparações de HTT positivas par a par.



Fonte: Elaborado pelo autor.

Figura 23 – Perfil de idade das famílias TE que sofreram transferências horizontais.



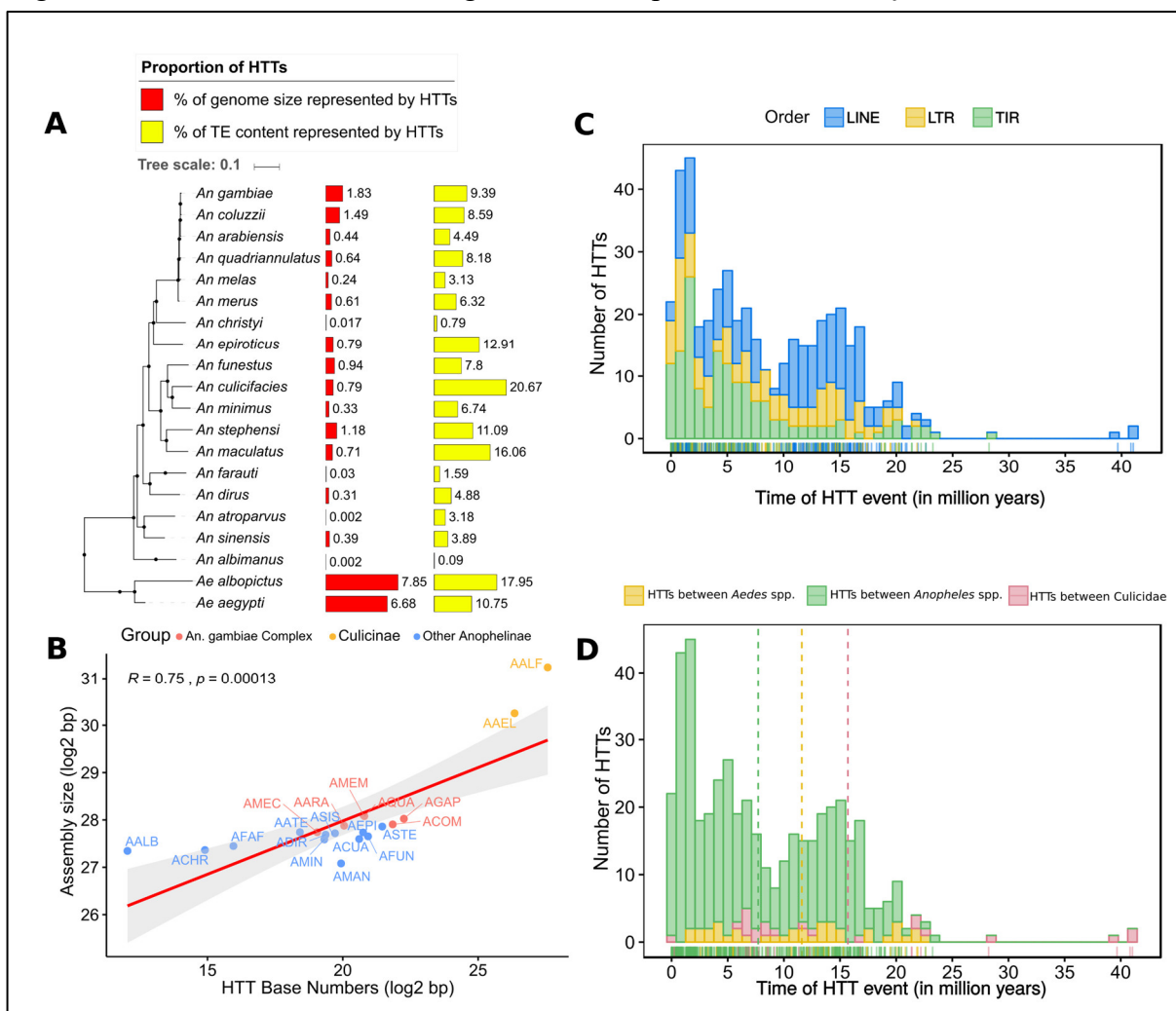
Fonte: Elaborado pelo autor.

Legenda: Cada ponto representa a idade relativa de cada família em um genoma de mosquito. Cada cor representa uma espécie distinta.

### 6.1.5 As transferências horizontais de transposon impactam diretamente no tamanho do genoma dos mosquitos

TEs transferidos horizontalmente entre espécies representam uma fração significativa de alguns genomas de mosquitos (Fig. 24A). Estimamos que cerca de 6,7% de *Ae. aegypti* e 7,8% de *Ae. albopictus* são cobertos por TEs transferidos horizontalmente, o que representa mais de 10% do conteúdo total de TE dessas espécies. No gênero *Anopheles*, o impacto da transferência horizontal no tamanho do genoma foi consideravelmente menor. Apenas *An. gambiae*, *An. coluzzii*, e *An. stephensi* mostraram mais de 1% de seus genomas cobertos com cópias derivadas de TEs transferidos horizontalmente.

Figura 24 – Influência das HTTs no genoma do hospedeiro e suas datações.



Fonte: Elaborado pelo autor.

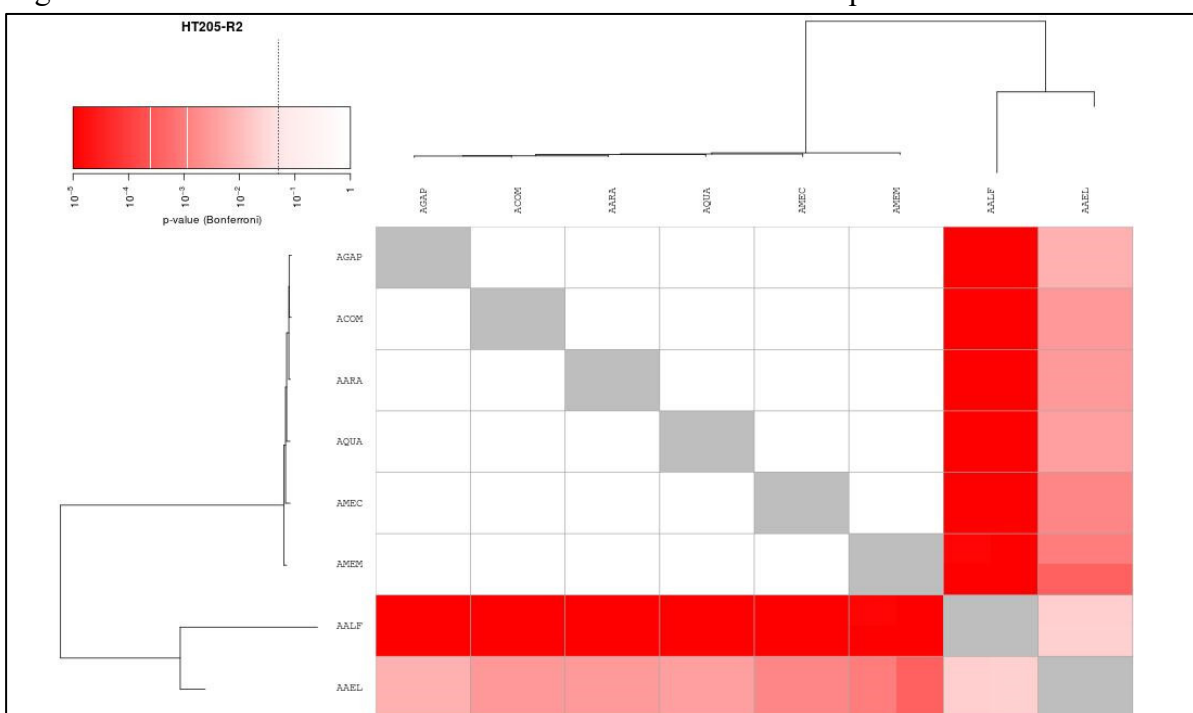
Legenda: (A) A proporção dos genomas do hospedeiro compostos de TEs envolvidos na HT. (B) A correlação entre o número de nucleotídeos cobertos por TEs envolvidos em eventos HTT e o tamanho do genoma dos mosquitos. A idade de cada evento HTT par a par é estimada e representada por grupos TE (C) e pelos táxons das espécies hospedeiras (D).

A proporção do genoma coberto por TEs transferidos horizontalmente mostrou uma correlação positiva com o tamanho do genoma da espécie hospedeira (Fig. 24B). Portanto, a invasão de TE por meio de transferência horizontal pode contribuir significativamente para a expansão genômica e também impactar o tamanho do genoma do mosquito. Assim, quanto mais frequentes os eventos HTT, maior o genoma da espécie.

### 6.1.6 As transferências horizontais de elementos transponíveis ocorreram nos últimos 30 milhões de anos

A datação das HTTs foi estimada com base em todas as 463 comparações significativas pareadas encontradas. A grande maioria dos eventos HTT detectados ocorreu nos últimos 30 milhões de anos (Fig. 24C e 24D), período durante o qual a maioria das espécies de *Anopheles* sofreu especiação (TANG *et al.*, 2018). Apenas uma transferência é mais antiga que esta: uma transferência de um clado R4 envolvendo espécies do *An. gambiae* e ambas as espécies de *Aedes* estudadas que ocorreram há cerca de 40 milhões de anos (Fig. 25). Elementos LINES e LTRs têm o número mais uniforme de HTTs ao longo do tempo, enquanto HTT de elementos TIRs se intensificaram mais recentemente (Fig. 24C). As transferências horizontais que ocorrem entre espécies do gênero *Anopheles* e entre espécies do gênero *Aedes* não mostraram grande diferença no tempo estimado (Fig. 24D).

Figura 25 – Transferência horizontal de um TE do clado R4 da superfamília R2.



**Fonte:** Elaborado pelo autor a partir do software VHICA.

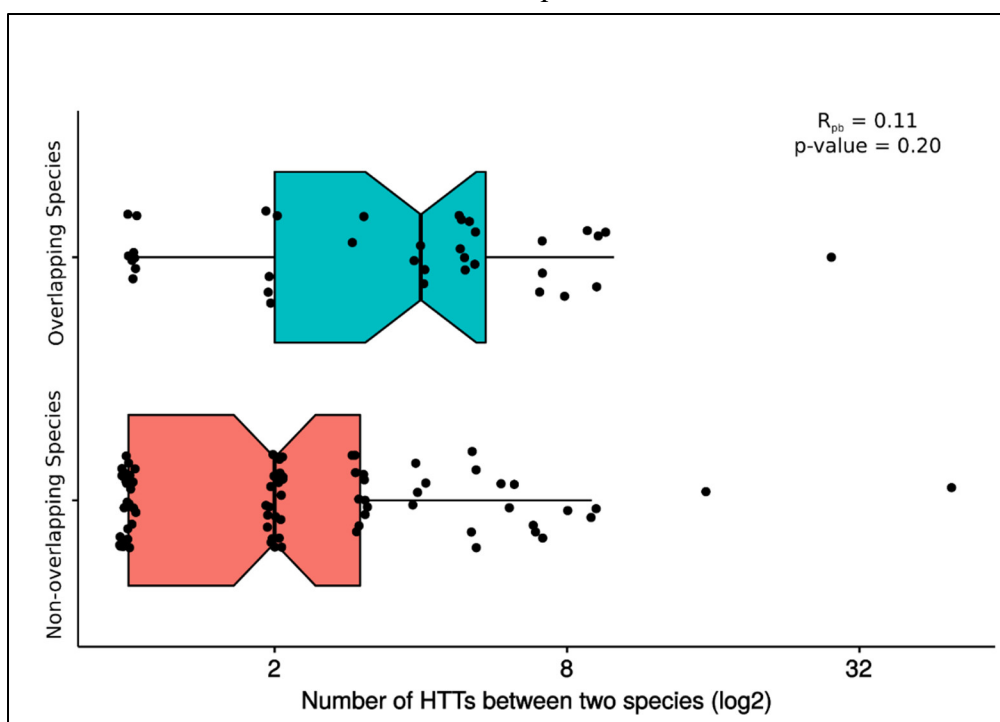
**Legenda:** Cada quadrado vermelho representa uma comparação de pares HTT significativa. A intensidade da coloração é proporcional ao nível de significância.



### 6.1.7 A localização geográfica não afeta o número de transferências entre as espécies de mosquitos

A fim de investigar se a sobreposição geográfica da área de distribuição dos mosquitos favorece a ocorrência de HTTs, comparamos a distribuição de espécies em todo o mundo com o número de HTTs entre cada par de espécies. Nenhuma correlação foi encontrada entre essas variáveis ( $R = 0,1$ ,  $p\text{-value} = 0,2$ ) (Fig. 26). É possível que espécies intermediárias tenham agido como vetores dessas transferências de elementos de transposição entre as espécies de mosquito no passado, ou que outras espécies que mosquitos estejam envolvidas nessa rede de HTTs. Nestes dois casos os genomas dessas espécies não seriam conhecidos até o atual momento.

Figura 26 – Falta de correlação entre sobreposição geográfica e número de casos de transferência horizontal em mosquitos.



Fonte: Elaborado pelo autor.

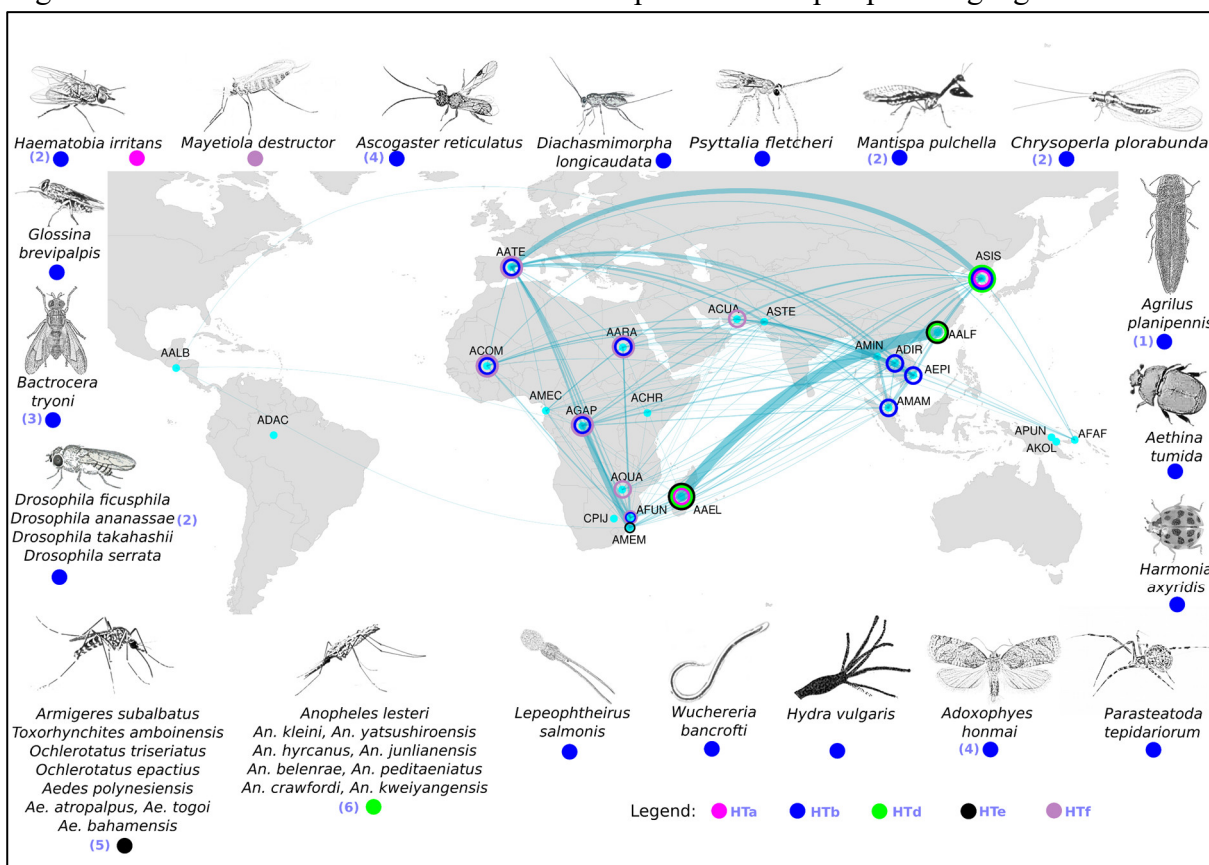
Legenda: Cada ponto representa o número de transferências horizontais entre as espécies.

### 6.1.8 Os elementos *Tc1-Mariner* foram transferidos horizontalmente para espécies eucarióticas distantemente relacionadas

Foi verificado, através de mapeamento dos TEs de mosquitos contra genomas de milhares de espécies, se alguns dos elementos transponíveis que são transferidos horizontalmente entre mosquitos também podem estar envolvidos em transferências para outros eucariotos ou vírus. Cinco famílias de TEs da superfamília *Tc1-Mariner* mostraram uma rede de transferência

horizontal mais ampla com outras espécies de eucariotos. Três delas já foram descritas em estudos anteriores, mostrando transferências entre pelo menos duas espécies (Fig. 27). Os atuais resultados expandem ainda mais essa rede, adicionando vinte novas espécies. Os outros eventos de HTT nas duas famílias TE restantes são descritos aqui pela primeira vez.

Figura 27 – Rede de transferências horizontais a partir de uma perspectiva geográfica.



**Fonte:** Elaborado pelo autor

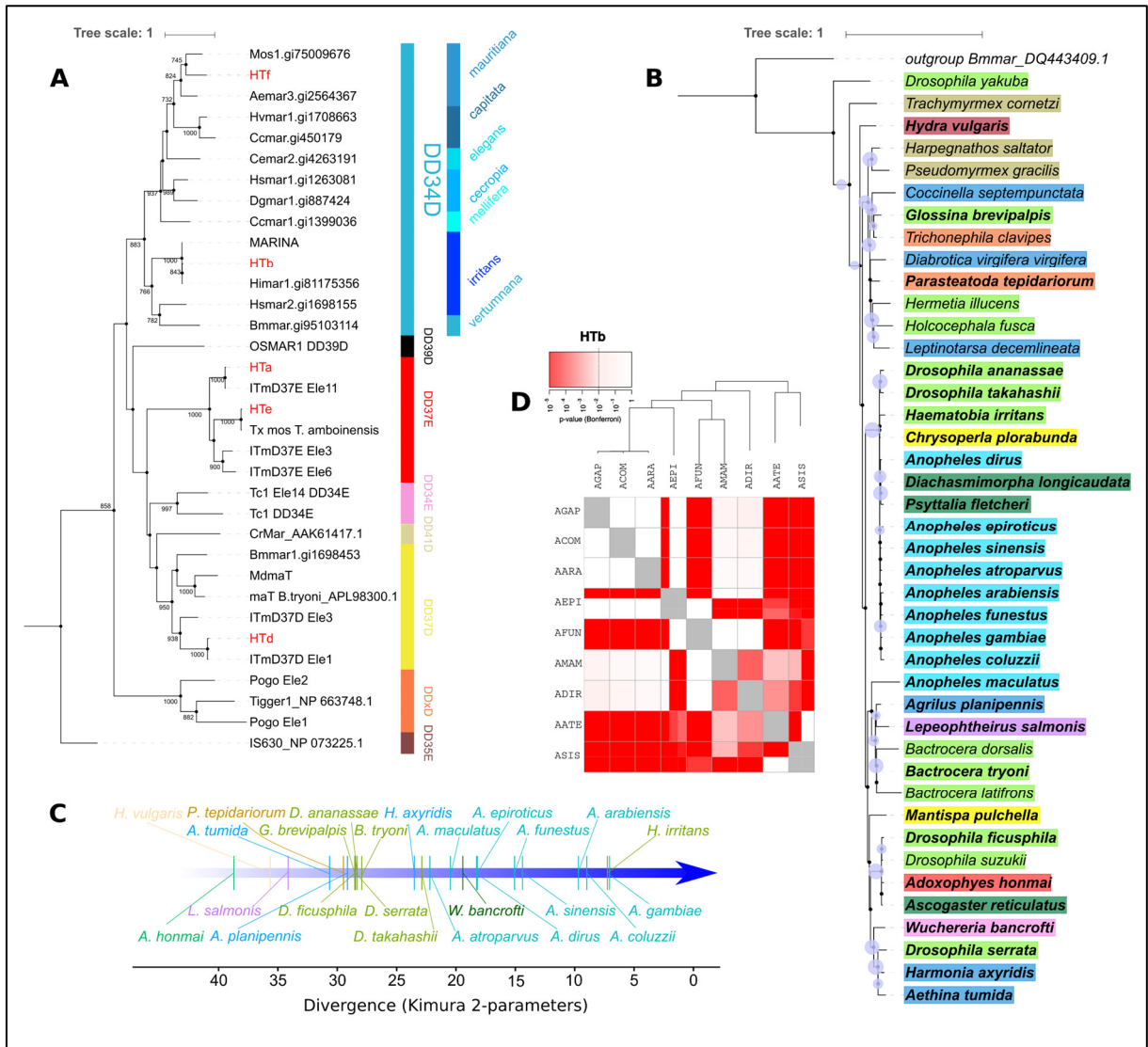
**Legenda:** Os pontos azuis claros representam o local de amostragem das espécies para os genomas usados neste estudo ou o local mais provável de origem de uma espécie. As arestas conectando os pontos representam eventos HTT par a par significativos, a espessura dessas arestas é proporcional ao número de eventos HTT. Cinco dos 172 TEs são transmitidos horizontalmente para outras espécies fora do grupo de 24 espécies estudadas, essas famílias TE são representadas por círculos coloridos: HTa (rosa), HTb (azul), HTd (verde), HTe (preto) e HTf (claro roxa). As espécies envolvidas em uma transferência horizontal com pelo menos um dos 24 mosquitos são mostradas na borda do mapa. O círculo colorido ao redor dos pontos azuis claros indica que esta espécie de mosquito está envolvida em uma rede HTT de longo alcance de pelo menos um TE específico (HTa a HTf). Os números entre parênteses representam algumas evidências preliminares de HTTs envolvendo esses TEs e espécies em outros artigos publicados. 1 representa o elemento APmar1 já descrito por Rivera-Veja e Mittapalli (2010), 2 representa transferências HTT entre espécies descritas por Robertson e Lampe (1995), 3 representa um elemento descrito por Green e Frommer (2001), 4 representa transferência HTT entre *Ascogaster reticulatus* e *Adoxophyes honmai* (YOSHIYAMA *et al.*, 2001), 5 representa eventos HTT já descrito em mosquitos (BIEDLER; SHAO; TU, 2007), 6 representa eventos HTT já descritos em outras espécies de Anophelinae (DIAO *et al.*, 2011) não investigadas neste estudo.

As espécies envolvidas nos eventos HTT pertencem a três filos de metazoários. O primeiro filo foi Arthropoda, que abrange a maioria das espécies. É representado pela classe Arachnida, envolvendo a aranha doméstica *Parasteatoda tepidariorum*, pela classe Hexanauplia, representada pelo copépode *Lepeophtheirus salmonis*, e pela classe Insecta, que inclui as espécies de mosquitos envolvidas nas HTTs. O segundo filo foi Cnidaria, com apenas um representante, *Hydra vulgaris*. O terceiro foi Nematoda, onde encontramos um caso de transferência horizontal envolvendo o verme parasita *Wuchereria bancrofti* e outras espécies de *Anopheles* que transmitem esse nematoide para humanos. Além da alta identidade entre os TEs de mosquitos e os de espécies distantes, às vezes separados por mais de 700 milhões de anos, também observamos uma distribuição desigual desses elementos, fornecendo suporte adicional para os eventos de transferência horizontal.

Duas dessas linhagens de TEs (HTa e HTe) pertencem ao grupo ITmD37E, que inclui TEs que possuem o motivo DD37E em suas sequências de transposase (Fig. 28A). Uma linhagem (HTd) agrupada com elementos descritos anteriormente da família ITmD37D, também conhecida como família *mat*. As outras duas linhagens (HTb e HTf) são da família DD34D ou *Mariner*. Os elementos da rede HTf pertencem à subfamília *mauritaniana*, e os elementos da rede HTb, que estão envolvidos na maioria dos eventos HTT com espécies fora da família Culicidae, pertencem à subfamília *irritans*.

Para obter um entendimento mais aprofundado da rede HTb, realizamos uma segunda busca de BLAST usando como *query* todas as sequências recuperadas da busca anterior, contra genomas de espécies fora da família Culicidae. Encontramos elementos relacionados à HTb presentes em 42 espécies de invertebrados divergentes, incluindo *W. bancrofti* (Fig. 28B). A datação intragenômica dos TEs transferidos horizontalmente permitiu uma interpretação mais precisa dos eventos de HTT, incluindo a fonte e a espécie receptora. A idade intragenômica relativa dos TEs mostrou que os elementos mais antigos pertencem à espécie de Lepidoptera *Adoxophyes honmai*. A maioria dos elementos das moscas são antigos e têm idades semelhantes. Os elementos dos mosquitos asiáticos (*An. maculatus*, *An. epiroticus*, *An. dirus*, *An. sinensis*) são mais antigos que os dos africanos (*An. gambiae*, *An. coluzzii*, *An. arabiensis*, *An. funestus*). Curiosamente, a idade do elemento *W. bancrofti* é muito semelhante à dos mosquitos asiáticos, sugerindo que *W. bancrofti* adquiriu o elemento das espécies asiáticas de *Anopheles* e doou para as espécies africanas de *Anopheles* mais recentemente (Fig. 28C). Outra evidência de que esse elemento é propenso a transpor horizontalmente é fornecida pelo sinal significativo de HTT múltiplo entre os mosquitos (Fig. 28D).

Figura 28 – Transposons transmitidos horizontalmente para outras espécies.



Fonte: Elaborado pelo autor.

**Legenda:** (A) Uma árvore filogenética baseada nas sequências de aminoácidos da transposase de vários elementos da superfamília ITM (valores de bootstrap são mostrados nos nós), incluindo os cinco TEs de Anophelinae (marcados em vermelho) envolvidos em eventos HTT com outras espécies de metazoários e muitos elementos que já foram descritos em estudos anteriores. A topologia em árvore indica que esses cinco TEs pertencem a famílias diferentes. HTb representa um elemento que é difundido em muitos genomas de metazoários, como mostrado na árvore filogenética baseada em nucleotídeos da parte B. O nome de espécie em negrito representa elementos que compartilham mais de 80% de semelhança com os elementos HTb de *Anopheles*. Brachycera é destacado em verde claro, cnidários em marrom escuro, formigas em marrom claro, aranhas em laranja, besouro em azul, elementos do gênero *Anopheles* em azul claro, vespas parasitóides em verde escuro, Neuroptera em amarelo, Lepidoptera em vermelho, crustáceos em roxo, nematoides em roxo claro. (C) Uma linha do tempo construída usando distâncias K2P mostrando a idade relativa média dos elementos no genoma de cada espécie. (D) Representação gráfica VHICA do elemento HTb em mosquitos. As caixas de cor vermelha representam sinais HTT estatisticamente significativos.

## 6.2 RESULTADOS DAS ANÁLISES FUNCIONAIS

Essa seção detalha os resultados das etapas que foram realizadas apenas para as espécies que possuíam algum dado de expressão gênica.

### 6.2.1 Centenas de TEs estão sobrepostos a genes nos mosquitos

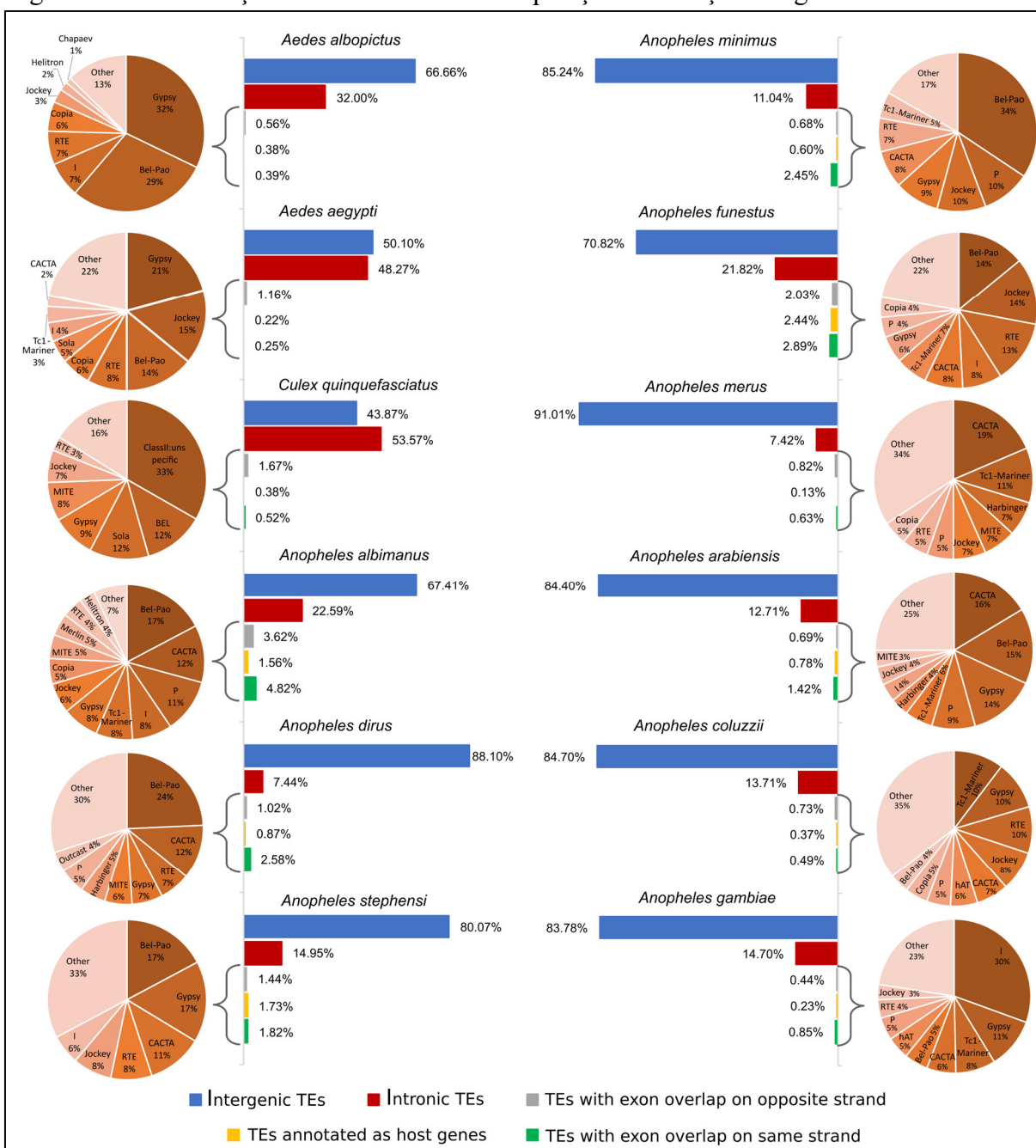
O primeiro passo para verificar os possíveis impactos dos elementos de transposição nos mosquitos, foi verificar se esses TEs se sobrepõem com estruturas genômicas codificantes, ou não codificantes, mas que são expressas em nível de RNA. Dentre essas estruturas, as mais conhecidas e estudadas são os genes. Dessa forma, comparou-se a anotação de TEs de cada espécie com a anotação de genes preditos, disponibilizada nas bases de dados online, para cada espécie de mosquito.

As análises mostraram, como esperado com base em dados de estudos em outras espécies, que a maior parte dos elementos de transposição encontram-se localizados em regiões intergênicas (Fig. 29). A única exceção é o *Cx. quinquefasciatus*, onde há mais TEs na região intrônica, que TEs na região intergênica. Inclusive, em todos os mosquitos da subfamília Culicinae, o número de TEs nos íntrons de genes é bem maior que em mosquitos da subfamília Anophelinae.

Alguns TEs possuem sobreposição total aos genes preditos. Esses casos respondem por menos de 1% dos TEs na maior parte das espécies, e indica que elementos de transposição estão anotados no arquivo de anotação gênica, provavelmente por possuírem a maior parte das características necessárias para sua expressão. No entanto, a maior parte deles está apenas listada como gene desconhecido, o que comprova o déficit das bases de dados tradicionais em realizar a anotação de TEs.

Interessantemente, uma pequena fração de elementos de transposição possui alguma sobreposição parcial com genes. Em várias espécies esse valor ultrapassa os 2% das cópias de elementos de transposição, sendo maior em *An. albimanus* (onde o número de cópias de TEs é muito pequeno) e menor em *Ae. albopictus*. Essas sobreposições com genes são divididas em dois tipos: I) sobreposições onde um gene e um fragmento de um elemento de transposição estão localizados em fitas opostas; II) sobreposições onde um gene e um fragmento de um elemento de transposição estão localizados na mesma fita. Neste último caso há a possibilidade de formação de transcritos quiméricos, onde parte do transcrito deriva da região gênica, e parte deriva da região do elemento de transposição, que pode estar passando por um processo de domesticação.

Figura 29 – Distribuição dos elementos de transposição em relação aos genes.



Fonte: Elaborado pelo autor

Legenda: Os gráficos de barras ao centro indicam a proporção das cópias de elementos de transposição que são distribuídas em cada categoria. Os gráficos de setor nas bordas da figura mostram as proporções das principais superfamílias de elementos de transposição que possuem algum tipo de sobreposição, total ou parcial, com regiões gênicas anotadas.

Dentre todos os genomas de mosquito, o genoma de *An. gambiae* é o mais bem anotado por ser estudado há quase duas décadas. Mesmo assim há 1,29% das cópias de elementos de transposição sobrepondo genes em alguma região. Isso pode ser utilizado como um parâmetro que indica fortemente que as sobreposições são elementos reais do genoma e não erros de anotação gênica. Os resultados da análise realizada ainda não são suficientes para indicar a

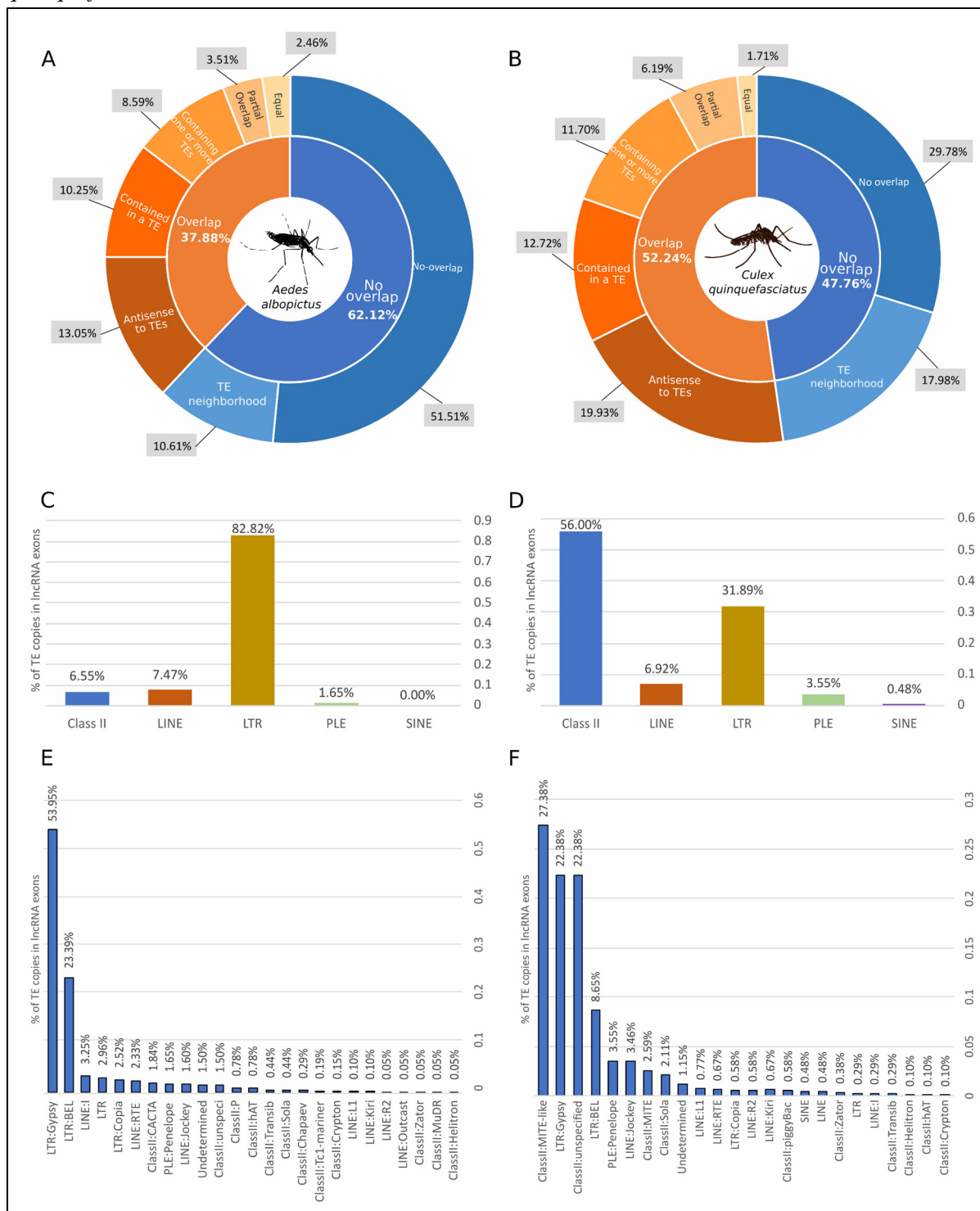
formação de transcritos quiméricos, mas foram capazes de gerar uma lista de genes que devem ser estudados com maior atenção no futuro, usando abordagens de análise de sequenciamento de RNA.

### 6.2.2 Os lncRNAs de algumas espécies de mosquitos são enriquecidos de TEs

Assim como estar sobrepondo estruturas gênicas codificantes, os TE também podem sobrepor e originar outras estruturas, como os RNAs longos não codificantes (lncRNA). Estudar os lncRNAs nos mosquitos, contudo, pode ser um desafio, pois para a maior parte das espécies há uma pobre descrição desses elementos. Para essa análise foram escolhidas duas espécies, *Ae. albopictus* e *Cx. quinquefasciatus*, cujo conjunto de lncRNAs foi sistematicamente caracterizado nos últimos anos (AZLAN *et al.*, 2021a, 2021b). Evitou-se o uso de outras espécies com lncRNAs menos padronizados, como forma de evitar a inserção de vieses.

Foi encontrada uma grande variabilidade na coocorrência de lncRNAs e TEs: I - um ou vários TEs podem ser inseridos em íntrons ou éxons de um único gene de lncRNA; II - sequências TE podem sobrepor junções íntron/éxon de um gene lncRNA ou III - um lncRNA único ou multiéxon pode ser inteiramente englobado por coordenadas TE. Descobrimos que 37,88% de todos os genes de lncRNAs de *Ae. albopictus* têm algum grau de sobreposição com elementos transponíveis (Fig. 30A). Enquanto esta proporção varia em ambas as versões de montagem de *Cx. quinquefasciatus*, de 32% em CpipJ2 a 52,24% em VPISU\_Cqui (Fig. 30B). Como esta última apresenta melhor contiguidade, ela foi escolhida como modelo de análise. Além disso, é interessante ressaltar que 10,61% e 17,98% dos loci de lncRNAs são expressos na vizinhança de um TE (distância máxima de 2 kb entre um locus de lncRNA e uma cópia de TE) em *Ae. albopictus* e *Cx. quinquefasciatus* respectivamente. Embora não sobreposta a uma região TE, a expressão desses lncRNAs pode ser modulada pela atividade transcricional dos TEs circundantes (LEE; ZHANG; KRAUSE, 2019).

Figura 30 - Sobreposição de lncRNAs e elementos transponíveis em *Aedes albopictus* e *Culex quinquefasciatus*.



Fonte: Elaborado pelo autor.

Legenda: Coocorrência de lncRNAs e elementos transponíveis em *Aedes albopictus* (A) e *Culex quinquefasciatus* (B). As sobreposições foram separadas em cinco tipos: sobreposição de lncRNAs com TEs, mas na fita oposta; lncRNA contendo um ou mais TEs inseridos em suas coordenadas; lncRNA apresentando alguma sobreposição parcial com TEs, neste caso, a interseção ocorre em 5' ou 3' do lncRNA; genes dos lncRNAs totalmente derivados de uma inserção de cópia TE, onde alguns representam o comprimento total do TE e outros são formados por uma região específica de um TE. A proporção de éxons de lncRNA derivados de TEs é mostrada pelas ordens TE em C e D para *Ae. albopictus* e *Cx. quinquefasciatus*, respectivamente. Essa proporção indica a prevalência de TE Gypsy em *Ae. albopictus* (E) e *Cx. quinquefasciatus* (F).



LncRNAs que ocorrem em íntrons de genes codificadores (intrônicos) e entre genes (intergênicos) também foram avaliados separadamente. A diferença na participação de TE na biogênese desses dois tipos de lncRNA foi muito pequena. Em *Ae. albopictus*, 39,34% dos lncRNAs intrônicos e 37,05% dos lncRNAs intergênicos se sobrepõem aos TEs. Essa variação é ainda menor em *Cx. quinquefasciatus* onde 52,86% e 52,36% dos lncRNAs intrônicos e intergênicos coocorrem com TEs, respectivamente.

Também foi avaliado se a interseção entre genes de lncRNA e TEs ocorre com maior frequência em íntrons ou éxons de lncRNAs. Observou-se que 77,56% e 85,17% dos genes de lncRNAs que apresentam algum grau de sobreposição com TEs possuem pelo menos um éxon envolvido com a inserção de TEs em *Ae. albopictus* e *Cx. quinquefasciatus* respectivamente. Isso representa 29,37% e 44,49% de todos os lncRNAs expressos nestas espécies, respectivamente.

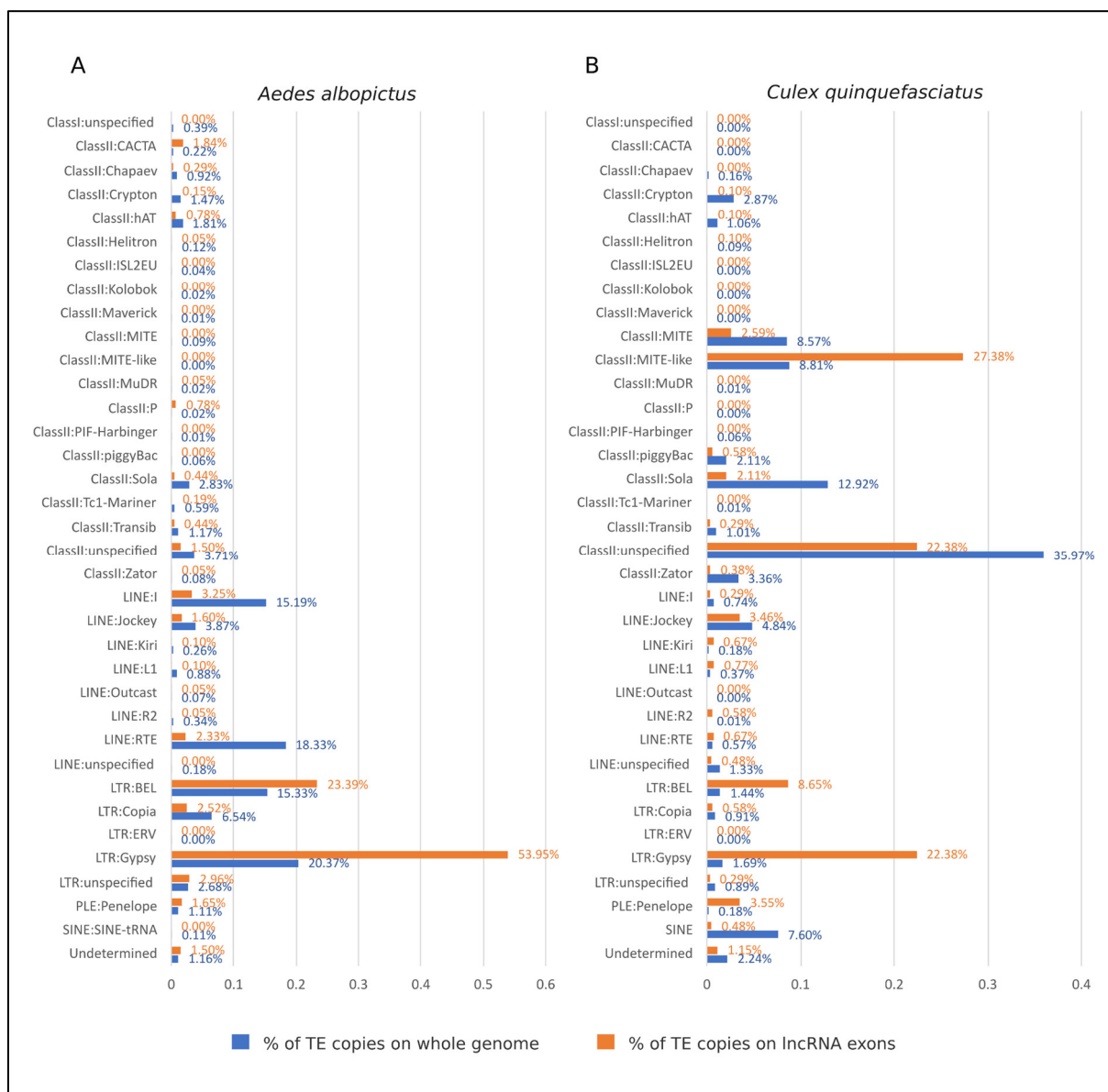
Curiosamente, uma fração significativa de cópias TE que se sobrepõem com genes de lncRNAs é derivada de retrotransposons LTR. Eles representam 82,82% e 31,89% dos éxons de lncRNA nas espécies *Ae. albopictus* e *Cx. quinquefasciatus* (Fig. 30C e Fig. 30D). Dentre esses retrotransposons, destaca-se a superfamília *Gypsy*. Em *Ae. albopictus*, a superfamília *Gypsy* sozinha é responsável por mais da metade de todos os éxons de lncRNAs derivados de TEs (Fig. 30E), e quase um quarto em *Cx. quinquefasciatus* (Fig. 30F).

Surpreendentemente, a proporção das superfamílias de TEs em éxons de lncRNAs não representa a proporção na qual as cópias das superfamílias TE são encontradas no mobiloma (teste qui-quadrado  $\chi^2$ , valor  $P < 2.2e-16$ ) (Fig. 31).

Realizando uma análise de enriquecimento foi observado que algumas superfamílias TE são enriquecidas em éxons lncRNA de ambas as espécies (Fig. 32B e Fig. 32D). Os elementos da superfamília *Gypsy* representam novamente um dos casos mais interessantes. Esta superfamília mostra o *geneRatio* mais alto e os valores *p* mais baixos na análise de super-representação entre todas as superfamílias TE de ambas as espécies (elementos semelhantes a MITE não são uma superfamília, em vez disso são elementos heterogêneos geralmente derivados de transposons de DNA que perderam sua região de codificação e são flanqueados por repetições invertidas terminais – TIRs). Os elementos *Gypsy* representam 53,95% dos éxons de lncRNA que se sobrepõem aos TEs, mas apenas 20,37% de todas as cópias TE encontradas no genoma de *Ae. albopictus* (Fig. 32A). Esta discrepância aumenta substancialmente em *Cx.*

*quinquefasciatus* onde *Gypsy* representa 22,38% dos éxons lncRNA, mas apenas 1,69% das cópias TE no genoma, uma diferença de 13 vezes (Fig. 32C).

Figura 31 – Diferentes distribuições de superfamílias TE no genoma e em éxons de lncRNAs.



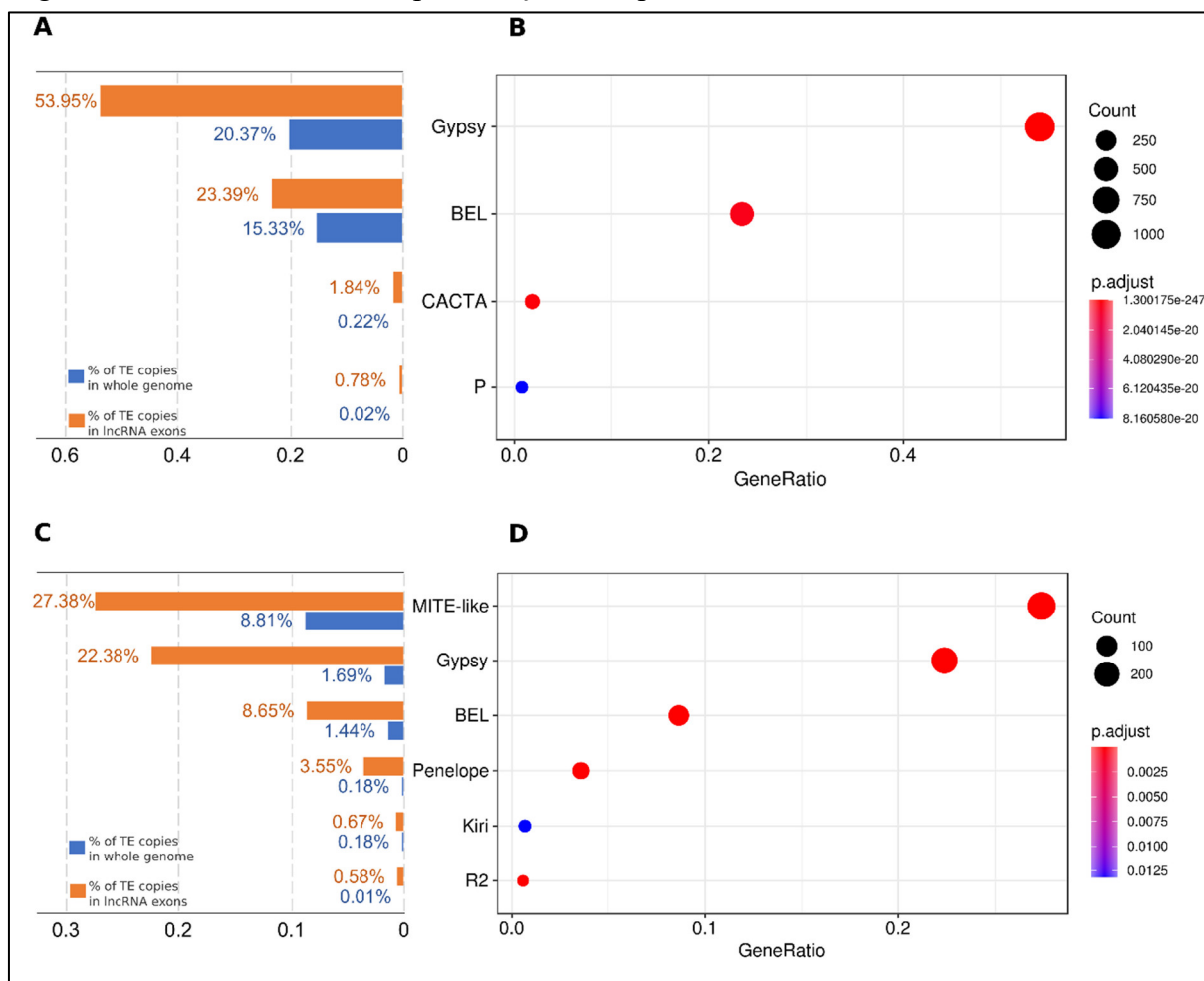
Fonte: Elaborado pelo autor.

**Legenda:** As barras laranjas representam a fração de cópias TE que se sobrepõem aos éxons de lncRNAs por cada superfamília TE (Nº de cópias por cada superfamília em éxons de lncRNAs/Nº total de cópias que se sobrepõe a éxons de lncRNAs), as barras azuis representam a fração de todas as cópias genômicas TE por cada superfamília (Número de cópias por cada superfamília/Número total de cópias do mobiloma).

Foi observado que cinco de treze lncRNAs que regulam a expressão de genes relacionados ao sistema imunológico envolvidos na defesa antiviral de arbovírus (AZLAN *et al.* 2021) são compostos por sequências de TEs. Dois lncRNAs têm um ou mais TEs dentro de seus íntrons (XR\_003898565.1, XR\_003896231.1). Os três restantes estão totalmente inseridos dentro dos limites das cópias dos TEs, duas delas na mesma fita (XR\_003894180.1,

lncRNA\_5135.1) e a outra na fita oposta (XR\_002127134.2). Esses três lncRNAs são derivados de retroelementos LTR (*Gypsy*, *Bel-Pao* e *Copia*). Além disso, um dos 13 também apresenta um LTR em sua região flanqueadora.

Figura 32 – Análise de sobre-representação de superfamílias TE em éxons de lncRNAs.



Fonte: Elaborado pelo autor

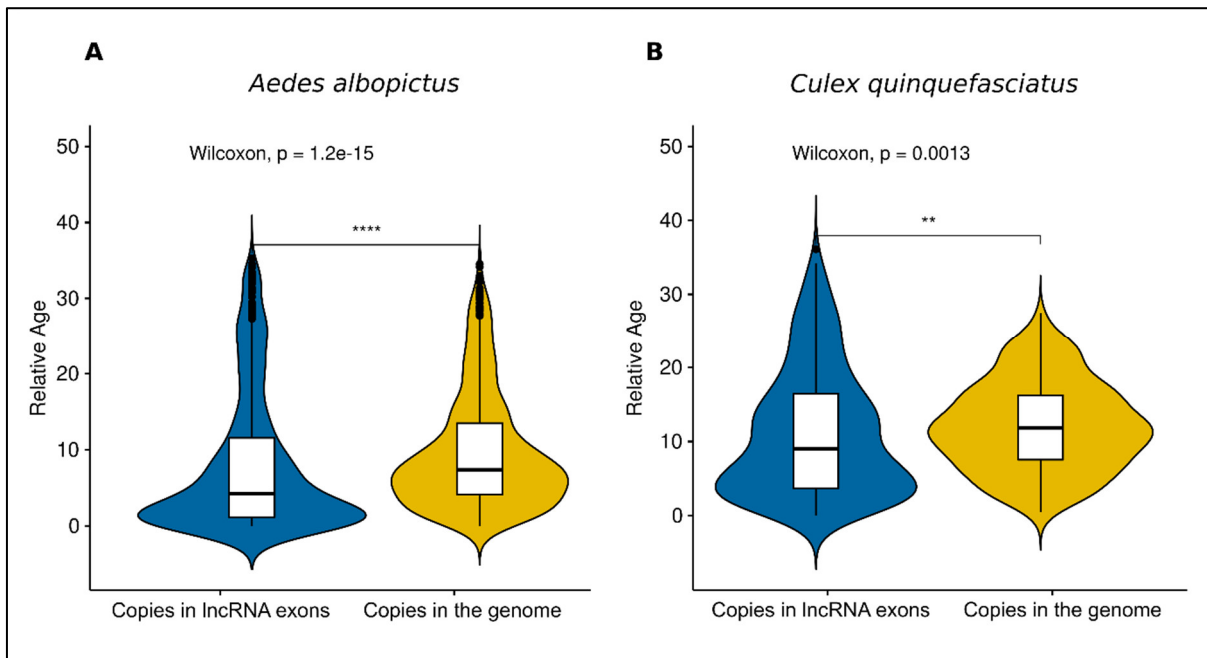
**Legenda:** Distribuição das superfamílias que são enriquecidas em *Ae. albopictus* (A) e *Cx. quinquefasciatus* (C). Os resultados da análise de super-representação mostram que *Gypsy* e *BEL* (ambos retrotransposons LTR) são as superfamílias mais enriquecidas tanto em *Ae. albopictus* (B) quanto em *Cx. quinquefasciatus* (D). Os elementos do tipo MITE não são uma superfamília TE, pois são um grupo heterogêneo de elementos defeituosos que possuem TIRs.

### 6.2.3 Cópias de TEs nos lncRNAs são mais jovens

Em média, as cópias de TEs em éxons de lncRNAs são mais semelhantes à sequência consenso TE (cópias mais jovens) do que as cópias TE que não ocorrem concomitantemente com lncRNAs (Fig. 33) para ambas as espécies *Ae. albopictus* e *Cx. quinquefasciatus*. Cerca de

69% e 60% das cópias TE envolvidas com éxons de lncRNA são mais jovens do que as demais cópias TE da mesma família dispersas no genoma dessas espécies.

Figura 33 – Idade relativa para cópias de TE dentro e fora de éxons de lncRNAs



**Fonte:** Elaborado pelo autor.

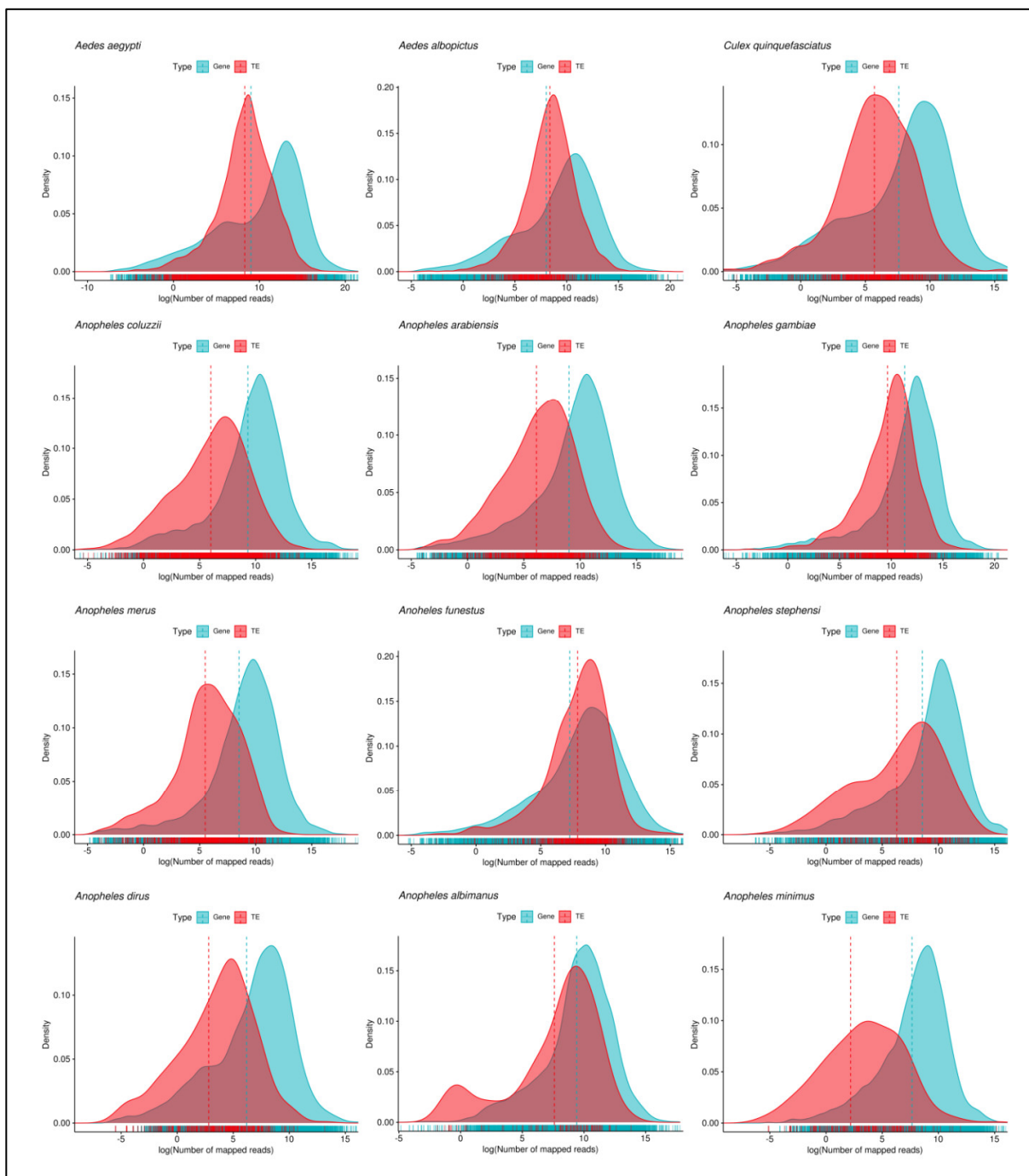
**Legenda:** A idade relativa foi calculada usando a divergência blastn das cópias para cada família TE contra o consenso da família TE. \*\* significa um valor  $p < 0,01$ , \*\*\*\* significa um valor  $p < 0,0001$ .

#### 6.2.4 Diversas famílias de TEs são expressas nos mosquitos

Foram analisadas centenas de corridas de RNA-Seq de 12 espécies de mosquitos (Fig. 34) que possuíam experimentos depositados em bases de dados públicas. Todas as 12 espécies analisadas possuem alguma família de TE expresso.

Um padrão comumente observado durante a análise dos dados de RNA-Seq de diferentes condições experimentais, é que para a maior parte das espécies de mosquito a média do número de *reads* (normalizado por CPM) que mapeiam nos elementos de transposição é menor que a média normalizada das *reads* que mapeiam nos genes, exceto *Ae. albopictus* e *An. funestus*. A figura 34 também deixa claro que a distribuição das *reads* mapeadas em TEs está sempre mais à esquerda da distribuição das *reads* mapeadas nos genes não relacionados aos elementos de transposição. Em conjunto, esses dados significam que no geral as famílias de TEs são proporcionalmente menos expressas que os genes dos mosquitos, embora indique a existência de famílias de elementos de transposição que são bastante expressas.

Figura 34 – Distribuição do número de *reads* mapeadas por genes e TEs dos mosquitos.

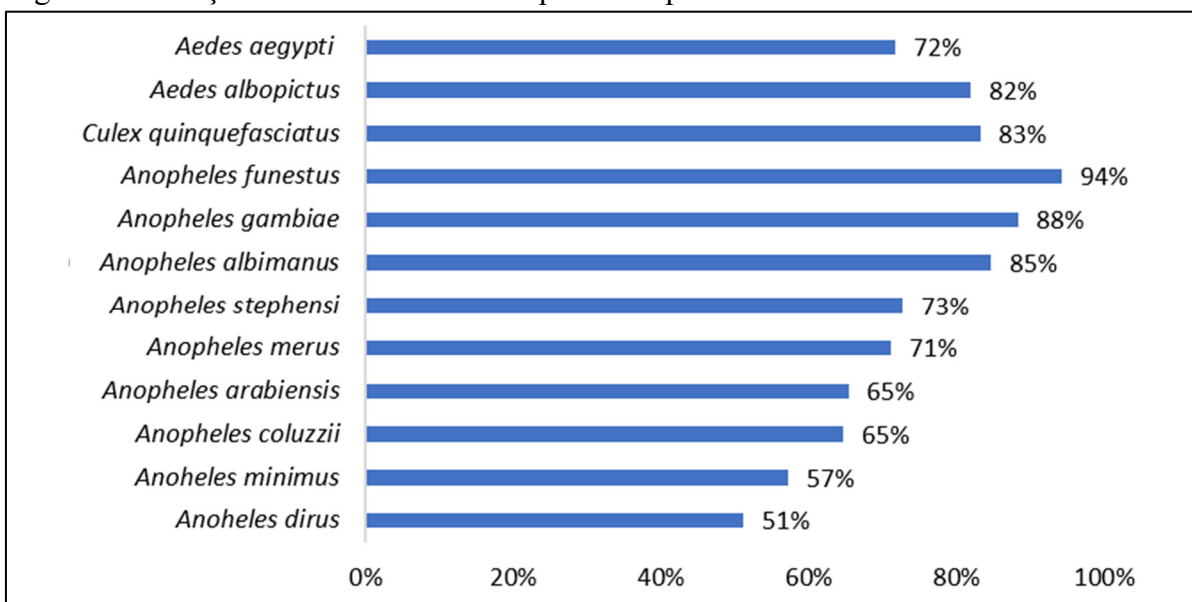


**Fonte:** Elaborado pelo autor.

**Legenda:** Cada gráfico apresenta a densidade (eixo Y) do número de *reads* mapeadas (eixo X) em cada família de TE (vermelho) ou cada cópia de gene (azul). A linha vertical pontilhada indica a média o número médio de *reads* para cada gene ou TE. Observa-se que a média de *reads* é menor para os TEs em quase todas as espécies.

O número de famílias de TEs que são expressas nas diversas condições nas diferentes espécies de mosquito é bastante elevado. Em todas as espécies estudadas, pelo menos 50% das famílias de elementos de transposição de uma determinada espécie possui ao menos uma cópia que é expressa em nível de RNA (Fig. 35).

Figura 35 – Frações das famílias de TEs que são expressas em nível de RNA.



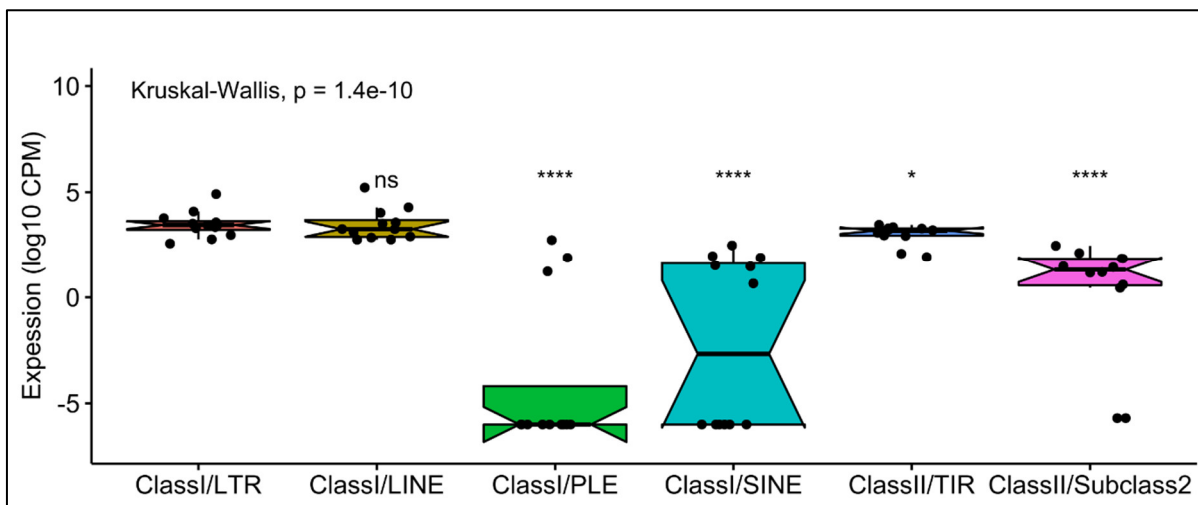
**Fonte:** Elaborado pelo autor

**Legenda:** Os valores em percentual representam a fração do mobiloma de cada espécie, que possui mapeamento de *reads* de RNA-Seq. Por exemplo, no caso do *Aedes aegypti*, 72% das famílias apresentam algum sinal de expressão.

Essa fração do número de famílias expressas parece não ter relação com a fração do genoma coberta por elementos de transposição já que espécies com poucos TEs, como *An. minimus*, possui uma fração de famílias expressas (57%) com uma pequena diferença para espécies como um maior número de TEs, como *An. coluzzii*, que possui 65% de todas as suas famílias de TEs com algum sinal de expressão em nível de RNA. Essa falta de relação também fica clara quando comparamos a espécie *Ae. aegypti*, cuja metade do genoma é formada por elementos de transposição, com a espécie *An. funestus*, que possui apenas ~5,5% do genoma coberto por TEs. Nesta última espécie, o número de famílias de TEs ainda com capacidade de expressão é bem maior (94%) que em *Ae. aegypti* (72%).

A maior parte dos elementos de transposição expressos nos mosquitos é formada por TEs de Classe I. Principalmente elementos da ordem LTR e elementos da ordem LINE, seguindo a classificação de Wicker *et al.* (2007). A figura 36 mostra o nível de expressão média normalizada por CPM de cada ordem de TE nas 12 espécies de mosquitos com dados de expressão gênica. Nota-se que os níveis de expressão das diferentes ordens são diferentes. Elementos da ordem PLE, SINE, TIR, Helitron e Maverick possuem níveis de expressão inferiores aos das ordens LTR e LINE. O nível médio de expressão entre essas duas ordens, por sua vez, não apresentou diferença significativa.

Figura 36 – Bloxplot da expressão dos TE nas 12 espécies de acordo com as diferentes ordens.



**Fonte:** Elaborado pelo autor.

**Legenda:** Cada ponto preto dentro dos boxes representa o valor da expressão normalizada de uma determinada ordem de TE numa determinada espécie de mosquito. A estatística de Kruskal-wallis foi aplicada como teste não paramétrico para comparação entre todas as médias. Cada grupo então foi comparado individualmente com a ordem LTR através do teste de Wilcoxon, Ns = não significante ( $p > 0.05$ ), \* significante ( $p \leq 0.05$ ), \*\*\*\* significante ( $p \leq 0.0001$ ).

A expressão dos TEs também foi avaliada para cada superfamília individualmente. A figura 37 mostra um *heatmap* com os níveis de expressão de cada tipo de TE por cada espécie de mosquito. Observa-se que as 3 superfamílias ubíquas nos genomas dos mosquitos (*Gypsy*, *Bel-Pao* e *Copia*) também são superfamílias com a maior expressão dentre todos os TEs. Sua expressão é particularmente alta nos mosquitos do gênero *Aedes*. É interessante destacar que a superfamília *Gypsy* é a superfamília mais expressa na espécie *Ae. albopictus*, onde está enriquecida nos RNAs longos não codificantes que são produzidos pela espécie.

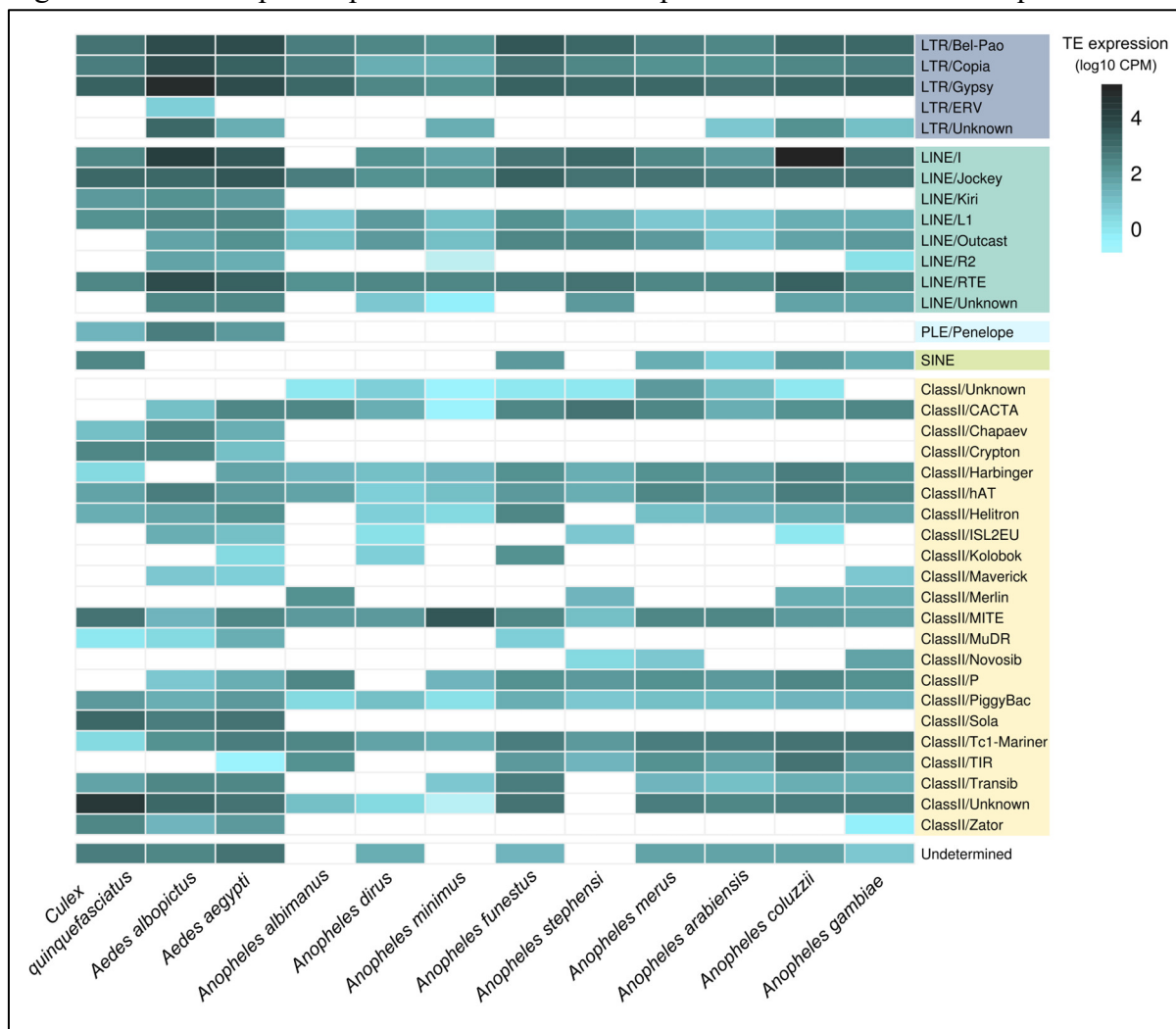
Dentre os LINEs, apenas 3 superfamílias (*Jockey*, *LI* e *RTE*) possuem cópias expressas em todas as espécies. Já os SINEs são expressos apenas em metade das espécies de mosquitos. Os elementos da superfamília Penelope por sua vez, são expressos em todas as espécies onde eles ocorrem, que são as 3 espécies da subfamília Culicinae.

O padrão de expressão dos elementos de Classe II é bem mais escasso que o observado para elementos de Classe I. Apenas as superfamílias *Tc1-Mariner*, *hAT* e *PiggyBac* são expressas em todas as espécies, no entanto o seu nível de expressão é bastante variado.

Assim como esperado devido ao grande número e diversidade de TEs compondo seus genomas, as espécies *Ae. aegypti* e *Ae. albopictus* são as que possuem uma maior diversidade de superfamílias de TEs expressas. Por outro lado, *An. albimanus*, a espécie com o menor número de TEs entre as 12 avaliadas nesta análise, é a espécie com a menor diversidade de

superfamílias de TEs expressas. Já o nível de expressão de TEs atinge o seu limiar inferior na espécie *An. dirus*, que possui apenas cerca de 1650 *reads* mapeados em cópias de TEs para cada milhão de *reads* de RNA-Seq mapeados.

Figura 37 – Heatmap da expressão das diferentes superfamílias de TEs nos mosquitos.



Fonte: Elaborado pelo autor.

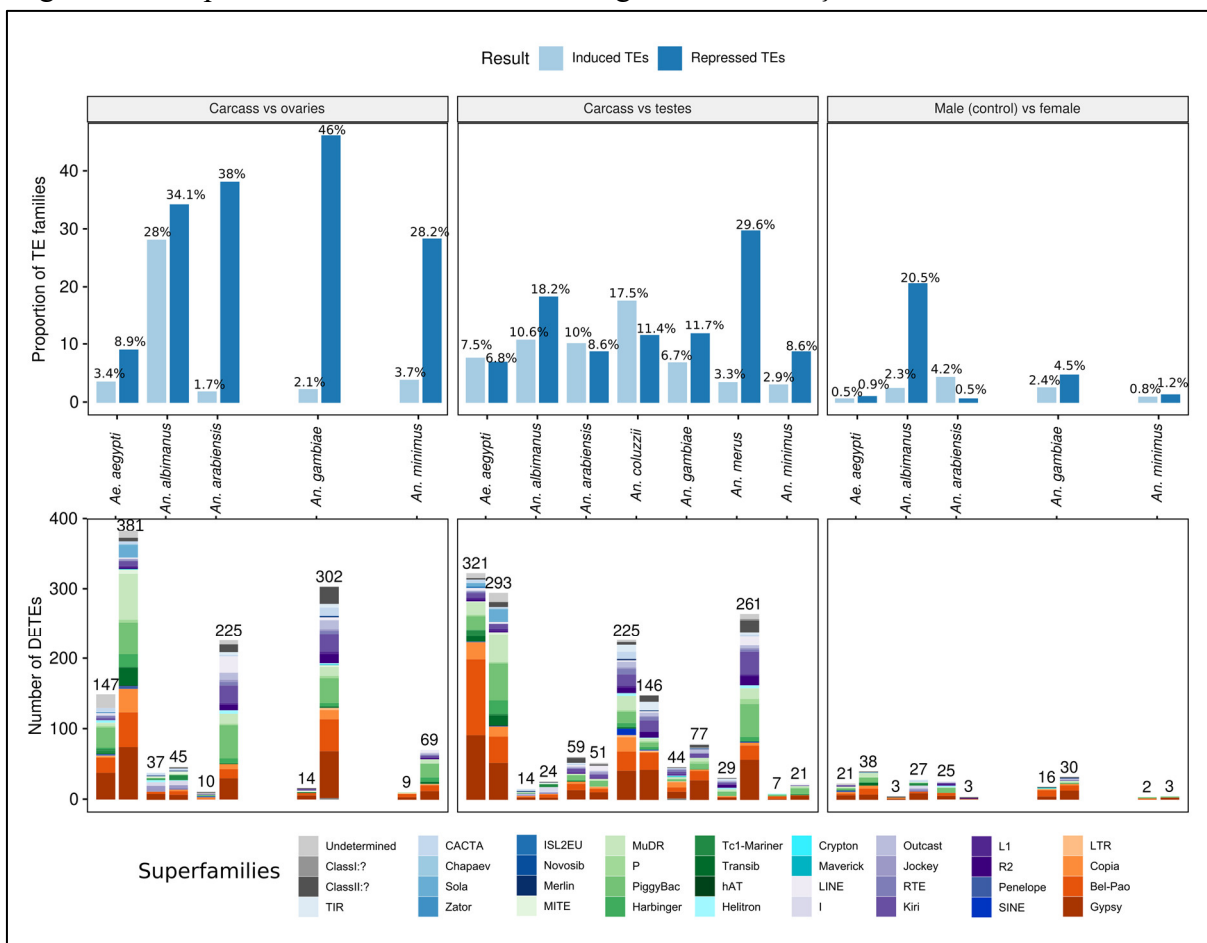
### 6.2.5 Os TEs são fortemente reprimidos nos ovários

No geral, foi observada uma maior quantidade de TEs sofrendo repressão em comparação com a quantidade de TEs com expressão induzida nos ovários em relação a carcaça (Fig. 38). Em algumas espécies, mais de um terço de todas as famílias de TEs expressas sofrem repressão no tecido germinativo feminino. Destaca-se especialmente a espécie *An. gambiae*, onde quase 50% dos TEs têm sua expressão reduzida.



Esse padrão de repressão dos TEs nos tecidos germinativos não é observado quando considerado o tecido germinativo masculino. Para 4 espécies, há mais TEs reprimidos que induzidos, para 3 outras espécies, o comportamento inverso é observado. No entanto, no geral, mais de 70% das famílias de elementos de transposição não tiveram sua expressão alterada nos testículos dos mosquitos quando comparadas aos tecidos somáticos.

Figura 38 – Expressão diferencial dos TEs nas gônadas e carcaças de machos e fêmeas.



**Fonte:** Elaborado pelo autor

**Legenda:** A comparação Carcaça x ovários e Carcaça x testículos, tem como referência (controle) a carcaça, já a comparação entre a carcaça de machos e fêmeas, tem o macho como referência. O painel superior mostra o quanto do total de famílias de TEs de uma espécie estão diferencialmente expressas (DETEs) numa determinada condição, já o painel inferior mostra o número absoluto de TEs diferencialmente expressos assim como sua composição por superfamílias. As colorações no painel superior são relativas à diferença de expressão dos TEs, as colorações do painel inferior mostram a distribuição de famílias que compõe o grupo dos TEs induzidos ou reprimidos em cada condição. No geral observa-se uma forte repressão da expressão dos TEs nos ovários. Esse padrão de repressão não é observado no testículo. Já a expressão dos tecidos somáticos varia muito pouco entre machos e fêmeas da mesma espécie.

A proporção de famílias cuja expressão não varia é maior quando se compara os tecidos somáticos de machos e fêmeas da mesma linhagem, para a maior parte das espécies esse valor fica acima de 90%. Na carcaça de *An. albimanus* e *An. gambiae*, os TEs são mais reprimidos nas fêmeas que nos machos, similarmente há uma maior expressão de alguns genes nos machos

que nas fêmeas (PAPA *et al.*, 2017). Já em *An. arabiensis*, eles são mais induzidos nas fêmeas, contrastando com uma proporção quase idêntica de genes que são induzidos em machos e fêmeas dessa espécie (PAPA *et al.*, 2017). Em *An. minimus* há uma variação muito pequena da expressão de TEs entre as carcaças dos diferentes sexos, essa variação pequena também é notada no conjunto total de genes (PAPA *et al.*, 2017).

### 6.2.6 Expressão dos TEs em situação de estresse e de infecção

Foram avaliados 4 tipos de estresse: o estresse populacional, quando há um maior número de mosquitos por metro quadrado que o habitual; o estresse de diapausa, quando os mosquitos são colocados em condições de indução de diapausa; o estresse de exposição a compostos químicos que afetam a sobrevivência dos mosquitos; e o estresse de exposição a diferentes graus de salinidade (Fig. 39).

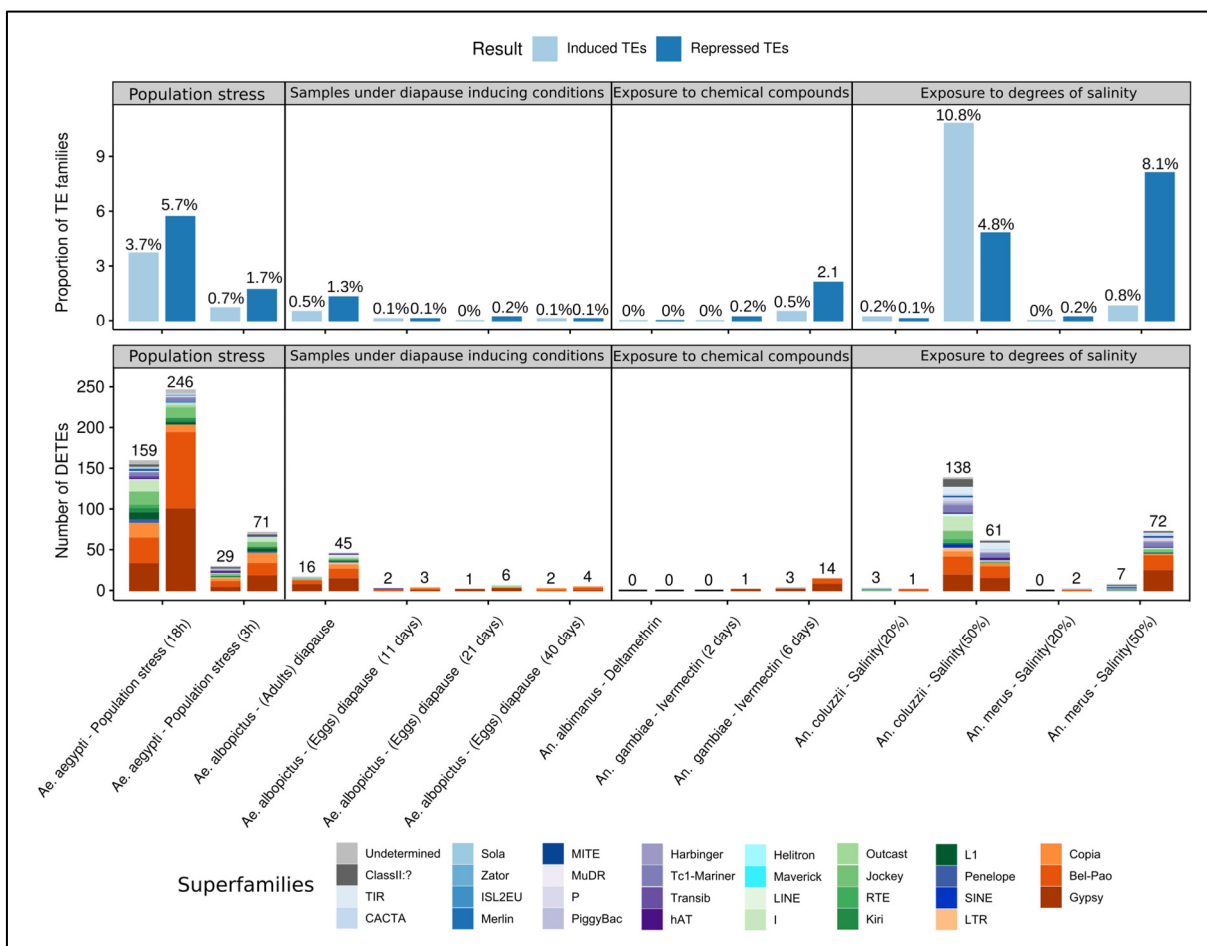
Interessantemente, a exposição a compostos químicos como ivermectina (alimentação com 11.75 ng/ml de sangue), um fármaco antiparasitário que também possui efeito na sobrevivência de mosquitos *Anopheles* (SEAMAN *et al.*, 2015), e deltametrina (exposição por 30 minutos à 12.5 µg num bioensaio de resistência no padrão do CDC - *CDC Bottle Bioassay*), um inseticida piretróide, em *An. gambiae* e *An. albimanus* respectivamente, praticamente não alterou a expressão dos elementos de transposição havendo apenas uma leve repressão da expressão de 14 famílias (2,1% do total) em *An. gambiae* com 6 dias de emergência e exposto à ivermectina.

A exposição a condições de indução de diapausa (ver materiais e métodos) também não foi capaz de alterar a expressão da maior parte dos elementos de transposição presentes em *Ae. albopictus*. Nos ovos do mosquito apenas 0,2% dos TEs tiveram sua expressão alterada pelas condições de diapausa. Em mosquitos adultos esse valor aumenta um pouco para 61 famílias diferencialmente expressas, o que corresponde a 1,8% do total de famílias de TEs dessa espécie.

Diferentemente das condições de estresse anteriores, a exposição ao estresse de salinidade (ver materiais e métodos) foi capaz de alterar a expressão de fração significativa dos elementos de transposição em mosquitos. Em *An. merus*, cerca de 9% das famílias de TEs respondem a mudança de salinidade no meio, já em *An. coluzzii* esse valor sobe para 15,5% do total de TEs. No entanto, a variação da expressão se mostra dependente da intensidade do estresse. Em baixo grau de salinidade, um número muito pequeno de TE tem sua expressão alterada, o que muda consideravelmente quando as espécies são expostas a um grau de salinidade de 50%. Em *An.*

*merus*, uma espécie mais tolerante à salinidade, a maior parte dos TEs diferencialmente expressos representam famílias cuja expressão é reprimida. Já em *An. coluzzii*, uma espécie que não tolera salinidade, a maior parte dos TEs diferencialmente expressos representam famílias cuja expressão é induzida, cerca de 11% dos TEs.

**Figura 39** – Expressão diferencial dos TEs em mosquitos sob diferentes situações de estresse.



**Fonte:** Elaborado pelo autor

**Legenda:** O painel superior mostra o quanto do total de famílias de TEs de uma espécie estão diferencialmente expressas (DETEs) numa determinada condição de estresse, já o painel inferior mostra o número absoluto de TEs diferencialmente expressos assim como sua composição por superfamílias. As colorações no painel superior são relativas à diferença de expressão dos TEs, as colorações do painel inferior mostram a distribuição de famílias que compõe o grupo dos TEs induzidos ou reprimidos em cada condição. A exposição a condições indutoras de diapause e a exposição a compostos indutores de mortalidade em mosquitos praticamente não afetou a expressão gênica. Enquanto isso o crescimento em condições de alta densidade populacionais e a exposição a salinidade afetam bastante a expressão de TEs.

O estresse populacional também alterou a expressão dos TEs, havendo mais TEs reprimidos que induzidos (Figura 39, painel esquerdo) em populações de mosquito cujas larvas cresceram sob alta densidade populacional. Neste caso, um segundo fator de estresse foi avaliado, a alimentação sanguínea. Embora não seja um tipo de estresse que afeta a

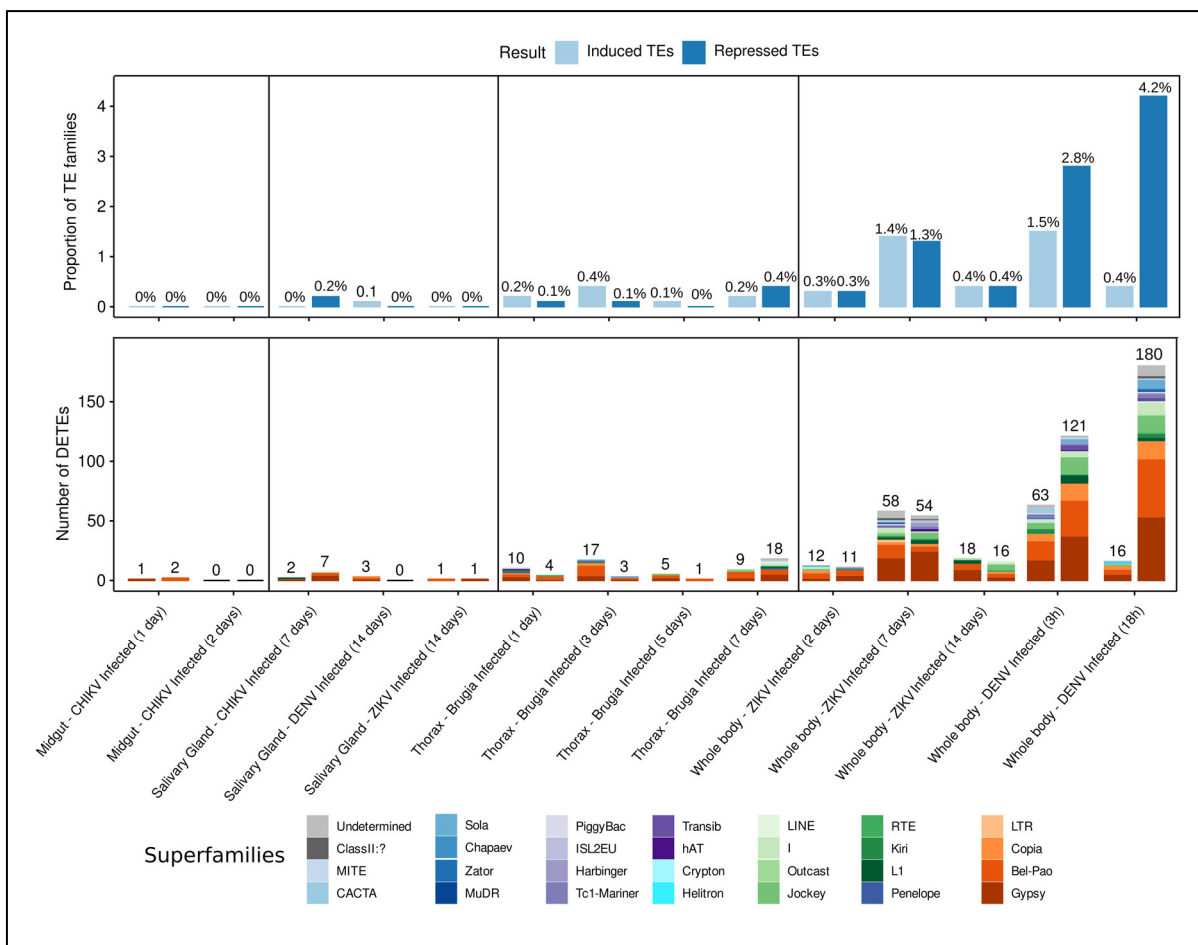
sobrevivência dos mosquitos, a alimentação interfere diretamente na fisiologia da fêmea ingurgitada. Em *Ae. aegypti* 3 horas após a ingestão de sangue, o número de TEs diferencialmente expressos (entre crescidos em condições ótimas e condições de estresse) representou 2,3% das famílias de TEs. Valor que atingiu 9,4% após 18 horas após a alimentação.

A grande quantidade de experimentos públicos de RNA-Seq em *Ae. aegypti* nos permitiu verificar o perfil de expressão dos TEs sob diversas condições e em diferentes estruturas (Fig. 40). A infecção com o vírus DENV foi capaz de alterar a expressão de parte dos TEs nos primeiros momentos da infecção, 184 e 196 famílias de TEs apresentaram variação nos níveis de expressão em 3 horas e 18 horas pós infecção, respectivamente, no corpo total. Neste caso, o número de famílias de TEs cuja expressão é reprimida aumenta com o tempo de infecção. O padrão contrário é observado para as famílias de TEs cuja expressão é induzida. Ainda considerando-se uma amostra de corpo total, a infecção com o vírus ZIKV altera a expressão de 23 famílias de TEs dois dias após o contato com o vírus. Esse número sobe para 112 famílias diferencialmente expressas após 7 dias de infecção, e cai para 34 após 14 dias. A alta em 7 dpi provavelmente está relacionada com o fato de ser um período em que o vírus ainda está se proliferando dentro dos tecidos dos mosquitos. Apesar da variação na expressão de algumas famílias, cerca de 95% dos TEs de *Ae. aegypti* permanecem com sua expressão inalterada, após a infecção com esses dois tipos virais.

Embora haja uma variação na expressão de TEs na glândula salivar (2 famílias) 14 dias pós-infecção, esse número é muito menor que o observado na amostra de corpo total (34 famílias). A infecção de glândula salivar com os vírus DENV e CHIKV também indicam que eles possuem muita pouca capacidade de induzir variações significativas nos níveis de expressão de elementos de transposição neste tipo de tecido. Praticamente também não há diferença na expressão dos TEs no intestino médio no tempo de 1 e 2 dias pós-infecção com CHIKV. A quase totalidade dos TEs que possuem alguma variação da expressão nestes dois tecidos são retrotransposons LTR.

Além da infecção viral, também foram avaliados dados de infecção contra parasitos. Analisando o curso da infecção do *Ae. aegypti* por *Brugia malayi*, nota-se que a expressão de mais de 99% dos TEs permanece inalterada. Dentre os TEs que sofrem variação da sua expressão, há uma maior indução do que repressão nos dias 1, 3 e 5 pós infecção. No dia 7, o número de famílias de TEs reprimidas é o dobro da quantidade de famílias induzidas.

Figura 40 – Expressão diferencial dos TEs em *Aedes aegypti* infectado com diferentes patógenos.



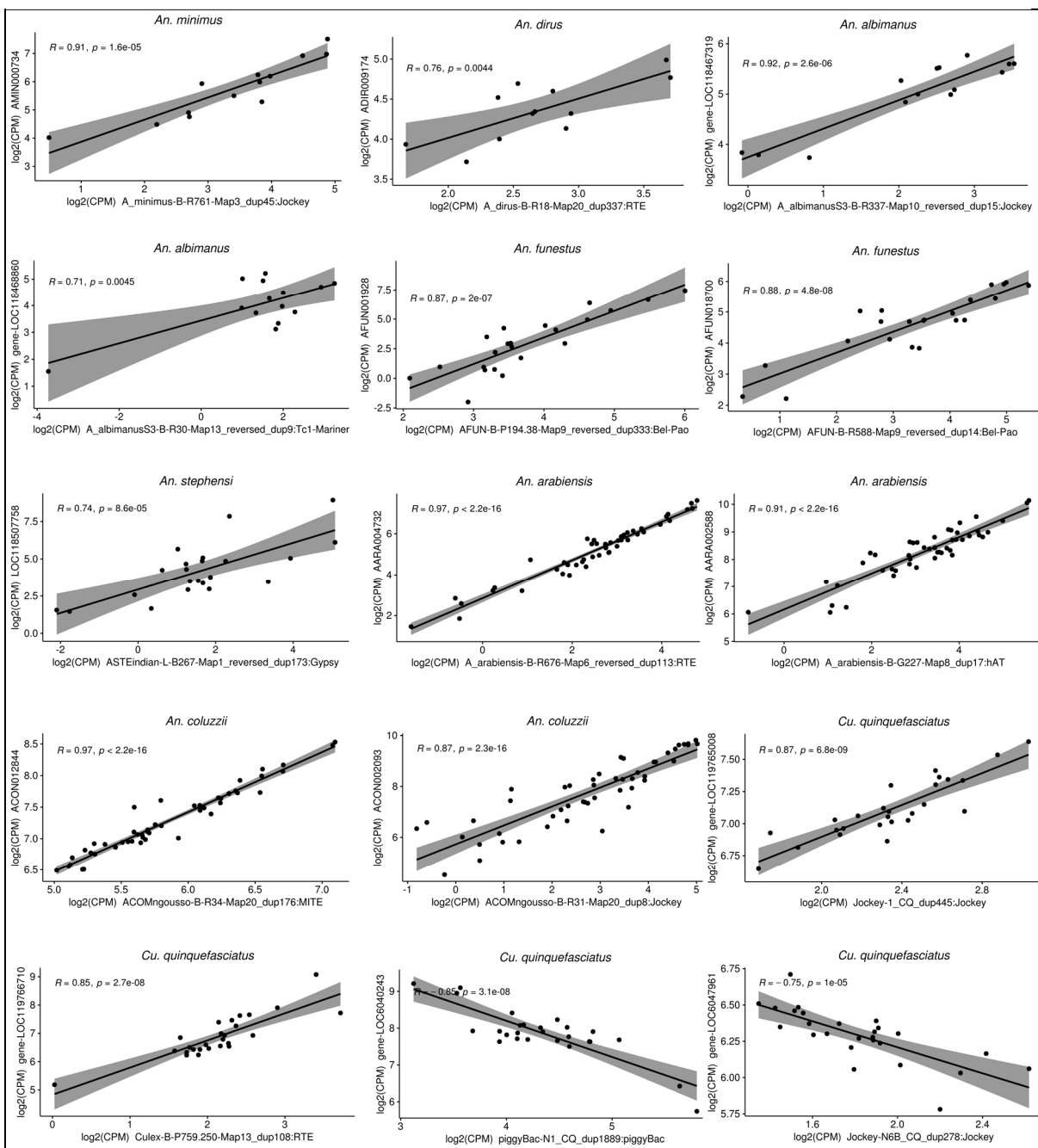
**Fonte:** elaborado pelo autor

**Legenda:** O painel superior mostra o quanto do total de famílias de TEs de uma espécie estão diferencialmente expressas (DETEs) numa determinada condição, já o painel inferior mostra o número absoluto de TEs diferencialmente expressos assim como sua composição por superfamílias. As colorações no painel superior são relativas à diferença de expressão dos TEs, as colorações do painel inferior mostram a distribuição de famílias que compõe o grupo dos TEs induzidos ou reprimidos em cada condição. Nota-se que, apesar da variação da expressão dos TEs em amostras de corpo total mosquitos infectados, com uma maior quantidade de TEs reprimidos em relação aos induzidos, a grande maioria dos TEs permanece com sua expressão inalterada.

## 6.2.7 Coexpressão entre TEs e genes

Além da sobreposição dos elementos de transposição sobre éxons de alguns genes dos mosquitos, o que pode significar um processo de exonização de TEs, também foi avaliado o efeito em *cis* da expressão dos TEs sobre os genes dos mosquitos. Para isso a expressão de TEs localizados até 10 kb *upstream* de cada gene foi correlacionada com a expressão do respectivo gene. Os resultados dessa análise mostraram que na maioria das espécies há presença de correlações significativas fortes ou moderadas entre as expressões dessas duas estruturas gênicas.

Figura 41 – Correlações fortes entre a expressão de genes e TEs nos mosquitos.



Fonte: Elaborado pelo autor.

Legenda: O valor de R representa o coeficiente de correlação de Pearson. O valor de  $p$  indica a significância estatística das correlações destacadas.

Na grande maioria (96,64%) das correlações, as variáveis, expressão de genes e expressão de TEs, crescem de forma diretamente proporcional, ou seja, quanto maior a expressão do TE, maior a expressão gênica. Porém em 4 casos, todos em *Cx. quinquefasciatus*, podemos observar o comportamento inverso, a expressão gênica é reduzida à medida que a expressão do TE na

região *upstream* aumenta. A Figura 41 mostra o perfil de correlações fortes entre genes e TEs nas espécies e também mostra correlações fortes inversamente proporcionais.

Ao todo foram encontrados 119 TEs capazes de influenciar forte (36 casos) ou moderadamente (83 casos) a expressão de genes próximos em 11 espécies de mosquitos (Quadro 7). Foi observada especificamente 1 associação em *An. minimus*, 1 em *An. dirus*, 3 em *An. albimanus*, 1 em *An. merus*, 5 em *An. stephensi*, 18 em *An. arabiensis*, 19 em *An. coluzzii*, 19 em *An. funestus*, 2 em *An. gambiae*, 37 em *Cx. quinquefasciatus* e 13 em *Ae. albopictus*. A exceção foi a espécie *Ae. aegypti*, cuja correlação mais forte entre as expressões de genes e TEs próximos foi de apenas  $R = 0,37$ . Interessantemente, as duas espécies com a maior carga de TEs são as espécies com menos associações fortes entre a expressão de genes e TEs. Com relação às classes e superfamílias dos TEs que influenciam genes dos mosquitos, a maior parte é composta por TEs de Classe I (64), um número bem próximo de TEs de Classe II (59). Entre os tipos de TEs destacam-se os *RTE* (19), os *Jockey* (17), os elementos de Classe II não caracterizados (15) e os MITEs (10).

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continua)

<i>Anopheles minimus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p-value</i>	Correlação
A_minimus-B-R761-Map3_dup45	<i>Jockey</i>	AMIN000734	<i>unspecified product</i>	0.83	1.59E-05	Muito Forte
<i>Anopheles dirus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p-value</i>	Correlação
A_dirus-B-R18-Map20_dup337	<i>RTE</i>	ADIR009174	<i>unspecified product</i>	0.57	4.36E-03	Forte
<i>Anopheles albimanus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p-value</i>	Correlação
A_albimanusS3-B-R337-Map10_reversed_dup15	<i>Jockey</i>	LOC118467319	<i>FAST kinase domain-containing protein 3, mitochondrial-like</i>	0.85	2.6E-06	Muito Forte
A_albimanusS3-B-R30-Map13_reversed_dup9	<i>Tcl-Mariner</i>	LOC118468860	<i>protein gurken-like</i>	0.50	4.5E-03	Forte
A_albimanusS3-L-B15-Map1_dup21	<i>BEL</i>	LOC118464205	<i>ras-related protein Rap-2a-like</i>	0.43	1.1E-02	Moderada
<i>Anopheles merus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p-value</i>	Correlação
A_merus-B-G1018-Map20_dup4	MITE	AMEM015760	<i>unspecified product</i>	0.28	5.65E-03	Moderada

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continuação)

<i>Anopheles funestus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
AFUN-B-P194.38-Map9_reversed_dup333 AFUN-L-B198-Map1_dup117	<i>RTE - BEL</i>	AFUN001928	<i>unspecified product</i>	0.82	3.5E-08	Muito Forte
AFUN-B-R588-Map9_reversed_dup14	<i>BEL</i>	AFUN018700	<i>unspecified product</i>	0.78	4.8E-08	Forte
AFUN-B-R1701-Map5_dup44	<i>CACTA</i>	AFUN003193	<i>ATP synthase subunit beta</i>	0.78	6.5E-08	Forte
AFUN-B-P194.38-Map9_reversed_dup505 AFUN-B-R3318-Map3_dup4	<i>RTE</i>	AFUN018534	<i>unspecified product</i>	0.55	1.8E-04	Forte
AFUN-B-R1067-Map4	<i>I</i>	AFUN019227	<i>CLIP domain-containing serine protease</i>	0.50	2.4E-04	Forte
AFUN-L-B71-Map1_reversed_dup222 AFUN-L-B71-Map1_reversed_dup223	Class II:?	AFUN021970	<i>DUF1279 domain-containing protein</i>	0.49	1.6E-03	Forte
AFUN-B-P81.167-Map5_dup20	<i>TIR</i>	AFUN009150	<i>serine/threonine-protein kinase ATR</i>	0.48	3.7E-04	Moderada
AFUN-B-R3322-Map8_dup20	<i>RTE</i>	AFUN019033	<i>unspecified product</i>	0.44	7.6E-04	Moderada
AFUN-B-R3322-Map8_dup21	<i>RTE</i>	AFUN018736	<i>unspecified product</i>	0.42	1.2E-03	Moderada
AFUN-B-R1067-Map4	<i>I</i>	AFUN019728	<i>CLIP domain-containing serine protease</i>	0.40	1.5E-03	Moderada
AFUN-B-R2473-Map5_dup36	<i>P</i>	AFUN008911	<i>calcium and integrin-binding protein</i>	0.39	2.0E-03	Moderada
AFUN-B-R524-Map8_dup16	<i>Tc1-Mariner</i>	AFUN007928	<i>unspecified product</i>	0.37	2.5E-03	Moderada
AFUN-B-R1065-Map20_dup10	<i>I</i>	AFUN002377	<i>Mesenchymal stem cell protein DSCD75</i>	0.37	2.6E-03	Moderada
AFUN-B-R2050-Map3_dup46	<i>MITE</i>	AFUN021710	<i>protease m1 zinc metalloprotease</i>	0.36	3.3E-03	Moderada
AFUN-B-R1065-Map20_dup10	<i>I</i>	AFUN002379	<i>Serine--pyruvate aminotransferase</i>	0.34	4.1E-03	Moderada
AFUN-B-R3322-Map8_dup23	<i>RTE</i>	AFUN018584	<i>unspecified product</i>	0.34	4.4E-03	Moderada
AFUN-B-R3322-Map8_dup17	<i>RTE</i>	AFUN019003	<i>unspecified product</i>	0.34	4.4E-03	Moderada
AFUN-L-B419-Map1_reversed_dup38	<i>Copia</i>	AFUN007624	<i>pigment dispersing hormone</i>	0.34	4.7E-03	Moderada
AFUN-B-R1067-Map4	<i>I</i>	AFUN019710	<i>Peptidase S1 domain-containing protein</i>	0.33	5.3E-03	Moderada



Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continuação)

<i>Anopheles stephensi</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
ASTEindian-L-B267-Map1_reversed_dup173	<i>Gypsy</i>	LOC118507758	<i>dihydroceramide fatty acyl 2-hydroxylase FAH2</i>	0.55	8.6E-05	Forte
ASTEindian-B-R45-Map20_dup466	<i>RTE</i>	LOC118511611	<i>YEATS domain-containing protein 4</i>	0.40	1.5E-03	Moderada
ASTEindian-B-R2417-Map4_reversed_dup12	<i>Gypsy</i>	LOC118512284	<i>uncharacterized LOC118512284</i>	0.34	4.7E-03	Moderada
ASTEindian-B-P175.14-Map12_dup819	<i>RTE</i>	LOC118506941	<i>copper homeostasis protein cutC homolog</i>	0.30	8.0E-03	Moderada
ASTEindian-B-P109.0-Map20_dup417	<i>I</i>	LOC118513387	<i>multidrug resistance-associated protein 1</i>	0.29	9.3E-03	Moderada
<i>Anopheles arabiensis</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
A_arabiensis-B-R676-Map6_reversed_dup113	<i>RTE</i>	AARA004732	<i>putative GHMP kinases C</i>	0.94	7.65E-30	Muito Forte
A_arabiensis-B-G455-Map17_dup65	<i>TIR</i>	AARA000328	<i>unspecified product</i>	0.83	4.53E-19	Muito Forte
A_arabiensis-B-G227-Map8_dup17	<i>hAT</i>	AARA002588	<i>Aminotran_1_2 domain-containing protein</i>	0.83	1.13E-18	Muito Forte
A_arabiensis-B-G24-Map20_dup57	<i>Gypsy</i>	AARA016808	<i>F-type H+-transporting ATPase subunit c</i>	0.74	7.40E-15	Forte
A_arabiensis-B-P116-87-Map5_dup5	MITE	AARA008738	<i>collagen type IV alpha</i>	0.68	8.25E-13	Forte
A_arabiensis-B-G772-Map20_dup57 A_arabiensis-B-G815-Map20_dup57	<i>P - Harbinger</i>	AARA003634	<i>LIM interaction domain (LID) protein</i>	0.59	2.93E-10	Forte
A_arabiensis-B-G807-Map9_dup3	MITE	AARA003160	<i>phosphatidate phosphatase</i>	0.57	9.17E-10	Forte
A_arabiensis-B-R399-Map13_dup1	<i>Jockey</i>	AARA001557	<i>oxysterol-binding protein</i>	0.49	4.67E-08	Moderada
A_arabiensis-B-G626-Map16_dup247	<i>Harbinger</i>	AARA006767	<i>unspecified product</i>	0.46	1.92E-07	Moderada
R_8_reversed_dup20	<i>Tc1-Mariner</i>	AARA014798	<i>unspecified product</i>	0.44	2.99E-07	Moderada
A_arabiensis-B-R6-Map20_dup37	MITE	AARA001530	<i>unspecified product</i>	0.43	4.94E-07	Moderada
A_arabiensis-B-R16-Map20_reversed_dup249	<i>RTE</i>	AARA011801	<i>unspecified product</i>	0.38	3.89E-06	Moderada
R_14_reversed_dup8	Class II:?	AARA002502	<i>unspecified product</i>	0.37	5.52E-06	Moderada
R_9_dup35	Class II:?	AARA001968	<i>unspecified product</i>	0.34	1.58E-05	Moderada
A_arabiensis-B-R410-Map11_reversed_dup1	<i>Jockey</i>	AARA000414	<i>unspecified product</i>	0.33	2.13E-05	Moderada
A_arabiensis-B-P85-114-Map20_dup34	MITE	AARA006342	<i>unspecified product</i>	0.33	2.16E-05	Moderada

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continuação)

<i>Anopheles stephensi</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
A_arabiensis-B-R16-Map20 reversed dup28	<i>RTE</i>	AARA001275	<i>unspecified product</i>	0.30	7.07E-05	Moderada
A_arabiensis-B-R230-Map11 dup11	<i>Jockey</i>	AARA016856	<i>unspecified product</i>	0.29	1.06E-04	Moderada
<i>Anopheles coluzzii</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
ACOMngouso-B-R34-Map20_dup176	MITE	ACON012844	<i>Calmodulin</i>	0.94	8.2E-30	Muito Forte
ACOMngouso-B-R31-Map20_dup8	<i>Jockey</i>	ACON002093	<i>Na_H_Exchanger domain-containing protein</i>	0.76	2.3E-16	Forte
ACOMngouso-B-R862-Map20_dup15	<i>Jockey</i>	ACON003092	<i>unspecified product</i>	0.65	3.2E-12	Forte
ACOMngouso-B-G321-Map20_dup2	Class II:?	ACON029616	<i>RRM domain-containing protein</i>	0.54	2.0E-09	Forte
ACOMngouso-B-R599-Map11_dup294	TIR	ACON011203	<i>Eukaryotic translation initiation factor 5</i>	0.51	7.9E-09	Forte
ACOMngouso-B-R3314-Map4 reversed dup2	<i>P</i>	ACON006809	<i>unspecified product</i>	0.50	1.6E-08	Forte
ACOMngouso-B-R605-Map5 reversed dup30	TIR	ACON004938	<i>Palmitoyltransferase</i>	0.45	1.2E-07	Moderada
ACOMngouso-B-G48-Map20_dup288 ACOMngouso-B-R3574-Map3_dup10	<i>Tc1-Mariner</i> - Class II:?	ACON010304	<i>unspecified product</i>	0.40	8.3E-06	Moderada
ACOMngouso-B-P8.36-Map20_dup57	<i>Tc1-Mariner</i>	ACON008988	<i>Glutamine synthetase</i>	0.38	2.8E-06	Moderada
ACOMngouso-B-P201.18-Map3	<i>Jockey</i>	ACON004189	<i>FAD-binding PCMH-type domain-containing protein</i>	0.37	3.4E-06	Moderada
ACOMngouso-B-R862-Map20_dup15	<i>Jockey</i>	ACON003093	<i>AAA domain-containing protein</i>	0.35	7.6E-06	Moderada
ACOMngouso-B-P201.18-Map3	<i>Jockey</i>	ACON004195	<i>unspecified product</i>	0.35	7.8E-06	Moderada
ACOMngouso-B-G4065-Map3_dup23	LTR	ACON013181	<i>Peptidase_M14 domain-containing protein</i>	0.34	9.4E-06	Moderada
ACOMngouso-B-R599-Map11_dup231	TIR	ACON000626	<i>MSP domain-containing protein</i>	0.34	1.0E-05	Moderada
ACOMngouso-B-P8.36-Map20_dup29	<i>Tc1-Mariner</i>	ACON010332	<i>C2H2-type domain-containing protein</i>	0.34	1.2E-05	Moderada
ACOMngouso-B-G4065-Map3_dup23	LTR	ACON001026	<i>unspecified product</i>	0.33	1.3E-05	Moderada
ACOMngouso-B-G48-Map20_dup85	<i>Tc1-Mariner</i>	ACON002858	<i>Sodium/potassium-transporting ATPase subunit alpha</i>	0.33	1.5E-05	Moderada

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continuação)

<i>Anopheles coluzzii</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
ACOMngouso-B-R3930-Map8 dup4	<i>hAT</i>	ACON009893	<i>Nocturnin</i>	0.32	2.3E-05	Moderada
ACOMngouso-B-P115.290-Map11 dup9	MITE	ACON011768	<i>unspecified product</i>	0.31	2.8E-05	Moderada
<i>Anopheles gambiae</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
R_60_dup103 Mariner-N22_AG_dup296	<i>hAT - Tc1-Mariner</i>	AGAP028897	<i>Eukaryotic large subunit ribosomal RNA</i>	0.39	2.36E-03	Moderada
A_gambiaeChr-B-R476-Map20 dup44	TIR	AGAP029751	<i>argonaute 2</i>	0.31	1.43E-04	Moderada
<i>Culex quinquefasciatus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
Jockey-1_CQ_dup445	<i>Jockey</i>	LOC119765008	<i>ankyrin repeat domain-containing protein 39</i>	0.76	6.80E-09	Forte
culex_third-B-P759.250-Map13 dup108	<i>RTE</i>	LOC119766710	<i>uncharacterized LOC119766710</i>	0.73	2.72E-08	Forte
<i>PiggyBac</i> -N1_CQ_dup1889	<i>PiggyBac</i>	LOC6040243	<i>protein LSM14 homolog B</i>	0.73	3.11E-08	Forte
culex_third-B-R6415-Map4_reversed_dup1	<i>Penelope</i>	LOC6041556	<i>CCR4-NOT transcription complex subunit 10</i>	0.66	5.31E-07	Forte
culex_third-B-G7237-Map20 dup5	SINE	LOC6037729	<i>1,4-alpha-glucan-branching enzyme</i>	0.66	5.41E-07	Forte
culex_third-B-P107.533-Map20 dup339	SINE	LOC6031960	<i>COP9 signalosome complex subunit 4</i>	0.65	7.65E-07	Forte
Jockey-N6B_CQ_dup278	<i>Jockey</i>	LOC6047961	<i>integrator complex subunit 2</i>	0.56	1.00E-05	Forte
Jockey-N1_CQ_dup76	<i>Jockey</i>	LOC6048396	<i>isoleucine--tRNA ligase, mitochondrial</i>	0.56	1.26E-05	Forte
Jockey-N8_CQ_dup160	<i>Jockey</i>	LOC6037735	<i>DNA-directed RNA polymerase III subunit RPC1</i>	0.55	1.43E-05	Forte
Jockey-1_CQ_dup445	<i>Jockey</i>	LOC6036652	<i>uncharacterized LOC6036652</i>	0.53	2.84E-05	Forte
culex_third-B-G1524-Map3 dup493	<i>Transib</i>	LOC6040991	<i>transcriptional repressor CTCF</i>	0.52	3.10E-05	Forte
culex_third-B-G3447-Map4 dup1101	MITE	LOC6047592	<i>RRP15-like protein</i>	0.50	5.43E-05	Forte
DNA-TA-7_CQ_dup580	Class II:?	LOC6043673	<i>chromatin-remodeling complex ATPase chain Iswi</i>	0.49	7.09E-05	Moderada
Jockey-N6B_CQ_dup286	<i>Jockey</i>	LOC6033499	<i>ATP synthase subunit alpha, mitochondrial</i>	0.46	0.000129	Moderada
I-3_CQ_dup199	<i>I</i>	LOC6041261	<i>protein alan shepard</i>	0.40	0.000644	Moderada

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continuação)

<i>Culex quinquefasciatus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	p-value	Correlação
culex_third-B-G4635-Map11 dup40	<i>Zator</i>	LOC6036109	<i>zinc finger protein 419</i>	0.38	0.00082	Moderada
culex_third-B-G5-Map20 reversed dup1078	Class II:?	LOC6039492	<i>DNA ligase 1</i>	0.38	0.000878	Moderada
culex_third-B-R214-Map20 dup441	MITE	LOC6049318	<i>uncharacterized LOC6049318</i>	0.37	0.001802	Moderada
culex_third-B-R2958-Map9 dup178	<i>Zator</i>	LOC6037734	<i>CXXC-type zinc finger protein 1</i>	0.36	0.001112	Moderada
Sola2-N1B_CQ_dup2348	<i>Sola</i>	LOC6034877	<i>uncharacterized LOC6034877</i>	0.36	0.001178	Moderada
culex_third-B-G4735-Map3 reversed dup119 culex_third-B-R2763-Map6 dup981	Class II:?	LOC6049542	<i>uncharacterized LOC6049542</i>	0.36	0.002433	Moderada
culex_third-B-G4413-Map4_dup25 culex_third-B-G4735-Map3 reversed dup56	Class II:?	LOC119767754	<i>uncharacterized serine-rich protein C215.13-like</i>	0.35	0.001535	Moderada
culex_third-B-G1702-Map20_dup119	SINE	LOC6035267	<i>vacuolar fusion protein MON1 homolog A</i>	0.35	0.001546	Moderada
culex_third-B-P759.250-Map13 dup137	<i>RTE</i>	LOC6041227	<i>pre-mRNA-splicing factor 38</i>	0.34	0.001703	Moderada
Sola2-N1_CQ_dup2461	<i>Sola</i>	LOC6037695	<i>protein argonaute-2</i>	0.34	0.001777	Moderada
DNA-TTAA-3_CQ_dup635 culex_third-B-G1236-Map5 dup209	Class II:? - <i>Zator</i>	LOC6039829	<i>beta carbonic anhydrase 1</i>	0.34	0.007105	Moderada
DYX-incomp_culex_third-B-G1400-Map6 dup246	<i>Crypton</i>	LOC6031239	<i>zinc finger protein 771</i>	0.33	0.002143	Moderada
culex_third-B-G4688-Map9 dup621	Class II:?	LOC6049721	<i>uncharacterized LOC6049721</i>	0.33	2.15E-03	Moderada
DNA8-15_CQ_dup397 DNA8-2_CQ_dup473 culex_third-B-R2763-Map6 dup1269	Class II:?	LOC6039422	<i>ATP-dependent DNA helicase Q4</i>	0.33	0.002179	Moderada
culex_third-B-R2763-Map6 dup175	Class II:?	LOC6036982	<i>ras-related protein Rab-10</i>	0.31	0.003392	Moderada
DNA-TA-7_CQ_dup580	Class II:?	LOC6043675	<i>chromatin-remodeling complex ATPase chain Iswi</i>	0.30	3.93E-03	Moderada
TE-2_CQ_dup562	Undetermined	LOC119766607	<i>uncharacterized LOC119766607</i>	0.29	0.004233	Moderada
DNA-TA-6_CQ_dup1509	Class II:?	LOC6035805	<i>neural/ectodermal development factor IMP-L2</i>	0.27	6.47E-03	Moderada

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Continuação)

<i>Culex quinquefasciatus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
RTE-2_CQ_dup63	<i>RTE</i>	LOC6052628	<i>transcriptional adapter 3</i>	0.26	0.007185	Moderada
Gypsy-33_CQ-I_dup3	<i>Gypsy</i>	LOC6048744	<i>zinc finger protein 135</i>	0.26	0.007548	Moderada
Sola2-N1B_CQ_dup2355	<i>Sola</i>	LOC6039093	<i>uncharacterized LOC6039093</i>	0.26	0.00778	Moderada
Jockey-N5_CQ_dup11	<i>Jockey</i>	LOC6039948	<i>1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase gamma-1</i>	0.25	0.00925	Moderada
<i>Aedes albopictus</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p</i> -value	Correlação
R_6978_reversed_dup24	<i>RTE</i>	LOC109409558	<i>probable ATP-dependent RNA helicase DDX28</i>	0.40	3.26E-04	Moderada
AALF-B-R2019-Map20_dup1005	<i>Chapaev</i>	LOC109417217	<i>uncharacterized LOC109417217</i>	0.39	3.89E-05	Moderada
AALF-L-B2569-Map1_reversed_dup13	<i>Gypsy</i>	LOC109416759	<i>nucleolar protein 12-like</i>	0.38	4.52E-05	Moderada
AALF-L-B6450-Map1_dup1760	<i>I</i>	LOC109411740	<i>beta-1,4-galactosyltransferase 7</i>	0.33	2.13E-04	Moderada
AALF-L-B3161-Map1_reversed_dup100	<i>BEL</i>	LOC109622242	<i>nucleoporin GLE1</i>	0.31	3.47E-04	Moderada
AALF-L-B8194-Map1_dup2	<i>Gypsy</i>	LOC109622551	<i>uncharacterized LOC109622551</i>	0.30	4.78E-04	Moderada
AALF-B-R1254-Map20_reversed_dup356	<i>RTE</i>	LOC115258003	<i>probable phospholipid-transporting ATPase NEO1</i>	0.30	5.03E-04	Moderada
AALF-L-B3627-Map1_dup102	<i>Copia</i>	LOC109411143	<i>kelch domain-containing protein 10 homolog</i>	0.30	5.23E-04	Moderada
AALF-B-R2142-Map20_dup109	<i>hAT</i>	LOC115265269	<i>U6 snRNA-associated Sm-like protein LSm6</i>	0.29	5.59E-04	Moderada
AALF-B-R78-Map20_reversed_dup3016	<i>RTE</i>	LOC109411700	<i>U5 small nuclear ribonucleoprotein 40 kDa protein</i>	0.29	6.28E-04	Moderada
AALF-L-B7720-Map1_reversed_dup5	<i>BEL</i>	LOC109401159	<i>chromosome transmission fidelity protein 8 homolog</i>	0.27	1.08E-03	Moderada
AALF-B-R78-Map20_reversed_dup601	<i>RTE</i>	LOC109425568	<i>heat shock 70 kDa protein cognate 4</i>	0.26	1.28E-03	Moderada
AALF-B-P14.92-Map20_dup8019	<i>I</i>	LOC109405821	<i>histone H2A-beta, sperm-like</i>	0.25	1.53E-03	Moderada

Quadro 7 – Elementos de transposição que potencialmente agem em *cis* modelando a expressão de genes adjacentes (Conclusão)

<i>Aedes aegypti</i>						
TE	Superfamília	Gene	Anotação	R <sup>2</sup>	<i>p-value</i>	Correlação
AaegL5-B-G11004-Map4 reversed dup46	<i>Zator</i>	AAEL005237	<i>unspecified product</i>	0.14	2.4E-05	Fraca*

**Fonte:** elaborado pelo autor

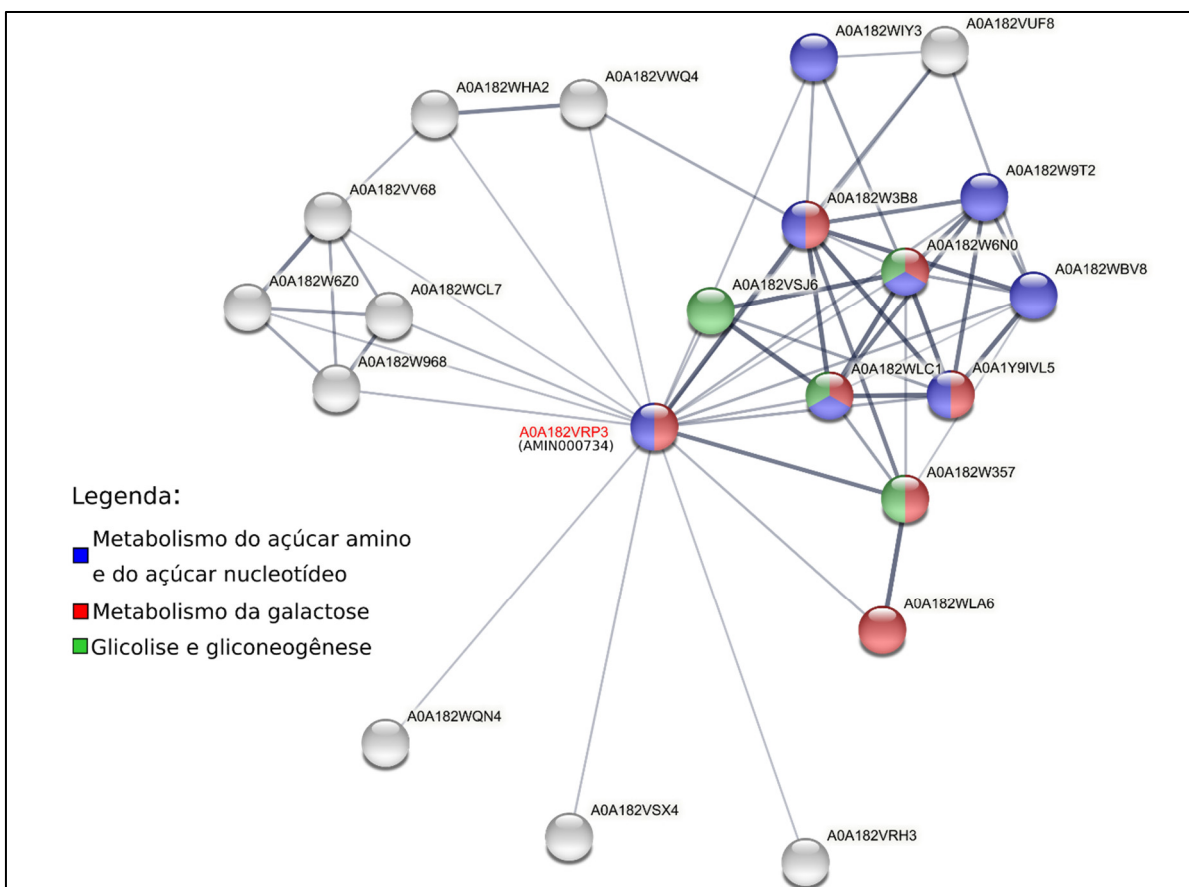
**Legenda:** O coeficiente de determinação (R<sup>2</sup>) indica a precisão da regressão em termos de predição da expressão do gene. O Um R<sup>2</sup> de 0,94 indica que o modelo de regressão poderia explicar 94 % da expressão do gene. Apenas correlações fortes e moderadas (R > 0,5) estão mostradas no quadro. Os códigos da coluna “Gene” são relativos aos códigos dos genes segundo o VectorBase (exemplo: formato AGAP00...) ou NCBI (formato LOC...). A anotação dos genes foi obtida dos respectivos bancos de dados. \* Indica a maior correlação entre TEs e genes em *Ae. aegypti*, embora essa correlação seja fraca. O termo “Class II:?” refere-se a uma família de TE cuja superfamília é apenas classificada como TE de Class II, mas não é ligada a nenhuma superfamília clássica. Isso é de ocorrência mais comum em *Cx. quinquefasciatus*, devido ao fato de diversas famílias de TEs dessa espécie não estarem classificadas em nível de superfamília no banco de dados Repbase.

Em seguida foram avaliadas as redes funcionais onde estão inseridas as proteínas que sofrem algum tipo de influência de TEs em *cis*. Com isso pode-se deduzir se há vias metabólicas ou processos biológicos que podem estar sofrendo impacto da expressão de elementos de transposição. Os resultados a seguir mostram que não há uma mesma via ou processo enriquecido em todas as espécies, sugerindo que a ação de TEs sobre genes é um processo espécie-específico. As vias relatadas aqui então podem ser avaliadas individualmente em estudos futuros focando em uma espécie específica

Em *An. minimus* foi encontrado apenas um elemento com associação a um gene, este, o gene AMIN000734, codifica a proteína A0A182VRP3, que está associada às funções moleculares de ligação de ATP e atividade de galactoquinase e ao processo metabólico de metabolismo da galactose. A rede de associações físicas e funcionais (obtidas a partir do STRING) com proteínas diretamente associadas a A0A182VRP3 mostra um enriquecimento das vias metabólicas KEGG da glicólise e gliconeogênese, assim como o metabolismo de alguns açúcares (Figura 42).

Em *An. dirus* foi encontrado apenas um elemento com associação a um gene, este, o gene ADIR009174, codifica a proteína A0A182NND9 sem nenhuma função inferida, mas que está diretamente associada a um grupo de proteínas que participam da via metabólica de vigilância de mRNA (*mRNA surveillance pathway*) (Figura 43), que é um mecanismo de controle de qualidade que detecta e degrada mRNAs anormais.

Figura 42 – Rede de associações físicas e funcionais com a única proteína de *An. minimus* cuja expressão é fortemente associada à expressão de um TE em *cis*.

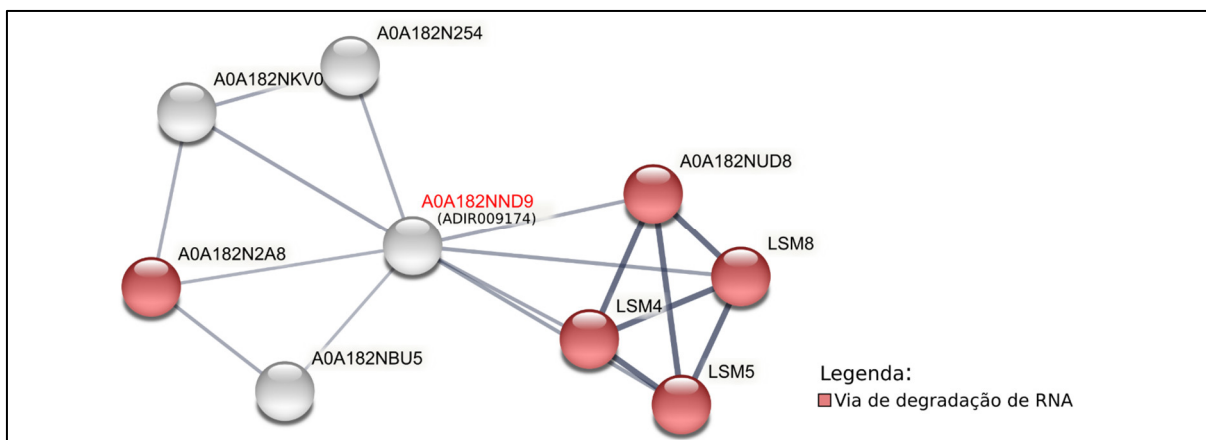


**Fonte:** elaborado pelo autor

Legenda: O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam vias KEGG enriquecidas.

Em *An. gambiae* também foi encontrado apenas um elemento com associação a um gene, este, o gene AGAP011537, codifica uma proteína Argonauta, que está ligada a vias de silenciamento gênico. A rede de associações físicas e funcionais (obtidas a partir do STRING) com proteínas diretamente associadas a AGAP011537 mostra o enriquecimento de processos biológicos de catabolismo de nucleotídeos, silenciamento gênico, RNA de interferência, Regulação pós-transcricional e produção de pequenos RNAs (Figura 44).

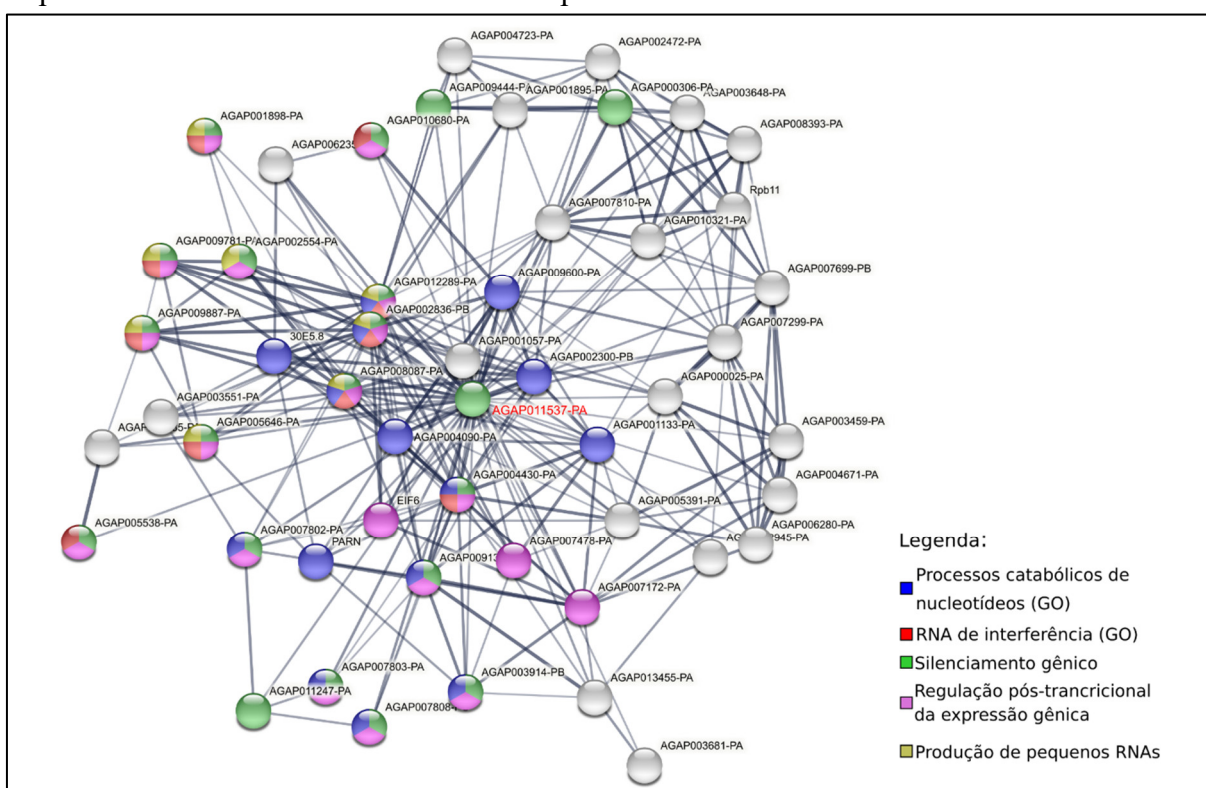
Figura 43 – Rede de associações físicas e funcionais com a única proteína de *An. dirus* cuja expressão é fortemente associada à expressão de TEs.



**Fonte:** elaborado pelo autor

**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam vias KEGG enriquecidas. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5).

Figura 44 – Rede de associações físicas e funcionais com a única proteína de *An. gambiae* cuja expressão é moderadamente associada à expressão de TEs.



**Fonte:** elaborado pelo autor

**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam termos GOs enriquecidos. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5).



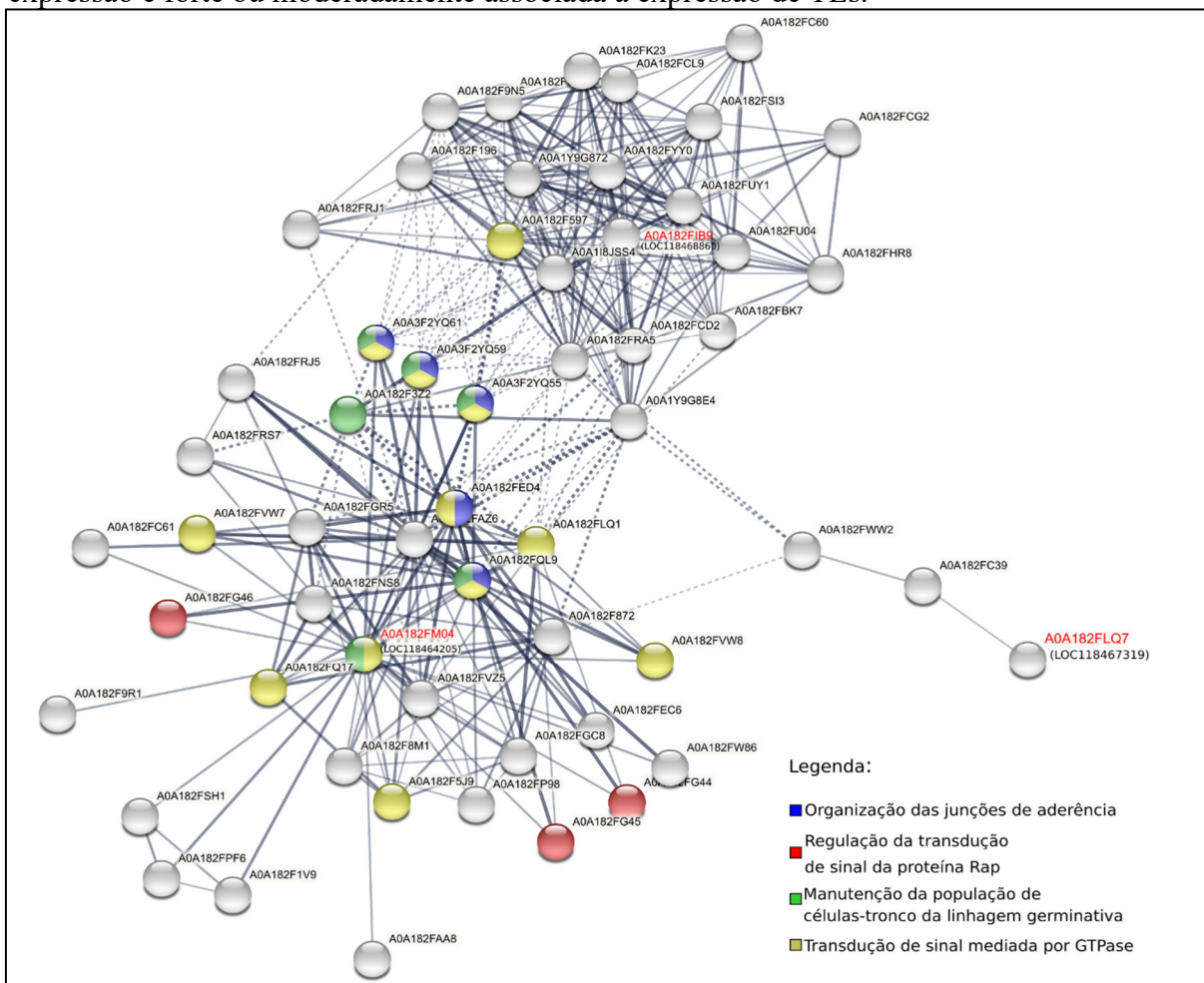
Em *An. albimanus* foram encontrados três TEs com associação a genes. Esses genes são: LOC118467319, que codifica a proteína A0A182FLQ7, que está associada ao processo biológico da regulação da estabilidade do mRNA mitocondrial; o gene LOC118468860, que codifica a proteína A0A182FIB9, que é um componente integral da membrana e que atua como um receptor de ligantes, o gene LOC118464205, que codifica a proteína A0A182FM04, que é uma proteína não caracterizada que participa do processo biológico de transdução de sinal. A rede de associações físicas e funcionais (obtidas a partir do STRING) com proteínas diretamente associadas a esses três genes não mostra um enriquecimento das vias metabólicas KEGG, mas indica um enriquecimento de termos GO envolvidos com a manutenção da população de células tronco germinativas, organização de junções de aderência celular, e transdução de sinal por meio de GTPase, incluindo a GTPase Rap1 (Figura 45).

Em *An. stephensi* foi encontrado apenas um TE com expressão fortemente correlacionada à expressão de um gene e 4 TEs com a expressão moderadamente associada. A rede de associações físicas e funcionais (obtidas a partir do STRING) com proteínas diretamente associadas a esses genes mostra um enriquecimento de funções relacionadas à acetilação das histonas, organização de subunidades do complexo proteína-DNA e metabolismo dos fatores de transcrição basais, principalmente em proteínas ligadas a A0A182Y8B3, e de funções relacionadas ao metabolismo dos esfingolípídios em proteínas associadas a A0A182Y6V3 (Figura 46).

Em *An. funestus* foi encontrado um TE com expressão extremamente correlacionada à expressão de um gene e 5 TEs com a expressão fortemente associada (outros 13 TEs estão moderadamente associados a genes). Dentre esses 6 genes com expressão fortemente associada a TEs, metade deles geram proteínas ainda não caracterizadas, os outros 3 geram as proteínas: A0A182RAH9 (subunidade beta da ATP sintetase) envolvida na síntese de ATP e no transporte de íons de hidrogênio, A0A4Y0BGJ4 (serina protease) e A0A4Y0BR95 (uma proteína com domínio transmembrana). A rede de associações físicas e funcionais mostra um enriquecimento de fosforilação oxidativa e biossíntese de ATP, bem como da via dos fagossomos (Figura 47).

Assim como em *An. funestus*, em *An. coluzzii* também foi encontrado um TE com expressão altamente correlacionada à expressão de um gene, e cinco TEs com a expressão fortemente associada (outros 13 TEs estão moderadamente associados a genes).

Figura 45 – Rede de associações físicas e funcionais das proteínas de *An. albimanus* cuja expressão é forte ou moderadamente associada à expressão de TEs.



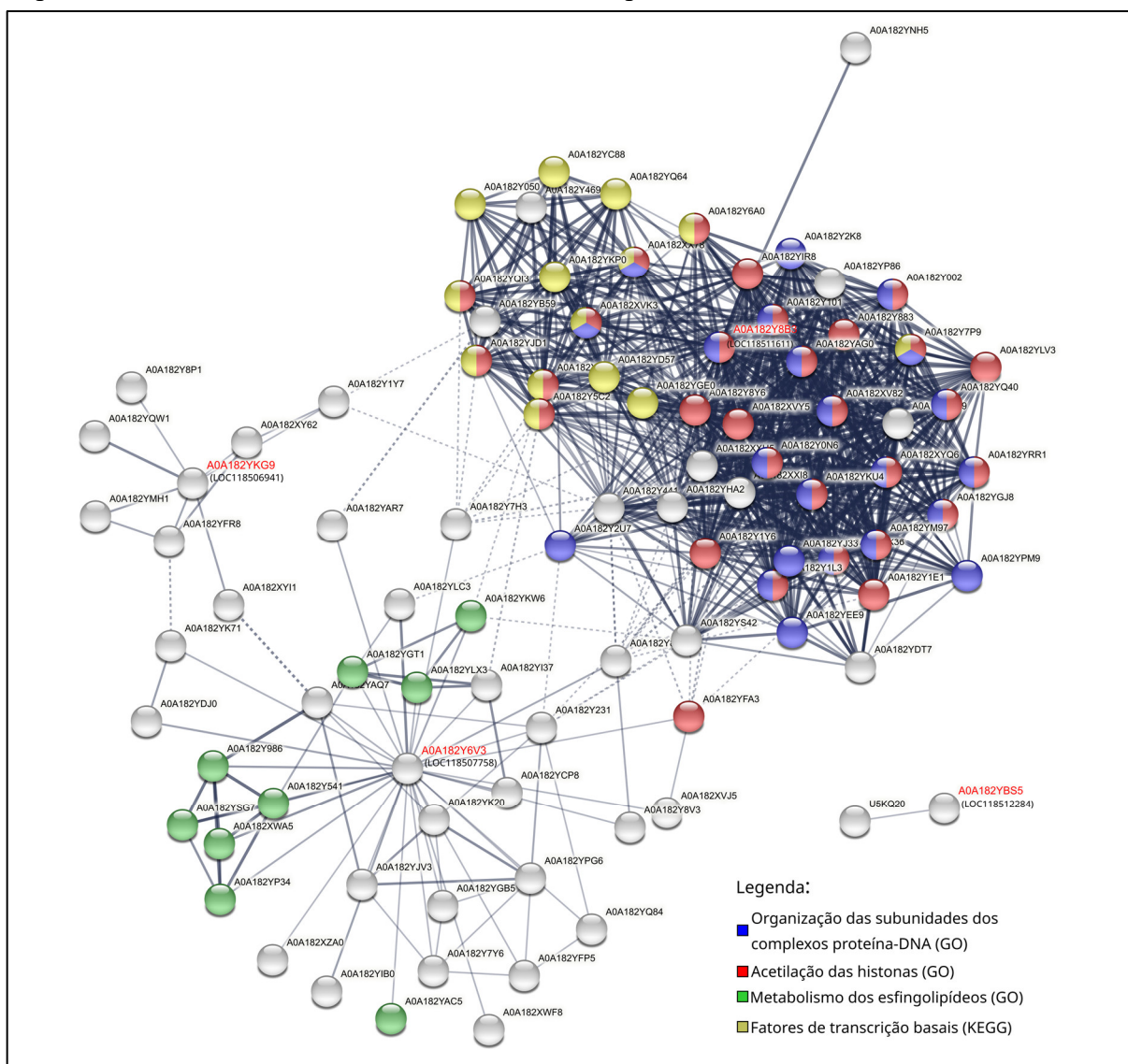
**Fonte:** elaborado pelo autor

**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam termos GO enriquecidos. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5).

Dentre esses seis genes de *An. coluzzii* com expressão fortemente associada a TEs, apenas um produz uma proteína não especificada, os outros 5 geram as proteínas: A0A6E8W673 (calmodulina), envolvida na transdução de sinal mediada por cálcio; A0A182LR05, envolvida no transporte transmembrana; A0A182KTE0, envolvida no processamento do mRNA; A0A182LHN7 (fator de iniciação da tradução 5) que catalisa a hidrólise do GTP vinculada ao complexo de iniciação ribossômica 40S; A0A182L3Z9 (Akirina), uma proteína que em *D. melanogaster* está associada à resposta imune contra patógenos. A rede de associações físicas e funcionais mostra um enriquecimento de diversas vias e processos biológicos, como *splicing* e regulação de *splicing* de RNA, iniciação da tradução e montagem do complexo de proteínas, assim como a via KEGG ribossomal (map03010). Também estão enriquecidos processos

relacionados à imunidade inata; à cascata de JNK, uma MAP quinase, relacionados à transdução de sinal dentro da célula; e também relacionados à transdução de sinal mediada por uma GTPase (enzimas que hidrolisam o GTP) monomérica (Figura 48).

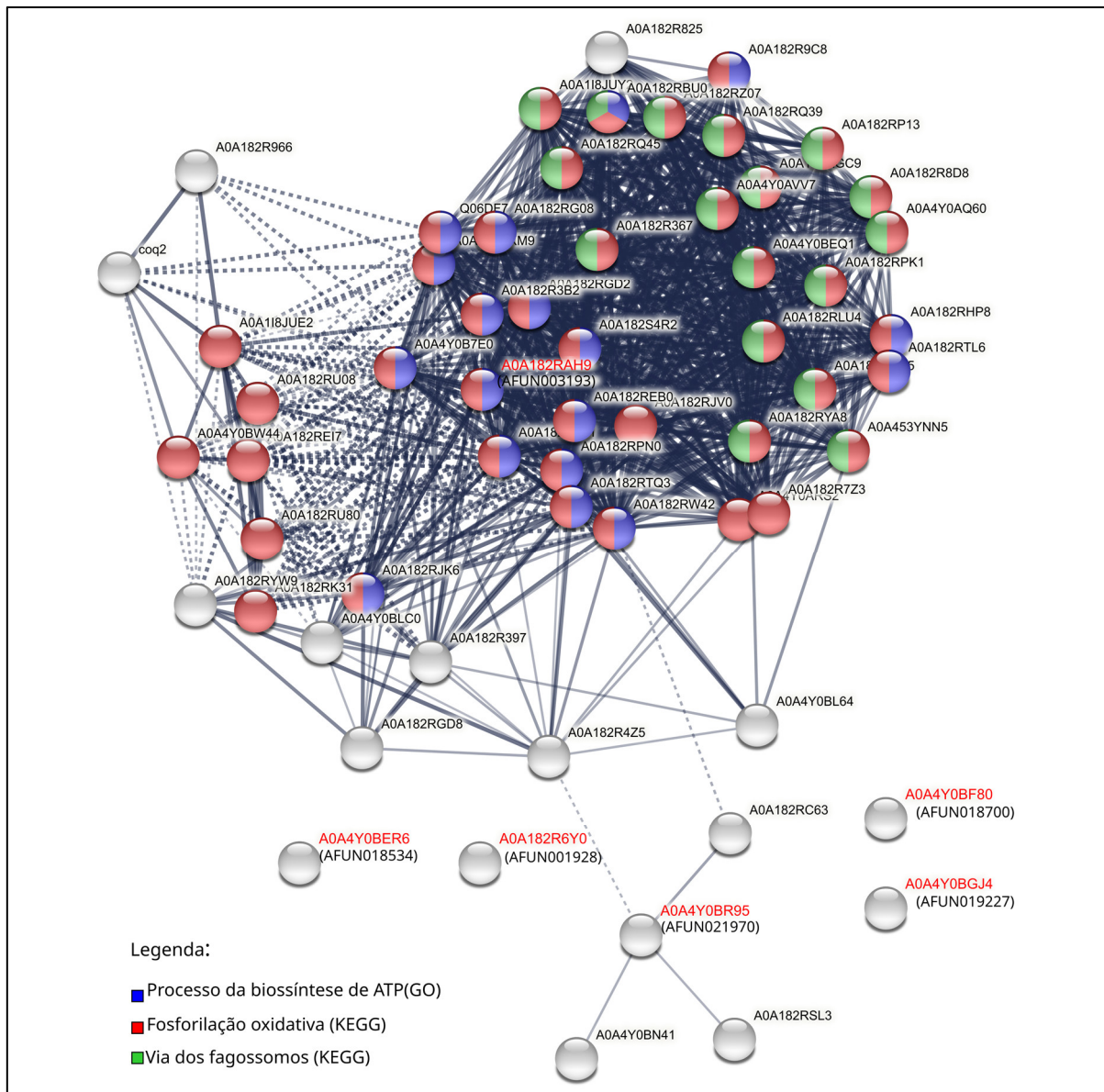
Figura 46 – Rede de associações físicas e funcionais das proteínas de *An. stephensi* cuja expressão é forte ou moderadamente associada à expressão de TEs.



**Fonte:** elaborado pelo autor

**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam vias KEGG ou termos GO enriquecidos. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5).

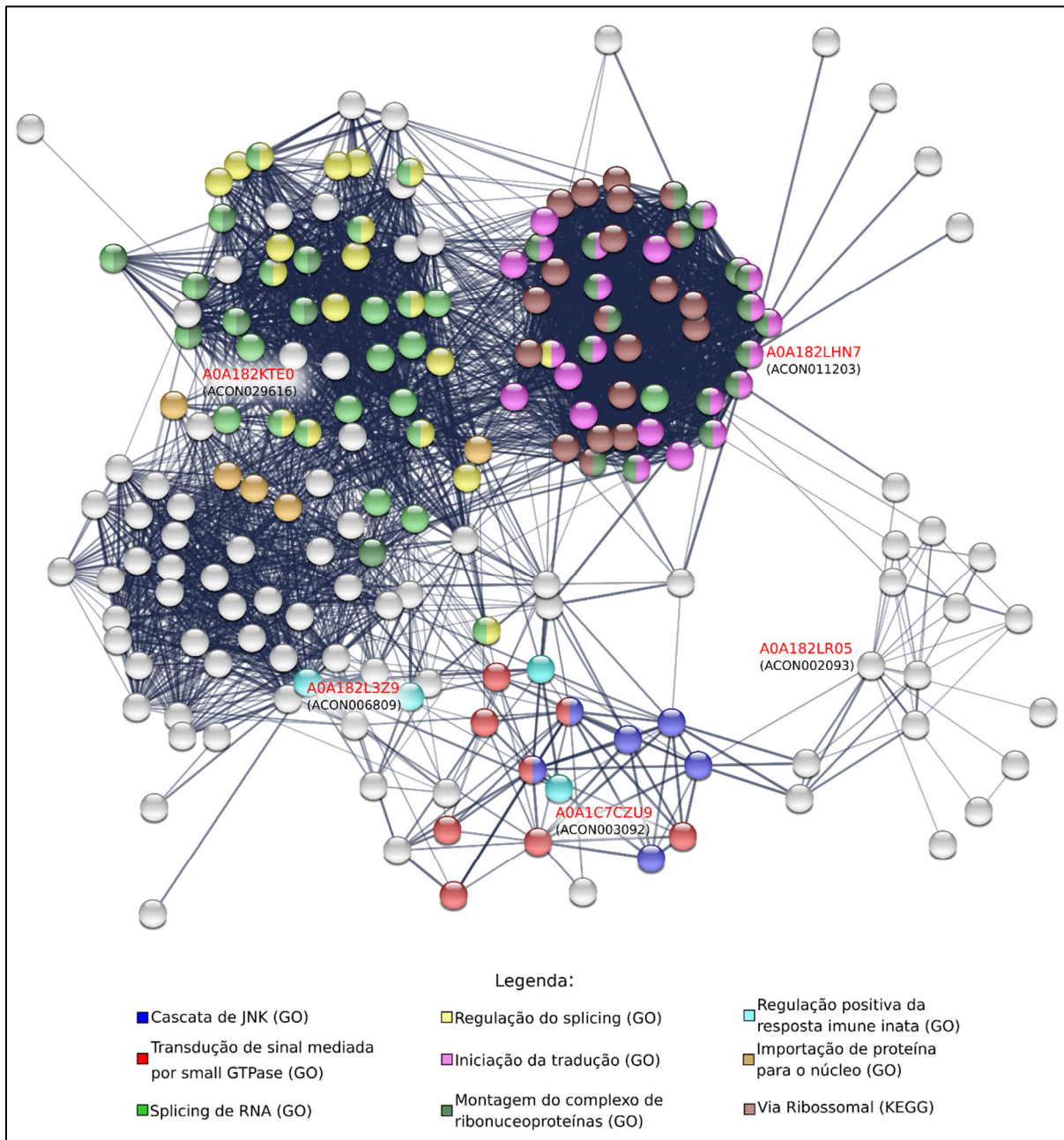
Figura 47 - Rede de associações físicas e funcionais das proteínas de *An. funestus* cuja expressão é fortemente associada à expressão de TEs.



**Fonte:** elaborado pelo autor

**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam vias KEGG e termos GO enriquecidos. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5).

Figura 48 – Rede de associações físicas e funcionais das proteínas de *An. coluzzii* cuja expressão é fortemente associada à expressão de TEs.



**Fonte:** elaborado pelo autor

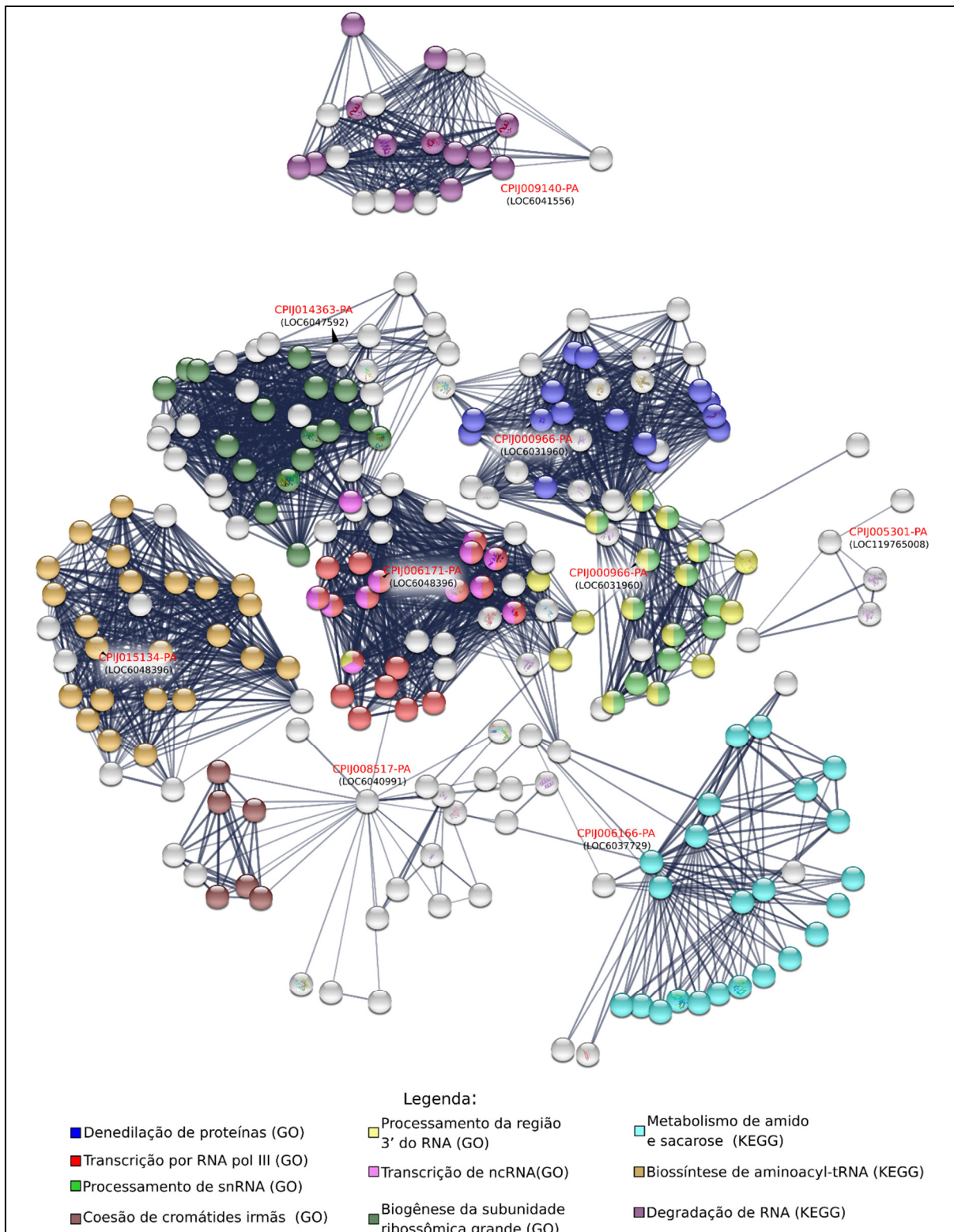
**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam vias KEGG e termos GO enriquecidos. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5).

Em *An. arabiensis* foram encontrados três TEs com expressão altamente correlacionada à expressão gênica, e quatro TEs com a expressão fortemente associada (outros 11 TEs estão moderadamente associados a genes). Dentre esses TEs com forte correlação a expressão de genes próximos, um deles se correlacionam com proteínas não caracterizadas, e outros seis com

os genes produtores das proteínas: A0A182HTX2, uma provável quinase GHMP, envolvida no processo metabólico da galactose; A0A182IAB9, uma ATPase, provavelmente envolvida no transporte de prótons e íons; A0A182I586, uma proteína com domínio de colágeno, constituinte da matriz extracelular estrutural; A0A182HMQ1, uma alanina transaminase; A0A182HQU1, uma proteína que contém um domínio de interação LIM (dois dedos contíguos de zinco separados por dois resíduos de aminoácidos hidrofóbicos), provavelmente envolvida em processos de regulação da transcrição pela RNA-polimerase II; A0A182HPH1, uma fosfatase envolvida em processos metabólicos dos fosfolípidios. Por não haver rede de interações das proteínas de *An. arabiensis* cadastrada no STRING, não foi possível inferir as vias ou processos biológicos enriquecidas, nesta espécie, que potencialmente estão sob influência em *cis* dos TEs.

Foi na espécie *Cx. quinquefasciatus* que foi encontrado o maior número de correlações fortes e moderadas, 12 e 25 respectivamente, entre a expressão de genes e TEs. A rede de associações físicas e funcionais (Figura 49) mostra um enriquecimento de diversos processos biológicos, como: denedilação de proteínas (GO:0010388), que é um processo de remoção de uma proteína semelhante à ubiquitina do tipo NEDD8 de uma proteína; transcrição por RNA polimerase III (GO:0006383); processamento de snRNA (GO:0016180), que abrange qualquer processo envolvido na conversão de uma transcrição primária de RNA nuclear (snRNA) em uma molécula snRNA madura; transcrição de RNAs não-codificantes (ncRNA) (GO:0098781); coesão de cromátides irmãs (GO:0007062), que é o processo do ciclo celular no qual as cromátides irmãs de um cromossomo replicado ficam amarrados uma à outra; Processamento da região 3' do RNA (GO:0031123), que abrange processos envolvidos na formação da extremidade madura de 3' de uma molécula de RNA; biogênese da subunidade ribossômica grande (GO:0042273); um processo celular que resulta na biossíntese de macromoléculas constituintes, montagem e arranjo de partes constituintes desta subunidade ribossômica, incluindo aí transporte para os locais de síntese proteica. Também estão enriquecidas as vias KEGG: Metabolismo de amido e sacarose (cqu00500), Biossíntese de aminoacyl-tRNA (cqu00970) e Degradação de RNA (cqu03018).

Figura 49 – Rede de associações físicas e funcionais das proteínas de *Cx. quinquefasciatus* cuja expressão é fortemente associada à expressão de TEs.



**Fonte:** elaborado pelo autor

**Legenda:** O código da proteína cuja expressão está associada a expressão de um TE está colorido de vermelho, o código do gene envolvido com essa proteína está mostrado entre parênteses. A espessura das arestas é proporcional ao número de evidências que associam essas proteínas. As colorações nos nós representam vias KEGG e termos GO enriquecidos. Apenas conexões de média e alta confiança são mostradas (*interaction score* > 0,5), conexões entre os clusters ao redor de cada proteína correlacionada com TEs foram ocultadas para facilitar a visualização.

Em *Ae. albopictus* não foi encontrada nenhuma correlação forte entre a expressão de gene e TE, apenas 13 correlações moderadas ( $0,4 < R^2 > 0,25$ ). Até recentemente a montagem disponível do genoma desta espécie encontrava-se muito fragmentada, dessa forma ainda há certa ausência de determinadas proteínas em algumas das bases: NCBI *protein*, STRING e UniProt *proteomes* (onde mais de 15% das proteínas essenciais da espécie estão ausentes). Por esses motivos, uma rede de associação das proteínas não foi extraída do STRING. Mas com base nas anotações disponíveis foi descoberto que esses genes com certa correlação a TEs produzem proteínas relacionadas com as funções de exportação de mRNAs contendo caudas poli(A) do núcleo para o citoplasma; glicosilação de proteínas, atividade de helicase e hidrolase; translocação de fosfolipídios; enovelamento de proteínas; proteínas que agem como componentes estruturais da cromatina; ou que desempenham um papel no *splicing* do pré-mRNA como um componente central das pequenas ribonucleoproteínas nucleares spliceossomais. Esse último papel é desempenhado por mais de uma proteína do grupo, de forma que 3 das 13 proteínas cujos genes possuem coexpressão com TEs participam da via KEGG da formação e atividade dos spliceossomos (map03040).

Após a verificação das vias associadas aos genes potencialmente influenciados por TEs, foi verificada a importância das proteínas produzidas por esses genes nas redes de interação proteína-proteína (PPI) de cada espécie mosquito. O conceito de *hub* numa rede PPI tem mais de uma definição, mas de forma simplificada um *hub* é uma proteína com alto número de conexões. Apesar disso, outras características como graus de centralidade podem ser utilizadas para definir um *hub*. Proteínas com alto grau de conexões e também com alta centralidade na rede não são tidas apenas como *hubs*, mas também como proteínas essenciais para o interactoma (VANDEREYKEN *et al.*, 2018).

Aqui as redes de interação proteína-proteína dos mosquitos foram obtidas do banco de dados STRING. Os graus e a centralidade de cada proteína das redes foram calculados e em seguida foi verificada a importância das proteínas produzidas pelos genes que se coexpressam com os elementos de transposição.

Em *An. minimus*, *An. dirus*, *An. merus*, *An. albimanus*, *An. funestus*, *An. gambiae* e *Ae. albopictus* as proteínas que se coexpressam com TEs não se comportam como hubs, ou foram eliminadas na construção da rede PPI devido ao baixo *score* de interação com outras proteínas.

Apenas uma das cinco proteínas (20%) que se coexpressam com TEs em *An. stephensi* é um hub. Essa, a proteína A0A182Y8B3, parece ser uma das proteínas chave no interactoma,



pois possui alto grau de conexões (132) e centralidade. Já em *An. coluzzii* esse número aumenta para três das 19 proteínas (15,19%), duas delas A0A1C7CZU8 e A0A182L4P7 também são proteínas chave no interactoma devido ao alto grau de conexões e centralidade. Em *Cx. quinquefasciatus*, quatro das 37 proteínas (10,81%) coexpressas com TEs são hubs na rede PPI desta espécie, no entanto nenhuma delas possui ao mesmo tempo as duas características de mais altos graus de conexão e centralidade.

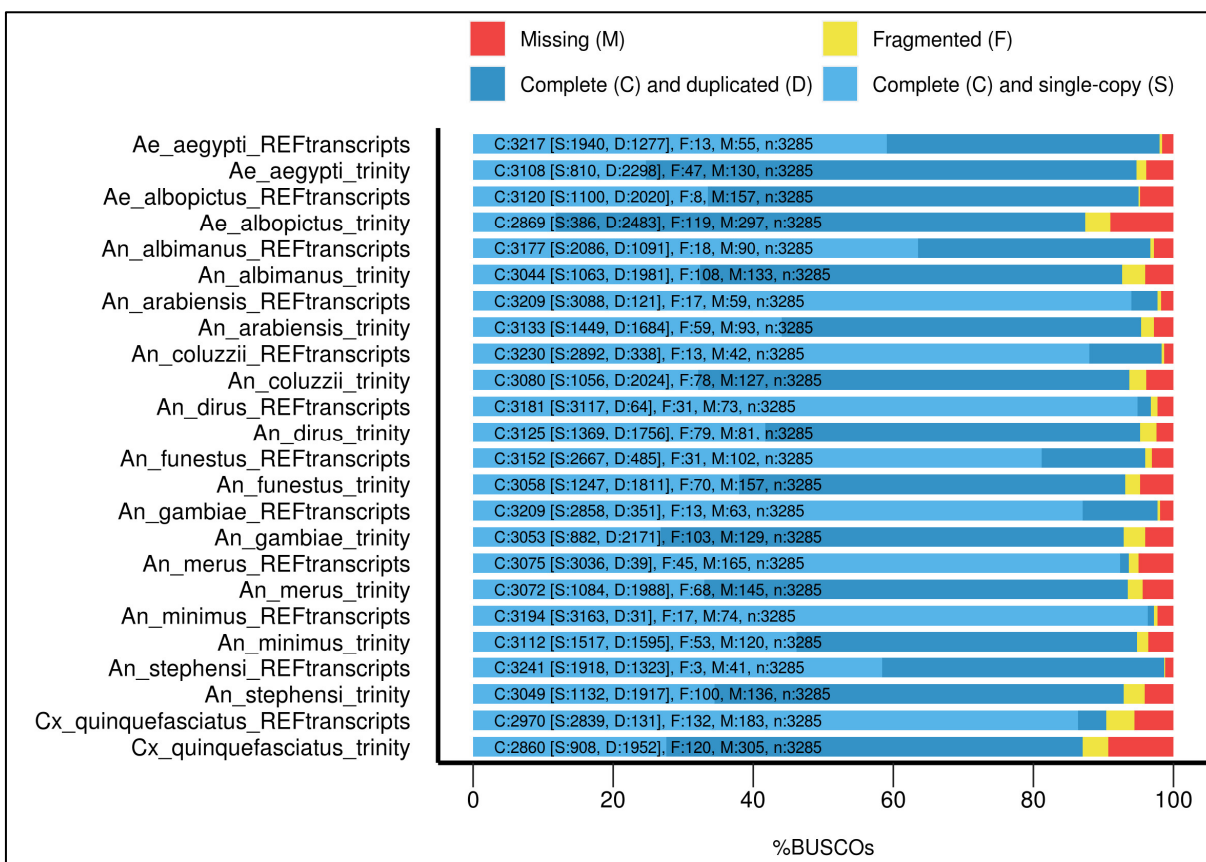
### 6.2.8 Remontagem dos transcritos de TEs

Embora várias famílias de TEs sejam expressas nas diversas espécies de mosquito, e embora algumas delas respondam a estímulos de estresse ou de infecção por patógenos, ou mesmo sejam reprimidas nos tecidos germinativos, uma pergunta que permanecia sem resposta: Os TEs expressos geram formas ainda capazes de codificar proteínas, ou geram apenas formas de RNAs não codificantes. Para solucionar essa questão, os transcriptomas das 12 espécies foram montados através de métodos *de novo*, porém utilizando-se das coordenadas genômicas para reduzir o número de artefatos e evitar a criação de grandes consensos de TEs ao invés de transcritos derivados de uma única cópia.

Para validar a completude da montagem dos transcriptomas foi utilizado o BUSCO. O resultado mostrou que todos os transcriptomas montados apresentaram uma alta completude, menos de 10% dos genes principais estão ausentes. Além disso, a integridade do transcriptoma montado foi muito semelhante ao transcriptoma de referência de cada mosquito (Fig. 50). A maior diferença entre o transcriptoma montado e o de referência, em termos de completude, foi observada em *Ae. albopictus*, porém essa diferença foi de apenas 7,6%. A completude alta é necessária para evitar a perda de muitos transcritos de elementos transponíveis.

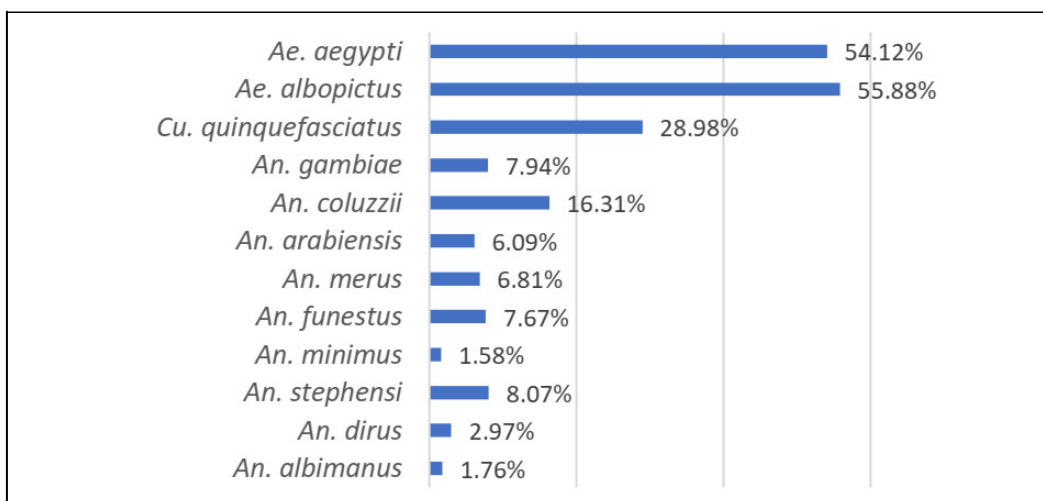
O número de transcritos derivados de TEs é altamente variável entre as espécies (Fig. 51). As três espécies da subfamília Culicinae apresentam grande parte do genoma coberto por elementos de transposição, e dessa forma apresentaram muito mais transcritos derivados de TEs que os membros da subfamília Anophelinae. Em *Ae. aegypti* e *Ae. albopictus*, mais da metade dos transcritos montados possuem relação com elementos de transposição. Esse número é muito menor em espécies pobres em TEs como *An. albimanus* e *An. minimus*, no entanto mesmo nessas espécies, ainda podem se encontrar TEs sendo transcritos.

Figura 50 – Completude dos transcriptomas montados e de referência verificada pelo BUSCO.



Fonte: Elaborado pelo autor.

Figura 51 – Fração do transcriptoma composta por transcritos relacionados a TEs

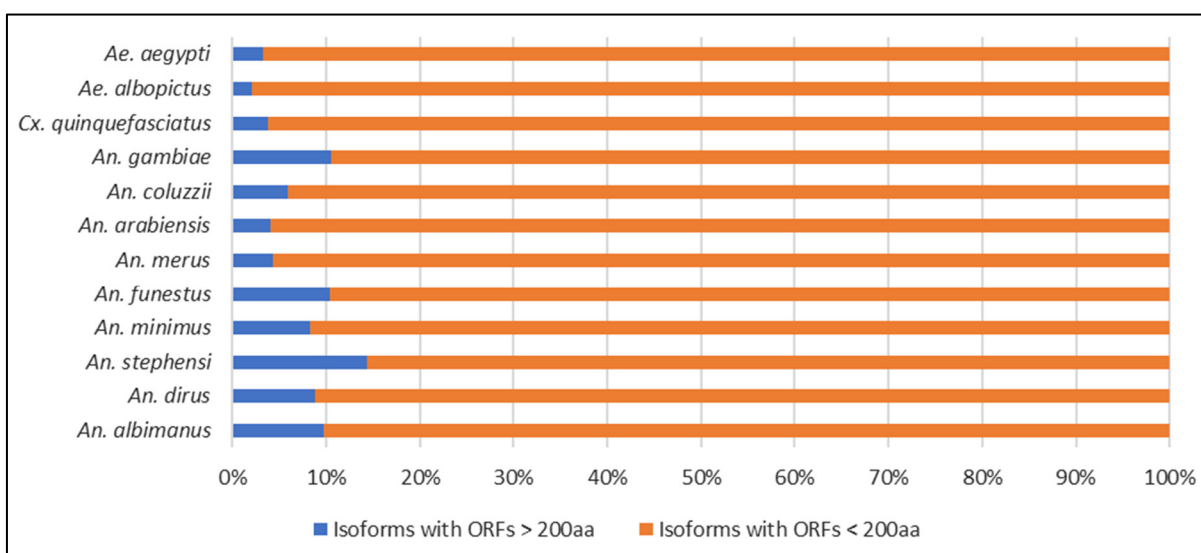


Fonte: Elaborado pelo autor.

A seguir foi investigado se essa grande quantidade de transcritos derivados de elementos de transposição são apenas transcritos ainda remanescentes de elementos já defectivos, gerando assim lncRNAs, ou se esses transcritos ainda possuem capacidade codificante. Foi observado

que para a maior parte das espécies, menos de 10% de todas as isoformas montadas de TEs ainda eram capazes de gerar ORFs maiores que 200 aminoácidos (Fig. 52). Embora pequena, essa proporção é similar em muitas espécies, indicando que a existência de TEs ainda potencialmente codificantes é comum nos mosquitos.

Figura 52 – Proporção de transcritos de TEs capazes de gerar ORFs significativas.

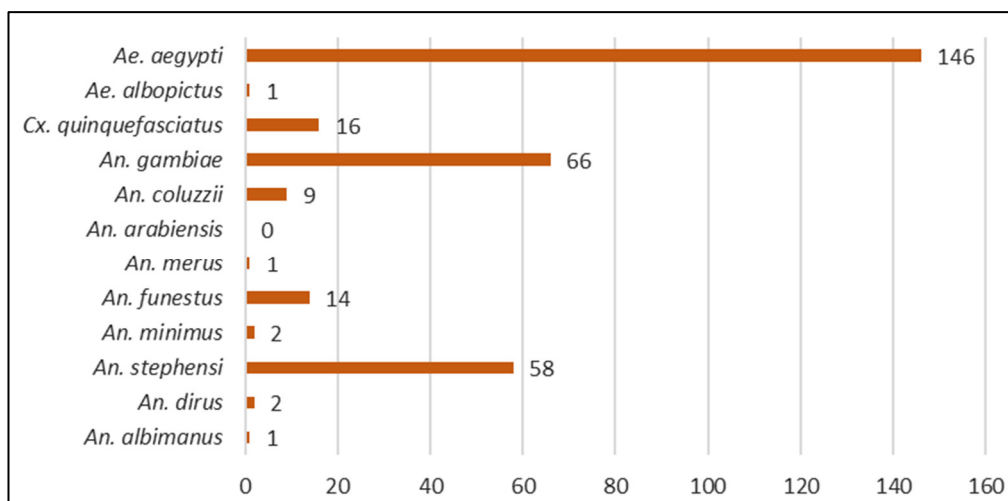


Fonte: Elaborado pelo autor

A próxima etapa foi então verificar quais desses elementos que ainda possuem a capacidade de codificar uma proteína potencialmente completa, e assim ser capaz de realizar a transposição de um elemento íntegro que esteja dentro do genoma. Foi encontrado um número muito pequeno de isoformas de TEs que potencialmente geram ORFs completas e que contém todos os domínios de elementos de transposição. Entre todos os mais de 500 mil transcritos relacionados a TE das 12 espécies de mosquito que possuem dados de transcriptoma, apenas pouco mais de 300 isoformas ainda possuem característica de um RNA mensageiro derivado de um elemento de transposição que é capaz de gerar uma proteína potencialmente completa (Fig. 53).

A espécie com o maior número de potenciais isoformas de TEs (146) também é a segunda espécie com o maior número de transcritos envolvidos com elementos de transposição (pouco mais de 205 mil), *Ae. aegypti*. No entanto a espécie com o maior número de transcritos envolvidos com TEs (pouco mais de 230 mil), *Ae. albopictus*, possui um número muito pequeno de potenciais isoformas completas. Entre os anofelinos, destacam-se as espécies *An. gambiae* e *An. stephensi*, com 66 e 61 potenciais isoformas completas, respectivamente.

Figura 53 – Número de potenciais isoformas completas de TEs dentre as ORFs preditas



Fonte: Elaborado pelo autor

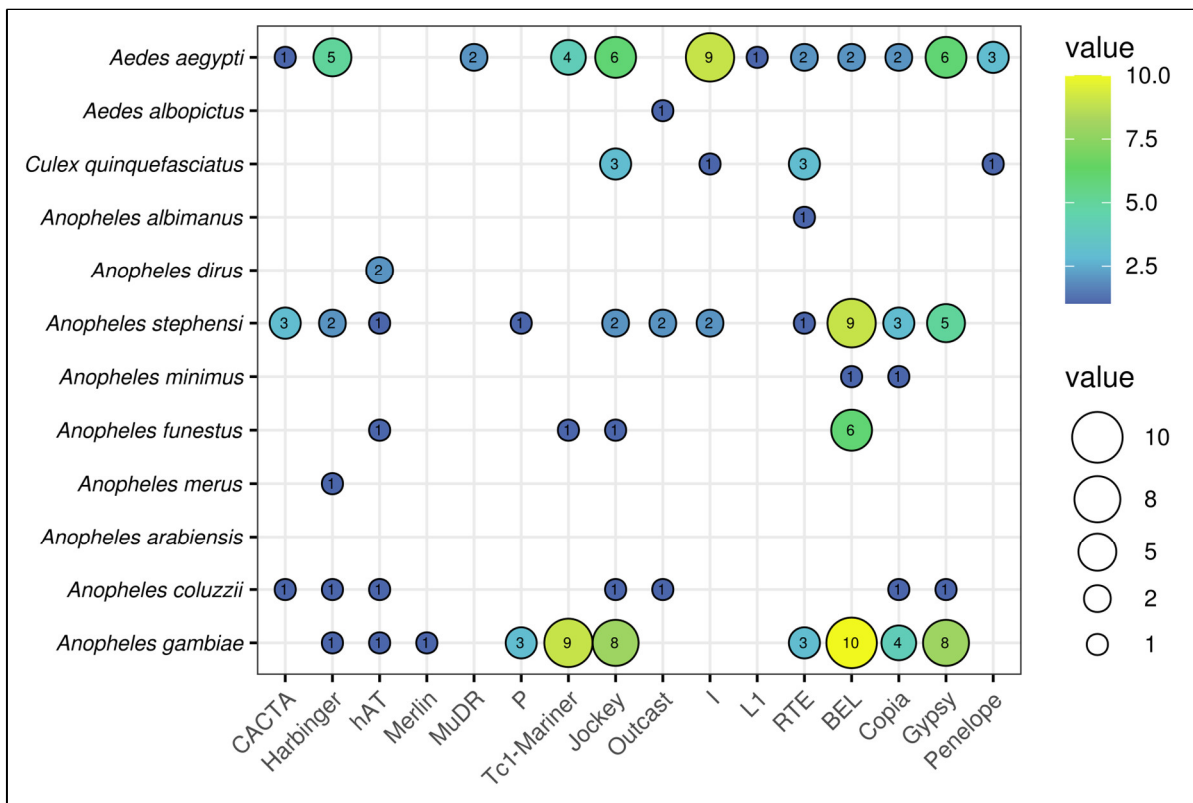
Dentre esse total de isoformas em cada espécie, algumas delas são derivadas da mesma família de TE. Dessa forma o número de famílias de TEs que ainda são capazes de produzir RNAs potencialmente completos é menor que o número total de isoformas. A Figura 54 mostra o número de famílias com potencial para gerar ORFs completas distribuídas entre as diferentes superfamílias de TEs das 12 espécies de mosquito. Percebe-se que em algumas espécies como *Ae. albopictus* e *An. albimanus*, apenas uma família de TE ainda pode gerar um transcrito potencialmente íntegro. Enquanto isso, em *Ae. aegypti* e *An. gambiae* 43 e 48 famílias ainda são capazes de gerar transcritos potencialmente íntegros, respectivamente, sendo esses os maiores números entre as espécies. Com relação às superfamílias, destacam-se os retrotransposons LTR *Gypsy* e *Bel-Pao* e o não-LTR *Jockey*.

### 6.2.9 Poucos TE chegam ao nível de expressão de proteínas

Ao final da cadeia de expressão, um gene no DNA da espécie é capaz de gerar proteínas. Isso também ocorre com os elementos de transposição, embora existam mecanismos do hospedeiro para silenciar os TEs. Para verificar a capacidade dos TEs dos mosquitos de produção de proteínas, foram investigados dados públicos de espectrometria de massas. Uma vez que métodos de proteômica de larga escala ainda não estão bem difundidos como os métodos de sequenciamento de RNA, o número de espécies com dados disponíveis caiu de 12 para 3 espécies (Fig. 55). No entanto, elas representam dois gêneros, e dentro do gênero

*Anopheles*, dois subclados distintos, fazendo com que represente bem o grupo dos mosquitos estudados.

Figura 54 – Famílias de potenciais isoformas completas de TEs dentro as ORFs preditas.



Fonte: Elaborado pelo autor

Legenda: A coloração e o tamanho dos balões são proporcionais ao número de famílias com isoformas completas dentro de cada superfamília de TE (eixo x). O valor dentro dos balões representa esse número de famílias.

As 3 espécies apresentaram proteínas de TEs expressas. No total foram encontradas 94 proteínas expressas de diferentes famílias de TEs somando-se todas as espécies. Dentre essas, destaca-se a expressão de TEs da ordem LTR, mais da metade (56,38%) das proteínas eram dessa ordem. Sendo 23 proteínas da superfamília *Gypsy*, 19 proteínas da superfamília *Bel-Pao* e 11 proteínas da superfamília *Copia*. Elementos não LTR também foram encontrados abundantemente expressos, 37,23% do total. Dentre eles todas as grandes superfamílias (utilizando a definição de Wicker *et al.* (2007)) foram encontradas expressas com pelo menos uma proteína, como no caso do grupo *R2*. As superfamílias de não LTR com proteínas mais expressas foram a *I* e a *Jockey*, com 12 e 11 proteínas expressas respectivamente.

Elementos de transposição de Classe II são pouco representados dentre o total de proteínas encontradas. Apenas 6 elementos desse tipo estão expressos, todos da ordem TIR.

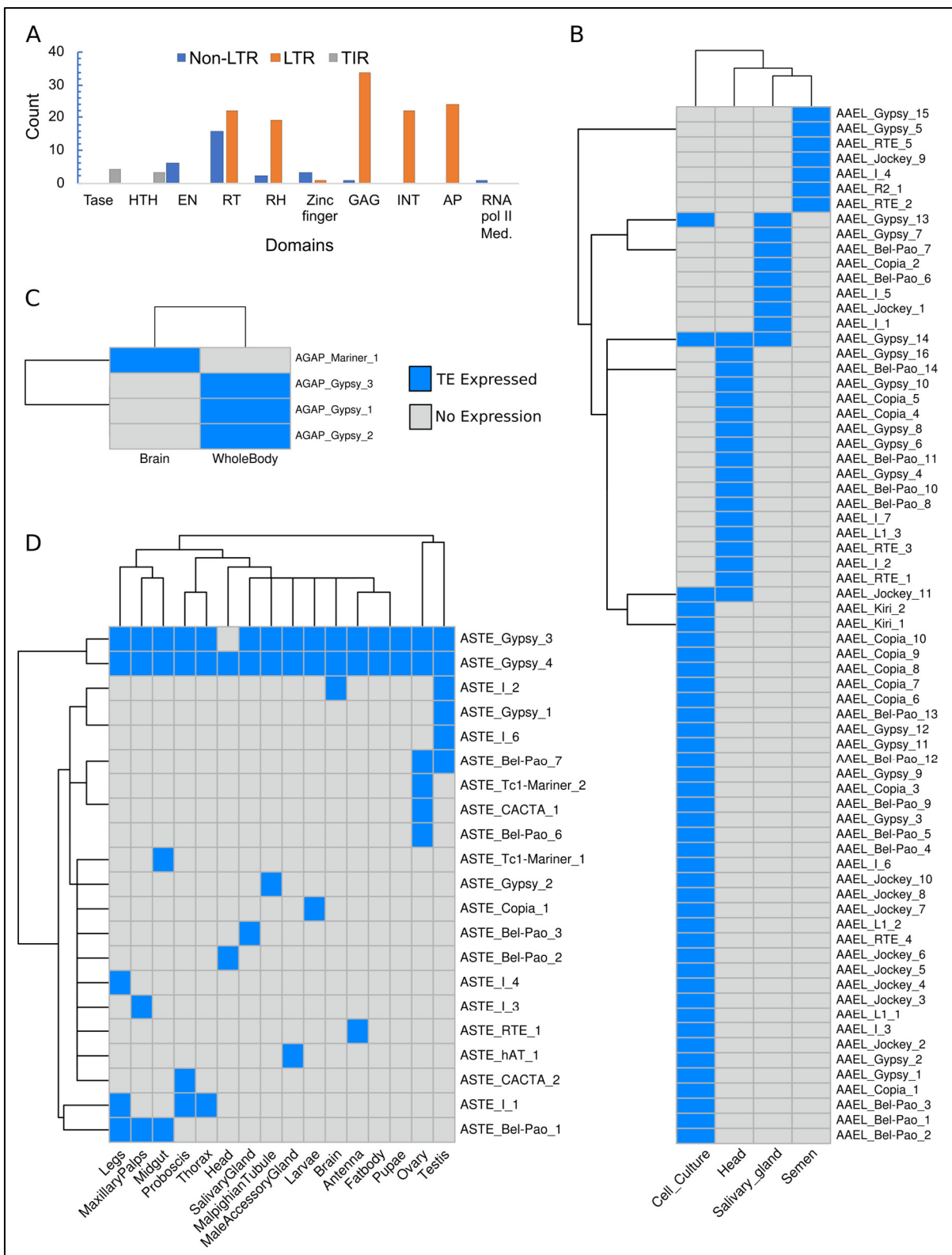
Eles são expressos apenas nos dois mosquitos do gênero *Anopheles*, e pertencem às superfamílias *Tc1-Mariner*, *CACTA* e *hAT*.

Nem todas as proteínas encontradas refletem elementos completos, a distribuição dos domínios proteicos é desigual (Fig. 55 A). Dentre os elementos LTR, o domínio mais comumente encontrado é o GAG, presente em 34 das 53 proteínas dessa ordem. Esse é o único domínio presente em 2 proteínas de TEs expressas em quase todos os tecidos do mosquito *An. stephensi* (Fig. 55 D). Dentre os elementos não LTR, a distribuição de domínios é ainda mais desigual. A maior parte das proteínas encontradas desses elementos possui apenas o domínio de transcriptase reversa (RT). Dentre os grupos proteicos analisados também foram encontradas 16 proteínas de TE sem domínios conservados, embora tenha sido confirmado que as ORFs sejam derivadas de elementos de transposição. A maior parte dessas proteínas sem domínios identificados são pertencentes a elementos não LTR, que são conhecidos por possuírem uma segunda ORF com função ainda não totalmente esclarecida.

*Aedes aegypti* foi a espécie com a maior quantidade de proteínas encontradas, 69 proteínas, embora houvesse apenas 4 tecidos analisados (Fig. 55 B). Em segundo lugar está o *An. stephensi*, com 21 proteínas encontradas (Fig. 55 D) e por último o *An. gambiae*, com apenas 4 proteínas encontradas (Fig. 55 C), embora esse baixo número possa ter alguma relação com a quantidade de tecidos disponíveis, apenas 2.

Analisando-se as três espécies, percebe-se que o padrão de expressão das proteínas derivadas de TEs é altamente tecido específico, com apenas uma proteína expressa em pelo menos três diferentes amostras de *Ae. aegypti* e três proteínas expressas em pelo menos três diferentes tecidos de *An. stephensi*, sendo duas delas expressas em quase todos os tecidos analisados. Outro padrão interessante encontrado quando *An. stephensi* é analisado é a presença de um perfil de expressão de TEs totalmente diferente nos tecidos germinativos, que se agrupam em um subclado próximo a raiz do dendrograma formado (Fig. 55 D).

Figura 55 – Proteínas de TEs identificadas por espectrometria de massas



Fonte: Elaborado pelo autor

Legenda: (A) Quantidade de domínios expressos nas proteínas de TEs encontradas nas 3 espécies, TASE = transposase, HTH = domínio de hélice-volta-hélice, EN = endonuclease, RT = transcriptase reversa, RH = RNase H, Zinc finger = dedos de zinco, INT = integrase, AP = endonuclease apurínica, RNA pol II Med = mediador da RNA polimerase II. As proteínas de TEs expressas em *Aedes aegypti* (B), *Anopheles gambiae* (C) e *Anopheles stephensi* (D).

## 7 DISCUSSÃO

Embora muito conhecimento tenha sido acumulado sobre o “ciclo de vida” dos TEs desde a revolução genômica, e se saiba que a abundância e diversidade do mobiloma no genoma de cada espécie é o resultado de uma corrida evolutiva sem fim entre TEs e hospedeiros (FRIEDLI; TRONO, 2015; PLATT; VANDEWEGE; RAY, 2018), ainda existem várias questões em aberto sobre o papel específico da transferência horizontal no rearranjo de TEs entre diferentes espécies eucarióticas e seu impacto nas espécies receptoras. Nesta tese, foi realizada uma caracterização detalhada do mobiloma em 24 genomas de mosquitos disponíveis, além disso, foram detectadas uma alta variabilidade do conteúdo de TE entre as espécies e uma série de transferências horizontais. Finalmente, a expressão dos TEs foi investigada e possíveis papéis funcionais foram identificados.

### 7.1 COMPOSIÇÃO DO MOBILOMA E TRANSFERÊNCIA HORIZONTAL

Os 24 genomas de mosquitos disponíveis durante a realização desse trabalho foram pouco estudados do ponto de vista do mobiloma. A caracterização de TEs, disponível para a maioria das espécies, geralmente data do momento em que a primeira descrição do genoma da espécie foi publicada. Uma exceção são as duas espécies modelo *An. gambiae* e *Ae. aegypti*, e também uma espécie de *Anopheles* encontrada no Brasil e maior vetor do *Plasmodium* spp. no Brasil, o *An. darlingi* (HOLT *et al.*, 2002; MARINOTTI *et al.*, 2013; NENE *et al.*, 2007). Para *An. gambiae*, uma abordagem baseada em homologia foi usada para comparar sequências de genoma com sequências e domínios de TEs conhecidos (HOLT *et al.*, 2002). Em *An. darlingi*, uma abordagem similar foi utilizada, foi feita uma predição ORF do genoma e estas foram anotadas por homologia com TEs presentes no NCBI (MARINOTTI *et al.*, 2013). Já o mobiloma de *Ae. aegypti* foi caracterizado pela primeira vez pelo RepeatScout e por homologia com elementos conhecidos (NENE *et al.*, 2007). Uma série de melhorias foram feitas desde que os genomas dessas espécies foram publicados, inclusive recentemente uma análise manual dos TEs de *An. darlingi* foi realizada, elevando de 2,3% para 5,6% a fração do genoma sabidamente composta por TEs (DIESEL *et al.*, 2019). No entanto, o mobiloma da maioria das espécies de mosquitos não foi extensivamente estudado, e embora algumas sequências de TE possam ser encontradas em bancos de dados, não há anotação padrão. Além disso, o uso de diferentes metodologias para detecção de TE (como exemplificado nos três casos anteriores) dificulta uma análise genômica comparativa em larga escala do mobiloma do mosquito. Na última década,



16 novos genomas de *Anopheles* foram publicados (NEAFSEY *et al.*, 2015), mas muito pouco se sabe sobre o mobiloma desses mosquitos, e os TEs anotados não foram disponibilizados para a comunidade científica.

Dois tipos principais de abordagem são usados para identificar TEs em sequências genômicas montadas: abordagens baseadas em homologia e métodos *de novo*. O método mais utilizado envolve a detecção de homologia entre TEs previamente caracterizados, normalmente obtidos de bancos de dados como Repbase, e as sequências recém-montadas (BAO; KOJIMA; KOHANY, 2015). O RepeatMasker é o software de primeira escolha para fornecer uma visão geral dos TEs em qualquer genoma através da comparação com bancos de dados (SMIT; HUBLEY; GREEN, 2015). Alternativamente, as abordagens *de novo* identificam TEs por suas características estruturais ou por características repetitivas/múltiplas no genoma do hospedeiro (GOUBERT *et al.*, 2015; XU; WANG, 2007; ZYTNICKI; AKHUNOV; QUESNEVILLE, 2014). Para melhorar a detecção e classificação padronizada de TEs, neste trabalho foi utilizada uma gama de softwares que realizam detecção *de novo* de TEs, assim como uma busca baseada em homologia. A isto se seguiu a classificação por três programas baseados em algoritmos distintos, PASTEC, TESorter e RepeatClassifier. Em consonância com outros estudos, as diferentes abordagens metodológicas geraram resultados diferentes, mas complementares (BERTHELIER *et al.*, 2018; GOERNER-POTVIN; BOURQUE, 2018). Se por um lado, abordagens *de novo* resultam em uma caracterização mais extensa do conteúdo de TEs no caso da maioria dos genomas de mosquitos, por outro lado, este método não recupera todo o mobiloma. Algumas famílias de TEs de baixa cópia e mais divergentes foram recuperadas apenas por busca de homologia. O impacto mais marcante da utilização de metodologias complementares em termos de conteúdo global de TEs ocorreu no caso do genoma de *Cx. quinquefasciatus*. Arensburger *et al.* (2010) mostraram que 29% do genoma de *Cx. quinquefasciatus* era composto por TEs. Um estudo adicional usando um método *de novo* baseado em pesquisa estrutural encontrou outros TEs não descritos no momento do sequenciamento deste genoma (MARSANO *et al.*, 2012). Este trabalho expandiu ainda mais a anotação de TEs para *Cx. quinquefasciatus* atingindo uma proporção de TEs de 43,55% e descreveu algumas superfamílias de TE que ainda não haviam sido descritas em algumas espécies de mosquitos, uma vez que relata pela primeira vez a ocorrência de elementos Penelope no genoma de *Cx. quinquefasciatus* e elementos semelhantes a *Crypton* no genoma de *Ae. albopictus*. *Cryptons* foram descritos pela primeira vez em fungos patogênicos (GOODWIN; BUTLER; POULTER, 2003) e mais tarde em stramenopilas e animais, incluindo

*Ae. aegypti* e *Cx. quinquefasciatus* (KOJIMA; JURKA, 2011). Assim, uma caracterização consistente do conteúdo de TEs de uma determinada espécie requer que abordagens *de novo* e baseadas em homologia (com diferentes bases de dados de TE de referência) sejam usadas para identificação e classificação de TEs.

No presente estudo, um conteúdo mais extenso de TE, além da detecção de superfamílias não descritas anteriormente podem ser explicados por vários fatores: a base de dados de TE de referência utilizada, a frequência com que é atualizada e a extensão em que a diversidade de sequências está disponível nesta base de dados. Além disso, a variedade de ferramentas e abordagens disponíveis empregadas em cada publicação é outro fator importante. Os resultados desta tese destacam o impacto de vários desses vieses no mesmo conjunto de genomas. Fica claro que a caracterização consistente do conteúdo de TEs de uma determinada espécie requer que ambas as abordagens, *de novo* e baseadas em homologia, sejam usadas para identificação e classificação de TEs.

A caracterização aprofundada do mobiloma realizada aqui nos permitiu investigar a influência dos TEs no tamanho dos genomas dos mosquitos. Foi observado que a fração do genoma do mosquito coberta pelo TE correlaciona-se significativamente com o tamanho do genoma, embora seja muito mais pronunciada em espécies do gênero *Aedes*. Esses resultados corroboram achados em artrópodes (PETERSEN *et al.*, 2019) e vertebrados (CHALOPIN *et al.*, 2015), no entanto, é interessante destacar que os mosquitos do complexo *Anopheles gambiae* apresentaram um maior tamanho de genoma para a sua fração do genoma composta por TEs do que as outras espécies de *Anopheles*. Isso sugere que outras sequências também são responsáveis pela expansão genômica, como microsatélites e duplicações segmentares, como mostrado para outras espécies (LU *et al.*, 2012; WANG *et al.*, 2013).

A influência de novos TEs invasivos, via transferências horizontais, tem sido proposta para vários elementos, como sugerido para BovB em genomas de vertebrados (BAIDOURI *et al.*, 2014) e An-RTEs em plantas (GAO *et al.*, 2018). Nesta tese, dada a alta variabilidade e distribuição desigual de várias superfamílias/famílias de TEs encontradas nos genomas de mosquitos, foi investigado o papel das transferências horizontais de transposon (HTT) na história evolutiva do mobiloma dos mosquitos. Mais de cinco mil HTTs foram relatadas entre eucariotos até agora, mas apenas quatro eventos de HTT envolvendo mosquitos foram bem documentados, enquanto 257 casos foram relatados entre drosofilídeos, que também pertencem à ordem Diptera (DOTTO *et al.*, 2018). Aqui, foi relatado um aumento de quarenta vezes (172) no número de eventos de HTT descritos entre mosquitos em relação aos dados anteriores da

literatura, ademais, várias hipóteses sobre a influência das características biológicas do hospedeiro e dos TEs nas HTTs foram investigadas. Um dos requisitos básicos para uma transferência de TEs para uma nova espécie hospedeira é a sobreposição espaço-temporal dos hospedeiros (WALLAU; VIEIRA; LORETO, 2018). Descobertas recentes sobre insetos apoiaram uma taxa de HTT mais alta entre *taxa* que compartilham o mesmo habitat (PECCOUD *et al.*, 2017). No entanto, em nossa análise refinada de genomas de mosquitos, não encontramos associação entre a faixa de habitat de espécies ancestrais e a taxa de HTT, mostrando que diferentes padrões podem surgir conduzindo a taxa de HTT em diferentes níveis taxonômicos do hospedeiro. A associação entre os tipos ou classes de TE e a taxa de transferência horizontal tem sido observada, mostrando que os transposons LTR e DNA transferem muito mais frequentemente do que os retrotransposons não LTR (GILBERT; FESCHOTTE, 2018; SILVA; LORETO; CLARK, 2004). Este trabalho descobriu, em mosquitos, que metade das superfamílias que sofreram HTT pertencem à ordem LTR de TEs, com um número quase igual de famílias de elementos TIR e não-LTR na fração restante. Este resultado difere um pouco de um outro estudo em insetos, que investigou um conjunto mais amplo de táxons e um período de tempo mais curto (últimos 10 MYA), e relatam uma maioria de TEs de Classe II envolvidos em eventos HTT seguidos por retrotransposons LTR e não LTR (PECCOUD *et al.*, 2017).

Outra característica intrínseca do hospedeiro que pode afetar a taxa de HTT é o parentesco filogenético das espécies hospedeiras. Vários estudos relataram que a transferência horizontal ocorre com mais frequência em espécies mais próximas (BAIDOURI *et al.*, 2014; PECCOUD *et al.*, 2017). No entanto, HTTs de longo alcance, entre táxons distantemente relacionados, também foram observados (GAO *et al.*, 2018; METZGER *et al.*, 2018). Estes ocorrem frequentemente com grupos de TEs específicos, como elementos de Classe II das superfamílias *Tc1-Mariner* (OLIVEIRA *et al.*, 2012) e *hAT* (GILBERT *et al.*, 2010). Os dados relatados aqui corroboram esses achados, mostrando que a maioria das HTTs de retrotransposons ocorreu entre espécies intimamente relacionadas do mesmo gênero, e as transferências de longo alcance entre espécies do gênero *Aedes* e *Anopheles* ocorreram principalmente com transposons de Classe II. Dos cinco TEs que sofreram transferência para espécies de metazoários ainda mais distantes, todos eram TEs Classe II da superfamília *Tc1-mariner*. Um padrão semelhante foi observado em um estudo que investigou HTTs em 195 espécies de insetos (PECCOUD *et al.*, 2017). Isso comprova que, em geral, HTTs ocorrem com mais frequência entre espécies intimamente relacionadas, mas tratando-se de transferências de longo alcance, há uma maior propensão do envolvimento de TEs de Classe II. O que provavelmente resulta da independência

de promotores e fatores de transposição do hospedeiro, tornando-os capazes de se transpor em um amplo conjunto de espécies hospedeiras divergentes.

Um aspecto interessante do grande número de eventos HTT encontrados entre os mosquitos é que ele está distribuído de forma heterogênea na árvore evolutiva desses insetos, mostrando uma prevalência muito maior de HTT entre as espécies do gênero *Aedes*. Embora haja um claro viés de mais genomas disponíveis do gênero *Anopheles*, ainda encontramos muito mais HTTs entre espécies do gênero *Aedes*, os maiores genomas de mosquitos sequenciados até agora. Além disso, apesar dessa influência do tamanho do genoma, não foi observado nenhum único caso de HTT envolvendo *Cx. quinquefasciatus*, uma espécie com tamanho de genoma intermediário (~600Mb), mas com um grande teor de TE (cerca de 44%). Estudos sem vieses, incluindo um conjunto diversificado de espécies de mosquitos, devem ser realizados para testar esses padrões, mas nossos resultados destacam que alguns táxons são mais propensos a trocar TEs horizontalmente do que outros, o que é corroborado por outros estudos de grande escala (REISS *et al.*, 2019).

Medir a influência direta de TEs transferidos horizontalmente é uma tarefa difícil, uma vez que é necessária uma anotação genética precisa de todos os genomas para associar inserções de TE com o impacto nos genes circundantes. No entanto, é possível estimar a proporção do genoma do hospedeiro que é derivado de TEs transferidos horizontalmente. Isso pode ser usado como um *proxy* para estimar o provável impacto/carga no genoma do hospedeiro. Neste trabalho foi descoberto que os TEs transferidos horizontalmente contribuem para o tamanho do genoma em vários graus, dependendo do genoma do mosquito. Para ambas as espécies do gênero *Aedes*, os TEs envolvidos em HTTs representam uma porção substancial do genoma (6,68–7,85%), o que representa cerca de 199 Mb no genoma de *Ae. albopictus* (2,5 GB). Isso é maior que o tamanho do genoma de *Drosophila melanogaster* (143,7 Mb) (THURMOND *et al.*, 2019) e do ácaro *Tetranychus urticae* (91 Mb) (GRBIĆ *et al.*, 2011). Esses resultados mostram que os TEs transferidos horizontalmente atingem um alto número de cópias, uma vez que se replicam sem controle do genoma do hospedeiro após sua invasão e podem, dessa forma, contribuir substancialmente para a expansão e estrutura do genoma do hospedeiro. Em consonância com isso, o tamanho do genoma também está correlacionado com a proporção do genoma coberto por TEs transferidos horizontalmente, confirmando que as HTTs contribuíram substancialmente para o tamanho do genoma dos mosquitos. Além disso, a datação dos eventos de HTTs sugere que a maior parte da invasão de TE em mosquitos ocorreu nos últimos vinte

milhões de anos, um período de tempo durante o qual a maioria das espécies de *Anopheles* sofreu especiação (TANG *et al.*, 2018).

Os resultados fornecem evidências de que HTTs geralmente ocorrem com mais frequência do que o esperado entre espécies intimamente relacionadas, mas alguns eventos HTT envolvendo hospedeiros mais divergentes também foram observados. Dois casos de HTT em mosquitos envolvem a mosca *H. irritans*, que também é um invertebrado sugador de sangue. Um desses casos envolve um elemento *Tc1-Mariner* da subfamília *irritans* que está envolvido em muitos eventos HTT em todos os filós. Curiosamente, foram encontradas várias cópias desses elementos em quatro conjuntos genômicos diferentes de *W. bancrofti*, um parasita nematoide transmitido por mosquitos que causa filariose linfática em humanos. Pelo menos um desses conjuntos foi reconstruído usando apenas sequências de DNA de vermes adultos isolados do sangue de um paciente humano infectado (GCA\_001555675.1), excluindo qualquer possibilidade de contaminação com DNA de mosquito. Para avaliar ainda mais essa possibilidade de contaminação nas montagens do genoma de *W. bancrofti*, uma busca utilizando o Blast foi realizada a partir dos quatro genomas disponíveis contra o genoma do *An. gambiae*. Isso confirmou que elementos com alta similaridade e com grande tamanho de alinhamento foram recuperados apenas para o elemento *Tc1-Mariner* (HTb). O conjunto dessas evidências confirma que esses elementos são verdadeiros componentes do genoma dessa espécie.

Ademais, foi observado que o elemento em *W. bancrofti* tem uma idade semelhante àquelas encontradas em espécies de *Anopheles* da Europa e Ásia, embora este elemento seja mais recente nas espécies de *Anopheles* africanas. Isso sugere que *W. bancrofti* adquiriu este elemento de espécies de *Anopheles* europeias e asiáticas e os transferiu, mais recentemente, para os *Anopheles* que habitam o continente africano. Considerando que as espécies de *Anopheles* não migram por longas distâncias, é possível que essas transferências intercontinentais de TEs tenham ocorrido por meio de vermes hospedados por animais migratórios, como pássaros (SUH *et al.*, 2016). Além disso, é sabido que esse verme filarial é transmitido por diferentes espécies dos gêneros *Aedes*, *Culex*, *Mansonia* e *Anopheles* (VAN DEN BERG; KELLY-HOPE; LINDSAY, 2013), sugerindo que *W. bancrofti* provavelmente mediou a transferência de elementos HTb entre mosquitos. Abundante evidência de troca de material genético entre vermes e hospedeiros corrobora esta hipótese (IMASE *et al.*, 2003; SUH *et al.*, 2016). A descoberta deste nematoide parasita reforça a evidência de espécies parasitárias como um importante agente de transferência de TE entre espécies eucarióticas multicelulares.

Outro caso interessante de HTT vem do retrotransposon não-LTR da superfamília *R2*. De acordo com a classificação de Wicker, a superfamília *R2* compreende quatro clados: *CRE*, *NeSL*, *R4* e *R2* (EICKBUSH; MALIK, 2002). Esse elemento se distribui de forma desigual entre os mosquitos, ocorrendo apenas nas espécies do gênero *Aedes* e nas espécies do complexo *An. gambiae*. Um estudo anterior sugeriu que esta distribuição é compatível com um evento de transferência horizontal (BIEDLER; CHEN; TU, 2015). A presente tese confirma este resultado, mostrando que os elementos *R2* provavelmente foram transferidos horizontalmente de *Ae. albopictus* ao ancestral do complexo de espécies *An. gambiae* e depois herdado verticalmente pelas seis espécies deste complexo.

## 7.2 EXPRESSÃO DOS TEs

De acordo com a literatura de elementos de transposição é esperado que a maioria das cópias de TEs não seja funcional. Contudo, nos dias atuais a visão reducionista original dos TEs como “DNA lixo” está ficando bem distante, e a comunidade está cada vez mais convencida de que os TEs representam elementos essenciais na regulação da expressão gênica (DRONGITIS *et al.*, 2019). Apesar disso, o primeiro passo na maior parte dos estudos é mascarar os elementos de transposição, devido a sua complexidade, de forma a tornar as análises focadas nos genes, o que deixa de mais da metade do genoma na escuridão (SLOTKIN, 2018).

Uma alta expressão e mobilidade dos TEs podem levar à instabilidade do genoma e impactos diretos no fenótipo do hospedeiro. Nos últimos anos, a expressão e os impactos da transposição dos TEs vêm sendo bastante estudados em humanos, principalmente por estar associado a diversos tipos de câncer. Em vários desses casos há formação de transcritos quiméricos de genes e TEs, geralmente TEs da ordem LTR, influenciando a oncogênese (JANG *et al.*, 2019).

Se por um lado os estudos sobre a expressão e os impactos dos elementos de transposição tem avançados em humanos, em outras espécies, o padrão de expressão dos TEs não é tão bem estudado. Isso se dá principalmente pela dificuldade em lidar com a natureza repetitiva dos elementos de transposição e principalmente pela falta de padronização e anotação de TEs na maior parte das espécies. Neste trabalho foi investigada a expressão de TEs em 12 das 24 espécies de mosquitos inicialmente estudadas. O primeiro passo foi realizar uma anotação de cada cópia de TE no genoma através de métodos específicos de anotação para TEs, rejeitando-se o método tradicional de anotação, onde o genoma de uma espécie era simplesmente

mascarado e múltiplas possíveis cópias de TEs, muitas vezes bastante fragmentadas, eram encontradas. Dessa forma, uma das contribuições desse trabalho é a disponibilização das anotações de TEs num formato amplamente utilizado, o formato GTF (*General Transfer Format*), o que permitirá que estudos futuros possam analisar não só os genes, mas também a expressão dessa grande fração do genoma, que são os elementos de transposição. Isto contribuirá para maximizar os resultados de trabalhos originais, que no geral são bastante caros em decorrência da alta quantidade de sequenciamentos realizados (SLOTKIN, 2018), uma vez que para avaliar a expressão de TEs precisa-se no geral de três fontes: I) software de análise, já desenvolvido por muitos grupos de pesquisas; II) anotação gênica, disponível em bancos de dados gerais como o NCBI, ou mais específicos, como o VectorBase no caso dos mosquitos; III) anotação dos TEs, essa é a mais difícil de se encontrar, pois está ausente para diversas espécies, ou feita de forma simplória para outras, através de métodos simples de mascaramento.

O resultado da análise de anotação dos TEs indicou que uma pequena fração das cópias de transposons, mas que ainda representam dezenas de TEs, possuem alguma sobreposição com genes dos mosquitos. Sobreposições ainda mais massivas de TEs nos genes de outras espécies já foram observadas, como por exemplo em fungos, onde TEs não autônomos são frequentemente encontrados em regiões gênicas (MUSZEWSKA *et al.*, 2019). Essas sobreposições podem impactar diretamente na expressão desses genes uma vez que a presença de um TE numa região pode afetar o estado da cromatina daquele local (DRONGITIS *et al.*, 2019). Além disso, os TEs que estão localizados na região antisense dos éxons dos genes (compondo mais da metade dos eventos de sobreposição de TEs e genes nos mosquitos) podem atuar regulando a expressão gênica a partir da formação dos transcritos antisensos, assim como foi observado em humanos, onde a maior parte dos locais de iniciação das transcrições relacionados a TEs está localizada em sentido oposto aos genes (CONLEY; MILLER; JORDAN, 2008). O restante das colocalizações representa sobreposições que ocorrem na mesma fita do DNA. Nestes casos percebe-se que os TEs influenciam estas espécies através da formação de novos éxons em alguns genes. Esse processo já é conhecido em algumas espécies, por exemplo, em *Arabidopsis thaliana* foi estimado que ~1,2% de todas as proteínas anotadas possuam influência de TEs em suas sequências (LOCKTON; GAUT, 2009), já em humanos estima-se que 5,8% dos genes possuem éxons envolvidos com transposons da ordem LTR (PIRIYAPONGSA *et al.*, 2007).

A presença de TEs coexpressos com genes próximos já foi observada em várias espécies, como em *Drosophila melanogaster*, quando exposta a diferentes condições de

estresse, onde a presença de um estressor aumentava ou reduzia a expressão de um gene e do TE que estava na sua vizinhança (OLIVEIRA *et al.*, 2021). A coexpressão de alguns TEs e genes próximos também foi observada no fungo patogênico *Zymoseptoria tritici*, onde cerca de 5% dos TEs, de ambas as Classes I e II, se coexpressam com genes da espécie (FOUCHÉ *et al.*, 2020). Neste caso, assim como no caso do retrotransposon *Hopscotch* no genoma do milho (STUDER *et al.*, 2011), é possível que a expressão dos TEs coexpressos sejam diretamente responsáveis pelo padrão de expressão de alguns genes (FOUCHÉ *et al.*, 2020).

Neste estudo foi verificada a coexpressão de TEs e genes vizinhos o que pode indicar uma possível influência em *cis* de TEs sobre genes. Para isso foi utilizado o pacote TEffectR. Como esse pacote possui a limitação de não tratar dados de diferentes experimentos para eliminar os *batch effects*, resultantes da regressão linear sobre experimentos da mesma espécie, mas inicialmente não relacionados, foi utilizado aqui o algoritmo ComBat (JOHNSON; LI; RABINOVIC, 2007). Ele foi escolhido pois, embora o TEffectR seja baseado no pacote *limma* (RITCHIE *et al.*, 2015), o ComBat teve um desempenho melhor do que a função `removeBatchEffect` do *limma* em estudos de Coexpressão de genes (VANDENBON, 2022). Além disso, embora existam algoritmos mais novos como o ComBat-seq (ZHANG; PARMIGIANI; JOHNSON, 2020) para correção de dados de RNA-Seq, o ComBat clássico foi escolhido por alguns motivos:

- a) O ComBat-seq é utilizado diretamente sobre a contagem bruta das *reads* mapeadas em cada estrutura do genoma. Ele utiliza uma transformação baseada numa distribuição binomial negativa (ZHANG; PARMIGIANI; JOHNSON, 2020). No entanto, um *benchmark* do efeito dessa transformação sobre as outras transformações realizadas nas etapas posteriores do TEffectR não foram avaliadas, e fogem ao escopo deste trabalho;
- b) O uso do ComBat-seq não trouxe melhorias significativas em relação a aplicação do ComBat (este é utilizado sobre as contagens de *reads* já normalizadas para a escala logCPM). O único momento em que qualidade dos resultados das transformações realizadas pelo ComBat-seq supera as transformações realizadas pelo ComBat foi observada em estudos cujo objetivo final eram análises de expressão diferencial (ZHANG; PARMIGIANI; JOHNSON, 2020). No entanto, a abordagem de expressão diferencial deste trabalho foi realizada com o uso do Tetranscripts e DEseq2, e não comparou a expressão de TEs entre estudos distintos, o que justifica o não uso de uma ferramenta de correção de *batch effects* naquela abordagem;



- c) Estudos anteriores focando em redes de coexpressão de humanos e camundongos indicaram que o ComBat-seq resulta em estimativas de coexpressão de qualidade inferior em comparação com o ComBat (VANDENBON, 2022).

Apesar da maioria das vias e processos enriquecidos nas redes construídas a partir das proteínas associadas aos genes coexpressos com TEs serem altamente variáveis, algumas similaridades entre espécies puderam ser observadas. Como exemplo temos vias de transdução de sinal mediadas por *small* GTPases, nas espécies *An. albimanus* e *An. coluzzii*; formação de ribossomos em *An. coluzzii* e *Cx. quinquefasciatus*; degradação de RNA em *An. dirus* e *Cx. quinquefasciatus*. Além disso, em múltiplas espécies os TEs estão associados a genes envolvidos em alguma etapa da transcrição, por exemplo, envolvimento com fatores de transcrição basais, *splicing*, silenciamento gênico, e transcrição de RNAs não codificantes.

Interessantemente, as espécies com o maior número de elementos de transposição não são as espécies que com o maior número de casos de coexpressão entre genes e TEs. Ao contrário disso, em *Ae. albopictus* não há nenhuma coexpressão forte, e em *Ae. aegypti*, nenhuma coexpressão forte ou moderada foi encontrada. Uma possível explicação para isso, nestes genomas onde a densidade de TEs é maior, pode ser o fato de que a expressão de um determinado gene pode ser influenciada por mais de uma cópia de TE, ou de inserções de TEs que estão em outra região fora dos limites dos 10 kb a montante dos genes (região avaliada).

Embora os resultados mostrados aqui não sejam suficientes para testar a ação direta da expressão em *cis* de TEs sobre os genes, a presença de TEs que se coexpressam como genes do hospedeiro em dezenas de corridas de RNA-Seq, e que além disso estão localizados 10 kb a montante ao gene e na mesma fita, justamente na região onde se localizam promotores e acentuadores ou isoladores da expressão, é um potencial indicativo de influência. Este trabalho fornece então uma lista de potenciais “alvos” (gene-TEs) para realização de futuras análises adicionais desses TEs próximos a esses genes, como a presença de evidência de seleção atuando no TE (SALCES-ORTIZ *et al.*, 2020).

Se considerarmos o total de 15% de proteínas marcadas como *hub* entre todas as proteínas nas redes PPI dos mosquitos ( $5800 \pm 551,74$ ), será obtido um número médio de 870 *hubs* por rede. No entanto, fica claro que dentre essas proteínas, um número ínfimo deriva de genes coexpressos com TE, e assim potencialmente influenciados em *cis*. Além disso, na maior parte das espécies nenhuma proteína envolvida com TEs se comporta como um *hub* da rede PPI. O

que leva a hipótese de que há uma seleção negativa agindo em nível de regulação da expressão, na ação em *cis* de TEs sobre genes de maior importância no interactoma dos mosquitos.

Embora alguns TEs componham as regiões expressas dos genes, a grande maioria (no geral mais de 95% das cópias) estão localizadas em regiões não codificantes. Mesmo assim, observou-se que a maior parte das famílias de TEs presentes nos mosquitos é expressa em nível de RNA. Esse resultado contrasta com uma visão de que a maior parte dos TEs é silenciada e apenas pouquíssimos elementos de transposição conseguem ser expressos (ZILBERMAN *et al.*, 2007). De fato, alguns estudos anteriores já indicavam que os TEs eram significativamente expressos em algumas espécies modelo, como por exemplo um estudo de CAGE-Seq em humanos, que mostrou que 20% das unidades transcricionais nesta espécie são iniciadas em elementos de transposição, o que deixa evidente a capacidade de TEs em iniciar a expressão gênica (CONLEY; MILLER; JORDAN, 2008). Contudo, uma avaliação da proporção de famílias expressas não havia sido ainda realizada.

Os resultados observados da expressão dos TEs nos mosquitos mostram um comportamento similar em relação aos níveis de expressão dos TEs observados em vertebrados, um valor baixo, 3,19% nos mosquitos contra 6,55% nos vertebrados (PASQUESI *et al.*, 2020), e com uma alta variabilidade entre espécies, com algumas espécies com abundância inferior a 1% e outras espécies com abundância superior a 10% do transcriptoma. Em camundongos, a expressão dos TEs também é bem menor que a dos genes, apenas 12-14% das *reads* de RNA-Seq de células embrionárias mapeiam em cópias de transposons (HE *et al.*, 2021). Percebe-se então, que no geral, para os eucariotos, a expressão dos TEs é pequena, porém ubíqua.

Em algumas linhagens do fungo patogênico *Z. tritici* a maior parte das famílias de TEs é expressa em condições naturais, com diferenças em sua expressão variando de acordo com a condição a qual o fungo é submetido. Independente disso, MITEs, TEs considerados não autônomos e sem proteínas codificantes, são altamente expressos em *Z. tritici* (FOUCHÉ *et al.*, 2020). Similarmente, MITEs são encontrados expressos em todas as espécies de mosquitos avaliadas, com expressão particularmente alta em *An. minimus*. Já que não há regiões codificantes nos MITEs, sua expressão nos organismos provavelmente indica a formação de RNAs longos não-codificantes.

Elementos de transposição não apresentam o mesmo nível de expressão em diferentes tecidos de um organismo. A expressão de TEs é tão distinguível e característica de cada momento e tipo celular, que em análises de *single cell* RNA-Seq em camundongos, o padrão

de expressão dos TEs pode ser utilizado para diferenciar tipos celulares. Cerca de 130 TEs estão especificamente expressos em tipos celulares distintos (HE *et al.*, 2021).

Em vários vertebrados, o nível de expressão de TEs é maior nos tecidos germinativos do que nos tecidos somáticos, embora em alguns casos a relação oposta seja observada. Em relação a esses tecidos germinativos, a expressão de TEs nos testículos é geralmente maior do que no tecido ovariano (PASQUESI *et al.*, 2020). Além disso, a expressão de TEs pode variar de acordo com a fase de desenvolvimento das células dentro do mesmo órgão germinativo, como a expressão de centenas de famílias de TEs que tem seu padrão alterado durante os diferentes períodos da espermatogênese em humanos (GUO *et al.*, 2018). Em *Drosophila buzzatii* e *D. koepferae* a expressão do transposon Helena é maior nos testículos que nos ovários, há aí então uma maior repressão nos ovários (ROMERO-SORIANO; GARCIA GUERREIRO, 2016). Os resultados observados no presente trabalho confirmam esse comportamento de alta repressão nos ovários (entre 8% e 46% dos TEs de uma espécie), o qual tinha sido observado anteriormente em alguns vertebrados e em algumas famílias específicas de TEs de *Drosophila spp.* No entanto, mostram um padrão contrário ao observado em algumas espécies de peixes, onde há uma maior quantidade de famílias superexpressas nos ovários em relação aos testículos (SONG *et al.*, 2019).

A repressão dos TEs nos testículos dos mosquitos, por outro lado, não é tão forte quanto nos ovários. Se em quatro espécies analisadas há mais famílias de TEs reprimidas que induzidas no testículo em relação à carcaça, em três espécies o padrão contrário é observado. Esse padrão de repressão mais forte nos ovários, que também ocorre em alguns vertebrados e em *Drosophila spp.* pode ser produzido uma maior produção de piRNAs, neste tecido, que silenciam os transcritos de TEs, assim como ocorre em *D. melanogaster*, *D. arizonae* e *D. mojavensis*, onde os piRNAs são superexpressos nos ovários quando comparado a expressão desses piRNAs nos testículos (BANHO *et al.*, 2021; SAINT-LEANDRE *et al.*, 2020). Apesar disso os piRNAs não são a única fonte de controle dos transcritos de TEs, como mostram os experimentos em *Paralichthys olivaceus*, na qual não há expressão de genes relacionados a via de piRNA nos ovários, mas ainda assim alguns TEs estavam reprimidos neste tecido em relação aos testículos onde os genes a via de piRNA estão expressos (SONG *et al.*, 2019).

No entanto, não há no momento dados que confirmem a hipótese de um silenciamento mais forte nos ovários dos mosquitos como resultado de uma elevada expressão de piRNA. Recentemente foi mostrado que há uma diferença considerável entre os piRNAs de *Drosophila*

e de mosquitos, enquanto em *Drosophila* os piRNAs são expressos apenas nas gônadas, em mosquitos os piRNAs são expressos em ambos os tecidos somáticos e germinativos (MA *et al.*, 2021). Além disso, a maior parte dos piRNAs dos mosquitos não parece mapear em TEs, estimativas de mais de 5 anos atrás mostravam que apenas 19% e 30% dos piRNAs de *Ae. aegypti* e *Ae. albopictus* mapeiam em transposons, respectivamente. Um número bem menor em comparação com *D. melanogaster* (uma espécie com uma menor proporção de TEs), onde 51% dos piRNAs mapeiam em TEs (ARENSBURGER *et al.*, 2011b; LIU *et al.*, 2016). Estudos mais recentes em versões menos fragmentadas dos genomas dos mosquitos sugerem que essa fração de piRNAs mapeando em TEs é ainda menor. Isso mostra que, mesmo fazendo parte da ordem Diptera, há diferenças substanciais entre essas duas espécies (MA *et al.*, 2021).

Adicionalmente, embora os mosquitos sejam espécies mais próximas do gênero *Drosophila* que do gênero *Paralichthys*, há uma ausência de experimentos de RNA-Seq de única origem, realizados na mesma linhagem e abrangendo, ao mesmo tempo, sequenciamento de pequenos RNAs e mRNAs em ovários, testículos e em algum outro tecido somático de mosquitos. Experimentos como esse poderiam indicar se as famílias de TEs cuja expressão está reduzida em tecidos germinativos são alvos de piRNAs, miRNAs, siRNAs ou nenhum desses RNAs regulatórios. O uso de dados públicos de experimentos de RNA-Seq de diferentes fontes, para testar a hipótese de que a forte repressão dos TEs nos ovários seja resultado de uma maior expressão dos piRNAs, pode causar enviesamento dos resultados. Sendo necessário um estudo específico para essa finalidade.

O número de transcritos derivados de TE também pode ser dependente da idade para alguns organismos (LI *et al.*, 2013a) estando superexpressos em organismos mais velhos (ELSNER; MEUSEMANN; KORB, 2018) ou variar em resposta ao estresse (OLIVEIRA *et al.*, 2021). A expressão de TEs em situações de estresse é matéria de debate. A grande maioria dos estudos indica que os TEs têm sua expressão aumentada sob condições de estresse. Contudo, muitos desses estudos focam em apenas poucos tipos de TEs em uma única espécie, como no caso do fim da repressão do elemento *L1* em camundongos sob condições de estresse pelo envelhecimento (VAN METER *et al.*, 2014); o aumento da expressão de MERV-L e IAPez em células de camundongo submetidas a um estresse de exposição a uma temperatura acima do normal (39.5 °C) (HUMMEL *et al.*, 2017); o aumento da expressão do TE EARE-1 (um *Copia*) em diversos tecidos da planta *Excoecaria agallocha* exposta a diversas condições de estresse (HUANG *et al.*, 2017). Outros estudos, embora não foquem em um TE específico,

não verificam o comportamento de todos os TEs da espécie, como no caso de algumas famílias superexpressas em *Pinus sylvestris* (VORONOVA *et al.*, 2014).

O uso de métodos de larga escala, como *microarrays*, para avaliar a expressão de TEs, mostrou que esta expressão era apenas moderadamente afetada pelo calor, com números quase similares de TEs induzidos e reprimidos em *Arabidopsis thaliana* (PECINKA *et al.*, 2010). Uma recente revisão da literatura identificou que nem sempre os TEs são ativados por situações de estresse, e que essa é uma relação complexa (HORVÁTH; MERENCIANO; GONZÁLEZ, 2017). Os resultados desta tese trazem mais conteúdo para esse debate, os resultados da exposição a situações de estresse nos mosquitos mostram os mais diversos cenários: I) há situações de estresse que não influenciam a expressão dos TEs, assim como ocorre em *Drosophila suzukii* (MARIN *et al.*, 2021) e *D. melanogaster* exposta a compostos químicos (STOFFEL *et al.*, 2020), II) embora haja TEs superexpressos, o número de famílias reprimidas foi mais alto que o número de famílias induzidas, III) apenas em condições de estresse extremas, onde o mosquito encontra-se perto da morte, é que o número de TEs induzidos superou o número de TEs reprimidos.

Os experimentos realizados aqui mostraram que diferentes tipos de estresse podem gerar respostas de TEs mais fortes ou mais fracas. Mostraram também que o padrão de expressão é dependente da intensidade do estresse (como o estresse de salinidade mostrado). Esses dados, em conjunto com dados de outras espécies, como os de células tronco hematopoiéticas humanas, que mostram que a expressão dos TEs varia de acordo com o tempo após exposição a um composto químico <sup>8</sup> (CLAPES *et al.*, 2021), fortalecem a ideia que os TEs não são sempre ativados frente a um estresse abiótico, mas sua indução ou repressão dependem: I) do tipo de estresse, II) da intensidade do estresse, III) do momento após a exposição ao estresse no qual o RNA é coletado.

Considerando a infecção com vírus/parasitos como uma condição de estresse biótico para o mosquito, também fica claro que alguns cenários não causam grande alteração da expressão dos TEs. Ainda que a infecção por *Brugia malayi* cause a indução da expressão de alguns TEs, isso representa menos de 1% do total de famílias de TEs em *Ae. aegypti*. Além disso, embora a dinâmica de variação da expressão gênica seja maior nos primeiros dias após a alimentação com sangue infectado (CHOI *et al.*, 2014) a maior variação da expressão dos TEs é observada

---

<sup>8</sup> Inicialmente não há nenhuma variação na expressão dos TEs, seguida do aumento da expressão de determinadas famílias em tempos intermediários, culminando com a repressão de outras famílias em tempos posteriores.

dias após a infecção, com o dobro de TEs reprimidos em relação aos induzidos. Isso pode indicar que os TEs provavelmente não participam de nenhuma via regulatória do sistema de defesa de *Ae. aegypti* frente à infecção por *B. malayi*.

Os transcritos de TEs também quase não sofrem variação diante da infecção pelos vírus ZIKV, CHIKV, e DENV na glândula salivar ou pelo vírus CHIKV no intestino médio dos mosquitos. Embora tenha sido reportado uma baixa variação da expressão gênica neste último caso (71 diferencialmente expressos (DE) em 1 dpi e 78 DE em 2 dpi) (DONG; BEHURA; FRANZ, 2017), a variação nos TEs é mínima (3 DE em 1 dpi e 0 DE em 2 dpi). Similarmente, em outros insetos como a mariposa *Trichoplusia ni*, a infecção viral também causa muito pouca variação na expressão dos TEs no intestino médio em 1 dpi e 2 dpi, ficando a maior parte da variação da expressão dos TEs restrita ao 3º dpi (MULLER *et al.*, 2021). Em *Ae. aegypti*, apesar do baixo número de TEs DE na glândula salivar de durante a infecção viral (nove TEs DE na infecção por CHIKV, dois na infecção por ZIKV e três na infecção por DENV), o número de genes diferencialmente expressos nesse tecido é alto, 966 DE na infecção por CHIKV, 396 DE na infecção por ZIKV e 202 na infecção por DENV (CHOWDHURY *et al.*, 2020), o que elimina a possibilidade de ausência de ativação de processos biológicos frente a infecção viral. Portanto, os TEs não parecem participar de vias regulatórias e nem a infecção pelos vírus ZIKV, CHIKV, e DENV parece ser capaz de ativar vias de repressão de TEs em tecidos como intestino médio e glândula salivar, dois órgãos onde há intenso contato viral com as células dos mosquitos. Além disso, a variação da expressão gênica e o comportamento estável da expressão dos TEs indicam que sua transcrição não é uma transcrição *readthrough* a partir da transcritos de genes.

Ainda que a variação da expressão dos TEs no intestino médio e na glândula salivar seja mínima, as análises indicaram haver alteração na expressão em outros órgãos ou tecidos, uma vez que em experimentos realizados com macerados de corpo total dos mosquitos resultaram num número maior de TEs diferencialmente expressos. A infecção de *Ae. aegypti* com ZIKV, por exemplo, provoca uma leve alteração na expressão dos TEs (pouco mais de uma centena de TEs), com valores de indução e repressão similares, enquanto há diferença de expressão de 1332 genes durante o curso da infecção. É possível que esses TEs componham parte do total de cerca de 500 lncRNAs que também estão diferencialmente expressos durante a infecção com o ZIKV, pois estes seguem a mesma dinâmica de expressão diferencial que os TEs, com uma menor variação em 2 dpi, seguindo de um grande aumento da variação em 7 dpi, que é reduzida em 14 dpi, mas ainda se mantendo num patamar mais alto que 2 dpi (ETEBARI *et al.*, 2017). Já os primeiros momentos da infecção com DENV provoca uma variação considerável na

expressão dos TEs. Esses dados estão em linha com o padrão de resposta em humanos, onde alguns tipos de vírus quase não provocam alteração na expressão dos TEs, enquanto outros alteram bastante sua expressão (MACCHIETTO; LANGLOIS; SHEN, 2020). Em *Ae. aegypti*, apenas 3 genes estão DE 3 horas pós-infecção (hpi) (KANG *et al.*, 2018), enquanto 196 TEs estão DE no mesmo período, com uma forte repressão na sua expressão. Logo, é possível que os TEs participem de vias regulatórias durante os primeiros momentos da infecção com DENV. Entretanto, dados de estudos em *Drosophila simulans* sugerem que a redução da expressão dos TEs, que ocorre durante infecções virais, é uma forte modulação decorrente da resposta antiviral mediada por pequenos RNAs (ROY *et al.*, 2020). Considerando esses cenários, não podemos distinguir entre essas duas hipóteses, pois embora o número de genes diferencialmente expressos nesse período seja quase irrelevante, descartando ativação de grandes cascatas imunológicas, uma análise de expressão de pequenos RNAs é necessária para testar quaisquer das duas hipóteses.

Considerando os diversos tipos de estresse, mostramos que o argumento, amplamente aceito, de que os TEs são mais expressos sob condições de estresse não é válido para todas as espécies/condições. Também fica claro que dados de larga escala, com enfoque no comportamento de todas as famílias de TEs podem dar respostas diferentes de estudos que focam em tipos específicos de TEs. Estes estudos provavelmente criaram um viés por serem reproduzidos muitas vezes nas mesmas espécies modelo e focando em tipos específicos de TEs.

Uma questão a ser respondida era se as moléculas de RNA geradas a partir dos TEs derivavam de cópias intactas de TEs ou se derivavam apenas de versões truncadas de RNA de TEs. Para reconstrução das cópias foi utilizado o software Trinity, numa versão guiada pelo genoma. Neste modo o software faz uma montagem *de novo* do transcriptoma, ou seja, ele capta as variações da linhagem sequenciada, evitando que apenas a sequência do genoma de referência seja considerada. Contudo a montagem não é realizada entre todas as *reads*, mas apenas nas que mapeiam em regiões próximas no genoma, isso evita que *reads* de cópias similares do mesmo TE, espalhadas pelo genoma, sejam montadas numa única cópia (HAAS, 2020). Os resultados indicaram que embora expressos em baixa abundância nos mosquitos, o número de isoformas geradas pelos TEs é bastante variável entre as espécies. Compondo mais de um quarto de todos os transcritos nos mosquitos da subfamília Culicinae, e no geral menos de 10% nos anofelinos, o que é um reflexo da maior quantidade de cópias de TEs nos mosquitos da subfamília Culicinae. Ainda assim a diferença entre os níveis de expressão e a diversidade

de transcritos pode ser bastante alta, em *Ae. aegypti* por exemplo, os TEs representam apenas 2,69% de todas as *reads* sequenciadas, mas representam cerca de 54% dos transcritos montados.

A dificuldade/impossibilidade de anotação de parte dos transcritos montados em estudos de montagem de transcriptoma *de novo*, possivelmente deve-se em grande parte à presença de transcritos de elementos de transposição. A grande fração de TE no transcriptoma de *Ae. aegypti* é similar a fração de 55% de *contigs* não anotados no transcriptoma (montado com Trinity) de *Ae. fluviatilis* (CARAGATA *et al.*, 2017). No artigo, no entanto, eles relatam que encontraram apenas 61 *contigs* de TEs (0.1% do total de transcritos), entretanto os autores utilizaram a base NR, que é sub-representada em TEs (Alguns desses TE estão com sua expressão alteradas nas diferentes condições). O uso de uma base específica de elementos de transposição para a anotação dos transcritos montados revelou que 1,9% do transcriptoma de *Pomacea canaliculata* é formado por TEs (SUN *et al.*, 2012). Essa proporção ascende para 3% dos transcritos no nematódeo *Heterodera avenae* (KUMAR *et al.*, 2014). E eleva-se ainda mais no anfíbio *Rhinella marina* para ~8,2% dos transcritos (LUDWIG *et al.*, 2021). Mas representa apenas 0,45% dos transcritos de *An. funestus* (FERNÁNDEZ-MEDINA *et al.*, 2017).

O uso de bases específicas de TEs durante a anotação é recomendado para aumentar a anotação dos transcritos montados (MORETON; IZQUIERDO; EMES, 2016). O uso de uma base ainda mais específica (transposons da própria espécie, ao invés de TEs do Repbase) é capaz de aumentar ainda mais a identificação dos TEs nos transcritos, como ocorreu com *Corydoras maculifer*, que passou de 1,17% para 4,68% do transcriptoma derivado de TEs, trocando uma base do Repbase por uma base de TEs específica do gênero (BELL *et al.*, 2022). Nesse estudo com os mosquitos nós reafirmamos os benefícios do uso de um banco de dados nucleotídico de TEs específico da espécie. Apesar disso, neste estudo, este banco espécie-específico foi utilizado em conjunto com o banco de dados de proteínas do Repbase, o que aumentou a identificação dos transcritos de TEs, provavelmente devido a possibilidade de captura de padrões que não são capturados por similaridade nucleotídica.

A partir da montagem dos transcritos pôde-se inferir a completude dos TEs expressos. Um número muito pequeno de transcritos potencialmente completos significa que apesar de muitas famílias de TEs serem amplamente expressas nos mosquitos, é muito pequena, embora existente, a chance de realização de movimentação dos TEs nessas espécies, uma vez que uma proteína íntegra é necessária para realização da correta transposição (GARCIA-PEREZ; WIDMANN; ADAMS, 2016; LEVIN; MORAN, 2011). Estudos de expressão de TEs em larga



escala são muito recentes e mais comuns para vertebrados, um resultado similar ao observado nos mosquitos foi também visto no anuro *Rhinella marina*, onde a maior parte dos transcritos derivados de TEs são derivados de cópias truncadas de TEs (LUDWIG *et al.*, 2021) e em humanos onde 99% dos transcritos do TE *L1* não são derivados de cópias completas com promotores íntegros (DEININGER *et al.*, 2017).

A inferência da completude a partir do transcriptoma montado com o Trinity, no entanto, possui uma limitação, as montagens dos transcritos foram geradas a partir de corridas de uma única linhagem da espécie, sob uma condição e cujo RNA foi obtido a partir do corpo total do mosquito, visando abranger todos os tipos de tecido. Isso foi realizado visando a redução da chance de formação de artefatos como transcritos quiméricos entre sequências de diferentes linhagens, que podem ser formados durante a montagem de regiões polimórficas (VINSON *et al.*, 2005). Além disso, embora o Trinity seja um software amplamente utilizado para montagens de transcriptoma, seu uso consome bastante recurso computacional e de tempo (MAO *et al.*, 2020), além de gerar um número de transcritos da ordem de  $10^5$  e  $10^6$ , de forma que se torna inviável a montagem *de novo* de milhares de corridas de RNA-Seq. Se por um lado houve certa perda de potenciais isoformas de TEs presentes em diferentes linhagens de mosquitos, por outro há uma maior garantia de que a chance de formação de artefatos seja menor.

A perda de variabilidade de transcritos em troca da redução do número de artefatos ficou um pouco evidente na etapa de análise proteômica. O uso de um banco de dados de TEs derivado apenas do transcriptoma montado foi capaz de gerar um número menor de proteínas de TEs identificadas, que quando comparado ao uso deste mesmo banco somado a um banco de dados com as ORFs de TEs derivadas das cópias de TEs encontradas no genoma. Isso provavelmente foi resultado de transcriptomas e proteomas derivados de condições experimentais distintas. Logo, se o transcriptoma montado e os dados proteômicos não derivarem da mesma amostra, o resultado da análise pode ser diferente do resultado mostrado por Maringer *et al.* (2017), onde os TEs são mais bem identificados com o uso da proteômica informada pela transcriptômica (PIT). Neste caso se recomenda o uso da PIT juntamente com as ORFs de TEs preditas diretamente do genoma da espécie.

O número de proteínas de TEs identificadas por espectrometria de massa foi pequeno quando comparado ao número de potenciais proteínas preditas a partir do transcriptoma montado ou mesmo preditas a partir das ORFs das cópias dos TEs no genoma: para *An. stephensi*, ORFs no genoma: 4769, ORFs no transcriptoma: 363, proteínas validadas: 21; para

*An. gambiae*, ORFs no genoma: 5558, ORFs no transcriptoma: 773, proteínas validadas: 4; para *Ae. aegypti*, ORFs no genoma: 37363, ORFs no transcriptoma: 3044, proteínas validadas: 69. Um baixo número de proteínas de TEs é esperado já que a produção descontrolada de proteínas de elementos de transposição pode causar uma série de danos ao hospedeiro (LI *et al.*, 2013a). Entretanto a grande diferença entre o número de transcritos de TEs com ORFs e expressos e o número de proteínas detectadas pode indicar uma regulação de TEs em nível pós transcricional ao invés de um silenciamento transcricional em nível de DNA, como por exemplo por metilação de regiões do DNA (COWLEY; OAKLEY, 2013; OKAMOTO; HIROCHIKA, 2001).

Os valores de expressão proteica de TEs em diferentes tecidos nunca foram mostrados. Estudos proteômicos abordando elementos de transposição são extremamente raros. Apenas o estudo de Maringer *et al.* (2017) usou tecnologia proteômica de *shotgun* para identificar TEs expressos em animais. Além disso, a identificação de TEs traducionalmente ativos também é difícil. Um outro estudo de proteômica de fungo *Coccidioides posadasii* conseguiu identificar apenas um único indício de TE expresso, neste caso um retrotransposon da superfamília *Gypsy* (MITCHELL *et al.*, 2018). No presente estudo, os espectros de massa gerados por Maringer *et al.* (2017) foram reanalisados como forma de comparação e validação do nosso método, assim como foi analisada uma série de outros experimentos gerados ao redor do mundo na última década. Foram identificadas menos proteínas derivadas de TEs nas células Aag2 de *Ae. aegypti* que o observado por Maringer *et al.* (2017), isto, no entanto, não se deve a uma limitação do nosso método, mas ao fato daqueles autores terem utilizado uma base de dados com um *cutoff* de tamanho de proteínas bem menor que o utilizado neste trabalho. Se por um lado o uso de proteínas com um menor tamanho pode gerar mais identificações, por outro o uso dessa abordagem pode gerar mais *matches* errôneos, devido a presença de ORFs espúrias na base de dados de proteínas. Considerando apenas proteínas de tamanho igual ou maior que 200 aminoácidos, os resultados desta tese validam o trabalho de Maringer *et al.* (2017) e são validados por ele, pois resultados semelhantes foram encontrados. Embora esses autores tivessem mostrado que algumas proteínas de elementos de transposição fossem de fato expressas, o estudo havia sido realizado em cultura de células, o que pode de fato gerar um resultado distinto do real comportamento dos TEs num organismo vivo.

Nossos resultados então trouxeram valiosas contribuições, pois deixam claro que o perfil de expressão dos TEs é altamente tecido-específico, com poucas proteínas sendo expressas em mais de um tecido. Embora as análises tenham sido realizadas apenas em poucos mosquitos, elas foram realizadas em espécies de gêneros distintos, podemos então deduzir que o comportamento

da expressão proteica do TEs é de fato tecido-específico nos Culicídeos. Ficou claro também o fato de proteínas de TEs estarem expressas em tecidos dos mosquitos vivos, embora em menor quantidade que em cultura de células. Em *An. stephensi*, a espécie com a maior quantidade de dados proteômicos, houve a formação de dois grandes grupos em relação ao padrão de expressão dos TEs, tecidos somáticos e tecidos germinativos. Isso confirma que o padrão de expressão dos TEs em tecidos somáticos não é diferente do tecido germinativo apenas em nível de RNA (HAIG, 2016), onde muitos TEs atuam como lncRNAs, mas também em nível proteico. A diferença nas proteínas de TEs é, provavelmente, um resultado da ativação de diferentes vias metabólicas e regulatórias entre os tipos de tecidos (PASQUESI *et al.*, 2020; RUSSO; HARRINGTON; STEINIGER, 2016), embora o controle exclusivo por piRNAs (silenciadores da expressão de TEs) nas estruturas germinativas seja matéria de debate e tenha sido recentemente mostrada a atuação de piRNAs também em células somáticas (LEWIS *et al.*, 2018).

Os resultados da análise proteômica indicaram raros casos de domesticação de proteínas de TEs. A maior parte das proteínas identificadas ainda possuíam domínios conservados dos elementos de transposição, contrastando com a degradação dos domínios proteicos característicos de TEs que tradicionalmente ocorre durante a domesticação de um TE para a formação de um novo gene do hospedeiro, além disso, as proteínas de TEs expressas numa espécie não possuem similaridade com proteínas de TEs expressas nas outras espécies de mosquito, diferenciando-se de outro sinal de domesticação, que é a presença de ortólogos deste TE em diversas espécies (SINZELLE; IZSVÁK; IVICS, 2009; VOLFF, 2006). Apenas duas proteínas detectadas em *An. stephensi* fogem da regra da expressão tecido específica e são expressas em quase todos os tecidos. Essas proteínas, XP\_035912068.1 e XP\_035912069.1, compõe o proteoma predito da espécie depositado no NCBI, onde estão classificadas como “*uncharacterized protein*”, e ainda possuem um domínio de GAG parcialmente conservado, e seus ortólogos estão presentes em diversas espécies dos gêneros *Anopheles*, *Culex* e *Aedes* e alguns outros artrópodes da classe Insecta.

Uma vez que foi possível detectar as proteínas de TEs expressas com sucesso, o uso do pipeline SearchGui em conjunto com o software PeptideShaker se mostrou uma alternativa viável ao uso do pipeline GIO (FAN *et al.*, 2015) proposto por Maringer *et al.* (2017), para o estudo de elementos de transposição. O SearchGui possui mais ferramentas de busca de proteínas (9) quando comparado ao GIO (2), isso pode facilitar a descoberta de TEs, uma vez que, durante as etapas de análise dos dados proteômicos, algumas combinações de determinados ferramentas de busca resultaram em melhor identificação de TEs quando comparado ao uso

exclusivo do buscador MSGF+, que é implementado no GIO. Surpreendentemente, o uso do Novor, um algoritmo para descoberta *de novo* de proteínas, não melhorou a descoberta de elementos de transposição. Em teoria o Novor geraria uma boa identificação de peptídeos de forma alternativa ao uso de engenhos de busca usando como banco de dados uma base muito grande de proteínas derivadas do proteoma predito, da proteômica informada pelo transcriptoma e das predições das cópias de TEs no genoma. No entanto, embora houvesse expectativa que o uso único das identificações realizadas a partir de apenas este algoritmo fosse insatisfatório (MUTH; RENARD, 2018), a mesclagem de seus resultados com os resultados de outras ferramentas gerou, em todas as ocasiões, uma identificação de proteínas menor que as análises em que os resultados do Novor eram excluídos.

### 7.3 ELEMENTOS DE TRANSPOSIÇÃO E RNAS LONGOS NÃO CODIFICANTES

Diferentes estudos mostraram que os lncRNAs afetam o fenótipo do espécime de várias maneiras, influenciando principalmente a expressão gênica, mas também regulando a arquitetura cromossômica, o recrutamento de modificadores de cromatina e modulando as interações cromossômicas (BONASIO; SHIEKHATTAR, 2014; YAO; WANG; CHEN, 2019). Além disso, experimentos de *knockdown* ou de *knockout* de lncRNAs resultam em mudanças morfológicas e comportamentais em diferentes organismos (MATTICK, 2018). Embora nosso conhecimento sobre o lncRNA tenha avançado, ainda existem várias perguntas sem resposta. Um exemplo é: de quais estruturas genômicas/*loci* surge um lncRNA? Surpreendentemente, os resultados de estudos em vertebrados mostraram um papel importante de cópias de elementos transponíveis (TEs) na emergência de lncRNAs (HEZRONI *et al.*, 2015; KAPUSTA *et al.*, 2013; KELLEY; RINN, 2012). Mais recentemente, estudos em plantas também mostram que alguns lncRNAs intrônicos se sobrepõem a TEs (PEDRO *et al.*, 2018). Embora existam poucos estudos em invertebrados, a contribuição de TEs para a biogênese do lncRNA parece variar amplamente entre os insetos, desde quase nenhuma sequência de TE sobrepondo-se a lncRNAs em himenópteros, até uma sobreposição considerável em *Drosophila melanogaster* (LOPEZ-EZQUERRA; HARRISON; BORNBERG-BAUER, 2017).

Neste estudo, foi observado que a maior parte (cerca de 90%) dos transcritos derivados de TE e montados nas 12 espécies de mosquitos é composta de RNAs com ORFs preditas de tamanho muito reduzido e dessa forma considerados lncRNAs. Conduziu-se então uma análise da relação do lncRNAs com os dados de TEs deste estudo. Infelizmente uma caracterização

sistemática dos lncRNAs existe apenas para pouquíssimas espécies de mosquito (AZLAN *et al.*, 2021a, 2021b). A comparação foi então realizada nas espécies *Ae. albopictus* e *Cx. quinquefasciatus*. Os resultados mostraram semelhança na sobreposição de TE/lncRNAs entre regiões intrônicas e intergênicas, diferindo substancialmente do padrão encontrado em lncRNAs de vertebrados, onde lncRNAs de regiões intergênicas e pobres em genes são mais ricos em TEs do que lncRNAs derivados da região intrônica e das regiões ricas em genes no genoma (KAPUSTA *et al.*, 2013).

Com relação à presença de TEs compondo os éxons dos lncRNAs, a alta fração de TEs encontrada nos éxons, observada nesse trabalho, contrasta com alguns outros insetos (LOPEZ-EZQUERRA; HARRISON; BORNBERG-BAUER, 2017). Assim, a depleção de TEs em éxons de lncRNAs, como sugerido por autores anteriores, não pode ser extrapolada para insetos em geral. Pelo menos, em espécies de insetos com grande genoma e alta densidade de TEs, a contribuição dos TEs para a evolução e propriedades dos lncRNAs pode ser tão importante quanto para os vertebrados. Em éxons de lncRNAs, os TEs podem ser diretamente responsáveis pelos domínios de atividade das espécies de RNA não codificante (JOHNSON; GUIGÓ, 2014). Além disso, eles também podem controlar a expressão de lncRNAs através da via PIWI-piRNA, uma vez que as regiões TE dentro dos lncRNAs podem ser alvo de piRNAs derivados de outras cópias TE dispersas no genoma (WANG; LIN, 2021).

A super-representação de alguns retrotransposons LTR, que aqui foi observada nos mosquitos, também foi observada em humanos e camundongos, mas não em insetos, onde estudos em algumas espécies mostraram que TEs em lncRNAs refletem a distribuição genômica dos TEs (LOPEZ-EZQUERRA; HARRISON; BORNBERG-BAUER, 2017). A predominância de um tipo específico de TE em lncRNAs também varia entre as espécies, LTR é a principal ordem de TE nos lncRNAs de *Ae. albopictus*, como em humanos, camundongos e insetos como *D. melanogaster* (KAPUSTA *et al.*, 2013; LOPEZ-EZQUERRA; HARRISON; BORNBERG-BAUER, 2017). Por outro lado, os transposons de Classe II são prevalentes nos lncRNAs de *Cx. quinquefasciatus*, assim como em outros insetos como *Tribolium castaneum* e *Sogatella furcifera* (CHANG *et al.*, 2020; LOPEZ-EZQUERRA; HARRISON; BORNBERG-BAUER, 2017). Isso sugere que o tipo de TE predominante em lncRNAs é específico de cada espécie ou gênero e não pode ser extrapolado para níveis familiares ou *taxa* superiores.

Os retroelementos LTR são conhecidos por possuírem repetições terminais longas (LTRs) nas bordas de sua região codificante. Além de cópias de retroelementos LTR completos e fragmentados, muitos solo-LTRs também são encontrados dispersos no genoma. O fato das

superfamílias *Gypsy* e *Bel-Pao* estarem enriquecidas nos éxons dos lncRNAs provavelmente está relacionado com as propriedades dessas LTRs. As regiões de LTRs contêm promotores de RNA polimerase II, bem como outros motivos de sequência regulatória, e vários estudos têm mostrado que LTRs são mais propensos a manter esses locais do que outros tipos de TE, aumentando a probabilidade de modular a expressão de genes flanqueadores (CHUONG; ELDE; FESCHOTTE, 2017). Além disso, uma pressão menos seletiva nos lncRNAs pode facilitar a domesticação de LTRs como promotores de lncRNA (THOMPSON; MACFARLAN; LORINCZ, 2016).

Com relação a idade dos TEs que compõem os lncRNAs, os nossos resultados indicaram uma prevalência de TEs menos divergentes em relação a cópia consenso de suas próprias famílias. Tais padrões de divergência levantam duas hipóteses não mutuamente exclusivas: I – cópias de TE coocorrendo com lncRNAs são de fato inserções mais jovens; II - essa baixa divergência é resultado de cópias antigas conservadas que estão sendo mantidas sob seleção purificadora. Ulitsky e Bartel (2013) chegaram à conclusão de que os éxons lncRNA são geralmente mais conservados do que as regiões intergênicas (evolução neutra) em vertebrados, dando suporte para a última hipótese. No entanto, os resultados desta tese são insuficientes para diferenciar e testar essas hipóteses devido à falta de *loci* ortólogos de lncRNAs entre *Ae. albopictus* e *Cx. quinquefasciatus*.

#### 7.4 LIMITAÇÕES E PERSPETIVAS FUTURAS

É importante destacar que a qualidade da montagem do genoma pode ter um grande impacto na caracterização de TEs, sua anotação e em toda análise posterior. Os genomas de mosquitos são altamente variáveis em termos de qualidade de montagem e muitos dos genomas de *Anopheles* inicialmente analisados eram fragmentados. No entanto, foram utilizados filtros muito rigorosos, além de validação com novas versões de conjuntos genômicos produzidos usando *reads* longas nas espécies com dados complementares disponíveis (*Ae. aegypti*, *Ae. albopictus*, *An. funestus*, *An. stephensi* e *An. coluzzii*), a fim de evitar resultados inconsistentes e confirmando assim a grande maioria dos achados baseados em versões mais fragmentadas do genoma. Na fase das análises de expressão, toda a caracterização de TEs foi adaptada a essas novas versões de montagem, mais completas e íntegras. Ainda assim, algumas espécies como *An. arabiensis* e *An. merus* não possuem novos dados de montagem genômica. Logo, a fragmentação desses genomas pode ter algum impacto nos resultados finais de expressão. No entanto, esse

impacto parece ser mínimo, já que a expressão dos TEs, nessas espécies, mostrou perfis similares aos encontrados em espécies com ótimas montagens, algumas em nível cromossomal.

Cada vez mais vem ficando evidente a necessidade de se inserir a análise dos elementos de transposição em estudos de larga escala de expressão gênica. Um dos resultados de maior potencial de reuso desse trabalho são os arquivos de anotação dos TEs. Até o momento, pouquíssimos estudos de expressão olharam para os elementos de transposição, embora existam mais de 2 mil corridas de RNA-Seq de mosquitos em bancos de dados públicos. Isso agora pode ser facilmente analisado utilizando-se o *pipeline* Tetranscripts (JIN *et al.*, 2015) em conjunto com os arquivos de anotação de TEs dos mosquitos.

Os TEs podem influenciar ou ser coexpressos com genes não apenas em *cis*, mas também em *trans* (FOUCHÉ *et al.*, 2020). Uma das limitações da nossa abordagem de avaliação de coexpressão é que foram considerados apenas TEs próximos a genes (com potenciais efeitos em *cis*), dessa forma coexpressões entre genes e TEs distantes não puderam ser verificadas. Além disso, é possível que a expressão de um gene sofra a influência de dois TEs, um agindo em *cis* e outro em *trans*, a análise apenas em *cis* poderia reduzir a correlação da expressão do gene com o TE. Uma alternativa a isso seria a construção de redes de coexpressão a partir de dados de RNA-Seq.

A construção de redes de coexpressão, no entanto, não é uma tarefa simples de ser realizada. Muitas variações de expressão (diferentes tecidos, estágios de desenvolvimento e estresses bióticos/abióticos) devem ser capturadas para a obtenção de uma rede de qualidade. Para isso um número grande de corridas deve ser utilizado, uma vez que o número de amostras é diretamente correlacionado com a qualidade da rede. Além disso, o número de amostras não deve ser menor que 20 e com profundidade de sequenciamento acima de 100 milhões de *reads*, muitas vezes necessitando do uso de meta-análise de múltiplos experimentos de RNA-Seq, cada um com mais de 10 corridas (BALLOUZ; VERLEYEN; GILLIS, 2015), impossibilitando a construção dessas redes para a grande maioria das espécies de mosquitos.

Há diversos métodos de inferência de correlação para as redes. Dentre os tipos de relação, os mais simples são as análises de correlação matemática. Neste caso o peso da aresta (relação) entre dois nós (genes) é dado pelo coeficiente de correlação, podendo ser aplicado o método de Pearson, de Spearman ou de Kendall. Genes com expressão correlacionadas, provavelmente interagem. Entretanto, uma vez que a análise de correlação pode não capturar todas as possíveis relações entre os genes, um segundo tipo de relação pode ser calculado, através dos métodos de

*mutual information*. Esses métodos verificam o grau de dependência entre duas variáveis aleatórias, neste caso, os genes (HUYNH-THU; SANGUINETTI, 2019). Eles são implementados em diferentes pacotes, sendo o ARACNE (MARGOLIN *et al.*, 2006) e o MRNET (MEYER *et al.*, 2007) bastante utilizados. No entanto, nem sempre os métodos mais complexos dão resultados suficientes, às vezes os métodos mais simples geram redes mais próximas à realidade (BALLOUZ; VERLEYEN; GILLIS, 2015).

O fato da rede ser construída a partir da junção do padrão de expressão nos genes em diferentes corridas de RNA-Seq poderia significar que quanto mais amostras fossem adicionadas à análise, melhor a detecção de relações. Entretanto, há um momento que a melhora atinge um platô, nesse momento, além de não melhorar a qualidade dos sinais de coexpressão, a adição de novas amostras pode aumentar o ruído e como consequência causar perda de alguns sinais de relação (LIESECKE *et al.*, 2019). Devido a todos esses fatores, como a profundidade de sequenciamento, da quantidade das corridas inseridas nas redes e também a necessidade do uso de diferentes métodos de construção, não foi possível construir redes de RNA-Seq para verificar a coexpressão de todos os TEs sobre os genes neste trabalho, isso irá requerer um esforço futuro e especificamente em uma espécie por vez.

Nesse trabalho também foram gerados milhares de arquivos de mapeamento dos dados de RNA-Seq em genomas das 12 espécies de mosquitos. Estes dados extremamente valiosos serão futuramente utilizados para validar a presença de transcritos quiméricos nas regiões de sobreposição de genes e TEs identificadas no presente estudo, e comparar a ocorrência desses transcritos em populações de mosquitos com diferentes fenótipos como susceptibilidade ou resistência a determinados compostos como inseticidas. Além disso, o grande volume de dados de expressão já calculados poderá servir de base para construção de redes de coexpressão da espécie *Ae. aegypti*.



## 8 CONCLUSÕES

Com a enorme quantidade de dados extraídos e informações geradas acerca dos elementos de transposição nos mosquitos, algumas conclusões sobre o comportamento dos TEs nos culicídeos podem ser geradas:

- a) Há grande variabilidade na fração do genoma composta por TEs nos mosquitos;
- b) Mesmo em mosquitos com poucos TEs, o mobiloma é diverso;
- c) O mobiloma sofreu muitas transferências horizontais entre as espécies até resultar no padrão atual de TEs nos genomas;
- d) Possivelmente, as transferências horizontais de transposons contribuíram para a diversificação das espécies de mosquitos;
- e) Transferências horizontais de LTRs são mais comuns do que transferências de TIRs entre espécies mais próximas. Entre espécies mais distantes, conceito aceito na literatura sobre a prevalência da transferência de TIRs continua válido;
- f) TEs não são estáticos a nível transcricional nos mosquitos, a maior parte das famílias é expressa;
- g) Os TEs mostram expressão diferencial em algumas situações, mas a ativação da expressão dos elementos durante o estresse não segue uma regra padrão como indicada na literatura;
- h) A maioria dos transcritos derivados de TEs forma lncRNAs, cobrindo uma fração significativa dos lncRNAs de algumas espécies, sendo este o maior meio de influência da expressão do TEs sobre o fenótipo da espécie;
- i) Especificamente em *Ae. albopictus*, alguns dos lncRNAs relacionados a TEs podem contribuir ativamente para a regulação da defesa do mosquito contra patógenos virais;
- j) A composição de TEs em lncRNAs não reflete a distribuição de TEs no genoma, há uma predileção de retrotransposons LTR;
- k) Uma pequena parte dos TEs dos mosquitos possui potenciais efeitos *cis*-regulatórios sobre alguns genes de 10 espécies;

- l) A coexpressão de TEs com genes que codificam proteínas chave dentro das redes PPI dos mosquitos é rara, podendo ser resultado de uma seleção negativa agindo a nível de regulação da expressão de TEs sobre genes de maior importância no interactoma dos mosquitos;
- m) Poucos elementos de transposição ainda são tradicionalmente ativos e sua expressão é altamente dependente do tecido.

O comportamento da expressão dos elementos de transposição para os eucariotos não pode ser extrapolado utilizando-se apenas espécies modelo, há muita variação específica para cada grupo de espécies. Nos mosquitos, a principal forma de alteração do fenótipo a partir da expressão dos TEs parece ser através dos lncRNAs derivados de TEs, ao invés de interação a partir de proteínas de TEs ou casos de proteínas de TEs domesticadas. Entretanto, do ponto de vista da domesticação, a fração de cópias de TEs que deram origem a lncRNAs funcionais faz com que parte desses elementos não sejam mais considerados parasitas genômicos ou elementos egoístas, mas como “genes” exógenos que foram cooptados para funções regulatórias.

## REFERÊNCIAS

AEBERSOLD, R.; MANN, M. Mass spectrometry-based proteomics. **Nature**, London, v. 422, n. 6928, p. 198–207, mar. 2003.

AGRAWAL, A.; EASTMAN, Q. M.; SCHATZ, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. **Nature**, London, v. 394, n. 6695, p. 744–51, 20 ago. 1998.

ANDREWS, S. **FastQC: a quality control tool for high throughput sequence data.**

Disponível em: <<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>>. Acesso em: 4 ago. 2018.

ARENSBURGER, P.; HICE, R. H.; ZHOU, L.; et al. Phylogenetic and functional characterization of the hAT transposon superfamily. **Genetics**, Austin, v. 188, n. 1, p. 45–57, 2011.

ARENSBURGER, P. et al. Sequencing of *Culex quinquefasciatus* Establishes a Platform for Mosquito Comparative Genomics. **Science**, New York, v. 330, n. 6000, p. 86–88, out. 2010. Disponível em: <<http://www.sciencemag.org/cgi/doi/10.1126/science.1191864>>. Acesso em: 20 mar. 2020.

ARENSBURGER, P.; HICE, R. H.; WRIGHT, J. a; et al. The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. **BMC Genomics**, London, v. 12, n. 1, p. 606, 15 dez. 2011. Disponível em: <<http://www.biomedcentral.com/1471-2164/12/606>>. Acesso em: 21 mar. 2020.

ARKHIPOVA, I. R. Distribution and phylogeny of penelope-like elements in eukaryotes. **Systematic Biology**, Washington, D.C, v. 55, n. 6, p. 875–885, 2006.

ARKHIPOVA, I. R.; YUSHENOVA, I. A.; ANGERT, E. Giant Transposons in Eukaryotes: Is Bigger Better? **Genome Biology and Evolution**, Oxford, v. 11, n. 3, p. 906–918, 2019.

ASHTIANI, M. et al. A systematic survey of centrality measures for protein-protein interaction networks. **BMC Systems Biology**, London, v. 12, n. 1, p. 80, 31 dez. 2018. Disponível em: <<https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-018-0598-2>>. Acesso em: 22 mar. 2020.

ATKINSON, P. W. hAT Transposable Elements. **Microbiology Spectrum**, Washington, D.C, v. 3, n. 4, p. 1–16, 2 jul. 2015. Disponível em: <<http://www.asmscience.org/content/journal/microbiolspec/10.1128/microbiolspec.MDNA3-0054-2014>>. Acesso em: 23 mar. 2020.

AZLAN, A. et al. Genome-wide identification of *Aedes albopictus* long noncoding RNAs and their association with dengue and Zika virus infection. **PLOS Neglected Tropical Diseases**, San Francisco, v. 15, n. 1, p. e0008351, 22 jan. 2021. Disponível em: <<http://dx.doi.org/10.1371/journal.pntd.0008351>>. Acesso em: 20 mar. 2021.

AZLAN, A. et al. Identification and characterization of long noncoding RNAs and their association with acquisition of blood meal in *Culex quinquefasciatus*. **Insect Science**, Victoria, v. 28, n. 4, p. 917-928, 22 jul. 2020.

BAIDOURI, M. El et al. Widespread and frequent horizontal transfers of transposable elements in plants. **Genome Research**, New York, v. 24, n. 5, p. 831–838, 2014.

BALLOUZ, S.; VERLEYEN, W.; GILLIS, J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. **Bioinformatics**, Oxford, v. 31, n. 13, p. 2123–2130, 1 jul. 2015.

BANHO, C. A. et al. Transposable Element Expression and Regulation Profile in Gonads of Interspecific Hybrids of *Drosophila arizonae* and *Drosophila mojavensis wrightleyi*. **Cells**, Basel, v. 10, n. 12, p. 3574, 18 dez. 2021. Disponível em: <<https://www.mdpi.com/2073-4409/10/12/3574>>. Acesso em: 30 mar. 2022.

BANTYSH, O. B.; BUZDIN, A. A. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. **Biochemistry**, Moscow, v. 74, n. 12, p. 1393–1399, 7 dez. 2009.

BAO, W. et al. New superfamilies of eukaryotic DNA transposons and their internal divisions. **Molecular Biology and Evolution**, New York, v. 26, n. 5, p. 983–993, 2009.

BAO, W.; KOJIMA, K. K.; KOHANY, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. **Mobile DNA**, London, v. 6, n. 1, p. 11, 2015. Disponível em: <<http://www.mobilednajournal.com/content/6/1/11>>. Acesso em: 1 mar. 2021.

BAO, Z.; EDDY, S. R. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. **Genome Research**, New York, v. 12, n. 8, p. 1269–1276, 1 ago. 2002.

BARAH, P. et al. Genome scale transcriptional response diversity among ten ecotypes of *Arabidopsis thaliana* during heat stress. **Frontiers in Plant Science**, Lausanne, v. 4, n. December, p. 532, 2013. Disponível em: <<http://journal.frontiersin.org/article/10.3389/fpls.2013.00532/abstract>>. Acesso em: 1 mar. 2021.

BARRÓN, M. G. et al. A new species in the major malaria vector complex sheds light on reticulated species evolution. **Scientific Reports**, London, v. 9, n. 1, p. 14753, 14 dez. 2019. Disponível em: <<http://www.nature.com/articles/s41598-019-49065-5>>. Acesso em: 1 mar. 2021.

BARSNES, H.; VAUDEL, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. **Journal of Proteome Research**, Washington, D.C., v. 17, n. 7, p. 2552–2555, 2018.

BARUCH, O.; KASHKUSH, K. Analysis of copy-number variation, insertional polymorphism, and methylation status of the tiniest class I (TRIM) and class II (MITE) transposable element families in various rice strains. **Plant Cell Reports, Cambridge**, Berlin, v. 31, n. 5, p. 885–893, 2012.

BATEMAN, A. et al. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, Oxford, v. 49, n. D1, p. D480–D489, 8 jan. 2021.

BELANCIO, V. P.; ROY-ENGEL, A. M.; DEININGER, P. L. All y'all need to know 'bout retroelements in cancer. **Seminars in Cancer Biology**, London, v. 20, n. 4, p. 200–210, ago. 2010.

BELL, E. A. et al. Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. **Molecular Ecology Resources**, Oxford, v. 22, n. 2, p. 823–833, 2 fev. 2022.

BENELLI, G.; MEHLHORN, H. Declining malaria, rising of dengue and Zika virus: insights for mosquito vector control. **Parasitology Research**, Berlin, v. 115, n. 5, p. 1747–1754, 2016.

BENNETZEN, J. L.; WANG, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. **Annual Review of Plant Biology**, Palo Alto, v. 65, n. 1, p. 505–530, 29 abr. 2014.

BERTHELIER, J. et al. A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*. **BMC Genomics**, London, v. 19, n. 1, p. 378, 22 dez. 2018. Disponível em: <<https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4763-1>>. Acesso em: 2 mar. 2021.

BIEDLER, J. K.; CHEN, X.; TU, Z. Horizontal transmission of an R4 clade non-long terminal repeat retrotransposon between the divergent *Aedes* and *Anopheles* mosquito genera. **Insect Molecular Biology**, London, v. 24, n. 3, p. 331–337, 2015.

BIEDLER, James K.; SHAO, H.; TU, Z. Evolution and horizontal transfer of a DD37E DNA transposon in mosquitoes. **Genetics**, Austin, v. 177, n. 4, p. 2553–2558, 2007.

BIÉMONT, C.; VIEIRA, C. Genetics: Junk DNA as an evolutionary force. **Nature**, London, v. 443, n. 7111, p. 521–524, 2006.

BOGU, G. K. et al. Atlas of transcriptionally active transposable elements in human adult tissues. **bioRxiv**, 2019. Disponível em: <<http://biorxiv.org/content/early/2019/07/24/714212.abstract>>. Acesso em: 2 mar. 2021.

BÖHNE, A. et al. Zisupton - A novel superfamily of DNA transposable elements recently active in fish. **Molecular Biology and Evolution**, New York, v. 29, n. 2, p. 631–645, 2012.

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

BONASIO, R.; SHIEKHATTAR, R. Regulation of Transcription by Long Noncoding RNAs. **Annual Review of Genetics**, Palo Alto, v. 48, n. 1, p. 433–455, 23 nov. 2014.

BONIZZONI, M. et al. The invasive mosquito species *Aedes albopictus*: Current knowledge and future perspectives. **Trends in Parasitology**, Oxford, v. 29, n. 9, p. 460–468, 2013.

BORKENT, A.; GRIMALDI, D. A. The Earliest Fossil Mosquito (Diptera: Culicidae), in Mid-Cretaceous Burmese Amber. **Annals of the Entomological Society of America**, College Park, v. 97, n. 5, p. 882–888, 2006.

BOUALLÈGUE, M. et al. Molecular evolution of *PiggyBac* superfamily: from selfishness to domestication. **Genome Biology and Evolution**, Oxford, v. 9, n. 2, p. evw292, 12 jan. 2017. Disponível em: <<https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evw292>>. Acesso em: 3 mar. 2021.

BOULESTEIX, M.; BIÉMONT, C. Transposable elements in mosquitoes. **Cytogenetic and Genome Research**, Basel, v. 110, n. 1–4, p. 500–509, 2005.

BRASIL. **Guia de Vigilância do Culex quinquefasciatus**. 3ª ed. Brasília: Ministério da Saúde, 2011.

BROECKER, F.; MOELLING, K. Evolution of Immune Systems From Viruses and Transposable Elements. **Frontiers in Microbiology**, Lausanne, v. 10, p. 51, 29 jan. 2019. Disponível em: <<https://www.frontiersin.org/article/10.3389/fmicb.2019.00051/full>>. Acesso em: 5 mar. 2021.

BUCHMANN, J. P. et al. Analysis of CACTA transposases reveals intron loss as major factor influencing their exon/intron structure in monocotyledonous and eudicotyledonous hosts. **Mobile DNA**, London, v. 5, n. 1, p. 24, 2014. Disponível em: <<http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-5-24>>. Acesso em: 5 mar. 2021.

BUREAU, T. E.; WESSLER, S. R. Tourist: A Large Family of Small Inverted Repeat Elements Frequently Associated with Maize Genes. **The Plant Cell**, Rockville, v. 4, n. 10, p. 1283, out. 1992. Disponível em: <<https://www.jstor.org/stable/10.2307/3869414?origin=crossref>>. Acesso em: 5 mar. 2021.

CARAGATA, E. P. et al. The transcriptome of the mosquito *Aedes fluviatilis* (Diptera: Culicidae), and transcriptional changes associated with its native *Wolbachia* infection. **BMC Genomics**, London, v. 18, n. 1, p. 6, 3 dez. 2017. Disponível em: <<http://dx.doi.org/10.1186/s12864-016-3441-4>>. Acesso em: 5 mar. 2021.

CASSONE, B. J. et al. Comparative transcriptomics of malaria mosquito testes: Function, evolution, and linkage. **G3: Genes, Genomes, Genetics**, Bethesda, v. 7, n. 4, p. 1127–1136, 2017.

CASTELLANA, N. E. et al. Discovery and revision of Arabidopsis genes by proteogenomics. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 105, n. 52, p. 21034–21038, 2008.

CHALOPIN, D. et al. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. **Genome Biology and Evolution**, Oxford, v. 7, n. 2, p. 567–580, 2015.

CHANG, Z. X. et al. Genome-wide characterization and developmental expression profiling of long non-coding RNAs in *Sogatella furcifera*. **Insect Science**, Victoria, v. 27, n. 5, p. 987–

997, 2020.

CHEN, S.; LI, X. Molecular characterization of the first intact Transib transposon from *Helicoverpa zea*. **Gene**, Amsterdam, v. 408, n. 1–2, p. 51–63, jan. 2008.

CHÉNAIS, B. et al. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. **Gene**, Amsterdam, v. 509, n. 1, p. 7–15, 2012.

CHOCU, S. et al. Forty-Four Novel Protein-Coding Loci Discovered Using a Proteomics Informed by Transcriptomics (PIT) Approach in Rat Male Germ Cells1. **Biology of Reproduction**, New York, v. 91, n. 5, p. 1–14, 1 nov. 2014.

CHOI, Y.-J. et al. Dual RNA-seq of Parasite and Host Reveals Gene Expression Dynamics during Filarial Worm–Mosquito Interactions. **PLOS Neglected Tropical Diseases**, San Francisco, v. 8, n. 5, p. e2905, 22 maio 2014. Disponível em: <<https://dx.plos.org/10.1371/journal.pntd.0002905>>. Acesso em: 5 mar. 2021.

CHOWDHURY, A. et al. JNK pathway restricts DENV2, ZIKV and CHIKV infection by activating complement and apoptosis in mosquito salivary glands. **PLOS Pathogens**, San Francisco, v. 16, n. 8, p. e1008754, 10 ago. 2020. Disponível em: <<https://dx.plos.org/10.1371/journal.ppat.1008754>>. Acesso em: 10 mar. 2021.

CHUNG, H. et al. Cis-Regulatory Elements in the Accord Retrotransposon Result in Tissue-Specific Expression of the *Drosophila melanogaster* Insecticide Resistance Gene *Cyp6g1*. **Genetics**, Austin, v. 175, n. 3, p. 1071–1077, mar. 2007.

CHUNG, S.; ZUKER, C.; LODISH, H. F. A repetitive and apparently transposable DNA sequence in *Dictyostelium discoideum* associated with developmentally regulated RNAs. **Nucleic Acids Research**, Oxford, v. 11, n. 14, p. 4835–4852, 1983.

CHUONG, E. B.; ELDE, N. C.; FESCHOTTE, C. Regulatory activities of transposable elements: from conflicts to benefits. **Nature Reviews Genetics, London**, v. 18, n. 2, p. 71–86, 21 fev. 2017.

CLAPES, T. et al. Chemotherapy-induced transposable elements activate MDA5 to enhance haematopoietic regeneration. **Nature Cell Biology, London**, v. 23, n. 7, p. 704–717, 2021.

CLAUDIANOS, C. et al. maT - A clade of transposons intermediate between mariner and Tc1. **Molecular Biology and Evolution**, New York, v. 19, n. 12, p. 2101–2109, 2002.

CLEMENTS, A. N. **The Biology of Mosquitoes: Transmission of Viruses and Interactions with Bacteria**. Cambridge: Cambridge University Press, 2012.

COATES, B. S. Horizontal transfer of a non-autonomous Helitron among insect and viral genomes. **BMC Genomics, London**, v. 16, n. 1, p. 137, 27 dez. 2015. Disponível em: <<https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1318-6>>. Acesso em: 10 mar. 2021.

CONLEY, A. B.; MILLER, W. J.; JORDAN, I. K. Human cis natural antisense transcripts

initiated by transposable elements. **Trends in Genetics**, Amsterdam, v. 24, n. 2, p. 53–56, 2008.

CORDAUX, R. et al. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 103, n. 21, p. 8101–8106, 23 maio 2006.

COWLEY, M.; OAKEY, R. J. Transposable Elements Re-Wire and Fine-Tune the Transcriptome. **PLoS Genetics**, San Francisco, v. 9, n. 1, p. e1003234, 24 jan. 2013. Disponível em: <<https://dx.plos.org/10.1371/journal.pgen.1003234>>. Acesso em: 10 mar. 2021.

COX, J. et al. Andromeda: A peptide search engine integrated into the MaxQuant environment. **Journal of Proteome Research**, Washington, D.C, v. 10, n. 4, p. 1794–1805, 2011.

COY, M. R.; TU, Z. Gambol and Tc1 are two distinct families of DD34E transposons: Analysis of the *Anopheles gambiae* genome expands the diversity of the IS630-Tc1-mariner superfamily. **Insect Molecular Biology**, London, v. 14, n. 5, p. 537–546, 2005.

CRAIG, R.; BEAVIS, R. C. TANDEM: matching proteins with tandem mass spectra. **Bioinformatics**, Oxford, v. 20, n. 9, p. 1466–1467, 12 jun. 2004.

DAVIDSON, A. D.; MATTHEWS, D. A.; MARINGER, K. Proteomics technique opens new frontiers in mobilome research. **Mobile Genetic Elements**, Austin, v. 7, n. 4, p. e1362494, 4 jul. 2017. Disponível em: <<https://doi.org/10.1080/2159256X.2017.1362494>>. Acesso em: 10 mar. 2021.

DE LA CHAUX, N.; WAGNER, A. BEL/Pao retrotransposons in metazoan genomes. **BMC Evolutionary Biology**, London, v. 11, n. 1, 2011.

DEGNER, E. C. et al. Proteins, Transcripts, and Genetic Architecture of Seminal Fluid and Sperm in the Mosquito *Aedes aegypti*. **Molecular & Cellular Proteomics**, Bethesda, v. 18, n. 1, p. S6–S22, mar. 2019.

DEININGER, P. et al. A comprehensive approach to expression of L1 loci. **Nucleic Acids Research**, Oxford, v. 45, n. 5, p. e31, 17 mar. 2017. Disponível em: <<https://academic.oup.com/nar/article/45/5/e31/2605729>>. Acesso em: 10 mar. 2021.

DEMARCO, R.; VENANCIO, T. M.; VERJOVSKI-ALMEIDA, S. SmTRC1, a novel *Schistosoma mansoni* DNA transposon, discloses new families of animal and fungi transposons belonging to the CACTA superfamily. **BMC Evolutionary Biology**, London, v. 6, n. 89, p. 89, 7 nov. 2006. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17090310>>. Acesso em: 14 mar. 2021.

DEUTSCH, E. W. et al. The ProteomeXchange consortium in 2020: Enabling “big data” approaches in proteomics. **Nucleic Acids Research**, Oxford, v. 48, n. D1, p. D1145–D1152, 2020.



DHAWAN, R. et al. Data from salivary gland proteome analysis of female *Aedes aegypti* Linn. **Data in Brief**, Amsterdam, v. 13, p. 274–277, ago. 2017.

DIAMENT, B. J.; NOBLE, W. S. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. **Journal of Proteome Research**, Washington, D.C, v. 10, n. 9, p. 3871–3879, 2 set. 2011.

DIAO, Y. et al. Next-generation sequencing reveals recent horizontal transfer of a DNA transposon between divergent mosquitoes. **PLoS ONE**, San Francisco, v. 6, n. 2, p. e16743, 2011. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0016743>>. Acesso em: 14 mar. 2021.

DIESEL, J. F. et al. A re-annotation of the *Anopheles darlingi* mobilome. **Genetics and Molecular Biology**, Ribeirão Preto, v. 42, n. 1, p. 125–131, 21 jan. 2019.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, Oxford, v. 29, n. 1, p. 15–21, jan. 2013.

DONG, S.; BEHURA, S. K.; FRANZ, A. W. E. The midgut transcriptome of *Aedes aegypti* fed with saline or protein meals containing chikungunya virus reveals genes potentially involved in viral midgut escape. **BMC Genomics**, London, v. 18, n. 1, p. 382, 15 dez. 2017. Disponível em: <<https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3775-6>>. Acesso em: 14 mar. 2021.

DORFER, V. et al. MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. **Journal of Proteome Research**, Washington, D.C, v. 13, n. 8, p. 3679–3684, 26 ago. 2014.

DOTTO, B. R. et al. HTT-DB: new features and updates. **Database**, Oxford, v. 2018, n. 2018, p. 1–5, 1 jan. 2018. Disponível em: <<https://academic.oup.com/database/article/doi/10.1093/database/bax102/4791205>>. Acesso em: 14 mar. 2021.

DRONGITIS, D. et al. Roles of Transposable Elements in the Different Layers of Gene Expression Regulation. **International Journal of Molecular Sciences**, Basel, v. 20, n. 22, p. 5755, 15 nov. 2019. Disponível em: <<https://www.mdpi.com/1422-0067/20/22/5755>>. Acesso em: 14 mar. 2021.

DU, J. et al. SoyTEdb: A comprehensive database of transposable elements in the soybean genome. **BMC Genomics**, London, v. 11, n. 1, 2010.

DUPEYRON, M. et al. Evolution of Mutator transposable elements across eukaryotic diversity. **Mobile DNA**, London, v. 10, n. 1, p. 12, 21 dez. 2019. Disponível em: <<https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-019-0153-8>>. Acesso em: 14 jun. 2021.

DWIVEDI, S. B. et al. Brain Proteomics of *Anopheles gambiae*. **OMICS: A Journal of Integrative Biology**, Larchmont, v. 18, n. 7, p. 421–437, jul. 2014.

EDDY, S. R. Accelerated Profile HMM Searches. **PLoS Computational Biology**, San

- Francisco, v. 7, n. 10, p. e1002195, 20 out. 2011. Disponível em: <<https://dx.plos.org/10.1371/journal.pcbi.1002195>>. Acesso em: 14 jun. 2021.
- EDGAR, R. C.; MYERS, E. W. PILER: identification and classification of genomic repeats. **Bioinformatics**, Oxford, v. 21, n. 1, p. i152–i158, 1 jun. 2005.
- EICKBUSH, T. H.; MALIK, H. S. Origins and Evolution of Retrotransposons. **Mob. DNA II**. Washington, DC: Wiley, 2002. p. 1111–1144.
- ELIAS, J. E.; GYGI, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. **Nature Methods**, New York, v. 4, n. 3, p. 207–214, 1 mar. 2007.
- ELIAS, J. E.; GYGI, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. **Methods Mol. Biol.** New York: Humana Press, 2010. v. 604. p. 55–71.
- ELLINGHAUS, D.; KURTZ, S.; WILLHOEFT, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. **BMC Bioinformatics**, Oxford, v. 9, n. 18, 14 dez. 2008. Disponível em: <<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-18>>. Acesso em: 14 jun. 2021.
- ELSNER, D.; MEUSEMANN, K.; KORB, J. Longevity and transposon defense, the case of termite reproductives. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 115, n. 21, p. 5504–5509, 22 maio 2018.
- EMMONS, S. W. et al. Evidence for a transposon in caenorhabditis elegans. **Cell**, Cambridge, v. 32, n. 1, p. 55–65, 1983.
- ENG, J. K.; JAHAN, T. A.; HOOPMANN, M. R. Comet: An open-source MS/MS sequence database search tool. **Proteomics**, Weinheim, v. 13, n. 1, p. 22–24, jan. 2013.
- ETEBARI, K. et al. Global Transcriptome Analysis of Aedes aegypti Mosquitoes in Response to Zika Virus Infection. **mSphere**, Washington, DC, v. 2, n. 6, p. 1648–1659, 27 dez. 2017.
- EVGEN'EV, M. B. et al. Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 94, n. 1, p. 196–201, 7 jan. 1997.
- EVGEN'EV, M. B.; ARKHIPOVA, I. R. Penelope-like elements – a new class of retroelements: distribution, function and possible evolutionary significance. **Cytogenetic and Genome Research**, Basel, v. 110, n. 1–4, p. 510–521, 2005.
- EWELS, P. et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, Oxford, v. 32, n. 19, p. 3047–3048, 1 out. 2016.
- FAN, J. et al. Galaxy integrated omics: Web-based standards-compliant workflows for proteomics informed by transcriptomics. **Molecular and Cellular Proteomics**, Bethesda, v. 14, n. 11, p. 3087–3093, 2015.

FANG, Y. J.; YAN, Z. T.; CHEN, B. Sialotranscriptome sequencing and analysis of *Anopheles sinensis* and comparison with *Psorophora albipes* sialotranscriptome (Diptera: Culicidae). **Insect Science**, Victoria, v. 25, n. 3, p. 368–378, 2018.

FATTASH, I. et al. Miniature inverted-repeat transposable elements: discovery, distribution, and activity. **Genome**, Ottawa, v. 56, n. 9, p. 475–486, set. 2013.

FEDOROFF, N. V. Molecular Genetics and Epigenetics of CACTA Elements. **Plant Transposable Elem.** Totowa: Humana Press, 2013. p. 177–192.

FERNÁNDEZ-MEDINA, R. D. et al. Transposable elements in the *Anopheles funestus* transcriptome. **Genetica**, Dordrecht, v. 145, n. 3, p. 275–293, 19 jun. 2017.

FESCHOTTE, C.; PRITHAM, E. Non-mammalian c-integrases are encoded by giant transposable elements. **Trends in Genetics**, Amsterdam, v. 21, n. 10, p. 551–552, out. 2005.

FESCHOTTE, Cédric. Merlin, a New Superfamily of DNA Transposons Identified in Diverse Animal Genomes and Related to Bacterial IS1016 Insertion Sequences. **Molecular Biology and Evolution**, New York, v. 21, n. 9, p. 1769–1780, set. 2004.

FESCHOTTE, Cédric; MOUCHÈS, C. Evidence that a Family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the *Arabidopsis thaliana* Genome Has Arisen from a pogo-like DNA Transposon. **Molecular Biology and Evolution**, New York, v. 17, n. 5, p. 730–737, 1 maio 2000.

FINNEGAN, D. J. et al. Repeated Gene Families in *Drosophila melanogaster*. **Cold Spring Harbor Symposia on Quantitative Biology**, , New York, v. 42, p. 1053–1063, 1 jan. 1978.

FLUTRE, T. et al. Considering transposable element diversification in de novo annotation approaches. **PLoS ONE**, San Francisco, v. 6, n. 1, p. e16526, 2011. Disponível em <<https://dx.plos.org/10.1371/journal.pone.0016526>>. Acesso em: 14 jun. 2021.

FLYNN, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 117, n. 17, p. 9451–9457, 28 abr. 2020.

FOSTER, W. A.; WALKER, E. D. Mosquitoes (Culicidae). In: MULLEN, G. R.; DURDEN, L. (Org.). **Med. Vet. Entomol.** London: Elsevier, 2019. p. 261–325.

FOUCHÉ, S. et al. Stress-Driven Transposable Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. **Molecular Biology and Evolution**, New York, v. 37, n. 1, p. 221–239, 1 jan. 2020.

FRIEDLI, M.; TRONO, D. The Developmental Control of Transposable Elements and the Evolution of Higher Species. **Annual Review of Cell and Developmental Biology**, Palo Alto, v. 31, n. 1, p. 429–451, 13 nov. 2015.

GALILI, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. **Bioinformatics**, Oxford, v. 31, n. 22, p. 3718–3720, 2015.

GAO, D. et al. Horizontal transfer of Non-LTR retrotransposons from arthropods to flowering plants. **Molecular Biology and Evolution**, New York, v. 35, n. 2, p. 354–364, 2018.

GAO, D. et al. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. **Genome Biology**, London, v. 17, n. 1, p. 7, 18 dez. 2016. Disponível em: <<http://dx.doi.org/10.1186/s13059-015-0867-y>>. Acesso em: 14 jun. 2021.

GARCIA-PEREZ, J. L.; WIDMANN, T. J.; ADAMS, I. R. The impact of transposable elements on mammalian development. **Development**, Cambridge, v. 143, n. 22, p. 4101–4114, 15 nov. 2016.

GEER, L. Y. et al. Open mass spectrometry search algorithm. **Journal of Proteome Research**, Washington, D.C, v. 3, n. 5, p. 958–964, 2004.

GENETIC INFORMATION RESEARCH INSTITUTE. **Rebase**. Disponível em: <<https://www.girinst.org/rebase/>>. Acesso em: 20 jun. 2022.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology**, London, v. 5, n. 10, p. R80, 2004. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/15461798>>. Acesso em: 14 jun. 2021.

GILBERT, C. et al. A role for host–parasite interactions in the horizontal transfer of transposons across phyla. **Nature**, London, v. 464, n. 7293, p. 1347–1350, abr. 2010.

GILBERT, C.; FESCHOTTE, C. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. **Current Opinion in Genetics and Development**, London, v. 49, n. October 2017, p. 15–24, 2018.

GIRALDO-CALDERÓN, G. I. et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. **Nucleic Acids Research**, Oxford, v. 43, n. D1, p. D707–D713, 28 jan. 2015.

GIRALDO-CALDERÓN, G. I. et al. VectorBase.org updates: bioinformatic resources for invertebrate vectors of human pathogens and related organisms. **Current Opinion in Insect Science**, London, v. 50, p. 100860, abr. 2022. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S2214574521001292>>. Acesso em: 30 abr. 2022.

GLADYSHEV, E. A.; ARKHIPOVA, I. R. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 104, n. 22, p. 9352–9357, 2007.

GOERNER-POTVIN, P.; BOURQUE, G. Computational tools to unmask transposable elements. **Nature Reviews Genetics**, London, v. 19, n. 11, p. 688–704, 19 nov. 2018.

GOLDBERG, M. L. et al. Unequal crossing-over associated with asymmetrical synapsis between nomadic elements in the *Drosophila melanogaster* genome. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 80, n. 16, p. 5017–5021, 1 ago. 1983.

GOODWIN, T. J. D.; BUTLER, M. I.; POULTER, R. T. M. Cryptons: A group of tyrosine-

recombinase- encoding DNA transposons from pathogenic fungi. **Microbiology**, Reading, v. 149, n. 11, p. 3099–3109, 2003.

GOUBERT, Clement et al. High-throughput sequencing of transposable element insertions suggests adaptive evolution of the invasive Asian tiger mosquito towards temperate environments. **Molecular Ecology**, Hoboken, v. 26, n. 15, p. 3968–3981, 2017.

GOUBERT, Clément et al. De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). **Genome Biology and Evolution**, Oxford, v. 7, n. 4, p. 1192–1205, 2015.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature Biotechnology**, New York, v. 29, n. 7, p. 644–652, 15 jul. 2011.

GRABUNDZIJA, I. et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. **Nature Communications, New York, New York**, v. 7, n. 1, p. 10716, 2 abr. 2016. Disponível em: <<http://www.nature.com/articles/ncomms10716>>. Acesso em: 30 abr. 2020.

GRBIĆ, M. et al. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. **Nature**, London, v. 479, n. 7374, p. 487–492, 23 nov. 2011.

GREEN, C. L.; FROMMER, M. The genome of the Queensland fruit fly *Bactrocera tryoni* contains multiple representatives of the mariner family of transposable elements. **Insect Molecular Biology**, London, v. 10, n. 4, p. 371–386, 2001.

GUINDON, S. et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. **Systematic Biology, Washington, D.C, Washington, D.C**, v. 59, n. 3, p. 307–321, 29 mar. 2010.

GUIO, L.; BARRÓN, M. G.; GONZÁLEZ, J. The transposable element Bari-Jheh mediates oxidative stress response in *Drosophila*. **Molecular Ecology**, Hoboken, v. 23, n. 8, p. 2020–2030, 2014.

GUO, J. et al. The adult human testis transcriptional cell atlas. **Cell Research**, Basingstoke, v. 28, n. 12, p. 1141–1157, 12 dez. 2018.

HAAPA-PAANANEN, S.; WAHLBERG, N.; SAVILAHTI, H. Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. **Molecular Phylogenetics and Evolution**, San Diego, v. 78, n. 1, p. 271–274, 2014.

HAAS, B. **Genome Guided Trinity Transcriptome Assembly**. Disponível em: <<https://github.com/trinityrnaseq/trinityrnaseq/wiki/Genome-Guided-Trinity-Transcriptome-Assembly>>. Acesso em: 8 fev. 2022.

HAHN, C.; BACHMANN, L.; CHEVREUX, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. **Nucleic Acids Research**, Oxford, v. 41, n. 13, p. e129, 1 jul. 2013. Disponível em: <<https://academic.oup.com/nar/article/41/13/e129/1129833>>. Acesso em: 30 abr. 2020.

HAIIG, D. Transposable elements: Self-seekers of the germline, team-players of the soma. **BioEssays**, Cambridge, v. 38, n. 11, p. 1158–1166, nov. 2016.

HAN, J. S. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. **Mobile DNA**, London, v. 1, n. 1, p. 15, 2010. Disponível em: <<http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-1-15>>. Acesso em: 30 abr. 2020.

HAN, M.-J. et al. Spy: A New Group of Eukaryotic DNA Transposons without Target Site Duplications. **Genome Biology and Evolution**, Oxford, v. 6, n. 7, p. 1748–1757, jul. 2014.

HAN, M. et al. The diversification of PHIS transposon superfamily in eukaryotes. **Mobile DNA**, London, v. 6, n. 1, p. 12, 24 dez. 2015. Disponível em: <<http://dx.doi.org/10.1186/s13100-015-0043-7>>. Acesso em: 30 abr. 2020.

HARBACH, R. E. **Mosquito Taxonomic Inventory**. Disponível em: <<http://mosquito-taxonomic-inventory.info/>>. Acesso em: 6 nov. 2019.

HE, J. et al. Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. **Nature Communications**, New York, v. 12, n. 1, p. 1456, 5 dez. 2021. Disponível em: <<http://dx.doi.org/10.1038/s41467-021-21808-x>>. Acesso em: 30 abr. 2020.

HEDGES, S. B. et al. Tree of life reveals clock-like speciation and diversification. **Molecular Biology and Evolution**, New York, v. 32, n. 4, p. 835–845, 2015.

HERINGER, P.; KUHN, G. C. S. Pif1 Helicases and the Evidence for a Prokaryotic Origin of Helitrons. **Molecular Biology and Evolution**, New York, v. 39, n. 1, 2022.

HEZRONI, H. et al. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. **Cell Reports**, Cambridge, v. 11, n. 7, p. 1110–1122, 2015.

HIKOSAKA, A.; KONISHI, S. Multiple massive domestication and recent amplification of Kolobok superfamily transposons in the clawed frog *Xenopus*. **Zoological Letters**, London, v. 4, n. 1, p. 17, 16 dez. 2018. Disponível em: <<https://zoologicalletters.biomedcentral.com/articles/10.1186/s40851-018-0100-4>>. Acesso em: 30 abr. 2020.

HIRANO, T.; SIOMI, H. Small RNAs: Artificial piRNAs for Transcriptional Silencing. **Current Biology**, Cambridge, v. 25, n. 7, p. R280–R283, mar. 2015.

HIRSCH, C. D.; SPRINGER, N. M. Transposable element influences on gene expression in plants. **Biochimica et Biophysica Acta**, Amsterdam, v. 1860, n. 1, p. 157–165, 2016.

HOEDE, C. et al. PASTEC: An Automatic Transposable Element Classification Tool. **PLoS ONE**, San Francisco, v. 9, n. 5, p. e91929, 2 maio 2014. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0091929>>. Acesso em: 1 jan. 2022.

HOF, A. E. V. van't et al. The industrial melanism mutation in British peppered moths is a transposable element. **Nature**, London, v. 534, n. 7605, p. 102–105, 1 jun. 2016.

HOLT, R. A. et al. The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*. **Science**, New York, v. 298, n. 5591, p. 129–149, 4 out. 2002.

HORVÁTH, V.; MERENCIANO, M.; GONZÁLEZ, J. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. **Trends in Genetics**, Amsterdam, v. 33, n. 11, p. 832–841, nov. 2017.

HOUWING, S. et al. A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. **Cell**, Cambridge, v. 129, n. 1, p. 69–82, 2007.

HUANG, J. et al. EARE-1, a Transcriptionally Active Ty1/Copia-Like Retrotransposon Has Colonized the Genome of *Excoecaria agallocha* through Horizontal Transfer. **Frontiers in Plant Science**, Lausanne, v. 8, n. January, p. 45, 24 jan. 2017. Disponível em: <<http://journal.frontiersin.org/article/10.3389/fpls.2017.00045/full>>. Acesso em: 1 jan. 2022.

HUANG, Xiaoqi. On global sequence alignment. **Bioinformatics**, Oxford, v. 10, n. 3, p. 227–235, 1994. Disponível em: <<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/10.3.227>>. Acesso em: 1 jan. 2022.

HUANG, Xin; POELCHAU, M. F.; ARMBRUSTER, P. A. Global Transcriptional Dynamics of Diapause Induction in Non-Blood-Fed and Blood-Fed *Aedes albopictus*. **PLOS Neglected Tropical Diseases**, San Francisco, v. 9, n. 4, p. e0003724, 21 abr. 2015. Disponível em: <<https://dx.plos.org/10.1371/journal.pntd.0003724>>. Acesso em: 1 jan. 2022.

HULSTAERT, N. et al. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. **Journal of Proteome Research**, Washington, D.C, v. 19, n. 1, p. 537–542, 3 jan. 2020.

HUMMEL, B. et al. The evolutionary capacitor HSP90 buffers the regulatory effects of mammalian endogenous retroviruses. **Nature Structural and Molecular Biology**, New York, v. 24, n. 3, p. 234–242, 2017.

HUSON, D. H.; SCORNAVACCA, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. **Systematic Biology**, Washington, D.C, v. 61, n. 6, p. 1061–1067, 2012.

HUYNH-THU, V. A.; SANGUINETTI, G. Gene Regulatory Network Inference: An Introductory Survey. **Gene Regul. Networks. Methods Mol. Biol.** New York: Humana Press, 2019. p. 1–23.

IMASE, A. et al. Existence of host DNA sequences in schistosomes—horizontal and vertical transmission. **Parasitology International**, Amsterdam, v. 52, n. 4, p. 369–373, dez. 2003.

ITO, H.; KAKUTANI, T. Control of transposable elements in *Arabidopsis thaliana*. **Chromosome Research**, Oxford, v. 22, n. 2, p. 217–223, 2014.

IVANCEVIC, A. M. et al. Horizontal transfer of BovB and L1 retrotransposons in

eukaryotes. **Genome Biology**, London, v. 19, n. 1, p. 85, 9 dez. 2018. Disponível em: <<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1456-7>>. Acesso em: 1 jan. 2022.

IVANCEVIC, A. M. et al. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. **Genome Biology and Evolution**, Oxford, v. 8, n. 11, p. 3301–3322, nov. 2016.

JACOBSON, J. W.; MEDHORA, M. M.; HARTL, D. L. Molecular structure of a somatically unstable transposable element in *Drosophila*. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 83, n. 22, p. 8684–8688, 1986.

JANG, H. S. et al. Transposable elements drive widespread expression of oncogenes in human cancers. **Nature Genetics**, New York, v. 51, n. 4, p. 611–617, 2019.

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

JARVIK, T.; LARK, K. G. Characterization of Soymar1, a mariner element in soybean. **Genetics**, Austin, v. 149, n. 3, p. 1569–1574, jul. 1998.

JIANG, N. et al. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). **Current Opinion in Plant Biology**, London, v. 7, n. 2, p. 115–119, 2004.

JIN, Y. et al. TETranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. **Bioinformatics**, Oxford, v. 31, n. 22, p. 3593–3599, 15 nov. 2015.

JOHNSON, R.; GUIGÓ, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. **RNA**, New York, v. 20, n. 7, p. 959–976, jul. 2014.

JOHNSON, W. E.; LI, C.; RABINOVIC, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. **Biostatistics**, Oxford, v. 8, n. 1, p. 118–127, 1 jan. 2007.

JOLY-LOPEZ, Z.; BUREAU, T. E. Exaptation of transposable element coding sequences. **Current Opinion in Genetics & Development**, London, v. 49, p. 34–42, abr. 2018.

JURKA, J. et al. Repbase Update, a database of eukaryotic repetitive elements. **Cytogenetic and Genome Research**, Basel, v. 110, n. 1–4, p. 462–467, 2005.

KAJIKAWA, M.; OKADA, N. LINEs mobilize SINEs in the eel through a shared 3' sequence. **Cell**, Cambridge, v. 111, n. 3, p. 433–444, 2002.

KALENDAR, R. et al. Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes. **Genetics**, Austin, v. 166, n. 3, p. 1437–1450, 2004.



- KAMAL, M. et al. Mapping the global potential distributions of two arboviral vectors *Aedes aegypti* and *Ae. albopictus* under changing climate. **PLoS ONE**, San Francisco, v. 13, n. 12, p. e0210122, 31 dez. 2018. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0210122>>. Acesso em: 12 jan. 2022.
- KANG, D. S. et al. A transcriptomic survey of the impact of environmental stress on response to dengue virus in the mosquito, *Aedes aegypti*. **PLOS Neglected Tropical Diseases**, San Francisco, v. 12, n. 6, p. e0006568, 11 jun. 2018. Disponível em: <<https://dx.plos.org/10.1371/journal.pntd.0006568>>. Acesso em: 12 jan. 2022.
- KAPITONOV, V. V.; JURKA, J. Rolling-circle transposons in eukaryotes. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 98, n. 15, p. 8714–8719, 2001.
- KAPITONOV, Vladimir V.; JURKA, J. A novel class of SINE elements derived from 5S rRNA. **Molecular Biology and Evolution**, New York, v. 20, n. 5, p. 694–702, 2003a.
- KAPITONOV, Vladimir V.; JURKA, J. Harbinger Transposons and an Ancient HARBI1 Gene Derived from a Transposase. **DNA and Cell Biology**, Larchmont, v. 23, n. 5, p. 311–324, maio 2004.
- KAPITONOV, Vladimir V.; JURKA, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 100, n. 11, p. 6569–6574, 2003b.
- KAPITONOV, Vladimir V.; JURKA, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. **PLoS Biology**, San Francisco, v. 3, n. 6, p. 0998–1011, 2005.
- KAPITONOV, Vladimir V.; KOONIN, E. V. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. **Biology Direct**, London, v. 10, n. 1, p. 20, 28 dez. 2015. Disponível em: <<https://biologydirect.biomedcentral.com/articles/10.1186/s13062-015-0055-8>>. Acesso em: 12 jan. 2022.
- KAPITONOV, V V; JURKA, J. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. **Genetica**, Dordrecht, v. 107, n. 1–3, p. 27–37, 1999.
- KAPITONOV, Vladimir V; JURKA, J. A universal classification of eukaryotic transposable elements implemented in Repbase. **Nature Reviews Genetics**, London, v. 9, n. 5, p. 411–412, maio 2008.
- KAPITONOV, Vladimir V; TEMPEL, S.; JURKA, J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. **Gene**, Amsterdam, v. 448, n. 2, p. 207–213, dez. 2009.
- KAPUSTA, A. et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. **PLoS Genetics**, San Francisco, v. 9, n. 4, p. e1003470, 25 abr. 2013. Disponível em: <<https://dx.plos.org/10.1371/journal.pgen.1003470>>. Acesso em: 12 jan. 2022.

KARAKÜLAH, G. et al. TEeffectR: An R package for studying the potential effects of transposable elements on gene expression with linear regression model. **PeerJ**, Corte Madera, v. 2019, n. 12, p. e8192, 2019. Disponível em: <<https://peerj.com/articles/8192>>. Acesso em: 12 jan. 2022.

KELLEY, D.; RINN, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. **Genome Biology**, London, v. 13, n. 11, p. R107, 2012. Disponível em: <<http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-11-r107>>. Acesso em: 12 jan. 2022.

KEMPKEN, F.; WINDHOFER, F. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. **Chromosoma**, Wien, v. 110, n. 1, p. 1–9, abr. 2001.

KIDWELL, M. G.; LISCH, D. R. Perspective: Transposable Elements, Parasitic DNA, and Genome Evolution. **Evolution**, Lancaster, v. 55, n. 1, p. 1–24, jan. 2001. Disponível em: <<http://doi.wiley.com/10.1111/j.0014-3820.2001.tb01268.x>>. Acesso em: 20 jan. 2022.

KIKUCHI, K. et al. The plant MITE mPing is mobilized in anther culture. **Nature**, London, v. 421, n. 6919, p. 167–170, jan. 2003.

KIM, S.; PEVZNER, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. **Nature Communications**, New York, v. 5, n. 1, p. 5277, 31 dez. 2014. Disponível em: <<http://www.nature.com/articles/ncomms6277>>. Acesso em: 20 jan. 2022.

KING, A. M. Q. et al. (Org.). **Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses**. [S.l.]: Elsevier, 2011. Disponível em: <<https://www.elsevier.com/books/virus-taxonomy/9780123846846>>. Acesso em: 20 jan. 2022.

KITTS, P. A. et al. Assembly: a resource for assembled genomes at NCBI. **Nucleic Acids Research**, Oxford, v. 44, n. D1, p. D73–D80, 4 jan. 2016.

KNUDSEN, G. M.; CHALKLEY, R. J. The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis. **PLoS ONE**, San Francisco, v. 6, n. 6, p. e20873, 14 jun. 2011. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0020873>>. Acesso em: 20 jan. 2022.

KOJIMA, K. K. A new class of SINEs with snRNA gene-derived heads. **Genome Biology and Evolution**, Oxford, v. 7, n. 6, p. 1702–1712, 2015.

KOJIMA, K. K. Human transposable elements in Repbase: genomic footprints from fish to humans. **Mobile DNA, London**, v. 9, n. 1, p. 2, 4 dez. 2018. Disponível em: <<https://mobileDNAjournal.biomedcentral.com/articles/10.1186/s13100-017-0107-y>>. Acesso em: 20 jan. 2022.

KOJIMA, K. K. Structural and sequence diversity of eukaryotic transposable elements. **Genes & Genetic Systems**, Mishima, v. 94, n. 6, p. 233–252, 1 dez. 2019.

KOJIMA, K. K.; FUJIWARA, H. An extraordinary retrotransposon family encoding dual endonucleases. **Genome Research**, New York, v. 15, n. 8, p. 1106–1117, 2005.

- KOJIMA, K. K.; JURKA, J. Crypton transposons: identification of new diverse families and ancient domestication events. **Mobile DNA**, London, v. 2, n. 1, p. 12, 2011. Disponível em: <<http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-2-12>>. Acesso em: 20 jan. 2022.
- KOJIMA, K. K.; KAPITONOV, V. V.; JURKA, J. Recent expansion of a new ingi-related clade of Vingi non-LTR retrotransposons in hedgehogs. **Molecular Biology and Evolution**, New York, v. 28, n. 1, p. 17–20, 2011.
- KONG, A. T. et al. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. **Nature Methods**, New York, v. 14, n. 5, p. 513–520, 2017.
- KOONIN, E. V.; DOLJA, V. V. Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements. **Microbiology and Molecular Biology Reviews**, Washington, DC, v. 78, n. 2, p. 278–303, 2014.
- KOONIN, E. V.; KRUPOVIC, M.; YUTIN, N. Evolution of double-stranded DNA viruses of eukaryotes: From bacteriophages to transposons to giant viruses. **Annals of the New York Academy of Sciences**, New York, v. 1341, n. 1, p. 10–24, 2015.
- KOZIOL, U. et al. A Novel Terminal-Repeat Retrotransposon in Miniature (TRIM) Is Massively Expressed in *Echinococcus multilocularis* Stem Cells. **Genome Biology and Evolution**, Oxford, v. 7, n. 8, p. 2136–2153, ago. 2015.
- KRAMEROV, D. A.; VASSETZKY, N. S. SINEs. **Wiley Interdisciplinary Reviews: RNA**, Hoboken, v. 2, n. 6, p. 772–786, nov. 2011.
- KRIVENTSEVA, E. V. et al. OrthoDB: the hierarchical catalog of eukaryotic orthologs. **Nucleic Acids Research**, Oxford, v. 36, n. Database issue, p. D271–275, 2007.
- KRUPOVIC, M. et al. Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. **Journal of Virology**, Baltimore, v. 92, n. 12, p. e00515–18, 15 jun. 2018. Disponível em: <<http://jvi.asm.org/lookup/doi/10.1128/JVI.00515-18>>.
- KRUPOVIC, M.; BAMFORD, D. H.; KOONIN, E. V. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. **Biology Direct**, London, v. 9, p. 6, 2014. Disponível em: <<http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-9-6>>. Acesso em: 20 jan. 2022.
- KRUPOVIC, M.; KOONIN, E. V. Self-synthesizing transposons: Unexpected key players in the evolution of viruses and defense systems. **Current Opinion in Microbiology**, London, v. 31, n. February, p. 25–33, 2016.
- KUMAR, D.; YADAV, A. K.; DASH, D. Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data. In: KEERTHIKUMAR, S.; MATHIVANAN, S. (Org.). . **Proteome Bioinforma**. [S.l: s.n.], 2017. p. 17–29.
- KUMAR, M. et al. De novo transcriptome sequencing and analysis of the cereal cyst

nematode, *Heterodera avenae*. **PLoS ONE**, San Francisco, v. 9, n. 5, p. e96311, 2014. Disponível em <<https://dx.plos.org/10.1371/journal.pone.0096311>>. Acesso em: 20 jan. 2022.

KUMAR, S. et al. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. **Molecular Biology and Evolution**, New York, v. 34, n. 7, p. 1812–1819, 2017.

LABBÉ, J. et al. Characterization of Transposable Elements in the Ectomycorrhizal Fungus *Laccaria bicolor*. **PLoS ONE**, San Francisco, v. 7, n. 8, p. e40197, 3 ago. 2012. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0040197>>. Acesso em: 20 jan. 2022.

LEE, H.; ZHANG, Z.; KRAUSE, H. M. Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? **Trends in Genetics**, Amsterdam, v. 35, n. 12, p. 892–902, dez. 2019.

LEEK, J. T. et al. **sva: Surrogate Variable Analysis**. R package version 3.32.1. [S.l: s.n.]. , 2019.

LEEK, J. T. et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. **Bioinformatics, Oxford**, v. 28, n. 6, p. 882–883, 15 mar. 2012.

LEEK, J. T.; STOREY, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. **PLoS Genetics**, San Francisco, v. 3, n. 9, p. e161, 28 set. 2007. Disponível em: <<https://dx.plos.org/10.1371/journal.pgen.0030161>>. Acesso em: 20 jan. 2022.

LEPREVOST, F. V. et al. PepExplorer: A Similarity-driven Tool for Analyzing de Novo Sequencing Results. **Molecular & Cellular Proteomics**, Bethesda, v. 13, n. 9, p. 2480–2489, set. 2014.

LEVIN, H. L.; MORAN, J. V. Dynamic interactions between transposable elements and their hosts. **Nature Reviews Genetics**, London, v. 12, n. 9, p. 615–627, set. 2011.

LEVIS, R.; DUNSMUIR, P.; RUBIN, G. M. Terminal repeats of the drosophila transposable element copia: Nucleotide sequence and genomic organization. **Cell**, Cambridge, v. 21, n. 2, p. 581–588, set. 1980.

LEWIS, S. H. et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. **Nature Ecology & Evolution**, London, v. 2, n. 1, p. 174–181, 4 jan. 2018.

LI, H. et al. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. **BMC Genomics**, London, v. 17, n. Suppl 13, 2016. Disponível em: <<http://dx.doi.org/10.1186/s12864-016-3327-5>>. Acesso em: 20 jan. 2022.

LI, Wanhe et al. Activation of transposable elements during aging and neuronal decline in *Drosophila*. **Nature Neuroscience**, New York, v. 16, n. 5, p. 529–531, 7 maio 2013.

LI, Weizhong; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, Oxford, v. 22, n. 13, p. 1658–1659, 1 jul.

2006.

LI, X. et al. A resurrected mammalian hAT transposable element and a closely related insect element are highly active in human cell culture. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 110, n. 6, p. E478–E487, 2013.

LIESECKE, F. et al. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. **Scientific Reports**, London, v. 9, n. 1, p. 14431, 8 dez. 2019. Disponível em: <<http://www.nature.com/articles/s41598-019-50885-8>>. Acesso em: 30 jan. 2022.

LIN, X.; FARIDI, N.; CASOLA, C. An ancient transkingdom horizontal transfer of penelope-like retroelements from arthropods to conifers. **Genome Biology and Evolution**, Oxford, v. 8, n. 4, p. 1252–1266, 2016.

LIPATOV, M. et al. Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. **BMC Biology**, London, v. 3, n. 1, p. 24, 12 dez. 2005. Disponível em: <<https://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-3-24>>. Acesso em: 30 jan. 2022.

LIPPMAN, Z. et al. Role of transposable elements in heterochromatin and epigenetic control. **Nature**, London, v. 430, n. 6998, p. 471–476, jul. 2004.

LISCH, D. Mutator transposons. **Trends in Plant Science**, New York, v. 7, n. 11, p. 498–504, 2002.

LISCH, D.; JIANG, N. Mutator and MULE transposons. **Handbook of Maize**. New York, NY: Springer New York, 2009. p. 277–306.

LIU, C.; YANG, Y.; SCHATZ, D. G. Structures of a RAG-like transposase during cut-and-paste transposition. **Nature**, London, v. 575, n. 7783, p. 540–544, 13 nov. 2019.

LIU, K.; WESSLER, S. R. Functional characterization of the active Mutator-like transposable element, Mut1 from the mosquito *Aedes aegypti*. **Mobile DNA**, London, v. 8, n. 1, p. 1, 11 dez. 2017. Disponível em: <<http://dx.doi.org/10.1186/s13100-016-0084-6>>.

LIU, P. et al. Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*. **Parasites & Vectors**, London, v. 9, n. 1, p. 524, 29 dez. 2016. Disponível em: <<http://dx.doi.org/10.1186/s13071-016-1815-8>>. Acesso em: 30 jan. 2022.

LIU, Z. et al. The Wukong Terminal-Repeat Retrotransposon in Miniature (TRIM) Elements in Diverse Maize Germplasm. **G3 Genes|Genomes|Genetics**, Bethesda, v. 5, n. 8, p. 1585–1592, 1 ago. 2015.

LLORENS, C. et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. **Nucleic Acids Research**, Oxford, v. 39, n. Database, p. D70–D74, 1 jan. 2011.

LOBO, N. et al. Mobility of the *PiggyBac* transposon in embryos of the vectors of Dengue fever (*Aedes albopictus*) and La Crosse encephalitis (*Ae. triseriatus*). **Molecular Genetics**

**and Genomics**, Berlin, v. 265, n. 1, p. 66–71, 27 mar. 2001.

LOCKTON, S.; GAUT, B. S. The Contribution of Transposable Elements to Expressed Coding Sequence in *Arabidopsis thaliana*. **Journal of Molecular Evolution**, Berlin, v. 68, n. 1, p. 80–89, 3 jan. 2009.

LOPEZ-EZQUERRA, A.; HARRISON, M. C.; BORNBERG-BAUER, E. Comparative analysis of lincRNA in insect species. **BMC Evolutionary Biology**, London, v. 17, n. 1, p. 155, 3 dez. 2017. Disponível em: <<http://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-017-0985-0>>. Acesso em: 30 jan. 2022.

LU, J. et al. Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. **BMC Genomics**, London, v. 13, n. 1, p. 246, 2012. Disponível em: <<http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-246>>.

LUDWIG, A. et al. Transposable elements expression in *Rhinella marina* (cane toad) specimens submitted to immune and stress challenge. **Genetica**, Dordrecht, v. 149, n. 5–6, p. 335–342, 12 dez. 2021.

MA, B. Novor: Real-Time Peptide de Novo Sequencing Software. **Journal of the American Society for Mass Spectrometry**, New York, v. 26, n. 11, p. 1885–1894, 2015.

MA, Q. et al. A mosquito small RNA genomics resource reveals dynamic evolution and host responses to viruses and transposons. **Genome Research**, New York, v. 31, n. 3, p. 512–528, mar. 2021.

MACCHIETTO, M. G.; LANGLOIS, R. A.; SHEN, S. S. Virus-induced transposable element expression up-regulation in human and mouse host cells. **Life Science Alliance**, Woodbury, v. 3, n. 2, p. e201900536, fev. 2020. Disponível em: <<https://www.life-science-alliance.org/lookup/doi/10.26508/lsa.201900536>>. Acesso em: 30 jan. 2022.

MACKENZIE-IMPOINVIL, L. et al. Contrasting patterns of gene expression indicate differing pyrethroid resistance mechanisms across the range of the New World malaria vector *Anopheles albimanus*. **PLoS ONE**, San Francisco, v. 14, n. 1, p. e0210586, 30 jan. 2019. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0210586>>. Acesso em: 30 jan. 2022.

MAGDELDIN, S. et al. Basics and recent advances of two dimensional- polyacrylamide gel electrophoresis. **Clinical Proteomics**, Totowa, v. 11, n. 1, p. 16, 2014. Disponível em: <<http://clinicalproteomicsjournal.biomedcentral.com/articles/10.1186/1559-0275-11-16>>. Acesso em: 30 jan. 2022.

MAO, S. et al. RefShannon: A genome-guided transcriptome assembler using sparse flow decomposition. **PLoS ONE**, San Francisco, v. 15, n. 6, p. e0232946, 2 jun. 2020. Disponível em: <<http://dx.doi.org/10.1371/journal.pone.0232946>>. Acesso em: 30 jan. 2022.

MARASCA, F. et al. The sophisticated transcriptional response governed by transposable elements in human health and disease. **International Journal of Molecular Sciences**, Basel,

v. 21, n. 9, p. 3201, 2020. Disponível em <<https://www.mdpi.com/1422-0067/21/9/3201>>. Acesso em: 30 jan. 2022.

MARCHESE, F. P.; RAIMONDI, I.; HUARTE, M. The multidimensional mechanisms of long noncoding RNA function. **Genome Biology**, London, v. 18, n. 1, p. 206, 31 dez. 2017. Disponível em: <<http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1348-2>>. Acesso em: 30 jan. 2022.

MARGOLIN, A. A. et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. **BMC Bioinformatics**, Oxford, v. 7, n. S1, p. S7, 20 mar. 2006. Disponível em: <<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S1-S7>>. Acesso em: 30 out. 2021.

MARIN, P. et al. Phenotypic and Transcriptomic Responses to Stress Differ According to Population Geography in an Invasive Species. **Genome Biology and Evolution**, Oxford, v. 13, n. 9, p. 1–17, 1 set. 2021.

MARINGER, K. et al. Proteomics informed by transcriptomics for characterising active transposable elements and genome annotation in *Aedes aegypti*. **BMC Genomics**, London, v. 18, n. 1, p. 101, 19 dez. 2017. Disponível em: <<http://dx.doi.org/10.1186/s12864-016-3432-5>>. Acesso em: 30 out. 2021.

MARINOTTI, O. et al. The Genome of *Anopheles darlingi*, the main neotropical malaria vector. **Nucleic Acids Research**, Oxford, v. 41, n. 15, p. 7387–7400, ago. 2013.

MARSANO, R. M. et al. Mosquitoes LTR Retrotransposons: A Deeper View into the Genomic Sequence of *Culex quinquefasciatus*. **PLoS ONE**, San Francisco, v. 7, n. 2, p. e30770, 24 fev. 2012. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0030770>>. Acesso em: 30 out. 2021.

MARTINS, M. et al. Comprehensive Quantitative Proteome Analysis of *Aedes aegypti* Identifies Proteins and Pathways Involved in *Wolbachia pipientis* and Zika Virus Interference Phenomenon. **Frontiers in Physiology**, Lausanne, v. 12, n. February, p. 642237, 25 fev. 2021. Disponível em: <<https://www.frontiersin.org/articles/10.3389/fphys.2021.642237/full>>. Acesso em: 30 out. 2021.

MASSON, P. et al. Essential large transcripts of the maize Spm transposable element are generated by alternative splicing. **Cell**, Cambridge, v. 58, n. 4, p. 755–765, ago. 1989.

MATSUTANI, S. et al. Isolation and characterization of IS elements repeated in the bacterial chromosome. **Journal of Molecular Biology**, Amsterdam, v. 196, n. 3, p. 445–455, ago. 1987.

MATTHEWS, B. J. et al. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. **Nature**, London, v. 563, n. 7732, p. 501–507, 14 nov. 2018.

MATTICK, J. S. The State of Long Non-Coding RNA Biology. **Non-Coding RNA**, Basel, v. 4, n. 3, p. 17, 10 ago. 2018. Disponível em: <<http://www.mdpi.com/2311-553X/4/3/17>>. Acesso em: 30 out. 2021.

MCCLINTOCK, B. The origin and behavior of mutable loci in maize. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 36, n. 6, p. 344–355, 1950.

MCCLINTOCK, Barbara. **Mutations in maize and chromosomal aberrations in Neurospora**. Washington: Cold Spring Harbor, 1954.

METCALFE, C. J.; CASANE, D. Modular organization and reticulate evolution of the ORF1 of Jockey superfamily transposable elements. **Mobile DNA**, London, v. 5, n. 1, p. 19, 2014. Disponível em: <<http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-5-19>>. Acesso em: 30 out. 2021.

METZGER, M. J. et al. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 115, n. 18, p. E4227–E4235, 1 maio 2018.

MEYER, P. E. et al. Information-Theoretic Inference of Large Transcriptional Regulatory Networks. **EURASIP Journal on Bioinformatics and Systems Biology**, Heidelberg, v. 2007, p. Article 79879, 2007. Disponível em: <<http://bsb.eurasipjournals.com/content/2007/1/79879>>. Acesso em: 30 out. 2021.

MISTRY, J. et al. Pfam: The protein families database in 2021. **Nucleic Acids Research**, Oxford, v. 49, n. D1, p. D412–D419, 8 jan. 2021.

MITCHELL, N. M. et al. Proteogenomic Re-Annotation of *Coccidioides posadasii* Strain Silveira. **Proteomics**, Weinheim, v. 18, n. 1, p. 1700173, jan. 2018. Disponível em: <<https://onlinelibrary.wiley.com/doi/10.1002/pmic.201700173>>. Acesso em: 30 out. 2021.

MOMOSE, M.; ABE, Y.; OZEKI, Y. Miniature inverted-repeat transposable elements of Stowaway are active in potato. **Genetics**, Austin, v. 186, n. 1, p. 59–66, 2010.

MORETON, J.; IZQUIERDO, A.; EMES, R. D. Assembly, Assessment, and Availability of De novo Generated Eukaryotic Transcriptomes. **Frontiers in Genetics**, Lausanne, v. 6, n. JAN, p. 361, 11 jan. 2016. Disponível em: <<http://journal.frontiersin.org/Article/10.3389/fgene.2015.00361/abstract>>. Acesso em: 30 out. 2021.

MUKAKA, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. **Malawi medical journal : the journal of Medical Association of Malawi**, Lilongwe, v. 24, n. 3, p. 69–71, set. 2012.

MULLER, H. et al. Assessing the Impact of a Viral Infection on the Expression of Transposable Elements in the Cabbage Looper Moth (*Trichoplusia ni*). **Genome Biology and Evolution**, Oxford, v. 13, n. 11, p. 1–17, 5 nov. 2021.

MUSZEWSKA, A. et al. Transposable elements contribute to fungal genes and impact fungal lifestyle. **Scientific Reports**, London, v. 9, n. 1, p. 4307, 13 dez. 2019. Disponível em: <<http://www.nature.com/articles/s41598-019-40965-0>>. Acesso em: 30 out. 2021.

MUTH, T. et al. A Potential Golden Age to Come—Current Tools, Recent Use Cases, and



Future Avenues for De Novo Sequencing in Proteomics. **Proteomics**, Weinheim, v. 18, p. 1700150, 2018.

MUTH, T.; RENARD, B. Y. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? **Briefings in Bioinformatics**, London, v. 19, n. 5, p. 954–970, 28 set. 2018.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **NCBI: Assembly**. Disponível em: <<https://www.ncbi.nlm.nih.gov/assembly/>>. Acesso em: 14 jun. 2022.

NEAFSEY, D. E. et al. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. **Science**, New York, v. 347, n. 6217, p. 1258522, 2 jan. 2015. Disponível em: <<https://www.sciencemag.org/lookup/doi/10.1126/science.1258522>>. Acesso em: 30 out. 2021.

NENE, V. et al. Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. **Science, New York**, v. 316, n. 5832, p. 1718–1723, 22 jun. 2007.

O'NEILL, J. R. An Overview of Mass Spectrometry-Based Methods for Functional Proteomics. In: WANG, X.; KURUC, M. (Org.). . **Functional Proteomics: Methods and Protocols** [S.l: s.n.], 2019. p. 179–196.

OHSHIMA, K.; OKADA, N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. **Cytogenetic and Genome Research**, Basel, Possui muitas referências sobre diversas superfamílias de LINEs, v. 110, n. 1–4, p. 475–490, 2005.

OKAMOTO, H.; HIROCHIKA, H. Silencing of transposable elements in plants. **Trends in Plant Science**, New York, v. 6, n. 11, p. 527–534, nov. 2001.

OKI, N. et al. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. japonica. **Genes & Genetic Systems**, Mishima, v. 83, n. 4, p. 321–329, 2008.

OLIVEIRA, D. S. et al. Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. **Journal of Evolutionary Biology**, Oxford, v. 34, n. 4, p. 628–638, 28 abr. 2021.

OLIVEIRA, S. G. et al. Horizontal transfers of Mariner transposons between mammals and insects. **Mobile DNA**, London, v. 3, n. 1, p. 14, 2012. Disponível em: <<http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-3-14>>. Acesso em: 30 out. 2021.

PACHULSKA-WIECZOREK, K.; LE GRICE, S.; PURZYCKA, K. Determinants of Genomic RNA Encapsidation in the *Saccharomyces cerevisiae* Long Terminal Repeat Retrotransposons Ty1 and Ty3. **Viruses**, Basel, v. 8, n. 7, p. 193, 14 jul. 2016. Disponível em: <<http://www.mdpi.com/1999-4915/8/7/193>>. Acesso em: 30 out. 2021.

PAPA, F. et al. Rapid evolution of female-biased genes among four species of Anopheles malaria mosquitoes. **Genome Research**, New York, v. 27, n. 9, p. 1536–1548, set. 2017.

PASQUESI, G. I. M. et al. Vertebrate Lineages Exhibit Diverse Patterns of Transposable Element Regulation and Expression across Tissues. **Genome Biology and Evolution**, Oxford, v. 12, n. 5, p. 506–521, 1 maio 2020.

PECCOUD, J. et al. Massive horizontal transfer of transposable elements in insects. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 114, n. 18, p. 4721–4726, 2 maio 2017.

PECINKA, A. et al. Epigenetic Regulation of Repetitive Elements Is Attenuated by Prolonged Heat Stress in Arabidopsis. **The Plant Cell**, Rockville, v. 22, n. 9, p. 3118–3129, 27 out. 2010.

PEDRO, D. L. F. et al. PlaNC-TE: a comprehensive knowledgebase of non-coding RNAs and transposable elements in plants. **Database**, Oxford, p. bay078, 1 jan. 2018. Disponível em: <<https://academic.oup.com/database/article/doi/10.1093/database/bay078/5064210>>. Acesso em: 30 out. 2021.

PEREZ-RIVEROL, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. **Nucleic Acids Research**, Oxford, v. 47, n. D1, p. D442–D450, 8 jan. 2019.

PERKINS, D. N. et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. **Electrophoresis**, Weinheim, v. 20, n. 18, p. 3551–3567, 1 dez. 1999.

PERTEA, G.; PERTEA, M. GFF Utilities: GffRead and GffCompare. **F1000Research**, London, v. 9, p. 304, 9 set. 2020. Disponível em: <<https://f1000research.com/articles/9-304/v2>>. Acesso em: 30 out. 2021.

PETERSEN, M. et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. **BMC Ecology and Evolution**, London, v. 19, n. 1, p. 11, 9 dez. 2019. Disponível em: <<https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-018-1324-9>>. Acesso em: 30 out. 2021.

PETERSON, P. A. The Pale Green Mutable System in Maize. **Genetics**, Austin, v. 45, n. 1, p. 115–33, jan. 1960.

PIRIYAPONGSA, J. et al. Exonization of the LTR transposable elements in human genome. **BMC Genomics**, London, v. 8, n. 1, p. 291, 28 dez. 2007. Disponível em: <<https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-8-291>>. Acesso em: 30 out. 2021.

PLASTERK, R. H. .; IZSVÁK, Z.; IVICS, Z. Resident aliens: the Tc1/ mariner superfamily of transposable elements. **Trends in Genetics**, Amsterdam, v. 15, n. 8, p. 326–332, ago. 1999.

PLATT, R. N.; VANDEWEGE, M. W.; RAY, D. A. Mammalian transposable elements and their impacts on genome evolution. **Chromosome Research**, Oxford, v. 26, n. 1–2, p. 25–43, 2018.

POELCHAU, M. F. et al. Transcriptome sequencing as a platform to elucidate molecular components of the diapause response in the Asian tiger mosquito *Aedes albopictus*. **Physiological Entomology**, Oxford, v. 38, n. 2, p. 173–181, 2013.

POINAR, J. et al. *Paleoculicis minutus* (Diptera: Culicidae) n. gen., n. sp., from Cretaceous Canadian amber, with a summary of described fossil mosquitoes. **Acta Geologica Hispanica**, Barcelona, v. 35, n. 1–2, p. 119–128, 2000.

POMASTOWSKI, P.; BUSZEWSKI, B. Trends in Analytical Chemistry Two-dimensional gel electrophoresis in the light of new developments. **Trends in Analytical Chemistry**, Amsterdam, v. 53, p. 167–177, 2014.

POULTER, R. T.M.; GOODWIN, T. J. D. DIRS-1 and the other tyrosine recombinase retrotransposons. **Cytogenetic and Genome Research, Basel**, v. 110, n. 1–4, p. 575–588, 2005.

POULTER, Russell T M; BUTLER, M. I. Tyrosine Recombinase Retrotransposons and Transposons. **Mob. DNA III**. Washington, DC: American Society of Microbiology, 2015. p. 1271–1291.

POWELL, J. R.; TABACHNICK, W. J. History of domestication and spread of *Aedes aegypti*--a review. **Memórias do Instituto Oswaldo Cruz**, Rio de Janeiro, v. 108, n. October, p. 11–17, 2013.

PRASAD, T. S. K. et al. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. **Genome Research, New York**, v. 27, n. 1, p. 133–144, jan. 2017.

PRICE, A. L.; JONES, N. C.; PEVZNER, P. A. De novo identification of repeat families in large genomes. **Bioinformatics**, Oxford, v. 21, n. 1, p. 351–358, 2005.

PRITHAM, E. J.; PUTLIWALA, T.; FESCHOTTE, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. **Gene**, Amsterdam, v. 390, n. 1–2, p. 3–17, 2007.

PUZAKOV, M. V.; PUZAKOVA, L. V. leidy Is a New Group of DD41D Transposons in *Mnemiopsis leidy* Genome. **Russian Journal of Genetics**, Moscow, v. 55, n. 7, p. 825–834, 2019.

QUESNEVILLE, H.; NOUAUD, D.; ANXOLABÉHÈRE, D. Detection of New Transposable Element Families in *Drosophila melanogaster* and *Anopheles gambiae* Genomes. **Journal of Molecular Evolution**, Berlin, v. 57, n. 1, p. 50–59, 2003.

QUESNEVILLE, H.; NOUAUD, D.; ANXOLABÉHÈRE, D. P elements and MITE relatives in the whole genome sequence of *Anopheles gambiae*. **BMC Genomics**, London, v. 7, p. 214, 18 ago. 2006. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/16919158>>. Acesso em: 12 fev. 2022.

RANWEZ, V. et al. MACSE: Multiple alignment of coding SEquences accounting for

frameshifts and stop codons. **PLoS ONE**, San Francisco, v. 6, n. 9, p. e22594, 2011. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0022594>>. Acesso em: 12 fev. 2022.

RAVINDRAN, S. Barbara McClintock and the discovery of jumping genes. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 109, n. 50, p. 20198–20199, 2012.

RECH, G. E. et al. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. **Nature Communications**, New York, v. 13, n. 1, p. 1948, 12 dez. 2022. Disponível em: <<https://www.nature.com/articles/s41467-022-29518-8>>. Acesso em: 12 fev. 2022.

REID, W. et al. An *Anopheles stephensi* Promoter-Trap: Augmenting Genome Annotation and Functional Genomics. **G3 Genes|Genomes|Genetics**, Bethesda, v. 8, n. 10, p. 3119–3130, 1 out. 2018.

REISS, D. et al. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. **PLoS Genetics**, San Francisco, v. 15, n. 2, p. e1007965, 1 fev. 2019. Disponível em: <<http://dx.plos.org/10.1371/journal.pgen.1007965>>. Acesso em: 12 fev. 2022.

RITCHIE, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Research**, Oxford, v. 43, n. 7, p. e47, 20 abr. 2015. Disponível em: <<http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for>>. Acesso em: 12 fev. 2022.

RIVERA-VEGA, L.; MITTAPALLI, O. Molecular characterization of mariner-like elements in emerald ash borer, *agrilus planipennis* (Coleoptera, Polyphaga). **Archives of Insect Biochemistry and Physiology**, New York, v. 74, n. 4, p. 205–216, 2010.

ROBERTSON, H. M.; LAMPE, D. J. Distribution of transposable elements in arthropods. **Annual review of entomology**, Palo Alto, v. 40, p. 333–357, 1995.

ROMERO-SORIANO, V.; GARCIA GUERREIRO, M. P. Expression of the Retrotransposon Helena Reveals a Complex Pattern of TE Deregulation in *Drosophila* Hybrids. **PLoS ONE**, San Francisco, v. 11, n. 1, p. e0147903, 26 jan. 2016. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0147903>>. Acesso em: 12 fev. 2022.

ROSSATO, D. O. et al. BuT2 is a member of the third major group of hAT transposons and is involved in horizontal transfer events in the genus *Drosophila*. **Genome Biology and Evolution**, Oxford, v. 6, n. 2, p. 352–365, 2014.

ROY, M. et al. Viral infection impacts transposable element transcript amounts in *Drosophila*. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 117, n. 22, p. 12249–12257, 2 jun. 2020.

RUGGLES, K. V. et al. Methods, tools and current perspectives in proteogenomics. **Molecular and Cellular Proteomics**, Bethesda, v. 16, n. 6, p. 959–981, 2017.

RUSSO, J.; HARRINGTON, A. W.; STEINIGER, M. Antisense Transcription of

Retrotransposons in *Drosophila* : An Origin of Endogenous Small Interfering RNA Precursors. **Genetics**, Austin, v. 202, n. 1, p. 107–121, 1 jan. 2016.

RUZZANTE, L.; REIJNDERS, M. J. M. F.; WATERHOUSE, R. M. Of Genes and Genomes: Mosquito Evolution and Diversity. **Trends in Parasitology**, Oxford, v. 35, n. 1, p. 32–51, 2019.

SABOT, F. et al. Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. **Molecular Genetics and Genomics**, Berlin, v. 274, n. 2, p. 119–130, 21 set. 2005.

SAHA, S.; MATTHEWS, D.; BESSANT, C. Proteomics Informed by Transcriptomics. In: BESSANT, C. (Org.). . **Proteome Informatics: New Developments in Mass Spectrometry**. London: Royal Society of Chemistry, 2016. p. 385–405.

SAINT-LEANDRE, B. et al. piRNA and transposon dynamics in *drosophila*: A female story. **Genome Biology and Evolution**, Oxford, v. 12, n. 6, p. 931–947, 2020.

SAITO, K.; SIOMI, M. C. Small RNA-Mediated Quiescence of Transposable Elements in Animals. **Developmental Cell**, Cambridge, v. 19, n. 5, p. 687–697, nov. 2010.

SALCES-ORTIZ, J. et al. Transposable elements contribute to the genomic response to insecticides in *Drosophila melanogaster*. **Philosophical Transactions of the Royal Society B: Biological Sciences**, London, v. 375, n. 1795, p. 20190341, 30 mar. 2020. Disponível em: <<https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0341>>. Acesso em: 12 fev. 2022.

SAMY, A. M. et al. Climate Change Influences on the Global Potential Distribution of the Mosquito *Culex quinquefasciatus*, Vector of West Nile Virus and Lymphatic Filariasis. **PLoS ONE**, San Francisco, v. 11, n. 10, p. e0163863, 3 out. 2016. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0163863>>. Acesso em: 12 fev. 2022.

SCHNABLE, P. S. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. **Science**, New York, v. 326, n. 5956, p. 1112–1115, 20 nov. 2009.

SEAMAN, J. A. et al. Age and prior blood feeding of *Anopheles gambiae* influences their susceptibility and gene expression patterns to ivermectin-containing blood meals. **BMC Genomics**, London, v. 16, n. 1, p. 797, 15 dez. 2015. Disponível em: <<http://dx.doi.org/10.1186/s12864-015-2029-8>>. Acesso em: 12 fev. 2022.

SHANNON, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome Research**, New York, v. 13, n. 11, p. 2498–2504, nov. 2003.

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**, Austin, v. 159, n. 3, p. 1103–1115, 2001.

SILVA, J. C.; LORETO, E. L.; CLARK, J. B. Factors that affect the horizontal transfer of transposable elements. **Current issues in molecular biology**, Basel, v. 6, n. 1, p. 57–71, jan. 2004.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, Oxford, v. 31, n. 19, p. 3210–3212, 1 out. 2015.

SIMÃO, M. C. et al. Helena and BS: two travellers between the genera *Drosophila* and *Zaprionus*. **Genome Biology and Evolution**, Oxford, v. 10, n. 10, p. 2671–2685, 25 ago. 2018.

SINGER, M. F. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. **Cell**, Cambridge, v. 28, n. 3, p. 433–434, mar. 1982.

SINKA, M. E. Global Distribution of the Dominant Vector Species of Malaria. **Anopheles mosquitoes - New insights into Malaria vectors**. [S.l.]: InTech, 2013. v. 1. p. 109–143.

SINKA, M. E.; BANGS, M. J.; et al. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. **Parasites & Vectors**, London, v. 3, n. 1, p. 117, 2010. Disponível em: <[http://link.springer.com/10.1007/978-4-431-54883-6\\_12](http://link.springer.com/10.1007/978-4-431-54883-6_12)>. Acesso em: 12 fev. 2022.

SINKA, M. E.; RUBIO-PALIS, Y.; et al. The dominant *Anopheles* vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. **Parasites & Vectors**, London, v. 3, n. 1, p. 72, 2010. Disponível em: <<http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=51446738>>. Acesso em: 23 fev. 2022.

SINKA, M. E. et al. The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. **Parasites & Vectors**, London, v. 4, n. 1, p. 89, 2011. Disponível em: <<http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=51446738>>. Acesso em: 23 fev. 2022.

SINZELLE, L.; IZSVÁK, Z.; IVICS, Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. **Cellular and Molecular Life Sciences**, Basel, v. 66, n. 6, p. 1073–1093, 12 mar. 2009.

SIOMI, M. C.; MIYOSHI, T.; SIOMI, H. PiRNA-mediated silencing in *Drosophila* germlines. **Seminars in Cell and Developmental Biology**, London, v. 21, n. 7, p. 754–759, 2010.

SLOTKIN, R. K. The case for not masking away repetitive DNA. **Mobile DNA**, London, v. 9, n. 1, p. 15, 1 dez. 2018. Disponível em: <<https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-018-0120-9>>. Acesso em: 23 fev. 2022.

SMIT, A. F. A.; RIGGS, A. D. Tiggers and other DNA transposon fossils in the human genome. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 93, n. 4, p. 1443–1448, 1996.

SMIT, A.; HUBLEY, R.; GREEN, P. **RepeatMasker Open-4.0**. Disponível em: <<http://www.repeatmasker.org>>. Acesso em: 23 fev. 2020.

SOGHIGIAN, J. et al. Genetic evidence for the origin of *Aedes aegypti*, the yellow fever mosquito, in the southwestern Indian Ocean. **Molecular Ecology**, Hoboken, n. July, p. mec.15590, 30 ago. 2020. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15590>>. Acesso em: 23 fev. 2021.

SONG, H. et al. Rapid evolution of piRNA pathway and its transposon targets in Japanese flounder (*Paralichthys olivaceus*). **Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics**, Amsterdam, v. 31, n. April, p. 100609, 2019. Disponível em: <<https://doi.org/10.1016/j.cbd.2019.100609>>. Acesso em: 23 fev. 2021.

STOFFEL, T. J. R. et al. Cyclophosphamide in *Drosophila* promotes genes and transposable elements differential expression and mitochondrial dysfunction. **Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology**, New York, v. 230, n. January, p. 108718, abr. 2020. Disponível em: <<https://doi.org/10.1016/j.cbpc.2020.108718>>. Acesso em: 23 fev. 2021.

STUDER, A. et al. Identification of a functional transposon insertion in the maize domestication gene *tb1*. **Nature Genetics**, New York, v. 43, n. 11, p. 1160–1163, 25 nov. 2011.

SUH, A. et al. Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. **Nature Communications**, New York, v. 7, n. 1, p. 11396, 21 set. 2016. Disponível em: <<http://www.nature.com/articles/pj201637>>. Acesso em: 23 fev. 2021.

SUN, J. et al. De novo assembly of the transcriptome of an invasive snail and its multiple ecological applications. **Molecular Ecology Resources**, Oxford, v. 12, n. 6, p. 1133–1144, 21 nov. 2012.

SUNDARAM, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. **Genome Research**, New York, v. 24, n. 12, p. 1963–1976, 2014.

SZKLARCZYK, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. **Nucleic Acids Research**, Oxford, v. 49, n. D1, p. D605–D612, 8 jan. 2021.

TANG, C. et al. Elevated atmospheric CO<sub>2</sub> promoted speciation in mosquitoes (Diptera, Culicidae). **Communications Biology**, London, v. 1, n. 1, p. 182, 5 dez. 2018. Disponível em: <<http://dx.doi.org/10.1038/s42003-018-0191-7>>. Acesso em: 23 fev. 2021.

TANG, Z. et al. Repeated horizontal transfers of four DNA transposons in invertebrates and bats. **Mobile DNA**, London, v. 6, n. 1, p. 3, 2015. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/25606061>>. Acesso em: 23 fev. 2021.

TARCHINI, R. et al. The Complete Sequence of 340 kb of DNA around the Rice *Adh1–Adh2* Region Reveals Interrupted Colinearity with Maize Chromosome 4. **The Plant Cell**, Rockville, v. 12, n. 3, p. 381–391, mar. 2000.

TELLIER, M.; BOUUAERT, C. C. Mariner and the ITm Superfamily of Transposons. **Mob. DNA III**. Washington, DC: American Society of Microbiology, 2015. p. 753–772.

TENZEN, T.; MATSUTANI, S.; OHTSUBO, E. Site-specific transposition of insertion sequence IS630. **Journal of Bacteriology**, Washington, DC, v. 172, n. 7, p. 3830–3836, 1990.

THE VEUPATHDB PROJECT TEAM. **VectorBase**. Disponível em: <<https://vectorbase.org/vectorbase/app>>. Acesso em: 14 jun. 2022.

THOMAS, J.; PRITHAM, E. J. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. **Microbiology Spectrum**, Washington, DC, v. 3, n. 4, 6 ago. 2015. Disponível em: <<http://www.asmscience.org/content/journal/microbiolspec/10.1128/microbiolspec.MDNA3-0049-2014>>. Acesso em: 14 jun. 2021.

THOMAS, J.; SCHAACK, S.; PRITHAM, E. J. Pervasive Horizontal Transfer of Rolling-Circle Transposons among Animals. **Genome Biology and Evolution**, Oxford, v. 2, n. 1, p. 656–664, 1 jan. 2010.

THOMPSON, P. J.; MACFARLAN, T. S.; LORINCZ, M. C. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. **Molecular Cell**, Cambridge, v. 62, n. 5, p. 766–776, 2016.

THURMOND, J. et al. FlyBase 2.0: the next generation. **Nucleic Acids Research**, Oxford, v. 47, n. D1, p. D759–D765, 8 jan. 2019.

TODD, C. D. et al. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. **eLife**, Cambridge, v. 8, 23 abr. 2019. Disponível em: <<https://elifesciences.org/articles/44344>>. Acesso em: 14 jun. 2021.

TU, Z. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 98, n. 4, p. 1699–1704, 13 fev. 2001.

TUDOR, M. et al. The pogo transposable element family of *Drosophila melanogaster*. **Molecular and General Genetics MGG**, Berlin, v. 232, n. 1, p. 126–134, mar. 1992.

ULITSKY, I.; BARTEL, D. P. lincRNAs: Genomics, Evolution, and Mechanisms. **Cell**, Amsterdam, v. 154, n. 1, p. 26–46, jul. 2013.

ULLU, E.; TSCHUDI, C. Alu sequences are processed 7SL RNA genes. **Nature**, London, v. 312, n. 5990, p. 171–172, 1984.

UYHELJI, H. A.; CHENG, C.; BESANSKY, N. J. Transcriptomic differences between euryhaline and stenohaline malaria vector sibling species in response to salinity stress. **Molecular Ecology**, Hoboken, v. 25, n. 10, p. 2210–2225, maio 2016.

VAN DEN BERG, H.; KELLY-HOPE, L. A.; LINDSAY, S. W. Malaria and lymphatic filariasis: The case for integrated vector management. **The Lancet Infectious Diseases**, New York, v. 13, n. 1, p. 89–94, 2013.



- VAN METER, M. et al. SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. **Nature Communications**, New York, v. 5, n. 1, p. 5011, 23 dez. 2014. Disponível em: <<http://www.nature.com/articles/ncomms6011>>. Acesso em: 14 jun. 2021.
- VANDENBON, A. Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data. **PLoS ONE**, San Francisco, v. 17, n. 1, p. e0263344, 28 jan. 2022. Disponível em: <<http://dx.doi.org/10.1371/journal.pone.0263344>>. Acesso em: 14 jun. 2021.
- VANDEREYKEN, K. et al. Hub Protein Controversy: Taking a Closer Look at Plant Stress Response Hubs. **Frontiers in Plant Science**, Lausanne, v. 9, p. 694, 5 jun. 2018. Disponível em: <<https://www.frontiersin.org/article/10.3389/fpls.2018.00694/full>>. Acesso em: 14 jun. 2021.
- VASSETZKY, N. S.; KRAMEROV, D. A. SINEBase: a database and tool for SINE analysis. **Nucleic Acids Research**, Oxford, v. 41, n. D1, p. D83–D89, 1 jan. 2013.
- VENDRELL-MIR, P. et al. Different Families of Retrotransposons and DNA Transposons Are Actively Transcribed and May Have Transposed Recently in *Physcomitrium* (*Physcomitrella*) patens. **Frontiers in Plant Science**, Lausanne, v. 11, p. 1274, 19 ago. 2020. Disponível em: <<https://www.frontiersin.org/article/10.3389/fpls.2020.01274/full>>. Acesso em: 14 jun. 2021.
- VICIENT, C. M. Transcriptional activity of transposable elements in maize. **BMC Genomics**, London, v. 11, n. 1, p. 601, 2010. Disponível em: <<https://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-601>>. Acesso em: 14 jun. 2021.
- VINSON, J. P. et al. Assembly of polymorphic genomes: Algorithms and application to *Ciona savignyi*. **Genome Research**, New York, v. 15, n. 8, p. 1127–1135, ago. 2005.
- VOLFF, J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. **BioEssays**, Cambridge, v. 28, n. 9, p. 913–922, set. 2006.
- VORONOVA, A. et al. Stress-induced transcriptional activation of retrotransposon-like sequences in the Scots pine (*Pinus sylvestris* L.) genome. **Tree Genetics & Genomes**, Heidelberg, v. 10, n. 4, p. 937–951, 16 ago. 2014.
- VOYTAS, D.F. Pseudoviruses. **Encyclopedia of Virology**. Washington, DC: Elsevier, 2008. p. 352–357.
- VOYTAS, Daniel F.; AUSUBEL, F. M. A copia-like transposable element family in *Arabidopsis thaliana*. **Nature**, London, v. 336, n. 6196, p. 242–244, 1988.
- WALKER, E. L. et al. Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. **Genetics**, Asustin, v. 146, n. 2, p. 681–693, 1997.

WALLAU, G. L. et al. VHICA, a New Method to Discriminate between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within *Drosophila*. **Molecular Biology and Evolution**, New York, v. 33, n. 4, p. 1094–1109, abr. 2016.

WALLAU, G. L.; VIEIRA, C.; LORETO, É. L. S. Genetic exchange in eukaryotes through horizontal transfer: connected by the mobilome. **Mobile DNA**, London, v. 9, n. 1, p. 6, 31 dez. 2018. Disponível em: <<https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-018-0112-9>>. Acesso em: 14 jun. 2021.

WANG, C.; LIN, H. Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. **Genome Biology**, London, v. 22, n. 1, p. 27, 8 dez. 2021. Disponível em: <<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02221-x>>. Acesso em: 14 jun. 2021.

WANG, N. et al. Patterns of Gene Duplication and Their Contribution to Expansion of Gene Families in Grapevine. **Plant Molecular Biology Reporter**, New York, v. 31, n. 4, p. 852–861, 18 ago. 2013.

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

WILKERSON, R. C.; LINTON, Y.-M.; STRICKMAN, D. **Mosquitoes of the World**. Baltimore: Johns Hopkins University Press, 2021.

WITTE, C.-P. et al. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 98, n. 24, p. 13778–13783, 2001.

XIONG, Y.; BURKE, W. D.; EICKBUSH, T. H. Pao, a highly divergent retrotransposable element from *Bombyx mori* containing long terminal repeats with tandem copies of the putative R region. **Nucleic Acids Research**, Oxford, v. 21, n. 9, p. 2117–2123, 1993.

XU, Z.; WANG, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. **Nucleic Acids Research**, Oxford, v. 35, n. Web Server, p. W265–W268, 8 maio 2007.

YANG, G. et al. Transposition of the rice miniature inverted repeat transposable element mPing in *Arabidopsis thaliana*. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 104, n. 26, p. 10962–10967, 2007.

YANG, Goujun et al. Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a Stowaway MITE. **Science**, New York, v. 325, n. 5946, p. 1391–1394, 11 set. 2009.

YANG, Guojun; WONG, A.; ROOKE, R. ATon, abundant novel nonautonomous mobile genetic elements in yellow fever mosquito (*Aedes aegypti*). **BMC Genomics**, London, v. 13, n. 1, p. 283, 2012. Disponível em: <<http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-283>>. Acesso em: 29 jun. 2021.

- YAO, R.-W.; WANG, Y.; CHEN, L.-L. Cellular functions of long noncoding RNAs. **Nature Cell Biology**, London, v. 21, n. 5, p. 542–551, 2 maio 2019.
- YOSHIYAMA, M. et al. Possible horizontal transfer of a transposable element from host to parasitoid. **Molecular Biology and Evolution**, New York, v. 18, n. 10, p. 1952–1958, 2001.
- YU, G. et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. **OMICS: A Journal of Integrative Biology**, Larchmont, v. 16, n. 5, p. 284–287, maio 2012.
- YUAN, Y. W.; WESSLER, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, DC, v. 108, n. 19, p. 7884–7889, 2011.
- YUSA, K. *PiggyBac* Transposon. **Microbiology spectrum**, Washington, DC, v. 3, n. 2, 2014, abr. 2015. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/26104701>>.
- ZAVORTINK, T. J.; POINAR, G. O. Anopheles ( Nyssorhynchus ) dominicanus sp . n . ( Diptera : Culicidae ) from Dominican Amber. **Annals of the Entomological Society of America**, College Park, v. 93, n. 6, p. 1230–1235, 2000.
- ZHANG, H.-H. et al. Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes. **Mobile DNA**, London, v. 9, n. 1, p. 19, 18 dez. 2018. Disponível em: <<https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-018-0125-4>>. Acesso em: 29 jun. 2021.
- ZHANG, H. H. et al. Identification and evolutionary history of the DD41D transposons in insects. **Genes and Genomics**, Seoul, v. 38, n. 2, p. 109–117, 2016.
- ZHANG, R.-G. et al. TESorter: lineage-level classification of transposable elements using conserved protein domains. **bioRxiv**, p. 800177, 1 jan. 2019. Disponível em: <<http://biorxiv.org/content/early/2019/10/10/800177.abstract>>. Acesso em: 29 jun. 2021.
- ZHANG, X. et al. P instability factor: An active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. **Proceedings of the National Academy of Sciences**, Washington, DC, v. 98, n. 22, p. 12572–12577, 23 out. 2001.
- ZHANG, Y.; PARMIGIANI, G.; JOHNSON, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. **NAR Genomics and Bioinformatics**, Oxford, v. 2, n. 3, p. 1–10, 1 set. 2020.
- ZHOU, Y.; CAHAN, S. H. A Novel Family of Terminal-Repeat Retrotransposon in Miniature (TRIM) in the Genome of the Red Harvester Ant, Pogonomyrmex barbatus. **PLoS ONE, San Francisco**, v. 7, n. 12, p. e53401, 28 dez. 2012. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0053401>>. Acesso em: 29 jun. 2021.
- ZILBERMAN, D. et al. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. **Nature Genetics**, New

York, v. 39, n. 1, p. 61–69, 26 jan. 2007.

ZOURÉ, A. A. et al. Proteomic Investigation on *Anopheles gambiae* in Burkina Faso Related to Insecticide Pressures from Different Climatic Regions. **Proteomics**, Weinheim, v. 20, n. 8, p. 1900400, 2 abr. 2020. Disponível em:

<<https://onlinelibrary.wiley.com/doi/10.1002/pmic.201900400>>. Acesso em: 29 jun. 2021.

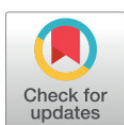
ZYTNICKI, M.; AKHUNOV, E.; QUESNEVILLE, H. Tedna: A transposable element de novo assembler. **Bioinformatics**, Oxford, v. 30, n. 18, p. 2656–2658, 2014.

## RESEARCH ARTICLE

## Mosquito genomes are frequently invaded by transposable elements through horizontal transfer

Elverson Soares de Melo , Gabriel Luz Wallau \*

Department of Entomology, Aggeu Magalhães Institute—Oswaldo Cruz Foundation (Fiocruz), Recife, Pernambuco, Brazil

\* [gabriel.wallau@cpqam.fiocruz.br](mailto:gabriel.wallau@cpqam.fiocruz.br)

## Abstract

Transposable elements (TEs) are mobile genetic elements that parasitize basically all eukaryotic species genomes. Due to their complexity, an in-depth TE characterization is only available for a handful of model organisms. In the present study, we performed a *de novo* and homology-based characterization of TEs in the genomes of 24 mosquito species and investigated their mode of inheritance. More than 40% of the genome of *Aedes aegypti*, *Aedes albopictus*, and *Culex quinquefasciatus* is composed of TEs, while it varied substantially among *Anopheles* species (0.13%–19.55%). Class I TEs are the most abundant among mosquitoes and at least 24 TE superfamilies were found. Interestingly, TEs have been extensively exchanged by horizontal transfer (172 TE families of 16 different superfamilies) among mosquitoes in the last 30 million years. Horizontally transferred TEs represents around 7% of the genome in *Aedes* species and a small fraction in *Anopheles* genomes. Most of these horizontally transferred TEs are from the three ubiquitous LTR superfamilies: Gypsy, Bel-Pao and Copia. Searching more than 32,000 genomes, we also uncovered transfers between mosquitoes and two different Phyla—Cnidaria and Nematoda—and two subphyla—Chelicerata and Crustacea, identifying a vector, the worm *Wuchereria bancrofti*, that enabled the horizontal spread of a Tc1-mariner element among various *Anopheles* species. These data also allowed us to reconstruct the horizontal transfer network of this TE involving more than 40 species. In summary, our results suggest that TEs are frequently exchanged by horizontal transfers among mosquitoes, influencing mosquito's genome size and variability.

 OPEN ACCESS

**Citation:** Melo ES, Wallau GL (2020) Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. *PLoS Genet* 16(11): e1008946. <https://doi.org/10.1371/journal.pgen.1008946>

**Editor:** Harmit S. Malik, Fred Hutchinson Cancer Research Center, UNITED STATES

**Received:** June 17, 2020

**Accepted:** October 19, 2020

**Published:** November 30, 2020

**Copyright:** © 2020 Melo, Wallau. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the manuscript, its [Supporting Information files](#) and in Figshare repository - [10.6084/m9.figshare.11336207](https://doi.org/10.6084/m9.figshare.11336207).

**Funding:** This project was supported by the National Council for Scientific and Technological Development by the productivity research fellowship level 2 for Wallau GL (303902/2019-1) and by Fundação Oswaldo Cruz and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under the project numbers PROEP/IAM (406667/2016-0 and 400742/2019-5).

## Author summary

A large fraction of most eukaryotic genomes is derived from DNA fragments that can multiply by inserting themselves in other regions of the genome, generating genomic instability and variability. These fragments are called Transposable Elements (TEs). Since they are a constituent part of the eukaryote genomes, these DNA pieces are normally inherited vertically along with other genes to the host offspring. Host genomes have an

## APÊNDICE B – ARTIGO PUBLICADO NA REVISTA GENETICS AND MOLECULAR BIOLOGY



*Genetics and Molecular Biology*, 45, 1, e20210215 (2022)  
Copyright © Sociedade Brasileira de Genética.  
DOI: <https://doi.org/10.1590/1678-4685-GMB-2021-0215>

Short Communication  
Genomics and Bioinformatics

### Mosquito long non-coding RNAs are enriched with Transposable Elements

Elverson Soares de Melo<sup>1</sup>  and Gabriel Luz Wallau<sup>1</sup> 

<sup>1</sup>Fundação Oswaldo Cruz, Instituto Aggeu Magalhães, Departamento de Entomologia e Núcleo de Bioinformática, Recife, PE, Brazil.

#### Abstract

Long non-coding RNAs (lncRNAs) lack coding capacity and mounting evidence suggests that they have a regulatory role in diverse organisms. Most knowledge about lncRNAs comes from studies on vertebrates, including a structural association between lncRNAs and transposable elements (TEs). TE sequences are genomic parasites found in all branches of life and are particularly active and abundant in insect genomes. Here we investigate the contribution of TEs to lncRNA biogenesis in *Aedes albopictus* and *Culex quinquefasciatus*. We found that a large fraction of lncRNA loci co-occurs with TE loci in both species. Around 40% of *A. albopictus* and 52% of *C. quinquefasciatus* lncRNAs show some association with TEs. Most of the lncRNA/TE associations are represented by TE-derived sequences that are expressed as one or all exons of lncRNAs, including five lncRNAs that seem to influence immune-related genes involved in antiviral response. The contribution of TEs to lncRNAs also varies among the different types of TEs. The Gypsy superfamily is particularly enriched in lncRNAs sequences. In sum, this study demonstrates that transposable elements substantially contribute to lncRNAs biogenesis in *A. albopictus* and *C. quinquefasciatus* and may have an impact on regulatory modulation in these species.

**Keywords:** lncRNAs, transposable elements, mosquitoes, *Aedes*, *Culex*.

Received: July 27, 2021; Accepted: November 29, 2021.

Mosquitoes of the genera *Aedes*, *Anopheles*, and *Culex* are vectors of several human pathogens. Many studies sought to investigate a range of biological phenomena in these species, such as insecticide resistance and pathogen infection, resulting in thousands of RNA sequencing datasets available. Some recently published articles used dozens of these datasets to characterize long non-coding RNAs (lncRNA) in mosquitoes (Azlan *et al.*, 2021a,b). Many lncRNAs are derived from genes that resemble protein-coding genes (mRNA producers). A mature lncRNA may be processed like a true protein-coding gene mRNA having a similar half-life, but normally they undergo different mRNA processing steps (Ulitsky and Bartel, 2013). However, unlike protein-coding gene mRNAs, lncRNAs are not capable or have a limited capacity of producing fully functional proteins (Mercer *et al.*, 2009). lncRNAs may contain spurious (not translated) or short Open Reading Frames (ORFs), that is, ORFs that might potentially give rise to short polypeptides (<100 amino acids), but these are commonly absent from the species proteome, or if detected, no function can be inferred or characterized (Ulitsky and Bartel, 2013). These small ORFs may have no functional consequences, but in some cases, they may evolve to a more complex polypeptide giving origin to new genes (Ruiz-Orera *et al.*, 2020). Therefore, lncRNAs are a group of RNA molecules larger than 200 nucleotides, expressed from intergenic, intronic or exonic loci, that lack true coding capacity to produce functional proteins (Ulitsky and Bartel, 2013; Cech and Steitz, 2014; Dhanoa *et al.*, 2018).

Different studies have shown that lncRNAs impact the specimen phenotype in several ways, mainly influencing gene expression, but also regulating chromosome architecture, recruitment of chromatin modifiers, and modulating chromosomal interactions (Bonasio and Shiekhattar, 2014; Yao *et al.*, 2019). In addition, lncRNA knockdowns or knockout experiments result in morphological and behavioral changes in different organisms (Mattick, 2018). Although our knowledge about lncRNA has advanced, there are still several unanswered questions. One example is: from which genomic structures/loci does lncRNA emerge? Strikingly, results from vertebrates showed an important role of transposable elements (TEs) copies in the emergence of lncRNAs (Kelley and Rinn, 2012; Kapusta *et al.*, 2013; Hezroni *et al.*, 2015). More recently, studies on plants also show that some intronic lncRNAs overlap with TEs (Pedro *et al.*, 2018). Although there are few studies on invertebrates, the TE contribution to lncRNA biogenesis appears to vary widely among insects, from almost no TE sequences overlapping with lncRNA in hymenopterans, to considerable overlap in *Drosophila melanogaster* (Lopez-Ezquerria *et al.*, 2017). A recent publication of our group characterized in detail the contribution of TEs to the genome of mosquitoes (Melo and Wallau, 2020). Here, we leverage such a rich TE dataset to evaluate the TE contribution to lncRNA loci in *Aedes albopictus* and *Culex quinquefasciatus*. We detected a broad contribution of some specific TE superfamilies in these two species' lncRNA profiles.

In order to investigate the role of TEs in lncRNA biogenesis in *A. albopictus* and *C. quinquefasciatus*, we first updated a previous TE characterization conducted by Melo and Wallau (2020) using the most up-to-date genome assembly of both species (GCF\_006496715.1 and GCF\_015732765.1, respectively). Briefly, we used TEdenovo to generate putative

Send correspondence to Gabriel Luz Wallau. Fundação Oswaldo Cruz, Instituto Aggeu Magalhães, Departamento de Entomologia e Núcleo de Bioinformática, Recife, PE, Brazil. E-mail: gabriel.wallau@fiocruz.br.

## APÊNDICE C – MONTAGENS DE GENOMA DE MOSQUITO USADOS NESTE ESTUDO

Abaixo (Quadro 8) estão descritas as versões dos genomas dos mosquitos que foram analisadas neste estudo. Algumas espécies possuem mais de uma versão estudada. Isso ocorreu quando versões de melhor qualidade foram disponibilizadas durante o longo período do estudo.

**Quadro 8** – Lista das montagens de genoma de mosquito usados neste estudo.

<b>Espécie</b>	<b>Número de acesso do GenBank</b>	<b>Número de contigs/scaffolds</b>	<b>Contigs/Scaffold N50</b>	<b>Fonte/Montagem</b>	<b>Cobertura</b>	<b>Origem</b>
<i>Anopheles gambiae</i>	GCF_000005575.2	8	49,364,325	Vectorbase/AgamP4	-	Linhagem PEST
<i>Anopheles albimanus</i>	GCA_000349125.2	201	37,976,048	Vectorbase/AalbS2	95.0x	Linhagem STECLA
	GCF_013758885.1*	7	89,049,450	NCBI/AalbS3	120.0x	Linhagem STECLA
<i>Anopheles koliensis</i>	GCA_000956275.1	41,925	4,659	NCBI/AKwgs3	39.0x	Papua Nova Guiné: Madang
<i>Anopheles dirus</i>	GCA_000349145.1	1,266	6,906,475	Vectorbase/AdirW1	184.0x	Linhagem WRAIR2
<i>Anopheles atroparvus</i>	GCA_000473505.1	1,371	9,206,694	Vectorbase/AatrE1	98.0x	Linhagem EBRO
<i>Anopheles arabiensis</i>	GCA_000349185.1	1,214	5,604,218	Vectorbase/AaraD1	100.0x	Linhagem Dongola
<i>Anopheles christyi</i>	GCA_000349165.1	30,369	9,057	Vectorbase/AchrA1	35.0x	Quênia
<i>Anopheles coluzzii</i>	GCA_000150765.1	10,521	4,437,438	Vectorbase/AcolM1	8x	Linhagem Mali
	GCA_004136515.2*	206	3,468,756	Vectorbase/AcolN1	48x	Linhagem Ngousso
<i>Anopheles culicifacies</i>	GCA_000473375.1	16,162	22,320	Vectorbase/AculA1	31.0x	Irã

<i>Anopheles darlingi</i>	GCA_000211455.3	2,220	115,072	Vectorbase/AdarC3	20.0x	Brasil: Coari
<i>Anopheles epiroticus</i>	GCA_000349105.1	2,673	366,526	Vectorbase/AepiE1	49.0x	Vietnã
<i>Anopheles farauti</i>	GCA_000473445.2	310	12,895,223	Vectorbase/AfarF2	233.0x	Linhagem FAR1
<i>Anopheles funestus</i>	GCA_000349085.1	1,392	671,960	Vectorbase/AfunF1	135.0x	Linhagem FUM0Z
	GCA_003951495.1*	4	93,833,897	Vectorbase/AfunF3	240x	Linhagem FUM0Z
<i>Anopheles maculatus</i>	GCA_000473185.1	47,797	3,841	Vectorbase/AmacM1	25.0x	Linhagem MACULATUS3
<i>Anopheles melas</i>	GCA_000473525.2	20,229	18,103	Vectorbase/AmelC2	147.0x	Camarões
<i>Anopheles merus</i>	GCA_000473845.2	2,027	1,489,982	Vectorbase/AmerM2	147.0x	Linhagem MAF
<i>Anopheles minimus</i>	GCA_000349025.1	678	10,313,149	Vectorbase/AminM1	211.0x	Linhagem MINIMUS1
<i>Anopheles punctulatus</i>	GCA_000956255.1	20,774	10,256	NCBI/APwgs2	50.0x	Papua Nova Guiné: Madang
<i>Anopheles quadriannulatus</i>	GCA_000349065.1	2,823	1,641,272	Vectorbase/AquaS1	93.0x	Linhagem SANGWE
<i>Anopheles sinensis</i>	GCA_000441895.2	9,592	814,231	Vectorbase/AsinC2	30.0x	China
<i>Anopheles stephensi</i>	GCA_000349045.1	1,110	837,295	Vectorbase/AsteS1	97.0x	Linhagem SDA-500
	GCF_013141755.1*	494	88,747,589	NCBI/UCI_ANSTEP_V1.0	180.0x	Linhagem Indian
<i>Culex quinquefasciatus</i>	GCF_000209185.1	3,172	486,756	Vectorbase/CpipJ2	6.14x	Linhagem Johannesburg
	GCF_015732765.1*	57	201,550,677	NCBI/VIPSU	76.0x	Linhagem Johannesburg
<i>Aedes aegypti</i>	GCA_000004015.3	4,757	1,547,048	Vectorbase/AaegL3	7.6X	Linhagem Liverpool



	GCF_002204515.2*	2,309	409,777,670	Vectorbase/AaegL5	110.0x	Linhagem Liverpool
<i>Aedes albopictus</i>	GCA_001444175.2	154,782	201,017	Vectorbase/AaloF1	229x	Linhagem Foshan
	GCF_006496715.1*	2,197	55,702,539	NCBI/Aalbo_primar y.1	75.0x	Linhagem Foshan

**Fonte:** Elaborado pelo Autor

**Legenda:** \*Versões mais atuais com montagens mais íntegras dos genomas

**APÊNDICE D – LISTA DE CORRIDAS DE RNA-SEQ UTILIZADAS NESTE ESTUDO**

**Quadro 9** – Lista de corridas de RNA-Seq investigadas neste estudo

Espécie	Código da Corrida no SRA	Tipo de amostra	Linhagem
<i>An. dirus</i>	SRR826831	-	WRAIR2
<i>An. dirus</i>	SRR5936564	Corpo total adulto	WRAIR2
<i>An. dirus</i>	SRR5936563	Corpo total adulto	WRAIR2
<i>An. dirus</i>	SRR5936566	Corpo total adulto	WRAIR2
<i>An. dirus</i>	SRR5936565	Corpo total adulto	WRAIR2
<i>An. dirus</i>	SRR5936562	Corpo total adulto	WRAIR2
<i>An. dirus</i>	SRR5936561	Corpo total adulto	WRAIR2
<i>An. dirus</i>	SRR7554466	Glândula Salivar	Anopheles dirus B
<i>An. dirus</i>	SRR7554467	Glândula Salivar	Anopheles dirus B
<i>An. dirus</i>	SRR7554464	Glândula Salivar	Anopheles dirus B
<i>An. dirus</i>	SRR7554465	Glândula Salivar	Anopheles dirus B
<i>An. dirus</i>	SRR908280	Glândula Salivar	-
<i>An. minimus</i>	SRR4104225	Carcaça	MINIMUS1
<i>An. minimus</i>	SRR4104223	Carcaça	MINIMUS1
<i>An. minimus</i>	SRR4104224	Carcaça	MINIMUS1
<i>An. minimus</i>	SRR4104279	Ovário	MINIMUS1
<i>An. minimus</i>	SRR4104278	Ovário	MINIMUS1
<i>An. minimus</i>	SRR4104308	Ovário	MINIMUS1
<i>An. minimus</i>	SRR4104221	Carcaça	MINIMUS1
<i>An. minimus</i>	SRR4104222	Carcaça	MINIMUS1
<i>An. minimus</i>	SRR4104220	Carcaça	MINIMUS1
<i>An. minimus</i>	SRR4104247	Testículos e Glândulas Acessórias	MINIMUS1
<i>An. minimus</i>	SRR4104246	Testículos e Glândulas Acessórias	MINIMUS1
<i>An. minimus</i>	SRR4104277	Testículos e Glândulas Acessórias	MINIMUS1
<i>An. minimus</i>	SRR826833	Corpo total adulto	MINIMUS1
<i>An. funestus</i>	SRR826832	Corpo total adulto	FUMUZ
<i>An. funestus</i>	ERR2240088	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR2240089	Corpo total adulto	Ghana
<i>An. funestus</i>	ERR2240090	Corpo total adulto	Ghana
<i>An. funestus</i>	ERR2240091	Corpo total adulto	Ghana
<i>An. funestus</i>	ERR2240092	Corpo total adulto	Cameroon
<i>An. funestus</i>	ERR2240093	Corpo total adulto	Cameroon
<i>An. funestus</i>	ERR2240094	Corpo total adulto	Cameroon
<i>An. funestus</i>	ERR981204	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR981205	Corpo total adulto	FANG colony

<i>An. funestus</i>	ERR981206	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR981207	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR981208	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR981210	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR981211	Corpo total adulto	FANG colony
<i>An. funestus</i>	ERR981209	Corpo total adulto	FANG colony
<i>An. funestus</i>	SRR10988461	-	-
<i>An. funestus</i>	SRR10988463	-	-
<i>An. funestus</i>	SRR10988464	-	-
<i>An. funestus</i>	SRR10988459	-	-
<i>An. funestus</i>	SRR10988460	-	-
<i>An. funestus</i>	SRR10988462	-	-
<i>An. albimanus</i>	SRR4104309	Carça	STECLA
<i>An. albimanus</i>	SRR4104310	Carça	STECLA
<i>An. albimanus</i>	SRR4104311	Carça	STECLA
<i>An. albimanus</i>	SRR4104312	Carça	STECLA
<i>An. albimanus</i>	SRR4104313	Carça	STECLA
<i>An. albimanus</i>	SRR4104314	Carça	STECLA
<i>An. albimanus</i>	SRR4104316	Testículos e Glândulas Acessórias	STECLA
<i>An. albimanus</i>	SRR4104317	Testículos e Glândulas Acessórias	STECLA
<i>An. albimanus</i>	SRR4104318	Testículos e Glândulas Acessórias	STECLA
<i>An. albimanus</i>	SRR4104319	Ovário	STECLA
<i>An. albimanus</i>	SRR4104340	Ovário	STECLA
<i>An. albimanus</i>	SRR4104341	Ovário	STECLA
<i>An. albimanus</i>	SRR8128646	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128631	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128630	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128633	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128632	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128635	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128638	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128645	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128644	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128641	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128640	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128639	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128643	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128642	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128647	Corpo total adulto	Peru
<i>An. albimanus</i>	SRR8128634	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128637	Corpo total adulto	Guatemala
<i>An. albimanus</i>	SRR8128636	Corpo total adulto	Guatemala

<i>An. albimanus</i>	SRR342223	-	-
<i>An. albimanus</i>	SRR489290	Intestino médio	white-stripe
<i>An. merus</i>	SRR606153	Corpo total adulto	-
<i>An. merus</i>	SRR606154	Corpo total adulto	-
<i>An. merus</i>	SRR606160	Corpo total adulto	-
<i>An. merus</i>	SRR606159	Corpo total adulto	-
<i>An. merus</i>	SRR1587090	Testículos	MAF
<i>An. merus</i>	SRR1587091	Testículos	MAF
<i>An. merus</i>	SRR1587092	Testículos	MAF
<i>An. merus</i>	SRR1587093	Testículos	MAF
<i>An. merus</i>	SRR1587098	Carcaça	MAF
<i>An. merus</i>	SRR1587099	Carcaça	MAF
<i>An. merus</i>	SRR1587100	Carcaça	MAF
<i>An. merus</i>	SRR1587101	Carcaça	MAF
<i>An. merus</i>	SRR2932565	Larva	MAF
<i>An. merus</i>	SRR2932577	Larva	MAF
<i>An. merus</i>	SRR2932589	Larva	MAF
<i>An. merus</i>	SRR2932566	Larva	MAF
<i>An. merus</i>	SRR2932578	Larva	MAF
<i>An. merus</i>	SRR2932590	Larva	MAF
<i>An. merus</i>	SRR2932567	Larva	MAF
<i>An. merus</i>	SRR2932579	Larva	MAF
<i>An. merus</i>	SRR2932591	Larva	MAF
<i>An. merus</i>	SRR2932568	Larva	MAF
<i>An. merus</i>	SRR2932580	Larva	MAF
<i>An. merus</i>	SRR2932592	Larva	MAF
<i>An. merus</i>	SRR2932569	Larva	MAF
<i>An. merus</i>	SRR2932581	Larva	MAF
<i>An. merus</i>	SRR2932593	Larva	MAF
<i>An. merus</i>	SRR2932570	Larva	MAF
<i>An. merus</i>	SRR2932582	Larva	MAF
<i>An. merus</i>	SRR2932594	Larva	MAF
<i>An. merus</i>	SRR7591638	Glândulas Acessórias masculinas	-
<i>An. merus</i>	SRR7591637	Glândulas Acessórias masculinas	-
<i>An. merus</i>	SRR7591636	Glândulas Acessórias masculinas	-
<i>An. merus</i>	SRR7591639	Testículos	-
<i>An. merus</i>	SRR7591641	Testículos	-
<i>An. merus</i>	SRR7591640	Testículos	-
<i>An. stephensi</i>	ERR1095200	Glândula Salivar	-
<i>An. stephensi</i>	ERR1095201	Glândula Salivar	-
<i>An. stephensi</i>	ERR1095202	Glândula Salivar	-
<i>An. stephensi</i>	ERR1095203	Glândula Salivar	-

<i>An. stephensi</i>	SRR606152	-	-
<i>An. stephensi</i>	SRR1447827	Intestino médio	Insectary colony (National Institute of Malaria Research)
<i>An. stephensi</i>	SRR1448245	Túbulo de Malpighi	Insectary colony (National Institute of Malaria Research)
<i>An. stephensi</i>	SRR1462338	Fat body	Insectary colony (National Institute of Malaria Research)
<i>An. stephensi</i>	SRR1448248	Ovário	Insectary colony (National Institute of Malaria Research)
<i>An. stephensi</i>	SRR5381624	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR5381623	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR5381625	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR5381621	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR5381620	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR5381622	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR5381626	Corpo total adulto	Hor strain
<i>An. stephensi</i>	SRR7061580	Embrião de 0-1 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061576	Embrião de 0-1 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061579	Embrião de 2-4 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061574	Embrião de 2-4 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061578	Embrião de 4-8 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061575	Embrião de 4-8 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061577	Embrião de 8-12 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR7061573	Embrião de 8-12 Horas	Indian Wild Type (Walter Reid)
<i>An. stephensi</i>	SRR8476333	Glândula Salivar	Insectary reared <i>Anopheles stephensi</i>
<i>An. stephensi</i>	SRR8476334	Glândula Salivar	India
<i>An. stephensi</i>	SRR8580010	Intestino médio	Insectary reared mosquito
<i>An. stephensi</i>	SRR8580011	Intestino médio	-
<i>An. stephensi</i>	SRR8580009	Intestino médio	-
<i>An. stephensi</i>	SRR8501108	Carcaça	China:Shanghai
<i>An. stephensi</i>	SRR8501107	Carcaça	China:Shanghai
<i>An. stephensi</i>	SRR8501110	Carcaça	China:Shanghai
<i>An. stephensi</i>	SRR8501109	Carcaça	China:Shanghai
<i>An. stephensi</i>	SRR8501112	Carcaça	China:Shanghai

<i>An. stephensi</i>	SRR8501111	Carcaça	China:Shanghai
<i>Cx. quinquefasciatus</i>	SRR364515	L4	-
<i>Cx. quinquefasciatus</i>	SRR364516	L4	-
<i>Cx. quinquefasciatus</i>	SRR991016	Antena	-
<i>Cx. quinquefasciatus</i>	SRR991017	Patras traseiras	-
<i>Cx. quinquefasciatus</i>	SRR1005577	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1005578	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1005579	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1005580	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1005581	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1005582	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1005583	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR1012914	L4	-
<i>Cx. quinquefasciatus</i>	SRR1012916	L4	-
<i>Cx. quinquefasciatus</i>	SRR1012913	L4	-
<i>Cx. quinquefasciatus</i>	SRR1012915	L4	-
<i>Cx. quinquefasciatus</i>	SRR1050976	Corpo total adulto	Vero Beach
<i>Cx. quinquefasciatus</i>	SRR1050983	Corpo total adulto	Gainesville
<i>Cx. quinquefasciatus</i>	SRR1462324	Corpo total adulto	-
<i>Cx. quinquefasciatus</i>	SRR8729867	Região anterior do embrião	-
<i>Cx. quinquefasciatus</i>	SRR8729868	Região anterior do embrião	-
<i>Cx. quinquefasciatus</i>	SRR8729870	Região anterior do embrião	-
<i>Cx. quinquefasciatus</i>	SRR8729863	Região posterior do embrião	-
<i>Cx. quinquefasciatus</i>	SRR8729864	Região posterior do embrião	-
<i>Cx. quinquefasciatus</i>	SRR8729869	Região posterior do embrião	-
<i>Cx. quinquefasciatus</i>	SRR9617920	L4	RIAB59
<i>Cx. quinquefasciatus</i>	SRR9617923	L4	RIAB59
<i>Cx. quinquefasciatus</i>	SRR9617924	L4	RIAB59
<i>Cx. quinquefasciatus</i>	SRR9617921	L4	CqSLab
<i>Cx. quinquefasciatus</i>	SRR9617922	L4	CqSLab
<i>Cx. quinquefasciatus</i>	SRR9617919	L4	CqSLab
<i>An. arabiensis</i>	SRR4104217	Carcaça	Dongola
<i>An. arabiensis</i>	SRR4104216	Carcaça	Dongola
<i>An. arabiensis</i>	SRR4104276	Carcaça	Dongola
<i>An. arabiensis</i>	SRR4104343	Carcaça	Dongola
<i>An. arabiensis</i>	SRR4104315	Carcaça	Dongola
<i>An. arabiensis</i>	SRR4104342	Carcaça	Dongola
<i>An. arabiensis</i>	SRR4104218	Ovário	Dongola
<i>An. arabiensis</i>	SRR4104219	Ovário	Dongola
<i>An. arabiensis</i>	SRR4104347	Ovário	Dongola
<i>An. arabiensis</i>	SRR4104345	Testículos e Glândulas Acessórias	Dongola
<i>An. arabiensis</i>	SRR4104346	Testículos e Glândulas Acessórias	Dongola

<i>An. arabiensis</i>	SRR4104344	Testículos e Glândulas Acessórias	Dongola
<i>An. arabiensis</i>	SRR5422863	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422865	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422875	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422862	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422874	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422864	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422877	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422876	Abdomen	Dongola
<i>An. arabiensis</i>	SRR5422907	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422896	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422894	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422906	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422895	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422909	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422908	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422897	Carcaça	Dongola
<i>An. arabiensis</i>	SRR5422890	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422878	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422892	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422891	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422880	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422893	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422879	Cabeça	Dongola
<i>An. arabiensis</i>	SRR5422881	Cabeça	Dongola
<i>An. arabiensis</i>	SRR7591626	Glândulas Acessórias masculinas	-
<i>An. arabiensis</i>	SRR7591624	Glândulas Acessórias masculinas	-
<i>An. arabiensis</i>	SRR7591625	Glândulas Acessórias masculinas	-
<i>An. arabiensis</i>	SRR7591627	Testículos	-
<i>An. arabiensis</i>	SRR7591629	Testículos	-
<i>An. arabiensis</i>	SRR7591628	Testículos	-
<i>An. arabiensis</i>	SRR7985448	Corpo total adulto	Asendabo
<i>An. arabiensis</i>	SRR7985447	Corpo total adulto	Asendabo
<i>An. arabiensis</i>	SRR7985449	Corpo total adulto	Asendabo
<i>An. arabiensis</i>	SRR7985446	Corpo total adulto	Asendabo
<i>An. arabiensis</i>	SRR7985453	Corpo total adulto	Chewaka
<i>An. arabiensis</i>	SRR7985452	Corpo total adulto	Chewaka
<i>An. arabiensis</i>	SRR7985451	Corpo total adulto	Chewaka
<i>An. arabiensis</i>	SRR7985450	Corpo total adulto	Chewaka
<i>An. arabiensis</i>	SRR7985456	Corpo total adulto	Mozambique
<i>An. arabiensis</i>	SRR7985455	Corpo total adulto	Mozambique

<i>An. arabiensis</i>	SRR7985454	Corpo total adulto	Mozambique
<i>An. arabiensis</i>	SRR7985457	Corpo total adulto	Mozambique
<i>An. arabiensis</i>	SRR7985458	Corpo total adulto	Sekoru
<i>An. arabiensis</i>	SRR7985461	Corpo total adulto	Sekoru
<i>An. arabiensis</i>	SRR7985459	Corpo total adulto	Sekoru
<i>An. arabiensis</i>	SRR7985460	Corpo total adulto	Sekoru
<i>An. arabiensis</i>	SRR7985462	Corpo total adulto	Tolay
<i>An. arabiensis</i>	SRR7985463	Corpo total adulto	Tolay
<i>An. arabiensis</i>	SRR7985464	Corpo total adulto	Tolay
<i>An. arabiensis</i>	SRR7985465	Corpo total adulto	Tolay
<i>An. arabiensis</i>	SRR342217		Dongola
<i>An. coluzzii</i>	SRR5526056	Intestino médio	-
<i>An. coluzzii</i>	SRR5526055	Intestino médio	-
<i>An. coluzzii</i>	SRR5526060	Intestino médio	-
<i>An. coluzzii</i>	SRR5526065	Intestino médio	-
<i>An. coluzzii</i>	SRR5526067	Intestino médio	-
<i>An. coluzzii</i>	SRR5526068	Intestino médio	-
<i>An. coluzzii</i>	SRR5526074	Intestino médio	-
<i>An. coluzzii</i>	SRR5526050	Intestino médio	-
<i>An. coluzzii</i>	SRR5526073	Intestino médio	-
<i>An. coluzzii</i>	SRR5526077	Intestino médio	-
<i>An. coluzzii</i>	SRR5526075	Intestino médio	-
<i>An. coluzzii</i>	SRR5526062	Intestino médio	-
<i>An. coluzzii</i>	SRR5526048	Intestino médio	-
<i>An. coluzzii</i>	SRR5526053	Intestino médio	-
<i>An. coluzzii</i>	SRR5526054	Intestino médio	-
<i>An. coluzzii</i>	SRR5526051	Intestino médio	-
<i>An. coluzzii</i>	SRR5526057	Intestino médio	-
<i>An. coluzzii</i>	SRR5526058	Intestino médio	-
<i>An. coluzzii</i>	SRR5526070	Intestino médio	-
<i>An. coluzzii</i>	SRR5526069	Intestino médio	-
<i>An. coluzzii</i>	SRR5526072	Intestino médio	-
<i>An. coluzzii</i>	SRR5526076	Intestino médio	-
<i>An. coluzzii</i>	SRR5526063	Intestino médio	-
<i>An. coluzzii</i>	SRR5526078	Intestino médio	-
<i>An. coluzzii</i>	SRR5526039	Intestino médio	-
<i>An. coluzzii</i>	SRR5526042	Intestino médio	-
<i>An. coluzzii</i>	SRR5526040	Intestino médio	-
<i>An. coluzzii</i>	SRR5526043	Intestino médio	-
<i>An. coluzzii</i>	SRR5526045	Intestino médio	-
<i>An. coluzzii</i>	SRR5526044	Intestino médio	-
<i>An. coluzzii</i>	SRR2932583	Larva	-
<i>An. coluzzii</i>	SRR2932595	Larva	-
<i>An. coluzzii</i>	SRR2932571	Larva	-
<i>An. coluzzii</i>	SRR2932584	Larva	-



<i>An. coluzzii</i>	SRR2932572	Larva	-
<i>An. coluzzii</i>	SRR2932596	Larva	-
<i>An. coluzzii</i>	SRR2932573	Larva	-
<i>An. coluzzii</i>	SRR2932585	Larva	-
<i>An. coluzzii</i>	SRR2932597	Larva	-
<i>An. coluzzii</i>	SRR2932574	Larva	-
<i>An. coluzzii</i>	SRR2932586	Larva	-
<i>An. coluzzii</i>	SRR2932598	Larva	-
<i>An. coluzzii</i>	SRR2932575	Larva	-
<i>An. coluzzii</i>	SRR2932587	Larva	-
<i>An. coluzzii</i>	SRR2932599	Larva	-
<i>An. coluzzii</i>	SRR2932588	Larva	-
<i>An. coluzzii</i>	SRR2932576	Larva	-
<i>An. coluzzii</i>	SRR2932600	Larva	-
<i>An. coluzzii</i>	SRR2932603	Larva	-
<i>An. coluzzii</i>	SRR2932601	Larva	-
<i>An. coluzzii</i>	SRR2932604	Larva	-
<i>An. coluzzii</i>	SRR2932602	Larva	-
<i>An. coluzzii</i>	SRR7591642	Glândulas Acessórias masculinas	-
<i>An. coluzzii</i>	SRR7591643	Glândulas Acessórias masculinas	-
<i>An. coluzzii</i>	SRR7591644	Glândulas Acessórias masculinas	-
<i>An. coluzzii</i>	SRR7591645	Testículos	-
<i>An. coluzzii</i>	SRR7591646	Testículos	-
<i>An. coluzzii</i>	SRR7591647	Testículos	-
<i>An. coluzzii</i>	ERR3150508	Intestino médio	Ngousso
<i>An. coluzzii</i>	ERR3150513	Intestino médio	Ngousso
<i>An. coluzzii</i>	ERR3150516	Intestino médio	Ngousso
<i>An. coluzzii</i>	SRR1586022	Testículos	Mali-NIH
<i>An. coluzzii</i>	SRR1587087	Testículos	Mali-NIH
<i>An. coluzzii</i>	SRR1587088	Testículos	Mali-NIH
<i>An. coluzzii</i>	SRR1587089	Testículos	Mali-NIH
<i>An. coluzzii</i>	SRR1587094	Carça	Mali-NIH
<i>An. coluzzii</i>	SRR1587095	Carça	Mali-NIH
<i>An. coluzzii</i>	SRR1587096	Carça	Mali-NIH
<i>An. coluzzii</i>	SRR1587097	Carça	Mali-NIH
<i>Ae. albopictus</i>	SRR458465	Embrião	-
<i>Ae. albopictus</i>	SRR458466	Embrião	-
<i>Ae. albopictus</i>	SRR458467	Embrião	-
<i>Ae. albopictus</i>	SRR458462	Embrião	-
<i>Ae. albopictus</i>	SRR458463	Embrião	-
<i>Ae. albopictus</i>	SRR458464	Embrião	-
<i>Ae. albopictus</i>	SRR458471	Embrião	-

<i>Ae. albopictus</i>	SRR458472	Embrião	-
<i>Ae. albopictus</i>	SRR458473	Embrião	-
<i>Ae. albopictus</i>	SRR458468	Embrião	-
<i>Ae. albopictus</i>	SRR458469	Embrião	-
<i>Ae. albopictus</i>	SRR458470	Embrião	-
<i>Ae. albopictus</i>	SRR652072	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652088	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652090	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652091	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652092	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652097	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652098	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652099	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652100	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652101	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652102	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652103	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652116	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652118	Ovos	Manassas
<i>Ae. albopictus</i>	SRR672801	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652119	Ovos	Manassas
<i>Ae. albopictus</i>	SRR652120	Ovos	Manassas
<i>Ae. albopictus</i>	SRR1663754	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663769	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663843	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663911	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663913	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663916	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1664190	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1664192	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663685	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663687	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663689	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663697	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663700	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663703	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663707	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1663709	Corpo total adulto	Manassas
<i>Ae. albopictus</i>	SRR1812886	Embrião	Guangzhou
<i>Ae. albopictus</i>	SRR1812887	Embrião	Guangzhou
<i>Ae. albopictus</i>	SRR1810034	Adult	Guangzhou
<i>Ae. albopictus</i>	SRR1812889	Carcaça	Guangzhou
<i>Ae. albopictus</i>	SRR1812891	L1-L2	Guangzhou
<i>Ae. albopictus</i>	SRR1812892	L3-L4	Guangzhou
<i>Ae. albopictus</i>	SRR1812890	Pupa	Guangzhou

<i>Ae. albopictus</i>	SRR1924197	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205660	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205661	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205662	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205663	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205664	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205665	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205666	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205667	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205668	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205669	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205670	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205671	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205672	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR3205673	Túbulo de Malpighi	MRA-804
<i>Ae. albopictus</i>	SRR2003390	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003289	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003389	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003288	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003350	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003280	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003387	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR2003282	Corpo total adulto	Athens
<i>Ae. albopictus</i>	SRR3208478	Corpo total adulto	Rome
<i>Ae. albopictus</i>	SRR3208451	Corpo total adulto	Rome
<i>Ae. albopictus</i>	SRR3208442	Cabeça	Rome
<i>Ae. albopictus</i>	SRR3208449	Cabeça	Rome
<i>Ae. albopictus</i>	SRR3208251	Antena	Rome
<i>Ae. albopictus</i>	SRR3208288	Antena	Rome
<i>Ae. albopictus</i>	SRR3208295	Palpos maxilares	Rome
<i>Ae. albopictus</i>	SRR3208407	Palpos maxilares	Rome
<i>Ae. albopictus</i>	SRR7006482	Cabeça e tórax	Hammond Island
<i>Ae. albopictus</i>	SRR7006483	Cabeça e tórax	Hammond Island
<i>Ae. albopictus</i>	SRR7006484	Cabeça e tórax	Hammond Island
<i>Ae. albopictus</i>	SRR7006486	Intestino médio	Hammond Island
<i>Ae. albopictus</i>	SRR7006487	Intestino médio	Hammond Island
<i>Ae. albopictus</i>	SRR7006481	Intestino médio	Hammond Island
<i>Ae. albopictus</i>	SRR7006488	Intestino médio	Hammond Island
<i>Ae. albopictus</i>	SRR8284027	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284028	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284035	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284036	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284037	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284038	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284039	Corpo total adulto	Shenzhen

<i>Ae. albopictus</i>	SRR8284040	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284041	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284042	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284043	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284044	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284025	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284026	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284029	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284030	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284031	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284032	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284033	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284034	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284045	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284046	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284047	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR8284048	Corpo total adulto	Shenzhen
<i>Ae. albopictus</i>	SRR7907916	Fat body	RIMINI strain
<i>Ae. albopictus</i>	SRR7907915	Fat body	RIMINI strain
<i>Ae. albopictus</i>	SRR7907924	Fat body	RIMINI strain
<i>Ae. albopictus</i>	SRR7907922	Fat body	RIMINI strain
<i>Ae. albopictus</i>	SRR7907919	Fat body	RIMINI strain
<i>Ae. albopictus</i>	SRR7907920	Fat body	RIMINI strain
<i>Ae. albopictus</i>	SRR7907937	Cabeça	RIMINI strain
<i>Ae. albopictus</i>	SRR7907938	Cabeça	RIMINI strain
<i>Ae. albopictus</i>	SRR7907930	Cabeça	RIMINI strain
<i>Ae. albopictus</i>	SRR7907927	Cabeça	RIMINI strain
<i>Ae. albopictus</i>	SRR7907925	Cabeça	RIMINI strain
<i>Ae. albopictus</i>	SRR7907926	Cabeça	RIMINI strain
<i>Ae. albopictus</i>	SRR7990521	Corpo total adulto	Foshan
<i>Ae. albopictus</i>	SRR7990522	Corpo total adulto	Foshan
<i>Ae. albopictus</i>	SRR7990519	Corpo total adulto	Foshan
<i>Ae. albopictus</i>	SRR7990520	Corpo total adulto	Foshan
<i>Ae. albopictus</i>	SRR7990523	Corpo total adulto	Foshan
<i>Ae. albopictus</i>	SRR7990524	Corpo total adulto	Foshan
<i>Ae. albopictus</i>	SRR8147016	Cabeça	-
<i>Ae. albopictus</i>	SRR8147017	Cabeça	-
<i>Ae. albopictus</i>	SRR8147019	Cabeça	-
<i>Ae. albopictus</i>	SRR8147018	Cabeça	-
<i>Ae. albopictus</i>	SRR8147020	Cabeça	-
<i>Ae. albopictus</i>	SRR8147021	Cabeça	-
<i>Ae. albopictus</i>	SRR8835873	Larva	-
<i>Ae. albopictus</i>	SRR8835875	Larva	-
<i>Ae. albopictus</i>	SRR8835872	Larva	-
<i>Ae. albopictus</i>	SRR8835870	Larva	-

<i>Ae. albopictus</i>	SRR8835871	Larva	-
<i>Ae. albopictus</i>	SRR8835874	Larva	-
<i>Ae. albopictus</i>	SRR1845684	Ovos	-
<i>Ae. albopictus</i>	SRR1845690	Ovos	-
<i>Ae. albopictus</i>	SRR1845688	Larva	-
<i>Ae. albopictus</i>	SRR1845689	Larva	-
<i>Ae. albopictus</i>	SRR1845686	Pupa	-
<i>Ae. albopictus</i>	SRR1845685	Corpo total adulto	-
<i>Ae. albopictus</i>	SRR1845687	Corpo total adulto	-
<i>An. gambiae</i>	ERR2192559	Hemócitos	-
<i>An. gambiae</i>	ERR2192534	Hemócitos	-
<i>An. gambiae</i>	ERR2192540	Hemócitos	-
<i>An. gambiae</i>	ERR2192538	Hemócitos	-
<i>An. gambiae</i>	ERR2192533	Hemócitos	-
<i>An. gambiae</i>	ERR2192544	Hemócitos	-
<i>An. gambiae</i>	ERR2192537	Hemócitos	-
<i>An. gambiae</i>	ERR2192550	Hemócitos	-
<i>An. gambiae</i>	ERR2192535	Hemócitos	-
<i>An. gambiae</i>	ERR2192553	Hemócitos	-
<i>An. gambiae</i>	ERR2192546	Hemócitos	-
<i>An. gambiae</i>	ERR2192555	Hemócitos	-
<i>An. gambiae</i>	ERR2192548	Hemócitos	-
<i>An. gambiae</i>	ERR2192552	Hemócitos	-
<i>An. gambiae</i>	ERR2192542	Hemócitos	-
<i>An. gambiae</i>	ERR2192551	Hemócitos	-
<i>An. gambiae</i>	ERR2192556	Hemócitos	-
<i>An. gambiae</i>	ERR2192557	Hemócitos	-
<i>An. gambiae</i>	ERR2192545	Hemócitos	-
<i>An. gambiae</i>	ERR2192549	Hemócitos	-
<i>An. gambiae</i>	ERR2192541	Hemócitos	-
<i>An. gambiae</i>	ERR2192560	Hemócitos	-
<i>An. gambiae</i>	ERR2192536	Hemócitos	-
<i>An. gambiae</i>	ERR2192547	Hemócitos	-
<i>An. gambiae</i>	ERR2192554	Hemócitos	-
<i>An. gambiae</i>	ERR2192558	Hemócitos	-
<i>An. gambiae</i>	ERR2192543	Hemócitos	-
<i>An. gambiae</i>	ERR2192539	Hemócitos	-
<i>An. gambiae</i>	ERR3661590	Linhagem germinativa masculina - células-tronco germinativas (GSCs) e espermatogônias	G3
<i>An. gambiae</i>	ERR3661591	Linhagem germinativa masculina - células-tronco germinativas (GSCs) e espermatogônias	G3

<i>An. gambiae</i>	ERR3661592	Linhagem germinativa masculina - células-tronco germinativas (GSCs) e espermatogônias	G3
<i>An. gambiae</i>	ERR3661593	Linhagem germinativa masculina - espermatócitos primários meióticos	G3
<i>An. gambiae</i>	ERR3661594	Linhagem germinativa masculina - espermatócitos primários meióticos	G3
<i>An. gambiae</i>	ERR3661595	Linhagem germinativa masculina - espermatócitos primários meióticos	G3
<i>An. gambiae</i>	ERR3661596	Linhagem germinativa masculina - espermatócitos secundários meióticos	G3
<i>An. gambiae</i>	ERR3661597	Linhagem germinativa masculina - espermatócitos secundários meióticos	G3
<i>An. gambiae</i>	ERR3661598	Linhagem germinativa masculina - espermatócitos secundários meióticos	G3
<i>An. gambiae</i>	ERR3661599	Linhagem germinativa masculina - espermátides pós-meióticas	G3
<i>An. gambiae</i>	ERR3661600	Linhagem germinativa masculina - espermátides pós-meióticas	G3
<i>An. gambiae</i>	ERR3661601	Linhagem germinativa masculina - espermátides pós-meióticas	G3
<i>An. gambiae</i>	ERR440788	Larva	G3
<i>An. gambiae</i>	ERR440789	Larva	G3
<i>An. gambiae</i>	ERR440790	Corpo total adulto	G3
<i>An. gambiae</i>	ERR440791	Corpo total adulto	G3
<i>An. gambiae</i>	ERR440792	Larva	G3
<i>An. gambiae</i>	ERR440793	Larva	G3
<i>An. gambiae</i>	ERR476719	Pupa	Pimperena
<i>An. gambiae</i>	ERR476718	Pupa	Pimperena
<i>An. gambiae</i>	ERR476716	Pupa	Pimperena
<i>An. gambiae</i>	ERR476720	Pupa	Pimperena
<i>An. gambiae</i>	ERR476717	Pupa	Pimperena
<i>An. gambiae</i>	ERR476715	Pupa	Pimperena
<i>An. gambiae</i>	ERR489306	Corpo total adulto	Pimperena
<i>An. gambiae</i>	ERR489301	Corpo total adulto	Pimperena
<i>An. gambiae</i>	ERR489303	L1	Pimperena
<i>An. gambiae</i>	ERR489305	L1	Pimperena
<i>An. gambiae</i>	ERR489299	L2	Pimperena

<i>An. gambiae</i>	ERR489300	L2	Pimperena
<i>An. gambiae</i>	ERR489302	L3	Pimperena
<i>An. gambiae</i>	ERR489298	L3	Pimperena
<i>An. gambiae</i>	ERR489304	L4	Pimperena
<i>An. gambiae</i>	ERR489297	L4	Pimperena
<i>An. gambiae</i>	ERR505239	Embrião	Pimperena
<i>An. gambiae</i>	ERR505238	Embrião	Pimperena
<i>An. gambiae</i>	ERR505234	Embrião	Pimperena
<i>An. gambiae</i>	ERR505237	Embrião	Pimperena
<i>An. gambiae</i>	ERR505235	Embrião	Pimperena
<i>An. gambiae</i>	ERR505240	Embrião	Pimperena
<i>An. gambiae</i>	ERR505236	Embrião	Pimperena
<i>An. gambiae</i>	ERR537779	Embrião	Pimperena
<i>An. gambiae</i>	ERR537782	Larva	Pimperena
<i>An. gambiae</i>	ERR537789	Larva	Pimperena
<i>An. gambiae</i>	ERR537787	Larva	Pimperena
<i>An. gambiae</i>	ERR537786	Larva	Pimperena
<i>An. gambiae</i>	ERR537781	Larva	Pimperena
<i>An. gambiae</i>	ERR537790	Larva	Pimperena
<i>An. gambiae</i>	ERR537780	Larva	Pimperena
<i>An. gambiae</i>	ERR537778	Larva	Pimperena
<i>An. gambiae</i>	ERR537783	Corpo total adulto	Pimperena
<i>An. gambiae</i>	ERR537785	Corpo total adulto	Pimperena
<i>An. gambiae</i>	ERR537788	Corpo total adulto	Pimperena
<i>An. gambiae</i>	ERR537784	Corpo total adulto	Pimperena
<i>An. gambiae</i>	SRR520427	Corpo total adulto	Emutete
<i>An. gambiae</i>	SRR520428	Corpo total adulto	Emutete
<i>An. gambiae</i>	SRR1727555	Antena	SUA2La
<i>An. gambiae</i>	SRR1727718	Antena	SUA2La
<i>An. gambiae</i>	SRR1727719	Antena	SUA2La
<i>An. gambiae</i>	SRR1727720	Antena	SUA2La
<i>An. gambiae</i>	SRR1727721	Antena	SUA2La
<i>An. gambiae</i>	SRR1727722	Antena	SUA2La
<i>An. gambiae</i>	SRR1727723	Antena	SUA2La
<i>An. gambiae</i>	SRR1727724	Antena	SUA2La
<i>An. gambiae</i>	SRR1727725	Antena	SUA2La
<i>An. gambiae</i>	SRR1727726	Antena	SUA2La
<i>An. gambiae</i>	SRR1179010	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179009	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179012	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179011	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179008	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179005	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179006	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1179007	Corpo total adulto	G3

<i>An. gambiae</i>	SRR1525402	Intestino médio	Kisumu
<i>An. gambiae</i>	SRR1525403	Glândula Salivar	Kisumu
<i>An. gambiae</i>	SRR1596491	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596489	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596487	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596485	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596483	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596481	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596479	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596477	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596475	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596473	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596471	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596470	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596469	Corpo total adulto	G3
<i>An. gambiae</i>	SRR1596472	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596474	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596476	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596478	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596480	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596482	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596484	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596486	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596488	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596490	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596492	Corpo total adulto	Transgenic strain I-PpoI-11A
<i>An. gambiae</i>	SRR1596468	L1	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596467	L1	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596466	L1	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596465	L1	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596464	L1	X-chromosome linked 3xP3



<i>An. gambiae</i>	SRR1596463	L1	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596462	L2-L3	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596461	L2-L3	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596460	L2-L3	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596459	L2-L3	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596458	L3-L4	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596457	L3-L4	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596456	L3-L4	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596455	L3-L4	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596454	L3-L4	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1596453	L3-L4	X-chromosome linked 3xP3
<i>An. gambiae</i>	SRR1805322	Carcaça	G3
<i>An. gambiae</i>	SRR1805324	Carcaça	G3
<i>An. gambiae</i>	SRR1805343	Carcaça	G3
<i>An. gambiae</i>	SRR1805328	Carcaça	G3
<i>An. gambiae</i>	SRR1805330	Carcaça	G3
<i>An. gambiae</i>	SRR1805349	Carcaça	G3
<i>An. gambiae</i>	SRR1805345	Testículos e Glândulas Acessórias	G3
<i>An. gambiae</i>	SRR1805347	Testículos e Glândulas Acessórias	G3
<i>An. gambiae</i>	SRR1805326	Testículos e Glândulas Acessórias	G3
<i>An. gambiae</i>	SRR1805341	Tecidos reprodutivos femininos	G3
<i>An. gambiae</i>	SRR1805332	Tecidos reprodutivos femininos	G3
<i>An. gambiae</i>	SRR1805334	Tecidos reprodutivos femininos	G3
<i>An. gambiae</i>	SRR1763906	Larva	Emutete
<i>An. gambiae</i>	SRR1763908	Larva	Bungoma
<i>An. gambiae</i>	SRR1763907	Larva	Busia
<i>An. gambiae</i>	SRR1756036	Larva	Kisumu
<i>An. gambiae</i>	SRR2094645	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094649	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094651	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094652	Corpo total adulto	G3

<i>An. gambiae</i>	SRR2094641	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094643	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094655	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094656	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094658	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094659	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094653	Corpo total adulto	G3
<i>An. gambiae</i>	SRR2094654	Corpo total adulto	G3
<i>An. gambiae</i>	SRR6127580	Testículos	Suakoko
<i>An. gambiae</i>	SRR6127581	Testículos	Suakoko
<i>An. gambiae</i>	SRR5219382	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219380	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219362	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219360	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219358	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219356	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219354	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219352	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219350	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219348	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219346	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219344	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219342	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219340	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219364	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219386	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219368	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219366	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219372	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219370	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219374	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219378	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219376	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR5219384	Corpo total adulto	Keele
<i>An. gambiae</i>	SRR8729858	Região anterior do embrião	G3
<i>An. gambiae</i>	SRR8729860	Região anterior do embrião	G3
<i>An. gambiae</i>	SRR8729859	Região anterior do embrião	G3
<i>An. gambiae</i>	SRR8729855	Região posterior do embrião	G3
<i>An. gambiae</i>	SRR8729856	Região posterior do embrião	G3
<i>An. gambiae</i>	SRR8729857	Região posterior do embrião	G3
<i>An. gambiae</i>	SRR7591630	Glândulas Acessórias masculinas	MRA-762 (KISUMU1)
<i>An. gambiae</i>	SRR7591631	Glândulas Acessórias masculinas	MRA-762 (KISUMU1)

<i>An. gambiae</i>	SRR7591632	Glândulas Acessórias masculinas	MRA-762 (KISUMU1)
<i>An. gambiae</i>	SRR7591633	Testículos	MRA-762 (KISUMU1)
<i>An. gambiae</i>	SRR7591634	Testículos	MRA-762 (KISUMU1)
<i>An. gambiae</i>	SRR7591635	Testículos	MRA-762 (KISUMU1)
<i>An. gambiae</i>	SRR7866327	Intestino médio e glândula salivar	-
<i>An. gambiae</i>	SRR7866328	Intestino médio e glândula salivar	-
<i>An. gambiae</i>	SRR7866329	Intestino médio e glândula salivar	-
<i>An. gambiae</i>	SRR7866330	Intestino médio e glândula salivar	-
<i>Ae. aegypti</i>	ERR1717910	Corpo total adulto	Bora Bora
<i>Ae. aegypti</i>	ERR1717911	Corpo total adulto	Bora Bora
<i>Ae. aegypti</i>	ERR1717912	Corpo total adulto	Bora Bora
<i>Ae. aegypti</i>	ERR1717913	Corpo total adulto	Liverpool
<i>Ae. aegypti</i>	ERR1717914	Corpo total adulto	Liverpool
<i>Ae. aegypti</i>	ERR1717915	Corpo total adulto	Liverspool
<i>Ae. aegypti</i>	ERR1717916	Corpo total adulto	New Orleans
<i>Ae. aegypti</i>	ERR1717917	Corpo total adulto	New Orleans
<i>Ae. aegypti</i>	ERR1717918	Corpo total adulto	New Orleans
<i>Ae. aegypti</i>	ERR1717919	Corpo total adulto	DeltaR
<i>Ae. aegypti</i>	ERR1717920	Corpo total adulto	DeltaR
<i>Ae. aegypti</i>	ERR1717921	Corpo total adulto	DeltaR
<i>Ae. aegypti</i>	ERR1717922	Corpo total adulto	Cayenne
<i>Ae. aegypti</i>	ERR1717923	Corpo total adulto	Cayenne
<i>Ae. aegypti</i>	ERR1717924	Corpo total adulto	Cayenne
<i>Ae. aegypti</i>	ERR1717925	Corpo total adulto	Saint-Georges
<i>Ae. aegypti</i>	ERR1717926	Corpo total adulto	Saint-Georges
<i>Ae. aegypti</i>	ERR1717927	Corpo total adulto	Saint-Georges
<i>Ae. aegypti</i>	ERR1717928	Corpo total adulto	Nakhonsawan
<i>Ae. aegypti</i>	ERR1717929	Corpo total adulto	Nakhonsawan
<i>Ae. aegypti</i>	ERR1717930	Corpo total adulto	Nakhonsawan
<i>Ae. aegypti</i>	ERR1717931	Corpo total adulto	Phetchaburi
<i>Ae. aegypti</i>	ERR1717932	Corpo total adulto	Phetchaburi
<i>Ae. aegypti</i>	ERR1717933	Corpo total adulto	Phetchaburi
<i>Ae. aegypti</i>	SRR789756	Corpo total adulto	CTM
<i>Ae. aegypti</i>	SRR789757	Corpo total adulto	RexD
<i>Ae. aegypti</i>	SRR789758	Corpo total adulto	LVP
<i>Ae. aegypti</i>	SRR924022	Carcaça	Liverpool Strain
<i>Ae. aegypti</i>	SRR924023	Carcaça	Liverpool Strain
<i>Ae. aegypti</i>	SRR924024	Ovário	Liverpool Strain

<i>Ae. aegypti</i>	SRR924021	Testículos	Liverpool Strain
<i>Ae. aegypti</i>	SRR1060807	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060808	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060996	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060823	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060824	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061012	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060839	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060848	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061020	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061044	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060855	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060856	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060869	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060867	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061049	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060882	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060880	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061062	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060895	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060893	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061075	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061073	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060908	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060906	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061088	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061086	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061161	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061162	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061164	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061170	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061168	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061169	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061184	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061183	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061182	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061196	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061197	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061198	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061210	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061211	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061212	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061224	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061225	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061226	Tórax	black-eyed Liverpool

<i>Ae. aegypti</i>	SRR1061238	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061240	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061239	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061252	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061254	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061253	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061266	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061268	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1061267	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR1060932	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060923	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061112	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060940	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060939	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061128	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060956	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060955	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061144	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060971	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1060972	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061152	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061278	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061277	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061280	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061286	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061285	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061284	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061299	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061300	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061298	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061314	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061313	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061312	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061328	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061327	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1061326	Tórax	Aedes aegypti RED
<i>Ae. aegypti</i>	SRR1167536	abdominal tips	Liverpool
<i>Ae. aegypti</i>	SRR1167534	abdominal tips	Liverpool
<i>Ae. aegypti</i>	SRR1167535	abdominal tips	Liverpool
<i>Ae. aegypti</i>	SRR1167564	abdominal tips	Liverpool
<i>Ae. aegypti</i>	SRR1167566	abdominal tips	Liverpool
<i>Ae. aegypti</i>	SRR1167568	abdominal tips	Liverpool
<i>Ae. aegypti</i>	SRR1167494	Antena	Liverpool
<i>Ae. aegypti</i>	SRR1167497	Antena	Liverpool
<i>Ae. aegypti</i>	SRR1167498	Antena	Liverpool

<i>Ae. aegypti</i>	SRR1167550	Antena	Liverpool
<i>Ae. aegypti</i>	SRR1167553	Antena	Liverpool
<i>Ae. aegypti</i>	SRR1167549	Antena	Liverpool
<i>Ae. aegypti</i>	SRR1167502	Cérebro	Liverpool
<i>Ae. aegypti</i>	SRR1167503	Cérebro	Liverpool
<i>Ae. aegypti</i>	SRR1167504	Cérebro	Liverpool
<i>Ae. aegypti</i>	SRR1167543	Cérebro	Liverpool
<i>Ae. aegypti</i>	SRR1167545	Cérebro	Liverpool
<i>Ae. aegypti</i>	SRR1167547	Cérebro	Liverpool
<i>Ae. aegypti</i>	SRR1167531	Patás dianteiras	Liverpool
<i>Ae. aegypti</i>	SRR1167532	Patás dianteiras	Liverpool
<i>Ae. aegypti</i>	SRR1167533	Patás dianteiras	Liverpool
<i>Ae. aegypti</i>	SRR1167554	Patás dianteiras	Liverpool
<i>Ae. aegypti</i>	SRR1167555	Patás dianteiras	Liverpool
<i>Ae. aegypti</i>	SRR1167556	Patás dianteiras	Liverpool
<i>Ae. aegypti</i>	SRR1167521	Patás traseiras	Liverpool
<i>Ae. aegypti</i>	SRR1167522	Patás traseiras	Liverpool
<i>Ae. aegypti</i>	SRR1167524	Patás traseiras	Liverpool
<i>Ae. aegypti</i>	SRR1167557	Patás traseiras	Liverpool
<i>Ae. aegypti</i>	SRR1167558	Patás traseiras	Liverpool
<i>Ae. aegypti</i>	SRR1167559	Patás traseiras	Liverpool
<i>Ae. aegypti</i>	SRR1167506	Palpos maxilares	Liverpool
<i>Ae. aegypti</i>	SRR1167508	Palpos maxilares	Liverpool
<i>Ae. aegypti</i>	SRR1167509	Palpos maxilares	Liverpool
<i>Ae. aegypti</i>	SRR1167526	Patás do meio	Liverpool
<i>Ae. aegypti</i>	SRR1167528	Patás do meio	Liverpool
<i>Ae. aegypti</i>	SRR1167529	Patás do meio	Liverpool
<i>Ae. aegypti</i>	SRR1167560	Patás do meio	Liverpool
<i>Ae. aegypti</i>	SRR1167561	Patás do meio	Liverpool
<i>Ae. aegypti</i>	SRR1167562	Patás do meio	Liverpool
<i>Ae. aegypti</i>	SRR1167538	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR1167540	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR1167541	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR1167518	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR1167516	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR1167520	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR1167510	probóscide	Liverpool
<i>Ae. aegypti</i>	SRR1167513	probóscide	Liverpool
<i>Ae. aegypti</i>	SRR1167514	probóscide	Liverpool
<i>Ae. aegypti</i>	SRR1578250	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1578251	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1578252	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1578254	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1578255	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1578256	Embrião	Liverpool

<i>Ae. aegypti</i>	SRR1568262	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1568264	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1568263	Embrião	Liverpool
<i>Ae. aegypti</i>	SRR1585314	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR1585315	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR1585316	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR1585317	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR1585318	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR1585319	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR2659965	Glândula Salivar	Liverpool
<i>Ae. aegypti</i>	SRR2659966	Glândula Salivar	Liverpool
<i>Ae. aegypti</i>	SRR2891750	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891754	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891734	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891739	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891731	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891736	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891737	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891742	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891733	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR2891738	Intestino da L4	Cuernavaca
<i>Ae. aegypti</i>	SRR5135764	L1	UGAL
<i>Ae. aegypti</i>	SRR5135761	L2	UGAL
<i>Ae. aegypti</i>	SRR5135765	L3	UGAL
<i>Ae. aegypti</i>	SRR5135760	Intestino da L1	UGAL
<i>Ae. aegypti</i>	SRR5135762	Intestino da L2	UGAL
<i>Ae. aegypti</i>	SRR5135763	Intestino da L3	UGAL
<i>Ae. aegypti</i>	SRR5288089	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288088	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288092	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288096	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288097	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288085	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288084	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288099	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288086	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288098	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288074	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288076	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288100	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288082	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288093	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288077	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288080	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288087	Intestino médio	Higgs White Eye

<i>Ae. aegypti</i>	SRR5288091	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288094	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288075	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288079	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288078	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288095	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288081	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288083	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5288090	Intestino médio	Higgs White Eye
<i>Ae. aegypti</i>	SRR5759451	Carcaça	Georgia
<i>Ae. aegypti</i>	SRR5759455	Carcaça	Georgia
<i>Ae. aegypti</i>	SRR5759452	Carcaça	Georgia
<i>Ae. aegypti</i>	SRR5759456	Carcaça	Georgia
<i>Ae. aegypti</i>	SRR5759453	Carcaça	Georgia
<i>Ae. aegypti</i>	SRR5759454	Carcaça	Georgia
<i>Ae. aegypti</i>	SRR5955023	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955022	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955024	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955026	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955025	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955027	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955010	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955011	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955012	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955013	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955015	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955014	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955017	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955016	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955018	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955019	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955021	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR5955020	Corpo total adulto	Galveston
<i>Ae. aegypti</i>	SRR6311360	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR6311357	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR6311358	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR6311355	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR6311356	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR6311353	Carcaça	Liverpool
<i>Ae. aegypti</i>	SRR6311351	Pupa jovem	Liverpool
<i>Ae. aegypti</i>	SRR6311352	Pupa jovem	Liverpool
<i>Ae. aegypti</i>	SRR6311316	Pupa jovem	Liverpool
<i>Ae. aegypti</i>	SRR6311315	Pupa jovem	Liverpool
<i>Ae. aegypti</i>	SRR6311313	Pupa jovem	Liverpool
<i>Ae. aegypti</i>	SRR6311314	Pupa jovem	Liverpool



<i>Ae. aegypti</i>	SRR6311380	L1	Liverpool
<i>Ae. aegypti</i>	SRR6311312	L1	Liverpool
<i>Ae. aegypti</i>	SRR6311374	L1	Liverpool
<i>Ae. aegypti</i>	SRR6311382	L2	Liverpool
<i>Ae. aegypti</i>	SRR6311371	L2	Liverpool
<i>Ae. aegypti</i>	SRR6311392	L2	Liverpool
<i>Ae. aegypti</i>	SRR6311348	L3	Liverpool
<i>Ae. aegypti</i>	SRR6311349	L3	Liverpool
<i>Ae. aegypti</i>	SRR6311341	L3	Liverpool
<i>Ae. aegypti</i>	SRR6311346	L3	Liverpool
<i>Ae. aegypti</i>	SRR6311350	L3	Liverpool
<i>Ae. aegypti</i>	SRR6311342	L3	Liverpool
<i>Ae. aegypti</i>	SRR6311363	L4	Liverpool
<i>Ae. aegypti</i>	SRR6311364	L4	Liverpool
<i>Ae. aegypti</i>	SRR6311369	L4	Liverpool
<i>Ae. aegypti</i>	SRR6311366	L4	Liverpool
<i>Ae. aegypti</i>	SRR6311367	L4	Liverpool
<i>Ae. aegypti</i>	SRR6311368	L4	Liverpool
<i>Ae. aegypti</i>	SRR6311361	Pupa tardia	Liverpool
<i>Ae. aegypti</i>	SRR6311362	Pupa tardia	Liverpool
<i>Ae. aegypti</i>	SRR6311324	Pupa tardia	Liverpool
<i>Ae. aegypti</i>	SRR6311322	Pupa tardia	Liverpool
<i>Ae. aegypti</i>	SRR6311321	Pupa tardia	Liverpool
<i>Ae. aegypti</i>	SRR6311328	Pupa tardia	Liverpool
<i>Ae. aegypti</i>	SRR6311326	Glândulas Acessórias masculinas	Liverpool
<i>Ae. aegypti</i>	SRR6311325	Glândulas Acessórias masculinas	Liverpool
<i>Ae. aegypti</i>	SRR6311332	Glândulas Acessórias masculinas	Liverpool
<i>Ae. aegypti</i>	SRR6311340	Pupa intermediária	Liverpool
<i>Ae. aegypti</i>	SRR6311337	Pupa intermediária	Liverpool
<i>Ae. aegypti</i>	SRR6311338	Pupa intermediária	Liverpool
<i>Ae. aegypti</i>	SRR6311339	Pupa intermediária	Liverpool
<i>Ae. aegypti</i>	SRR6311333	Pupa intermediária	Liverpool
<i>Ae. aegypti</i>	SRR6311336	Pupa intermediária	Liverpool
<i>Ae. aegypti</i>	SRR6311330	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR6311331	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR6311393	Ovário	Liverpool
<i>Ae. aegypti</i>	SRR6311396	Testículos	Liverpool
<i>Ae. aegypti</i>	SRR6311395	Testículos	Liverpool
<i>Ae. aegypti</i>	SRR6311391	Testículos	Liverpool
<i>Ae. aegypti</i>	SRR6473142	Corpo total adulto	VietNam
<i>Ae. aegypti</i>	SRR8729853	Região anterior do embrião	Liverpool "black eye"
<i>Ae. aegypti</i>	SRR8729854	Região anterior do embrião	Liverpool "black eye"

<i>Ae. aegypti</i>	SRR8729851	Região posterior do embrião	Liverpool "black eye"
<i>Ae. aegypti</i>	SRR8729852	Região posterior do embrião	Liverpool "black eye"
<i>Ae. aegypti</i>	SRR7356168	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356169	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356170	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356171	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356172	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356173	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356174	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356175	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356176	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356177	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356178	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356179	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356180	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356181	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356182	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356183	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356184	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356185	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356186	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356187	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356188	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356189	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356190	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7356191	Corpo total adulto	Trinidad
<i>Ae. aegypti</i>	SRR7938863	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938853	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938852	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938856	Embrião	jane
<i>Ae. aegypti</i>	SRR7938866	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938855	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938858	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938857	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938860	Embrião	Jane
<i>Ae. aegypti</i>	SRR7938859	Embrião	jane
<i>Ae. aegypti</i>	SRR7727875	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727876	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727877	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727878	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727879	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727880	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727887	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727888	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727889	Corpo total adulto	Key West

<i>Ae. aegypti</i>	SRR7727890	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727891	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727892	Corpo total adulto	Key West
<i>Ae. aegypti</i>	SRR7727872	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727873	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727874	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727881	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727882	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727883	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727884	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727885	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727886	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727893	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727894	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR7727895	Corpo total adulto	Orlando
<i>Ae. aegypti</i>	SRR8234180	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR8234181	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR8234182	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR8234183	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR8234184	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR8234185	Tórax	black-eyed Liverpool
<i>Ae. aegypti</i>	SRR8274727	Espermateca	Rio de Janeiro
<i>Ae. aegypti</i>	SRR8274728	Espermateca	Rio de Janeiro
<i>Ae. aegypti</i>	SRR8921132	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921131	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921126	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921125	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921124	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921123	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921130	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921129	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921128	Glândula Salivar	Singapore
<i>Ae. aegypti</i>	SRR8921127	Glândula Salivar	Singapore

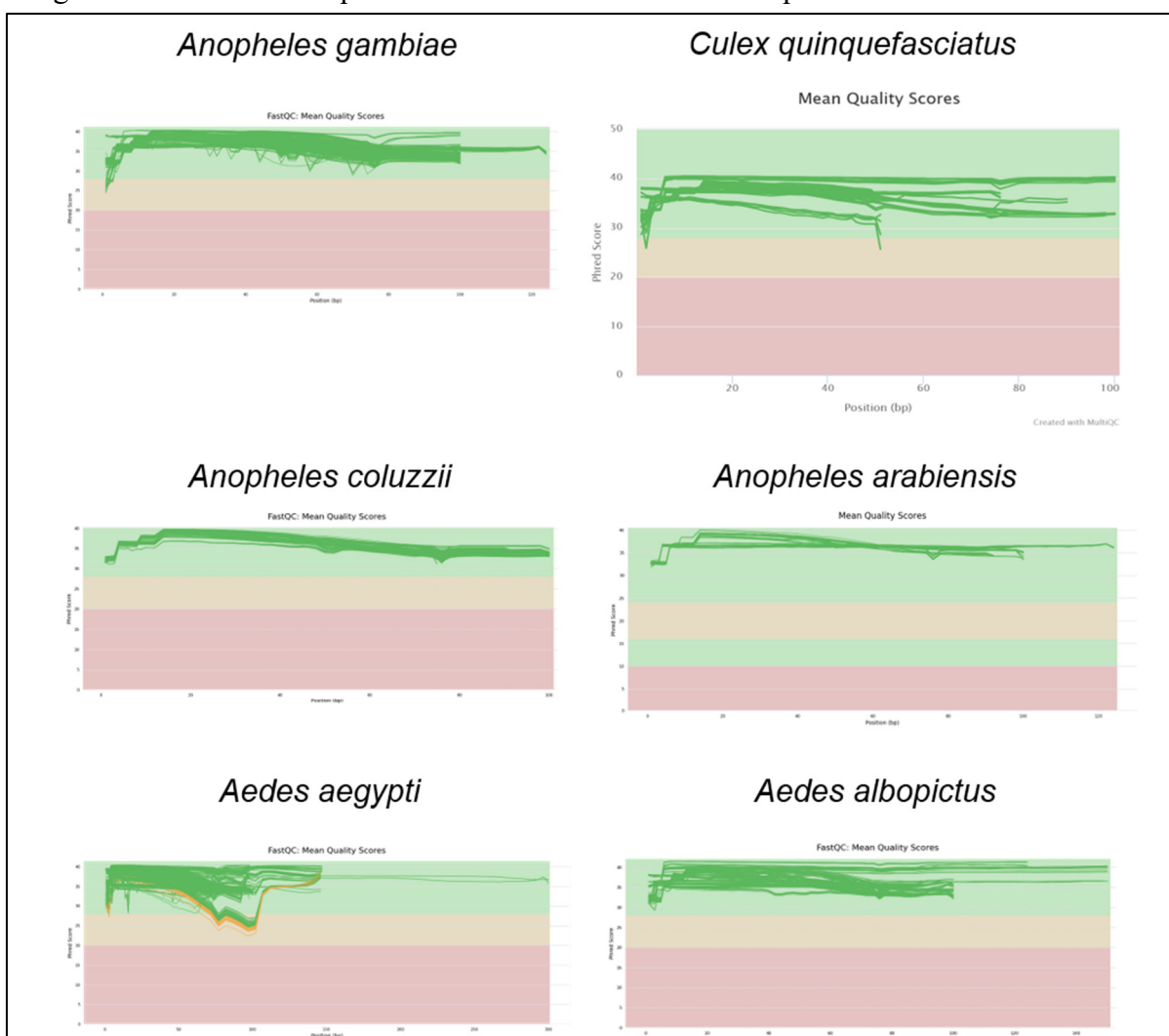
**Fonte:** Elaborado pelo autor

**Nota:** Uma série de outras características como sexo, infecção, idade entre outros fatores foram ocultados dessa tabela por questões de espaço e poluição visual. Contudo essas informações podem ser encontradas no endereço: [www.ncbi.nlm.nih.gov/sra](http://www.ncbi.nlm.nih.gov/sra)

## APÊNDICE E – GARANTIA DA QUALIDADE DAS CORRIDAS DE RNA-SEQ UTILIZADAS

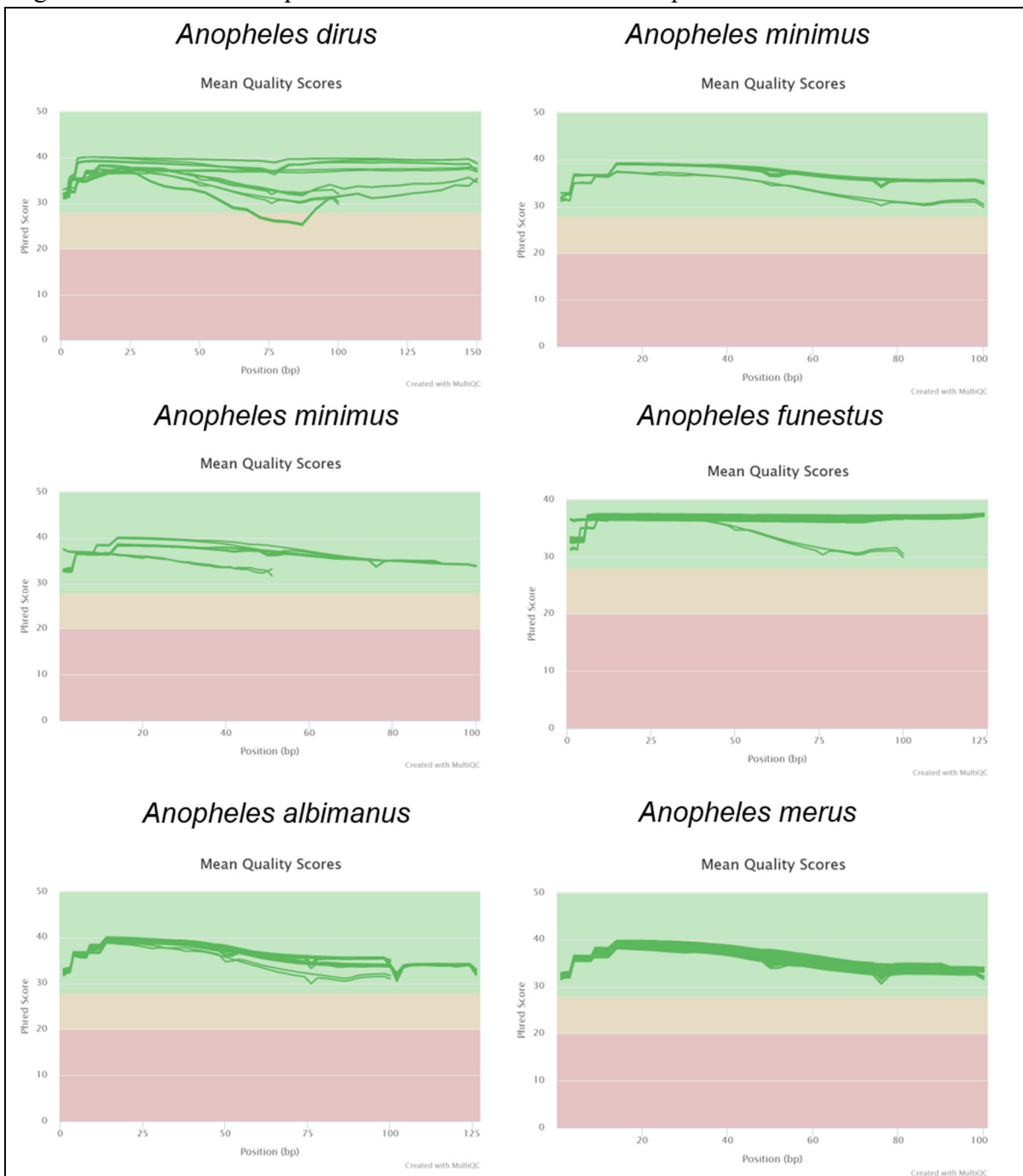
Neste trabalho foram utilizados dados de terceiros que estão publicamente acessíveis nos bancos de dados de sequências genômicas. Esses dados possuíam alta heterogeneidade o que impactava diretamente a qualidade de muitas das corridas de RNA-Seq. A baixa qualidade impactaria toda a etapa da análise de expressão desse estudo, logo uma cuidadosa avaliação de qualidade seguida de eliminação das sequências de baixo score foi realizada. A comprovação da qualidade do material utilizado está mostrada nas figuras abaixo, que representam gráficos de Phred score gerados pelo programa MultiQC.

Figura 56 – Gráficos de qualidade das amostras de RNA-Seq analisadas



Fonte: Elaborado pelo Autor

Figura 57 – Gráficos de qualidade das amostras de RNA-Seq analisadas



Fonte: Elaborado pelo autor.