EDGAR ANDRÉS OCHOA CRUZ

**Caracterização de Transposases
da Família SChAT em cana-de-açúcar:
estudo molecular e funcional**

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, para obtenção do Título de Doutor em Biotecnologia.

São Paulo
2012

EDGAR ANDRÉS OCHOA CRUZ

**Caracterização de Transposases
da Família SChAT em cana-de-açúcar:
estudo molecular e funcional**

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, para obtenção do Título de Doutor em Biotecnologia.
Área de concentração: Biotecnologia.
Orientadora: Profa. Dra. Marie-Anne Van Sluys
Versão original

São Paulo
2012

Ficha catalográfica

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

© reprodução total

Cruz, Edgar Andrés Ochoa.

Caracterização de transposases da família SChAT em cana-de-açúcar: estudo molecular e funcional / Edgar Andrés Ochoa Cruz. -- São Paulo, 2012.

Orientador: Profª Dra. Marie Anne Van Sluys.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Biotecnologia USP/IPT/Instituto Butantan. Área de concentração: Biotecnologia. Linha de pesquisa: Elementos de transposição.

Versão do título para o inglês: Characterization of SChAT family transposases in sugarcane: molecular and functional studies.

Descritores: 1. Cana-de-açúcar 2. Elementos de transposição 3. Domesticação 4. Evolução do genoma I. Sluys, Profª Dra. Marie Anne Van II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Biotecnologia USP/IPT/Instituto Butantan III. Título.

ICB/SBIB030/2012

UNIVERSIDADE DE SÃO PAULO
Programa de Pós-Graduação Interunidades em Biotecnologia
Universidade de São Paulo, Instituto Butantan, Instituto de Pesquisas Tecnológicas

Candidato(a): Edgar Andrés Ochoa Cruz.

Título da Tese: Caracterização de transposases da família SChAT em cana-de-açúcar: estudo molecular e funcional.

Orientador(a): Profª Dra. Marie Anne Van Sluys.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a/...../....., considerou

() Aprovado(a)

() Reprovado(a)

Examinador(a): Assinatura:
Nome:
Instituição:

Examinador(a): Assinatura:
Nome:
Instituição:

Examinador(a): Assinatura:
Nome:
Instituição:

Examinador(a): Assinatura:
Nome:
Instituição:

Presidente: Assinatura:
Nome:
Instituição:

Dedico este trabajo a mis padres Esperanza y Edgar que siempre creyeron en mi, creyeron tanto en mi que yo también termine creyendo.
A mis hermanas Ana y Ximena que son mi orgullo y a mi abuelo Pedro
por ser mi abuelo

AGRADECIMENTOS

Para mis padres, hermanas y abuelo por ser la razón por la que me despierto todos los días.

À Profa. Dra. Marie-Anne Van Sluys, pelo apoio, paciência e confiança. Por pular de alegria quando o resultado dava certo e pela ajuda quando o resultado não dava tão certo.

A todas as pessoas que passaram pelo GaTE-Lab ao longo desses quatro anos porque foram fundamentais para o desenvolvimento deste trabalho e fizeram do laboratório um excelente lugar para estar: Alessandro, Bruno, Cushla, Danielle Quintanilha, Douglas, Érika, Fabiana, Guilherme (companheiro espartano de experimentos e cervejas), Hana, Jonas, Juliana Nico, Kleber, Luisa, Marcelo, Marisa, Mayra, Nathália e Robson.

Quero agradecer especialmente a Leonor, Silvia e Tatiana por sempre prestar um serviço além de suas responsabilidades no GaTE –Lab, fazendo possíveis tarefas chaves para o desenvolvimento das pesquisas do grupo.

À Profa. Dra. Kátia Scortecci por ceder os tecidos usados nos ensaios de expressão e ao Prof. Dr. Reinhard Kunze pelo apoio e orientação durante minha estadia na Alemanha.

Ao Prof. Dr. Gabriel Padilla por ter sido um grande apoio na minha chegada ao Brasil, sempre tendo uma palavra amigável e inteligente para ser dita.

A Lara Zimmermann pelo apoio e carinho ao longo da confecção deste documento, trabalho chave para evitar que o autor perdesse a cabeça no caminho.

Agradeço à CAPES e CNPq pelo apoio financeiro e concessão de bolsa que permitiram a realização deste trabalho.

Finalmente, agradeço a todas as pessoas mencionadas aqui pela amizade e bons momentos que fizeram desta experiência algo que realmente valeu a pena.

**“This is ten percent luck, twenty percent skill
Fifteen percent concentrated power of will
Five percent pleasure, fifty percent pain”**

Fort Minor

**“El colegio poco me enseñó. Si es por esos libros nunca aprendo a:
coger el cielo con las manos
a reír y a llorar lo que te canto
a Coser mi alma rota
a perder el miedo a quedar como un idiota
y a empezar la casa por el tejado”**

Fito e Fitipaldis

RESUMO

Ochoa Cruz EA. Caracterização de Transposases da Família SChAT em cana-de-açúcar: estudo molecular e funcional. [Tese (Doutorado em Biotecnologia)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2012.

Os elementos de transposição (TEs) se movimentam de um *locus* para outro no genoma afetando a estrutura e evolução destes. Podem ocasionar recombinação, aumento do tamanho do genoma e sofrer eventos de domesticação. No processo de domesticação, os domínios protéicos de TEs são recrutados pelo genoma para a construção de novos genes. A superfamília de transposases hAT é definida pelos elementos que compartilham os domínios de dimerização e ligação ao DNA (tipo *Zinc Finger*) com os transposons previamente descritos: *hobo* (*Drosophilla melanogaster*), *Activator* (*Zea mays*) e *Tam3* (*Antirrhinum majus*). Análises prévias encontraram algumas evidências da presença genômica e ativação transcricional de TEs relacionados à superfamília hAT em cana-de-açúcar (denominados de família SChAT) e pelo menos três linhagens evolutivas foram postuladas. O objetivo deste trabalho é caracterizar versões genômicas das linhagens de transposons (191 e 257) e linhagem domesticada (074). Pretende-se estudar as relações evolutivas, distribuição em gramíneas, identificar os padrões de expressão e propriedades funcionais. Regiões genômicas clonadas em vetor BAC carregando elementos das três linhagens foram selecionados e sequenciados de uma biblioteca comercial da variedade R570. O número de cópias de cada uma destas linhagens foi estimado no genoma da cana-de-açúcar sendo elas linhagem 074 – 6 cópias; linhagem 191 – 29 cópias e linhagem 257 – 22 cópias. Dezesesseis clones foram sequenciados usando o *GS-FLX pyrosequencer* (Roche) e destes BACs foram extraídos sete versões completas que apóiam a definição de pelo menos três linhagens evolutivas; SChAT160F01 (SChAT1), SChAT167G07 (SChAT2) e SChAT185O22 (SChAT3) (linhagem 191); SChAT152D05 e SChAT170C07 (linhagem 257) e dois genes SCtpase92C13 e SCtpase124F07 (linhagem 074). Posteriormente foram selecionados mais oito clones BAC com 5 genes completos da linhagem 074 por varredura com iniciadores específicos na mesma biblioteca comercial. Regiões de sintenia foram estabelecidas para estes BACs em *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* e *Zea mays*. Elementos relacionados com as três linhagens foram procurados nestes genomas. A caracterização funcional (estrutura do gene, número de cópias, sintenia entre gramíneas e padrão de expressão) apóia a hipótese que a linhagem 074 seja um elemento domesticado. Também foi demonstrado que esta linhagem possui ao menos dois *loci* expressos no genoma da cana-de-açúcar, um deles foi estudado, sendo encontrado que este elemento foi fixado no genoma há pelo menos 12 MYA, antes de ocorrer a divergência entre *Zea mays* e *Sorghum/Saccharum*. Por outro lado, os elementos das linhagens 191 e 257 apresentaram um número de cópias, estrutura, localização genômica e padrão de expressão típico de transposons. A linhagem 257 apresentou elementos preditos como ativos, o que levou à expressão de uma de suas transposases neste trabalho. Finalmente, foi feita varredura procurando por domínios conservados das transposases hAT em 270 BACs da base de dados do BIOEN, sendo achados mais cinco elementos que não pertencem às três linhagens anteriormente descritas, ampliando assim a visão da história de diversificação da superfamília hAT no genoma da cana-de-açúcar. A recuperação das versões genômicas apresentadas neste trabalho permitiu explorar a estrutura de

elementos da família, o contexto genômico, sintenia das regiões com outras gramíneas e estabelecer relações filogenéticas com TEs relacionados à superfamília hAT presentes neste e em outros genomas. Os resultados aqui apresentados contribuem na compreensão do impacto das transposases da superfamília hAT na organização e evolução do genoma das gramíneas, em particular da família SChAT de cana-de-açúcar.

Palavras-chaves: Elementos de transposição. Cana-de-açúcar. Domesticação. Evolução do genoma.

ABSTRACT

Ochoa Cruz EA. Characterization of SChAT family transposases in sugarcane: molecular and functional studies. [Ph.D. thesis]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2012.

Transposable elements (TEs) are able to move from one locus to another within a genome. TE mobilization affects genome structure and evolution by phenomena as genome shuffling and host genome expansion. Domestication events of protein domains from TEs can appear, this allows the construction of new genes in the host genome. The hAT transposase superfamily is defined as elements that share the dimerization and zinc-finger DNA ligation domains with the previously described *hobo* (*Drosophilla melanogaster*), *Activator* (*Zea mays*) and *Tam3* (*Antirrhinum majus*) transposon elements. Previous analyses found some genomic and transcriptional evidences of TEs related to hAT superfamily in sugarcane (named SChAT family); furthermore, at least three evolutionary lineages were proposed in the family. The aim of this work is to characterize full-length genomic versions of the transposons lineages (191 and 257) and from the domesticated lineage (074). It is proposed to study the evolutionary relationship, distribution along grasses genomes, identify expression patterns and functional capacities of the SChAT elements. Genomic regions cloned into BAC vector carrying genomic versions were chosen and sequenced from a commercial BAC library. The commercial library was also used for calculating the relative number of copies of the three lineages (lineage 074 – 6 copies, lineage 191 – 29 copies and lineage 257 – 22 copies). Sixteen positive clones selected by hybridization with group-specific probes were sequenced using GS-FLX pyrosequencer (Roche). From these 16 BACs seven complete elements were recovered allowing to establish at least three evolutionary lineages; SCtpase92C13 and SCtpase124F07 (the 074 lineage); SChAT160F01 (SChAT1), SChAT167G07 (SChAT2) and SChAT185O22 (SChAT3) (the 191 lineage); SChAT152D05 and SChAT170C07 (the 257 lineage). Another five elements related to the lineage 074 were selected by scan with group-specific primers. Syntenic regions for the BACs containing elements from the three lineages were mapped in *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* and *Zea mays*. Related elements were search on the same genomes. The characterization of the elements showed that lineage 074 has the expected structure, number of copies, genomic distribution, expression pattern and distribution within the grasses expected for a domesticated element. It was also found that this lineage has two expressed loci in the sugarcane genome, one of them was studied showing that this element have being fixed in this region at least since 12MYA, before the division between *Zea mays* and *Sorghum/Sugarcane* occurred. In contrast, the 191 and 257 lineage elements had the expected structure, number of copies, genomic location, distribution through grasses and expression pattern of *strict sense* transposons. The 257 lineage contains elements predicted as active, one of these transposases was expressed in this work. Finally, five elements related to the hAT superfamily but which did not fall into the three previously characterized lineages were identified by using the superfamily domains as queries against 270 BACs in the BIOEN database. The identification of elements outside the three lineages gives a broader view of the diversification history of the SChAT family in sugarcane. The recovery of full-length genomic copies of SChAT elements allowed us to explore the structure of these

elements, their genomic context, syntenic regions of these elements in other grass genomes and to established the relationships of TEs related to the hAT superfamily in sugarcane and in other genomes. The results presented here contribute to our understanding of the impact of transposons on the evolution and organization of the grass genomes, focusing on the SChAT family from sugarcane

Keywords: Transposable elements. Sugarcane. Domestication. Genome evolution.

LISTA DE FIGURAS

Figura 1 - Estrutura de uma família de transposons.	17
Figura 2 - Mecanismos de transposição.	19
Figura 3 - Diferentes fenótipos em grãos de milho devido à ação de TEs.	21
Figura 4 - Sistema de classificação proposto para TEs	22
Figura 5 - Estrutura do elemento Ac.	30
Figura 6 - Diagrama esquematizando a organização do genoma de um cultivar atual de cana-de-açúcar.	32
Figura 7 - Resumo dos ciclos de PCR utilizados em diferentes procedimentos do trabalho.	43
Figura 8 - Hibridação de duas sondas do agrupamento II da família SChAT com membrana de DNA genômico digerido com a enzima DraI.	64
Figura 9 - Hibridação das sondas das três linhagens da família SChAT com três membranas da coleção de BACs e Número relativo de cópias dos elementos caracterizados.	66
Figura 10 - Padrão de agrupamento dos 17 elementos completos, elemento SChATG1 antes caracterizado e os cDNAs da família.	72
Figura 11 - Alinhamento das TIRs dos 10 transposons completos com a TIR da transposase do elemento Ac de milho e o elemento SChATG1.	74
Figura 12 - Cladograma das 17 transposases SChAT caracterizadas neste trabalho.	76
Figura 13 - Alinhamento das transposases da linhagem 191 e o elemento de referencia Ac	78
Figura 14 - Anotação do BAC SHCRBa_160_F01.	80
Figura 15 - Anotação do BAC SHCRBa_167_G07.	80
Figura 16 - Anotação e sintenia do BAC SHCRBa_185_O22.	81
Figura 17 - Alinhamento das transposases da linhagem 257 e o elemento de referencia Ac	83
Figura 18 - Anotação e sintenia do BAC SHCRBa_152_D05.	85
Figura 19 - Anotação e sintenia do BAC SHCRBa_170_C07.	85
Figura 20 - <i>Western Blot</i> usando o anticorpo anti-histidina em diferentes extratos protéicos de E coli. transformada com o plasmideo pET257	87
Figura 21 - <i>Western Blot</i> usando o anticorpo anti-histidina em diferentes extratos protéicos de S. cerevisiae transformada com os plasmídeos; GS, G1, NGS e NG1.	88
Figura 22 - Alinhamento das transposases domesticadas da linhagem 074, o elemento domesticado DAYSLEEPER e a transposase de Ac.	91
Figura 23 - Anotação e sintenia dos BACs de um dos locus da linhagem 074 em cana-de-açúcar.	95
Figura 24 - Análise de expressão do meristema apical de duas variedades de cana-de-açúcar.	98
Figura 25 - Procedimentos realizados para sequenciar transcritos da linhagem 074 do meristema apical de duas variedades de cana-de-açúcar, RB75-126 e SP81-3250.	98
Figura 26 - Alinhamento das ORFs dos elementos SCtpase92C13, SCtpase124F07, os TEs da linhagem 074 e os amplificadores dos cDNAs do meristema apical de duas variedades de cana-de-açúcar, RB75-126 e SP81-3250.	99

Figura 27 - Filogenia simplificada da família <i>Poaceae</i> .	102
Figura 28 - Alinhamento das transposases dos cinco ortólogos da família SChAT encontrados em <i>Sorghum bicolor</i> com a transposase do elemento SChAT160F01.	104
Figura 29- Cladograma das transposases SChAT, transposase do elemento Ac e transposases ortólogas à linhagem 191 e 257, encontrada em <i>Sorghum bicolor</i> .	105

LISTA DE QUADROS

Quadro 1 - Membros da Superfamília hAT.	28
Quadro 2 - Material vegetal utilizado para os diferentes experimentos deste trabalho.	37
Quadro 3 - Sequências dos iniciadores utilizados para amplificar as sondas.	55
Quadro 4 - BACs selecionados por hibridação e sequenciados.	67
Quadro 5 - BACs selecionados e sequenciados.	69
Quadro 6 - TIRs e TSDs identificados para os transposons completos usando o programa TIR_FINDER (Rubin et al., 2001).	73
Quadro 7 - Descrição dos elementos da linhagem 191.	77
Quadro 8 - Descrição dos elementos da linhagem 257.	82
Quadro 9 - Construções feitas para expressar a transposase codificada pelo TE257.	87
Quadro 10 - Descrição dos elementos da linhagem 074.	90
Quadro 11 - Ortólogos da linhagem 074.	93
Quadro 12 - Descrição dos outros elementos hAT.	100
Quadro 13 - Ortólogos da família SChAT em <i>Sorghum bicolor</i> .	102

ABREVIATURAS

BAC – do inglês *bacteria artificial chromossome*

Contigs - sequência nucleotídica consenso de um conjunto de leituras de nucleotídeos que se agrupam por sobreposição

DNA – ácido desoxirribonucleico

DNase - desoxirribonuclease

dNTP- desoxirribonucleotídeos trifosfato

EST – do inglês *expressed sequence tag*

ISs – Sequências de inserção

IPTG – isopropil tio- β -D-galactosídeo

Kb – kilo pares de base

μ L, μ g – microlitro, micrograma

MITE – do inglês *Miniature inverted-repeat transposable elements*

mL, mg – mililitro, miligrama

ng - nanograma

ORF – do inglês *open reading frame*

pb – pares de base

PCR – do inglês *polymerase chain reaction*

PVP – Polivinilpirrolidona

PVPP – Polivinilpolipirrolidona

RNA – ácido ribonucleico

RNase - ribonuclease

SDS – dodecil sulfato de sódio

TBE - Tris-borato-EDTA

TEs – do inglês *transposable elements*

TIRs - do inglês *terminal inverted repeats*

Tpases – Transposases.

TSB – do inglês *target site duplication*

IN – Integrase

RT – do inglês *Reverse Transcriptase*

SUMÁRIO

1 INTRODUÇÃO	15
1.1 Elementos de Transposição, uma História de Diversidade	15
1.2 Impacto dos Elementos de Transposição: Diversidade, Obesidade dos Genomas, Criação de Redes Regulatórias e Domesticação	23
1.3 Superfamília hAT	27
1.4 Família SChAT e a História do Estudo do Potencial Genômico da Cana de Açúcar	30
2 OBJETIVOS	36
2.1 Objetivo Geral	36
2.2 Objetivos Específicos	36
3 MATERIAIS E MÉTODOS	37
3.1 Material Vegetal	37
3.2 Cepa Bacteriana Utilizada (<i>Escherichia coli</i>)	38
3.3 Plasmídeos	38
3.3.1 Clonagem de produtos de PCR e RT-PCR	38
3.4 Antibióticos	38
3.5 Manipulações Gerais com Bactérias (<i>Escherichia coli</i>)	38
3.5.1 Meios de Cultura para Bactérias	38
3.5.2. Obtenção de Bactérias Químio Competentes	39
3.6 Clones da Biblioteca Comercial de BACs de Cana-de-açúcar	39
3.6.1 Avaliação do número de cópias relativo das linhagens 74,191 e 257	40
3.7 Eletroforese e Géis de Agarose	40
3.8 Procedimentos Gerais de Manipulação de DNA	41
3.8.1 Extração de DNA Genômico de Plantas	41
3.8.2 Quantificação de DNA	42
3.8.3 Amplificação de Fragmentos por PCR (<i>Polymerase Chain Reaction</i>)	42
3.8.4 Purificação de Fragmentos de DNA	43
3.8.5 Ligação de Fragmentos de DNA	43
3.8.6 Transformação de Bactérias Químio Competentes	44
3.8.7 Extração de DNA Plasmidial em Pequena Escala (Miniprep)	44
3.8.8 Extração de DNA Plasmidial em Pequena Escala (Miniprep) em Placas de 96 Poços	45
3.8.9 Extração de BACs dos Clones da Biblioteca	46
3.8.10 Sequenciamento Automático de Nucleotídeos	47
3.8.11 Sequenciamento em Alta Escala de BACs.	48
3.9 Procedimentos Gerais de Manipulação de RNA	49
3.9.1 Extração de RNA Total	50
3.9.2 Quantificação de RNA	50
3.9.3 Tratamento do RNA Total para a Remoção de DNA Contaminante	50
3.9.4 Síntese da Primeira Fita de cDNA	50
3.9.5 RT-PCR	50
3.10 Procedimentos Gerais de Manipulação de Proteínas	51

3.10.1 Plasmídeos Usados para a Expressão da Transposase	51
3.10.2 Expressão e extração da Transposase	52
3.10.2.1 Expressão e extração em <i>E. coli</i>	52
3.10.2.2 Expressão e extração em <i>Saccharomyces cerevisiae</i>	53
3.10.3 Eletroforese em Gel de Poliacrilamida	53
3.10.4 Western Blot	54
3.11 Hibridação Molecular	54
3.11.1 Sondas	54
3.11.2 Hidrólise do DNA Genômico	55
3.11.3 Gel de Agarose	55
3.11.4 Hibridação Molecular de DNA	55
3.11.4.1 Transferência Alcalina	55
3.11.4.2.1 Pré-Hibridação para Membranas de DNA Genômico	56
3.11.4.2.2 Pré-Hibridação para Membranas da Biblioteca Comercial de BACs de Cana-de-açúcar	56
3.11.4.3 Marcação das Sondas	56
3.11.4.4 Hibridação	57
3.11.4.5 Lavagens das Membranas	57
3.11.4.6 Exposição e Revelação	57
3.11.5 Hibridação Molecular de RNA	58
3.11.5.1 Gel de Formaldeído	58
3.11.5.2 Transferência e Hibridação de RNA	58
3.12 Caracterização dos Elementos usando Ferramentas Bioinformáticas	59
3.12.1 Identificação dos Elementos SChAT nos BACs Sequenciados	59
3.12.2 Análise Comparativa dos Elementos SChAT	60
3.12.3 Procura de Elementos Relacionados à família SChAT em Outros Genomas	61
3.12.4 Geração de Dendogramas a partir de Alinhamentos	61
3.13 Identificação de Regiões de sintonia em Genomas Relacionados	61
 4 RESULTADOS E DISCUSSÃO	 63
4.1 A família SChAT de cana de açúcar	64
4.1.1 Identificando as primeiras linhagens (TE074; TE191 e TE257)	64
4.1.2 Estimativa do Número de Cópias no Genoma da Cana-de-açúcar (Variedade R570) das linhagens 74, 191 e 257 da família SChAT	65
4.1.3 Diversidade e Relações Filogenéticas Observadas nas Linhagens	70
4.1.4 Estrutura Genômica dos Elementos SChAT	73
4.1.5 Transposases da Família SChAT	75
4.2 Linhagem 191	77
4.2.1 Análises das Regiões Sintênicas e Contexto Genômico	79
4.3 Linhagem 257	82
4.3.1 Análises das Regiões Sintênicas e Contexto Genômico	84
4.3.2 Expressão da Transposase	86
4.4 Linhagem 074, um Possível Gene Domesticado.	89
4.4.1 Linhagem Domesticada e sua Conservação ao Longo de Vários Genomas	92
4.4.2 Análises das Regiões Sintênicas e Contexto Genômico	94

4.4.3 Análises de Expressão da transposase domesticada	96
4.5 Outros Elementos hAT no Genoma da Cana-de-açúcar	100
4.6 Elementos Relacionados em Outros Genomas	101
5 CONCLUSÕES	106
REFERÊNCIAS	108
ANEXOS	
Anexo A. Alinhamento múltiplo de aminoácidos de todas as versões hAT identificadas no presente trabalho	
Anexo B. Diversification of hAT transposase paralogues in the sugarcane genome.	

1 INTRODUÇÃO

1.1 Elementos de Transposição, uma História de Diversidade

Apesar da grande variação de tamanho entre os genomas eucariotos, pouca diferença é observada no número de genes (Feschotte e Pritham, 2007; Gregory, 2005). A diferença de tamanho está relacionada com as regiões não-codificantes e com a presença dos elementos de transposição. Dentre as regiões não-codificantes do genoma encontramos regiões intergênicas, regiões regulatórias e regiões repetitivas como os microsátélites.

Os elementos de Transposição (TEs) são sequências genômicas repetitivas codificantes, com a capacidade inerente de se movimentar de um *locus* para outro no genoma. Eles também são conhecidos por estarem presentes na maioria dos organismos, sugerindo uma origem precoce na história evolutiva das espécies. Estes elementos têm um papel fundamental, como provedores de plasticidade genômica nos três domínios biológicos (Bennetzen, 2000; Feschotte et al., 2002; Haren et al., 1999.).

Os transposons [incluindo as sequências de inserção (ISs), como são conhecidos os elementos de transposição de bactérias] são encontrados nos Domínios *Bactéria*, *Archea* e *Eukarya*. Os retrotransposons, entretanto, são encontrados basicamente em *Eukarya*.

Apenas em bactérias já foram reportadas mais de 17 famílias de ISs com mais de 500 membros. Em *Escherichia coli* são encontradas várias cópias (<10) de qualquer uma das famílias mais comuns de ISs, com uma taxa de transposição média de 10^3 a 10^4 por geração. Também é interessante observar que no Domínio *Archea* e *Bacteria*, os transposons predominantemente são os ISs (Filée et al., 2007; Toro, 2003).

Os TEs são ubíquos entre os seres vivos. A capacidade de se mobilizar também faz deles candidatos para disseminação horizontal interespecífica (Monteiro-Vitorello et al., 2005). Em bactérias, por exemplo, os TEs podem ser encontrados em plasmídeos, o que aumenta a dificuldade para datar estes elementos na história evolutiva de alguns grupos (Bennetzen, 2000). Sendo observados como parasitos ou genes egoístas, a mobilização dos TEs faria deles sequências capazes de amplificação dentro do genoma sofrendo uma seleção

neutra, sempre e quando a amplificação não comprometa a adaptação ou fenótipo do hospedeiro (Bennetzen, 2000).

Segundo Finnegan (1989), os TEs podem ser divididos em duas classes, dependendo do substrato intermediário da transposição. Os retrotransposons (Classe I) utilizam uma molécula de RNA, enquanto os transposons (Classe II) utilizam uma molécula de DNA. Além disso, cada classe é composta por elementos autônomos e não-autônomos (Figura 1). Os primeiros são elementos com a capacidade de codificar produtos funcionais necessários na sua transposição, os segundos, são incapazes de codificar os produtos funcionais necessários para a transposição, mas possuem as sequências *in cis* necessárias para seu reconhecimento pelo sistema de transposição de TEs autônomos. Os elementos não-autônomos são dependentes dos membros autônomos da família porque perderam a capacidade de realizar sua própria transposição, devido a deleções ou mutações na região codificante (Feschotte e Pritham, 2007).

A principal categoria de TEs não-autônomos pertencentes à Classe II são os MITEs (*Miniature inverted-repeat transposable elements*) (Figura 1). Estes elementos apresentam entre 150 e 600 pares de base (pb), estão presentes em alto número de cópias no genoma, apresentam uma preferência de inserção por sítios TA ou TAA e possuem TIRs (*terminal inverted repeats*) em ambas as extremidades da sequência (Bennetzen, 2000; Macas et al., 2003; Moreno Vasquez et al., 2005). Para alguns MITEs também foi predita a formação de estruturas secundárias estáveis, por exemplo, o elemento da superfamília hAT conhecido como hATpin apresenta estrutura secundária de grampo (Moreno Vasquez et al., 2005).

Outra categoria importante de elementos não-autônomos são os elementos domesticados (Figura 1), também relacionados à Classe II. Os elementos domesticados são reconhecidos pela sua similaridade com domínio protéico de um domínio “transposase”, no entanto não possuem TIRs, não apresentam TSD (*target site duplication*) vizinhas as TIRs e apresentam posição genômica conservada entre espécies evolutivamente distantes (Feschotte, 2008).

O mecanismo de transposição dos elementos de Classe I é conhecido como “cópia e cola” e o dos elementos de Classe II “corta e cola”. Ainda assim, existem alguns elementos da Classe II que podem também apresentar o mecanismo de “copia e cola” (Feschotte et al., 2002; Feschotte e Pritham, 2007).

Figura 1-Estrutura de uma família de transposons.

Figura demonstrando as diferentes formas em que podem ser encontrados os transposons. O elemento autônomo possui uma ORF (*open reading frame*) íntegra que codifica uma transposase funcional e em suas extremidades encontram-se as TIRs (*terminal inverted repeats*). Os elementos não-autônomos apresentam uma ORF defeituosa conservando as TIRs. Os elementos domesticados não possuem TIRs podendo conservar ou não a integridade da ORF

Fonte: Modificado a partir de De Jesus (2007).

A proteína fundamental para o processo de transposição dos elementos de Classe II é a transposase (Tpase). Chamada de integrase (IN) para os elementos de Classe I, a RT (*Reverse Transcriptase*) também é fundamental para os elementos da Classe I (Haren et al., 1999).

O mecanismo de transposição apresenta três passos segundo revisado em Haren et al. (1999) e requer a formação de um complexo altamente organizado de nucleoproteínas (transposome). Este complexo está constituído por a Tpase ou IN, os extremos finais do TE e o DNA são os alvos de inserção. Os passos são descritos a continuação (Figura 2):

- a) saída do elemento de seu local original: É produzida uma clivagem das pontes fosfodiester por hidrólise dos extremos 3' do elemento na Classe I e II, com posterior clivagem das extremidades 5' na maioria dos casos da Classe II (Figura 2a, 2b). O espaço deixado pelo corte é reparado após a saída do elemento, algumas vezes por ligação simples e outras por recombinação genética usando como molde o outro alelo homólogo ou a

- cromátide irmã (Bennetzen, 2000). Na Classe I os extremos 3' livres do elemento são utilizados para realizar a transcrição do elemento em um intermediário de RNA, que posteriormente é retrotranscrito pela RT (codificada pelo mesmo elemento) em um DNA dupla fita;
- b) transferência do elemento (DNA dupla fita) para outro local do genoma, passo compartilhado por Classe I e II;
 - c) integração do elemento: Durante a inserção este apresenta regiões flanqueantes fita única que precisam ser completadas pela maquinaria de reparo/replicação do hospedeiro. O reparo destas sequências fita única das extremidades do elemento gera as conhecidas marcas de inserção dos TEs, chamadas de TSD.

Historicamente, os mecanismos de transposição têm sido separados em dois tipos, conservativos e replicativos, diferenciados pela replicação do elemento no deslocamento, o que aumenta o número de cópias deste no genoma. A transposição conservativa é típica dos elementos de Classe II (Figura 2a, 2b), com a exceção daqueles em que no processo de excisão corta-se apenas um extremo do elemento em cada fita de DNA. Nesse caso, cortar apenas uma das fitas produz uma transposição replicativa (Figura 2c). Entre os elementos com transposição conservativa encontramos Tn7, IS10, IS50, a família Tc/mariner, o elemento P e o elemento Ac. Este mecanismo conserva o número de cópias do elemento no genoma por ser uma excisão completa do elemento com posterior inserção em outro local do genoma, embora o número de cópias dos elementos que usam este mecanismo possa aumentar por replicação passiva, isto quer dizer replicação do genoma hospedeiro com posterior transposição do elemento da fita replicada (Haren et al., 1999).

A transposição replicativa (Figura 2c) é dada pela quebra de um único extremo do elemento em cada fita de DNA. O elemento não é excisado do local, porém, os extremos livres atacam outra região do genoma produzindo uma forquilha de replicação que será resolvida deixando uma cópia do elemento em cada local do genoma (Haren et al., 1999).

O mecanismo replicativo aumenta o número de cópias dos elementos, assim como o mecanismo de retrotranscrição usado pelos elementos Classe I (Figura 2d).

Figura 2-Mecanismos de transposição.

Nesse esquema o transposon está representado por caixas vazias, o local de origem do elemento está representado por caixas cinza e o local de inserção por caixas pretas. O local de quebra da cadeia de DNA é indicado por pequenas setas verticais. Os ataques nucleofílicos das pontes fosfodiéster também são apresentados. a)Elementos que sofrem dupla quebra da cadeia de DNA no local de origem, representados pelo elemento Tn7. b)Elementos que sofrem dupla quebra de DNA no local de origem com um intermediário em forma de *hairpin*, representados pelo elemento Ac. c)Elementos que passam por uma fase intermediária de cointegração, como o fago Mu. d)Elementos que sofrem transcrição com posterior retrotranscrição, retrovírus.

Fonte: Modificado a partir de Haren et al. (1999).

Os primeiros TEs descobertos em plantas pertenciam à Classe II, foram identificados pelas altas taxas de alterações que produziam no genoma e em fenótipos de tecidos somáticos e germinativos (Bennetzen, 2000). Na Classe II são encontrados TEs como; Ac, Em/Spm e MuDR de milho. (Bennetzen, 2000; Haren et al., 1999; Moreno Vasquez et al., 2005). A maioria dos elementos da Classe II tem um tamanho que varia de poucos pares de bases até 10Kb, são encontrados em

todos os organismos e constituem a maior classe de DNAs transponíveis dos procariontes (Bennetzen, 2000).

Estudos das sequências de DNA das extremidades dos transposons Classe II têm mostrado uma preferência de inserção em regiões ativas do genoma. Isto acontece, por exemplo, com MuDR (membro da família Mutator) e com o elemento Spm/En. Ambos têm preferência por se inserir em lugares não metilados do genoma do milho (Bennetzen, 2000).

Os TEs foram descobertos por Barbara McClintock em 1948, como importantes agentes mutagênicos. Em 1984 ela recebeu o Prêmio Nobel de Medicina pela profundidade de suas observações e a visão à frente de seu tempo. McClintock foi capaz de perceber a influência destes elementos na evolução e diversidade dos genomas, o que está sendo demonstrado por outros pesquisadores, apenas décadas depois (McClintock, 1948).

O trabalho de McClintock analisou mutantes de milho com padrões segregantes aberrantes para alguns *loci*. Uma destas segregações seria relacionada com o padrão da coloração do endosperma da semente do milho, que seria alterado por dois *loci* capazes de se transpor de uma região cromossômica para outra ao longo das divisões celulares do endosperma (McClintock, 1948).

Estes dois *loci* estudados por McClintock depois ganharam o nome de Ac (*Activator*) e Ds (*Dissociator*), sendo Ac reconhecido como o primeiro TE da história (McClintock, 1948).

Mais tarde, foi entendido que a variação do padrão de coloração seria causada por um elemento que interrompe um gene envolvido na biossíntese de antocianinas, gerando um grão sem pigmento e a excisão do elemento restauraria o fenótipo (Figura 3). A instabilidade dos fenótipos observados por McClintock se deve à inserção ou excisão do elemento. Setores revertentes (com pigmento) resultam da excisão do elemento em células da camada de aleurona. O tamanho do setor reflete o momento do desenvolvimento quando a excisão aconteceu, como pode ser observado na figura 3 (Feschotte et al., 2002).

O sequenciamento dos genomas demonstrou que os TEs são o maior componente do material genético na maioria dos eucariontes. Estes elementos representam cerca de 45% no genoma humano, e entre 50 e 80% nos genomas de algumas plantas (Feschotte et al., 2002). A descoberta diária de novos TEs, estimulada pelo aumento de genomas sequenciados e depositados nas bases de

dados, faz com que seja necessária uma classificação mais completa destes elementos, não apenas baseada em seus mecanismos de transposição.

Figura 3-Diferentes fenótipos em grãos de milho devido à ação de TEs.

O fenótipo colorido é recuperado pela excisão do TE em diferentes estágios de desenvolvimento do grão.

Fonte: Adaptado de Feschotte et al. (2002).

A mais recente das propostas para a classificação dos TEs de eucariotos é o sistema unificado do Wicker et al. (2007). Este sistema funciona de forma hierárquica integrando os conhecimentos estruturais e de similaridade entre as sequências dos elementos de transposição. O objetivo desta classificação é encontrar um sistema de fácil aplicação para pesquisadores não especializados em TEs, auxiliando assim na anotação dos novos genomas sequenciados. A Figura 4 permite uma visão mais detalhada da classificação postulada por Wicker e seus colaboradores para os elementos de transposição.

Figura 4–Sistema de classificação proposto para TEs.

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	Copia	→ GAG AP INT RT RH →	4–6	RLC	P, M, F, O
	Gypsy	→ GAG AP RT RH INT →	4–6	RLG	P, M, F, O
	Bel-Pao	→ GAG AP RT RH INT →	4–6	RLB	M
	Retrovirus	→ GAG AP RT RH INT ENV →	4–6	RLR	M
	ERV	→ GAG AP RT RH INT ENV →	4–6	RLE	M
DIRS	DIRS	→ GAG AP RT RH YR ←	0	RYD	P, M, F, O
	Ngaro	→ GAG AP RT RH YR → → →	0	RYN	M, F
	VIPER	→ GAG AP RT RH YR → → →	0	RYV	O
PLE	Penelope	← RT EN →	Variable	RPP	P, M, F, O
LINE	R2	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	tRNA		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	Tc1-Mariner	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PtgyBac	Tase	TTAA	DTB	M, O
	PIF-Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	Crypton	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	Helitron	RPA Y2 HEL	0	DHH	P, M, F
Maverick	Maverick	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features					
→	Long terminal repeats	→ ←	Terminal inverted repeats	█	Coding region
—	Non-coding region	—	Region that can contain one or more additional ORFs		
—	Diagnostic feature in non-coding region				
Protein coding domains					
AP, Aspartic proteinase	APE, Apurinic endonuclease	ATP, Packaging ATPase	C-INT, C-integrase	CYP, Cysteine protease	EN, Endonuclease
ENV, Envelope protein	GAG, Capsid protein	HEL, Helicase	INT, Integrase	ORF, Open reading frame of unknown function	
POL B, DNA polymerase B	RH, RNase H	RPA, Replication protein A (found only in plants)	RT, Reverse transcriptase	Y2, YR with YY motif	
Tase, Transposase (* with DDE motif)		YR, Tyrosine recombinase			
Species groups					
P, Plants	M, Metazoans	F, Fungi	O, Others		

Fonte: Figura extraída de Wicker et al.(2007).

1.2 Impacto dos Elementos de Transposição: Diversidade, Obesidade dos Genomas, Criação de Redes Regulatórias e Domesticação

Os TEs são agentes mutagênicos, sua atividade produz mudanças em genes particulares, como também grandes rearranjos no genoma. Podem alterar os padrões de expressão quando se inserem em regiões não codificantes, causar mutações que levem à inativação de genes quando a inserção ocorre em regiões codificantes e causar recombinação ou deleções cromossômicas por quebras do DNA no momento da excisão. Essas alterações podem afetar o fenótipo, mas este novo fenótipo é instável na maioria dos casos, sendo recuperado o fenótipo original com a excisão do elemento (Cordaux et al., 2006; Feschotte et al., 2002). Diferente dos TEs da Classe I, os TEs da Classe II têm a propriedade de se excisar. A excisão de elementos nem sempre ocasiona a recuperação do fenótipo inicial. As excisões parciais alteram a sequência original da região, produzindo pequenas deleções, inversões e introduções de DNA randômico para preencher a região. Gerando assim, variações alélicas vinculadas à excisão dos elementos (Feschotte e Pritham, 2007).

Entre os fenômenos ocasionados pelos TEs esta o aumento do tamanho nos genomas eucariotos. A maioria dos TEs produz uma duplicação do genoma hospedeiro no local onde aconteceu a inserção (TSD-*Target site duplication*). O tamanho desta duplicação depende da família de elemento de transposição (Cordaux et al., 2006; Feschotte et al., 2002). O genoma também é aumentado pelas cópias deixadas pelo próprio elemento, fenômeno especialmente encontrado no caso dos retrotransposons. Os transposons de DNA que usam o mecanismo de transposição replicativa também podem acarretar em aumento do tamanho do genoma.

Os fenômenos de transposição replicativa, assim como a criação de TSDs não têm um impacto dramático no aumento do genoma quando comparados com a propagação dos retrotransposons. Os retrotransposons são responsáveis pelo aumento considerável do tamanho de alguns genomas (obesidade genômica). São encontrados, por exemplo, os elementos das famílias RIRE1, Wallabi e Kangaroo que duplicaram o genoma de *Oryza australiensis* ao longo dos 3 últimos milhões de anos, aproximadamente. A família Gran3 aumentou o tamanho de *Oryza granulata*

em aproximadamente 25% (Ammiraju J. et al., 2007; Bennetzen, 2000; Bundock e Hooykaas, 2005; Feschotte e Pritham, 2007; Piegu et al., 2006).

Estudos em milho e arroz mostram que a amplificação massiva de retrotransposons com LTR, causou um aumento do genoma em até o dobro do tamanho original (Ammiraju J. et al., 2007; Piegu et al., 2006). Alguns transposons de DNA também podem contribuir com o aumento substancial do genoma, sendo um fenômeno menos comum, como é o caso do genoma de *Trichomonas vaginalis*. O 65% do genoma deste eucarioto unicelular está ocupado por transposons de Classe II, expansão causada em grande parte pelos transposons da superfamília Maverick que ocupam ~37% deste genoma (Feschotte e Pritham, 2007).

Os transposons de DNA são em princípio movimentados por mecanismo “corta e cola”, para aumentar seu número de cópias dependem de mecanismos indiretos, isto é, usam a maquinaria do hospedeiro. O primeiro dos mecanismos indiretos usa a transposição durante a replicação do DNA hospedeiro; o elemento é excisado de uma cromátide replicada para um sítio na cromátide que ainda não foi replicado, isto faz que o transposon seja replicado duas vezes, como foi documentado para os elementos Ac/Spm. O segundo mecanismo indireto depende da reparação da quebra na dupla cadeia de DNA que se produz após a excisão do elemento; se o elemento se insere em um cromossomo homólogo, a reparação é feita por recombinação homóloga resultando na reintrodução do elemento no local de excisão. Se a transposição é feita durante a fase S (*Synthesis/Replication phase*) do ciclo celular, a reparação pode ser feita por recombinação com a cromátide irmã, resultando também na reintrodução do elemento no local de excisão (Feschotte e Pritham, 2007).

O aumento do tamanho do genoma, além da variabilidade ocasionada pela transposição dos elementos poderia aumentar as possibilidades de codificação do genoma. Estes dois fenômenos poderiam conferir funções adaptativas na evolução dos eucariontes como propôs originalmente McClintock (Feschotte et al., 2002; McClintock, 1984).

Por outro lado, a atividade de transposição dos TEs está restrita apenas às últimas divisões celulares do desenvolvimento dos tecidos somáticos ou germinativos, além do mais, não possui correlação com os níveis de transcrição dos genes da célula (Rudenko e Walbot, 2001; Rudenko et al., 2003). Existem evidências da expressão dos TEs em momentos específicos da vida do hospedeiro,

como é o caso do elemento Tnt1 (retrotransposon) em tabaco. Normalmente este elemento é fortemente regulado no genoma do tabaco, apresentando expressão apenas em condições de estresse biótico e abiótico. Também foi mostrada expressão em condições heterólogas, quando o promotor (contido na LTR) é transferido para tomate ou *Arabidopsis* (Grandbastien, 2004). Ainda não está claro se a expressão do elemento corresponde a um mecanismo de defesa da planta, ou se a expressão do elemento é causada por falha nos mecanismos de regulação causada pelo mesmo estresse, quebrando mecanismo que normalmente manteriam o elemento silenciado.

Também é claro que mesmo existindo mecanismos de regulação dos TEs nos genomas eucariontes, alguns deles conseguem ser expressos, o que não significa que estes elementos estão conseguindo transpor-se, especialmente no caso dos retroelementos. Mesmo assim, existem casos como o elemento Mutator (transposon Classe II) que é expresso constitutivamente em meristema, órgãos imaturos e tecidos totalmente diferenciados de várias linhagens de milho (Rudenko e Walbot, 2001; Rudenko et al., 2003).

A célula possui variados mecanismos de silenciamento epigenético dos TEs, como; metilação de citosinas, iRNA, *imprinting* genômico e heterocromatina (Okamoto e Hirochika, 2001). Piriyaongsa J. e Jordan K. (2007), propõem que estes mecanismos surgiram para controlar os TEs na célula e posteriormente foram usados para controlar os padrões de expressão dos próprios genes do hospedeiro. Isto pode ter aumentado a complexidade regulatória e fenotípica, o que é observado nos organismos eucariotos.

Os elementos autônomos não são os únicos conhecidos por ter impacto nos genomas eucariotos. Os MITEs, descritos anteriormente, poderiam estar envolvidos em processos celulares como regulação. O que foi observado para os elementos Made1 de mamíferos. Os autores Piriyaongsa J. e Jordan K. em 2007 mostraram a existência em humanos de miRNAs (*micro-RNAs*) derivados da família Made1 atuando como iRNAs (*interference RNA*), servido assim à maquinaria celular hospedeira de regulação (Smalheiser e Torvik, 2005; Piriyaongsa J. e Jordan K., 2007).

Outro dos fenômenos importantes no estudo das modificações que os elementos de transposição podem causar no genoma hospedeiro é a Domesticação Molecular (Bundock e Hooykaas, 2005). Este fenômeno oferece um ponto de vista

dos TEs totalmente distinto ao papel de parasitas do DNA ou *Junk-DNA*, que se lhes conferiu nos primeiros estudos. Fazendo uma exceção com os estudos de McClintock, quem denominou aos TEs de “elementos controladores”. Denominação que se faz clara na Domesticação Molecular, onde as sequências de TEs perdem a capacidade de transpor-se devido à perda ou degeneração das TIRs. Além do mais, os TEs são recrutados pelo genoma hospedeiro para servir em funções celulares. Este recrutamento de novas sequências aumenta a capacidade de codificação do genoma. Este fenômeno faz dos TEs, elementos importantes no processo evolutivo do genoma hospedeiro (Cordaux et al., 2006; Cowan et al., 2005). Este poder de codificação de novas funções poderia dar vantagens como a resistência a estresses de tipo biótico ou abiótico, entre outros (Cordaux et al., 2005, 2006). Este é o caso da proteína RAG1, que faz parte do complexo RAG1/RAG2 que origina a recombinação V(D)J para a produção da cadeia leve dos anticorpos nos linfócitos B e T (Agrawal et al., 1998)

A perda ou degeneração de TIRs, assim como propriedades intrínsecas da estrutura dos TEs os fazem blocos interessantes para a construção de novos genes. O TEs de Classe II possuem em sua T_{pase} o domínio de ligação ao DNA, este domínio é compartilhado por vários genes que apresentam funções críticas para o hospedeiro. Uma destas funções é a regulação transcricional do genoma feita pelos seus fatores de transcrição, entre outras moléculas.

Os elementos domesticados também apresentam altos níveis de expressão comparáveis com os níveis de expressão dos genes, o que é incomum para um TE. Este padrão de expressão relaciona aos TEs com funções celulares (Feschotte e Pritham, 2007).

Para considerar um elemento domesticado este deve cumprir, segundo o revisado em Feschotte e Pritham (2007), seis características:

- a) ausência ou degeneração de sequência flanqueantes características (TIRs) e nenhuma evidencia de mobilização recente;
 - b) correspondência filogenética da proteína codificada (pelo TE domesticado) com proteínas codificadas por transposons;
 - c) capacidade intacta de codificação e evolução sob alta pressão funcional (distinto ao ocorrido com as regiões codificantes de TEs, as quais sofrem de uma pressão evolutiva neutral);
-

- d) detecção de ortólogos intactos em regiões sintênicas de espécies distantes relacionadas;
- e) evidências de transcrição;
- f) evidencia genética de função biológica *in vivo*.

Entre os casos relatados de domesticação, estão os genes *FAR1* (*far-red impaired response protein 1*), *FHY3* (*far-red elongated hypocotyl 3*), *FRS* (*FAR1-related sequences*) e o elemento *MUG1* (*MUSTANG*). Todos, relacionados à família de transposons MULEs (*Mutator like*) (Cowan et al., 2005; Hudson et al., 2003). Também está o elemento DAYSLEEPER, conhecido como o primeiro transposon domesticado derivado da família hAT (Bundock e Hooykaas, 2005).

A história de domesticação de TEs parece ser muito antiga na evolução das espécies. O mecanismo de defesa generalizado de vertebrados é a produção de imunoglobulinas e receptores de células T. Estas moléculas são geradas por recombinação V(D)J, um processo que gera uma combinação quase interminável de alelos responsáveis pela diversidade da identificação do sistema imunológico. Recentemente esta tomando força a hipótese que diz que o mecanismo de recombinação V(D)J é derivado de transposons Classe II domesticados (Feschotte e Pritham, 2007).

A domesticação das Tpsases também pode gerar a criação de redes regulatórias. O gene gerado pela domesticação conserva a capacidade de interagir com sequências de reconhecimento (TIRs) e sequências subterminais de TEs previamente dispersos pelo genoma. A interação do elemento domesticado com TEs adjacentes a genes pode gerar diferentes efeitos nestes; ativação, repressão ou nenhum efeito. As interações benéficas para o genoma seriam selecionadas e as deletérias seriam eliminadas (Feschotte e Pritham, 2007).

1.3 Superfamília hAT

A superfamília hAT é definida por elementos que compartilham os domínios de dimerização e ligação ao DNA semelhantes aos elementos **h**obo (encontrado em *Drosophilla melanogaster*), **A**ctivator (*Zea mays*) e **I**am3 (*Antirrhinum majus*), previamente descritos na literatura. Outros elementos importantes que fazem parte desta superfamília são Hobo (*Drosophilla melanogaster*), Crypt1 (do fungo

Cryptonectria parasítica), DodoPi (do fungo fitopatológico *Phytophth infestans*) e o elemento Hermes (*Musca domestica*), entre outros (Moreno Vasquez et al., 2005; Rubin et al., 2001).

Análises filogenéticas e de agrupamento das sequências da superfamília hAT sugerem que sua origem é bastante antiga, aparecendo na história evolutiva das espécies provavelmente antes da separação de planta-fungo-animal. Também foi mostrado, que não existe evidência de HGT (*Horizontal gene transfer*) influenciando na evolução e ampla distribuição da superfamília (Quadro 1) (Rubin et al., 2001).

Quadro 1—Membros da Superfamília hAT

Organismo	Elemento de Transposição
Fungos	Fot1 em <i>Fusarium oxysporum</i> , Restless em <i>Tolypocladium inflatum</i>
Plantas	<u>Monocotiledôneas</u> ; Ac/Ds em <i>Zea mays</i> , Millet-Ac-like em <i>Pennisetum glaucum</i> . <u>Dicotiledôneas</u> ; Tam3 em <i>Antirrhinum majus</i> , Slide em <i>Nicotiana tabacum</i> , Tag1 em <i>Arabidopsis</i> .
Animais	<u>Invertebrados</u> ; hobo em <i>Drosophila melanogaster</i> , Hermes em <i>Musca domestica</i> . <u>Vertebrados</u> ; 1723 em <i>Xenopus laevis</i> .

Fonte: Revisado de Moreno Vasquez et al. (2005).

A característica mais conservada na superfamília hAT é um domínio de ~50 aminoácidos localizado na região C-terminal entre os resíduos 682 a 720 da Tpase (Pfam PF05699). Este domínio está envolvido na sua dimerização, a superfamília se mostrou ativa quando forma oligômeros (Essers et al., 2000). No extremo N-terminal da Tpase está presente outro domínio característico da superfamília hAT, com função de ligação ao DNA (do tipo *Zinc Finger*, Pfam PF02892) (Haren et al., 1999). Estes dois domínios são usados para caracterizar e agrupar os elementos da família SChAT, que será apresentada posteriormente.

As famílias de transposons contidas na superfamília hAT são definidas por compartilhar sequências de TIRs similares em suas extremidades. Cada família pode apresentar um ou mais membros autônomos (Bennetzen, 2000; Moreno Vasquez et al., 2005), geralmente é restrita a uma espécie, como por exemplo a família Ac presente no milho (Moreno Vasquez et al., 2005). Análise dos genomas

de plantas tem mostrado que os membros desta superfamília normalmente estão em baixo número de cópias (menos de 400), como foi visto em Milho, *Arabidopsis thaliana* e *B. oleracea* (Feschotte et al., 2002; Moreno Vasquez et al., 2005; Rubin et al., 2001).

Os membros da superfamília hAT também são caracterizados por produzir uma TSD de 8pb. Os elementos desta superfamília possuem dois motivos *in cis* de reconhecimento, necessários para a formação da maquinaria de transposição: TIRs, curtas com a sequência consenso (T/C)A(A/G)NG; STRs (*Subterminal regions*), com motivos repetitivos curtos. Geralmente, diferenças nestes dois motivos repetitivos tem sido usadas para caracterizar as diferentes famílias dentro de hAT. Estes motivos *in cis* têm de 4 a 12 pb e podem ser usados como sítios de ancoragem pela transposase que promove a dupla quebra de DNA no local de origem, este DNA passa por um intermediário em forma de *hairpin* e se insere posteriormente em outro local do genoma (Moreno Vasquez et al., 2005; Rubin et al., 2001).

No estudo de Rubin et al. (2001) também foi observado que a superfamília hAT possui 6 blocos conservados de aminoácidos ao longo de sua Tpase, assim como TIRs pouco conservadas.

Dentro da família hAT também são encontradas Tpases domesticadas como DAYSLEEPER, que demonstrou ter função essencial para o crescimento e desenvolvimento de *Arabidopsis*. A proteína codificada por este elemento atua como um fator de transcrição se ligando às regiões promotoras de diversos genes. O silenciamento desta Tpase gera deformidades no desenvolvimento das plântulas (Bundock e Hooykaas, 2005). Com base na sequência de DAYSLEEPER, foram descritos outros TEs domesticados associados à superfamília hAT em trigo e cevada. Estes foram chamados de w-gary em trigo e b-gary em cevada (Muehlbauer et al, 2006).

O elemento Ac de milho, descoberto inicialmente por McClintock em 1948, foi o primeiro TE caracterizado da história. Depois de sua descoberta, Fedoroff et al. (1983) isolaram molecularmente os elementos Ac e Ds (membro não-autônomo da família). Um ano depois Pohlman et al. (1984) determinaram a sequência nucleotídica do elemento Ac identificando o gene que codifica a transposase e as TIRs imperfeitas de 11 pb.

Um único transcrito foi identificado para o elemento Ac, com 3,5 Kb de tamanho e formado por cinco éxons. Foi encontrada uma ORF de 2,421 pb relativa à transposase, depois chamada de ORFa (Figura 5) (Kunze et al., 1987).

Figura 5-Estrutura do elemento Ac.

Nessa figura podem ser observadas as TIRs (triângulos pretos nos extremos do elemento), a região transcrita (mRNA), os 4 íntrons, as duas ORFs codificadas e os domínios de dimerização (DD) e de localização nuclear (N1,N2,N3) do elemento Ac. Fonte: Extraído de Kunze R. et al. (1987) e Weil e Kunze (2000).

Becker e Kunze (1997) demonstraram que a transposase do elemento Ac apresenta um domínio de ligação ao DNA (do tipo *Zinc finger*) que reconhece tanto as TIRs quanto pequenos motivos de trinucleotídeos, repetidos nos 250 pb subterminais de cada extremidade do elemento. A primeira demonstração de que o elemento Ac poderia se mobilizar em um sistema heterólogo foi realizada com protoplastos de *Nicotiana tabacum* (Baker et al., 1986). Em seguida, Van Sluys et al. (1987) demonstraram a transposição de Ac em raízes e raízes derivadas de calos de plantas transgênicas de *Arabidopsis thaliana* e *Daucus carota*.

O elemento Ac foi utilizado neste trabalho como base para a comparação da estrutura dos elementos da superfamília hAT presentes no genoma da cana-de-açúcar, chamados de SChAT e descritos neste projeto.

1.4 Família SChAT e a História do Estudo do Potencial Genômico da Cana-de-açúcar

A cana-de-açúcar é uma cultura de alta importância econômica a nível mundial, conhecida pela capacidade de acumular sacarose em seu talo. Além de ser importante para a indústria alimentícia, também é fundamental para a indústria de produção do etanol. Cerca de dois terços do açúcar produzido no mundo são

derivados da cana-de-açúcar, que possui cerca de 20 milhões de hectares plantados em regiões tropicais e subtropicais do globo (Menossi et al., 2008).

A importância econômica desta cultura a faz alvo de inúmeros projetos de pesquisa que procuram entender e manipular seu potencial genômico tendo em vista o aumento da produtividade.

O presente trabalho pretende ampliar o conhecimento sobre o impacto dos elementos de transposição sobre o genoma das gramíneas, em especial da cana-de-açúcar. É necessária a junção entre o entendimento do genoma e sua regulação, as vias metabólicas, os aspectos fisiológicos e produtivos da cana-de-açúcar para se conseguir um nível de manipulação e desenvolvimento que nos permita suprir a futura demanda deste cultivar por parte da indústria alimentícia e energética mundial.

A cana-de-açúcar (*Saccharum* spp) é uma planta monocotiledônea pertencente à família *Poaceae* (gramíneas), como o arroz, subfamília *Panicoideae*, tribo *Andropogoneae*, como o milho e o sorgo, subtribo *Saccharinae*. Segundo a classificação taxonômica do complexo *Saccharum* revisada por Amalraj e Balasundaram (2006), este gênero possui seis espécies: *S. officinarum*, *S. barberi*, *S. sinense*, *S. edule*, *S. robustum* e *S. spontaneum*.

São observados dois principais centros de origem do gênero *Saccharum*; o Sudeste Asiático, mais especificamente na Ilha da Nova Guiné (*S. officinarum*, *S. edule*, *S. robustum*), e as espécies com origem no Norte da Índia e da China (*S. barberi* e *S. sinense*). *S. spontaneum* seria a espécie com a maior distribuição geográfica, com origem na Índia e se estendendo da região do Mediterrâneo, passando pela África e chegando até a Nova Guiné (Grivet et al., 2004).

A cana-de-açúcar é caracterizada por possuir um dos genomas mais complexos entre as culturas domesticadas pelo homem. As variedades modernas de cana são derivadas de híbridos produzidos por cruzamentos interespecíficos entre *Saccharum officinarum* e *Saccharum spontaneum*. *Saccharum officinarum* apresenta número cromossômico básico de $x=10$ e $2n=80$ cromossomos, o que indica que é octaploide. *Saccharum spontaneum* apresenta número cromossômico básico de $x=8$ e $2n=40-128$ cromossomos indicando diferenças de ploidia de 5 a 16 (D'Hont, 2005). Estes cruzamentos procuram aproveitar as características agronômicas de ambas espécies; *Saccharum officinarum*, rica em açúcar, e *Saccharum spontaneum* que aporta maior robustez e resistência às doenças.

As hibridações trouxeram ganhos efetivos na produção. Baseado em evidências moleculares, D'Hont (2005) estima que os cultivares modernos de cana-de-açúcar são decaploides, com $2n=100-130$, o que daria um número cromossômico básico de $X=10-13$. O tamanho do genoma total seria de aproximadamente 10 GB. Nos quais 75 a 85% dos cromossomos são originários de *S. officinarum* e 15 a 25% dos cromossomos são originários de *S. spontaneum*.

A figura 6 representa a organização cromossômica inferida para os cultivares modernos de cana. Destaca-se a alta poliploidia, a aneuploidia, a origem bi-parental dos cromossomos (com no mínimo 3/4 herdados de *S. officinarum*), a ocorrência de diferenças estruturais entre os cromossomos de origens distintas e a existência de cromossomos recombinantes interespecíficos (Grivet e Arruda, 2002).

Figura 6-Diagrama esquematizando a organização do genoma de um cultivar atual de cana-de-açúcar.

Nesse diagrama cada barra representa um cromossomo; em amarelo estão representadas as regiões provenientes de *S. officinarum* e em verde as provenientes de *S. spontaneum*. Os cromossomos alinhados na mesma coluna pertencem ao mesmo grupo homeólogo.
Fonte: Grivet e Arruda (2002).

Avaliações citogenéticas em diferentes cultivares modernos propõem que menos de 5% dos cromossomos de cana sofreram recombinações e possuem trechos de ambas das espécies progenitoras (Cuadrado et al., 2004; D'Hont, 2005).

Estima-se que um conjunto básico de cromossomos de cana corresponda a entre 760 e 926 Mpb (Menossi et al., 2008). Embora seja consideravelmente maior que o genoma do arroz, que é uma espécie-modelo de monocotiledôneas (389 Mpb), o tamanho é similar ao do sorgo (760 Mpb) (Menossi et al., 2008), que entre plantas com interesse agrônômico é a mais próxima filogeneticamente da cana-de-açúcar.

Devido à complexidade do genoma, à inexistência do sequenciamento em larga escala, as primeiras tentativas de entender e aproveitar o potencial genômico da cana-de-açúcar iniciaram pela localização de genes e regiões genômicas com interesse agrônômico usando o mapeamento genético. Foram mapeados genes como o relacionado à resistência à ferrugem em cana (Asnaghi et al., 2000), QTLs (*Quantitative trait loci*) relacionados à quantidade de açúcar (Ming et al., 2001) e à produtividade em cana (Hoarau et al., 2002). Ainda que tenham sido obtidos resultados positivos, a complexidade do mapeamento genético em cana torna pouco atrativo este método para o estudo deste genoma.

Posteriormente, a pesquisa da cana-de-açúcar entrou na era do sequenciamento de larga escala. Seu potencial genético foi acessado pela estratégia de transcriptoma com o projeto SUCEST (*Sugarcane EST Project*), através do sequenciamento de ESTs (*Expressed Sequence Tag*). Este projeto foi desenvolvido majoritariamente no estado de São Paulo e foram sequenciados parcialmente mais de 237,000 RNAs mensageiros de cana (Vettore et al., 2003). O projeto construiu um total de 26 bibliotecas de cDNA a partir de diferentes tecidos de 13 cultivares de cana-de-açúcar, cuja descrição detalhada encontra-se em Vettore et al. (2001).

O agrupamento destas sequências resultou em uma base de dados com cerca de 33,000 genes únicos, que se estima representam mais de 90% dos genes da cana-de-açúcar (Vettore et al., 2003), dos quais cerca de 70% apresentam um correspondente em *Arabidopsis thaliana* e mais de 81% apresentam um homólogo em arroz (Vincentz et al., 2004).

Dentro da caracterização *in silico* dos hipotéticos genes de cana, também se identificou a presença de sequências similares a TEs das Classes I e II. As famílias de retrotransposons (Classe I) mais representadas foram Hopscotch, Bs1, Opie2 e Maggy. Para os transposons, ou elementos da Classe II, as famílias mais representadas foram MuDr, Ac e Tam3. Estes TEs foram encontrados expressos em

diferentes tecidos e ao longo de diferentes estágios do crescimento da planta (Araujo et al., 2005; Rossi et al., 2001, 2004; Vettore et al., 2003; Vincentz et al., 2004).

A partir de inferências *in silico* com base nos bancos de dados de ESTs, estima-se que pouco mais de 2% do transcriptoma de cana é composto de sequências relacionadas a elementos genéticos móveis expressos (Vettore et al., 2003), porcentagem semelhante foi observada em trigo, outra gramínea poliplóide (Li et al., 2004). Comparando-se estes dados aos do transcriptoma de milho (gramínea diplóide), que apresentou menos de 0,1% de transcritos relacionados aos TEs (Meyers et al., 2001), poderia se relacionar os recentes eventos de poliploidia e hibridação à ativação transcricional de elementos de transposição em gramíneas.

Trabalhos prévios do GaTE_Lab realizaram uma análise detalhada de algumas famílias de TEs encontradas no SUCEST, como em Rossi et al. (2004) e Saccaro JR et al. (2007), bem como avaliaram o padrão transcricional destes elementos (Araújo et al., 2005). Os elementos da Classe II relacionados à superfamília hAT compõem o segundo grupo mais representado quantitativamente na base de dados do SUCEST, depois de Mutator, e apresentaram alta expressão em calo e flores (Araújo et al., 2005).

Os elementos expressos similares à superfamília hAT encontrados na cana-de-açúcar foram chamados de elementos SChAT (*Sugar Cane hAT-like elements*). Possuem Tpsases codificadas por elementos do tipo hAT e domínios protéicos típicos da superfamília (Kempken et al., 2001; De Jesus et al., 2012).

No trabalho desenvolvido por De Jesus et al. (2012), onde foram encontrados os elementos SChAT, estes foram separados em três linhagens em função da similaridade interna das Tpsase preditas.

Os membros da linhagem I (074) apresentam características similares às características dos elementos descritos como domesticados, como são um baixo número de cópias, evidência de transcrição e correspondência filogenética com outras transposases domesticadas como é a transposase do elemento DAYSLEEPER (De Jesus et al., 2012), o que os faz candidatos interessantes para estudos moleculares e funcionais. Estes estudos poderiam mostrar se estas sequências são realmente elementos domesticados, se têm algum papel na regulação do genoma da cana-de-açúcar e se este papel é fundamental para o desenvolvimento da planta.

A recuperação das sequências da linhagem II, que posteriormente seria dividida em linhagens 191 e 257, também possibilitariam a realização de estudos moleculares e funcionais que mostrem a capacidade de transposição dos elementos da família SChAT correspondentes a estas duas linhagens de elementos *strictu sensu*.

Também é fundamental estudar se existe algum tipo de interação entre as proteínas da linhagem I (domesticado) e II (transposases), o que poderia dar indícios sobre a regulação das transposases da família SChAT no genoma da cana-de-açúcar.

O presente projeto dá continuidade ao trabalho desenvolvido por De Jesus (2007), onde a família SChAT foi dividida em duas linhagens (elementos domesticados e transposases) de elementos e que permitiu a recuperação de 3 integrantes da família; SChAT-G1, SChAT-G2 e um possível MITE. O objetivo é recuperar versões genômicas dos integrantes da família que permitam sua caracterização molecular e funcional, assim como dos elementos domesticados relacionados. Parte destes resultados estão consignados em De Jesus et al. (2012).

Os TEs têm um importante papel estrutural, funcional e regulatório nos genomas. As respostas às questões propostas por este projeto deverão, portanto contribuir para a compreensão do impacto dos TEs sobre o genoma complexo e poliplóide da cana-de-açúcar. Conhecimento relevante para o desenvolvimento de tecnologias que visem aprimorar a produtividade desta importante cultura.

2 OBJETIVOS

2.1 Objetivo Geral

Contribuir para a compreensão do impacto das transposases da superfamília hAT na organização e evolução do genoma das gramíneas, em particular da família SChAT de cana-de-açúcar (*Saccharum* spp.).

2.2 Objetivos Específicos

- Recuperar e caracterizar cópias genômicas das linhagens da família SChAT.
 - Analisar a diversidade e relação evolutiva das versões identificadas.
 - Identificar os padrões de expressão da linhagem domesticada 074 da família SChAT.
 - Analisar sítios de integração e padrões de expressão dos transposons das linhagens 191 e 257 da família SChAT.
 - Análise de mobilidade dos elementos da linhagem 191 e 257 da família.
-

3 MATERIAIS E MÉTODOS

3.1 Material Vegetal

As variedades de cana-de-açúcar aqui utilizadas variaram em função da disponibilidade de material e/ou objetivo do experimento. Os materiais de cana, bem como os de sorgo, *Miscanthus* e *Erianthus* foram gentilmente fornecidos pelo CTC (Centro de Tecnologia Canavieira – Piracicaba, SP).

Os RNAs de meristemas apicais das variedades RB75-126 e SP81-3250 foram gentilmente cedidos pela Profa. Dra. Kátia Castanho Scortecci (Universidade Federal do Rio Grande do Norte). Os outros RNAs usados neste trabalho pertencem às variedades;

- a) SP80-3280, cultivar utilizado nas bibliotecas de cDNA do projeto SUCEST (Vettore et al., 2001) ;
- b) R570, cultivar utilizado na biblioteca comercial de clones genômicos de BAC de cana-de-açúcar (Tomkins et al., 1999).

Quadro 2–Material vegetal utilizado para os diferentes experimentos deste trabalho.

Variedade	Espécie	Experimento utilizado	Objetivo
SP80-3280	<i>Saccharum</i> híbrido	Amplificação de MITE , Hibridação (<i>Southern Blot</i>), Hibridação (<i>Northern Blot</i>), RT-PCR e Real time PCR	2, 3 e 4
R570	<i>Saccharum</i> híbrido	Hibridação (<i>Southern Blot</i>), Hibridação (<i>Northern Blot</i>), RT-PCR e Real time PCR Biblioteca comercial BACs	1, 2, 3 e 4
SP81-3250	<i>Saccharum</i> híbrido	Hibridação (<i>Southern Blot</i>), Hibridação (<i>Northern Blot</i>), RT-PCR e Real time PCR	3 e 4
RB75-126	<i>Saccharum</i> híbrido	Hibridação (<i>Southern Blot</i>), Hibridação (<i>Northern Blot</i>), RT-PCR e Real time PCR	3 e 4
SP80-1836	<i>Saccharum</i> híbrido	Amplificação de MITE	4
Mandalay	<i>Saccharum spontaneum</i> híbrido	Hibridação (<i>Southern Blot</i>)	2, 4
Badila	<i>Saccharum officinarum</i> híbrido	Hibridação (<i>Southern Blot</i>)	2, 4
Coimbatore	<i>Saccharum</i> híbrido	Hibridação (<i>Southern Blot</i>)	2, 4
SP71-1088	<i>Saccharum</i> híbrido	Hibridação (<i>Southern Blot</i>)	2, 4
CP78-2114	<i>Miscanthus</i> spp.	Hibridação (<i>Southern Blot</i>)	2, 4
IK7648	<i>Erianthus</i> spp.	Hibridação (<i>Southern Blot</i>)	2, 4
Sorgo	<i>Sorghum</i> spp.	Hibridação (<i>Southern Blot</i>)	2, 4

Fonte: Ochoa (2012).

3.2 Cepa Bacteriana Utilizada (*Escherichia coli*)

DH10B – genótipo F⁻ *mcrA* $\Delta(mrr-hsdRMS-mcrBC)$ $\Phi80dlacZ\Delta M15$ $\Delta lacX74$ *endA1* *recA1* *deoR* $\Delta(ara,leu)7697$ *araD139* *galU* *galK* *nupG* *rpsL* λ^- . A cepa foi mantida em placas com meio sólido LB e repicadas mensalmente. Para conservar a cepa por períodos mais longos (até 1 ano), foram feitos estoques de glicerol. Estes são obtidos adicionando-se 1 mL de meio LB com 40% de glicerol a 1 mL de cultura líquida bacteriana crescida por no máximo 12 h em LB.

3.3 Plasmídeos

3.3.1 Clonagem de produtos de PCR e RT-PCR

PGEM: pGEM®-T Easy Vector System (Promega, código A1360), vetor para clonagem de produtos amplificados em reação de PCR. Possui 3015 pb e confere resistência à ampicilina.

3.4 Antibióticos

Ampicilina (100 mg/mL) e Clorafenicol (25 mg/mL): Solução preparada em água filtrada (filtro 0,22 μ m) estéril e armazenada a -20 °C. A concentração empregada no meio de cultura para a seleção de bactérias contendo plasmídeo de interesse foi de 100 μ g/mL e 25 μ g/mL, respectivamente.

3.5 Manipulações Gerais com Bactérias (*Escherichia coli*)

3.5.1 Meios de Cultura para Bactérias

Os meios de cultura foram autoclavados a 120 °C por 20 min. Para os meios sólidos foram acrescentados 20 g de ágar por litro. Quando necessário, os meios foram acrescidos do antibiótico de seleção, IPTG na concentração final de 200 μ g/mL e X-Gal na concentração final de 50 μ g/mL.

- **LB** – 10 g bactotripton; 5 g extrato de levedura; 5 g NaCl; pH ajustado para 7,4 com NaOH; água destilada para completar o volume de 1 litro.
-

- **SOC** – 5 g bactotripton; 1,25 g extrato de levedura; 500 µl NaCl 5M; 620 µl KCl 1M; 2,5 mL MgCl₂ 1M; 2,5 mL MgSO₄ 1M; glicose para uma concentração de 20 mM; pH ajustado para 6,8; água destilada para completar o volume de 250 mL.
- **Circle Grow (BIO 101 – Inc.)** – de acordo com as instruções do fabricante.

3.5.2 Obtenção de Bactérias Químio Competentes

Para a obtenção de bactérias competentes, um inóculo de colônia isolada de DH10B foi colocado para crescer em 2 mL de meio SOC, a 37 °C em agitação por 18 horas. Em seguida esta cultura foi reinoculada em 50 mL do mesmo meio, e incubada sobre agitação a 37 °C até atingir uma medida de absorbância entre 0,5 e 0,6, medida em espectrofotômetro com o comprimento de onda de 600 nm. Foi adicionado 0,5 mL de MgCl₂ 1M à cultura, que ficou no gelo por 15 min. As bactérias foram centrifugadas a 3000 RPM por 20 min, a 4 °C. O sobrenadante foi descartado, e o precipitado ressuspenso em 15 mL de solução RFI gelada, ficando por 10 min no gelo. Em seguida uma nova centrifugação foi realizada, e as bactérias foram ressuspenso em 2 mL de solução RFII gelada. Este volume foi dividido em alíquotas de 50 µL, que foram estocadas a -70 °C. Toda a manipulação foi realizada em fluxo laminar, com o uso de material estéril.

As soluções de MgCl₂, RFI e RFII foram preparadas com água deionizada e autoclavadas a 120 °C por 15 min.

- **Solução RFI** - acetato de potássio (30 mM); cloreto de manganês (50 mM); cloreto de potássio (100 mM); cloreto de cálcio (10 mM) e glicerol (15%).
- **Solução RFII** - MOPS pH 7,0 (10 mM); cloreto de cálcio (75 mM) cloreto de potássio (10 mM) e glicerol (15%).

3.6 Clones da Biblioteca Comercial de BACs de Cana-de-açúcar

Uma biblioteca comercial de clones genômicos de BAC de cana-de-açúcar (Tomkins et al., 1999), da variedade R570, foi comprada da *Clemson University Genomics Institute* e escaneada por hibridação molecular com sondas derivadas de uma região conservada de 2 representantes das linhagens de elementos SChAT

(TE191 e TE257) e um representante da linhagem dos elementos domesticados (TE074) relacionados à família (ver 3.11.1).

A biblioteca foi construída no vetor pbeloBAC11 e possui 103.296 clones com tamanho médio de 130 Kb (Tomkins et al., 1999), representando aproximadamente 1,3 vezes o genoma total da variedade R570 (10 Gb com $2n = 115$ segundo D'Hont, 2005).

3.6.1 Avaliação do número de cópias relativo das linhagens 74,191 e 257

Foram avaliados os 103.296 clones da biblioteca por PCR com iniciadores específicos para cada linhagem (Figura 8 e Quadro 3). Para se chegar ao número de cópias no genoma de cana, foi usado o seguinte cálculo:

$$z = x/1,3 \quad (1)$$

Onde x representa o número total de clones positivos para a biblioteca; 1,3 a cobertura da biblioteca; z , o número de cópias para um genoma total da variedade R570.

3.7 Eletroforese e Géis de Agarose

Os géis de agarose utilizados no decorrer deste projeto foram preparados da seguinte maneira: agarose (Gibco) numa concentração de 0,7% a 2%, dissolvida em tampão TBE 1/2X, levando-se ao microondas até a mistura entrar em ebulição. Em seguida os géis foram vertidos em molde apropriado para polimerizar à temperatura ambiente. Os DNAs, as reações de PCR e de hidrólise com endonucleases de restrição foram misturados em tampão de carregamento para eletroforese de DNA, e então foram aplicados no gel. As cubas utilizadas para eletroforese foram os modelos "Horizon 58" e "Horizon 11.14" (Gibco BRL). Após a eletroforese, os géis foram corados em uma solução de brometo de etídeo a 5 µg/mL, lavados em água, visualizados no "Image Master®VDS" (*Pharmacia Biotech, US*), e fotografados no "Thermal Imaging System FTI-500" (*Fuji Film*).

- **TBE 10X** – 108 g de Tris base; 55 g de Ácido bórico; 40 mL de EDTA 0,5 M pH8,0. Completar o volume para 1L com água deionizada. Autoclavar a 120 °C por 20 min.

- **Tampão de amostra para eletroforese de DNA** - azul de bromofenol 0,25%(p/v); xileno cianol 0,25% (p/v) e glicerol 30% (p/v). Acertar o volume com água deionizada.

3.8 Procedimentos Gerais de Manipulação de DNA

3.8.1 Extração de DNA Genômico de Plantas

O DNA genômico de sorgo, *Miscanthus*, *Erianthus*, e as variedades de cana-de-açúcar usadas neste trabalho, foi extraído de folhas, com base no protocolo abaixo, adaptado de Stewart e Via (1993).

Macerou-se 3 g de folhas em nitrogênio líquido, adicionando Poly(vinylpolypyrrolidone) (0,5 g por grama de material), aos quais foi adicionado 10 mL/g de tecido de Tampão CTAB [CTAB (brometo de cetiltrimetilamônio) 2%, NaCl 1,4 M, EDTA (ácido etilenodiamino tetra-acético) 20 mM, Tris pH8,0 (trishidroximetilaminometano) 100 mM, Poly(vinylpolypyrrolidone) 1% (p/v)]. Para cada 100 mL de tampão CTAB acrescentou-se 200 µL de β-mercaptoetanol, apenas no momento da utilização. O conjunto macerado e solução foi incubado por 2 a 5 min a 60-65 °C e agitado manualmente a cada minuto. Em seguida, adicionou-se um volume de clofofórmio:álcool isoamílico (24:1 v/v) igual ao volume total obtido até o momento, que em seguida foi agitado delicadamente e centrifugado a 5.000 rpm por 20 min (*Hettich Mikro 22R*).

A fase aquosa foi transferida para um novo tubo e a ela acrescentada um volume de fenol (pH 8,0):clofofórmio:álcool isoamílico (25:24:1 v/v) igual ao volume total obtido até o momento e agitado manualmente. Mais uma vez, o conteúdo foi centrifugado a 5.000 rpm por 20 min. Esta fase foi repetida uma segunda vez com a fase aquosa aqui obtida.

A seguir, transferiu-se a fase aquosa para um novo tubo, ao qual foi adicionado 2/3 do volume de isopropanol 100%, Este conteúdo foi levemente agitado. A partir dele foi possível "pescar" o precipitado (DNA) com um gancho feito com uma pipeta Pasteur, e mergulhá-lo em etanol 70%, que foi posteriormente seco no gancho à temperatura ambiente. Ao final do processo, foi adicionado 100 µL de TE (Tris-HCl 10mM e EDTA 1mM) e deixado durante 14-16 h a 4 °C para

ressuspensão. Antes de sua utilização em diferentes manipulações, o DNA genômico foi tratado com 50 µg de RNase H e incubado por 1 h a 37 °C.

- **Tris-HCl** - Tris 10mM. Acertar o pH 7.8 com HCL e o volume final com água deionizada.
- **EDTA** - EDTA 1mM. Acertar o pH 7.8 com NaOH e o volume final com água deionizada.
- **RNase H** -
- **CTAB** – 2% CTAB; 100mM Tris-HCL pH 8; 20 mM EDTA; 1,4M NaCl. Na hora de usar adicionar 0,2% de β-Mercapto.

3.8.2 Quantificação de DNA

A quantificação do DNA genômico utilizado em diferentes experimentos foi feita no fluorômetro HoeferDyNA-Quant de acordo com as especificações do fabricante, e com o Spectrophotometer Nanodrop®, US,ND-100, segundo especificações do fabricante.

- **Solução de ensaio para o uso do fluorômetro** - 10 µL de H33258 (10 mg/µL); TNE10X 10 mL. Com água destilada levar até um volume de 100 mL.
- **TNE 10X** - Tris 100 mM; EDTA Na₂.2H₂O 10 mM; NaCl 2 M; Tris pH 8,0; 1 mM EDTA pH 8,0.

3.8.3 Amplificação de Fragmentos por PCR (*Polymerase Chain Reaction*)

Para as reações de PCR foi utilizado 60 ng de DNA genômico, água destilada, ultrafiltrada e autoclavada, tampão de PCR 1X, 1,5 mM MgCl₂, 0,2 mM de dNTPs (dATP, dCTP, dGTP e dTTP), 0,2 µM de cada iniciador e dois tipos de DNA polimerase: para fragmentos até 2000 pb, foi utilizado 1,25 unidades de Taq Polymerase Recombinante (Invitrogen Brasil código 11615-010); para fragmentos maiores que 2000 pb foi utilizada a Elongase Enzyme Mix (Invitrogen Brasil código 10480-010), de acordo com as recomendações do fabricante.

Sempre foi feito um controle negativo contendo todos os reagentes menos a amostra de DNA, para garantir a ausência de qualquer tipo de contaminação. Todas as reações foram feitas em um termociclador PTC-100 (MJ Research, Inc.).

Os ciclos de amplificação variaram em função da polimerase utilizada finalidade do PCR e da temperatura de anelamento, que são relatadas em cada um dos experimentos ao longo deste capítulo.

A figura 7 resume os ciclos de PCR utilizados. Posteriormente as amostras eram mantidas a 4 °C.

Figura 7-Resumo dos ciclos de PCR utilizados em diferentes procedimentos do trabalho.

Fonte: Ochoa (2012).

3.8.4 Purificação de Fragmentos de DNA

A purificação de fragmentos de PCR, extraídos dos géis de agarose que foram submetidos à eletroforese foi feita através do kit de purificação Illustra GFX PCR DNA and Gel Band Purification kit (GE Healthcare, código 27-9602-01), com posterior quantificação, como a realizada para as extrações de DNA genômico no item 3.8.2.

3.8.5 Ligação de Fragmentos de DNA

Os fragmentos originados em reações de amplificação por PCR foram ligados ao vetor pelo sistema pGEM®-T Easy, US, Vector System I (Promega, código A1360), utilizando-se as instruções do fabricante. Para a realização dos ensaios de ligação, foi utilizada a proporção inserto:vetor de 3:1 incubando por 12h a 16 °C. Após a ligação, foram feitos os procedimentos de transformação descritos abaixo e a seleção de colônias para a posterior extração do DNA plasmidial.

3.8.6 Transformação de Bactérias Químio Competentes

As células competentes preparadas no item 3.5.2 são colocadas em gelo e misturadas com 2uL de ligação (100-150 ng/uL), deixando incubar no gelo por 20 min. Posteriormente são levadas a 42 °C por dois minutos e de novo colocadas em gelo por 10 min. Depois é adicionado 800 uL de meio SOC (temperatura ambiente) e incubadas por 1 h. A 37 °C a 100 rpm. Finalmente, as células são colocadas no meio LB com o antibiótico de marcação.

3.8.7 Extração de DNA Plasmidial em Pequena Escala (Miniprep)

As bactérias contendo os clones de interesse foram colocadas para crescer um dia antes da extração de DNA. Em fluxo laminar, um inóculo de cada clone foi coletado com palito de madeira autoclavado, colocado em tubo de vidro para cultura de bactérias contendo 3 mL de meio Circle Grow com o antibiótico de seleção na concentração desejada. A cultura foi levada ao agitador de tubos (170 rpm) a 37 °C por um período de 22 h. Em seguida transferiu-se 1,4 mL das bactérias para tubo de 1,5 mL e centrifugou-se a 20.000g, por 60 s, em temperatura ambiente, para que precipitassem (Centrífuga Mikro 22 R, Rotor 1151 – *Hettich Zentrifugen*). Descartou-se o sobrenadante e adicionou-se o restante das bactérias, repetindo a centrifugação e descartando-se o sobrenadante completamente. Ressuspendeu-se as bactérias em 200 µL da solução P1 com auxílio de agitador de tubos (Vortex AP 56 – *Phoenix*). Adicionou-se 200 µL de P2 aos tubos, que foram invertidos 6 vezes gentilmente para homogeneizar seu conteúdo, e deixados a temperatura ambiente por 5 min. Adicionou-se 200 µL da solução P3 e inverteu-se os tubos gentilmente até que se formou um precipitado branco. Nesta etapa os tubos foram mantidos no gelo. Os tubos foram centrifugados em seguida, a 20.000 g por 15 min a 4 °C. O precipitado branco se separou da fase aquosa. O sobrenadante foi coletado com cuidado para não tocar no precipitado. Adicionou-se à fase aquosa 400 µL de isopropanol, e homogeneizou-se o conteúdo dos tubos por inversão e imediatamente os tubos foram centrifugados a 20.000g, à temperatura ambiente por 10 min. Descartou-se o sobrenadante e foi adicionado 500 µL de etanol 70%, seguido por inversão dos tubos. Estes foram centrifugados a 20.000 g, a 4 °C, por 5 min. e o etanol foi descartado em seguida. O DNA precipitado no fundo do tubo foi

seco completamente através de uma câmara de vácuo (*Centrivac Console* – *Labconco*), e ressuspendido em 30 µL de água esteril. Estas minipreparações foram então tratadas com RNase I-A a 50 µg/µL e incubadas por 1 h a 37 °C, seguido por armazenamento a -20 °C.

- **P1** - Tris 50 mM pH 8,0; EDTA 10mM; Rnase 100 g/mL.
- **P2** – 200 mM NaOH; 1% SDS.
- **P3** - KoAc 3 M pH 5,5 acertado com ácido acético glacial.

3.8.8 Extração de DNA Plasmidial em Pequena Escala (Miniprep) em Placas de 96 Poços

Em fluxo laminar foram utilizadas placas de 96 poços, próprias para o crescimento de bactérias, em cada poço da placa contendo o antibiótico desejado foi adicionado 1 mL de meio Circle Grow. Com um carimbo próprio para fazer réplicas de placas de 96 poços, devidamente flambado e resfriado, o inóculo das bactérias coletado das placas de estoque de glicerol foi depositado nas placas de cultivo (Mega Titer Plate). Estas foram seladas com adesivos (“Adhesive Plate Sealer” – AB Gene e “Adhesive Sealing Films for Microplates” – Simport). Com agulha flambada, um furo foi feito em cima de cada poço, para aeração das bactérias e as placas foram levadas a uma incubadora horizontal com agitação, a 37 °C, por um período de aproximadamente 22 hs a 300 rpm. Em seguida as placas foram centrifugadas por 6 min a 2.000 g (Centrífuga “Universal 30 F”, Rotor “1422” – *Hettich Zentrifugen*), os adesivos foram removidos e o sobrenadante descartado, vertendo-se as placas sobre papel absorvente por 5 min. Adicionou-se 240 µL de GTE em cada poço através de um micropipetador multicanal *Finnpipette*, em seguida as placas foram seladas com adesivos e levadas ao agitador de tubos por 2 min para ressuspender as bactérias precipitadas. Centrifugou-se as placas por 6 min a 2.000 g, e o sobrenadante foi descartado, vertendo-se as placas sobre papel absorvente por 5 min. Adicionou-se 80 µL de GTE em cada poço e as placas foram novamente seladas e levadas ao agitador de tubos para ressuspender o precipitado. Foram adicionados 2,5 µL de RNase a 10 mg/mL (RNase A – USB – fonte: pâncreas bovino) em cada poço das placas de polipropileno de 96 poços com fundo em “U” (Costar). Transferiu-se 60 µL da suspensão de células para cada poço das novas placas contendo RNase. Em seguida adicionou-se a cada poço 60 µL de uma

solução de NaOH 0,2N SDS 1% em água deionizada e as placas foram seladas com adesivos, invertidas por 10 vezes para homogeneização e deixadas à temperatura ambiente por 10 min. Adicionou-se então, a cada poço, 60 µL de uma solução de KoAc 3M gelada e novamente as placas foram seladas com adesivos, invertidas 10 vezes e deixadas à temperatura ambiente por 10 min. Os adesivos foram removidos, as placas foram levadas à estufa a 90 °C por 30 min e colocadas no gelo por 10 min. As placas foram centrifugadas por 4 min a 2.000 g a temperatura ambiente e todo o volume de cada poço (180 µL), foi transferido para placas de 96 poços (“MultiScreen Filter Plate” (cód. MAGVN 22-5) – Millipore), previamente fixadas no topo de placas de polipropileno de 96 poços com fundo em “V” (Costar), de maneira que os poços das placas ficassem alinhados. O conjunto foi centrifugado e o precipitado ficou retido nas placas Millipore, enquanto a fase aquosa contendo o DNA foi filtrada para as placas com fundo em “V”. Adicionou-se, então, 110 µL de isopropanol 100% gelado a cada poço contendo o filtrado. As placas foram seladas com adesivos, invertidas por 15 vezes e centrifugadas a 2.000 g por 45 min à temperatura ambiente. O sobrenadante foi descartado, invertendo-se as placas sobre papel absorvente. Seguido pela adição de 200 µL de etanol 70% gelado a cada poço. As placas foram centrifugadas a 2.000 g por 10 min à temperatura ambiente e vertidas sobre papel absorvente para a remoção do sobrenadante. As placas foram secas em estufa a 37 °C por 30 min e o DNA precipitado no fundo das placas foi ressuscitado em 80 µL de água deionizada. A manipulação de bactérias foi realizada em fluxo laminar.

- **GTE** – 23 mL de Glicose Anidra 20%; 10 mL de EDTA 0,5M pH8,0; 13 mL de Tris-HCl 1M pH 7,0; água deionizada para completar um volume de 500 mL.

3.8.9 Extração de BACs dos Clones da Biblioteca

Os clones de BAC são mantidos em bactérias estocadas em solução de glicerol 50% a -70° C. Inicialmente foi necessário estriar os clones em placas de Petri com meio LB solido, ao qual foi colocado o antibiótico de seleção cloranfenicol (12,5 mg/L), para deixá-los em crescimento em um período de 12 a 16 h (37 °C) a fim de se obter colônias isoladas. Com um palito foi selecionada uma colônia isolada, a qual foi colocada para crescer em 2 mL de meio LB contendo cloranfenicol a 12,5 mg/L durante 4-6 h (pré-inóculo), seguido pela transferência de 1 mL do pré-

inóculo para 500 mL de LB com o mesmo antibiótico e incubado por 12 h. A extração foi feita com o *QIAGEN Large-Construct Kit* (Cat.No. 12462) segundo o protocolo do fabricante.

3.8.10. Sequenciamento Automático de Nucleotídeos

Em todos os procedimentos de sequenciamento, as amostras de DNA utilizadas foram provenientes de extrações de DNA plasmídial em pequena escala, descrita anteriormente no item 3.8.8. Os princípios do sequenciamento utilizado é o descrito por Sanger *et al.* (1977). Cada reação utiliza: 100-200 ng de DNA, 5 µM do respectivo iniciador, 2 µL de BigDye™ Terminator V3.1 Cycle Sequencing Ready Kit (Applied Biosystem código 4336919), e água deionizada para completar um volume final de 10 µL.

A reações de sequenciamento foram levados para o termociclador “PTC 100TM Programmable Thermal Controller” (MJ Research, Inc.), passando pelo ciclo que consistiu de: 3 min a 96 °C, 25 vezes a repetição da seguinte sequência: 96 °C por 10 s, 48 °C por 20 s, 60 °C por 4 min O DNA marcado com as moléculas fluorescentes foi então precipitado através da adição em cada reação de 100 µL de isopropanol 65% seguido pela agitação dos tubos ou placas em agitador. Os mesmos foram deixados por 20 min em temperatura ambiente ao abrigo da luz e em seguida centrifugados a 2.000 g por 50 min a temperatura ambiente. O sobrenadante foi descartado vertendo-se os tubos ou placas sobre papel absorvente e em seguida foi realizada uma lavagem do precipitado com 150 µL de etanol 60%, e centrifugação a 2000 g por 10 min em temperatura ambiente. O sobrenadante foi descartado vertendo-se os tubos ou placas sobre papel absorvente e os mesmos foram levados à estufa a 37 °C por 15 min, para que o etanol evaporasse completamente. Os DNAs marcados foram ressuspensos em 10 µL de formamida (“Hi-Di Formamide” P/N 4311320 – Perkin Elmer Biosystems), desnaturados (4 min a 96 °C), e deixados no gelo por 10 min. Em seguida, as amostras foram lidas no sequenciador “3700 Genetic Analyser” ou “3100 Genetic Analyser” - ABI PRISM. Todas as sequências analisadas passaram por pelo menos duas leituras e possuem qualidade consenso de Phred (Ewing *et al.*, 1998) de no mínimo 20.

3.8.11 Sequenciamento em Alta Escala de BACs

Todos os BACs analisados neste trabalho foram sequenciados usando inicialmente a estratégia de pirosequenciamento. Os clones da biblioteca de R570 (ver 3.5) foram comprados da CUGI (*Clemson University Genomics Institute*), os BACs foram extraídos como foi descrito em 3.9.3, e posteriormente sequenciados.

Este material foi usado em dois sequenciamentos; o primeiro feito no centro da *Roche Applied Science* em Penzberg, Alemanha, sendo uma colaboração com esta empresa. O segundo foi realizado no CATG (Centro Avançado de Tecnologias em Genômica) da USP, São Paulo, sendo realizado pelo GaTE_Lab com a colaboração do Dr. Carlos Takeshi Hotta e o apoio econômico da FAPESP.

Ambos sequenciamentos foram feitos no sistema *Genome Sequencer FLX*. A metodologia usada foi a de sequenciamento por síntese (pirosequenciamento) descrita por Margulies et al. (2005), Velasco et al. (2007) e Droege e Hill (2008). O protocolo de sequenciamento 454 utilizando o sistema *Genome Sequencer FLX*, requer a geração de uma biblioteca de fragmentos de fita simples de DNA de 500 a 600 pb obtidos por nebulização os quais são posteriormente ligados a adaptadores comerciais.

Esta biblioteca é então utilizada para a geração de *beads* de sequenciamento que serão empregados em uma PCR em emulsão. O perfil de sequenciamento por síntese a partir desse ponto, consiste num fluxo sequencial de cada nucleotídeo, em um padrão repetido de TACG por 200 vezes.

O sistema FLX possui uma série de programas padrão que reconhecem a qualidade das sequências (de 100 a 200 pb) e convertem os sinais da emissão de luz do pirosequenciamento para parâmetros análogos à qualidade usada no sequenciamento tipo Sanger.

A *Roche Applied Science* e o CATG forneceram as sequência brutas das leituras e sequências-consenso (*contigs*), com seus respectivos valores de qualidade. No GaTE_Lab essas sequências foram importadas para o pacote Phred/Phrap/Consed (Ewing et al., 1998; Ewing e Green, 1998; Gordon et al., 1998) onde foi possível avaliar a qualidade da montagem e resolver de maneira manual as leituras mal-alinhadas no conjunto de sequências.

3.9 Procedimentos Gerais de Manipulação de RNA

3.9.1 Extração de RNA Total

Foi extraído RNA de calo, folhas e raízes das variedades SP80-3280 e R570. Os meristemas apicais das variedades RB75-126 e SP81-3250 foram cedidos pela Profa. Dra. Kátia Castanho Scortecci (Universidade Federal do Rio Grande do Norte). Colmos da variedade SP80-3280 e R570 foram cultivados em casa de vegetação usando vasos com terra e vermiculita (1:1) e posteriormente, foram transferidos para terra e mantidas em condições naturais. Foi feita a coleta de todos os tecidos após 3 e 6 meses de semeados, com o intuito de realizar uma análise de expressão dos elementos estudados neste trabalho a longo dos diferentes tecidos da planta em dois momentos de desenvolvimento. Os meristemas apicais cedidos pela Profa. Dra. Kátia correspondem a duas variedades: RB75-126 e SP81-3250, as quais são conhecidas por nunca florescer e florescer precocemente, respectivamente. Estas plantas foram acompanhadas por 6 meses de desenvolvimento em condições naturais de cultura. Isto, com o intuito de realizar uma análise de expressão dos elementos estudados neste trabalho ao longo de um período de tempo sobre o mesmo tecido da planta, neste caso o meristema apical.

O material vegetal após a coleta foi imediatamente limpo, congelado em nitrogênio líquido e acondicionado em tubos falcon de 50 mL os quais foram mantidos em freezer a -70 °C até sua utilização na extração. Posteriormente, o material foi macerado em nitrogênio líquido até a obtenção de um pó fino. Para cada extração de RNA foi utilizado 1 g do material macerado (aproximadamente 5 mL do volume do tubo de 50 mL).

Ao conteúdo macerado foi adicionado 1 mL de trizol ou tri-reagente para cada 100-200 mg de material vegetal. A mistura foi ressuspendida de maneira rápida e logo congelada em nitrogênio líquido para aumentar a lise celular. A mistura foi descongelada e incubada a temperatura ambiente por 5 min e 200 uL de clorofórmio foram adicionados. Com ajuda de um vortex a solução foi misturada por 15 s e incubada a temperatura ambiente por 10 min seguido por centrifugação a 12000 g por 15 min. O sobrenadante foi transferido para novo tubo falcon com 500 uL de isopropanol por 1ml de trizol utilizado e incubado a -20 °C por 12 h, sendo centrifugado a 12000 g x 10 min a 4 °C. O sobrenadante foi descartado e o

precipitado lavado com álcool 70% duas vezes. Os tubos permaneceram secando à temperatura ambiente e logo em seguida o conteúdo foi ressuspendido em 110 μ L de água DEPC autoclavada. As amostras foram alíquotadas em 10 μ L para a quantificação e preparo de gel e os outros 100 μ L foram guardados a -70 °C.

- **Água DEPC** –0,01% DEPC em água deionizada. Esta solução é mantida por 16 h a 37 °C. Posteriormente foi autoclavada por 20 min a 120 °C.

3.9.2 Quantificação de RNA

A quantificação do RNA total usado em diferentes experimentos foi realizada com o Spectrophotometer Nanodrop®, US, ND-100, segundo especificações do fabricante.

3.9.3 Tratamento do RNA Total para a Remoção de DNA Contaminante

O RNA total extraído foi tratado com o kit TURBO DNA-free (Ambion, código AM1907) antes da síntese da primeira fita de cDNA e utilização em reações de RT-PCR, seguindo as recomendações do fabricante. No final do tratamento as amostras foram novamente quantificadas.

3.9.4 Síntese da Primeira Fita de cDNA

A síntese da primeira fita de cDNA foi feita a partir de 10 μ g de RNA total descontaminado do DNA genômico, como explicado no item 3.9.3. Foi utilizado o kit High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, código 4368814) para a síntese da primeira fita, de acordo com as especificações do fabricante. Para inibir ribonucleases foram utilizadas 40 unidades de RNaseOUT™ Recombinant Ribonuclease Inhibitor (Invitrogen, código 10777-019). No final do processo, as amostras foram quantificadas mais uma vez.

3.9.5 RT-PCR

Para as reações de RT-PCR foi utilizado 1 μ L de primeira fita de cDNA (feito com 2 μ g de RNA purificado), água destilada, ultrafiltrada e autoclavada, tampão de

PCR 1X, 1,5 mM MgCl₂, 0,2 mM de dNTPs (dATP, dCTP, dGTP e dTTP), 0,2 μM dos indicadores e 1,25 unidades de Taq Polymerase Recombinante (Invitrogen Brasil código 11615-010), para uma reação com volume final de 50 μL.

Foram feitos dois controles negativos para garantir a ausência de DNA genômico, assim como de qualquer tipo de contaminação: um contendo amostra de cDNA sintetizado sem transcriptase reversa e posteriormente tratada para a remoção do DNA; e o outro controle que não continha cDNA. As amplificações foram realizadas em um termociclador PTC-100 (MJ Research, Inc.) como descrito na figura 9. Os produtos das amplificações foram posteriormente submetidos a eletroforese em gel de agarose 0,8% e usando tampão de corrida TBE 0,5X para a separação e observação dos fragmentos. Os fragmentos amplificados foram posteriormente sequenciados.

3.10 Procedimentos Gerais de Manipulação de Proteínas

3.10.1 Plasmídeos Usados para a Expressão da Transposase

Foram usados dois plasmídeos para expressar a transposase codificada pelo transcrito TE257: o plasmídeo pET15b para expressão na linhagem BL21(DE3) de *E. coli*, e os plasmídeos pGAL416GAL1 (promotor forte) e pGAL416GALS (promotor fraco) para expressão em *Saccharomyces cerevisiae* linhagem BY4723 (MATa, his3Δ0, ura3Δ0). Todos cedidos pelo professor Dr Reinhard Kunze.

A ORF do TE257 possui todos os domínios conservados necessários para sua função, além de não possuir códons de parada prematuros, porém não possui códon de início de tradução. Por esta razão foi adicionado um códon de início de tradução no marco de leitura correto para produzir uma proteína fusionada à cauda de histidina presente no vetor pET15b, o qual já possui códon de início de tradução próprio, este vetor foi chamado de pET257. Esta construção foi mantida neste vetor para ser expressa em *E. coli*, e posteriormente a parte correspondente à transposases fusionada com a calda de histidina foi transferida para os vetores pGAL416GAL1 e pGAL416GALS para a expressão em *Saccharomyces cerevisiae*.

Para a expressão em *Saccharomyces cerevisiae* foram usadas duas estratégias diferentes, em uma delas foi adicionada aos vetores a sequência NLS

(*Nuclear localization signal*) e na outra estratégias esta sequência não foi adicionada (Quadro 9).

3.10.2 Expressão e extração da Transposase

3.10.2.1 Expressão e extração em *E. coli*

A linhagem de *E. coli* BL21(DE3) transformada com o plasmídeo pET257 foi crescida em 50 mL de meio de cultura LB com ampicilina, como antibiótico de seleção, até chegar na fase exponencial (OD_{600} 0,6-0,8) a 37 °C em agitação de 210rpm. Uma vez na fase exponencial, a cultura foi colocada no gelo por 5 min e em seguida foi adicionado 1mM de IPTG para induzir a expressão e a cultura foi deixada a 20 °C crescendo por 16 h a 210 rpm.

Neste processo foram coletadas cinco amostras. A primeira amostra foi da cultura antes de ser induzida com o IPTG, as outras três após a indução. A segunda amostra foi do meio de cultura após a separação da fase celular. A terceira amostra foi da fase celular total. Finalmente, a quarta amostra foi tomada após a extração do extrato protéico solúvel e a quinta amostra corresponde ao extrato protéico insolúvel.

Para a extração do extrato protéico solúvel a cultura foi centrifugada a 3000 g por 15 min a 4 °C. O sobrenadante foi descartado (Amostra 2) e o conteúdo celular foi precipitado e resuspendido com a solução US2 (10 mL por grama de células). A seguir foi adicionado 1/100 volume de lisozima (100 mg/ml) e incubado por 30 min a temperatura ambiente. Posteriormente, foi adicionado 1/20 volume de triton X-100 e incubado por 5 min a temperatura ambiente seguida pela adição de 25U/ml de benzonase na amostra para posterior quebra por sonicação em banho de gelo por 10 min. Uma vez centrifugado por 15 min a 20000 g a 4 °C foi coletado o sobrenadante (fase solúvel, amostra 4) e o precipitado celular (fase insolúvel, amostra 5) e guardados a -20 °C.

- **US2** – 50 mM TrisHCl pH 8.5, 20 mM NaCl, 2 mM $MgCl_2$, 5 mM DTT (Ditiotreitol) e 1 mM de PMSF (Phenylmethylsulfonylfluorid). Adicionar os dois últimos na hora de usar a solução. Também foi testada esta solução adicionando o conjunto de inibidores de proteases comercial *Proteoblock protease inhibitor cocktail* 100X (Fermentas) para reforçar a atividade do
-

PMSF. Não foi observado diferenças nos resultado usando o *Proteoblock protease inhibitor cocktail*.

3.10.2.2 Expressão e extração em *Saccharomyces cerevisiae*

A linhagem de *Saccharomyces cerevisiae* BY4723 (MATa, his3 Δ 0, ura3 Δ 0) foi transformada usando o protocolo de choque térmico para bactérias quimio competentes e colocada para crescer em 10 mL de meio sintético mínimo SD-ura suplementado com todos os aminoácidos. Foram usadas duas fontes de carbono glicose (inibe a produção da proteína) para os controles negativos e galactose para induzir a produção da proteína. As culturas foram incubadas por 16 h a 30 °C a 220 rpm. Para a extração da fração solúvel e insolúvel protéica foi usado o regente Y-PER (*Thermo Scientific*), segundo indicações do fabricante. Amostras foram guardadas a - 20 °C.

3.10.3 Eletroforese em Gel de Poliacrilamida

Para realizar a separação das proteínas por massa molecular aparente, as amostras foram submetidas a uma eletroforese em gel de poliacrilamida. Esse gel é composto por duas partes. A parte inferior do gel, chamado gel de corrida contém Tris 0,375 M (pH 8,8), SDS 0,1%, acrilamida 12%, Persulfato de Amônio (APS) 0,03% e ,N,N,N tetra metilenodiamina (TEMED, Sigma); e a parte superior, chamado gel de empilhamento contém, Tris 0,125 M (pH 6,8), SDS 0,1%, acrilamida 4%, APS 0,045%, e TEMED. O gel é submetido inicialmente a uma diferença de potencial constante de 50 V até que as amostras atinjam o gel de corrida e em seguida, a voltagem pode ou não ser elevada até 80 V. A eletroforese das amostras ocorre completamente submersa em tampão de corrida.

Após a corrida o gel foi corado com Coomassie blue G ou as proteínas foram transferidas para uma membrana de PVDF.

Para a coloração com Coomassie blue G, o gel foi colocado em solução fixadora (50% etanol e 10% ácido acético) por 1 hora sob agitação à temperatura ambiente. Em seguida, o gel foi colocado em solução de Coomassie G-250 (0,1% Coomassie G-250, 2% ácido fosfórico, 10% sulfato de amônia e 20% metanol) e

incubado sob agitação à temperatura ambiente por 16 horas. No dia seguinte, o gel foi colocado em solução 25% metanol sob agitação para descoloração do mesmo.

Após a descoloração o gel foi seco, guardado e fotografado.

3.10.4 Western Blot

O processo de transferência para membrana de PVDF foi feito sob voltagem constante de 100 V por uma hora, imerso em tampão de transferência (25 mM Tris, 192 mM Glicina e 20% metanol).

Após a transferência, a membrana de PVDF foi incubada com TBS 1x, 0,05% Tween 20, 5% leite em pó. Essa incubação tem como objetivo bloquear os sítios inespecíficos na membrana.

Após o bloqueio, a membrana foi incubada com o anticorpo anti-histidina por uma hora, em seguida foram feitas três lavagens de 5 min com TBS-T à temperatura ambiente e uma incubação com o anticorpo secundário marcado com HRP por mais uma hora.

Finalmente, foram feitas três lavagens com TBS-T e membrana foi revelada com o kit de revelação para ECL (Amersham).

3.11 Hibridação Molecular

3.11.1 Sondas

Foram realizados ensaios de hibridação molecular com DNA genômico e RNA, utilizando-se diferentes sondas. Nos ensaios foram utilizados como sondas regiões conservadas de cDNAs representantes da linhagem I (TE074), linhagem II (TE191) e linhagem III (TE257) da família SChAT. Conforme o descrito em de Jesus, (2007) e apresentadas no quadro 3. As regiões correspondentes às sondas são específicas dos TEs de cada linhagem, isto para garantir uma hibridação estridente.

Os produtos das amplificações foram migrados em gel de agarose e as bandas de interesse foram isoladas do gel como explicado anteriormente em 3.9. Para as hibridações foram usados 50 ng de sonda purificada.

Quadro 3–Sequências dos iniciadores utilizados para amplificar as sondas.

Nome do iniciador	Sequência do iniciador	Região de anelamento	Tamanho do cDNA molde
SHAT-2-F	5' CCACGTTTCAAGATGAAGC 3'	280	1022
SHAT-2-R	5' CCCTTGACTGTTATTTCGCA 3'	442	1022
SHAT-3-F	5' GGAGAGTATGGAAGTGTCCC 3'	1074	2195
SHAT-4-R	5' CCTATCATACTCGCTGTTTTCT 3'	1261	2195
SHAT-5-F	5' GAGAATGCAGAAGCGGAA 3'	1377	2445
SHAT-6-R	5' CCACGAAGTCCAGAAGAA 3'	1523	2445

Fonte: Ochoa (2012).

3.11.2 Hidrólise do DNA Genômico

Para os ensaios de hibridação, os DNAs genômicos foram hidrolisados com a endonuclease Dral (New England Biolabs). As reações foram preparadas como descrito a seguir: 15 µg de cada DNA, endonuclease (10 U/µg de DNA), tampão da enzima para 1X e água deionizada para completar um volume final de 30 µL. As reações de hidrólise ficaram incubadas a 37 °C por um período de 12 h.

3.11.3 Gel de Agarose

Os DNAs após hidrólise foram migradas em gel de agarose 0,8% em TBE 0,5X. Realizou-se a corrida por eletroforese a 45 V, 20 mA, por 12 h. Após a eletroforese, os géis foram corados em uma solução de brometo de etídeo a 5 µg/mL, visualizados no “Image Master®VDS” (Pharmacia Biotech), e fotografados no “Thermal Imaging System FTI-500” (Fuji Film).

3.11.4 Hibridação Molecular de DNA

3.11.4.1 *Transferência Alcalina*

Após fotografado, o gel foi submerso por 15 min em solução de ácido clorídrico 0,25 N, passando por água destilada para enxágue e submerso por 30 min em NaOH 0,4 N. O DNA foi transferido do gel para a membrana de náilon por capilaridade. A transferência foi preparada colocando-se o gel sobre papéis de filtro,

em contato com o tampão de transferência (NaOH 0,4 N) e sobre o gel foi colocada a membrana de náilon (“Gene Screen Plus Hybridization Transfer Membrane” – NENTM Life Science), finalmente sobre o gel são colocados mais dois papéis de filtro e toalhas de papel.

A transferência transcorreu por um período de 16 h. Após transferência, a membrana foi submersa em uma solução de 2X SSC por 5 min, colocada em lâmpada UV por 1 min e seca por 2 h a 80 °C.

SSC – 150mM NaCl; 15mM de citrato de sódio.

3.11.4.2.1 Pré-Hibridação para Membranas de DNA Genômico

As membranas foram colocadas em garrafas próprias para hibridação, contendo 50 µL da solução de pré-hibridação por cm² de membrana e levadas ao forno a 65 °C, com rotação por 7 h.

- **Solução de pré-hibridação** - 10% dextran sulfato; 1% SDS, 6% NaCl. Dissolver em água deionizada.

3.11.4.2.2 Pré-Hibridação para Membranas da Biblioteca Comercial de BACs de Cana-de-açúcar

As membranas da biblioteca foram colocadas em garrafas próprias para hibridação, contendo 20 mL da solução de pré-hibridação e levadas ao forno a 65°C, com rotação por 4 h.

- **Solução de pré-hibridação** - 50% NaPO₄ 1 M; 7% SDS; 0,2% EDTA 0,5 M. Dissolver em água deionizada.

3.11.4.3 Marcação das Sondas

As sondas foram marcadas com P α³² dATP, utilizando-se o “Kit Random Primers DNA Labelling System” da Invitrogen, de acordo com as instruções do fabricante. Adicionou-se à sonda 100 µL de DNA de esperma de salmão para cada 10 mL da solução de pré-hibridação. A desnaturação desta mistura foi feita deixando-se por 5 min à 92 °C e colocando-se no gelo por 2 min.

3.11.4.4 Hibridação

As sondas, devidamente marcadas, foram colocadas nas garrafas de hibridização com as membranas e a solução de pré-hibridação. Estas retornaram para o forno a 65 °C, com rotação, por 36 h.

3.11.4.5 Lavagens das Membranas

Após o período de hibridação as membranas passaram por três lavagens, duas com a solução 1 e uma com a solução 2. Foram adicionados 50 mL de solução por garrafa e as lavagens com a solução 1 foram feitas à temperatura ambiente, por 5 min com rotação. A lavagem com a solução 2 foi de 30 min no forno a 65 °C com rotação e em seguida as lavagens as membranas foram enxaguadas em cuba contendo 0,1X SSC e secas com toalhas de papel. Para a desibridização (e reutilização), as membranas foram submersas em uma cuba contendo SDS 0,5% pré-aquecido até o ponto de ebulição e levadas ao forno a 65 °C por 30 min. Este procedimento foi repetido. As membranas foram enxaguadas com água destilada e secas entre toalhas de papel sobre a bancada, ficando prontas para uma nova hibridação.

- **Solução 1** - 2X SSC em água deionizada.
- **Solução 2** – 2X SSC; 1% SDS. Dissolver em água deionizada.

3.11.4.6 Exposição e Revelação

As membranas foram envolvidas com filme plástico e colocadas em cassetes de exposição, juntamente com um intensificador e uma chapa de filme de raio X (Kodak), ficando mantidas a -70 °C. Esta chapa de filme ficou exposta à membrana pelo tempo necessário para a sua impressão. A revelação do filme foi feita em tanque de revelação em sala escura, utilizando-se revelador e fixador Kodak, para isso, o filme foi mantido no revelador por 3 min, na água por 2 min para enxágue, no fixador por 5 min e novamente na água por dois min. O filme foi finalmente enxaguado em água corrente e colocado para secar.

3.11.5 Hibridação Molecular de RNA

As manipulações de RNA requerem um tratamento diferenciado de todo o material a ser utilizado, para a eliminação das RNases. Toda a vidraria, assim como cadinhos, bastões e espátulas foram lavados com água deionizada e esterilizados por autoclavagem de 20 min a 120 °C. Cubas de eletroforese, pentes e suportes para o preparo dos géis devem ser tratados por 4 h com detergente e enxaguados com água deionizada. Toda a água deionizada empregada no preparo das soluções utilizadas no procedimento foi tratada com água DEPC e todo o procedimento foi feito com luvas.

- **Água DEPC** – Foi adicionado 0,01% DEPC na água deionizada e esta foi mantida a 37 °C por 16 h. Após este período a água DEPC foi autoclavada por 20 min a 120 °C, este tratamento inativa RNases presentes na água.

3.11.5.1 Gel de Formaldeído

Os RNAs foram migrados em gel desnaturante de agarose, constituído de 1,5% agarose; 5,4% Formaldeído 37%; MOPS 1X e água deionizada. A agarose foi adicionada à água e levada ao microondas para dissolver. Após esfriar para 60°C, o MOPS 10 X e o formaldeído 37% foram adicionados ao gel para que ficassem nas concentrações mencionadas acima.

O gel ficou polimerizando dentro de uma capela com exaustão, para evitar que vapores do formaldeído saíssem para o ambiente e desta forma, a eletroforese foi realizada em tampão MOPS 1 X. As amostras a serem aplicadas no gel constituíram-se de 15 µg de RNA em 10 µL, acrescidos de 4µL de MOPS 10 X; 8,65µL de formaldeído 37%, 14 µL de formamida e 0,5 µL brometo de etídeo a 4µg/mL. Estas permaneceram incubadas por 10 min a 65 °C e colocadas 2 min no gelo e em seguida aplicadas no gel e realizou-se uma eletroforese a 80 V por 3 a 4 h.

- **MOPS 10X** – MOPS pH 7,0 (0,4 M); Acetato de sódio pH 7,0 (0,1 M) e EDTA (10mM).
-

3.11.5.2 Transferência e Hibridação de RNA

A transferência foi feita em SSC 10X por capilaridade e foi montada colocando-se o gel sob papel filtro em contato com o SSC 10X da cuba de transferência. Uma membrana de náilon (*Gene Screen Plus Hybridization Transfer Membrane – NENTM Life Science*), previamente umedecida com água deionizada e mantida por 10 min em SSC 10 X foi depositada sobre o gel e coberta com papel filtro umedecido com SSC 10 X e toalhas de papel. A transferência ocorreu por um período de 18 h. Após este período a membrana foi lavada rapidamente em SSC 2X e levada ao forno a 80 °C por 2 h e seguido por exposição à luz UV por 1 minuto, com a face contendo os RNAs para cima para permitir sua fixação.

Foi feita hibridação com a sonda amplificada do ribossomal 25S como controle positivo da membrana.

Os passos de marcação das sondas, pré-hibridação, hibridação, lavagem das membranas, exposição e revelação foram feitos de maneira igual aos descritos para o DNA no itens 3.11.4.2.2, 3.11.4.3, 3.11.4.4, 3.11.4.5 e 3.11.4.6.

3.12 Caracterização dos Elementos usando Ferramentas Bioinformáticas

Para analisar a diversidade da superfamília SChAT, foi feita uma comparação entre as sequências dos elementos recuperados dos BACs de cana-de-açúcar, assim como foi realizada a procura de elementos relacionados em outros genomas de gramíneas (Paterson et al., 2009; Schnable et al., 2009; Yu et al., 2002).

3.12.1 Identificação dos Elementos SChAT nos BACs Sequenciados

As sequências de cada BAC foram alinhadas para construir *contigs* (sequência consenso de um conjunto de fragmentos que possuem sobreposição), para isto foi usada a combinação dos *contigs* obtidos usando o pacote *Phred*, *Phrap*, *Consed* e o os *contigs* obtidos usando o programa *Newbler*. Estas ferramentas eliminam automaticamente as sequências com baixa qualidade e que apresentam similaridade com DNA genômico de *E. coli*, resultante da contaminação do processo de replicação dos BACs no microorganismo.

Foi feita a procura nos *contigs*, assim como nas sequências que não entraram nos *contigs*, por sequências relacionadas com as sondas das três linhagens estudadas nos 25 BACs selecionados por hibridização molecular para estas mesmas sondas.

Quando foram encontradas sequências com similaridade às sondas que não entraram nos *contigs*, estas sequências foram incluídas manualmente dentro dos *contigs*, garantindo assim, que a totalidade da informação sequenciada referente aos elementos estivesse nos *contigs*. Para realizar a procura foi usada a ferramenta *Crossmatch* (*minmatch*: 8, *minscore*: 15).

Foi feita uma segunda procura usando o algoritmo *hmmer* 3.0 (Finn R.D. et al., 2012) usando os modelos *hmm* construídos para o domínio de dimerização (Pfam PF05699) e o domínio *Zinc Finger* (Pfam PF02892), que estão nos extremos da transposase dos elementos desta superfamília. No total foram usados 270 BACs de cana-de-açúcar sequenciados pelo GaTE-Lab para esta busca.

3.12.2 Análise Comparativa dos Elementos SChAT

As sequências nucleotídicas dos elementos encontrados foram comparadas usando as ferramentas de alinhamento *MUSCLE* (Edgar, 2004). Para os elementos completos foram preditas ORFs usando a ferramenta *ORF Finder* e a comparação com a transposase do elemento de referência, Ac do milho. Foram feitas previsões de sequências promotoras usando a ferramenta *PLACE* (*A Database of Plant Cis-acting Regulatory DNA Elements*) e finalmente, foram preditas as TIRs usando o programa *TIR_FINDER* (Rubin et al., 2001) ajustado aos parâmetros da superfamília hAT descritos por Kempken F. e Windhofer F., 2001; TSDs de 7-8 pb, TIRs imperfeitas de 5-27 pb, elementos com tamanho de 5 kb até 12 kb.

Tendo uma melhor idéia da estrutura dos elementos, foi feita a comparação das transposases preditas, através da análise dos segmentos relativos a dois domínios protéicos característicos da superfamília hAT: o domínio de dimerização e o domínio *Zinc Finger*, conhecido por sua função de ligação ao DNA. Também foram comparadas outras regiões conservadas de aminoácidos que são importantes para o funcionamento desta proteína descritas por Rubin et al. (2001).

A caracterização de três destes elementos foi incluída no manuscrito de De Jesus et al. (2012).

3.12.3 Procura de Elementos Relacionados à família SChAT em Outros Genomas

Usando a sequência completa de cada elemento como isca, foram procurados elementos relacionados à família SChAT nos genomas de *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* e *Zea Mays*. Estes organismos estão completamente sequenciados e depositados nas bases de dados do NCBI.

Para realizar a procura foi utilizado o algoritmo de alinhamento BLASTN (*Matrix: blastn matrix, Gap Penalties: Existence: 5, Extension: 2*) da ferramenta CoGeBLAST. Foram escolhidos alinhamentos com pelo menos 10% de cobertura sobre a isca e estes foram analisados manualmente. Aumentando 10 Kb para cada lado do alinhamento inicial estes elementos foram isolados e suas transposases comparadas. Também foi feita procura, usando o algoritmo de alinhamento BLASTN, no genoma de *Miscanthus giganteus* (Swaminathan K. et al., 2010), gramínea altamente relacionada com cana e sorgo que teve sequenciado um 1,2% do seu genoma

3.12.4 Geração de Dendogramas a partir de Alinhamentos

Os dendogramas apresentados neste trabalho foram gerados a partir dos diversos alinhamentos descritos através do programa Mega 5 (Kumar et al, 2004). As árvores foram geradas através do algoritmo de *Neighbor joining* (Saitou e Nei, 1987), com teste de *bootstrap* de 1000 réplicas e a opção *Pairwise deletion*.

3.13 Identificação de Regiões de sintenia em Genomas Relacionados

Para caracterizar alguns TEs dos BACs e obter a porcentagem de elementos repetitivos de cada BAC as sequências repetitivas dos BACs foram identificadas usando a ferramenta *RepeatMasker* com três bases de dados de elementos de transposição: *GaTEs_DataBase* (Base de elementos de cana-de-açúcar do *GaTE_Lab*), *TIGR_Gramineae* e *Repeats.v3.3_repbases_viridiplantae*.

Uma vez eliminadas as sequências repetitivas, os genes de cada BAC foram identificados por alinhamento dos BACs com dois genomas anotados altamente relacionados a cana, o genoma do arroz e do sorgo. O alinhamento foi feito por meio

de um *pipeline* criado pelo grupo de bioinformática do GaTE_Lab, usando BLASTN e BLASTX, o critério de seleção do alinhamento foi $e\text{-value} < 1e-10$.

Este *pipeline* também agrupa outras fontes de informação importantes para a identificação de genes como são: os transcritos de cana sequenciados no projeto SUCEST (Vettore et al., 2001) e os domínios protéicos da base de dados Pfam.

A ferramenta Gbrowser implementada para visualizar o resultado do *pipeline*, foi usada para a seleção de regiões correspondentes a genes em cada BAC. Estas regiões eram avaliadas segundo a presença e alinhamento com transcritos em cana, sorgo e arroz.

Uma vez que os genes de cada BAC foram selecionados, estes eram usados para procurar regiões sintênicas usando CoGeBlast e os diferentes genomas de interesse.

4 RESULTADOS E DISCUSSÃO

A complexidade do genoma da cana-de-açúcar dificulta seu sequenciamento assim como o entendimento de sua estrutura genômica. Para este projeto optou-se pela estratégia de sequenciamento de BACs para recuperar regiões genômicas contendo os elementos da família SChAT e assim complementar estudos prévios realizados a partir de cDNAs.

O conjunto do trabalho realizado está organizado em etapas que visam identificar versões genômicas associadas às linhagens previamente descritas em De Jesus (2007). Parte do trabalho realizado já encontra-se publicado em Jesus et al. (2012) (vide anexo B). Três estratégias de identificação de cópias genômicas de elementos hAT foram utilizadas, a saber: (a) hibridação de membrana de BAC; (b) PCR em pool 3D; e (c) varredura procurando domínios conservados das transposases hAT em uma coleção de BACs sequenciados.

A seleção de regiões genômicas de cana-de-açúcar (variedade R570) para sequenciamento foi feita através da hibridação de biblioteca de BACs com sondas específicas para cada um dos três grupos de elementos. As sondas utilizadas estão descritas em De Jesus et al. (2012). Além da seleção inicial por hibridação, foi realizada varredura total da biblioteca de BACs com iniciadores específicos para cada um dos três grupos de elementos (Quadro 3) e os BACs positivos para elementos da linhagem 074 foram sequenciados. A partir de (a) foram selecionados 16 clones BAC sequenciamento e (b) outros 8 clones BAC. Em (c) foi feita uma varredura por similaridade de sequência em 270 BACs da base de dados do BIOEN usando o algoritmo hmmer 3.0 (Finn R.D. et al., 2011) com os modelos hmm construídos para o domínio de dimerização (Pfam PF05699) e o domínio *Zinc Finger* (Pfam PF02892). Desta análise outros 6 clones de BAC foram identificados.

A recuperação das versões genômicas permitiu explorar a estrutura de elementos das três linhagens quanto o contexto genômico, sintenia das regiões com outras gramíneas, e estabelecer relações filogenéticas com TEs relacionados à superfamília hAT presentes neste e em outros genomas sequenciados (Paterson et al., 2009; Schnable et al., 2009; Yu et al., 2002). Também permitiu delimitar melhor os agrupamentos da família SChAT, assim como iniciar análises funcionais dos elementos relacionados à superfamília hAT encontrados no genoma da cana-de-açúcar.

4.1 A família SChAT de cana de açúcar

O universo total explorado neste projeto inicialmente foi de 30 clones de BAC que contém transposases semelhantes a hAT. No entanto, apenas 17 clones foram considerados para efeito de descrição detalhada uma vez que os demais apresentam a transposase incompleta seja por estar na extremidade do clone BAC ou porque o clone de BAC não encontra-se com a sequência consenso fechada.

4.1.1 Identificando as primeiras linhagens (TE074; TE191 e TE257)

Com base em estudos anteriores (De Jesus, 2007) três cDNAs do SUCEST foram selecionados para o presente trabalho. Estes são denominados TE074, TE191 e TE257. A partir de alinhamentos protéicos e relações filogenéticas, determinamos que se tratam de três linhagens evolutivas (De Jesus et al., 2012) cada uma com características específicas.

A figura 8 ilustra o perfil genômico de hibridização molecular em (A) com a sonda 191 e em (B) com a sonda 257.

Figura 8-Hibridação das sondas das linhagens 191 e 257 da família SChAT com membrana de DNA genômico digerido com a enzima DraI.

São avaliadas oito variedades de cana-de-açúcar: Mandalay (**M**), Badila (**B**), Coimbatore (**C**), SP71-1088 (**S1**), SP80-3280 (**S2**), R570 (**R**), CP78-2114 (**CP**), IK7648 (**IK**). Assim como, a gramínea sorgo (**S**). A) sonda do TE191 B) sonda do TE257.
Fonte: Ochoa (2012).

O padrão de hibridização aqui apresentado é distinto entre as duas linhagens, sendo que a linhagem 191 apresenta uma distribuição heterogênea entre as linhagens ancestrais quando comparada com as variedades cultivadas (S1 e S2). A linhagem 257 apresenta um sinal de hibridização mais intenso o que sugere amplificação desta linhagem em todos os genomas. Em ambos os casos verifica-se que sorgo (S) apresenta um sinal de hibridização fraco se comparado com as demais canaletas. Os resultados aqui apresentados complementam os publicados recentemente pois a análise encontra-se restrita ao grupo *Saccharineae* e confirma a que as linhagens 191 e 257 estão limitadas ao subgrupo *Saccharum-Erianthus-Miscanthus*, portanto mais recentes que 8-9 MYA.

Os genomas de plantas contêm diferentes membros da superfamília hAT. Em milho existem ao menos três tipos de elementos hAT, além do elemento de referência Ac. Arroz tem pelo menos quatro e *Arabidopsis thaliana* pelo menos cinco tipos. Alguns deles formam agrupamentos pouco sustentados, sugerindo uma divergência ancestral da superfamília antes da divergência de Eudicotiledonea e Monodicotiledonea. (Xu e Dooner, 2005).

Por outro lado, a linhagem 074 da família corresponde a TEs domesticados desde antes da divergência de Eudicotiledonea e Monodicotiledonea (De Jesus et al., 2012). Esta linhagem apresenta baixo número de cópias e uma distribuição conservada entre as gramíneas conforme o observado por experimentos de hibridação apresentados em De Jesus et al., (2012) e resultados apresentados nos capítulos a seguir deste trabalho.

4.1.2 Estimativa do Número de Cópias no Genoma da Cana-de-açúcar (Variedade R570) das linhagens 74, 191 e 257 da família SChAT

A seleção dos BACs foi feita sobre uma biblioteca comercial de clones de BACs de cana-de-açúcar (Tomkins et al., 1999) da variedade R570. São apresentados resultados da hibridação de três das membranas da biblioteca hibridadas com a sonda do TE074, representante da linhagem 074; a sonda do TE191, representante da linhagem 191 e a sonda TE257, representante da linhagem 257 (Figura 9A). Posteriormente, em colaboração com a Dr. N. de Setta foi realizada a varredura total da biblioteca de BACs usando iniciadores específicos e PCR (Figura 9B).

A linhagem 074, considerada domesticada, apresenta 6,2 cópias no genoma total segundo a varredura da biblioteca, este número equivale a um *locus* com sete alelos, como foi posteriormente demonstrado através do sequenciamento destes oito clones. O alinhamento de todos os transcritos relacionados com esta linhagem propõe a existência de dois *loci* sendo transcritos (De Jesus et al., 2012), um deles estudado neste trabalho.

Figura 9-Hibridação das sondas das três linhagens da família SChAT com três membranas da coleção de BACs e Número relativo de cópias dos elementos caracterizados.

A) Hibridação das sondas das três linhagens da família SChAT com três membranas da coleção de BACs. Para ser considerada uma hibridação positiva devem marcar dois clones no mesmo quadrante, ressaltados na imagem com círculos. **B)** Número relativo de cópias dos elementos caracterizados. Foram usados iniciadores específicos para a linhagem 074 (SCHAT-2-F, SCHAT-2-R), linhagem 191 (SCHAT-3-F, SCHAT-4-R) e linhagem 257 (SCHAT-5-F, SCHAT-6-R) para achar o número de clones. O número de cópias é igual ao número de clones dividido pela cobertura da biblioteca (1,3 vezes o genoma completo da variedade R570).

Fonte: Ochoa (2012).

A quantificação apresentada na figura 9B é relativa, sendo correspondente apenas às cópias que compartilham as regiões onde os iniciadores ancoram dentro dos TEs. Isto pode subestimar o número de cópias real. Ainda assim, o padrão de representatividade do número de elementos de cada grupo corresponde ao esperado e corrobora as análises de hibridação genômica apresentadas em De Jesus et al. (2012), onde as versões da linhagem 074 estão representadas no

genoma de arroz, milho e do grupo *Saccharum*. Representação que concorda com outros estudos que reportam que cada transposon hAT apresenta um número típico de cópias entre 1 e 50 por genoma, em locais altamente polimórficos entre espécies (Kunze et al., 1997).

Em outros genomas de plantas foram reportados números compatíveis com os aqui relatados para elementos relacionados à superfamília hAT. No genoma da uva do vinho (*Vitis vinifera* L.) foram reportados cinco elementos diferentes possivelmente domesticados contra 1459 elementos *strictu sensu*, incluindo versões completas e incompletas. Cada um dos cinco elementos domesticados possuía uma cópia no genoma (Benjak et al., 2008).

A triagem por hibridação das membranas permitiu o endereçamento de 25 clones (Figura 9A). Foi confirmada a presença de sequências homólogas às sondas moleculares usadas para a triagem em 16 destes clones. Os 16 clones foram preparados para sequenciamento neste trabalho usando a plataforma *Genome Sequencer FLX*. O quadro 4 apresenta os valores de qualidade e concentração dos 16 BACs antes do preparo das bibliotecas para sequenciamento. Para a técnica é recomendado o uso de 3-5 ug de material purificado.

Quadro 4 - BACs selecionados por hibridação e sequenciados.

BAC	Concentração (ng/μl)	260/280	Volume (ul)	Massa (ug)
SHCRBa_092_C13	1487,4	1,76	2.7	4
SHCRBa_120_I22	317,14	1,88	12.7	4
SHCRBa_124_F07	96,98	1,93	41.2	4
SHCRBa_145_O09	3348,2	1,89	1.2	4
SHCRBa_160_F01	1306	1,81	3.1	4
SHCRBa_161_O13	58,24	1,82	68.7	4
SHCRBa_162_H04	162,22	1,63	24.6	4
SHCRBa_167_G07	1548,6	1,85	2.6	4
SHCRBa_185_O22	1543,2	1,83	2.6	4
SHCRBa_152_D05	75.52	1.83	53	4
SHCRBa_160_F12	319.49	1.78	12.5	4
SHCRBa_160_O06	93.76	1.74	42.6	4
SHCRBa_163_G19	93.63	1.8	42.7	4
SHCRBa_170_C07	67.68	1.87	59.1	4
SHCRBa_172_L15	49.99	1.85	80	4
SHCRBa_165_F05	37.5	1.84	106.6	4

São apresentados os resultados de qualidade e concentração do material usado no preparo das bibliotecas para sequenciamento. A quantidade padrão requerida pela técnica é de 3-5 ug.
Fonte: Ochoa (2012).

Também foram selecionados os oito clones de BAC da linhagem 074 encontrados na varredura total da biblioteca com os iniciadores específicos para cada linhagem (Figura 9B). O sequenciamento deste oito clones foi realizado pelo GaTE-Lab como parte do projeto de sequenciamento do genoma da cana-de-açúcar. Estes BACs são: SHCRBa_037_O20, SHCRBa_099_K03, SHCRBa_145_D15, SHCRBa_171_I09, SHCRBa_179_E09, SHCRBa_188_F16, SHCRBa_190_D07, SHCRBa_262_C04. Após analisados estes BACs e os transcritos da linhagem 074 foi verificado que estes iniciadores eram específicos para um dos dois *loci* expressos, razão pela qual estes oito clones correspondem a um único *locus*.

O quadro 5 apresenta os resultados do sequenciamento dos 16 BACs selecionados por hibridização e sequenciados neste trabalho. A complexidade do genoma da cana-de-açúcar e a presença de repetições constituem regiões difíceis de montar em um único *contig*.

Sete dos BACs foram fechados em um único *contig* (sequência nucleotídica consenso de um conjunto de leituras nucleotídicas que se agrupam por sobreposição). Os clones SHCRBa_120_I22, SHCRBa_124_F07, SHCRBa_152_D05 e SHCRBa_170_C07 foram aplicados em duas corridas de *Genome Sequencer FLX*.

A presença dos elementos nos BACs recuperados por hibridação foi confirmada usando sondas eletrônicas sobre as sequências (sobre as leituras independentes e sobre os *contigs*), assim como pela técnica de PCR com os iniciadores específicos para cada sonda (Quadro 5). A PCR dos BACs selecionados por hibridação da linhagem 074 não funcionou com os iniciadores específicos para a sonda. Dois destes BACs continham cópias truncadas na segunda metade do elemento (SHCRBa_92_C13 e SHCRBa_124_F07), fazendo que só um dos iniciadores fosse capaz de se ancorar nelas. O terceiro BAC SHCRBa_120_I22 carregava apenas 165 pb do elemento domesticado em uma das suas extremidades.

Quadro 5 - BACs selecionados e sequenciados.

TE	BAC	Leituras na Montagem	#contigs (>1000pb)	Sequências Repetitivas (%)	Tamanho sequenciado (bp)	Estratégia para fechar	Validação da presença do TE	
							Sonda eletrônica Tamanho/ Presença	PCR
TE074	SHCRBa_092_C13	20284	1	44.91	149674	Fechado	1005pb / 1-416	ND
TE074	SHCRBa_124_F07	45903	1	59.27	149301	Fechado	1005pb / 1-417	ND
TE074	SHCRBa_120_I22	13846	16	30.83	90468	Sanger	1005pb / 252-417	ND
TE191	SHCRBa_145_O09	2399	30	24.37	111854	Sanger	2179pb / 546-1549	+
TE191	SHCRBa_160_F01	11067	1	32	93390	Fechado	2179pb / 2-1580	+
TE191	SHCRBa_162_H04	3601	24	19.67	123523	Sanger	2179pb / 1051-1237	+
TE191	SHCRBa_167_G07	17494	1	45.08	109854	Fechado	2179pb / 1-1580	+
TE191	SHCRBa_185_O22	11987	1	40.13	120019	Fechado	2179pb / 545-1549	+
TE191	SHCRBa_161_O13	17022	17	28.52	63665	Sanger	2179pb / 1-1580	+
TE 257	SHCRBa_152_D05	38882	1	49.24	172838	Fechado	2430pb / 1-1905	+
TE 257	SHCRBa_160_F12	5848	17	42.62	76994	SOLID	2430pb / 12-1919	+
TE 257	SHCRBa_160_O06	3980	34	32.76	107016	Sanger	2430pb / 1-1919	+
TE 257	SHCRBa_163_G19	2846	23	50.85	71389	Sanger	2430pb / 1-1919	+
TE 257	SHCRBa_170_C07	49301	1	40.51	153538	Fechado	2430pb / 1-2429	+
TE 257	SHCRBa_172_L15	9700	34	50.50	133494	Sanger	2430pb / 1-2429	+
TE 257	SHCRBa_165_F05	3108	26	35.48	69574	Sanger	2430pb / 1-1919	+

Neste quadro são apresentados; estado de montagem do BAC, resultados da verificação de presença do elemento por PCR e procura com sonda eletrônica e a estratégia para fechar o BAC. O conteúdo de bases repetitivas de cada BAC foi gerado por mascaramento de todos os contigs, usando a ferramenta RepeatMasker com três bases de dados de elementos de transposição; GaTEs_DataBase (Base de elementos de cana-de-açúcar do GaTE_Lab), TIGR_Gramineae e Repeats.v3.3_repbse_viridiplantae.

ND = Não determinado.

Fonte: Ochoa (2012).

4.1.3 Diversidade e Relações Filogenéticas Observadas nas Linhagens

Os elementos analisados receberam o nome do BAC, quando encontrados como uma única versão dentro do clone foi adicionado o prefixo SChAT no caso dos transposons ou SCtpase no caso dos genes com domínio da transposase domesticada. No caso do BAC SHCRBa_152_D05, onde foram encontradas duas cópias do mesmo elemento inseridas em duas regiões idênticas de 30 kb, estes elementos foram diferenciados pela terminação A e B, foi caracterizado apenas o elemento A por ser uma cópia exata de B, assim como as regiões de 30 kb onde os elementos estavam inseridos.

Dos 16 BACs identificados por hibridização só sete possuíam elementos completos em um único *contig*: cinco transposons e dois elementos domesticados. Dos cinco transposons completos encontrados por hibridização os três da linhagem 191 fazem parte De Jesus et al. (2012), assim como o elemento SChATG1 descrito em trabalhos prévios (De Jesus, 2007). Para facilitar a descrição dos elementos no manuscrito citado acima a nomenclatura foi simplificada: SChAT160F01 adotou o nome de SChAT1, SChAT 167G07 adotou o nome de SChAT2 e SChAT185O22 adotou o nome de SChAT3. As outras oito versões incompletas estão contidas em *contigs* menores a 10 kb e em alguns casos quebradas em mais de um *contig*, estes elementos não foram usados nas análises posteriores. O elemento domesticado do BAC SHCRBa_120_I22 não foi caracterizado porque corresponde apenas a uma pequena sequência de 165 pb na extremidade do BAC.

Posteriormente, mais oito BACs foram selecionados por varredura com iniciadores específicos da biblioteca de BACs. Nestes oito clones só se tem cinco elementos domesticados diferentes. Situação também encontrada nos BACs nos que se identificaram elementos completos da superfamília hAT não relacionados às linhagens de elementos 191 e 257, nestes seis clones só se tem cinco elementos SChAT devido a que os BAC SHCRBa_007_C22 e SHCRBa_102_H05 se sobrepõem e provavelmente correspondem ao mesmo alelo.

Para as análises de diversidade e relações filogenéticas foram considerados apenas os elementos completos: sete elementos domesticados, três transposons relacionados à linhagem 191, dois à linhagem 257 e cinco não relacionados a nenhum dos três agrupamentos iniciais. As sequências nucleotídicas destes 17 elementos, do elemento SChATG1 (De Jesus et al, 2012) e dos cDNAs da família

foram alinhados usando o programa *MUSCLE*, os agrupamentos foram definidos pelo algoritmo *Neighbor-Joining* e as distâncias evolutivas foram computadas usando *Maximum Composite Likelihood* com a opção *Pairwise deletion* do programa MEGA 5 (Kumar et al, 2004). A figura 10 ilustra as relações de agrupamento entre as 17 versões genômicas e as sequências transcritas.

Na figura 10 é possível observar que a linhagem 074 apresenta uma divisão interna em dois grupos. Postulamos que cada grupo corresponda a um *loci* distinto. Do *locus* com maior número de transcritos foram caracterizados os oito clones recuperados por varredura da biblioteca de BACs e os dois clones recuperados por hibridização, dos quais foram recuperados um total de sete elementos domesticados diferentes, cada um correspondente a um alelo distinto.

A presença no genoma da cana de açúcar de pelo menos três linhagens da família hAT relacionadas a transcritos do SUCEST é confirmada pelos resultados de sequenciamento de BACs e filogenia. A linhagem 074 e as outras duas originadas da divisão do grupo de transposons *strictu sensu* (linhagem 191 e 257). A primeira linhagem de transposons está mais proximamente relacionada com os TE191 e TE221 e a segunda linhagem com os TE257, TE001 e TE154. A identificação de um quarto ramo potencial composto pelo elemento SChAT28_C07 e o TE116 deve ser melhor analisada.

Nota-se também que alguns transcritos como TE170 e TE243, permanecem isolados dos agrupamentos principais, podendo representar variantes ainda não identificadas no genoma.

Os elementos SChAT203_M05, SChAT007_C22, SChAT001_L05 e SChAT137_O14 não apresentam transcritos relacionados.

Para os cinco transposons completos encontrados nos BACs por hibridização (linhagens 191 e 257) e os cinco elementos completos que não pertencem às linhagens anteriormente descritas por De Jesus et al. (2012), foi realizada uma segunda análise sequencial que permitiu gerar um *pipeline* de trabalho para identificar sequências intrônicas e exônicas, ORFs, predições de sequências promotoras e TIRs. Esta análise de diversidade será discutidas posteriormente para cada linhagem por separado.

Figura 10-Padrão de agrupamento dos 17 elementos completos, o elemento SChATG1 antes caracterizado e os cDNAs da família.

Cladograma feito pelo algoritmo *Neighbor-Joining*, a partir do alinhamento dos nucleotídeos totais dos elementos usando MUSCLE. As distâncias evolutivas foram computadas usando *Maximum Composite Likelihood* com a opção *Pairwise deletion*. As setas vermelhas indicam as três linhagens estudadas (074 de elementos domesticados, 191 e 257 de transposons) e a seta amarela indica uma possível terceira linhagem de transposons. Os círculos representam cada um dos *loci* da linhagem domesticada. O *loci* azul foi caracterizado neste trabalho.

Fonte: Ochoa (2012).

Nos genomas de plantas é comum encontrar membros distintos da superfamília hAT. Em milho existem três tipos diferentes além do elemento Ac, em arroz são encontrados pelo menos quatro e em *Arabidopsis* pelo menos cinco. Alguns destes elementos não formam clados relacionados o que sugere uma divergência antiga da superfamília, mesmo antes da divergência de monocotiledôneas e dicotiledôneas. Ao menos um par de transposons hAT de milho não interagem entre si, o que implica que evoluíram para atuar independentemente (Xu e Dooner, 2005). A relação entre as transposases das linhagens precisa de ensaios funcionais para ser entendida, tendo a possibilidade de interação entre as transposases codificadas por estes elementos, mesmo pertencendo a diferente linhagens evolutivas. Também poderia existir uma relação entre os elementos domesticados e as linhagens e os transposons.

4.1.4 Estrutura Genômica dos Elementos SChAT

A recuperação de versões genômicas dos elementos permitiu a predição de TIRs, informação valiosa para a identificação de elementos ativos. Os elementos ativos conservariam a estrutura consenso das TIRs (*terminal inverted repeats*) da família à que pertencem, ou ao menos da linhagem da qual fazem parte.

No local de inserção, do lado das TIRs são encontrados as TSDs (*target site duplication*), estas são as sequências geradas no momento da inserção do elemento no DNA. São difíceis de identificar por seu pequeno tamanho (8 pb no caso do elemento Ac) e também podem sofrer mutações. Estas sequências dão informação do local de inserção, mas não dão informação sobre a capacidade de transposição do elemento inserido nesse local.

Para a identificação de possíveis TIRs e TSDs nos transposons identificados nos BACs (linhagem 191, 257 e elementos hAT não relacionados às linhagens) foi usado o programa TIR_FINDER (Rubin et al., 2001) ajustado com os parâmetros da superfamília hAT descritos por Kempken F. e Windhofer F. (2001): TSDs de 7-8 pb, TIRs imperfeitas de 5-27 pb e unidades com tamanho de 5 kb até 12 kb. A caracterização dos 10 transposons completos e o elemento SChATG1 permitiu a proposta de TIRs e TSDs, os quais estão nas extremidades dos elementos (Quadro 6, Figura 11).

O quadro 6 apresenta os resultados de identificação das TIRs e TSDs das unidades de transposons presentes nos BACs analisados. Foram encontrados TSDs para todas as TIRs preditas nos elementos aqui estudados. Nota-se que nenhuma cópias compartilha a TSD ou TIR sugerindo que trata-se de cópias independentes. Há conservação em tamanho, sendo a linhagem 191 de 8 Kb e a linhagem 257 de 5kb. As versões dos transposons SChAT001L05 e SChAT137O14 não contém TIRs e portanto não foram inclusas no quadro abaixo

Quadro 6 - TIRs e TSDs identificados para os transposons completos usando o programa TIR_FINDER (Rubin et al., 2001).

Elemento	Linhagem	TSD (pb)	TSD	TIR (pb)	TIR	TIR	Tamanho (pb)
SChAT160F01	191	7	CAAGTTA	7	GTAACGG	CCATTCC	8673
SChAT167G07	191	7	TCTTCTA	8	GAACAAAG	CATTGTTT	7118
SChAT185O22	191	7	ACAAGCT	8	TAGGGAGC	TCTCCCGA	8405
SChAT152D05	257	7	GCACTCT	8	TAACGGAG	CTTCATTA	6432
SChAT170C07	257	7	ACGTATA	6	TAGAGA	TTTCAA	5875
SChAT007C22	-	8	ACAATAAC	11	TAAAGATGGCA	TGCCATCCTTA	4783
SChAT028C07	116	7	AGCGCGA	7	CAATGCA	TGGAGTG	4183
SChAT203M05	-	7	TACTAGA	10	GCAAATAAAG	CTTGATTTAC	11795

Fonte: Ochoa (2012).

A Figura 11 apresentado apresenta as TIRs destes elementos alinhadas e nota-se que conservam o padrão proposto para a superfamília hAT (T/C)A(A/G)NG (Rubin et al., 2001).

Figura 11-Alinhamento das TIRs dos 8 transposons, do elemento Ac de milho e do elemento SChATG1.

Fonte: Ochoa (2012).

Tanto a proposta de TIRs para a linhagem 191 como para a linhagem 257 devem ser validadas por análises funcionais. Para estas análises se delimitaram regiões onde as TIRs poderiam estar, que compreendem 1 a 2 kb à montante e a jusante da ORF predita para cada elemento. Geralmente esta região de 2kb está delimitada por retrotransposons e genes adjacentes aos elementos. Estas regiões serão descritas com detalhe para cada linhagem

Deve ser lembrado que as TIRs podem ser repetições imperfeitas como é o caso do elemento Ac, o que aumenta mais a dificuldade de identificação (Kunze et al., 1987). A TIR do elemento Ac apresenta como sequência consenso TAGGGATGAAA (Rubin et al., 2001).

As TIRs são reconhecidas pela transposase no momento da excisão do elemento (Feschotte, 2008). O pequeno tamanho das TIRs (11pb no caso do elemento Ac), dificulta a sua identificação, bem como as mutações pontuais que podem ocorrer, tornando estas sequências degeneradas. Estudos realizados por Wang e Kunze (1998), mostraram que não só as TIRs como também a porção 5' terminal do elemento Ac é crucial para a transposição, por possuir sítios de ligação do domínio *Zinc Finger* da transposase.

4.1.5 Transposases da Família SChAT

Um total de 30 clones de BAC foram analisados para a presença e estrutura das transposases semelhantes a hAT. Apenas 17 foram consideradas a título de comparação na figura 12. As transposases e genes domesticados dos 17 elementos foram alinhados (vide anexo A) usando o programa *MUSCLE*, os agrupamentos foram definidos pelo algoritmo *Neighbor-Joining* e as distâncias evolutivas foram computadas usando *Poisson correction* com a opção *Pairwise deletion* do programa MEGA 5 (Kumar et al, 2004).

Na figura 12 pode ser observado a diferenciação das linhagens de elementos pertencentes à superfamília hAT no genoma da cana-de-açúcar. Enquanto a linhagem 074 está relacionada com elementos domesticados, o grupo II (linhagem 191 e 257) está relacionado com elementos de transposição como Ac e Tam3. A formação de dois clados no interior do grupo II assim como as demais análises aqui apresentadas dão suporte à existência de pelo menos duas linhagens evolutivas de transposons. Estas linhagens poderiam ter histórias distintas nos genomas das

gramíneas e ser importantes para entender a diversidade da família SChAT. Existem também outros elementos completos que não pertencem às linhagens 191 e 257 mas fazem parte da historia evolutiva da superfamília hAT no genoma da cana-de-açúcar.

Figura 12 - Cladograma das 17 transposases SChAT caracterizadas neste trabalho.

Cladograma das 17 transposases SChAT caracterizadas neste trabalho, com as transposases de referência da superfamília hAT; hobo (*Drosophilla melanogaster*), Activator (*Zea mays*) e Tam3 (*Antirrhinum majus*), assim como, a transposase SChATG1 e o elemento domesticado DAYSLEEPER de *Arabidopsis*, feito por *Neighbor-Joining*. As distâncias evolutivas foram computadas usando *Poisson correction* com a opção *Pairwise deletion*.

Fonte: Ochoa (2012).

4.2 Linhagem 191

Para entender a diversidade da linhagem 191 foram preditas as sequências intrônicas, as ORFs (*Open reading frame*) e identificadas as sequência TATA que fazem parte do promotor para os três elementos da linhagem: SChAT160F01, SChAT167G07 e SChAT185O22 (Quadro 7). Também foi calculada a similaridade com o elemento de referencia Ac (5 kb) (Kunze R. et al., 1987). Estes elementos fazem parte dos resultados apresentados em De Jesus et al.,(2012) (vide anexo B).

Quadro 7 - Descrição dos elementos da linhagem 191

Elemento	Tamanho (pb)	TATA	ORF (pb)	Sequências intrônicas	Similaridade com Ac	Cobertura
SChAT160F01	8673	TTATTT	1824	1	34%	74%
SChAT167G07	7118	N.I	1812	Não	51%	75%
SChAT185O22	8405	TTATTT	1038	1	47%	38%

N.I indica não identificada.
Fonte: Ochoa (2012).

Na predição das ORFs das versões genômicas da linhagem 191, foi identificado que o elemento SChAT167G07 possui uma mudança de fase de leitura, assim como dois códons de parada internos. Não foi identificada nenhuma sequência intrônica para este elemento.

As ORFs de SChAT160F01 e SChAT185O22 não possui códons de parada internos nem mudança de fase de leitura. SChAT160F01 possui uma sequência intrônica na UTR3' (*untranslated region*). No entanto, SChAT185O22 possui uma sequências intrônica na região que é processada e traduzida para produzir a transposase. A divergência entre a posição das sequências intrônicas é uma característica comum dos elementos da superfamília hAT em plantas (Muehlbauer et al, 2006).

Dois dos elementos desta linhagem compartilham o mesmo TATA na sequência promotora. Os TATA e as sequências promotoras aqui preditas devem ser validadas por ensaios funcionais.

Posteriormente, a ORF predita foi utilizada para gerar um alinhamento das transposases destes elementos (Figura 13).

Figura 13 - Alinhamento das transposases da linhagem 191 e o elemento de referência Ac.

Alinhamento das transposases da linhagem 191 com a transposase de Ac. Com setas vermelhas são indicados os aminoácidos conservados DDE, também são indicadas as seis regiões conservadas A-F e os domínios conservados *Zinc Finger* (Caixa Azul) e hATC (Caixa Verde).

Fonte: Ochoa (2012).

Na figura 13 pode ser observado que a transposase SChAT185O22 não possui o domínio *Zinc finger* nem as regiões A e B, o que também faz que o primeiro D da assinatura DDE não esteja presente (vide anexo A). Este elemento possui uma grande deleção interna na sua porção 5', o que poderia impossibilitar sua transposição tanto autônoma como não autônoma. O elemento SChAT160F01 possui todas as regiões das transposases importantes para a função da proteína. No entanto, SChAT167G07 possui os domínios e regiões necessários para a função da proteína, faltando só um aminoácido da assinatura DDE, mas a sua ORF possui uma mudança de fase de leitura e códons de parada internos, razão pela qual esta proteína poderia não ser funcional.

A assinatura DDE das transposases é crítica para sua correta função, assim como o motivo CXXH da região B. Mutação no H resultam em defeitos severos da atividade (Muehlbauer et al, 2006).

Em estudos feitos sobre o elemento Ac foi observado que mutação no domínio de dimerização (domínio hATC) praticamente eliminam a transposição do elemento. A inserção de dois aminoácidos na região adjacente ao domínio de dimerização não afeta a transposição quando o elemento foi expresso em altos níveis (Weil e Kunze, 2000). Outros estudos, como o de Esser et al. (2000), descrevem que mutações em 5 dos 6 resíduos que fazem parte da assinatura do domínio hATC não afetam sua função de dimerização *in vitro*.

O elemento SChAT160F01 poderia ser um elemento autônomo, enquanto as cópias SChAT167G07 e SChAT185O22 poderiam ser elementos não autônomos. Estes poderiam ser mobilizados *in trans* pela proteína dos outros integrantes da mesma linhagem. Para todos os elementos desta linhagem foram encontrados TIRs com TSDs. Informação que apresenta mais um indicio de que esta linhagem poderia possuir elementos completos e ativos. Finalmente, os transcritos TE 191 e TE221, representantes da linhagem, poderiam traduzir transposases potencialmente ativas.

4.2.1 Análises das Regiões Sintênicas e Contexto Genômico

Foi estudado o contexto genômico em que os elementos da linhagem 191 estão inseridos. Foram feitos estudos de sintenia nos genomas de *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* e *Zea Mays*. Os

genes e TEs destes BACs também foram anotados como explicado na metodologia (Figura 14 - 16).

Foram encontradas regiões sintênicas para um dos três BACs que contêm elementos desta linhagem (Figura 16). Nos casos em que não foi encontrada sintenia a anotação dos BACs é apresentada (Figura 14 e 15). Para nenhum dos BACs foi encontrada sintenia com *Arabidopsis thaliana*.

Figura 14 - Anotação do BAC SHCRBa_160_F01.

Os genes são apresentados em cores diferentes. Os TEs mais relevantes quanto ao tamanho são apresentados na cor branca. O elemento SChAT160F01 é apresentado em vermelho.

Fonte : Ochoa (2012).

Figura 15 - Anotação do BAC SHCRBa_167_G07.

Os genes são apresentados em cores diferentes. Os TEs mais relevantes quanto ao tamanho são apresentados na cor branca. O elemento SChAT167G07 é apresentado em vermelho.

Fonte: Ochoa (2012).

Para o BAC SHCRBa_160_F01 foi encontrada uma região sintênica com arroz e sorgo para os últimos três genes, correspondente aproximadamente a 15kb (dado não apresentado). O elemento SChAT160F01 está do lado do gene que

codifica para uma proteína similar a um receptor de quinase 1, ambos no mesmo sentido. Isto abre a possibilidade de que o promotor do elemento possa interferir na expressão do gene (Figura 14).

No caso do BAC SHCRBa_167_G07 é impossível encontrar uma regiões sinténicas, este BAC esta composto por TEs e um único gene repetido. A primeira cópia do gene glicosil hidrolase esta degenerada e no sentido complementar, a segunda cópia está bem conservada (Figura 15).

A estrutura destes dois elementos os torna candidatos a elementos ativos (Figura 13), porém a falta de sinténia entre os BACs onde estão inseridos impossibilita verificar se estes dois elementos tiveram uma transposição recente.

Para o BACs SHCRBa_185_O22 foi possível encontrar sinténia com todas as gramíneas estudadas, até *Brachypodium* ssp., gramínea basal (Figura 16). Os dados de sinténia conferem com a sinténia postulada por Paterson et al., 2009 para arroz e sorgo.

Figura 16 - Anotação e sinténia do BAC SHCRBa_185_O22.

Os genes são apresentados em cores diferentes. Os TEs mais relevantes quanto ao tamanho são apresentados na cor branca. O elemento SChAT185O22 é apresentado em vermelho.

Fonte: Ochoa (2012).

O elemento SChAT185O22 teria se mobilizado recentemente, após a divergência entre cana e sorgo (Figura 16). Este elemento apresenta uma deleção na porção 5' motivo pelo qual pode ser predito como um elemento ativo mas não autônomo (Figura 13). A identificação das TIRs e TSDs deste elemento também apóiam esta hipótese (Quadro 6). No entanto, a sua inserção dentro de um *locus* em cana-de-açúcar, e a sua ausência na região sintênica correspondente em sorgo, arroz, milho e brachypodium, indicam que este elemento foi ativo na história evolutiva recente da cana. O elemento SChAT185O22 está localizado a montante do gene serina/treonina quinase, ambos no mesmo sentido (Figura 16). Existem ainda outros elementos presentes entre o gene serina/treonina quinase e o transposon, porém, como a maioria está incompleta não são apresentados na figura.

Para o elemento SChAT185O22 não foram encontrados ortólogos na mesma região de sintenia com o genoma de sorgo. Isto sugere uma transposição relativamente recente, posterior à divergência de sorgo e cana-de-açúcar, há aproximadamente oito milhões de anos.

4.3 Linhagem 257

Posteriormente foram analisados os elementos da linhagem 257. As ORFs destes elementos não apresentam códons de parada internos, e ambos apresentam duas sequências exônicas na região traduzida. O elemento SChAT170C07 também apresenta uma sequência intrônica na UTR3'. Para um deles foi possível identificar o TATA da sequência promotora, o qual é igual ao identificado em dois elementos da linhagem 191 (Quadro 8). O TE257 é o cDNA correspondente ao elemento SChAT170C07 com um 99% de identidade, não temos cDNA correspondente para o elemento SChAT152D05 na coleção do SUCEST.

Quadro 8 - Descrição dos elementos da linhagem 257.

Elemento	Tamanho (pb)	TATA	ORF (pb)	Sequências intrônicas	Identidade com Ac	Cobertura
SChAT152D05	6432	N.I	1986	1	29%	73%
SChAT170C07	5875	TTATTT	1983	2	62%	74%

N.I indica não identificada.
Fonte: Ochoa (2012).

As transposases de dois elementos completos desta linhagem foram caracterizadas, SChAT170C07 e SChAT152D05_A. Na figura 17 foi observado que as transposases destes elementos possuem o domínio e regiões necessários para a função da proteína, faltando só um aminoácido da assinatura DDE na transposase de SChAT152D05.

Figura 17 - Alinhamento das transposases da linhagem 257 e o elemento de referencia Ac.

Alinhamento das transposases da linhagem 257 com a transposase de Ac. Com setas vermelhas são indicados os aminoácidos conservados DDE, também são indicadas as seis regiões conservadas A-F e os domínios conservados *Zinc Finger* (Caixa Azul) e hATC (Caixa Verde).

Fonte: Ochoa (2012).

Os elementos desta linhagem têm um alto potencial de serem elementos ativos no genoma, ou pelo menos de ter transposases funcionais. Possuindo as regiões A-F, assim como os domínios conservados *Zinc finger* e hATC, a assinatura DDE e o motivo CXXH da região B. O TE257 parece corresponder a um transcrito capaz de ser traduzido em uma transposase ativa. Análises funcionais dariam informação conclusiva sobre estas hipóteses.

4.3.1 Análises das Regiões Sintênicas e Contexto Genômico

Para as duas versões da linhagem 257 foram feitos estudos de sintenia nos genomas de *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* e *Zea Mays*. Os genes e TEs destes BACs também foram anotados como explicado na metodologia (Figura 18 e 19).

Para os BACs SHCRBa_152_D05 e SHCRBa_170_C07 foi possível encontrar sintenia com todas as gramíneas estudadas, até *Brachypodium* ssp., gramínea basal. Os dados de sintenia conferem com a sintenia postulada por Paterson et al., 2009 para arroz e sorgo. Para nenhum dos BACs foi encontrada sintenia com *Arabidopsis thaliana*.

O elemento SChAT152D05 está localizado numa região repetida exata dentro do BAC com tamanho de 30Kb. Esta região repetida não contém nenhum gene, nela estão presentes o elemento SChAT e o elemento LTR/COPIA/Maximus-Sire R51. A segunda repetição do elemento SChAT152D05 localiza-se a montante do gene da glucanase/ribito desidrogenase, ambos no mesmo sentido. Não existem outros elementos presentes entre o gene e o transposons, tendo a possibilidade de que o promotor do elemento possa interferir na expressão do gene.

Na figura 18 também pode ser observado que o gene que codifica para uma proteína similar às de ligação de Poli(A) está interrompido pela primeira cópia do elemento LTR/COPIA/Maximus-Sire R51, que tem três cópias idênticas dentro deste BAC.

Figura 18 - Anotação e sintenia do BAC SHCRBa_152_D05.

Os genes são apresentados em cores diferentes. Os TEs mais relevantes quanto ao tamanho são apresentados na cor branca. O elemento SchAT152D05 é apresentado em vermelho. Os colchetes indicam sequências que foram extraídas da sintenia, cada colchete apresenta as kilobases extraídas para adequar a proporção da figura. Fonte: Ochoa (2012).

Figura 19 - Anotação e sintenia do BAC SHCRBa_170_C07.

Os genes são apresentados em cores diferentes. Os TEs mais relevantes quanto ao tamanho são apresentados na cor branca. O elemento SChAT170C07 é apresentado em vermelho. Os colchetes indicam sequências extraídas da sintenia, cada colchete apresenta as kilobases extraídas, a fim de manter a proporção do diagrama. Fonte: Ochoa (2012).

Na sintonia do BAC SHCRBa_170_C07 pode ser observado que o elemento SChAT170C07 está localizado após o gene que codifica uma glioxalase, estão em sentidos opostos e a uma distância de mais de 10 kb, fazendo pouco provável uma interferência do elemento sobre este gene (Figura 19).

Na figura 19 também pode ser observado que o gene que codifica para uma proteína Alba de ligação a DNA/RNA está interrompido pelo elemento homólogo a LTR/COPIA/IVANA R7.

Para as duas versões de elementos da linhagem 257, não foram encontrados ortólogos na mesma região de sintonia com o genoma de sorgo. Isto sugere uma transposição relativamente recente, posterior à divergência de sorgo e cana-de-açúcar, há aproximadamente oito milhões de anos. O que concorda com as previsões feitas sobre suas transposases (Figura 17), postulando-lhes como elementos ativos, e autônomo no caso do SChAT170C07.

4.3.2 Expressão da Transposase

O TE257, quem possui o domínio *Zinc finger* e o domínio hATC assim como os outros aminoácidos importantes para a função da transposase, foi escolhido para ser expresso usando 5 construções (Quadro 9). No uso de um transcrito não se teria a preocupação de eliminar sequências intrônicas, problema que poderia ser enfrentado se fosse escolhida alguma das versões genômicas estudadas neste trabalho. Este experimento foi realizado no laboratório do professor Dr. Reinhard Kunze em Berlin, Alemanha. Sendo a primeira transposase de cana-de-açúcar expressa na história.

Nas construções expressas em *Saccharomyces cerevisiae* foi adicionada a sequência NLS (*Nuclear localization domain*) de levedura. Também foram usados dois promotores derivados do promotor GAL, induzido por galactose, um forte (GAL1) e um fraco (GALS). A transposase Ac de milho, uma das mais estudadas da superfamília SChAT, forma complexos protéicos formados por homodímeros para realizar a transposição, também forma grandes agregados insolúveis quando expressa em níveis altos de forma heteróloga (Kunze et al., 1993). Razão pela qual, decidiu-se usar os dois tipos de promotores derivados de GAL para a expressão da transposase de cana-de-açúcar na levedura.

Quadro 9 - Construções feitas para expressar a transposase codificada pelo TE257.

Organismo	NLS	Promotor	Plasmideo	Nome	Tamanho (KDa)
<i>E. coli</i>			pET15b	pET257	75,3
<i>S. cerevisiae</i>	+	Forte	p416GAL1	NG1	78,4
<i>S. cerevisiae</i>	+	Fraco	p416GALS	NGS	78,4
<i>S. cerevisiae</i>		Forte	p416GAL1	G1	75,3
<i>S. cerevisiae</i>		Fraco	p416GALS	GS	75,3

Fonte: Ochoa (2012).

A expressão foi verificada usando a técnica de *Western Blot*. Não existe anticorpo contra a transposase, razão pela que nas construções foi inserida uma cauda de histidina no início da proteína (Figura 20 e 21).

Figura 20 - *Western Blot* usando o anticorpo anti-histidina em diferentes extratos protéicos de *E. coli*. transformada com o plasmideo pET257.

A) Filme com um minuto de exposição. B) Filme com 2 minutos de exposição. Ac) Parte da transposase Ac de milho (80KDa) purificada e expressa em *Saccharomyces cerevisiae* cedida pelo professor Reinhard Kunze. M) Marcador de peso molecular. 1) Linhagem de *E. coli* não transformada. 2) Linhagem de *E. coli* não induzida. 3) *E. coli* após induzida, extrato total. 4) Meio de cultura. 5) *E. coli* após induzida, fração solúvel. 6) *E. coli* após induzida, fração insolúvel.

Fonte: Ochoa (2012).

A expressão em *E. coli* apresenta varias bandas em cada um dos extratos protéicos (Figura 20). A proteína esperada teria ao redor de 75 KDa, incluindo a cauda de histidina, as outras bandas corresponderiam a outros produtos desta proteína que foram terminados prematuramente, degradados ou iniciados prematuramente. Caso já visto na expressão deste tipo de proteínas em *E. coli*. quando a transposase de Ac foi expressa neste mesmo sistema (Feldmar S. e Kunze R., 1991).

Na Figura 20 pode ser vista uma grande parte da proteína com o peso aparente esperado (73 KDa) na fração insolúvel (Figura 20A, amostra 6), mas também pode ser observado que uma parte menor da proteína esta na fração solúvel (Figura 20B, amostra 5). Não foram encontradas proteínas com calda de histidina sendo excretadas no meio de cultura (Amostra 4), nem expressão no sistema sem induzir (Amostra 2). Foi observada um pouco de reação cruzada do anticorpo anti-histidina com o marcador de peso molecular (M). As transposases são proteínas pouco solúveis que apresentam uma alta tendência a formar agregados, este é o caso da transposase Ac de milho produzida neste mesmo sistema, a qual se acumulava quase exclusivamente em corpos protéicos insolúveis (Feldmar S. e Kunze R., 1991).

Na expressão em *S. cerevisiae* não foi observada proteína compatível com o tamanho esperado no extrato de proteína solúvel (Figura 21A). Na fração insolúvel foi observada a expressão de uma proteína com peso aparente ao redor de 58 KDa unicamente nas amostras induzidas com galactose (Figura 21B, amostras 2-5). Também foi observada reação cruzada do anticorpo com o marcador de peso molecular e alguma reação com a *S. cerevisiae* sem transformar (Amostra 1). A reação cruzada com a levedura sem transformar foi pouca e não estava na mesma altura da proteína de 58 KDa observada nas amostras induzidas para expressão.

Figura 21 - Western Blot usando o anticorpo anti-histidina em diferentes extratos protéicos de *S. cerevisiae* transformada com os plasmídeos; GS, G1, NGS e NG1.

A) Extrato de proteínas solúveis. B) Extrato de proteínas insolúveis. Nas amostras 2-5 foi induzida a expressão da transposase 257 usando galactose como fonte de carbono em quanto nas amostras 6-8 foi reprimida usando glicose como fonte de carbono. Ac) Parte da transposase Ac de milho (80KDa) purificada e expressa em *Saccharomyces cerevisiae* cedida pelo professor Reinhard Kunze. M) Marcador de peso molecular. 1) *S. cerevisiae* sem transformar. 2) G1. 3) GS. 4) NG1. 5) NGS. 6) G1. 7) GS 8) NG1.

Fonte: Ochoa (2012).

Na figura 21B também pode ser observado que não teve variação aparente de expressão entre as construções com NLS (Amostras 4 e 5) e sem (Amostras 2 e 3), mas teve uma maior expressão aparente quando usado o promotor forte (Amostras 2 e 4). Não foi possível produzir proteína solúvel com esta metodologia em *S. cerevisiae*. O peso aparente de 58 KDa é menor ao esperado para qualquer das transposases expressas por estas quatro construções (Quadro 9). Isto pode ser devido, como no caso de *E. coli*, a produtos desta proteína que foram terminados prematuramente ou degradados. Para verificar se a proteína estava sendo degradada no processo de extração foi adicionando um conjunto de inibidores de proteases comercial, não sendo observado diferença no resultado da extração. O peso real da proteína poderia ser verificado por outras técnicas como espectrometria de massa.

Mesmo usando gel de poliacrilamida com SDS (dodecil sulfato de sódio), como foi usado neste experimento, anomalias podem ocorrer na migração de proteínas como glicoproteínas, lipoproteínas e nucleoproteínas, por causa de reduções incompletas ou ligações não homogêneas do detergente com os componentes não protéicos deste tipo de proteínas (Shi e Jackowski, 1998). Também ocorre em proteínas com alto conteúdo de lisina ou prolina, que não é o caso da transposase que possui apenas 6% e 3% destes aminoácidos, respectivamente (Shi e Jackowski, 1998).

A expressão das diferentes versões desta transposase abre as portas para outros ensaios que validem as previsões *in silico* feitas neste trabalho como: capacidade de ligação no DNA, capacidade de dimerização, capacidade de transposição e promotores. Esses experimentos complementaríamos as previsões funcionais feitas neste trabalho para as transposases da superfamília hAT em cana-de-açúcar.

4.4 Linhagem 074, um Possível Gene Domesticado

Foram estudados sete versões da linhagem domesticada 074; SCtpase92C13, SCtpase124F07, SCtpase037O20, SCtpase099K03, SCtpase145D15, SCtpase171I09 e SCtpase188F16.

O alinhamento dos cDNAs apresentado no documento De Jesus et al. (2012) assim como o agrupamento da Figura 10 dividem em dois grupos a linhagem 074,

que poderiam ser provenientes de dois *loci* no genoma. Razão pela qual se optou por sequenciar os 8 clones identificados na biblioteca por varredura com iniciadores específicos. Deste oito clones foram obtidos cinco elementos, devido a que alguns clones faziam parte do mesmo alelo com sobreposição no elemento SChAT. A estes cinco elementos se juntam as duas versões identificadas por hibridação, SCtpase92C13 e SCtpase124F07.

As sequências deste sete elementos foram caracterizadas. As ORFs destes elementos não apresentam códons de parada internos, com exceção do elemento SCtpase124F07 quem apresenta três códons de parada internos. Nenhum apresenta sequências intrônicas. Para quase todos foi possível identificar o TATA da sequência promotora, o qual é igual ao identificada em dois elementos da linhagem 191 e um elemento da linhagem 257 (Quadro 10).

Quadro 10 - Descrição dos elementos da linhagem 074.

Elemento	Tamanho (pb)	TATA	ORF (pb)	Sequências intrônicas	Identidade com Ac	Cobertura
SCtpase92C13	3288	N.I	1563	0	50%	51%
SCtpase124F07	4840	TTATTT	1350	0	55%	38%
SCtpase037O20	5255	TTATTT	2181	0	33%	65%
SCtpase099K03	4003	TTATTT	2181	0	42%	65%
SCtpase145D15	4175	TTATTT	2181	0	42%	65%
SCtpase171I09	4251	TTATTT	2181	0	33%	65%
SCtpase188F16	4395	TTATTT	2181	0	33%	65%

Fonte: Ochoa (2012).

A linhagem, representada pelo TE074, é encontrada em um número relativo menor de cópias que as linhagens de elementos 191 e 257 (Figura 9A e B). Posteriormente foi visto que estes sete elementos eram versões alélicas do mesmo *locus* como vai ser apresentado posteriormente na sintonia dos BACs.

Não foram recuperadas versões do outro *locus*, relacionado aos TEs 048, 265, 207 e 096, devido a que os iniciadores usados para a varredura da biblioteca resultaram ser específicos para um único dos *loci*.

Na figura 22 é apresentada a análise das transposases domesticadas codificadas por estes elementos.

Figura 22 - Alinhamento das transposases domesticadas da linhagem 074, o elemento domesticado DAYSLEEPER e a transposase de Ac.

Alinhamento das transposases domesticadas da linhagem 074 com a transposase de Ac e a transposase domesticada DAYSLEEPER. Com setas vermelhas são indicados os aminoácidos conservados DDE, também são indicadas as seis regiões conservadas A-F e os domínios conservados Zinc Finger (Caixa Azul) e hATC (Caixa Verde).
Fonte: Ochoa (2012).

No caso das versões SCtpase92C13 e SCtpase124F07, suas transposases conservam identidade com a transposase de Ac, conservam as regiões A, B e C mas perderam as regiões D,E e F assim como o domínio hATC. Isto como parte de uma deleção nestas duas versões que não é vista nas outras cinco versões deste *locus*. As outras cinco versões alélicas e a transposase domesticada DAYSLEEPER apresentam todas as regiões A, B, C, D, E, F e os domínios conservados *Zinc Finger* e hATC (dimerização) (Figura 22).

A historia é diferente quando analisada a assinatura DDE, fundamental para a função da proteína como transposase. O primeiro D é conservado em DAYSLEEPER mas não aparece em nenhuma das versões da linhagem 074. O segundo D aparece em todas as versões da linhagem 074 mas não no elemento DAYSLEEPER. Finalmente, o E não aparece em nenhuma das versões da linhagem nem no elemento DAYSLEEPER. Esta informação demonstra que os elementos da linhagem 074 possuem relação com transposases, mas também poderia indicar que as proteínas codificadas poderiam não ter mais a função de transposases, como é o caso do elemento DAYSLEEPER, conhecido por ser um fator de transcrição que se liga na *Kubox1* que esta a montante do gene de reparo *Ku70* e que é essencial para o desenvolvimento de *Arabidopsis* (Bundock e Hooykaas, 2005).

Como possíveis elementos domesticados, as versões desta linhagem perderiam partes de suas estruturas o que impediria sua transposição, mas conservariam alguns domínios de interesse para a composição de novos genes. Existe o caso da formação de novos fatores de transcrição, que ocorre quando o domínio usado na criação do novo gene é um domínio de ligação ao DNA, como no caso do DAYSLEEPER descrito por Bundock e Hooykaas em 2005.

Também deve ser contemplada a possibilidade de que este elemento tenha papel regulatório sobre os elementos SChAT das outras linhagens. Estas hipóteses só poderão ser avaliadas com análises funcionais.

4.4.1 Linhagem Domesticada e sua Conservação ao Longo de Vários Genomas

Usando os elementos da linhagem 074 como sondas, que correspondem a um único *locus* em cana-de-açúcar, foi feita uma busca de outros elementos nos genomas de *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*,

Oryza sativa e *Zea Mays*. Foram encontradas quatro cópias em *Brachypodium*, três cópias em sorgo, cinco em arroz e sete em milho (Quadro 11). Os ortólogos encontrados em sorgo são diferentes dos encontrados para as linhagens de elementos 191 e 257.

Os resultados apresentados no quadro 11 corroboram resultados de *Southern Blot* (De Jesus et al., 2012) em que foram encontrados ortólogos da linhagem 074 ao longo das gramíneas. Não foram achados ortólogos em *Arabidopsis*, o que concorda com os mesmos resultados de *Southern Blot* em que não foram encontrados ortólogos nas dicotiledôneas. Por outro lado, quando sequências de elementos domesticados presentes em dicotiledôneas, como é o caso de DAYSLEEPER, são alinhadas com os cDNAs do grupo I representado pelo TE074, encontramos a formação de um clado que diferencia a linhagem 074 das linhagens 191 e 257. Mesmo que o clado de sequências domesticadas também esteja subdividido em sequências de dicotiledôneas e monocotiledôneas (De Jesus et al., 2012).

Quadro 11 -Ortólogos da linhagem 074.

Organismo	Cromossomo	Posição (bp)	Tamanho (bp)	Identidade %	E-value
<i>Brachypodium distachyon</i>	Bd1	6924543	1532	85	0
<i>Brachypodium distachyon</i>	Bd3	42912835	1559	82	0
<i>Brachypodium distachyon</i>	Bd3	14232724	1567	81	0
<i>Brachypodium distachyon</i>	Bd1	6917639	1559	81	0
<i>Oryza sativa</i>	5	29119473	1548	85	0
<i>Oryza sativa</i>	3	30804550	1548	85	0
<i>Oryza sativa</i>	5	8492121	1548	84	0
<i>Oryza sativa</i>	5	29110173	1421	78	0
<i>Oryza sativa</i>	1	20049174	975	74	7E-103
<i>Sorghum bicolor</i>	2	77207647	2105	91	0
<i>Sorghum bicolor</i>	1	7832603	1539	84	0
<i>Sorghum bicolor</i>	8	6909325	1265	74	3E-146
<i>Zea mays</i>	5	8599245	1658	93	0
<i>Zea mays</i>	7	168323052	1664	92	0
<i>Zea mays</i>	1	272321084	1539	83	0
<i>Zea mays</i>	5	8594688	1539	83	0
<i>Zea mays</i>	10	126221307	691	85	0
<i>Zea mays</i>	10	1384762	690	85	0
<i>Zea mays</i>	10	16558635	998	71	3E-73

Em vermelho são apresentados os ortólogos localizados em regiões próximas entre si. Os ortólogos localizados na região sintênica ao locus estudado em cana-de-açúcar são destacados em amarelo.

Fonte: Ochoa (2012).

Foi feita procura no genoma de *Miscanthus giganteus* (Swaminathan K. et al., 2010), gramínea altamente relacionada com cana e sorgo, da qual foi sequenciado um 1,2% do genoma. Foi usado o algoritmo de BLASTN da página do NCBI. Não foram encontradas sequências homologas aos elementos desta linhagem, o que era esperado devido à pequena porcentagem do genoma que foi sequenciada até o momento.

O número de cópias menor, a conservação ao longo das gramíneas e a correspondência filogenética destes elementos com a transposase de Ac, sustentam a postulação desta linhagem como uma linhagem de elementos domesticados. Esta linhagem teria origem anterior à divergência de monocotiledôneas e dicotiledôneas como postulado em De Jesus et al. (2012).

Esta linhagem corresponderia a pelo menos dois *loci* com capacidade de transcrição em cana (Figura 10), que são conservados ao longo de várias gramíneas (Quadro 11).

Os elementos desta linhagem apresentam todas as características de uma linhagem de elementos domesticados enunciadas por Feschotte et al. (2007). No entanto devem ser apresentadas provas de evidencia funcional de função biológica e evidências de expressão.

4.4.2 Análises das Regiões Sintênicas e Contexto Genômico

Foram completamente sequenciados 10 BACs com cópias de elementos da linhagem domesticada (Figura 23). Para estes BACs foram feitos estudos de sintenia nos genomas de *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* e *Zea Mays*. Os genes destes BACs também foram anotados como explicado na metodologia.

Na Figura 23 são observados os sete alelos para um dos *locus* domesticados em cana-de-açúcar. Alguns dos BACs poderiam representar o mesmo alelo por isso são representados entre colchetes na figura. Também são apresentadas as regiões sintênicas nas outras gramíneas estudadas. Não foi encontrada região sintênica em *Arabidopsis*.

Dois dos ortólogos, encontrados nos genomas de sorgo e milho, foram encontrados na região sintênica ao *locus* estudado em cana-de-açúcar, o que poderia indicar uma incapacidade de transposição destes elementos, anterior à

divergência de pelo menos estas três espécies de gramíneas, faz 12 milhões de anos aproximadamente (Quadro 11 e Figura 23).

Figura 23 - Anotação e sintenia dos BACs de um dos *locus* da linhagem 074 em cana-de-açúcar.

Os TEs não são apresentados nesta figura. O elemento domesticado é apresentado em vermelho. Os colchetes pretos indicam sequências que foram extraídas da sintenia, cada colchete apresenta as kilobases extraídas, a fim de manter a proporção do diagrama. Os colchetes vermelhos indicam BACs que poderiam fazer parte do mesmo contínuo alélico.
 Fonte: Ochoa (2012).

O gene ortólogo aos elementos SChAT da linhagem 074 que está presente na região sintênica de sorgo é SORBIDRAFT_02g043380, anotado como codificante de uma proteína hipotética com domínio de ligação ao DNA, com expressão em

folhas de sorgo. Em milho também o gene ortólogo é anotado como codificante para uma proteína hipotética predita.

4.4.3 Análises de Expressão da transposase domesticada

Os TEs estudados apresentam características de distribuição e estrutura bem definidas, as quais foram descritas nos capítulos anteriores. Foi observado que os TEs preditos como canônicos conservam melhor sua estrutura e apresentam maior número de cópias relativa no genoma hospedeiro, caso da linhagem 191 e 257. Para a linhagem 074 o panorama foi exatamente o contrário, menor conservação da estrutura, menor número de cópias no genoma hospedeiro e maior distribuição ao longo das gramíneas.

Além das características anteriormente descritas, os TEs também apresentam padrões de expressão que são regulados pelo genoma hospedeiro. Por exemplo, os elementos domesticados apresentam níveis de expressão comparáveis com os níveis de expressão dos genes, o que é incomum para um TE. Este padrão de expressão relaciona os TEs com funções na célula (Feschotte e Pritham, 2007).

Para constatar os padrões diferenciados de expressão entre os elementos das linhagens 074 e os transposons (linhagem 191 e 257) foram usadas análises de hibridação com RNA (*Northern Blot*). Foram realizados dois ensaios para ver expressão na cana-de-açúcar: um ao longo de diferentes tecidos no que foram feitas duas coletas e um segundo no mesmo tecido ao longo do tempo.

O primeiro destes ensaios para constatar expressão foi realizado usando colmos da variedade SP80-3280 e R570 cultivados em casa de vegetação e posteriormente passados para terra, sendo mantidos em condições naturais. Foram feitas duas coletas, após três meses e após seis meses da semeadura, para constatar expressão ao longo da planta. Foi feita coleta dos seguintes tecidos: raiz, base floral, talo (apenas na coleta de seis meses), folhas superiores, folhas inferiores e meristema apical (apenas na coleta de seis meses).

No experimento de hibridação por *Northern Blot* das plantas com três meses de idade não foi encontrada expressão de nenhuma das três linhagens em talo, raiz, folhas superiores nem inferiores. Por outro lado, no experimento de seis meses foi detectada hibridação da sonda da linhagem 074. Este resultado é apresentado no manuscrito De Jesus et al. (2012) (vide Anexo B). Onde foi observado que a

variedade R570 parece ter maior expressão dos elementos da linhagem 074 ao longo de todos seus tecidos quando comparada com a variedade SP-803280. No entanto, em meristema apical a expressão foi mais intensa na variedade SP-803280 que em R570, o que poderia estar relacionado com uma função específica do gene composto pela transposase domesticada nesta determinada variedade. As sondas das linhagens 191 e 257, relacionadas com elementos de transposição, não apresentaram hibridação neste experimento em três ou seis meses.

Para os TEs domesticados, é de se esperar níveis de expressão similares aos encontrados para genes. Isto porque o elemento teria sido aproveitado pelo genoma e transformado num gene, ou faria parte de um ou vários genes, tendo papéis regulatórios. Este tipo de fenômenos foram discutidos na introdução, quando foi falado de domesticação de TEs e da formação de redes regulatórias por estes (Bundock e Hooykaas, 2005; Cowan et al., 2005; Feschotte e Pritham, 2007 Hudson et al., 2003). No caso dos TEs é de se esperar níveis de expressão mais baixos, devido a que sua expressão é regulada no genoma por diferentes mecanismos, já explicados na introdução.

O segundo ensaio de expressão foi feito com meristemas apicais cedidos pela Profa. Dra. Kátia Scortecci que correspondem a duas variedades, RB75-126 e SP81-3250, conhecidas por nunca florescer e florescer precocemente, respectivamente. Estas plantas foram acompanhadas por oito meses de desenvolvimento em condições naturais de cultura. Isto, com o intuito de fazer uma análise de expressão ao longo de um período de tempo sobre o mesmo tecido. Neste segundo ensaio foi observada expressão em todos os meristemas apicais, nas duas variedades e ao longo dos 8 meses (Figura 24).

Os resultados de expressão destes elementos poderiam indicar expressão relacionada ao momento de desenvolvimento da planta, já que a expressão nos diferentes tecidos não foi encontrada nas amostras de três meses, mas sim nas de seis meses, quando analisadas as variedades SP803280 e R570. Na coleta de três meses não foi possível coletar meristema apical pelo pequeno tamanho das plantas, isto limita a análise deste tecido em particular, o qual foi encontrado como expresso ao longo de todo o desenvolvimento das variedades RB75-126 e SP81-3250.

Posteriormente, os RNAs do segundo experimento foram usados para fazer cDNA e foram amplificados e sequenciados, para assim observar os polimórficos

que poderiam estar acontecendo nas regiões expressas. O que poderia dar uma visão da diversidade alélica deste *locus* (Figura 25 e Figura 26).

Os transcritos amplificados e subclonados foram sequenciados. Suas sequências foram então alinhadas com as sequências de TEs e as ORFs dos elementos SCtpase92C13 e SCtpase124F07 da linhagem domesticada (Figura 26).

Figura 24 - Análise de expressão do meristema apical de duas variedades de cana-de-açúcar.

Amostra	1	2	3	4	5	6	7	8	9	10
florescer precocemente	+	-	+	+	-	-	+	-	+	-
nunca floresce	-	+	-	-	+	+	-	+	-	+
mês	1	1	2	2	2	2	6	6	8	8

Análise de expressão do meristema apical de duas variedades de cana-de-açúcar, RB75-126 e SP81-3250, conhecidas por nunca florescer e florescer precocemente, respectivamente. Estas plantas foram acompanhadas por oito meses em condições naturais.

Fonte: Ochoa (2012).

Figura 25 - Procedimentos realizados para sequenciar transcritos da linhagem 074 do meristema apical de duas variedades de cana-de-açúcar, RB75-126 e SP81-3250

A) RNAs utilizados para *Northern Blot*. B) RNAs tratados com DNase I usados para gerar cDNA. C) Amplificação usando cDNA e iniciadores específicos para a linhagem 074. D) Amplificados purificados para serem usados no sequenciamento. Canaletas 1,3,4,7 e 9 correspondem à variedade SP81-3250, conhecida por florescer precocemente. Canaletas 2,5,6,8, e 10 correspondem à variedade RB75-126, conhecida por não florescer. M) marcador de peso molecular 100bp. Na imagem A e B as bandas observadas são as correspondentes a 28SrRNA e 18SrRNA. Na imagem B, as canaletas marcadas com b são replicatas. Na imagem C, + corresponde a DNA do TE074, - reação sem usar cDNA, - reação usando amostra de RNA para verificar que não tinha contaminação com DNA genômico.

Fonte: Ochoa (2012).

Figura 26 - Alinhamento das ORFs dos elementos SCtpase92C13, SCtpase124F07, os TEs da linhagem 074 e os amplificadores dos cDNAs do meristema apical de duas variedades de cana-de-açúcar, RB75-126 e SP81-3250.

O alinhamento foi feito utilizando-se o programa *MUSCLE*, usando as sequências nucleotídicas. O número de cada cDNA corresponde à ordem da figura 24, os cDNAs com as terminações B ou C são replicatas dos clones sequenciados.
Fonte: Ochoa (2012).

No alinhamento da figura 26 pode ser observado que os amplificadores correspondem a um dos *loci* que possui a linhagem domesticada, o mesmo dos elementos SCTpase92C13 e SCTpase124F07, assim como dos outros cinco BACs sequenciados. Isto devido a que os iniciadores usados para amplificar estes transcritos só amplificam um dos *loci*. Não foi avaliada a expressão do outro *locus* da linhagem 074 neste trabalho.

Os elementos da linhagem 074 se comportam como outros descritos na literatura como domesticados, como são o elemento DAYSLEEPER ou os elementos da uva do vinho; Vinesleeper 1-5 (Benjak et al., 2008; Bundock e Hooykaas em 2005). Para Vinesleeper 1 as sequências expressas foram encontradas num único tecido, flor. Em tanto, Vinesleeper 2 se expressa em diferentes tecidos, em diferentes momentos do desenvolvimento da planta. Ambos os padrões são compatíveis com genes relacionados a desenvolvimento (Benjak et al., 2008).

4.5 Outros Elementos hAT no Genoma da Cana-de-açúcar

Cinco elementos completos relacionados à superfamília SChAT que não fazem partes das linhagens 074, 191 e 257 foram achados quando foi feita a procura com o algoritmo hmmer 3.0 (Finn et al., 2011) usando os modelos hmm construídos para o domínio de dimerização (Pfam PF05699) e o domínio *Zinc Finger* (Pfam PF02892) nos 270 BACs de cana-de-açúcar sequenciados pelo GaTE-Lab. A estrutura destes elementos foi analisada (Quadro 12).

Quadro 12 - Descrição dos outros elementos hAT.

Elemento	Tamanho (pb)	TATA	ORF (pb)	Sequências intrônicas	Identidade com Ac	Cobertura
SChAT001L05	5276	TATATAA	2265	2	32%	78%
SChAT007C22	4783	N.I	1887	4	32%	82%
SChAT028C07	4183	N.I	2001	2	38%	79%
SChAT137O14	5246	TTATTT	2109	2	35%	76%
SChAT203M05	11795	N.I	1866	4	50%	77%

Fonte: Ochoa (2012).

As ORFs destes elementos não apresentam códons de parada internos. Apresentam diferentes números de sequências intrônicas. Para dois foi possível identificar o TATA da sequência promotora, um dos quais é igual ao identificado antes para os elementos antes descritos (Quadro 12).

O elemento SChAT028C07 poderia conformar uma terceira linhagem de transposons, além das duas anteriormente estudadas (191 e 257). Esta linhagem possui transcritos relacionados e apresenta todos os domínios e aminoácidos com exceção do bloco A (vide anexo A). Para este elemento também foram encontradas TIRs com TSDs (Quadro 6).

Os demais elementos que não estão relacionados a transcritos do SUCEST são; SChAT001L05, SChAT007C22, SChAT137O14 e SChAT203M05, para estes foram encontrados todos os domínios e aminoácidos necessários para serem ativos (vide anexo A). Também foram encontradas TIRs com TSDs para dois destes elementos; SChAT007C22 e SChAT203M05 (Quadro 6). O que poderia indicar que estes elementos estão completos e poderiam ser ativos, por outro lado a falta de transcritos relacionados no SUCEST poderia indicar que existe uma regulação do genoma sobre estes elementos para não serem expressos, ou que simplesmente o SUCEST não possui transcritos das condições em que estes elementos são expressos.

4.6 Elementos Relacionados em Outros Genomas

Finalmente, foram procurados elementos relacionados às linhagens 191 e 257 de transposons em outros genomas. Foram usadas as sequências dos cinco elementos destas duas linhagens para procurar nos genomas de *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* e *Zea Mays*. A figura 27 apresenta filogenia simplificada da família Poaceae (modificado de Bouchenak et al., 2008).

Foram encontrados elementos relacionados à família SChAT, linhagem 191 e 257 unicamente no genoma de *Sorghum bicolor* (Quadro 13). A existência destes ortólogos é corroborada na figura 8, onde são observadas pelo menos duas bandas na canaleta contendo DNA genômico de sorgo hibridado com as sondas de cada uma destas linhagens.

Figura 27 - Filogenia simplificada da família *Poaceae*

Fonte: Modificado de Bouchenak et al. (2008).

No caso da linhagem 191 foram encontrados dois ortólogos no mesmo cromossomo em regiões distantes. Na linhagem 257 foram encontrados três ortólogos em diferentes cromossomos (Quadro 13).

Quadro 13 - Ortólogos da família SChAT em *Sorghum bicolor*.

	Organismo	Cromossomo	Posição (pb)	Tamanho (pb)	Identidade %	Nome
Linhagem 191	<i>Sorghum bicolor</i>	4	2555868	2048	91	SBhAT1
Linhagem 191	<i>Sorghum bicolor</i>	4	58113860	945	89	SBhAT2
Linhagem 257	<i>Sorghum bicolor</i>	7	10035157	1511	93	SBhAT3
Linhagem 257	<i>Sorghum bicolor</i>	2	5410182	2020	83	SBhAT4
Linhagem 257	<i>Sorghum bicolor</i>	5	12977246	780	92	SBhAT5

A identidade apresentada foi calculada com relação ao elemento canônico de cada linhagem, SChAT160F01 e SChAT152D05_A.

Fonte: Ochoa (2012).

Os ortólogos das duas linhagens cobrem diferentes regiões ao longo dos elementos da família SChAT de cada linhagem. A hipótese de separação do grupo II da família SChAT em pelo menos duas linhagens de elementos de transposição é apoiada pela análise de ortólogos, que apresenta diferentes ortólogos para cada linhagem proposta. Indicando diferentes histórias de amplificação nos genomas destas duas gramíneas. Não foram encontrados ortólogos para estas duas linhagens

em outras gramíneas, corroborando resultados de *Southern Blot* De Jesus *et al.* (2012).

O fato de encontrar elementos relacionados à família SChAT em sorgo, proporciona indícios do aparecimento da família ao longo da história evolutiva destas duas gramíneas. A família SChAT poderia ter um parental comum compartilhado com os elementos achados em sorgo, razão pela qual são chamados de linhagens de elementos de transposição, presente desde antes da divergência destas duas espécies vegetais. A família SChAT poderia não ter tido êxito de replicação no sorgo, em tanto em cana obteve um numero de cópias concordante ao encontrado em outras plantas para elementos relacionados com a superfamília hAT.

Também foram realizadas buscas no genoma de *Miscanthus giganteus* (Swaminathan K. *et al.*, 2010), gramínea altamente relacionada com cana e sorgo, da qual foi sequenciado 1,2% do genoma. Foi usado o algoritmo de BLASTN da pagina do NCBI. Não foram encontrados leituras homologas aos elementos da família SChAT, o que era esperado, devido à pequena porcentagem do genoma que foi sequenciado até o momento.

Nenhum dos ortólogos (Quadro 13) foi encontrado na região sintênica dos BACs até aqui estudados para ao genoma de sorgo. Isto poderia indicar que a transposição destes elementos da família SChAT aconteceu apos a divergência destas duas espécies de gramíneas faz oito a nove milhões de anos.

As OFRs dos cinco elementos achados em sorgo foram preditas por alinhamento com Ac. Os elementos SBhAT2 e SBhAT4 possuem uma mudança de fase de leitura e o elemento SBhAT5 tem uma grande deleção interna quando comparado com a transposase de Ac, que origina também uma mudança de fase de leitura. Uma vez ajustadas as ORFs foi gerada uma comparação das transposases destes elementos (Figura 28 e 29).

Na figura 28 pode ser observado que entre os elementos de sorgo unicamente SBhAT1 e SBhAT4 possuem o domínio *Zinc Finger*, estes elementos também possuem o domínio hATC, assim como as regiões A-F. Destes dois elementos somente SBhAT1 possui completa a assinatura DDE. SBhAT4 só possui o segundo D da assinatura. Estas duas transposases são potencialmente funcionais, principalmente a do elemento SBhAT1 que é ortólogo da linhagem 191. Nenhum dos outros elementos possui assinatura DDE completa. O segundo D é encontrado em todos os elementos com exceção do elemento SBhAT2, no entanto a região E é

encontrada somente nos elementos SBhAT1 e SBhAT2. O domínio hATC não é encontrado em SBhAT3 nem SBhAT5. SBhAT2 não possui as regiões A nem B. Isto sugere que SBhAT2, SBhAT3 e SBhAT5 não possuem transposases ativas.

Figura 28 - Alinhamento das transposases dos cinco ortólogos da família SChAT encontrados em *Sorghum bicolor* com a transposase do elemento SChAT160F01.

Com setas vermelhas é indicada a assinatura DDE, também são indicadas as seis regiões conservadas A-F e os domínios conservados Zinc Finger (Caixa Azul) e hATC (Caixa Verde).

Fonte: Ochoa (2012).

Finalmente, o cladograma da figura 29 mostra como as duas linhagens da família SChAT e seus ortólogos se agrupam em clados diferentes, sustentando a hipótese de uma historia evolutiva diferente destas duas linhagens nos genomas destas duas gramíneas altamente relacionadas

Figura 29 - Cladograma das transposases SChAT, transposase do elemento Ac e transposases ortólogas à linhagem 191 e 257, encontrada em *Sorghum bicolor*.

Cladograma feito por *Neighbor-Joining*. As distancias evolutivas foram computadas usando *Poisson correction* com a opção *Pairwise deletion*.
Fonte: Ochoa (2012).

5 CONCLUSÕES

Nos capítulos anteriores a estrutura dos elementos pertencentes à superfamília hAT presentes no genoma da cana-de-açúcar foi caracterizada. Foi estudado o padrão de agrupamento de suas sequências, bem como a distribuição no genoma da cana e de outras gramíneas. A presença no genoma da cana de açúcar de pelo menos três linhagens relacionadas a transcritos do SUCST da família hAT é confirmada pelos resultados de sequenciamento de BACs e filogenia.

Os resultados apresentados sugerem que ao menos dois dos elementos de transposição das linhagens de transposons caracterizadas poderiam conservar a capacidade autônoma de transposição, SChAT160F01 e SChAT170C07. Três poderiam conservar a capacidade de transposição, de forma não autônoma (SChAT167G07, SChAT185O22, SChAT152D05). Os resultados da análise de sintenia aqui apresentados sugerem que os elementos da família SChAT, das linhagens 191 e 257 apresentam uma atividade recente de transposição no genoma da cana-de-açúcar.

Os elementos da linhagem 257 têm um alto potencial de serem elementos ativos no genoma, ou pelo menos de ter transposases funcionais. Possuindo as regiões A-F, assim como os domínios conservados *Zinc finger* e hATC, as assinaturas DDE e o motivo CXXH da região B. Para as duas versões de elementos da linhagem 257, não foram encontrados ortólogos na mesma região de sintenia com o genoma de sorgo. Isto sugere uma transposição relativamente recente, posterior à divergência de sorgo e cana-de-açúcar, há aproximadamente oito milhões de anos. O que concorda com as predições feitas sobre suas transposases, postulando-lhes como elementos ativos, e autônomo no caso do SChAT170C07. O TE257 parece corresponder a um transcrito capaz de ser traduzido em uma transposase ativa da linhagem 257, razão pela qual esta transposase foi expressa visando seu uso em análises funcionais que num futuro validem as propostas de função e estrutura feitas por este trabalho.

Para a linhagem domesticada foi observado que as versões encontradas no genoma da cana-de-açúcar pertenceriam a sete variantes alélicas do mesmo *locus*. As análises de sintenia apresentam evidência de imobilização desde antes da divergência de sorgo, milho e cana, 12MYA. O fato de ter encontrado dois *loci* da linhagem domesticada de SChAT relacionados com DAYSLEEPER, um gene de

cópia única, sugere que este gene poderia ter sido duplicado durante a evolução das gramíneas podendo dar origem a uma diferenciação de funções. Finalmente, os níveis de expressão verificados para esta linhagem concordam com o esperado para elementos domesticados.

Este trabalho ampliou a visão da diversificação da família SChAT na cana-de-açúcar apresentando novos transposons não relacionados com as duas linhagens inicialmente propostas (191 e 257), alguns destes elementos apresentam a possibilidade de ser ativos. Para quatro destes elementos não se tem evidência de transcrição. Esta descoberta deixa a porta aberta para uma posterior caracterização funcional destes elementos que poderiam estar sendo fortemente regulados pelo genoma. Em contraste, o elemento SChAT028C07 possui transcritos relacionados e poderia conformar uma terceira linhagem de transposons.

Nos BACs estudados do genoma da cana-de-açúcar foi evidente o alto impacto dos retrotransposons no aumento do tamanho das regiões genômicas, assim como no silenciamento de genes.

A análise comparativa entre os BACs de cana e os genomas de outras gramíneas mostra uma maior divergência na organização genômica entre cana e milho. Enquanto estas duas espécies são mais correlacionadas filogeneticamente, *Brachypodium*, uma espécie mais distante, apresentou uma organização genômica mais similar a do genoma da cana. Isto pode ser devido ao alto conteúdo de elementos de transposição em milho 85% (Schnable et al., 2009), os quais poderiam gerar um impacto significativo em fenômenos como sequestro de genes (Pack-MULES e Helitrons), aumento do tamanho do genoma (Retrotransposons) e geração de novos genes (caso do elemento MUSTANG) após a divergência destas gramíneas.

REFERÊNCIAS

- Agrawal A, Eastman QM, Schatz DG. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, 1998; 394(6695):744–751. doi:10.1038/29457
- Amalraj VA, Balasundaram N. On the taxonomy of the members of 'Saccharum complex'. *Genetic Resources and Crop Evolution*, 2006; 53:35-41. doi: 10.1007/s10722-004-0581-1.
- Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, SanMiguel PJ, Jiang N, Jackson SA, Panaud O, Wing RA. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *The Plant Journal*, 2007; 52(2):342–351. doi: 10.1111/j.1365-313X.2007.03242.x
- Araujo PG, Rossi M, Jesus EM, Saccarojr NL, Kajihara D, Massa R, De Felix JM, Drummond RD, Falco MC, Chabregas SM, Ulian EC, Menossi M, Van Sluys MA. Transcriptionally active transposable elements in recent hybrid sugarcane. *The Plant Journal*, 2005; 44(5):707-717. doi: 10.1111/j.1365-313X.2005.02579.x
- Asnaghi C, Paulet F, Kaye C, Grivet L, Glaszmann JC, D'hont A. Application of synteny across the *Poaceae* to determine the map location of a rust resistance gene of sugarcane. *Theoretical and Applied Genetics*, 2000; 101(5-6):962–969. doi: 10.1007/s001220051568
- Baker B, Schell J, Lorz H, Fedoroff N. Transposition of the maize controlling element "Activator" in tobacco. *Proceedings of the National Academy of Sciences of the United States of America*, 1986; 83(13):4844-4848.
- Becker HA, Kunze R. MAIZE Activator transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. *Molecular and General Genetics*, 1997; 254(3):219-230. doi: 10.1007/s004380050410
- Benjak A, Forneck A, Casacuberta JM. Genome-wide analysis of the "cut-and-paste" transposons of grapevine. *PLoS One*, 2008; 3(9):e3107. doi:10.1371/journal.pone.0003107
- Bennetzen JL. Transposable element contributions to plant genome evolution. *Plant Molecular Biology*, 2000; 42(1):251–269. doi: 10.1023/A:1006344508454
- Bundock P, Hooykaas P. An Arabidopsis hAT-like transposase is essential for plant development. *Nature*, 2005; 436(7048): 282-284. doi:10.1038/nature03667
- Casacuberta JM, Santiago N. Plant LTR-retrotransposons and *MITEs*: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, 2003; 311:1–11. doi: 10.1016/S0378-1119(03)00557-2
- Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proceedings of the National Academy of Sciences of the United States of America*, 2006; 103(21):8101–8106. doi: 10.1073/pnas.0601161103
- Cowan RK, Hoen DR, Schoen DJ, Bureau TE. MUSTANG Is a Novel Family of Domesticated Transposase Genes Found in Diverse Angiosperms. *Molecular Biology and Evolution*, 2005; 22(10):2084–2089. doi: 10.1093/molbev/msi202
- Cuadrado A, Acevedo R, Moreno Díaz De La Espina S, Jouve N, De La Torre C. Genome remodelling in three modern *S. officinarum* x *S. spontaneum* sugarcane cultivars. *Journal of Experimental Botany*, 2004; 55(398):847-54. doi: 10.1093/jxb/erh093

De acordo com:
International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to Biomedical Journal: sample references. Available from <http://www.icmje.org> [2007 May 22].

- De Jesus E, Ochoa Cruz EA, Cruz GM, Van Sluys MA. Diversification of hAT transposase paralogues in the sugarcane genome. *Molecular Genetics and Genomics*, 2012. Epub ahead of print. doi: 10.1007/s00438-011-0670-8
- De Jesus E. Estudo de Dois Grupos de Elementos de Cana-de-Açúcar Homólogos à superfamília *hAT* de Transposons. [tese (Doutorado em Ciências)] São Paulo (Brasil): Instituto de Biociências, Universidade de São Paulo; 2007.
- D'hont A. Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenetic and Genome Research*, 2005; 109(1-3):27-33. doi: 10.1159/000082378
- Droege M, Hill B. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology*, 2008; 136(1-2):3-10. doi: 10.1016/j.jbiotec.2008.03.021
- Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 2004; 5:113. doi: 10.1186/1471-2105-5-113
- Essers L, Adolphs RH, Kunze R. A highly conserved domain of the maize Activator transposase is involved in dimerization. *The Plant Cell*, 2000; 12(2):211-224. doi: 10.1105/tpc.12.2.211
- Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 1998; 8(3):186-194. doi: 10.1101/gr.8.3.186
- Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 1998; 8(3):175-185. doi: 10.1101/gr.8.3.175
- Fedoroff NV, Wessler S, Shure M. Isolation of the transposable maize controlling elements Ac and Ds. *Cell*, 1983; 35(1):235-242. doi:10.1016/0092-8674(83)90226-X
- Feldmar S, Kunze R. The ORFa protein, the putative transposase of maize transposable element Ac, has a basic DNA binding domain. *The EMBO Journal*, 1991; 10(13):4003-4010.
- Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, 2002; 3(5):329-341. doi:10.1038/nrg793
- Feschotte C, Pritham EJ. Computational analysis and paleogenomics of interspersed repeats in eukaryotes. In: Stojanovic N, editor. *Computational genomics*. Norwich (UK): Horizon Scientific Press, 2007; 31-53.
- Feschotte C. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review Genetics*, 2007; 41:331-368. doi: 10.1146/annurev.genet.40.110405.090448.
- Feschotte C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 2008; 9(5):397-405. doi:10.1038/nrg2337
- Filée J, Siguier P, Chandler M. Insertion Sequence Diversity in Archaea. *Microbiology and Molecular Biology Reviews*, 2007; 71(1):121-157. doi: 10.1128/MMBR.00031-06
- Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 2011; 39(Web Server Issue):W29-W37. doi: 10.1093/nar/gkr367
- Finnegan DJ. Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, 1989; 5(4):103-107. doi: 10.1016/0168-9525(89)90039-5
- Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. *Genome Research*, 1998; 8(3):195-202. doi: 10.1101/gr.8.3.195

- Grandbastien MA, Audeon C, Bonnivard E, Casacuberta JM, Chalhou B, Costa AP, Le QH, Melayah D, Petit M, Poncet C, Tam SM, Van Sluys MA, Mhiri C. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic and Genome Research*, 2005; 110(1-4): 229–241. doi: 10.1159/000084957
- Grandbastien MA. Activation of plant retrotransposons under stress conditions. *Trends in Plant Science*, 1998; 3(5):181–187. doi: 10.1016/S1360-1385(98)01232-1
- Gregory TR. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics*, 2005; 6(9):699–708. doi:10.1038/nrg1674
- Grivet L, Arruda P. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology*, 2002; 5(2):122–127. doi: 10.1016/S1369-5266(02)00234-0
- Grivet L, Daniels C, Glaszmann JC, D'hont A. A review of recent molecular genetics evidence for sugarcane evolution and domestication. *Ethnobotany Research & Applications*, 2004; 2:9–17.
- Hanahan, D. Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology*, 1983; 166(4):557–580. doi: 10.1016/S0022-2836(83)80284-8
- Haren L, Ton-Hoang B, Chandler M. Integrating DNA: Transposases and Retroviral Integrases. *Annual Review of Microbiology*, 1999; 53:245–281. doi: 10.1146/annurev.micro.53.1.245
- Hoarau JY, Grivet L, Offmann B, Raboin LM, Diorflar JP, Payet J, Hellmann M, D'hont A, Glaszmann JC. Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). Detection of QTLs for yield components. *Theoretical and Applied Genetics*, 2002; 105(5-6):1027–1037. doi: 10.1007/s00122-002-1047-5
- Hudson ME, Lisch DR, Quail PH. The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. *The Plant Journal*, 2003; 34(4):453–471. doi: 10.1046/j.1365-313X.2003.01741.x
- Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences of the United States of America*, 2000; 97(12):6603–6607. doi: 10.1073/pnas.110587497
- Kempken F, Windhofer F. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma*, 2001; 110(1):1–9. doi: 10.1007/s004120000118
- Kumar S, Tamura K, Nei M. Mega3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics*, 2004; 5(2):150–163. doi: 10.1093/bib/5.2.150
- Kunze R, Behrens U, Courage-Franzkowiak U, Feldmar S, Kuhn S, Lutticke R. Genetics Dominant transposition-deficient mutants of maize Activator (Ac) transposase (Zea mays/transient expression/petunia/negative complementation/transposable element). *Proceedings of the National Academy of Sciences of the United States of America*, 1993; 90(15):7094–7098.
- Kunze R, Saedler H, Lonig WE. Plant transposable elements. *Advances in Botanical Research*, 1997; 27:331–469.
- Kunze R, Stochaj U, Laufs J, Starlinger P. Transcription of transposable element Activator (Ac) of *Zea mays* L. *The EMBO Journal*, 1987; 6(6):1555–1563.
- Li W, Zhang P, Fellers JP, Friebe B, Gill BS. Sequence composition, organization and evolution of the core *Triticeae* genome. *The Plant Journal*, 2004; 40(4):500–511. doi: 10.1111/j.1365-313X.2004.02228.x

- Macas J, Neumann P, Pozárková D. Zaba: a novel miniature transposable element present in genomes of legume plants. *Molecular Genetics and Genomics*, 2003; 269(5):624–631. doi: 10.1007/s00438-003-0869-4
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005; 437(7057): 376-380. Erratum in: *Nature*, 2006; 441(7089):120. doi: 10.1038/nature03959
- Mcclintock B. Mutable loci in maize. *Carnegie Inst. Washington Yearbook*, 1948; 47:155-169.
- Mcclintock B. The significance of responses of the genome to challenge. *Science*, 1984; 226(4676):792-801. doi: 10.1126/science.15739260
- Menossi M, Silva-Filho MC, Vincentz M, Van Sluys MA, Souza GM. Sugarcane functional genomics: gene discovery for agronomic trait development. *International Journal of Plant Genomics*, 2008; ID 458732. doi: 10.1155/2008/458732
- Meyers BC, Tingey SV, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research*, 2001; 11(10):1660-1676. doi: 10.1101/gr.188201
- Ming R, Liu SC, Moore PH, Irvine JE, Paterson A. H. QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. *Genome Research*, 2001; 11(12):2075-2084. doi: 10.1101/gr.198801
- Monteiro-Vitorello CB, De Oliveira MC, Zerillo MM, Varani AM, Civerolo E, Van Sluys MA. Xylella and Xanthomonas Mobil'omics. *OMICS: A Journal of Integrative Biology*, 2005; 9(2):146-159. doi:10.1089/omi.2005.9.146
- Moreno-Vázquez S, Ning J, Meyers BC. *hATpin*, a family of MITE-like *hAT* mobile elements conserved in diverse plant species that forms highly stable secondary structures. *Plant Molecular Biology*, 2005; 58(6):869–886. doi: 10.1007/s11103-005-8271-8
- Muehlbauer GJ, Bhau BS, Syed NH, Heinen S, Cho S, Marshall D, Pateyron S, Buisine N, Chalhoub B, Flavell AJ. A hAT superfamily transposase recruited by the cereal grass genome. *Molecular Genetics and Genomics*, 2006; 275(6):553-563. doi: 10.1007/s00438-006-0098-8
- Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, Inoue H, Tanisaka T. Mobilization of a transposon in the rice genome. *Nature*, 2003; 421(6919):170–172. doi: 10.1038/nature01219
- Ochoa Cruz EA. Caracterização de Transposases da Família SChAT em cana-de-açúcar: estudo molecular e funcional. [Tese (Doutorado em Biotecnologia)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2012.
- Okamoto H, Hirochika H. Silencing of transposable elements in plants. *Trends in Plant Science*, 2001; 6(11):527–534. doi: 10.1016/S1360-1385(01)02105-7
- Ouyang S, Buell CR. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, 2004; 32:D360–D363. doi: 10.1093/nar/gkh099
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otiillar RP, Penning

- BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, Mccann MC, Ming R, Peterson DG, Mehboob-Ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 2009; 457(7229): 551-556. doi: 10.1038/nature07723
- Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research*, 2006; 16(10):1262–1269. doi: 10.1101/gr.5290206
- Piriyapongsa J, Jordan IK. A Family of Human *MicroRNA* Genes from Miniature Inverted-Repeat Transposable Elements. *PLoS One*, 2007; 2(2):e203. doi: 10.1371/journal.pone.0000203
- Pohlman RF, Fedoroff NV, Messing J. The nucleotide sequence of the maize controlling element activator. *Cell*, 1984; 37(2):635-643. doi: 10.1016/0092-8674(84)90395-7
- Rossi M, Araujo PG, Jesus EM, Varani AM, Van Sluys MA. Comparative analysis of Mutator-like transposases in sugarcane. *Molecular Genetics and Genomics*, 2004; 272(2):194–203. doi: 10.1007/s00438-004-1036-2
- Rossi M, Araujo PG, Van Sluys MA. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). *Genetics and Molecular Biology*, 2001; 24(1-4):147–154. doi: 10.1590/S1415-47572001000100020
- Rubin E, Lithwick G, Levy AA. Structure and Evolution of the hAT Transposon Superfamily. *Genetics*, 2001; 158(3):949–957.
- Rudenko GN, Ono A, Walbot V. Initiation of silencing of maize MuDR/Mu transposable elements. *Plant Journal*, 2003; 33(6):1013–1025. doi: 10.1046/j.1365-3113X.2003.01683.x
- Rudenko GN, Walbot V. Expression and post-transcriptional regulation of maize transposable elements MuDR and its derivatives. *Plant Cell*, 2001; 13(3): 553–570. doi: 10.1105/tpc.13.3.553
- Saccaro NL Jr, Van Sluys MA, De Mello Varani A, Rossi M. MudrA-like sequences from rice and sugarcane cluster as two bona fide transposon clades and two domesticated transposases. *Gene*, 2007; 392(1-2):117-25. doi: 10.1016/j.gene.2006.11.017
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 1987; 4(4):406-425.
- Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 1977; 74:5463–5467.
- Santiago N, Herraiz C, Goñi JR, Messeguer X, Casacuberta JM. Genome-wide analysis of the Emigrant family of MITEs: of *Arabidopsis thaliana*. *Journal of Molecular Biology and Evolution*, 2002; 19(12):2285-2293.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddelloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel

- DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, 2009; 326: 1112.
- Shi Q, Jackowski G. One-dimensional polyacrylamide gel electrophoresis. In: Hames BD. *Gel electrophoresis of proteins: a practical approach*. 3rd edition. Oxford: Oxford University Press; 1998. Cap.1, p. 1-52.
- Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. *Trends in Genetics*, 2005; 21:322–326. doi: 10.1016/j.tig.2005.04.008
- Stewart CNJr, Via LE. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *BioTechniques*, 1993; 14:748-51.
- Swaminathan K, Alabady MS, Varala K, De Paoli E, Ho I, Rokhsar DS, Arumuganathan AK, Ming R, Green PJ, Meyers BC, Moose SP, Hudson ME. Genomic and small RNA sequencing of *Miscanthus x giganteus* shows the utility of sorghum as a reference genome sequence for Andropogoneae grasses. *Genome Biology*, 2010; 11(2): R12. doi: 10.1186/gb-2010-11-2-r12
- Takahashi S, Furukawa T, Asano T, Terajima Y, Shimada H, Sugimoto A, Kadowaki K. Very close relationship of the chloroplast genomes among *Saccharum* species. *Theoretical and Applied Genetics*, 2005; 110(8):1523–1529. doi: 10.1007/s00122-005-1990-z
- Takeda S, Sugimoto K, Otsuki H, Hirochika H. Transcriptional activation of the tobacco retrotransposon Tto1 by wounding and methyl jasmonate. *Plant Molecular Biology*, 1998 36(3):365–376. doi: 10.1023/A:1005911413528
- Tan S, Wu W, Liu J, Kong Y, Pu Y, Yuan R. Efficient expression and secretion of recombinant hirudin III in *E. coli* using the L-asparaginase II signal sequence. *Protein Expression and Purification*, 2002; 25(3):430–436. doi: 10.1016/S1046-5928(02)00023-2
- Tomkins JP, Yu Y, Miller-Smith H, Frisch DA, Woo SS, Wing RA. A bacterial artificial chromosome library for sugarcane. *Theoretical and Applied Genetics*, 1999; 99:419–24.
- Toro N. Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environmental Microbiology*, 2003; 5(3):143–151. doi: 10.1046/j.1462-2920.2003.00398.x
- Van Sluys MA, Tempé J, Fedoroff N. Studies on the introduction and mobility of the maize Activator element in *Arabidopsis thaliana* and *Daucus carota*. *EMBO Journal*, 1987; 6(13):3881–3889.
- Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, 2007; 2(12):e1326. doi: 10.1371/journal.pone.0001326
- Vettore AL, Da Silva FR, Kemper EL, Arruda P. The libraries that made SUCEST. *Genetics and Molecular Biology*, 2001; 24(1-4):1-7. Doi: 10.1590/S1415-47572001000100002
- Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Gigliotti EA, Lemos MV, Coutinho LL, Nobrega MP, Carrer H, França SC, Bacci Júnior M, Goldman MH, Gomes SL, Nunes LR, Camargo LE, Siqueira WJ, Van Sluys MA, Thiemann OH, Kuramae EE, Santelli RV, Marino CL, Targon ML, Ferro JA, Silveira HC, Marini DC, Lemos EG, Monteiro-Vitorello CB, Tambor JH, Carraro DM, Roberto PG, Martins VG, Goldman GH, de Oliveira RC, Truffi D, Colombo CA, Rossi

- M, de Araujo PG, Sculaccio SA, Angella A, Lima MM, de Rosa Júnior VE, Siviero F, Coscrato VE, Machado MA, Grivet L, Di Mauro SM, Nobrega FG, Menck CF, Braga MD, Telles GP, Cara FA, Pedrosa G, Meidanis J, Arruda P. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Research*, 2003; 12:2725–2735. doi: 10.1101/gr.1532103
- Vincentz M, Cara FA, Okura VK, da Silva FR, Pedrosa GL, Hemerly AS, Capella AN, Marins M, Ferreira PC, França SC, Grivet L, Vettore AL, Kemper EL, Burnquist WL, Targon ML, Siqueira WJ, Kuramae EE, Marino CL, Camargo LE, Carrer H, Coutinho LL, Furlan LR, Lemos MV, Nunes LR, Gomes SL, Santelli RV, Goldman MH, Bacci M Jr, Gigliotti EA, Thiemann OH, Silva FH, Van Sluys MA, Nobrega FG, Arruda P, Menck CF. Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiology*, 2004; 134(3):951–959. doi: 10.1104/pp.103.033878
- Wang L, Kunze R. Transposase binding site methylation in the epigenetically inactivated Ac derivative Ds-cy. *Plant Journal*, 1998; 13(4):577–582. doi: 10.1046/j.1365-313X.1998.00060.x
- Weil C, Kunze R. Transposition of maize Ac/Ds transposable elements in the yeast *Saccharomyces cerevisiae*. *Nature Genetics*, 2000; 26(2):187–190. doi: 10.1038/82827
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, Sanmiguel P, Schulman AH. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 2007; 8(12):973–82. doi: 10.1038/nrg2165
- Witte CP, Le QH, Bureau T, Kumar A. Terminal-repeat retrotransposons in miniature (*TRIM*) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 2001; 98(24):13778–13783. doi: 10.1073/pnas.241341898
- Xu Z, Dooner HK. *Mx-rMx*, a family of interacting transposons in the growing hAT superfamily of maize. *Plant Cell*, 2005; 17(2):375–388. doi: 10.1105/tpc.104.027797
- Yang G, Hall TC. MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Research*, 2003; 31(13):3659–3665. doi: 10.1093/nar/gkg531
- Yang G, Zhang F, Hancock CN, Wessler SR. Transposition of the rice miniature inverted repeat transposable element mPing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 2007; 104(26):10962–10967. doi: 10.1073/pnas.0702080104
- Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. indica). *Science*, 2002; 296(5565):79–92. doi: 10.1126/science.1068037
- Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 2003; 31(13):3406–3415. doi: 10.1093/nar/gkg595

[illegible]

	540	560	580	600	
Hobo	TLKSACFTR--WNSNYKMSKIDNRVSDYKILGEAD	TH--VDFNKSSL-KVVVDIIGDFERIFKKLTSSSP	IC		428
Tam3	HYLDVHR--WNYAYRMTIERKP--SEPIFFCHRSLS	GP--NSLLADNINIC-DILVYVYVYEEFTKIMSSCTYPSN			495
Days leeper	VLSLDDCTQ--WNTYLMIVARSE-LKEVFSCLDTAD-DLKK	PP--SAEDRKHV-EALCTFLKPLFAVSTLQSTGNPSAV			464
SChAT124F07	XNCYTIILHIASS--	QOQTOLTS			340
SChAT92C13	TLGLDVTIQ--WNTYLMILAAID-YKQFTTILETCDD-N	NE-AP--SAEDRKKV-EAACNYKLLYDSAHSIMAAANP	AN		410
SChAT188 F16	TLGLDVTIQ--WNTYLMILAAID-YKQFTTILETCDD-N	NE-AP--SAEDRKKV-EAACNYKLLYDSAHSIMAAANP	AN		473
SChAT171 I09	TLGLDVTIQ--WNTYLMILAAID-YKQFTTILETCDD-N	NE-AP--SAEDRKKV-EAACNYKLLYDSAHSIMAAANP	AN		473
SChAT037 O20	TLGLDVTIQ--WNTYLMILAAID-YKQFTTILETCDD-N	NE-AP--SAEDRKKV-EAACNYKLLYDSAHSIMAAANP	AN		473
SChAT099 K03	TLGLDVTIQ--WNTYLMILAAID-YKQFTTILETCDD-N	NE-AP--SAEDRKKV-EAACNYKLLYDSAHSIMAAANP	AN		473
SChAT145 D15	TLGLDVTIQ--WNTYLMILAAID-YKQFTTILETCDD-N	NE-AP--SAEDRKKV-EAACNYKLLYDSAHSIMAAANP	AN		473
Ac	GISYDVSTR--WNSYLDLRDAIY-YKPALIRLKTSPDRR	DAICP--KAEEKKMA-LITFKCKKFFDLTELLSGTQYSAN			530
SChAT152 D05A	GLVLVDVHR--WNAAYDILNEAIK-YKAALNRFAAEQ	--QD-VP--SEDDQKA-ESLHEPEQFSEATKAFSADRHP	AH		404
SChAT170 C07	GLILVDVHR--WNAAYDILNEAIK-YKAALNRFAAEQ	--QD-VP--SEDDQKA-ESLHEPEQFSEATKAFSADRHP	AH		404
SChAT185 O22	APQLDINHC--WNSYDILDSAEI-FKEAFLELVKVDK-S	TY-CP--SSQDEERA-SAVCKLVKFKKAMEVSGTSYPSN			111
SChAT167 G07	VPKIDVENH--WNSCDLISAP-FKEAFLELVKVDK-K	TY-CP--SSQDEERA-NAVCKLVKFKKAMEVSGTSYPSN			387
SChAT160 F01	VPKIDVENH--WNSCDLISAP-FKEAFLELVKVDK-N	SY-CP--SSQDEERA-SAVCKLVKFKKAMEVSGTSYPSN			389
SChATG1	VPKIDVENH--WNSCDLISAP-FKEAFLELVKVDK-N	SY-CP--SSQDEERA-NAVCKLVKFKKAMEVSGTSYPSN			389
SChAT028 C07	KLGLDCHTR--WNSYKMSKISIAIP-YKAVFNRAHVEK-L	DC-CP--SEEEEFARREVAGRLKFNIDITEVFSRTNYV	AN		442
SChAT203 M05	KLHLDCHTR--WNSYKMSKISIAIP-YKAVFELGLDR-N	DC-CP--AADDITFA-TNCDCKGIFYELLTELFSGTQYV	AN		361
SChAT007 C22	RIALDCHTR--WNSYTMILITTV--QDVFERLARGP	--C-VP--LAEQKKA-KELCERKMFYDTELLSGTKY	--		434
SChAT137 O14	RPSLDVHTR--WNTYLMISKATE-YRAVFDVMSQDP-N	ID-KP--SNTDQMA-DLLCNVFKPFYDATNVSGTLGY	DTN		488
SChAT001 L05	TFGLDMVTR--WNSYLMILKHVP-HGEIHFHTFIES--	N-P--RP--RGGAFLLLTDHFMV-TKLKFLFLFYDSTVALSGVYYP	AL		487

	620	*	640	*	660	*	680	*	700																																				
Hobo	LE	FVLPSISK	LE	CEPNILDSAAAL	KERTILENIRKIMMANIS	TWHKAAFFLYPPAAH	QEEDEL	LE	KVFCISQIQVPS	ISYTSLS	: 515																																		
Tam3	IMLL	YMYSV	VR	LFQHNHRHA	--TLKN	IKKEKKVKYKKVPNV	CLSSCL	--DP	RV	LIGTLELEKXHSALNNV	YNG--EER	: 575																																	
Daysleeper	TF	FEHVEWTKQSD	SR	ATAGEEDP	PTVTK	AKTQVQK	YDKYWRDCSLVLA	AVVM	--DP	RFK	MKLVFEFSFKIFG	EDAGK--NI	: 541																																
SchAt124F07	--	FFHEAKK	QLE	LANGTGH	--SLAKD	HERPDR	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEYSYKYGAEAAK	--YV	: 410																																
SchAt92C13	IFF	HETWKQ	QLE	LANVTGH	--SLAKD	HERPDR	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEFSYSKNYGAEAAK	--YV	: 481																																
SchAt188_F16	IFF	HEAWK	QLE	LANGTGHE	DP	TFSS	AKD	HERPDK	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEFSYSKYGAEAAK	--YV	: 550																														
SchAt171_I09	IFF	HEAWK	QLE	LANGTGHE	DP	TFSS	AKD	HERPDK	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEFSYSKYGAEAAK	--YV	: 550																														
SchAt037_O20	IFF	HEAWK	QLE	LANGTGHE	DP	TFSS	AKD	HERPDK	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEFSYSKYGAEAAK	--YV	: 550																														
SchAt099_K03	IFF	HEAWK	QLE	LANGTGHE	DP	TFSS	AKD	HERPDK	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEFSYSKYGAEAAK	--YV	: 550																														
SchAt145_D15	IFF	HEAWK	QLE	LANGTGHE	DP	TFSS	AKD	HERPDK	WKDCSLVLA	AVVM	--DP	RFK	MKLVFEFSYSKYGAEAAK	--YV	: 550																														
Ac	LY	KGFC	E	KDL	IQDWCE	HKFVIR	RAVA	SEK	EKKYKVSNTALA	ACFL	--DP	RYK	KILTEFYMKK	FHGDSYKV--HV	: 607																														
SchAt152_D05A	LE	FMKLMA	RV	LD	ETWNTN	ELLNE	LEA	AT	QKQW	WAAPS	SVMLTA	IAVL	--DP	SMK	ADP	RFFLYLTVENTEA	EAK--KM	: 481																											
SchAt170_C07	LE	FMKLMA	RV	LD	ETWNTN	ELLNE	LEA	AT	QKQW	WAAPS	SVMLTA	IAVL	--DP	SMK	ADP	RFFLYLTVENTEA	EAK--KM	: 481																											
SchAt185_O22	LY	FHETSSR	--	GRFS	NETIAS	VEF	QON	DK	KMLSYLT	NPV	PIVL	--DP	RFK	YGF	TEFLRKQAFGE	YGSV--PHL	: 184																												
SchAt167_G07	LY	FHEIWS	KQV	VEE	EAFSPNE	II	IVA	VSK	QK	IDK	KMLSYLT	NPV	PIVL	--DP	RFK	YGF	TEFLRKQAFGE	YGSV--PHL	: 465																										
SchAt160_F01	LY	FHEIWS	KQV	VEE	EAFSPNE	II	IVA	VSK	QK	IDK	KMLSYLT	NPV	PIVL	--DP	RFK	YGF	TEFLRKQAFGE	YGSV--PHL	: 467																										
SchAtG1	LY	FHEIWS	KQV	VEE	EAFSPNE	II	IVA	VSK	QK	IDK	KMLSYLT	NPV	PIVL	--DP	RFK	YGF	TEFLRKQAFGE	YGSV--PHL	: 467																										
SchAt028_C07	IQ	LLKICE	AK	EQ	RQWAV	CNP	II	EQ	SF	EM	KE	DK	YK	IQ	GS	ML	AT	IVL	--DP	RFK	IDP	LVG	FLETT	EGSH	LECLEK	--YV	: 522																		
SchAt203_M05	MF	FPKVR	E	KLK	NSWGH	ENET	IRNKSAA	IQ	Y	KD	W	KL	DI	GLMA	IAVL	--DP	RLK	MTM	HACYALTE	GE	VDG--TYV	: 439																							
SchAt007_C22	--	--	--	--	--	--	--	--	KPNK	Y	TS	D	VH	GLMA	IAVL	--DP	RYK	LHL	NALF	SO--NS	: 470																								
SchAt137_O14	RC	FHVLW	E	W	KG	ET	LES	HT	DIS	TV	AK	K	L	K	E	K	W	IC	ML	Q	IC	VP	IVL	--DP	RFK	F	NS	F	RL	NAG	F	G	D	K	G	T	A	--YV	: 565						
SchAt001_L05	LM	VH	IV	IR	AT	HK	--YON	DE	H	IR	V	Q	M	D	K	Y	N	K	Y	AT	IP	PL	Y	AT	IVL	--DP	AKG	F	N	K	L	R	K	L	Q	L	T	D	A	D	Y	S	N	--K	: 566

[illegible]

Alinhamento das transposases dos 17 elementos SchAT com as transposases de referencia da superfamília hAT; hobo (*Drosophilla melanogaster*), Activator (*Zea mays*) e Tam3 (*Antirrhinum majus*), assim como, a transposase SchATG1 e o elemento domesticado DAYSLEEPER de Arabidopsis. Com setas vermelhas são indicados os aminoácidos conservados DDE, também são indicadas as seis regiões conservadas A-F e os domínios conservados Zinc Finger (Caixa Azul) e hATC (Caixa Verde).

Fonte: Ochoa (2012).

ANEXO B. Diversification of hAT transposase paralogues in the sugarcane genome

Diversification of *hAT* transposase paralogues in the sugarcane genome

Erika M. de Jesus · Edgar A. Ochoa Cruz ·
Guilherme M. Q. Cruz · Marie-Anne Van Sluys

Received: 24 May 2011 / Accepted: 18 December 2011
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract Transposons are abundant components of eukaryotic genomes, and play important role in genome evolution. The knowledge about these elements should contribute to the understanding of their impact on the host genomes. The *hAT* transposon superfamily is one of the best characterized superfamilies in diverse organisms, nevertheless, a detailed study of these elements was never carried in sugarcane. To address this question we analyzed 32 cDNAs similar to that of *hAT* superfamily of transposons previously identified in the sugarcane transcriptome. Our results revealed that these *hAT*-like transposases cluster in one highly homogeneous and other more heterogeneous lineage. We present evidences that support the hypothesis that the highly homogeneous group is a domesticated transposase while the remainder of the lineages are composed of transposon units. The first is common to grasses, clusters significantly with domesticated transposases from *Arabidopsis*, rice and sorghum and is expressed in different tissues of two sugarcane cultivars analyzed. In contrast, the more heterogeneous group represents at least two transposon lineages. We recovered five genomic versions of one lineage, characterizing a novel transposon family with conserved DDE motif, named SChAT. These results indicate the presence of at least three

distinct lineages of *hAT*-like transposase paralogues in sugarcane genome, including a novel transposon family described in *Saccharum* and a domesticated transposase. Taken together, these findings permit to follow the diversification of some *hAT* transposase paralogues in sugarcane, aggregating knowledge about the co-evolution of transposons and their host genomes.

Keywords Sugarcane · Transposon · *hAT* transposase · Domesticated transposase

Introduction

The activity of transposable elements (TE) and their inherent ability to change location within chromosomes results in mutational activity mediating structural and functional changes in genomes. Gene disruption upon insertion, gene capture or modified gene expression patterns are examples of TE-genome interplay (Wessler 2006; Lisch and Bennetzen 2011). These changes provide genetic variability and are thought to influence genome evolution and species formation (McClintock 1984). The recruitment of transposase functional domains to create new host genes exemplifies the importance of these elements as active players in the evolution of the genomes (Britten 2004; Volff 2006; Casola et al. 2007; Feschotte 2008).

The simplest autonomous transposons are DNA sequences that encode a transposase protein responsible for their mobilization in the genome and are flanked by terminal inverted repeats (TIRs). The transposase has a DDE catalytic domain that recognizes the TIRs and promotes the excision and insertion of the element in a different site of the genome (Yuan and Wessler 2011). A non-autonomous element can be mobilized in trans through the recognition

Communicated by M.-A. Grandbastien.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-011-0670-8) contains supplementary material, which is available to authorized users.

E. M. de Jesus · E. A. Ochoa Cruz · G. M. Q. Cruz ·
M.-A. Van Sluys (✉)
GaTE Lab, Departamento de Botânica, Instituto de Biociências,
Universidade de São Paulo, Rua do Matão, 277, Sala 153,
Butantã, 05508-090 São Paulo, SP, Brazil
e-mail: mavsluys@usp.br

of its TIRs by a functional transposase encoded by an autonomous element. The Ac/Ds family comprises autonomous and non-autonomous elements and was originally described as “controlling elements of the gene” by Barbara McClintock in maize (McClintock 1946). Ac is one of the defining members of the *hAT* superfamily of transposons, together with *Tam3* from *Antirrhinum majus* (Hehl et al. 1991), and *hobo* from *Drosophila melanogaster* (Streck et al. 1986; Calvi et al. 1991). *hAT* transposases exhibit conserved domains that have been identified in plants, fungi and animals as well as in humans (Kempken and Windhofer 2001). The most characteristic domains are the zinc finger at the amino terminal region of the protein, and two other conserved domains at the carboxy terminal region, one involved in dimerization (Essers et al. 2000) and other of whose function is still unknown.

The mechanism through which a transposase is incorporated by the genome to the formation of a new gene is called domestication. The *hAT*-like domesticated transposase DAYSLEEPER was described as a factor that binds to the Kubox1 motif present in the upstream region of some genes, including the DNA repair gene *Ku70*, and it is essential for the development of *Arabidopsis thaliana* (Bundock and Hooykaas 2005). Other plant genomes have similar transposase-derived genes such as wheat, barley and rice (Muehlbauer et al. 2006) with detectable expression, but no function has yet been assigned.

Sugarcane is one of the most economically important worldwide crops, providing raw material to the sugar and ethanol industry. Commercial sugarcane cultivars ($2n = 100\text{--}130$) are the recent hybrids with a highly polyploid and aneuploid genome (D’Hont et al. 1996; Grivet et al. 2004). Plants can support large numbers of highly heterogeneous TE families, thought to be due to the elevated frequency of polyploidy (Kraitshtein et al. 2010). In this context, sugarcane is proving to be an interesting model to study TE evolution and expression. Exploring TE diversification and distribution represents an interesting opportunity to unravel new mobile elements and potentially new gene functions.

Previous studies described the expression of 162 fully sequenced sugarcane cDNAs similar to several TE families (Araujo et al. 2005). Of these, 32 were similar to the *hAT* superfamily of transposons (Rubin et al. 2001). The present work describes the diversity, structure and expression of sugarcane *hAT* transposase paralogues. Our results reveal three distinct lineages of *hAT*-like transposase paralogues in sugarcane genome. One lineage presents characteristics that suggest domestication. We also recovered genomic copies of other lineage, identifying a novel transposon family present in *Saccharum*, named SchAT. Analyses of TE content in sugarcane should contribute to the understanding of their impact on the evolution and functioning of polyploid genomes.

Materials and methods

Sequence analyses: clustering of protein sequences

Nucleotide sequences of sugarcane *hAT*-like cDNA clones previously sequenced and annotated in our lab were analyzed. The annotation criterion to characterize it as *hAT*-like was a cutoff of -50 in a BLASTX against the transposase of the element Ac. These cDNAs came from several sugarcane hybrid cultivars (Table S1 in the supplementary material). These sequences were translated to amino acids using the program BioEdit Sequence Alignment Editor (Hall 1999). *hAT*-like sequences from sorghum, rice, maize, tomato, tobacco, grape, arabidopsis and populus were also retrieved from GenBank through BLASTX against Ac transposase aminoacid sequence, and introduced into the alignment. Accession numbers are as follows: Sorghum sequences (1–4): XP_002463929.1, XP_002442948.1, XP_002444199.1 and XP_002455637.1. Rice sequences (1–6): AC134929.2, AK067574, AK102230, AK105085, AK121128 and BAC10751. Maize sequence AY111033. Tomato sequences (Lycopl–3): AK319242.1, BT014568.1 and AC234415.1. Tobacco sequences (1–3): X97569.1, BT014568.1 and AK319242.1. Grape sequences (Vitis1–3): AM487463.2, AM486739.1 and XP_002275062.1. Arabidopsis sequences (Ath1–5): NM_106651, AL137079, AC009322, AC011717 and AAD24567. Populus sequences (1–4): XP_002325202.1, XP_002327524.1, XP_002332734.1 and XP_002331299.1. *hobo* transposase (P12258), Ac transposase (CAA29005), *Tam3* transposase (CAA38906), DAYSLEEPER (AY728267.1). Two conserved motifs were aligned, one containing residues 555–594 of Ac transposase, and the dimerization domain (PFAM accession number PF05699) containing residues 682–719 of Ac transposase. The alignment was carried out using the program ClustalX 1.81 (Thompson et al. 1997), with default multiple alignment parameters. The program MEGA4.0 was used to generate the corresponding trees (Tamura et al. 2007), using the distance method and neighbor joining algorithm, with default parameters. Bootstrap test of 1,000 replicates was made. The GenBank accession numbers of genomic copies SCHAT-G1, SCHAT-G2, SCHAT1, SCHAT2 and SCHAT3 are HM067360, HM067361, HM067362, HM067363 and HM067364, respectively.

Evolutionary analysis

The nucleotide cDNA sequences from the three lineages were aligned separately, as follows: domesticated lineage 074: TE016, TE017, TE035, TE048, TE074, TE096, TE124, TE132, TE203, TE207, TE217, TE223 and TE265. Lineage 191: TE191 and TE221. Lineage 257: TE001,

TE154 and TE257. Sequence alignments were performed using ClustalW (Thompson et al. 1994) in the Bioedit program (Hall 1999) and corrected manually if necessary, removing the nucleotides relative to stop codons or frameshifts. The number of synonymous substitutions per synonymous site (dS) and non-synonymous substitutions per non-synonymous site (dN), the dN:dS ratios and codon-based Z test were calculated through the modified Nei-Gojobori method with the Jukes-Cantor correction, implemented in the MEGA4.0 software (Tamura et al. 2007). Time divergence between the two loci of the domesticated lineage was estimated using $T = dS/2k$, where dS is the estimated number of synonymous substitutions per synonymous site, and k is the average synonymous substitution rate, assuming a rate of 0.0065 mutations per site per MY, as previously reported to the *Adh* loci of grasses (Gaut et al. 1996).

Molecular hybridization

Probes were based on the nucleotide sequences of cDNA clones TE074, TE191 and TE257, and were 182, 187 and 146 bp in length, respectively. To avoid cross hybridization, primers were designed outside the conserved domains to amplify unique probes for each clone. PCR-synthesized probes were labeled with α^{32} dCTP using a Random Primers DNA Labeling System kit (Invitrogen # 18187-013). Molecular hybridizations were performed on 10 μ g of *Xba*I-digested genomic DNA from the following species: potato (*Solanum tuberosum*), orchid (*Catasetum fimbriatum*), pineapple (*Ananas comosus*), rice (*Oryza sativa*), maize (*Zea mays*), two ancestral species of the genus *Saccharum*, *Saccharum spontaneum* (variety Mandalay) and *Saccharum officinarum* (variety Badila), and sugarcane hybrid cultivars SP-89-1115 and SP-80-3280. DNA was extracted using a modified version of the CTAB method (Porebski et al. 1997); 10 units of *Xba*I (New England Biolabs) were used for each 1 μ g of genomic DNA, according to the manufacturer's instructions. DNA was digested at 37°C for 16 h. The hydrolyzed DNA was electrophoresed (1.8 V/cm, 16 h) through a 0.8% agarose gel, in 0.5× TBE buffer. The gel was washed in 0.25 N HCl for 15 min, in 0.4 N NaOH for 15 min, and then transferred by capillary action to nylon membrane (Gene Screen Plus Hybridization Transfer Membrane—Life Science) in 0.4 N NaOH buffer for 16 h. After transfer, membranes were washed in 2× SSC solution for 15 min and dried at room temperature. Membranes were placed for pre-hybridization at 65°C under rotation for 4 h in a solution of 10% Dextran, 1% SDS, 6% NaCl in deionized water. After the addition of the probe the hybridization time was 16 h. Membranes were then washed twice with 0.5× SSC/0.1% SDS for 30 min at 65°C, and twice with

0.1× SSC/0.1% SDS for 30 min at 65°C. The membranes were exposed to X-Ray films (Kodak T-MAT G/RA Film) in cassettes at −70°C; films were developed according to the manufacturer's instructions.

RNA expression analysis

The probes used on RNA blots were TE074 (182 bp) and TE191 (187 bp), i.e., the same probes as for DNA hybridization. A probe for the ubiquitin gene was used as a control for constitutive expression. RNA was extracted from the leaves and roots of sugarcane plants of the variety R570 grown in the greenhouse, from in vitro calli from the variety SP-89-1115. Six-months-old plants grown under field conditions were used for total RNA extraction from apical meristem, stem, mature and immature leaves. RNA was extracted using the lithium chloride method, adapted from Sambrook et al. (1989). To samples containing 10 μ g of RNA, MOPS was added to a final concentration of 1×, formaldehyde 6.5% and formamide 48%. Samples were incubated at 60°C for 15 min, and buffer (bromophenol blue 0.25%, xylene cyanol 0.25%, EDTA 1 mM and glycerol 50%) was added subsequently to 16% and ethidium bromide to 0.35%. Samples were electrophoresed in denaturing agarose gels (1.5% agarose, 6.7% formaldehyde and MOPS 1×) at 80 V for 3 h. After electrophoresis, the gel was washed three times in deionized water for 10 min, and once in 10× SSC for 40 min, followed by capillary transfer to nylon membrane (Gene Screen Plus Hybridization Transfer Membrane—Life Science), in 10× SSC solution for 18 h. After transfer, the membrane was rinsed quickly in 10× SSC, baked at 80°C for 2 h and exposed to UV light for 1 min to fix the RNA. The membrane was placed in a hybridization bottle for pre-hybridization at 65°C under rotation for 4 h in a solution of 0.5 M phosphate buffer, 7% SDS and 1 mM EDTA in deionized water. Hybridization with probes occurred at 65°C under rotation for 40 h and the membrane was then washed with 2× SSC for 5 min at room temperature, twice with 2× SSC/SDS 0.5% for 30 min at 65°C, and finally with 2× SSC for 5 min at room temperature. X-Ray films (Kodak T-MAT G/RA Film) were exposed to the membranes in cassettes at −70°C, and developed according to the manufacturer's instructions. Signals on the films were quantified by densitometric analysis. Signal intensity was determined using the “Kodak Digital Science 1D” program. After subtraction of background noise, data were normalized to the ubiquitin hybridization signal.

TIR and subsequent genomic copy recovery

A two-step approach was used to access the genomic copies. First, an inverse PCR (iPCR) was performed using

10 µg of genomic DNA from cultivar SP80-3280 digested with *EcoRI* (New England Biolabs) and ligated to form circular molecules using T4 DNA Ligase (Fermentas), according to manufacturer's instructions. These circularized fragments of genomic DNA enabled outward amplification from primers designed on the outermost portion of TE221. A nested primers strategy was used in the iPCR. In the first round of amplification 100 ng of circularized genomic DNA was used as template, and the reactions were submitted to the following thermal cycle: 2 min at 94°C, 35 rounds of 30 s at 94°C, 30 s at 57°C and 1 min at 72°C, and a final extension time of 72°C for 10 min. In the second round 1 µl of the first round PCR, with a final volume of 50 µl, was used as template to the second amplification, which was submitted to the same thermal cycle. The primers used are presented in Table 1. The product of the iPCR was cloned in the pGEM-T-Easy, according to manufacturer's instructions, and sequenced using primers of the vector. The TIRs of the element were identified in these sequenced fragments amplified in the iPCR. New primers were designed on the sequences of the TIRs, and a regular PCR was performed. This amplification was made using "Elongase Enzyme Mix" (Invitrogen), according to the manufacturer's instructions and utilizing 80 ng of genomic DNA of the hybrid sugarcane variety SP-80-3280 as template, in a final volume of 50 µl. The reaction was submitted to the following thermal cycle: 2 min at 94°C, 35 rounds of 30 s at 94°C, 30 s at 55°C and 4 min at 68°C, and a final extension time of 68°C for 10 min. PCR products were resolved in low-melting agarose gels, the bands of interest were excised and purified through "GFX PCR DNA and Gel Purification Kit" (Amersham), cloned in the vector "pGEM-T-Easy" (Promega) and sequenced in both directions with the primers anchored on the vector. The sequences obtained, SChAT-G1 and SChAT-G2, were analyzed using BLASTN, BLASTX and ORF Finder on Genbank (<http://www.ncbi.nlm.nih.gov>) and TIGR (<http://www.tigr.org>).

Recover of SChAT elements from sequenced BACs

BACs were chosen by hybridization of lineage-specific probes (TE074 and TE191) against a commercial BAC library (Clemson University Genomics Institute) according to the manufacturer's instructions. Positive clones, BACs SHCRBa167_G07 (containing SChAT2), SHCRBa160_F01

(containing SChAT1) and SHCRBa185_O22 (containing SChAT3) were sequenced using GS-FLX pyrosequencer (Roche), according to the manufacturer's instructions. The SChAT family elements were characterized and compared with Ac transposase and the SChAT-G1 element. Two conserved motifs were aligned, one containing the residues 140–196, i.e., the Zinc finger DNA ligation domain of Ac transposase (Pfam accession number PF05699), and the second motif was the dimerization domain, containing the residues 682–719 of Ac transposase (Pfam accession number PF05699). The alignment was carried out using the multiple alignment program ClustalX (MUSCLE tool), with default parameters.

Results

Uncovering *hAT*-like transposases in sugarcane

A local sugarcane mobilome database of expressed TEs had previously been constructed, based on the full-length sequencing of 162 cDNAs related to TEs (Rossi et al. 2001; Araujo et al. 2005). Thirty-two cDNAs were identified as similar to the *hAT* superfamily of transposons and were analyzed further despite frameshifts and in-frame stop codons being detected. The diversity of the sequences was evaluated through alignment of two conserved domains characteristic of *hAT* superfamily transposases (Fig. 1a). These domains are present in the predicted sequence of 21 sugarcane clones, which were aligned at the protein level with related sequences recovered from sorghum, rice, maize, arabidopsis, tomato, tobacco, grape and populus, and the reference sequences from Ac and Tam3 of the *hAT* superfamily. The resulting analysis revealed a subdivision of the sugarcane sequences into distinct lineages. A total of 13 clones cluster in a particular lineage, that ranges from 86 to 100% nucleotide identity and which are almost identical at the amino acid level in the region of the conserved domains. This set of sequences, comprising TE016, TE017, TE035, TE048, TE074, TE096, TE124, TE132, TE203, TE207, TE217, TE223 and TE265, was named arbitrarily lineage 074, based on the cDNA chosen to design the probe used for further hybridization assays (Fig. 1a). Lineage 074 clusters with the domesticated transposase DAYSLEEPER from arabidopsis (Bundock and Hooykaas 2005) in a branch supported by a bootstrap

Table 1 Primers used in the iPCR assay

	Nested PCR	Name	Sequence
External and internal nested PCR primers are indicated	External primer	IPPTE221-1F	5' GACCAGAAGGCCTGCAAC 3'
		IPPTE221-1R	5' GACTAATGACTGTAACTTATCAACG 3'
	Internal primer	IPPTE221-2F	5' GCTCAAAGGGATTATCATTCC 3'
		IPPTE221-2R	5' GCATCTTCAGTCTTGGCTCA 3'

Fig. 1 Multiple-alignment of conserved *hAT* superfamily transposase domains. **a** Alignment of a conserved domain that encompasses amino acids 555–594 of *Ac* transposase (Pfam PB001541), and the dimerization domain (Pfam PF05699) encompassing amino acids 682–719. The alignment comprises amino acid sequences of the transposons Activator and Tam3, sugarcane cDNAs, the monocots *Zea mays* (Maize1), *Sorghum bicolor* (Sorghum1-4), *Oryza sativa* (Rice1-6), and the dicotyledonous *Arabidopsis thaliana* (Ath1-5), *Nicotiana tabacum* (Tobacco1-3), *Vitis vinifera* (Vitis1-4), *Solanum lycopersicon* (Lycop1-3) and *Populus trichocarpa* (Populus1-4) and DAYSLEEPER. The conserved lineage 074 is indicated with a bar. **b** Distance tree generated from aligned *hAT*-related sequences. The tree refers to the concatenated sequences from both alignments presented in **a**. Boxes highlight the founding elements of the *hAT* superfamily, *Ac* and Tam3, and the cDNAs used as probes, by small ellipses. Large ellipses define monocots and eudicots. Bootstrap values are indicated in the branches as numbers

A *hAT*-like transposase model

Fig. 2 Molecular hybridization of TE074, TE191 and TE257 probes with blots of angiosperm genomic DNA digested with the endonuclease *Xba*I. Species used were: sugarcane hybrid variety SP-89-1115 (Sc1); sugarcane hybrid variety SP-80-3280 (Sc2); *Saccharum officinarum* variety Badila (So); *Saccharum spontaneum* variety Mandalay (Ss); *Zea mays* (Zm); *Oryza sativa* (Os); *Ananas comosus* (Ac); *Catassatum fimbriatum* (Cf) and *Solanum tuberosum* (St). Molecular weight markers are indicated along the top of each blot

of 88%. This branch also displays a clear division between monocot and eudicot sequences (Fig. 1b). Lineage 074 sequences exhibit 89% amino acid similarity with DAYSLEEPER, along a stretch of 75 amino acids concatenated from two protein domains. The remaining sequences are heterogeneous lineages and cluster with the original plant transposable elements Ac and Tam3 (Fig. 1b). Two of these lineages were named lineage 191 and lineage 257, according to the cDNA used as probe for further hybridizations. An ellipse in Fig. 1b highlights these cDNAs.

Since different lineages were identified in sugarcane cDNAs, a genomic hybridization assay was performed to check the relative distribution and abundance between *Saccharum* species and hybrid cultivars, as well as among different grasses and other angiosperms (Fig. 2). Lineages 074, 191 and 257 were chosen to genomic hybridization assay. Under stringent conditions, probes from any of these lineages did not hybridize with distantly related monocots that neither belong to the *Poaceae* family, pineapple (*Bromeliaceae*) or orchid (*Orchidaceae*), nor to potato (*Solanaceae*; a dicot genome). Lineage 074 hybridized to all grasses analyzed (sugarcane modern cultivars, *S. officinarum*, *S. spontaneum*, maize and rice) with a low copy number profile. The 146 bp probe of TE257 hybridized with all *Saccharum* genomes, while the 187 bp probe of TE191 hybridized exclusively with sugarcane hybrid genomes and with *S. officinarum*. The TE257 probe is located between the nucleotides 1377 and 1523 of the cDNA, and the TE191 probe between the nucleotides 1074 and 1261, both non-conserved regions of the cDNAs outside of the conserved domains to avoid cross hybridization. We also used the whole nucleotide sequence of TE074 as a query sequence in BLASTN to search through three fully sequenced grass

genomes: rice (International Rice Genome Sequencing Project 2005), sorghum (Paterson et al. 2009) and maize (Schnable et al. 2009). We found three copies in sorghum (chromosomes 1, 2 and 8), five copies in rice (chromosomes 1, 3 and 3 copies in the chromosome 5), and six copies in maize (chromosomes 1, 5 and 7 and 3 copies in the chromosome 10), indicating a low copy number in grasses and showing to be localized in more than one genomic locus.

The expression of the lineage 074 was examined by hybridizing total RNA prepared from leaves and roots of greenhouse grown R570 plants and in vitro calli from SP89-1115 cultivar (Table 2) as well as RNA from apical meristem, stem, mature leaf and young leaf of field grown plants from R570 and SP80-3280 cultivars (Fig. 3).

Table 2 Relative expression levels of a TE074-related element in different tissues

Probe	Absolute value (a)	Background (b)	a – b	Normalization to ubiquitin
TE074				
Calli	234.7109	181.0375	53.6734	1.8253
Roots	185.5706	181.0375	4.5331	0.1734
Leaves	188.0968	181.0375	7.0593	0.2424
Ubiquitin				
Calli	220.2342	190.8299	29.4043	–
Roots	216.9712	190.8299	26.1413	–
Leaves	219.9508	190.8299	29.1209	–

Values were obtained from a semi-quantitative Northern blot hybridization assay. Total RNA from leaves, roots and calli of sugarcane hybrid cultivar was hybridized with a probe corresponding to TE074 and a probe from the constitutive gene ubiquitin. The intensity of the bands (a) and the background (b) were measured by densitometry. Values are normalized dividing the TE074 value by its corresponding ubiquitin value

Fig. 3 TE074 expression pattern analysis. Six-month-old plant tissues of sugarcane cultivars R570 and SP80-3280 were harvested and total RNA extracted. A northern blot assay was carried out using apical meristem (AM), stem (St), mature leaf (ML) and immature leaf (YL). P-32 labeled probes of TE074 clone and 25S were used for the hybridization assay

Expression was detected in all tissues and the developmental stages analyzed, indicating a wide distributed expression. In addition to meristems, callus, leaves and roots were also analyzed and a sevenfold higher expression was observed in callus compared to the latter tissues. By contrast, probing the membranes with TE191 revealed no expression (data not shown).

Selective constraints were evaluated to explore the hypothesis that lineage 074 contains sequences from a domesticated transposase in opposition to the other two transposon lineages identified. It is expected that a functional gene, as proposed to the domesticated lineage, should be under a stronger selective pressure to maintain its coding sequence, while the transposon lineages are more tolerant to mutations (Feschotte and Pritham 2007). The synonymous (dS) and non-synonymous (dN) substitution rates, as well as dN:dS values, were calculated for cDNA sequences from the lineage 074 and the other two lineages, 191 and 257. Assuming that synonymous substitutions are under neutral evolution, dN:dS = 1 will represent neutral evolution, dN:dS < 1 will correspond to purifying selection, and dN:dS > 1, positive Darwinian selection.

Among thirteen lineage 074 cDNAs compared, nine presented dN:dS values under unity when compared with all the other cDNAs from the same lineage. For the lineage 074 subgroup formed by cDNAs TE048, TE096, TE207 and TE265, dN:dS ratios ranged from 0.223 to 0.256, with Z test *P* values ranging from 0.000 to 0.047. This means that these sequences are under purifying selection. Although the Z test *P* values were above the cutoff 0.05 to the other cDNAs from lineage 074, the dN:dS values were under unity to the majority of the sequences. In only three cases this ratio was above unity (Fig. S2; supplementary material). These results suggest a purifying selection to the majority of the lineage 074 sequences, corroborating a characteristic of a functional gene. Nevertheless, the purifying selection over the other lineage 074 sequences is more relaxed than the one over the subgroup formed by the cDNAs TE048, TE096, TE207 and TE265. Among the cDNAs from lineages 191 and 257 analyzed, no statistically validated data were obtained.

To examine the diversity among lineage 074 cDNA clones, a nucleotide sequence alignment was performed. Although these sequences are very similar at the amino acid level, nucleotide alignment revealed their remarkable subdivision into two highly conserved groups (Fig. 4). TE048, TE096, TE207 and TE265 diverge from the other group, suggesting that these sequences could be transcribed from two distinct loci. The separation time between these two hypothetical loci was calculated, and range from 32.23 to 35.6 MYA.

SChAT, a novel family of transposons present in the sugarcane genome

Two independent strategies were used to recover full-length copies of *hAT* transposons from the sugarcane genome: amplification from genomic DNA and searches in sugarcane BACs sequenced in our laboratory. Inverse PCR (iPCR) experiments enabled us to access genomic sequences adjacent to transposons related to lineage 191. Primers were designed on both subterminal regions of the cDNA TE221, directed toward the ends. TE221 was chosen because its sequence is the longer among lineage 191 cDNAs. This approach leads to two possibilities of amplification: the first is the amplification of two extremities of one copy of the element, including untranscribed regions and flanking genomic regions. The second possibility is amplification of the extremities of two identical or very similar tandem copies of the element instead of two flanks of the same copy. Two bands, with sizes of 1,062 and 755 bp, were amplified by iPCR (Fig. S1a; supplementary material). Analysis of these sequenced fragments revealed the absence of the restriction site for the endonuclease *EcoRI*, which was used to cleave the genomic

◀ **Fig. 4** Multiple alignment of a 1,008 bp nucleotide segment shared by the lineage 074 cDNAs. Alignment was performed with ClustalX using default parameters. *Shaded boxes* indicate conserved domains of the *hAT* transposase superfamily in the nucleotide sequence as presented. *Boxed names on the left* highlight a subgroup within lineage 074

DNA. This information indicates that the terminal sequences obtained, including the putative TIRs, correspond to copies sitting in tandem in the genome. The pair of TIRs isolated is composed of 21 reverse complementary base pairs, with 15 matches among them. No target site duplication was found. Primers were designed based on these TIR sequences, directed inwards to amplify a full-length copy. Amplifications making use of these primers allowed the recovery of two genomic versions of the sugarcane *hAT* element from the genome of cultivar SP80-3280. These elements, 4,616 bp and 3,838 bp in length, were named SChAT-G1 and SChAT-G2, respectively (Fig. S1b; supplementary material). SChAT-G2 is a deleted copy of SChAT-G1 with a deletion of 777 bp in the 3' portion of SChAT-G1 comprising positions 2,822–3,624 (Fig. S1c; supplementary material).

Searching sugarcane BACs revealed three copies of the SChAT family, named SChAT1, SChAT2 and SChAT3. Figure 5a presents a schematic representation of the coding region structure among the elements comprising the SChAT family. The cDNAs TE191, TE221 and TE257 are also included as well as the Ac element from maize. All these elements have a dimerization domain in their carboxy terminal portion and a zinc finger domain in the amino terminal region, except SChAT3, which is truncated. SChAT2 has two frameshifts in its predicted coding sequence, represented by the black vertical bars in Fig. 5a. SChAT-G1 and SChAT1 are larger elements and both present the potential to be autonomous elements. SChAT-G1 shares 51% similarity with the Ac transposase, on a segment covering 77% of the full Ac sequence. Nevertheless, there are 2,540 nucleotides downstream of the putative stop codon of SChAT-G1 predicted transposase that present similarity to the SNF2 helicase domain. In addition, there is probably a deletion in the 5' untranscribed region of SChAT-G1, between the putative TATA box and the putative start codon, which are separated by only two nucleotides. SChAT1 exhibits 51% protein similarity along a region covering 77% of the full Ac transposase. ORF analysis indicates putative start and stop codons located at positions 1,049 and 2,870 of its nucleotide sequence, respectively.

SChAT1, SChAT2, SChAT3, SChAT-G1 and SChAT-G2 are the first DNA transposons transcriptionally active isolated from sugarcane genome. The detailed alignments of the zinc finger and the dimerization domains are presented in Fig. 5b, c. A phylogenetic tree constructed based

on the alignment of the concatenated sequences of these domains reveals that TE191, TE221 and the SChAT elements cluster in a branch supported by 100% bootstrap, while TE257 is more distantly related (Fig. 5d). These results suggest the presence of at least two distinct *hAT*-like transposon lineages inhabiting the sugarcane genome, one called lineage 191 related to TE191 cDNA, and a second related to TE257 cDNA, that were not recovered by amplification or by searches in BAC sequences in the present study.

To strengthen the observation of recovery of a potentially active version of SChAT 191 lineage and the existence of a domesticated variant, the identification of the DDE motif described in Muehlbauer et al. (2006) was made. Figure 6 displays the alignment of four lineage 074 cDNAs, w-gary1, w-gary2, b-gary, the two recovered genomic versions SChAT1 and SChAT2 along with the Ac transposase. All the six domains (A–F) previously described in Muehlbauer et al. (2006) were identified. Arrows in the figure highlight the predicted DDE motif necessary for the enzymatic cleavage activity. Only the canonical Ac transposase and SChAT1 possesses the expected amino acid triad at the correct position. The domesticated SChAT versions found in sugarcane present a GDG motif in the corresponding positions. Another conserved motif of *hAT* transposases is a CxxH downstream of the second D residue of the DDE motif, required to the catalytic activity of the enzyme (Zhou et al. 2004). Three of the four lineage 074 cDNAs analyzed presented the correct motif: TE096, TE048 and TE124 as well as SChAT1, SChAT2 and the Ac transposase. In summary, we recovered five genomic copies of a novel *hAT* superfamily of transposons in sugarcane, named SChAT. It was the first family of transposons isolated from sugarcane genome. Among these copies there is one potential autonomous element, presenting a full-length ORF and the intact motifs characteristic of the superfamily.

Discussion

The *hAT* superfamily of transposons presents remarkable historical significance for transposable elements studies because Ac, a belonging member of the superfamily, was the first transposable element observed in the early studies of Barbara McClintock. It is also one of the most studied superfamily of transposons. However, a deepened analysis of these elements was never made in sugarcane. In order to access the information about these elements present in sugarcane genome, we used a strategy based in the analysis of cDNAs. In the first part of this work we analyzed the diversity, genomic distribution, expression and selective constraints relative to a set of thirty-two cDNAs previously

Fig. 5 Comparison of *hAT* transposons recovered from sugarcane genome and the *Ac* transposase. **a** Schematic representation comparing the amino acid sequence of the functional transposase of the *Ac* element from maize, the sugarcane cDNAs TE191, TE221 and TE257, and the predicted amino acid sequences of the transposase of the five SChAT elements found in sugarcane: SChAT1, SChAT2, SChAT3, SChAT-G1 and SChAT-G2. The numbers in brackets in

each line correspond to the size of the predicted ORF. The three *hAT* superfamily domains are highlighted in gray boxes. Black bars indicate the presence of frameshift mutations on the copy of SChAT2. **b** Detailed alignment of the zinc finger domain. **c** Detailed alignment of the *hAT* dimerization domain. **d** Distance tree based on the concatenated sequences aligned in **b** and **c**. Bootstrap value above 50% is presented

annotated as *hAT*-like by our group (Rossi et al. 2001; Araujo et al. 2005). The alignment of two conserved protein domains characteristic of *hAT* transposases present in the cDNAs with homologous of several species (see “Materials and methods”) resulted in the separation of distinct lineages among the sugarcane cDNAs (Fig. 1). Lineage 074 was particularly interesting, presenting a highly conserved aminoacid sequence among its sequences and clustering with the domesticated transposase DAYS-LEEPER from *Arabidopsis thaliana* (Bundock and Hoo-yaas 2005). The branch containing these sequences is supported by 88% bootstrap (Fig. 1b). The clustering pattern suggests that all these transposases may be DAYS-LEEPER homologous, also domesticated transposases. The division of this branch in monocot and eudicot suggests that the domestication of these transposases might occur before the separation between mono and eudicots.

Subsequently three lineages were investigated further regarding their genomic distribution: 074, 191 and 257. None of the lineages was present in the non-grasses genomes examined, potato (dicot), pineapple and orchid (monocots). While lineage 257 was present in all *Saccharum* genomes, lineage 191 was even more restrict and present only in the genomes of sugarcane hybrids and *S. officinarum* (Fig. 2). The contribution of *S. spontaneum* to chromosome content in the hybrids described to be about 15–25% (D’Hont et al. 1996). Thus it is reasonable that sequences which evolved after the divergence of ancestral *Saccharum* into *S. spontaneum* and *S. officinarum* might be present in *S. officinarum* and its hybrids, but not in *S. spontaneum*. On the other hand, lineage 074 was present in the genomes of all the grasses analyzed. Since all the probes were made carefully to avoid cross hybridization, this result reveals a higher degree of sequence conservation of lineage 074 when compared with the other two lineages. A search for lineage 074 homologous in maize (Schnable et al. 2009), rice (International Rice Genome Sequencing Project 2005) and sorghum (Paterson et al. 2009) genomes also found a low copy number: six, five and three copies, respectively. These results confirm a low copy number in grasses and the presence of more than one locus. Again, the results suggest that lineage 074 could be a domesticated transposase, since it is expected that a gene with cellular function should be more conserved in sequence and present in the genome as a lower copy number than transposons (Feschotte and Pritham 2007). In summary, the genomic hybridization results, in addition to the clustering pattern of TE074, TE191 and TE257 in the phylogenetic tree, indicate that *hAT* lineages display distinctly unique patterns, suggesting independent evolution of these paralogues, and the lineage 074 presents characteristics of a domesticated transposase.

We carried out the expression analysis of lineages 074 and 191 in four tissues of two sugarcane cultivars.

Expression of lineage 074 was confirmed in all the samples analyzed. In opposition, no expression was detected to lineage 191 (Table 2; Fig. 3). Araujo et al. (2005) presented a global analysis of sugarcane cDNAs homologous to many TE families, including a macroarray in which they report the expression of several *hAT*-like clones in distinct tissues. Of the 13 clones from 074 lineage, several were expressed in tissues such as calli (TE017, TE048, TE096, TE203 and TE207), apical meristem (TE017, TE203 and TE207), leaf roll (TE017 and TE203), and mature and immature inflorescence (TE017 and TE203). Of the other lineages, only TE191 and TE221 were expressed in calli. Transcripts of lineage 074 are generally more abundant than those from the other lineages, indicating higher expression in the tissues investigated. These previous studies are in agreement with the expression results presented here, which demonstrate higher expression levels and distribution among diverse tissues in lineage 074, while no detectable expression was found to lineage 191. It has been extensively demonstrated that transposons are under strict transcriptional control mediated by their methylation status (Van Sluys et al. 1993; Scortecci et al. 1997; Bird 2007). These data support our hypothesis that lineage 074 refers to a domesticated transposase, presenting higher levels of expression, while the other lineages correspond to transposons, which are expected to be under transcriptional repression. The analysis of selective constraints also supports the hypothesis of lineage 074 as a domesticated transposase. An evaluation of dN:dS for the lineages 074, 191 and 257 revealed a purifying selection to the majority of the lineage 074 sequences, while to the other lineages no statistical supported data were obtained, may be due to the low number of sequences available to compare (Figure S2; supplementary material). The purifying selection acting over lineage 074 sustains a feature of a functional gene.

A nucleotide sequence alignment was performed to analyze in detail the diversity among the lineage 074. The alignment revealed a well-defined subdivision of the sequences into two highly conserved groups (Fig. 4). TE048, TE096, TE207 and TE265 diverge from the other group. This pattern suggests that these cDNAs could be originated from two distinct loci. In addition to the split up of lineage 074 in two subgroups of sequences, it is known that duplication events have occurred during the *Poaceae* evolution (Paterson et al. 2004; Kim et al. 2009) and particularly in sugarcane (Ming et al. 1998). Homologues of TE074 were found in sorghum and rice in three chromosomes, and in maize in four chromosomes, denoting the existence of distinct loci. These data strongly suggest that the differences between the two subgroups of lineage 074 are due to its origin from two loci, and not a polymorphism between two alleles of the same locus. We calculated the

Fig. 6 Alignment of the predicted aminoacid sequences of sugarcane cDNAs, gary transposases from barley and wheat, and the Ac transposase. Sugarcane sequences TE096, TE048, TE074 and TE124 refer to domesticated transposase, and SChAT1 and SChAT2 are transposons. The key DDE aminoacid motifs are indicated with arrows. The six regions (A–F) characteristics of the hAT superfamily are underlined

separation time between the two hypothetical loci and the range obtained was from 32.23 to 35.6 MYA. Since the speciation of the *Saccharum* genus is estimated in 8–9 MYA (Jannoo et al. 2007; Wang et al. 2010) and the origin of the grasses is between 75 and 55 MYA (Kellogg 2001), thus it is reasonable to infer that the locus referent to lineage 074 was duplicated after the differentiation of the grasses, and significantly before the differentiation of *Saccharum*.

Based on the results presented, we propose the occurrence of a domestication event during the evolution process of the *hAT*-like transposases, before the divergence of mono and eudicots, approximately from 70 to 55 MYA (Kellogg 2001), and further demonstrate that these loci are expressed in sugarcane. Moreover, expression level is higher in undifferentiated tissues (calli and apical meristems) of different sugarcane cultivars in agreement with the aberrant phenotype of DAYSLEEPER arabidopsis mutants (Bundock and Hooykaas 2005). The results presented here corroborate the hypothesis that lineage 074 is a domesticated transposase.

The second part of this work presents a new transposon family in sugarcane. We recovered two genomic copies of lineage 191 elements from sugarcane genome using iPCR strategy. These copies were named SChAT-G1 and SChAT-G2. Subsequently, three copies of the same lineage were identified in sugarcane BACs: SChAT1, SChAT2 and SChAT3. From these, two elements presented the potential to be autonomous, SChAT-G1 and SChAT1. Both elements present the dimerization domain in the carboxy terminal portion and the zinc finger domain in the amino terminal region (Fig. 5a). SChAT-G1 and SChAT1 also share 51% protein similarity along a region covering 77% of the full Ac transposase. However SChAT-G1 presents 2,540 nucleotides downstream of the putative predicted stop codon that present similarity to the SNF2 helicase domain. Further analysis will be necessary to evaluate if this region represents an exon capture or if it has been functionally co-opted by the element. Further transcript analysis may contribute to solving this question. In addition, SChAT-G1 probably presents a deletion between the putative TATA box and the putative start codon, what probably makes it a non-functional element. SChAT1, on the other hand, exhibits all the six protein domains previously described in Muehlbauer et al. (2006), the DDE motif necessary for the enzymatic cleavage transposase activity,

and the conserved motif CxxH downstream of the second D residue of the DDE motif, also required to the catalytic activity of the enzyme (Zhou et al. 2004) (Fig. 6). ORF analysis of SChAT1 indicates putative start and stop codons located at positions 1,049 and 2,870 of its nucleotide sequence, respectively, and the absence of premature stop codons or frameshifts. Although SChAT1 has characteristics indicative of an autonomous element, further studies in vitro are necessary to evaluate the functionality of the transposase encoded by this element.

With the aim to corroborate the correlation of the genomic copies isolated with lineage 191, on which the iPCR primers were designed, we carried out an alignment of the zinc finger and the dimerization domains (Fig. 5b–d). The clustering pattern obtained revealed that TE191, TE221 and the genomic SChAT elements cluster in a branch supported by 100% bootstrap, while TE257 and the maize Ac element represent independent branches. This result suggests the presence of at least two distinct *hAT*-like transposon lineages inhabiting the sugarcane genome, lineage 191 and lineage 257. No copy from lineage 257 was recovered by amplification or by searches in BAC sequences in the present study. Further analysis will be necessary to search for these elements as the sugarcane BAC sequence database expands. Five genomic copies recovered belong to lineage 191, supporting the existence of a true transposon lineage composed by the elements SChAT1, SChAT2, SChAT3, SChAT-G1 and SChAT-G2. To date these represent the first collection of DNA transposons recovered from the sugarcane genome. Nevertheless, transposition activity remains to be demonstrated.

Here we report a detailed analysis of *hAT* transposase paralogues in the sugarcane genome, using a collection of cDNAs as a starting point. We identified a putative domesticated transposase present in two loci, and at least two lineages of transposons. Previous studies have reported *hAT*-like elements transcriptionally active in rice (Jiao and Deng 2007) and grapevine (Benjak et al. 2008). Nevertheless, here we presented a combined approach to reveal these actively expressed *hAT* paralogues from sugarcane tissues. Our data suggest the presence of a domesticated *hAT*-like transposase in two loci of the sugarcane genome with clear transcriptional activity in several tissues of distinct sugarcane cultivars. This transposase is highly conserved among monocots and eudicots, suggesting early specialization. We also identified at least two lineages of transcriptionally active transposons restricted to *Saccharum*, including non-autonomous copies that are also expressed. For one of the lineages, we recovered five genomic copies including one potential autonomous element from a novel transposon family, SChAT. One of the isolated copies carries a SNF2-helicase domain. These findings provide the basis for further functional studies on

the domesticated transposase gene as well as on the transposition mechanisms of the SChAT family, which will contribute to our understanding of the impact of *hAT* superfamily transposons on the sugarcane genome.

Acknowledgments We thank Dr. Magdalena Rossi and Paula Araújo for early discussion of the experiments, Dr. Josep Casacuberta for careful reading of the manuscript and suggestions, Dr. Nathalia de Setta for useful discussions about the evolutionary analysis, and Silvia Regina Blanco and Myna Nakabashi for technical assistance. This work had the following financial support: FAPESP and CNPq.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Araujo PG, Rossi M, de Jesus EM et al (2005) Transcriptionally active transposable elements in recent hybrid sugarcane. *Plant J* 44(5):707–717
- Benjak A, Forneck A, Casacuberta JM (2008) Genome-wide analysis of the “Cut-and-Paste” transposons of grapevine. *PLoS ONE* 3(9):e3107
- Bird A (2007) DNA methylation patterns and epigenetic memory. *Genes Dev* 16:6–21
- Britten RJ (2004) Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc Natl Acad Sci USA* 101(48):16825–16830
- Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. *Nature* 436(7048):282–284
- Calvi BR, Hong TJ, Findley SD, Gelbart WM (1991) Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: hobo, Activator and Tam3. *Cell* 66:465–471
- Casola C, Lawing AM, Betran E, Feschotte C (2007) PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol Biol Evol* 24(8):1872–1888
- D’Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC (1996) Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol Gen Genet* 250:405–413
- Essers L, Adolphs RT, Kunze R (2000) A highly conserved domain of the maize Activator transposase is involved in dimerization. *Plant Cell* 12:211–223
- Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9:397–405
- Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331–368
- Gaut BS, Morton BR, McCraig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc Natl Acad Sci* 93:10274–10279
- Grivet L, Daniels C, Glaszmann JC, D’Hont A (2004) A review of recent molecular genetics evidence for sugarcane evolution and domestication. *Ethnobot Res Appl* 2:9–17
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 41:95–98
- Hehl R, Nacken WK, Krause A, Saedler H, Sommer H (1991) Structural analysis of Tam-3, a transposable element from *Antirrhinum majus*, reveals homologies to the Ac element from Maize. *Plant Mol Biol* 16(2):369–371
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice. *Nature* 436:793–800
- Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann JC, Arruda P, D’Hont A (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J* 50:574–585
- Jiao Y, Deng XW (2007) A genome-wide transcriptional activity survey of rice transposable element-related genes. *Genome Biol* 8:R28
- Kellogg E (2001) Evolutionary history of the grasses. *Plant Physiol* 125:1198–1205
- Kempken F, Windhofer F (2001) The *hAT* family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma* 110(1):1–9
- Kim C, Tang H, Paterson AH (2009) Duplication and divergence of grass genomes: interacting the Chloridoids. *Tropical Plant Biol.* 2:51–62
- Kraitshtein Z, Yaakov B, Khasdan V, Kashkush K (2010) Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. *Genetics* 186:801–812
- Lisch D, Bennetzen JL (2011) Transposable element origins of epigenetic gene regulation. *Curr Opin Plant Biol* 14:156–161
- McClintock B (1946) Maize genetics. Year Book Carnegie Inst Washington 45:176–186
- McClintock B (1984) The significance of responses of the genome to challenge. *Science* 226:792–801
- Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A et al (1998) Detailed alignment of *Saccharum* and *Sorghum* chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* 150:1663–1682
- Muehlbauer GJ, Bhau BS, Syed NH et al (2006) A hAT superfamily transposase recruited by the cereal grass genome. *Mol Genet Genomics* 275(6):553–563
- Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci* 101(26):9903–9908
- Paterson AH, Bowers JE, Bruggmann R et al (2009) The *Sorghum* bicolor genome and the diversification of grasses. *Nature* 457(29):551–556
- Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Rep* 15:8–15
- Rossi M, Araújo PG, Van Sluys MA (2001) Survey of transposable elements in sugarcane expressed sequence tags (ESTs). *Genet and Mol Biology* 24(1–4):147–154
- Rubin E, Lithwick G, Levy AA (2001) Structure and evolution of the *hAT* transposon superfamily. *Genetics* 158(3):949–957
- Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor
- Schnable PS, Ware D, Fulton RS et al (2009) The B73 Maize Genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Scortecchi KC, Dessaux Y, Petit A, Van Sluys MA (1997) Somatic excision of the Ac transposable element in transgenic *Arabidopsis thaliana* after 5-azacytidine treatment. *Plant Cell Physiol* 38(3):336–343
- Streck RD, MacGaffey JE, Beckendorf SK (1986) The structure of hobo transposable elements and their insertion sites. *The EMBO Journal* 5(13):3615–3623
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599

- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 24:4876–4882
- Van Sluys MA, Scortecci KC, Tempé J (1993) DNA methylation associated to Ac element in *Arabidopsis*. *Plant Physiol Biochem* 31(6):805–813
- Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* 28(9):913–922
- Wang J, Roe B, Macmill S, Yu Q, Murray JE, Tang H, Chen C, Najar F, Wiley G, Bowers J, Van Sluys MA, Rokhsar DS, Hudson ME, Moose SP, Paterson AH, Ming R (2010) Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11:261
- Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. *Proc Natl Acad Sci* 103(47):17600–17601
- Yuan Y, Wessler SR (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci* 108(19):7884–7889
- Zhou L, Mitra R, Atkinson PW, Burgess Hickman A, Dyda F, Craig NL (2004) Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* 432:995–1001