UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

GISELE STRIEDER PHILIPPSSEN

Estudo da influência de elementos transponíveis nos
genomas das algas *C. reinhardtii* e *V. carteri*

São Carlos
2014

GISELE STRIEDER PHILIPPSEN

Estudo da influência de elementos transponíveis nos
genomas das algas *C. reinhardtii* e *V. carteri*

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Física Aplicada
Opção: Física Computacional
Orientador: Prof. Dr. Ricardo De Marco

Versão corrigida
(Versão original disponível na Unidade que aloja o Programa)

São Carlos
2014

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação do IFSC,
com os dados fornecidos pelo(a) autor(a)

Philippsen, Gisele Strieder

Estudo da influência de elementos transponíveis
nos genomas das algas *C. reinhardtii* e *V. carteri* /
Gisele Strieder Philippsen; orientador Ricardo De
Marco - versão corrigida -- São Carlos, 2014.
151 p.

Tese (Doutorado - Programa de Pós-Graduação em
Física Aplicada Computacional) -- Instituto de Física
de São Carlos, Universidade de São Paulo, 2014.

1. Elementos transponíveis. 2. *C. reinhardtii*. 3.
V. carteri. 4. Evolução de genes. I. De Marco,
Ricardo, orient. II. Título.

Aos meus pais, Carlito e Melita.

Agradecimentos

Ao professor Ricardo De Marco, pela oportunidade concedida e pela orientação sempre presente no decorrer deste trabalho. Agradeço também pelos ensinamentos correlatos à Biologia Molecular e à pesquisa científica.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq - pelo suporte financeiro deste trabalho.

À Ana Eliza Zeraik, Débora Orcia e Natália O. Alves, pela convivência enriquecedora e pelas conversas sobre a área experimental da Biologia Molecular.

À professora Ana Paula Ulian de Araújo e à Juliana Sposto Avaca Crusca, pela realização do experimento relacionado à validação do gene Cre06.g262800.

Aos funcionários do Instituto de Física de São Carlos, que de forma direta ou indireta, contribuíram para a realização deste projeto.

Aos meus pais, Carlito e Melita, pelo apoio e incentivo ao estudo e, principalmente, pelo amor que fortalece em todos os momentos da vida. Às manas, Adriana e Juliana, pela alegria de tê-las como irmãs e por estarem sempre presentes, apesar da distância; agradeço também pelas discussões acerca da Estatística. Ao meu sobrinho Gabriel, por continuar trazendo tamanha alegria a este mundo.

Aos amigos Cláudia Weber, Fernanda Mônico, Elizandra Sehn, Francielle Sato, Kelly C. da Silva, Maria J. N. Amaral, Manoel M. A. de Jesus, Ivan Laczkowski, Simone F. Rodrigues, Andreza Beezão, Mayron Moreira, Jéssica Santos, Rafael Viegas, Ingrid Esteves, Renata Batista, Carolina Menezes Silvério e Moussa Mansour, pela amizade que segue com o tempo. Em especial, à Débora C. Corrêa e Daniele Jacinto, pelos momentos de risos soltos, de fortalecimento e de discussões científicas no transcorrer deste estudo.

Ao Arthur, pela ternura do seu companheirismo em todos os momentos.

A Deus, por ter tornado tudo isso possível.

“Transposable elements should not be viewed as genomic parasites, but rather as genomic symbionts that create a ‘genomic scrap yard’, the source of ‘junk’ that natural selection utilizes in its evolutionary experiments.”

Wojciech Makalowski (2000)

Resumo

PHILIPPSEN, G. S. **Estudo da influência de elementos transponíveis nos genomas das algas *C. reinhardtii* e *V. carteri***. 2014. 151p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2014.

Elementos transponíveis (TEs) são sequências de DNA que possuem a capacidade de transposição no genoma hospedeiro. O principal objetivo deste trabalho reside na investigação em torno de possíveis contribuições de TEs nos genomas das algas *C. reinhardtii* e *V. carteri*, mais especificamente, na arquitetura dos genes ortólogos nestas espécies. Neste contexto, análises em *silico* em larga escala foram realizadas, buscando-se identificar associações entre TEs e os genes ortólogos. Os resultados indicaram que os genes em *C. reinhardtii* tendem a acumular mais cópias de TEs em relação aos seus ortólogos em *V. carteri*. *C. reinhardtii* apresentou maior densidade de cópias de TEs para as regiões flangeadora 5' , flangeadora 3' e intrônica quando comparada a *V. carteri*; o inverso foi verificado quando analisada a densidade de TEs nas regiões codificantes. Análises para apurar a distribuição dos elementos em regiões intergênicas e intragênicas foram estabelecidas, nas quais a frequência observada dos elementos foi comparada à frequência esperada segundo a distribuição randômica de TEs no genoma, simulada computacionalmente. Foram constatadas regiões em que a presença dos elementos encontra-se significativamente abaixo do esperado, a exemplo de intervalos adjacentes ao início e ao término dos genes, o que provavelmente reflete a seleção negativa de eventos de integração nestas delimitações, em virtude dos efeitos deletérios associados à disrupção de estruturas de regulação da expressão gênica. De forma geral, nas regiões flangeadoras 5' e 3', foi identificada a tendência de elevação da frequência padronizada de TEs à medida que a classe de distância avaliada se distancia do início e do término do gene, respectivamente. A baixa representatividade dos elementos também foi constatada em regiões intragênicas. O estudo da distribuição de TEs nos íntrons dos genes ortólogos indicou a preservação destas regiões quanto à fixação de TEs, sendo a representatividade abaixo do esperado mais evidente em intervalos adjacentes ao éxon, o que minimiza a chance de ruptura no padrão de *splicing* dos genes. Em sequências codificantes, a escassez de TEs - esperada devido ao provável efeito deletério destes eventos para a função do gene - foi constatada nos ortólogos das duas espécies. No entanto, inovações decorrentes da integração dos elementos em regiões codificantes podem resultar em efeitos evolutivos positivos, embora estes eventos sejam raros. Nas espécies analisadas foram identificados dois casos, de especial interesse, em que um domínio da sequência peptídica encontra-se localizado em região derivada de TE: o primeiro refere-se ao gene Cre06.g262800, em *C. reinhardtii*, no qual foi identificado o domínio PHD-finger associada ao elemento Gypsy-5-LTR_CR; o segundo remete ao gene Vocar20001092m.g, em *V. carteri*, no qual o domínio zinc knuckle foi reconhecido em região derivada do elemento Gypsy3-LTR_VC. Estes genes constituem exemplos da contribuição de TEs na evolução de sequências codificadoras nas espécies *C. reinhardtii* e *V. carteri*, corroborando a hipótese de que os TEs podem contribuir na evolução da arquitetura dos genes, apesar do efeito disruptivo inerente à integração dos mesmos em regiões gênicas.

Palavras-chave: Elementos transponíveis. *C. reinhardtii*. *V. carteri*. Evolução de genes.

Abstract

PHILIPPSEN, G. S. **Influence of transposable elements in the genomes of *C. reinhardtii* and *V. carteri* algae.** 2014. 151p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2014.

Transposable elements (TEs) are DNA sequences able to transpose in the host genome. The aim of this study resides in the investigation of TEs contributions in the algae *C. reinhardtii* and *V. carteri* genomes, more specifically in the architecture of orthologous genes in these species. In this context, large scale *in silico* analysis were performed to identify associations between TEs and orthologous genes. The results indicated that genes in the *C. reinhardtii* specie tend to accumulate more TEs copies than orthologous genes in *V. carteri*. *C. reinhardtii* showed higher density of TEs copies in the 5' flanking, 3' flanking and intronic regions when compared to *V. carteri*; the opposite was observed in coding regions. Investigation of the elements distribution in the intergenic and intragenic regions was performed, in which the observed TE frequency was compared to expected TE frequency from the simulated random distribution of the elements in the genome. It was verified regions where TE frequency was significantly lower than expected, as in gene boundaries adjacencies, probably reflecting a negative selection of the TE integration events in these delimitations due to deleterious effects associated with disruption of gene regulatory structures. In general terms, it was observed an increasing standardized frequency in the 5' and 3' flanking regions as the distance from gene start and gene end, respectively, increases. TEs underrepresentation was also verified in the intragenic regions. The study of TEs distribution in the introns of orthologous genes revealed the preservation of these structures in relation to TEs fixation, with a stronger underrepresentation near exon, which minimizes the chance of gene splicing pattern disruption. In the coding sequences, the TEs scarcity - expected due the likely deleterious effects to gene function - was verified in the orthologous of both species. However, in rare instances, innovations mediated by TEs integration in the coding regions can lead to positive evolutionary effects. In the species analyzed two instances of particular interest were observed, in which the domain of peptide sequence is located in the region derived from TE. The first one refers to the Cre06.g262800 gene, in the *C. reinhardtii* specie, which has a PHD-finger domain associated with Gypsy-5-LTR_CR element. The second one refers to the Vocar20001092m.g gene, in *V. carteri*, in which the zinc knuckle was recognized in region derived from Gypsy3-LTR_VC element. These genes are examples of TEs contributions in the evolution of coding sequences in the *C. reinhardtii* and *V. carteri* species, corroborating the hypothesis that TEs can contribute to the evolution of gene architecture, despite the inherent disruptive effect in their integration in the gene regions.

Keywords: Transposable elements. *C. reinhardtii*. *V. carteri*. Gene evolution.

Lista de Figuras

Figura 1.1 -	Campos de pesquisa relacionados a TEs	26
Figura 2.1 -	Ilustração das etapas do processo de transposição de retrotransposons	29
Figura 2.2 -	Etapas do processo de transposição no mecanismo <i>cut-and-paste</i> .	31
Figura 2.3 -	Impacto de TEs na expressão gênica com enfoque na população .	32
Figura 2.4 -	Alterações na regulação gênica decorrentes da inserção de TEs . .	33
Figura 2.5 -	Representação da ocorrência de unicelularidade e multicelularidade em eucariotos	36
Figura 2.6 -	Organismos modelo	37
Figura 2.7 -	Ciclo de vida da espécie <i>C. reinhardtii</i>	38
Figura 2.8 -	Ciclo de vida assexuado da espécie <i>V. carteri</i>	39
Figura 2.9 -	Modelo do programa genético de diferenciação celular em <i>V. carteri</i>	40
Figura 2.10 -	Reprodução sexuada na espécie <i>V. carteri</i>	41
Figura 3.1 -	Passos do protocolo adotado para o mapeamento de cópias de TEs no genoma	49
Figura 3.2 -	Representação esquemática de alinhamentos no genoma	50
Figura 3.3 -	Representação de sobreposição de inserções no genoma	51
Figura 3.4 -	Frequência de ocorrência absoluta das famílias de TEs no genoma da espécie <i>C. reinhardtii</i>	53
Figura 3.5 -	Distribuição de frequências de ocorrência das famílias de TEs que apresentam cópias íntegras no genoma de <i>C. reinhardtii</i>	54
Figura 3.6 -	Frequência de ocorrência absoluta das famílias de TEs no genoma da espécie <i>V. carteri</i>	55
Figura 3.7 -	Distribuição de frequências de ocorrência das famílias de TEs que apresentam cópias íntegras no genoma de <i>V. carteri</i>	56
Figura 3.8 -	Box-plot da distribuição de comprimentos das cópias de TEs nos genomas das duas espécies	57
Figura 3.9 -	Representação esquemática da região gênica	58
Figura 3.10 -	Box-plot da distribuição de densidades de TEs em regiões gênicas dos ortólogos	61
Figura 3.11 -	Fração de genes ortólogos com presença de TE na região flanke- adora 5', codificante, intrônica e flaqueadora 3'	61
Figura 3.12 -	Distribuição de frequências de ocorrência das classes de TEs em regiões gênicas dos ortólogos	62
Figura 3.13 -	Densidade de cópias de TEs nas sub-localizações da região gênica de genes ortólogos	63
Figura 4.1 -	Dogma central da expressão gênica	66
Figura 4.2 -	Unidades reguladoras da transcrição em eucariotos	67

Figura 4.3 -	Reprentação esquemática da região flanqueadora 5'	70
Figura 4.4 -	Reprentação esquemática de um contexto genômico analisado . .	71
Figura 4.5 -	Distribuição de frequências relativas do número de cópias randômicas de TEs na região flanqueadora 5'	74
Figura 4.6 -	Distribuição de frequências relativas do número de cópias randômicas de TEs, em classes de distância na região flanqueadora 5'	75
Figura 4.7 -	Distribuição de frequências de TEs na região flanqueadora 5' de genes ortólogos em <i>C. reinhardtii</i>	76
Figura 4.8 -	Distribuição de frequências de TEs na região flanqueadora 5' de genes ortólogos em <i>V. carteri</i>	77
Figura 4.9 -	Frequência de ocorrência absoluta das famílias de TEs na região flanqueadora 5' dos genes ortólogos	79
Figura 4.10 -	Posições, em termos da porcentagem do comprimento total da sequência consenso do TE, inseridas na região flanqueadora 5' dos genes ortólogos	79
Figura 4.11 -	Fatores envolvidos no processo de poliadenilação em mamíferos .	81
Figura 4.12 -	Reprentação esquemática da região flanqueadora 3' do gene . . .	83
Figura 4.13 -	Formas de associação do sítio poli(A) com TE	84
Figura 4.14 -	Distribuição de frequências de TEs na região flanqueadora 3' de genes ortólogos em <i>C. reinhardtii</i>	86
Figura 4.15 -	Distribuição de frequências de TEs na região flanqueadora 3' de genes ortólogos em <i>V. carteri</i>	87
Figura 4.16 -	Associação entre sítios poli(A) e TEs na região flanqueadora 3' de genes ortólogos	88
Figura 4.17 -	Exemplos de associações entre TEs e sítios poli(A) em genes da espécie <i>C. reinhardtii</i>	89
Figura 4.18 -	Exemplo de sítio poli(A) associado a duas cópias de TEs	90
Figura 5.1 -	Representação esquemática de formas de incorporação de TEs em sequências codificantes nos genes	92
Figura 5.2 -	Representação esquemática da condição requerida para validação, via EST, da presença de TE em sequência codificante predita . .	93
Figura 5.3 -	Representação esquemática dos genes de <i>C. reinhardtii</i> com presença de TE em sequência codificante confirmada por EST	96
Figura 5.4 -	Representação esquemática dos genes de <i>V. carteri</i> com presença de TE em sequência codificante confirmada por EST	97
Figura 5.5 -	Domínio PHD-finger identificado na sequência peptídica do gene Cre06.g262800 em <i>C. reinhardtii</i>	98
Figura 5.6 -	Representação esquemática do transcrito derivado do gene Cre06.g262800 de <i>C. reinhardtii</i>	99
Figura 5.7 -	Alinhamentos relativos ao domínio PHD-finger identificado em íntron do gene Vocar20002708m.g em <i>V. carteri</i>	100
Figura 5.8 -	Domínio zinc knuckle (zf-CCHC) identificado na sequência peptídica do gene Vocar20001092m.g em <i>V. carteri</i>	101
Figura 5.9 -	Filogenia dos domínios zinc knuckle (C2HC) identificados no elemento Gypsy3-LTR_VC e nos genes ortólogos das espécies <i>C. reinhardtii</i> e <i>V. carteri</i>	102

Figura 5.10 - Distribuição de frequências da posição (em relação à extremidade gênica mais próxima) ocupada pelo éxon contendo o fragmento de TE	103
Figura 5.11 - Mecanismo canônico de <i>splicing</i> e snRNAs (<i>small nuclear RNAs</i>) envolvidos no mesmo	104
Figura 5.12 - Exemplo de mapa de cores para frequência relativa de TEs em íntrons de genes ortólogos	107
Figura 5.13 - Representação esquemática dos modelos de reconhecimento do pré-mRNA definidos no íntron e no éxon	110
Figura 5.14 - Distribuição de frequências de TEs nos íntrons de genes ortólogos em <i>C. reinhardtii</i>	111
Figura 5.15 - Orientação senso e antissenso de TEs nos íntrons de genes ortólogos	112
Figura 5.16 - Mapa de cores para a frequência relativa de cópias de TEs em íntrons de genes ortólogos	114
Figura 5.17 - Distribuição de comprimentos de íntrons em três conjuntos de dados: pares de íntrons sem cópia de TEs nas duas espécies (A), pares em que apenas o íntron de <i>C. reinhardtii</i> apresenta cópia de TE (B) e pares em que apenas o íntron de <i>V. carteri</i> abriga TE (C)	116
Figura 5.18 - Box-plot da distribuição de razões de conteúdo GC para pares de íntrons ortólogos em <i>C. reinhardtii</i> e <i>V. carteri</i>	117
Figura 5.19 - Distribuição de distâncias par a par para o conjunto de cópias de TEs intrônicas e não-intrônicas, computadas para as três famílias mais frequentes em íntrons de genes ortólogos em <i>C. reinhardtii</i> .	119
Figura 5.20 - Distribuição de distâncias par a par para o conjunto de cópias de TEs intrônicas e não-intrônicas, computadas para as três famílias mais frequentes em íntrons de genes ortólogos em <i>V. carteri</i> . . .	120
Figura 5.21 - Razões entre a frequência de distâncias menores ou iguais a 1,0 e frequência de distâncias acima de 1,0	121

Lista de Tabelas

Tabela 2.1 -	Dados relativos ao genoma e aos modelos gênicos para as espécies <i>V. carteri</i> e <i>C. reinhardtii</i>	42
Tabela 3.1 -	Resultados do mapeamento de TEs nos genomas de <i>C. reinhardtii</i> e <i>V. carteri</i>	52
Tabela 3.2 -	Número de cópias anotadas para cada classe de TE nos genomas de <i>C. reinhardtii</i> e <i>V. carteri</i>	52
Tabela 4.1 -	Frequências observada e esperada para as classes de TEs na região flanqueadora 5' dos genes ortólogos em cada espécie	73
Tabela 4.2 -	Frequências observada e esperada para as classes de TEs na região flanqueadora 3' dos genes ortólogos em cada espécie	85
Tabela 5.1 -	Frequências observada e esperada para as classes de TEs na região codificante dos genes ortólogos em cada espécie.	94
Tabela 5.2 -	Domínios proteicos associados aos genes que contém cópias de TEs, comprovadas via EST, na região codificante.	95
Tabela 5.3 -	Frequências observada e esperada para as classes de TEs na região intrônica dos genes ortólogos em cada espécie.	109
Tabela 5.4 -	P-valores obtidos no teste binomial exato para orientação das cópias de TEs, nas classes de distância em íntrons de genes ortólogos	113
Tabela A.1 -	Dados relativos aos TEs descritos para a espécie <i>C. reinhardtii</i>	139
Tabela A.2 -	Dados relativos aos TEs descritos para a espécie <i>V. carteri</i>	143
Tabela B.1 -	Valores da estatística z para as classes de distância da região flanqueadora 5' dos genes ortólogos na espécie <i>C. reinhardtii</i> . . .	146
Tabela B.2 -	Valores da estatística z para as classes de distância da região flanqueadora 5' dos genes ortólogos na espécie <i>V. carteri</i>	147
Tabela B.3 -	Valores da estatística z para as classes de distância da região flanqueadora 3' dos genes ortólogos na espécie <i>C. reinhardtii</i> . . .	148
Tabela B.4 -	Valores da estatística z para as classes de distância da região flanqueadora 3' dos genes ortólogos na espécie <i>V. carteri</i>	149
Tabela B.5 -	Valores da estatística z para as classes de distância em íntrons nos genes ortólogos na espécie <i>C. reinhardtii</i>	150
Tabela B.6 -	Valores da estatística z para as classes de distância em íntrons nos genes ortólogos na espécie <i>V. carteri</i>	151

Sumário

1	Introdução	23
2	Conceitos iniciais	27
2.1	Elementos transponíveis	27
2.2	Organismos modelo	35
3	Genes ortólogos e associação destes com TEs	45
3.1	Mapeamento das inserções de TEs nos genomas	46
3.1.1	Introdução	46
3.1.2	Métodos	47
3.1.3	Resultados	51
3.2	Associação de TEs com genes ortólogos	57
3.2.1	Métodos	57
3.2.2	Resultados	60
4	Estudo de TEs em regiões intergênicas	65
4.1	Região flanqueadora 5' dos genes ortólogos	67
4.1.1	Introdução	67
4.1.2	Métodos	70
4.1.3	Resultados	73
4.2	Região flanqueadora 3' dos genes ortólogos	80
4.2.1	Introdução	80
4.2.2	Métodos	82
4.2.3	Resultados	85
5	Estudo de TEs em regiões intragênicas	91
5.1	Região codificante	91
5.1.1	Introdução	91
5.1.2	Métodos	92
5.1.3	Resultados	94
5.2	Região intrônica	104
5.2.1	Introdução	104

5.2.2	Métodos	106
5.2.3	Resultados	108
6	Conclusões	123
	Referências	127
	Apêndice A Elementos de transposição	139
	Apêndice B Resultados de análises estatísticas	145

Capítulo 1

Introdução

Na década de 1.940, ao realizar experimentos com milho, Barbara McClintock deparou-se com a descoberta da “mobilidade de componentes específicos do genoma”, componentes estes inicialmente por ela denominados “elementos de controle”, em função da hipótese de que regulariam a expressão gênica.(1) Posteriormente, estes elementos passaram a ser referenciados como elementos transponíveis (TEs).(2, 3) Por meio de seu achado, a pesquisadora concluiu que os genomas eucarióticos contêm sequências que possuem a habilidade de transposição de uma localização cromossômica para outra, estabelecendo uma nova perspectiva em que os genomas não seriam estáticos.(4, 5) A mobilidade de segmentos genômicos contrariava a conjectura de estaticidade dos fatores genéticos vigente na época, de modo que a comunidade científica permanecera cética a respeito desta descoberta.(4–6)

A identificação de um elemento de DNA móvel em *E. coli* em 1.969, bem como estudos conduzidos em *Drosophila* na década de 1.970, modificaram este cenário.(3, 5) A existência dos TEs fora reconhecida, mas a aparente falta de função dos mesmos para o hospedeiro conduziu-os ao status de DNA lixo, contrariamente ao apontado inicialmente por Barbara McClintock.(5) Nesta época, a compreensão de que as proteínas codificadas pelos TEs eram necessárias somente à sua transposição - visando a sobrevivência dos mesmos no genoma hospedeiro - e não a funções celulares, contribuiu para a designação destes elementos como genes egoístas.(7) Mesmo considerando possível que em alguns casos sequências de TEs poderiam adquirir uma função que resultasse em vantagem seletiva ao organismo hospedeiro, a hipótese dos mesmos como parasitas genômicos foi indicada como a mais razoável por pesquisadores como Doolittle, Sapienza, Orgel e Crick.(5, 7, 8)

Esta visão prevaleceu até a década de 1.990, na qual o desenvolvimento das técnicas de sequenciamento viabilizou a análise do genoma de diversas espécies e expôs o fato de que praticamente todos os organismos estudados abrigam TEs, os quais amplificaram suas cópias no decorrer da evolução.(2, 5, 6, 9) Esta constatação, juntamente com o número crescente de evidências em torno da habilidade dos TEs em introduzir inovações nos aspectos regulatório e codificante dos genes (10–12), suscitou o debate acerca da

importância dos mesmos na evolução das espécies.

Atualmente, a teoria que define os TEs como fonte de variabilidade genética apta a influenciar trajetórias evolutivas, apesar da natureza parasitária dos mesmos, é aceita.(2, 5, 10–20) O aspecto que os caracteriza como relevantes na questão evolutiva relaciona-se à sua capacidade de transposição: devido a eventos de inserção, genes podem ser acometidos por modificações em sua estrutura e regulação, ou ainda conduzidos à disrupção de sua função.(11, 12, 15, 21) Estudos que investigam a contribuição de elementos móveis no genoma humano indicaram que aproximadamente 24% dos promotores possuem sequências derivadas de TEs, assim como 4% dos genes possuem TEs em regiões codificantes.(11, 22) Em termos de composição, aproximadamente 45% do genoma humano deriva de elementos de transposição.(3)

Nekrutenko e Li (11) apontam que integrações de TEs, comparadas a mutações decorrentes da substituição de nucleotídeos, podem acelerar a evolução dos genes e desencadear a rápida divergência entre as espécies. Como exemplo, os autores citam o elemento Alu, verificado em sequências codificantes de aproximadamente 1,3% dos genes humanos, porém não encontrado em não primatas; é possível que este elemento de transposição tenha contribuído para a divergência entre primatas e outros mamíferos. Levin e Moran (9) sugerem que, possivelmente, apenas “a ponta do *iceberg*” tenha sido vislumbrada e que a influência dos TEs na variação estrutural dos genomas mamíferos ainda seja subestimada. Muitos exemplos de contribuições desta natureza nunca serão identificados em função da divergência em relação às sequências originais dos TEs, o que impede o reconhecimento destes eventos.(10)

Acredita-se que a transposição da maioria dos TEs objetiva a manutenção dos próprios elementos no genoma, sendo raros os eventos que culminam na incorporação de sequências provenientes dos mesmos para funções celulares, resultando em efeitos positivos ao hospedeiro.(5, 9, 11) Para minimizar efeitos deletérios, alguns elementos transponíveis desenvolveram mecanismos que direcionam sua integração para intervalos “seguros” no genoma, como regiões de baixa densidade gênica (heterocromatina) ou distantes da região promotora dos genes.(9) O elemento LTR *Ty5*, em *S. cerevisiae*, realiza inserções preferencialmente em regiões de heterocromatina, devido à afinidade do domínio fosforilado de sua integrase pelo componente estrutural Sir4 (do inglês, *silent information regulator 4*) da heterocromatina.(9, 23) Já o elemento P, em *D. melanogaster*, evita a disrupção dos genes por meio da integração em um ponto que antecede em 500 pb o sítio de início de transcrição; no entanto, o mecanismo que resulta neste padrão de inserção ainda é desconhecido.(9)

Em contrapartida, hospedeiros evoluíram estratégias que restringem a atividade dos TEs nas células germinativas e somáticas.(6, 9) Uma das formas de silenciamento envolve

o controle epigenético^a, em que a metilação da sequência genômica dos TEs impede a transcrição dos mesmos.(6, 9, 20, 24) Outro recurso utilizado pela célula para o controle da proliferação de TEs remete à interferência de RNA.(6, 9, 24) Neste, pequenos RNAs de interferência (do inglês, *small interfering RNAs* - siRNAs) interagem com o RNA mensageiro oriundo da transcrição do TE, conduzindo-o à uma via de degradação e ao consequente impedimento da transposição do elemento.(6, 9, 24)

Estudos indicam também a existência de mecanismos celulares aptos a influenciar a atividade dos TEs em condições adversas ao desenvolvimento do hospedeiro.(9) O trabalho realizado por Dai et al. (23) mostrou que o já citado retrotransposon *Ty5*, em *S. cerevisiae*, modifica seu padrão de inserção quando a célula é exposta a condições de stress, a exemplo da privação de aminoácidos ou nitrogênio. Nestas situações adversas, a enzima integrase codificada pelo TE não é fosforilada - devido à ausência de modificação pós-transcricional realizada pela célula hospedeira em condições normais - perdendo a especificidade de integração por regiões de heterocromatina, de forma que o elemento passa a integrar regiões ricas em genes, frequentemente causando mutações.(23) Segundo os autores, a especificidade de integração do elemento *Ty5* é alterada mediante a privação de nutrientes, o que é consistente com a hipótese de Barbara McClintock de que os TEs podem introduzir mudanças no genoma em resposta a mudanças ambientais.(9, 23)

Ainda na conjectura de que os TEs podem representar uma fonte de inovações mediante situações adversas ao hospedeiro, cita-se o trabalho de Lagemaat et al..(12) Este indicou que genes associados a respostas ao stress e a estímulos externos, em humano e camundongo, mostraram maior acúmulo de TEs quando comparados a outras classes de genes, a exemplo dos correlatos à manutenção celular. Este resultado reforça a hipótese de que eventos de integração em regiões gênicas possam estar relacionados a respostas adaptativas, contribuindo para o aumento da diversidade genética e chance de sucesso do hospedeiro.

O desenvolvimento da pesquisa relacionada a TEs conduziu ao cenário atual, em que diversos campos de estudo correlatos a estes elementos encontram-se sob exploração científica. A Figura 1.1 ilustra algumas das linhas de estudo, dentre as quais cita-se a relação entre TEs e a evolução dos genomas, contribuições de TEs nos aspectos regulatório e codificante dos genes, estudo da evolução e biologia dos próprios TEs e desenvolvimento de tecnologias para aplicação de TEs em terapias gênicas.(6) De forma geral, todos estes campos convergem para o entendimento da estrutura e função biológica dos elementos de transposição nos genomas.

É justamente no contexto de investigação a respeito da influência dos TEs na evolução de genomas, mais especificamente na evolução de estruturas gênicas ortólogas nas espécies *C. reinhardtii* e *V. carteri*, que se insere este trabalho. Lopes et al. (17) apontam que contribuições de TEs no genoma humano têm sido amplamente investigadas, enquanto

^a Entende-se por mecanismo epigenético o controle da expressão de genes ou TEs, sem modificações na sequência de DNA destas estruturas.(5)

Figura 1.1 – Campos de pesquisa relacionados a TEs. Em preto, etapas e fluxo de informação para o estudo de regiões codificantes de genes; em azul, áreas de estudo relativas aos TEs. Fonte: JANICKI;ROOKE;YANG(6)

em plantas análises mais detalhadas são ainda necessárias. Trabalhos recentes visam o aprofundamento do conhecimento nesta área, a exemplo de estudos realizados nos genomas de arroz (18), café (17), tomate e pimenta.(25) A necessidade de análises genômicas em diversas espécies, para o estudo da contribuição de TEs em eventos evolutivos, também fora apontada por Biémont.(3) Dessa forma, a publicação recente dos genomas das algas *C. reinhardtii* e *V. carteri* (26, 27) propicia a oportunidade de estudo da influência dos TEs na arquitetura dos genes ortólogos nestas espécies, as quais constituem um importante modelo de estudo em diversas questões na biologia.(28–30) Conceitos relativos a estas algas, assim como aos TEs, são apresentados no Capítulo 2.

Análises em sílico em larga escala foram estabelecidas com o intuito de identificar possíveis contribuições dos TEs nos genomas de *C. reinhardtii* e *V. carteri*. Primeiramente, as bibliotecas de TEs descritas para cada espécie - depositadas no Repbase (31) - foram interrogadas junto aos genomas. Estratégias de análise definidas para o mapeamento das cópias de TEs, bem como para a identificação das instâncias destas que possuem relação com genes ortólogos, são descritas no Capítulo 3. Aspectos tangentes à influência dos TEs na arquitetura destes genes são explorados nos capítulos 4 e 5. O Capítulo 4 retrata a distribuição de TEs em regiões flangeadoras 5' e 3' dos genes ortólogos, bem como possíveis contribuições dos mesmos no enfoque regulatório. Ocorrências dos elementos móveis em regiões codificantes são descritas no Capítulo 5, no qual explora-se também a distribuição dos TEs nos íntrons. Por meio dos resultados apresentados no decorrer deste trabalho, esperamos contribuir no conhecimento acerca da distribuição de TEs e influência destes elementos nos genomas das algas *C. reinhardtii* e *V. carteri*.

Capítulo 2

Conceitos iniciais

De acordo com o exposto na introdução, o principal objetivo deste trabalho reside na investigação acerca de possíveis influências dos TEs nos genomas das algas *C. reinhardtii* e *V. carteri*. Por este motivo, o presente capítulo destina-se à apresentação de conceitos biológicos relativos aos TEs - com enfoque nos mecanismos de transposição e possibilidades de mutação desencadeadas por estes eventos - e às espécies que constituem o modelo de estudo.

2.1 Elementos transponíveis

Classificação e mecanismos de transposição

Elementos de transposição são sequências genéticas que podem se deslocar de um ponto a outro no genoma hospedeiro, sendo classificados em dois grandes grupos de acordo com o mecanismo de transposição: retrotransposons e transposons de DNA.(2, 13) Enquanto os retrotransposons realizam novas inserções no genoma por meio de um RNA intermediário, os elementos pertencentes ao segundo grupo deslocam seu DNA sem o emprego de RNA.(2, 9, 13)

Outro aspecto interessante relacionado ao mecanismo de transposição refere-se ao fato de que os elementos transponíveis, independentemente do grupo ao qual pertençam, podem apresentar-se em duas variações: autônomos e não autônomos. Elementos autônomos codificam a maquinaria protéica necessária à transposição, ao passo de que elementos não-autônomos fazem uso da maquinaria produzida por outros elementos transponíveis com os quais possuem parentesco.(2)

Retrotransposons

Os retrotransposons eucarióticos são divididos em quatro classes: LTR (*long terminal repeat*), non-LTR (*non-long terminal repeat*), Penelope e DIRS (*Dictyostelium intermediate repeat sequence*), sendo as duas primeiras as principais e mais estudadas classes de TEs cuja transposição envolve RNA intermediário.(2)

Retrotransposons LTR A nomenclatura desta classe tem origem nos longos terminais repetitivos (LTRs), sequências repetitivas diretas verificadas nas extremidades dos elementos.(24) TEs LTR autônomos codificam as proteínas Gag e pol, sendo a última constituída por domínios protease, transcriptase reversa e integrase.(9, 24) Como exemplo de elementos pertencentes a esta classe, cita-se o Ty, em *S. cerevisiae*, e o gypsy, em *D. melanogaster*.(24)

As etapas inerentes à transposição dos elementos LTR estão representadas na Figura 2.1-A. O LTR 5' contém um promotor que é reconhecido pela RNA polimerase II do hospedeiro.(9) Após a transcrição do mRNA intermediário, o mesmo é exportado ao citoplasma onde ocorre sua tradução (não ilustrado na figura).(24) Constitui-se então o invólucro protéico VLP (*virus-like particle*), composto por proteínas Gag, que acaba por encapsular o mRNA que será transcrito reversamente, originando o cDNA.(2, 9, 24) Este último é então importado pelo núcleo, onde será inserido no genoma hospedeiro pela ação da integrase.(9, 24)

Elementos LTR são filogeneticamente e estruturalmente próximos de retrovírus, os quais codificam ainda um envelope constituído por glicoproteínas (gene env), cuja função é associada ao reconhecimento de receptores de membrana da célula hospedeira no início da infecção, conferindo-lhes a habilidade de migração entre células vizinhas ou entre organismos.(14, 32) Este quadro remete à uma questão interessante que reside na relação evolutiva entre retrovírus e retrotransposons LTR: enquanto uma vertente propõe que os retrovírus tenham emergido de elementos retrotransponíveis ao adquirirem o gene env (32), a hipótese contrária, em que a intracelularização de um retrovírus seguida da deleção do gene env teria originado elementos retrotransponíveis, é também igualmente plausível.(33)

Retrotransposons non-LTR De maneira semelhante à classe anterior, a nomenclatura adotada para esta classe deriva da estrutura dos elementos, que exhibe ausência de LTRs.(24)

Elementos autônomos non-LTR, comumente designados por LINE (*long interspersed nuclear element*), possuem um ou dois quadros de leitura abertos (do inglês, *open reading frame* - ORF), além de um promotor interno na região terminal 5' que controla a transcrição do retrotransposon (Figura 2.1-B).(2, 24) O primeiro quadro de leitura

Figura 2.1 – Ilustração das etapas do processo de transposição de retrotransposons. (A) Estruturas LTR de elementos desta classe, representados pelas setas em preto, flanqueiam a região codificante Gag, protease, transcriptase reversa e integrase.(9) Após a transcrição do mRNA intermediário e tradução deste no citoplasma, ocorre também nesta região celular a edificação da estrutura VLP, constituída por proteínas Gag (círculos em rosa).(24) É no interior deste invólucro que ocorre a transcrição reversa do RNA, mediada pela transcriptase reversa (representada em laranja), originando o cDNA que será transportado ao núcleo para integração no DNA hospedeiro por meio da integrase (representada pelos círculos em lilás).(9, 24) O reparo da região clivada pode gerar sequências TSD (*target-site duplication*), indicadas pelas linhas horizontais em preto adjacentes à nova cópia (retângulo em azul marinho).(9) (B) A transposição de non-LTR autônomos tem início com a transcrição, produzindo o mRNA que será exportado ao citoplasma onde originará a ribonucleoproteína com domínios relativos à endonuclease e à transcriptase reversa (ORF 2).(2, 24) Este complexo é transferido para o núcleo celular para a realização do processo TPRT.(2, 24) O novo elemento é geralmente truncado na região 5', como representado na ilustração; segmentos adjacentes à cópia representam sequências TSD ou pequenas deleções no ponto clivado do DNA.(9) Fonte: LEVIN;MORAN(9)

codifica a proteína chaperona, enquanto o segundo codifica uma proteína com domínios relativos à endonuclease (EN) e à transcriptase reversa (TP).(24) A transposição tem início com a transcrição do mRNA, o qual é exportado ao citoplasma onde ocorrerá a

tradução das proteínas necessárias ao processo (não ilustrado na figura), constituindo assim a ribonucleoproteína que será transportada ao núcleo para que ocorra o mecanismo TPRT (*target-primed reverse transcription*).⁽²⁴⁾ Neste, a inserção tem início pela ação da endonuclease que cliva o DNA alvo, dando origem a um grupo hidroxil 3' livre que será utilizado como *primer* para a transcrição reversa do RNA intermediário.^(2, 9, 24) O processo de inserção é finalizado após a síntese da segunda fita de cDNA, seguida da integração completa da cópia via mecanismos de reparo de DNA da célula.⁽²⁴⁾ Alguns elementos non-LTR não possuem cauda poli(A).⁽⁹⁾

Como exemplo desta classe, cita-se o retrotransposon L1 (LINE-1), sendo este o elemento LINE mais abundante em humanos, constituindo aproximadamente 17% do genoma desta espécie.^(9, 24) Apesar de encontrar-se frequentemente truncado na região 5', devido à baixa processividade da enzima transcriptase reversa, o elemento L1 aparentemente existente na forma ativa em todos os mamíferos, sendo o número médio de cópias ativas deste TE no genoma humano estimado entre 80 e 100.^(9, 13, 34) Elementos não autônomos desta classe são referenciados como SINEs (*short interspersed nuclear elements*), sendo o Alu o representante desta categoria mais comum em humano.⁽²⁴⁾

Transposons de DNA

Elementos pertencentes a este grupo realizam transposição por meio da transferência de sua cópia genômica de uma localização para outra no genoma hospedeiro, sem a síntese de um RNA molde intermediário.⁽²⁾ São divididos em três classes segundo seu mecanismo de transposição: *cut-and-paste*, *Helitrons* e *Polintons*.⁽²⁾ Respectivamente, os mecanismos de transposição são *cut-and-paste*, *rolling-circle replicative* e *self-synthesizing*, sendo o primeiro considerado um mecanismo não replicativo devido ao caráter “recorta e cola” inerente ao mesmo.^(2, 9)

Transposons *Cut-and-paste* Como já citado, deslocam seu DNA por meio do processo *cut-and-paste* (Figura 2.2), no qual os terminais de repetição invertida (do inglês, *terminal inverted repeats* - TIRs) do elemento são reconhecidos pela enzima transposase, codificada por elementos autônomos desta classe.^(2, 9) Após a interação da enzima com os TIRs ou regiões adjacentes aos mesmos, o DNA hospedeiro sofre uma quebra de dupla fita nas regiões que flanqueiam o TE, liberando-o para a inserção em um outro ponto alvo do genoma.^(9, 35) Ambas as reações de quebra e ligação de DNA envolvidas neste procedimento são catalizadas pela enzima transposase.⁽²⁾

Como exemplos desta classe, citam-se os transposons *Mariner*, *Merlin*, *piggyBac* e *Harbinger*, tendo sido este último o primeiro elemento transponível descoberto via estudos computacionais.⁽²⁾

Figura 2.2 – Etapas do processo de transposição no mecanismo *cut-and-paste*. A enzima transposase (representada pelos círculos em lilás) liga-se nas proximidades dos terminais de repetições invertidas (setas em preto), onde efetua uma reação de quebra de dupla fita que remove o transposon de sua posição inicial.(9, 35) O TE excisado será integrado pela transposase em uma nova posição do genoma hospedeiro (retângulo em cinza escuro), em uma reação que pode originar ainda sequências TSD (representadas pelas linhas horizontais em preto, adjacentes ao elemento na nova posição).(2, 9) O trecho do DNA hospedeiro em que ocorreu a remoção será alvo da maquinaria de reparo celular.(36) Fonte: LEVIN;MORAN(9)

Elementos transponíveis como mediadores do processo evolutivo

Estudos envolvendo TEs indicam que os mesmos desempenham um importante papel na dinâmica da evolução das espécies: no decurso do deslocamento da sequência genômica destes elementos, genes podem ter sua regulação ou sequência codificante alterada, passar à inatividade ou ainda novos genes podem ser originados.(2, 11–13, 15, 21, 37–40) Um evento de integração de TE pode ocasionar efeitos positivos, negativos ou neutros ao hospedeiro.(5, 14) Inserções em estruturas gênicas são, geralmente, deletérias para o indivíduo; no entanto, sob a perspectiva da população, inserções deletérias são selecionadas

negativamente e as inserções retidas tendem à neutralidade ou a efeitos mutacionais benéficos (Figura 2.3).(5, 6, 14) As variações mediadas por TEs podem, portanto, ser vantajosas para a sobrevivência e proliferação da espécie hospedeira.(5, 6)

Figura 2.3 – Impacto de TEs na expressão gênica com enfoque na população. Em células germinativas (A), eventos de transposição em indivíduos da população (representado pelos círculos maiores) podem resultar em efeitos benéficos (setas em verde), deletérios (setas em vermelho) ou neutros.(5) Por meio da seleção purificadora, as cópias de TEs com efeitos negativos não são retidas na população, enquanto aquelas com efeitos positivos ou neutros permanecem, o que pode conduzir ao advento de inovações benéficas para a população ou subpopulação (círculo menor).(5, 6) O mesmo não ocorre em células somáticas (B), uma vez que os eventos de integração ocorridos no indivíduo não serão transmitidos à geração seguinte.(5) Apesar do exemplo referenciar a influência de TEs na regulação gênica, a mesma lógica é aplicável a eventos de integração em regiões codificantes e intrônicas. Fonte: REBOLLO;ROMANISH;MAGER(5)

Mutações regulatórias são aquelas que alteram o padrão espacial ou temporal da expressão do gene, sem necessariamente alterar a sequência que codifica a proteína.(21) Estas podem manifestar seus efeitos na transcrição ou na fase que a sucede (15), como ilustra a Figura 2.4. Nos aspectos correlatos à transcrição, a inserção de um TE em uma região adjacente ao início de um gene pode introduzir um promotor alternativo ao pré-existente - instituindo sítio de iniciação de transcrição alternativo (a), pode danificar estruturas regulatórias existentes (b) ou ainda originar uma nova estrutura regulatória (c).(15) Uma inserção no interior de um íntron pode alterar o sentido da transcrição (d), ou atuar como centro de formação de heterocromatina, acarretando assim o silenciamento da transcrição do gene e possivelmente de genes adjacentes a esta região (e).(15) No âmbito pós-transcricional, uma inserção no terminal 3' UTR (*untranslated region*) pode introduzir um sítio alternativo de poliadenilação (f) ou um sítio ligante para um micro RNA (g).(15) Já uma inserção no interior de um íntron pode interferir no padrão normal de *splicing* do pré-mRNA (h) ou originar um processo em que o TE é incorporado como um éxon alternativo: neste caso, uma nova proteína pode ser originada ou o mRNA

pode ser conduzido à degradação via mecanismo celular NMD (*nonsense-mediated decay*), especialmente nos casos em que se origina um códon de parada prematuro, representado na figura por um asterisco (*). (15)

Figura 2.4 – Alterações na regulação gênica decorrentes da inserção de TEs. O elemento transposto é representado em marrom. Fonte: FESCHOTTE(15)

O gene *Amy 1* exemplifica a alteração do padrão de expressão gênica mediada por TEs. Na maioria dos mamíferos, genes membros da família amilase são expressos exclusivamente no pâncreas; em humanos, o gene *Amy 1*, pertencente a esta família gênica, possui um retrotransposon inserido na região flanqueadora 5' que conduziu à expressão do mesmo em glândulas salivares. (21)

Sequências de TEs codificadoras de transposase, integrase e transcriptase reversa foram domesticadas^a por genomas hospedeiros eucarióticos ao longo da evolução. (14, 41) Alzohairy et al. (41) apontam que cerca de 50 a 100 casos de genes codificadores de proteínas, originados a partir de sequências codificantes de TEs das classes retrotransposon ou DNA, foram identificados em mamíferos e plantas.

Segundo Volff (14), o caso mais estudado envolvendo domesticação de transposase refere-se à proteína Rag1, inerente ao sistema imune adaptativo em vertebrados mandibulados.

^a O termo domesticação é empregado para designar casos em que ORFs de TEs são co-adaptados para desempenhar uma função relativa ao genoma hospedeiro. (5) TEs domesticados diferem de TEs autônomos por apresentarem perda da capacidade de transposição; além disso, em muitos casos, apresentam-se conservados em *loci* ortólogos em diferentes espécies. (14, 41, 42)

As proteínas Rag1 e Rag2 iniciam a recombinação V(D)J, que consiste em um mecanismo de recombinação sítio específica em células B e T, em que os segmentos V, D e J de genes distintos são combinados em um único éxon que codificará a região variável do receptor que se liga ao antígeno.(14, 41, 43) Este mecanismo permite a criação de uma variedade de receptores aptos ao reconhecimento de antígenos.(14) A enzima Rag1 possui atividade nuclease e transposase, introduzindo quebra de dupla fita próxima aos sinais específicos de recombinação.(14, 41, 43) Evidências indicam que tanto o cerne da proteína Rag1, quanto os sinais de recombinação, teriam sua origem no elemento Transib, um transposon de DNA.(43)

Outro exemplo interessante remete aos fatores FHY3 (*far-red-elongated hypocotyl 3*) e FAR1 (*far-red-impaired response 1*) em plantas, descobertos em Arabidopsis.(14, 41) Estes fatores atuam na ativação da transcrição dos genes FHY1 (*far-red-elongated hypocotyl 1*) e FHL (*FHY1-like*), cujos produtos interagem com o fotorreceptor phyA (*phytochrome A*), associado ao monitoramento da luz FR (do inglês, *far-red*, relativo à faixa de 700 a 750 nm).(44) Ao ser ativada pela luz, a proteína phyA altera sua conformação, interagindo com os fatores FHY1 e FHL, disparando uma via que acaba por conduzi-la do citoplasma ao núcleo da célula, onde desencadeia respostas à luz FR, a exemplo da germinação de sementes e floração.(44) Suscintamente, os fatores FHY3 e FAR1 são relacionados à regulação do acúmulo de phyA no núcleo, por meio da ativação da expressão dos fatores FHY1 e FHL.(44) Os fatores FHY3 e FAR1 possuem domínios de ligação ao DNA correlatos à transposase de elementos Mutator.(44) Segundo Lin et al. (44), este constitui um exemplo de domesticação molecular acompanhada de perda da habilidade de transposição.

Um possível exemplo de domesticação da enzima transcriptase reversa de elementos non-LTR relaciona-se à telomerase, associada à replicação dos telômeros em animais, plantas, fungos e protozoários.(14) A telomerase contém uma subunidade de RNA utilizada como molde para a replicação do telômero e pertence à família de enzimas transcriptase reversa, assim como transcriptases reversas codificadas por retrotransposons, com as quais possui relação filogenética.(14) Além disso, a telomerase utiliza o término 3' hidroxil do cromossomo como *primer* para a transcrição reversa, em um mecanismo muito semelhante ao utilizado por elementos non-LTR no processo TPRT após a clivagem do DNA alvo.(14) Segundo Volff (14), não é possível definir se a telomerase realmente deriva de um TE ou se o gene desta proteína fora capturado e incorporado por retroelementos.

Outro processo evolutivo que pode ser mediado por TEs é a duplicação gênica^b, resultante do carregamento de cópias dos genes no deslocamento do TE ou via transcrição reversa do mRNA do gene pela ação da transcriptase reversa codificada para a transposição, originando os denominados retrogenes.(45) Em eucariotos superiores, é possível que os

^b Na duplicação gênica, enquanto uma das cópias preserva a funcionalidade pré-existente, a outra pode divergir no aspecto regulatório ou codificante, possibilitando o surgimento de novas funções.(41, 45)

genes sem íntrons dos receptores associados à proteína G tenham sua origem em processos desta natureza, o que sugere que a diversidade verificada no sistema nervoso dos vertebrados tenha sido atingida por meio de trajetórias evolutivas influenciadas por TEs.(37, 46)

Transposições podem acarretar também efeitos deletérios ao organismo hospedeiro.(5, 14, 16) Kazazian et al. (47) constataram que a hemofilia A em humano está associada a inserções *de novo* do retrotransposon L1 no éxon 14 do gene que codifica o fator coagulante VIII. Outro estudo, realizado por Garvey et al. (48), mostrou que a distrofia muscular em ratos teria sua origem em uma mutação advinda de eventos que incluem a inserção do elemento L1.

A relação entre TEs e câncer também tem sido apontada na literatura.(9, 49) Como exemplo, cita-se o gene APC, considerado um supressor do câncer colorretal: Miki et al. (50) reportaram a disrupção do mesmo, devido à inserção somática do retrotransposon L1 no último éxon, em tecido tumoral. Este fora o primeiro caso de disrupção da sequência codificante de um gene supressor de tumor, mediada por TE, reportado na literatura.(50) Além disso, existem indícios do aumento da expressão do elemento L1 em tumores; no entanto, ainda não foi compreendido se o ambiente hipometilado verificado em células tumorais, ou ainda a diminuição da quantidade de microRNAs nestas células - conduzindo à redução do silenciamento dos TEs por interferência de RNA - são fatores responsáveis pela elevação da atividade destes elementos, ou se os TEs agem como promotores da gênese de tumores.(9, 49)

Os exemplos ora mencionados ilustram o potencial dos TEs no âmbito de mutações genéticas, desencadeadas pelos mais variados rearranjos decorrentes da transposição destes elementos. Segundo Volff (14), as evidências identificadas em diversos organismos indicam que os TEs podem atuar como reservatórios de sequências para inovação genética, tanto no aspecto regulatório como codificante, o que confere aos mesmos um papel importante na evolução. Ponto de vista semelhante é sustentado por outros autores na literatura.(5, 10, 15, 16, 42)

2.2 Organismos modelo

As algas verdes *Volvox carteri* e *Chlamydomonas reinhardtii*, pertencentes ao grupo volvocine algae, constituem modelos importantes para o estudo de questões relevantes na Biologia, a exemplo da divisão celular assimétrica, diferenciação celular, morfogênese embrionária, evolução da multicelularidade com divisão de trabalho (dicotomia germinativa-somática) e evolução da reprodução sexuada.(27–30) Além disso, *C. reinhardtii* é apontada ainda como um modelo adequado para o estudo do aparato fotossintético.(51)

Estas espécies são indicadas como adequadas à pesquisa por serem excelentes modelos

experimentais, por apresentarem biologia simples quando comparadas a organismos mais complexos e pela divergência relativamente recente (Figura 2.5), no contexto de estudo da evolução da multicelularidade, estimada em trabalhos na literatura como ocorrida a 75-230 milhões de anos.(29, 52, 53) De fato, o sequenciamento do genoma destas algas (26, 27) proporcionou um cenário adequado para o estudo dos aspectos moleculares correlatos às questões biológicas apontadas, bem como para a investigação da contribuição de TEs na arquitetura dos genes ortólogos em cada espécie.

Figura 2.5 – Representação da ocorrência de unicelularidade e multicelularidade em eucariotos. Nas classes representadas na árvore filogenética, a ocorrência exclusiva de unicelularidade é representada pela letra U e a ocorrência exclusiva de multicelularidade/colônia pela letra M. Classes em que as duas formas são observadas são representadas por ambas as letras; casos em que a unicelularidade é predominante, com poucas espécies multicelulares, são indicados pela letra U em tamanho maior. Fonte: HALLMANN(28)

A alga *C. reinhardtii* (Figura 2.6-A) é constituída por uma única célula biflagelada.(26, 28) A função dos flagelos relaciona-se à motilidade da alga e à transdução sensorial.(54) *V. carteri* (Figura 2.6-B), ao contrário, é um exemplo claro da dicotomia de células germinativas e somáticas.(28) Esta espécie é constituída por aproximadamente 2000 células biflageladas somáticas - inseridas em uma superfície esférica transparente composta por matriz extracelular (ECM) rica em glicoproteínas - e aproximadamente 16 células germinativas de tamanho maior, denominadas gonídios.(27, 28) Células gonídio são especializadas no crescimento e na reprodução; as somáticas, por sua vez, são especializadas na biossíntese da ECM e na motilidade, mas incapazes de se reproduzir.(28)

Figura 2.6 – Organismos modelo. A) *C. reinhardtii*: imagem relativa à fase somática, em que é possível a visualização dos flagelos (f) associados à motilidade. B) *V. carteri*: constituída por dois tipos de célula, somática (s) e germinativa (g). Fonte: PROCHNIK et al.(27)

Ambas as espécies, em condições normais de desenvolvimento, são haploides e reproduzem-se de forma assexuada; no entanto, em condições adversas (a exemplo da privação de nitrogênio no caso de *C. reinhardtii* ou choque térmico em *V. carteri*), passam à reprodução sexuada, apontada por pesquisadores como sendo uma forma de resposta emergencial do grupo volvocine algae.(28, 30) Outra característica comum a estas espécies remete ao fato de que o ciclo de vida com reprodução assexuada não difere para organismos *mating type* + (mt+) ou *mating type* - (mt-).(28) O ciclo sexuada origina zigotos diploides dormentes que podem resistir a condições adversas por um longo período de tempo.(28)

C. reinhardtii O ciclo de vida assexuada para a espécie *C. reinhardtii*, ilustrado nas figuras 2.7-A e 2.7-B, pode ser compreendido como um programa sequencial definido pelas fases somática e reprodutiva. A primeira fase é caracterizada pela presença dos flagelos; na sequência, estes são reabsorvidos pela célula, em preparação à fase de divisão celular que dará origem aos novos indivíduos.(28, 54) Na divisão, a célula procede ao denominado

modo de fissões múltiplas, em que ocorrem rodadas sequenciais de replicação de DNA e mitose, produzindo 2^n descendentes assexuados, sendo n o número de rodadas.(28, 54) Para o grupo volvocine algae, n varia de 2 a 15 dependendo da espécie, geralmente assumindo os valores 2 ou 3 para *C. reinhardtii*.(28) A nova geração de indivíduos é liberada quando apresenta os flagelos desenvolvidos.(28)

Figura 2.7 – Ciclo de vida da espécie *C. reinhardtii*. Sob condições favoráveis ao desenvolvimento, *C. reinhardtii* reproduz-se de forma assexuada para ambos os organismos *mating type* + (A) e *mating type* - (B); em condições adversas, passa à reprodução sexuada (C). Fonte: HALLMANN(28)

Quando alterna para o modo de reprodução sexuada, os organismos unicelulares *mt+* e *mt-* sofrem diferenciação celular em gametas, que podem aglutinar, passando à fusão, originando o zigoto que, após maturação, transforma-se em um zigósporo dormente (Figura 2.7-C).(28) Na germinação do esporo, ocorre a meiose, originando quatro indivíduos haploides aptos ao ciclo de vida assexuado.(28)

V. carteri Em *V. carteri*, a ocorrência simultânea de células somáticas e germinativas conduz à divisão do trabalho, de forma que as funções somática e reprodutiva ocorrem em paralelo, ao contrário do verificado em *C. reinhardtii*.(28, 54) No ciclo de vida assexuado desta espécie (Figura 2.8), cada célula gonídio passa à fase de clivagem, em que ocorrem sucessivas divisões celulares: primeiramente ocorrem 5 divisões simétricas, que originam 32 células idênticas, dentre as quais 16 passam à divisão assimétrica, em que cada célula origina uma célula maior - denominada gonídio inicial - e uma célula menor - denominada somática inicial.(28) As células gonídio iniciais realizam ainda duas rodadas de divisão

cuja expressão nas células maiores – reprodutivas – acaba por reprimir o desenvolvimento somático; *somatic regenerator A* (regA), cujo produto age nas células menores reprimindo o desenvolvimento reprodutivo, originando células somáticas.(28–30) É provável que a ativação específica de cada gene esteja relacionada com o tamanho da célula; no entanto, a via que media este efeito ainda não é conhecida.(28)

Figura 2.9 – Modelo do programa genético de diferenciação celular em *V. carteri*. Representação esquemática do programa de diferenciação celular que envolve os genes *glsA*, *lag* e *regA*. A atuação do primeiro tem como efeito a divisão assimétrica das células na fase de clivagem, originando células gonídio e somática iniciais. Nas células gonídio iniciais (maiores) atua o gene *lag*, que reprime o desenvolvimento somático; nas células somática iniciais (menores), atua o gene *regA* que reprime o desenvolvimento reprodutivo. Fonte: HALLMANN(28)

Quando expostos à condições adversas ao desenvolvimento, organismos da espécie *V. carteri* produzem uma glicoproteína que induz a reprodução sexuada (Figura 2.10).(28, 29)

As células gonídio em indivíduos adultos passam ao processo de divisão celular assimétrica que originará, após a diferenciação celular, organismos femininos com ~ 32 ovos e ~ 2000 células somáticas, e organismos masculinos com 128 pacotes de esperma e 128 células somáticas.(28) O indivíduo masculino libera os pacotes de esperma biflagelados, que por sua vez nadam até contatar uma fêmea, liberando o esperma que permeia a ECM, encontrando os ovos para fertilização.(28) O zigoto resultante será diploide e ficará latente até que condições favoráveis sejam restabelecidas, quando passará à meiose, originando três pequenos corpúsculos não viáveis e uma única célula germinativa viável haploide feminina ou masculina.(28)

Figura 2.10 – Reprodução sexuada na espécie *V. carteri*. Condições adversas ao desenvolvimento desta espécie induzem a produção e liberação do feromônio, conduzindo à reprodução assexuada. Fonte: NISHII;MILLER(29)

Sequenciamento dos genomas

De acordo com o exposto no início desta seção, o sequenciamento dos genomas de *C. reinhardtii* e *V. carteri* proporcionou a oportunidade para o estudo molecular de questões biológicas importantes relativas a estes modelos. Os resultados obtidos no trabalho de Prochnik et al. (27) sugerem que, quando comparada a outras espécies, *V. carteri* apresenta limitadas inovações em termos de proteínas. Segundo os autores, apenas 32 modelos gênicos comprovados por ESTs (do inglês, *expressed sequence tag*) não tiveram homólogos detectados em *C. reinhardtii* ou outras espécies. Além disso, o número de genes codificadores de proteínas, bem como a porcentagem do genoma atribuída a regiões codificantes, é semelhante nas duas algas (Tabela 2.1).(27) A maioria das famílias de proteínas relacionadas a processos de desenvolvimento (a exemplo de sinalização, excreção

e organização do citoesqueleto) mostraram-se similares nas duas espécies, em termos de estrutura e número.(27, 29)

Tabela 2.1 – Dados relativos ao genoma e aos modelos gênicos para as espécies *V. carteri* e *C. reinhardtii*.

	<i>V. carteri</i>	<i>C. reinhardtii</i>
Extensão do genoma (Mpb)	138	118
Número de cromossomos	14*	17
Número de genes codificadores de proteínas	14.520	14.516
% codificante no genoma	18,0	16,3
Número médio de éxons	7,78	8,42
Extensão média dos éxons (pb)	194	232
Número médio de íntrons	7,05	7,4
Extensão média dos íntrons (pb)	358	174
% modelos gênicos com alinhamento EST	34	51

Fonte: PROCHNIK et al.(27).

* O genoma de *V. carteri* encontra-se dividido em 434 *scaffolds* na versão utilizada neste trabalho, disponível em: <ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Vcarteri/>.

Em relação à multicelularidade no grupo volvocine algae, postula-se que dois aspectos tenham sido especialmente importantes para a sua evolução: o primeiro relaciona-se à constituição das paredes celulares que permitiu o arranjo para compartilhamento de matriz extracelular; o segundo remete à questão da diferenciação celular com divisão de trabalho.(55)

Quanto ao primeiro aspecto, acredita-se que a evolução seja resultante das diferenças nas famílias de proteínas pherophorin e VMP (*Volvox* metaloprotease) envolvidas na biossíntese da ECM e na sinalização sexual: enquanto *C. reinhardtii* possui 27 pherophorins e 8 VMPs, *V. carteri* possui 49 e 42, respectivamente.(27) Segundo Nishii e Miller (29), estes dados sugerem que a transformação da parede celular em ECM e subsequente expansão desta matriz, bem como a evolução da sinalização sexual em *V. carteri*, possam estar associadas ao processo de amplificação e diversificação dos genes pherophorin e VMP.

O segundo aspecto relaciona-se à divisão e diferenciação celular verificada em *V. carteri*. Além dos genes *glsA*, *lag* e *regA*, é possível que diferenças na família de proteína ciclina D1 tenham contribuído para a divergência entre os programas de divisão celular nos dois organismos: além das três ciclinas D existentes em *C. reinhardtii* (*Cycd2*, *Cycd3* e *Cycd4*), *V. carteri* possui ainda quatro proteínas relacionadas à ciclina D1 (*Cycd1.1* a *Cycd1.4*), enquanto *C. reinhardtii* possui apenas uma.(27, 29) Ciclinas D associam-se a proteínas quinases dependentes de ciclina e as direcionam à fosforilação de proteínas retinoblastoma, associadas ao controle da divisão celular.(27) Prochnik *et. al* (27) sugerem que a expansão da família ciclina D, em *V. carteri*, possa ter contribuído para a regulação do programa de

divisão celular nesta espécie.

A disponibilidade dos genomas, assim como dos modelos gênicos relativos às duas espécies, oferece uma importante oportunidade para o estudo da influência dos TEs nos genomas destas algas pertencentes ao grupo volvocine algae. Os capítulos seguintes deste trabalho relatam métodos computacionais, bem como resultados advindos destes, estabelecidos com o intuito de identificar possíveis contribuições de elementos de transposição na arquitetura dos genes ortólogos em *C. reinhardtii* e *V. carteri*.

Capítulo 3

Genes ortólogos e associação destes com TEs

Análises que identificam a região de inserção de TEs no genoma são de grande valia para o entendimento de mecanismos pelos quais estes elementos podem influenciar a regulação e estrutura dos genes de um organismo. De acordo com o exposto no Capítulo 2, inserções intergênicas podem introduzir novas sequências promotoras ou regulatórias, ou ainda causar danos às estruturas desta natureza nativas da região de integração da nova cópia. No âmbito intragênico, as inserções podem acarretar a inversão do sentido de transcrição do gene, introduzir sítios alternativos de *splicing* ou de poliadenilação, dar origem a um novo éxon ou ainda silenciar o gene. O conhecimento acerca do potencial de TEs em eventos evolutivos, juntamente com o aumento do número de genomas sequenciados, tem incentivado o estudo da associação entre elementos de transposição e estruturas gênicas.(11, 17, 18, 25)

Quando genomas apresentam divergência recente, é esperado que características relativas ao conteúdo gênico dos mesmos - como número de genes e estrutura destes, a exemplo da organização de éxons e íntrons - permaneçam altamente conservadas.(56) Este fato é importante para estudos que almejam identificar modificações introduzidas por TEs, uma vez que estabelece um cenário adequado em que podem ser realizadas comparações de contextos gênicos. As espécies *C. reinhardtii* e *V. carteri* se enquadram nesta conjectura devido à divergência relativamente recente na escala evolutiva (~ 75 -230 milhões de anos (52, 53)).

Neste capítulo são apresentadas abordagens de análise, bem como resultados advindos destas, que objetivam a identificação da associação entre TEs e genes ortólogos nas espécies *C. reinhardtii* e *V. carteri*. Neste contexto, a primeira seção aborda o mapeamento das cópias de elementos de transposição nos genomas destas algas; a seção seguinte destina-se à definição das regiões gênicas pertinentes aos genes ortólogos e ao reconhecimento das cópias de TEs associadas às mesmas.

3.1 Mapeamento das inserções de TEs nos genomas

3.1.1 Introdução

A comparação de sequências, possível a partir de ferramentas que realizam o alinhamento das mesmas, é essencial para estudos na área de bioinformática. Uma das ferramentas mais utilizadas em buscas por alinhamentos é o BLAST (*Basic Local Alignment Search Tool*).⁽⁵⁷⁾ Seu processamento tem início com a busca por alinhamentos locais com comprimentos equivalentes ao estabelecido pelo parâmetro *word size* e que apresentem uma pontuação mínima (T).⁽⁵⁸⁾ Ao identificar uma região alinhada que atenda a estes requisitos, o algoritmo procede à extensão do alinhamento em ambas as direções, considerando o *score* cumulativo resultante da soma dos *scores* obtidos em cada posição, de acordo com a matriz de *score* empregada na busca (no caso de nucleotídeos, considera-se apenas ganho de pontuação para pareamentos e penalizações para não-pareamentos); são incluídas ainda penalidades associadas à abertura de lacunas (*gaps*), que representam inserções ou deleções.⁽⁵⁸⁾ Desta forma, é estabelecido um mecanismo de pontuação que permite o delineamento do alinhamento que maximize a similaridade entre as duas sequências. Esta região alinhada é denominada HSP (*high-scoring segment pair*).

O parâmetro *e-evaluate* (E - Equação 3.1), associado ao HSP, reporta o número de alinhamentos - com *score* maior ou igual ao verificado no resultado reportado - esperados meramente ao acaso.^(58, 59) Este é estimado em função do tamanho do espaço de busca (definido pelo produto mN , sendo m o número de resíduos da sequência submetida à busca e N o número de resíduos que compõem o banco consultado), dos parâmetros k e λ (inerentes ao modelo estatístico e ao sistema de pontuação utilizado) e do *score* (S) obtido pelo HSP. Quanto menor o valor do *e-evaluate*, mais significativo é o alinhamento.⁽⁵⁸⁾

$$E = kmNe^{-\lambda S} \quad (3.1)$$

Além do *e-evaluate*, também o parâmetro *bit score* (S' - Equação 3.2) é reportado com o HSP.⁽⁶⁰⁾ Este representa o *score* normalizado do alinhamento, segundo os parâmetros k e λ . Esta normalização, em relação ao sistema de pontuação utilizado, permite a comparação de *scores* de alinhamentos advindos de diferentes buscas.

$$S' = (\lambda S - \ln k) / \ln 2 \quad (3.2)$$

Sendo o DNA uma molécula de dupla fita, estruturas gênicas podem existir na fita positiva (5'-3') ou na fita negativa (3'-5'). Por este motivo, ao efetuar a busca por uma sequência, o BLASTn (*nucleotide BLAST*) examina ambas as fitas, mais especificamente,

a *query* e seu complemento reverso.⁽⁵⁹⁾ A convenção adotada pelo NCBI-BLASTn para a descrição de alinhamentos identificados com a sequência complementar reversa da *query* é a apresentação das coordenadas de início e término no banco de dados invertidas ($s_start > s_end$).⁽⁵⁹⁾ Dessa forma, se a sequência existente no banco de dados representa a fita positiva de uma molécula de DNA, um alinhamento contra o mesmo descrito com coordenadas invertidas do banco indica identificação da *query* na fita negativa do DNA.

Os métodos de análise descritos neste capítulo resultam da conjugação do BLAST^a com protocolos definidos segundo os objetivos do estudo. Estes foram implementados utilizando-se a linguagem Python (61, 62), bem como as bibliotecas BioPython (63) e Matplotlib (64).

3.1.2 Métodos

Identificação de famílias de TEs

Neste estudo foram utilizadas bibliotecas de TEs descritos para os genomas de *C. reinhardtii* e *V. carteri*, disponíveis no banco de dados de acesso público Repbase (31), um banco de elementos repetitivos eucarióticos. Uma biblioteca de TEs pode conter sequências consenso^b de famílias de TEs, bem como instâncias de cópias de um elemento específico.⁽³⁴⁾ Elementos LTR encontram-se depositados neste banco de forma particionada, ou seja, a parte interna da sequência é depositada separadamente da sequência LTR. Em função deste fato, um processamento inicial de junção destas sequências foi realizado para obtenção da sequência de uma cópia integral.

Com o objetivo de eliminar sequências consenso de TEs redundantes entre si, cada família teve sua sequência submetida à busca em um banco de dados contendo todas as demais famílias, via o programa BLASTn. Sequências que apresentaram alinhamento com porcentagem de identidade e comprimento superiores a 80% e 100 pb, respectivamente, e *e-value* inferior a 10^{-10} , foram agrupadas em uma mesma família. Após o agrupamento, foi selecionado como representante de cada família a sequência consenso com maior quadro de leitura aberto, constituindo assim o conjunto final de sequências correspondentes às famílias de TEs participantes do estudo.

^a A versão local do BLAST (versão 2.2.27+) está disponível em <www.ncbi.nlm.nih.gov/books/NBK52640>.

^b Sequências consenso representam a melhor aproximação disponível dos TEs ativos que originaram as cópias observadas nos genomas atualmente.⁽³⁴⁾

Mapeamento de TEs no genoma

A identificação de TEs nos genomas apresenta dois objetivos principais: o primeiro antecede o processo de predição dos genes, em que as cópias de TEs devem ser identificadas e mascaradas para evitar predições gênicas espúrias (34, 65); após a etapa de anotação do genoma, a identificação de TEs é importante para o estudo biológico destes elementos, assim como para estudos correlatos à associação destes com estruturas gênicas específicas. Os softwares mais utilizados para detecção de cópias de TEs nos genomas são o Censor e o RepeatMasker, cujas versões mais recentes utilizam o WU-BLAST ou o NCBI-BLAST como ferramenta básica para a busca por alinhamentos.(6, 34)

Grandes inserções ou deleções podem ocorrer na sequência dos TEs após a integração no genoma, o que pode originar dois alinhamentos para uma sequência de busca, ao invés de um único alinhamento com um grande *gap*.(34, 66) Além disso, sequências de TEs inseridas em regiões do genoma hospedeiro de baixa conservação são mais suscetíveis ao acúmulo de mutações, o que também pode conduzir à detecção fragmentada do elemento originalmente inserido. Por este motivo, um pós-processamento denominado desfragmentação faz-se necessário para a consideração conjunta de alinhamentos neste contexto, de forma que os mesmos representem um único evento de inserção com significado biológico.(34) Esta consideração é necessária também para evitar a superestimação do número de cópias no genoma.

A estratégia de anotação das cópias de TEs adotada neste trabalho (Figura 3.1) é semelhante à empregada pelo software Censor, em que a saída do BLAST é processada para remoção de sobreposições e desfragmentação de sequências repetitivas detectadas.(66) Primeiramente, as sequências de TEs são interrogadas contra o genoma, via execução do BLASTn, sendo considerados neste estudo somente alinhamentos com *e-value* inferior a 10^{-10} . A segunda etapa ocupa-se do pós-processamento dos alinhamentos gerados na primeira, sendo aplicados dois *scripts* em série: o primeiro objetiva a desfragmentação das cópias; o segundo verifica possíveis sobreposições das cópias desfragmentadas no genoma. Eventos de sobreposição podem ocorrer porque a expansão de famílias de TEs no genoma, decorrente de eventos mutacionais, estabelece um cenário em que o limite para a distinção entre as mesmas é bastante tênue.

O processamento da segunda etapa do método de anotação de TEs (Figura 3.1) tem início com a separação dos alinhamentos segundo a fita do DNA - senso ou antissenso - em que foram identificados. Em seguida, estes são ordenados, para cada família de TE em cada cromossomo ou *scaffold*, segundo ordem crescente de início no genoma no caso de alinhamentos na fita senso, e em ordem decrescente de início para alinhamentos na fita antissenso; caso ocorra sobreposição da coordenada de início, a função de ordenamento considera ainda ordem crescente de *e-value*.

Figura 3.1 – Passos do protocolo adotado para o mapeamento de cópias de TEs no genoma.

Na primeira etapa as sequências de TEs são submetidas à busca por alinhamentos no genoma, via execução do BLASTn; na segunda etapa são aplicados os *scripts* de desfragmentação das cópias de TEs e de remoção de possíveis sobreposições destas no genoma. Fonte: Elaborada pela autora.

O próximo passo remete à execução do *script* de desfragmentação, cuja ideia principal consiste em analisar se dois alinhamentos sucessivos de um TE (Figura 3.2) podem ser considerados pertencentes à mesma inserção. Para tanto, são verificados três aspectos relativos aos alinhamentos analisados:

- i* A distância entre o término do alinhamento i e o início do alinhamento $i+1$ no genoma (*subject*) deve ser menor do que 100 pb.
- ii* Estando os alinhamentos i e $i+1$ a uma distância coerente com a especificada para o compito de uma única inserção, verifica-se se há ocorrência de sobreposição entre os mesmos no genoma. Caso ocorra e a sobreposição esteja dentro de um limite máximo de 50 pb, o alinhamento $i+1$ é considerado pertencente à inserção; caso contrário, o

alinhamento $i+1$ substituirá o alinhamento i se duas condições forem satisfeitas: o alinhamento $i+1$ deve possuir menor *e-value* e o alinhamento i deve estar presente na inserção sem que tenha substituído algum alinhamento anterior. As mesmas condições são válidas para casos de sobreposição em que um dos alinhamentos esteja completamente sobreposto, mesmo sendo esta sobreposição menor do que o limite aceito.

- iii O início do alinhamento $i+1$, no referencial definido pelas coordenadas da sequência do TE (*query*), deve ser maior do que o término do alinhamento i , salvo a sobreposição aceita com extensão máxima de 50 pb.

Figura 3.2 – Representação esquemática de alinhamentos no genoma. São ilustrados alinhamentos sucessivos na fita positiva (5'-3') (A) e na fita negativa (3'-5') (B) do genoma. Para os exemplos em cada fita, se os alinhamentos i e $i + 1$ atenderem os requisitos especificados no *script* de desfragmentação, podem ser considerados representantes de um único evento de integração de TE. Fonte: Elaborada pela autora.

A execução do *script* de desfragmentação origina dois conjuntos de cópias de TEs, identificados na fita positiva e negativa. A partir deste ponto, as cópias são descritas obedecendo o formato GFF3^c, em que a coordenada de início é sempre menor do que a coordenada de término. A fita em que a inserção fora anotada é identificada pelo atributo *strand*, designado por positivo ou negativo de acordo com a fita. Além disso, o *score* normalizado da cópia desfragmentada é definido a partir dos *scores* normalizados dos alinhamentos que a compõem: para cada alinhamento, recupera-se o *score* não normalizado (S) considerando-se os parâmetros *lambda* e *k* e a relação expressa pela Equação 3.2; a soma destes fornece o *score* não normalizado da inserção que, por sua vez, é submetido à normalização.

Finalmente, as cópias resultantes da desfragmentação identificadas em ambas as fitas (positiva e negativa) são concatenadas em um único conjunto. Na sequência, são

^c Generic Feature Format Version 3. Especificações podem ser verificadas em: <<http://www.sequenceontology.org/gff3.shtml>>.

submetidas ao ordenamento em cada cromossomo ou *scaffold* segundo ordem crescente de coordenada de início no genoma; ocorrendo sobreposição da posição inicial, a função de ordenamento considera também ordem decrescente de pontuação. O conjunto de cópias ordenadas constitui a entrada para o *script* que avalia possíveis sobreposições das inserções no genoma. A identificação de sobreposição conduz à análise do seguinte critério (Figura 3.3): se a região sobreposta não exceder 50 pb, ambas as cópias são consideradas; caso contrário, a inserção $i+1$ substituirá a inserção i se possuir maior pontuação e se a inserção i não tiver ocasionado substituição anterior.

Figura 3.3 – Representação de sobreposição de inserções no genoma. Se a sobreposição estiver acima do limite aceito, a cópia $i+1$ substituirá a cópia i se os requisitos especificados no *script* correlato à análise de sobreposição forem satisfeitos. Fonte: Elaborada pela autora.

É importante ressaltar que um dos objetivos da desfragmentação das cópias é evitar a superestimação do número de inserções de TEs no genoma. Dessa forma, estudos envolvendo análise de frequência de ocorrência consideram as cópias desfragmentadas; em contrapartida, em estudos envolvendo a análise do conteúdo da sequência do TE introduzido em estruturas gênicas, os alinhamentos constituintes da inserção são recuperados para que seja apurada somente a região realmente atribuída ao elemento de transposição, e não sequências genômicas intervenientes.

3.1.3 Resultados

Identificação de famílias de TEs

O emprego do método exposto inicialmente na Seção 3.1.2 para os TEs descritos para *C. reinhardtii* resultou no conjunto de famílias apresentado na Tabela A.1, enquanto que a aplicação do mesmo aos elementos descritos para *V. carteri* resultou no conjunto apresentado na Tabela A.2, ambas localizadas no Apêndice A. Nestas tabelas encontram-se a identificação, a descrição e o comprimento das sequências das famílias de TEs que foram interrogadas junto aos genomas analisados, constituindo, portanto, a base para a obtenção dos resultados expressos a seguir.

Mapeamento de TEs nos genomas dos organismos modelo

A aplicação do método de mapeamento de TEs aos genomas de *C. reinhardtii*^d (26) e *V. carteri*^e (27), obtidos na base de dados Phytozome (67), permitiu a inferência da frequência e da representatividade destes elementos nos genomas destas espécies (Tabela 3.1), bem como da distribuição de frequências de ocorrência das classes de TEs (Tabela 3.2). Resultados relativos à frequência de cada família encontram-se disponíveis nas tabelas A.1 e A.2.

Tabela 3.1 – Resultados do mapeamento de TEs nos genomas de *C. reinhardtii* e *V. carteri*.

Espécie	Extensão do genoma (Mpb)	Número de cópias	Representatividade no genoma (%)
<i>C. reinhardtii</i>	111	12.402	5,92
<i>V. carteri</i>	131	11.686	6,00

Fonte: Elaborada pela autora.

Dados relativos ao mapeamento de sequências de TEs descritas para cada espécie, depositadas no Repbase. A extensão do genoma ora apresentada refere-se à versão do mesmo utilizada no trabalho, depositada no Phytozome.

Tabela 3.2 – Número de cópias anotadas para cada classe de TE nos genomas de *C. reinhardtii* e *V. carteri*.

Classe	Número de cópias	% do número total de cópias no genoma	Representatividade no genoma (%)
<i>C. reinhardtii</i>			
LTR	1.086	8,76	0,66
non-LTR	7.184	57,92	3,72
DNA	3.789	30,55	1,45
Outras	343	2,77	0,09
Total	12.402	100,00	5,92
<i>V. carteri</i>			
LTR	4.761	40,74	2,23
non-LTR	6.234	53,35	3,17
DNA	243	2,08	0,10
Outras	448	3,83	0,50
Total	11.686	100,00	6,00

Fonte: Elaborada pela autora.

^d Disponível em <ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/early_release/Creinhardtii_v5.3/>.

^e Disponível em <ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Vcarteri/>.

Chlamydomonas reinhardtii O mapeamento de TEs realizado no genoma deste organismo resultou na detecção de 12.402 cópias (Tabela 3.1), sendo a classe mais abundante a non-LTR, com 57,92% das inserções, seguida pela DNA-transposon, com 30,55% (Tabela 3.2). Esta análise indicou ainda que aproximadamente 5,92% do genoma de *C. reinhardtii* é constituído por resíduos das famílias de TEs analisadas (Tabela 3.1).

A Figura 3.4 exibe a frequência de ocorrência de famílias com número de cópias superior a 100; dados de todas as famílias estão disponíveis na Tabela A.1, no Apêndice A. Dentre as famílias analisadas, a de maior ocorrência no genoma deste organismo é a L1-3_CR (1.196 cópias), pertencente à classe non-LTR. A segunda mais frequente é a L1-1_CR (998 cópias), também non-LTR.

Figura 3.4 – Frequência de ocorrência absoluta das famílias de TEs no genoma da espécie *C. reinhardtii*. Apresentam-se, nesta figura, famílias com frequência superior a 100. Fonte: Elaborada pela autora.

Foram identificadas também as famílias de TEs que possuem cópias íntegras no genoma analisado, bem como o número de ocorrências destas (Figura 3.5). As cópias íntegras são identificadas por meio da seleção de alinhamentos com identidade mínima de 99%

e desvio percentual máximo em relação ao comprimento de 1%, quando comparado à sequência consenso representante da família de TE. É interessante notar que as famílias mais representativas neste genoma, L1-3_CR e L1-1_CR, não apresentam ocorrências íntegras.

Figura 3.5 – Distribuição de frequências de ocorrência das famílias de TEs que apresentam cópias íntegras no genoma de *C. reinhardtii*. Considera-se como íntegra a cópia que, comparada à sequência consenso da família de TE, apresenta identidade mínima de 99% e desvio percentual máximo de 1% em relação ao comprimento. Fonte: Elaborada pela autora.

Dentre as 12.402 cópias identificadas nesta espécie, 11.803 (95,17%) foram designadas a partir de um único alinhamento entre a sequência consenso e o genoma, 562 (4,53%) pela junção de dois alinhamentos e 37 por mais de três alinhamentos (0,30%). O número máximo de alinhamentos atribuídos a uma mesma cópia foi 5, tendo sido delineados dois eventos com esta estrutura.

Volvox carteri O mapeamento realizado neste genoma indicou a presença de 11.686 cópias de TEs (Tabela 3.1), dentre as quais 53,35% pertencem à classe non-LTR e 40,74%

à LTR (Tabela 3.2). Outro dado obtido nesta etapa refere-se à representatividade das famílias de TEs estudadas no genoma de *V. carteri*: aproximadamente 6,00% (Tabela 3.1).

Os valores de frequência de ocorrência absoluta de famílias com contagem superior a 100 cópias são expostos na Figura 3.6. Dados relativos a todas as famílias estão disponíveis na Tabela A.2, no Apêndice A. Neste genoma, dentre as famílias analisadas, a mais representativa é a Gypsy5-LTR_VC (1.148 cópias), seguida pela RTE5_VC (1.081 cópias); enquanto a primeira pertence à classe LTR, a segunda pertence à non-LTR.

Figura 3.6 – Frequência de ocorrência absoluta das famílias de TEs no genoma da espécie *V. carteri*. São apresentadas somente famílias com frequência superior a 100. Fonte: Elaborada pela autora.

A Figura 3.7 indica as famílias de TEs que possuem cópias íntegras no genoma de *V. carteri*, bem como o número de ocorrências destas. Neste genoma a família mais representativa, Gypsy5-LTR_VC, apresenta uma ocorrência íntegra de acordo com os critérios adotados; já a RTE5_VC, segunda mais frequente, possui duas ocorrências íntegras.

Dentre as 11.686 ocorrências de TEs mapeadas para *V. carteri*, 10.742 (91,92%) são

Figura 3.7 – Distribuição de frequências de ocorrência das famílias de TEs que apresentam cópias íntegras no genoma de *V. carteri*. Considera-se como íntegra a cópia que, comparada à sequência consenso da família de TE, apresenta identidade mínima de 99% e desvio percentual máximo de 1% em relação ao comprimento. Fonte: Elaborada pela autora.

constituídas por um alinhamento, 763 (6,53%) por dois e 181 (1,55%) por mais de três alinhamentos. Nesta espécie, o número máximo de alinhamentos atribuídos a uma mesma inserção foi 6, havendo apenas uma cópia com esta característica.

Breve comparativo Apesar de apresentar um número maior de inserções e um genoma mais compacto, *C. reinhardtii* apresentou uma representatividade de TEs muito próxima à verificada em *V. carteri* (~ 6% - Tabela 3.1). Tal fato deve-se à extensão das cópias (Figura 3.8) que, em média, é menor em *C. reinhardtii*.

Outro aspecto que chama a atenção no resultado do mapeamento de TEs é a modesta ocupação destes elementos em ambos os genomas. De fato, o estudo extensivo dos genomas de *C. reinhardtii* e *V. carteri*, realizado por Prochnik et al. (27), indicou uma

Figura 3.8 – Box-plot da distribuição de comprimentos das cópias de TEs nos genomas das duas espécies. Estatísticas para comprimentos são: *C. reinhardtii* - média = 529,55 ± 1028,84 pb e mediana = 201,0 pb (mínimo: 34 pb, máximo: 14.080 pb); *V. carteri* - média = 674,02 ± 1045,73 pb e mediana = 302,0 pb (mínimo: 35 pb, máximo: 11.520 pb). *Out-liers* não são mostrados. Fonte: Elaborada pela autora.

representatividade de 6,9% e 7,2% de elementos de transposição, valores relativamente próximos aos obtidos no mapeamento das famílias estudadas neste trabalho. Os índices de representatividade de TEs nos organismos modelo deste estudo assemelham-se aos verificados em espécies de levedura; como exemplo, cita-se a *C. neoformans* (~ 5%) e a *S. cerevisiae* (~ 4%).(3) A baixa representatividade de TEs nestas espécies torna-se mais perceptível quando comparada à representatividade verificada em vertebrados mamíferos (*H. sapiens* ~ 45%; *M. musculus* ~ 40%) e não-mamíferos (*G. gallus* ~ 10%; *D. rerio* ~ 26%), bem como a espécies de invertebrados (*D. melanogaster* ~ 18%; *C. elegans* ~ 12%).(3, 68) Citam-se ainda espécies de plantas, como *O. sativa* (~ 30%) e *Z. mays* (~ 65%) (3), que também possuem uma ocupação de TEs mais pronunciada em seus genomas.

3.2 Associação de TEs com genes ortólogos

3.2.1 Métodos

A análise informática de sequências genômicas possibilita o mapeamento em larga escala das cópias de TEs e a avaliação da posição das mesmas em relação aos genes. Assim sendo, torna-se possível precisar se a inserção ocorrera em região adjacente às fronteiras do gene, ou ainda em regiões codificantes ou intrônicas do mesmo.

Neste estudo o gene foi definido como a sequência genômica delimitada pelo sítio de início de tradução (SIT) e pelo final do códon de terminação (Figura 3.9), sendo estas coordenadas obtidas a partir da predição gênica realizada previamente nos dois organismos estudados^f. Esta convenção também foi adotada em estudo semelhante realizado no genoma de *Oryza sativa ssp. japonica* (18), sendo derivada do fato que regiões transcritas não traduzidas não são facilmente preditas em genomas pelos métodos computacionais, na ausência de sequências derivadas de transcritos.

Figura 3.9 – Representação esquemática da região gênica. São ilustrados o gene, considerado como a sequência genômica existente entre o SIT e o final do códon de terminação, e a região gênica, definida pelo gene e pelas regiões intergênicas adjacentes ao mesmo. Fonte: Elaborada pela autora.

As coordenadas dos genes, juntamente com as coordenadas das cópias das diferentes famílias de TEs nos genomas, constituem a entrada do método que visa detectar a associação entre estes elementos e estruturas gênicas. O primeiro passo consiste na verificação de ocorrências de sobreposição dos genes preditos, uma vez que a sobreposição destas estruturas são presumidas como raras em genomas eucarióticos.(56) Para tanto, os genes foram ordenados segundo ordem crescente de coordenada de início; em seguida, genes que se sobrepunham ao anterior foram desconsiderados do conjunto de dados. O passo seguinte ocupa-se da definição da região gênica correlata ao gene (Figura 3.9). A regra adotada para o estabelecimento das regiões flangeadoras considerou a distância genômica entre dois genes sucessivos: sendo esta maior ou igual a 4 kpb, define-se, para cada um dos gene, um flanco de 2 kpb; caso contrário, o flanco atribuído a cada gene remete à 50% da distância que os separa. Após este processamento, cada gene possui as delimitações da região flangeadora 5' e da região flangeadora 3'.

Uma vez definidas as regiões gênicas, o passo seguinte da análise considerou a busca por cópias de TEs associadas às mesmas. Nesta etapa as coordenadas derivadas do mapeamento de TEs foram comparadas às coordenadas das regiões gênicas para as duas espécies, buscando-se o reconhecimento do posicionamento dos elementos em relação às estruturas gênicas. Dessa forma, foram constatadas associações entre TEs e regiões flangeadoras 5', aos genes propriamente, às regiões flangeadoras 3' ou ainda possibilidades

^f As coordenadas relativas ao SIT e ao final do códon de terminação foram estabelecidas a partir das coordenadas genômicas de sequências codificantes dos genes (CDS), depositadas no repositório Phytozome sob o formato GFF3.

de extensão da cópia ao longo destes intervalos. Casos em que a cópia do TE se sobrepunha a praticamente toda a extensão da região gênica - 1^o flanco, gene e 2^o flanco - conduziram à desconsideração do gene, por suposto erro de predição do mesmo. Erros de predição gênica podem ocorrer em consequência de falhas no processo de identificação e mascaramento de TEs na fase de anotação de genomas.(65) Além disso, apesar do avanço em técnicas de predição de estruturas gênicas, estas continuam ainda muito suscetíveis a erros, o que é explicado também pela complexidade dos genomas eucariotos.(56, 69)

Cópias de TEs podem estar relacionadas a mais de uma região gênica; neste caso, o posicionamento das mesmas foi analisado em relação a cada contexto gênico associado. Outro procedimento adotado para cópias associadas a mais de um gene foi o estabelecido do gene mais correlato, definido como sendo o que possui maior extensão de sobreposição com a cópia. Este procedimento fora adotado, em estudos de distribuição de frequências, para evitar contagem dupla de elementos associados a mais de uma região gênica.

Definição de genes ortólogos

Devido à diferença notável entre o número de genes preditos nos genomas de *C. reinhardtii* and *V. carteri* (17.287 e 14.295 respectivamente, após remoção dos modelos dúbios), foram considerados nas análises apenas os genes para os quais foi possível a identificação de um ortólogo no outro genoma. Outro aspecto que conduz à adoção desta abordagem reside na questão biológica que motiva este trabalho: uma vez identificado um evento relevante em um determinado gene, torna-se mais simples verificar se o mesmo também ocorrera em seu ortólogo. Além disso, análises dirigidas a ortólogos minimizam a chance de que o gene avaliado seja artefato de predição, uma vez que estruturas muito semelhantes foram identificadas em dois genomas com processos de anotação distintos.

O termo ortólogo, introduzido por Walter Fitch em 1970, é empregado para designar genes em espécies distintas, derivados de um gene em um último ancestral comum às mesmas.(58, 70, 71) Ortólogos tendem a manter similaridade na estrutura de domínio, na estrutura tridimensional e na função biológica; em contraste, genes parálogos, originados em eventos de duplicação gênica, tendem a divergir da função original com o passar do tempo por estarem livres da seleção purificadora.(58, 71) Ambas as classificações pertencem à definição mais geral, denominada homologia, que por sua vez designa uma relação envolvendo ancestral comum, sem especificar o cenário evolutivo envolvido.(71)

Uma das abordagens mais utilizadas para a inferência de prováveis genes ortólogos em estudos de genômica comparativa é a BBHs (*bidirectional best hits*), em que são indicados como ortólogos os genes cujas sequências peptídicas apresentam o melhor alinhamento de forma simétrica ou recíproca.(70) Exemplificando, para que duas sequências peptídicas A e B sejam consideradas ortólogas, é necessário que A apresente seu melhor alinhamento

com B e que B, por sua vez, apresente seu melhor alinhamento com A. Esta abordagem é ancorada na consideração de que sequências peptídicas de genes ortólogos são mais similares entre si do que a sequências peptídicas de outros genes dos organismos comparados.

As sequências peptídicas dos genes de *C. reinhardtii* e *V. carteri* foram obtidas no repositório Phytozome. Para o estabelecimento dos pares de melhores alinhamentos bidirecionais, foram realizadas duas etapas de execução do BLASTp (*protein* BLAST), uma primeira em que o conjunto de sequências peptídicas de *C. reinhardtii* é interrogado contra o conjunto de sequências peptídicas de *V. carteri*, e uma segunda em que o conjunto de *V. carteri* é interrogado contra o conjunto de *C. reinhardtii*. Foram selecionados, em cada execução, apenas resultados correspondentes ao melhor alinhamento entre duas sequências (*best-hits*), com $e\text{-value} \leq 10^{-10}$, conduzindo à identificação dos pares de alinhamentos recíprocos. O emprego desta abordagem resultou em um conjunto de 8.625 ortólogos, identificados em cada espécie.

3.2.2 Resultados

Considerando o conjunto de genes ortólogos entre as espécies *C. reinhardtii* e *V. carteri*, foram analisadas as ocorrências de TEs relacionadas aos mesmos. Do total de cópias de elementos de transposição identificadas no genoma de *C. reinhardtii*, 4.158 (33,53%) encontram-se mapeadas em regiões correlatas a genes ortólogos; no caso de *V. carteri*, 3.033 cópias (25,95%) possuem esta relação.

Dentre os ortólogos em *C. reinhardtii*, 1.885 genes possuem associação com uma cópia de TE, 634 genes com duas cópias e 332 genes exibem esta relação com mais de duas cópias; em *V. carteri*, estes números são, sucessivamente, 1.580, 417 e 202. Assim sendo, 2.851 (33,06%) ortólogos na primeira espécie e 2.199 (25,5%) na segunda exibem presença de TEs em pelo menos uma das sub-localizações em estudo. Este dado, associado à distribuição de densidades de TEs nos genes ortólogos em cada espécie (Figura 3.10), sugere que os genes em *C. reinhardtii* tendem a acumular mais cópias quando comparado à *V. carteri*. A densidade de TEs em cada gene foi determinada pela razão entre o número de cópias e a extensão da região gênica.

A fração de genes ortólogos com TE na região flanqueadora 5' é bastante semelhante nas duas espécies (Figura 3.11): 8,08% em *C. reinhardtii* e 8,19% em *V. carteri*. Em relação às sequências codificantes, *V. carteri* apresenta uma maior porcentagem de genes com TEs nesta região: 1,19%; em *C. reinhardtii*, apenas 0,41% dos genes se enquadra nesta categoria. Cerca de 12,35% dos ortólogos em *C. reinhardtii* e 10,84% em *V. carteri* abrigam TEs em íntrons. Em relação à região intergênica adjacente à terminação 3' dos genes ortólogos, *C. reinhardtii* apresenta uma maior porcentagem de genes com presença

Figura 3.10 – Box-plot da distribuição de densidades de TEs em regiões gênicas dos ortólogos. Estatísticas para densidades são: *C. reinhardtii* - média = $8,42 \times 10^{-2} \pm 15,73 \times 10^{-2}$ cópias/kpb e mediana = 0 cópias/kpb (mínimo: 0 cópias/kpb, máximo: 1,76 cópias/kpb); *V. carteri* - média = $4,89 \times 10^{-2} \pm 10,96 \times 10^{-2}$ cópias/kpb e mediana = 0 cópias/kpb (mínimo: 0 cópias/kpb, máximo: 1,36 cópias/kpb). Participam desta análise 8.625 regiões gênicas em cada espécie. *Out-liers* não são mostrados. Fonte: Elaborada pela autora

de TEs neste segmento: 17,77%; em *V. carteri*, apenas 10,23% dos ortólogos apresentou cópias neste intervalo. Estes índices são modestos quando comparados ao observado no genoma humano, o qual apresenta 4% dos genes codificantes para proteínas abrigando TEs em sequências codificantes (11) e 85-90% abrigando TEs em íntrons.(40)

Figura 3.11 – Fração de genes ortólogos com presença de TE na região flangeadora 5' (5'), codificante (éxon), intrônica e flangeadora 3' (3'). Um total de 8.625 genes foi considerado em cada espécie. Fonte: Elaborada pela autora

A distribuição de frequências das classes de TEs, em cada sublocalização da região gênica, encontra-se representada na Figura 3.12. Esta análise considerou cópias associadas a um único gene e, dentre as associadas a mais de um gene, somente aquelas mais correlatas

aos ortólogos.

Figura 3.12 – Distribuição de frequências de ocorrência das classes de TEs em regiões gênicas dos ortólogos. Dentre as cópias mapeadas no genoma de cada espécie, 3.903 possuem relação com genes ortólogos em *C. reinhardtii* (A) e 2.978 em *V. carteri* (B). Fonte: Elaborada pela autora

No caso da espécie *C. reinhardtii* (Figura 3.12-A), observa-se a maior ocorrência da classe non-LTR em todas as regiões estudadas, seguida pela DNA-transposon. No caso de *V. carteri* (Figura 3.12-B), o padrão emergente é de maior ocorrência de elementos de transposição da classe non-LTR, seguida pela LTR. Em ambas as espécies, a distribuição de frequências das classes de TEs ora evidenciada reproduz o padrão observado na distribuição no genoma como um todo (Tabela 3.2). Ocorrências designadas por sobreposição constituem casos em que o mapeamento da cópia de TE se estende por mais de uma sublocalização da região gênica. Para *C. reinhardtii* foram verificadas 19 ocorrências desta natureza, enquanto para *V. carteri* 59. Devido ao fato de que estas cópias possuem relação com regiões preditas como codificantes, serão contempladas nesta categoria nas análises que seguem. A distribuição de TEs em cada região será discutida em maiores detalhes nas seções seguintes.

Definindo-se a densidade de cópias de TEs em cada sublocalização da região gênica (Figura 3.13), verifica-se que *C. reinhardtii* apresenta índices mais elevados para a densidade relativa à região flanqueadora 5', intrônica e flanqueadora 3' do gene, quando comparado à *V. carteri*; no entanto, na região codificante, a segunda espécie apresenta maior densidade de cópias de TEs. A densidade foi determinada pela razão entre o número de cópias de TEs e a extensão total de cada sublocalização, definida pela soma relativa à mesma em cada gene ortólogo.

Figura 3.13 – Densidade de cópias de TEs nas sub-localizações da região gênica de genes ortólogos. Dados relativos à região flangeadora 5' (5'), codificante (E), intrônica (I) e flangeadora 3' (3') em genes ortólogos nas espécies *C. reinhardtii* e *V. carteri*. Fonte: Elaborada pela autora

No decorrer deste trabalho análises mais detalhadas serão abordadas, buscando-se o estudo do padrão de distribuição de TEs em regiões gênicas específicas, bem como a detecção de possíveis influências destes elementos na arquitetura dos genes analisados.

Capítulo 4

Estudo de TEs em regiões intergênicas

O processo de transferência da informação armazenada no DNA, culminando na síntese proteica, tem o RNA mensageiro (mRNA) como molécula intermediária.(72) Para que a informação genética seja transcrita para o mRNA e este possa transportá-la de forma adequada aos ribossomos (Figura 4.1), é necessária a interação de complexos moleculares com estruturas sinalizadoras codificadas nas adjacências do gene.

Os fatores de transcrição II (TFIIA-J), juntamente com a RNA polimerase II, integram o complexo de início de transcrição (CIT).(73) Estes fatores ligam-se ao DNA por meio do reconhecimento de motivos característicos localizados nas imediações da extremidade 5' do gene.(36) Além disso, sequências regulatórias - não necessariamente próximas ao gene - constituem sítios de ligação para fatores de regulação que atuam no ajuste do nível de expressão gênica.(36, 73) De acordo com o exposto na Seção 2.1, a integração de TEs em fronteiras de estruturas gênicas podem introduzir novas sequências promotoras ou regulatórias, ou ainda causar a disrupção de estruturas desta natureza pré-existentes, alterando o padrão de expressão do gene.

Tendo iniciado a extensão do precursor de mRNA (pré-mRNA), a RNA polimerase II dissocia-se dos fatores de transcrição; em seguida, acumulam-se na cauda desta enzima componentes necessários aos mecanismos de maturação do transcrito primário, processo que ocorre concomitantemente à transcrição.(36) A primeira modificação consiste na adição de um resíduo guanina metilada à extremidade 5' do pré-mRNA, em um processo denominado capeamento.(36) A segunda alteração remete ao mecanismo de *splicing*, em que as sequências intrônicas são removidas do mRNA nascente.(74) Finalmente, a extremidade 3' do pré-mRNA é reconhecida por proteínas associadas ao processo de poliadenilação, em que a cauda poli(A) não codificada no genoma é polimerizada.(75) As extremidades modificadas do mRNA indicam à célula que a molécula encontra-se íntegra e apta ao transporte da informação para a síntese proteica.(75, 76)

Figura 4.1 – Dogma central da expressão gênica. Representação do processo de expressão gênica em eucariotos, em que o gene é transcrito do DNA para o pré-mRNA. O processamento do pré-mRNA, concomitante à síntese do mesmo, inclui as etapas de *splicing*, capeamento 5' e poliadenilação, e origina o mRNA. Esta molécula é exportada ao citoplasma, onde será traduzida pelos ribossomos, promovendo a síntese de uma proteína.(56) Embora não indicado na figura, no período que antecede a tradução o mRNA pode ser alvo de diversas vias de regulação pós-transcricional. Como exemplos, cita-se o silenciamento da tradução, mediada por interações com microRNAs (miRNAs), e o armazenamento do mRNA em um estado dormente, via deadenilação da cauda poli(A), no qual ficará latente até o momento em que deverá ser traduzido.(76, 77) Para se tornar novamente apto à tradução, o mRNA dormente é submetido à poliadenilação citoplasmática. Abreviaturas: SITr - sítio de início de transcrição; STT - sítio de término de transcrição. Fonte: ZHANG(56)

Ainda em relação ao mecanismo de poliadenilação, a extremidade 3' do mRNA é especificada por sinais de DNA codificados no genoma que, ao serem transcritos no pré-mRNA, são reconhecidos pelas proteínas correlatas a esta etapa de maturação.(76) Inserções de TEs no contexto genômico que armazena estes sinais podem alterar o padrão de poliadenilação do gene, conduzindo a formas alternativas do transcrito.

Neste trabalho a abordagem utilizada para o estudo de associações entre TEs e genes ortólogos considera o gene como sendo a sequência genômica delimitada pelo códon de iniciação e o códon de terminação (Seção 3.2.1). Deste modo, a região intergênica adjacente à extremidade 5' do gene é a potencial contenedora de estruturas promotoras e regulatórias associadas ao início do processo de transcrição, enquanto que a região intergênica adjacente à extremidade 3' abriga principalmente sequências sinalizadoras para o mecanismo de poliadenilação e para a terminação da transcrição. Por este motivo, ambas representam zonas de interesse para o estudo da distribuição de TEs, bem como de aspectos relativos à possíveis contribuições destes elementos nas estruturas de DNA envolvidas nos processos mencionados.

4.1 Região flanqueadora 5' dos genes ortólogos

4.1.1 Introdução

A iniciação da transcrição é um processo complexo que tem início com o remodelamento da cromatina, que passa a um estado mais relaxado, seguido pela constituição do CIT no promotor do gene.(56) O cerne do promotor (no inglês, *core promoter*) é classicamente definido como a região em torno do sítio de início de transcrição (SITr), com extensão aproximada de 200 pb, que comporta sequências às quais se ligam os fatores necessários para a formação do CIT.(56, 73)

Como exemplo, cita-se a arquitetura de promotor contendo o motivo *TATA-box* e sequências com elevado percentual AT que atuam como sítios ligantes para a proteína de ligação TATA (TBP).(73) Esta proteína agrega-se a fatores de associação específicos (TAFs), constituindo a subunidade TFIID - integrante do CIT.(73) Elementos frequentemente identificados no promotor, além do *TATA-box*, são o BRE (*TFIIB recognition element*) - existente em promotores com *TATA-box*, o Inr (*initiator element*) e o DPE (*downstream promoter element*), que atua em cooperação com o Inr (73) (Figura 4.2). Outros elementos menos frequentes são MTE (*motif ten element*), DCE (*downstream core elements*, subunidades 1, 2 e 3) e XCPE1 (*X core promoter element 1*).(73) A extensão do promotor completo pode atingir 1000 pb na região que antecede o SITr.(56)

Figura 4.2 – Unidades reguladoras da transcrição em eucariotos. Representação esquemática de sequências sinal envolvidas no processo de transcrição em eucariotos simples (A, levedura) e superiores (B, mamíferos); em C, representação do cerne do promotor em mamíferos. Éxons encontram-se representados por retângulos em cinza. A curva no DNA (B) representa a possibilidade de aproximação do *enhancer* em relação ao cerne do promotor ou a estruturas UPE. Abreviaturas: UAS (*upstream activating sequences*); UPE (*upstream promoter elements*) e DPE (*downstream promoter element*). Fonte: RIETHOVEN(73)

Além do promotor, existem sequências no DNA às quais se ligam elementos capazes de regular a expressão gênica. Os *enhancers* são motivos de DNA que, mesmo à distância do gene, podem recrutar fatores de transcrição que atuam elevando a taxa de formação e ligação do CIT ao promotor.(73, 78) Estas sequências podem ser encontradas em regiões que antecedem ou sucedem o SITr - incluindo éxons, íntrons e UTRs - em distâncias que podem chegar à 100.000 pb em humanos.(73) Os *silencers*, por sua vez, possuem ação inversa: atuam na regulação negativa da expressão gênica, devido ao recrutamento de fatores que interferem negativamente na formação ou na ligação do CIT ao SITr.(73) A classe de motivos denominada *insulator* engloba seqüências de DNA *cis* que interagem com fatores que atuam no bloqueio da ação de *enhancers* e *silencers*.(73)

Apesar do avanço em técnicas de reconhecimento de promotores e estruturas regulatórias, a acurácia destas ainda é considerada deficiente.(79, 80) Zeng et al. (79) indicam que três características são importantes para a localização do promotor: sequências sinal dos motivos regulatórios, o contexto genômico destas e a estrutura tridimensional do DNA. Os autores apontam ainda que é necessário o aprofundamento no estudo destes aspectos - em especial, dos dois últimos - para o aprimoramento de abordagens de detecção de estruturas regulatórias.

Grande parte dos preditores de promotores, dentre os quais citam-se os de maior relevância Eponine e FirstEF, destinam-se ao reconhecimento de estruturas relativas ao genoma humano (81), o que é explicado pelo estudo extensivo de motivos moleculares relativos a esta espécie nos últimos anos.(79) Este fato torna inadequada a aplicação dos mesmos a espécies não pertencentes ao grupo de metazoários, como é o caso de *C. reinhardtii* e *V. carteri*. O mesmo ocorre para preditores de sítios de ligação de fatores de transcrição, que exibem ainda o agravante de conservação moderada até mesmo em espécies relativamente próximas, a exemplo da humana e de camundongo (aproximadamente 50%).(81)

A contribuição de TEs no contexto regulatório, mediada principalmente pela inserção destes elementos na região não codificante 5' dos genes, é reconhecida na literatura.(5, 82, 83) Contribuições desta natureza devem-se ao fato de que os TEs intrinsecamente possuem motivos regulatórios, que acabam sendo dispersados no genoma em virtude da transposição destes elementos.(5) Jordan et al. (22) analisaram sequências de promotores humanos caracterizadas experimentalmente, disponíveis no banco HPD (do inglês, *Human Promoter Database*), e identificaram que aproximadamente 24% destas exibem presença de TEs. Os autores analisaram também sítios de ligação de fatores de transcrição caracterizados experimentalmente, disponíveis no banco de dados TRANSFAC (*Transcription Factor Database*), e observaram que cerca de 2,5% destes motivos - correspondendo a ~ 4,5% dos genes associados aos fatores - eram derivados de TEs.

Em outro estudo disponível na literatura, Lagemaat et al. (12) exploraram sequências

de transcritos com anotação de UTRs armazenados no RefSeq^a. Para humano, 12.179 *loci* foram analisados, dos quais 27,4% apresentaram no mínimo um mRNA com trechos derivados de TEs na UTR 5' ou 3'. Segundo os autores, esta seria uma forte evidência de que elementos de transposição podem influenciar a expressão dos genes por meio da doação de sinais regulatórios da transcrição.

Perspectiva semelhante também fora apontada em um estudo mais recente que utilizou dados de sequenciamento obtidos pelo método CAGE (do inglês, *Cap Analysis Gene Expression*), no qual o DNA complementar (cDNA) completo é clivado na posição 20 ou 21 a partir do extremo 5', sendo sequenciado em seguida.⁽⁸²⁾ Sequências geradas por esta técnica, quando mapeadas no genoma, permitem a determinação do SITr.⁽⁸²⁾ O estudo realizado por Faulkner et al. ⁽⁸²⁾ mostrou que, dentre os SITr identificados com o uso desta estratégia, 31,4% e 18,1% em humano e camundongo, respectivamente, ocorrem no interior de elementos repetitivos - sequências satélite, sequências repetitivas simples ou TEs. Os autores observaram também que aproximadamente 35% dos SITr associados com retrotransposons exibem expressão associada a tecidos específicos, enquanto no conjunto que engloba todos os SITr este índice fica em torno de 17%. Estes dados referem-se à espécie humana e indicam a relação de TEs com a regulação do padrão espacial ou temporal da expressão dos genes.⁽⁸²⁾

A integração de TEs pode introduzir promotores alternativos, conduzindo a padrões diferenciados de expressão em genes ortólogos.⁽¹²⁾ O gene CYP19, codificador da enzima aromatase P450 necessária à síntese do hormônio estrogênio, exemplifica este efeito.⁽¹²⁾ Este gene é expresso principalmente nas gônadas e no cérebro da maioria dos mamíferos, enquanto nos primatas é expresso também em níveis elevados em uma camada constituinte da placenta (*syncytiotrophoblast layer*).⁽¹²⁾ A expressão específica na placenta é mediada por um promotor alternativo localizado a aproximadamente 100 kb do início do gene, cuja origem remete ao elemento MER21A, um retrotransposon pertencente à classe LTR.⁽¹²⁾ Este promotor foi caracterizado para o gene CYP19 em humano, mas não foi identificado no ortólogo em camundongo, o qual não é expresso na placenta.⁽¹²⁾ O gene CYP19 também é considerado um importante exemplo da contribuição de TEs na indução da expressão gênica em tecidos específicos.⁽¹²⁾

Estes resultados documentados na literatura corroboram a hipótese de que sequências advindas de TEs podem contribuir na regulação da expressão gênica. Outros exemplos acerca da influência de TEs na regulação gênica podem ser encontrados no catálogo C-GATE^b.⁽⁵⁾

^a RefSeq (do inglês, *Reference Sequence*) é uma coleção de sequências anotadas, incluindo DNA genômico, transcritos e proteínas. Maiores detalhes em: <<http://www.ncbi.nlm.nih.gov/refseq/about/>>.

^b Disponível em: <<https://sites.google.com/site/tecatalog/>>.

4.1.2 Métodos

Distribuição de TEs em regiões flangeadoras 5'

A região flangeadora 5' do gene (Figura 4.3), com extensão máxima de 2 kpb (definida de acordo com o especificado na Seção 3.2.1), foi considerada como a partição genômica com potencial de influenciar o promotor do gene. Este intervalo foi adotado devido ao fato de que regiões regulatórias funcionais tendem à localização próxima do sítio de início de transcrição.(18, 81)

Figura 4.3 – Representação esquemática da região flangeadora 5'. São indicadas a região flangeadora 5' para um gene na fita 5'-3' (A) e para um gene na fita 3'-5' (B). A extensão máxima da região flangeadora 5', considerada neste trabalho, é 2 kpb. Fonte: Elaborada pela autora.

Define-se como contexto genômico a consideração conjunta de regiões genômicas que possuem a mesma disposição em relação ao gene correlato. Exemplificando, a Figura 4.4 ilustra a análise do contexto genômico relativo à região flangeadora 5' dos genes ortólogos, em que são consideradas conjuntamente as regiões flangeadoras 5' - de diferentes extensões - para a determinação da frequência observada de cada classe de TE. Contudo, o número observado de cópias de elementos de transposição não é elucidativo; para que esta informação possa ser explorada no aspecto biológico, é necessária a comparação com o número esperado de cópias no contexto genômico sob estudo.

É conhecido o fato de que a superposição dos efeitos decorrentes da seleção natural, da deriva genética e da preferência por sítios de inserção específicos para algumas famílias de TEs contribuem para a determinação da distribuição de TEs no genoma.(40) No entanto, considerando-se a situação hipotética livre da atuação destes fatores, as inserções supostamente seriam distribuídas de forma randômica no genoma (84), devido ao fato

Figura 4.4 – Representação esquemática de um contexto genômico analisado. Ilustração da consideração conjunta de regiões flanqueadoras 5' de genes distintos, para estudo da frequência de TEs no contexto genômico flanqueador 5' nos genes ortólogos. Fonte: Elaborada pela autora.

de que todas as posições do mesmo teriam igual probabilidade de serem acometidas por um evento de inserção. Esta conjectura foi utilizada como base para a avaliação da frequência observada de TEs, buscando-se inferir se a frequência observada difere do que seria esperado na ausência de fatores de restrição de inserção e fixação de elementos de transposição.

A frequência esperada foi estimada por meio da simulação computacional da distribuição randômica de TEs no genoma, tendo sido aplicado para este fim um método similar ao descrito por Zhang et al.(40) Em cada iteração da simulação, para cada família de TE, foi gerado um número de *loci* randômicos correspondente ao número de cópias mapeadas no genoma. Finalmente, em cada partição genômica de interesse, torna-se possível a inferência do número esperado de cópias, para cada família ou classe de TE, via determinação do número médio de inserções randômicas após a execução de 20.000 iterações da simulação.

Com o intuito de analisar se o número de cópias observado difere do esperado com significância estatística, a frequência observada de TEs (f_{obs}) foi comparada à distribuição de frequências do número de cópias randômicas resultante da simulação, através da aproximação da mesma à uma distribuição Normal. Esta variável aleatória pode ser denotada por $X \sim N(f_{esp}, \sigma^2)$, sendo f_{esp} a frequência esperada - frequência média obtida na simulação - e σ o desvio padrão, e pode ser transformada na variável aleatória $Z \sim N(0, 1)$, denominada Normal Padrão ou Reduzida, por meio da Equação 4.1.(85) Dessa forma, cada elemento da distribuição X pode ser padronizado, conduzindo à determinação da estatística z , a partir da qual torna-se possível a inferência da probabilidade de ocorrência de uma frequência de TE mais extrema do que a observada (p-valor), sob a consideração

da distribuição randômica de TEs no genoma. Neste trabalho a frequência observada será considerada significativa quando $|z| > 2,3263$, o que implica em p-valor $< 0,01$.(18, 86)

$$Z = \frac{X - f_{esp}}{\sigma} \quad (4.1)$$

Ainda em relação à análise acerca da presença de TEs, uma forma de expressar o nível de representatividade destes elementos em um determinado contexto genômico remete à frequência padronizada (Equação 4.2), em que a frequência observada é comparada diretamente à frequência esperada.(40) Esta variável assume valor positivo para a representação dos elementos acima do esperado e valor negativo para a representação abaixo do esperado, indicando o nível de ocorrência de TEs.

$$f_p = \log_{10}(f_{obs}/f_{esp}) \quad (4.2)$$

Identificação de sequências regulatórias

A estratégia utilizada neste trabalho para o reconhecimento de motivos regulatórios, possivelmente inseridos por TEs na região não codificante 5' dos genes, utilizou a base de dados PLACE, um banco de sequências nucleotídicas existentes em elementos de DNA regulatórios encontrados em plantas.(87) A atualização do repositório, em abril de 2006, passou a incluir motivos identificados na espécie *C. reinhardtii*^c. O banco dispõe uma ferramenta, o software SIGNAL SCAN^d, que permite a submissão de sequências para a busca (baseada em homologia) por motivos regulatórios armazenados no mesmo.(87)

O procedimento adotado teve início com a identificação da família de TE com maior frequência de inserção na região flanqueadora 5' para a espécie *C. reinhardtii* e a família com a mesma característica para a espécie *V. carteri*. Para cada família, foram recuperados os alinhamentos relativos às cópias na região flanqueadora 5'. A partir destes alinhamentos foi identificado o trecho da sequência consenso com maior frequência de inserção na região flanqueadora 5' dos ortólogos. Este trecho da sequência consenso foi submetido à busca por motivos regulatórios no PLACE, via execução do software SIGNAL SCAN.

^c Informação disponível em: <<http://www.dna.affrc.go.jp/PLACE/info.html>>.

^d Disponível em: <<http://www.dna.affrc.go.jp/htdocs/PLACE/>>.

4.1.3 Resultados

Distribuição de TEs na região flanqueadora 5' de genes ortólogos

As classes non-LTR e DNA transposon são as mais frequentes na região flanqueadora 5' dos genes ortólogos em *C. reinhardtii*, enquanto que em *V. carteri* as mais abundantes são a non-LTR e a LTR (Tabela 4.1). Este padrão é igual ao observado no genoma como um todo (Tabela 3.2).

Tabela 4.1 – Frequências observada e esperada para as classes de TEs na região flanqueadora 5' dos genes ortólogos em cada espécie.

Classe TE	Frequência observada	Frequência esperada	z
<i>C. reinhardtii</i>			
LTR	55	46,68 ± 6,72	1,24
non-LTR	443	308,89 ± 17,02	7,88
DNA-transposon	265	162,92 ± 12,31	8,29
outras	29	14,75 ± 3,74	3,81
<i>V. carteri</i>			
LTR	277	300,95 ± 16,68	-1,44
non-LTR	470	394,26 ± 19,01	3,98
DNA-transposon	3	15,38 ± 3,79	-3,27
outras	31	28,4 ± 5,17	0,5

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

De acordo com o exposto na Seção 4.1.2, a distribuição randômica de TEs no genoma - derivada da simulação - foi considerada como base de comparação para a frequência observada de TEs. A Figura 4.5 ilustra a distribuição de frequências relativas do número de cópias randômicas de TEs na região flanqueadora 5', derivada de 20.000 iterações na simulação, a qual aproxima-se de uma distribuição Normal, o que permitiu a determinação da estatística z (Tabela 4.1).

No caso de *V. carteri*, as classes LTR e DNA-transposon exibem representatividade abaixo do esperado segundo a distribuição randômica de TEs no genoma, sendo a representatividade associada à segunda classe diminuída significativamente em relação ao esperado (p-valor < 0,01). Apenas a classe non-LTR exibe frequência acima do esperado com significância estatística em regiões flanqueadoras 5' desta espécie. É interessante notar que, para *C. reinhardtii*, as quatro classes exibem frequência maior do que o esperado,

Figura 4.5 – Distribuição de frequências relativas do número de cópias randômicas de TEs na região flaqueadora 5'. Dados relativos às classes LTR, non-LTR, DNA-transposon e outras (respectivamente, da esquerda para direita) para as espécies *C. reinhardtii* (A) e *V. carteri* (B). Histogramas representam a distribuição de frequências obtida na simulação, com 20.000 iterações, e a curva em vermelho a distribuição Normal, com média e desvio padrão oriundos da simulação (Tabela 4.1). Fonte: Elaborada pela autora.

sendo esta representatividade estatisticamente significativa para três classes: non-LTR, DNA-transposon e outras.

Figura 4.6 – Distribuição de frequências relativas do número de cópias randômicas de TEs, em classes de distância na região flanqueadora 5'. Dados relativos à 1^a, 2^a e 3^a classes de distância (respectivamente, da esquerda para a direita), para TEs non-LTR (A) e DNA-transposon (B) em *C. reinhardtii* (Tabela B.1, Apêndice B), e para TEs non-LTR (C) e LTR (D) em *V. carteri* (Tabela B.2, Apêndice B). Histogramas representam a distribuição de frequências obtida na simulação, com 20.000 iterações, e a curva em vermelho a distribuição Normal, com média e desvio padrão oriundos da simulação. Fonte: Elaborada pela autora.

Visando uma análise mais detalhada acerca da distribuição de TEs na região flanqueadora 5' dos genes ortólogos, esta foi inspecionada em classes de distância de extensão variável e crescente à medida que a classe se distancia do SIT. Exemplos de distribuições de frequências do número de cópias, nas diferentes classes de distância obtidas na simulação

de inserções randômicas, são ilustradas na Figura 4.6 para as espécies *C. reinhardtii* e *V. carteri*. Embora sejam expostos apenas os perfis obtidos para a 1ª, 2ª e 3ª classes de distância, as demais classes apresentam o mesmo comportamento, sendo possível atribuir a distribuição Normal para todas elas. Em cada classe de distância, foram identificadas a frequência observada e a frequência esperada de cópias de TEs (figuras 4.7-A e 4.8-A), para que, a partir destes dados, fossem acessados valores da estatística z e da frequência padronizada (figuras 4.7-B e 4.8-B). Somente as duas classes de TEs mais abundantes em cada espécie foram submetidas a esta análise.

O primeiro aspecto que chama a atenção no estudo da distribuição de frequências padronizadas remete ao fato de que ambas as classes examinadas em *C. reinhardtii* (Figura 4.7-B) e *V. carteri* (Figura 4.8-B) encontram-se fortemente reprimidas no início da região flanqueadora 5', mais especificamente, nas proximidades do SIT. Esta zona de baixa representatividade (p-valor < 0,01) compreende o intervalo de 1 a 300 pb para as classes na primeira espécie; em relação à segunda espécie, este intervalo pode ser definido na região de 1 a 300 pb para a non-LTR e 1 a 600 pb para a classe LTR.

Figura 4.7 – Distribuição de frequências de TEs na região flanqueadora 5' de genes ortólogos em *C. reinhardtii*. São analisadas apenas as classes de TEs mais abundantes na espécie. (A) Em cada classe de distância, são apresentadas a frequência observada (barra da esquerda) e a frequência média obtida na simulação (barra da direita, com coloração mais suave). As barras de erro correspondem ao desvio padrão da frequência de TEs obtida na simulação. Classes de TEs, cuja frequência observada encontra-se indicada com *, exibem $|z| > 2,3263$ ($p < 0,01$) na região analisada. Os valores de z, relativos a cada classe de distância, encontram-se na Tabela B.1, localizada no Apêndice B. (B) A frequência padronizada esperada, de acordo com a distribuição randômica de cópias de TEs no genoma, é representada pela linha em vermelho. Barras de erro representam a incerteza da frequência padronizada, definida pela propagação do desvio padrão da frequência esperada.(88) Fonte: Elaborada pela autora.

Figura 4.8 – Distribuição de frequências de TEs na região flanqueadora 5' de genes ortólogos em *V. carteri*. São analisadas apenas as classes de TEs mais abundantes na espécie. (A) Em cada classe de distância, são apresentadas a frequência observada (barra da esquerda) e a frequência média obtida na simulação (barra da direita, com coloração mais suave). As barras de erro correspondem ao desvio padrão da frequência de TEs obtida na simulação. Classes de TEs, cuja frequência observada encontra-se indicada com *, exibem $|z| > 2,3263$ ($p < 0,01$) na região analisada. Os valores de z , relativos a cada classe de distância, encontram-se nas Tabela B.2, localizada no Apêndice B. (B) A frequência padronizada esperada, de acordo com a distribuição randômica de cópias de TEs no genoma, é representada pela linha em vermelho. Barras de erro representam a incerteza da frequência padronizada, definida pela propagação do desvio padrão da frequência esperada. (88) Fonte: Elaborada pela autora.

Além disso, os resultados indicam que ambas as espécies possuem intervalos de elevada representatividade de TEs, quando comparado ao esperado segundo a distribuição randômica das cópias no genoma. Para *C. reinhardtii*, as classes analisadas encontram-se representadas acima do esperado (p-valor $< 0,01$) no intervalo de 301 a 2000 pb. Para *V. carteri*, observa-se que a classe non-LTR exibe frequências significativamente acima do esperado no intervalo de 601 a 2000 pb; no caso da classe LTR, apenas uma classe de distância pode ser citada neste quesito: a que compreende o intervalo de 1001-1500 pb.

Apesar da representatividade de TEs acima do esperado na região flanqueadora 5' como um todo (Tabela 4.1) para as classes non-LTR e DNA-transposon em *C. reinhardtii* e para a classe non-LTR em *V. carteri*, a inspeção mais detalhada expôs a ausência desta super-representação na proximidade do SIT em ambas as espécies (figuras 4.7 e 4.8). Considerando o comprimento médio da UTR 5' em plantas vasculares, aproximadamente 106 pb (89), este resultado sugere que a representação reprimida de TEs neste intervalo pode ser resultado da seleção negativa de eventos de inserção, devido à proximidade ao sítio de início de transcrição. Em termos gerais, os resultados relativos à distribuição de frequências

padronizadas sugerem que existe uma tendência de aumento na representatividade de TEs na região flanqueadora 5' à medida que a distância em relação ao SIT aumenta, tendência esta mais evidente para a espécie *C. reinhardtii*.

Jordan et al. (22) constataram um padrão semelhante. No estudo realizado com promotores humanos, estas sequências foram divididas em regiões de 100 pb de extensão, constituindo seis classes de distância definidas em torno do SITr (-500 a +100), nas quais foi avaliada a proporção de promotores contendo TEs. Foi observado um aumento na proporção à medida que a classe examinada apresentava maior distância em relação ao SITr.(22) Segundo os autores, este padrão sugere que inserções próximas ao SITr possuem maior potencial deletério para a função do gene, o que resulta na remoção mais frequente destas pela seleção purificadora, quando comparado a inserções mais distantes deste sítio.

Identificação de sequências regulatórias

A ausência de bancos de dados com sequências de promotores caracterizados especificamente para as espécies *C. reinhardtii* e *V. carteri* impossibilita a análise acerca da associação entre estas estruturas e os TEs, como realizado em estudos para a espécie humana - a exemplo do realizado por Jordan et al.(22)

No entanto, com o objetivo de investigar uma possível contribuição dos TEs no contexto regulatório dos genes ortólogos destas espécies, foi avaliado o conteúdo da sequência consenso inserida. Esta inspeção foi limitada à família de TE mais abundante na região flanqueadora 5' em cada espécie, L1-3_CR para *C. reinhardtii* (Figura 4.9-A) e RTE5-VC para *V. carteri* (Figura 4.9-B).

A Figura 4.10 ilustra o conteúdo das sequências consenso das famílias L1-3_CR e RTE5-VC inserido na região flanqueadora 5' dos genes ortólogos. A tendência de inserção da terminação 3' da sequência dos TEs analisados deve-se, muito provavelmente, à baixa processividade da enzima transcriptase reversa envolvida no processo de transposição destes elementos. Assim sendo, é muito pouco provável que a predominância de inserção da terminação 3' das sequências dos TEs seja um efeito resultante da pressão seletiva evolutiva. Apesar deste fato, é possível que o trecho inserido com maior frequência esteja agregando à região flanqueadora 5' novas estruturas regulatórias.

Para cada família de TE, o intervalo da sequência consenso com maior frequência de inserção (fração correspondente ao extremo 3' - 98% à 100% do comprimento total, o que equivale à delimitação de 1932 a 1991 pb para a família L1-3_CR e 5166 a 5325 pb para a RTE5-VC) foi analisado quanto à presença de motivos regulatórios, detectados utilizando-se a base de dados PLACE. Esta busca revelou que o segmento analisado para L1-3_CR contém o motivo E-box (CANNTG), uma sequência consenso de reconhecimento da proteína bHLH (basic helix-loop-helix). De acordo com Heim et al. (90), uma função

Figura 4.9 – Frequência de ocorrência absoluta das famílias de TEs na região flanqueadora 5' dos genes ortólogos. São indicadas apenas as famílias com frequência acima de 20 cópias nesta localização, para as espécies *C. reinhardtii* (A) e *V. carteri* (B). Fonte: Elaborada pela autora.

Figura 4.10 – Posições, em termos da porcentagem do comprimento total da sequência consenso do TE, inseridas na região flanqueadora 5' dos genes ortólogos. A família mais frequente na região flanqueadora 5' para *C. reinhardtii* é a L1-3_CR (A) e para *V. carteri* é a RTE-5_VC (B), ambas pertencentes à classe non-LTR. Os perfis apresentados referem-se a cópias com orientação senso em relação ao gene, tendo sido englobadas nesta análise 103 ocorrências para a família L1-3_CR e 48 para a RTE-5_VC. Fonte: Elaborada pela autora.

anciã desta proteína retida pelas plantas está relacionada a respostas ao ambiente como, por exemplo, respostas à luz. O segmento analisado para o elemento RTE5-VC, por sua vez, contém o motivo ACGTATERD1 (ACGT). No processo de senescência induzido por privação de luz, esta sequência está associada à regulação positiva do gene *erd1* na planta *Arabidopsis thaliana*.(91)

A presença destes motivos regulatórios em TEs integrados em regiões flangeadoras 5' sugere que estes elementos possam estar influenciando a regulação gênica. No entanto, estudos experimentais específicos são necessários para a verificação desta hipótese.

4.2 Região flangeadora 3' dos genes ortólogos

4.2.1 Introdução

Os sinais que especificam a terminação 3' do mRNA encontram-se codificados no DNA: estes são transcritos para o pré-mRNA, onde são reconhecidos pelas proteínas correlatas à poliadenilação (36) (Figura 4.11). Duas proteínas de especial importância neste processo, denominadas CPSF (*cleavage and polyadenylation specificity factor*) e CstF (*cleavage stimulation factor F*), são transferidas da cauda da RNA polimerase II à região de terminação do pré-mRNA no momento em que esta emerge da enzima polimerase.(36, 72) Estando estas proteínas associadas ao transcrito em síntese, passam a integrar o complexo os fatores de clivagem I e II (CFI e CFII) e a enzima poli-A polimerase (PAP), finalizando a constituição da maquinaria de poliadenilação.(72) Esta é inteiramente edificada no pré-mRNA antes que o processo tenha início.

O mecanismo de poliadenilação envolve duas etapas: na primeira, o pré-mRNA é clivado em um sítio específico denominado sítio poli(A) pela ação da nuclease CPSF; em seguida, a enzima poli-A polimerase adiciona resíduos adenina (A) à terminação 3' do transcrito resultante da clivagem.(72, 75, 92) Estes resíduos são adicionados um a um na cadeia, sintetizando assim a cauda poli(A) cujo comprimento pode variar entre espécies: enquanto em humanos são polimerizadas de 250 a 300 adeninas em média, em leveduras este número se restringe ao intervalo aproximado de 70 a 80 resíduos.(75) A maioria dos elementos regulatórios *cis* envolvidos neste processo associam-se à sequências auxiliares localizadas em regiões entre -100 nt e +100 nt em relação ao sítio poli(A) (Figura 4.11); já as sequências sinalizadoras, essenciais ao processo porque interagem com as proteínas CPSF e CstF, localizam-se em intervalos de -40 nt a +40 nt.(92)

Aproximadamente metade dos genes humanos possuem sítios alternativos de poliadenilação, o que propicia a variabilidade de transcritos.(93) Estes sítios podem estar

Figura 4.11 – Fatores envolvidos no processo de poliadenilação em mamíferos. A região genômica que abriga o sítio poli(A) e sequências regulatórias correlatas ao mesmo é definida como região do sítio poli(A).(92) O sinal canônico de poliadenilação AAUAAA (ou um variante próximo), ao qual se liga a proteína CPSF, e a sequência rica em resíduos U/GU, que interage com a CstF, localizam-se, respectivamente, em regiões entre -40 nt e +40 nt em relação ao sítio poli(A).(92) Sequências auxiliares, que interagem com elementos regulatórios que modulam a poliadenilação, localizam-se na região de -100 nt a +100 nt. Estão representados ainda os fatores de clivagem I e II (CFI e CFII) e a enzima poli-A polimerase (PAP) que, juntamente com as proteínas CPSF e CstF, compõem a maquinaria de poliadenilação. Fonte: LUTZ(72)

localizados após o códon de terminação (configuração mais frequente associada à poliadenilação alternativa), em íntrons ou em regiões codificantes dos genes.(75) O emprego de sítios poli(A) alternativos no processamento do pré-mRNA pode acarretar alterações na extensão da UTR3' no mRNA, ou na sequência codificante no transcrito maduro.(93)

Variações na UTR3' podem influenciar a estabilidade do mRNA, o direcionamento do mesmo a determinadas regiões celulares e a eficiência na tradução, uma vez que esta região do transcrito abriga sequências que são reconhecidas em vias correlatas a estas regulações, a exemplo de miRNAs e proteínas que se ligam ao RNA.(76, 77) O gene que codifica o fator BDNF (*brain-derived neurotrophic factor*) constituiu a primeira evidência, em neurônios, de que a poliadenilação alternativa pode atuar no direcionamento da localização do mRNA.(76) Embora a via envolvida na regulação do transporte e da tradução local não tenha sido totalmente elucidada, estudos com células do hipocampo indicam que transcritos de BDNF com UTR3' mais longa são especificamente direcionados para os dendritos, onde atuam na regulação da formação e eliminação destas ramificações, enquanto transcritos com UTR3' mais curta permanecem, em geral, no corpo celular.(76)

Outra forma de alteração da expressão gênica por meio da poliadenilação alternativa remete à modificação do produto original do gene. Um exemplo interessante refere-se à imunoglobulina M (IgM), um anti-corpo produzido pelas células B.(75) No estágio de desenvolvimento desta célula, o emprego do sítio poli(A) mais próximo da extremidade 3' do gene produz a forma do IgM que se liga à membrana da célula, onde atua como

receptor no reconhecimento de antígenos.(93) Durante a ativação do linfócito B, o sítio poli(A) localizado no íntron passa a ser utilizado na maturação do pré-mRNA, originando uma forma da proteína IgM que será secretada devido à deleção da região hidrofóbica C-terminal, responsável pela interação com a membrana.(75, 93) Esta troca de sítio de poliadenilação é essencial para o mecanismo de resposta imune.

É possível que a determinação do sítio alternativo utilizado na maturação do pré-mRNA esteja relacionada à regulação do desenvolvimento celular e a tecidos específicos.(76) No entanto, apesar do avanço em estudos que buscam a identificação de processos biológicos que utilizam sítios alternativos, ainda não é possível compreender completamente o papel biológico de eventos de poliadenilação alternativa.(75)

Segundo Lee et al. (92), genes com história evolutiva mais longa exibem maior probabilidade de possuir poliadenilação alternativa, sugerindo ganho de sítios poli(A) no decorrer da evolução. O estudo realizado pelos pesquisadores, envolvendo genes de humano e camundongo, revelou que os sítios poli(A) não conservados entre as duas espécies encontram-se mais frequentemente associados a TEs, quando comparado aos sítios poli(A) conservados.(92) Este resultado indica que os TEs podem contribuir para a criação e modulação de sítios poli(A) no decorrer da evolução, ou ainda para o surgimento de sítios específicos em uma determinada espécie.(92) Segundo os autores, elementos de transposição podem interferir na evolução da terminação 3' dos genes por meio da exaptação de segmentos de suas sequências em sinais de poliadenilação.(92) Outros trabalhos na literatura também apontam a relevância de eventos de transposição nesta região como fonte de novos sinais de poliadenilação.(10, 12)

Em função da relevância do mecanismo de poliadenilação, esta seção destina-se à descrição dos métodos estabelecidos para o estudo, e resultados advindos destes, da distribuição de TEs na região adjacente ao término 3' dos genes ortólogos em *C. reinhardtii* e *V. carteri*, bem como para o estudo de formas de associação entre estes elementos e sítios poli(A).

4.2.2 Métodos

Distribuição das cópias de TEs

A região flanqueadora 3' corresponde, neste estudo, ao intervalo intergênico tangente ao término do gene, com extensão máxima de 2 kpb (Figura 4.12). Como já citado, é neste segmento genômico que encontram-se os sinais relacionados ao processo de poliadenilação. Assim como para a região flanqueadora 5', este intervalo pode apresentar extensão menor do que 2 kpb devido à proximidade entre dois genes. A partir da definição da região

flanqueadora 3' dos genes ortólogos, foi possível a determinação das frequências observada e esperada neste contexto genômico, o que permitiu o estudo acerca da distribuição de TEs no mesmo.

Figura 4.12 – Representação esquemática da região flanqueadora 3' do gene. São indicadas a região flanqueadora 3' para um gene na fita 5'-3' (A) e para um gene na fita 3'-5' (B). A extensão máxima da região flanqueadora 3', considerada neste trabalho, é 2 kpb. Fonte: Elaborada pela autora.

Associação de TEs com sítios poli(A)

O trabalho realizado por Shen et al. (94) resultou em uma biblioteca de sítios poli(A)^e para a espécie *C. reinhardtii*. Os autores utilizaram ESTs para o mapeamento dos sítios no genoma (versão v.3.1), sendo o sítio poli(A) definido como o último nucleotídeo, antecedente à cauda poli(A), mapeado no genoma.(94) Uma vez identificado este nucleotídeo, a região genômica flanqueadora 5' (com extensão de 300 nt) e a região genômica flanqueadora 3' (com extensão de 100 nt, incluindo o sítio poli(A)) foram extraídas, originando desta forma um banco com sequências genômicas com extensão de 400 nt, que constituem evidência dos sítios poli(A).(94) Outro aspecto interessante apresentado no trabalho dos autores refere-se ao sinal de poliadenilação predominante verificado em *C. reinhardtii*, UGUAA, o qual difere do canônico AAUAAA descrito em mamíferos.(94) Estudos anteriores já haviam apontado esta sequência como sinalizadora do processo de poliadenilação em *C. reinhardtii*, mas a verificação em larga escala foi possível somente após o sequenciamento do genoma desta alga.(94) A ausência de repositório semelhante para *V. carteri* inviabilizou a análise para esta espécie.

O método adotado neste estudo, com o objetivo de analisar a associação entre TEs e sítios de poliadenilação, tem como primeira etapa a definição das coordenadas dos

^e Disponível em: <http://www.users.muohio.edu/liq/links.html>.

sítios poli(A) nos genes ortólogos em *C. reinhardtii*. Com este intuito, as sequências do repositório foram mapeados no genoma desta espécie, via execução do software BLASTn, sendo selecionados apenas os alinhamentos correspondentes à sequência exata dos sítios de referência. As coordenadas derivadas deste mapeamento foram comparadas às coordenadas dos genes, exigindo-se também orientação senso entre o sítio poli(A) e o gene correlato, conduzindo à identificação de sítios localizados em regiões intragênicas e flangeadoras 3'. Sendo identificados mais de um sítio poli(A) no gene, são selecionados somente aqueles que possuem uma distância mínima de 20 pb entre si.

Após a definição dos sítios poli(A) nos genes ortólogos, torna-se possível a identificação de associações entre os mesmos e os TEs. Esta consitui a segunda etapa do método ora abordado, na qual, para cada gene ortólogo com sítio poli(A), são recuperadas as inserções relacionadas ao referido gene, verificando-se se estas apresentam sobreposição com a região genômica compreendida pelas coordenadas definidas a -40 pb e $+40$ pb em relação ao sítio poli(A). Casos de inserções designadas por mais de um alinhamento foram avaliadas segundo as coordenadas no genoma de cada alinhamento, evitando que sítios poli(A) entre os mesmos fossem atribuídos ao TE em questão. Possibilidades de associação entre TEs e sítios poli(A) estão representadas na Figura 4.13.

Figura 4.13 – Formas de associação do sítio poli(A) com TE. Representação esquemática de possibilidades de associação em que o sítio poli(A), bem como as estruturas regulatórias do processo de poliadenilação, encontram-se relacionadas ao TE (tipo de associação 1), e possibilidades em que apenas as estruturas regulatórias apresentam esta relação (tipos de associação 2 e 3). A linha horizontal em preto representa a região genômica contendo o sítio poli(A), enquanto as barras horizontais representam TEs. Fonte: LEE;JI;TIAN(92)

4.2.3 Resultados

Distribuição de TEs na região flanqueadora 3' de genes ortólogos

As classes non-LTR e DNA-transposon são as mais frequentes na região flanqueadora 3' dos genes ortólogos em *C. reinhardtii*, enquanto em *V. carteri* a non-LTR e a LTR exibem maior frequência de ocorrência (Tabela 4.2). Novamente, este padrão reproduz o observado no genoma (Tabela 3.2).

Tabela 4.2 – Frequências observada e esperada para as classes de TEs na região flanqueadora 3' dos genes ortólogos em cada espécie.

Classe TE	Frequência observada	Frequência esperada	z
<i>C. reinhardtii</i>			
LTR	129	71,91 ± 8,16	7,0
non-LTR	920	475,99 ± 20,88	21,26
DNA-transposon	656	250,87 ± 15,24	26,58
outras	56	22,71 ± 4,6	7,24
<i>V. carteri</i>			
LTR	287	362,01 ± 18,17	-4,13
non-LTR	651	474,08 ± 20,61	8,58
DNA-transposon	13	18,49 ± 4,1	-1,34
outras	31	34,08 ± 5,65	-0,55

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

A comparação entre a frequência observada e a esperada - segundo a distribuição randômica de cópias no genoma - evidencia a representatividade significativamente acima do esperado (p -valor $< 0,01$) para as classes de TEs em *C. reinhardtii*. A espécie *V. carteri*, por sua vez, exibe representatividade acima do esperado somente para a classe non-LTR; a classe LTR, contrariamente, exibe representatividade significativamente abaixo do esperado, enquanto as demais não exibem distanciamento significativo em relação ao esperado.

Com o intuito de explorar em maiores detalhes a distribuição de TEs em regiões flanqueadoras ao término dos genes, as duas classes de TEs mais frequentes em cada espécie foram submetidas à inspeção da frequência padronizada em intervalos genômicos definidos a partir do códon de terminação (figuras 4.14-B e 4.15-B). Os elementos non-LTR, em *C. reinhardtii* (Figura 4.14), exibem frequência próxima do esperado nas classes de

distância próximas ao término do gene (1-300 pb); em regiões mais distantes (301-2000 pb), a frequência padronizada mostra-se significativamente acima do esperado. Os elementos DNA-transposon exibem um padrão de distribuição bastante similar ao da classe non-LTR, exceto pelo fato de que encontram-se com representatividade significativamente abaixo do esperado na primeira classe de distância avaliada (1-100 pb).

Figura 4.14 – Distribuição de frequências de TEs na região flanqueadora 3' de genes ortólogos em *C. reinhardtii*. São analisadas as classes de TE mais abundantes na espécie. (A) Em cada classe de distância, são apresentadas a frequência observada (barra da esquerda) e a frequência média obtida na simulação (barra da direita, com coloração mais suave). As barras de erro correspondem ao desvio padrão da frequência de TEs obtida na simulação. Classes de TEs, cuja frequência observada encontra-se indicada com *, exibem $|z| > 2,3263$ ($p < 0,01$) na região analisada. Os valores de z , relativos a cada classe de distância, encontram-se na Tabela B.3, localizada no Apêndice B. (B) A frequência padronizada esperada, de acordo com a distribuição randômica de cópias de TEs no genoma, é representada pela linha em vermelho. Barras de erro representam a incerteza da frequência padronizada, definida pela propagação do desvio padrão da frequência esperada. (88) Fonte: Elaborada pela autora.

Comparada à espécie *C. reinhardtii*, *V. carteri* (Figura 4.15) apresenta níveis de frequência padronizada mais modestos nesta região. Elementos non-LTR exibem frequência padronizada negativa significativa no intervalo genômico adjacente ao término do gene, passando a uma frequência próxima do esperado em regiões intermediárias (101-600 pb), e a frequências positivas significativas em regiões mais distantes (601-2000 pb). A classe LTR encontra-se significativamente reprimida em um intervalo bastante extenso seguinte ao códon de terminação (1-600 pb), passando a níveis de frequência padronizada que não divergem significativamente do esperado nas demais classes de distância avaliadas.

É interessante notar que, assim como no estudo da distribuição de TEs na região promotora, existe uma tendência de elevação nos níveis da frequência padronizada à

Figura 4.15 – Distribuição de frequências de TEs na região flanqueadora 3' de genes ortólogos em *V. carteri*. São analisadas as classes de TE mais abundantes na espécie. (A) Em cada classe de distância, são apresentadas a frequência observada (barra da esquerda) e a frequência média obtida na simulação (barra da direita, com coloração mais suave). As barras de erro correspondem ao desvio padrão da frequência de TEs obtida na simulação. Classes de TEs, cuja frequência observada encontra-se indicada com *, exibem $|z| > 2,3263$ ($p < 0,01$) na região analisada. Os valores de z , relativos a cada classe de distância, encontram-se na Tabela B.4, localizada no Apêndice B. (B) A frequência padronizada esperada, de acordo com a distribuição randômica de cópias de TEs no genoma, é representada pela linha em vermelho. Barras de erro representam a incerteza da frequência padronizada, definida pela propagação do desvio padrão da frequência esperada.(88) Fonte: Elaborada pela autora.

medida que a distância em relação ao limite do gene (neste caso, o códon de terminação) aumenta. Novamente, este padrão é mais evidente para a espécie *C. reinhardtii*.

Associação entre TEs e sítios poli(A)

O mapeamento das sequências relativas aos sítios poli(A), para *C. reinhardtii*, resultou na identificação de 2.799 sítios no contexto genômico dos ortólogos nesta espécie. Destes, 27 foram mapeados em regiões intragênicas e 2.772 em regiões flanqueadoras 3'. O número de sítios intragênicos é modesto se comparado ao montante mapeado na região flanqueadora 3'; provavelmente, a ocorrência moderada de sítios poli(A) em regiões intragênicas deve-se à modificação da região codificante do transcrito mediada por estes sinais.(75) Esta análise resultou em um conjunto de 1.893 genes ortólogos em *C. reinhardtii* com, no mínimo, um sítio poli(A) atribuído; dentre estes, 595 (~31,43%) apresentaram sítios alternativos de poliadenilação.

A inserção de elementos de transposição pode agregar ao gene uma sequência contendo o sítio poli(A) propriamente, ou uma sequência contendo apenas os motivos sinalizadores do processo de poliadenilação. A Figura 4.16-A ilustra a primeira forma de contribuição destes elementos, em que o cerne da região do sítio poli(A) encontra-se associado ao TE (forma de associação 1); a segunda forma, em que os TEs podem contribuir com as sequências sinalizadoras *upstream* ou *downstream*, encontra-se representada pelas associações 2 e 3.

Figura 4.16 – Associação entre sítios poli(A) e TEs na região flanqueadora 3' de genes ortólogos. (A) Tipos de associação entre TEs e sítios de poliadenilação. A linha horizontal em preto representa a região genômica contendo o sítio poli(A), enquanto as barras horizontais representam TEs. Fonte: LEE;JI;TIAN(92). (B) Frequência de ocorrência dos tipos de associação entre os sítios poli(A) e as classes de TEs para a espécie *C. reinhardtii*. Encontram-se representadas 247 associações identificadas na região flanqueadora 3' dos genes ortólogos; a associação entre o sítio poli(A) e o elemento L1-6_CR no terceiro íntron do gene Cre12.g510900 (associação tipo 2) não foi incluída nesta representação. Fonte: Elaborada pela autora.

A averiguação da associação entre regiões de sítio poli(A) e TEs resultou na detecção de 243 sítios com esta característica, dos quais 242 localizam-se em regiões flanqueadoras 3' e um em região intragênica, mais especificamente, no terceiro íntron do gene Cre12.g510900. Estes sítios conduzem a um cenário em que 174 genes apresentam sítio poli(A) associado a TE (173 na região flanqueadora 3' e 1 na região intragênica), dos quais 91 possuem sítio alternativo de poliadenilação (~ 52,30%). Este percentual, quando comparado ao percentual de genes ortólogos com sítios de poliadenilação alternativos (31,43%), sugere que os TEs possam estar contribuindo para o advento de sinais alternativos de poliadenilação.

Os resultados desta análise indicam que ~ 8,68% das regiões de sítio poli(A) mapeadas exibem associação com TEs, correspondendo a ~ 9,19% dos genes com mapeamento de sítio poli(A). No trabalho de Lee et al. (92), em que os autores analisaram a associação entre sítios poli(A) e TEs em humano, estes índices correspondem a ~ 8% e ~ 16%,

respectivamente.

A forma de associação preponderante entre TEs e regiões de sítio poliA é a envolvendo o cerne da região do sítio poli(A), seguida pela associação envolvendo a sequência sinal *downstream* para todas as classes de TEs analisadas (Figura 4.16-B). O número de associações entre regiões de sítio poli(A) e TEs é maior para a classe non-LTR, seguida pela DNA-transposon, o que reproduz o padrão de frequência de ocorrência observado no genoma para as famílias de TEs estudadas. Exemplos de cada forma de associação são ilustrados na Figura 4.17.

Figura 4.17 – Exemplos de associações entre TEs e sítios poli(A) em genes da espécie *C. reinhardtii*. Nos diagramas acima, regiões de sítio poli(A) - definidas pelas coordenadas -40 pb a + 40 pb em relação ao sítio - são representados em vermelho; sequências relativas a TEs são indicadas em marrom. (A) Representação esquemática do gene Cre01.g054500, no qual foi identificada a associação tipo 1, em que o cerne da região do sítio poli(A) encontra-se em região derivada do elemento hAT-N7_CR (cromossomo 1: 7.554.988-7.555.685). (B) Representação esquemática do gene Cre12.g498850, onde foi registrada ocorrência da associação tipo 2, na qual o elemento L1-2_CR (cromossomo 12: 3.213.260-3.213.338) encontra-se associado à sequência sinalizadora de poliadenilação *upstream*. (C) Representação esquemática do gene Cre01.g016300, no qual existe uma associação do tipo 3, em que o elemento DNA-1-9_CR (cromossomo 1: 2.754.316-2.754.437) encontra-se associado à sequência sinalizadora de poliadenilação *downstream*. Fonte: Elaborada pela autora.

As associações indicadas na Figura 4.16-B derivam de 184 cópias que possuem relação com região de sítio poli(A) em localizações flanqueadoras 3'; existe ainda a cópia do elemento L1-6_CR, vinculada à região do sítio poli(A) intrônico no gene Cre12.g510900,

perfazendo um total de 185 cópias associadas a regiões de sítio poli(A). O número total de cópias é menor do que o número de regiões de sítio poli(A) associadas a TEs, devido ao fato de que uma mesma cópia pode estar associada a mais de um sítio. De fato, dentre as 185 cópias, 137 relacionam-se a um único sítio poli(A), 37 a dois sítios, 7 a três sítios e 4 cópias a quatro sítios. Como exemplo do último caso, cita-se o elemento CRTOC1 (LTR), para o qual foram identificadas quatro associações do tipo 1 - extrapolando os limites -40 nt a +40 nt em relação ao sítio poli(A) - no gene Cre05.g245900. Este quadro conduz a um número de 248 associações entre sítios poli(A) e TEs, número este maior do que o número de regiões de sítio poli(A), indicando a ocorrência de associações entre um único sítio e duas cópias de TEs. Esta última configuração foi observada para cinco sítios poli(A), que apresentam influência de TEs distintos nas regiões *upstream* e *downstream* do sítio, como exemplifica a Figura 4.18.

Figura 4.18 – Exemplo de sítio poli(A) associado a duas cópias de TEs. Representação esquemática do gene Cre10.g432050 em *C. reinhardtii*, no qual o 4º sítio poli(A) encontra-se associado a duas cópias de TEs. Regiões de sítio poli(A) - definidas pelas coordenadas -40 pb a + 40 pb em relação ao sítio - são representados em vermelho; sequências relativas aos elementos L1-6_CR (cromossomo10: 1.932.779-1.932.900) e TE2-7b_CR (cromossomo10: 1.932.936-1.933.130) são indicadas em marrom. Fonte: Elaborada pela autora.

Os resultados apresentados nesta seção indicam contribuições dos elementos de transposição na introdução de regiões de sítios poli(A), ou estruturas sinalizadoras do processo de poliadenilação, em genes da espécie *C. reinhardtii*. Eventos desta natureza na região flangeadora 3' resultam em formas alternativas dos transcritos, em termos da extensão da UTR3', o que pode influenciar as vias de regulação da tradução e do direcionamento do mRNA na célula, em virtude de sequências armazenadas na UTR3' correlatas a estas vias.(76, 77) A identificação de contribuições dos TEs na sinalização do processo de poliadenilação em genes de *C. reinhardtii* corrobora a hipótese de que estes elementos possam influenciar a evolução do término 3' dos genes nesta alga, de maneira semelhante ao observado em outras espécies.(92)

Capítulo 5

Estudo de TEs em regiões intragênicas

No capítulo anterior foram explorados aspectos relativos à ocorrência de elementos de transposição em regiões genômicas flanqueadoras às extremidades dos genes ortólogos. O capítulo que ora se inicia destina-se ao estudo da ocorrência destes elementos em regiões intragênicas, mais especificamente, em regiões codificantes e intrônicas dos ortólogos em *C. reinhardtii* e *V. carteri*.

5.1 Região codificante

5.1.1 Introdução

A relevância dos TEs na evolução de sequências codificadoras de proteínas vem sendo reconhecida à medida que eventos desta natureza são constatados.(11, 14, 41) Como exemplo, citam-se os casos de domesticação de TEs identificados na proteína Rag1 e nos fatores FHY3 e FAR1, já citados na Seção.2.1 Contribuições mais singelas, decorrentes da inserção com subsequente adoção da sequência do TE ainda que não codificante - em um processo denominado exaptação^a, podem influenciar a evolução do gene por meio da extensão da sequência codificante, introdução de códons de iniciação ou terminação alternativos ou ainda sítios alternativos de *splicing*, os quais conduzem à formas alternativas do transcrito e a isoformas da proteína.(10, 11, 18, 68)

Dois mecanismos possibilitam aos TEs integrar regiões codificantes (Figura 5.1): a inserção direta no DNA codificante e a inserção no DNA não codificante, seguida pela

^a Brosius e Gould sugeriram o termo exaptação para eventos em que uma sequência "lixo" do DNA, como um segmento de TE, adquire uma nova função no genoma.(95)

exaptação destas sequências em novos éxons, o que é possível devido aos sítios de *splicing* existentes em TEs.(11, 16) O estudo realizado por Nekrutenko e Li (11) indicou que aproximadamente 4% dos genes humanos possuem TEs em regiões codificantes, sendo a segunda via de integração a mais frequente, representando 89,5% dos éxons com presença destes elementos. Segundo os autores, o segundo mecanismo é vantajoso em relação ao primeiro porque permite que o novo éxon, em circunstâncias específicas, não esteja presente no mRNA devido ao *splicing* alternativo, não prejudicando a função original do gene.

Figura 5.1 – Representação esquemática de formas de incorporação de TEs em sequências codificantes nos genes. A integração do TE pode ocorrer pela inserção direta em um éxon codificante (A), ou pela inserção em um íntron seguida da exaptação da sequência do TE, ou de parte da sequência, em um éxon codificante (B). Fonte: NEKRUTENKO;LI(11)

Ocorrências de TEs em regiões codificantes são pouco frequentes devido ao elevado potencial deletério destes eventos para a função do gene, o que conduz à eliminação das cópias inseridas nestas delimitações.(11, 84) No entanto, cópias que passam a integrar sequências codificantes podem conduzir a efeitos evolutivos positivos, embora estes eventos sejam raros, o que motiva estudos acerca da participação dos TEs em parcelas codificantes dos genes.(11, 17, 18) O efeito da integração de TEs nestas regiões têm sido explorado em genes de plantas como o arroz (18) e o café (17), assim como em mamíferos, a exemplo da espécie humana e de camundongos.(11)

5.1.2 Métodos

O mapeamento de TEs no genoma permitiu a identificação de ocorrências destes elementos em regiões codificantes dos genes preditos. Valores da estatística z (definida na Seção 4.1.2), relativas às frequências observadas de cada classe de TEs neste contexto genômico, foram obtidas com base na frequência esperada segundo a distribuição randômica dos elementos no genoma.

Apesar do fato de que uma parcela das instâncias identificadas como ocorrências de TEs em regiões codificantes possa ser reflexo de artefatos de predição gênica, é possível que em alguns casos os elementos possam estar realmente presentes. Para validar a incorporação de sequências de TEs em regiões codificantes, genes ortólogos com este atributo foram submetidos à busca por ESTs que comprovassem a sequência derivada de TE no transcrito. Foi realizada a busca, via execução do BLASTn ($e\text{-value} \leq 10^{-10}$), de sequências CDS preditas - potencialmente contenedoras de fragmentos de TEs - contra bibliotecas de EST depositadas para *C. reinhardtii* e *V. carteri* no NCBI (National Center of Biotechnology Information). Somente ocorrências que puderam ser associadas a um alinhamento com EST na região de junção da cópia do TE e da sequência codificante, com extensão mínima de 20 pb em cada lado da fronteira definida pela referida junção (Figura 5.2), foram consideradas para inspeções mais detalhadas.

Figura 5.2 – Representação esquemática da condição requerida para validação, via EST, da presença de TE em sequência codificante predita. A cópia de TE (retângulo marrom) no interior da sequência codificante do gene predito (retângulo azul) é confirmada quando o alinhamento com a EST (barra horizontal em preto) sobrepõe a junção do TE e da sequência CDS com uma extensão mínima de 20 pb em cada lado da fronteira definida pela referida junção. Existem três possibilidades para confirmação de TEs em CDS, ilustradas em A, B e C. Fonte: Elaborada pela autora.

Genes, cuja presença do TE em região codificante fora confirmada, tiveram sua sequência peptídica analisada quanto à identificação de domínios, utilizando-se o banco de dados de famílias de proteínas Pfam.(96) Domínios detectados na região gênica derivada do TE foram analisados, de forma mais detalhada, quanto à sua origem. Inicialmente, o software HMMER 3.0 (97) foi empregado para a busca do domínio sob análise nas sequências peptídicas dos genes ortólogos, na sequência consenso do TE relacionado ao domínio no gene e nas cópias do mesmo mapeadas no genoma; nos dois últimos casos, as sequências nucleotídicas foram examinadas nos seis possíveis quadros de leitura. O alinhamento múltiplo dos domínios reconhecidos nesta busca foi gerado utilizando-se o software Clustal X 2.1 (98) (parâmetros padrão), constituindo a entrada para a análise

filogenética^b, na qual aplicou-se o método Maximum Likelihood (99) disponível no software MEGA 5.1 (100, 101), com *bootstrap* de 1000 repetições.

5.1.3 Resultados

O número de cópias de TEs em sequências codificantes de genes ortólogos é significativamente inferior ao esperado (p -valor $< 0,01$), segundo a distribuição randômica de TEs no genoma, para as quatro classes avaliadas nas duas espécies (Tabela 5.1). A representatividade fortemente inferior à esperada pode ser compreendida considerando-se a provável disrupção na função gênica induzida pela inserção de TEs nestas regiões, o que conduz à seleção negativa destes eventos. (11, 84) Em termos gerais, esta análise revelou uma frequência de TEs mais elevada em sequências codificantes para os genes ortólogos de *V. carteri* (109 cópias) quando comparada à frequência verificada em *C. reinhardtii* (36 cópias).

Tabela 5.1 – Frequências observada e esperada para as classes de TEs na região codificante dos genes ortólogos em cada espécie.

Classe TE	Frequência observada	Frequência esperada	<i>z</i>
<i>C. reinhardtii</i>			
LTR	11	183,84 $\pm$ 12,34	-14,01
non-LTR	12	1216,96 $\pm$ 31,32	-38,47
DNA-transposon	13	641,73 $\pm$ 22,9	-27,46
outras	0	58,07 $\pm$ 6,94	-8,37
<i>V. carteri</i>			
LTR	34	619,8 $\pm$ 22,93	-25,55
non-LTR	72	811,77 $\pm$ 26,39	-28,03
DNA-transposon	0	31,66 $\pm$ 5,26	-6,02
outras	3	58,43 $\pm$ 7,21	-7,69

Fonte: Elaborada pela autora.

Estimativas para p -valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

Apesar do efeito majoritariamente deletério, em raras instâncias integrações de TEs em regiões codificantes podem modificar de forma benéfica a função original do gene, podendo ser selecionadas positivamente, ou ainda introduzir modificações neutras do ponto de vista evolutivo, o que também justifica a ausência de exclusão e permanência dos TEs

^b A filogenia pode ser definida como o estudo de relações evolutivas, o que caracteriza a análise filogenética como o meio de inferir ou estimar estas relações. (58, 99)

nestas delimitações.(11, 17, 18) Com o propósito de investigar em maiores detalhes as possíveis contribuições dos elementos móveis em sequências codificantes, foi realizada a busca por ESTs que validassem os modelos gênicos preditos que sugeriram a utilização de sequências de TEs na codificação de proteínas do genoma hospedeiro, possibilitando a análise mais rigorosa destes casos. Esta inspeção confirmou a ocorrência de TEs na região codificante de 4 genes ortólogos para *C. reinhardtii* (Figura 5.3) e 8 em *V. carteri* (Figura 5.4). É possível que o moderado número de casos validados seja consequência do número relativamente limitado de sequências de ESTs para as duas espécies. Informações sobre os domínios proteicos associados a estes genes, obtidas no banco de dados Pfam, estão concisamente apresentadas na Tabela 5.2.

Tabela 5.2 – Domínios proteicos associados aos genes que contém cópias de TEs, comprovadas via EST, na região codificante.

Identificador do gene	Domínio Pfam
<i>C. reinhardtii</i>	
Cre06.g262800	Condensin2nSMC (PF12422); PHD-finger (PF00628)
g8391	Ank_2 (PF12796)
Cre08.g383800	-
Cre10.g453250	-
<i>V. carteri</i>	
Vocar20001092m.g	zf-CCHC (PF00098); WD40 (PF00400); zf-MYND (PF01753)
Vocar20006942m.g	zf-BED (PF02892); DUF659 (PF04937)
Vocar20009410m.g	-
Vocar20002210m.g	CAP_N (PF01213); CAP_N (PF01213); CAP_C (PF08603)
Vocar20012917m.g	PAP2 (PF01569)
Vocar20010806m.g	GDPD (PF03009)
Vocar20001492m.g	Glyco_transf_90 (PF05686)
Vocar20005694m.g	-

Fonte: Elaborada pela autora.

Dentre os genes com confirmação de TE em região codificante, dois casos mostram-se particularmente interessantes porque representam contribuições em que o elemento de transposição fornece à proteína um novo domínio proteico com função conhecida. O primeiro refere-se ao gene Cre06.g262800 de *C. reinhardtii* (Figura 5.3-A). A sequência peptídica deste gene (Figura 5.5) codifica uma proteína contendo o domínio Condensin2nSMC (Pfam PF12422), relacionado à manutenção da integridade estrutural dos cromossomos.(102) Além deste, o domínio PHD-finger (Pfam PF00628) foi detectado na região do gene (19º éxon codificante) derivada do elemento Gypsy-5-LTR_CR. Domínios PHD-finger são abundantes em reguladores de cromatina; acredita-se que este domínio tenha afinidade por histonas metiladas, embora ligue-se também a nucleossomos sem modificações pós-transcricionais.(103–105)

Figura 5.3 – Representação esquemática dos genes de *C. reinhardtii* com presença de TE em sequência codificante confirmada por EST. Retângulos em marrom representam TEs e retângulos em azul representam CDSs. Linhas azuis entre CDSs indicam regiões intrônicas. Fonte: Elaborada pela autora.

Segundo o alinhamento delineado pelo BLASTn que identificou a ocorrência do elemento Gypsy-5-LTR_CR no gene Cre06.g262800, a região derivada do TE (correspondente ao intervalo 911-947 na sequência peptídica do gene, ilustrado na Figura 5.5-A) constitui parte representativa do domínio PHD-finger. O nível de identidade no alinhamento foi relativamente baixo (79,65%), indicando que a cópia fixada nesta região deve representar um membro ancestral e que, portanto, apresenta um grau elevado de divergência em relação à cópia moderna que serviu como modelo na busca.

O acúmulo de mutações na sequência nucleotídica relativa ao TE no gene pode restringir a região identificada como similar à sequência consenso do mesmo. No entanto, em nível proteico, mutações no DNA podem ser sinônimas ou silentes - nos casos em que o nucleotídeo que sofreu mutação não modifica o aminoácido especificado pelo códon - ou ainda, havendo mutações não-sinônimas, estas podem ser conservativas - quando os aminoácidos pré e pós mutação possuem propriedades químicas semelhantes.⁽⁵⁹⁾ Estas características do

Figura 5.4 – Representação esquemática dos genes de *V. carteri* com presença de TE em sequência codificante confirmada por EST. Retângulos em marrom representam TEs e retângulos em azul representam CDSs. Linhas azuis entre CDSs indicam regiões intrônicas.
 Fonte: Elaborada pela autora.

código genético permitem que, em algumas situações, o alinhamento delineado a partir de sequências peptídicas consiga identificar similaridade em sequências divergentes em nível nucleotídico, o que de fato foi constatado neste caso. Interrogando-se a sequência consenso do elemento Gypsy-5-LTR_CR contra a sequência peptídica do gene Cre06.g262800 por meio do programa BLASTx^c, o alinhamento gerado indica maior extensão derivada do TE (904-955; Figura 5.5-C), englobando a região completa do domínio PHD-finger (Figura 5.5-D).

Figura 5.5 – Domínio PHD-finger identificado na sequência peptídica do gene Cre06.g262800 em *C. reinhardtii*. (A) Representação esquemática da sequência peptídica (retângulo cinza) e dos domínios Condensin2nSMC e PHD-finger (retângulos laranja) identificados neste gene. Segundo alinhamento delineado pelo BLASTh, o segundo domínio possui uma contribuição substancial do elemento Gypsy-5-LTR_CR (retângulo hachurado). (B) Alinhamento múltiplo de sequências peptídicas correspondentes à região da assinatura do domínio PHD-finger no modelo HMM, no elemento Gypsy-5-LTR_CR e no gene Cre06.g262800. As colunas sombreadas evidenciam a assinatura do domínio PHD-finger (C4HC3), localizado na sequência peptídica do gene no intervalo de 907-951. (C) Alinhamento gerado via BLASTx, no qual a similaridade com o elemento Gypsy-5-LTR_CR foi identificada no intervalo de 904-955 na sequência peptídica do gene Cre06.g262800. (D) Representação esquemática da sequência peptídica com delimitação do TE (retângulo hachurado) segundo alinhamento gerado pelo BLASTx, indicando que o domínio PHD-finger encontra-se em sua totalidade em região derivada do elemento Gypsy-5-LTR_CR. Fonte: Elaborada pela autora.

A existência do transcrito contendo ambos os domínios - Condensin2nSMC e o PHD-finger derivado do TE - em *C. reinhardtii* foi confirmada pela reação de RT-PCR (do

^c O programa BLASTx tem como *query* a sequência nucleotídica, a qual é traduzida nos seis quadros de leitura para a busca contra o banco de sequências peptídicas.(59)

inglês, *reverse transcription polymerase chain reaction*). Por meio desta reação, a região do transcrito correspondente ao intervalo 602-2989 nt na sequência codificante do gene - no qual encontra-se a codificação dos domínios (Figura 5.6) - foi amplificada, tendo sido a molécula de cDNA resultante subsequentemente clonada, confirmando a sequência predita do transcrito produzido pelo gene Cre06.g262800. Este experimento foi realizado pela Dra. Juliana Sposto Avaca Crusca, no laboratório do Grupo de Biofísica Molecular do Instituto de Física de São Carlos.

Figura 5.6 – Representação esquemática do transcrito derivado do gene Cre06.g262800 de *C. reinhardtii*. O intervalo definido por 602-2989 nt na sequência codificante do gene (em amarelo) foi confirmado pela reação de RT-PCR, comprovando o modelo predito para o gene Cre06.g262800. Nesta delimitação estão contidos o domínio Condensin2nSMC (652-1104 nt) e o domínio localizado na região gênica derivada do elemento Gypsy-5-LTR_CR, o PHD-finger (2719-2859 nt). Fonte: Elaborada pela autora.

Ocorrência similar do elemento Gypsy-5-LTR_CR, associada ao domínio PHD-finger, foi detectada na região gênica do ortólogo em *V. carteri* (gene Vocar20002708m.g). Alinhamentos gerados pelo BLASTn e pelo tBLASTx^d identificaram similaridade entre região do 18º íntron do gene Vocar20002708m.g (scaffold_2: 3713704-3713775) e região do TE (3874-3945) (Figura 5.7-A). A região alinhada corresponde a um trecho do domínio PHD-finger, identificado via HMMER no íntron do gene (Figura 5.7-B), o qual apresenta elevada divergência em relação ao modelo HMM, o que é evidenciado pela ausência de conservação dos resíduos relativos à assinatura do domínio, mais especificamente, as cisteínas 1, 2, 6 e 7. A presença dos domínios PHD-finger em região derivada do TE nos genes ortólogos Cre06.g262800 e Vocar20002708m.g sugere que o evento de integração do elemento Gypsy-5-LTR_CR precede a divergência das duas espécies. A preservação dos resíduos correlatos à assinatura do domínio PHD-finger no gene de *C. reinhardtii* indica que o mesmo está associado à função do gene, enquanto o domínio verificado no íntron da segunda espécie mostra-se menos conservado, possivelmente devido à inexistência de funcionalidade. No entanto, a exclusão do domínio PHD-finger do transcrito derivado do gene Vocar20002708m.g não pôde ser confirmada, uma vez que não foi identificada evidência de EST que validasse o modelo gênico predito.

^d O programa tBLASTx tem como *query* a sequência nucleotídica, a qual é traduzida nos seis quadros de leitura para a busca contra o banco de sequências nucleotídicas, também traduzido nos seis quadros de leitura.(59)

Figura 5.7 – Alinhamentos relativos ao domínio PHD-finger identificado em íntron do gene Vocar20002708m.g em *V. carteri*. (A) Alinhamento gerado via tBLASTx, no qual foi identificada a similaridade entre a região do íntron 18 do gene Vocar20002708m.g (scaffold_2: 3713704-3713775) com o elemento Gypsy-5-LTR_CR (3874-3945). (B) Alinhamento múltiplo de sequências peptídicas correspondentes à região da assinatura do domínio PHD-finger no modelo HMM, no elemento Gypsy-5-LTR_CR e no gene Vocar20002708m.g. As colunas sombreadas evidenciam a assinatura do domínio PHD-finger (C4HC3). Fonte: Elaborada pela autora.

O segundo caso de presença de TE no contexto genômico de um domínio relaciona-se ao gene Vocar20001092m.g da espécie *V. carteri* (Figura 5.4). Este abriga um fragmento do elemento Gypsy3-LTR_VC no primeiro éxon codificante, o qual encontra-se associado ao primeiro domínio detectado na sequência peptídica do gene (Figura 5.8), o zf-CCHC (Pfam PF00098), um tipo de zinc finger originário da proteína retroviral gag referenciado na literatura como zinc knuckle.(106) Esta estrutura está relacionada com interações proteína-proteína, proteína-DNA e proteína-RNA. A sequência peptídica do gene contém também o domínio WD40 (Pfam PF00400), o qual está associado a uma vasta gama de processos celulares envolvendo interações proteína-proteína ou proteína-DNA (107), e o domínio zf-MYND (Pfam PF01753), existente em proteínas que atuam na regulação transcricional.(105)

Embora o alinhamento definido pelo BLASTn tenha identificado similaridade ao elemento Gypsy3-LTR_VC apenas na região final do domínio zinc knuckle no gene Vocar20001092m.g (intervalo 37-64 na sequência peptídica, indicado na Figura 5.8-A, o qual compreende o intervalo 37-42 que corresponde ao trecho final da assinatura do domínio, apontado na Figura 5.8-B), o alinhamento delineado pelo BLASTx indica similaridade entre o TE e a sequência peptídica do gene em uma delimitação que extrapola os limites do domínio (intervalo 3-64 na sequência peptídica, figuras 5.8-C e 5.8-D), corroborando a hipótese de que o zinc knuckle seja derivado do TE. O domínio zinc knuckle, proveniente do elemento Gypsy3-LTR_VC, não está presente no gene ortólogo em *C. reinhardtii* (g12386), o que sugere que o evento de inserção do TE é posterior à especiação das espécies.

Joly-Lopez *et. al* (42) apontam que, devido à habilidade de transposição, as sequências de TEs possuem a capacidade de codificação de produtos com propriedades moleculares (como aquelas correlatas a interações com DNA ou proteínas), o que constitui uma fonte

Figura 5.8 – Domínio zinc knuckle (zf-CCHC) identificado na sequência peptídica do gene Vocab20001092m.g em *V. carteri*. (A) Representação esquemática da sequência peptídica (retângulo cinza) e dos domínios zf-CCHC, WD40 e zf-MYND (retângulos em laranja) identificados neste gene. Segundo alinhamento delineado pelo BLASTn, o primeiro domínio possui uma contribuição do elemento Gypsy3-LTR_VC (retângulo hachurado). (B) Alinhamento múltiplo de sequências peptídicas correspondentes à região da assinatura do domínio zinc knuckle no modelo HMM, no Gypsy3-LTR_VC e no gene Vocab20001092m.g. As colunas sombreadas evidenciam a assinatura do domínio zinc knuckle (C2HC), localizado na sequência peptídica do gene no intervalo de 29-42. Os resíduos indicados pela linha contínua (intervalo 37-42 na sequência peptídica) referem-se à região da assinatura do domínio no gene atribuída ao TE, segundo alinhamento gerado no BLASTn; a linha descontínua indica região da assinatura equivalente nas duas sequências, porém não compreendida no alinhamento. (C) Alinhamento gerado via BLASTx, no qual a similaridade com o elemento Gypsy3-LTR_VC foi identificada no intervalo de 3-64 na sequência peptídica do gene Vocab20001092m.g. (D) Representação esquemática da sequência peptídica com delimitação do TE (retângulo hachurado) segundo alinhamento gerado pelo BLASTx. Fonte: Elaborada pela autora.

de material genético que pode ser adaptada para a execução de novas funções. Neste contexto, o gene Cre06.g262800 pode constituir um exemplo de exaptação de uma sequência proveniente de um TE (neste caso, o domínio PHD-finger) que, agregada ao contexto gênico prévio à inserção, pode ter aprimorado o potencial de interação da proteína codificada pelo referido gene com o DNA. Lembrando que o primeiro domínio (Condensin2nSMC) é correlato a uma família proteica cromossômica, torna-se possível inferir ainda que este gene pode representar também um exemplo de contribuição de TEs na evolução de proteínas cromossômicas, cuja relevância já fora reportada na literatura.(108)

Em relação ao segundo caso de contribuição de TE em um domínio proteico, apesar de não ter sido possível identificar a função precisa do gene Vocab20001092m.g, a existência do domínio WD40 indica que o produto gênico pode ser uma proteína que interage em

diversos processos celulares. Devido a este fato, também este pode ser um exemplo de aperfeiçoamento da afinidade de interação da proteína.

Na tentativa de investigar mais profundamente a origem dos domínios zinc knuckle e PHD-finger identificados nos genes em região derivada de TE, foi realizado o estudo filogenético dos mesmos. Para cada domínio avaliado, a assinatura HMM correspondente foi interrogada junto às sequências peptídicas dos genes ortólogos, à sequência consenso do TE relacionado ao domínio no gene e às cópias do elemento no genoma.

Figura 5.9 – Filogenia dos domínios zinc knuckle (C2HC) identificados no elemento Gypsy3-LTR_VC e nos genes ortólogos das espécies *C. reinhardtii* e *V. carteri*. O domínio zinc knuckle foi identificado, via busca HMM (código de acesso HMM para C2HC: PF00098.18) em seis genes para *C. reinhardtii*, em oito genes para *V. carteri*, na sequência canônica do TE Gypsy3-LTR_VC e em duas cópias deste elemento no genoma de *V. carteri*. A árvore filogenética foi inferida pelo método *maximum likelihood* no software MEGA, com *bootstrap* de 1000 repetições. Em destaque (cor laranja), domínios relativos à sequência consenso do TE, cópias do mesmo no genoma e ao gene Vocar20001092m.g de *V. carteri*. Fonte: Elaborada pela autora.

No caso do domínio zinc knuckle (Figura 5.9), foi possível confirmar a relação filogenética devido à formação de um ramo monofilético (*bootstrap* igual a 88%) entre os domínios presentes no retrotransposon Gypsy3-LTR_VC e no gene Vocar20001092m.g, confirmando a proximidade evolutiva dos referidos domínios. No caso do domínio PHD-finger, apesar do agrupamento dos domínios relativos ao elemento Gypsy-5-LTR_CR e ao gene Cre06.g262800 no mesmo ramo filogenético, não foi possível confirmar a proximidade

destas sequências com alto suporte estatístico, uma vez que o valor de *bootstrap* obtido foi igual a 35%.

Mesmo nos genes representados pictoricamente nas figuras 5.3 e 5.4, que possuem cópia de TE em sua sequência codificante porém não constituinte de domínio propriamente dito, foi possível identificar a introdução de trechos codificantes, códons de iniciação/terminação de tradução ou ainda sítios de *splicing*, sugerindo que estas inserções podem ter contribuído de alguma forma para a evolução da estrutura gênica.

Ainda no contexto das cópias de TEs verificadas em sequências codificantes, uma observação interessante remete à posição - em relação à terminação mais próxima do gene - ocupada pelo éxon contendo o fragmento do elemento. O padrão emergente da distribuição dos valores assumidos por esta variável (Figura 5.10) expressa uma prevalência de eventos de integração de TEs nas extremidades da sequência codificante dos genes, provavelmente devido ao menor risco de disrupção dos domínios proteicos pré-existentes nos mesmos. Este padrão é observado para o conjunto de genes com validação por EST e também para o conjunto de genes sem a referida validação nas duas espécies, o que corrobora a confiabilidade dos modelos gênicos preditos.

Figura 5.10 – Distribuição de frequências da posição (em relação à extremidade gênica mais próxima) ocupada pelo éxon contendo o fragmento de TE. Distribuições para o conjunto de casos com evidência EST da presença de TE na região codificante (painel à esquerda) e para o conjunto de casos sem a referida confirmação (painel à direita) para a espécie *C. reinhardtii* (A) e *V. carteri* (B). Fonte: Elaborada pela autora.

5.2 Região intrônica

5.2.1 Introdução

Apesar de constituírem uma grande parte dos genomas, pouco é conhecido a respeito da importância funcional dos íntrons.(109) De certa forma, os mesmos podem influenciar a diversidade de proteínas codificadas por um gene devido ao processo de *splicing* alternativo, por meio do qual o conjunto de éxons remanescentes no mRNA maduro pode ser modificado.(36, 74) A Figura 5.11 ilustra o processo de *splicing* de um íntron em um pré-mRNA.

Figura 5.11 – Mecanismo canônico de *splicing* e snRNAs (*small nuclear RNAs*) envolvidos no mesmo. O pareamento de bases dos snRNAs U1 e U6 ao sítio de *splicing* 5' do íntron e do snRNA U2 ao ponto da forquilha (nucleotídeo adenina indicado por *) estão representados na parte superior da figura.(110) Os círculos preenchidos em preto, no final dos snRNAs, indicam a estrutura cap na terminação 5' destas moléculas. As setas apontam o curso do ciclo de *splicing*, bem como o acoplamento e liberação dos snRNAs em cada etapa do mesmo. Apesar de não representado na figura, é importante lembrar que cada snRNA agrega-se a, no mínimo, sete subunidades protéicas, compondo assim as snRNPs (*small nuclear ribonucleoproteins*). (36) Estas formações constituem a unidade básica do spliceossomo, o grande complexo molecular de RNA e proteínas que promove o *splicing*. De forma sucinta, após a interação do pré-mRNA com os snRNAs U1, U2, U4, U5 e U6 (A e B), ocorre a excisão da região terminal 5' do íntron (C), que passa a ser covalentemente ligada ao nucleotídeo adenina do ponto da forquilha (D), formando uma alça. A extremidade 3' livre do éxon remanescente após o corte reage com o início do éxon adjacente à extremidade 3' do íntron (E), unindo os dois éxons em uma sequência contínua.(74) Fonte: LORKOVIĆ;KIRK;LAMBERMON;FILIPOWICZ(110)

Outro aspecto relativo à funcionalidade destas estruturas, que vem se tornando cada vez mais evidente, relaciona-se à capacidade de regulação da expressão gênica. Este potencial pode ser manifestado por estruturas regulatórias inseridas nos íntrons ou por sequências intrônicas aptas à formação de estruturas secundárias no pré-mRNA, que podem influenciar o processo de *splicing*.(109)

O gene *regA* de *V. carteri* representa um exemplo muito interessante de regulação gênica mediada por sequências regulatórias localizadas em íntrons. De acordo com o apresentado na Seção 2.2, o mesmo desempenha um importante papel na diferenciação celular nesta espécie, atuando nas células somáticas de forma a reprimir o desenvolvimento reprodutivo. Este gene possui elementos regulatórios localizados nos íntrons 3, 5 e 7 (111); em células somáticas, os íntrons 3 e 5 ligam-se a fatores que cooperam para ativar o promotor do gene; nas células reprodutivas, também o íntron 7 liga-se a um fator específico de célula gonídio, que tem como efeito evitar a ativação do promotor do gene. Este mecanismo apresenta grande relevância para a questão correlata à diferenciação celular em *V. carteri*, e consequentemente, à questão da multicelularidade neste modelo de estudo.

Já no contexto de regulação via formação de estruturas secundárias do pré-mRNA, cita-se o exemplo do gene *Adh* da espécie *D. melanogaster*.(112) Este possui em seu primeiro íntron uma sequência que, quando transcrita, origina uma estrutura secundária *hairpin*. O estudo promovido por Chen e Stephan (112) identificou que mutações que induzem a disrupção desta estrutura têm como efeito a diminuição da eficiência do processo de *splicing* do pré-mRNA e, consequentemente, da expressão gênica. Os autores do trabalho apontam que a estrutura *hairpin*, localizada na vizinhança 5' do ponto da forquilha, deve facilitar o pareamento de bases entre o snRNA (*small nuclear RNAs*) U2 e a sequência do ponto da forquilha no íntron, aumentando a eficiência do processo de *splicing*.

Exemplos desta natureza conduzem à reflexão de que parâmetros como constituição e comprimento da sequência do íntron podem apresentar relação com a regulação da expressão gênica. Apesar do acúmulo de evidências em torno da importância funcional dos íntrons, Haddrill et al. (109) apontam que ainda não foram estabelecidos os pontos restritivos que regem sua evolução.

No aspecto evolutivo, já fora abordado neste trabalho os possíveis efeitos de inserções de TEs nos íntrons (ver Figura 2.4). Zhang et al. (40) afirmam que o conhecimento acerca da distribuição de TEs no interior dos íntrons ainda é bastante limitado, assim como o motivo pelo qual alguns elementos no contexto intrônico perturbam a transcrição do gene enquanto outros não o fazem. Neste cenário, o intuito desta seção é apresentar alguns resultados advindos do estudo da distribuição de TEs nos íntrons dos genes ortólogos de *C. reinhardtii* e *V. carteri*.

5.2.2 Métodos

Distribuição de TEs nos íntrons

A frequência observada de cópias de TEs nos íntrons foi comparada à esperada - segundo a distribuição randômica de TEs no genoma - de forma análoga ao descrito na Seção 4.1.2, permitindo o estabelecimento da estatística z para as classes de TEs neste contexto genômico.

Com o objetivo de investigar detalhadamente a distribuição de TEs ao longo da extensão dos íntrons, as cópias intrônicas, para as classes de TEs mais abundantes em cada espécie, foram analisadas quanto à distância em relação ao éxon mais próximo. Resumidamente, dada uma inserção em um íntron, esta é indexada à classe de distância na qual localiza-se seu início, considerando-se como referencial o éxon mais próximo. Desta forma, cópias localizadas nas terminações 5' e 3' do íntron são atribuídas às primeiras classes de distância, como realizado no trabalho de Zhang et. al.(40) Diante destas considerações, torna-se possível a determinação das frequências observada e esperada em cada classe de distância, conduzindo, finalmente, ao valor da frequência padronizada que expressa o nível de representatividade dos TEs nos compartimentos genômicos analisados. As classes de distância mais próximas aos éxons foram definidas com menor comprimento, quando comparadas às mais distantes, com o objetivo de realizar uma inspeção com maior resolução nestas regiões. A última classe de distância contempla regiões intrônicas com distâncias superiores a 400 pb em relação ao éxon mais próximo. Este valor foi adotado em função do comprimento médio dos íntrons nas duas espécies (*C. reinhardtii*: 174 pb; *V. carteri*: 358 pb (27)).

O aspecto relacionado à orientação das cópias de TEs em relação ao gene no qual residem também foi avaliado. Neste ensaio, a frequência padronizada relaciona-se à razão entre a frequência de cópias com orientação senso e a frequência de cópias com orientação antissenso (Equação 5.1). Dessa forma, valores positivos desta variável refletem a predominância de orientação senso, enquanto valores negativos a predominância de orientação antissenso, na classe de distância avaliada.

$$f_p = \log_{10}(f_{senso}/f_{antissenso}) \quad (5.1)$$

Considerando-se a probabilidade de orientação senso como sendo igual à probabilidade de orientação antissenso ($p = 0,5$) para uma determinada cópia de TE, é possível definir, em um conjunto de n cópias, se a frequência de orientação senso difere com significância estatística do que seria esperado mediante a igualdade de probabilidade dos eventos senso e antissenso. O p-valor pôde ser aferido utilizando-se como estrutura probabilística o

modelo Binomial, mais especificamente, a função teste binomial exata^e no R^f.

Ainda no contexto da distribuição das cópias de TEs nos íntrons, outro ponto tangente a esta questão relaciona-se à posição do íntron - que abriga o elemento - em relação ao gene. Para explorar este quesito, os genes que abrigam TEs em seus íntrons foram agrupados em conjuntos de acordo com o número de íntrons que possuem. A partir desta organização, realizou-se a contagem do número de cópias verificadas em cada íntron, para cada conjunto de genes. Por meio desta abordagem, torna-se possível uma representação matricial em que o número de íntrons do gene é expresso pelo índice da linha na matriz e a posição do íntron no gene é expresso pelo índice da coluna na matriz. Deste modo, o valor armazenado na célula $m[i,j]$ refere-se à frequência de cópias de TEs identificadas no j -ésimo íntron de genes com i íntrons. Finalmente, resta estabelecer ainda que a frequência que será armazenada em cada posição na matriz será a relativa, ou seja, obtida pela razão entre o número de cópias atribuídas ao íntron e o número total de cópias intrônicas contempladas na análise (estas não participam em sua totalidade porque genes com TEs e um número de íntrons maior do que 40 foram desconsiderados). O mapa de cores (Figura 5.12) simboliza a matriz de frequências relativas. Nesta ilustração, as linhas da matriz encontram-se representadas no sentido ascendente do eixo vertical do mapa, com o objetivo de que seja mantida a convenção mais usual (ordem crescente dos valores no sentido ascendente do eixo).

Figura 5.12 – Exemplo de mapa de cores para frequência relativa de TEs em íntrons de genes ortólogos. Valores no eixo y expressam o número de íntrons no gene e valores no eixo x a posição do íntron do gene. A frequência relativa de TEs é computada em relação ao número total de cópias considerado nesta análise. Células em branco refletem frequência nula. Fonte: Elaborada pela autora.

^e Disponível em: <<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/binom.test.html>>.

^f Informações sobre a linguagem de programação estatística R: <<http://www.r-project.org/>>.

Distâncias entre cópias de TEs

As famílias de TEs mais frequentes nos íntrons foram examinadas em termos da distância entre suas cópias. Cada família sob análise teve suas cópias divididas em dois conjuntos: no primeiro, encontram-se as ocorrências anotadas em regiões intrônicas; no segundo, as cópias não-intrônicas (genômicas exceto as intrônicas). Inserções designadas por mais de um alinhamento foram representadas pelo alinhamento de menor *e-value*, evitando que segmentos genômicos não homólogos ao TE sob estudo fossem considerados na análise.

Para cada conjunto avaliado, foi gerado o alinhamento múltiplo das sequências nucleotídicas - relativas às cópias do TE - utilizando-se o software Clustal X 2.1 (parâmetros padrão). Este alinhamento constituiu a entrada para o software MEGA, no qual foram definidas as distâncias par a par entre as cópias empregando-se o modelo Tajima-Nei (113) (parâmetros padrão e *pairwise deletion of missing data*), como realizado no trabalho de Venancio et al.(19) Os dados armazenados na matriz de distâncias são descritos, segundo a distribuição de frequências dos mesmos, por um histograma.

5.2.3 Resultados

Distribuição de TEs nos íntrons

O padrão de maior frequência para a classe non-LTR, seguida pela DNA transposon, observado no genoma de *C. reinhardtii* foi mantido nos íntrons (Tabela 5.3). Também a espécie *V. carteri* preservou nesta região o padrão observado no genoma como um todo: maior frequência para a classe non-LTR, seguida pela LTR. Todas as classes, em ambas as espécies, apresentaram frequência de ocorrência nos íntrons significativamente abaixo do esperado ($p\text{-valor} < 0,01$) segundo uma distribuição randômica de TEs no genoma.

Por meio dos dados apresentados nas tabelas 5.3 e 3.1, é possível definir que 10,59% das cópias de TEs em *C. reinhardtii* e 9,46% em *V. carteri* residem em íntrons de genes ortólogos. Considerando-se que o conjunto de ortólogos constitui 59,42% e 59,40% do total de genes codificadores de proteínas em cada espécie[§], é possível estimar o número de cópias de TEs localizadas em íntrons, o que corresponderia a $\sim 18\%$ em *C. reinhardtii* e $\sim 16\%$ em *V. carteri*. Estes índices são moderados quando comparados aos verificados em outras espécies de vertebrados mamíferos (*H. sapiens* $\sim 60\%$) e não-mamíferos (*G. gallus* - 33,2%; *D. rerio* - 47,3%), bem como de invertebrados: *D. melanogaster* (60%) e

[§] Número de genes codificadores de proteínas em cada espécie: 14.516 em *C. reinhardtii* e 14.520 em *V. carteri*.(27)

Tabela 5.3 – Frequências observada e esperada para as classes de TEs na região intrônica dos genes ortólogos em cada espécie.

Classe TE	Frequência observada	Frequência esperada	z
<i>C. reinhardtii</i>			
LTR	74	207,2 ± 12,89	-10,33
non-LTR	1070	1369,38 ± 32,64	-9,17
DNA-transposon	156	722,37 ± 23,9	-23,7
outras	14	65,34 ± 7,29	-7,04
<i>V. carteri</i>			
LTR	219	1046,03 ± 28,41	-29,11
non-LTR	885	1369,32 ± 32,17	-15,06
DNA-transposon	1	53,4 ± 6,47	-8,1
outras	1	98,6 ± 8,8	-11,09

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

C. elegans (53%).(68)

Fatores proibitivos associados à fixação de TEs em íntrons, como a interrupção do padrão de *splicing* e de estruturas regulatórias localizadas nestas regiões, podem explicar esta subrepresentação dos elementos. Além disso, outro aspecto decorrente da integração de TEs que pode conduzir a efeito deletério é a alongação dos íntrons. Levy et al. citam em seu trabalho que elementos L1, inseridos em íntrons, decrescem significativamente a expressão de genes em mamíferos devido à alongação inadequada do transcrito.(16) Em genomas eucariotos inferiores, em que é predominante o modelo de *splicing* definido no íntron (114) (Figura 5.13), é possível que a alongação destas estruturas seja ainda mais crítica.

As distribuições de frequências padronizadas de TEs ao longo da extensão dos íntrons, para as duas classes de TEs mais abundantes em cada espécie, são ilustradas nas figuras 5.14-B e 5.14-D. Como pode ser observado, para *C. reinhardtii* (figuras 5.14-A e 5.14-B) a classe non-LTR exibe frequência significativamente abaixo do esperado ($p\text{-valor} < 0,01$) nos intervalos de 1 a 20 pb, 51 a 200 pb (3ª e 4ª classes de distância) e em regiões cuja distância ao éxon mais próximo assume valores maiores do que 400 pb. Nos intervalos de 21 a 50 pb e 201 a 400 pb, a frequência dos TEs encontra-se muito próxima do esperado. A classe DNA-transposon, por sua vez, apresenta frequência significativamente inferior ao esperado ($p\text{-valor} < 0,01$) no intervalo de 1 a 400 pb, especialmente nas primeiras classes de distância. Quanto à espécie *V. carteri* (figuras 5.14-C e 5.14-D), pôde-se verificar que ambas as classe non-LTR e LTR exibem uma distribuição de frequências padronizadas

Figura 5.13 – Representação esquemática dos modelos de reconhecimento do pré-mRNA definidos no íntron e no éxon. O pareamento de fatores de *splicing* (a exemplo do U2AF) e snRNPs (apenas U1 encontra-se indicado na figura) é muito semelhante nas perspectivas orientadas ao íntron e ao éxon.(74, 114) A figura à esquerda ilustra o modelo definido no íntron, no qual arranjos envolvendo ribonucleoproteínas e proteínas estabelecem contatos entre os sítios de *splicing* ao longo do íntron, quando os éxons são separados por íntrons curtos; à direita, o modelo definido no éxon, em que a comunicação entre os sítios de *splicing* ocorre ao longo dos éxons quando estes são separados por íntrons longos.(74) Segundo Berget (114), o primeiro modelo é predominante em eucariotos inferiores, enquanto o segundo o seria para vertebrados. Informações mais detalhadas acerca dos modelos podem ser verificadas na literatura.(74, 110, 114) Fonte: CONTI;BARALLE;BURATTI(74)

significativamente abaixo do esperado em toda a extensão intrônica.

É interessante o fato de que as classes DNA-transposon em *C. reinhardtii*, non-LTR e LTR em *V. carteri* exibem representação fortemente diminuída em regiões intrônicas adjacentes ao éxon (1 a 100 pb); no caso da classe non-LTR em *C. reinhardtii*, a região de maior restrição à ocorrência de TEs restringe-se ao intervalo de 1 a 20 pb. De forma geral, o padrão emergente expressa uma elevação na frequência padronizada de TEs à medida que a distância em relação ao éxon aumenta, embora seja ainda negativa. Este padrão é coerente, uma vez que a maioria das sequências sinalizadoras de *splicing* localizam-se nas terminações 5' e 3' dos íntrons (Figura 5.11). Assim sendo, inserções nas adjacências dos éxons apresentam grande potencial de disrupção do padrão de *splicing* do gene, o que as torna alvo de seleção purificadora. Resultados semelhantes foram encontrados em genomas de mamíferos (humano e camundongo).(40)

A orientação das cópias de TEs, em relação ao gene no qual residem, foi analisada utilizando-se o conceito de frequência padronizada, associada neste caso à razão entre o número de cópias com orientação senso e o número de cópias com orientação antissenso. Para retrotransposons non-LTR, em ambas as espécies (Figura 5.15), a orientação das cópias mostrou uma prevalência na orientação antissenso nas classes de distância analisadas, sendo esta tendência mais evidente - e significativa estatisticamente (Tabela 5.4) - nas classes de distância próximas ao éxon. A baixa ocorrência de TEs non-LTR com orientação senso nas proximidades do éxon sugere que elementos nesta configuração possam ser prejudiciais ao processo de *splicing*, embora não seja possível precisar a causa deste efeito. Em relação aos elementos DNA-transposon em *C. reinhardtii* (Figura 5.15-A) e LTR em *V. carteri* (Figura 5.15-B), observa-se uma predominância de frequências padronizadas positivas que

Figura 5.14 – Distribuição de frequências de TEs nos íntrons de genes ortólogos em *C. reinhardtii*. São analisadas as classes de TEs mais abundantes em cada espécie. Em cada classe de distância, são apresentadas a frequência observada (barra da esquerda) e a frequência média obtida na simulação (barra da direita, com coloração mais suave) para a espécie *C. reinhardtii* (A) e *V. carteri* (C). As barras de erro correspondem ao desvio padrão da frequência de TEs obtida na simulação. Classes de TEs, cuja frequência observada encontra-se indicada com *, exibem $|z| > 2,3263$ ($p < 0,01$) na região analisada. Os valores de z , relativos a cada classe de distância, encontram-se na Tabela B.5 para *C. reinhardtii* e na Tabela B.6 para *V. carteri*, ambas localizadas no Apêndice B. Distribuições de frequências padronizadas para *C. reinhardtii* (B) e *V. carteri* (D): a frequência padronizada esperada, de acordo com a distribuição randômica de cópias de TEs no genoma, é representada pela linha em vermelho. Barras de erro representam a incerteza da frequência padronizada, definida pela propagação do desvio padrão da frequência esperada.(88) Fonte: Elaborada pela autora.

expressam orientação senso das cópias, no entanto sem significância estatística.

Figura 5.15 – Orientação senso e antissenso de TEs nos íntrons de genes ortólogos. Valores no eixo y expressam o logaritmo da razão entre o número de cópias de TEs na orientação senso e o número de cópias de TEs na orientação antissenso. Distribuições das duas maiores classes para *C. reinhardtii* (A) e *V. carteri* (B) são apresentadas separadamente. P-valores, relativos à cada classe de distância, podem ser verificados na Tabela 5.4. Fonte: Elaborada pela autora.

A distribuição das cópias de TEs nos íntrons dos genes ortólogos em *C. reinhardtii* (Figura 5.16-A) evidencia uma tendência de ocupação dos íntrons mais próximos ao término do gene, o que pode ser observado por valores mais elevados de frequência relativa em células mais próximas da diagonal no mapa. Sob outra perspectiva, a tendência mais evidente, de fato, refere-se aos índices mais brandos de frequência de TEs no primeiro íntron dos genes avaliados. Possivelmente, este padrão deve-se ao fato de que íntrons que abrigam elementos regulatórios tendem a ser os primeiros no gene (109), o que os torna muito suscetíveis à disrupção de sua função mediante um evento de inserção. Provavelmente, devido a esta conjectura, a fixação das cópias de TEs nestas regiões deve ter sido contida pela ação da pressão evolutiva seletiva. Padrão muito similar é verificado para a espécie *V. carteri* (Figura 5.16-B). A baixa ocorrência de TEs em genes com elevado número de íntrons nas duas espécies deve-se ao fato de que genes com esta característica não são frequentes, uma vez que o número médio de íntrons por gene é 7,4 para *C. reinhardtii* e 7,05 para *V. carteri*.(27)

Tabela 5.4 – P-valores obtidos no teste binomial exato para orientação das cópias de TEs, nas classes de distância em íntrons de genes ortólogos. Dados relativos às duas classes de TEs mais frequentes em íntrons para cada espécie. Valores em negrito são significativos (p-valor < 0,01).

Distância ao éxon mais próximo (pb)										
	f_{senso}	n	r	p-valor	f_{senso}	n	r			p-valor
<i>C. reinhardtii</i>										
	non-LTR				DNA					
1-20	38	124	0,31	$9,73 \times 10^{-6}$	4	8	0,5	0,64		
21-50	119	332	0,36	$1,40 \times 10^{-7}$	17	30	0,57	0,29		
51-100	107	274	0,39	$1,74 \times 10^{-4}$	19	31	0,61	0,14		
101-200	69	222	0,31	$8,66 \times 10^{-9}$	21	50	0,42	0,16		
201-400	37	93	0,40	0,03	12	20	0,6	0,25		
>400	12	25	0,48	0,5	8	17	0,47	0,5		
<i>V. carteri</i>										
	non-LTR				LTR					
1-20	7	30	0,23	$2,61 \times 10^{-3}$	2	6	0,33	0,34		
21-50	29	112	0,26	$1,65 \times 10^{-7}$	13	17	0,76	0,02		
51-100	79	186	0,42	0,02	27	42	0,64	0,04		
101-200	128	291	0,44	0,02	33	55	0,6	0,09		
201-400	87	188	0,46	0,17	31	58	0,53	0,35		
>400	36	78	0,46	0,29	21	41	0,51	0,5		

Fonte: Elaborada pela autora.

Para cada classe de distância são apresentados valores relativos ao número de cópias com orientação senso (f_{senso}), ao número total de cópias (n), à razão entre o número de cópias senso e o número total de cópias (r) e ao p-valor obtido no teste binomial exato. Para a realização do teste são considerados a f_{senso} , n e a probabilidade de sucesso - orientação senso - igual a 0,5, o que representa igual probabilidade de ocorrência para eventos de integração senso e antisenso.

Figura 5.16 – Mapa de cores para a frequência relativa de cópias de TEs em íntrons de genes ortólogos. A frequência relativa em cada célula do mapa foi obtida a partir do número de cópias identificadas nos íntrons - cuja posição no gene é representada pelo índice da respectiva célula - e do número de cópias intrônicas que participaram da análise. Células em branco representam frequência relativa nula. Participam desta análise 1288 cópias de TEs para *C. reinhardtii* (A) e 1087 cópias para *V. carteri* (B); as demais 26 (1,98%) e 19 (1,72%) cópias intrônicas em cada espécie, respectivamente, não estão contempladas porque se localizam em íntrons cujo gene possui mais de 40 íntrons. Fonte: Elaborada pela autora.

Contribuição das cópias de TEs na extensão e no conteúdo GC dos íntrons

O percentual de íntrons - relativos aos genes ortólogos - com incidência de TEs em *C. reinhardtii* é 1,39% e 1,25% em *V. carteri*. Novamente, pôde-se observar que estes índices são modestos quando comparados aos verificados em vertebrados mamíferos (*H. sapiens* - 63,40%; *M. musculus* - 60,20%) e não-mamíferos (*G. gallus* - 21,30%; *D. rerio* - 44,30%).(68) Na verdade, os índices encontrados para as espécies sob estudo aproximam-se mais dos verificados para duas espécies de invertebrados: *D. melanogaster* (1,70%) e *C. elegans* (5,60%).(68)

Analisando diferentes espécies, Sela et al. (68) verificaram em seu estudo uma correlação positiva entre o tamanho médio dos íntrons e a fração de íntrons com incidência de TEs. Segundo os autores, a transição evolutiva entre o mecanismo de *splicing* definido no íntron e definido no éxon teria permitido um relaxamento na pressão seletiva sobre o comprimento dos íntrons. Por este motivo, a modificação do modelo de reconhecimento do pré-mRNA teria possibilitado inserções de TEs em íntrons com menor potencial deletério, o que pode ter contribuído para a maior fração de íntrons abrigando TEs em vertebrados.

Com o objetivo de avaliar a influência de TEs no comprimento e na composição das sequências intrônicas de *C. reinhardtii* e *V. carteri*, foram comparados pares de íntrons ortólogos. São definidos como ortólogos os íntrons que residem em genes ortólogos com idêntica quantidade de íntrons.

Do total de 8.625 genes ortólogos, apenas 2.109 (24,45%) apresentam a mesma quantidade de íntrons. Este, portanto, é o número de genes que participam desta análise, contribuindo com um total de 12.180 íntrons em cada espécie. Foram estabelecidos três conjuntos de dados: o primeiro engloba pares de íntrons em *C. reinhardtii* e em *V. carteri* que não possuem cópias de TEs em seu interior; já o segundo grupo contempla pares em que o íntron de *C. reinhardtii* abriga TEs, ao passo de que seu par em *V. carteri* não; finalmente, no terceiro conjunto encontram-se os pares cujo íntron relativo a *V. carteri* abriga TEs, enquanto o de *C. reinhardtii* não. Os três conjuntos possuem, respectivamente, 11.903, 138 e 137 pares de íntrons, sendo dois pares não contemplados em nenhum dos conjuntos por abrigarem cópias de TEs em ambos os íntrons. A definição destes conjuntos deriva da anotação de 147 cópias intrônicas de TEs para *C. reinhardtii* e 144 para *V. carteri*.

A distribuição de comprimentos dos íntrons no conjunto com ausência de cópias de TEs exibe um padrão muito semelhante nas duas espécies (Figura 5.17-A), assim como no conjunto de pares em que apenas o íntron relativo a *C. reinhardtii* apresenta ocorrência de TE (Figura 5.17-B). Apesar disso, é possível notar que íntrons de *V. carteri* tendem a ser maiores que os de *C. reinhardtii* no primeiro conjunto analisado, tendência que é invertida quando considera-se o conjunto de íntrons de *C. reinhardtii* com cópias de TEs.

Figura 5.17 – Distribuição de comprimentos de íntrons em três conjuntos de dados: pares de íntrons sem cópia de TEs nas duas espécies (A), pares em que apenas o íntron de *C. reinhardtii* apresenta cópia de TE (B) e pares em que apenas o íntron de *V. carteri* abriga TE (C). (A) *C. reinhardtii*: média = $260,8 \pm 250,4$ pb e mediana = 226 pb (mínimo: 38 pb, máximo: 20.300 pb); *V. carteri*: média = $360,6 \pm 269,3$ pb e mediana = 314 pb (mínimo: 5 pb, máximo: 3003 pb). (B) *C. reinhardtii*: média = $649,3 \pm 428,5$ pb e mediana = 501 pb (mínimo: 97 pb, máximo: 2642 pb); *V. carteri*: média = $422,3 \pm 320,2$ pb e mediana = 349,5 pb (mínimo: 4 pb, máximo: 2649 pb). (C) *C. reinhardtii*: média = $262,3 \pm 111,1$ pb e mediana = 240 pb (mínimo: 73 pb, máximo: 562 pb); *V. carteri*: média = $910,4 \pm 525,9$ pb e mediana = 786 pb (mínimo: 148 pb, máximo: 2674 pb). Fonte: Elaborada pela autora.

A semelhança no padrão das distribuições não é verificada no terceiro conjunto avaliado, em que somente íntrons relativos a *V. carteri* abrigam cópias de TEs (Figura 5.17-C), para o qual é evidente a mudança na distribuição dos dados em virtude da maior frequência de íntrons mais longos para *V. carteri*. Se comparadas as medianas associadas às distribuições em cada espécie, dentro de cada conjunto (dados apresentados na legenda da Figura 5.17), estabelecendo-se a distância entre as mesmas, percebe-se que o terceiro conjunto analisado apresenta a maior amplitude de distância (88 pb, ~ 152 pb e 546 pb, respectivamente), refletindo o deslocamento da distribuição para maiores comprimentos de íntrons.

Estes dados sugerem que as inserções de TEs podem ter contribuído para um aumento

do comprimento dos íntrons nas duas espécies, sendo este efeito mais evidente nos íntrons de *V. carteri*, provavelmente, devido à maior extensão média das cópias de TEs no genoma desta espécie (ver Figura 3.8).

Para averiguar o efeito da presença de TEs no conteúdo GC dos íntrons, realizou-se uma amostragem em que foram selecionados aleatoriamente 138 pares de íntrons do conjunto com ausência de TEs nas duas espécies. Os pares ortólogos de cada conjunto foram avaliados segundo a razão entre o conteúdo GC do íntron em *C. reinhardtii* e o conteúdo GC de seu par em *V. carteri*. Como o percentual GC do genoma da primeira espécie é maior do que o da segunda (64% e 56%, respectivamente (27)), é esperada uma predominância de razões superiores a um.

Figura 5.18 – Box-plot da distribuição de razões de conteúdo GC para pares de íntrons ortólogos em *C. reinhardtii* e *V. carteri*. Estatísticas para amostra do conjunto de pares de íntrons com ausência de TEs em ambas as espécies (n = 138): média = $1,13 \pm 0,17$ e mediana = 1,13 (mínimo: 0,11, máximo: 1,76); conjunto de pares com presença de TEs em íntrons de *C. reinhardtii*: média = $1,13 \pm 0,28$ e mediana = 1,11 (mínimo: 0,83, máximo: 3,73); conjunto de pares com presença de TEs em íntrons de *V. carteri*: média = $1,15 \pm 0,13$ e mediana = 1,15 (mínimo: 0,88, máximo: 1,63). P-valor = 0,03 (teste de Kruskal Wallis). *Out-liers* não são mostrados. Fonte: Elaborada pela autora.

A Figura 5.18 retrata a distribuição das razões obtidas para os pares de íntrons ortólogos nos três conjuntos analisados. Apesar de apresentarem-se distintas do ponto de vista estatístico (p-valor = 0,03; teste de Kruskal-Wallis^h), as referidas distribuições exibem diferenças muito sutis entre suas medianas. Comparando-se o conjunto em que os íntrons de *C. reinhardtii* possuem cópias de TEs com o conjunto em que os pares de íntrons não possuem, percebe-se um menor valor para a mediana (1,11 e 1,13, respectivamente). Este resultado pode traduzir uma tendência de que a inserção de sequências de TEs nos íntrons

^h O teste não paramétrico Kruskal-Wallis foi utilizado para avaliar se há evidência estatística de que as medianas das distribuições de razões entre conteúdos GC, nos conjuntos avaliados, diferem. O mesmo possui função implementada no R para sua aplicação (<<http://www.r-tutor.com/elementary-statistics/non-parametric-methods/kruskal-wallis-test>>).

de *C. reinhardtii* tenha reduzido o percentual GC dos mesmos, o que diminui a razão. Por outro lado, comparando-se o conjunto em que os íntrons de *V. carteri* possuem cópias de TEs com o conjunto em que os pares de íntrons não possuem, nota-se uma elevação no valor da mediana (1,15 e 1,13, respectivamente). Novamente, a introdução de sequências relativas a TEs pode ter causado a diminuição do conteúdo GC nos íntrons de *V. carteri*, o que resulta na elevação dos valores da razão. Apesar da verificação de uma possível tendência de redução no conteúdo GC nos íntrons desencadeada pela incorporação de sequências de TEs, as diferenças entre os conjuntos avaliados é muito tênue, o que não sugere relevância biológica para este fenômeno.

Distâncias entre cópias de TEs

As famílias de TEs que apresentaram maior frequência de ocorrência nos íntrons - NonLTR-1_CR, NonLTR-6_CR e NonLTR-6A_CR, para *C. reinhardtii*, e L1-1_VC, RTE-2_VC e L1-6_VC, para *V. carteri* - foram examinadas em termos da distância par a par entre suas cópias. Para cada família, as cópias foram divididas em dois conjuntos: cópias intrônicas e não-intrônicas.

As figuras 5.19 e 5.20 exibem a distribuição de frequências de distâncias entre cópias, para cada conjunto analisado em *C. reinhardtii* e *V. carteri*, respectivamente. De forma geral, as famílias de TEs apresentaram nos dois conjuntos uma maior frequência de distâncias próximas de zero, diferentemente do observado, por exemplo, para cópias do elemento Perere-3/SR3 em *S. mansoni* e *S. japonicum* (19), em que máximos de frequência são observados para distâncias em torno de 0,33 e 0,36, respectivamente. A predominância de distâncias próximas de zero já foi apontada para as famílias de TEs nos genomas de *C. reinhardtii* e *V. carteri* por Prochnik et al. (27)

Além do máximo de frequência para distâncias próximas de zero, é notável um segundo pico de frequência para distâncias maiores do que 1,0, mais evidente no conjunto de cópias não-intrônicas, em especial para o conjunto de dados relativo à espécie *V. carteri* (Figura 5.20). Em outros termos, existe uma tendência bimodal para a distribuição de distâncias, mais notória no conjunto relativo a cópias não-intrônicas para as famílias analisadas.

Com o objetivo de quantificar esta tendência, para cada conjunto de dados foi estabelecida a razão entre a frequência de distâncias menores ou iguais a 1,0 e a frequência de distâncias acima de 1,0 (*razão_frequências*). Desta forma, valores mais elevados da variável *razão_frequências* traduzem maior ocorrência de distâncias menores entre as cópias analisadas e, portanto, maior proximidade entre as mesmas. Foi possível constatar que as famílias de TEs analisadas apresentam valores para a variável *razão_frequências* maiores para cópias intrônicas, quando comparado ao conjunto de cópias não-intrônicas (Figura 5.21). Estes dados mostram um menor grau de divergência entre as cópias intrônicas, o

Figura 5.19 – Distribuição de distâncias par a par para o conjunto de cópias de TEs intrônicas (painel à esquerda) e não-intrônicas (painel à direita), computadas para as três famílias mais frequentes em íntrons de genes ortólogos em *C. reinhardtii*. (A) NonLTR-1_CR: 368 cópias intrônicas e 57 não-intrônicas; (B) NonLTR-6_CR: 324 cópias intrônicas e 110 não-intrônicas; (C) NonLTR-6A_CR: 158 cópias intrônicas e 47 não-intrônicas. Fonte: Elaborada pela autora.

que indica uma predominância de cópias mais recentes nos íntrons.

O máximo de frequência para distâncias próximas de zero, associado à modesta representatividade dos TEs em *C. reinhardtii* e *V. carteri* (Tabela 3.1), sugere que os genomas destas espécies estão submetidos a um processo de eliminação gradual das cópias, provavelmente via recombinação, que conduz à remoção das cópias mais ancestrais. Este

Figura 5.20 – Distribuição de distâncias par a par para o conjunto de cópias de TEs intrônicas (painel à esquerda) e não-intrônicas (painel à direita), computadas para as três famílias mais frequentes em íntrons de genes ortólogos em *V. carteri*. (A) L1-1_VC: 218 cópias intrônicas e 500 não-intrônicas; (B) RTE2-2_VC: 171 cópias intrônicas e 683 não-intrônicas; (C) L1-6_VC: 138 cópias intrônicas e 322 não-intrônicas. Fonte: Elaborada pela autora.

cenário possibilitaria a manutenção do baixo número de cópias existentes nestes genomas, uma vez que o surgimento de novas cópias seria compensado pela eliminação das cópias mais antigas. Tal alternativa parece mais adequada do que a hipótese de que o grande número de cópias recentes seja consequência de um fenômeno explosivo de aumento de transposições, visto que a predominância de cópias recentes foi observada nas três famílias

Figura 5.21 – Razões entre a frequência de distâncias menores ou iguais a 1,0 e frequência de distâncias acima de 1,0. As distâncias par a par foram calculadas para cópias de TEs intrônicas e não-intrônicas, para as três famílias mais frequentes em íntrons de genes ortólogos em (A) *C. reinhardtii* e (B) *V. carteri*. Fonte: Elaborada pela autora.

de TEs analisadas em cada espécie e que os genomas sob estudo parecem restritivos ao número total de cópias de TEs.

A tendência de eliminação de cópias ancestrais mais severa nos íntrons, indicada pelo menor grau de divergência destas cópias em relação a cópias não intrônicas, sugere forte restrição à permanência de TEs nestas regiões. Uma possível explicação remete à hipótese apontada por Orgel e Crick, na década de 1980, de que a maioria do "DNA egoísta" estaria acumulado em regiões não-transcritas do genoma, devido ao custo energético associado à transcrição do DNA.⁽⁸⁾ Ao realizar esta predição, os mesmos apontaram ainda que, na maioria dos casos, o "DNA egoísta" seria excluído de regiões codificantes em virtude de consequências prejudiciais. Nesta perspectiva, é possível que o custo relativo à transcrição seja um dos fatores que contribui para o processo de eliminação das cópias de TEs nos íntrons em *C. reinhardtii* e *V. carteri*.

Capítulo 6

Conclusões

A extensa pesquisa relacionada a elementos transponíveis expôs evidências acerca da influência destes na evolução dos genes em diversas espécies, o que motivou o estudo em torno da associação entre TEs e genes ortólogos nas espécies *C. reinhardtii* e *V. carteri*.

O primeiro aspecto avaliado remete ao percentual do genoma de cada alga derivado de TEs: os índices verificados apontam uma modesta ocupação dos elementos ($\sim 6\%$), comparável ao reconhecido em espécies de leveduras. A identificação de associações entre TEs e genes ortólogos indicou que os genes em *C. reinhardtii* tendem a acumular mais cópias do que os genes em *V. carteri*. Quando examinada a densidade de cópias de TEs em cada contexto gênico, a primeira espécie apresentou índices mais elevados para as regiões flangeadora 5', intrônica e flangeadora 3'; em regiões codificantes, no entanto, *V. carteri* exibe maior densidade de TEs, o que sugere que os elementos de transposição possam ter maior contribuição na evolução de sequências codificadoras nesta espécie do que em *C. reinhardtii*.

Analisando-se a distribuição de TEs em regiões flangeadoras 5' dos genes ortólogos, foi possível constatar um intervalo comum às duas espécies - definido de 1 a 300 pb em relação ao sítio de início de tradução - em que a frequência dos elementos relativos às classes de TEs mais abundantes em cada espécie encontra-se significativamente abaixo do esperado, segundo a distribuição randômica de TEs no genoma. Este resultado sugere que ocorrências dos elementos nestas imediações foram selecionadas negativamente, em virtude da possível disrupção de estruturas regulatórias associadas à transcrição dos genes. Os resultados indicaram uma tendência de aumento da frequência padronizada dos TEs, à medida que o intervalo analisado na região flangeadora 5' se distancia do sítio de início de tradução.

Padrão semelhante foi verificado no estudo da distribuição de TEs na região flangeadora 3' dos genes ortólogos, segundo o qual a frequência padronizada dos elementos - relativos às classes de TEs mais abundantes em cada espécie - aumenta à medida que a região analisada se distancia do término do gene. Os resultados indicaram um intervalo de baixa

representatividade de TEs - comum às espécies *C. reinhardtii* e *V. carteri* - no segmento adjacente ao término do gene (1-100 pb), embora sem significância estatística para a classe non-LTR na primeira espécie. Apesar da semelhança entre o comportamento das distribuições, o intervalo de baixa representatividade de TEs mostrou-se mais estreito na região flaqueadora 3' do que na região flaqueadora 5', o que indica que a região crítica a eventos de inserção é mais extensa na adjacência ao início do gene do que ao término do mesmo, provavelmente em função da natureza das unidades regulatórias existentes em cada contexto. Em relação à espécie *C. reinhardtii*, a identificação de regiões de sítios poli(A) com contribuições de sequências derivadas de TEs, a exemplo dos genes Cre01.g054500, Cre12.g498850, Cre01.g016300 e Cre10.g432050, indicam a relevância destes elementos na evolução do término 3' dos genes nesta alga, como verificado em outras espécies na literatura.

A presença de TEs em regiões codificantes nas duas espécies mostrou-se altamente reprimida, o que era esperado mediante o elevado potencial deletério de eventos de integração nestas regiões, o que conduz à eliminação das cópias em decorrência da seleção negativa. No entanto, apesar do provável efeito deletério, é possível que a inovação introduzida por um TE no aspecto codificante do gene possa desencadear benefícios evolutivos ao organismo, o que conduz à fixação da sequência derivada do elemento no gene. Neste trabalho foram identificados (e validados por EST) instâncias de genes que abrigam TEs em regiões codificantes: quatro genes em *C. reinhardtii* e oito em *V. carteri*. Dentre estes, um caso em cada espécie mostrou-se particularmente interessante por possuir um domínio proteico em região derivada do TE. O primeiro caso remete ao gene Cre06.g262800, em *C. reinhardtii*, no qual foi identificado o domínio PHD-finger na região em que reside a cópia do elemento Gypsy-5-LTR_CR. A presença do domínio Condensin2nSMC na sequência codificante do gene, associado à manutenção da estrutura de cromossomos, sugere que o gene Cre06.g262800 representa um exemplo de contribuição de elemento de transposição na evolução de proteínas cromossômicas. O segundo caso de contribuição de TE em sequência codificante descrito neste trabalho refere-se ao gene Vocar20001092m.g, em *V. carteri*. A sequência codificante deste gene possui três domínios, zinc knuckle (zf-CCHC), WD40 e zf-MYND, estando o primeiro localizado em região derivada do elemento Gypsy3-LTR_VC. Os três domínios identificados possuem em comum o caráter de interação proteína-DNA, o que sugere que a incorporação do domínio zinc knuckle possa ter contribuído no ajuste da afinidade de interação e, portanto, na evolução da sequência codificadora do gene.

Análises relativas à distribuição de TEs nos íntrons mostraram que estas regiões também são fortemente preservadas, quanto à ocupação de TEs, nas espécies *C. reinhardtii* e *V. carteri*. De forma geral, todas as classes de TEs exibem frequência abaixo do esperado neste contexto genômico; quando avaliadas as principais classes em cada espécie, a representatividade abaixo do esperado é mais evidente em intervalos adjacentes ao

éxon. Outro ponto remete à preservação dos primeiros íntrons - principais contenedores de sequências regulatórias localizadas em regiões intrônicas - quanto à ocorrência de TEs. Consideradas conjuntamente, as tendências identificadas sugerem que elementos de transposição inseridos nestas regiões foram selecionados negativamente, provavelmente devido ao potencial de ruptura do modelo de *splicing* ou de regulação dos genes. Outro aspecto que pode estar associado à seleção negativa de TEs em regiões intrônicas, mesmo quando a integração não ocorre em intervalos críticos, é a restrição à elongação dos íntrons, deletéria ao processo de *splicing* no modelo definido no íntron. Finalmente, o menor grau de divergência entre cópias intrônicas, comparado a cópias não-intrônicas, sugere uma tendência de eliminação de cópias ancestrais mais severa nestas regiões, possivelmente devido ao custo de transcrição, constituindo portanto mais um fator relacionado à baixa representatividade de TEs nos íntrons das espécies analisadas.

As constatações decorrentes do estudo da influência de TEs nos genes ortólogos de *C. reinhardtii* e *V. carteri* corroboram a hipótese de que estes elementos podem contribuir na evolução da arquitetura dos genes, a exemplo de sequências provenientes de TEs incorporadas em regiões codificantes e sinalizadoras de poliadenilação, mas são majoritariamente carreadores de efeitos deletérios, o que é evidenciado pelos padrões que expressam baixa representatividade destes elementos em regiões críticas para a expressão adequada do gene. Como apontado por Makałowski (10), é possível que os elementos transponíveis constituam, de fato, uma fonte de recursos aptos a gerar diversidade genética, utilizada pela seleção natural em seus experimentos evolutivos.

Referências

- 1 MCCLINTOCK, B. The significance of responses of the genome to challenge. *Science*, v. 226, n. 4676, p. 792–801, 1984.
- 2 JURKA, J.; KAPITONOV, V. V.; KOHANY, O.; JURKA, M. V. Repetitive sequences in complex genomes: structure and evolution. *Annual Review of Genomics and Human Genetics*, v. 8, n. 1, p. 241–259, 2007.
- 3 BIÉMONT, C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics*, v. 186, n. 4, p. 1085–1093, 2010.
- 4 PRAY, L.; ZHAUROVA, K. Barbara McClintock and the discovery of jumping genes (transposons). *Nature Education*, v. 1, n. 1, p. 169, 2008.
- 5 REBOLLO, R.; ROMANISH, M. T.; MAGER, D. L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics*, v. 46, n. 1, p. 21–42, 2012.
- 6 JANICKI, M.; ROOKE, R.; YANG, G. Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Research*, v. 19, n. 6, p. 787–808, 2011.
- 7 DOOLITTLE, W. F.; SAPIENZA, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, v. 284, n. 5757, p. 601–603, 1980.
- 8 ORGEL, L. E.; CRICK, F. H. C. Selfish DNA: the ultimate parasite. *Nature*, v. 284, n. 5757, p. 604–607, 1980.
- 9 LEVIN, H. L.; MORAN, J. V. Dynamic interactions between transposable elements and their hosts. *Nature Reviews Genetics*, v. 12, n. 9, p. 615–617, 2011.

-
- 10 MAKALOWSKI, W. Genomic scrap yard: how genomes utilize all that junk. *Gene*, v. 259, n. 1-2, p. 61–67, 2000.
- 11 NEKRUTENKO, A.; LI, W. H. Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics*, v. 17, n. 11, p. 619–621, 2001.
- 12 LAGEMAAT, L. N. van de; LANDRY, J. R.; MAGER, D. L.; MEDSTRAND, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics*, v. 19, n. 10, p. 530–536, 2003.
- 13 KAZAZIAN JUNIOR, H. H. Mobile elements: drivers of genome evolution. *Science*, v. 303, n. 5664, p. 1626–1632, 2004.
- 14 VOLFF, J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays*, v. 28, n. 9, p. 913–922, 2006.
- 15 FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, v. 9, n. 5, p. 397–405, 2008.
- 16 LEVY, A.; SELA, N.; AST, G. TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. *Nucleic Acids Research*, v. 36, p. D47–D52, 2008. doi: 10.1093/nar/gkm949.
- 17 LOPES, F.; CARAZZOLLE, M.; PEREIRA, G.; COLOMBO, C.; CARARETO, C. Transposable elements in Coffea (Gentianales: Rubiaceae) transcripts and their role in the origin of protein diversity in flowering plants. *Molecular Genetics and Genomics*, v. 279, n. 4, p. 385–401, 2008.
- 18 KROM, N.; RECLA, J.; RAMAKRISHNA, W. Analysis of genes associated with retrotransposons in the rice genome. *Genetica*, v. 134, n. 3, p. 297–310, 2008.
- 19 VENANCIO, T. M.; WILSON, R. A.; VERJOVSKI-ALMEIDA, S.; DEMARCO, R. Bursts of transposition from non-long terminal repeat retrotransposon families of the RTE clade in *Schistosoma mansoni*. *International Journal for Parasitology*, v. 40, n. 6, p. 743–749, 2010.
- 20 FEDOROFF, N. V. Transposable elements, epigenetics, and genome evolution. *Science*, v. 338, n. 6108, p. 758–767, 2012.

- 21 MCDONALD, J. F. Transposable elements: possible catalysts of organismic evolution. *Trends in Ecology & Evolution*, v. 10, n. 3, p. 123–126, 1995.
- 22 JORDAN, I. K.; ROGOZIN, I. B.; GLAZKO, G. V.; KOONIN, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*, v. 19, n. 2, p. 68–72, 2003.
- 23 DAI, J.; XIE, W.; BRADY, T. L.; GAO, J.; VOYTAS, D. F. Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin. *Molecular Cell*, v. 27, n. 2, p. 289–299, 2007.
- 24 BEAUREGARD, A.; CURCIO, M. J.; BELFORT, M. The take and give between retrotransposable elements and their hosts. *Annual Review of Genetics*, v. 42, n. 1, p. 587–617, 2008.
- 25 PARK, M.; JO, S.; KWON, J.-K.; PARK, J.; AHN, J. H.; KIM, S.; LEE, Y.-H.; YANG, T.-J.; HUR, C.-G.; KANG, B.-C.; KIM, B.-D.; CHOI, D. Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. *BMC Genomics*, v. 12, n. 85, p. 1–13, 2011.
- 26 MERCHANT, S. S.; PROCHNIK, S. E.; VALLON, O.; HARRIS, E. H.; KARPOWICZ, S. J.; WITMAN, G. B.; TERRY, A.; SALAMOV, A.; FRITZ-LAYLIN, L. K.; MARÉCHAL-DROUARD, L.; MARSHALL, W. F.; QU, L.-H.; NELSON, D. R.; SANDERFOOT, A. A.; SPALDING, M. H. et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, v. 318, n. 5848, p. 245–251, 2007.
- 27 PROCHNIK, S. E.; UMEN, J.; NEDELCO, A. M.; HALLMANN, A.; MILLER, S. M.; NISHII, I.; FERRIS, P.; KUO, A.; MITROS, T.; FRITZ-LAYLIN, L. K.; HELLSTEN, U.; CHAPMAN, J.; SIMAKOV, O.; RENSING, S. A.; TERRY, A. et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science*, v. 329, n. 5988, p. 223–226, 2010.
- 28 HALLMANN, A. Evolution of reproductive development in the volvocine algae. *Sexual Plant Reproduction*, v. 24, n. 2, p. 97–112, 2011.
- 29 NISHII, I.; MILLER, S. M. Volvox: simple steps to developmental complexity? *Current Opinion in Plant Biology*, v. 13, n. 6, p. 646–653, 2010.
- 30 KIRK, D. L. Seeking the ultimate and proximate causes of *Volvox* multicellularity and cellular differentiation. *Integrative and Comparative Biology*, v. 43, n. 2, p. 247–253, 2003.

-
- 31 JURKA, J.; KAPITONOV, V.; PAVLICEK, A.; KLONOWSKI, P.; KOHANY, O.; WALICHIEWICZ, J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, v. 110, n. 1-4, p. 462–467, 2005.
- 32 KIM, F. J.; BATTINI, J.-L.; MANEL, N.; SITBON, M. Emergence of vertebrate retroviruses and envelope capture. *Virology*, v. 318, n. 1, p. 183–191, 2004.
- 33 RIBET, D.; HARPER, F.; DUPRESSOIR, A.; DEWANNIEUX, M.; PIERRON, G.; HEIDMANN, T. An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Research*, v. 18, p. 597–609, 2008. doi: 10.1101/gr.073486.107.
- 34 BERGMAN, C. M.; QUESNEVILLE, H. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, v. 8, n. 6, p. 382–392, 2007.
- 35 CRAIG, N. L. Unity in transposition reactions. *Science*, v. 270, n. 13, p. 253–254, 1995.
- 36 ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. *Molecular biology of the cell*. 5 th. ed. New York: Garland Science, 2008.
- 37 BROSIUS, J. Retrotransposons-seeds of evolution. *Science*, v. 251, n. 4995, p. 753, 1991.
- 38 MORAN, J. V.; DEBERARDINIS, R. J.; KAZAZIAN JUNIOR, H. H. Exon shuffling by L1 retrotransposition. *Science*, v. 283, n. 5407, p. 1530–1534, 1999.
- 39 LONG, M. Evolution of novel genes. *Current Opinion in Genetics & Development*, v. 11, n. 6, p. 673–680, 2001.
- 40 ZHANG, Y.; ROMANISH, M. T.; MAGER, D. L. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Computational Biology*, v. 7, n. 5, p. 1–13, 2011.
- 41 ALZOHAIERY, A. M.; GYULAI, G.; JANSEN, R. K.; BAHIELDIN, A. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. *Plasmid*, v. 69, n. 1, p. 1–15, 2013.

- 42 JOLY-LOPEZ, Z.; EWA, F.; R., H. D.; NIKOLETA, J.; E., B. T. A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in *Arabidopsis thaliana*. *PLoS Genetics*, v. 8, n. 9, p. 1–10, 2012.
- 43 KAPITONOV, V. V.; JERZY, J. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biology*, v. 3, n. 6, p. e181, 2005.
- 44 LIN, R.; DING, L.; CASOLA, C.; RIPOLL, D. R.; FESCHOTTE, C.; WANG, H. Transposase-derived transcription factors regulate light signaling in Arabidopsis. *Science*, v. 318, n. 5854, p. 1302–1305, 2007.
- 45 HANCOCK, J. M. Gene factories, microfunctionalization and the evolution of gene families. *Trends in Genetics*, v. 21, n. 11, p. 591–595, 2005.
- 46 PROBST, W. C.; SNYDER, L. A.; SCHUSTER, D. I.; BROSIUS, J.; SEALFON, S. C. Sequence alignment of the G-Protein coupled receptor superfamily. *DNA and Cell Biology*, v. 11, n. 1, p. 1–20, 1992.
- 47 KAZAZIAN, H. H.; WONG, C.; YOUSSEF, H.; SCOTT, A. F.; PHILLIPS, D. G.; ANTONARAKIS, S. E. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, v. 332, n. 6160, p. 164–166, 1988.
- 48 GARVEY, S. M.; RAJAN, C.; LERNER, A. P.; FRANKEL, W. N.; COX, G. A. The muscular dystrophy with myositis (mdm) mouse mutation disrupts a skeletal muscle-specific domain of titin. *Genomics*, v. 79, n. 2, p. 146–149, 2002.
- 49 CHÉNAIS, B. Transposable elements and human cancer: a causal relationship? *Biochimica et Biophysica Acta*, v. 1835, n. 1, p. 28–35, 2013.
- 50 MIKI, Y.; NISHISHO, I.; HORII, A.; MIYOSHI, Y.; UTSUNOMIYA, J.; KINZLER, K. W.; VOGELSTEIN, B.; NAKAMURA, Y. Disruption of the APC gene by a retrotransposon insertion of L1 sequence in a colon cancer. *Cancer Research*, v. 52, n. 3, p. 643–645, 1992.
- 51 SHRAGER, J.; HAUSER, C.; CHANG, C.-W.; HARRIS, E. H.; DAVIES, J.; MCDERMOTT, J.; TAMSE, R.; ZHANG, Z.; GROSSMAN, A. R. *Chlamydomonas reinhardtii* genome project. a guide to the generation and use of the cDNA information. *Plant Physiology*, v. 131, n. 2, p. 401–408, 2003.

- 52 RAUSCH, H.; LARSEN, N.; SCHMITT, R. Phylogenetic relationships of the green alga *Volvox carteri* deduced from small-subunit ribosomal RNA comparisons. *Journal of Molecular Evolution*, v. 29, n. 3, p. 255–265, 1989.
- 53 HERRON, M. D.; HACKETT, J. D.; AYLRARD, F. O.; MICHOD, R. E. Triassic origin and early radiation of multicellular volvocine algae. *Proceedings of the National Academy of Sciences*, v. 106, n. 9, p. 3254–3258, 2009.
- 54 MILLER, S. M. Volvox, Chlamydomonas, and the evolution of multicellularity. *Nature Education*, v. 3, n. 9, p. 65, 2010.
- 55 KIRK, D. L. Evolution of multicellularity in the *volvocine algae*. *Current Opinion in Plant Biology*, v. 2, n. 6, p. 496–501, 1999.
- 56 ZHANG, M. Q. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, v. 3, n. 9, p. 698–709, 2002.
- 57 TEUFEL, A.; KRUPP, M.; WEINMANN, A.; GALLE, P. R. Current bioinformatics tools in genomic biomedical research. *International Journal of Molecular Medicine*, v. 17, n. 6, p. 967–973, 2006.
- 58 BAXEVANIS, A. D.; OUELLETTE, B. F. F. *Bioinformatics: a practical guide to the analysis of genes and proteins*. 3 th. ed. Hoboken: John Wiley & Sons, 2005.
- 59 KORF, I.; YANDELL, M.; BEDELL, J. *BLAST*. Sebastopol: O'Reilly & Associates, 2003.
- 60 MADDEN, T. The BLAST sequence analysis tool. In: MCENTYRE, J.; OSTELL, J. (Ed.). *The NCBI handbook*. [S.l.: s.n.], 2002. Disponível em: <<http://www.ncbi.nlm.nih.gov/books/NBK21097/>>. Acesso em: 28 jan. 2014.
- 61 KINSER, J. M. *Python for bioinformatics*. Massachusetts: Jones and Bartlett Publishers, 2009.
- 62 PYTHON. Disponível em: <<http://www.python.org/>>. Acesso em: 20 fev. 2014.
- 63 BIOPYTHON. Disponível em: <<http://biopython.org/DIST/docs/tutorial/Tutorial.html>>. Acesso em: 20 fev. 2014.

- 64 MATPLOTLIB. Disponível em: <<http://matplotlib.org/>>. Acesso em: 20 fev. 2014.
- 65 YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, v. 13, n. 5, p. 329–342, 2012.
- 66 KOHANY, O.; GENTLES, A. J.; HANKUS, L.; JURKA, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, v. 7, n. 474, p. 1–7, 2006.
- 67 GOODSTEIN, D. M.; SHU, S.; HOWSON, R.; NEUPANE, R.; HAYES, R. D.; FAZO, J.; MITROS, T.; DIRKS, W.; HELLSTEN, U.; PUTNAM, N.; ROKHSAR, D. S. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, v. 40, n. D1, p. D1178–D1186, 2012.
- 68 SELA, N.; KIM, E.; AST, G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biology*, v. 11, n. 6, p. R59, 2010.
- 69 SLEATOR, R. D. An overview of the current status of eukaryote gene prediction strategies. *Gene*, v. 461, n. 1–2, p. 1–4, 2010.
- 70 KRISTENSEN, D. M.; WOLF, Y. I.; MUSHEGIAN, A. R.; KOONIN, E. V. Computational methods for Gene Orthology inference. *Briefings in Bioinformatics*, v. 12, n. 5, p. 379–391, 2011.
- 71 KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, v. 39, n. 1, p. 309–338, 2005.
- 72 LUTZ, C. S. Alternative polyadenylation: a twist on mRNA 3' end formation. *ACS Chemical Biology*, v. 3, n. 10, p. 609–617, 2008.
- 73 RIETHOVEN, J. M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. In: LADUNGA, I. (Ed.). *Computational biology of transcription factor binding*. [S.l.]: Humana Press, 2010. v. 674, p. 33–42.
- 74 CONTI, L. D.; BARALLE, M.; BURATTI, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA*, v. 4, n. 1, p. 49–60, 2013.

- 75 ELKON, R.; UGALDE, A. P.; AGAMI, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, v. 14, n. 7, p. 496–506, 2013.
- 76 ANDREASSI, C.; RICCIO, A. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends in Cell Biology*, v. 19, n. 9, p. 465–474, 2009.
- 77 MOORE, M. J. From birth to death: the complex lives of eukaryotic mRNAs. *Science*, v. 309, n. 5740, p. 1514–1518, 2005.
- 78 LIN, C. J.; SILVA, B. A. Técnicas de análise da regulação da transcrição gênica e suas aplicações na endocrinologia molecular. *Arquivos Brasileiros de Endocrinologia & Metabologia*, v. 46, n. 4, p. 330–340, 2002.
- 79 ZENG, J.; ZHU, S.; YAN, H. Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. *Briefings in Bioinformatics*, v. 10, n. 5, p. 498–508, 2009.
- 80 ZHANG, M. Q. Computational analyses of eukaryotic promoters. *BMC Bioinformatics*, v. 8, p. 1–5, 2007. doi: 10.1186/1471-2105-8-S6-S3.
- 81 WASSERMAN, W. W.; SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, v. 5, n. 4, p. 276–287, 2004.
- 82 FAULKNER, G. J.; KIMURA, Y.; DAUB, C. O.; WANI, S.; PLESSY, C.; IRVINE, K. M.; SCHRODER, K.; CLOONAN, N.; STEPTOE, A. L.; LASSMANN, T.; WAKI, K.; HORNIG, N.; ARAKAWA, T.; TAKAHASHI, H.; KAWAI, J. et al. The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, v. 41, n. 5, p. 563–571, 2009.
- 83 THORNBURG, B. G.; GOTEA, V.; MAKALOWSKI, W. Transposable elements as a significant source of transcription regulating signals. *Gene*, v. 365, n. 0, p. 104–110, 2006.
- 84 SAKAI, H.; TANAKA, T.; ITOH, T. Birth and death of genes promoted by transposable elements in *Oryza sativa*. *Gene*, v. 392, n. 1-2, p. 59–63, 2007.
- 85 MAGALHÃES, M. N.; LIMA, A. C. P. d. *Noções de probabilidade e estatística*. 6. ed. São Paulo: Editora da Universidade de São Paulo, 2007.

- 86 DHADI, S. R.; KROM, N.; RAMAKRISHNA, W. Genome-wide comparative analysis of putative bidirectional promoters from rice, *Arabidopsis* and *Populus*. *Gene*, v. 429, n. 1–2, p. 65–73, 2009.
- 87 HIGO, K.; UGAWA, Y.; IWAMOTO, M.; KORENAGA, T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research*, v. 27, n. 1, p. 297–300, 1999.
- 88 VUOLO, J. H. *Fundamentos da teoria de erros*. 2. ed. São Paulo: Edgard Blücher LTDA, 1992.
- 89 LYNCH, M.; SCOFIELD, D. G.; HONG, X. The evolution of transcription-initiation sites. *Molecular Biology and Evolution*, v. 226, n. 4, p. 1137–1146, 2005.
- 90 HEIM, M. A.; JAKOBY, M.; WERBER, M.; MARTIN, C.; WEISSHAAR, B.; BAILEY, P. C. The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Molecular Biology and Evolution*, v. 20, n. 5, p. 735–747, 2003.
- 91 SIMPSON, S. D.; NAKASHIMA, K.; NARUSAKA, Y.; SEKI, M.; SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. Two different novel cis-acting elements of *erd1*, a *clpA* homologous *Arabidopsis* gene function in induction by dehydration stress and dark-induced senescence. *The Plant Journal*, v. 33, n. 2, p. 259–270, 2003.
- 92 LEE, J. Y.; JI, Z.; TIAN, B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Research*, v. 36, n. 17, p. 5581–5590, 2008.
- 93 TIAN, B.; HU, J.; ZHANG, H.; LUTZ, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, v. 33, n. 1, p. 201–212, 2005.
- 94 SHEN, Y.; LIU, Y.; LIU, L.; LIANG, C.; LI, Q. Q. Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii*. *Genetics*, v. 179, n. 1, p. 167–176, 2008.
- 95 SOUZA, F. S. de; FRANCHINI, L. F.; RUBINSTEIN, M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Molecular Biology and Evolution*, v. 30, n. 6, p. 1239–1251, 2013.

- 96 PUNTA, M.; COGGILL, P. C.; EBERHARDT, R. Y.; MISTRY, J.; TATE, J.; BOURSNEILL, C.; PANG, N.; FORSLUND, K.; CERIC, G.; CLEMENTS, J.; HEGER, A.; HOLM, L.; SONNHAMMER, E. L. L.; EDDY, S. R.; BATEMAN, A.; FINN, R. D. The Pfam protein families database. *Nucleic Acids Research*, v. 40, n. D1, p. D290–D301, 2012.
- 97 EDDY, S. R.; WHEELER, T. J. *HMMER user's guide*. [S.l.], 2013. Disponível em: <<http://hmmerr.org/>>. Acesso em: 28 jan. 2014.
- 98 LARKIN, M.; BLACKSHIELDS, G.; BROWN, N.; CHENNA, R.; MCGETTIGAN, P.; MCWILLIAM, H.; VALENTIN, F.; WALLACE, I.; WILM, A.; LOPEZ, R.; THOMPSON, J.; GIBSON, T.; HIGGINS, D. Clustal W and Clustal X version 2.0. *Bioinformatics*, v. 23, n. 21, p. 2947–2948, 2007.
- 99 HOLDER, M.; LEWIS, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, v. 4, n. 4, p. 275–284, 2003.
- 100 HALL, B. G. Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution*, v. 30, n. 5, p. 1229–1235, 2013.
- 101 TAMURA, K.; PETERSON, D.; PETERSON, N.; STECHER, G.; NEI, M.; KUMAR, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, v. 28, n. 10, p. 2731–2739, 2011.
- 102 HUDSON, D. F.; MARSHALL, K. M.; EARNSHAW, W. C. Condensin: architect of mitotic chromosomes. *Chromosome Research*, v. 17, n. 2, p. 131–144, 2009.
- 103 BECKER, P. B. Gene regulation: a finger on the mark. *Nature*, v. 442, n. 7098, p. 31–32, 2006.
- 104 SHI, X.; KACHIRSKAIA, I.; WALTER, K. L.; KUO, J.-H. A.; LAKE, A.; DAVRAZOU, F.; CHAN, S. M.; MARTIN, D. G. E.; FINGERMAN, I. M.; BRIGGS, S. D.; HOWE, L.; UTZ, P. J.; KUTATELADZE, T. G.; LUGOVSKOY, A. A. et al. Proteome-wide analysis in *Saccharomyces cerevisiae* identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *Journal of Biological Chemistry*, v. 282, n. 4, p. 2450–2455, 2007.
- 105 MATTHEWS, J. M.; BHATI, M.; LEHTOMAKI, E.; MANSFIELD, R. E.; CUBEDDU, L.; MACKAY, J. P. It takes two to tango: the structure and function of

LIM, RING, PHD and MYND domains. *Current Pharmaceutical Design*, v. 15, n. 31, p. 3681–3696, 2009.

106 KRISHNA, S. S.; MAJUMDAR, I.; GRISHIN, N. V. Structural classification of zinc fingers. *Nucleic Acids Research*, v. 31, n. 2, p. 532–550, 2003.

107 XU, C.; MIN, J. Structure and function of WD40 domain proteins. *Protein Cell*, v. 2, n. 3, p. 202–214, 2011.

108 ARAVIND, L. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends in Biochemical Sciences*, v. 25, n. 9, p. 421–423, 2000.

109 HADDRILL, P. R.; CHARLESWORTH, B.; HALLIGAN, D. L.; ANDOLFATTO, P. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*, v. 6, n. 8, p. R67.1–R67.8, 2005.

110 LORKOVIĆ, Z. J.; KIRK, D. A. W.; LAMBERMON, M. H.; FILIPOWICZ, W. Pre-mRNA splicing in higher plants. *Trends in Plant Science*, v. 5, n. 4, p. 160–167, 2000.

111 STARK, K.; KIRK, D. L.; SCHMITT, R. Two enhancers and one silencer located in the introns of *regA* control somatic cell differentiation in *Volvox carteri*. *Genes & Development*, v. 15, n. 11, p. 1449–1460, 2001.

112 CHEN, Y.; STEPHAN, W. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster* *Adh* gene. *Proceedings of the National Academy of Sciences*, v. 100, n. 20, p. 11499–11504, 2003.

113 TAJIMA, F.; NEI, M. Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, v. 1, n. 3, p. 269–285, 1984.

114 BERGET, S. M. Exon recognition in vertebrate splicing. *Journal of Biological Chemistry*, v. 270, p. 2411–2414, 1995. doi: 10.1074/jbc.270.6.2411.

Apêndice A Elementos de transposição

Tabela A.1 – Dados relativos aos TEs descritos para a espécie *C. reinhardtii*, cujas sequências encontram-se depositadas no Repbase (31). As três primeiras colunas apresentam informações referentes às sequências, enquanto que a quarta indica a frequência de ocorrência absoluta dos elementos no genoma de *C. reinhardtii*.

Identificador	Descrição	Comprimento	Frequência
Copia1-LTR_CR	Copia	6502	56
Copia10-LTR_CR	Copia	3339	27
Copia2-LTR_CR	Copia	6119	5
Copia3-LTR_CR	Copia	5994	73
CRTOC1	CRTOC1	5662	77
DIRS-1_CR	DIRS	6454	16
DNA-1-7_CR	DNA transposon	710	448
DNA-1-9_CR	DNA transposon	630	87
DNA-4-7_CR	DNA transposon	709	325
DNA-8-2_CR	P	2426	18
DNA-8-3_CR	P	2870	20
DNA3-1_CR	DNA transposon	1989	82
DNAX1_CR	DNA transposon	1281	33
DualenCr4	RandI	10346	36
EnSpm-N1a_CR	EnSpm	5347	28
EnSpm-N2_CR	EnSpm	6161	112
Gulliver	hAT	7144	75
Gypsy-1-LTR_CR	Gypsy	7669	75
Gypsy-10-LTR_CR	Gypsy	8069	32
Gypsy-11-LTR_CR	Gypsy	15836	81
Gypsy-12-LTR_CR	Gypsy	7071	26
Gypsy-2-LTR_CR	Gypsy	5482	8

Continua

Continuação

Identificador	Descrição	Comprimento	Frequência
Gypsy-3-LTR_CR	Gypsy	881	43
Gypsy-4-LTR_CR	Gypsy	5427	42
Gypsy-5-LTR_CR	Gypsy	5664	92
Gypsy-6-I_CR	Gypsy	6090	20
Gypsy-8-LTR_CR	Gypsy	15640	38
Gypsy-9-LTR_CR	Gypsy	18794	33
Gypsy14-LTR_CR	Gypsy	3891	7
Gypsy15-LTR_CR	Gypsy	6683	16
Gypsy16-LTR_CR	Gypsy	4290	106
Gypsy17-LTR_CR	Gypsy	5100	5
Gypsy18-LTR_CR	Gypsy	6761	18
Gypsy19-LTR_CR	Gypsy	7069	11
hAT-N10_CR	hAT	574	9
hAT-N11a_CR	hAT	249	134
hAT-N11_CR	hAT	268	104
hAT-N1_CR	hAT	893	32
hAT-N2_CR	hAT	238	290
hAT-N3_CR	hAT	468	108
hAT-N4_CR	hAT	645	94
hAT-N5_CR	hAT	783	95
hAT-N7_CR	hAT	1125	305
hAT-N8_CR	hAT	646	35
hAT-N9_CR	hAT	694	31
Helitron-1N1_CRe	Helitron	660	372
Helitron-1N2_CRe	Helitron	679	371
L1-1_CR	L1	7343	998
L1-2_CR	L1	5292	539
L1-3_CR	L1	1991	1196
L1-4_CR	L1	1609	451
L1-5_CR	L1	852	81
L1-6_CR	L1	5286	242
L1-7_CR	L1	2678	269
LINE-8_CR	Non-LTR	1881	100
LTR1_CR	LTR	491	99
LTR2_CR	LTR	593	22
LTR3_CR	LTR	357	35

Continua

Continuação

Identificador	Descrição	Comprimento	Frequência
LTR4_CR	LTR	352	30
Mariner-N2_CR	Mariner/Tc1	519	14
Mariner-N3_CR	Mariner/Tc1	654	12
Mariner-N4_CR	Mariner/Tc1	568	24
Mariner-N5_CR	Mariner/Tc1	742	16
Mariner-N6_CR	Mariner/Tc1	414	58
MarinerL-1_CR	Mariner/Tc1	15988	135
NonLTR-1_CR	Non-LTR	761	549
NonLTR-2_CR	Non-LTR	1207	76
NonLTR-3_CR	Non-LTR	1070	77
NonLTR-4_CR	Non-LTR	588	262
NonLTR-5_CR	Non-LTR	5274	315
NonLTR-6A_CR	Non-LTR	980	298
NonLTR-6_CR	Non-LTR	1226	667
Novosib-1_CR	Novosib	4776	10
Novosib-2_CR	Novosib	3280	9
Novosib-3_CR	Novosib	6736	7
Novosib-4_CR	Novosib	6379	10
Novosib-5_CR	Novosib	5285	17
Novosib-6_CR	Novosib	3392	23
Novosib-N1	Novosib	787	86
Novosib-N1a	Novosib	781	57
P-1_CR	P	9810	33
P-2_CR	P	9309	65
Pioneer1	Transposable Element	2816	5
RandI-1	RandI	10812	136
RandI-2	RandI	9543	158
RandI-3	RandI	11237	55
RandI-4	RandI	7807	120
RandI-5	RandI	7466	54
RandI-6	RandI	13822	48
RTEX-1_CR	RTEX	5811	229
SINE-6_CR	Non-LTR	405	21
SINEX-2_CR	SINE	484	75
SINEX-3_CR	Non-LTR	507	87
SINEX-4_CR	SINE	518	36

Continua

Continuação

Identificador	Descrição	Comprimento	Frequência
SINEX-5_CR	Non-LTR	404	9
TCR1_LTR	LTR	164	9
TE2-1_CR	Transposable Element	471	24
TE2-2_CR	Transposable Element	764	39
TE2-3_CR	Transposable Element	542	23
TE2-6_CR	Transposable Element	784	115
TE2-7a_CR	Transposable Element	743	66
TE2-7b_CR	Transposable Element	772	60

Fonte: Elaborada pela autora.

Tabela A.2 – Dados relativos aos TEs descritos para a espécie *V. carteri*, cujas sequências encontram-se depositadas no Repbase (31). As três primeiras colunas apresentam informações referentes às sequências, enquanto que a quarta indica a frequência de ocorrência absoluta dos elementos no genoma de *V. carteri*.

Identificador	Descrição	Comprimento	Frequência
Copia2-LTR_VC	Copia	5809	283
Copia4-LTR_VC	Copia	6787	278
Copia6-LTR_VC	Copia	4597	85
Gypsy10-LTR_VC	Gypsy	6623	102
Gypsy11-LTR_VC	Gypsy	6462	167
Gypsy12-LTR_VC	Gypsy	7244	233
Gypsy13-LTR_VC	Gypsy	7163	216
Gypsy14-LTR_VC	Gypsy	6194	73
Gypsy16-LTR_VC	Gypsy	7409	44
Gypsy2-LTR_VC	Gypsy	8296	89
Gypsy3-LTR_VC	Gypsy	6386	220
Gypsy5-LTR_VC	Gypsy	11414	1148
Gypsy6-LTR_VC	Gypsy	6484	291
Gypsy7-LTR_VC	Gypsy	7168	584
Gypsy9-LTR_VC	Gypsy	7424	17
JORDAN	Transposable Element	1595	243
KANGAROO-1_VC	DIRS	8903	448
L1-1_VC	L1	7976	781
L1-2_VC	L1	8021	690
L1-3_VC	L1	8387	438
L1-4_VC	L1	6116	363
L1-5_VC	L1	6922	259
L1-6_VC	L1	2741	510
Lueckenbuesser_LTR	Copia	4603	72
Lusen_LTR	Copia	4346	791
OSSER	Copia	4885	68
RandI-1_VC	RandI	11537	89
RTEX-1_VC	RTEX	5067	186
RTEX-2_VC	RTEX	5366	910
RTEX-3_VC	RTEX	4991	108
RTEX-4_VC	RTEX	5587	819
RTEX-5_VC	RTEX	5325	1081

Fonte: Elaborada pela autora.

Apêndice B Resultados de análises estatísticas

Tabela B.1 – Valores da estatística z para as classes de distância da região flaqueadora 5' dos genes ortólogos na espécie *C. reinhardtii*. São apresentados somente os dados relativos às classes de TEs de maior frequência de ocorrência.

Distância relativa ao SIT (pb)	non-LTR			DNA		
	Freq. observada	Freq. esperada	z	Freq. observada	Freq. esperada	z
1-100	4	55,71 ± 7,4	-6,99	3	29,37 ± 5,37	-4,91
101-300	31	92,1 ± 9,52	-6,42	12	48,59 ± 6,87	-5,33
301-600	130	74,55 ± 8,58	6,46	74	39,3 ± 6,23	5,57
601-1000	150	46,11 ± 6,68	15,55	76	24,4 ± 4,93	10,47
1001-1500	84	26,77 ± 5,21	10,98	69	14,09 ± 3,72	14,76
1501-2000	44	13,63 ± 3,68	8,25	31	7,16 ± 2,67	8,93

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

Tabela B.2 – Valores da estatística z para as classes de distância da região flaqueadora 5' dos genes ortólogos na espécie *V. carteri*. São apresentados somente os dados relativos às classes de TEs de maior frequência de ocorrência.

Distância relativa ao SIT (pb)	non-LTR		LTR			
	Freq. observada	Freq. esperada	z	Freq. observada	Freq. esperada	z
1-100	15	40,89 ± 6,42	-4,03	7	31,08 ± 5,51	-4,37
101-300	34	71,87 ± 8,43	-4,49	20	54,98 ± 7,36	-4,75
301-600	82	79,71 ± 8,79	0,26	37	60,87 ± 7,72	-3,09
601-1000	116	78,67 ± 8,75	4,27	69	60,03 ± 7,66	1,17
1001-1500	114	70,19 ± 8,3	5,28	90	53,64 ± 7,22	5,04
1501-2000	109	52,94 ± 7,15	7,84	54	40,36 ± 6,33	2,15

Fonte: Elaborada pela autora.

Estimativas para p -valor: $|z| > 3,0902$: $p < 0,001$; $|z| > 2,3263$: $p < 0,01$; $|z| > 1,6449$: $p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

Tabela B.3 – Valores da estatística z para as classes de distância da região flaqueadora 3' dos genes ortólogos na espécie *C. reinhardtii*. São apresentados somente os dados relativos às classes de TEs de maior frequência de ocorrência.

Distância relativa ao códon de terminação (pb)	non-LTR			DNA		
	Freq. observada	Freq. esperada	z	Freq. observada	Freq. esperada	z
1-100	50	55,65 ± 7,37	-0,77	15	29,38 ± 5,37	-2,68
101-300	120	108,74 ± 10,33	1,09	64	57,27 ± 7,48	0,9
301-600	237	131,6 ± 11,27	9,35	142	69,32 ± 8,26	8,8
601-1000	269	98,45 ± 9,75	17,49	212	51,99 ± 7,05	22,7
1001-1500	168	55,68 ± 7,33	15,32	145	29,28 ± 5,4	21,43
1501-2000	76	25,87 ± 5,04	9,95	78	13,63 ± 3,66	17,59

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

Tabela B.4 – Valores da estatística z para as classes de distância da região flaqueadora 3' dos genes ortólogos na espécie *V. carteri*. São apresentados somente os dados relativos às classes de TEs de maior frequência de ocorrência.

Distância relativa ao códon de terminação (pb)	non-LTR			LTR		
	Freq. observada	Freq. esperada	z	Freq. observada	Freq. esperada	z
1-100	26	40,9 ± 6,31	-2,36	6	31,22 ± 5,59	-4,51
101-300	59	78,16 ± 8,78	-2,18	25	59,69 ± 7,63	-4,55
301-600	116	102,23 ± 9,9	1,39	41	77,99 ± 8,72	-4,24
601-1000	154	102,52 ± 9,95	5,17	69	78,41 ± 8,74	-1,08
1001-1500	180	87,57 ± 9,16	10,09	85	66,87 ± 8,14	2,23
1501-2000	116	62,7 ± 7,86	6,78	61	47,83 ± 6,77	1,95

Fonte: Elaborada pela autora.

Estimativas para p -valor: $|z| > 3,0902$: $p < 0,001$; $|z| > 2,3263$: $p < 0,01$; $|z| > 1,6449$: $p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

Tabela B.5 – Valores da estatística z para as classes de distância em íntrons nos genes ortólogos na espécie *C. reinhardtii*. São apresentados somente os dados relativos às classes de TEs de maior frequência de ocorrência.

Distância relativa ao éxon mais próximo (pb)	non-LTR			DNA		
	Freq. observada	Freq. esperada	z	Freq. observada	Freq. esperada	z
1-20	124	198,87 ± 13,89	-5,39	8	104,93 ± 10,01	-9,68
21-50	332	294,47 ± 16,55	2,27	30	155,36 ± 12,14	-10,33
51-100	274	407,62 ± 19,43	-6,88	31	215,01 ± 14,18	-12,98
101-200	222	333,06 ± 17,71	-6,27	50	175,75 ± 12,84	-9,79
201-400	93	95,39 ± 9,67	-0,25	20	50,29 ± 6,94	-4,36
> 400	25	39,97 ± 6,1	-2,45	17	21,04 ± 4,56	-0,89

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902 : p < 0,001$; $|z| > 2,3263 : p < 0,01$; $|z| > 1,6449 : p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.

Tabela B.6 – Valores da estatística z para as classes de distância em íntrons nos genes ortólogos na espécie *V. carteri*. São apresentados somente os dados relativos às classes de TEs de maior frequência de ocorrência.

Distância relativa ao éxon mais próximo (pb)	non-LTR			LTR		
	Freq. observada	Freq. esperada	z	Freq. observada	Freq. esperada	z
1-20	30	134,45 ± 11,48	-9,1	6	102,72 ± 10,03	-9,64
21-50	112	194,21 ± 13,79	-5,96	17	148,38 ± 11,88	-11,06
51-100	186	285,43 ± 16,49	-6,03	42	218,04 ± 14,48	-12,16
101-200	291	393,5 ± 19,01	-5,39	55	300,45 ± 16,79	-14,62
201-400	188	257,24 ± 15,54	-4,46	58	196,56 ± 13,64	-10,16
>400	78	104,5 ± 10,03	-2,64	41	79,88 ± 8,76	-4,44

Fonte: Elaborada pela autora.

Estimativas para p-valor: $|z| > 3,0902$: $p < 0,001$; $|z| > 2,3263$: $p < 0,01$; $|z| > 1,6449$: $p < 0,05$. Valores em negrito são significativos ($p < 0,01$) em relação ao esperado segundo a distribuição randômica de TEs no genoma.