**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Rearranjos cromossômicos e sua influência no aparecimento de novas
famílias de DNA satélite em *Arabidopsis***

Gabriela Moreira Machado

Dissertação apresentada para obtenção do título de Mestra
em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2019**

Gabriela Moreira Machado
Bacharela em Ciências Biológicas

**Rearranjos cromossômicos e sua influência no aparecimento de novas famílias de DNA
satélite em *Arabidopsis***

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **MATEUS MONDIN**

Dissertação apresentada para obtenção do título de Mestra
em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Machado, Gabriela Moreira

Rearranjos cromossômicos e sua influência no aparecimento de novas famílias de DNA satélite em *Arabidopsis* / Gabriela Moreira Machado. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

37 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. DNA satélite 2. *A. thaliana* 3. Evolução de satDNA 4. Recombinação cromossômica
5. *A. lyrata* I. Título

AGRADECIMENTOS

Ao CNPq pela oportunidade.

À Escola Superior de Agricultura “Luiz de Queiroz” - ESALQ-USP pelo acolhimento.

Ao Departamento de Genética pela disponibilização da infraestrutura.

Ao Prof. Dr. Mateus Mondin, pela orientação, fidelidade, sabedoria e confiança.

À técnica do Laboratório de Citogenômica e Epigenética – CYNGELA, Silvia, por toda ajuda com os equipamentos, e todos os cafés.

Aos colegas do Laboratório de Citogenômica e Epigenética – CYNGELA pelo companheirismo em momentos felizes e tristes.

Ao Gabriel, por todas as incontáveis horas de discussão sobre o trabalho.

A meus queridos pais, Darlete e Adilson, e amada irmã Eloá, por todo amor e carinho.

Ao meu amado noivo Ramon, por toda paciência, amor e companheirismo em todas as fases.

.

SUMÁRIO

RESUMO.....	5
ABSTRACT.....	6
LISTA DE FIGURAS.....	7
LISTA DE TABELAS.....	8
1. INTRODUÇÃO	9
REFERÊNCIAS.....	9
2. REVISÃO BIBLIOGRÁFICA.....	11
2.1. DNA satélite.....	11
2.2. Evolução em concerto	16
2.3. Sequências de satDNA em <i>A. thaliana</i>	17
REFERÊNCIAS.....	18
3. IMPACT OF CHROMOSOMAL REARRANGEMENTS IN THE EVOLUTION ON SATDNA FAMILIES IN <i>ARABIDOPSIS</i>	23
ABSTRACT.....	23
3.1. Introduction.....	23
3.2. Methodology.....	25
3.2.1. Identification and extraction of satDNA sequences from <i>A. thaliana</i>	25
3.2.2. Identification and extraction of satDNA sequences from <i>A. lyrata</i>	26
3.2.3. Alignment and similarity tree construction	26
3.2.4. DotPlot analysis.....	27
3.3. Results	27
3.3.1. Description of the sequences and their similarities.....	27
3.3.2. Similarity tree	30
3.4. Discussion.....	31
REFERENCES	34
4. CONCLUSÃO	37

RESUMO

Rearranjos cromossômicos e sua influência no aparecimento de novas famílias de DNA satélite em *Arabidopsis*

O centrômero é composto por heterocromatina constitutiva, que atua como uma estrutura cromossômica importante. Além de ser um ponto de referência para a separação dos braços cromossômicos, ele é responsável pela correta segregação das cromátides durante o ciclo celular. Esta estrutura indica onde as fibras do fuso mitótico devem se ligar ao cromossomo, formando uma estrutura proteica chamada de cinetócoro. Os centrômeros são compostos principalmente de sequências altamente repetitivas. Uma parte destas sequências é caracterizada como DNA satélite, e aparece organizada em *tandem*. Estas sequências passam por um processo evolutivo chamado de Evolução em Concerto, que é impulsionado pela acumulação de mutações devido aos mecanismos de deriva molecular. Este processo acaba homogeneizando os DNAs satélites de uma espécie. Algumas famílias de satDNAs são similares em um determinado grupo de espécies, como é no caso do gênero *Arabidopsis*. Diversos rearranjos cromossômicos ocorreram no processo evolutivo deste grupo. Quatro destas alterações ocorreram em regiões centroméricas e pericentroméricas, somente nos cromossomos ancestrais 6, 7 e 8, similares ao cariótipo moderno de *A. lyrata*, que deram origem aos cromossomos modernos 4 e 5 de *A. thaliana*. As três famílias de satDNA de 180pb neste cariótipo ancestral são diferentes da única família presente no cariótipo moderno de *A. thaliana*. Ademais, esta única família aparece em todos os cinco cromossomos desta espécie. Isso levanta algumas questões, como por exemplo: como a família de 180pb de *A. thaliana* aparece em todos os cromossomos, sendo que as alterações aconteceram somente nos cromossomos 4 e 5? E o que aconteceu com as outras famílias presentes no cariótipo ancestral? Para responder estas questões, foram realizados procedimentos bioinformáticos. O primeiro passo foi localizar os centrômeros nos sequenciamentos disponíveis em bancos de dados (<https://plants.ensembl.org/>). No caso de *A. thaliana*, o software JDotter foi utilizado para construir um DotPlot para localizar as sequências repetitivas, e no caso de *A. lyrata*, foi realizado um Blast dentro da própria base de dados, com as sequências disponíveis na literatura. Então, o software Tandem Repeats Finder foi utilizado para localizar e extrair as famílias de satDNA do sequenciamento. Estas sequências foram organizadas utilizando o BioEdit, e alinhadas no software MEGA7. Uma árvore de máxima verossimilhança foi construída, e apresentou uma separação clara de ambas as espécies, e que as sequências dos cromossomos 4 e 5 de *A. thaliana* se apresentaram misturadas em algumas regiões da árvore. Isto mostra que existe uma diferença na homogeneidade destas sequências, mesmo que ambas sejam da mesma família. Este resultado sugere que mudanças no contexto genômico devido aos rearranjos cromossômicos influenciam o processo de homogeneização nestas sequências de satDNA, portanto, influenciando a evolução em concerto.

Palavras-chave: DNA satélite, *A. thaliana*, Evolução de satDNA, Recombinação cromossômica, *A. lyrata*

ABSTRACT

Chromosomal rearrangements and their effects on the appearance of novel satellite DNA families in *Arabidopsis*

The centromere is composed of constitutive heterochromatin, taking part as an important chromosomal structure. Apart from being a landmark for the separation of both chromosome arms, it is responsible for the correct segregation of chromatids during the cell cycle. This structure indicates where the spindle fibers should connect to the chromosome, forming a proteic structure called kinetochore. The centromeres are composed mainly by repetitive sequences, organized in tandem arrays, called satellite DNAs. These sequences go through an evolutionary process called Concerted Evolution, which is driven by accumulation of mutations due to molecular drive mechanisms. This process ends up in a homogenization of the satDNA in a species. Some satDNA families are similar in a given group of species, such as in *Arabidopsis* genera. Several chromosomal rearrangements took place in this group's evolution. Four of these alterations occurred in centromeric and pericentromeric regions, only on ancestral chromosomes 6, 7 and 8, similar to *A. lyrata* modern karyotype, that gave place to modern *A. thaliana* chromosomes 4 and 5. The three 180bp satDNA families present in this ancestral karyotype differ from the single one present in modern *A. thaliana* karyotype. Moreover, this single family appears in all five chromosomes of this species. This gives rise to a few questions, of how did the *A. thaliana* 180bp satDNA family present appeared in all of its five chromosomes, since the alterations only occurred in chromosomes 4 and 5? And what happened to the other families present in the ancestral karyotype? In order to answer these questions, bioinformatic procedures were performed. The first step was to locate the centromeres on the sequencing available on the database (<https://plants.ensembl.org/>). In the case of *A. thaliana*, the software JDotter was used to construct a DotPlot to locate repetitive sequences, and in the case of *A. lyrata* a Blast was performed on the database with the sequences presented in the literature. Then, the software Tandem Repeats Finder was used to locate and extract the 180bp satDNA families. The sequences were organized using BioEdit and aligned on MEGA7 software. A Maximum Likelihood tree was constructed, and showed a clear separation between both species, and that in *A. thaliana*, sequences from chromosomes 4 and 5 intermingled in some regions. This shows that there is a different homogeneity between these sequences, even though they are members of the same family. This result hints that changes in the genomic context due to the chromosomal rearrangements influence the homogenization process in these satDNA sequences, and thus, influencing the concerted evolution.

Keywords: Satellite DNA, *A. thaliana*, satDNA evolution, Chromosomal recombination, *A. lyrata*

LISTA DE FIGURAS

- Figura 1. Comparative analysis representation of the alignment of all extracted motifs from centromeric and pericentromeric regions of chromosomes 6, 7 and 8 of *A. lyrata*, the chromosomes 4 and 5 of *A. thaliana*, and the consensus sequences presented by KAWABE and NASUDA (2005) of the subfamilies of 180 bp family. (CON) Represent the consensus sequences of all sequences; (a) pAL1 subfamily consensus sequence; (b) consensus sequence of 180 bp satDNA family from chromosomes 4 and 5 of *A. thaliana*; (c) pAa consensus sequence; (d) consensus sequence of 180 bp satDNA family of *A. lyrata* chromosome 7; (e) pAge1 consensus sequence; (f) consensus sequence of 180 bp satDNA family from chromosome 6 from *A. lyrata*; (g) consensus sequence of 180 bp satDNA family from chromosome 8 from *A. lyrata*; (h) pAge2 consensus sequence; (i) consensus sequence of 180 bp family from chromosome 6 of *A. lyrata*. 28
- Figura 2. DotPlot graphic analysis between the consensus sequences of centromeric and pericentromeric satDNA families of *A. lyrata* (pAa and pAge2) and *A. thaliana* (pAa). Only the chromosomes 4 and 5 of *A. thaliana* and chromosomes 6, 7 and 8 of *A. lyrata* were used in this analysis. In both x and y axis, the order of sequences is the same. It is observable a similarity between the chromosomes of *A. thaliana* and chromosome 7 of *A. lyrata*. However, the chromosome 6 and 8 of *A. lyrata* don't present clear similarity with the other analyzed sequences, including the consensus of chromosome 7 of the same species. 29
- Figura 3. Maximum Likelihood tree of satDNA sequences between chromosomes 6 (blue), 7 (pink) and 8 (light green) of *A. lyrata* and chromosomes 4 (red) and 5 (dark green) of *A. thaliana*. Each terminal represents a single satDNA sequence. The tree represents a ML tree using all extracted sequences of each chromosome for each species. 31
- Figura 4. Chromosome evolution diagram of *A. thaliana* (AT4 and AT5) as a result of chromosomal rearrangements of an ancestral karyotype (AK6 - green, AK7 - blue and AK8 - pink). T1 represents the first translocation in the centromere region between chromosomes AK6 and AK7, resulting in two chromosomes AK6/7. One of these chromosomes suffered an inversion (I1) involving the centromere, resulting in the chromosome AT4. The other chromosome suffered a translocation (T2) with an inverted AK8 (I2), resulting in one microchromosome which was lost during the process, and the AT5 chromosome. The shadowed chromosomes indicate where a change in the context in which the chromosome is inserted happened. This image is an adaption of LYSAK et al. (2006). 32

LISTA DE TABELAS

Tabela 1. Sequencing of the species used in this work, and database used to gather the data to harvest the satDNA sequences. 26

Tabela 2. Chromosomes searched for the satDNA sequences of each species, number of repeats found and mean repeat length of each satDNA family. The last column is the similarity of the sequences provided by KAWABE and NASUDA (2005) from each of the families in relation to the consensus of the pAL1 family. 27

1. INTRODUÇÃO

A descoberta dos DNA satélites (satDNA) nos genomas das mais variadas espécies estimulou a curiosidade para estas sequências, já que não se sabia qual a função destas dentro do genoma. Estas sequências estão presentes em regiões altamente repetitivas do genoma, como os centrômeros e telômeros. Estas regiões, também chamadas de heterocromatinas constitutivas, são repletas de sequências não codantes, como os próprios satDNAs e retrotransposons.

Ainda hoje não se sabe exatamente a função destes satélites, mas sabe-se que mudanças em suas características podem resultar em alterações fenotípicas. Dentro de tais circunstâncias, foram criadas hipóteses de seu processo evolutivo no contexto genômico, o que levou aos questionamentos de porque estas sequências são repetitivas e organizadas in tandem.

Embora estes estudos genéticos e moleculares tenham aprofundado o conhecimento sobre tais sequências, questões importantes permanecem abertas, como por exemplo: como e quando estas sequências surgem em um determinado grupo?

A dissertação aqui presente responde estas questões com base na evolução cariotípica de *A. thaliana* e *A. lyrata*, de forma que os sucessivos eventos de recombinação presentes nesta história evolutiva tenham influenciado (ou tenham sido influenciados) pela presença de satDNA nas regiões onde estes eventos ocorreram.

Além disso, esta dissertação faz parte de um conjunto de trabalhos deste grupo de pesquisa ainda não publicados sobre o tema, portanto, certos pontos que parecem incompletos aqui descritos são descritos nestes trabalhos a serem publicados.

2. REVISÃO BIBLIOGRÁFICA

2.1. DNA satélite

A descoberta da estrutura molecular do DNA no início da década de 1950 possibilitou diversos estudos acerca desta molécula. Um destes estudos foi o de (MESELSON; STAHL; VINOGRAD, 1957), que descobriram que ao centrifugar uma solução de CsCl era obtido um gradiente de concentração no recipiente onde se encontrava esta solução. A partir disto, ao adicionar uma macromolécula à solução era possível visualizar uma banda em algum local específico do tubo. Aplicando esta metodologia às moléculas de DNA de bacteriófagos e moléculas de DNA com 5-bromouracila no lugar de timina, perceberam que poderiam ser formadas uma ou mais bandas, caracterizando padrões unimodal, bimodal ou polimodal. Estas observações foram fundamentais para que se fosse possível determinar o peso molecular das macromoléculas testadas, dependendo da posição onde eram formadas as bandas dentro do gradiente de densidade.

Já no final da década, ROLFE e MESELSON (1959), utilizando a mesma metodologia, observaram que *Escherichia coli* apresentava um padrão de densidade relativamente homogêneo no gradiente. A soma desta informação ao fato de que as bases guanina e citosina são molecularmente mais ricas em nitrogênio que adenina e timina, levaram à conclusão de que a homogeneidade na densidade de material genético de *E. coli* refletia a homogeneidade de sua composição de bases nucleotídicas. Foi assim demonstrado que é possível correlacionar a densidade do DNA com sua composição de nucleotídeos.

SUEOKA (1961) testou esta metodologia com o DNA de três espécies de caranguejos do gênero *Cancer*. Duas destas espécies apresentaram bandas secundárias com densidade menor que a banda principal, 1,683 e aproximadamente 1,705, respectivamente (*C. borealis* representando 30% do DNA total; *C. irroratus* representando 10% do DNA total). O autor sugeriu que estas bandas deviam ser uma fração do DNA ligada a materiais menos densos (proteínas, por exemplo), DNA de algum organismo simbiote, ou então alguma anomalia metabólica. Mas ainda assim, o autor não descartou a possibilidade de ser um elemento genético. Como mencionado anteriormente, a densidade da banda secundária era de 1,683, e no trabalho de SUEOKA, MARMUR e DOTY (1959) foi demonstrado que uma sequência sintetizada enzimaticamente chamada de deoxipoli-AT (dAT) tinha uma densidade de 1,68. Então, como as duas sequências tinham densidades muito parecidas, o autor sugeriu que esta banda menor pudesse ser esta sequência dAT natural.

Ainda no final da década de 1950, MARMUR e DOTY (1959) perceberam que a desnaturação com o aumento de temperatura de moléculas de DNA seguia um padrão definido. Ao submeter amostras de DNA natural e de DNA sintetizado (era constituído somente por AT) a um aumento de temperatura, era possível observar uma curva de desnaturação, que consistia na sobreposição de transições de moléculas únicas de DNA. Como este processo envolve a quebra de ligações entre as bases nucleotídicas, a quantidade destas ligações por pares de bases (adenina e timina possuem menos ligações de hidrogênio que citosina e guanina) influenciava diretamente sua estabilidade. Isto resultava na correlação da temperatura de desnaturação com a composição nucleotídica de um determinado segmento de DNA. Então, quanto maior a porcentagem de GC presente dentro de um segmento de DNA, maior vai ser a temperatura necessária para que ocorra a sua desnaturação.

Além disso, a dispersão da curva, ou seja, a diferença entre a temperatura final e inicial de desnaturação também indicava se uma amostra é mais heterogênea ou não em sua composição. No caso das espécies estudadas por MARMUR e DOTY (1959), foi observado que quanto maior a diferença entre as temperaturas, mais heterogênea é a composição de nucleotídeos daquela amostra. Outra característica importante observada pelos autores foi que, enquanto as amostras de DNA natural apresentavam irreversibilidade no processo de desnaturação, a amostra de DNA-AT sintético era completamente reversível, e foi explicado como sendo uma característica de um DNA repetitivo.

A reversibilidade foi estudada mais a fundo alguns anos mais tarde. Já era sabido na época que ao desnaturar moléculas de DNA e resfriá-las rapidamente, o estado desnaturado seria mantido, mesmo que alguns pareamentos de bases não específicos fossem restabelecidos. No entanto, MARMUR e DOTY (1961) perceberam que se a amostra desnaturada fosse resfriada lentamente, ela sofreria o processo de renaturação, o que resultava na restauração, mesmo que parcial, de suas características. E assim como no processo de desnaturação, a temperatura de renaturação de uma amostra também dependia da porcentagem de CG presente no fragmento utilizado (MARMUR; LANE, 1960).

A denominação “banda de DNA satélite” da banda deslocada da porção principal de DNA devido à centrifugação em gradiente de CsCl apareceu pela primeira vez em KIT (1961), porém foi somente no final da década seguinte (PECH; IGO-KEMENES; ZACHAU, 1979) que entrou-se em convenção de que este termo seria utilizado para a denominação deste tipo de sequências. Mas ainda sem entender a razão desta banda existir, KIT (1962) propôs, sem sucesso, explicações para este fenômeno. Como por exemplo, (i) as diferenças nas densidades seriam relacionadas à composição nucleotídica das sequências presentes em cada banda; (ii) a banda

satélite deslocada poderia representar DNA ligado a proteínas; (iii) o grau de hidratação das bandas era diferente; (iv) presença de metilcitosina ou outras bases nucleotídicas traço no DNA da banda satélite.

A banda satélite também representava, em alguns organismos, o DNA mitocondrial, como no caso de levedura, galinha e pombo (BOND et al., 1967; BORST; RUTTENBERG, 1966; CORNEO et al., 1966). Nestes organismos, a banda satélite de origem mitocondrial possuía uma densidade maior do que a da banda principal de origem nuclear. No caso de mamíferos, como de camundongo, do porquinho-da-índia, da vaca, da ovelha e do coelho, o DNA mitocondrial possui a mesma densidade que o DNA nuclear da banda principal. No entanto, em ambos os casos, a curva de renaturação também se diferenciava, pois o DNA mitocondrial se renatura mais rapidamente que o DNA nuclear presente na banda principal, o que também indica que o DNA mitocondrial é mais homogêneo (BOND et al., 1967; BORST; RUTTENBERG, 1966; CORNEO et al., 1966).

Assim como que para o DNA mitocondrial, começou-se a perceber que fragmentos da porção satélite reassociavam-se mais rapidamente que fragmentos da banda principal de DNA, o que também indicava sua homogeneidade (BOND et al., 1967; FLAMM et al., 1966; FLAMM; MCCALLUM; WALKER, 1967). Além disso, já se havia provado que o DNA mitocondrial e o DNA satélite eram moléculas diferentes (BOND et al., 1967; FLAMM et al., 1966; WARING; BRITTEN, 1966), portanto uma nova hipótese para explicar o DNA satélite surgiu. A fim de que a reassociação ocorresse, era necessário que fitas complementares se colidissem. Esta taxa de reassociação era influenciada, então, pela concentração de fragmentos complementares na amostra (MARMUR; LANE, 1960). Estes experimentos levaram à conclusão de que a banda de DNA satélite era composta por sequências pequenas de nucleotídeos, repetidas milhares de vezes, pouco variáveis e organizadas *in tandem* (FLAMM; MCCALLUM; WALKER, 1967; WARING; BRITTEN, 1966).

Além disso, baseada também no princípio de desnaturação e renaturação do DNA, foi desenvolvida a metodologia de formação de híbridos de moléculas de DNA normais e marcadas com isótopos (SCHILDKRAUT; MARMUR; DOTY, 1961). Esta metodologia foi primeiramente aplicada em genomas pequenos, como em vírus e bactérias. Estes resultados mostraram a formação de três tipos diferentes de moléculas: o primeiro sendo moléculas compostas de fitas normais de DNA, o segundo sendo moléculas híbridas, e o terceiro, moléculas compostas por duas fitas contendo isótopos radioativos. Quanto mais diferentes eram as duas moléculas testadas, menos híbridos eram formados, até que quando não existisse nenhuma

relação genética entre as duas, nenhum híbrido era formado (HOYER; MCCARTHY; BOLTON, 1964; SCHILDKRAUT et al., 1962; SCHILDKRAUT; MARMUR; DOTY, 1961).

A partir disto, a metodologia foi extrapolada para os DNA satélites. FLAMM, MCCALLUM e WALKER (1967) estudaram a formação de duplexes de DNA entre amostras de DNA isolado da banda satélite de camundongo, com amostras de DNA isolados de porquinho-da-índia e ratos. A não formação de duplexes híbridos entre as amostras de DNA das espécies demonstrou que cada uma possuía sequências espécie específicas. Este estudo também possibilitou a descoberta de que estas sequências possuem pouca informação genética, que levou os autores ao questionamento de que o genoma possui outras funções além de codificar proteínas.

Um parâmetro importante para se estimar o tempo de conclusão da renaturação das moléculas de DNA foi observado por BRITTEN e KOHNE (1968), que o chamaram de C_{ot} . Este parâmetro representa o produto da concentração de DNA dentro de uma amostra vezes o tempo de incubação. Ou seja, quanto maior o C_{ot} , ou o tempo de incubação necessário para que se complete a reação, mais heterogênea será a amostra. A observação dos dados de reassociação sob este parâmetro possibilitou a percepção de que o genoma era composto por sequências repetitivas em diferentes escalas, além de sequências únicas que não se repetem. Com base nos padrões de reassociação destas sequências, foi então possível identificar as sequências altamente e moderadamente repetitivas, purificá-las e posteriormente usá-las em outros estudos.

A heterocromatina e a eucromatina foram isoladas em camundongos, utilizando a técnica de sonicação, que consistia no uso de ondas sonoras para fragmentar uma amostra de DNA. Quando estas amostras de DNA fragmentado de camundongo foram centrifugadas em $CsCl_2$, observou-se que a heterocromatina constitutiva era composta primariamente por DNA satélite (YASMINEH; YUNIS, 1969, 1970). Esta mesma observação foi constatada em estudos com porquinho-da-índia e bezerro (YASMINEH; YUNIS, 1970, 1971; YUNIS; YASMINEH, 1970).

A partir de técnicas de hibridação de rRNAs marcados com isótopos radioativos e fitas de DNA imobilizadas em meios sólidos, semi-sólidos e membranas (BOLTON; MCCARTHY, 1962; GILLESPIE; SPIEGELMAN, 1965), foram propostas metodologias de localização cromossômica de sequências específicas de DNA. No primeiro momento, foram localizados os DNA ribossomais a partir da hibridação de fragmentos de rRNAs em oócitos de *Xenopus* (JOHN; BIRNSTIEL; JONES, 1969; PARDUE; GALL, 1969) e em células HeLa (JOHN; BIRNSTIEL; JONES, 1969). Em *Xenopus*, esta hibridação ocorreu somente em regiões do *cap*

extracromossômico, e então, esta metodologia foi utilizada para a localização de sequências teste de DNA cromossômico na mesma espécie (PARDUE; GALL, 1969).

Esta técnica possibilitou a localização dos DNAs repetitivos, que em camundongo mostrou que estão localizados em regiões de heterocromatina constitutiva centromérica do genoma (JONES, 1970; PARDUE; GALL, 1970a). Além de camundongos, *Rhynchosciara hollaenderi*, *Triturus viridescens* e *Drosophila melanogaster* também apresentaram DNAs satélites nas regiões centroméricas (PARDUE; GALL, 1970b). E ainda, corroborando com os experimentos de formação de duplexes interespecíficos de FLAMM, MCCALLUM e WALKER (1967), JONES (1970) observou que não houve a hibridação de DNAs satélites de camundongo nos cromossomos de coelho.

Além disso, metodologias com técnicas mais complexas de centrifugação em gradiente de CsCl aumentaram a resolução das bandas, o que revelou famílias de DNAs repetitivos que antes se mostravam crípticas (BARNES; WEBB; DOVER, 1978; SINGER, 1982). Esta metodologia também impulsionou o descobrimento das famílias de sequências repetitivas de DNA presentes no grupo *Drosophila* (BARNES; WEBB; DOVER, 1978), o que foi muito importante para o amadurecimento das teorias de evolução deste tipo de sequências que aconteceriam na década seguinte.

Ao mesmo tempo, começaram a aparecer estudos moleculares. Com a descoberta das enzimas de restrição (ARBER, 1978), percebeu-se que quando o DNA de espécies que continham sequências repetitivas era digerido por determinadas enzimas de restrição, padrões específicos de bandas apareciam. No caso, as bandas de maior quantidade de pares de bases eram de comprimentos múltiplos do comprimento da banda menor, padrão que representava sequências repetitivas *in tandem* (PECH; IGO-KEMENES; ZACHAU, 1979; SINGER, 1982; SOUTHERN, 1975). Entretanto, foi somente com os experimentos de hibridização *in situ* de PARDUE e GALL (1970) que a associação do DNA satélite e sequências centroméricas pôde ser comprovada.

A partir de então, uma série de estudos foram realizados para se entender a natureza e a diversidade das sequências de satDNA ou ainda das regiões heterocromáticas entre as diferentes espécies, utilizando-se a técnica de bandamento cromossômico, derivados dos trabalhos de PARDUE e GALL (1970). Estas técnicas mais modernas, em companhia do avanço das técnicas de sequenciamento (SANGER et al., 1973; SANGER; COULSON, 1975) proporcionaram estudos mais específicos e complexos das sequências repetitivas, como por exemplo a alta variabilidade deste tipo de sequências.

2.2. Evolução em concerto

Com o aprofundamento e detalhamento dos estudos principalmente das famílias de satDNA no subgrupo de espécies *melanogaster* dentro de *Drosophila*, BARNES, WEBB e DOVER (1978) perceberam que existem diferenças na presença ou ausência das famílias dentro de cada espécie. Como a filogenia do grupo já era bem conhecida por conta de estudos com cromossomos politênicos, a aparição de uma determinada família de satDNA pôde ser relacionada com a história evolutiva do grupo como um todo.

Espécies com mais relação evolutiva possuem semelhanças no aparecimento de famílias de satDNA em seus genomas. Isso indica que de alguma maneira estas famílias foram preservadas durante o processo evolutivo, porém, como as sequências repetitivas ainda não possuíam função definida, a evolução por seleção natural não era suficiente para explicar este fenômeno (BARNES; WEBB; DOVER, 1978).

Teorias que explicavam a evolução de sequências repetitivas já haviam sido apresentadas, como é o caso de processos moleculares como o *crossing over* desigual, conversão gênica ou ação de elementos de transposição (DOVER, G. A., 1982). Estes fatores dão origem ao mecanismo de deriva molecular, que por sua vez, direcionam o processo de evolução em concerto (DOVER, G., 2002; DOVER, G. A., 1982).

STRACHAN et al. (1985) estudando o DNA ribossomal de espécies de *Drosophila*, notou que os motivos repetitivos seguiam um padrão no modo de se dispersar e se modificar, até que tais mudanças alcançassem todos os arranjos repetitivos, tornando-os idênticos.

A evolução em concerto se dá por meio de seis passos bem estabelecidos, sendo comparado cada posição nucleotídica dentro de um conjunto de sequências de satDNA de duas espécies diferentes (STRACHAN; WEBB; DOVER, 1985). O primeiro passo é quando a mesma posição nas duas espécies está completamente homogênea com a mesma base nucleotídica; o segundo se dá quando a base é alterada em 25% das sequências de uma das espécies; os passos 3 e 4 são quando esta base nova é passada de 25% até 50% e de 50% até 99% das repetições; o quinto passo é visível quando todas as bases nucleotídicas de uma espécie em determinada posição são substituídas por um nucleotídeo diferente do que está presente na outra espécie; e o sexto passo representa todas as mutações presentes após a ocorrência do quinto passo.

A este fenômeno, STRACHAN et al. (1985) atribuíram a denominação de homogeneização. A homogeneização das sequências repetitivas *in tandem* leva à evolução em concerto, que a partir de mecanismos de recombinação, como o *crossing over* desigual, causam a

evolução simultânea de todos os motivos repetitivos de uma família, mantendo-os similares enquanto ocorre a divergência na estrutura e número dos mesmos (ZIMMER et al., 1980).

Desde então, a evolução em concerto de sequências repetitivas de satDNA foi utilizada para entender o nível de divergência molecular de populações de espécies diferentes, porém relacionadas, como no caso de humanos e gorilas (DURFY; WILLARD, 1990). Este tipo de comparação foi também utilizado, sem sucesso extenso, para aferir o nível de divergência da homogeneização interespecífica de satDNA. Tal característica foi somente encontrada em uma população de *Cyprinodon wriegau* (ELDER; TURNER, 1994).

2.3. Sequências de satDNA em *A. thaliana*

A popularização do sequenciamento de DNA no final dos anos 1990 e início de 2000, marcaram uma revolução nos estudos genômicos. A espécie *Arabidopsis thaliana* foi a primeira planta a ter o genoma completamente sequenciado em 2000, através de um consórcio internacional (THE ARABIDOPSIS GENOME INITIATIVE, 2000), que aliado a diversos outros fatores, como tamanho e duração do ciclo de vida, fez com que a espécie se consolidasse como modelo de estudos em plantas, tal como o camundongo para estudos em animais.

Embora o sequenciamento da *A. thaliana* tenha trazido grandes revelações sobre sua constituição genética, muito já se sabia sobre a organização do seu genoma. O número cromossômico da espécie foi definido por LAIBACH (1907). Entretanto, devido ao tamanho reduzido do genoma, 130Mpb, e aos instrumentos de observação cromossômicas da época não serem tão refinados como os de agora, a estrutura cariotípica não pôde ser descrita. Ainda assim, a sequência repetitiva de 180pb foi isolada por MARTINEZ-ZAPATER et al. (1986), mas somente foi caracterizada como centromérica por MALUZYNSKA e HESLOP-HARRISON (1991) através de hibridação fluorescente *in situ*. A família de sequências de 180pb mostrou-se ser a maior componente dos centrômeros de *A. thaliana*, presente nos cinco pares de cromossomos da espécie, sendo considerada a família de satDNA mais importante da espécie (BRANDES et al., 1997; THE ARABIDOPSIS GENOME INITIATIVE, 2000).

Esta família de satDNA está presente em algumas espécies do gênero *Arabidopsis*, estando altamente conservada em *A. arenosa*, bem como em espécies mais distantes como *A. pumila* e *A. griffithiana*. Entretanto, em outros gêneros da família *Brassicaceae* esta família não está presente ou estão em estágios avançados de diversificação (HESLOP-HARRISON; BRANDES; SCHWARZACHER, 2003; KAWABE; NASUDA, 2005).

SIMOENS et al. (1988) demonstraram que além da sequência de 180pb, outros satélites estavam presentes no genoma de *A. thaliana*. O motivo de 160pb foi um destes satDNAs descritos, sendo que este não apresentou homologia à família de 180pb, indicando uma provável origem independente. Além desta, uma família de satDNA com motivo de 500pb foi descrita e apresentando regiões de homologia com a sequência de 180pb, provavelmente originada de rearranjos e posterior amplificação.

A montagem do genoma de *A. thaliana* e a grande disponibilidade de sequências nos bancos de dados, associado à facilidade de sequenciar grandes volumes de DNA permitiram um aprofundamento sobre a diversidade da própria sequência de 180pb de forma intraespecífica ou restrita a um único ecótipo da espécie (Columbia).

NAGAKI et al. (2003), a partir de testes de imunoprecipitação de cromatina (ChIP), mostrou a interação da proteína centromérica CENH3 com diversas sequências centroméricas testadas, e dentro destas, a família de 180pb. Esta família de elementos repetitivos foi a que se mostrou em maior frequência na precipitação da proteína centromérica quando comparado com a outras famílias testadas, incluindo elementos transponíveis.

Os possíveis sítios de interação com a proteína CENH3 foram primeiramente propostos por HESLOP-HARRISON et al. (1999) que identificou dois boxes, A e B, com nível de conservação de 99% no ecótipo Columbia. Entretanto, HALL et al. (2003) em um estudo com 41 ecótipos da espécie *A. thaliana*, apontou sítios diferentes daqueles sugeridos por HESLOP-HARRISON et al. (1999). Embora os trabalhos não sejam concordantes com relação ao sítio de interação com a CENH3, ficou evidente a possível existência de uma pressão de seleção para a conservação do sítio de interação DNA-proteína, o que indica a funcionalidade destas sequências.

Corroborando com a teoria de pressão seletiva intensa, MAHESHWARI et al. (2017), em um estudo com milhares de monômeros de DNA satélites centroméricos da família 180pb, puderam separar seis grandes grupos por similaridade. Destes, apenas um grupo de monômeros aparece nos cinco centrômeros da espécie e ele mostra preferência por ligação à CENH3, provavelmente decorrente das interações com proteínas nesta região do genoma.

Referências

- ARBER, W. Restriction endonucleases. **Angewandte Chemie International Edition in English**, v. 17, n. 2, p. 73–79, 1978.
- BARNES, S. R.; WEBB, D. A.; DOVER, G. A. The distribution of satellite and main-band DNA components in the melanogaster species subgroup of *Drosophila*. **Chromosoma**, v. 67, n. 4, p. 341–363, 1978.

- BOLTON, E. T.; MCCARTHY, B. J. A general method for the isolation of RNA complementary to DNA. **Proceedings of the National Academy of Sciences**, v. 48, n. 8, p. 1390–1397, 1962.
- BOND, H. E. et al. Mouse satellite DNA. **Journal of Molecular Biology**, v. 27, n. 2, p. 289–302, 1967.
- BORST, P.; RUTTENBERG, G. J. C. M. Renaturation of mitochondrial DNA. **Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis**, v. 114, n. 3, p. 645–647, 1966.
- BRANDES, A. et al. Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidopsis thaliana* L. **Chromosome Research**, v. 5, p. 238–246, 1997.
- BRITTEN, R. J.; KOHNE, D. E. Repeated Sequences in DNA. **Science**, v. 161, n. 3841, p. 529–540, 1968.
- CORNEO, G. et al. Mitochondrial DNA in yeast and some mammalian species. **Science**, v. 151, n. 3711, p. 687–689, 1966.
- DOVER, G. Molecular Drive. **Trends in Genetics**, v. 18, n. 11, p. 587–589, 2002.
- DOVER, G. A. Molecular drive: a cohesive mode of species evolution. **Nature**, v. 299, n. 5879, p. 111–117, 1982.
- DURFY, S. J.; WILLARD, H. F. Concerted evolution of primate alpha satellite DNA. **Journal of Molecular Biology**, v. 216, n. 3, p. 555–566, 1990.
- ELDER, J. F.; TURNER, B. J. Concerted evolution at the population level: Pupfish HindIII satellite DNA sequences. **Proceedings of the National Academy of Sciences**, v. 91, p. 994–998, 1994.
- FLAMM, W. G. et al. Satellite DNA isolated from mouse liver: Some physical and metabolic properties. **Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis**, v. 123, n. 3, p. 652–654, 1966.
- FLAMM, W. G.; MCCALLUM, M.; WALKER, P. M. B. The isolation of complementary strands from a mouse DNA fraction. **Proceedings of the National Academy of Sciences**, v. 57, n. 6, p. 1729–1734, 1967.
- GILLESPIE, D.; SPIEGELMAN, S. A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. **Journal of Molecular Biology**, v. 12, n. 3, p. 829–842, 1965.
- HALL, S. E.; KETTLER, G.; PREUSS, D. Centromere Satellites From *Arabidopsis* Populations: Maintenance of Conserved and Variable Domains. **Genome Research**, v. 13, n. 2, p. 195–205, 2003.

- HESLOP-HARRISON, J. S. et al. Polymorphisms and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis* chromosomes. **The Plant Cell**, v. 11, p. 31–42, 1999.
- HESLOP-HARRISON, J. S.; BRANDES, A.; SCHWARZACHER, T. Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. **Chromosome Research**, v. 11, n. 3, p. 241–253, 2003.
- HOYER, B. H.; MCCARTHY, B. J.; BOLTON, E. T. A molecular approach in systematics of higher organisms. **Science**, v. 144, n. 3621, p. 959–967, 1964.
- JOHN, H. A.; BIRNSTIEL, M. L.; JONES, K. W. RNA-DNA hybrids at the cytological level. **Nature**, v. 223, n. 5206, p. 582–587, 1969.
- JONES, K. W. Chromosomal and nuclear location of mouse satellite DNA in individual cells. **Nature**, v. 225, n. 5236, p. 912–915, 1970.
- KAWABE, A.; NASUDA, S. Structure and genomic organization of centromeric repeats in *Arabidopsis* species. **Molecular Genetics and Genomics**, v. 272, n. 6, p. 593–602, 2005.
- KIT, S. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. **Journal of Molecular Biology**, v. 3, n. 6, p. 711–IN2, 1961.
- _____. Species differences in animal deoxyribonucleic acids as revealed by equilibrium sedimentation in density gradients. **Nature**, v. 193, n. 4812, p. 274–275, 1962.
- LAIBACH, F. Zur Frage nach der individualität der chromosomen im pflanzenreich. **Beihefte zum Botanischen Zentralblatt**, v. 22, p. 191–210, 1907.
- MAHESHWARI, S. et al. Centromere location in *Arabidopsis* is unaltered by extreme divergence in CENH3 protein sequence. **Genome Research**, v. 27, n. 3, p. 471–478, 2017.
- MARMUR, J.; DOTY, P. Heterogeneity in deoxyribonucleic acids: I. Dependence on composition of the configurational stability of deoxyribonucleic acids. **Nature**, v. 183, n. 4673, p. 1427–1429, 1959.
- _____. Thermal renaturation of deoxyribonucleic acids. **Journal of Molecular Biology**, v. 3, n. 5, p. 585–594, 1961.
- MARMUR, J.; LANE, D. Strand separation and specific recombination in deoxyribonucleic acids: biological studies. **Proceedings of the National Academy of Sciences**, v. 46, n. 4, p. 453–461, 1960.
- MARTINEZ-ZAPATER, J. M.; ESTELLE, M. A.; SOMERVILLE, C. R. A highly repeated DNA sequence in *Arabidopsis thaliana*. **Molecular & General Genetics**, v. 204, n. 3, p. 417–423, 1986.

- MESELSON, M.; STAHL, F. W.; VINOGRAD, J. Equilibrium sedimentation of macromolecules in density gradients. **Proceedings of the National Academy of Sciences of the United States of America**, v. 43, n. 7, p. 581–588, 1957.
- NAGAKI, K. et al. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. **Genetics**, v. 163, n. 3, p. 1221–1225, 2003.
- PARDUE, M. Lou; GALL, J. G. Chromosomal localization of mouse satellite DNA. **Science**, v. 168, n. 3937, p. 1356–1358, 1970.
- _____. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. **Proceedings of the National Academy of Sciences**, v. 64, n. 2, p. 600–604, 1969.
- PECH, M.; IGO-KEMENES, T.; ZACHAU, H. G. Nucleotide sequence of a highly repetitive component of rat DNA. **Nucleic Acids Research**, v. 7, n. 2, p. 417–432, 1979.
- ROLFE, R.; MESELSON, M. The relative homogeneity of microbial DNA. **Proceedings of the National Academy of Sciences**, v. 45, n. 7, p. 1039–1043, 1959.
- SANGER, F. et al. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage ϕ 1 DNA. v. 70, n. 4, p. 1209–1213, 1973.
- SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v. 94, n. 3, p. 441–448, 1975.
- SCHILDKRAUT, C. L. et al. A study of the base sequence homology among the T series of bacteriophages. **Virology**, v. 18, n. 1, p. 43–55, 1962.
- SCHILDKRAUT, C. L.; MARMUR, J.; DOTY, P. The formation of hybrid DNA molecules and their use in studies of DNA homologies. **Journal of Molecular Biology**, v. 3, n. 5, p. 595–IN16, 1961.
- SIMOENS, C. R. et al. Characterization of highly repetitive sequences of *Arabidopsis thaliana*. **Nucleic Acids Research**, v. 16, n. 14, p. 6753–6766, 1988.
- SINGER, M. F. Highly repeated sequences in mammalian genomes. **Int. Rev. Cytol.** [S.l: s.n.], 1982. v. 76. p. 67–112.
- SOUTHERN, E. M. Long range periodicities in mouse satellite DNA. **Journal of Molecular Biology**, v. 94, n. 1, p. 51–69, 1975.
- STRACHAN, T. et al. Modes and rates of change of complex DNA families of *Drosophila*. **Journal of Molecular Biology**, v. 158, n. 1, p. 37–54, 1982.
- STRACHAN, T.; WEBB, D.; DOVER, G. A. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. **The EMBO journal**, v. 4, n. 7, p. 1701–1708, 1985.

- SUEOKA, N. Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. **Journal of Molecular Biology**, v. 3, n. 1, p. 31-IN15, 1961.
- SUEOKA, N.; MARMUR, J.; DOTY, P. Heterogeneity in deoxyribonucleic acids: II. Dependence of the density of deoxyribonucleic acids on guanine–cytosine content. **Nature**, v. 183, n. 4673, p. 1429–1431, 1959.
- THE ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome. **Nature**, v. 408, n. 6814, p. 796–815, 2000.
- WARING, M.; BRITTEN, R. J. Nucleotide sequence repetition: A rapidly reassociating fraction of mouse DNA. **Science**, v. 154, n. 3750, p. 791–794, 1966.
- YASMINEH, W. G.; YUNIS, J. J. Localization of mouse satellite DNA in constitutive heterochromatin. **Experimental Cell Research**, v. 59, n. 1, p. 69–75, 1970.
- _____. Satellite DNA in calf heterochromatin. **Experimental Cell Research**, 1971.
- _____. Satellite DNA in mouse autosomal heterochromatin. **Biochemical and Biophysical Research Communications**, v. 35, n. 6, p. 779–782, 1969.
- YUNIS, J. J.; YASMINEH, W. G. Satellite DNA in constitutive heterochromatin of the guinea pig. **Science**, 1970.
- ZIMMER, E. A. et al. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. **Proceedings of the National Academy of Sciences of the United States of America**, v. 77, n. 4, p. 2158–62, 1980.

3. IMPACT OF REARRANGEMENTS DURING CHROMOSOMAL EVOLUTION AND THE APPEARANCE OF NEW 180bp satDNA FAMILIES IN *Arabidopsis*

Authors: Gabriela Moreira Machado, Gabriel Fernando da Silva, Mateus Mondin

ABSTRACT

Satellite DNA (satDNA) sequences are one of the main components of the centromeres in higher eukaryotes. During the several cycles of replication and cell division, the mechanism of molecular drive acts on the satDNA, serving as a mutation provider. The accumulation of mutations inside the repeats results in a process called Concerted Evolution, which leads to the homogenization of these sequences. It is thought that the karyotype of *A. thaliana* arose from several structural alterations of a chromosome set of an ancestral species, which were modeled based on *A. lyrata* chromosomes. Only four rearrangements occurred within the centromeric regions of the ancestral chromosomes 6, 7 and 8, originating modern *A. thaliana* chromosomes 4 and 5. In *Arabidopsis* genera there are at least four known families of 180 bp satDNA, which are spread throughout the related species of this genera. *A. thaliana* has only one known family of satDNA of 180 bp, present in all five centromeres of the species. Unlike *A. thaliana*, *A. lyrata* has three families of satDNA, and probably the ancestral species also had three families, none of them being the same as the one found in *A. thaliana*. The question is: how is it possible to a species such as *A. thaliana* end up possessing only one family of satDNA, while *A. lyrata*, which had a more stable karyotypic evolution, thus have a similar karyotype to the ancestral species, has three families? Do the breakages inside the centromere help with the emergence and homogenization of satDNA families? To answer those questions, we constructed a similarity tree of all repeats present in the sequencing of both species. With these analyses it was possible to see the pattern of distribution of these repeats around the tree. A characteristic pattern of grouping was observed in each satDNA family of each chromosome, making it possible to infer a differentiation degree due to karyotypic evolution of these species, and to observe the homogenization process of the satDNA sequences.

Keywords: Satellite DNA, Concerted evolution, Chromosomal recombination

3.1. Introduction

The mutation and evolution mechanisms of the satDNA motifs are distinct from the Mendelian and Neo-Darwinian theories, and it is called Concerted Evolution in multigene families (DOVER, G. A., 1982). This process takes place in repetitive sequences on the genomic DNA of all living organisms (LIAO, 2008). Furthermore, its main characteristic is the high similarity rate in the repetitive sequences within a species in relation to organisms of different species (ELDER; TURNER, 1995; GARRIDO-RAMOS, 2017).

The Concerted Evolution is divided into two components: the rate of change and the rate of fixation. The first is the rate of nucleotide mutation in a sequence, which can then engage

on the homogenization and fixation process. The rate of fixation, however, depends on three processes: (i) stochastic rate of fixation in sequence arrays of one chromosome (ii) stochastic rate of fixation in sequence arrays of one chromosome in a population and (iii) transfer rate between different chromosomes (COEN; STRACHAN; DOVER, 1982).

The aforementioned characteristics are promoted through molecular drive mechanisms, such as transposition elements, unequal crossing-over and genic conversion (DOVER, G. A., 1982). These factors lead to a propagation of a mutation within a multigene family because of a directional and non-random process, resulting in the molecular drive. And after the spread in an individual, the mutation will then be fixed in population through of sexual reproduction (DOVER, G. A., 1982; GARRIDO-RAMOS, 2017).

The directional fixation between sequences of different chromosomes was observed in ribosomal genes in members of the *Drosophila* group. Coen et al. (1982) observed exchanges of sequences between X and Y chromosomes in this group, which lead to a high homogenization level in sequences within sexual chromosomes in relation to other chromosomes.

Moreover, studies with *Drosophila* satDNA showed that different families have a specific rate of fixation and it can be associated with the location of these families in the genome (STRACHAN et al., 1982). HALL et al. (2003) found high degree of similarity among repetitive sequences located in one same sequenced BAC and greater divergence between motifs from different BACs in centromere sequences from species of the genus *Arabidopsis*.

Furthermore, KUHN et al. (2012) also observed interference of genome localization in the molecular drive mechanism. In this study, sequences of the 1.688 family located in euchromatin have different and independent rates of homogenization in relation to the ones located in heterochromatin.

The molecular drive has a higher rate of mutation when compared to mutation inside of genes. The high rate of mutation leads to the differentiation into new subfamilies inside a genome. But the presence of subfamilies is not restricted to one genome only, they can also appear in other species of a clade (GARRIDO-RAMOS, 2015).

This is better understood in *Drosophila melanogaster* species subgroup, where in this subgroup there are at least 3 families of satellite DNA. These families are named after their length in base pairs, 180, 360 and 500. Each of these families, excluding 180 family, have subfamilies, which are shared within the *D. melanogaster* species subgroup (STRACHAN et al., 1982).

The same also occurs in the *Arabidopsis* genus. According to KAWABE and NASUDA (2005), there are four known families of centromeric satDNA in this genus. In *A. thaliana*, for

instance, only one family is present in the genome, called pAL1. This family is well known and is spread through the centromeres of all the 5 chromosomes of the species.

On the other hand, *A. lyrata*, a close species to *A. thaliana*, has three satDNA families in its genome. The family pAa, which is also found by itself in *A. arenosa*, in *A. lyrata* is found in chromosomes AL1, AL3, AL4, and AL7 (KAWABE; NASUDA, 2005; LYSAK et al., 2006). The other two families of *A. lyrata* were first described in *A. halleri* subsp. *gemmifera*, thus the name pAge1 and pAge2 (KAWABE; NASUDA, 2005). While pAge1 was found in many chromosomes of *A. lyrata* (AL2, AL5, AL6 and AL8), the pAge2 subfamily was only found in the chromosome AL6 (LYSAK et al., 2006).

Even though all four families cited above are part of the same family of 180 bp, pAL1 and pAa have higher similarity of nucleotides and length, while pAge1 and pAge2 have different lengths due to indels (KAWABE; NASUDA, 2005).

The centromeric regions of the chromosomes 4 and 5 of *A. thaliana* suffered chromosomal alterations during the karyotypic evolution of the species. According to LYSAK et al. (2006), the chromosomes 4 and 5 were originated by a series of chromosomal rearrangements of the chromosomes 6, 7 and 8 of an ancestral karyotype (denominated AK). The reference karyotype for the ancestral karyotype was based on *A. lyrata* karyotype.

However, these rearrangements involving the centromeric regions could be linked to alterations of the satDNA, and therefore, influence in the mechanisms of the concerted evolution, or the molecular drive. Moreover, it is still not understood or discussed how chromosomes containing different families of satDNA such as chromosomes AL6, AL7 and AL8 could give origin to chromosomes AT4 and AT5, which contains only pAL1 satDNA. From these hypotheses, in order to discuss whether the rearrangements involving the centromere influence the molecular drive or not, we discuss the evolutionary behavior and the diversity patterns from the centromeric and pericentromeric satDNA sequences from AT4 and AT5, and *A. lyrata* chromosomes 6, 7 and 8.

3.2. Methodology

3.2.1. Identification and extraction of satDNA sequences from *A. thaliana*

The Ensembl Plants database (<https://plants.ensembl.org/>) was used for the compilation of satellite DNA repetitive sequences from BACs for *A. thaliana*. In this platform, in order to select BACs only from the pericentromeric region, we used specific characteristics from

the centromeres of *Arabidopsis*, such as low quantity of genes and high percentage of CG repeats (Table 1).

Tabela 1. Sequencing of the species used in this work, and database used to gather the data to harvest the satDNA sequences.

Species	Database	Name	Publication	Total length of the genome (bp)
<i>A. lyrata</i>	EnsemblPlants	v1.0	Hu et al. 2011	206,667,935
<i>A. thaliana</i>	EnsemblPlants	TAIR10	Lister et al, 2008; Filichkin et al, 2010	135,670,229

All BACs located inside these windows were collected and a DotPlot analysis was made using the software JDotter (BRODIE; ROPER; UPTON, 2004), each BAC against itself. The ones presenting repetitive sequence graphic patterns (a dark polygon) were selected and were then analyzed on the Tandem Repeats Finder software (BENSON, 1999). This software locates the exact position of each repeat on the file. The sequences presenting a motif length according to the 180 bp families of *Arabidopsis* centromeric regions presented on KAWABE and NASUDA (2005) and (MARTINEZ-ZAPATER et al. (1986) were selected, extracted and organized using Bioedit software (HALL, T. A., 1999), generating a sequence bank in FASTA format.

3.2.2. Identification and extraction of satDNA sequences from *A. lyrata*

Repetitive sequences from *A. lyrata* were collected from the same database as *A. thaliana*. However, we used a different process to get to the sequences. Based on the location of the families in each chromosome of *A. lyrata* provided by LYSAK et al. (2006) and using the consensus of these families provided by KAWABE and NASUDA (2005), we could use the BLAST tool on the database. For each chromosome we only collected scaffolds located by the BLAST of the family presented on that chromosome by LYSAK et al. (2006). Then, those scaffolds were also analyzed on the Tandem Repeats Finder software and further steps were the same as in *A. thaliana*.

3.2.3. Alignment and similarity tree construction

The extracted sequences were aligned in MEGA7 (KUMAR; STECHER; TAMURA, 2016), using the MUSCLE algorithm. The sequences with abnormal indels and unusual size were discarded in this step. The similarity tree was constructed based on the Maximum Likelihood model using the Tamura-3 parameter, with gamma distribution, with bootstrap of 1000 repetitions, according to default parameters.

3.2.4. DotPlot analysis

Consensus sequences were generated on the Consensus Maker tool on the HIV Sequences Database. All consensus sequences were used in a DotPlot against themselves, made with JDotter software (BRODIE; ROPER; UPTON, 2004). This analysis allowed the comparison of the similarity between the sequences.

3.3. Results

3.3.1. Description of the sequences and their similarities

In total, we found 2270 satDNA repeats located in centromeres of chromosomes AT4 and AT5 from *A. thaliana* and AL6, AL7 and AL8 from *A. lyrata*, according to the Ensembl Plants database (Tables 1 and 2).

Tabela 2. Chromosomes searched for the satDNA sequences of each species, number of repeats found and mean repeat length of each satDNA family. The last column is the similarity of the sequences provided by KAWABE and NASUDA (2005) from each of the families in relation to the consensus of the pAL1 family.

Species	Chromosome	satDNA Family*	Number of Repeats	Mean repeat length (bp)	Similarity with pAL1 (%)
<i>A. thaliana</i>	4	pAL1	650	178	100
	5	pAL1	991	178	100
	6	pAge1	64	168	66.1
	6	pAge2	461	176	68.3
<i>A. lyrata</i>	7	pAa	51	179	81.1
	8	pAge1	53	168	66.1
Total			2270		

*Note: Families as in KAWABE and NASUDA (2005).

Mean repeat length of each chromosome shows an approximate value of 177 ± 1 bp, except chromosome AL8, which shows a mean length of 168 bp. Moreover, 72,3% of the motifs (1641) were found in *A. thaliana*, and all from only one satDNA family, which is present in all studied chromosomes.

In the other hand, we only found 629 repeats in *A. lyrata*. This species has three different satDNA families: pAge1, located in chromosomes AL6 and AL8; pAge2, located in chromosome AL6; and pAa, located in chromosome AL7 (Table 2). The families were confirmed after the alignment of the consensus of each family from each chromosome with the sequences present in KAWABE and NASUDA (2005) (Figure 1).

```

CON TTTTGGAGCTTCT-AACAAAGAAACACTACTTTAGCTTTTAGGATCC-GGTTGCGGTTCTAG 60
a C...TGT.....G.....AG.....A.....TA.. 60
b C...T.T.....G.....AG.....A..... 60
c .....A...GG.A.....G..... 61
d .....A...GG.....G.....G..... 61
e ....C.T.---.-----G.....AGCC.....G.....AT...CT 49
f ....C.T.---.-----T.....A-CC.....G.....AT...ACT 48
g ....C.T.---.-----G.....AACC.....G.....AT...CT 49
h -----T.C.....C.C.....G.....A.....AC...AAGC... 57
i -----G.AC.....C.C.....G.....A...C.A.....ACCA.AAGC... 56

CON TTCTTATACTCAATCATACACATGAGATCTAGTCATATTTGACTCCAAAAACACTAACCAA 121
a .....C...A.....C.....-..... 120
b .....C.....-..... 120
c .....C.....A...-..... 121
d .....-.....C.....A... 121
e .....G.....CC.-...T.....G...C...CA... 109
f .....CC.-...T.....G.....C...CA... 108
g .....G.....CC.-...T.....G.....G...C...CA... 109
h .....T...A...G-...T.....C.G...G... 117
i .....T.T...A...ATG...T.....A.....C.G...G... 117

CON GCTTCTTATTGCTTCTCAAAGCTTTGTGGGTGTGGCCGAAGTCC-TATGAG--TTTTTTGG 179
a C.....C.....-.....CAT.....A...A.....A.....-C..... 177
b .....C.....AT.....A.....G.....-C...T 178
c .....T.....-.....C... 178
d .....C.....T.....-..... 178
e ...A...G...C.C...T...AA...A...A.....C.....GT.....T 170
f ...A...G...TC.T...T...AT...A...A.....C.....T 167
g ...A...G...CAC...T...AA...A...A.....C.....T 168
h ..A.TG..A..G...C...T...A...GA.GA...CT...A..TG...T 176
i ...TG..A..G...C.....A.....G..GA...T...A..AG..... 176

```

Figura 1. Comparative analysis representation of the alignment of all extracted motifs from centromeric and pericentromeric regions of chromosomes 6, 7 and 8 of *A. lyrata*, the chromosomes 4 and 5 of *A. thaliana*, and the consensus sequences presented by KAWABE and NASUDA (2005) of the subfamilies of 180 bp family. (CON) Represent the consensus sequences of all sequences; (a) pAL1 subfamily consensus sequence; (b) consensus sequence of 180 bp satDNA family from chromosomes 4 and 5 of *A. thaliana*; (c) pAa consensus sequence; (d) consensus sequence of 180 bp satDNA family of *A. lyrata* chromosome 7; (e) pAge1 consensus sequence; (f) consensus sequence of 180 bp satDNA family from chromosome 6 from *A. lyrata*; (g) consensus sequence of 180 bp satDNA family from chromosome 8 from *A. lyrata*; (h) pAge2 consensus sequence; (i) consensus sequence of 180 bp family from chromosome 6 of *A. lyrata*.

The satDNAs found in each chromosome were compared in a nucleotide basis using a DotPlot tool (Figure 2). This analysis showed a clear similarity the sequences from chromosomes AT4, and AT5, which displays a continuous diagonal line. The absence of this diagonal line when compared to each of the other sequences, shows that each chromosome has its own nucleotidic constitution.

Figure 2. DotPlot graphic analysis between the consensus sequences of centromeric and pericentromeric satDNA families of *A. lyrata* (pAa and pAge2) and *A. thaliana* (pAa). Only the chromosomes 4 and 5 of *A. thaliana* and chromosomes 6, 7 and 8 of *A. lyrata* were used in this analysis. In both x and y axis, the order of sequences is the same. It is observable a similarity between the chromosomes of *A. thaliana* and chromosome 7 of *A. lyrata*. However, the chromosome 6 and 8 of *A. lyrata* do not present clear similarity with the other analyzed sequences, including the consensus of chromosome 7 of the same species.

This absence is observed when sequences from *A. thaliana* and *A. lyrata* are compared. Different degrees of similarity are shown in this case. The consensus of chromosomes AL6 and AL8 have a specific base composition due to few similarity regions between them and the consensus of other chromosomes. Differently, the AL7 consensus has a higher identity degree with chromosomes of *A. thaliana*, which confirms the similarity between the pAa and pAL1, as shown in KAWABE and NASUDA (2005).

3.3.2. Similarity tree

We constructed a similarity tree in order to comprehend the distribution and grouping of the repeats of both species among themselves, and understand, beyond this result, their evolutive behavior according to LYSAK et al. (2006).

The tree has a default characteristic according to the results shown in the DotPlot analyses (Figures 2 and 3). The pattern of distribution is clear, given that the motifs grouped according to their respective families in different branches of the tree. This distribution was expected, because the families present clear differences, such as length and base composition (Table 2, Figure 1). Repeats from pAge2 (from *A. lyrata*) and pAL1 (from *A. thaliana*) are located in the extremities of the tree, which complements the low similarity degree between these both families.

Repeats from the families pAa and pAge1 are in an intermediary position between the other repeats, and both families do not share any repeats between themselves. It is also observable that repeats from the family pAa are closer to repeats from chromosomes from *A. thaliana*.

However, the only intermingled motifs between themselves are the motifs coming from *A. thaliana*, which is a coherent result, giving that they are repeats from the same satDNA family. But, despite this, only few motifs present this distribution. Most repeats are distributed in different branches, at which point chromosome AT4 is closer to chromosome AL7 than AT5, which is in the opposite position. This distribution was contrary to what was expected, since repetitive sequences are thought to have homogeneity among the repeats. And this result could indicate different degrees of homogeneity between both *A. thaliana* chromosomes.

Figura 3. Maximum Likelihood tree of satDNA sequences between chromosomes 6 (blue), 7 (pink) and 8 (light green) of *A. lyrata* and chromosomes 4 (red) and 5 (dark green) of *A. thaliana*. Each terminal represents a single satDNA sequence. The tree represents a ML tree using all extracted sequences of each chromosome for each species.

3.4. Discussion

A. thaliana went through a series of chromosomal alterations during its karyotypic evolution, which resulted the chromosome set of $2n=10$. These chromosomal rearrangements occurred in the evolution of all five pairs of chromosomes. However, those that originated chromosomes AT4 and AT5 have an important feature: they occurred in the centromeric region of chromosomes AK6 and AK7 from the ancestral karyotype.

The first rearrangement was a robertsonian translocation between chromosomes AK6 and AK7. After this event, a paracentric inversion occurred in one of the translocated AK6/7 chromosome, originating *A. thaliana* chromosome 4 (AT4). On the other translocated AK6/7, another reciprocal translocation occurred involving chromosome AK8, in a close region of the first chromosome, resulting on the chromosome 5 from *A. thaliana* (AT5) (LYSAK et al., 2006).

Even though both *A. thaliana* chromosomes have similar origin regarding chromosome breakage, there are significant differences between these processes. In chromosome 4, for instance, the two rearrangements that occurred in the centromere region altered the genomic context in which the centromere was embedded: the translocation and the pericentric inversion (T1 and I1, respectively - Figure 4). Yet, on chromosome 5, only the first translocation altered the genomic context (T1 - Figure 4). The second translocation (T2 - Figure 4) did not alter the immediate genomic context of the centromere and pericentromeric region.

The alterations in both chromosomes were not only necessary for the physical formation of the chromosomes. They also contributed for the differences on the genomic content of the centromere.

Figure 4. Chromosome evolution diagram of *A. thaliana* (AT4 and AT5) as a result of chromosomal rearrangements of an ancestral karyotype (AK6 - green, AK7 - blue and AK8 - pink). T1 represents the first translocation in the centromere region between chromosomes AK6 and AK7, resulting in two chromosomes AK6/7. One of these chromosomes suffered an inversion (I1) involving the centromere, resulting in the chromosome AT4. The other chromosome suffered a translocation (I2) with an inverted AK8 (I2), resulting in one microchromosome which was lost during the process, and the AT5 chromosome. Translocated chromosomes 6/7 present an unknown, possibly hybrid satDNA family. The shadowed chromosomes indicate where a change in the context in which the chromosome is inserted happened. This image is an adaption of LYSAK et al. (2006).

As previously established, the presence of repetitive sequences activity on the genome may be a trigger for DNA damaging, resulted by chromosomal alterations, for example (BRADLEY; WICHMAN, 1994; NAVAJAS-PÉREZ; QUESADA DEL BOSQUE; GARRIDO-RAMOS, 2009; WICHMAN et al., 1991; ZELLER; GASSER, 2017). So, is it possible that the amount of these alterations on the centromere also interferes with the dynamics in the molecular drive on the sequences in this region? Moreover, is it possible that an alteration on these sequences can alter the concerted evolution?

The behavior of the results of the similarity analyses helped us answer these questions. For instance, in the similarity tree, the motifs of both chromosomes of *A. thaliana* grouped in different branches of the tree. This behavior suggests that there is difference of homogeneity between both chromosomes set of sequences (KUHN et al., 2012), even though they are part of the same satDNA family (pAL1). Such feature infers that the change of the genomic context of these sets, resulted from the rearrangements that occurred during the karyotypic evolution, could be the reason for this difference.

In fact, chromosome AT5 seems to be more homogenized (Figure 3), which is explained by its evolutionary story, in which only one recombination event triggered one main molecular drive process, controlled by the mechanisms mentioned before. Differently, on chromosome AT4, the two alterations triggered this mechanism both times they occurred, which explains the lower homogenization rate in relation to AT 5. This is explained by that depending on the genomic region a particular array of repeats is located, its homogenization may take place in a different rate of an array located in other location (KUHN et al., 2012). This result corroborates to the results in NAVAJAS-PÉREZ et al. (2009), that compared species from the *Rumex* genus, which have an evolutionary history marked with chromosomal rearrangements. The chromosomes that had a breakage in its evolutionary history showed a higher rate of concerted evolution.

Likewise, the structural alterations the karyotype during the evolution of *A. thaliana* also explain why there are different satDNA families between both studied species. These differences demonstrate that there were both mutation, fixation steps and accumulation of repeats in these type of sequences during the evolutive process (BRADLEY; WICHMAN, 1994).

Another question on this subject to be covered in future papers, is when analyzing all five *A. thaliana* chromosomes at once, whether the homogenization of chromosomes 4 and 5 also differs and stay isolated in the similarity tree from the chromosomes 1, 2 and 3.

References

- BENSON, G. Tandem Repeats Finder: a program to analyse DNA sequences. **Nucleic Acids Research**, v. 27, n. 2, p. 573–578, 1999.
- BRADLEY, R. D.; WICHMAN, H. A. Rapidly evolving repetitive DNAs in a conservative genome: A test of factors that affect chromosomal evolution. **Chromosome Research**, v. 2, n. 5, p. 354–360, 1994.
- BRODIE, R.; ROPER, R. L.; UPTON, C. JDotter: A Java interface to multiple dotplots generated by dotter. **Bioinformatics**, v. 20, n. 2, p. 279–281, 2004.
- COEN, E.; STRACHAN, T.; DOVER, G. A. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of *Drosophila*. **Journal of Molecular Biology**, v. 158, n. 1, p. 17–35, 1982.
- DOVER, G. Molecular Drive. **Trends in Genetics**, v. 18, n. 11, p. 587–589, 2002.
- DOVER, G. A. Molecular drive: a cohesive mode of species evolution. **Nature**, v. 299, n. 5879, p. 111–117, 1982.
- ELDER, J. F.; TURNER, B. J. Concerted Evolution of Repetitive DNA Sequences in Eukaryotes. **The Quarterly Review of Biology**, v. 70, n. 3, p. 297–320, 1995.
- GARRIDO-RAMOS, M. A. Satellite DNA: An evolving topic. **Genes**, v. 8, n. 9, p. 230, 2017.
- _____. Satellite DNA in plants: More than just rubbish. **Cytogenetic and Genome Research**, v. 146, n. 2, p. 153–170, 2015.
- HALL, S. E.; KETTLER, G.; PREUSS, D. Centromere Satellites From *Arabidopsis* Populations: Maintenance of Conserved and Variable Domains. **Genome Research**, v. 13, n. 2, p. 195–205, 2003.
- HALL, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. **Nucleic Acids Symposium Series**, v. 41, p. 95–98, 1999.
- KAWABE, A.; NASUDA, S. Structure and genomic organization of centromeric repeats in *Arabidopsis* species. **Molecular Genetics and Genomics**, v. 272, n. 6, p. 593–602, 2005.
- KUHN, G. C. S. et al. The 1.688 repetitive DNA of *Drosophila*: Concerted evolution at different genomic scales and association with genes. **Molecular Biology and Evolution**, v. 29, n. 1, p. 7–11, 2012.
- KUMAR, S.; STECHER, G.; TAMURA, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Molecular Biology and Evolution**, v. 33, n. 7, p. 1870–1874, 2016.

- LIAO, D. Concerted Evolution. **Encycl. Life Sci.** Chichester, UK: John Wiley & Sons, Ltd, 2008. .
- LÓPEZ-FLORES, I.; GARRIDO-RAMOS, M. A. The Repetitive DNA Content of Eukaryotic Genomes. **Genome Dyn.** Basel: Karger, 2012. v. 7. p. 1–28.
- LYSAK, M. a et al. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 13, p. 5224–5229, 2006.
- MARTINEZ-ZAPATER, J. M.; ESTELLE, M. A.; SOMERVILLE, C. R. A highly repeated DNA sequence in *Arabidopsis thaliana*. **Molecular & General Genetics**, v. 204, n. 3, p. 417–423, 1986.
- NAVAJAS-PÉREZ, R.; QUESADA DEL BOSQUE, M. E.; GARRIDO-RAMOS, M. A. Effect of location, organization, and repeat-copy number in satellite-DNA evolution. **Molecular Genetics and Genomics**, v. 282, n. 4, p. 395–406, 2009.
- STRACHAN, T. et al. Modes and rates of change of complex DNA families of *Drosophila*. **Journal of Molecular Biology**, v. 158, n. 1, p. 37–54, 1982.
- STRACHAN, T.; WEBB, D.; DOVER, G. A. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. **The EMBO journal**, v. 4, n. 7, p. 1701–1708, 1985.
- WICHMAN, H. A. et al. Genomic distribution of heterochromatic sequences in equids: Implications to rapid chromosomal evolution. **Journal of Heredity**, v. 82, n. 5, p. 369–377, 1991.
- ZELLER, P.; GASSER, S. M. The Importance of Satellite Sequence Repression for Genome Stability. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 82, p. 15–24, 2017.

4. CONCLUSÃO

As recombinações cromossômicas que aconteceram durante o processo evolutivo do cariótipo de *A. thaliana*, ocorridas em regiões pericentroméricas, alteraram o contexto genômico ao redor das sequências repetitivas de tais regiões. Este tipo de dano cromossômico acentua o processo de propagação das sequências de satDNA, que influencia no processo de homogeneização destas sequências.