FUNDAÇÃO UNIVERSIDADE FEDERAL DO ABC

JULIANA MELO RODRIGUES

DIVERSIDADE GENÉTICA DERIVADA DE ELEMENTOS DE TRANSPOSIÇÃO E SEU IMPACTO NO GENOMA FUNCIONAL VEGETAL: EXISTE CORRELAÇÃO ENTRE O PADRÃO DE INSERÇÃO DESSAS SEQUÊNCIAS E AS FUNÇÕES GÊNICAS?

Santo André
2023

JULIANA MELO RODRIGUES

DIVERSIDADE GENÉTICA DERIVADA DE ELEMENTOS DE TRANSPOSIÇÃO E SEU IMPACTO NO GENOMA FUNCIONAL VEGETAL: EXISTE CORRELAÇÃO ENTRE O PADRÃO DE INSERÇÃO DESSAS SEQUÊNCIAS E AS FUNÇÕES GÊNICAS?

Trabalho de Conclusão de Curso apresentado ao Centro de Ciências Naturais e Humanas da Universidade Federal do ABC como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas.

Orientador: Profª. Drª. Nathalia De Setta Costa (UFABC)

Santo André

2023

JULIANA MELO RODRIGUES

DIVERSIDADE GENÉTICA DERIVADA DE ELEMENTOS DE TRANSPOSIÇÃO E SEU IMPACTO NO GENOMA FUNCIONAL VEGETAL: EXISTE CORRELAÇÃO ENTRE O PADRÃO DE INSERÇÃO DESSAS SEQUÊNCIAS E AS FUNÇÕES GÊNICAS?

Trabalho de Conclusão de Curso apresentado ao Centro de Ciências Naturais e Humanas da Universidade Federal do ABC como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas.

Santo André, 14/08/2023

BANCA EXAMINADORA

_____

Profª. Drª Nathalia De Setta Costa (UFABC)
Universidade Federal do ABC

_____

Profa. Dra. Marcella Pecora Milazzoto
Universidade Federal do ABC

_____

Prof. Dr. Danilo Trabuco do Amaral
Universidade Federal do ABC

Dedico este trabalho aos meus pais e amigos que sempre me incentivaram.

AGRADECIMENTOS

Agradeço à Universidade Federal do ABC e ao Bacharelado em Ciências Biológicas pela oportunidade e aprendizado.

À minha orientadora, Profª. Drª. Nathalia de Setta Costa, por todo aprendizado, dedicação e paciência. Sendo, além de uma orientadora e professora excepcional, uma inspiração para minha vida e carreira.

Aos meus amigos que me acompanharam durante minha trajetória na universidade, Lara, Felipe, Helena, Maria, Henrique, por todos os momentos de compreensão e apoio.

Aos meus pais, José e Carmen, e a minha irmã, Gabriela, pelo apoio em todos os momentos e por sempre acreditarem no meu potencial!

RESUMO

Presentes em grande quantidade nos genomas vegetais, os elementos de transposição (TEs) variam em termos estruturais e funcionais e podem trazer mudanças significativas na informação genética, resultando na geração de novos fenótipos. Uma das formas pela qual os TEs geram variabilidade genética é pela alteração nos padrões de expressão gênica quando se inserem dentro ou na vizinhança de genes, não só afetando qualitativamente e quantitativamente a produção de proteínas, mas, também, impactando nas vias metabólicas associadas. O objetivo deste trabalho foi avaliar se existe correlação entre a distribuição dos TEs e a função dos genes nos quais eles estão inseridos nos genomas vegetais. Para isso, avaliamos 12 espécies vegetais, explorando, primeiramente, o mobiloma nos três contextos gênicos. Os resultados mostraram a superfamília *Helitron* com o maior número de TEs entre as espécies analisadas e para todos os contextos gênicos, seguida da superfamília *Mariner*. Em contraste, a superfamília *Politon*, presente apenas em *Arabidopsis thaliana*, obteve os valores mais baixos, o que é interessante já que diversos estudos relatam a presença de *Politons* em diversos eucariotos, mas nunca em *A. thaliana* e em outras plantas até então. *Oryza sativa e Zea mays* são as espécies que se destacam em número de TEs por superfamília nas regiões genômicas analisadas. Em seguida foi realizada uma análise da distribuição dos TEs em função das funções gênicas. Análise funcional mostrou que o grupo das monocotiledôneas apresentou a maior riqueza de TEs entre as diferentes categorias funcionais. Em comparação, entre as dicotiledôneas, *Populus trichorcarpa* foi a única espécie com maior destaque em inserções de TEs, sendo agrupada junto das monocotiledôneas. Avaliando o enriquecimento das categorias funcionais, *Solanum lycopersicum* foi a única espécie com grande destaque em inserções de TEs. Ainda, *Z. mays* foi a espécie com mais categorias empobrecidas entre todas as espécies analisadas, o que, em contraste com o tamanho de seu genoma, evidencia que o processo de expansão do genoma pelo aumento da quantidade de sequências repetitivas não envolveu um aumento da frequência de inserções em regiões gênicas. Finalmente, as categorias relacionadas ao estresse e metabolismo secundário apresentaram resultados interessantes, já que o enriquecimento nessas rotas pode estar relacionado a respostas adaptativas a condições de estresse biótico e abiótico.

Palavras-chave: evolução genômica; elementos de transposição; rotas metabólicas; adaptabilidade; funções gênicas.

ABSTRACT

Transposable elements (TEs) are present in significant quantities in plant genomes, varying in structural and functional terms and bringing meaningful changes in genetic information and phenotypes. TEs can generate genetic variability by altering gene expression patterns when they insert within or in the vicinity of genes, qualitatively and quantitatively affecting protein production, and also impacting the associated metabolic pathways. This work aimed to evaluate whether there is a correlation between the distribution of TEs and the function of the genes in which they are inserted in plant genomes. To this purpose, we evaluated 12 plant species genomes, by exploring the mobilome in three gene contexts. *Helitron* superfamily had the highest number of TEs in all species analyzed for all gene contexts, followed by the *Mariner* superfamily. In contrast, the *Politon* superfamily, present only in *Arabidopsis thaliana*, showed the lowest values in all gene contexts, which is interesting since no *Politons* have been identified in *A. thaliana* and other plants so far. *Oryza sativa and Zea mays* are the species that stand out in number of TEs per superfamily in the genomic regions analyzed. Also, *O. sativa* was the only species that showed a high number of *Mariner* superfamily elements. The functional analysis showed that the monocot species had the highest TE richness among the different categories. In comparison, among dicots, *Populus trichorcarpa* was the only species exhibiting a higher incidence of TE insertions, grouping together with monocots. The enrichment of gene functional categories analysis *Solanum lycopersicum* as the species with higher prominence of TE insertions. Also, *Z. mays* was the species with more underrepresented categories among all species analyzed, which, in contrast with its large genome size, evidences that the process of genome expansion by increasing the mobilome did not involve an increase in TE frequency of insertions in gene-rich regions. Finally, the functional categories related to stress and secondary metabolism showed interesting results, since the overrepresentation in these pathways may be related to adaptive responses to biotic and abiotic stress conditions.

Keywords: Genomic evolution; Transposable elements; Metabolic pathways; Adaptability; Gene functions.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

BINs	Categorias funcionais
cDNA	DNA complementar
EDTA	The Extensive de novo TE Annotator
FDR	Taxa de falsas descobertas
mRNA	RNA mensageiro
PR	Pathogenesis-related
PTGS	Silenciamento pós-transcricional
RNAseq	Sequenciamento de RNA
TEs	Elementos de transposição
TSD	Sítio alvo de duplicação

SUMÁRIO

1 INTRODUÇÃO

Ao observar uma alteração do padrão de coloração dos grãos de milho após cruzamentos, Barbara MacClintock, em 1940, fez uma descoberta revolucionária que mudaria a visão sobre o comportamento e composição dos genomas. Diferente do que se acreditava na época, o DNA, apesar de portador da herança genética e, portanto, responsável pela informação molecular a ser transmitida de forma mais integral possível para as próximas gerações (Jesus et al., 2017), não seria uma molécula completamente estável. Grande parte dessa instabilidade está relacionada com a existência dos, primeiramente, chamados elementos controladores, os quais possuem habilidade de se mover nos cromossomos e assim alterar a atividade gênica. Atualmente, com as diversas técnicas moleculares avançadas que possibilitam o sequenciamento de genomas completos, sabe-se que os antigos elementos controladores, agora denominados elementos de transposição (TEs), estão presentes em praticamente todos os seres vivos, podendo compreender uma grande parte do genoma para algumas espécies (Biémont; Viera, 2006).

1.1 DIVERSIDADE E CLASSIFICAÇÃO DOS ELEMENTOS DE TRANSPOSIÇÃO

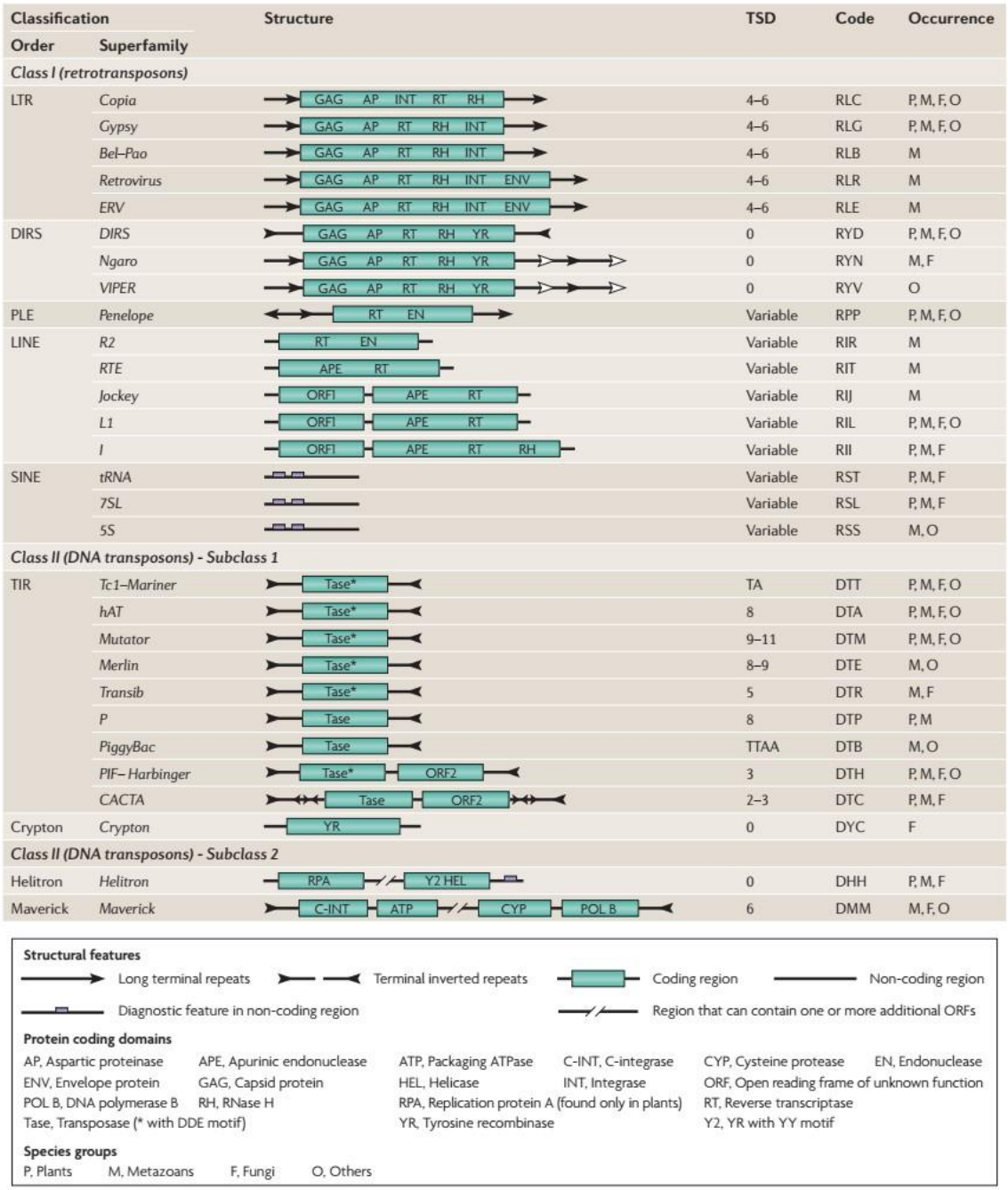
A estrutura básica de um TE é dada por repetições terminais, diretas ou invertidas, e um ou mais genes responsáveis pela codificação de proteínas necessárias à transposição. Além dessas proteínas, os TEs utilizam o aparato de transcrição e tradução da célula durante a sua mobilização. Caso ocorram mutações na sequência dos elementos, eles podem utilizar proteínas produzidas por outras inserções, se replicando de forma não-autônoma, em contraposição à elementos autônomos completos e intactos (Chénais et al.; 2012).

Os TEs são classificados em um sistema análogo ao taxonômico (Wicker et al., 2007). Nessa classificação, o nível mais amplo é o de classe, baseado no modo de transposição, sendo agrupados em Classe I (ou retrotransposons) ou em Classe II (ou transposons de DNA) (Figura 1). Os retrotransposons se mobilizam por meio da produção de um mRNA, que participará tanto da tradução das enzimas responsáveis pela transposição, como será o molde na transcrição reversa para a

produção de um DNA complementar (cDNA). O cDNA será guiado por uma das proteínas codificadas pelo elemento, a integrase, e inserido no DNA em uma nova posição do genoma. A inserção original do elemento é mantida e, dessa forma, o elemento é apenas copiado, sendo, portanto, um ciclo replicativo, fato que lhe concedeu a alcunha de mecanismo ‘copia e cola’ (Rubin et al., 2001). Por outro lado, os elementos de DNA excisam-se de suas posições originais com o auxílio de uma proteína codificada pelo próprio elemento, a transposase; o intermediário é, portanto, a própria sequência de DNA. Como não há a produção de uma nova cópia, esse mecanismo é denominado conservativo ou ‘recorta e cola’ (Rubin et al., 2001).

Níveis mais específicos de classificação dos TEs são as ordens, superfamílias e famílias. As ordens são dadas por diferenças nos mecanismos de inserção e consequentemente, na presença e organização dos genes dos elementos. Nas superfamílias, a distinção se encontra no comprimento dos elementos e no tipo do sítio alvo de duplicação (TSD, do Inglês, *target site duplication*). Finalmente, as famílias são classificadas de acordo com a similaridade na sequência codificante e nas repetições terminais (Wicker et al., 2007). As enzimas codificadas por elementos de mesma família apresentam alta taxa de conservação, enquanto as sequências intergênicas e as repetições terminais exibem menor conservação.

Figura 1 — Distribuição, classificação e estrutura dos TEs.



Fonte: Retirada de Wicker et al., 2007.

1.2 PARTICIPAÇÃO DOS TES EM GENOMAS VEGETAIS

Os genomas vegetais são compostos em grande parte de TEs. Alguns grupos de espécies, como as monocotiledôneas da família Poaceae (gramíneas), possuem representantes com altas proporções de TEs, chegando até a 80% no caso do genoma de milho (Schnable et al., 2009). A maior parte dos TEs vegetais costumam ser da ordem retrotransposons com LTR, mais especificamente da superfamília *Gypsy*, seguida pela superfamília *Copia*, transposons de DNA e finalmente de retrotransposons sem LTR. Estes elementos estão distribuídos ao longo de todo o genoma, embora exista uma maior frequência de inserções nas regiões centroméricas e pericentroméricas (Bennetzen et al., 2012). Os elementos das superfamílias *Gypsy* e *Copia* vegetais costumam ainda ser classificados em linhagens, um nível de classificação que é intermediário entre superfamílias e famílias, baseado nas relações filogenéticas das sequências codificantes. As linhagens são amplamente identificadas nos genomas analisados e são conhecidas como: *Ale/Retrofit, Angela/Tork, Bianca, Ivana/Oryco, Maximus/Sire e TAR/Tork* para a superfamília *Copia e CRM/CR, DEL/Tekay, Galadriel, Reina e TAT/Athila* para a superfamília *Gypsy* (Suguiyama et al., 2019).

1.3 IMPACTO DOS TES NA DIVERSIDADE E FUNCIONAMENTO DOS GENES

Devido à sua natureza móvel, os TEs possuem a capacidade de gerar mutações, e, portanto, têm o potencial de influenciar a evolução dos genomas. A mobilização dos TEs é regulada pelo genoma do hospedeiro, com o objetivo de minimizar mutações deletérias. Esse processo é denominado silenciamento de TEs. Em plantas, o silenciamento dos TEs pode ser dado pelo acúmulo de mutações que levam a perda de função gênica e, principalmente, por silenciamento epigenético (Fedoroff, 2012). Embora os mecanismos de silenciamento dos TEs sejam complexos e eficientes, existem situações em que eles podem escapar dessa regulação, e quando isso acontece, sua mobilização pode gerar diferentes impactos, a depender da região da nova inserção. Enquanto inserções nos éxons são, em sua maioria, deletérias, as inserções nas regiões regulatórias e em introns parecem ter

um menor impacto negativo, já que são mais frequentemente fixadas e, provavelmente por isso, se apresentam em maior frequência nos genomas eucariotos (Kim, 2017). Muitas vezes, as inserções em introns não apresentam efeitos significativos na sequência das proteínas, mas podem interferir nos padrões de transcrição, resultando em alterações no fenótipo. Já uma inserção na região promotora, por exemplo, pode afetar diretamente a taxa de transcrição, alterando a expressão do gene de forma relevante, por meio da interrupção ou geração de novas sequências regulatórias.

As espécies eucariotas evoluíram mecanismos de silenciamento dos TEs baseados em compactação da cromatina dirigida por marcas epigenéticas (silenciamento transcricional) e degradação de mRNAs pela maquinaria de RNA de interferência (silenciamento pós-transcricional), ambos dependentes de pequenos RNAs regulatórios homólogos a TEs (Matzke; Mosher, 2014). Embora o silenciamento dos TEs possa controlar os efeitos mutagênicos da mobilização, ele também pode ter impacto na funcionalidade do genoma. Quando um TE se insere em uma região gênica, promotora ou codificante, seu silenciamento epigenético pode afetar os padrões de expressão do gene hospedeiro, gerando epialelos derivados de TEs. Por exemplo, se o TE se inserir no promotor do gene e for silenciado por meio de remodelamento da cromatina, o gene vizinho ao TE também poderá ser silenciado (Feschotte, 2008). Um exemplo da existência de epialelos derivados de TEs foi descrito em tomateiro. O incremento de vitamina E em tomates, observado em plantas submetidas a condições de estresse abiótico, foi associado com o aumento na transcrição do gene VTE3, mediado pela redução nos níveis de metilação de um TE do tipo SINE inserido na região promotora desse gene (Quadrana et al., 2014). Ainda, se o TE se inserir dentro de um gene e passar a fazer parte de sua informação codificante, esse gene poderá ser silenciado pelo mecanismo pós-transcricional, em que os transcritos são degradados em uma via homóloga à de microRNAs (Feschotte, 2008). O mecanismo de silenciamento pós-transcricional (PTGS) é estudado em diversos organismos como um sistema de defesa a invasão por vírus e TEs, e capaz de limitar a expansão dessas sequencias repetidas. Um exemplo disso foi observado em um estudo com o fungo *Neurospora crassa*, no qual o mecanismo de PTGS funciona independentemente da metilação

do DNA e está envolvido da repressão do retrotransposon LINE1- like (Nolan et al.; 2005).

Ao alterar a expressão de um gene, os TEs podem estar diretamente relacionados a mudanças mais complexas, podendo afetar não somente uma proteína em particular, mas toda a cadeia de reações dependente dela. Quando vias metabólicas - essenciais para garantir a homeostasia do organismo - são afetadas de modo relevante, essas podem causar mudanças na função e, consequentemente, na adaptabilidade do organismo. Portanto, o potencial mutagênico presente nos TE pode influenciar de forma positiva ou negativa a adaptação do hospedeiro, podendo resultar em um valor adaptativo maior ou menor dependendo da intervenção das forças evolutivas – seleção natural e deriva genética (Bourque, 2018). Uma questão que ainda não foi detalhadamente analisada é se existe algum tipo de viés entre a fixação de inserções de TEs e a função dos genes nos quais eles se inserem. Será que genes que participam de rotas metabólicas primárias apresentam maiores restrições seletivas para a inserção de TEs do que genes de rotas metabólicas secundárias? Este projeto de pesquisa abordará essa temática em uma tentativa de produzir informações sobre esse tópico.

2 OBJETIVO

2.1 OBJETIVOS GERAIS

O objetivo geral deste trabalho foi avaliar se existe correlação entre a distribuição dos TEs e a função dos genes nos quais eles estão inseridos em genomas vegetais.

2.2 OBJETIVOS ESPECÍFICOS

- Minerar informações sobre a anotação dos genes codificantes de proteínas e TEs em genomas vegetais disponíveis no repositório Phytozome;
- Categorizar os genes anotados em rotas metabólicas e grupos funcionais;
- Analisar a posição relativa de inserção dos TEs nos genes codificantes de proteínas;
- Buscar por correlações entre a inserção de TEs e a categorização funcional dos genes.

3 METODOLOGIA

3.1 MATERIAL

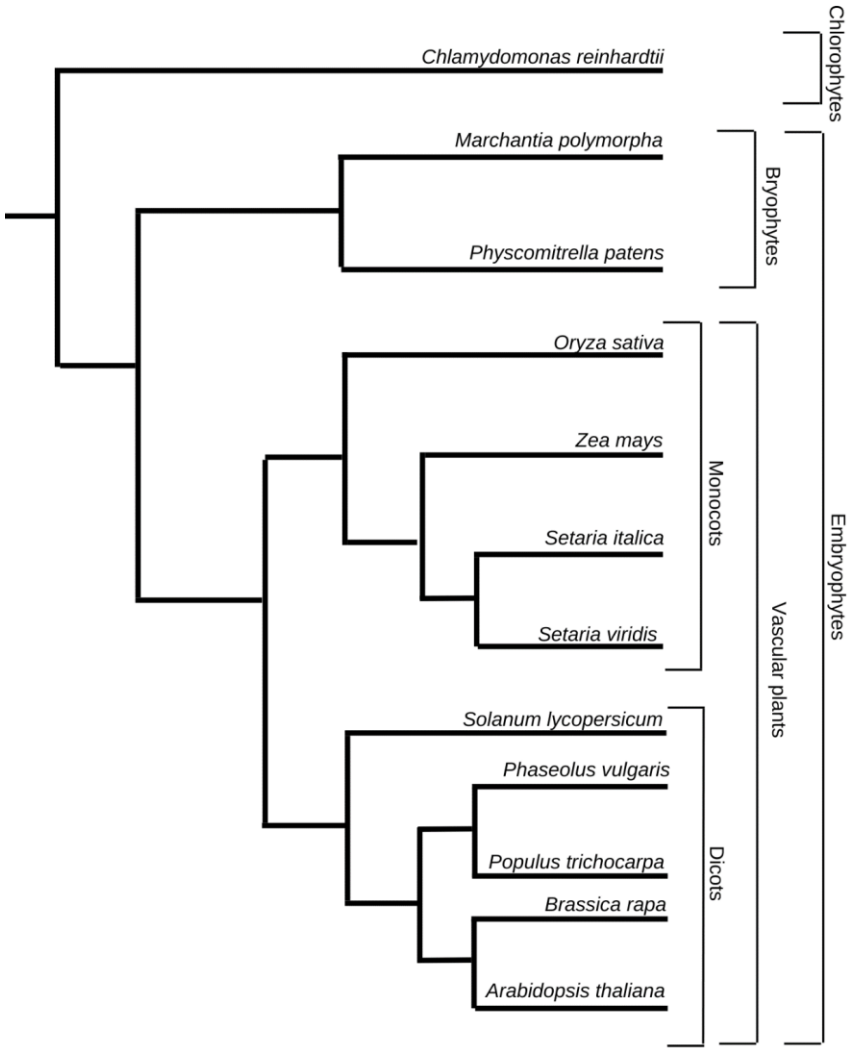
Este trabalho analisou 12 espécies vegetais (Figura 2). Os genomas foram obtidos no repositório de dados Phytozome (phytozome.jgi.doe.gov/pz/portal.html#), que permite acessar, visualizar e analisar genomas de plantas já sequenciados por diversos grupos de pesquisa ao redor do mundo. Foram baixados todos os genomas (formato .fasta) a serem analisados (Tabela 1), bem como arquivos dos genes codificantes de proteínas em aminoácidos e as anotações mais recentes dos genes (formato .gff3). A partir dos dados de anotação dos genes obtidos no Phytozome, construímos arquivos (.gff3) com as coordenadas de três contextos gênicos, utilizando o programa Excel: (i) região 5', correspondente a 2 kb *upstream* do sítio de início da transcrição; (ii) região corpo do gene: região entre os sítios de início e terminação da transcrição, contendo regiões transcritas e não-traduzidas 5' e 3', sequências codificantes e introns, e; (iii) região 3': correspondente a 2 kb *downstream* do sítio de finalização da transcrição.

Tabela 1 — Lista das espécies analisadas, suas respectivas classificações taxonômicas, nomes comuns e versões dos genomas disponíveis no Phytozome. As classificações foram obtidas nas bases de dados World Flora Online (www.worldfloraonline.org), NCBI Taxonomy Browser (www.ncbi.nlm.nih.gov/taxonomy), AlgaeBase (www.algaebase.org), Integrated Taxonomic Information System (www.itis.gov) e Global Biodiversity Information Facility (www.gbif.org/pt/).

Divisão	**Classe**	**Família**	**Espécie**	**Nome comum**	**Genoma**
Bryophyta	Bryopsida	Chlamydomonadaceae	*Chlamydomonas reinhardtii*	-	v5.6
		Funariaceae	*Physcomitrella patens*	Musgo	v3.3
Magnoliophyta	Liliopsida	Poaceae	*Oryza sativa*	Arroz	v7.0
			Setaria italica	Painço Moha	v2.2
			Setaria viridis	Green foxtail millet	v2.1
			Zea mays	Milho	*Ensembl18*
	Magnoliopsida	Brassicaceae	*Arabidopsis thaliana*	Rockcress	TAIR 10
			Brassica rapa	Nabo	FPsc V1.3
		Fabaceae	*Phaseolus vulgaris*	Feijão	v2.1
		Salicaceae	*Populus trichocarpa*	Poplar	v4.1
		Solanaceae	*Solanum lycopersicum*	Tomate	ITAG3.2
Marchantiophyta	Marchantiopsida	Marchantiaceae	*Marchantia polymorpha*	Common liverwort	v3.1

Fonte: O autor (2023).

Figura 2 — Árvore filogenética indicando as relações filogenéticas entre as 12 espécies avaliadas no projeto.

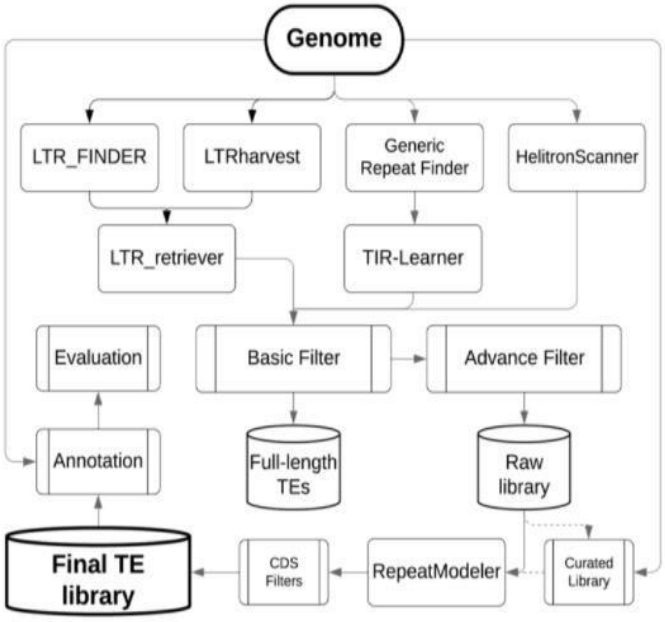


Fonte: O autor (2023).

3.2 IDENTIFICAÇÃO DOS TES

O mapeamento dos TEs nos genomas foi realizado com o *pipeline* EDTA (do Inglês, *The Extensive de novo TE Annotator*), em uma colaboração com o Dr Nicolas Bellora, pesquisador do Institute of Nuclear Technologies for Health (INTECNUS, Argentina). A Figura 3 ilustra o processo do funcionamento do EDTA. Os genomas são primeiramente utilizados para anotação dos retrotransposons com LTR com os programas LTR_FINDER (Xu; Wang, 2007), LTRharvest (Ellinghaus et al.; 2008) e LTR_retrivier (Ou; Jiang, 2018). O uso de cada um desses programas possui pontos positivos e negativos, por exemplo, LTR_FINDER possui o melhor balanço de desempenho, porém é muito mais lento quando comparado aos outros. Já o LTR_retriever possui uma alta especificidade, precisão e uma baixa taxa de falsas descobertas. Utilizando os três programas em conjunto é possível combinar um melhor desempenho com um menor processamento. Ao mesmo tempo em que os anotadores de retrotransposons com LTR são aplicados, é realizada a anotação dos transposons de DNA utilizando os programas Generic Repeat Finder (Shi; Liang, 2019) associado ao TIR-Learner (Weijia et al.; 2019) e o HelitronScanner (Xiong et al.; 2014).

Figura 3 — Esquema ilustrativo do funcionamento do EDTA.



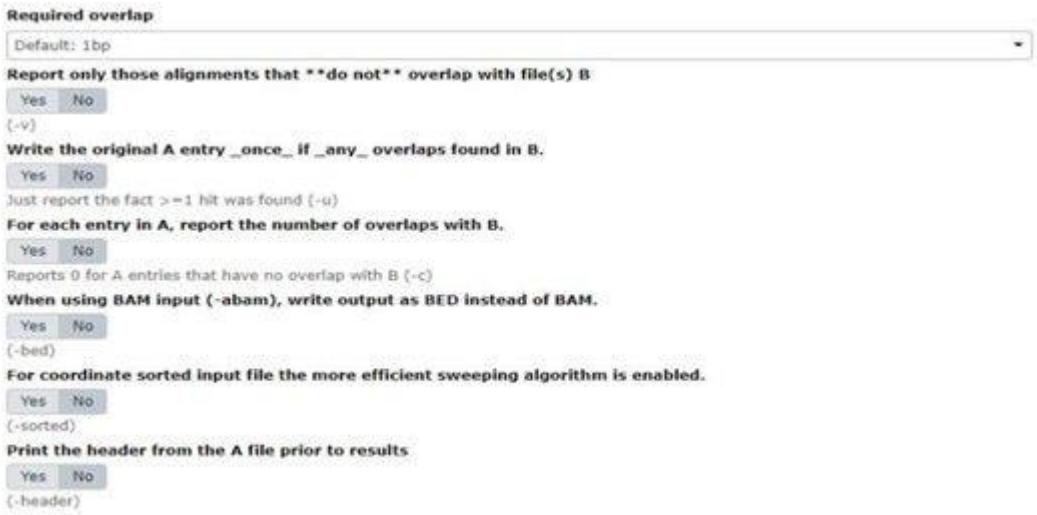
Fonte: Retirado de Ou et al., 2019.

Após a etapa de anotação são aplicados filtros. O filtro básico é aplicado para remover sequências curtas e repetições em tandem. O filtro avançado é aplicado para remover falsos positivos por meio de uma análise das repetições terminais dos TEs. A biblioteca (*raw library*) produzida é então introduzida a mais um anotador de repetições geral, o RepeatModeler (Hubley; Smit, 2015), o qual possui a mais alta performance entre os anotadores, sendo capaz de produzir resultados compactos e gerar uma classificação das anotações dos TEs. O resultado é uma biblioteca condensada de TEs que dará origem a uma anotação mais precisa, confiável e com uma menor taxa de erros (Ou et al.; 2019). Ao final da aplicação do *pipeline* é gerado um arquivo tabular do tipo gff3 que contém as seguintes informações para cada um dos TEs anotados: localização cromossômica, nucleotídeos de início e fim, classificação, sentido, comprimento, identidade com o *query* e método de identificação.

3.3 INTERSECÇÃO DOS DADOS DE ANOTAÇÃO ESTRUTURAL DOS GENES E DOS TES

Os dados gerados pelo pipeline EDTA foram associados aos dados de anotação estrutural dos genes codificantes de proteínas obtidos no Phytozome para que possamos quantificar quantos e quais TEs são inseridos em cada um dos três contextos gênicos. Para isso nós utilizamos a ferramenta ‘bedtools Intersect Intervals’ do servidor Galaxy Pasteur (Quinlan et al., 2010) e a aplicamos com êxito para todas as espécies, utilizando os parâmetros descritos abaixo (Figura 4). A partir das intersecções foi possível identificar o número, o tamanho e a classificação dos TEs para todas as espécies analisadas. Para ilustrar esses resultados foram produzidos *heatmaps*, utilizando a ferramenta ‘Heatmap 2’ do servidor Galaxy Main (https://usegalaxy.org/), tanto para o número, como para o comprimento de elementos por superfamília (Figura 7-8).

Figura 4 — Parâmetros utilizados na ferramenta bedtools intersect intervals.



Fonte: Quinlan et al., 2010

3.4 CATEGORIZAÇÃO FUNCIONAL DOS GENES

A categorização funcional dos genes foi feita pela plataforma MapMan (Usadel et al., 2009), a qual permite a criação de diagramas das vias metabólicas e processos biológicos no qual os genes estão envolvidos. O MapMan foi desenvolvido para a análise de dados de expressão gênica, como os gerados pela metodologia de microarranjos e RNAseq, no entanto, neste trabalho, utilizamos essa ferramenta para mostrar a frequência de inserção dos TEs nas regiões gênicas e relacionar as vias metabólicas em que eles podem estar atuando na geração de variabilidade genética. A primeira etapa de aplicação do MapMan é realizada por meio da anotação funcional dos genes dos genomas por meio do sistema de anotação funcional Mercator 3.6 (Lohse et al., 2014). Para isso, utilizamos o arquivo de predição de proteínas adquirido no Phytozome e os parâmetros *default*. O Mercator gera um arquivo de anotação funcional dos genes, organizando as proteínas em categorias funcionais (BINs), que são parte em uma estrutura hierárquica de contexto biológico de acordo com a função das proteínas *query* das bases de dados (Lohse et al., 2014). De posse da anotação funcional do Mercator e dos dados de número de inserções

de TEs por gene, construído para cada contexto gênico, foi gerada uma análise da ‘riqueza´, referente aos valores da razão do número de genes com TEs em cada BINs e SUB-BINs pela quantidade total de genes na categoria. A partir dos valores de riqueza foram gerados *heatmaps,* utilizando a ferramenta ‘Heatmap 2’ do servidor Galaxy Main (https://usegalaxy.org) com parâmetros *default,* indicando as categorias com maior número de inserções de TEs nos genes (em vermelho).

A ferramenta PageMan (Usadel et al., 2005) do programa MapMan permite a análise de representação de uma série de dados por meio de diferentes métodos estatísticos e da correção de teste múltiplos, gerando um resultado que permite a visualização de quais categorias funcionais (BINs e SUB-BINs) possuem mais genes com TEs do que o esperado. Para realizar esta análise é necessária a utilização do arquivo de categorização funcional do Mercator e o arquivo com o número de TEs por gene. A análise de representação usa o Teste Exato de Fisher usado para determinar se existe ou não uma associação significativa entre duas variáveis categóricas, permitindo testar se para determinada classe o número de objetos é dado ao acaso (Usadel et al., 2005). Como múltiplas hipóteses são testadas de uma só vez é essencial a utilização de algum método de correção de testes múltiplos. Neste trabalho utilizamos a correção de Benjamini-Yekutieli, a qual permite controlar a taxa de falsas descobertas (FDR) sem a necessidade de suposições sobre como as várias comparações se correlacionam entre si (Benjamini; Yekutilie, 2001). Quando utilizadas correções de controle de taxa de falsas descobertas, o teste gera um ‘P-valor´ ajustado, que é o valor FDR (do Inglês, *False Discovery Rate*). O PageMan ainda converte o p-valor em um valor de z-score, representando o número de desvios padrão em relação à média de um ponto de informação. Neste trabalho usamos como *cut-off* z-*scores* ≤ -1,96 e ≥ 1,96, que equivalem a um P-valor de 0,05.

4 RESULTADOS E DISCUSSÃO

Todas as análises automatizadas descritas acima foram realizadas para todas as 12 espécies propostas no trabalho, englobando representantes do grupo das clorófitas, briófitas, monocotiledôneas e dicotiledôneas (Tabela 1 e Figura 2). Os resultados serão apresentados em duas partes, na primeira parte teremos a exploração dos mobilomas, destrinchando o número e comprimento dos TEs nos três contextos gênicos (região 5', corpo do gene e região 3') e, também, realizando uma análise comparativa entre as superfamílias das espécies avaliadas. Na segunda parte avaliaremos a relação entre distribuição dos elementos pelas categorias funcionais e a função gênica, revelando categorias funcionais e regiões gênicas diferencialmente representadas com inserções de TEs quando comparado com o genoma completo.

4.1 ANÁLISE DO MOBILOMA

A anotação dos TEs com o *pipeline* EDTA permitiu identificar inserções nos três contextos genômicos (região 5', corpo do gene e região 3') das espécies alvo do estudo. Identificamos *Marchantia polymorpha* como a espécie com menor quantidade de TEs para todos os contextos gênicos, sendo 7.031 na região 5', 3.914 no corpo do gene e 7.882 na região 3' (Anexo A). *M. polymorpha* foi a primeira espécie de hepática com o genoma sequenciado, sendo uma linhagem basal de plantas terrestres não vasculares (Bowman et al., 2017). A baixa quantidade de TEs encontrados pode estar relacionada tanto com a complexidade de seu genoma, como com o seu tamanho (225,761 Mb), sendo o terceiro menor genoma avaliado neste estudo (Bowman et al., 2017). Além disso, seus elementos repetitivos equivalem a apenas 22% do genoma autossômico, com os retrotransposons com LTR representando a maior fração com 9.7% (Bowman et al., 2017). Neste trabalho, avaliando apenas regiões gênicas, os TEs representam 7.7%, sendo 2,7% referente a retrotransposons com LTR (Anexo A).

Por outro lado, a espécie que se destaca com o maior número de TEs foi *O. sativa* (371.381), seguida de *Z. mays* (300.663). Apesar de ter um genoma quase

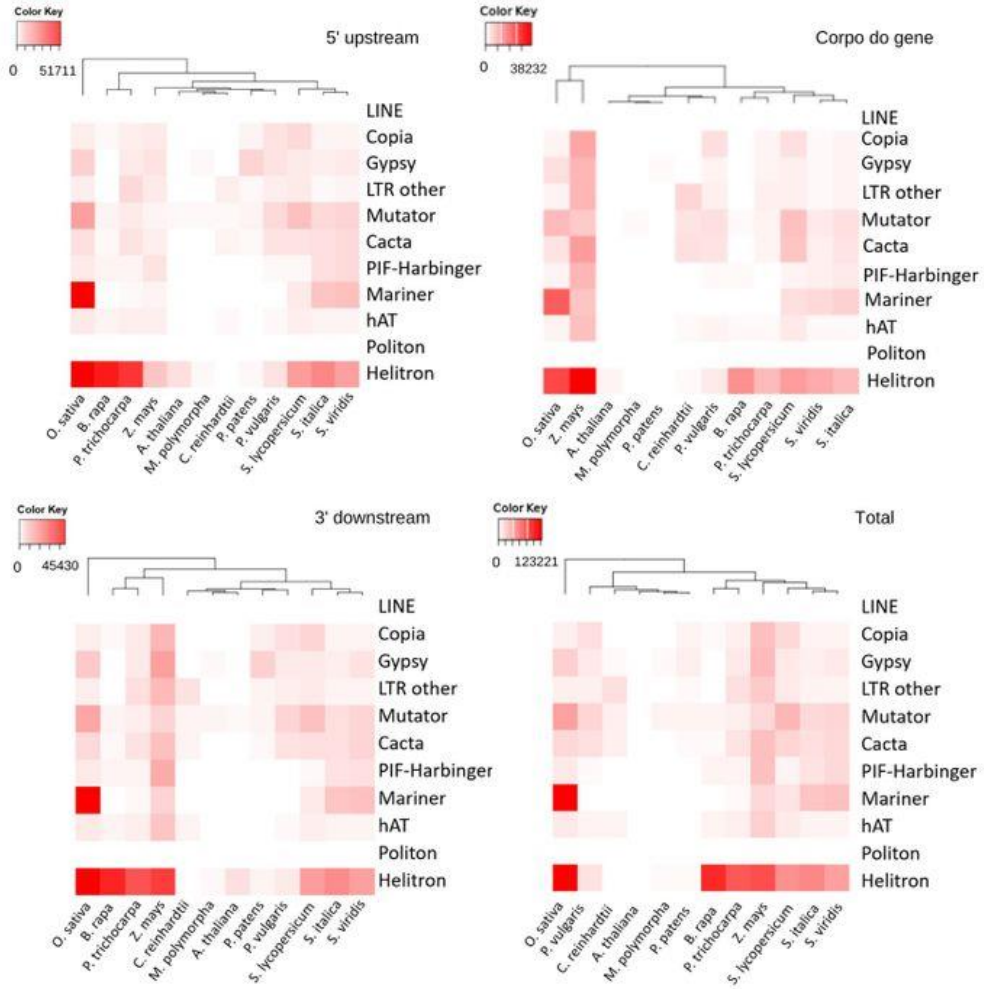
seis vezes menor, *O. sativa* exibiu uma quantidade substancialmente maior de TEs em comparação a *Z. mays* (Tabela 2). É comumente aceito que a variação no tamanho do genoma vista entre as espécies eucarióticas está mais estreitamente correlacionada com a quantidade de DNA repetitivo (TEs, DNAs satélites e repetições em tandem) do que com o número de genes codificadores (Kidwell, 2002). Portanto, o tamanho do genoma de *Z. mays* - sendo o maior entre os genomas avaliados, deve estar diretamente relacionado a quantidade de DNA repetitivo e, mais especificamente, o alto número de TEs presentes dentro e na vizinhança das sequências codificadoras de proteínas. Embora *O. sativa* tenha um número maior de TEs em seu genoma em comparação com *Z. mays*, é importante destacar que a correlação entre a quantidade de TEs e o tamanho do genoma parece não ser uma relação rígida, especialmente quando consideramos as regiões gênicas. Essa exceção na correlação entre o tamanho do genoma e a quantidade de TEs vista para a espécie pode ser atribuída a casos de amplificação seletiva de TEs em regiões específicas do genoma de O.*sativa,* assim como a mecanismos de regulação que controlam o funcionamento e a propagação desses elementos.

Em relação ao agrupamento das espécies no *heatmap*, as monocotiledôneas *S. italia e S. viridi*s são agrupadas juntas em todos os contextos gênicos (Figura 5), o que é esperado já que espécies irmãs compartilhando similaridades evolutivas. Ainda, analisando as demais monocotiledôneas, *O. sativa e Z. mays*, são agrupadas próximas em quase todos os contextos gênicos, porém distante das outras monocotiledôneas (Figura 5). Isso pode ser explicado pelo destaque em número de TEs observado para as duas espécies, sendo as duas com maior número de TEs entre todas as espécies avaliadas. Ainda, *Solanum lycopersicum* é agrupada mais próxima com as monocotiledôneas, *S. italica e S. viridis*, do que com o restante das dicotiledôneas (Figura 5). Isso porque, *S. lycopersicum* (180.653) possui um valor de número de TEs mais próximo de *S. italica* (157.103) e *S. viridis* (169.645), do que em comparação com as dicotiledôneas, *P. trichorcapa* (155.619), *B. rapa* (130.408), *P. vulgaris* (97.043), *A. thaliana* (23.887), sendo, entre esse grupo, a espécie com maior número de TEs (Anexo A).

Em relação as superfamílias, o maior número de inserções de TEs é encontrado para a superfamília *Helitron* em todas as espécies analisadas e para todos os contextos gênicos, totalizando 586.370 inserções (Figura 5 e Anexo A).

Analisando a somatória dos três contextos gênicos para o número de elementos *Helitron* por espécies, temos uma variação de 85.090 elementos em *Z. mays* a 2.441 elementos em *Chlamydomonas reinhardtii* (Anexo A). Cabe ressaltar que *Helitron* é uma ordem de elementos, sendo que o *pipeline* EDTA não subdivide esses TEs em superfamília por falta de informações sobre a diversidade dessas sequências nas bases de dados. Portanto, o alto número de inserções desses elementos, observada nos resultados, pode estar ligada a falta de uma classificação mais específica, o que acaba levando a serem agrupados como um todo, como feito frequentemente nos estudos que os descrevem (Thomas; Pritham, 2015).

Figura 5 — Número de TEs por superfamília em cada contexto gênico, clusterizando os dados pelas espécies.



Fonte: O autor (2023).

Por outro lado, diversos estudos relataram um número variável de *Helitrons* no genoma de diferentes espécies de plantas. Em *Arabidopsis*, esse número chega a 2%, enquanto para *Z. mays* temos um valor de 6.6% e, para outras espécies, uma variação de 0.1–4.3% de elementos *Helitron* na composição do genoma completo (Hu et al., 2019). Ainda, para *A. thaliana* já foi relatada uma maior abundância desses elementos nas regiões pobres em genes, enquanto para a espécie *Z. mays*, *Helitrons* estão presentes principalmente nas regiões ricas em genes (Hu et al, 2019). Esses dados podem explicar os resultados obtidos em relação as duas espécies, já que neste projeto avaliamos apenas as regiões gênicas, observando um maior número de *Helitrons* para *Z. mays* (85.090) do que para *A. thaliana* (14.718) (Anexo A). A abundância de *Helitrons* traz um resultado interessante para esse estudo, já que são agentes importantes em relação a evolução dos genomas vegetais, participando dos rearranjos e duplicação de regiões genômicas (Bennetzen et al., 2005). Alguns estudos relataram a capacidade desses elementos em alterar fenótipos quando há inserção em regiões promotoras, levando a mudanças nos padrões de expressão (Hu et al., 2019). Um exemplo disso, foi demonstrado para a espécie *Brassica rapa*, em que a inserção de um *Helitron* de 4.3 Kb em um íntron do gene *BrTT8* resultou em uma casca de semente amarelada (Li et al., 2012).

A superfamília *Mariner* é a segunda mais abundante em número de TEs em todos os contextos gênicos, totalizando 219.394 inserções (Anexo A). Analisando a soma do número de TEs nas 12 espécies temos, 86.364 TEs para a região 5', 52.597 para o corpo do gene e 80.433 para a região 3' (Anexo A). Cabe ressaltar que *O. sativa* foi a espécie que mostrou os maiores valores para número de elementos da superfamília *Mariner*, sendo 51.711 na região 5', 24.067 no corpo do gene e 45.388 na região 3', o que pode explicar a presença dessa espécie como grupo externo nos *heatmaps* para as regiões 5' e 3', sendo agrupada distante das outras monocotiledôneas nesses dois contextos (Figura 5). Os elementos *Mariner* são classificados em DNA transposons (Classe II) e em *O. sativa* e outras espécies de gramíneas são a superfamília mais abundante dessa classe no genoma completo (Roffler; Wicker, 2005).

Em contraste com o alto número de TEs de *Helitrons* e *Mariners*, a superfamília *Politon*, presente apenas em *A. thaliana*, obteve os valores mais baixos nos três contextos gênicos, sendo dois TEs na região 5', 16 no corpo do gene e 12

na região 3’ (Anexo A). Os elementos *Politon*, também chamados de *Mavericks*, são considerados os mais complexos elementos da classe de DNA transposons em eucariotos, sendo capazes de codificar diferentes proteínas, como a DNA polimerase B, retroviral-like integrase e adenoviral-like protease (Kapitonov; Jurka, 2006). Neste estudo, os elementos *Politon* foram encontrados em pequenos números e apenas na espécie *A. thaliana*. Esse resultado é de interesse já que diversos estudos relatam a presença de *Politons* em diversos eucariotos, como invertebrados, vertebrados não mamíferos, fungos e eucariotos unicelulares, mas eles nunca foram descritos em *A. thaliana* (Quesneville, 2020) e em outras plantas até então (Pritham; Putiwala; Feschotte, 2007).

Tabela 2 — Composição dos genomas analisados neste estudo.

Species	Total scaffold length (Mb)	% TEs	% TEs found in this study	Number of TEs in the genome	Number of TEs found in this study
A. thaliana	119,667	21% [a]	25%	32,000 [a]	23,887
B. rapa	315,053	43% [b]	32%	387,420 [b]	130,408
C. reinhardtii	111,100	-	18%	-	46,243
M. polymorpha	225,761	-	8%	-	18,827
O. sativa	374,471	30% [c]	69%	-	371,381
P. vulgaris	537,218	-	9%	-	97,043
P. patens	473,226	57% [d]	8%	-	42,803
P. trichocarpa	392,162	44% [e]	22%	427,901 [e]	155,619
S. italica	405,737	40% [f]	25%	-	157,103
S. viridis	395,731	46% [g]	29%	-	169,645
S. lycopersicum	828,076	66% [h]	11%	665,122 [h]	180,653
Z. mays	2.135,083	85% [i]	11%	338,224 [i]	300,663

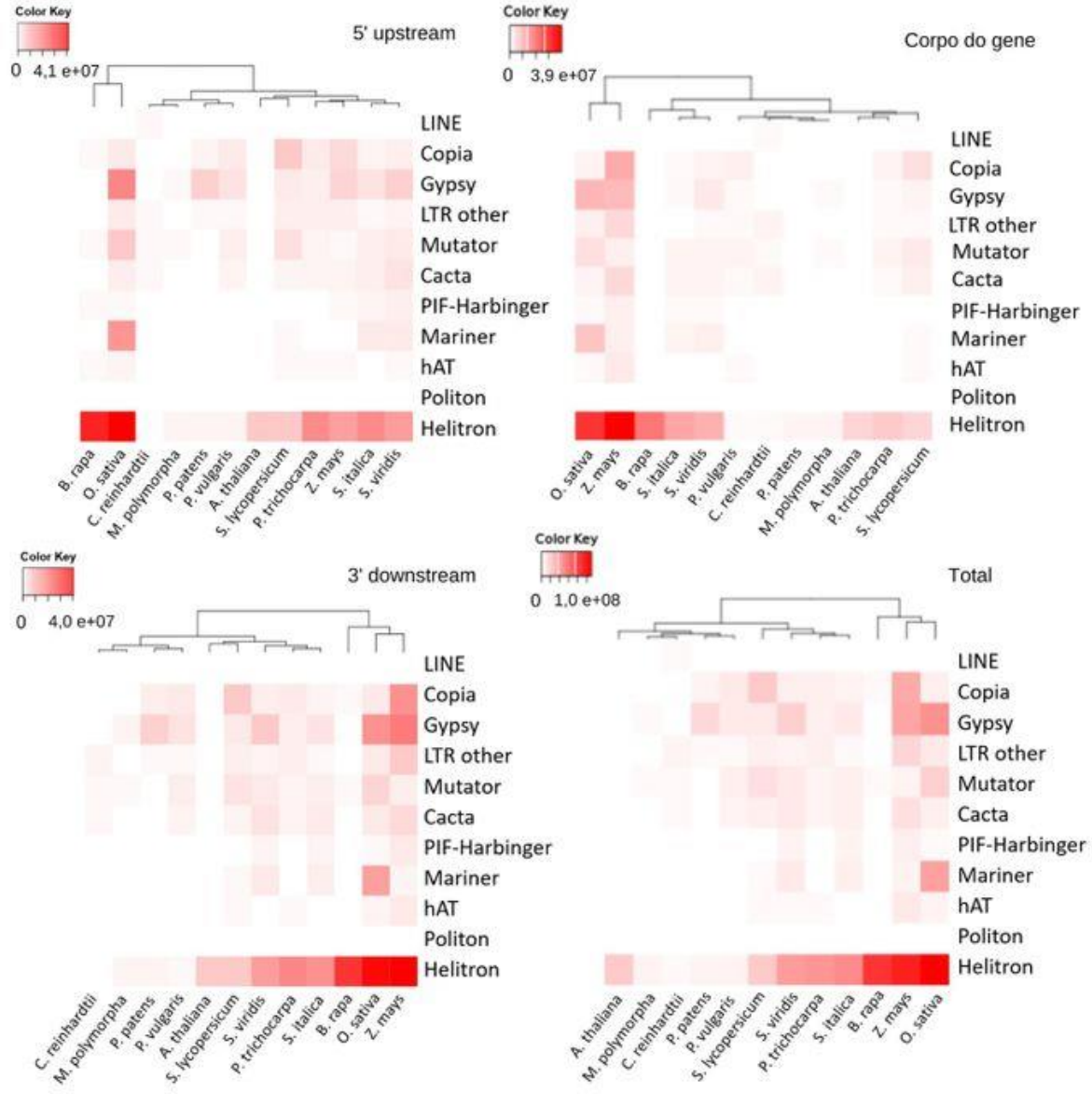
[a]Quesneville, 2020; [b]Panget al., 2015; [c]Vendrell-mir et al., 2020; [c]Vendrell-mir et al., 2020; [e]Zhao et al., 2022; [f]Bennetzen et al., 2012; [g]https://phytozome-next.jgi.doe.gov/; [h]Su et al., 2021; Domínguez et al., 2020; [i]Anderson et al, 2019.

Em geral, os genomas completos apresentaram uma maior porcentagem de número de TEs do que o encontrado para as regiões gênicas (Tabela 2). Esse resultado é esperado, já que deve existir uma pressão de seleção purificadora maior nas regiões gênicas, eliminando com maior eficiência novas inserções. Todas as espécies analisadas seguiram esse padrão, com exceção de duas espécies, *A. thaliana e O. sativa*. É possível que a existência de características específicas ou eventos de remodelamento do genoma possam ter favorecido a mobilidade e inserção de TEs nessas áreas genômicas, podendo estar relacionadas a sistemas

de regulação ou modificações epigenéticas que tornem as regiões gênicas mais acessíveis aos TEs nessas espécies.

Em seguida, analisamos os dados do comprimento dos TEs em cada contexto gênico. Nesse sentido, não foi observada uma diferença importante (Anexo A). Para a região 5', os maiores e menores valores foram para *O. sativa* e *M. polymorpha*, com 99.956 e 6.140 Kb, respectivamente (Anexo A). Para o corpo do gene, *Z. mays* apresentou o maior valor, 87.507 Kb, enquanto *Physcomitrella patens* apresenta o menor valor, 4.843 Kb. Ainda, para a região 3', temos *Z. mays* com 106.023 Kb e *C. reinhardtii* com 6.440 Kb, sendo o maior e menor comprimento, respectivamente. Podemos observar que o grupo das briófitas (*M. polymorpha e P. patens*) e o grupo das clorófitas (*C. reinhardtii*) foram agrupadas próximas em todas os contextos gênicos (Figura 6). Isso porque, tiveram os menores valores para comprimento dos TEs, o que era esperado, já que estas são as espécies mais basais, com os menores genomas avaliados, com exceção de *P. patens* (Tabela 2).

Figura 6 — Comprimento dos TEs por superfamília em cada contexto gênico, clusterizando os dados por espécies.



Fonte: O autor (2023).

Os resultados de comprimento também mostraram uma distribuição que favorece as superfamílias *Helitron* (95.457 Kb), *Copia* (37.163 Kb) e *Gypsy* (39.040 Kb) em *Z. mays* e *Helitron* (110.818 Kb), *Gypsy* (48.160 Kb) e *Mariner* (41.363 Kb) em *O. sativa* (Anexo A e Figura 6). Interessantemente, os *Helitrons* são muito mais representativos que todas as outras superfamílias, em todos os contextos gênicos (Figura 6). Os retrotransposons representam a classe mais frequente de elementos encontrados para a maioria das espécies quando se analisa o genoma completo. Por

exemplo, em *Z. mays*, eles compreendem 50% do genoma (Anderson et al., 2019) e em *A. thaliana* compõem a maior parte das sequências de TEs, seguida dos *Helitrons* e DNA transposons (Quesneville, 2020). Em *O. sativa* os retrotransposons com LTR compõem pelo menos 17% do genoma (Ou; Jiang, 2018) e em *S. italica* essa mesma ordem constitui entre 25% e 30% do conteúdo nuclear total (Suguiyama et al., 2019). Diferente dos resultados encontrados para o genoma completo, nossos resultados, avaliando o contexto dos genes, mostraram que a maioria das espécies avaliadas possui uma maior quantidade de DNA transposons do que retrotransposons, com exceção apenas de *Phaseolus vulgaris* e *P. patens* (Anexo A). Mesmo não incluindo os *Helitrons* na somatória do comprimento dos TEs, os valores de comprimento de retrotransposons e elementos de DNA ficam próximos, sem um padrão claro de preponderância (Anexo A).

4.2 ANÁLISES FUNCIONAIS

Por conta da sua natureza mutagênica, os TEs existem nos genomas eucariotos majoritariamente em uma forma silenciada, mas reversível, dirigida pelo mecanismo de silenciamento epigenético (Fedoroff et al., 2012). Portanto, é observada que a atividade dos TEs está, em grande parte, inativa durante o ciclo de vida dos organismos, mas pode retornar ao seu estado ativo, em situações não muito bem delimitadas, mas que em geral envolvem algum tipo de estresse (Wessler, 1996). Como parte dos TEs estão inseridos dentro ou próximo de genes, modificações epigenéticas acabam afetando não somente a atividade dos TEs, mas também, dos genes localizados em sua vizinhança. Dessa forma, assim como o silenciamento dos TEs pode inibir a atividade dos genes próximos, ou em que os elementos estão inseridos, quando os TEs voltam a sua forma ativa os genes também podem passar a ser mais expressos. Essa capacidade dos TEs em agir como controladores da expressão gênica torna interessante o estudo da sua distribuição em relação as categorias gênicas funcionais.

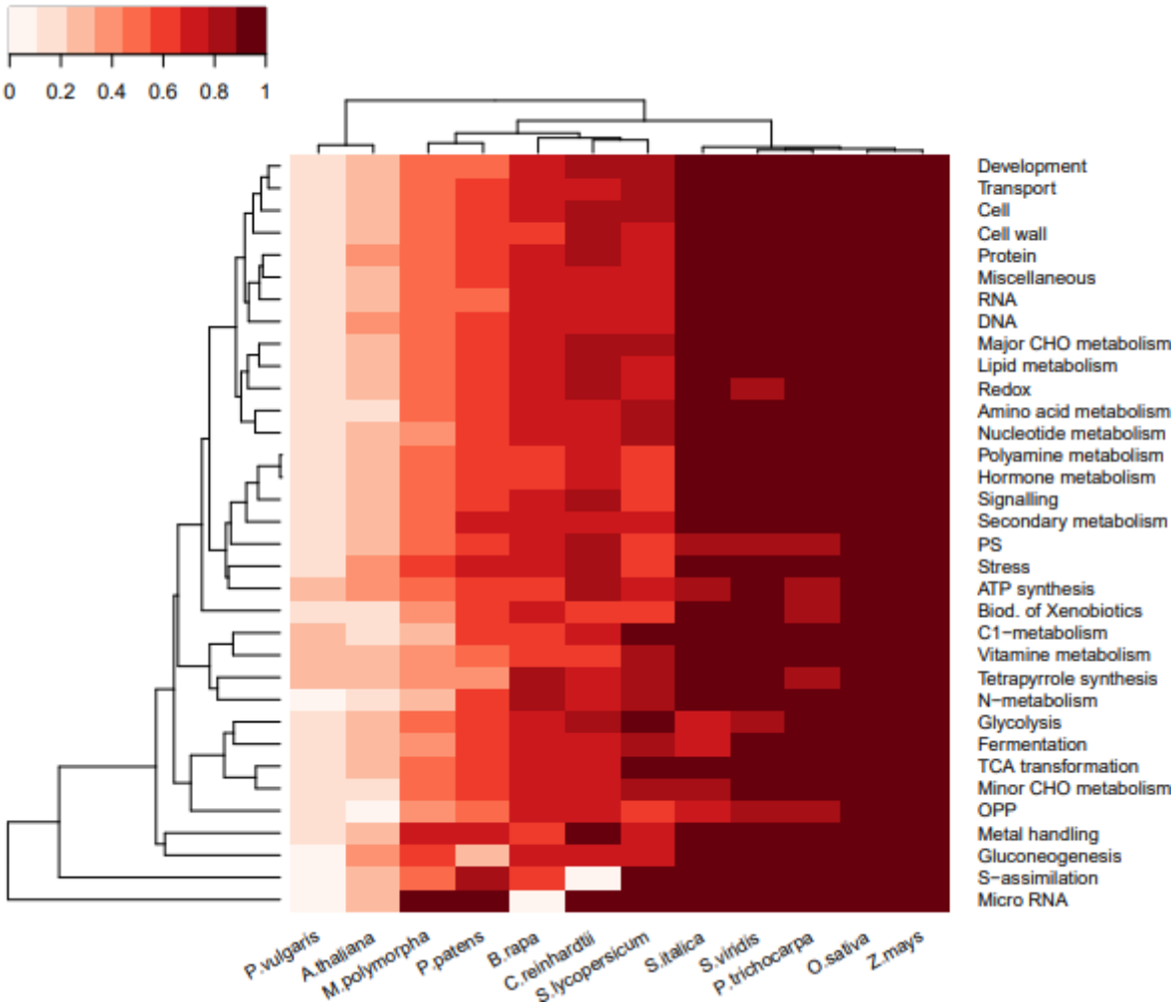
A ferramenta MapMan possibilitou fazer uma avaliação em larga-escala dos padrões de inserção de TEs de acordo com as diferentes categorias funcionais gênicas. A primeira etapa desta análise foi categorizar os genes anotados

estruturalmente disponíveis no Phytozome utilizando a ferramenta Mercator, e, em seguida, avaliamos o número de inserções de TEs em cada gene categorizado (Tabela 3). A Figura 7 permite uma ampla visualização da distribuição funcional dos TEs em genes funcionalmente categorizados e traz, ainda, uma análise de distribuição em relação às espécies. O grupo das monocotiledôneas (*Z. mays, O. sativa, S. viridis e S. italica*) apresentou a maior frequência de TEs inseridos em genes funcionalmente categorizados (Figura 7). Em comparação, entre as dicotiledôneas, *P. trichorcarpa* foi a única espécie com maior destaque em inserções de TEs, estando junto das monocotiledôneas, com uma distribuição semelhante para todos os BINs/SUB-BINs (Figura 7). Nessa análise mais ampla e qualitativa não foi possível identificar categorias funcionais com uma representação diferencial de TEs inseridos nas regiões gênicas.

Tabela 3. Número de genes codificantes e com inserções de TEs. BIN 26: Miscellaneous e BIN 35: not assigned.

Species	Total Number of Coding Genes	Number of coding genes functionally annotated with TE insertions (including BINs 26 and 35)	Number of coding genes functionally annotated with TE insertions (excluding BINs 26 and 35
A. thaliana	27,416	8,945	5,188
B. rapa	40,492	28,310	17,462
C. reinhardtii	17,741	13,966	4,736
M. polymorpha	19,287	9,220	3,923
O. sativa	42,189	42,108	16,678
P. vulgaris	27,433	5,607	3,267
P. patens	32,926	20,243	7,846
P. trichocarpa	34,699	31,129	18,602
S. italica	34,584	27,591	16,210
S. viridis	38,334	30,314	15,709
S. lycopersicum	35,768	19,605	12,564
Z. mays	39,498	32,463	17,603

Figura 7 — *Heatmap* indicando a riqueza de TEs (número de genes com TEs em cada BINs dividido pela quantidade total de genes na categoria), destacando as categorias com maior número de inserções. Os valores foram clusterizados por espécies e categoria funcional. Os dados brutos de anotação do Mercator e o número de inserções por *locus* podem ser acessados nos Anexos B e C. Nesta análise os BINs 26 (Miscellaneous) e 35 (not assigned) não foram incluídos.



Fonte: O autor (2023).

A ferramenta PageMan permitiu filtrar, por meio de análise estatística, quais categorias funcionais são diferentemente representadas por genes abrigando inserções de TEs. Já que não houve uma diferença importante na frequência e comprimento dos TEs nos três contextos gênicos (Anexo A), e que, inserções de TEs nos três contextos podem impactar nos padrões de silenciamento gênico epigenético da mesma forma, a análise funcional foi realizada de forma integrada. Assim, a Tabela 4 faz um recorte destacando as categorias com valores significativos de *z-score*, destacando, em verde, as categorias funcionais em que foi observado

enriquecimento e, em vermelho, as categorias empobrecidas. As espécies mostraram diferenças na quantidade e tipo de categorias diferencialmente representadas. Para as espécies mais basais, a briófita *M. polymorpha* é a única que se destaca com categorias diferencialmente representadas, com enriquecimento das categorias *Cell wall.Modification* (BIN 10.7), *Stress* (BIN 20), *Stress abiotic* (BIN 20.2), *Miscellaneous* (BIN 26) e *Signalling.Receptor kinases* (BIN 30.2).

Já para as dicotiledôneas, todas as espécies apresentaram ao menos uma categoria diferencialmente representada, com exceção de *B. rapa*. *Solanum lycopersicum* se destacou, com 29 categorias enriquecidas e apenas uma empobrecida (Tabela 4). Um estudo identificou 6.906 polimorfismos de inserções de TE (TIPs) em *S. lycopersicum*, sendo a maioria localizada dentro ou na proximidade de genes envolvidos a respostas do ambiente (Domínguez et al., 2020). O alto número de categorias enriquecidas encontrado na espécie pode estar relacionado com a presença dos TIPs em rotas relacionadas a diversas respostas ambientais. Isso porque, os genes que abrigam TIPs estão super representados em funções relacionadas à resposta a patógenos ou outros estresses ambientais (Domínguez et al., 2020). Apesar das categorias funcionais de *Stress* (BIN 20), *Stress biotic* (BIN 20.1) e *Stress abiotic* (BIN 20.2) não apresentarem valores estatisticamente significativos, outras categorias relacionadas apresentam um enriquecimento, como é o caso das categorias funcionais de Secondary metabolism (BIN 16) e Secondary metabolism Isoprenoids (BIN 16.1). Além disso, outras categorias funcionais enriquecidas podem estar indiretamente envolvidas, como *Amino acid metabolism* (BIN 13), *Nucleotide metabolism* (BIN 23), RNA (BIN 27), *RNA Processing* (BIN 27.1), *RNA Regulation of transcription* (BIN 27.2), DNA (BIN 28), *DNA Repair* (BIN 28.2) e *Protein* (BIN 29), tendo um papel fundamental na regulação e produção de proteínas associadas às respostas a patógenos e estresses ambientais.

Tabela 4 — *Z-scores* obtidos na análise de representação das inserções de TEs em genes de acordo com as categorias funcionais do MapMan para as 12 espécies estudadas. Vermelho e verde claro indicam categorias funcionais empobrecidas e enriquecidas com inserções de TEs, respectivamente (z-score ≤ -1,96 ou ≥ 1,96, respectivamente). São mostrados apenas com BINs e SUB-BINs que se mostraram diferencialmente representados para ao menos uma espécie. Os BINs 26 (Miscellaneous) e 35 (Not assigned) não foram incluídos nesta análise.

BIN	BINs/subBINs	*A. thaliana*	*B. rapa*	*P. trichocarpa*	*P. vulgaris*	*S. lycopersicum*	*S. viridis*	*S. italica*	*Z. mays*	*O. sativa*	*P. patens*	*M. polymorpha*	*C. reinhardtii*
2	Major CHO metabolism	-	-	-	-	2,2	-	-	-	-	-	-	-
3	Minor CHO metabolism	-	-	-	-	5,8	-	-	-6,4	-	-	-	-
3.3	Minor CHO metabolism.Sugar alcohols	-	-	-	-	-	-	-	-6,3	-	-	-	-
3.99	Minor CHO metabolism.Misc	-2,2	-	-	-	-	-	-	-9,4	-	-	-	-
4	Glycolysis	-	-	-	-	5,6	-	-	-	-	-	-	-
4.1	Glycolysis.Cytosolic branch	-	-	-	-	4,4	-	-	-	-	-	-	-
8	TCA / org transformation	-	-	-	-	4,3	-	-	-	-	-	-	-
9	ATP synthesis	-	-	-	-	3,7	-	-	-	-	-	-	-
10	Cell wall	-	-	-	-	6,0	-	-	-	-	-	-	-
10.6	Cell wall.Degradation	-	-	-	-	3,5	-	-	-	-	-	-	-
10.7	Cell wall.Modification	-	-	-	-	-	-	-	-	-	-	3,5	-
11	Lipid metabolism	-	-	-	-	5,3	-	-	-	-	-	-	-
11.9	Lipid metabolism.Lipid degradation	-	-	-	-	2,7	-	-	-	-	-	-	-
13	Amino acid metabolism	-	-	-	-	8,1	-	-	-2,9	-	-	-	-
13.1	Amino acid metabolism.Synthesis	-	-	-	-	-	-	-	-2,9	-	-	-	-
13.2	Amino acid metabolism.Degradation	-	-	-	-	3,9	-	-	-	-	-	-	-
16	Secondary metabolism	-	-	-	-	3,2	3,7	-	-3,7	-	-	-	-
16.1	Secondary metabolism.Isoprenoids	-	-	-	-	2,6	-	-	-	-	-	-	-
16.2	Secondary metabolism.Phenylpropanoids	-	-	-	-	-	-	-	-2,0	-	-	-	-
16.20	Secondary metabolism.Amino acid derivatives	-	-	-	-	-	-	-	-2,2	-	-	-	-
18	Vitamine metabolism	-	-	-	-	2,4	-	-	-	-	-	-	-
20	stress	4.1	-	-	-	-	-	-	-	-	-	7,3	-
20.1	Stress.Biotic	7,8	-	-	-	-	-	-	-	-	-	-	-
20.2	Stress.Abiotic	-	-	-	-	-	-	-	-	-	-	7,3	-
23	Nucleotide metabolism	-	-	-	-	5,0	-	-	-	-	-	-	-
23.1	Nucleotide metabolism.Synthesis	-	-	-	-	-	-	-	-2,4	-	-	-	-
26	Miscellaneous	-	-	-	-	6,9	4,2	-	-	-	-	3,1	-
26.1	Misc.Cytochrome P450	-	-	-	-	-	3,0	-	-	-	-	-	-

27	RNA	-	-	-	-	16,3	-	-	-	-	-	-	-
27.1	RNA.Processing	-	-	-	-	7,7	-	-	-	-	-	-	-
27.3	RNA.Regulation of transcription	-	-	-	-	12,5	-	-	-	-	-	-	-
28	DNA	-	-	-	-	5,8	-	-	2,0	-	-	-	-
28.2	DNA.Repair	-	-	-	-	3,9	-	-	-	-	-	-	-
29	Protein	4,4	-	-	-	16,1	-	-	5,6	-	-	-	-
29.2	Protein.Synthesis	-	-	-	3,5	-	-	-	-	-	-	-	-
29.4	Protein.Postranslational modification	-	-	-	-	-	-	-	2,4	-	-	-	-
29.5	Protein.Degradation	7,8	-	-	-	-	-	-	-	-	-	-	-
30.2	Signalling.Receptor kinases	-	-	2,1	-	-9,7	-	-	-	-	-	3,6	-
31	Cell	-	-	-	-	12,4	-	-	2,6	-	-	-	-
31.1	Cell.Organisation	-	-	-	-	10,4	-	-	-	-	-	-	-
31.2	Cell.Division	-	-	-	-	2,6	-	-	-	-	-	-	-
31.3	Cell.Cycle	-	-	-	-	2,3	-	-	-	-	-	-	-
33	Development	-	-	-	-	8,7	-	-	-	-	-	-	-
34	Transport	-	-	-	-	13,3	-	-	-	-	-	-	-

Fonte: O autor (2023)

Entre as monocotiledôneas, *Z. mays* e *S. viridis* são as únicas espécies que apresentaram representação significativa, sendo *Z. mays* a espécie com mais categorias empobrecidas entre todas as espécies analisadas. É comum que TEs sejam encontrados em regiões pobres em gene (heterocromatina), já que dessa forma é menos provável que causem mutações com efeitos deletérios no hospedeiro (Dimitri; Junakovic, 1999). Em plantas, como *A. thaliana* e *Z. mays*, já foi relatado que heterocromatina é rica na presença de DNA satélite e TEs (Kidwell, 2002), sendo que em milho os retrotransposons com LTRs são encontrados em abundância em certas regiões da heterocromatina (Ananiev et al.; 1998). Portanto, a quantidade de categorias empobrecidas observadas para *Z. mays* em contraste com o tamanho de seu genoma (Tabela 2) evidencia que o processo de expansão do genoma pelo aumento da quantidade de sequências repetitivas não envolveu um aumento da frequência de inserções em genes ou na sua vizinhança, pelo contrário, existe até um empobrecimento da quantidade de inserções em diversas categorias funcionais, especialmente nas categorias "*Minor CHO metabolism*" (BIN 3), "*Minor CHO metabolism.Sugar alcohols*" (BIN 3.3), "*Minor CHO metabolism.Misc*" (BIN 3.99).

Entre os resultados, vale a pena destacar o enriquecimento das categorias de Stress (BIN 20), Stress biotic (BIN 20.1) e Stress abiotic (BIN 20.2) nas espécies *A. thaliana e M. polymorpha*. No geral, para estresse, observamos genes que expressam proteínas em resposta a estímulos do meio, como calor, estímulo ao ácido abcísico, alta intensidade de luz e presença de peróxido de hidrogênio (Anexo B). Um exemplo observado é o gene 'at4g34190 - *stress enhanced protein* 1 (SEP1)´ de *A. thaliana*, o qual codifica uma proteína ativada em situações de estresse que se localiza na membrana do tilacóide e cujo mRNA tem sua regulação aumentada em resposta à alta intensidade luminosa (Anexo B). Já para estresse biótico identificamos genes que, em sua maioria, codificam proteínas relacionadas a defesa contra patógenos (fungos e bactérias), como PR (pathogenesis-related) e proteínas da superfamília thaumatina (Anexo B). O gene 'at4g26090 - *resistant to pseudomonas syringae* 2 (RPS2)´ codifica uma proteína de plasma de membrana com repetições ricas em leucina e zíper de leucina que conferem resistência a infecção por *Pseudomonas syringae*, por meio da interação com o gene de virulência avrRpt2. Ainda, para *M. polymorpha* foram identificados nove genes com 22 TEs inseridos relacionados ao estresse abiótico, envolvidos com stress oxidativo, reparação UV e fosforilação de aminoácidos (Anexo

C). O gene ‘mapoly0027s0062 - Concanavalin A-like lectin protein ´, por exemplo, codifica uma enzima envolvida na degradação inicial dos conjugados de glutationa, a qual possui papel na biotransformação, na eliminação de xenobióticos e na defesa das células contra o estresse oxidativo (Anexo B).

A fixação de inserções de TEs nos genes dessas rotas pode estar ligada ao fato de terem seu silenciamento modulado por mudanças ambientais, sendo possivelmente ativados por estresse e, com isso, podendo gerar adaptabilidade a condições adversas (Casacuberta; González, 2013). Já foi relatada a regulação diferencial de genes contendo TEs em sua proximidade genômica, a partir de diferentes condições de estresse. Como as superfamílias *Copia e Gypsy*, em *A. thaliana*, pelo estresse de calor, hAT por infecção e, para a espécie *S. lycopersicum*, as superfamílias *Harbinger e LINE* por estresse luminoso (Deneweth et al., 2022). Ainda em *A. thaliana*, foi demonstrado que o retrotransposon com LTR da família *ONSEN* é responsivo ao estresse de térmico e, devido a fatores de resposta ao calor que reconhecem a sequência regulatória no promotor do elemento, ocorre um aumento da regulação transcricional dos genes vizinhos sob esse tipo de estresse (Cavrak, 2014).

Além disso, representação diferencial de rotas relacionadas ao metabolismo secundário de *S. viridis* (BIN 16) e *S. lycopersicum* (BIN 16 - 16.1) também trazem informações interessantes para o estudo. A produção de metabólitos secundários está diretamente ligada às condições do meio, agindo como resposta ao estresse biótico e abiótico, tendo um papel essencial na adaptação e sobrevivência aos mais diversos ambientes (Ramakrishna; Ravishankar, 2011). Um estudo demonstrou que o remodelamento da cromatina em *A. thaliana* leva o retrotransposon LINE EPCOT3 a funcionar como um intensificador. Dessa forma, esse elemento passa a intermediar a ligação ao fator de transcrição WRKY33, ligado a biossíntese dos metabólitos camalexina e 4-hydroxyindole-3-carbonylnitrile; media a transcrição do gene CYP82C2, responsivo a patógenos e, ainda, está relacionado com defesa antibacteriana (Barco et al., 2019). Portanto, o enriquecimento observado nessas categorias, para as duas espécies, pode estar relacionado a respostas adaptativas a condições de estresse biótico e abiótico.

Sendo assim, nosso estudo traz resultados interessantes acerca do tema, permitindo avaliar as inserções de TEs nos genomas explorados e a distribuição de

categorias enriquecidas/empobrecidas, possibilitando a discussão em relação à capacidade dos TEs em agir como elementos controladores da expressão gênica. Trazendo, ainda, um destaque para as categorias de estresse e metabolismo secundário, sendo rotas que podem estar relacionadas a respostas adaptativas e, consequentemente, a evolução dos organismos.

5 CONCLUSÕES

Este trabalho estudou os padrões de inserção de TEs em regiões gênicas de 12 espécies vegetais, englobando representantes do grupo das clorófitas, briófitas, monocotiledôneas e dicotiledôneas, permitindo avaliar a diversidade genética derivada de TEs e seu possível impacto no genoma funcional. Para isso foi realizada a exploração do mobiloma entre as espécies avaliadas, destrinchando o número e comprimento dos TEs em três contextos gênicos (região 5', corpo do gene e região 3'). A análise comparativa do mobiloma no contexto gênico nos permitiu identificar vieses de inserção entre as espécies estudadas e as superfamílias de elementos, demonstrando que essas sequências não evoluem da mesma forma em diferentes genomas.

Ainda, foi avaliada relação entre distribuição dos elementos pelas categorias funcionais, revelando rotas metabólicas e atividades funcionais com diferentes padrões de inserções de TEs quando comparado com o conjunto gênico total. Esses dados também evidenciaram viés entre espécies. Pensando nos TEs como agentes que geram variabilidade genética, nossos resultados reafirmam hipóteses anteriores, mas também, trazem novas informações a respeito do mobiloma e inserções de TEs que contribuem para o entendimento do papel dos TEs na evolução das espécies vegetais.

REFERÊNCIAS

ANANIEV, V. E.; PHILLIPS, L. R.; RINES, W. H. A knob-associated tandem repeat in maize capable of forming fold-back DNA segments: Are chromosome knobs megatransposons? **Proceedings of the National Academy of Sciences**, v. 95, p. 10785–10790, 1998.

ANDERSON, N. S. et al. Transposable elements contribute to dynamic genomecontent in maize. **The plant journal**, v. 100, p. 1052–1065, 2019.

BARCO, B.; KIM, Y.; CLAY, K. N. Expansion of a core regulon by transposable elements promotes Arabidopsis chemical diversity and pathogen defense. **Nature Communications**, V. 10, 2019.

BOWMAN, L. J. et al. Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. **Cell**, v.171, p. 287 - 304, 2017.

BENNETZEN, L. J. et al. Reference genome sequence of the model plant Setaria. **Nature Biotechnology**, v. 30, p. 555-561, 2012.

BENNETZEN, L. J. et al. Transposable elements, gene creation and genome rearrangement in flowering plants. **Elsevier**, v. 15, p. 621-627, 2005.

BENJAMINI, Y.; YEKUTILIE, D. The control of the false discovery rate in multiple testing under dependency. **Annals of Statistics**, v.29, p. 1165-1188, 2001.

BIÉMONT, C; VIEIRA, C. Junk DNA as an evolutionary force. **Nature**, v. 443, p. 521-524, 2006.

BOURQUE, G. et al. Ten things you should know about transposable elements. **Genome Biology**, v. 19, n. 199, 2018.

CARARETO, C. M. A. et al. Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling Drosophila species. **Elsevier**, v. 537, n. 1, p. 93-99, 2017.

CASACUBERTA, E.; GONZÁLEZ, J. The impact of transposable elements in environmental adaptation. **Molecular Ecology**, v.22, p. 1503-1517, 2013.

CATLIN, S, N.; JOSEPHS, B, E. The important contribution of transposable elements to phenotypic variation and evolution. **Elsevier**, v. 65, 2022.

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

CAVRAK, V. V. How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation. **PLoS Genetics**, v. 10, 2014.

DENEWETH J.; PEER, V. Y.; VERMEIRSSEN, V. Nearby transposable elements impact plant stress gene regulatory networks: a meta-analysis in A. thaliana and S. lycopersicum. **BMC Genomics**, v. 23, 2022.

DIMITRI, P.; JUNAKOVIC, N. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. **Trends in genetics**, v . 15, p. 123-124, 1999.

DOMÍNGUEZ, M. et al. The impact of transposable elements on tomato diversity. **Nature Communications**, v. 11, n. 4058, 2020.

ELLINGHAUS, D.; KURTZ, D.; WILLHOEFT, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. **BMC Bioinformatics**, v. 9**,** 2008.

FEDOROFF, N. et al. Transposable Elements, Epigenetics, and Genome Evolution. **Science**, v. 338, n. 6108, p. 758-767, 2012.

FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. **Nature Reviews Genetics**, v. 9, p. 397-405, 2008.

GAO, L.; et al. Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. **BMC Genomics**, v.5, 2004.

GIDEON, G. et al. The Stem Cell State in Plant Development and in Response to Stress. **Frontiers in Plant Science**, v. 2, 2011.

GRZEBELUS, D. The Functional Impact of Transposable Elements on the Diversity of Plant Genomes. **Diversity**, v. 10, n. 2, 2018.

HUBLEY, R.; SMIT, A. RepeatModeler, 2015. Disponível em: https://www.repeatmasker.org/RepeatModeler/

HU, K. et al. Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. **BMC Bioinformatics**, v. 20, 2019.

ITO, H; KAKUTANI, T. Controlo f transposable elements in Arabidopsis thaliana. **Chromossome Research**, v. 22, p. 217 -223, 2014.

JESUS, E. et al. Genoma móvel: Mecanismos de transposição e impacto evolutivo. In: MENCK, C. F. M.; SLUYS, M. (Org.). Genética molecular básica: dos genes aos genomas. Rio de Janeiro: Guanabara Koogan, 2017.

KAPITONOV, V. V.; JURKA, J. Self-synthesizing DNA transposons in eukaryotes. **Biological sciences**, v. 103, p. 4540-4545, 2006.

KIM, N. The genomes and transposable elements in plants: are they friends or foes? **Genes & Genomics,** v. 39, p. 359–370, 2017.

KIDWELL, G. M. Transposable elements and the evolution of genome size in eukaryotes. **Genetica**, v. 115, p. 49-63, 2002.

KHAN, A. et al. Activation of Tag1 transposable elements in Arabidopsis dedifferentiating cells and their regulation by CHROMOMETHYLASE 3-mediated CHG methylation. **Biochimica et Biophysica Acta (BBA)**, v. 1859, p. 1289-1298, 2016.

LE, HIEN, Q. et al. Transposon diversity in Arabidopsis thaliana. **Proceedings of the National Academy of Sciences**, v. 97, p. 7376 -7381, 2000.

LI, XIA. et al. A Large Insertion in bHLH Transcription Factor BrTT8 Resulting in Yellow Seed Coat in Brassica rapa. **Plos One**, 2012.

LOHSE, M. et al. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. **Plant, Cell & Environment**, v. 35, n. 5, p. 1250-1258, 2014.

MARTIN, A. et al. A transposon-induced epigenetic change leads to sex determination in melon. **Nature.** v. 461, p. 1135-1138, 2009.

MATZKE, M., MOSHER, R. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. **Nature Reviews Genetics**, v.15, p.394–408, 2014.

MIGICOVSKY, Z; KOVALCHUK, I. Transgenerational changes in plant physiology and in transposon expression in response to UV-C stress in Arabidopsis thaliana. **Plant Signal Behav**, v. 9, 2014.

MORGANTE, M. et al. Transposable elements and the plant pan-genomes. **Elsevier**, Itália, v. 10, n. 2, p. 149–155, 2007.

MUYLE, M. A.; et al. Gene-body methylation in plants: mechanisms, functions and important implications for understanding evolutionary processes. **Genome Biology and Evolution**, 2022.

NOLAN, T. et al. The post-transcriptional gene silencing machinery functions independently of DNA methylation to repress a LINE1-like retrotransposon in *Neurospora crassa*. **Nucleic Acids Research**, v. 33, p. 1564–1573, 2005.

OU, S.; JIANG, N. A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons, **Plant Physiology**, v. 176, p. 1410–1422, 2018.

OU, S., Su, W., LIAO, Y. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. **Genome Biology 20,** 275, 2019.

PANG, E. et al. Crop Genome Annotation: A Case Study for the Brassica rapa Genome. **The Brassica rapa Genome**, p.53–64, 2015.

PRITHAM, J. E.; PUTLIWALA, T.; FESCHOTTE, C. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viroses. **Elsevier**, v. 390, p. 3- 17, 2007.

QUADRANA, L. et al. Natural occurring epialleles determine vitamin E accumulation in tomato fruits. **Nature Communications**, v. 5, n. 4027, 2014.

QUESNEVILLE, H. Twenty years of transposable element analysis in the Arabidopsis thaliana genome*.* **Mobile DNA**, 11, 28, 2020.

QUINLAN, AARON R.; HALL, IRA M. BEDTools: a flexible suite of utilities for comparing genomic features. ***In*** **Bioinformatics***, 26 (6), p.*841–842*, 2010.*
RAMAKRISHNA, A.; RAVISHANKAR, A. G. Influence of abiotic stress signals on secondary metabolites in plants. **Plant signalling & behavior**, v. 6, p. 1720–1731, 2011.

ROFFLER, S.; WICKER, T. Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. **Mobile DNA**, v.6, 2015.

RUBIN, E. et al. Structure and evolution of the hAT transposon superfamily. **Genetics**, v. 158, n. 3, p. 949-957, 2001.

SANSEVERINO, W. et al. Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome. **Molecular Biology and Evolution**, v. 32, p. 2760–2774, 2015.

SCHNABLE, P. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. **Science**, v. 326, n. 5956, p. 1112-1115, 2009.

SHI, J.; LIANG, C. Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection. **Plant Physiology**, v. 180, p. 1803-1815, 2019.

SUGUIYAMA, V. et al. The population genetic structure approach adds new insights into the evolution of plant LTR retrotransposon lineages. **PLoS ONE**, v. 14, n. 5, 2019.

SU, X. et al. A high-continuity and annotated tomato reference genome. **BMC genomics**, v. 22, 2021.

TENAILLON, I. M.; et at. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and *Zea luxurians.* **Genome Biology and Evolution**, v.3, p. 219-29, 2011.

THIMM, O. et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. **The Plant Journal**, v. 37, p. 914-939, 2004.

THOMAS, H.; PRITHAM, J. E. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. **Microbiology Spectrum, v**. 3, 2015.

VENDRELL-MIR, P. et al. Different Families of Retrotransposons and DNA Transposons Are Actively Transcribed and May Have Transposed Recently in Physcomitrium (Physcomitrella) patens. **Frontiers in Plant Science** , v. 11, 2020.

WELLS, J. N.; FESCHOTTE, C. A Field Guide to Eukaryotic Transposable Elements. **Annual Review of Genetics**, v. 54, p. 539-561, 2020.

WEIJIA, S.; GU, X.; PETERSON, T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. **Molecular Plant,** v.12, p.447-460, 2019.

WESSLER, R. S. Plant retrotransposons: Turned on by stress. **Current Biology**, v.6. p. 959–961, 1996.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Perspectives**, v. 8, p. 973-982, 2007.

XIONG, W. et al. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. **Biological Sciences**, v. 111, 2014.

XU, Z.; WANG, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. **Nucleic Acids Research**, v. 35, p. 265–268, 2007.

USADEL, B. et al. A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. **Plant, Cell & Environment**, v. 31, n. 9, p. 1211-1229, 2009.

USADEL, B. et al. Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. **Plant Physiology**, v. 138, p. 1195-204, 2005.

YADAV, H. et al. Genome-wide development of transposable elements-based markers in foxtail millet and construction of an integrated database. **DNA Research**, v. 22, p. 79-90, 2015.

ZHAO, Y. et al. Transposable Elements: Distribution, Polymorphism, and Climate Adaptation in Populus. **Frontiers in Plant Science**, v. 13, 2022.

ANEXO A — Documento de planilha feito na ferramenta Excel referente ao número e comprimento dos TEs agrupados por superfamília nas 12 espécies analisadas para os três contextos gênicos.

ANEXO B — Documento de planilha feito na ferramenta Excel referente a anotação do programa Mercator anotando os *locus* dos genomas das 12 espécies para um ou mais BINs.

ANEXO C — Documento de planilha feito na ferramenta Excel referente ao número de elementos de transposição por *locus* em cada espécie analisada.