

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

**Impactos da localização genômica no comportamento evolutivo das
sequências repetitivas *in tandem* em milho e espécies relativas**

Gabriel Fernando da Silva

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

**Piracicaba
2021**

Gabriel Fernando da Silva
Engenheiro Agrônomo

Impactos da localização genômica no comportamento evolutivo das sequências repetitivas
***in tandem* em milho e espécies relativas**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2021

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

Silva, Gabriel Fernando da

Impactos da localização genômica no comportamento evolutivo das sequências repetitivas *in tandem* em milho e espécies relativas / Gabriel Fernando da Silva. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2021.

77 p.

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

1. Evolução em concerto 2. DNA satélite 3. Centrômero 4. Knob 5. Milho I.
Título

DEDICATÓRIA

À Deus e Nossa Senhora, aqueles que são donos da minha vida.

Aos meus pais Ana e Joaquim e ao meu irmão Gustavo pelo amor e incentivo.

AGRADECIMENTOS

À meus pais, Joaquim Ferreira da Silva e Ana Lúcia do Nascimento da Silva por serem a fonte de amor e apoio humano fundamental durante os dois anos de dissertação. Mas, também, sou grato pela educação que recebi e continuo recebendo que foram essencial para o convívio pessoal e profissional. E ao meu irmão, Gustavo Felipe da Silva, pela sua companhia, amizade e até mesmo como ouvidos para discussão das minhas ideias provindas desta dissertação.

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ) e a Universidade de São Paulo (USP) pela contribuição na minha formação como Engenheiro Agrônomo e pelo apoio com a infra-estrutura de todo o desenvolvimento técnico-científico do Mestrado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro e técnico-científico no projeto 2018/09760-1 que resultou na maior qualidade no desenvolvimento e nos resultados finais desta dissertação. Meu Muito Obrigado pela confiança neste trabalho.

Ao meu orientador Prof. Dr. Mateus Mondin pelo suporte científico e profissional que recebi durante a Graduação e Mestrado. Sou grato pelo desenvolvimento acadêmico realizado com muito cuidado e cautela para que eu pudesse terminar este tempo melhor do que eu comecei. Fica aqui o meu Muito Obrigado.

Aos funcionários e alunos do Laboratório de Citogenômica e Epigenética, Prof^a. Dra. Margarida Lopes Rodrigues de Aguiar-Perecin, Silvia Cristina Menuzzo Molina, Renata Flavia Carvalho, Yajahaira Nevenka Carbajal Gonzales, Pedro Alexander, Frederico Hickmann e Gabriela Moreira Machado pelas vossas companhias e aprendizados cotidianos. Vossas amizades ficarão preservada para sempre.

Ao Ministério Universidades Renovadas (MUR) e toda amizade que encontrei neste grupo. Sou grato por ensinarem a amar e mostrar que a Fé e a Ciência andam de mão dadas. Fica aqui os meus sinceros agradecimentos por cada um que passou em minha vida diante deste grupo.

À República Sagrada Família, Gabriel, Márcio e Mateus obrigado pela vossa companhia e por todos os momentos de ajuda e fortaleza em minha vida. Os senhores foram essenciais na construção desta dissertação.

EPÍGRAFE

“Não é reprovável a Ciência ou qualquer outro conhecimento das coisas, pois é BOA em si e ordenada por DEUS (...). Muitos, porém, estudam mais para SABER, que para bem VIVER; por isso erram a miúdo e pouco ou nenhum fruto colhem.”.

Tomás de Kempis

Retirado do livro Imitação de Cristo

SUMÁRIO

RESUMO	8
ABSTRACT.....	9
1. INTRODUÇÃO	11
2. REVISÃO BIBLIOGRÁFICA	13
2.1. Origem	16
2.1.1. Elementos de transposição	16
2.1.2. DNA ribossomal	17
2.1.3. Regiões de íntrons.....	18
2.2. Funções	18
2.2.1. Centromérica.....	18
2.2.2. Formação da heterocromatina.....	19
2.3. Evolução.....	20
2.3.1. Exemplos de famílias de DNAs satélite na evolução em concerto.....	21
2.3.2. Casos específicos na evolução em concerto	22
2.4. DNAs Satélites de Milho	23
Referências.....	25
3. CAPÍTULO 1: LOCALIZAÇÃO GENÔMICA INFLUENCIA NO COMPORTAMENTO EVOLUTIVO E NA ORGANIZAÇÃO DAS FAMÍLIAS DE DNA SATÉLITES EM MILHO.....	29
RESUMO.....	29
ABSTRACT.....	29
3.1. Introdução	30
3.2. Materiais e Métodos.....	32
3.2.1. Materiais.....	32
3.2.2. Métodos.....	32
3.3. Resultados	33
3.3.1. Diversidade nucleotídica em arranjos de knobs e centrômeros	34
3.3.2. Organização dos motivos centroméricos	36
3.3.3. Árvore com motivos de todos os cromossomos.....	39
3.3.4. Diversidade nucleotídica dos motivos de K180 entre os braços cromossômicos	43

3.3.5. Diferenças de diversidade nucleotídica entre os knobs caracterizados e a região com menor quantidade de motivos de K180	42
3.4. Discussão	45
3.4.1. Organização do Centrômero	45
3.4.2. Centrômero 5	46
3.4.3. Homogeneização	46
3.4.4. Knob	47
3.4.5. Dinâmica Evolutiva em Diferentes Contextos Genômicos	47
Referências	48
Figuras Suplementares.....	51
4. CAPÍTULO 2: NOVA FAMÍLIA DO KNOB HETEROCROMÁTICO POSSUI UMA ASSINATURA EVOLUTIVA ENTRE AS ESPÉCIES PRÓXIMAS DE MILHO.....	56
RESUMO.....	56
ABSTRACT	56
4.1. Introdução	57
4.2. Materiais e Métodos	58
4.2.1. Materiais	58
4.2.2. Métodos	59
4.3. Resultados.....	61
4.3.1. Descrição de K180_B	61
4.3.2. K180_B está presente em espécies ancestrais	64
4.3.3. Análise de Diversidade de K180_B.....	64
4.3.4. Árvore de Similaridade – Evolução em Concerto	66
4.3.5. Comportamento evolutivo das sequências de K180_B nos braços cromossômicos.....	68
4.4. Discussão	70
4.4.1. Evolução Sequências Repetitivas K180_B.....	71
4.4.2. Evolução em Concerto de K180_B – Homogeneização Localmente.....	72
Referências	72
Materiais Suplementares.....	76

RESUMO

Impactos da localização genômica no comportamento evolutivo das sequências repetitivas *in tandem* em milho e espécies relativas

A evolução dos motivos de DNAs satélites é conduzido por meio dos mecanismos da evolução em concerto. A evolução em concerto conduz a homogeneização dos motivos em nível de arranjos, onde estes tendem a ser mais semelhantes entre os motivos mais próximos de si do que em posições mais distantes no genoma. É observado que esta evolução sobre o DNA satélite possui diferenças no seu comportamento entre os motivos presentes em famílias de espécies diferentes e possui poucos relatos sobre esta característica evolutiva entre famílias de um mesmo genoma. Para aprofundar nos estudos evolutivos de famílias de DNA satélite presente no mesmo genoma, o objetivo deste trabalho foi estudar o comportamento evolutivo de duas famílias denominadas de CentC e K180 presente no milho e que estão localizadas em dois contextos genômicos diferentes, o centrômero e o knob. Para isso, iniciou-se este trabalho com a busca dos motivos destas famílias de DNAs satélites no genoma montado da linhagem modelo B73 de milho. Posteriormente, foram coletados, organizados e alinhados para obtenção dos valores de diversidade nucleotídica, construção da MLTree e observação da organização dos motivos de acordo com a sua posição no genoma. A partir dessas análises, foram comparados os resultados entre as duas famílias de DNA satélite e foi possível visualizar que os motivos localizado na região centromérica (CentC) são mais conservados em relação aos motivos da região do knob (K180) evidenciando que, o contexto genômico pode influenciar o comportamento evolutivo dos DNAs satélites em milho. Também foi observado que a família CentC possui três grupos de motivos que possuem diversidades nucleotídicas distintas (altamente conservado de 0% a 1%, moderadamente conservado de 4% a 8% e mais polimórficos de 13% a 17%) e que, possivelmente, pode estar associado com as funções do centrômero e com associação das proteínas centroméricas. No contexto genômico do knob, foi localizado uma nova sub-família derivada da K180 que aumenta esta grande diversidade que existe neste DNA satélite. Esta sub-família foi denominada de K180_B e está presente nos knobs de diversas espécies do gênero *Zea* e na espécie *Tripsacum dactyloides* e ausente nas espécies de *Coix aquatica* e *Coix lacryma-jobi*.

Palavras-chave: Evolução em concerto, DNA satélite, Centrômero, Knob, Milho

ABSTRACT

Impacts of genomic localization on the evolutionary behavior of repetitive sequences *in tandem* in maize and relative species

The satellite DNA evolution is led through the mechanisms of Concerted Evolution. Concerted evolution leads to the homogenization of motifs at the level of arrays. It is observed that concerted evolution has differences in its behavior between the motifs present in families of different species and has few reports on this evolutionary characteristic between families of the same genome. In order to increase the knowledge of the evolutionary studies of satellite DNA families present in the same genome, the goals of this work was to study the evolutionary behavior of two families named CentC and K180 present in maize and located in two different genomic contexts, the centromere and the knob, respectively. This work started with the search for the motifs of these satellite DNA families in the assembled genome of the lineage model of maize B73. Subsequently, the motifs were collected, organized and aligned to obtain the nucleotide diversity values, construction of the MLTree and observation of the motifs organization according to their position in the genome. From these analyzes, the results between these two satellite DNA families were compared and it was possible to see that the motifs located in the centromeric region (CentC) are more conserved in relation to the motifs in the knob region (K180) showing that the genomic context may influence the evolutionary behavior of satellites DNAs in maize. It was also observed that the CentC family has three groups of motifs that have distinct nucleotide diversity (highly conserved from 0% to 1%, moderately conserved from 4% to 8% and more polymorphic from 13% to 17%) and that, possibly, may be associated with the functions centromere and with the association of centromeric proteins. Moreover, a new subfamily derived from K180 has been found that increases the great diversity that exists in this satellite DNA. This sub-family was called K180_B and is present in knobs of several species of the genus *Zea* and in the species *Tripsacum dactyloides* and absent in the species of *Coix aquatica* and *Coix lacryma-jobi*.

Keywords: Concerted Evolution, Satellite DNA, Centromere, Knob, Maize

1. INTRODUÇÃO

DNAs satélites são sequências altamente repetitivas *in tandem* que estão localizadas em grandes blocos no genoma, principalmente nas heterocromatinas das regiões teloméricas, centroméricas e àquelas adjacentes à RON. As bases nucleotídicas destas sequências possuem alta taxa de mutação e fixação, controladas por mecanismos evolutivos específicos conhecidos como *Molecular Drive* (Evolução em Concerto).

A Evolução em Concerto conduz a taxa de mutação e fixação de famílias de DNA satélite em uma homogeneização em nível de arranjos. Exemplos podem ser verificados no DNA satélite centromérico de diversas espécies de plantas e animais. Em arroz é verificado que os motivos presentes no cromossomo 1 possuem uma alta homogeneização, indicando uma semelhança entre si do que quando comparados com os motivos presentes no cromossomo 8, por exemplo. Em *Drosophila* foi verificado que motivos presentes na região de heterocromatina possuem uma taxa de homogeneização diferente dos poucos motivos que estão na região de eucromatina.

Ainda em *Drosophila* foi observado que a taxa de mutação e a homogeneização dos motivos é diferente entre as diversas famílias de DNAs satélites (STRACHAN; WEBB; DOVER, 1985; DOVER, 1982). Também é observado que as famílias de DNAs satélites localizadas somente em um cromossomo possui taxas de homogeneização distintas de famílias que estão presentes em todos os cromossomos de *Drosophila* (STRACHAN; WEBB; DOVER, 1985).

Em batatas, esta diferença do comportamento evolutivo dos motivos foi observado nas famílias de DNAs satélites subteloamérica (TORRES et al., 2012). A família mais próxima dos telômeros possui maiores taxas de polimorfismos em relação as famílias mais distante desta região. Assim, possivelmente, a ação de rearranjos que ocorrem próximo aos telômeros podem estar associado com a maior taxa de evolução destas sequências subteloamérica (TORRES et al., 2012).

Esses trabalhos evidenciam que cada família de DNA satélite possui diferentes comportamentos na evolução em concerto em relação a sua posição no genoma. Entretanto, nenhum trabalho buscou realizar uma análise sobre as diferenças na evolução em famílias de DNAs satélites que estão localizados em posição e contextos genômicos diferentes em milho.

O milho possui em seu genoma pelo menos quatro grandes famílias de DNAs satélites: CentC, Cent4, TR-1 e K180, sendo que os dois primeiros estão localizados na região

centromérica e os demais em regiões de heterocromatinas denominadas de knobs. Os knobs e a região centromérica estão localizados em diferentes contextos genômicos e com distintas funções. Os knobs são heterocromatinas constitutivas, localizada em uma região rica em gene, sem uma função bem definida e com o tempo de replicação do DNA tardia, enquanto centrômero está localizado em uma região com pouco genes, associado com a função de união da cromátide irmã, segregação dos cromossomos (MA et al., 2007) e possui forte ligação com diversas proteínas centroméricas (MCKINLEY; CHEESEMAN, 2016). Além disso, o número e as posições dos knobs heterocromáticos são muito variáveis no complemento cromossômico de milho, enquanto os centrômeros estão presentes em todos os cromossomos em todas as linhagens estudadas.

Tais características presentes no milho fazem deste organismo ser um bom modelo de estudo para a verificação da homogeneização dos motivos de DNAs satélites presentes em diferentes contextos genômicos.

Objetivo

O objetivo deste trabalho é analisar o comportamento da dinâmica evolutiva dos motivos de DNAs satélites presentes no knob e no centrômero por meio das famílias mais abundantes das respectivas regiões (K180 e CentC).

Objetivo específicos:

1) Comparar padrões de diversidade nucleotídica e homogeneização dos motivos dentro de ambas famílias para verificar se existem padrões de evolução em concerto no nível de arranjos;

2) Comparar padrões de diversidade nucleotídica e homogeneização dos motivos entre a família K180 e CentC com o objetivo de entender, se diferentes contextos genômicos afetam o padrão de diversidade das sequências de DNA satélites e;

3) Analisar o comportamento organizacional dos motivos em cada família

2. REVISÃO BIBLIOGRÁFICA

O genoma dos organismos eucarióticos está organizado em estruturas denominadas de cromossomos. Os cromossomos são constituídos por uma molécula de DNA única, compactada na forma de cromatina em uma ação conjunta, principalmente, com proteínas histônicas. Esta estrutura denominada cromossomo está compartimentalizada, de acordo com a natureza e o tipo de sequência de DNA presente em uma determinada região do genoma (SCHMIDT et al., 1998).

O genoma é constituído em sua grande parte por sequências repetitivas que podem ser classificadas em: sequências repetitivas dispersas e sequências repetitivas *in tandem*. Dentro das sequências repetitivas dispersas pode-se destacar os elementos de transposição. Os elementos de transposição podem ser divididos em dois grandes grupos: Transposons e Retrotransposons. Os Transposons são constituídos pelas famílias de elementos primeiramente descritas por Barbara McClintock (1950) e são encontrados no genoma em alto número de cópias, assim como os Retrotransposons que constituem a maior classe de sequências de DNA, sendo os maiores responsáveis pelo tamanho do genoma devido ao seu mecanismo de transposição via RNA intermediário, que multiplicam as suas cópias por meio da Transcriptase Reversa (SCHMIDT et al. , 1998).

Embora algumas classes de genes como os Ribossomais, Histonas, algumas proteínas de reserva, globulinas e outros, sejam considerados repetitivos *in tandem*, esta classe de DNAs repetitivos é principalmente caracterizada pelos DNAs satélites. Os DNAs satélites são sequências nucleotídicas não codantes caracterizadas pelas repetições de suas cópias *in tandem* (HEMLEBEN et al. , 2007, GARRIDO-RAMOS, 2017) e podem ser sub classificados em Satélites, Mini-satélites, Micro-satélites. Os micro-satélites são organizados por motivos entre 1 e 20 nucleotídeos que podem se repetir por 10 a 100 números de vezes. Os mini-satélites são caracterizados por motivos entre 30 a 90 nucleotídeos com repetições de 500 a 50000 pb (HEMLEBEN et al., 2007). Estas duas classes de DNA satélites têm como característica comum estarem dispersas ao longo do genoma. O DNA satélite é, principalmente, encontrado nas regiões de heterocromatina constitutiva e são formados por motivos de 120 a 740 pb (HEMLEBEN et al., 2007, MEHROTRA e GOYAL, 2014).

Por estarem principalmente localizados nas heterocromatins constitutivas não apresentam uma característica de dispersão pelo genoma, tal como as duas primeiras classes, mas podem estar localizados em diferentes cromossomos de acordo com a distribuição desta estrutura da cromatina (MEHROTRA e GOYAL, 2014). Embora as heterocromatins constitutivas possam estar em diferentes regiões do genoma, três delas são bastante conhecidas e caracterizadas, tais como a heterocromatina telomérica, a heterocromatina ao redor da RON e a heterocromatina centromérica (SCHMIDT et al., 1998).

Os DNA satélites foram descobertos por meio do uso de gradiente de densidade por Cloreto de Césio (CsCl). Neste tipo de separação, uma partícula de determinada densidade encolherá durante a centrifugação até atingir uma posição onde a densidade da solução ao redor é exatamente igual à densidade da partícula. Os primeiros estudos de gradiente de densidades observaram bandas características para cada espécie e obtiveram um comportamento bimodal em camundongos e porquinhos-da-índia, ou seja, com a presença de duas bandas bem evidentes (KIT, 1961). Observava-se uma banda maior, também chamada de principal e adjacente a ela e observado uma banda menor, que poderia estar abaixo ou acima de acordo com sua constituição de nucleotídeos. Esta banda menor devido a seu comportamento no gradiente de densidade em relação a banda principal, foi denominada de satélite. Kit (1961) observou que os valores médios de densidade são variáveis de acordo com as espécies, porém as densidades não possuem diferenças de valores entre os diversos tecidos dentro de uma mesma espécie. Estes resultados evidenciam que existem diferenças estruturais destes DNAs entre as espécies de camundongos, porquinho-da-índia, jacarés e macacos; porém há uma alta similaridade dos satélites localizados em tecidos de uma mesma espécie.

Entretanto, mesmo sabendo que existiam uma parte do DNA que se diferenciavam no gradiente de densidade, ainda não se sabia que esta região era repetitiva. A descoberta das sequências repetitivas em DNAs satélites foi descoberta por meio de estudos de cinética de reassociação do DNA (BRITTEN e KOHNE, 1968). Estes estudos promovem a desnaturação e a reassociação (renaturação) de todo o DNA de um determinado organismo com o respectivo cálculo do tempo destes processos. Em camundongos foi observado que 10% do DNA foi renaturado extremamente rápido e que esta característica está associada com as sequências de DNA altamente repetitivas (BRITTEN e WARING, 1965; BRITTEN e KOHNE, 1968), por causa da maior quantidade deste DNA disponível para a renaturação do DNA.

Posteriormente, novos métodos mais simples e eficientes para a descoberta, isolamento e quantificação de famílias de DNAs satélites foram sendo descobertos e substituindo a centrifugação por cloreto de césio, como o tratamento com endonucleases de restrição do DNA genômico, dot-blot, hibridização por Southern blot, *Polymerase Chain Reaction* (PCR) e hibridização *in situ* (GARRIDO RAMOS, 2017). A união destas técnicas foi importante para descobrir a localização e a organização dos motivos das famílias de DNA satélite no genoma.

Por exemplo, foi descoberta a principal família de DNA satélite centromérico em humanos que é denominada de alfa-satélite. A organização destas sequências é baseada em motivos de 171 pb arranjados em uma organização denominada de High Order Repeats (HOR - WILLARD et al., 1987; RUDD et al., 2004 e SUJIWATTANARAT, 2015). Esta família foi bastante estudada e possui muitas informações sobre a sua constituição, organização e função, sendo base de estudo para outras espécies.

Os camundongos possuem duas famílias de DNA satélites presentes na região pericentromérica denominados de Minor Satellite e Major Satellite (JOSEPH et al., 1989). A organização destas famílias é distinta no centrômero, enquanto a Major Satellite está presente nos centrômeros de todos os cromossomos e sobre toda a heterocromatina; a Minor Satellite está presente em menos da metade dos cromossomos e em regiões de heterocromatina mais descondensada (JOSEPH et al., 1989).

Em cavalos, foi observado que apenas os motivos da família 37Cen de DNA satélite estão associados com a região funcional centromérica. Aproximadamente, 97% destas sequências constituem a interação com a histona centromérica CENP-A e estão dispostas em uma organização head-to-tail (CERUTTI et al., 2016). As outras famílias de DNA satélites estão localizadas em regiões pericentroméricas e não possui uma função centromérica (CERUTTI et al., 2016).

Em espécies de plantas como: arroz, banana, milho, *Medicago truncatula*, tabaco também foram descritas e estudadas famílias de DNA satélite (MEHROTRA e GOYAL, 2014).

Em *M. truncatula* as sequências MtR3 constituem a região centromérica, enquanto, que as sequências MtR1 e MtR2 são encontradas adjacentes a MtR3, na região pericentromérica (KULIKOVA et al., 2004). No gênero de *Arabidopsis*, também, é observado diferentes famílias de DNA Satélite localizadas no centrômero: pAL, pAa, pAge1, pAge2 (KAWABE; NASUDA, 2005; LYSAK et al., 2006). Cada uma dessas famílias possui comportamentos distintos em cada espécie de gênero, por exemplo, em *Arabidopsis thaliana* está presente somente a família pAL e localizada em todos os cromossomos (MARTINEZ-ZAPATER et al., 1986; The Arabidopsis Genome 2000), em *Arabidopsis arenosa* somente a pAa está presente e, também, localizada em todos os cromossomos (KAWABE; NASUDA, 2005). Entretanto, em *Arabidopsis lyrata* cada cromossomo possui uma família destes DNA Satélites localizados no centrômero (BERR et al., 2006), exceto a família pAL. A família pAa está localizada nos cromossomos 1, 3, 4 e 7; enquanto a pAge1 nos cromossomos 2, 5, 6 e 8 e a pAge2 somente no cromossomo 6.

Atualmente, com as novas ferramentas de bioinformática associada com as técnicas de hibridização *in situ*, foram encontradas novas características sobre a localização e a organização de famílias de DNA satélites em um mesmo genoma (RUIZ et al., 2016). Em gafanhotos foram encontradas mais de 60 famílias de DNAs satélites em diferentes regiões do cromossomo (RUIZ et al., 2016).

Em *Lathyrus sativus*, uma espécie vegetal da família da fabaceae, encontraram 11 famílias de DNA satélite (VONDRAK et al., 2019). Entre estas famílias, somente duas estão organizadas em longos arranjos típicos de DNA satélites, enquanto as outras 9 famílias possuem pequenos arranjos de motivos.

O uso destas mesmas ferramentas em *Vicia faba* localizaram 23 novas famílias de DNA satélite. Com a caracterização de todas essas famílias por FISH, ROBLDILLO et al. (2018) especificaram regiões de cromossomos e loci de heterocromatina que possuem alta quantidade de

motivos e consideraram estas regiões como possíveis *hotspots* para a acumulação desta classe de DNA repetitivo.

2.1. Origem

A origem dos DNAs satélites no interior do genoma ainda não está bem estabelecida. Diferentes estudos levantam diversas origens possíveis desta classe de DNA, entre elas estão: amplificação de várias sequências genômicas originadas de 1) partes de DNA repetitivos dispersos; 2) sequências de cópias simples de íntrons; 3) expansão de pequenos arranjos presentes nos espaçadores do DNA ribossomal; 4) reparo da replicação de DNA e 5) região hipervariável de LTR (VONDRAK et al., 2019). Destacam-se as três primeiras teorias citadas acima.

2.1.1. Elementos de transposição

Diversos estudos levantaram a hipótese de que determinados organismos originaram uma família de DNA satélite por meio dos elementos de transposição. A *Lathyrus sativus* é um exemplo destes organismos, onde foram encontradas 9 famílias de DNAs satélites derivados de pequenos arranjos de elementos de retrotransposição LTR (VONDRAK et al., 2019).

Vondrak et al. (2019) definiram 2 diferentes cenários entre esta associação de DNAs satélites com elementos de retrotransposição: 1) um longo arranjo de DNA satélite originou-se por expansão de sequências *in tandem* que, originalmente, estavam presentes dentro do elemento *Ogre*; 2) os grandes arranjos de DNA satélites são ancestrais e não relacionados com elementos *Ogre*, porém seus fragmentos foram capturados por alguns elementos de transposição e a seguir foram dispersos no genoma.

Em outras espécies de plantas, também, foram visualizadas a associação de sequências repetitivas *in tandem* com elementos de retrotransposição. Em um *screening* para mapear a presença de motivos repetitivos em regiões internas de retrotransposons de plantas, foi observado que a grande maioria destes motivos estão localizados na linhagem TAT da família TY3/*gypsy* (MACAS et al., 2009). Cerca de 42% das sequências de DNAs satélites deste *screening* estão presentes em TAT, enquanto somente 9,5% nas outras linhagens de *gypsy* e 7,5% em TY1/*copia* (MACAS et al., 2009). Os motivos *in tandem* estão localizados na região de 3' UTR. Esta região possui alta variabilidade em suas sequências e no tamanho dos motivos em cada espécie analisada, evidenciando a sua característica espécie-específica que é semelhante aos DNAs satélites (MACAS et al., 2009).

No gênero *Solanum* também foi observada a associação entre DNAs satélites e elementos de retrotransposição do tipo LTR. Entre os 6 centrômeros com a presença de DNAs satélites, 4 desses compartilham semelhanças (GONG et al., 2012) Essas similaridades estão associadas aos retrotransposons do tipo LTR, não autônomo e sequências de *solo-LTR* (GONG et al., 2012). A região

semelhante a estes elementos de transposição corresponde a um segmento truncado no gene GAG (GONG et al., 2012). Desde modo, Gong et al. (2012) concluíram que os DNAs satélites tiveram a sua origem por meio de sequências LTR ou incluíram-se em uma região perto deste elemento.

Alguns autores afirmam que os elementos de transposição são responsáveis pela origem e também pela expansão dos DNAs satélites no genoma (McGURK e BARBASH, 2018; VONDRAK et al., 2019). A expansão ocorreu preferencialmente nas heterocromatinas e nas constrições secundárias (VONDRAK et al., 2019). A expansão se deve, possivelmente, pela preferência de alguns locais do genoma à inserção dos elementos de transposição. Entre estas regiões, está o *target site duplications* (TSD), uma região do genoma que possui um sítio de nucleotídeos específicos, onde são inseridos dois elementos (McGURK e BARBASH, 2018). TSD é o principal mecanismo pelo qual dímeros de elementos de transposição surgem e criam uma grande quantidade dos mesmos. Essa abundância de dímeros possui o potencial de fornecer material de origem para futura criação de arranjos de DNAs satélites (McGURK e BARBASH, 2018).

2.1.2. DNA ribossomal

Teorias semelhantes ao verificado com elementos de transposição foram observados nas sequências ribossomais em diversas espécies.

Sequências de DNAs satélites de tomateiro são semelhantes às sequências espaçadoras intergênicas do DNA ribossomal (IGS). Essa semelhança sugere uma hipótese de que os DNAs satélites foram originados dos genes ribossomais. Outros resultados a partir de estudos genômicos e de FISH complementam esta hipótese. Nas análises com FISH do rDNA de 45S e IGS foram encontrados 4 sinais evidentes no genoma de tomate; enquanto o FISH utilizando-se somente a sonda do rDNA de 45S permitiu a visualização de apenas 1 grande sinal (JO et al., 2009). Possivelmente, os outros 3 sinais que foram observados no primeiro experimento sejam correspondentes ao IGS que teria dado origem aos motivos de DNAs satélites (JO et al., 2009).

Foi observado, também, em leguminosas como *Vicia faba*, *Vigna radiata* e *Phaseolus vulgaris* uma grande amplificação de motivos de DNAs satélites com homologia às sequências IGS do DNA ribossomal (FALQUET et al., 1997). Em particular, *Vicia sativa* possui de 83% a 84% de similaridade entre o clone de DNA satélite denominado de S12 com os subrepeats de IGS, confirmando a alta similaridade entre estas sequências (MACAS, NAVRATILOVA e MESZAROS, 2003). Macas, Navratilova e Meszaros (2003) compararam sequências disponíveis em banco de dados de DNAs satélites com os subrepeats IGS entre diversas espécies de plantas e observaram que os motivos são sempre mais similares aos subrepeats dentro de uma mesma espécie em relação às mesmas sequências comparada entre espécies distintas. Este resultado foi semelhante ao observado na comparação de sequências de elementos de transposição e motivos de DNAs satélites.

Ainda em *Vicia sativa* foi descrito algumas variantes de DNA satélite que mostraram pequenas sequências divergentes do consenso, revelando que essas subfamílias são amplificadas somente no conjunto do DNA satélite S12 e não são encontradas na região IGS dos genes ribossomais (MACAS, NAVRATILOVA e MESZAROS, 2003). E para complementar, as sequências que estão localizadas no DNA ribossomal estão presentes em todos os arranjos de S12, demonstrando que as variantes evoluíram depois de uma possível propagação dos motivos de S12 no genoma de *V. sativa* (MACAS, NAVRATILOVA e MESZAROS, 2003).

2.1.3. Regiões de íntrons

Além das sequências de elementos de transposição e espaçadores intergênicos de genes ribossomais, outras regiões do genoma também foram vistas como a possível origem de DNAs satélites, como as sequências de íntrons em espécies de primatas do novo mundo.

O macaco-prego-das-Guianas (*Sapajus apella*) e outros primatas do novo mundo contém uma família de DNA satélite conhecida de CapA (1500 pb). Valeri et al. (2018) mostraram que CapA é um abundante DNA satélite presente nestas espécies de primatas, porém nos genomas de diversos mamíferos do clado Eutheria – incluindo os seres humanos – o motivo desta família está presente somente como uma cópia única localizada no interior de um grande íntron de um gene associado à síntese de óxido nítrico (VALERI et al., 2018)

A partir deste resultado, os autores criaram a hipótese de que essa sequência gênica originou a CapA e propagou-a em um processo de transferência deste segmento para regiões de heterocromatina de um genoma ancestral dos primatas do novo mundo e, posteriormente, a ocorrência de uma expansão por *crossing-over* desigual (VALERI et al., 2018).

2.2. Funções

Ainda não existem estudos que abordam funções de DNAs satélites que é exercido diretamente por estes motivos. As funções que serão citadas abordarão efeitos indiretos destas sequências dentro do funcionamento do genoma e dos cromossomos.

2.2.1. Centromérica

Os DNAs satélites constituem a região centromérica de diversas espécies contendo uma alta correlação com as funções do centrômero. Os centrômeros são regiões heterocromáticas que formam uma constrição primária localizada nos cromossomos de organismos eucariotos com função bem definida, sendo responsáveis pela ligação dos cromossomos às fibras do fuso. Esta etapa é essencial para a dinâmica da divisão celular, pois as fibras irão se ligar em uma série de proteínas que se depositam sobre a heterocromatina centromérica, denominada de cinetócoro. Esta complexa região centromérica

ainda é responsável por manter as cromátides irmãs unidas após a duplicação do DNA, característica fundamental para a segregação normal dos cromossomos durante as divisões meiótica e mitóticas (MA et al., 2007).

A cromatina da região centromérica é diferencialmente destacada das outras regiões do cromossomo devido a presença da histona variante H3, conhecida como CENP-A em humanos ou CENH3 em vegetais (HENIKOFF, AHMAD e MALIK, 2001; McKINLEY e CHESSEMAN, 2016; COMAI, MAHESHWARI, MARIMUTHU, 2017). Esta histona é responsável em interagir com as sequências de DNA que constitui a região centromérica. Em muitos organismos, como descrito acima, estas sequências de DNA são caracterizadas pelos DNAs satélites (COMAI, MAHESHWARI, MARIMUTHU, 2017). Existem evidências que a CENH3 se liga no DNA baseado em um sinal epigenético ainda não conhecido, possivelmente do próprio CENH3 e que a alta concentração em uma região de CENH3 é suficiente para a função do cinetócoro (COMAI, MAHESHWARI, MARIMUTHU, 2017).

A comparação entre a CENH3 de *Drosophila* (Cid) com a de outras espécies do gênero demonstram que a região da cauda N-terminal e o loop 1, uma região dentro do domínio *core* da Cid, possuem mudanças adaptativas durante a evolução (HENIKOFF, AHMAD e MALIK, 2001). Loop 1 é a região do *core* que possui ligação direta entre a histona H3 e o DNA, sugerindo, então, que a histona centromérica é uma sequência dependente de uma região específica do DNA (HENIKOFF, AHMAD e MALIK, 2001).

2.2.2. Formação da heterocromatina

Heitz em 1928 identificou a heterocromatina como um material do núcleo que possui um padrão de coloração diferenciado durante o ciclo celular (PASSARGE, 1978). Atualmente, foi acrescentado no conceito de heterocromatina, características epigenéticas como a presença de uma extensa metilação da lisina 9 da histona H3 (H3K9) em que a metilação nesta histona ocorre por uma metiltransferase conservada durante a evolução, a SU(VAR)3-9 e a presença da proteína heterocromática (HP1 e Swis4) (HUSINGA, BROWER-TOLAND e ELGIN, 2006). A formação de heterocromatina é bem descrita em *Schizosaccharomyces pombe* e contém o auxílio de transcritos de DNAs satélites para esta função.

O processo de formação da heterocromatina neste organismo possui uma essencial participação das sequências repetitivas *in tandem* centroméricas. Este processo inicia com a transcrição dos motivos de DNA satélite *dg-dh* da região centromérica da levedura para um RNA de fita simples e com a posterior geração de RNA de fita dupla conduzida pelo complexo denominado de RDRC (VOLPE et al., 2002; HUSINGA, BROWER-TOLAND e ELGIN, 2006). A enzima Dicer é responsável em realizar pequenos cortes nos RNAs de fita dupla gerando diversos micro RNAs de interferências (microRNAi) (VOLPE et al., 2002; HUSINGA, BROWER-TOLAND e ELGIN, 2006). Estes

microRNAi serão incorporados em um conjunto de proteínas denominado de RITS. A união dos microRNAi com o complexo RITS irá atuar na incorporação de modificações responsáveis em determinar a característica de heterocromatina no centrômero de *S. pombe* (VOLPE et al., 2002; HUSINGA, BROWER-TOLAND e ELGIN, 2006). Caso não haja a incorporação dos microRNAi no complexo RITS, não ocorrerá a deposição do radical metil na lisina 9 da histona H3 e conseqüentemente não haverá formação de heterocromatina (HUSINGA, BROWER-TOLAND e ELGIN, 2006).

Os microRNAi também possuem interação direta com as metiltransferases Suv39h. Um domínio básico da estrutura proteica da Suv39h2 possui alta afinidade com a família major satellite em camundongos e intermediária e baixa afinidade para os motivos da família minor satellite e elementos transponíveis (CAMACHO et al., 2017). A função de Suv39h é perdida em tratamento com RNase, sugerindo que um RNA funciona como um mecanismo essencial para a organização estável desta enzima na cromatina (CAMACHO et al., 2017). Assim, Camacho et al., 2017 sugerem que os motivos de major satellite possui uma função importante para a organização da cromatina centromérica na atuação como template fisiológica para a enzima Suv39h.

Na fase de desenvolvimento do embrião já se observa a importância da ação dos microRNAi para a constituição da heterocromatina. Após a fecundação ocorre a reprogramação epigenética em que as marcas de modificações de histonas dos parentais (núcleos masculino e feminino) são extensivamente alteradas. Na primeira e segunda divisão após a fecundação já se observa a importância dos transcritos de major satellite de camundongos, pois estes transcritos são requeridos para a reorganização da heterocromatina pericentromérica materna e paterna após o processo de reprogramação epigenética (CASANOVA et al., 2013).

Em cavalos, também é observado a importância dos transcritos de DNA satélite para o organismo, uma vez que a principal família de DNA satélite possui uma taxa de transcrição de seus motivos que representa um valor maior de transcritos em relação aos genes de tubulina e componente de RNA telomerase (CERUTTI et al., 2016).

2.3. Evolução

Como visto, as sequências de DNA satélites possuem diferentes constituições entre as espécies. Os motivos das sequências centroméricas das espécies vegetais são alguns exemplos desta diversidade. Os motivos da família de DNA satélite centromérica de arroz (Cent-O) possuem regiões semelhantes aos motivos de DNA satélite de milho (Cent-C) (CHENG et al., 2002). A região semelhante entre estas duas famílias equivalem a mais do que a metade do motivo consenso (CHENG et al., 2002). Essa semelhança se deve pela proximidade evolutiva destas duas espécies. Entretanto, a comparação destas mesmas sequências com uma espécie distante filogeneticamente, como por exemplo a *Arabidopsis thaliana*, é observada a ausência de qualquer similaridade entre estas.

Em humanos, também se observa este comportamento nas sequências repetitivas do centrômero. As sequências de DNA satélites alfa possuem alto grau de semelhanças entre humanos, chimpanzés, orangotangos e gorilas. As outras espécies de primatas do velho mundo tem um maior grau de polimorfismos em relação aos primatas superiores (ALKAN et al., 2007), enquanto que a maior divergência entre os mesmos alfa-satélites se encontra nos primatas do novo mundo (Américas) que contém cerca de 64% de identidade em relação às espécies do velho mundo (ALVES, SEUÁNEZ e FANNING, 1994). Em uma comparação com espécies mais distantes, como por exemplo os roedores, não se visualizou qualquer relação entre as sequências de DNA satélite centromérico. Novamente, evidenciando que estas sequências são altamente heteromórficas quanto mais distante são as espécies, filogeneticamente.

Inversamente aos motivos de DNA satélites, muitos genes de cópia única são altamente conservados entre diversos organismos. Conclui-se, então, que as sequências repetitivas *in tandem* possuem uma alta taxa de mutação dos nucleotídeos comparada às sequências gênicas codantes de cópias únicas. A forma como as mutações ocorrem e os mecanismos evolutivos envolvidos na diversificação dos DNA satélites é conduzido por um tipo de evolução denominado como Evolução em Concerto (DOVER, 1982). As sequências repetitivas deste DNA repetitivo *in tandem* sofrem frequentes mutações em uma única sequência e, ao longo de ciclos de alterações não recíprocas entre cromossomos, as mutações são propagadas para outras regiões de um mesmo ou de cromossomo distintos (STRACHAN; WEBB; DOVER, 1985).

2.3.1. Exemplos de famílias de DNAs satélite na evolução em concerto

Como abordado anteriormente, a evolução em concerto parece possuir um padrão específico de fixação de suas mutações, onde a fixação primeiramente ocorrerá a nível de arranjo e somente depois haverá a fixação em regiões mais distantes. Estudos recentes que abordam a evolução em concerto demonstraram resultados que confirmam esta teoria.

Kuhn et al. (2012) em suas análises com *Drosophila*, verificaram que o processo de homogeneização está mais evidente em motivos presente no mesmo arranjo do genoma, indicando que este ocorre diferencialmente e independentemente em diferentes regiões genômicas. Esse processo também é semelhante aos motivos de alfa satélite presente em humanos (SCHUELER et al., 2005).

A mesma homogeneização de motivos está presente em anfíbios da família Leptodactylidae e Hylodidae. A árvore de similaridade de diversas espécies dessas famílias apresentou agrupamentos de motivos de DNA satélite de uma mesma espécie dentro de um mesmo clado, evidenciando a similaridade entre motivos de uma mesma espécie em relação a motivos de outras espécies (VITORAZZI, LOURENÇO e PIMENTEL, 2014).

As análises de Hall et al. (2005) também verificaram a forte homogeneização dos motivos entre sequências vizinhas em DNAs satélites de plantas. Na sua pesquisa feita com espécies próximas do gênero *Arabidopsis*, concentrando-se em DNA satélites centroméricos, os autores encontraram alto grau de similaridade entre os motivos localizados no mesmo BAC de sequenciamento e uma menor similaridade entre motivos que estão presentes em regiões mais distantes do genoma.

Os mesmos resultados foram visualizados nas sequências repetitivas centroméricas de arroz. Análises de similaridade entre sequências de DNA satélites do cromossomo 1 e 8 de arroz comprovaram que motivos centroméricos do cromossomo 1 formam *clusters* em um mesmo braço da árvore de similaridade (LEE et al., 2006). O mesmo resultado ocorreu com os motivos centroméricos do cromossomo 8, enquanto não foi observado nenhum motivo de um cromossomo presente em um cluster de outro cromossomo (LEE et al., 2006). Os mesmos autores complementam esta análise por meio de uma comparação de grau de identidade entre os motivos de cada região centromérica estudada e, observaram que os motivos centroméricos do cromossomo 1 possui maior identidade entre os seus próprios motivos em relação ao cromossomo 8 e o inverso também foi observado (LEE et al., 2006).

A não aleatoriedade dessas sequências, conduzido por meio de um processo estocástico foi verificado em motivos centroméricos em ecótipos de *Arabidopsis thaliana* (HALL et al., 2003). Hall et al. (2003) encontraram regiões conservadas e variáveis nos motivos do ecótipo Columbia, e, estas mesmas regiões identificadas como altamente conservadas ou variável na espécie mostrou padrões similares nas sequências consensos entre os outros 41 ecótipos estudados. As regiões conservadas e variáveis, possivelmente, podem estar associadas com funções específicas dentro do centrômero desta espécie.

2.3.2. Casos específicos na evolução em concerto

Como dito anteriormente, algumas posições dentro do genoma possuem comportamentos distintos que influenciam na taxa de fixação da evolução em concerto. Nos cromossomos sexuais de *Drosophila* observou-se que sequências dentro de cada arranjo dos cromossomos 2 e 3 possui diversidade nucleotídica entre 5 a 6%, respectivamente e no cromossomo sexual X esta diferença alcança valores de até 11%, demonstrando que os cromossomos sexuais, de alguma forma, podem influenciar a diversidade nucleotídica destas sequências (KUHN et al., 2012). Diferenças na homogeneização nos cromossomos sexuais também foi verificado em genes ribossomais (genes que também possuem organização de suas sequências *in tandem*). Coen et al. (1982) observaram trocas de sequências entre os cromossomos X e Y nesse grupo, o que leva a um alto nível de homogeneização nas sequências dentro dos cromossomos sexuais em relação a outros cromossomos.

O cromossomo Y de *Rumex* também possui um efeito dos processos evolutivos da evolução em concerto. Este cromossomo possui menor taxa evolutiva em relação aos autossomos e possivelmente

isto ocorre por causa da não recombinação nesse cromossomo (NAVAJAS-PÉREZ et al., 2005; GARRIDO-RAMOS, 2017). A não recombinação nestes organismos não permite a ação do *crossing-over* desigual e, assim, diminui o processo de homogeneização destas sequências (NAVAJAS-PÉREZ et al., 2005). Outra possível consequência da ausência de *crossing-over* nos cromossomos sexuais são a presença exclusiva de famílias de DNAs satélites nestes cromossomos (NAVAJAS-PÉREZ et al., 2005).

O eussocialismo presentes em formiga permite uma melhor percepção deste efeito, uma vez que, pode influenciar no tempo evolutivo da evolução em concerto. O eussocialismo é um grande fator de impacto para a baixa taxa de evolução dos DNA satélites, pois permite pouca divisão meiótica em sua população, uma vez que somente a rainha realizará esse processo durante a sua vida. As outras castas, que compõem a grande maioria dos indivíduos, não possuem este tipo divisão e consequentemente ficam inabilitados ao processo de *crossing-over* desigual e sua transmissão para a próxima geração (LORITE et al., 2017).

2.4. DNAs Satélites de Milho

O milho é uma importante espécie para pesquisas fundamentais de citogenética e epigenética. O genoma deste organismo possui ~2,3 Gb e é composto por, aproximadamente, 32.000 genes, sendo que 85% do total do genoma é composto de elementos de transposição (SCHNABLE et al., 2009). No milho há 4 famílias distintas de DNAs satélites denominados de CentC, Cent4, 180bp-Knob e TR-1, onde cada família está localizada em diferentes regiões do genoma (PEACOCK et al., 1981; ANANIEV; PHILLIPS; RINES, 1998a; KATO; LAMB; BIRCHLER, 2004; MONDIN et al., 2014).

As famílias CentC e Cent4 estão localizadas nas regiões centroméricas dos cromossomos de milho (ANANIEV; PHILLIPS; RINES, 1998a; KATO; LAMB; BIRCHLER, 2004). Esta região centromérica é responsável por manter as cromátides irmãs unidas após a duplicação do DNA, característica fundamental para a segregação normal dos cromossomos durante as divisões meiótica e mitóticas (MA et al., 2007). A cromatina da região centromérica é diferencialmente destacada das outras regiões do cromossomo devido a presença da histona variante H3, conhecida como CENP-A em humanos ou CENH3 em vegetais (MCKINLEY; CHEESEMAN, 2016) que em milho interagem diretamente com as sequências da família de DNA satélite Cent-C para a formação do complexo centromérico (ZHANG et al. 2008).

As sequências de CentC possuem, aproximadamente, 156 pb de comprimento e são encontradas em todos os cromossomos de milho. O centrômero de cada cromossomo possui diferentes proporções de números de cópias desta sequência, que estão interdispersos entre vários tipos de retrotransposons centroméricos (ANANIEV; PHILLIPS; RIEV, 1998b). Motivos específicos de CentC e dos retrotransposons CRM são responsáveis pela associação com as proteínas centroméricas CenH3,

histona esta responsável pela conexão com as proteínas do cinetocóro (ZHONG et al., 2002; JIN et al., 2004).

Diferentemente da família CentC, a Cent4 é um DNA satélite cromossomo específico localizado na região pericentromérica do cromossomo 4 do genoma do milho (PAGE; WANOUS; BIRCHLER, 2001). Os motivos do DNA satélite Cent4 são muito mais abundantes em número de cópias do que o CentC dentro do cromossomo 4; sendo que estas duas famílias não possuem qualquer tipo de interação física. (JIN et al., 2004).

Durante a história evolutiva destas duas famílias de DNA satélite, observou-se o aumento do número de cópias da família CentC e uma redução da família Cent4, sugerindo que, a partir da domesticação desta espécie, houve uma alteração na proporção destes DNAs satélites (BILINSKI et al., 2015). Contudo, não há até o momento estudos comparativos que trace uma relação ou ancestralidade entre estas duas famílias de DNA satélites e demonstre que estas fazem parte de um mesmo grupo.

As famílias de DNA satélites K180 e TR-1 estão localizadas nas regiões de heterocromatinas denominadas de knobs. Estas heterocromatinas foram descritas primeiramente em cromossomos na fase de paquíteno da meiose e posteriormente na forma de bandas-C em metáfases mitóticas através desta metodologia específica de coloração (AGUIAR-PERECIN et al., 2000).

Os knobs heterocromáticos de milho foram descritos praticamente em todos os cromossomos, ocupando posições intersticiais, no braço curto e no braço longo de cada cromossomo; sendo a única exceção o cromossomo 10 que nunca apresenta regiões de heterocromatina knob na sua versão normal. Porém, o número e as posições destes knobs heterocromáticos são muito variáveis no complemento cromossômico de milho, tanto que estes têm servido como marcadores para caracterizar variedades e suas correlações biogeográficas. (McCLINTOCK et al. 1981; KATO; LAMB; BIRCHLER, 2004; ALBERT et al., 2013; MONDIN et al., 2014).

A família de satDNA TR-1 também está presente nos knobs heterocromáticos. Entretanto, enquanto a família de K180 é extremamente abundante e caracteriza a maior família presente em todos os knobs heterocromáticos, a família TR-1 é menos frequente e tem sido encontrada em apenas alguns knobs de cromossomos específicos. (KATO; LAMB; BIRCHLER, 2004; ALBERT et al., 2013; MONDIN et al., 2014). Na grande maioria das linhagens, as cópias de TR-1 estão localizadas no knob do braço longo do cromossomo 4 e no braço curto do cromossomo 6.

A sequência consenso da família TR-1 possui duas regiões de homologia com K180 que compartilham 64% e 76% de similaridade, respectivamente. A primeira região de homologia compartilha um trecho do motivo de 180 pb, que também aparece conservado em teosinto e *Tripsacum*, duas espécies ancestrais do milho moderno. O alto nível de homologia entre estas duas regiões nestas diferentes espécies, indicam uma conexão na história evolutiva destas famílias de DNA satélite.

Enquanto a segunda região de homologia compartilha trechos somente com as sequências K180 do milho moderno. (ANANIEV; PHILLIPS; RINES, 1998a).

Referências

- AGUIAR-PERECIN, M.L.R. Heterochromatin of maize chromosomes: structure and genetics effects. **Genetics and Molecular Biology**, v. 23, n. 4, p. 1015-1019, 2000.
- ALBERT, P. S. et al. Diversity of chromosomal karyotypes in maize and its relatives. **Cytogenetic and Genome Research**, v. 129, n. 1–3, p. 6–16, 2010.
- ALKAN, C.; VENTURA M.; ARCHIDIACONO, N. et al. Organization and Evolution of Primate Centromeric DNA from Whole-Genome Shotgun Sequence Data. **PLoS Comput Biol**, n. 9, v. 3, p. 1807 – 1818, 2007.
- ALVES, G.; SEUÁNEZ, H.N.; e FANNING, T. Alpha satellite DNA in neotropical primates (*Platyrrhini*). **Chromosoma**, n. 103, p. 262 – 267, 1994.
- ANANIEV, E. V.; PHILLIPS, R. L.; RINES, H. W. Complex structure of knob DNA on maize chromosome 9: Retrotransposon invasion into heterochromatin. **Genetics**, v. 149, p. 2025–2037, 1998a.
- ANANIEV, E. V.; PHILLIPS, R. L.; RINES, H. W. Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. **Proc. Nat. Acad. Sci.**, v. 95, n. 22, p. 13073–13078, 1998b.
- BERR, A.; PECINKA, A.; MEISTER, A. et al. Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*. **The Plant Journal**, v. 48, p.771–783.
- BILINSKI, P.; GAO, Z.; DANILOVA, V. et al. Diversity and evolution of centromere repeats in the maize genome. **Chromosoma**, v. 124, p. 57-65, 2015.
- BRITEEN, E. J.; KOHNE, D. E. Repeated Sequences in DNA. v. 161, n. 384, p. 5329-540, 1968.
- CAMACHO, O.; GALAN, C.; SWIST-ROSOWSKA, K. et al. Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA:DNA hybrid formation. **Elife**, v. 6, 2017.
- CASANOVA, M.; PATERSANK, M.; EL MARJOU, F. et al. Heterochromatin reorganization during early mouse development requires a single-stranded noncoding transcript. **Cell Rep.**, v. 4, p. 1156-1167, 2013.
- CERUTTI, F.; GAMBA, R.; MAZZAGATTI, A. et al. The major horse satellite DNA family is associated with centromere competence. **Mol Cytogenet.**, v. 9 , n. 35, 2016.
- CHENG, Z.; DONG, F.; LANGDON, T. et al. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. **The Plant Cell**, v. 14, p.1691–1704, 2002.
- COEN; E., STRACHAN, T.; DOVER, G. 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.
- COMAI, L.; MAHESHWARI, S.; MARIMUTHU, M.P.A. Plant centromeres. **Current Opinion in Plant Biology**, v. 36, p. 158–167, 2017.
- DOVER, G. Molecular drive: a cohesive mode of species evolution. **Nature**, v. 299, p.111-117, 1982.

FALQUET, J.; CREUSOT, F.; DRON, M. Molecular analysis of *Phaseolus vulgaris* rDNA unit and characterization of a satellite DNA homologous to IGS subrepeats. **Plant Physiol Biochem.**, v. 35, p. 611–622, 1997.

GARRIDO-RAMOS, M. A. Satellite DNA: An Evolving Topic. **Genes**, v. 8, n. 9, p. 230, 2017

GONG, Z.; WU, Y.; KOBLÍZKOVÁ, A. et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. **Plant Cell**, v. 24, n. 9, p. 3559–3574, 2012.

HALL S.E.; KETTLER, G.; PREUSS, D. Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. **Genome Research**, v. 13 p. 195–205, 2003.

HALL S.E.; LUO, S.; HALL, A.E.; PREUSS, D. Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. **Genetics**, v. 170 p. 1913–1927, 2005

HEMLEBEN, V.; KOVARIK, A.; TORRES-RUIZ, R.A.; et al . Plant highly repeated satellite DNA: molecular evolution, distribution and use for identification of hybrids. **Systematics and Biodiversity**, Reino Unido, v. 5, n. 3, p.277-289, 2007.

HENIKOFF S., AHMAD K., MALIK H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. **Science**, v. 293, n. 5532, p. 1098-1102, 2001.

HUISINGA, K. L.; BROWER-TOLAND, B.; ELGIN, S. C. The contradictory definitions of heterochromatin: transcription and silencing. **Chromosoma**, v. 115, n. 2 p. 110-122, 2006.

JIN, W.; MELO, J. R.; NAGAKI, K. et al. Maize centromeres: Organization and functional adaptation in the genetic background of oat. **The Plant Cell**, v. 16, p. 571-581, 2004.

JO, S.; KOO, D.; KIM, J.F. et al. Evolution of ribosomal DNA-derived satellite repeat in tomato genome. **BMC Plant Biol**, v. 9 , n. 42, 2009.

JOSEPH, A.; MITCHELL, A. R.; MILLER, O. J. The organization of the mouse satellite DNA at centromeres. **Experimental Cell Research**, v. 183, n. 2, p. 494-500, 1989.

KATO, A.; LAMB, J. C.; BIRCHLER, J. A. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. **Proc. Nat. Acad. Sci.**, v. 101, n. 37, p. 13554–13559, 2004.

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. 711-716, 1961.

KUHN, G.C.S.; KÜTTLER, H.; MOREIRA-FILHO, O. 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.

KULIKOVA, O.; GEURTS, R.; LAMINE, M. et al . Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. **Chromosoma**, v. 113, p. 276-283, 2004.

LEE, H.; NEUMANN, P.; MACAS, J. et al. Transcription and Evolutionary Dynamics of the Centromeric Satellite Repeat CentO in Rice. **Molecular Biology and Evolution**, v. 12, n. 23, p. 2505 -2520, 2006.

LYSAK, M.; BERR, A.; PECINKA, A. et al. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. **Proc. Nat. Acad. Sci.**, v. 103, n. 13, p. 5224–5229, 2006.

LORITE, P.; MUÑOZ-LÓPEZ, M.; CARRILLO, J. et al. Concerted evolution, a slow process for ant satellite DNA: study of the satellite DNA in the *Aphaenogaster* genus (Hymenoptera, Formicidae). **Org Divers Evol**, v. 17, p. 595-606, 2017.

- MA, J.; ROD, A.W.; BENNETZEN, J.L. et al. Plant centromere organization: a dynamic structure with conserved functions. **Trends in genetics**, v. 23, p. 134-139, 2007.
- MACAS, J.; NAVRÁTILOVÁ, A.; & MÉSZÁROS, T. Sequence subfamilies of satellite repeats related to rDNA intergenic spacer are differentially amplified on *Vicia sativa* chromosomes. **Chromosoma**, v. 112, p. 152–158, 2003.
- MACAS, J.; KOBLÍZKOVÁ, A.; NAVRÁTILOVÁ, A. Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. **Gene**, v. 448, n. 2, p. 198-206, 2009.
- 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.
- McCLINTOCK, B.; KATO Y., T. A.; BLUMENSCHNEIN, A. Constitución cromosómica de las razas de maíz: su significado en la interpretación de relaciones entre las razas y variedades en las Américas. Chapingo: Colegio de Postgraduados, p. 521, 1981.
- McCLINTOCK, B. The origin and behavior of mutable loci in maize. **Proc. Nat. Acad. Sci. USA** 36, 344-355, 1950
- McGURK, M.P.; BARBASH, D.A. Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. **Genome Res.**, v. 28, p. 714–25, 2018.
- McKINLEY, K.L.; CHEESEMAN, I.M. The molecular basis for centromere identity and function. **Nature Reviews Molecular Cell Biology**, v. 17, n. 1, p. 16-29, 2016.
- MEHROTRA, S.; GOYAL, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. **Genomics Proteomics Bioinformatics**, Pequim, v. 12, p. 164-171, 2014.
- MONDIN, M.; SEREJO, J. A. dos S.; BERTÃO, M. R. et al. Karyotype variability in tropical maize sister inbred lines and hybrids compared with KYS standard line. **Frontiers in plant science**, v. 5, n. October, p. 544, 2014.
- NAVAJAS-PÉREZ, R.; LA HERRÁN, R.; JAMILENA, M. et al. Reduced Rates of Sequence Evolution of Y-Linked Satellite DNA in *Rumex* (Polygonaceae). **J. Mol Evol.**, v. 60, p. 391-399, 2005.
- PAGE, B.T.; WANOUS, M.K.; BIRCHLER, J.A. Characterization of a maize chromosome 4 centromeric sequence: Evidence for an Evolutionary relationship with the B chromosome centromere. **Genetics**, v. 159, p. 291-302, 2001.
- PASSARGE, E. Emil Heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. **Am J Hum Genet**, v. 31, n. 2, p. 106–115, 1978.
- PEACOCK, W. J.; DENNIS, E. S.; RHOADES, M. M. et al. Highly repeated DNA sequence limited to knob heterochromatin in maize. **Proc. Nat. Acad. Sci.**, v. 78, n. 7, p. 4490–4, 1981.
- ROBLEDILLO, L. A.; KOBLÍKOVÁ, A.; NOVÁK, P. et al. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. **Scientific Reports**, v. 8, n. 5838, 2018.
- RUDD, M.K.; WILLARD, H.F. Analysis of the centromeric regions of the human genome assembly. **Trends in genetics**, v. 20, n. 11, p. 529-33, 2004.
- RUIZ-RUANO, F.J.; LÓPEZ-LEÓN, M. D.; CABRERO, J. et al. High-throughput analysis of the satelliteome illuminates satellite DNA evolution. **Scientific Report**, v. 6, 2016
- SCHIMDT, T.; HESLOP-HARRISON, J. S. Genomes, genes and junk: the large-scale organization of plant chromosomes. **Trends in Plant Science**, v. 3, n. 5, p. 195-199, 1998.
- SCHNABLE, P.; WARE, D.; FULTON, R. F. et al. The B73 maize genome: Complexity, diversity, and dynamics. **Science**, v. 326, n. 5956, p. 1112-1115, 2009.

- SCHUELER, M.G.; DUNN, J.M.; BIRD C.P. et al. Progressive proximal expansion of the primate X chromosome centromere. **Proc. Nat. Acad. Sci.**, v. 102, n. 30, p. 10563-10568, 2005.
- SMITH, G. P. Evolution of repeated DNA sequences by unequal crossover. **Science**, v. 191, p. 528-535, 1976.
- STRACHAN, T.; WEBB, D.; DOVER, G.A. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*, **The EMBO Journal**, Oxford, v. 4, p. 1701-1708, 1985.
- SUJIWATTANARAT, P.; THAPANA, W.; SRIKULNATH, K. et al. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. **Scientific Reports**, v. 5, n. 1, p. 10315, 2015.
- THE ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v. 408, p. 796-815, 2000.
- TORRES, G. A.; GONG, Z.; IOVENE, M. et al. Organization and Evolution of Subtelomeric Satellite Repeats in the Potato Genome. **G3**, v. 1, p. 85 – 92, 2011.
- VALERI, M. P.; DIAS, G. B.; CAMPOS SILVA KUHN, G. et al. An eutherian intronic sequence gave rise to a major satellite DNA in Platyrrhini. **Biology letters**, v. 14, n. 1, 2018.
- VITTORAZZI, S.E. LOURENÇO, L.B; & RECCO-PIMENTEL, S.M. Long-time evolution and highly dynamic satellite DNA in leptodactylid and hyloidid frogs. **BMC Genet**, v. 15, n. 111, 2014.
- VOLPE, T. A.; KIDNER, C.; HALL, I. M. et al. Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi. **Science**, p. 1833-1837, 2002.
- VONDRAK, T.; ROBLEDDILLO, L. A.; NOVAK, P. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon derived tandem repeats. **The plant journal**, . v. 101, p. 584-500, 2019.
- WILLARD. H.F.; WAYE, J.S. Hierarchical order in chromosome-specific human alpha satellite DNA. **Trends in Genetics**, v. 3, n. 7, p. 192-198, 1987.
- ZHANG, W.; LEE, H.; KOO, D. et al. Epigenetic Modification of Centromeric Chromatin: Hypomethylation of DNA Sequences in the CENH3-Associated Chromatin in *Arabidopsis thaliana* and Maize. **The Plant Cell**, v.20, p. 25-34, 2008.
- ZHONG, C.; MARSHAL, J. B.; TOPP, C. et al. Centromeric retroelements and satellites Interact with maize kinetochore protein CENH3. **The Plant Cell**, v. 14, n. 11, p. 2825-2836.

3. CAPÍTULO 1: LOCALIZAÇÃO GENÔMICA INFLUENCIA NO COMPORTAMENTO EVOLUTIVO E NA ORGANIZAÇÃO DAS FAMÍLIAS DE DNA SATÉLITES EM MILHO

RESUMO

A evolução das sequências de DNA de satélite é controlada por mecanismos de Molecular Drive. O mecanismo de Molecular Drive é bem estudado em várias espécies que abrangem diferentes famílias de sequências repetitivas. No entanto, a relação entre a localização do DNA satélite e a dinâmica da evolução em concerto ainda é pouco estudada. Algumas características da organização do genoma do milho são interessantes para investigar essa relação. Em primeiro lugar, o milho possui uma família de DNA satélite denominada de CentC localizada nas regiões centroméricas que interage com as histonas CENH3 e garantindo a segregação do cromossomo. Em segundo lugar, a outra família de DNA satélite denominada de K180 que é o principal componente dos knobs heterocromáticos, uma região muito visível de cromatina condensada, imersa em eucromatina rica em genes, mas sem uma função bem definida. Uma análise da frequência nucleotídica, da diversidade dessas duas famílias e uma comparação entre elas podem fornecer novos conhecimentos sobre o efeito da localização do DNA satélite na dinâmica do Molecular Drive. Para realizar essas análises, foram extraídos os motivos pertencentes ao CentC e ao K180 do genoma de referência da linhagem B73. A partir dessas análises, observou-se que a localização de motivos de DNA de satélite dentro do genoma do milho pode influenciar seu comportamento evolutivo e na organização espacial. Esses comportamentos são observados devido à maior conservação dos motivos centroméricos em relação aos motivos do knob heterocromático e pela presença de uma disposição bem organizada entre os motivos centroméricos, enquanto o knob não foi observado nenhum tipo de organização. Estes relatos auxiliam evidenciar a importância do contexto genômico para a evolução em concerto em milho.

Palavras-chave: Evolução em Concerto, DNA Satélite, Centrômero, Knob, Milho

ABSTRACT

The evolution of satellite DNA (satDNA) sequences is controlled by Molecular Drive mechanisms. The mechanism of the Molecular Drive is well-studied in several species covering different families of repetitive sequences. However, the relationship between satDNA localization and the concerted evolution dynamics is still little studied. Some features of the maize genome organization are interesting to investigate this relationship. Firstly, a satDNA family named as CentC resides in the centromeric regions, interacting with CenH3 histones and guarantying chromosome segregation. Second, another satDNA family named as K180 is the major component of the heterochromatic knobs, a very conspicuous region of condensed chromatin, immersed into gene-rich euchromatin, but without a well-defined function. An analysis of nucleotide frequency and diversity of these both families and a comparison of them might provide insights about the effect of satDNA localization in the molecular drive dynamics. Motifs belonging to CentC and K180 were data-mined from reference genome of the inbred-line B73. From these analyses it was observed that the location of satellite DNA motifs within the maize genome can influence their evolutionary behavior and spatial organization. These behaviors are observed due to the greater conservation of centromeric motifs in relation to the motifs of the heterochromatic knob. Moreover, it is also observed that there is a well-organized disposition between the centromeric motifs, while the knob was not observed any type of

organization. These reports help to highlight the importance of the genomic context for the concerted evolution in maize.

Keywords: Concerted Evolution, Satellite DNA, Centromere, Knob, Maize

3.1. Introdução

DNAs satélites são constituídos por sequências repetitivas *in tandem* localizadas, principalmente, em regiões heterocromáticas (HEMLEBEN et al., 2007; GARRIDO-RAMOS, 2017). As sequências de DNAs satélites possuem mecanismos de evolução diferente daqueles encontrados nas sequências codantes. Este tipo de evolução, denominado de Evolução em Concerto é dividida em duas variáveis: a taxa de mutação e a taxa de fixação.

Estes passos no caminho evolutivo do DNA satélite gera a homogeneização dos motivos dentro do genoma. Kuhn et al. (2012) trabalhando com drosófilas, demonstraram que o processo de homogeneização leva a uma maior similaridade na sequência de nucleotídeos dos motivos repetitivos dentro de uma determinada região do genoma, mais do que entre as regiões espalhadas ao longo dos cromossomos. As análises de Hall et al. (2005) também verificaram este mesmo padrão de homogeneização dos motivos entre as sequências vizinhas em DNA satélites de plantas.

Estudos reportaram que existem diferentes padrões de comportamento na homogeneização entre DNAs satélites de espécies distintas (STRACHAN; WEBB; DOVER, 1985; DOVER, 1982), entretanto, poucos relatam a diferença desta homogeneização dos motivos entre famílias de DNAs satélites presentes no mesmo genoma que estão em contexto genômico diferente (STRACHAN; WEBB; DOVER, 1985; TORRES et al., 2012). O milho é uma espécie cultivada e modelo para estudos genéticos, sendo um excelente material para este tipo de estudo, pois possui regiões de heterocromatina com características distintas e compostas de DNAs satélites: o knob e o centrômero.

O knob é uma região de heterocromatina bem visível na cromatina que possui replicação tardia e é constituída principalmente por duas famílias de DNAs satélites denominados de K180 e TR-1 (PEACOCK et al., 1981; ANANIEV; PHILLIPS; RINES, 1998). Os knobs heterocromáticos de milho foram descritos praticamente em todos os cromossomos, ocupando posições intersticiais no braço curto e no braço longo; sendo a única exceção o cromossomo 10 que nunca apresenta knob na sua versão normal. Porém, o número e as posições destes knobs são muito variáveis no complemento cromossômico de milho, tanto que estes têm servido como marcadores para caracterizar variedades e suas correlações biogeográficas. (McCLINTOCK et al. 1981; KATO; LAMB; BIRCHLER, 2004; ALBERT et al., 2013; MONDIN et al., 2014; REALINI et al., 2018). Essa variação de número e posição de knobs possuem correlação com a origem de cada linhagem de acordo com a sua posição latitudinal, de altitude e climática (KREINER e WRIGHT; 2018). Entretanto, Lamb et al (2007) observaram que existem uma pequena quantidade de motivos presentes nas regiões onde deveriam existir knobs heterocromáticos.

Os knobs heterocromáticos interrompem regiões de eucromatina ricas em genes, suprimindo o *crossing-over* e afetando o funcionamento normal dos genes que estão nas proximidades. A função dos knobs no genoma ainda não é bem definida, porém existem algumas correlações entre a sua composição com o tempo de florescimento do milho (RAYBURN et al., 1985; TITO, POGGIO E NARANJO, 1991; REALINI et al., 2016; JIAN et al., 2017).

No entanto, diferentemente do knob, o centrômero possui uma função bem definida, está presente em todos os cromossomos de todas as linhagens, variedades e raças e possui baixa quantidade de genes. Esta estrutura tem a função de manter as cromátides irmãs unidas após a duplicação do DNA, característica fundamental para a segregação normal dos cromossomos durante as divisões meióticas e mitóticas (MA et al., 2007). A posição funcional do centrômero está localizada dentro da região de heterocromatina denominada de pericentromérica. A região funcional é diferencialmente destacada das outras regiões do cromossomo devido a presença da histona variante H3, também conhecida como CENP-A em humanos ou CENH3 em vegetais (MCKINLEY; CHEESEMAN, 2016), que em milho interagem diretamente com as sequências da família de DNA satélite CentC e os elementos de retrotransposição CRM para a formação do complexo centromérico (ZHANG et al. 2008). A presença da CENH3 e a formação de um complexo proteico nos motivos de CentC pode atuar como uma pressão de evolução seletiva nestas sequências repetitivas *in tandem* (HALL et al., 2005; ZHONG et al., 2002).

Em alguns organismos os motivos de DNA satélite são altamente organizados na região centromérica. Nos seres humanos, o centrômero é constituído pelas sequências de alfa-satélite, onde a organização dos motivos ocorre de duas formas distintas: High Order Repeats (HOR) e monômeros (RUDD et al., 2004).

São chamados de HOR os arranjos *in tandem* que consistem de múltiplos motivos (WILLARD et al., 1987; SUJIWATTANARAT, 2015) e que estão organizados em grandes arranjos formando uma unidade com 97 a 100% de similaridade. Esta conformação em HOR também é visualizada nas espécies de primatas do velho mundo e nas poucas observações feitas em primatas do novo mundo (SUJIWATTANARAT et al., 2015). Enquanto os motivos denominados de monômeros não possuem nenhum tipo de organização complexa dos seus motivos. Cada uma dessas organizações citadas está localizada em posições diferentes ao longo do centrômero e possui relações específicas com as respectivas funções da região centromérica (RUDD et al., 2004).

Em plantas há poucos estudos sobre a organização das sequências de DNA satélite em HOR. Alguns estudos isolados, como em arroz já foram publicados (LEE et al., 2006). Nesta espécie foi observada uma pequena região no centrômero do cromossomo 8 que possui uma organização em HOR, porém, a grande maioria dos motivos são monômeros e associados com a CENH3 (LEE et al., 2006). No cromossomo 1 não foi localizado nenhum arranjo organizado em HOR (LEE et al., 2006).

Dentro deste contexto, o presente estudo tem como objetivo comparar os padrões de evolução de famílias de DNAs satélites localizadas em dois contextos genômicos distintos, sendo um destes o knob heterocromático que interrompe regiões de eucromatina rica em genes e não possui uma função definida e o centrômero que forma uma heterocromatina em uma região com baixa frequência de genes que por outro lado possui uma função constitutiva bem definida e interage com um conjunto específico de proteínas para o seu funcionamento. Ambas as regiões serão analisadas para o entendimento de sua organização estrutural ao longo do arranjo das sequências repetitivas *in tandem* utilizando a família de DNA satélite mais abundante em cada região (CentC e K180).

3.2. Materiais e Métodos

3.2.1. Materiais

Foi utilizado o genoma montado da linhagem B73 de milho (*Zea mays*) disponível no banco de dados MAIZEGDB (www.maizegdb.org) com o respectivo código de identificação Zm-B73-REFERENCE-GRAMENE-4.0 obtido a partir do sequenciamento de PacBio Single Molecule Real Time (SMRT) com cobertura de 60x e o erro de correção desse mecanismo foi feita pelo sequenciamento por Illumina. A montagem foi realizada com o auxílio de um mapeamento de restrição de genoma inteiro de alta resolução (óptica) obtendo pseudomoléculas com uma montagem próxima de uma estrutura end-to-end. No total foi montado aproximadamente 2,1GB do genoma da linhagem B73 de um total real de 2,4GB (87,5%).

3.2.2. Métodos

3.2.2.1. Data mining dos motivos dos knobs heterocromáticos e da região centromérica

O *data mining* foi realizado por meio da ferramenta *Basic Local Alignment Search Tool* (BLAST) presente no próprio banco de dado MAIZEGDB. Os parâmetros utilizados para esta ferramenta foram *max target sequence* = 500, *e-value* = 1, *word size* = 11 e *match/mismatch scores* = 2, -3. Foi utilizado o código de acesso AF030935.1 e AF078918.1 para busca dos motivos da família K180 e CentC, respectivamente. A partir do BLAST, foi localizada a posição de cada motivo das famílias nos cromossomos do respectivo genótipo. Os contigs que possuíam os motivos de cada família foram selecionados e a seguir processados no *Tandem Repeats Finder* (TRF - BENSON et al., 1999) utilizando a linha de comando padrão.

3.2.2.2. Análises das Sequências

Os motivos de cada família foram separados e organizados de acordo com a sua posição no genoma. Para realizar as análises da dinâmica evolutiva foi necessário fazer o alinhamento múltiplo. Os alinhamentos múltiplos dos motivos das duas famílias de DNA satélites foram realizados pelo algoritmo *Muscle* no programa MEGA 7.0 (KUMAR, STECHER e TAMURA, 2016). O alinhamento de cada arranjo foi processado pelo programa *Jalview* (CLAMP et al., 2004) para gerar a imagem do alinhamento destacando os nucleotídeos polimórficos de cada motivo.

3.2.2.3. Diversidade Nucleotídica

Para realizar as análises de comparação das duas famílias de DNA satélite, nós geramos a frequência nucleotídica para K180 e CentC utilizando o programa *Jalview*. Foi realizado o cálculo da diversidade nucleotídica por meio do modelo de Jukes Cantor (JC) disponível no programa MEGA 7.0. A construção de um mapa *heatmap* foi feita para facilitar a visualização das variações das diversidades presentes nestas duas famílias. O *heatmap* foi construído pela metodologia de percentil em Excel (Microsoft). A região do mapa condicionada a coloração azul indica pequenos valores de diversidade e a região em vermelho indica altos valores de diversidade.

E para contribuir com os estudos da dinâmica evolutiva foi gerada figuras WebLogo (CROOKS et al., 2004). As configurações utilizadas foram 180 nucleotídeos por linha e coloração clássica (NA).

3.2.2.4. Evolução em Concerto – MLTree

As árvores MLTree foram construídas por meio do método estatístico de Máxima Verossimilhança (*Maximum Likelihood*) e com bootstrap de 100 repetições no MEGA 7.0. O modelo selecionado foi Tamura-3-Parameter de acordo com as recomendações do pacote MEGA 7.0.

3.3. Resultados

Foi utilizado para essas análises a linhagem B73 de milho por ser o genoma modelo para estudos genéticos e epigenéticos desta espécie e que contém o genoma sequenciado e montado com alta qualidade em relação as demais linhagens.

No primeiro passo foram coletados os motivos de duas famílias de DNA satélites da linhagem B73 de milho. As famílias coletadas foram CentC e K180 presentes na região centromérica e knob heterocromático, respectivamente, pois estas são as famílias mais abundante presente em cada região. No total foram coletados 5945 motivos de K180 e 1950 motivos de CentC.

Os motivos de ambas as famílias foram organizados e alinhados; e posteriormente foram realizadas análises de variabilidade nucleotídica para verificar a dinâmica mutacional dos motivos nestes

contextos genômicos distintos. O primeiro passo foi a análise do comportamento da frequência nucleotídica destas duas famílias de DNA satélite.

Em uma análise exploratória observa-se frequências maiores (próxima de 100% de identidade) nas posições nucleotídicas de CentC em relação a K180, indicando possivelmente uma maior conservação destes motivos (Figura 1). Para validar esta observação foram calculados os valores de diversidade nucleotídica destas duas famílias de DNAs satélites (Tabela 1). Nesta análise, também, foi observado que os motivos centroméricos são mais conservados do que os motivos do knob. A maior conservação dos motivos centroméricos pode estar relacionada a uma possível pressão seletiva nestes motivos, devido a interação destas sequências de DNA com as proteínas centroméricas (HALL et al., 2005; ZHONG et al., 2002).

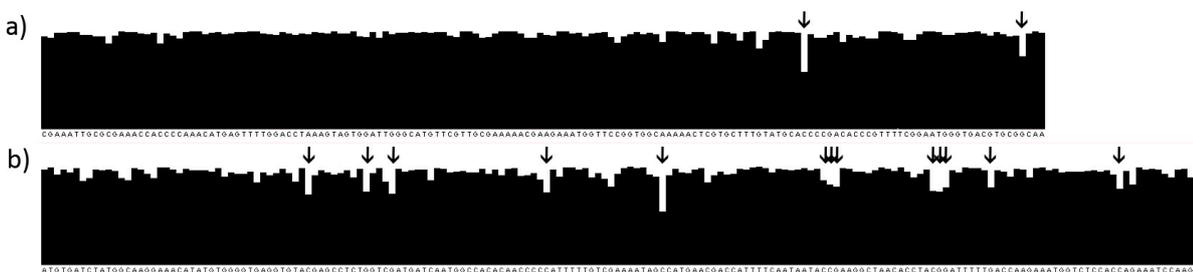


Figura 1. Frequência nucleotídica dos motivos de CentC e K180 na linhagem B73. a) Cada barra vertical representa a frequência nucleotídica de uma posição da família CentC, iniciando na posição 1 e terminando na posição 156; b) Barra vertical representa a frequência nucleotídica de uma posição da família K180, iniciando na posição 1 e terminando na posição 180. As setas indicam posições que possuem valores de frequência nucleotídica inferior a 80%.

3.3.1. Diversidade nucleotídica em arranjos de knobs e centrômeros

Para comparação entre os knobs e os centrômeros considerando seus respectivos contextos genômicos foi calculada a diversidade geral dos nucleotídeos (Figura 2). Os arranjos foram separados de acordo com o braço cromossômico para os knobs e para os motivos de CentC pela região centromérica de cada um dos cromossomos. Após os cálculos dos valores, foi construído um *heatmap* para facilitar a visualização das diferenças entre as diversidades (Figura 2a). De maneira geral, observa-se nitidamente que os arranjos centroméricos são mais conservados em relação aos arranjos dos knobs, assim como já descrito anteriormente.

Tabela 1. Diversidade nucleotídica de CentC e K180 na linhagem B73

DNA Satélite	Diversidade
CentC	0,028 +/- 0,005
K180	0,079 +/- 0,039

O arranjo mais conservado está presente no centrômero do cromossomo 8 que possui 0,007 de diversidade, enquanto o arranjo menos conservado está localizado no knob do braço curto do cromossomo 7 (K7S) que possui 0,192 de diversidade.

Também foi observado que os valores de diversidade dos arranjos de K180 possui uma maior amplitude em relação aos arranjos de CentC (Figura 2b) indicando que os knobs heterocromáticos contêm diferenças expressivas de diversidade, enquanto os centrômeros possuem valores de diversidade mais próximos entre si. Provavelmente, cada knob pode estar sofrendo um tipo distinto de pressão seletiva que pode levar a diferentes efeitos no comportamento evolutivo dos motivos, enquanto, as repetições de CentC devem possuir um efeito evolutivo semelhante entre os centrômeros de milho devido a sua interação com um grupo específico de proteínas. Este resultado reforça, novamente, que existe uma diferença no comportamento evolutivo entre os motivos destas duas famílias de DNAs satélites.

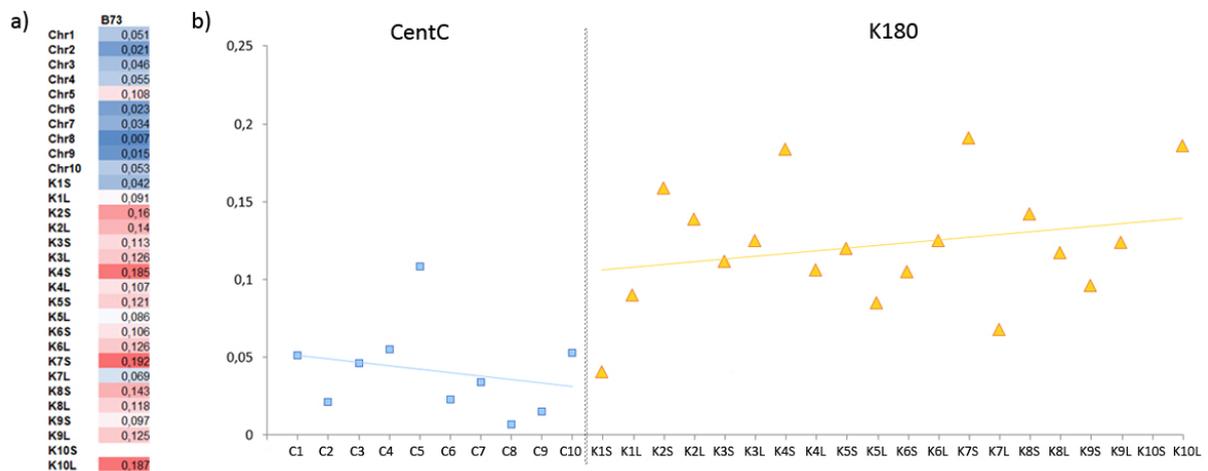


Figura 2. a) HeatMap das diversidades nucleotídicas dos arranjos das famílias K180 e CentC. Os blocos de coloração azul indicam valores menores de diversidade, os blocos em vermelho indicam valores maiores e a coloração branca indicam valores intermediários. Os valores de diversidade nucleotídica são indicados dentro de cada bloco. b) Gráfico de dispersão dos valores de diversidade nucleotídica de cada arranjo da família CentC e K180 na linhagem B73 e suas respectivas linhas de regressão. Os pontos de coloração azul e amarelo indicam os valores de diversidade dos arranjos de CentC e K180, respectivamente. No eixo x os centrômeros estão indicados por C seguido do número do respectivo cromossomo, enquanto os knobs indicados por K, sendo o braço curto indicado por S e o braço longo por L com os números indicando os respectivos cromossomos. No eixo y estão os valores de diversidade nucleotídica de 0 a 0,25 (0% a 25%).

O arranjo centromérico do cromossomo 5 e o knob do braço curto do cromossomo 1 contêm valores de diversidade diferentes do comportamento geral das respectivas famílias de DNA satélites (Figura 2). Os motivos de CentC do centrômero 5 tem um valor de diversidade de 0,108 que é aproximadamente 2x maior do que a média observada para os demais centrômeros e muito semelhante

aos valores de diversidade observados para os knobs. Já os motivos do knob do braço curto do cromossomo 1 possuem uma diversidade igual a 0,042, um valor muito próximo aos observados para os arranjos centroméricos. Esses resultados sugerem que estas duas regiões podem estar sofrendo padrões distintos de modificação dos seus nucleotídeos em relação ao comportamento geral das duas famílias.

3.3.2. Organização dos motivos centroméricos

No alinhamento dos motivos de CentC de cada cromossomo (Figura 3b e 4b, suplementar 1b e suplementar 2b), observamos que existem três grupos de motivos com diferentes comportamentos mutacionais. O grupo 1 é marcado por uma alta conservação dos nucleotídeos ao longo dos arranjos de todos os centrômeros e possui valores de diversidade nucleotídica entre 0% a 1,2%. O grupo 2 apresenta um maior grau de diversidade de nucleotídeos quando comparado ao grupo 1, com valores variando entre 4% e 8%. Este grupo sempre aparece adjacente ou flanqueando os arranjos classificados como grupo 1. O grupo 3 apresenta o maior grau de diversidade com valores variando entre 13% e 17%. Este grupo nunca é observado flanqueando arranjos classificados como grupo 1.

Cada cromossomo possui uma organização destes grupos de uma maneira específica. Os cromossomos foram divididos em três classes para melhor explicar estas organizações.

A classe número 1 é localizada no centrômero dos cromossomos 2, 3, 4, 8 e 9. Foi escolhido como base para o estudo desta classe o cromossomo 4. Neste cromossomo observa-se que o primeiro conjunto de motivos é altamente conservado (Grupo 1) devido a pouca mutação presente no alinhamento (Figura 3b). Estes motivos estão localizados entre as posições 96.446.151 a 96.449.404 no genoma da linhagem B73 (Figura 3b). Logo após, entre as posições 96.457.173 a 96.460.890 e 109.338.530 a 109.459.374, estão presentes os demais motivos menos conservados (Grupo 2 - Figura 3b). Na MLTree (Figura 3a), estes dois conjuntos estão separados em diferentes ramos, evidenciando a clara diferença de homogeneização destes motivos. Os motivos com as características do grupo 3 não estão presentes nestes cromossomos.

Foram obtidos WebLogos para verificar se houve alguma diferença significativa na fixação de nucleotídeos entre estes dois conjuntos (Figura 3c). Nesta análise, verifica-se a diferença de conservação entre os grupos citados anteriormente, entretanto, não houve uma inversão de frequência entre nucleotídeos presentes em uma mesma posição.

O cromossomo 8 possui uma pequena diferença no primeiro grupo. O Grupo 1 possui dois subgrupos de motivos com composições nucleotídicas distintas. O subgrupo 1.A é semelhante aos motivos dos demais cromossomos, enquanto o subgrupo (1.B) contém duas substituições de Citosina para Timina nas posições 11 e 17. Estas substituições nucleotídicas são as causas da separação dos motivos conservados na MLTree (Figura Suplementar 1).

A classe número 2 é constituída pelos centrômeros dos cromossomos 1, 7 e 10. O cromossomo 1 é utilizado para exemplificar o comportamento desta classe. O cromossomo 1 possui uma região com motivos de CentC altamente conservados que se inicia na posição 137.076.097 e termina na posição 137.083.076 (Grupo 1 – Figura 4b). Os motivos que estão localizados após as duas extremidades desta região, são menos conservados (Figura 4b) e pertencem aos Grupos 2 e 3.

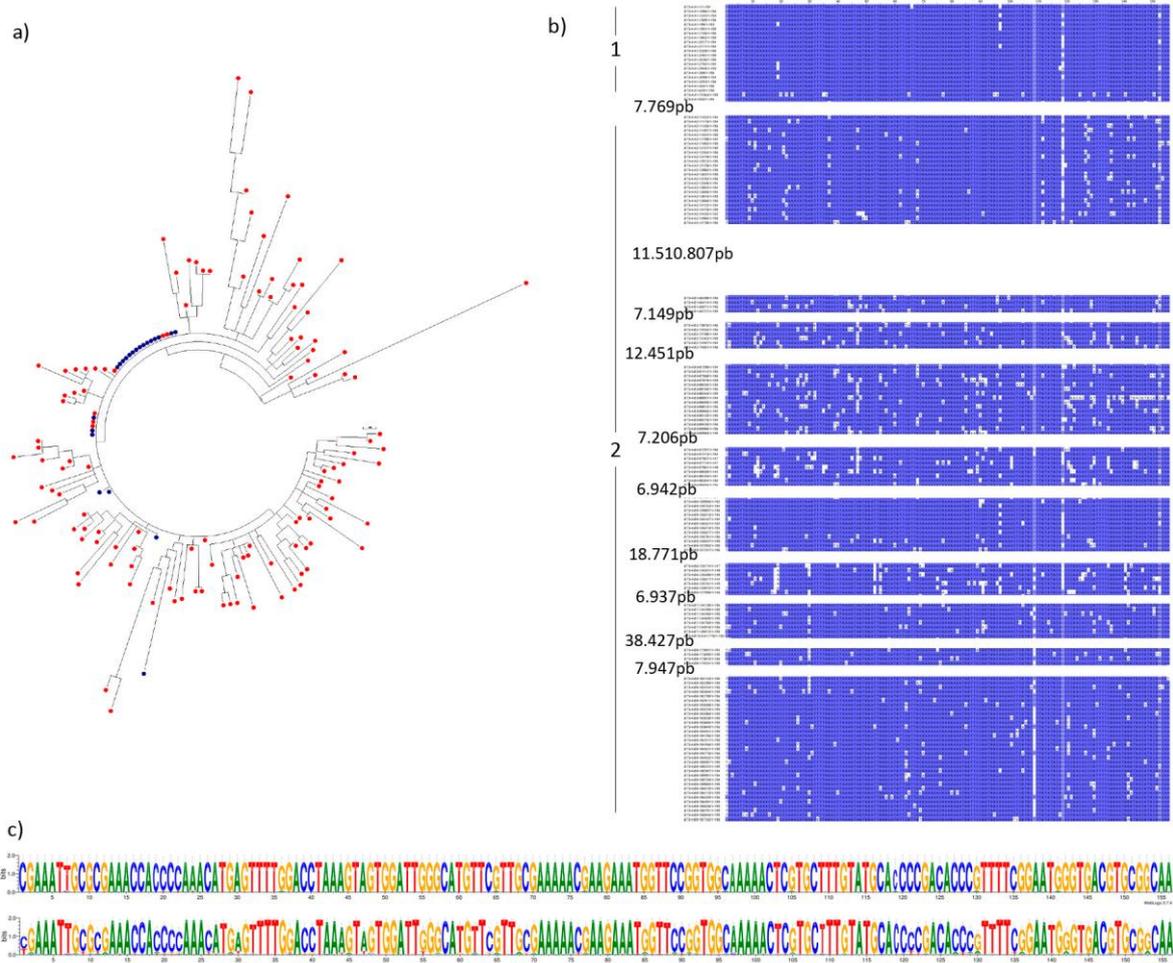


Figura 3. a) MLTree contendo todos os motivos de CentC presentes no cromossomo 4. Os motivos estão representados em cada terminal no formato circular. Os círculos da coloração azul e vermelho representam os motivos presentes nos Grupos 1 e 2, respectivamente. b) Alinhamento de todos os motivos de CentC presentes no cromossomo 4. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. Os arranjos de CentC estão separados em grandes blocos de motivos com a respectiva distância entre cada um. Os blocos de motivos estão marcados com o tipo de Grupo que os representam (1 e 2). c) WebLogos dos motivos presentes em cada Grupo. Os Grupos 1 e 2 são representados pelos WebLogos superior e inferior, respectivamente.

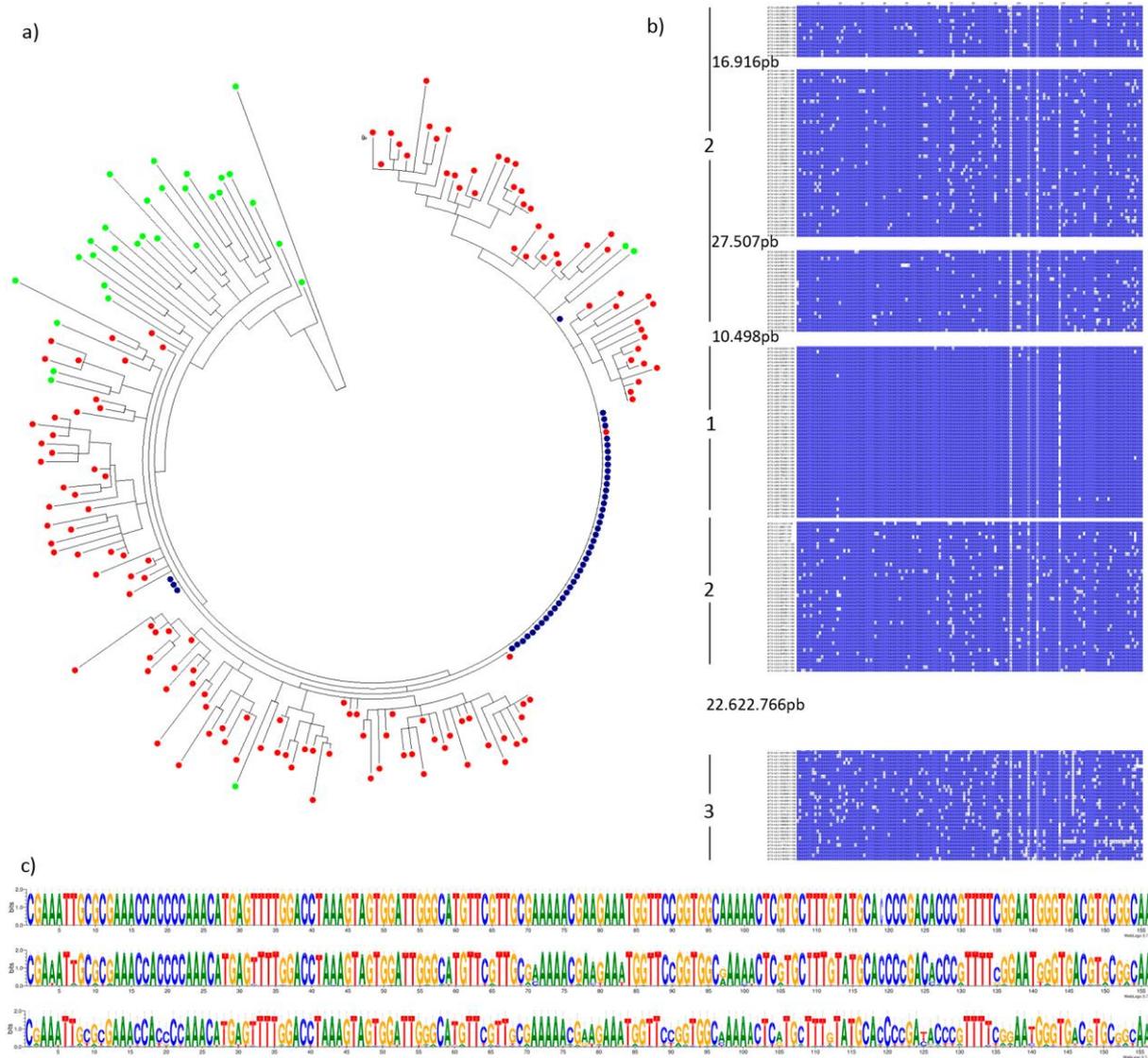


Figura 4. a) MLTree contendo todos os motivos de CentC presentes no cromossomo 1. Os motivos estão representados em cada terminal no formato circular. Os círculos da coloração azul, vermelho e verde representam os motivos presentes nos Grupos 1, 2 e 3, respectivamente. b) Alinhamento de todos os motivos de CentC presentes no cromossomo 1. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. Os arranjos de CentC estão separados em grandes blocos de motivos com a respectiva distância entre cada um. Os blocos de motivos estão marcados com o Grupo de conservação dos motivos (1, 2 ou 3) ao lado esquerdo. c) WebLogos dos motivos presentes em cada Grupo de conservação. Os Grupos 1, 2 e 3 são representados pelos WebLogos superior, intermediário e inferior, respectivamente.

O Grupo 2 é constituído por motivos que estão presentes nas posições 137.002.521 a 137.068.186 e 137.083.575 a 137.116.312. Os motivos do Grupo 3 estão localizados na posição aproximada de 160.000.000 em uma região mais distante das demais repetições de CentC. Na MLTree

observa-se que cada grupo está localizado em ramos diferentes de forma isolada (Figura 4a). Estes resultados evidenciam que os três Grupos estão em níveis de homogeneização diferentes, indicando que o posicionamento dos arranjos de CentC na região centromérica pode afetar o comportamento evolutivo das suas sequências repetitivas.

O centrômero do cromossomo 7 possui uma derivação na composição nucleotídica entre os subgrupos do Grupo 1. Verifica-se que a posição nucleotídica 153 dos motivos do subgrupo 1.C, que estão localizados entre as posições 57.868.109 a 57.886.815, possui na maioria dos motivos a base Guanina em relação ao subgrupo 1.A (Figura Suplementar 2b e 2c). Na MLTree, a maioria dos motivos do subgrupo 1.C estão localizados em um ramo isolado, enquanto, os subgrupos 1.A estão clusterizados em outros ramos com característica bem conservadas (Figura Suplementar 2a).

Os motivos de CentC do centrômero do cromossomo 5 está classificado na classe 3 e possui um comportamento diferente dos demais cromossomos, pois não existem motivos altamente conservados do Grupo 1 no centrômero. Esta característica pode estar associada com as características presentes especificamente neste cromossomo como a modificação da posição funcional do centrômero de uma região rica para uma pobre em motivos de CentC (WOLFGRUBER et al., 2009, SCHNEIDER et al., 2016).

Entretanto, verifica-se a presença dos motivos dos Grupos 2 e 3. Estes dois grupos estão localizados em posições diferentes dentro do genoma, o Grupo 2 aparece entre as posições 108.318.510 a 108.377.382 e o Grupos 3 entre as posições de 104.074.361 a 104.074.516 (Figura 5b). Esta diferença de diversidade é observada na quantidade de nucleotídeos polimórficos presentes no alinhamento dos motivos deste cromossomo (Figura 5b). Na MLTree existe uma clara separação destes dois grupos, resultado que, também, evidencia o efeito da homogeneização em arranjos no cromossomo 5, mesmo com a ausência dos motivos altamente conservados do Grupo 1 (Figura 5a).

3.3.3. Árvore com motivos de todos os cromossomos

Para investigar possíveis assinaturas nucleotídicas que pudessem caracterizar cromossomos específicos, foi construída uma árvore com todos os motivos de CentC (Figura 6). A MLTree resultante desta análise não identificou nenhum tipo de cluster que permitisse a identificação de cromossomos específicos. Entretanto, há agrupamentos que respeitam a divisão dos motivos de CentC em Grupo 1, 2 e 3.

Observa-se que os motivos CentC altamente conservados, classificados como Grupo 1, estão presentes clusterizados no mesmo ramo com todos os cromossomos representados. Este resultado evidencia duas propriedades evolutivas deste grupo: 1) essas repetições são altamente conservadas entre si em relação aos motivos dos outros dois grupos e 2) a homogeneização ocorre dentro e entre os cromossomos simultaneamente. Deste modo, deve haver algum mecanismo que conduz a

homogeneização dos motivos do Grupo 1 para dentro do seu arranjo e depois compartilhando para os arranjos dos demais cromossomos ou estas regiões compartilham de uma mesma composição estrutural protéica para organização do cromossomo que direciona todos os motivos a se comportarem de forma semelhante.

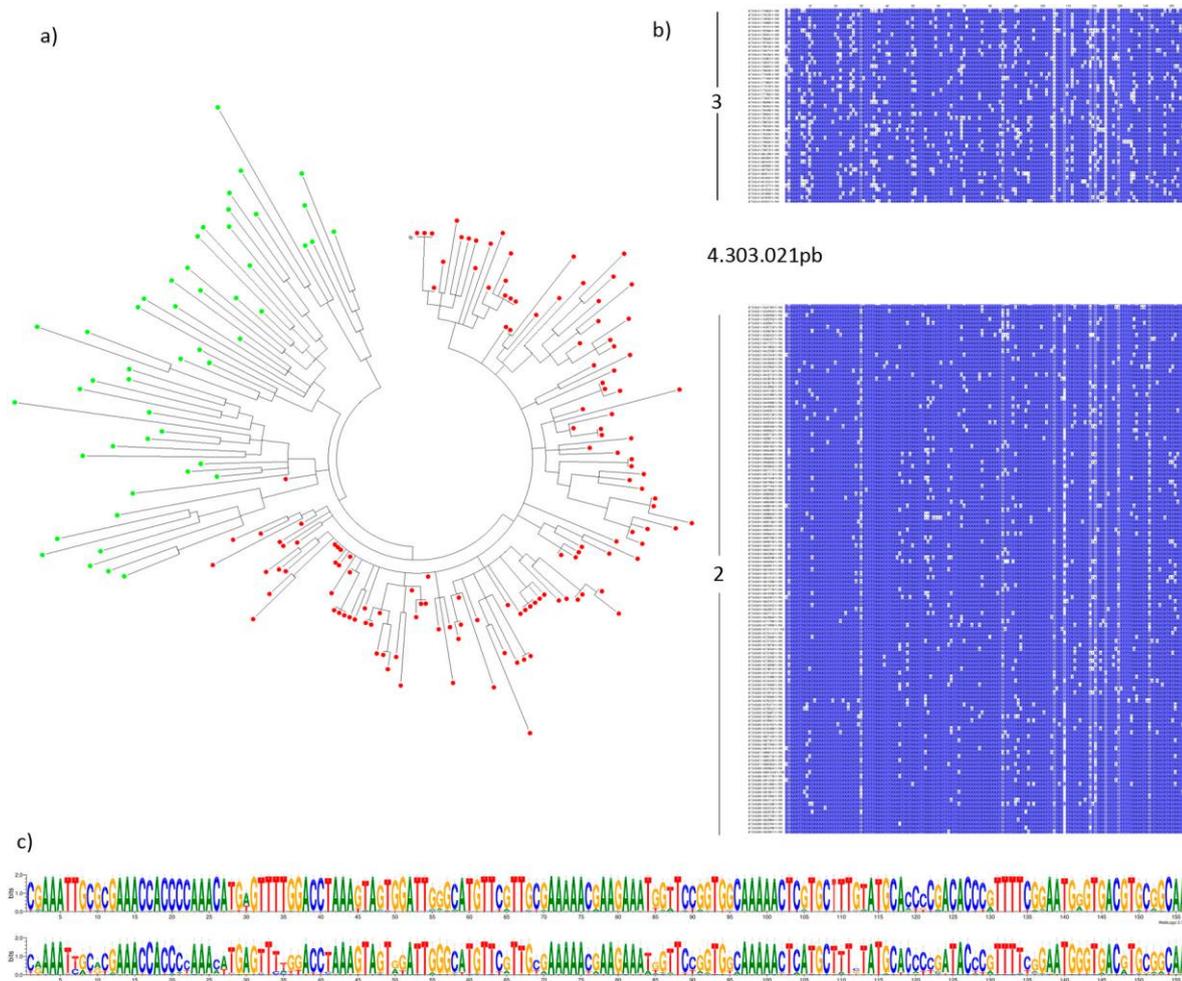


Figura 5. a) MLTree contendo todos os motivos de CentC presentes no cromossomo 5. Os motivos estão representados em cada terminal no formato circular. Os círculos da coloração vermelha e verde representam os motivos presentes nos Grupos 2 e 3, respectivamente. b) Alinhamento de todos os motivos de CentC presentes no cromossomo 5. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. Os arranjos de CentC estão separados em grandes blocos de motivos com a respectiva distância entre cada um. Os blocos de motivos estão marcados com o tipo de Grupo de organização ao lado esquerdo (2 e 3). c) WebLogos dos motivos presentes em cada Grupo de organização. Os Grupos 2 e 3 são representados pelos WebLogos superior e inferior, respectivamente.

Os motivos do Grupo 3 são encontrados clusterizados em um ramo localizado na ponta esquerda da árvore. Os outros ramos são constituídos pelos motivos do Grupo 2 e estão espalhados em toda extensão da MLTree.

Também existem algumas características específicas do Grupo 1 que se organizaram, diferentemente, na MLTree. Os motivos do subgrupo 1.C presentes no centrômero do cromossomo 7 encontram-se em um ramo menor, junto com motivos dos cromossomos 8 e 9, que também possuem motivos com esta característica. Os motivos do subgrupo 1.B presentes no cromossomo 8 também aparecem isolados na MLTree devido aos diferentes nucleotídeos que compõem as posições 11 e 17.

O cromossomo 6 possui uma pequena quantidade de motivos presentes somente em uma região do genoma, o que dificultou a sua caracterização, entretanto, na árvore com todos os cromossomos, podemos observar que existem motivos nos ramos do Grupo 1 e do Grupo 2.

3.3.4. Diversidade nucleotídica dos motivos de K180 entre os braços cromossômicos

Os motivos presentes nos arranjos de knobs não possuem qualquer tipo de alta organização semelhante ao que foi identificado nos centrômeros de B73. Somente dois braços cromossômicos possuem um comportamento semelhante, entretanto, distante do que foi observado para o centrômero. O knob presente no braço longo do cromossomo 1 possui uma pequena região com motivos altamente conservados (Figura suplementar 3) e o braço longo do cromossomo 4 possui motivos que são diferentes dos demais motivos da mesma região (Figura Suplementar 4).

Contudo, foram realizadas análises do comportamento evolutivo para se identificar algum tipo de padrão entre os arranjos dos knobs.

Foi utilizado o WebLogo para verificar se existe alguma diferença nas posições nucleotídicas entre os arranjos dos knobs na linhagem B73. A posição 7 foi a posição que mostrou uma expressiva diferença no passo evolutivo de fixação (Figura 8). Nesta posição, a Timina é o nucleotídeo mais abundante entre os motivos do braço curto do cromossomo 4 (K4S), do braço curto do cromossomo 7 (K7S) e do braço longo do cromossomo 10 (K10L), enquanto, a Citosina é a mais abundante nos motivos presentes nos outros braços. Provavelmente, os processos de Evolução em Concerto destes braços estão sendo diferentemente conduzidos por alguma influência específica destas regiões.

Estes mesmos arranjos (K4S, K7S e K10L) da linhagem B73 possuem os maiores valores de diversidade nucleotídica entre todos os braços cromossômicos (Figura 2a), indicando que existe uma correlação entre a alta variabilidade dos motivos com a diferença entre os passos de fixação da posição 7.

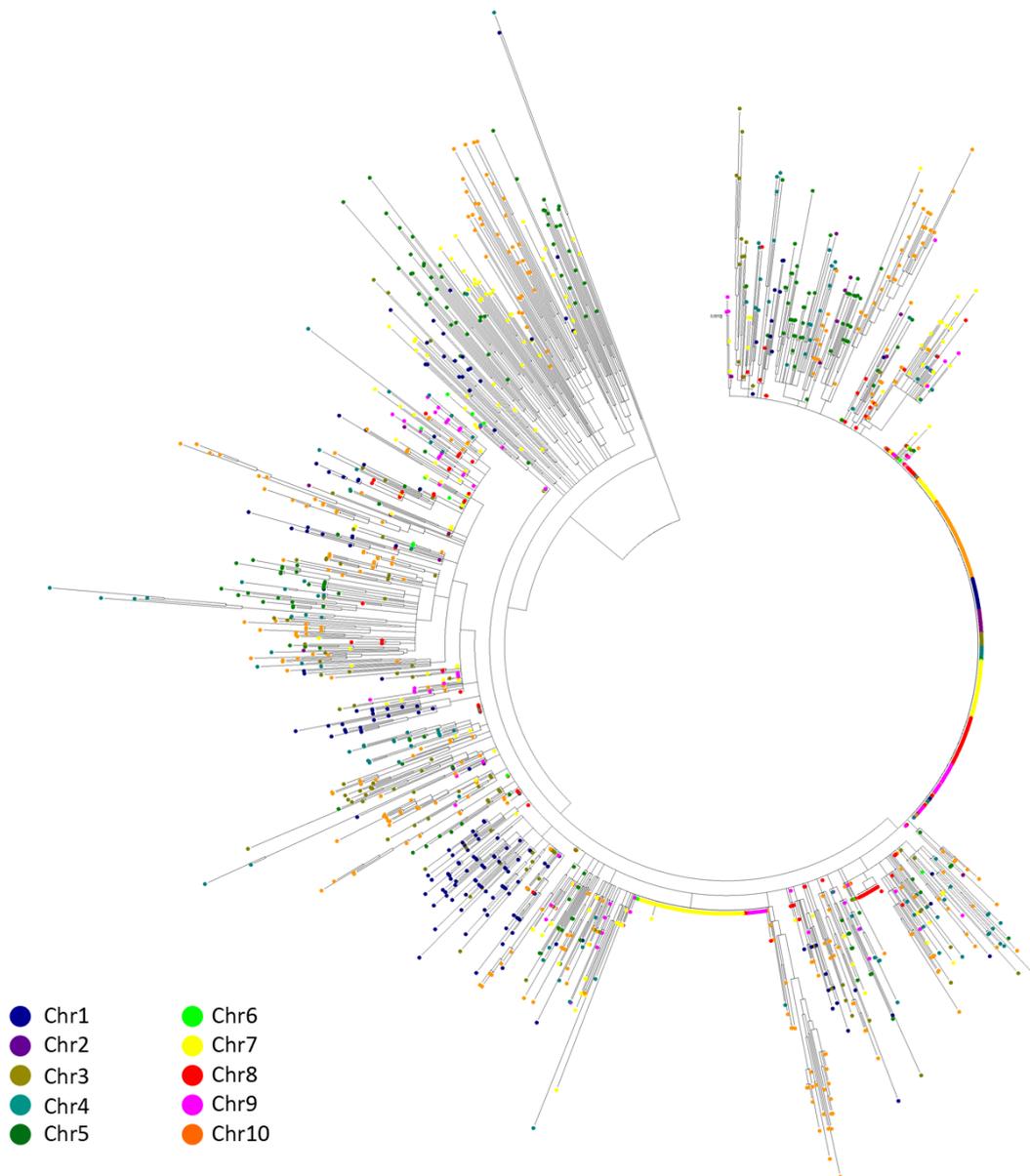


Figura 6. MLTree constituído pelos motivos do centrômero de todos cromossomos da montagem da linhagem B73. Cada terminal no formato circular indica um motivo. Círculos da coloração azul escura, roxo, marrom, verde água, verde escuro, verde claro, amarelo, vermelho, rosa, e laranja representam os motivos dos cromossomos números 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10, respectivamente.

3.3.5. Diferenças de diversidade nucleotídica entre os knobs caracterizados e a região com menor quantidade de motivos de K180

Estudos de FISH com técnicas que aumentam a intensidade do sinal da sonda alvo comprovam que existem motivos de K180 em regiões que ainda não foram caracterizadas como knobs heterocromáticos (LAMB et al., 2007) e que estão presentes em todos os braços cromossômicos de linhagens de milho. As principais diferenças entre estas novas regiões de motivos e com as regiões

caracterizadas são: 1) a quantidade de sequências de DNAs satélites, onde a primeira região contém menor quantidade do que a segunda e 2) a forma heterocromática, pois a segunda possui uma heterocromatina bem formada e visível, enquanto que a primeira região não constitui desta estrutura.

Contudo, foi investigado se existe efeito do tamanho do arranjo do knob heterocromático sobre a diversidade nucleotídica dos motivos. Para isso, os motivos das respectivas regiões foram agrupados em duas classes, sendo a primeira formada pelos motivos de regiões com knobs citologicamente visíveis e o segundo pelos motivos visualizados com técnicas diferenciadas de microscopia ou seja regiões que não formam knobs citologicamente visíveis, mas que apresentam sequências de DNA satélite da família K180 (Figura 7) e suas diversidades nucleotídicas foram calculadas.

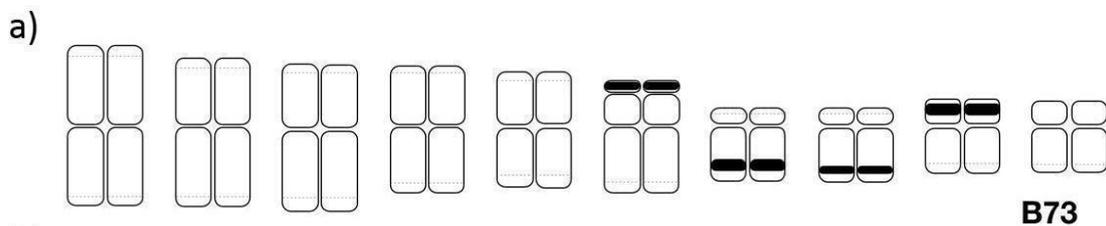


Figura 7. Ideograma da linhagem B73 com as indicações das posições dos knobs em seus respectivos braços cromossômicos. Os blocos pretos no interior de cada cromossomo indicam os knobs caracterizados por Albert et al (2010) e as linhas pontilhadas representam regiões com pequenas quantidades de sequências de K180 observadas por FISH capturados em exposições longas (LAMB et al., 2007).

De uma maneira geral, foi observado que existe uma tendência para a maior variabilidade entre os motivos presentes nos menores arranjos (Tabela 2). Conclui-se, que existe uma pequena diferença entre estas duas regiões contendo motivos de K180 em relação a sua diversidade nucleotídica.

Em nível de arranjo não houve uma correlação entre aumento ou diminuição da diversidade de acordo com o tamanho do knob (Figura 2a). Somente é observada uma tendência de os knobs maiores possuírem menores valores de diversidade nucleotídica na linhagem B73.

Tabela 2. Diversidade nucleotídica da região constituída por uma pequena quantidade de motivos de K180 (OKnob) e a região caracterizada de knobs heterocromáticos (Knob) na linhagem B73.

Condição	Diversidade Nucleotídica
Oknob	0,146 +/- 0,009
Knob	0,123 +/- 0,008

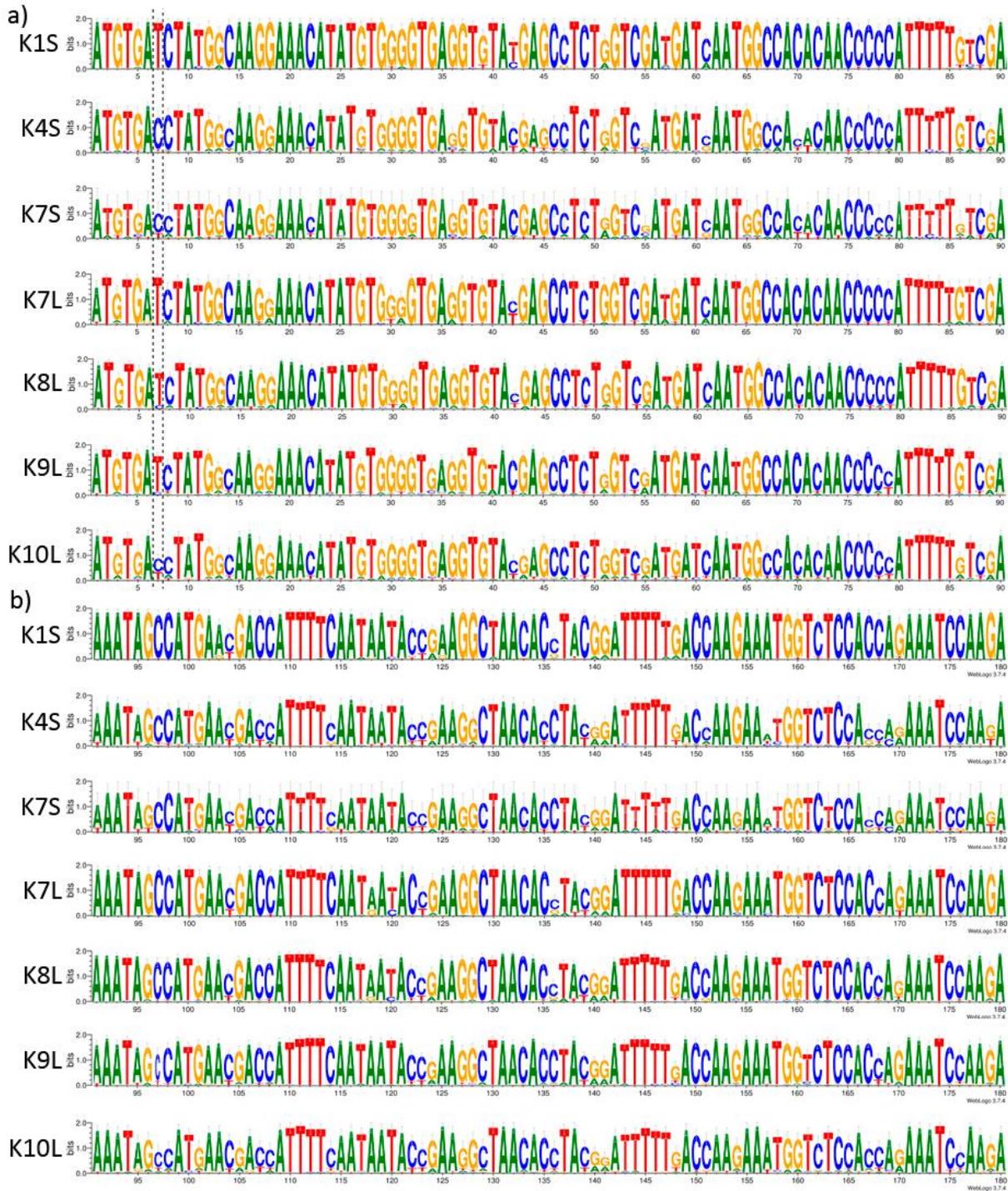


Figura 8. a) WebLogo dos arranjos da família K180 nos knobs na linhagem B73. WebLogo da posição 1 a 79 no arranjo K1S, K4S, K7S, K7L, K8L, K9L e K10L e b) da posição 91 a 180 no arranjo K1S, K4S, K7S, K7L, K8L, K9L e K10L. A posição entre as linhas pontilhadas indica a posição que está em um diferente passo de evolução. Os demais arranjos são apresentados na figura suplementar 5.

As análises WebLogo demonstram que não existem diferenças nos passos evolutivos entre estes knobs, evidenciando que essa pequena variação não é suficiente para alterar, consideravelmente, os passos de fixação.

3.4. Discussão

3.4.1. Organização do Centrômero

As sequências alfa-satélite, responsável pela organização do centrômero humano, é uma família de DNA satélite de constituição, organização e função bem caracterizada, sendo base de estudos para outras espécies. As sequências possuem 171 pb e são organizadas em dois tipos de estruturas denominadas de *High Order Repeats* (HOR) e monômeros (RUDD et al., 2004). Cada uma dessas organizações citadas está localizada em posições diferentes no centrômero e possui relação específica com as respectivas funções desta região do cromossomo (RUDD et al., 2004).

Este tipo de organização é bem caracterizado nas espécies de mamíferos e com poucos estudos realizados em organismos do reino vegetal, como, por exemplo, no arroz. Em arroz foi observado uma pequena região no centrômero do cromossomo 8 que possui uma organização em HOR, porém, a grande maioria dos motivos são organizados como monômeros (LEE et al., 2006).

Em nossas análises com a família de DNA satélite do centrômero de milho denominado de CentC, observamos que, também existem mais do que um tipo de organização entre os motivos centroméricos, entretanto, estes são diferentes daqueles presentes nas sequências alfa-satélites dos seres humanos.

A família de DNA satélite presente nos centrômeros de milho é composta por 3 grupos de motivos, onde cada um possui comportamento organizacional distinto. Estes grupos estão localizados em posições diferentes dentro da organização centromérica não havendo sobreposição entre os seus motivos. Possivelmente, esta disposição organizacional, destes grupos, pode estar associada com as funções centroméricas e a evolução adaptativa entre DNA satélite e as proteínas centroméricas.

Em plantas, sabe-se que a proteína CENP-C indica uma suposta pressão evolutiva e adaptativa na região responsável em realizar a ligação com as sequências centroméricas (TALBERT, BRYSON e HENIKOFF, 2004). Além disso, espécies do domínio Eukaria possuem diversas outras proteínas que atuam diretamente nas sequências centroméricas denominadas do grupo *constitutive centromere-associated network* (CCAN) (MCKINLEY; CHEESEMAN, 2016, COMAI, MAHESHWARI e MARIMUTHU, 2017). Já GENT et al. (2017) encontraram que não existem diferenças de identidade nucleotídica entre os motivos de CentC imunoprecipitado com CENH3 e os motivos presente no restante do genoma, afirmando que não existe uma pressão evolutiva da CENH3 com as sequências centroméricas de milho.

Portanto, é possível inferir que o grupo constituído por motivos altamente conservados estão diretamente relacionado com grupos de proteínas centroméricas, os quais estão envolvidos em uma evolução conjunta adaptativa, exceto as histonas CENH3 que não devem possuir grande influência. Esta evolução permite que os motivos continuem conservados para que as proteínas possam realizar a sua devida ligação e função sobre os centrômeros dos cromossomos.

É importante destacar que, nos cromossomos 7 e 8 que possuem mais do que um grupo de motivos conservados, somente um destes são 100% semelhantes aos motivos dos centrômeros dos demais cromossomos. Os outros grupos possuem um ou dois nucleotídeos que diferenciam do grupo principal e estão localizados em uma região próxima entre si (diferença de 20 pb entre as posições). Talvez essas diferenças nucleotídicas levem a uma perda de capacidade de ligação das proteínas centromérica a estes motivos ao longo da evolução.

3.4.2. Centrômero 5

O único centrômero que não possui esta organização é o número 5. Neste cromossomo, observamos que existe somente o grupo de motivos mais polimórficos e, conseqüentemente, é o centrômero que possui o maior valor de diversidade nucleotídica. Esta característica do cromossomo 5 pode estar associada com as particularidades presentes neste centrômero: 1) reposicionamento do centrômero para uma nova região que ocorreu antes da domesticação do milho, onde foi inserida sequências de retrotransposons centroméricos de milho (CRM) nesta nova região e a eliminação dos motivos CentC na antiga região (WOLFGRUBER et al., 2009, SCHNEIDER et al., 2016), 2) esta posição possui vários haplótipos entre as linhagens de milho evidenciando que não está sobre um grande efeito de domesticação, assim como os cromossomo 7 e 10 (SCHNEIDER et al., 2016) e 3) o centrômero 5 da linhagem B73 possui duas regiões centroméricas diferentes (WOLFGRUBER et al., 2009).

Estes resultados permitem inferir que possivelmente houve eliminação preferencial dos motivos conservados. Embora não seja possível determinar causa e consequência, curiosamente este é o centrômero que sofreu um processo de reposicionamento, onde os retrotransposons centroméricos parecem ter assumido o papel antes pertencente aos motivos CentC.

3.4.3. Homogeneização

A evolução em concerto conduz a homogeneização dos motivos localmente no genoma. Este padrão é visualizado nos motivos centroméricos de arroz (Cent-O), onde o cromossomo 1 possui um conjunto de motivos que são mais semelhantes entre si em relação aos motivos da mesma família presentes no cromossomo 8 (LEE et al., 2006).

Na linhagem B73 de milho não existe uma homogeneização padronizada que diferencia os motivos por cromossomo, entretanto, podemos observar uma homogeneização dentro da região centromérica de cada cromossomo. A MLTree demonstra que cada conjunto de motivos localizados nos Grupos denominados de 1, 2 e 3 estão localizados no mesmo ramo da árvore, indicando que os motivos dentro de cada grupo possuem o mesmo padrão de homogeneização.

3.4.4. Knob

Os motivos presentes nos knobs heterocromáticos não possuem uma alta organização definida na montagem do sequenciamento da linhagem B73 e também não é observado um grau de homogeneização dos motivos. Esta configuração pode estar associada com a sua organização dentro do genoma e a função não bem definida. Também pode-se destacar que os knobs em milho são constituídos diferentemente em cada linhagem de acordo com a sua origem latitudinal, altitudinal e climática (KATO, LAM, BIRCHLER; 2004, ALBERT et al., 2010), demonstrando que o efeito da domesticação do homem fez com que houvesse mudanças no comportamento do knob enquanto que esse mesmo efeito não ocorre na região centromérica desta espécie.

3.4.5. Dinâmica Evolutiva em Diferentes Contextos Genômicos

Os knobs e a região centromérica estão localizados em diferentes contextos genômicos e com distintas funções. Os knobs são regiões citologicamente visíveis, heterocromáticas de replicação tardia, constituídas primariamente por duas famílias de DNAs satélites (K180 e TR1), que interrompem regiões de eucromatina rica em genes e que suprimem a recombinação, sendo altamente polimórficas entre linhagens, cultivares, híbridos e variedades de milho, com função não definida, mas com evidências de correlações com o tempo de florescimento e um possível papel adaptativo. Enquanto centrômero está localizado em uma região com pequena quantidade de genes e associado com funções bem definidas como responsável na união das cromátides irmãs e na segregação dos cromossomos (MA et al., 2007) possuindo uma forte ligação com diversas proteínas centroméricas (MCKINLEY; CHEESEMAN, 2016).

No nosso trabalho, percebemos que a posição da família de DNA satélite em contextos genômicos diferentes pode afetar o comportamento evolutivo das sequências repetitivas na linhagem B73 de milho. Esta afirmação é baseada em dois diferentes comportamentos presentes entre os motivos localizados na região centromérica e do knob heterocromático. 1) Maior conservação da diversidade nucleotídica dos motivos centroméricos em relação ao knob e 2) os valores de diversidade dos arranjos dos knobs são mais diversos, enquanto CentC possui diversidade semelhante entre os seus arranjos.

Referências

- ALBERT, P. S. et al. Diversity of chromosomal karyotypes in maize and its relatives. **Cytogenetic and Genome Research**, v. 129, n. 1–3, p. 6–16, 2010.
- ANANIEV, E. V.; PHILLIPS, R. L.; RINES, H. W. Complex structure of knob DNA on maize chromosome 9. Retrotransposon invasion into heterochromatin. **Genetics**, v. 194, n. 4, p. 2025-2037, 1998.
- BRODIE, R.; ROPER, R.L. JDotter: a Java interface to multiple dotplots generated by dotter. **Bioinformatics**, v.20, n.2, p.279-281, 2004.
- CLAMP, M. et al. The Jalview Java alignment editor. **Bioinformatics**, v.20, n.3, p.426-427, 2004.
- CROOKS, G.E. et al. WebLogo: A sequence logo generator. **Genome Research** v. 14 p. 1188-1190, 2004.
- COEN, E.; STRACHAN, T.; DOVER, G. Dynamics of Concerted Evolution of Ribosomal DNA and Histone Gene Families in the melanogaster Species Subgroup of *Drosophila*. **Journal of Molecular Biology**, v. 158, p. 17-35, 1982
- COMAI, L.; MAHESHWARI, S.; MARIMUTHU, M.P.A. Plant centromeres. **Current Opinion in Plant Biology**, v. 36, p. 158–167, 2017.
- DOVER, G. Molecular drive: a cohesive mode of species evolution. **Nature**. v. 299, p.111-117, 1982.
- GARRIDO-RAMOS, M. A. Satellite DNA: An Evolving Topic. **Genes**. v. 8, n. 9, p. 230, 2017
- GENT, J.I.; WANG, N e SAWE, R.K. Stable centromere positioning in diverse sequence contexts of complex and satellite centromeres of maize and wild relatives. **Genome Biology**, v.18, n. 121, 2017.
- HALL S.E.; LUO, S.; HALL, A.E.; PREUSS, D. Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. **Genetics** v. 170 p. 1913–1927, 2005.
- HEMLEBEN, V.; KOVARIK, A.; TORRES-RUIZ, R.A.; et al. Plant highly repeated satellite DNA: molecular evolution, distribution and use for identification of hybrids. **Systematics and Biodiversity**, Reino Unido, v. 5, n. 3, p. 277-289, 2007.
- KATO, A.; LAMB, J. C.; BIRCHLER, J. A. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, n. 37, p. 13554–13559, 2004.
- KREINER, J. M. e WRIGHT, S. I. A less selfish view of genome size evolution in maize. **PLOS Genetics**, v. 14, n. 5, e1007249, 2018.

- KUHN, G.C.S.; KÜTTLER, H.; MOREIRA-FILHO, O. *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.; STECHET G.; TAMURA K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Mol. Biol. Evol.** Oxford University, v. 33, n. 7, p. 1870-1874, 2016.
- KOMISSAROV, S. A., GAVRILOVA, E.V., DEMIN, S.J. *et al.* Tandemly repeated DNA families in the mouse genome. **BMC Genomics**, v. 12, n. 531, 2011.
- JIAN, Y. Maize (*Zea mays* L.) genome size indicated by 180-bp knob abundance is associated with flowering time. **Scientific Report** v. 7, p. 5954, 2017.
- LEE, H.; NEUMANN, P.; MACAS, J. *et al.* Transcription and Evolutionary Dynamics of the Centromeric Satellite Repeat CentO in Rice. *Molecular Biology and Evolution*, n. 12, v. 23, p. 2505 - 2520, 2006.
- LAMB, J.C. Distinct chromosomal distributions of highly repetitive sequence in maize. **Chromosome Research**, v. 15, p. 33-49, 2007.
- MA, J; WING, R. A.; BENNETZEN, J. L. *et al.* Plant centromere organization: a dynamic structure with conserved functions. **Trends in Genetics**, v. 23; n. 3, p. 134-139, 2007.
- McKINLEY, K.L.; CHEESEMAN, I.M. The molecular basis for centromere identity and function. **Nature Reviews Molecular Cell Biology**. v. 17, n. 1, p. 16-29, 2016.
- PEACOCK, W. J. *et al.* Highly repeated DNA sequence limited to knob heterochromatin in maize. **Proceedings of the National Academy of Sciences of the United States of America**, v. 78, n. 7, p. 4490-4, 1981.
- RAYBURN, A.L; PRICE, J. H.; SMITH, J. D. *et al.* C-BAND Heterochromatin and DNA Content in *Zea mays*. **American Journal of Botany**, v. 72, n. 10, p. 1610-1617, 1985.
- REALINI, M.F.; POGGIO, L.; CAMARA HERNANDEZ, J. *et al.* Intra-specific variation in genome size in maize: cytological and phenotypic correlates. **AoB PLANTS**, v. 8, 2016
- REALINI, M.F.; POGGIO, L.; CAMARA HERNANDEZ, J. *et al.* Exploring karyotype diversity of Argentinian Guaraní maize landraces: Relationship among South American maize. **PLoS ONE**. v. 13, n. 6, p. e0198398, 2018.
- RUDD, M.K.; WILLARD, H.F. Analysis of the centromeric regions of the human genome assembly. **Trends in genetics**, v. 20, n. 11, p. 529-33, 2004.
- SCHNEIDER, K. L.; XIE, Z. WOLFBRUGER, T. K. *et al.* Inbreeding drives maize centromere evolution. **PNAS**, v. 113, n. 8, p. E987 – E996, 2016.
- STRACHAN, T.; WEBB, D.; DOVER, G.A. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*, **The EMBO Journal**, Oxford, v. 4, p. 1701-1708, 1985.

SUJIWATTANARAT, P.; THAPANA, W.; SRIKULNATH, K. et al. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. **Scientific Reports**, v. 5, n. 1, p. 10315, 2015.

TALBERT, P., BRYSON, T. e HENIKOFF, S. Adaptive evolution of centromere proteins in plants and animals. **Journal of biology**, v. 3, n. 18, 2004

TITO, C.M.; POGGIO, L. e NARANJO, C. A. Cytogenetics studies in the genus *Zea* 3. DNA content and heterochromatin in species and hybrids. **Theo. Appl. Genet.**, v. 83, p. 58-64, 1991.

TORRES, G. A.; GONG, Z.; IOVENE, M. et al. Organization and Evolution of Subtelomeric Satellite Repeats in the Potato Genome. **G3**, v. 1, p. 85 – 92, 2011.

WILLARD, H.F.; WAYE, J.S. Hierarchical order in chromosome-specific human alpha satellite DNA. **Trends in Genetics**, v. 3, n. 7, p. 192-198, 1987.

WOLFGRUBER, T. K.; SHARMA, A., SCHNEIDER K. L. Maize Centromere Structure and Evolution: Sequence Analysis of Centromeres 2 and 5 Reveals Dynamic Loci Shaped Primarily by Retrotransposons. **PLOS Genetics**, v. 5, n. 11, 2009

ZHANG, W.; LEE, H.; KOO, D. et al. Epigenetic Modification of Centromeric Chromatin: Hypomethylation of DNA Sequences in the CENH3-Associated Chromatin in *Arabidopsis thaliana* and Maize. **The Plant Cell**, v. 20, p. 25-34, 2008.

ZHONG, C.; MARSHALL, J. B.; TOPP, C. et al. Centromeric retroelements and satellites Interact with maize kinetochore protein CENH3. **The Plant Cell**, v. 14, n. 11, p. 2825-2836, 2002.

Figuras Suplementares

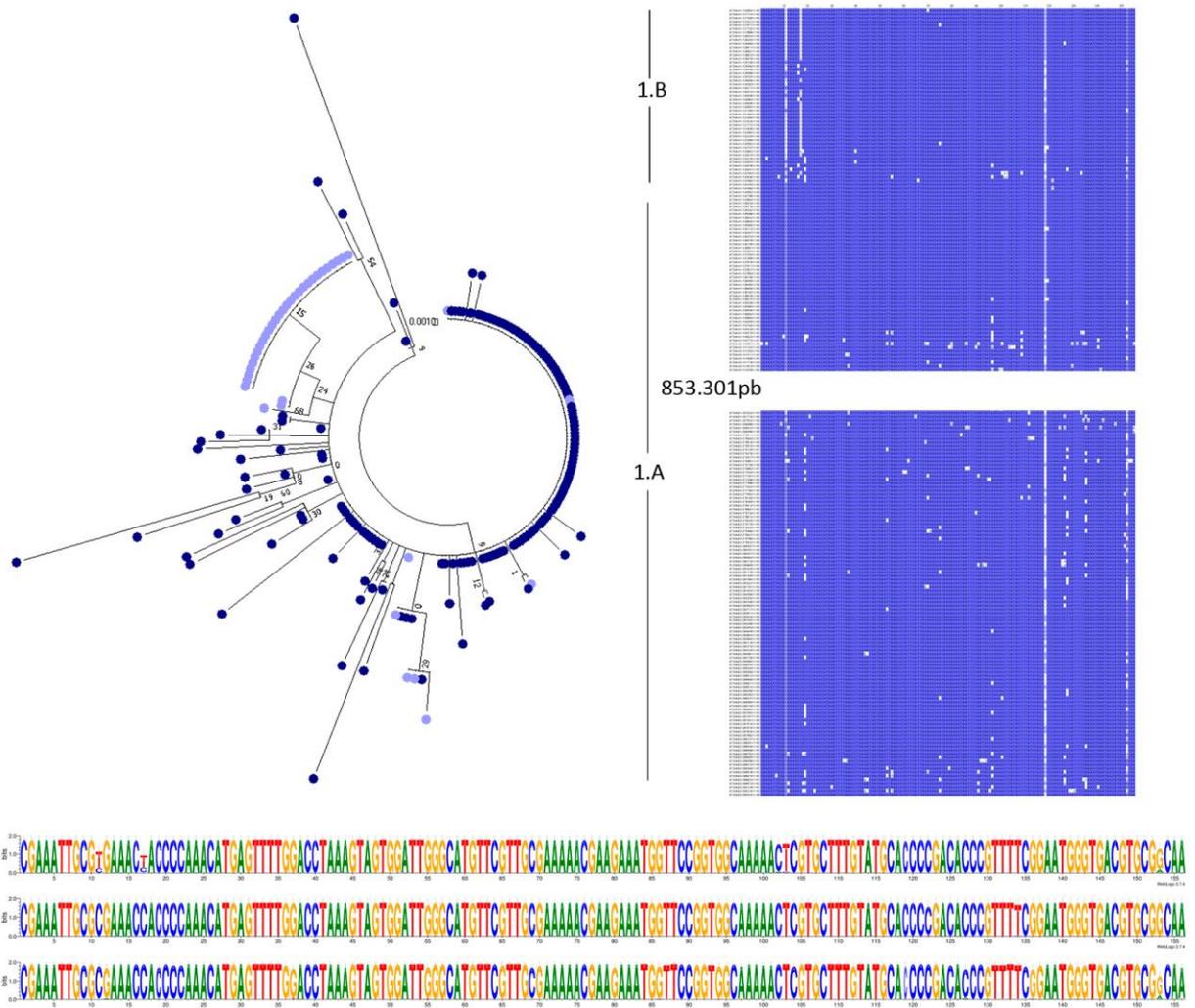


Figura Suplementar 1. a) MLTree contendo todos os motivos de CentC presentes no cromossomo 8. Os motivos estão representados em cada terminal no formato circular. Os círculos da coloração azul escuro e azul claro representam os motivos presentes nos Grupos 1.A e 1.B, respectivamente. b) Alinhamento de todos os motivos de CentC presentes no cromossomo 8. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. Os arranjos de CentC estão separados em grandes blocos de motivos com a respectiva distância entre cada um. Os blocos de motivos estão marcados com o tipo de Grupo de organização ao lado esquerdo (1.A e 1.B). c) WebLogos dos motivos presentes em cada Grupo de organização. Os Grupos 1.A e 1.B são representados pelos WebLogos superior e inferior, respectivamente.

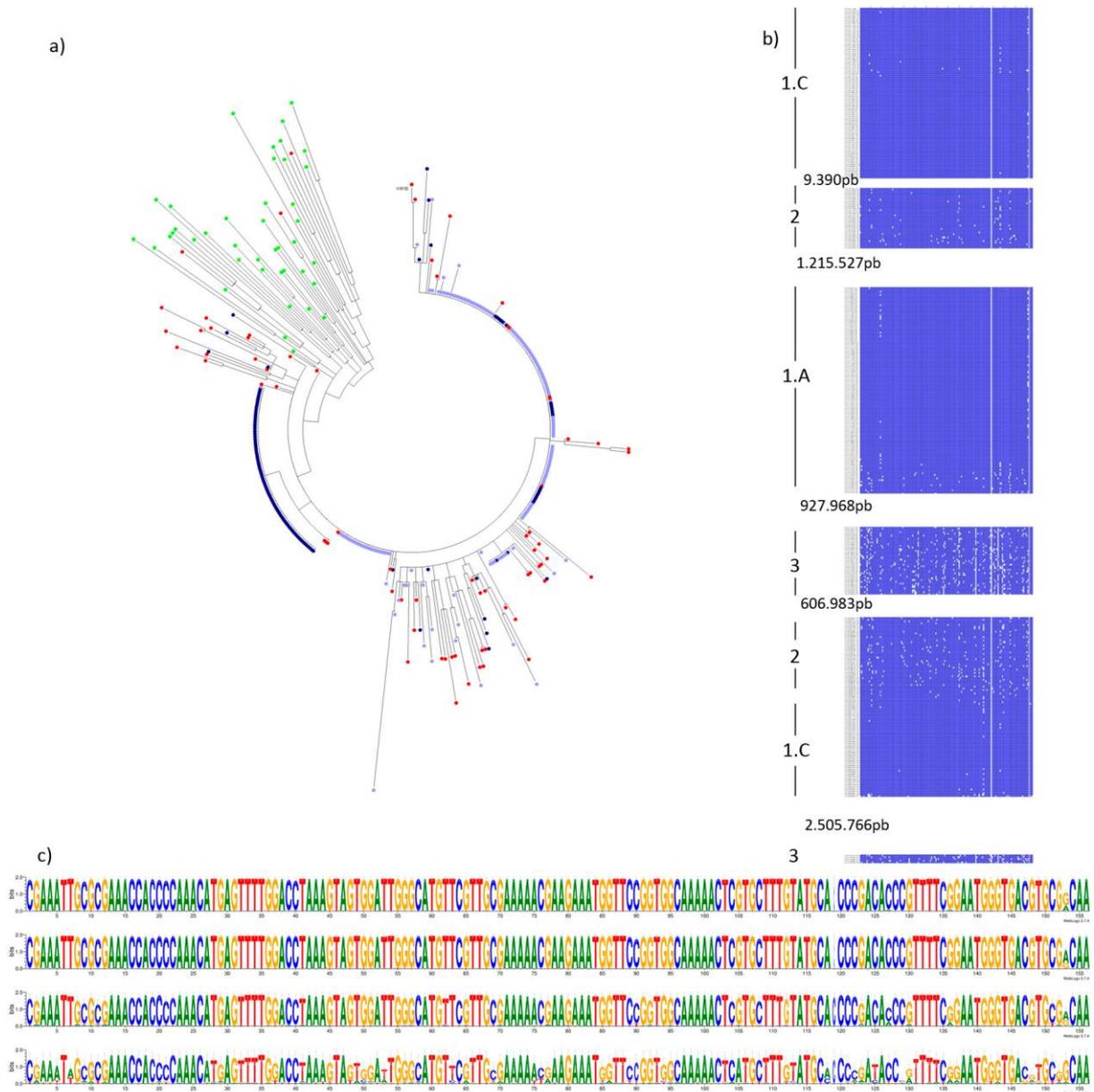


Figura Suplementar 2. a) MLTree contendo todos os motivos de CentC presentes no cromossomo 7. Os motivos estão representados em cada terminal no formato circular. Os conjuntos A e C do Grupo 1 possuem a mesma constituição e são similares aos motivos do Grupo 1 dos outros cromossomos e do conjunto 1.A do cromossomo 7, deste modo, denominados de 1.A. O conjunto B não é similar a nenhum conjunto de motivos e denominamos de 1.C. Os círculos da coloração azul escuro, azul claro, vermelho e verde representam os motivos presentes nos Grupos 1.A, 1.C, 2 e 3, respectivamente. b) Alinhamento de todos os motivos de CentC presentes no cromossomo 7. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. Os arranjos de CentC estão separados em grandes blocos de motivos com a respectiva distância entre cada um. Os blocos de motivos estão marcados com o tipo de Grupo de organização ao lado esquerdo (1.A, 1.C, 2 ou 3). c) WebLogos dos motivos presentes em cada Grupo de organização. Os Grupos 1.A, 1.C, 2 e 3 são representados pelos WebLogos superior para o inferior, respectivamente.

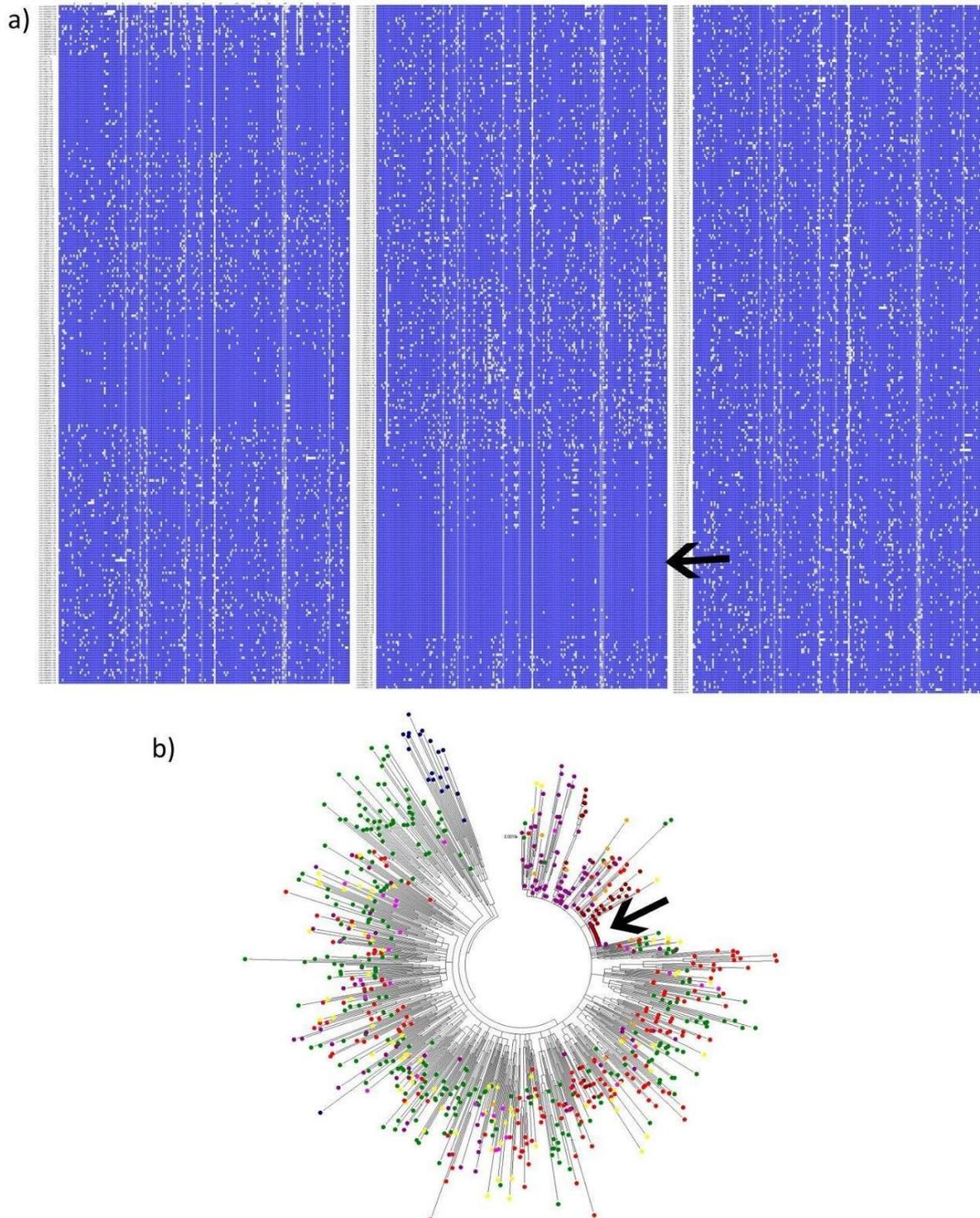


Figura Suplementar 3. a) Alinhamento de todos os motivos de K180 presentes no braço longo de cromossomo 1. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. b) MLTree de todos os motivos presentes no braço longo do cromossomo 1. Cada terminal no formato circular representa um motivo e as respectivas colorações indicam o contig onde está inserido cada motivo dentro do genoma: Azul (B73V4_ctg2725), Roxo (B73V4_ctg2848), Verde (B73V4_ctg1098), Vinho (B73V4_ctg949), Rosa (B73V4_ctg1013), Vermelho (B73V4_ctg764), Laranja (B73V4_ctg2371), Amarelo (B73V4_ctg3065). As setas apontam a região que contém os motivos altamente conservados.

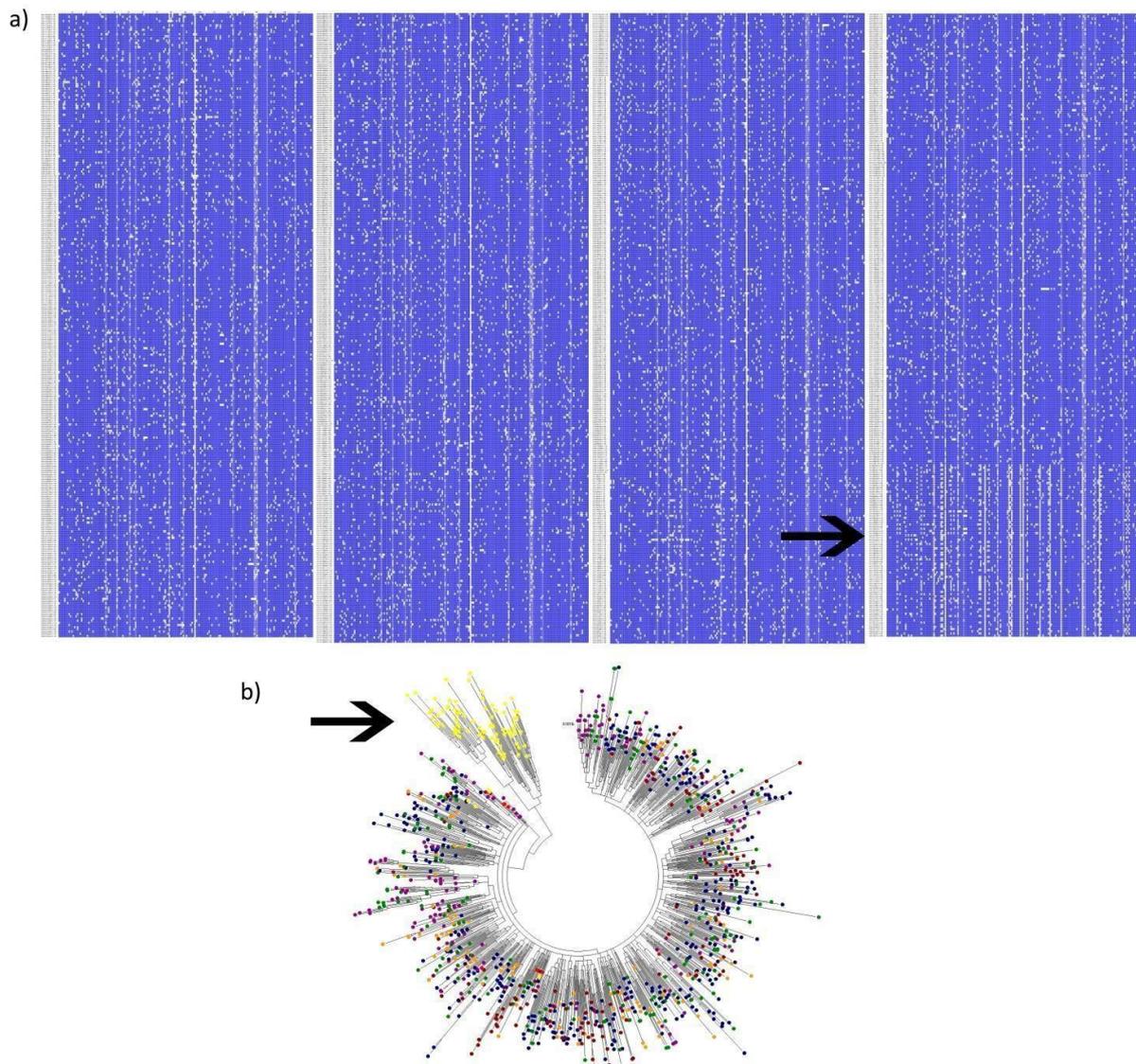


Figura Suplementar 4. a) Alinhamento de todos os motivos de K180 presentes no braço longo de cromossomo 4. Os motivos estão separados em cada linha na horizontal. As posições nucleotídica que estão destacadas em azul são as posições semelhantes a sequência consenso, enquanto as posições em branco possuem uma substituição de nucleotídeo em relação à sequência consenso. b) MLTree de todos os motivos presentes no braço longo do cromossomo 4. Cada terminal no formato circular representa um motivo e as respectivas colorações indicam o contig onde está inserido cada motivo dentro do genoma: Azul (B73V4_ctg885), Roxo (B73V4_ctg3041), Verde (B73V4_ctg1314), Vinho (B73V4_ctg2882), Vermelho (B73V4_ctg1941), Laranja (B73V4_ctg1841) e Amarelo (B73V4_ctg2618 e B73V4_ctg2789). As setas apontam a região que contém os motivos diferentes dos demais.



Figura Suplementar 5. WebLogo dos arranjos da família K180 nos knobs da linhagem B73. a) K1S, b) K1L, c) K2S, d) K2L, e) K3S, f) K3L, g) K4S, h) K4L, i) K5S, j) K5L, l) K6S, m) K6L, n) K7S, o) K7L, p) K8S, q) K8L, r) K9S, s) K9L e t) K10L.

4. CAPÍTULO 2: NOVA FAMÍLIA DO KNOB HETEROCROMÁTICO POSSUI UMA ASSINATURA EVOLUTIVA ENTRE AS ESPÉCIES PRÓXIMAS DE MILHO

RESUMO

Os knobs são blocos heterocromáticos descritos em 23 posições diferentes de todos os 10 cromossomos de milho. Esses knobs variam em número, tamanho e posição entre variedades, linhagens e raças de milho. Devido a essa variabilidade ser distribuída em um padrão não aleatório para altitude e latitude, alguns efeitos genéticos têm sido atribuídos aos knobs no milho. Dois DNAs satélites foram descritos como os principais componentes dos knobs (TR-1 e K180). A família K180 é a maior família de DNA satélite presente em knobs heterocromáticos, enquanto a família TR-1 é menos frequente e foi encontrada em alguns knobs específicos de cromossomo. Desde sua descoberta, a família K180 revelou alguma variabilidade nucleotídica dentro dos motivos, no entanto, uma análise profunda dessa variabilidade nunca foi feita. Para melhor compreender o comportamento da família de DNA satélite de K180 foi realizado um *data-mining* de todos os motivos no genoma sequenciado da linhagem B73 no banco de dados do *EnsemblPlants*. A partir das análises de variabilidade de nucleotídeos realizada por *p-distance*, *sliding-window* e *DotPlot* destes motivos coletados foi encontrada uma nova sub-família de K180, cuja possui um alto nível de variação na última metade do motivo. Essa sub-família foi denominada de K180_B. Os motivos de K180_B também foram localizados na linhagem Mo17 de *Zea mays*, nas espécies de *Zea mays subsp mexicana*, *Zea diploperennis*, *Zea luxurians* e *Tripsacum dactyloides* e nenhuma sequência foi localizada em *Coix aquatica* e *Coix lacryma-jobi*. Em *T. dactyloides* possui um grupo de motivos que contém metade da sequência com um padrão similar as sequências de K180 e uma outra metade similar a K180_B. Esta característica pode ser uma evidência de que a origem da K180_B pode ter sido iniciada nesta espécie e diversificado durante a evolução. A presença desta sub-família entre várias espécies evidencia um papel conservador desses motivos durante a evolução.

Palavras-chave: DNA Satélite, Knob, K180, Milho

ABSTRACT

Knobs are heterochromatic blocks described in 23 different positions of all 10 maize chromosomes. These heterochromatic knobs vary in number, size, and position among maize landraces, varieties, and inbred lines. Due to this variability being distributed in a non-random pattern for altitude and latitude, some genetic effects have been attributed to knobs in maize. Two main satellite DNAs have been described as the major component of the knobs. The K180 family is the largest satDNA family present in heterochromatic knobs, while the TR-1 family is less frequent and has been found in a few specific chromosome knobs. Since its discovery, the K180 family revealed some nucleotide variability inside the motifs. However, a deep analysis of this variability has not been done. The repetitive sequences were data-mined from the B73 inbred line genome from the EnsemblPlants genomic-centric portal. Nucleotide variability analysis performed by *p-distance*, *window-sliding* and *DotPlot* revealed a high level of variation in the last half part of the K180 on a set of motifs and was considered as a new sub-family. This subfamily was called K180_B. The K180_B was also located in the Mo17 inbred lines of *Zea mays*, in the species of *Zea mays subsp mexicana*, *Zea diploperennis*, *Zea luxurians* and *Tripsacum dactyloides* and no sequence were located in *Coix aquatica* and *Coix lacryma-*

jobi. In *T. dactyloides*, a group of motifs that have half of the sequence with a pattern similar to the K180 sequences and another half similar to K180_B were observed. This characteristic may be evidence that the origin of the K180_B may have been initiated in this species and diversified during evolution. The presence of a new sub-family among several species evidences a conservative role of these motifs during evolution.

Keywords: Satellite DNA, Knob, K180, Maize

4.1. Introdução

DNAs satélites são sequências nucleotídicas não codantes caracterizadas pelas repetições de suas cópias *in tandem* e localizadas em grandes blocos de aproximadamente 10Mb no genoma (HEMLEBEN et al., 2007; GARRIDO-RAMOS, 2017). Estas sequências são formadas por motivos de 120 a 740 pb e são, principalmente, encontradas na heterocromatina constitutiva em regiões centroméricas, pericentroméricas e sub-teloméricas (HEMLEBEN et al., 2007, MEHROTRA e GOYAL, 2014). Atualmente, novas ferramentas de bioinformática estão localizando pequenas famílias de DNAs satélites, de menor representatividade no genoma, em regiões não heterocromáticas (RUIZ-RUANO et al., 2016; VONDRAK et al., 2019).

Diversas espécies possuem diferentes tipos de famílias de DNAs satélites que podem estar presentes no mesmo genoma (KULIKOVA et al., 2004, KAWABE; NASUDA, 2005, MEHROTRA; GOYAL, 2014, RUIZ-RUANO et al., 2016; VONDRAK et al., 2019). A distribuição de cada família dentro do genoma é variável, podendo ser encontrado somente em um cromossomo, em uma parte do conjunto cromossômico ou em todos os cromossomos (KATO; LAMB; BIRCHLER, 2004; ALBERT, P. S., 2010; MONDIN, M. et al., 2014; RUIZ-RUANO et al., 2016; VONDRAK et al., 2019).

As funções que são associadas aos DNAs satélites são efeitos indiretos destas sequências dentro do funcionamento do genoma e dos cromossomos, como, por exemplo, as funções centroméricas, do telômero e a formação da heterocromatina (HENIKOFF, AHMAD e MALIK, 2001; VOLPE et al., 2002; HUSINGA, BROWER-TOLAND e ELGIN, 2006; SUJIWATTANARAT et al., 2015; McKINLEY e CHESSEMAN, 2016).

Estas sequências possuem uma alta taxa de mutação de nucleotídeos comparada às sequências gênicas codantes de cópias únicas. A forma como as mutações ocorrem e o mecanismo evolutivo envolvido na diversificação dos DNAs satélites é distinto da teoria neo-Darwiniana e das Leis Mendelianas e é conhecido como Evolução em Concerto (DOVER, 1982). Os motivos deste DNA repetitivo *in tandem* sofrem frequentes mutações em uma única sequência e, ao longo de ciclos de divisões celulares, as mutações são propagadas para outras regiões de um mesmo ou de diferentes cromossomos (STRACHAN; WEBB; DOVER, 1985). A mutação é transmitida, primeiramente, para as sequências localizadas em um mesmo arranjo do genoma para que depois possa ocorrer à propagação desta mutação para posições mais distantes. Tal homogeneização são medidas pelos passos de Strachan

(STRACHAN; WEBB; DOVER, 1985). Os mecanismos gerais mais aceitos do espalhamento destas variantes ao longo dos arranjos repetitivos são o *crossing-over* desigual e a conversão gênica (DOVER, 1982).

O milho (*Zea mays*) é um importante modelo para pesquisas fundamentais de citogenética e epigenética. O genoma deste organismo possui ~2,3 Gb e é composto por, aproximadamente, 32.000 genes, sendo que mais de 85% do total do genoma é composto por sequências repetitivas (SCHNABLE et al., 2009). O gênero *Zea* é constituído por diferentes espécies, entre elas: *Z. mays*, *Z. mays spp. parviglumis*, *Z. Mays spp mexicana*, *Z. mays spp. huehuetenangensis*, *Z. luxurians*, *Z. diploperennis* e *Z. perennis* em que possuem alta similaridade genética (DOEBLEY, 1990; BUCKER IV e HOLTSFORD, 1996). Este gênero está presente na subtribo *Maydeae* que contém também os gêneros *Coix*, *Polytoca* e *Tripsacum*. Entre estes gêneros, o *Tripsacum* forma um grupo monofilético com *Zea*, enquanto *Coix*, que já foi vista como um gênero próximo de *Zea*, não forma um grupo monofilético e está mais próximo das espécies do gênero *Sorghum* (BUCKER IV e HOLTSFORD, 1996).

No milho há 4 famílias distintas de DNAs satélites denominados de CentC, Cent4, 180bp-knob (K180) e TR-1, onde cada família está localizada em diferentes regiões do genoma (PEACOCK et al., 1981; ANANIEV; PHILLIPS; RINES, 1998b; KATO; LAMB; BIRCHLER, 2004; ALBERT, P. S., 2010; MONDIN, M. et al., 2014). A família mais abundante é a K180 que está presente nas regiões heterocromáticas denominadas de knobs. Estas regiões foram descritas primeiramente em cromossomos na fase de paquíteno da meiose e posteriormente na forma de bandas-C em metáfases mitóticas através desta metodologia específica de coloração (AGUIAR-PERECIN et al., 2000). *Coix*, *Tripsacum* e outras espécies do gênero *Zea* também possuem knobs heterocromáticos em seus genomas, onde os motivos da família de DNA Satélite K180 constituem somente os knobs dos dois últimos gêneros (DENNIS e PEACOCK, 1984, ALBERT et al., 2010).

Neste estudo, descrevemos uma nova subfamília variante de K180 presente no genoma de duas linhagens de *Zea mays* (B73 e Mo17) e avaliamos a sua dinâmica evolutiva das espécies ancestrais do milho que, também, são constituídos por knob heterocromático.

4.2. Materiais e Métodos

4.2.1. Materiais

Foram utilizados genomas montados e sequenciados de milho e de espécies relacionadas. Os genomas montados utilizados no trabalho foram duas linhagens modelos de *Zea mays* – B73 (Zm-B73-REFERENCE-GRAMENE-4.0) e Mo17 (Zm-Mo17-REFERENCE-YAN-1.0) – e *Z. mays spp mexicana* (Zx-PI566673-REFERENCE-YAN-1.0) coletados no banco de dados MAIZEGDB (www.maizegdb.org). O genoma de referência *Zm-B73-REFERENCE-GRAMENE-4.0* obtido a partir do sequenciamento de PacBio Single Molecule Real Time (SMRT) com cobertura de 60x e o erro de

correção desse mecanismo foi feita pelo sequenciamento por Illumina. A montagem foi realizada com o auxílio de um mapeamento de restrição de genoma inteiro de alta resolução (óptica) obtendo pseudomoléculas com uma montagem próxima de uma estrutura end-to-end. No total foi montado aproximadamente 2,1GB do genoma da linhagem B73 de um total real de 2,4GB (87,5%). Os demais genomas de referência também foram montados a partir do sequenciamento de PacBio Single Molecule Real Time (SMRT) com a correção dos erros por meio do sequenciamento via Illumina. Os genomas de *Z. luxurians* (SRR088692), *Z. diploperennis* (SRR5466388), *Tripsacum dactyloides* (SRR5127199), *Coix aquatica* (SRR1634981) e *C. lacryma-jobi* (SRR1634980) não possuem o genoma montado e, portanto, foi utilizado o sequenciamento disponível no NCBI Sequence Read Archive (SRA) nos respectivos acessos descritos acima.

4.2.2. Métodos

4.2.2.1. Data mining dos motivos dos knobs heterocromáticos

O data mining foi realizado por meio da ferramenta BLAST. Para os genomas montados foi utilizado a ferramenta BLAST presente no próprio banco de dado. Os parâmetros utilizados para esta ferramenta foram *max target sequence* = 500, *e-value* = 1, *word size* = 11 e *match/mismatch scores* = 2, -3.

Foram utilizadas duas sequências referências para a busca destas famílias de DNA satélite nos genomas. Foi utilizado o seguinte código de acesso: AF030935.1 (ANANIEV; PHILLIPS; RINES, 1998a) para localizar os motivos de K180 e para K180_B utilizou-se a sequência consenso dos motivos encontrados nas sequências de K180 na linhagem B73. Motivos que apresentavam expressivas distorções mutacionais e número de nucleotídeos inferior ou superior a 3 pb do valor de 178 pb foram excluídos das análises.

Para os genomas sequenciados, foram coletadas apenas a região de 87 pb uma vez que os reads não possuem comprimento total do motivo. Fizemos os downloads dos sequenciamentos de cada genótipo pelo software *Fastq-dump* (<http://ncbi.github.io/sra-tools/> e SRA Toolkit Development Team). Com o *Cutadapt* (MARTIN, 2011) os reads com baixa qualidade (q 20,20) foram filtrados e o BLASTn foi utilizado para realizar a busca dos motivos. Todos os motivos alinhados resultantes do BLASTn com tamanho menor a 85 pb ou superior a 89 pb foram excluídos, novamente, utilizando o *Cutadapt*.

4.2.2.2. Análises das Sequências - Descrição

Após a coleta, os motivos foram separados por genótipos e organizados de acordo com a sua posição no genoma para aqueles que apresentavam montagens. Os alinhamentos múltiplos dos motivos das duas famílias de DNA satélites foram realizados pelo algoritmo *Muscle* do pacote MEGA 7.0 (KUMAR, STECHER e TAMURA, 2016).

Para realizar as análises de comparação das duas famílias de DNA satélite, foram geradas as sequências consensos de uma amostra dos motivos de K180 e K180_B da linhagem B73 utilizando o software *Jalview* (CLAMP et al., 2004) e depois geramos um gráfico DotPlot destes consensos utilizando a ferramenta *Jdotter* (BRODIE, ROPER e UPTON, 2004).

4.2.2.3. Análises das Sequências - Diversidade

As análises de diversidade contribuíram para discutir a evolução em concerto da família K180_B. A primeira análise realizada foi o *Sliding-Window* construído pelo *DnaSP* (LIBRADO e ROSAS, 2009), onde é apresentado a diversidade nucleotídica em Pi de cada genoma ao longo de toda a sequência. A configuração utilizada para este procedimento foi o comprimento da janela em 5pb e o tamanho do espaço de 1 pb. Para complementar as análises de diversidade também foi calculada a diversidade nucleotídica pelo modelo de Jukes Cantor (JC) para cada genoma deste estudo dentro do pacote MEGA 7.0. Outros modelos de diversidade, como *p-distance* e *Tamura-3-Parameter*, também foram calculados com objetivo de comparar os resultados obtidos pelo modelo de Jukes Cantor. Não foram observadas diferenças entre os modelos utilizados (Tabela Suplementar 1).

4.2.2.4. Análises das Sequências – Evolução em Concerto

Foram construídas duas árvores MLTree. A primeira foi construída com os motivos de todas as espécies e a segunda foi obtida somente com os motivos das espécies com o genoma montado.

As duas árvores foram construídas por meio do método estatístico de Máxima Verossimilhança (*Maximum Likelihood*) e com *bootstrap* de 100 repetições no MEGA 7.0. O modelo selecionado para as árvores foi *Tamura-3-Parameter* de acordo com as recomendações do pacote MEGA 7.0. Os resultados foram apresentados na forma de árvores de similaridade, utilizando-se uma estrutura circular e os terminais apresentados por meio de círculos coloridos onde os genomas foram representados por coloração diferente.

Também foi construída uma árvore filogenética a partir das sequências consensos do trecho de 87 pb dos motivos de K180_B para verificar a presença de mutações fixadas durante a evolução destas espécies. O modelo utilizado foi Jukes Cantor com a apresentação da árvore no formato tradicional retangular e com os terminais apresentados pelo nome da espécie. O modelo, também, foi selecionado de acordo com as recomendações do pacote MEGA 7.0.

As figuras WebLogo de cada genoma foram construídas pelo próprio software WebLogo disponível online (CROOKS et al., 2004). As configurações utilizadas foram 90 nucleotídeos por linha e coloração clássica (NA).

4.3. Resultados

4.3.1. Descrição de K180_B

Após um alinhamento com diversos motivos de K180, observou-se um pequeno grupo de sequências que continham diversos polimorfismos em relação às demais. Estas sequências foram separadas e posteriormente foram realizadas análises para verificar a similaridade com a família K180. Este conjunto de sequências foi nomeado de K180_B.

As análises comparativas foram realizadas com os motivos da linhagem B73, uma linhagem modelo para estudo de genética em milho. Foi estabelecido uma sequência consenso que caracteriza esse conjunto variante K180_B e foi comparada com uma sequência consenso dos motivos padrões da família K180 por DotPlot. Nesta análise observam-se duas regiões que apresentaram características distintas entre estes dois conjuntos de motivos representado pelos seus consensos (Figura 1a). O motivo de ambas, K180 e K180_B, apresentam alto nível de conservação em sua primeira metade, do nucleotídeo 1 ao 93. Entretanto, do nucleotídeo 94 a 180, os motivos dos dois grupos de sequências apresentam uma baixa porcentagem de similaridade.

Podemos visualizar este resultado, claramente, no alinhamento dos dois consensos. A segunda metade do alinhamento aparece destacado (Fig 1b) para indicar o alto nível de polimorfismo. No total foram mapeadas 48 posições polimórficas que equivalem a uma dissimilaridade de 26%. Somente na região destacada existem 33 polimorfismos que correspondem 72% do total.

A partir das análises de *Sliding-Window* (Figura 1c) observam-se dois pontos importantes. Primeiro, a subfamília K180_B possui maior variabilidade dos nucleotídeos entre os seus motivos em relação a subfamília K180. O outro ponto a se destacar é o maior grau de diversidade localizado na região final das duas famílias, caracterizado com um pico no gráfico (Figura 1c). Esta mesma região é o local onde está presente a baixa porcentagem de similaridade entre K180 e K180_B. O alto grau de polimorfismo na segunda metade dos motivos de K180_B indicam uma predisposição ao acúmulo de mutações. Tal acúmulo de mutação pode ser a causa que conduziu a diferenciação desta região entre as famílias de K180 e K180_B.

As análises de diversidade nucleotídica apresentam resultados que complementam as análises de *Sliding-Window*. Os motivos de K180 possuem cerca de 9% de diversidade entre as suas sequências e, enquanto, K180_B apresenta 15% de diversidade (Tabela 1a). Os valores em *Tamura-3-parameter* reforçam esses resultados, uma vez que os valores foram proporcionalmente semelhantes (Tabela 1b). Assim pode-se inferir que os motivos de K180_B apresentam uma maior variabilidade entre os seus motivos, enquanto os motivos de K180 são mais conservados. Nesta análise verifica-se que a diversidade nucleotídica entre todos os motivos de K180 em relação aos motivos de K180_B é de aproximadamente 25%, um valor relativamente alto para sequências de uma mesma família de DNA satélite (KAWABE, NASUDA; 2005).

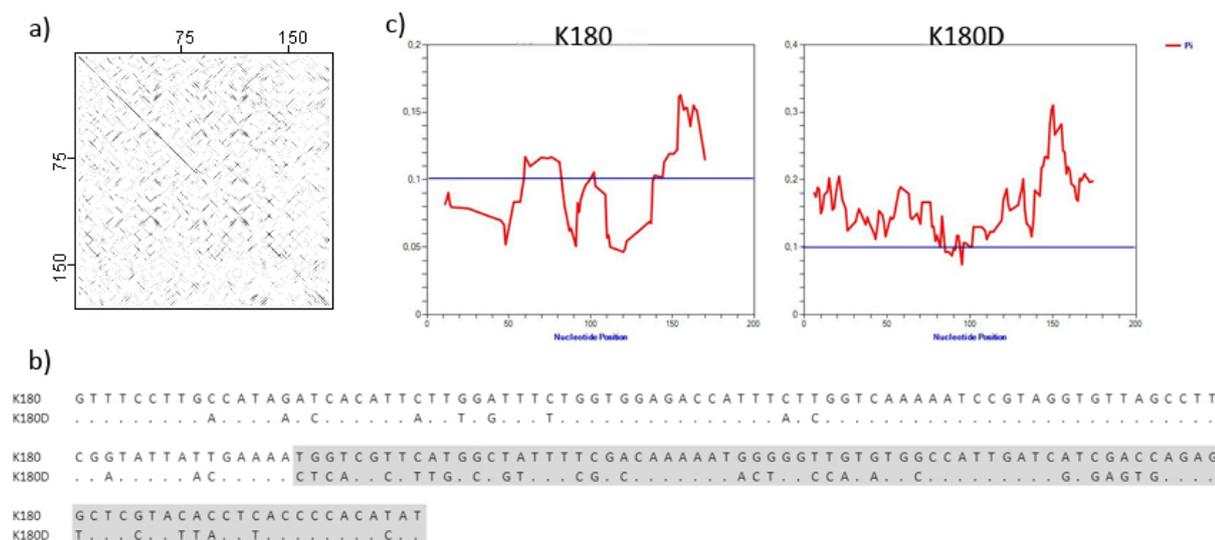


Figura 1. Análises descritivas e comparativas entre K180 e K180_B. a) DotPlot entre as sequências consensos de K180 (Horizontal) e K180_B (Vertical) da linhagem de milho B73. b) Alinhamento das sequências consensos entre K180 e K180_B. No quadro cinza está destacado a região com maior grau de diversidade. c) Análise das duas famílias em *Window-Sliding* obtidos através do software *DnaSP*. Observa-se que a linha média é superior em K180_B, indicando uma maior diversidade de nucleotídeos.

Para analisar o comportamento evolutivo entre K180 e K180_B foi construída uma MLtree. O resultado mostra a formação de dois *clusters* de motivos separados com bootstrap de 100%, deixando explícita sua divergência evolutiva (Fig 2a).

Essa clara separação permite a subdivisão destes motivos em duas subfamílias. Como os polimorfismos estão concentrados em uma região de 87 pb, com alto nível de conservação nos primeiros 93 pb, a proposta de divisão em subfamília é mais parcimoniosa.

a)	K180	K180_B
K180	0,09 ± 0,01	
K180D	0,25 ± 0,04	0,15 ± 0,02
b)	K180	K180_B
K180	0,12 ± 0,02	
K180D	0,42 ± 0,14	0,20 ± 0,03

Tabela 1. Diversidade nucleotídica de K180 e K180_B calculado em a) p-distance e b) Tamura-3-paramater.

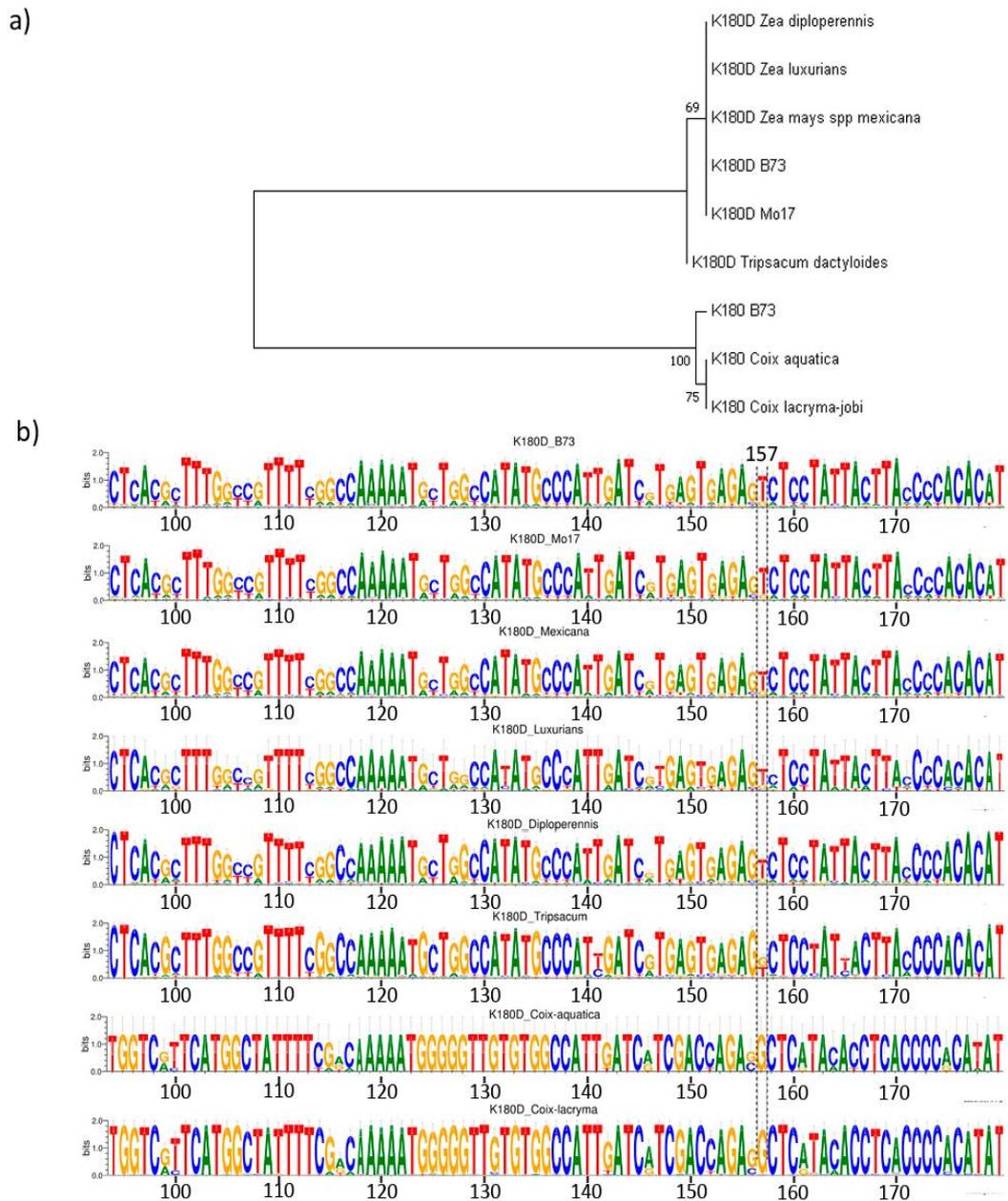


Figura 2. a) Árvore filogenética construída apenas com as sequências consenso de K180_B das linhagens B73 e Mo17 de *Zea mays*, *Zea mays spp mexicana*, *Zea diploperennis*, *Zea luxurians*, *Tripsacum dactyloides* e com as sequências K180 da linhagem B73, *Coix aquatica* e *Coix lacryma-jobi* como grupo externo b) WebLogo correspondente a segunda metade das sequências de K180_B das linhagens B73 e Mo17 de *Zea mays*, *Zea mays spp mexicana*, *Zea diploperennis*, *Zea luxurians*, *Tripsacum dactyloides* e com as sequências K180 de *Coix aquatica* e *Coix lacryma-jobi*, respectivamente. Destaque para a posição 157 que possui um nucleótido diferente fixado entre *Tripsacum* e o gênero *Zea*.

4.3.2. K180_B está presente em espécies ancestrais

Para localizar o ponto inicial da separação evolutiva entre ambas famílias de DNAs satélites e seus respectivos comportamentos evolutivos, foi realizada a busca da variante K180_B nos genomas de mais uma linhagem (Mo17) e de espécies ancestrais de milho (*Z. mays spp mexicana*, *Z. luxurians*, *Z. diploperennis* e *Tripsacum dactyloides*). Nas espécies *Coix aquatica* e *C. lacryma-jobi* não foi encontrada a variante K180_B, somente uma pequena quantidade de K180.

Para se mapear o comportamento dos motivos de K180_B ao longo da história evolutiva da origem do milho, um trecho de 87 pb correspondente a região polimórfica foi analisado a partir do consenso obtido para as sequências de cada uma das espécies, utilizando-se *Coix aquatica* e *Coix lacryma-jobi* como grupo externo por conter apenas cópias do motivo padrão (Figura 2a).

A árvore revela que as sequências consensos das espécies do gênero *Zea* não se resolveram, pois estas não obtiveram nenhuma diferença nucleotídica em toda extensão do consenso. Houve separação da sequência consenso de *Tripsacum dactyloides* com as espécies do gênero *Zea* representado apenas por uma pequena diferença presente em uma posição nucleotídica do consenso.

Na topologia, *T. dactyloides* aparece como uma espécie irmã dentro do clado *Zea-Tripsacum*. E por fim, como esperado o gênero *Coix* possui 100% de bootstrap separando integralmente dos motivos de K180_B.

A diferença entre as sequências consensos de *Tripsacum* e *Zea* está localizada na posição 157 (destacado na figura 2b), onde *Zea* possui uma Timina e *T. dactyloides* uma Guanina. Interessantemente, foi observado que a taxa de fixação da base Timina na posição 157 aumenta com a derivação das espécies. No gênero *Coix*, a base Guanina está presente na grande maioria das sequências, em *T. dactyloides*, a quantidade desta base diminui consideravelmente com o aparecimento da Timina. Em *Z. diploperennis*, *Z. luxurians* e *Z. mays spp mexicana* ocorreu uma mudança na proporção das bases, onde a Timina está presente em maior proporção em relação a Guanina e por fim, as duas linhagens de *Z. mays* tiveram um aumento na quantidade de Timina nesta posição (Figura 2b).

Estes resultados demonstram que os motivos de K180_B de *T. dactyloides*, a espécie mais distante do milho cultivado, possui maior identidade com motivos de K180 na posição 157 em relação aos próprios motivos de K180_B das outras espécies.

4.3.3. Análise de Diversidade de K180_B

Para analisar com maiores detalhes a variabilidade de K180_B foram gerados *Sliding-Window* do conjunto de motivos de cada uma das espécies (Figura 3). Os gráficos de *Sliding-Window* dos genótipos B73, Mo17 e *Z. mays spp mexicana* (Figura 3a, 3b, 3c) possuem uma grande identidade os valores de diversidade nucleotídica. As principais características de similaridade entre estes genótipos são os dois picos presentes no centro e na porção final do consenso. A semelhança observada para estes três materiais se deve a sua proximidade genética, fazendo com que ocupem o mesmo clado dentro da

árvore filogenética do grupo de espécies do gênero *Zea* (BUCKER IV e HOLTSFORD, 1996). Enquanto as outras espécies possuem variações na diversidade que pode ser consequência de suas distâncias filogenéticas dentro do gênero *Zea*.

Tabela 2. Diversidade nucleotídica calculado em Jukes-Cantor da região de 87pb de K180_B nas linhagens B73 e Mo17 de *Zea mays*, *Zea mays spp mexicana*, *Zea diploperennis*, *Zea luxurians*, *Tripsacum dactyloides* e da mesma região em K180 de *Coix aquatica* e *Coix lacryma-jobi*

	Diversidade (Pi)
B73	0,226 +/- 0,026
Mo17	0,221 +/- 0,025
<i>Zea mays spp mexicana</i>	0,238 +/- 0,028
<i>Zea luxurians</i>	0,239 +/- 0,027
<i>Zea diploperennis</i>	0,223 +/- 0,020
<i>Tripsacum dactyloides</i>	0,157 +/- 0,017
<i>Coix aquatica</i>	0,062 +/- 0,015
<i>Coix lacryma-jobi</i>	0,067 +/- 0,014

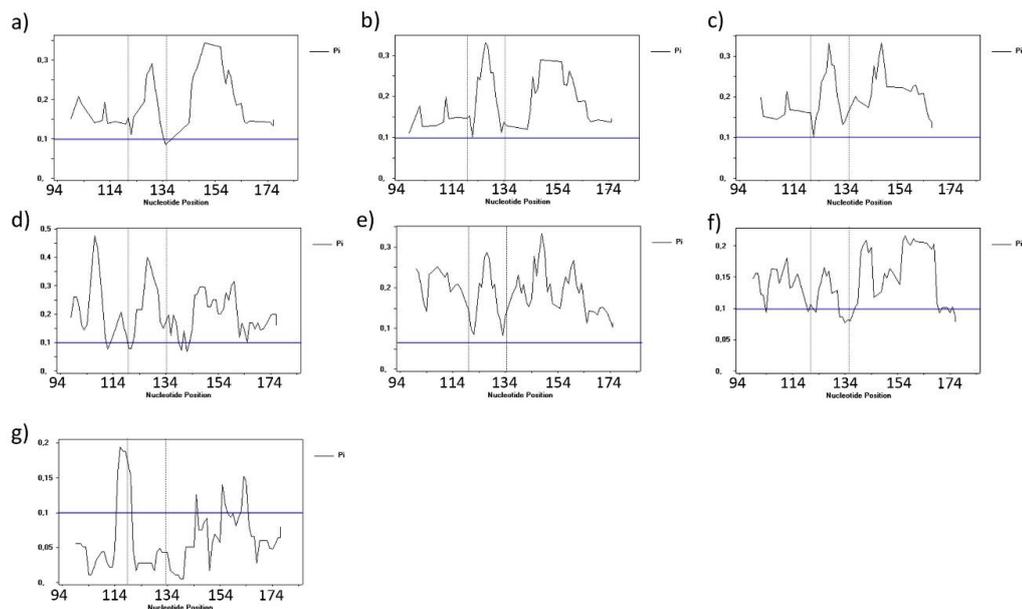


Figura 3. Sliding-window obtido a partir dos valores de diversidade nucleotídica de K180_B de cada espécie em *p-distance*. a) B73, b) Mo17, c) *Zea mays spp mexicana*, d) *Zea diploperennis*, e) *Zea luxurians* e f) *Tripsacum dactyloides*. A letra g) é o sliding-window das sequências K180 em *Coix lacryma-jobi* e *Coix aquatica*. As linhas pontilhadas em cada sliding-window indica uma região variável que é conservada entre as espécies que contém K180_B.

A região entre as posições 114 e 134 pb possui um pico de grande variabilidade que é característico em todas as espécies do gênero *Zea*. Esta região é composta pelos nucleotídeos TGCTGGCGATATG. Possivelmente esta região pode estar sofrendo algum tipo de pressão evolutiva

que gera este comportamento de alta variabilidade. Uma hipótese seria a sua localização dentro de uma região de DNA-linker entre nucleossomos, uma vez que o DNA satélite possui o tamanho de um nucleossomo. Deste modo esta região estará mais acessível ao meio e, conseqüentemente, exposta para ação de agentes mutacionais e ocasionando a alta variabilidade.

Também é observado que as espécies do gênero *Zea* possuem uma alta diversidade e os seus valores são semelhantes entre si (Tabela 2). Inversamente, *T. dactyloides* possui um valor menor de diversidade e que difere das outras espécies, indicando que os motivos de K180_B de *T. dactyloides* são mais conservados dos que os motivos do gênero *Zea*.

4.3.4. Árvore de Similaridade – Evolução em Concerto

Para estudar a dinâmica evolutiva de K180_B em todas as espécies, foi construída uma árvore de similaridade com todos os motivos, acrescentado como grupo externo todos os motivos K180 das duas espécies de *Coix* (Figura 4a). As sequências pertencentes a *Coix* são separadas em um clado externo conforme esperado e indicando a origem monofilética de K180_B.

As sequências das espécies do gênero *Zea* tendem a se agrupar em um mesmo ramo da árvore, indicando que estas possuem padrão semelhante de homogeneização da subfamília K180_B. Tal semelhança de homogeneização pode ser consequência do pouco tempo de especiação que estas espécies possuem na história evolutiva (BUCKER IV e HOLTSFORD, 1996).

Importante destacar o comportamento dos motivos de *T. dactyloides* na árvore. A maioria dos motivos estão agrupados em um grande cluster compartilhado com uma pequena quantidade de motivos de espécies do gênero *Zea*, principalmente *Z. diploperennis*. Provavelmente esta maior quantidade de motivos de *Z. diploperennis* no clado é consequência da sua maior proximidade com *T. dactyloides*.

Uma pequena quantidade de motivos de *T. dactyloides* estão espalhados em outros ramos agrupados com uma maioria de *Zea*. Estes resultados demonstram que os motivos de *T. dactyloides* são mais relacionados entre si do que os motivos de outras espécies.

Também é observado um grupo de sequências de *T. dactyloides* que está inserido dentro do ramo das sequências das espécies do gênero *Coix*. Em uma análise detalhada verifica-se que estas sequências possuem regiões similares a K180_B e a K180, simultaneamente (Figura 4b). Portanto, a região polimórfica de K180_B de *T. dactyloides* pode ser subdividido em duas subregiões, a primeira entre os nucleotídeos 94 e 133 conservado com K180 e a segunda entre os nucleotídeos 146 e 180 idênticos aos K180_B de *Zea*. Isso explicaria o posicionamento intermediário de *T. dactyloides* entre *Coix* e *Zea*. Essa característica dos motivos de *T. dactyloides* indicam que a subfamília K180_B surgiu a partir de dois eventos distintos. O primeiro evento durante a divergência que deu origem ao *Tripsacum* e um segundo evento durante a separação do gênero *Zea*. Esta assinatura evolutiva no DNA repetitivo sugere o gênero *Tripsacum* como o elo evolutivo entre *Zea* e *Coix*.

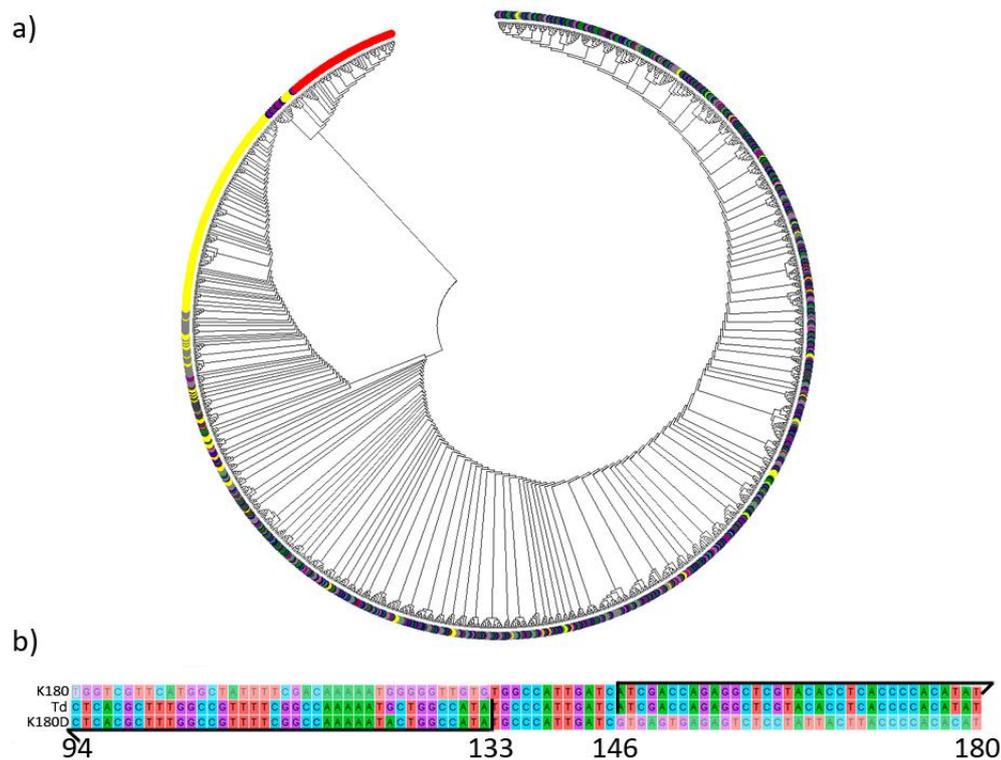


Figura 4. Análise da dinâmica evolutiva dos motivos K180_B nas espécies estudadas. a) Árvore de similaridade de seqüências de K180_B das linhagens B73 (círculo azul) e Mo17 (círculo verde) de *Zea mays*, *Zea mays subsp mexicana* (círculo roxo), *Zea diploperennis* (círculo cinza), *Zea luxurians* (círculo vinho), *Tripsacum dactyloides* (círculo amarelo) e com as seqüências K180 de *Coix aquatica* e *Coix lacrymajobi*(círculo vermelho). b) Alinhamento das seqüências K180 e K180_B com a seqüência de *T. dactyloides* (Td) presente na região intermediária do ramo entre o cluster de K180 e o cluster de K180_B. A primeira, segunda e terceira linha está representada pelas seqüências K180, Td e K180_B, respectivamente. As setas na região inicial e na final indicam regiões de similaridade entre Td com K180_B e K180, respectivamente.

Para se entender o padrão de especificidade da diversidade nucleotídica de cada região cromossômica, foram obtidas e analisadas as posições polimórficas em WebLogos (Figura 5). Existem algumas posições nucleotídicas que estão em diferentes passos de fixação de mutação em relação a mesma posição localizada em outro braço cromossômico. A posição nucleotídica 124 possui esta característica, por exemplo, pode-se observar que no braço curto do cromossomo 7 da linhagem B73 (Figura 5a) a base mais presente dos motivos nesta posição é a Adenina, enquanto os outros braços cromossômicos (Figura 5b, 5c, 5d) é principalmente constituído por Guanina. Outras posições como 107, 146, 148 e 157 também possuem esta característica.

Também é observado uma variabilidade da frequência de uma base nucleotídica dentro de uma mesma posição. Isto é verificado nas posições 163 e 179, onde um braço cromossômico possui uma

diminuição expressiva da base Adenina nos motivos. Todos estes resultados também foram encontrados em Mo17 e *Z. mays spp mexicana*.

Porém, quando é analisada a variabilidade destas mesmas posições de nucleotídeos e as suas mutações de um mesmo braço cromossômico em outro genótipo, observa-se o mesmo padrão de comportamento. Isto evidencia que o padrão de fixação de uma mutação está conservado entre os três genótipos no mesmo braço e que determinada mutação está sendo conservada durante o processo de diversificação destas espécies e entre linhagens.

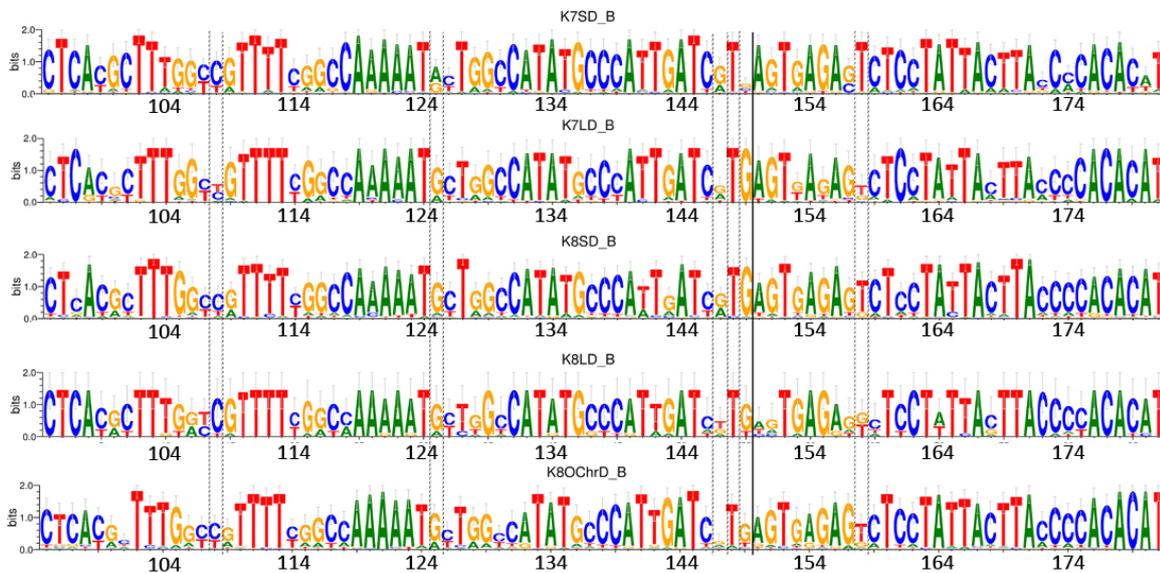


Figura 5. WebLogo das sequências presentes no braço curto e longo dos cromossomos 7 e 8 e localizadas em outros cromossomos da linhagem B73. K7SD representa braço curto do cromossomo 7; K7LD representa o braço longo do cromossomo 7; K8SD representa o braço curto do cromossomo 8; K8LD representa o braço longo do cromossomo 8 e KOChrD representa os demais cromossomos. Linhas pontilhadas destacam as posições com heteromorfismos

4.3.5. Comportamento evolutivo das sequências de K180_B nos braços cromossômicos

Para se entender o comportamento evolutivo dos motivos da subfamília K180_B dentro de suas posições nos braços cromossômicos foi construída uma árvore de similaridade (Figura 7). Primeiro destaca-se os motivos a nível cromossômico e nota-se uma tendência de agrupamentos dos motivos dos cromossomos 7 e 8 em ramos distintos (Figura 7a). E somente uma pequena quantidade dos motivos do cromossomo 8 estão dispersos dentro do agrupamento do cromossomo 7. Estes dois agrupamentos evidenciam que os motivos do cromossomo 7 são mais relacionados com os seus próprios motivos, assim como também ocorre com os motivos do cromossomo 8.

Quando analisados os motivos a nível de braço cromossômico percebe-se que as sequências do braço curto do cromossomo 7 possuem um cluster específico e separado das sequências do braço longo (Figura 7b). Tal resultado indica que as sequências destes dois braços estão em níveis diferentes de homogeneização. O mesmo não ocorre no cromossomo 8, onde as sequências compartilham o mesmo grau de similaridade entre os dois braços cromossômicos. Este comportamento também é observado para os outros dois genótipos (Figura Suplementar 1 e 2).

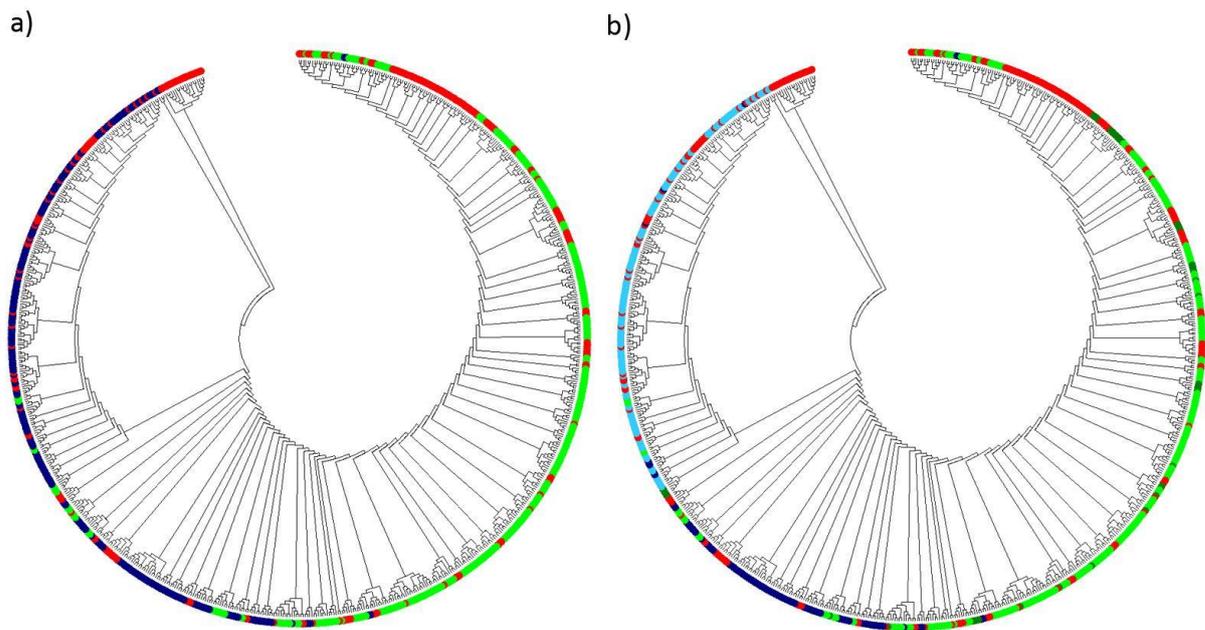


Figura 7. a) Árvore de Similaridade das sequências de K180_B linhagens B73 e Mo17 de *Zea mays* e *Zea mays spp mexicana* com separação dos terminais à nível de cromossomo. Cromossomo 7 representado pelo círculo azul, cromossomo 8 representado pelo círculo verde e os demais cromossomos representados pelo círculo vermelho. b) Árvore de Similaridade das sequências de K180_B linhagens B73 e Mo17 de *Zea mays* e *Zea mays spp mexicana* com separação dos terminais a nível de braços cromossômicos. Braço curto do Cromossomo 7 representado pelo círculo azul escuro, braço longo do Cromossomo 7 representado pelo círculo azul claro, braço curto do cromossomo 8 representado pelo círculo verde escuro, braço longo do cromossomo 8 representado pelo círculo verde claro e os demais cromossomos representados pelo círculo vermelho.

Já os motivos dos cromossomos restantes estão dispersos nos dois grandes ramos do cromossomo 7 e 8, ausente de qualquer comportamento lógico. Também é observado um pequeno *cluster* de sequências dos cromossomos restantes com 33% de *bootstrap* formado como um grupo

externo dos demais. Em uma análise detalhada, verificou-se que algumas sequências deste grupo possuem outros tipos de mutações que não são semelhantes à família K180 e K180_B.

4.4. Discussão

Foi caracterizada uma nova subfamília, denominada de K180_B, presente nos knobs das linhagens de milho e das respectivas espécies ancestrais. K180_B possui aproximadamente 25% de diversidade em relação às sequências da família K180, o DNA satélite mais abundante no knob heterocromático. Este valor é aproximadamente semelhante aos valores presentes na diversidade entre as subfamílias de DNA satélites do gênero *Arabidopsis* (KAWABE, NASUDA; 2005). Os valores encontrados nestas subfamílias foram entre 28 a 33% e foram suficientes para separar visualmente as subfamílias por FISH (KAWABE, NASUDA; 2005). A quantidade de motivos desta nova subfamília encontrada no sequenciamento foi muito inferior em relação a K180, representando uma pequena fração na composição dos knobs.

DENNIS and PEACOCK (1984) observaram uma grande quantidade de motivos de K180 no gênero *Zea* e *Tripsacum* e não localizaram nenhuma sequência nas espécies de *Coix*. Em nossas análises, inversamente ao que foi observado por DENNIS e PEACOCK (1984), podemos observar uma pequena quantidade de K180 nas duas espécies de *Coix*, com base nos dados de sequenciamento disponíveis. Possivelmente, pela baixa quantidade destes motivos e pela baixa resolução do método de ISH com sondas radiotivas, não foi possível localizar estas sequências de K180 nas espécies do gênero *Coix*.

Destaca-se que *Tripsacum dactyloides* é a espécie mais ancestral que possui os motivos K180_B entre as espécies estudadas. Esta espécie tem uma história filogenética em comum com as espécies do gênero *Zea* formando um grupo monofilético dentro da família *Poaceae* (BUCKER IV e HOLTSFORD, 1996; DENNIS and PEACOCK, 1984). Os organismos com genomas sequenciados mais próximos de *Zea mays* após *Tripsacum* é *Coix* (BUCKER IV e HOLTSFORD, 1996), entretanto, esta espécie não possui os motivos de K180_B, reforçando a hipótese de origem da K180_B durante o processo de diversificação das três espécies.

As sequências consensos das espécies de *Zea* não se resolveram dentro da árvore filogenética, resultado distinto ao que foi observado na filogenia feita com sequências de ITS destas mesmas espécies (BUCKER IV e HOLTSFORD, 1996). Entretanto, esta característica da árvore foi similar ao visto na árvore filogenética construída por sequências gênicas de cloroplasto (DOEBLEY, 1990). Portanto, os resultados evolutivos de K180_B foram semelhantes a uma sequência gênica e diferente de uma sequência repetitiva.

Conclui-se, então, que os motivos de K180_B no grupo *Zea* e *Tripsacum* possui uma alta variabilidade, porém não são o suficiente para divergirem em uma outra família de DNA satélite. Portanto, os motivos de K180_B possuem algum mecanismo que permite uma taxa de variabilidade alta

sem perder a sua característica principal. Algo semelhante ao que acontece com as sequências que estão organizadas em *High Order Repeats* (HOR – WILLARD et al., 1987; SUJIWATTANARAT, 2015).

4.4.1. Evolução Sequências Repetitivas K180_B

Os motivos de *T. dactyloides* formaram um agrupamento isolado com uma pequena quantidade de sequências de espécies do gênero *Zea*, assim como os motivos de *Zea* foram agrupados em um grande clado que contém uma pequena quantidade de motivos de *T. dactyloides*. A presença destes pequenos grupos de motivos de um gênero em um ramo com uma grande quantidade de motivos do outro gênero indica a conservação destas sequências durante a diversificação de *Zea-Tripsacum*. Tais motivos são conjuntos ancestrais que foram diferencialmente amplificados quando houve a separação evolutiva de cada gênero conduzidos pelos mecanismos de Evolução em Concerto. Este modelo de evolução é denominado de biblioteca (FRY and SALSER, 1977; RUIZ-RUANO et al, 2016; GARRIDO-RAMOS, 2017).

A presença expressiva de motivos de *Zea diploperennis* no ramo *Tripsacum* indica que a amplificação dos motivos de *T. dactyloides* devem ter iniciado anteriormente a diferenciação da espécie *Zea diploperennis*.

A origem da família K180_B provavelmente teve início em *Tripsacum* conduzida pelos mecanismos de Evolução em Concerto. Isso se deve a presença de K180_B neste gênero e a sua ausência nas espécies de *Coix*, a próxima espécie mais distante de milho com knob. O complemento desta afirmação também se deve pela presença de motivos de K180_B em *T. dactyloides* que possuem uma região com a inserção de uma pequena parte da sequência de K180. A evolução dever ter iniciado com uma mutação nesta região final das sequências K180, que, por alguma razão, possui maior suscetibilidade a geração de uma mutação em relação a primeira parte. Posteriormente esta mutação foi fixada e amplificada no genoma. O mesmo ocorreu nos outros nucleotídeos da porção final de K180_B.

A data de divergência do gênero *Tripsacum* e *Zea* foi estimada em 4,5 milhões de anos atrás (HILTON e GAUT, 1998) e possivelmente esta mesma data possa ser sugerida para a origem da K180_B no genoma do ancestral comum dos dois gêneros. Após estes 4,5 milhões de anos somente um nucleotídeo foi fixado (a posição 157), a partir dos resultados das sequências consensos. O tempo de evolução desta família de DNA satélite é relativamente baixa em relação às sequências centroméricas do gênero *Arabidopsis* que alteraram a composição dos motivos em 73% após 5 milhões de anos (HALL et al., 2005; KAWABE, NASUDA; 2005) e também do gênero *Nicotiana*, onde os motivos alteraram 72% após 4,5 milhões de anos (LIM et al., 2007). Tal conservação dos motivos de DNAs satélites de milho pode ser atribuída ao sistema reprodutivo por alogamia presente nestas espécies e alta compatibilidade entre os seus gametas. Este comportamento causa a troca de informações genéticas entre espécies próximas.

Entretanto, quando comparamos as sequências centroméricas de arroz e milho (duas espécies da família *Poaceae*) verifica-se que houve uma grande similaridade entre os motivos durante 50 milhões de anos (LAI et al., 2004). Deste modo, o tempo de divergência entre estas duas subfamílias parecem ter um valor intermediário, demonstrando que cada DNA satélite possui um comportamento diferente em indivíduos diferentes.

4.4.2. Evolução em Concerto de K180_B – Homogeneização Localmente

A Evolução em concerto conduz a homogeneização dos motivos localmente no genoma. Kuhn et al. (2012) em suas análises com *Drosophila*, verificaram que o processo de homogeneização ocorre diferencialmente entre os motivos da família 1.688 localizados no cromossomo 2, 3 e X, onde cada cromossomo está em um passo de fixação distinto. Esse processo também é semelhante aos motivos de alfa satélite presente em humanos (SCHUELER et al., 2005) e Cent-O em arroz (LEE et al, 2006).

Os motivos de K180_B dos cromossomos 7 e 8 também possuem um modelo de evolução sujeito a homogeneização localmente. Os motivos de cada cromossomo formaram dois ramos isolados com pouco compartilhamento do cromossomo 8 no ramo do cromossomo 7. A homogeneização dos motivos destes dois cromossomos não é tão expressiva quanto observado na família Cent-O em arroz. Em arroz, os motivos dos cromossomos 1 e 8 separam na árvore com valores superiores a 55% de *bootstrap*, enquanto o K180_B possui valores menores de 2%. Isto indica que a sequência variante do knob foi menos afetada pela homogeneização em concerto em relação aos motivos de Cent-O em arroz.

Ainda no cromossomo 7, também existe uma homogeneização local entre os braços curto e longo, observado na separação destes em dois ramos distintos na MLtree. Enquanto no cromossomo 8, os motivos dos dois braços estão espalhados dentro de um mesmo ramo. Este comportamento evidencia que existe algum fenômeno que está atuando na aceleração da homogeneização nos motivos do cromossomo 7 entre os dois braços. Essa característica pode ser conduzida pela troca de informação considerando-se a teoria da equidistância que pode ser observada entre regiões heterocromáticas presente no cromossomo 7 e ausente no cromossomo 8.

Referências

- AGUIAR-PERECIN, M.L.R. Heterochromatin of maize chromosomes: structure and genetics effects. **Genetics and Molecular Biology**, v. 23, n. 4, p. 1015-1019, 2000.
- ALBERT, P. S. et al. Diversity of chromosomal karyotypes in maize and its relatives. **Cytogenetic and Genome Research**, v. 129, n. 1–3, p. 6–16, 2010.

- ANANIEV, E. V.; PHILLIPS, R. L.; RINES, H. W. Complex structure of knob DNA on maize chromosome 9: Retrotransposon invasion into heterochromatin. **Genetics**, v. 149, p. 2025–2037, 1998a.
- ANANIEV, E. V.; PHILLIPS, R. L.; RINES, H. W. Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 95, n. 22, p. 13073–13078, 1998b.
- BUCKLER IV, E.S.; HOLTSFORD, T.P. *Zea* Systematics: Ribosomal ITS Evidence. **Molecular Biology Evolution**, v.13, n.4, p.612-622, 1996
- BRODIE, R.; ROPER, R.L. JDotter: a Java interface to multiple dotplots generated by dotter. **Bioinformatics**, v.20, n.2, p.279-281, 2004.
- CLAMP, M. et al. The Jalview Java alignment editor. **Bioinformatics**, v.20, n.3, p.426-427, 2004.
- CROOKS, G.E. et al. WebLogo: A sequence logo generator. **Genome Research** v. 14 p. 1188-1190, 2004.
- DOEBLEY, J.F. Molecular Systematics of *Zea* (Graminea). **Maydica**, v.35, n.2, p.143-150, 1990
- DOVER, G. Molecular drive: a cohesive mode of species evolution. **Nature**. v. 299, p.111-117, 1982.
- DENNIS, E.S.; PEACOCK, W.J. Knob Heterochromatin Homology in Maize and Its Relatives. **Journal of Molecular Evolution**, v. 20, p. 341-350, 1984.
- FRY, K.; SALSER, W. Nucleotide Sequences of HS-a Satellite DNA from Kangaroo Rat *Dipodomys ordii* and Characterization of Similar Sequences in Other Rodents. **Cell**, v. 12, p. 1069-1084, 1977.
- GARRIDO-RAMOS, M. A. Satellite DNA: An Evolving Topic. **Genes**. v. 8, n. 9, p. 230, 2017
- HALL S.E.; LUO, S.; HALL, A.E.; PREUSS, D. Differential rates of local and global homogenization in centromere satellites from Arabidopsis relatives. **Genetics** v. 170 p. 1913–1927, 2005.
- HILTON, H.; GAUT, B.S. Speciation and Domestication in Maize and Its Wild Relatives: Evidence From the *Globulin-1* Gene. **Genetic**, v.150, n2, p. 863-872, 1998.
- HEMLEBEN, V.; KOVARIK, A.; TORRES-RUIZ, R.A.; *et al.* Plant highly repeated satellite DNA: molecular evolution, distribution and use for identification of hybrids. **Systematics and Biodiversity**, Reino Unido, v. 5, n. 3, p. 277-289, 2007.
- HENIKOFF, S.; AHMAD, K.; MALIK, H.S. The centromere paradox: stable inheritance with rapidly evolving DNA. **Science**, v. 293, n. 5532, p.1098-1102, 2001.
- HUISINGA, K.L.; BROWER-TOLAND, B.; ELGIN, S.C. The contradictory definitions of heterochromatin: transcription and silencing. **Chromosoma**. v.115, n.2, p.110-122, 2006.

- KATO, A.; LAMB, J. C.; BIRCHLER, J. A. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, n. 37, p. 13554–13559, 2004.
- KAWABE, A.; NASUDA, S. Structure and genomic organization of centromeric repeats in *Arabidopsis* species. **Mol Genet Genomics**, v. 272, p. 593–602, 2005.
- KUHN, G.C.S.; KÜTTLER, H.; MOREIRA-FILHO, O. *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.
- KULIKOVA, O.; GEURTS, R.; LAMINE, M. *et al.* Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. **Chromosoma**, v. 113, p. 276–283, 2004.
- KUMAR, S.; STECHET G.; TAMURA K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Mol. Biol. Evol.** Oxford University, v. 33, n. 7, p. 1870-1874, 2016.
- LAI, J; DEY, N.; KIM, C. Characterization of the Maize Endosperm Transcriptome and Its Comparison to the Rice Genome. **Genome Research**, v. 14, p. 1932-1937, 2004.
- LEE, H.; NEUMANN, P.; MACAS, J. *et al.* Transcription and Evolutionary Dynamics of the Centromeric Satellite Repeat CentO in Rice. **Molecular Biology and Evolution**, n. 12, v. 23, p. 2505 -2520, 2006.
- LIBRADO, P.; ROZAS, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. **Bioinformatics**, v. 25, n. 11, p. 1451-1452, 2009.
- LIM, K. Y.; KOVARIK, A.; MATYASEK, R. *et al.* Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. **New Phytologist** v. 175, p. 756–763, 2007.
- MAHESHWARI, S.; ISHII, T.; BROWN C.T. Centromere location in *Arabidopsis* is unaltered by extreme divergence in CENH3 protein sequence. **Genome Research**, v. 27, p. 471-478, 2017.
- McKINLEY, K.L.; CHEESEMAN, I.M. The molecular basis for centromere identity and function. **Nature Reviews Molecular Cell Biology**. v. 17, n. 1, p. 16-29, 2016.
- MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, v.17, n.1, p. 10-12, 2011.
- MEHROTRA, S.; GOYAL, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. **Genomics Proteomics Bioinformatics**, Pequim, v. 12, p. 164-171, 2014.
- MONDIN, M. *et al.* Karyotype variability in tropical maize sister inbred lines and hybrids compared with KYS standard line. **Frontiers in plant science**, v. 5, n. October, p. 544, 2014.
- PEACOCK, W. J. *et al.* Highly repeated DNA sequence limited to knob heterochromatin in maize. **Proceedings of the National Academy of Sciences of the United States of America**, v. 78, n. 7, p. 4490–4, 1981.

- RUDD, M.K.; WILLARD, H.F. Analysis of the centromeric regions of the human genome assembly. **Trends in genetics**, v. 20, n. 11, p. 529-33, 2004.
- RUIZ-RUANO, F.J.; LÓPEZ-LEÓN, M. D.; CABRERO, J. et al. High-throughput analysis of the satellitome illuminates satellite DNA evolution. **Sci Rep**, v. 6, 2016
- SCHNABLE, P. et al. The B73 maize genome: Complexity, diversity, and dynamics. **Science**, v. 326, n. 5956, p. 1112-1115, 2009.
- SCHUELER, M.G.; DUNN, J.M.; BIRD C.P. *et al.* Progressive proximal expansion of the primate X chromosome centromere. **PNAS**, v. 102, n. 30, p. 10563-10568, 2005.
- STRACHAN, T.; WEBB, D.; DOVER, G.A. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*, **The EMBO Journal**, Oxford, v. 4, p. 1701-1708, 1985.
- SUJIWATTANARAT, P.; THAPANA, W.; SRIKULNATH, K. *et al.* Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. **Scientific Reports**, v. 5, n. 1, p. 10315, 2015.
- VOLPE, T. A.; KIDNER, C.; HALL, I. M. et al. Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi. **Science**, p. 1833-1837.
- VONDRAK, T.; ROBLEDILLO, L. A.; NOVAK, P. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon derived tandem repeats. *The plant journal*.
- WILLARD. H.F.; WAYE, J.S. Hierarchical order in chromosome-specific human alpha satellite DNA. **Trends in Genetics**, v. 3, n. 7, p. 192-198, 1987.

Materiais Suplementares

Tabela suplementar 1. Diversidade nucleotídica calculado em a) *p-distance* e b) *Tamura-3-parameter* da região de 87pb de K180_B nas linhagens B73 e Mo17 de *Zea mays*, *Zea mays spp mexicana*, *Zea diploperennis*, *Zea luxurians*, *Tripsacum dactyloides* e da mesma região em K180 de *Coix aquatica* e *Coix lacryma-jobi*

a)	Diversidade (Pi)	b)	Diversidade (Pi)
B73	0,191 +/- 0,019	B73	0,236 +/- 0,032
Mo17	0,188 +/- 0,021	Mo17	0,228 +/- 0,030
<i>Zea mays spp mexicana</i>	0,198 +/- 0,019	<i>Zea mays spp mexicana</i>	0,246 +/- 0,030
<i>Zea luxurians</i>	0,204 +/- 0,017	<i>Zea luxurians</i>	0,242 +/- 0,026
<i>Zea diploperennis</i>	0,191 +/- 0,016	<i>Zea diploperennis</i>	0,227 +/- 0,025
<i>Tripsacum dactyloides</i>	0,140 +/- 0,014	<i>Tripsacum dactyloides</i>	0,159 +/- 0,018
<i>Coix aquatica</i>	0,059 +/- 0,016	<i>Coix aquatica</i>	0,063 +/- 0,017
<i>Coix lacryma-jobi</i>	0,062 +/- 0,011	<i>Coix lacryma-jobi</i>	0,067 +/- 0,014

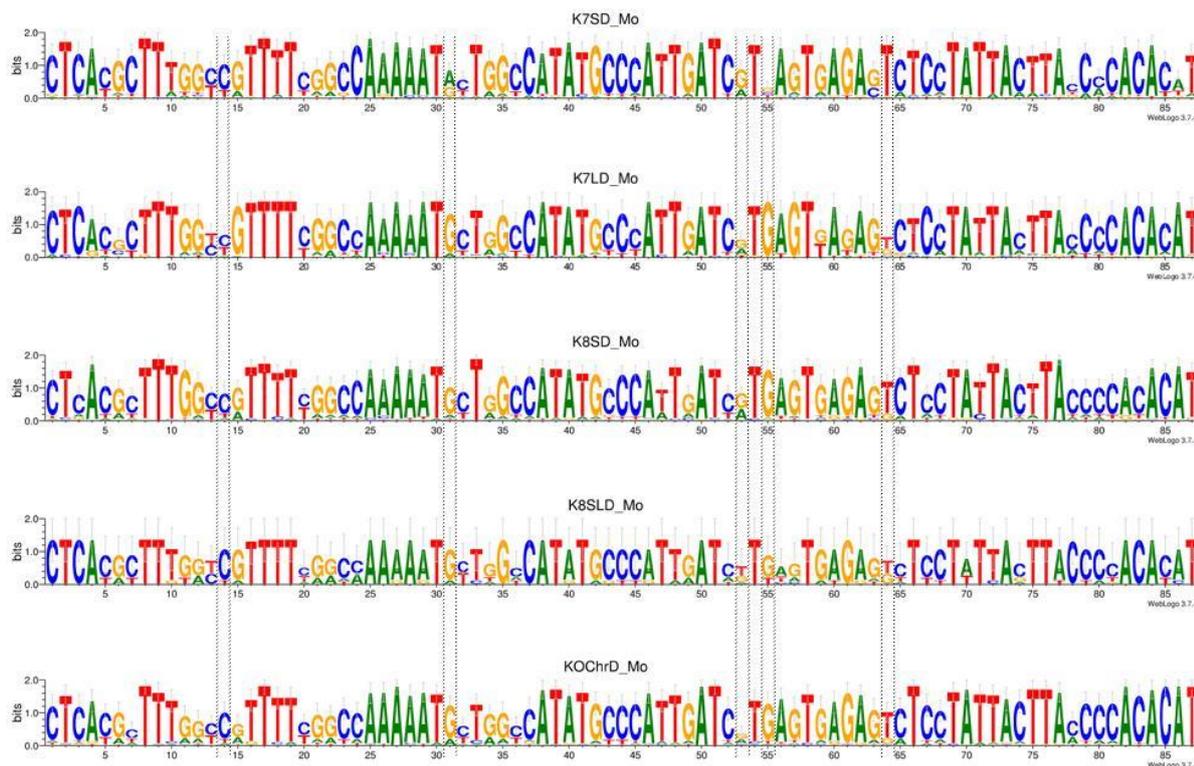


Figura Suplementar 1. WebLogo das sequências presentes no braço curto e longo dos cromossomos 7 e 8 e localizadas em outros cromossomos da linhagem Mo17. K7SD representa braço curto do cromossomo 7; K7LD representa o braço longo do cromossomo 7; K8SD representa o braço curto do cromossomo 8; K8LD representa o braço longo do cromossomo 8 e KOChrD representa os demais cromossomos.

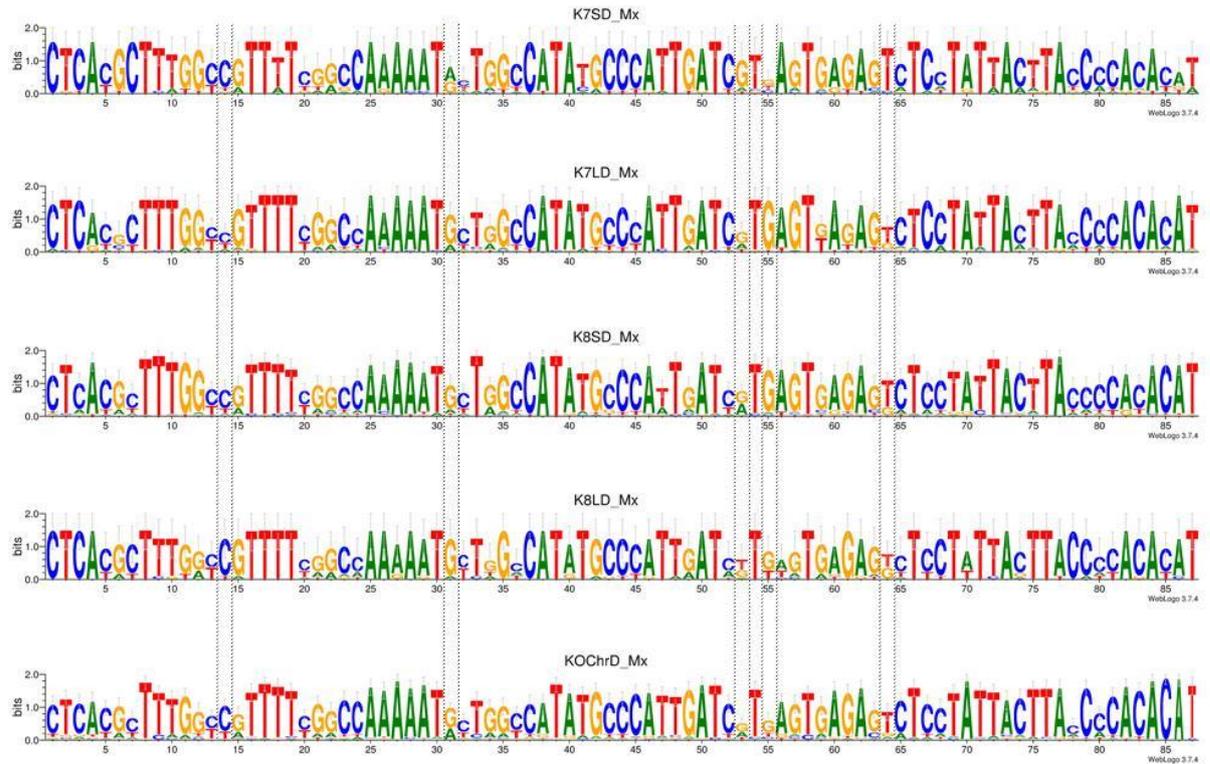


Figura Suplementar 2. WebLogo das seqüências presentes no braço curto e longo dos cromossomos 7 e 8 e localizadas em outros cromossomos da linhagem *Z. mays spp mexicana*. K7SD representa braço curto do cromossomo 7; K7LD representa o braço longo do cromossomo 7; K8SD representa o braço curto do cromossomo 8; K8LD representa o braço longo do cromossomo 8 e KOChrD representa os demais cromossomos.