UNIVERSIDADE FEDERAL DE ALAGOAS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM
AGRONOMIA - PRODUÇÃO VEGETAL

GLEICA MARIA CORREIA MARTINS

Análises evolutivas em *Spondias* L.: uma abordagem genômica do gênero

Rio Largo, AL
2019

GLEICA MARIA CORREIA MARTINS

Análises evolutivas em *Spondias* L.: uma abordagem genômica do gênero

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Agronomia-Produção Vegetal da Universidade Federal de Alagoas, como requisito parcial para obtenção do título de Doutor em Agronomia.

Orientadores: Prof. Dr. Cícero Carlos de Souza Almeida
Dr. André Seco Marques da Silva

Rio Largo, AL
2019

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Setorial do Centro de Ciências Agrárias
Bibliotecária Responsável: Myrtes Vieira do Nascimento

M386a Martins, Gleica Maria Correia
Análises evolutivas em *Spondias* L.: uma abordagem genômica do gênero – 2019.
103 f.; il.

Tese (Doutorado em Agronomia: Produção Vegetal) - Universidade Federal de Alagoas, Centro de Ciências Agrárias. Rio Largo, 2019.
Orientação: Prof. Dr. Cícero Carlos de Souza Almeida
Dr. André Seco Marques da Silva

Inclui bibliografia

1. Mitogenoma. 2. Fração repetida. 3. Evolução genética

I. Título

CDU: 57

TERMO DE APROVAÇÃO

GLEICA MARIA CORREIA MARTINS
(Matrícula 16140032)

"Análises evolutivas em *Spondias* L.: uma abordagem genômica do gênero"

Tese apresentada e avaliada pela banca examinadora em cinco de abril de 2019, como parte dos requisitos para obtenção do título de Doutora em Agronomia, área de concentração em Produção Vegetal do Programa de Pós-graduação em Agronomia (Produção Vegetal) da Unidade Acadêmica Centro de Ciências Agrárias da UNIVERSDADE FEDERAL DE ALAGOAS.

Prof. Dr. Cícero Carlos de Souza Almeida
Presidente

Prof. Dr. André Seco Marques da Silva
Membro

Prof. Dr. Luiz Gustavo Rodrigues Souza
Membro

Prof. Dr. Leonardo Bruetto
Membro

Rio Largo - AL
Abril-2019

DEDICATÓRIA

Dedico este trabalho a minha família, em especial minha avó Ilva Isidoro do Bonfim Correia (*in memorian*) que sempre me ajudou, motivou e acreditou em mim.

AGRADECIMENTOS

À Deus, por toda força, luz e sabedoria e por sempre dar condições para que eu jamais desistisse dos meus objetivos;

Ao meu orientador prof. Dr. *Cícero Carlos Almeida*, por todos os ensinamentos, e pela motivação para sempre seguir em buscar de novos conhecimentos e me fazendo entender o verdadeiro sentido de ser um bom pesquisador;

Ao meu co-orientador Prof. Dr. *André Marques*, pelas ideias e contribuições científicas para realização deste trabalho;

Aos meus queridos pais (*Maria José* e *Paulo César*) e meus irmãos (*Carla Maria*, *Paulo Filho* e *Natalie Maria*) por todo suporte, compreensão e motivação de sempre, jamais teria conseguido cehgar aqui sem o apoio de vocês;

A equipe LARGE, em especial: *Eliane Balbino*, *Elvia Jéssica*, *Graziele Clemente* e *Jack Terto e Suzyanne Morais*, por todo apoio, motivação, alegria, parceria e ensinamentos, que foram essenciais para meu amadurecimento profissional.

Aos meus queridos amigos, *Dailson Oliveira*, *Juliany Barros* e *Renato Carvalho*, por toda parceria, pelo apoio e pelos momentos infinitos de aprendizado e alegrias, que deixaram tudo mais leve durante o Doutorado. Admiro demais vcs e sou muita grata pela amizade de vocês!

Ao meu namorado, amigo e parceiro, *Ricardo Barros*, por todo carinho, compreensão e apoio, que foram essenciais para finalização deste trabalho.

À minha querida amiga e Comadre *Edilma Silva* por todo carinho, compreensão e motivação;

À minha querida amiga *Beatriz Caetano*, pelas conversas, pelo carinho, força e apoio, sempre me fazendo acreditar no meu potencial;

A todos os professores do Programa de Pós-Graduação em Agronomia (Produção Vegetal) por todos os momentos de aprendizagem e pela disposição em solucionar as dúvidas;

Ao secretário do PPGA, *Gustavo Nepomuceno*, por toda competência e disponibilidade de sempre.

A todos aqueles que apesar de não terem sido citados, contribuíram para realização deste trabalho.

Muito Obrigada!

RESUMO

O gênero *Spondias* é um grupo de árvores frutíferas da família Anacardiaceae que compreende espécies de importância econômica, com distribuição nas regiões neotropiciais. Estudos de caracterização genômica no gênero tem potencial de fornecer importantes informações para a compreensão da origem e evolução das espécies. Nessa perspectiva, neste estudo objetivou-se sequenciar os mitogenomas, avaliar o conteúdo de DNA e caracterizar a fração repetitiva do genoma de espécies do gênero *Spondias*, a fim de descrever a composição das sequências repetitivas nos genomas e estabelecer relações evoluitvas com a fiogenia no gênero *Spondias.* Para isso foi feita extração de DNA de *S. tuberosa*, *S. mombim*, *S. bahiensis*, *Spondias* sp.(*S. tuberosa* x *S.bahiensis*) e *Spondias dulcis* e preparadas bibliotecas de *single-end e paired-end reads* e feita análises de bioinformática em softwares para montagem de genomas mitocondriais e caracterização da fração repetitiva dos genomas, além de quantificação do conteúdo de DNA genômico usando citometria de fluxo. Os mitogenomes apresentaram 779,106 bp e 674,156 bp, para *S. tuberosa* e *S. mombin*, exibindo um total de 74 genes e 68 genes, respectivamente, com estrutura com muitos rearranjos mediados pela presença de DNA repetitivo e alta incorporação de DNA do cloroplasto. A análise do conteúdo de DNA e da fração repetitiva do genoma demonstraram que os tamanhos dos genomas de *Spondias* analisados variaram de 460 a 530 Mb e a análise filogenética revelou uma tendência ao aumento no tamanho do genoma em espécies mais diversificadas. As análises de composição do genoma resultaram na identificação e caracterização do DNA repetitivo, perfazendo aproximadamente 35,57% em *S. tuberosa*, 33,43% em *S. mombim*, 34,60% em *Spondias sp*, 31,02% *S. bahiensis e* 49,66% em *S.dulcis.* Para satDNA foram observados dois satélites principais distribuídos em todas as espécies, sendo o SatSpo1 com alta diversificação no gênero. Além disso, para *S. tuberosa* foi detectado um satélite no espaçador intergênico de rDNA. A hibridização *in situ* do satDNA (SatSpo1 e SatSpo2) evidenciou a baixa diversificação de DNA repetitivo, a presença de satélites ricos em GC que provavelmente estaõ distrubuídos em regiões de heterocromatina. A partir disso, conclui-se que o gênero *Spondias* tem uma diversificação recente do DNA repetitivo, com distribuição similar de TEs e famílias de satDNA entre as espécies.

Palavras-chave: Mitogenoma. Fração repetitiva. Evolução. Tamanho do genoma.

ABSTRACT

The genus Spondias is a group of fruit trees of the family Anacardiaceae that includes species of economic importance, with distribution in the neotropical regions. Studies of genomic characterization in the genus have the potential to provide important information for the understanding of the origin and evolution of the species. In this perspective, the objective of this study was to sequence the mitogenomas, to evaluate the DNA content and to characterize the repetitive fraction of the genome of species of the genus Spondias, in order to describe the composition of the repetitive sequences in the genomes and to establish evolutionary relationships with foligenia in the genus Spondias. For this, extraction of S. tuberosa, S. mombim, S. bahiensis, Spondias sp. and Spondias dulcis and prepared single-end and paired-end reads libraries and performed bioinformatics analyzes on computational software for assembly of mitochondrial genomes and characterization of the repetitive fraction of genomes, as well as quantification of genomic DNA content using flow cytometry. The mitogenomes presented 779,106 bp and 674,156 bp, for S. tuberosa and S. mombin, exhibiting a total of 74 genes and 68 genes, respectively, with structure with many rearrangements mediated by the presence of repetitive DNA and high incorporation of chloroplast DNA. Analysis of the DNA content and the repetitive fraction of the genome demonstrated that the sizes of the Spondias genomes analyzed ranged from 460 to 530 Mb and phylogenetic analysis revealed a tendency to increase the size of the genome in derived species. The analyzes in the Repeat Explorer resulted in the identification and characterization of the repetitive DNA, making up approximately 15.11% in S. tuberosa, 22.41% in S. mombim, 18.21% in Spondias sp, 14.12% S. bahiensis and 29.65% in S.dulcis. For SatDNA, two main satellites distributed in all species were observed, with SatSpo1 being highly diversified in the genus. In addition, for S. tuberosa a satellite was detected in the intergenic spacer of rDNA. In situ genomic hybridization of satDNA (SatSpo1 and SatSpo2) evidenced the low repetitive DNA diversification, the presence of GC-rich satellites that are likely to be distributed in heterochromatin regions. From this, it is concluded that the genus Spondias has a recent diversification to repetitive DNA, with similar distribution of TEs and families of satDNA among the species.

Keywords: Genomic evolution. Mitogenoma. Repetitive fraction.

LISTA DE ILUSTRAÇÕES

REVISÃO DE LITERATURA

ARTIGO I

ARTIGO II

LISTA DE TABELAS

ARTIGO I

ARTIGO II

LISTA DE ABREVIATURAS E SIGLAS

AFLP- Amplified Fragment Length Polymorphism

BLAST- *Basic Local Alignment Search Tool*

cp- Cloroplasto

CTAB- Cetyl trimethylammonium bromide

EST- expressed sequence tag

LTR- Long terminal repeats

mt- Mitocôndria

NCBI- National Center for Biotechnology Information

NGS- sequenciamento de Nova Geração

ORF- *open reading frame*

RFLP- Restriction Fragment Length Polymorphism

Sat- satélite

SNP-Polimorfismo em nucleotídeo único

SSR- Simple Sequence Repeats

TE-elementos transponíveis

SUMÁRIO

\

1. INTRODUÇÃO

O tamanho do genoma é um traço de biodiversidade que mostra diversidade notável em eucariotos, variando mais de 64.000 vezes (PELLICER et al., 2018). De todos os principais grupos taxonômicos, as plantas terrestres se destacam devido à sua enorme diversidade em tamanho do genoma e está se tornando cada vez mais evidente que esta característica não desempenha apenas um papel importante na evolução dos genomas das plantas, como também pode influenciar as espécies no nível de ecossistema. Os avanços recentes e as melhorias nas novas tecnologias de sequenciamento, bem como ferramentas analíticas, tornam possíveis *insights* críticos sobre os mecanismos genômicos e epigenéticos que sustentam as mudanças no tamanho do genoma, os efeitos sobre a trajetória evolutiva de uma planta e sua capacidade de responder às mudanças ambientais (KELLY et al., 2012; PELLICER et al., 2018). A partir disso emerge a necessidade de que os estudos de caracterização genômica se tornem frequentes, no intuito de fornecer bases de dados para inferências sobre a dinâmica e evolução genômica nas espécies.

Embora as plantas com flores (as angiospermas) variem significativamente no conteúdo de DNA nuclear, a maior parte dessa variação não é associada a diferenças no número de genes ou no tamanho do gene (BENNETZEN et al., 2005). Os principais mecanismos que contribuem para a expansão do genoma em plantas são a poliploidia e o acúmulo de sequências repetitivas de DNA que compõem a maior parte o genoma, combinada com baixas taxas de remoção de DNA (HIDALGO et al., 2017; PELLICER et al., 2010; SCHUBERT; VU, 2016).

As sequências repetitivas podem dar uma grande contribuição para a evolução do genoma em uma variedade de níveis. Isto inclui a organização estrutural e ampla composição do genoma, efeitos epigenéticos, e também a regulação mais apurada da expressão gênica e do espaço gênico (DODSWORTH et al., 2015).

Os genomas modernos das plantas derivam de processos iniciados por uma história de repetidos eventos de duplicação do genoma completo, em que a variação extraordinária no tamanho do genoma através das espécies de plantas reflete em grande parte as diferenças na proliferação e sobrevivência de várias classes e famílias de elementos transponíveis (TEs), e eventos regulatórios mediados por pequenos RNAs. Esses eventos são todos moldados por interações bióticas e abióticas mais complexas

entre os organismos e seus ambientes (BENNETZEN et al., 2005; PANAUD et al., 2014; VERDE et al., 2013; WENDEL et al., 2016).

Além do DNA nuclear, as plantas apresentam genomas organelares no cloroplasto (cp) e na mitocôndria (mt). Esses genomas apesar de ainda serem considerados complexos em estrutura e organização, especialmente o da mitocôndria, eles vêm sendo comumente utilizados em estudos evolutivos para inferir sobre eventos da história evolutiva de grupos (DONG et al., 2018; OLMSTEAD; PALMER,1994; ZANG et al., 2012). A evolução lenta do cpDNA indica que estes são mais adequados para estudos filogenéticos para uma gama de espécies e o mtDNA por ter uma rápida mudança em sua estrutura, tem facilitado bastante a compreensão dos eventos evolutivos de espécies intimamente relacionadas (NAITO et al., 2013).

Muitos estudos têm se voltado para caracterização e comparação dos genomas dessas duas organelas (DONNELLY et al., 2016; GUI et al., 2016; KERSTEN et al., 2016; NAITO et al., 2013; PARK et al., 2014; STRAUB, 2013; ZHANG et al., 2012), o que possibilita discutir sobre os padrões de organização dos genomas organelares dentro do panorama evolutivo das espécies. Nesse sentido, o conhecimento obtido em estudos sobre a composição e organização do genoma eucariótico é importante para entender como os genomas funcionam e evoluem (KELLY et al., 2015; MACAS et al., 2015; ROBLEDILLO et al 2018; VAN-LUME et al., 2017), o que fornece base para projetos de estratégias de manipulação de genomas sobretudo em espécies de grande utilização pelas populações que sofrem grande pressão extrativista ou que estão com *status* de consevação ameaçado.

Dentre os gêneros de grande importância, *Spondias* destaca-se por ser um grupo de árvores frutíferas que se destaca por apresentar elevada importância econômica para as comunidades do semiárido nordestino (ALBUQUERQUE et al., 2005; ALMEIDA, et al., 2010; SANTOS et al., 2010; SILVA et al, 2014). Dentre as espécies deste gênero, o umbu (*S. tuberosa*), o cajá (*S. mombim*), a siriguela (*S. purpurea*) e a cajarana (*S. dulcis* L.) apresentam ampla distribuição no Brasil, sendo o umbu considerado nativa desta região. Estas espécies são comumente relatadas como envolvidas em processos de hibridização natural (MACHADO et al. 2015; MITCHAEL; DALY, 2015), um exemplo muito relatado é o umbucajá (*S. banhiensis*), que é um hibrido muito dissemiando pleo nordeste do Brasil resultante do cruzamento de *Spondias tuberosa* (NOBRE etl al, 2018), com outra *Spondias* até etão não esclarecida . Os frutos das

Spondias são utilizados para diversas finalidades (LIRA JUNIOR, 2005), inclusive com finalidade medicinal (SILVA et al, 2014)..

Estudos de genômica no grupo ainda são muito escassos, e não se tem publicações que abordem a taxonomia e filogenia do grupo de maneira clara. Até então não foram registradas dados de montagens de genomas mitocondriais, nem de caracterização do genoma nuclear para espécies do gênero, se encontram sequenciados apenas o genoma plastidial de três espécies de *Spondias* (SANTOS; ALMEIDA, 2018). Nessa perspectiva, se faz necessário o desenvolvimento de estudos para cacaterização de mais genomas, a fim de possibilitar o esclarecimento da origem híbrida de algumas espécies, e estabelecer filogenias mais claras para o gênero.

Partindo da hipótese que o sequenciamento de genomas mitocondriais, a análise do conteúdo de DNA e a caracterização das sequências repetitivas possibilita compreender a organização de rearranjos do genoma, estabelecendo relações sobre a evolução estrutural de genomas do grupo. O presente trabalho objetivou sequenciar mitogenomas, avaliar o conteúdo de DNA e caracterizar a fração repetitiva do genoma de espécies do gênero *Spondias*, no intuito de esclarecer as relações filogenéticas e os aspectos evolutivos no gênero.

2. REVISÃO BIBLIOGRÁFICA

2.1.Gênero *Spondias*

2.1.1. Taxonomia e filogenia

O gênero Spondias L. destaca-se como um dos primeiros gêneros de *Anacardiaceae* descritos por Linnaeus (1737: 365), com a espécie *Spondias mombin* tendo sido publicada em 1753 (MITCHELL; DALY, 2015). Atualmente esta família compreende 83 gêneros e aproximadamente 860 espécies (CHRISTENHUSZ; BYNG, 2017) que se destacam pela presença de frutos comestíveis.

Spondias L. é um gênero de árvores frutíferas, típico da subfamília Spondioideae (MITCHELL et al., 2006; PELL, 2004; PELL et al., 2010), que compreende 20 espécies nativas da América tropical, Ásia e Madagascar. Destas espécies, dez são nativas do Neotrópico, distribuídas do México ao sul do Brasil, uma nativa de Madagascar e sete são nativas da Ásia e do Pacífico Sul, da Malásia (sensu *Flora Malesiana*) para a China tropical, Sri Lanka Indochina, Tailândia, Índia (exceto o extremo norte), Myanmar (Birmânia), Ilhas Salomão a leste da Polinésia. *Spondias dulcis* é cultivada na América tropical e nas Antilhas. *Spondias mombin* e *Spondias purpurea* são ambos introduzidos em toda a África Ocidental Tropical e na Ásia (e nas Índias Ocidentais, onde não são encontradas na vegetação primária e podem não ser nativas); e *Spondias mombin* é frequentemente relatada na África Ocidental Tropical (MACHADO et al., 2015; MITCHEL; DALY, 2015).

No Brasil, o gênero *Spondias* é representado por espécies de distribuição nas regiões Norte e Nordeste, entre elas destacam-se: *S. mombin* L. (Cajazeira), *S. purpurea* L. (Sirigueleira), *S. dulcis* Sonn. (Cajaraneira), *S. tuberosa* Arr. Cam (Umbuzeiro), *Spondias venulosa* Mart. Ex Engl (cajazinho), e *Spondias bahiensis* (umbu-cajá) (LIRA JUNIOR, 2005; SILVA et al., 2015; SOUZA et al., 2006). *Spondias mombin* e *S. dulcis* ocorrem principalmente em áreas próximas ao litoral, assim como *S. purpurea*, que também pode ser encontrada em perímetros irrigados com significativa produtividade. *Spondias tuberosa* é a espécie mais característica do semiárido nordestino, seguida pelo umbu-cajazeira (*S. bahiensis*), considerada um híbrido natural entre *S. tuberosa* e *S. mombim* (GIACOMETTI, 1993; SILVA JÚNIOR et al., 2004).

A taxonomia do gênero não é clara, com a presença de supostos híbridos e espécies com pouca diferenciação morfológica, que aos poucos vem sendo estudadas a fim de esclarecer sua origem e as relacões taxonômicas (ALMEIDA et al., 2007; MACHADO et al., 2015; SILVA, 2015).

A ocorrência de espécies com características intermediárias entre as espécies já descritas de *Spondias* vem sendo relatadas em vários estudos que tentam determinar a existência de híbridos no gênero. Machado et al. (2015) descreveram uma nova espécie conhecida com o híbrido umbucajá (*S. banhiensis*), que é uma junção dos nomes umbu (*S. tuberosa*) e cajá (*S. mombin*), relatado pelos moradores da região em que são encontradas, como um híbrido natural dessas duas espécies. No entanto, estudos recentes sugerem que *S. tuberosa* é um dos parentais desse híbrido, mas *S. mombin* não apresenta relação parental com essa espécie de híbrido (NOBRE et al., 2018).

Outro possível híbrido é a umbuguela (*Spondias* sp.), que tem o nome formado pela junção de umbu e siriguela (SOUZA,1998). Além destes, outros potenciais híbridos vem sendo discutidos em vários relatos e publicações (MITCHAEL; DALY, 2015; SANTOS, 2016). No entanto, ainda há certa confusão sobre a determinação exata de quais fenótipos são realmente resultantes de hibridação ou se são espécies ainda não descritas, visto que ainda não se tem evidências genéticas sobre muitas dessas espécies, que possam explicar sua origem e classificação.

A existência desses possíveis híbridos vem dificultando o desenvolvimento de estudos que possam revelar a relação filogenética do grupo. As principais contribuições foram sistematizadas a partir de uma revisão organizada por Mitchell e Dally (1998) que apresentaram uma chave para classificação de espécies neotropicais do gênero *Spondias* baseada em características morfológicas, fornecendo informações para classificação e diferenciação das espécies: *S. purpurea L., S. dulcis* Parkinson, *S. tuberosa* Arruda, *S, radlkoferi* Donn. Sm., *S. macrocarpa* Engl., *S. testudinis* J. D. Mitch. & Daly, *S. venulosa* Mart. ex Engl., e *S. mombin* L. complex. No entanto, o trabalho não aponta muitas evidências moleculares dessa classificação, sendo ainda necessários estudos para esclarecer melhor as relações filogenéticas do gênero.

Santos e Oliveira (2008) em estudo para avaliar as inter-relações genéticas entre espécies do gênero *Spondias* com base em marcadores AFLP, construiu um fenograma, que revelou algumas evidências sobre as relações filogenéticas entre seis espécies deste gênero, e reuniu indícios sobre a ocorrência de híbridos naturais no gênero. As análises obtidas por eles demonstraram que todos os indivíduos de *S. dulcis, S. tuberosa* e *S.*

purpurea foram posicionados em clados monofiléticos, enquanto que indivíduos de *S. mombin*, *Spondias* sp. (umbu-cajá) e *Spondias* sp. (umbuguela) não formaram grupos-espécie específicos. Os autores sugeriram que a posição do umbu-cajá entre umbuzeiro e cajazeira, e a similaridade de sequência em torno de 50% podem indicar que este material pode ser um híbrido das duas espécies, como indicado no seu nome. A posição da umbuguela entre umbuzeiro e cajazeira, com similaridade em torno de 60% sugerem que a umbuguela possa ser um híbrido dessas duas espécies. O fenograma de AFLP mostra que *S. dulcis* foi a espécie mais divergente dentro as 6 analisadas do gênero *Spondias,* podendo ser considerada uma espécie mais basal.

Estudos recentes desenvolvidos indicam que o gênero *Spondias* possui diversificação há aproximadamneto 20 Ma, com uma filogenia contendo espécies muito próximas, mostrando que as espécies *S. tuberosa* e *S. venulosa* são as espécies mais diversificadas no gênero, enquanto *S. dulcis* destaca-se como uma espécie mais basal (MACHADO et. al., 2015; SILVA et. al, 2015). Uma filogenia baseada em plastomas completos corrobora com a diversificação das espécies do gênero, mpstarndo *S. dulcis* como a primeira espécie a se divergir no grupo (Figura 1)

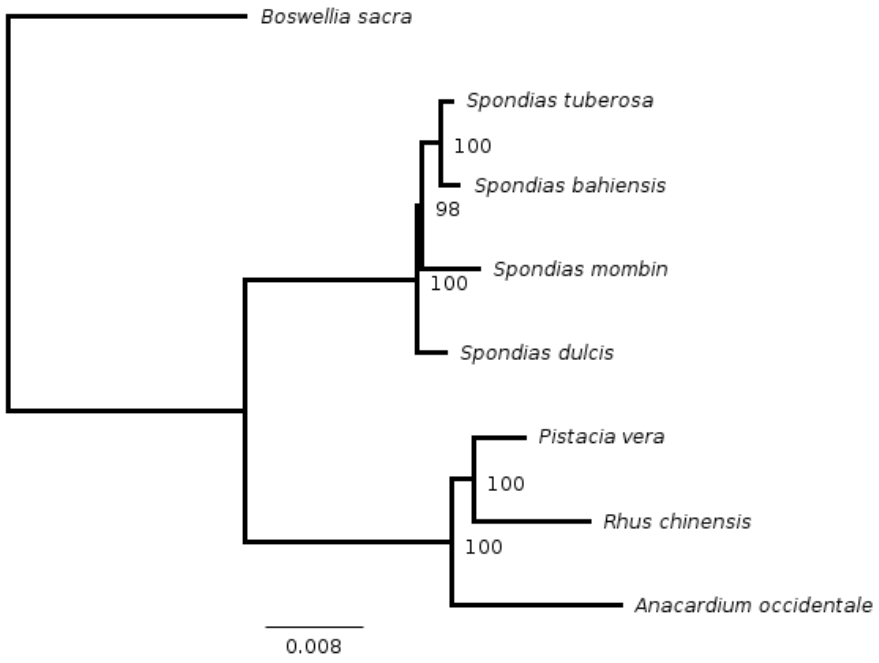


Figura 1. Filogenia de plastomas completos para família Anacardiaceae (Dados não publicados).

Apesar do considerável número de estudos filogenéticos para o gênero que já vem sendo desenvolvidos (ALMEIDA et al., 2007; MACHADO et al., 2015; PELL, 2004; SANTOS; ALMEIDA, 2019; SANTOS; OLIVEIRA, 2008; SILVA et al., 2015).,

ainda necessita-se de estudos mais conclusivos que possam esclarecer sua filogenia. Estudos da genômica do grupo tem o potencial para resolver conflitos sobre a evolução do gênero e esclarecer a possível origem híbrida de indivíduos relatados dentro deste gênero, o que pode assegurar a correta identificação das espécies de *Spondias*, já que a identificação pelo uso de raízes, sementes, pólen ou uma mistura de partes de plantas torna-se muito difícil usando características fenotípicas (CBOL, 2009). Estudos voltados para caracterização molecular e esclarecimento das relações filogéneticas foram desenvolvidos por Silva et al. (2015), no entanto ainda são necessários estudos mais frequentes para suplantar as discussões.

Mitchell e Daly (2015) em estudo ressaltaram a dificuldade em purificar e amplificar o DNA, mesmo a partir de amostras de folhas frescas de *Spondias* (S. Pell e A. Miller, com. Pers.), o que dificulta a condução de estudos que possam estabelecer de maneira mais clara por meio de eventos moleculares as relações de filogenia desse grupo. Neste sentido, se faz necessário o desenvolvimento de trabalhos voltados para caracterização molecular e bioquímica de espécies do gênero, a fim de esclarecer processos fisiológicos e estabelecer protocolos que possam otimizar os procedimentos moleculares e facilitar o desenvolvimento de pesquisas com estas espécies.

2.1.2. Importância econômica

Spondias tem uma história de uso que surgiu pelo menos até 6500 a.c., no Tehuacan Vale do México (SMITH, 1967), em que quatro espécies de Spondias eram reconhecidas como economicamente importantes na América tropical: *S. dulcis*, *S. mombin*, *S. purpurea* e *S. tuberosa* (MITCHELL; DALY, 2015).

No Brasil, o gênero Spondias é representado por várias espécies entre elas destacam-se: Cajazeira, Sirigueleira, Cajaraneira, Umbuzeiro e Umbu-cajá, espécies utilizadas popularmente no preparo de sucos, doces, picolés e sorvetes a partir do fruto, por apresentarem odor agradável e sabor agridoce (LIRA JUNIOR, 2005), se destacando com elevada importância econômica e potencial de uso para populações nordestinas.

Todas as espécies deste gênero tem um mesocarpo comestível e alguns são altamente valorizados (MITCHELL et al., 2012). Quase todas as espécies de *Spondias* tem um endocarpo fibroso (MILLER; SCHAAL, 2005), sendo algumas espécies

consideradas com grande relevância econômica e social para as comunidades de ambientes áridos.

No Nordeste do Brasil, algumas *Spondias* são cultivadas em fundos de quintais ou em pequenos pomares. Entre seus principais usos, podem ser citados: madeira, lenha/carvão, alimentação humana, medicina caseira, higiene corporal, ornamental, criação de abelhas, forragem e sombreamento (MAIA, 2004). Os frutos das espécies deste gênero são consumidos *in natura*, vendidos em mercados locais ou nas margens de algumas rodovias brasileiras, ou processados na forma de polpas, sucos e outros produtos alimentícios (ALMEIDA, et al., 2010; SANTOS; OLIVEIRA, 2008; SANTOS et al., 2010; SILVA et al, 2014).

Entre as espécies cultivadas da região semiárida do Nordeste brasileiro, as *Spondias* destacam-se pela possibilidade de serem cultivadas em larga escala, podendo ser aproveitadas tanto para alimentação humana quanto para a suplementação alimentar de animais, especialmente caprinos e ovinos, que constituem os rebanhos predominantes nessa região (ALBUQUERQUE et al., 2005; CAVALCANTI et al., 2000).

Além disso, as sementes têm potencial como fonte de óleo de cozinha devido ao alto teor de óleo, alta concentração de minerais e composição de ácidos graxos (BORGES et al., 2007). As raízes têm sido usadas como alimento de fome em tempos de seca (NASCIMENTO et al., 2012).

No Brasil, notadamente no Nordeste, estas espécies têm considerável importância social e econômica, fato comprovado pela crescente comercialização de seus frutos e produtos processados em mercados, supermercados e restaurantes da região (SOUZA, 1998). Segundo estimativa de Santos (1999) o mercado brasileiro de algumas *Spondias* pode ficar em torno de cerca de seis milhões de dólares por ano, incluindo colheita, venda e transformação de frutos resultantes do extrativismo. Em algumas regiões do Estado da Bahia, o comércio de frutas frescas e processadas está em rápida expansão, com muitas famílias de pequenos produtores e/ou assalariados agrícolas envolvidos (SANTOS; OLIVEIRA, 2008). Em Alagoas, em 2013 foram registrados 11 municípios produtores de umbu, com uma produção anual de 32 toneladas, sendo considerado um dos estados com menor produtividade da região nordeste (BATISTA et al., 2015).

Outra característica que revela a importância econômica de espécies deste gênero são as propriedades medicinais evidenciadas em alguns estudos científicos que

relatam o uso de *Spondias* por pessoas de várias partes do mundo para tratar doenças (SILVA et al, 2014). Em alguns locais as folhas são usadas para banhos medicinais e chás da casca são usados para tratar resfriados e disenteria. As espécies podem ser uma adição valiosa à flora e dieta econômicas de muitas outras regiões tropicais secas (MITCHELL; DALY, 2015).

As propriedades farmacológicas atribuídas a algumas espécies do grupo são devido em grande parte aos compostos fenólicos (taninos e flavonoides), na maioria, presentes nas folhas. No entanto, outros metabólitos secundários também podem contribuir para essas atividades, pois vitamina C, saponinas, alcaloides, terpenos e carotenoides têm sido identificados nessas espécies (SILVA et al, 2014).

Nos últimos anos, descobriu-se que o extrato das folhas e dos ramos de algumas *Spondias* continham taninos elágicos com propriedades medicinais para o controle de bactérias gram negativas e positivas (AJAO et al., 1985), do vírus da herpes simples (CORTHOUT et al., 1994) e da herpes dolorosa (MATOS, 1994); inclusive já existem produtos à base do extrato das folhas e dos ramos da cajazeira, industrializados e comercializados na cidade de Fortaleza, CE, para combate à herpes labial (SOUZA, 1998).

Recentemente, os compostos purificados de algumas *Spondias* foram testados para uma variedade de atividade biológica em animais de laboratório, incluindo antifúngicos, antimicrobianos (CORTHOUT et al., 1994; RODRIGUES; HASSE, 2000), antihelmínticos (ADEMOLA et al., 2005), anti-virais e propriedades psicoativas (AYOKA et al., 2006, 2008).

A madeira de *Spondias* pode ser usada como um material combustível, mas é de má qualidade porque é suscetível a podridão e ataque por insetos (TER WELLE et al., 1997). Mesmo assim, é usado ocasionalmente para construção, carpintaria e postes de fenda (AYOKA et al., 2008). A casca grossa é esculpida para fazer artesanato.

De acordo com Batista et al (2015) apesar da importância das fruteiras de *Spondias*, sobretudo do umbuzeiro e do seu elevado potencial sócio-econômico, poucos estudos têm sido realizados visando aumentar a base de informações e ampliar suas possibilidades de uso. Os frutos do umbuzeiro apresentam apelo exótico para mercados de outras regiões do Brasil como sul e sudeste, e também para o mercado externo, o que de certa forma pode incentivar o aumento da produção. Ainda não devidamente caracterizado, particularmente no que se refere ao seu potencial agroindustrial, o umbu é uma fruta que demanda pesquisas, principalmente adequação de tecnologias

convencionais e desenvolvimento de novas, voltadas para o processamento, de forma a promover um aproveitamento mais rentável, mediante agregação de valor ao produto.

2.2.*GENÔMICA VEGETAL*

A célula vegetal contém três diferentes genomas: o do cloroplasto, o mitocondrial e o nuclear. Os genomas do cloroplasto e mitocondrial são de herança uniparental (normalmente materna em angiospermas) e o genoma nuclear é biparental. Os três genomas diferem grandemente em tamanho, sendo o nuclear o maior – medido em megabases de DNA. O genoma mitocondrial inclui centenas de kilobases (Kb) de DNA (200-2.500 Kb), o que o torna pequeno com relação ao genoma nuclear, mas muito grande com relação ao genoma mitocondrial de animais (o qual tende a apresentar cerca de 16 Kb) (JANSEN et al., 2005; LANG et al., 2012). O genoma do cloroplasto é o menor dos três genomas, variando, na maioria das plantas de 120–200 Kb (SMITH, 2017).

O conhecimento obtido em estudos sobre a organização do genoma eucariótico é importante para entender como os genomas funcionam e evoluem (LAPITAN, 1992). Para espécies do gênero *Spondias* já se encontram sequenciados e montados o genoma plastidial de *S. tuberosa, S. bahiensis,* e *S. mombin* (SANTOS; ALMEIDA, 2018.). Ainda não foram registrados montagens de genomas mitocondriais, nem caracterização de genoma nuclear no gênero, o que evidencia a importância de desenvolver estudos para descrição de mais genomas, posto que a caracterização genômica ajuda a esclarecer a origem híbrida de algumas espécies dentro deste gênero, além de contribuir para reconstrução de filogenias no grupo.

2.2.1. Genoma Mitocondrial

As mitocôndrias são organelas semiautônomas envolvidas em processos celulares de produção de energia, desempenhando um papel significativo na produtividade e desenvolvimento das plantas (YASUNARI et al., 2005). Esta organela apresenta um sistema genético próprio, adquirido por endosimbiose de bactérias anteriormente vivas (GREINER; BOCK, 2013). Sua informação genética é transmitida pelos genes de maneira uniparental (BIRKY JÚNIOR, 1995).

Uma das características mais notáveis do genoma mitocondrial é a presença de DNA que poder ser adquirido do plastídio, núcleo e até mesmo material genético de outras espécies, incluindo bactérias, vírus e plantas (ALVERSON et al., 2010; KNOOP, 2004; PARK et al., 2014; SHEARMAN et al., 2016). Essa incorporação de DNA é relatada como um dos principais responsáveis pela expansão desse genoma em algumas espécies (HANDA, 2003).

Outros eventos que podem estar relacionados com a expansão do genoma mitocondrial são ao acúmulo de sequências repetidas, e a expansão dos íntrons (BULLERWELL; GRAY, 2004; TURMEL et al., 2003). A acumulação de sequências repetitivas nos genomas mitocondriais das plantas causam eventos de recombinação frequentes e rearranjos do genoma dentro de uma espécie, o que leva à geração de múltiplas fitas de DNA circulares com sequências sobrepostas e cópias de diferentes números (ALLEN et al., 2007; CHANG et al., 2011; GUO, 2017; PARK et al., 2014). Neste sentido, os genomas mitocondriais nas plantas superiores não são apenas maiores em tamanho, mas também contêm rearranjos estruturais devido a eventos homólogos de recombinação intra ou molecular (HANDA, 2003).

O tamanho e o número desses elementos repetitivos no genoma mitocondrial de plantas são importantes porque refletem os locais de recombinação intramolecular, o que, em última instância, constituem a base para grande parte da diversidade estrutural conhecida nos genomas mitocondriais de plantas (ALVERSON et al., 2011).

Apesar da variação em tamanho, o conteúdo dos genes mitocondriais permanece baixo e restrito a alguns polipeptídios necessários para a biogênese dos complexos da cadeia de fosforilação oxidativa, proteínas ribossômicas, RNA de transferência e RNAs ribossômicos (GUALBERTO et al., 2014). Essa alta conservação do conteúdo gênico apoia as evidências de que a origem da mitocôndria provavelmente só ocorreu uma única vez na história evolutiva dos organismos, pois essa conservação de conteúdo em uma ampla gama de organismos dificilmente seria explicada por convergência evolutiva (ARIAS et al. 2003).

Os genomas mitocondriais possuem uma organização dinâmica multipartida devido à recombinação associada a regiões repetidas (GOREMYKIN et al., 2009; OGIHARA et al., 2005*)*. A estrutura física dos genomas mitocondriais das plantas é geralmente representada por um círculo de DNA de cadeia dupla, chamado "cromossomo master", abrigando o conjunto completo de genes mitocondriais. Os mtDNA de plantas geralmente contêm sequências repetidas dispersas em todo o

genoma, que pode estar em orientação direta ou invertida. Eles podem emparelhar e recombinar, redistribuindo sequências e gerando moléculas de DNA circulares sub-genômicas (STUPAR et al., 2001; GUALBERTO et al., 2014).

A estrutura dos genomas mitocondriais das plantas superiores tem um número de características únicas. Considerando que a maioria dos animais possui mitogenoma de tamanho 15 e 17 kb, os genomas mitocondriais das plantas são muito maiores e diferem muito em tamanho, mesmo entre espécies muito próximas ou dentro de espécies. Eles geralmente variam entre 200 e 750 kb em angiospermas, mas com um enorme avanço em pares de base em algumas linhagens (ALLEN et al., 2007; GUALBERTO et al., 2014; KUBO, 2008).

Os mitogenomas são geralmente maiores do que os genomas plastidiais, variando de 200 Kb em *Brassica hirta* (PALM; HERBON, 1987) a 11.3 Mb em *Silene conica* (SLOAN et al., 2012). No entanto, os genes funcionais nestes genomas são bastantes conservados (BI et al., 2016). Apesar do seu tamanho relativamente grande, os genomas mitocondriais contêm menos genes do que os seus homólogos plastidiais; 37-83 genes diferentes, incluindo os codificadores de proteína, tRNA e rRNA genes (GUALBERTO et al., 2014; MOWER et al., 2012).

Em plantas os mitogenomas têm uma alta frequência de edição de RNA que contribui para a conservação funcional das proteínas (GUALBERTO et al., 1989; HIESEL et al., 1994). Em plantas, a edição de RNA afeta transcritos mitocondriais e plastidiais por modificação específica de sítios (HIESEL et al., 1994; WU et al., 2015). A identificação de sítios de edição de RNA podem fornecer pistas importantes para prever funções gênicas (HANDA, 2003; YE et al., 2017;).

Os genes mitocondriais são separados por amplas regiões de DNA não-codificante (íntrons), que apresentam ordens variáveis no genoma. Rearranjos do genoma mitocondrial ocorrem tão frequentemente em indivíduos vegetais que eles não caracterizam ou diferenciam espécies ou grupos de espécies e, por consequência, não são utilizados para inferir relações filogenéticas (JUDD et al., 2009).

O emprego do mtDNA em estudos populacionais e evolutivos se vale principalmente do fato dele possuir uma baixa taxa de substituições de base, apresentar alterações no tamanho total da molécula devido a inserções e deleções, principalmente nas regiões ricas em A+T, e mais recentemente, do fato de que translocações de genes codificadores de tRNA parecem ser mais frequentes do que se imaginava, acarretando

assim em alterações na ordem gênica entre organismos filogeneticamente relacionados (ARIAS et al., 2003).

O mtDNA de plantas constitui um dos principais recursos para estudos evolutivos, devido a evolução lenta das regiões de codificação. Nesse sentido, a evolução estrutural e a plasticidade dos mtDNA tem potencial para torná-los um modelo poderoso para explorar as forças que afetam sua divergência e recombinação (WANG, et al., 2012). Atualmente a reconstrução de mitogenomas tem sido muito utilizado para fazer inferênciais sobre eventos de filogenia antigos.

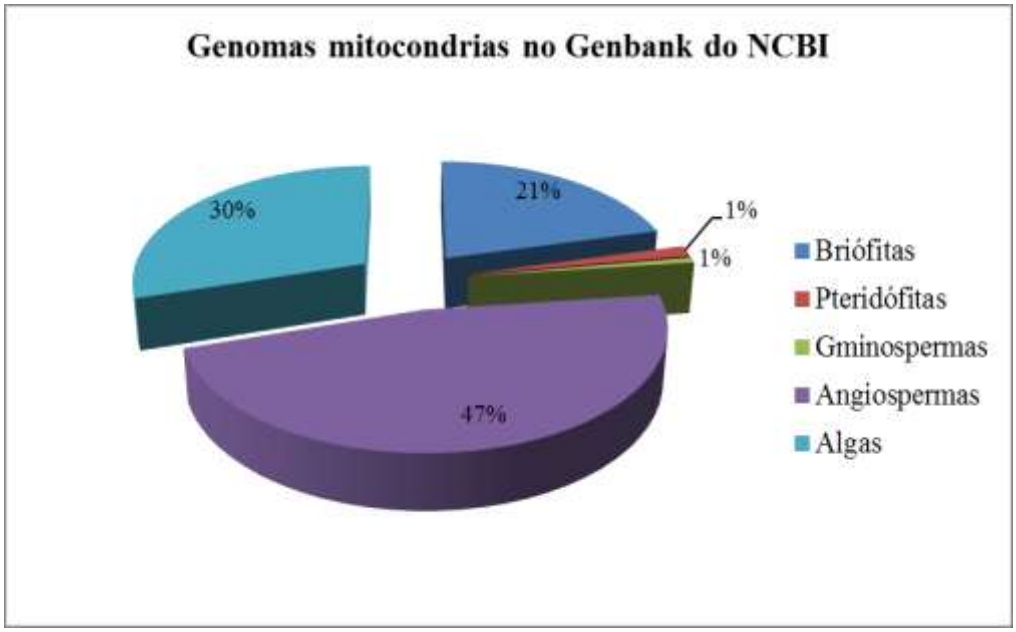
A análise e a comparação de características genômicas com outros genomas mt de plantas devem contribuir então para uma compreensão abrangente da evolução do mtDNA de plantas (YE et al., 2017), aumentando a compreensão acerca do rearranjo do genoma, dos mecanismos de transferência de DNA e da diversidade filogenética (WEI et al., 2016). Dessa forma, as variações significativas na organização, no conteúdo de genes e na edição de RNA do genoma mitocondrial são muito úteis para estudar a evolução das estruturas e sequências do genoma (GUI et al., 2016).

O mitogenoma é descrito como sendo de bastante complexidade, o que dificulta o desenvolvimento de alguns estudos de caracterização que exigem sequências com alta qualidade. De acordo com Shearman et al (2016) as principais dificuldades para a montagem de genomas mitocondriais de plantas se deve principalmente pela presença de sequências de DNA adquiridas do genomas nucleares e plastidiais, longas regiões de repetições que podem confundir os esforços de montagem justamente por introduzir pontos de ramificação que levam a montagem de múltiplas sequências quiméricas, incluindo sequências mitocondrial, nuclear ou de cloroplasto. Portanto, esse compartilhamento de sequências, a natureza altamente repetitiva e o tamanho relativamente grande dos genomas mitocondriais das plantas são os principais fatores que dificultam sua montagem.

Já se encontram sequenciados no Genbank do NCBI 194 genomas de plantas completas, sendo a maioria destes representado por espécies de angiospermas (Figura 3). No Brasil, o primeiro genoma mitocondrial de uma planta nativa e endêmica só foi sequenciado e montado em 2017 (SILVA et al., 2017). Nessa perspectiva, tendo em vista a carência de pesquisas e a ausência de um padrão de referência para sequenciar e montar os genomas mitocondriais de plantas brasileiras, o sequenciamento do genoma de outras espécies, como as pertencentes ao gênero *Spondias* se faz muito necessário

com vista à obtenção de dados que possam subsidiar potenciais estudos de conservação da flora nativa.

Figura 2. Proporção de grupos de plantas com genomas mitocondriais depositadas no Genbank do NCBI[1].



Fonte: dados da pesquisa, 2018.

2.2.2. Genoma Nuclear

Nos eucariotos a maior parte do genoma está concentrada nos núcleos, cada um com o mesmo conteúdo de DNA. O DNA nuclear é dividido em segmentos separados fisicamente, cada um sendo uma longa dupla hélice. Um cromossomo individual contém uma dessas duplas hélices altamente espiralada. O conjunto de cromossomos em organismos da mesma espécie tem um número de cromossomos e aspecto característicos. O número de cromossomos no conjunto genômico básico é denominado número haploide (designado *n)*. Os organismos que contêm duas cópias completas do genoma em seus núcleos são denominados diploides. Em um organismo diploide, os dois membros de um par de cromossomos são denominados cromossomos homólogos ou, às vezes, apenas homólogos. As sequências de DNA de um par de cromossomos homólogos são praticamente as mesmas, embora frequentemente haja variação mínima na sequência de nucleotídeos (HESLOP-HARRISON; SCHIMIDT, 2012).

[1] Apesar das algas de acordo com os atuais sistemas de classificação não serem agrupadas dentro do grupo das plantas, no NCBI, ambas são representadas no mesmo grupo de dados.

Cada molécula de DNA cromossômico contém muitas regiões funcionais denominadas *gene*. Assim como nos demais organismos eucariotos, os genomas de plantas são compostos basicamente por moléculas de DNA associadas a proteínas, as quais formam unidades denominadas nucleossomos. Os nucleossomos por sua vez estão dispostos em intervalos de aproximadamente 200 pb, compactados em arranjos que formam a cromatina (HESLOP-HARRISON; SCHWARZACHER, 2011; MICHEL; JACKSON, 2013). A cromatina é organizada em cromossomos, não sendo conhecido outro padrão de organização até o momento. Cada cromossomo consiste em uma única molécula de DNA linear e ininterrupta de uma extremidade à outra. Além das proteínas que se ligam ao DNA, nos cromossomos também ocorrem proteínas associadas aos processos de expressão gênica, replicação e reparo (LEE; KIM, 2014).

Os genomas nucleares de plantas apresentam uma extensa variação estrutural em tamanho, número de cromossomo, número e disposição de genes e número de cópias de genoma por núcleo. Um resultado derivado principalmente de sua origem poliplóide frequente e da amplificação de retrotransposons (BENNETZEN, 2002; KELLOGG; BENNETZEN, 2004). Essa variação também é resultante da inserção de DNAs organelares inseridos de forma independente no DNA nuclear. Em análises no genoma de arroz, constatou-se que 0,20-0,24% e 0,18-0,19% do genoma nuclear é composto por inserções de DNA plastidial e mitocondrial, respectivamente (MATSUMOTO et al., 2005; RIBEIRO, 2016).

Muitos genes ao serem adquiridos pelo genoma nuclear passam a ser expressos corretamente neste compartimento, sendo seus produtos transportados para organela de origem. Isso acontece tendo em vista a adaptação dos genes transferidos, que se ajustam à composição das bases nucleares e ao uso de códons, e passam a ser mais semelhantes aos genes nucleares do que as seus respectivos organelares (OLIVER; MARTÍNEZ-ZAPATER, 1990).

Os genomas nucleares são em grande parte colineares entre espécies estreitamente relacionadas, mas observam-se mais rearranjos com o aumento da distância filogenética. No entanto, não é possível correlacionar a quantidade de rearranjo e o tempo decorrido, uma vez que a divergência não é perfeita. Os rearranjos do genoma ao alterar os padrões de expressão gênica, desencadeiam novas combinações de genomas (híbridos), que podem atuar como uma força motriz na evolução (JUDD et al., 2009; KELLOGG; BENNETZEN, 2004).

Os genomas nucleares nas plantas superiores são relativamente complexos com genes espalhados por múltiplos cromossomos. Além dos genes, os cromossomos de plantas contêm um excesso de DNA repetitivo. Mesmo nos genomas menores, estima-se que mais de 20% do DNA é composto por vários elementos repetitivos. Essas repetições incluem elementos transponíveis e vários tipos de simples repetições em tandem, incluindo DNA satélite e simples repetições de sequência (SSRs). Na maioria das angiospermas, elementos transponíveis, especialmente *long terminal repeats* (LTR) retrotransposons compreendem a grande maioria deste DNA repetitivo (KELLOGG; BENNETZEN, 2004; RIBEIRO, 2016).

O genoma das plantas é estudado por uma gama de técnicas de microscopia e moleculares. O DNA é empacotado em cromatina e pode ser observado por microscopia de luz em contraste de fase e após coloração, por microscopia de epifluorescênciaou por microscopia eletrônica. Seu tamanho pode ser estimado por citometria de fluxo e medido em valores de 1C(valor para genoma haploide). Sua estrutura pode ser analisada por técnicas bioquímicas e moleculares envolvendo restriçãopor digestão enzimática, reação em cadeia da polimerase (PCR), métodos químicos, eletroforese e hibridização (HESLOP-HARRISON; SCHIMIDT, 2012). Cada vez mais, a estrutura, organização e composição do genoma vegetal, são estudadas por sequenciamento de DNA e análises de bioinformática, com dezenas de softwares que se destinam a caracterização dos genomas, o que vêm possibilitando desvendar a estrutura e funcionamento dos elementos do genoma, e partir disso tem sido possível delinear a história evolutiva de grupos e planejar estratégias de manipulação de genomas.

2.2.2.1. Fração repetitiva

O termo "sequências repetitivas" refere-se a fragmentos de DNA homólogos que são presente em múltiplas cópias no genoma. As sequências de DNA repetitivo estão presentes em todas as plantas superiores e podem representar até 90% do tamanho do genoma em algumas espécies. Essas sequências repetitivas são consideradas responsáveis por gerar grandes diferenças entre genomas, que podem refletir distâncias evolutivas entre espécies (MEHROTRA; GOYAL, 2015). A origem dessas repetições não está clara, mas a característica que as torna úteis é que, em indivíduos diferentes, existem geralmente números diferentes de cópias (RAO et al., 2010), o que vem sendo

foco de estudos que buscam testar os elementos repetitivos como sinal filogenético para evolução de grupos (DODSWORTH et al., 2014; MACAS et al., 2015).

As sequências repetidas podem ser classificadas em duas famílias grandes, a dos elementos arrajandos em tandem (*tandem repeats*) e *repetições dispersas* com moderado ou alto grau de repetitividade (LÓPEZ- FLORES; GARRIDO- RAMOS, 2012). As repetições dispersas inclui todos os transposons, genes de tRNA, e pseudogenes, enquanto repetições em tandem inlcuem gene tandems, DNA ribossômico (rDNA) e DNA satélite. De acordo com o comprimento da unidade de repetição e o tamanho do arranjo, as sequências de DNA satélite podem ser classificadas em três grupos: (i) microsatélites, que são regiões com repetições de 2-6 pb e um tamanho de arranjo da ordem de 10-100 unidades, (ii) minisatélites com 6-100 pb (geralmente em torno de 15 pb) e (iii) DNA satélite (satDNA) com uma unidade de repetição variável que muitas vezes forma arranjos de até 100 Mb. O comprimento das sequências do monômero de satDNA variam de 150 a 400 pb na maioria das plantas e animais (RICHARD et al., 2008).

O DNA de satélite (satDNA) é uma classe de DNA repetitivo caracterizada por sua organização genômica em longas matrizes de unidades dispostas em série chamadas monômeros. Monômeros (unidades de repetição) são sequências repetidas em tandem, geralmente com mais de 200 nucleotídeos de comprimento e tipicamente organizadas em matrizes de centenas ou milhares de cópias que podem ocupar até vários megabases dentro dos genomas. SatDNAs são o principal componente da heterocromatina, podendo formar blocos que aparecem como cromocentros nucleares e bandas cromossômicas (LÓPEZ- FLORES; GARRIDO- RAMOS, 2012; ROBLEDILLO et al., 2018).

As sequências satDNA se encontram principalmente em regiões centroméricas e subteloméricas nos cromossomos, mas também em posições intercaladas. Essas sequências de DNA satélite são consideradas como um componente que evolui rapidamente em genomas eucariotos, e por isso são muito utilizados em estudos de filogenia (MEHROTRA; GOYAL, 2014).

Os microssatélites são considerados uma ferramenta muito útil para estimativa dos níveis de variabilidade genética dentro das populações e permite fazer análises das relações genéticas entre elas. Estes dados são muito úteis para estimar a diversidade genética e endogamia em populações. Assim, os microssatélites fornecem informações que possibilita estimar as distâncias genéticas entre populações ou entre indivíduos, e a

filogenia como bem como a análise estrutural de cada população (LÓPEZ- FLORES; GARRIDO- RAMOS, 2012).

Outra classe de elementos repetitivos muito relatada são os elementos transponíveis, que tem sequências reconhecidas pela estrutura característica, que tipicamente inclui repetições terminais em ambas as extremidades com capacidade de se mover em um cromossomo de uma posição para outra dentro do mesmo genoma.Os elementos transponíveis (TEs) incluem vários grupos representados especialmente por retrotransposons LTR, não-LTR retrotransposons e transposons de DNA (WICKER et al. 2007).

As várias famílias de elementos transponíveis em uma espécie podem representar 10% ou mais do DNA total e podem ter efeito na repressão ou na expressão de genes de acordo com as condições. Muitos tipos de elementos transponíveis são duplicados por transcrição reversa, sendo denominados Retrotransposons (BALTIMORE, 1985). Tal como os retrovírus, com os quais compartilham características estruturais semelhantes, esses elementos duplicam-se pela transcrição da sua sequência de DNA inserida em RNA. Esta molécula é então transcrita reversamente em DNA (denominado cDNA) pela ação da enzima transcriptase reversa. Em geral, os TEs afetam o genoma pela sua capacidade de se mover e replicar, gerando assim plasticidade no genoma (WICKER et al., 2007).

Os pseudogenes apresentam ORF (*open reading frame)* ou ORF parciais que, a princípio, parecem genes, mas não são funcionais, ou são inativos, devido ao modo de sua origem ou a mutações. Essa família de elementos repetitivos possui numerosas substituições comparadas com os seus equivalentes funcionais, incluindo mutações por deslocamento do quadro de leitura ("*frameshift mutation*") e códons de terminação que impedem a tradução em polipeptídios funcionais. Um tipo muito relatado de pseudogenes são os processados, que são sequências de DNA que sofreram transcrição reversa do RNA e foram inseridos aleatoriamente no genoma. Os pseudogenes processados são muito numerosos e por serem desprovidos de íntrons, correspondem a quantidades menos importantes para o total de DNA da célula (LÓPEZ- FLORES; GARRIDO- RAMOS, 2012).

As sequências repetitivas têm sido muito utilizadas em estudos evoluitvos, por acumular variações nas sequências, podendo produzir várias cópias de si durante a evolução, sendo consideradas ferramentas importantes para estudos taxonômicos e filogenéticos, conhecidos como "botões de ajuste" na evolução. Nesse sentido, o

conhecimento das sequências repetitivas ajuda a compreender a organização, evolução e comportamento dos genomas eucarióticos. As sequências repetitivas têm efeitos citoplasmáticos, celulares e de desenvolvimento e desempenham um papel em recombinações cromossômicas (MEHROTRA; GOYAL, 2014).

As frações repetitivas do genoma podem ser específicas para uma espécie ou um gênero ou mesmo para uma família. Já foram relatadas muitas repetições de satélite que não são apenas específicas de um gênero ou família, mas existem em outras famílias, gêneros e espécies também (MEHROTRA et al., 2014). Nessa perspectiva, o estudo e a caracterização de sequências repetitivas podem fornecer informações sobre a organização, função e evolução do genoma de plantas.

2.2.2.3. Variação genômica

As diferenças entre genomas de plantas variam desde polimorfismos de nucleotídeos únicos até duplicações em grande escala, amplificações, deleções e rearranjos. Essas variações no genoma das plantas existem em muitas formas, podendo ser benéficas, neutras ou prejudiciais para a planta, e são descritas como as variações mais frequentes no genoma de qualquer organismo (SAXENA et al., 2014; SMITH, 2002).

Os polimorfismos de nucleotídios únicos ou SNPs (*single nucleotide polymorphisms*) são os tipos mais prevalentes de polimorfismo na maioria dos genomas. Os SNPs se destacam como a forma de variação mais simples que se poderia observar em um *locus*, caracterizando-se como uma diferença no nucleotídio presente em um sítio de nucleotídio único, seja adenina, citosina, guanina ou timina (LEACHÉ; OAKS, 2017).

Inserções/exclusões de base única (indels) não são geralmente incluídos na definição de um SNP. Os SNPs ocorrem em genes, inclusive éxons, íntrons e regiões reguladoras. Dois tipos de substituições de bases nucleotídicas resultando são descritas como resultantes em SNPs: uma substituição de transição ocorre entre purinas (A, G) ou entre pirimidinas (C, T), este tipo de substituição constitui dois terços de todos os SNPs; e uma substituição por transversão que ocorre entre uma purina e uma pirimidina (BROOKES, 2007; SMITH, 2002).

Em regiões codificadoras de proteínas os SNPs podem ser classificados em três grupos: *sinônimos*, se os alelos diferentes codificarem o mesmo aminoácido; *não*

sinônimos se os dois alelos codificarem aminoácidos diferentes e *sem sentido (nonsense)* se um alelo codificar um códon de parada e o outro alelo codificar um aminoácido. Nesse sentido, às vezes é possível associar um SNP à variação funcional em proteínas e qualquer alteração associada no fenótipo. Os SNPs localizados fora de sequências codificadoras denominam-se SNP *silenciosos* e podem ser muito úteis na genética de populações, mesmo que sejam encontrados em DNA não funcional, visto que podem ser usados como marcadores para tratar de questões sobre processos na genética de populações, como o fluxo gênico entre populações (BRUMFIELD et al., 2003).

Além de SNPs outras variações são bastante estudadas, tais como inversões, translocações, deleções ou duplicações e análises da presença ou ausência de um elemento de transposição em um *locus* particular no genoma. Outra forma comum de variação *é* o polimorfismo por inserção/deleção, ou *indel.* Esse tipo de polimorfismo envolve a presença ou ausência de um ou mais nucleotídios de um *locus* em um alelo com relação a outro (KIM et al., 2014). As variações genômicas podem ser evidenciadas no genoma nuclear, no plastidial ou mitocondrial. Quando são analisadas as variações de tamanho, organização e sequência do genoma desses três compartimentos nas plantas, o cpDNA é considerado evolutivamente mais conservado que os demais, com uma variação em tamanho relativamente pequena. Em contrapartida, o genoma nuclear de plantas apresenta variações de uma ou mais ordens de magnitude em tamanho. Por sua vez, o genoma mitocondrial é, via de regra, substancialmente maior (e mais variável) do que o genoma plastidial (MARTINS, 2013).

Os polimorfismos de nucleotídeo único (SNPs) representam o tipo mais difundido de variação de sequência em genomas, se destacando como um valioso marcador genético para revelar a história evolutiva das populações (BRUMFIELD et al., 2003). Nesse sentido, o estudo das variações em sequências de DNA nos diferentes compartimentos celulares pode dar margem para várias análises, que podem suplantar discussões sobre a origem de espécies, a história evolutiva de grupos, bem como sua distribuição geográfica.

2.3.Next-Generation Sequencing (NGS)

A necessidade em se reconstruir o genoma dos organismos e investigar a composição das sequências para esclarecimento de padrões de herança e da evolução genômica, fez com os esforços para o desenvolvimento de tecnologias para explorar as sequências de DNA fossem uma realidade iminente, com o intuito de obter métodos cada vez sofisticados e mais precisos, que produzam um volume de dados maiores com menores custos. A capacidade de medir ou inferir sobre as sequências de DNA se tornou então a base da pesquisa biológica (HEATHER; CHAIN, 2016).

No início dos anos 90, a produção de sequências de DNA começaram a ser realizadas com métodos semi-automatizados, baseados na clonagem de DNA em um plasmídeo por capilares de Sanger (SENDURE; JI, 2008). O método automatizado de Sanger é considerado uma tecnologia de "primeira geração" (METZKER, 2010) que possibilitou a obtenção de importantes informações sobre a caracterização gênica, o tamanho e conteúdo repetitivo do DNA, a uma taxa de erro muito baixa (KOBOLDT, 2013).

No entanto, a exigência de métodos rápidos, de baixo custo e altamente confiáveis desencadeou o desenvolvimento de tecnologias de seqüenciamento de DNA de nova geração, que além de superar as restrições nos métodos de sequenciamento Sanger, promovem o sequenciamento de DNA em plataformas capazes de gerar informação sobre milhões de pares de bases em uma única corrida (CARVALHO; SILVA, 2010; KÜREKÇI; DINÇER, 2014). A velocidade e a quantidade de informações geradas por essas novas tecnologias de sequenciamento estão revolucionando a investigação biológica e permitindo o acesso a genomas de diferentes espécies e sequenciamentos de diferentes linhagens (MARDIS, 2008; SHENDURE, AIDEN, 2012).

No sequenciamento de nova geração (*Next generation DNA sequencing - NGS*) para cada base lida, a informação é obtida por meio de *reads*, fragmentos de DNA gerados que serão utilizados nas abordagens genômicas subsequentes. Os fragmentos (*reads*) gerados por esses sequenciadores são fragmentos curtos (*short read sequence - SRS*) comparados com os fragmentos produzidos pela tecnologia *Sanger*. O tamanho dos fragmentos produzidos por NGS variam de acordo com as especificidades de organização de cada biblioteca, o que, apesar dos avanços, vem constituindo desafio para a bioinformática na montagem de genomas (MARDIS, 2008; METZKER, 2010).

As bibliotecas podem ser constituídas por diferentes métodos de leituras de sequências; o modo como estas sequências são lidas, pode faclitar ou limitar a

montagem de genomas. Os *reads Single-end*, são construídos usando leituras de uma das extremidades da sequência de DNA, este tipo de leitura é considerada a mais simples. *Reads paired-end* são obtidos a partir da leitura pareada de cada uma das extremidades das sequências de DNA, de modo que para cada *read* têm-se o par complementar. Reads *mate-pair,* são constituídos a partir da leitura de fragmentos de DNA de cadeia dupla que são circularizados de modo que as extremidades distantes são fisicamente ligadas e lidas juntas; este tipo de leitura tem potencial para resolver problemas de montagem que envolva sequências de DNA repetitivo (GLENN, 2011).

Uma variedade de recursos de NGS tornou possível a estruturação de várias plataformas no mercado, que oferecem uma série de particularidades. As principais tecnologias disponíveis comercialmente são as plataformas: Roche/454, Illumina/Solexa, Life/APG, Helicos e Pacific BioSciences (METZKER, 2010; MARDIS, 2008). A estratégia específica empregada em cada plataforma determina a qualidade, a quantidade e os vieses dos resultados das sequências, bem como a utilidade da plataforma para os estudos de genômica (MARDIS, 2013; REUTER et al., 2015). Os volumosos dados produzidos por estas plataformas colocam demandas substanciais na informação tecnológica em termos de armazenamento de dados, rastreamento e controle de qualidade.

As plataformas de sequenciamento de nova geração são uma alternativa poderosa para estudos de genômica estrutural e funcional (CARVALHO; SILVA, 2010). Na genômica de plantas, os trabalhos com as novas plataformas têm sido destinados ao sequenciamento de novo de genomas organelares, sequenciamento de transcritos, e mapeamento de elementos repetitios do genoma (DONG et al., 2018; FENG, et al., 2018; MACAS et al., 2015).

As plataformas de nova geração estão ajudando a abrir áreas inteiramente novas de investigação biológica, incluindo a investigação de genomas antigos, a caracterização da diversidade ecológica e identificação de agentes etiológicos desconhecidos. O NGS traz enorme mudança na pesquisa genética e biológica (MARDIS, 2008), uma vez que potencializou o sequenciamento de genomas completos de muitos organismos relacionados, permitindo estudos comparativos em larga escala, e o desenvovlimento de estudos evolutivos, o que antes era praticamente impossivel de ser realizado (METZKER, 2009). Neste sentido, as tecnologias de NGS representaram um grande impacto na capacidade de explorar e responder questões biológicas em todo o genoma (SENDURE; JI, 2008).

A diversidade e a rápida evolução do sequenciamento de nova geração vêm colocando desafios para a bioinformática em áreas que incluem a obtenção de sequências de qualidade, alinhamento, montagem e liberação de dados (SENDURE; JI, 2008). O que impõe um desafio aos profissionais da bioinformática, que necessitam aprimorar as ferramentas computacionais no sentido de acompanhar a evolução em volume de dados e sequências proporcionadas por esssa tecnologias se sequenciamento (ANSORGE, 2009).

Para espécies vegetais, a utilização das novas plataformas de sequenciamento se encontra em expansão, mas ainda são relatados muito desafios. Os principais são ainda em relação aos custos de montagem das bibliotecas, e o tamanho dos *reads* produzidos, o que muitas vezes é considerado incompatível com a montagem dos genomas nucleares gigantescos e altamente repetitivos das plantas (CARVALHO; SILVA, 2010).

Novas tecnologias de sequenciamento e análise de dados vêm continuamente sendo desenvolvidas, e espera-se que em um futuro próximo estas estejam mais acessíveis aos pesquisadores e que a estrutura da grande maioria dos genomas das plantas seja descrita.

2.4. BIOINFORMÁTICA APLICADA À GENÔMICA

A bioinformática é considerada uma linha de pesquisa que envolve aspectos multidisciplinares e que surgiu a partir do momento em que se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos, bioquímicos e de biologia molecular (PROSDOCIMI, 2007). Nesse aspecto, a bioinformática surge a partir da biologia molecular e dela ainda é inseparável, tendo como finalidade principal decifrar o grande volume de dados que vem sendo obtido através de sequências de DNA e proteínas.

Ferramentas de bioinformática para análise de dados são rotineiramete desenvolvidas e aprimoradas, e apesar das limitações comercais ainda encontradas, já se tem disponivel uma variedade de ferramentas de software disponível para análise de dados de sequenciamento de nova geração. Suas funções se encaixam em várias categorias gerais, incluindo: alinhamento de sequências de reads em uma referência; detecção de polimorfismos; montagem de novo de *reads* pareados ou não pareados, anotação de genoma, e análises filogenéticas (SENDURE; JI, 2008; VERLI, 2014).

2.4.1. Análise, montagem e anotação de genomas

O avanço nas técnicas de sequenciamento do DNA tem permitido um crescente aumento no número de genomas disponíveis em bancos de dados públicos. Esta maior disponibilidade exigiu um grande aumento na capacidade computacional de armazenamento e no investimento em desenvolvimento de técnicas de processamento adequadas para a análise destes dados (JUNQUEIRA et al., 2014). No entanto, os custos de sequenciamento geralmente estabelecem limites para o valor sequências que podem ser geradas e, consequentemente, os resultados biológicos que podem ser alcançado a partir de um desenho experimental (SIMS et al., 2014).

Como resultado do sequenciamento obtêm-se uma grande lista de sequências nucleotídicas - os *reads* - de tamanhos que podem variar de 50 a 800 pb. Para montagem das sequências genômicas a partir destes *reads*, diferentes estratégias são utilizadas, dependendo da metodologia empregada. Em geral, as estratégias se voltam para o alinhamento de *reads* entre si na procura de regiões de identidade ou de sobreposição, de maneira a construir fragmentos contíguos (*contigs*), os quais podem ser definidos como a união de duas ou mais sequências (*reads*) formadas por sobreposição de elementos comuns a pelo menos duas sequências (KÜREKÇI; DINÇER, 2014; STAATS et al., 2014). Montagens de alta qualidade são muitas vezes produzidos utilizando abordagens combinadas, em que as vantagens de sequenciamento de alta profundidade e de reads curtos são complementados com aqueles de menor profundidade, mas de leitura mais longa (SCHATZ et al., 2012).

Os genomas de eucariotos, em especial de eucariotos superiores, tem processo de montagem mais complexo devido à quantidade considerável de sequências repetitivas e a extensão em número de pares de bases do genoma (STAATS et al., 2014). Para sobrepujar estas dificuldades, passos intermediários se tornam necessários, como a construção de sub-bibliotecas genômicas. O grande desafio na montagem das sequências genômicas com alto conteúdo de elementos repetitivos se refere à correta quantificação e localização destes elementos nos cromossomos. Isso vem sendo superado por meio da obtenção de bibliotecas com *reads* maiores, como as de *mate-pair* (PROSDOCIMI, 2007).

A qualidade de uma montagem é avaliada pela cobertura do genoma obtida. A cobertura de estudos de sequenciamento é frequentemente citada como a profundidade

de reads brutos ou alinhada, que denota a cobertura com base no número e no comprimento dos reads de alta qualidade antes ou depois do alinhamento de referência. Embora os termos profundidade e cobertura possam ser usados com finalidades similares, a cobertura também tem sido usada para denotar a amplitude de cobertura de um genoma alvo, que é definido como a porcentagem de bases alvo que são sequenciados um dado número de vezes (SIMS et al., 2014); e a profundidade necessária para um estudo de sequenciamento do genoma é determinada pela taxa de erro do método de sequenciamento, dos algoritmos de montagem usados, da complexidade dos repeats do genoma particular sob estudo e o tamanho do *read* (SCHATZ et al., 2010).

O passo seguinte à montagem dos genomas é sua anotação, que constitui em um conjunto de protocolos e fluxos de trabalho utilizados para delimitar, em uma determinada sequência genômica, possíveis genes e predizer a sua função com base na similaridade com sequências conservadas (STAATS et al., 2014). Na anotação do genoma procura-se encontrar a localização física (posição cromossômica) de cada parte da sequência e descobrir onde estão os genes, RNAs e elementos repetitivos.

2.4.2. Análise de SNPs

A comparação entre genomas de diferentes espécies, ou até mesmo de indivíduos da mesma espécie, possibilita a análise de variações (mutações ou polimorfismos) nas sequências e, em alguns casos, permite a identificação de relações entre variações no DNA e os efeitos para genômica do grupo (VERLI, 2014). Os SNPs são polimorfismos causados pela alteração em nucleotídeos únicos, que vem sendo muito utilizados em estudos genômicos, para esclarecimento de filogenias e determinação da distribuição de espécies.

Há duas maneiras de detectar um SNP. A primeira consiste em sequenciar um segmento de DNA em cromossomos homólogos e comparar os segmentos homólogos para descobrir diferenças. Uma segunda maneira é possível no caso de SNP localizados em um sítio-alvo de uma enzima de restrição: esses SNP são os polimorfismos de comprimento de fragmento de restrição (RFLP). Em tais casos, existirão dois "alelos" de RFLP, ou morfos, dos quais um tem o sítio-alvo da enzima de restrição e o outro não. A enzima de restrição irá cortar o DNA no SNP que contém o alvo e ignorar o

outro SNP. Os SNP são, então, detectados como bandas diferentes em um gel de eletroforese. Os sítios de RFLP podem estar entre genes ou dentro deles. Atualmente já são desenvolvidos alguns softwares específicos para análise dos SNPs, como o SNPHunter (WANG et al., 2005), SNPdetector (ZHANG et al., 2005) Seq4SNPs (FIELD et al., 2009), Bioedit (HALL, 2011), SNPator (MORCILLO-SUAREZ, 2008), e DiscoSNP (URICARU et al., 2015).

Os métodos de detecção de sítios únicos são, potencialmente, mais fáceis de automatizar e de aplicar na análise genética em larga escala. Deste modo, a informação gerada sobre os SNPs pode ser posteriormente, utilizada em vários níveis nas análises do DNA como: marcadores no mapeamento genético; material de estudo na identificação de SNPs funcionais e seu respectivo fenótipo; e ferramenta para estudos populacionais, por comparação do genoma de indivíduos de diferentes populações (ZHANG E HEWITT, 2003).

ESCREVER SOBRE O TRABALHO DA LILÁS

2.4.3. Análises filogenéticas

As análises filogenéticas são feitas com base no alinhamento de sequências que possibilitam a comparação entre pares de bases similares e a partir disso se possibilita a inferência sobre a proximidade filogenética das espécies. Os alinhamentos são técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas (VERLI, 2014). São frequentemente utilizados na comparação de sequências contra grandes bancos de dados, exatamente como faz o BLAST, que procura a similaridade de uma sequência de entrada contra milhões de outras presentes em seu banco de dados. Estes alinhamentos são feitos também comumente no software Geneious, que é utilizado para executar diversos comandos que envolvem estudos genômicos.

Depois que o alinhamento foi proposto, diversos métodos podem ser usados para estimar a filogenia das sequências estudadas. Podemos dividir estes métodos em dois principais grupos: métodos quantitativos e métodos qualitativos. Estes grupos diferem na forma como os dados são tratados, refletindo diretamente como os dados do alinhamento serão inicialmente processados. Os principais métodos utilizados são o de máxima parcimônia, máxima verossimilhança e inferência Bayesiana, por se tratarem de métodos que buscam uma única filogenia entre diversas árvores (BRAUN et al., 2014).

Estes são executados em geral através dos programas, Geneious (KEARSE et al., 2012), MEGA (TAMURA, et al., 2013), Mr. Bayes (RONQUIST; HUELSENBECK, 2003) BEAST (SUCHARD et al., 2018) PAUP (SWOFFORD, 2002) que em geral representam o contexto evolutivo dos organismos de forma gráfica em árvores filogenéticas.

Para as análises filogenéticas podem ser utilizadas sequências de aminoácidos, de nucleotídeos, genes e até o genoma completo Quando se analisa sequências de nucleotídeos ou aminoácidos para inferir uma filogenia, se utiliza de informações derivadas das taxas evolutivas para determinar a sequência de eventos que levaram ao surgimento de novos organismos. A taxa de evolução molecular refere-se à velocidade na qual os organismos acumulam diferenças genéticas ao longo do tempo. Essa taxa é frequentemente definida pelo número de substituições por sítio (ou posição no alinhamento de sequências) por unidade de tempo e, portanto, são usadas para descrever a dinâmica das mudanças em uma linhagem ao longo de várias gerações (BRAUN et al., 2014).

3. REFERÊNCIAS BIBLIOGRÁFICAS

AJAO, A.O.; SHONUKA, O.; FEMI-ONADEKO, B. (1985): Antibacterial effect of aqueous and alcohol extracts of *Spondia monbin* and *Alchornea cordifolia* – Two local antimicrobial remedies. **Int. J. Crude Drug Res**., v. 23, n. 2, p. 67-72, 1985. Doi: 10.3109/13880208509069004

ALBUQUERQUE, U.P., ANDRADE, L.H.C., SILVA, A. C. O. Use of plant resources in a seasonal dry forest (northeastern Brazil). **Acta Botanica Brasilica,** v. 19, 1–16, jan/mar. 2005. doi: 10.1590/S0102-33062005000100004

ALLEN, J. O, et al. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. **Genetics**, v. 177, n.2, p. 1173-1192, 2007. doi: 10.1534/genetics.107.073312

ALMEIDA, C. C. de S.; CARVALHO, P. C. De L.; GUERRA, M. Karyotype differentiation among *Spondias* species and the putative hybrid Umbu-cajá (Anacardiaceae). **Botanical Journal of the Linnean Society**, v. 155, p. 541–547, jul. 2007. doi: 10.1111/j.1095-8339.2007.00721.x

ALMEIDA, C. F. C. B. R. et al. A comparison of knowledge about medicinal plants for three rural communities in the semi-arid region of northeast of Brazil. **Journal of Ethnopharmacology**, v. 127, p.674–684, jan. 2010. doi:10.1016/j.jep.2009.12.005

ALVERSON, A. J. et al. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). **MBE Advance**, jan. 2010.

ALVERSON, A. J. et al. The Mitochondrial Genome of the Legume Vigna radiate and the Analysis of Recombination across Short Mitochondrial Repeats. **Plos One**, v.6, n. 1, jan. 2011. doi:10.1371/journal.pone.0016404

ANSORGE, W. J. Next-generation DNA sequencing techniques. **New Biotechnology**, v. 25, n. 4, p. 195-203, apr. 2009. doi: 10.1016/j.nbt.2008.12.009.

ARIAS, M. C.; FRANCISCO, F. de O.; SILVESTRE, D. O DNA mitocondrial em estudos populacionais e evolutivos de meliponíneos. *In* G. A. R. Melo & I. Alves-dos-Santos, **Apoidea Neotropica: Homenagem aos 90 Anos de Jesus Santiago Moure**. Editora UNESC, Criciúma, p. 301-309, 2003.

AYOKA, A. O. et al. Medicinal and economic value of *Spondias mombin*. African **Journal of Biomedical Research**, v. 11, p. 129–136, 2008.

AYOKA, A. O. et al. Sedative, antiepileptic and antipsychotic effects of *Spondias mombin* L. (Anacardiaceae) in mice and rats. **Journal of Ethnopharmacology**, 103: 166–175, sep, 2006. doi: 10.1016/j.jep.2005.07.019

BALTIMORE, D. Retroviruses and retro-transposons: The role of reverse transcripition in shapping the eukaryotic genoma. **Cell**, v. 48, p. 481-482, mar.1985. doi: 10.1016/0092-8674(85)90190-4

BATISTA, F. R. da C. et al. O umbuzeiro e o semiárido brasileiro. Campina Grande: **INSA**, 2015. 72p.

BENNETZEN J. L. Opening the door to comparative plant biology. **Science**, v. 296, p. 60–63, apr. 2002. doi: 10.1126/science.1071402

BENNETZEN, J. L.; MA, J.; DEVOS, K. M. Mechanisms of Recent Genome Size Variation in Flowering Plants. **Annals of Botany**, v. 95, n. 1, p.127–132, 2005. doi:10.1093/aob/mci008

BI, C. et al. Analysis of the Complete Mitochondrial Genome Sequence of the Diploid Cotton *Gossypium raimondii* by Comparative Genomics Approaches. **BioMed Research International**, p. 1-19, 2016. doi: 10.1155/2016/5040598

BIRKY JÚNIOR, C.W. Uniparental in heritance of mitochondrial and chloroplast genes: Mechanisms and evolution. **Proc. Natl. Acad. Sci**. **USA**, v. 92, n. 25, p. 11331-11338, dec.1995.

BORGES, S.V. et al. Chemical composition of umbu (*Spondias tuberosa* Arr. Cam.) seeds. **Química Nova**, v. 30, n. 1, p. 49-52, 2007. doi: 10.1590/S0100-40422007000100011.

BRAUN, R. L.; JUNQUEIRA, D. M.; VERLI, H. Filogenia Molecular In: VERLI, H. (Org.). **Bioinformática da Biologia à flexibilidade molecular**. São Paulo: SBBq, 2014, cap. 5, p. 81- 114.

BROOKES, A. J. Single Nucleotide Polymorphism (SNP). **eLS**, dec. 2007,doi: 10.1002/9780470015902.a0005006.pub2

BRUMFIELD, R.T. et al. The utility of single nucleotide polymorphisms in inferences of population history. **Trends Ecol. Evol.** v. 18, n. 5, p.249–56, may. 2003. doi:10.1016/S0169-5347(03)00018-1

BULLERWELL, C. E.; GRAY, M.W. Evolution of the mitochondrial genome: protist connections to animals, fungi and plants. **Current Opinion in Microbiology**, v. 7, n. 5 p. 528–534, oct. 2004. doi: 10.1016/j.mib.2004.08.008

CARVALHO, M. C. da C. G. de; SILVA, D. C. G. da. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, v.40, n.3, p.735-744, mar, 2010. 10.1590/S0103-84782010000300040

CAVALCANTI, N. B et al. Ciclo reprodutivo do umbuzeiro (*Spondias tuberosa* Arruda) no semiárido do Nordeste brasileiro. **Revista Ceres,** v. 47, n. 272, p. 421–439, 2000.

CBOL Plant working group. A DNA barcode for land plants. **Proc. Natl. Acad. Sci. U.S.A.** 106, n. 31, p.12794-12797, aug. 2009. doi: 10.1073/pnas.0905845106

CHANG S. et al. Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in Brassica. **BMC Genomics**, v. 12, n. 2, p.497, feb. 2011. doi:10.1371/journal.pone.0056502

CHRISTENHUSZ, M. J. M.; BYNG, J.W. The number of known plant species in the world and its annual increase. **Phytotaxa, v.** 261, n. 3, p. 201–217, may 2017. doi: 10.11646/phytotaxa.261.3.1

CORTHOUT, J. L. et al. Antibacterial and molluscicidal phenolic acids from *Spondias mombin*. **Planta Medica**, v. 60, n.5, p. 460-463, oct. 1994. doi: 10.1055/s-2006-959532
DODSWORTH, S. et al., Genomic Repeat Abundances Contain Phylogenetic Signal. **Systematic Biology**, v. 64, n. 1, p.112–126, sep. 2015. doi: 10.1093/sysbio/syu080

DONG, S. et al. Complete mitochondrial genome sequence of *Anthoceros angustus*: conservative evolution of the mitogenomes in hornworts. **The Bryologist**, v. 121, n. 1, p.14-22, jan. 2018. doi: 10.1639/0007-2745-121.1.014

DONNELLY, K. et al. Reconstructing the plant mitochondrial genome for marker discovery: a case study using Pinus. **Molecular Ecology Resources**, v. 17, n. 5, p. 943–954, jan. 2017. doi: 10.1111/1755-0998.12646

DODSWORTH, S. Genomic Repeat Abundances Contain Phylogenetic Signal. **Syst. Biol.**, v. 64, n. 1, p. 112–126, sep. 2015. doi: 10.1093/sysbio/syu080.

FENG, B. et al. Development of novel EST-SSR markers for ploidy identification based on *de novo* transcriptome assembly for *Misgurnus anguillicaudatus*. **PLOS ONE**, v. 3, n. 4, p. 1-15, apr. 2018. doi:10.1371/journal.pone.0195829

FIELD, H. I. et al. Seq4SNPs: new software for retrieval of multiple, accurately annotated DNA sequences, ready formatted for SNP assay design**. BMC Bioinformatics**, v.10, n.180, p.1-10, jun. 2009. doi:10.1186/1471-2105-10-180

FUJII, S. et al. Discovery of global genomic re-organization based on comparison of two newly sequenced rice mitochondrial genomes with cytoplasmic male sterility-related genes. **BMC Genomics**, v. 11, n. 209, p. 1-15, mar. 2010. doi: 10.1186/1471-2164-11-209.

GIACOMETTI, D.C. Recursos genéticos de fruteiras nativas do Brasil. In: **Anais do Simpósio Nacional de Recursos Genéticos de Fruteiras Nativas**. Cruz das Almas/BA. Cruz das Almas: Embrapa CNPMF, 1993, p. 13–27.

GLENN, T. C. Field guide to next-generation DNA sequencers. **Molecular Ecology Resources**, v. 11, n. 5, p. 759–769, sep. 2011 doi: 10.1111/j.1755-0998.2011.03024.x

GOREMYKIN, V. V. et al. Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. **Mol Biol Evol**, v. 26, p. 99-110, jan. 2009. doi: 10.1093/molbev/msn226

GREINER, S.; BOCK, R. Tuning a menage a trois: co-evolution and co-adaptation of nuclear and organelar genomes in plants. **Bioessays**, v. 35, n. 4, p. 354-365, jan. 2013. doi: 10.1002/bies.201200137

GUALBERTO, J. M. et al. RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature, v.341, n.6243, p. 660-662, oct.1989. doi:10.1038/341660a0

GUALBERTO, J. M. et al. The plant mitochondrial genome: Dynamics and maintenance. **Biochimie**, v. 100, p. 107-120, sep. 2014. doi: 10.1016/j.biochi.2013.09.016

GUI, S. et al. O mapa do genoma mitocondrial de *Nelumbo nucifera* revela características evolutivas antigas. **Sci. Rep**. v. 6, n. 30158, p. 1-11, jul. 2016. doi: 10.1038/ srep30158 .

GUO, W., et al. Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. **New Phytol**. 213, p. 391–403, jul. 2017. doi: 10.1111/nph.14135

HALL, T. BioEdit: An important software for molecular biology. **GERF Bulletin of Biosciences**, v. 2, n.1, p.60-61, jun. 2011.

HANDA, H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the

mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. **Nucleic Acids Res,** v. 31, n. 20, p. 5907-5916, oct. 2003. doi: 10.1093/nar/gkg795

HESLOP-HARRISON, J. S.; SCHMIDT, T. (August 2012) Plant Nuclear Genome Composition. **eLS**, p. 1-9, aug , 2012. Doi:10.1002/9780470015902.a0002014.pub2.

HESLOP-HARRISON, J.S.; SCHWARZACHER, T. Organisation of the plant genome in chromosomes. **The Plant Journal**, v. 66, n. 1, p. 18-33, apr. 2011. doi: 10.1111/j.1365-313X.2011.04544.x

HEATHER, J. M.; CHAIN, B. The Sequence of Sequencers: The History of Sequencing DNA, **Genomics**, v. 107, n.1, p. 1-8, jan. 2016. doi:10.1016/j.ygeno.2015.11.003

HIDALGO, O. et al., Genomic gigantism in the whisk-fern family (Psilotaceae): *Tmesipteris obliqua* challenges record holder *Paris japonica*. **Botanical Journal of the Linnean Society**, v. 183, 509–514, jan. 2017. doi: 10.1093/botlinnean/box003

HIESEL, R.; COMBETTES, B.; BRENNICKE, A. Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta. **Proc. Natl. Acad. Sci. USA**, v. 91, n. 2, p. 629-33, jan.1994.

JANSEN, R. K. Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences. **Methods in enzymology**, v. 395, 2005.

JUDD, W. S. et al. **Sistemática Vegetal: um enfoque filogenético**. Tradução de André Olmos Simões, et al. 3ed. São Paulo: Artmed, 2009, 632p.

JUNQUEIRA, D. M.; BRAUN, R. L.; VERLI, H. Alinhamentos. In: VERLI, H. (Org.). **Bioinformática da Biologia à flexibilidade molecular**. São Paulo: SBBq, 2014, cap. 3, p. 39-61.

KEARSE, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. **Bioinformatics Applications Note**. v. 28, p. 1647-1649, 2012.

KELLOGG, E. E. A.; BENNETZEN, J. J. L. The evolution of nuclear genome structure in seed plants. **American Journal of Botany**, v. 91, n. 10, p. 1709–1725, out. 2004. doi: 10.3732/ajb.91.10.1709.

KELLY, L. J. et al. Why size really matters when sequencing plant genomes. **Plant Ecol. & Diversity**, v. 5, 415–425, oct. 2012. doi: 10.1080/17550874.2012.716868

KELLY, L. J. et al. Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. **New Phytologist**, v.208, p. 596–607, apr. 2015. doi: 10.1111/nph.13471

KERSTEN, B. et al. Genome Sequences of Populus tremula Chloroplast and Mitochondrion: Implications for Holistic Poplar Breeding. **PLOS ONE**, v. 11, n. 1, p. 1-21, jan. 2016. doi:10.1371/journal.pone.0147209.

KIM, S. et al. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. **Genome Res**., v. 24, n. 6, p. 1012-9, jun. 2014. doi: 10.1101/gr.171322.113.

KNOOP, V. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. **Curr Genet**, v. 46, p.123-139, aug. 2004. doi: 10.1007/s00294-004-0522-8

KUBO, T.; NEWTON, K.J. Angiosperm mitochondrial genomes and mutations. **Mitochondrion**, v. 8, n. 1, p. 5-14, jan. 2008. doi: 10.1016/j.mito.2007.10.006

KOBOLDT, D. C. The Next-Generation Sequencing Revolution and Its Impact on Genomics. **Cell**., v. 155, n.1, p. 27–38, sep. 2013. doi:10.1016/j.cell.2013.09.006.

KÜREKÇI, G. K.; DINÇER, P. Next-Generation DNA Sequencing Technologies. **Erciyes Med J**, v, 36, n. 3, p. 99-103, 2014. doi: 10.5152/etd.2014.7803

LANG, A. S.; ZHAXYBAYEVA, O.; BEATTY, J. T. Gene transfer agents: phage-like elements of genetic exchange. **Nat Rev Microbiol.,** v. 10, n.7, p. 472–482, jun. 2012. doi: 10.1038/nrmicro2802.

LAPITAN, N. L. V. Organization and evolution of higher plant nuclear genomes. **Genome**, v. 35, n. 2, p. 171–81, oct. 1992. doi: 10.1139/g92-028

LEACHÉ; OAKS, 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. **Annual Review of Ecology, Evolution, and Systematics**, v. 48, p 69-84, nov. 2017. doi: 10.1146/annurev-ecolsys-110316-022645

LEE, S.; KIM, N-S. Transposable Elements and Genome Size Variations in Plants. Sep. 2014. **Genomics Inform**, v.12, n.3, p.87-97, jul. 2014. doi: 10.5808/GI.2014.12.3.87

LIRA JÚNIOR, J.S. et al. Caracterização física e físico-química de frutos de cajá-umbu (*Spondias spp.*). **Ciências de tecnologia de alimentos,** v. 25, n. 4, p. 757-761, dez. 2005. doi: 10.1590/S0101-20612005000400021.

LÓPEZ- FLORES, I.; GARRIDO- RAMOS, M.A. The Repetitive DNA Content of Eukaryotic Genomes. **Genome Dyn. Basel**, Karger, v. 7, p. 1–28, 2012,

MACHADO, M. C.; CARVALHO, P. C. L.; VAN DEN BERG, C. Domestication, hybridization, speciation, and the origins of an economically import ant tree crop of Spondias (Anacardiaceae) from the Brazilian Caatinga dry forest. **Neodiversity**. v. 8, p. 8-49, jun. 2015. doi: 10.13102/neod.81.2

MACAS, J. et al. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. **PLOS ONE**, v. 10, n.11, p. 1-23, nov. 2015. doi:10.1371/journal.pone.0143424

MAIA, G. N. Caatinga: árvores e arbustos e suas utilidades. São Paulo: **D&Z Computação Gráfica e Editora**, 413p, 2004.

MARDIS, E. R. Next-Generation DNA Sequencing Methods. **Annu. Rev. Genomics Hum. Genet**., v. 9, p. 387–402, jun. 2008. doi: 10.1146/annurev.genom.9.081307.164359

MARDIS, E. R. Next-Generation Sequencing Platforms. **Annu. Rev. Anal. Chem**., v. 6, n. 8, p. 287–303, mar. 2013. doi: 10.1146/annurev-anchem-062012-092628

MARTINS, A. M. **Sequenciamento de DNA, montagem de novo do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis***. 2013, 198 f. Tese (Doutorado) – Instituto de Ciências Biológicas, Universidade de Brasília, Brasília.

MATOS, F. J. A. Cajazeira *Spondias mombin* Jacq. (Anacardiaceae). In: MATOS, F.J.A. **Farmácia viva:** sistema de utilização de plantas medicinais projetado para pequenas comunidades. 2. ed. Fortaleza: EUFC, p. 67-68, 1994.

MATSUMOTO, T. et al;. The map-based sequence of the rice genome. **Nature**, v. 436, n. 7052, p. 793–800, aug. 2005. doi: 10.1038/nature03895

MEHROTRA, S. et al. Significance of Satellite DNA Revealed by Conservation of a Widespread Repeat DNA Sequence Among Angiosperms. **Appl Biochem Biotechnol**, v. 173, n. 7, p. 1790-17801, aug. 2014. DOI: 10.1007/s12010-014-0966-3.

MEHROTRA, S.; GOYAL, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. **Genomics Proteomics Bioinformatics**, v. 12, p. 164–171, 2015 doi: 10.1016/j.gpb.2014.07.003

METZKER. M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, v. 11, p. 31-46, jan. 2010 doi:10.1038/nrg2626

MICHAEL, T. P.; JACKSON. S. The first 50 plant genomes. **Plant Genome**, v.6, n. 2, aug. 2013. doi:10.3835/plantgenome2013.03.0001in

MILLER, A.; SCHAAL, B. Domestication of a Mesoamerican cultivated fruit tree, *Spondias purpurea*. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 36, p. 12801–12806, sep. 2005. DOI: 10.1073_pnas.0505447102

MITCHELL, J. D.; DALY, D. C. A revision of *Spondias* L. (Anacardiaceae) in the Neotropics. **PhytoKeys**, v. 55, p. 1-92, 2015. doi: 10.3897/phytokeys.55.8489

MITCHELL, J. D.; DALY, D. C. The "Tortoise's Cajá"- A new species of *Spondias* (Anacardiaceae) from Southwestern Amazonia. **Brittonia**, v. 50, n. 4, p. 447-451, oct. 1998. DOI: 10.2307/2807753

MITCHELL, J. D. et al. *Poupartiopsis* gen. nov. and its context in Anacardiaceae classifi cation. **Systematic Botany**, v. 31, n. 2, p. 337–348, apr. 2006. doi: 10.1600/036364406777585757

MORCILLO-SUAREZ, C. et al. SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. **Bioinformatics applications note**, v. 24, n. 14, p. 1643–1644, may. 2008. doi:10.1093/bioinformatics/btn241

MOWER J. P.; SLOAN, D. B.; ALVERSON, A. J. Plant mitochondrial genome diversity: the genomics revolution. **Plant Genome Diversity**, v.1, p. 123-144, mar. 2012. doi: 10.1007/978-3-7091-1130-7_9

NAITO et al. De novo assembly of the complete organelle genome sequences of azuki bean (*Vigna angularis*) using next-generation sequencers. **Breeding Science**, v. 63, n.2, p.176–182, feb. 2013. doi:10.1270/jsbbs.63.176

NASCIMENTO, V.T. et al. Famine foods of Brazil's seasonal dry forests: Ethnobotanical and nutritional aspects. **Economic Botany**, v. 66, p. 22–34, mar. 2012.

NOBRE, L. L. de M. et al. Phylogenomic and SNP analysis reveals the hybrid origin of Spondias bahiensis (Anacardiaceae): De novo genome sequencing and comparative genomics. **Genetics and Molecular Biology**, 2018.

OGIHARA , Y. et al. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. **Nucleic Acids Res,** v. 33, n. 19, p. 6235–6250, oct. 2005. DOI: 10.1093/nar/gki925

OLIVER J. L.; MARÍN, A.; MARTÍNEZ-ZAPATER, J. M. Chloroplast Genes Transferred to the Nuclear-Plant Genome Have Adjusted to Nuclear-Base Composition and Codon Usage. **Nucleic Acids Research**, v. 18, n. 1, p. 65-73, jan. 1990.

OLMSTEAD, R. G.; PALMER, J. D. Chloroplast DNA systematics: a review of methods and data analysis. **American Journal of Botany**, v. 81, n.9, p. 1205-1224, sep. 1994. doi: 10.1002/j.1537-2197.1994.tb15615.x

PALMER, J.; NUGENT, J.; HERBON, L. Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. **Proc Natl Acad Sci USA**, v. 84, n. 3, p. 769–773, 1987.

PANAUD, L. et al. Drivers and dynamics of diversity in plant genomes. **New Phytologist**, v. 202, p. 15–18, 2014. doi: 10.1111/nph.12633

PELL, S. K. Molecular systematic of the cashew family (Anacardiaceae). **PhD Thesis, Louisiana State University**, Baton Rouge, 2004.

PELL, S. K. et al. Anacardiaceae. In: Kubitzki K (Ed.) **The families and genera of vascular plants**. Springer-Verlag, Berlin, v.10, p. 7–50, 2010.

PELLICER, J.; FAY, M. F.; LEITCH, I. J. The largest eukaryotic genome of them all? **Bot. J. Linn. Soc.**, 164, 10–15, sep. 2010. doi: 10.1111/j.1095-8339.2010.01072.

PELLICER, J. et al.. Genome Size Diversity and Its Impact on the Evolution of Land Plants. **Genes**, v. 9, n. 88, p. 1-14, feb. 2018. doi:10.3390/genes9020088

PROSDOCIMI, F. **Introdução à Bioinformática**. Brasília: Biotecnologia Ciencia & Desenvolvimento, 2007, 77p.

REUTER, J. A.; SPACEK, D. V.; SNYDER, M. P. High-Throughput Sequencing Technologies. **Molecular Cell** , v. 58, n. 4, p. 586-597, may. 2015. doi: 0.1016/j.molcel.2015.05.004

RIBEIRO, S.B. **Sequenciamento e caracterização parcial do genoma de cagaiteira** (*Eugenia dysenterica* DC.). 2016. 76 f. Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Goiânia. Orientador: Alexandre Siqueira Guedes Coelho.

RICHARD, G-F; KERREST, A.; DUJON, B. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. **Microbiol Mol Biol Rev.**, v.72, n.4, Dec. 2008, p. 686–727. doi: 10.1128/MMBR.00011-08.

ROBLEDILLO, L. A. 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, apr. 2018. doi:10.1038/s41598-018-24196-3

RODRIGUES, K. F.; HASSE, M.; WERNER, C. Antimacrobial activities of secondary metabolites produced by endophytic fungi from *Spondias mombin*. **Journal of Basic Microbiology**, v. 40, n. 4, p. 261 – 267, apr. 2000. doi: 10.1002/1521-4028

RONQUIST, F.; HUELSENBECK, J. MRBAYES 3: Bayesian Phylogenetic inference under mixed model. **Bioinformatics**, v. 19, n. 12, p. 1572–1574, sep. 2003. Doi: 10.1093/bioinformatics/btg180

SANTOS, C.A.F. *In situ* evaluation of fruit yield and estimation of repeatability coefficient for major fruit traits of umbu tree [*Spondias tuberosa* (Anacardiaceae)] in the semi-arid region of Brazil. **Genetic Resources and Crop Evolution**, v. 46, n. 5, p. 455-460, 1999.

SANTOS, J. O. D. dos **Análise genômica revela origem híbrida entre *Spondias tuberosa* X *Spondias bahiensis* (Anacardiaceae) e alta similaridade na fração repetitiva**, 2016, 43f. Dissertação (Mestrado em Agricultura e Ambiente) – Universidade Federal de Alagoas, Arapiraca. Orientador: Prof. Dr. Cícero Carlos de Souza Almeida.

SANTOS, L. A. et al. Coeficientes de correlação entre caracteres químicos e físico-químicos em frutos de umbucajazeira. In: **Reunião Regional da SBPC, 2010, Cruz das Almas**. Anais/Resumos da Reunião Regional da SBPC no Recôncavo da Bahia/BA. Cruz das Almas: Sociedade Brasileira para o Progresso da Ciência/SBPC.

SANTOS, V.; ALMEIDA, C.C. de. The complete chloroplast genome sequences of three *Spondias* species reveal high relationships among species. **Genetics and Molecular Biology,** 2018.

SANTOS, C. A. F.; OLIVEIRA, V. R. Inter-relações genéticas entre espécies do gênero *Spondias* com base em marcadores AFLP. **Revista Brasileira de Fruticultura**, v. 30, n. 3, p. 731-735, set. 2008. doi: 10.1590/S0100-29452008000300028.

SAXENA, R. K.; DAVID EDWARDS, D.; VARSHNEY, R. K. Structural variations in plant genomes. **BRIEFINGS IN FUNCTIONAL GENOMICS**. v. 13. n. 4., p. 296-307, jun. 2014. doi: 10.1093/bfgp/elu016

SCHATZ, M. C., DELCHER, A. L.; SALZBERG, S. L. Assembly of large genomes using second-generation sequencing. *Genome Res. v.***20**, n.9, p. 1165–1173, sep. 2010. doi: 10.1101/gr.101360.109.

SCHATZ, M. C., WITKOWSKI, J.; MCCOMBIE, W. R. Current challenges in *de novo* plant genome sequencing and assembly. **Genome Biol**., v.13, n.4, p. 243, abr.2012. doi: 10.1186/gb-2012-13-4-243

SCHUBERT, I.; VU, G. T. H. Genome Stability and Evolution: Attempting a Holistic View. **Trends in Plant Science**, v. 21, n. 9, p. 749-757, sep. 2016. Doi: 10.1016/j.tplants.2016.06.003 2016

SHENDURE, J; JI, H. Next-generation DNA sequencing. **Nature biotechnology**, v. 26, n.10, p. 1135- 1145, oct. 2008. doi:10.1038/nbt1486

SHENDURE, J; AIDEN, E. L. The expanding scope of DNA sequencing. **Nat Biotechnol.**, v.30, n. 11, p. 1084-94, nov. 2012. doi: 10.1038/nbt.2421.

SHEARMAN, J. R. The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. **Scientific Reports**, v. 6, n. 31533, p.1-7, aug. 2016. DOI: 10.1038/srep31533

SILVA, G.A.; BRITO, N.; SANTOS, E. C. G.. Gênero *Spondias*: Aspectos botânicos, composição química e potencial farmacológico. **BioFar**, v. 10, n. 1, jan. 2014.

SILVA, J. N. et al. DNA barcoding and phylogeny in neotropical species of the genus *Spondias*. **Biochemical Systematics and Ecology**. v. 61, p. 240-243, aug. 2015. doi: 10.1016/j.bse.2015.06.005

SILVA, S. R. et al. The mitochondrial genome of the terrestrial carnivorous plant *Utricularia reniformis* (Lentibulariaceae): Structure, comparative analysis and evolutionary landmarks. **PLOS ONE**, v. 12, n. 7, p. 1-26, jul. 2017. doi: 10.1371/journal.pone.0180484

SILVA JÚNIOR, J. F. et al. Collecting, *ex situ* conservation and characterization of 'cajá-umbu' (*Spondias mombin* x *Spondias tuberosa*) germplasm in Pernambuco State, Brazil. **Genetic Resources and Crop Evolution**, v. 51, p. 343–349, 2004.

SIMS, D. et al. Sequencing depth and coverage: key considerations in genomic analyses. **Nat Rev Genet**. v. 15, n. 2, p. 121-32, feb. 2014. doi: 10.1038/nrg3642.

SLOAN, D. B., et al. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. **PLOS BIOL**, v. 10, p. 1-17, jan. 2012. doi:10.1371/journal.pbio.1001241

SMITH, C. E. Plant remains. In: MacNeish R. S (Ed.) **The Prehistory of the Tehuacan Valley 5**. University of Texas Press, Austin, p. 220–255, 1967.

SMITH, D. R. Does Cell Size Impact Chloroplast Genome Size? **Front. Plant Sci**., v. 8, n. 2116, p. 1-6, dec. 2017. doi: 10.3389/fpls.2017.02116

SOUZA, F. X. de. *Spondias* agroindustriais e os seus métodos de propagação**.** Fortaleza: **Embrapa**-CNPAT /SEBRAE/CE, 1998, 26p.

SOUZA, F. X. de.; COSTA, J. T. A.; LIMA, R. N. de; CRISOSTOMO, J. R. Crescimento e desenvolvimento de clones de cajazeira cultivados na chapada do Apodi, Ceará. **Revista Brasileira de Fruticultura**, v. 28, p. 414-420, 2006.

STAATS, C. C.; MORAIS, G. L. de; MARGIS, R. Projetos Genoma. In: VERLI, H. (Org.). **Bioinformática da Biologia à flexibilidade molecular**. São Paulo: SBBq, 2014, cap. 4, p. 63-79.

STRAUB, S. C. K. Horizontal Transfer of DNA from the Mitochondrial to the Plastid Genome and Its Subsequent Evolution in Milkweeds (Apocynaceae). **Genome Biol. Evol., v.** 5, n. 10, p.1872–1885, sep. 2013. doi:10.1093/gbe/evt140

STUPAR, R.M., et al. Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. **Proceedings of the National Academy of Sciences USA**, v. 98, n. 9, p. 5099–5103, apr. 2001. doi: 10.1073ypnas.091110398

SUCHARD, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. **Virus Evolution**, v. 4, n.1, p.1-5, jul. 2018. doi: 10.1093/ve/vey016

SWOFFORD, D. L. PAUP: Phylogenetic Analysis Using Parsimon, **PAUP* 4.0 beta documentation,** Sinauer Associates, 144p. feb. 2002.

TAMURA, K, et al. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. **Molecular Biology and Evolution**, v. 30, n.12, p. 2725-2729, oct. 2013. doi:10.1093/molbev/mst197

TER WELLE, J. H; DETIENNE, P; TERRAZAS, T. Wood and timber. In: Gorts-van Rijn ARA, Jansen-Jacobs MJ (Eds) Flora of the Guianas, Series A (Phanerogams), Anacardiaceae. **Royal Botanic Gardens**, v. 19, p. 48–67, 1997.

TURMEL, M.; OTIS, C.; LEMIEUX, C. The mitochondrial genome of Chara vulgaris: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. **Plant Cell, v.** 15, n. 8, p.1888-1903, aug. 2003. doi: 10.1105/tpc.013169

URICARU, R. et al. Reference-free detection of isolated SNPs. **Nucleic Acids Research**, v. 43, n. 2, p. 1-11, nov. 2015. doi: 10.1093/nar/gku1187

VAN-LUME, J. et al.Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. **Perspectives in Plant Ecology, Evolution and Systematics**, v. 29, p. 51–63, nov. 2017. doi: 10.1016/j.ppees.2017.11.004

VERDE, I. et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. **Nat Genet.**, v. 45, n. 5, p. 487–94, mar. 2013. doi: 10.1038/ng.2586

VERLI, H. O que é bioinformática? In: VERLI, H. (Org.). **Bioinformática da Biologia à flexibilidade molecular**. São Paulo: SBBq, 2014, cap. 1, p. 2-12.

WANG, D. et al. Plastid Sequences Contribute to Some Plant Mitochondrial Genes. **Mol. Biol. Evol,** v. 29, n. 7, p. 1707–1711, jan. 2012. doi: 10.1093/molbev/mss016.

WANG, L. et al. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. **BMC Bioinformatics**, v. 6, n. 60, p. 1-7, mar.2005. doi:10.1186/1471-2105-6-60

WEI, S. et al. Assembly and analysis of the complete *Salix purpurea* L. (Salicaceae) mitochondrial genome sequence. **Springer Plus,** v. 5, n. 1, p.1-10, oct. 2016. doi: 10.1186/s40064-016-3521-6

WENDEL, J. F. et al. Evolution of plant genome architecture. **Genome Biol**., v. 17, n. 37, p. 1–14, mar. 2016. doi: 10.1186/s13059-016-0908-1

WICKER, A unified classification system for eukaryotic transposbale elments. **Nature Reviews Genetics**, v. 8, n. 12, p. 973-982, dec. 2007. doi: 10.1038/nrg2165 ·

WU, Z. et al. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. **Proceedings of the National Academy of Sciences**, v. 112, n. 33, p. 10185-91, may. 2015. doi: 10.1073/pnas.1421397112

YASUNARI, O. et al. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. **Nucleic Acids Research**, v. 33, n. 19, p. 6235-6250, oct. 2005. doi:10.1093/nar/gki925.

YE, N.et al. Assembly and comparative analysis of complete mitochondrial genome sequence of an economic plant Salix suchowensis. **PeerJ**, v. 5, p. 1-27, mar. 2017. doi: 7717/peerj.3148.

ZHANG, B. H. et al. Identification and characterization of new plant microRNAs using EST analysis. **Cell Res**, v**.** 15, n. 5, p. 336-360, may. 2005. doi: 10.1038/sj.cr.7290302

ZHANG, J. et al. SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. **PLoS Computational Biology**, v. 1, n. 5, p. 395- 404, oct. 2005. doi: 10.1371/journal.pcbi.0010053

ZHANG, DE-X.; HEWITT, G. M. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. **Molecular Ecology**, v. 12, n.3, p. 563-584, apr. 2003. doi: 10.1046/j.1365-294X.2003.01773.x

ZANG, T. et al. The Complete Chloroplast and Mitochondrial Genome Sequences of Boea hygrometrica: Insights into the Evolution of Plant Organellar Genomes. **PLOS ONE**, v.7, n.1, jan. 2012. doi:10.1371/journal.pone.0030531

Artigo 01

Complete mitochondrial genomes of the *Spondias tuberosa* Arr. Cam and *Spondias bmombin* L. reveal highly repetitive DNA sequences[2]

Gleica Martins[1], Eliane Balbino[1], André Marques[1], Cicero Almeida[1]

[1]*Laboratory of Genetics Resources, Campus Arapiraca, Universidade Federal de Alagoas, Brazil*

Short running title: *Mitogenomes of Spondias*

Corresponding author:

Cícero Almeida

Campus de Arapiraca

Universidade Federal de Alagoas

Avenida Manoel Severino Barbosa s/n, Rodovia AL 115, km 6,5. Bairro Bom Sucesso

Arapiraca, AL, Brazil

Phone number: 55 823482-1831

E-mail: cicero@arapiraca.ufal.br

[2] Artigo nas normas da revista GENE.

ABSTRACT

Mitogenomes in plants are well-known as exhibiting high diversity in genome size architecture and repetitive DNA sequences. In this research study, we report on the complete mitochondrial genomes of *S. tuberosa* and *S. mombin* using Illumina paired-end and mate-pair end reads. These genomes were obtained by a combination of methods of *de novo* assembly and contig extension. The mitogenomes of *S. tuberosa* and *S. mombin* showed 779,106 bp and 674,156 bp in length, with a total of 74 genes and 68 genes, respectively. Genome comparisons showed many rearrangements that were mediated by repetitive DNA, and also high incorporation of DNA from chloroplast. In summary, we demonstrate: (1) first complete mitochondrial genomes for the genus *Spondias*; (2) the synteny between *S. tuberosa* and *S. mombin* showed rearrangements, mediated by repetitive DNA; (3) that gene content in *Spondias* mitogenomes is highly conserved; and (4) the high incorporation DNA from chloroplast genome.

Keywords: Mitogenomes. Evolution. *Spondias*.

1. Introduction

The plant mitochondrion is an organelle important to produce energy (Ogihara et al., 2005) and its structure is formed by double helix DNA in a chromosome master or master circle configuration. It might also be composed of other small circular chromosomes or sub genomes (Wu et al., 2015; Shearman et al., 2016). The mitochondrial genome (mitogenome) structure is highly diverse in flowering plants, including having sizes of high variability, repetitive DNA sequences and frequency of recombination across large repeats, low gene densities, and low rates of nucleotide substitution (Gou et al., 2017; Gualberto et al., 2014). The plant mitogenome expansion is primarily due to accumulation of repetitive sequences and incorporation of DNA from chloroplast and nuclear genomes (Alverson et al., 2010), resulting in complex genomes that are difficult to be assembled with short paired-end reads sequencing technologies.

The mitogenomes of plants are large and complex compared to animals, of which all animal mitogenomes are about 15-17 kb in length, while plant mitogenome range 200 kb to 11 Mb in length, however, the plant mitogenomes do not contain significantly more genes than animals mitogenomes (Gualberto et al., 2017). The plant mitogenomes have low mutation rates when compared to plastid and nuclear genomes, however, the rearrangements are frequent and is an important characteristic, which can be used for evolutive inferences among close species. The achievement of complete mitogenomes allows for the analysis of rearrangements, incorporation of DNA from chloroplast and nuclear genomes, as well as for description of repeat sequences in recent species, as genus *Spondias*, which had recent diversification (approximately 20 million years ago) (Machado et al., 2015) or intraspecific variation, as described in different accessions of maize, of which revealed important structural changes in the mitogenome

(Allen et al., 2007). For other hand, the abundance of repeat sequences difficult the genome assembly, which is necessary long-read (PacBio) or long insert libraries (mate-pair) to obtain the master circle.

The genus *Spondias* belongs to the family Anacardiaceae, and comprises 20 species from the Neotropical, Asian, and Madagascar regions (Mitchell and Daly, 2015; Machado et al., 2015). In Brazil, the most common species are *S. tuberosa* Arr. Cam (Umbuzeiro), *S. mombin* L. (Cajazeira), *S. purpurea* L. (Sirigueleira), *S. dulcis* Sonn. (Cajaraneira), and e *Spondias venulosa* Mart. Ex Engl (Cajazinho) (Silva et al., 2015)]. Recently, several studies have characterized the genus *Spondias* by means of phylogeography (Balbino et al., 2018), complete chloroplast genome assembly (Santos and Almeida, 2019), SNP analysis (Nobre et al., 2018), and molecular phylogeny (Machado et al., 2015; Silva et al., 2015). Aside from this, the taxonomy of the genus is still not clear and the mitogenomes can help to understand the putative hybrid origin for many species within the genus. However, there is no report to date for complete mitogenomes in the genus. With the advent of next-generation DNA sequencing technologies, there has been an increase in the number of mitogenomes sequenced; however, within the order Sapindales, which comprises 460 genera and 5,670 species, only one mitogenome (*Citrus sinensis* L.) has been completely sequenced (Yu, et al., 2018). This study aimed to sequence the mitogenomes of *S. tuberosa* and *S. mombin* using high-throughput sequencing technology and analyses of repeat abundance and rearrangements.

2. Materials and Methods

2.1. Plant material, DNA isolation, and high-throughput DNA sequencing

Plant material of *S. tuberosa* and *S. mombin* was collected in the state of Alagoas, Brazil, and total DNA was extracted (including nuclear, chloroplast, and mitochondrial DNA) through use of approximately 2 cm² of the leaves following the cetyltrimethylammonium bromide (CTAB) extraction method (Doyle and Doyle, 1987). The quantity and quality of the extracted DNA were verified by visualization on 1% agarose gel. The DNA sample was fragmented mechanically into 400–600 bp to construct the sequencing paired-end library. The fragments were ligated with adapters using the "Nextera DNA Sample Preparation" (Illumina) the 2x100 bp paired-end were sequenced on the Illumina HiSeq2500 platform. Mate pair library was prepared with high molecular weight genomic DNA using Illumina Nextera Mate Pair Sample Preparation Kit, according to the manufacturer's instructions for a gel-free preparation of 2-4 kb effective insert size library and sequenced on the Illumina HiSeq2500 platform. The sequencing was performed at the Central Laboratory for High Performance Technologies in Life Sciences (LacTad-*Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida*) at the State University of Campinas-UNICAMP, São Paulo.

2.2. Mitogenome assembly, annotation, and comparison

The Illumina reads were filtered using BBDuk (implemented in Geneious R9 software) to remove the Illumina adapters, artifacts and to quality-trim. To generate the mitogenomes of *S. tuberosa* and *S. mombin*, a total number of 245,955,843 reads and 136,312,628 reads, respectively, were assembled in the Ray software version 2.3.1 (Doisvert et al., 2012) and SPAdes software version 3.12.0 (Bankevich et al., 2012) (Table 1) on Linux architecture (distribution ubuntu 16.04 LTS). The *de novo* contigs were then manually merged using mate pair reads in the Geneious R9 software. Genome

annotation was achieved using *Citrus sinensis* as the reference by the Geneious annotation tool and was checked with the annotation achieved using Mitofy (Alverson et al., 2010). For annotation using Geneious, a minimum of 70% identity cutoff between the genomes was considered. The annotations were individually checked and, where necessary, were manually corrected for start and stop codons. A graphic representation of the mitogenomes was created using Organellar Genome DRAW (Lohse et al., 2013). Comparison among *Spondias* mitogenomes and chloroplast genomes was performed using AliTV (Ankenbrand et al., 2017) and repeat analysis was achieved by Circoletto (Darzentas et al., 2010).

2.3. Microsatellites analysis

The mitogenomes were analyzed using Phobos software (Mayer, 2010) implemented from Geneious R9, searching for di, tri, tetra, penta, and hexa- and hepta-nucleotide motifs, with a minimum of five repeats. Seventeen microsatellites were selected widespread in the mitogenome of *S. tuberosa* (Fig. S2), and primers (Table S1) were designed using Primer3 software (Koressaar et al., 2007) for polymorphism analysis in 20 of the individuals of *S. tuberosa*. For primer design, we used the following criteria: annealing temperature of 50–60 °C; a maximum difference of 2 °C between primer pairs; and a GC content of between 40% and 60%. PCR amplification from 20 samples was performed using a volume of 50 µl containing 5 µl of a reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 1 U Taq DNA polymerase, 0.5 µM of each primer, and 200 ng of DNA. Amplification was achieved with an initial denaturation of 94 °C for 3 min, followed by 35 cycles at 94 °C for 30 s, annealing at 55–60 °C for 30 s, and final extension at 72 °C for 10 min. The PCR reactions were performed in a

BioCycler thermocycler and the PCR products were subjected to electrophoresis on 3% acrylamide gel to confirm amplification.

3. Results

3.1. Mitogenome assembly

De novo contigs using 24,6 Gb for *S. tuberosa* and 13,6 Gb for *S. mombin* (Table 1) resulted in eight scaffolds for *S. tuberosa* and seven scaffolds for *S. mombin*. The contigs were merged by mate-pair reads in the circular chromosomes corresponding to the circular mitogenomes of *S. tuberosa* and *S. mombin.* The mitogenomes showed 779,106 bp and 674,156 bp for *S. tuberosa* and *S. mombin*, respectively (Fig. 1). Both mitogenomes consisted of several repeated regions, which are in greatest abundance in *S. tuberosa*, with large repeats (> 150kbp), while *S. mombin* showed minor repeat blocks (Figs. 2A and 2B). A total of 74 genes were identified in *S. tuberosa* mitogenome, including 30 protein-coding genes, 39 tRNAs, and four rRNAs. For *S. mombin*, we identified a total of 68 genes, including 30 protein-coding genes, 34 tRNAs, and three rRNAs (Table S2).

3.2. Mitogenomes Evolution

The chloroplast and mitochondrial genomes of both species have high and widespread homologies which are most likely as a result of horizontal sequence transfer, of which the *S. tuberosa* mitogenome incorporated more chloroplast DNA than did the *S. mombin* mitogenome (Figs. 2C and 2D), resulting in extensive repetitive sequences in both mitogenomes. We observed that most repeat regions showed high coverage (Figs. 2A and 2B) because of the similarity with chloroplast sequences (Figs. 2C and 2D). Both *Spondias* mitogenomes shared high sequence homology with each

other (approximately 85%), besides the great extensive rearrangements observed (Fig. 3A). When the *Spondias* mitogenomes were compared with *Citrus sinensis* mitogenome, minor sequence homologies and high rearrangements were observed (Figs. 3B and 3C), which indicated low synteny among these mitogenomes.

3.4. Microsatellites analysis

SSR analysis showed an extensive number of di- to hepta-nucleotides in both *Spondias* mitogenomes (Fig. S1), of which all motifs were found in both mitogenomes, except for pentanucleotide in the *S. tuberosa* (Fig. S1). Among microsatellites motifs, dinucleotides were the most frequent in both mitogenomes, whereas in *S. tuberosa* the average number of SSR motifs was more abundant than in *S. mombin*. Furthermore, we tested these SSRs for intraspecific polymorphism, but the results revealed no polymorphism, indicating no intraspecific variation in these loci. However, *in silico* analysis between *S. tuberosa* and *S. mombin* showed polymorphism for four microsatellites loci (Fig. 4).

4. Discussion

The mitogenomes have been characterized by repeat abundance, incorporation of chloroplast DNA, and extensive rearrangements, which hinder the genome assembly (Shearman et al., 2016; Silva et al., 2017). Generally, mitogenome *de novo* assemblies need long insert libraries like mate pair or PacBio reads to manually merge the scaffolds. In the present study, we demonstrated the first *de novo* assembly for two mitogenomes of the genus *Spondias* and provided a robust analysis for the evolution of mitogenomes in the Sapindales order using a combined set of single-end and mate-pair Illumina short read libraries. Our results revealed that the mitogenomes of *S. tuberosa*

and *S. mombin* show many rearrangements, possibly due to the high abundance of repeat found and the high incorporation of chloroplast DNA. The incorporation chloroplast DNA has been reported in the other studies, as in the *Rhazya stricta* (Park et al., 2014) and *Saccharum officinarum* (Shearman et al., 2016). The results suggest that incorporation of chloroplast DNA may contribute to mitogenome expansion in the genus *Spondias*.

Plant mitogenomes are characterized by many rearrangements and show fast evolution, in which even very closely-related plant species show differences in their genome structure (Tang et al., 2015). Indeed, this conclusion is observed in the present study as we observed that both mitogenomes of the close species of *Spondias* showed notable differences in their genome structure, regardless of the recent diversification that *Spondias* has experienced (Muellner-Riehl et al., 2016). Although the two *Spondias* mitogenomes showed high sequence identity, the main differences were observed in the gene order and repeat content incorporated by chloroplast DNA insertion agreeing with the plant pattern for mitogenomes evolution (Gualberto et al., 2014). High conservation of sequence homology was observed in the 17 microsatellites loci; this polymorphism was not detected in the 20 individuals of *S. tuberosa*, suggesting high intraspecific sequence conservation, and only four loci showed polymorphism between *S. tuberosa* and *S. mombin*. Notably, these microsatellites loci can be used rather for species identification instead of population genetics, in which fragment analysis is very fast and cheap when compared with sequence data that have been determined for *Spondias* (Balbino et al., 2018).

Recent phylogenies based on chloroplast (Silva et al., 2015) and nuclear sequences (Machado et al., 2015; Nobre et al., 2018) were reported for *Spondias*, which shows *S. tuberosa* in the derived clade while *S. mombin* is shown in the basal clade. The

nuclear genome content determined by flow cytometry showed that species in the derived clade has genomes greater than basal species, suggesting the increase in the genome size in derived species. This trend was also observed in the mitogenomes of *Spondias*, wherein the *S. tuberosa* mitogenome is greater than the *S. mombin* mitogenome, which suggests that mechanisms of genomes increasing may act in both nuclear and mitochondrial genomes. In summary, we demonstrate: (1) first complete mitochondrial genomes for the genus *Spondias*; (2) that the synteny between *S. tuberosa* and *S. mombin* revealed notable rearrangements; (3) gene content and sequences identity in *Spondias* mitogenomes are highly conserved; and (4) the incorporated chloroplast DNA contributes to the increase of mitogenome size.

Author contributions

Martins G performed the DNA library, de novo assembles and wrote the manuscript. Balbino E, performed the microsatellites analysis. Marques A and Almeida C analyzed data and wrote the manuscript

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Federal University of Alagoas for the laboratories and scientific support and the Fundação de Apoio à Pesquisa de Alagoas (FAPEAL) for funding this Project, and the National Council for the Improvement of Higher Education (CAPES).

References

Koressaar, T., Remm, M., 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23, 1289-1291. https://doi.org/10.1093/bioinformatics/btm091

Allen, J.O., Fauron, C.M., Minx, P., Roark, L., Oddiraju, S., Guan, N.L., Meyer, L., Sun, H., Kim, K., Wang, C., Du, F., Xu, D., Gibson, M., Cifrese, J., Clifton, S.W., Newton, K.J., 2007. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 177, 1173–1192. https://doi.org/10.1534/genetics.107.073312

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K., Palmer, J.D., 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol. Biol. Evol. 27, 1436-48. https://doi.org/10.1093/molbev/msq029

Ankenbrand, M.J., Hohlfeld, S., Hackl, T., Förster, F., 2017. AliTV — interactive visualization of whole genome comparisons 1–10. https://doi.org/10.7717/peerj-cs.116

Balbino, E., Caetano, B., Almeida, C., 2018. Phylogeographic structure of Spondias tuberosa Arruda Câmara (Anacardiaceae): seasonally dry tropical forest as a large and continuous refuge.

Bankevich, A., Nurk, S., Pevzner P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J., 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13, R122. https://doi.org/10.1186/gb-2012-13-12-r122

Darzentas, N., 2010. Circoletto: Visualizing sequence similarity with Circos. Bioinformatics 26, 2620–2621. https://doi.org/10.1093/bioinformatics/btq484

Doyle, J.J., Doyle, J.L., (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19, 11-15.

Gualberto, J.M., Mileshina, D., Wallet, C., Niazi, A.K., Weber-Lotfi, F., Dietrich, A., 2014. The plant mitochondrial genome: Dynamics and maintenance. Biochimie 100, 107–120. https://doi.org/10.1016/j.biochi.2013.09.016

Gualberto, J.M., Newton, K.J., 2017. Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu. Rev. Plant Biol. 68, 225–252. https://doi.org/10.1146/annurev-arplant-043015-112232

Guo, W., Zhu, A., Fan, W., Mower, J.P., 2017. Complete mitochondrial genomes from the ferns Ophioglossum californicum and Psilotum nudum are highly repetitive with the largest organellar introns 391–403.

Lohse, M., Drechsel, O., Kahlau, S., Bock, R., 2013. OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 41, 575–581. https://doi.org/10.1093/nar/gkt289

Machado, M., Carvalho, P., van den Berg, C., 2015. Domestication, hybridization, speciation, and the origins of an economically important tree crop of Spondias (Anacardiaceae) from the Brazilian Caatinga dry forest. Neodiversity 8, 8–49.

Mayer, C., 2008. Phobos, a tandem repeat search tool for complete genomes, Version 3.3.12, Available from: http://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.

Mitchell, J.D., Daly, D.C., 2015. A revision of Spondias (Anacardiaceae) in the Neotropics. Phytotaxa 92, 1–92. https://doi.org/10.3897/phytokeys.55.8489

Muellner-Riehl, A.N., Weeks, A., Clayton, J.W., Buerki, S., Nauheimer, L., Chiang, Y.C., Cody, S., Pell, S.K., 2016. Molecular phylogenetics and molecular clock dating of Sapindales based on plastid rbcL, atpB and trnL-trnF DNA sequences. Taxon 65, 1019–1036. https: //doi.org/10.12705/655.5

Nobre L.L.M., D.M., Daniel, J., Leite, R., Almeida, C., 2018. Phylogenomic and single nucleotide polymorphism analyses revealed the hybrid origin of Spondias bahiensis (family Anacardiaceae): de novo genome sequencing and comparative genomics.

Ogihara, Y., Yamazaki, Y., Murai, K., Kanno, A., Terachi, T., Shiina, T., Miyashita, N., Nasuda, S., Nakamura, C., Mori, N., Takumi, S., Murata, M., Futo, S., Tsunewaki, K., 2005. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res. 33, 6235–6250. https://doi.org/10.1093/nar/gki925

Park, S., Ruhlman, T.A., Sabir, J.S.M., Mutwakil, M.H.Z., Baeshen, M.N., Sabir, M.J., Baeshen, N.A., Jansen, R.K., 2014. Complete sequences of organelle genomes

from the medicinal plant Rhazya stricta (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. BMC Genomics 15, 1–18. https://doi.org/10.1186/1471-2164-15-405

Santos, V., Almeida, C., 2019. The complete chloroplast genome sequences of three Spondias species reveal close relationship among the species. Genetics and Molecular Biology 138, 132–138. https://doi.org/10.1590/1678-4685-GMB-2017-0265

Shearman, J.R., Sonthirod, C., Naktang, C., Pootakham, W., Yoocha, T., Sangsrakru, D., Jomchai, N., Tragoonrung, S., 2016. The two chromosomes of the mitochondrial genome of a sugarcane cultivar : assembly and recombination analysis using long PacBio reads. Nat. Publ. Gr. 1–7. https: //doi.org/10.1038/srep31533

Silva, J.N., Bezerra da Costa, A., Silva, J.V., Almeida, C., 2015. DNA barcoding and phylogeny in neotropical species of the genus Spondias. Biochem. Syst. Ecol. 61, 240–243. https://doi.org/10.1016/j.bse.2015.06.005

Silva, S.R., Alvarenga, D.O., Aranguren, Y., Penha, H.A., Fernandes, C.C., Pinheiro, D.G., Oliveira, M.T., Michael, T.P., Miranda, V.F.O., Varani, A.M., 2017. The mitochondrial genome of the terrestrial carnivorous plant Utricularia reniformis (Lentibulariaceae): Structure, comparative analysis and evolutionary landmarks. PLoS One 12, e0180484. https://doi.org/10.1371/journal.pone.0180484

Stupar, R.M., Lilly, J.W., Town, C.D., Cheng, Z., Kaul, S., Buell, C.R., Jiang, J., 2001. Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: Implication of potential sequencing errors caused by large-

unit repeats. Proc. Natl. Acad. Sci. 98, 5099–5103.https://doi.org/10.1073/pnas.091110398

Tang, M., Chen, Z., Grover, C.E., Wang, Y., Li, S., Liu, G., Ma, Z., Wendel, J.F., Hua, J., 2015. Rapid evolutionary divergence of Gossypium barbadense and G. hirsutum mitochondrial genomes. BMC Genomics 16, 1–16. https://doi.org/10.1186/s12864-015-1988-0

Wu, Z., Cuthbert, J.M., Taylor, D.R., Sloan, D.B., 2015. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes 112, 10185–10191. https://doi.org/10.1073/pnas.1421397112

Yu, F., Bi, C., Wang, X., Qian, X., Ye, N., 2018. The complete mitochondrial genome of Citrus sinensis. Mitochondrial DNA Part B Resour. 3, 592–593. https://doi.org/10.1080/23802359.2018.1473738

Legends figures

Figure 1. Genome map of *Spondias* mitogenomes. (A) Circular chromosome map for S. *tuberosa,* and (B) Circular chromosome map for *S. mombin*. In the both maps, the genes on the inside of outer circles are transcribed in a clockwise direction, while genes on the outside of outer circles are transcribed in an anticlockwise direction. Genes belonging to the same protein complex are in the same color as described in the legend.

Figure 2. Comparative analyses of mitogenomes. (A) satellite regions in the genome of *S. tuberosa,* (B) satellite regions in the mitogenomes of *S. mombin*, (C) sharing of regions between cpDNA and mtDNA of *S. tuberosa*, (D) comparison between the mitogenoma *S. mombin* and cpDNA. The inner circle reveals the distribution of repeats in two mitogenomes, with curved lines and ribbons connecting pairs of repeats and proportional width to repeat size. The blue lines represent the repeats distributed in the genome (A, B). The green ribbons represent similarly high regions between cpDNA and mtDNA (C, D). The red lines represent regions with similarly low and intermediate colors representing regions that have undergone some change.

Figure 3. Dynamic of rearrangements between three species of Sapindales mitochondrial genomes, with curved ribbons connecting pairs of syntenic blocks and width proportional to block size. (A) *S. tuberosa* vs. *S. mombin*; (B) *C. sinensis* vs. S. *tuberosa;* and (C) *C. sinensis* vs. *S. mombin*.

Figure 4. Distribution of the microsatellite in the mitogenoma of *S. tuberosa* and *S. mombin* based on: (A) number of SSR motifs; (B) number of repeats

Figure S1. Microsatellites in the mitogenoma of *S. tuberosa* and *S. mombin* according to SSR motifs

Figure S2. Distribution of 70 microsatellites in the mitogenome of *S. tuberosa*

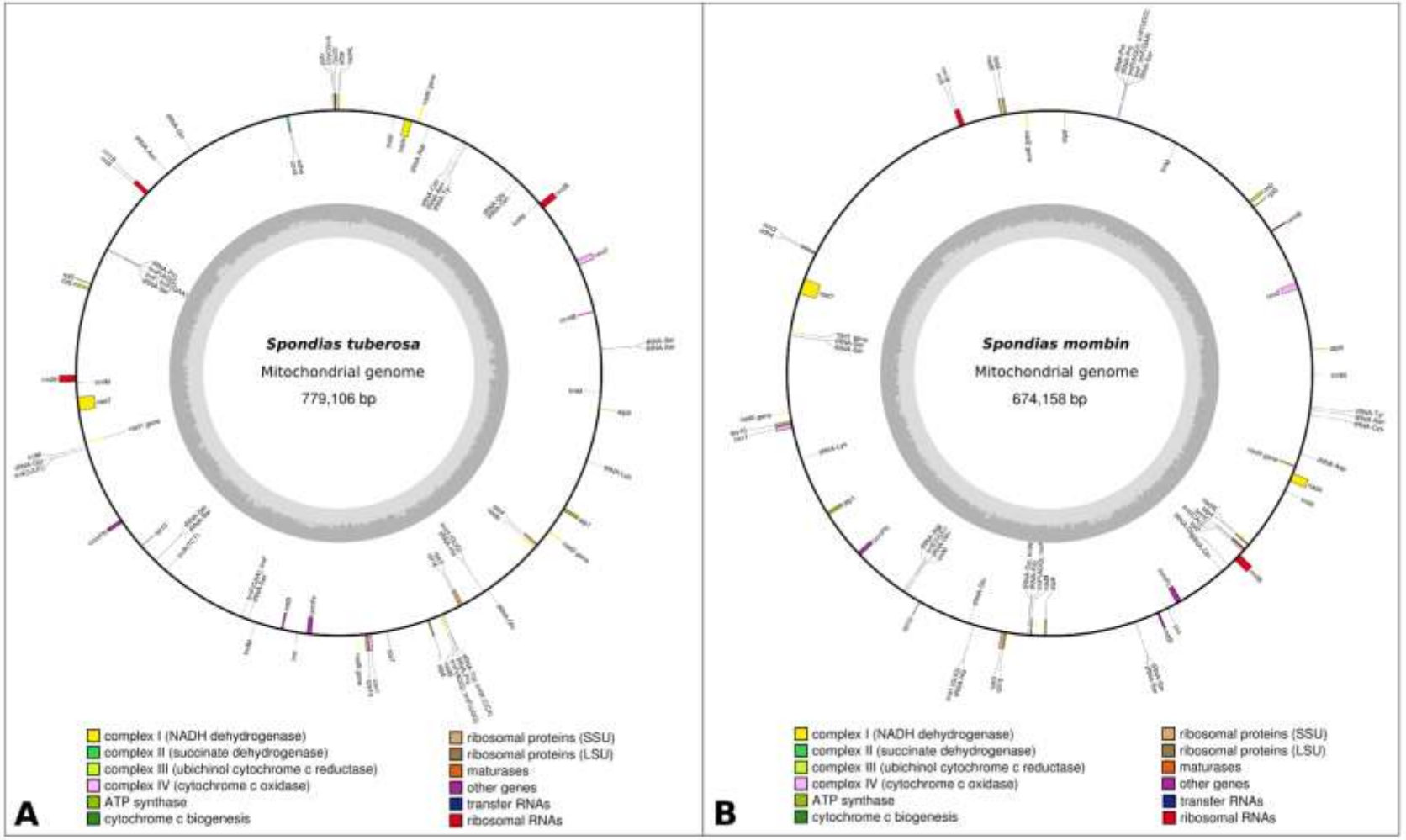


Figure 1. Genome map of *Spondias* mitogenomes. (A) Circular chromosome map for S. *tuberosa* and (B) Circular chromosome map for *S. mombin*. In the both maps, the genes on the inside of outer circles are transcribed in a clockwise direction, while genes on the outside of outer circles are transcribed in a reverse direction. Genes belonging to the same protein complex are in the same color as described in the legend.

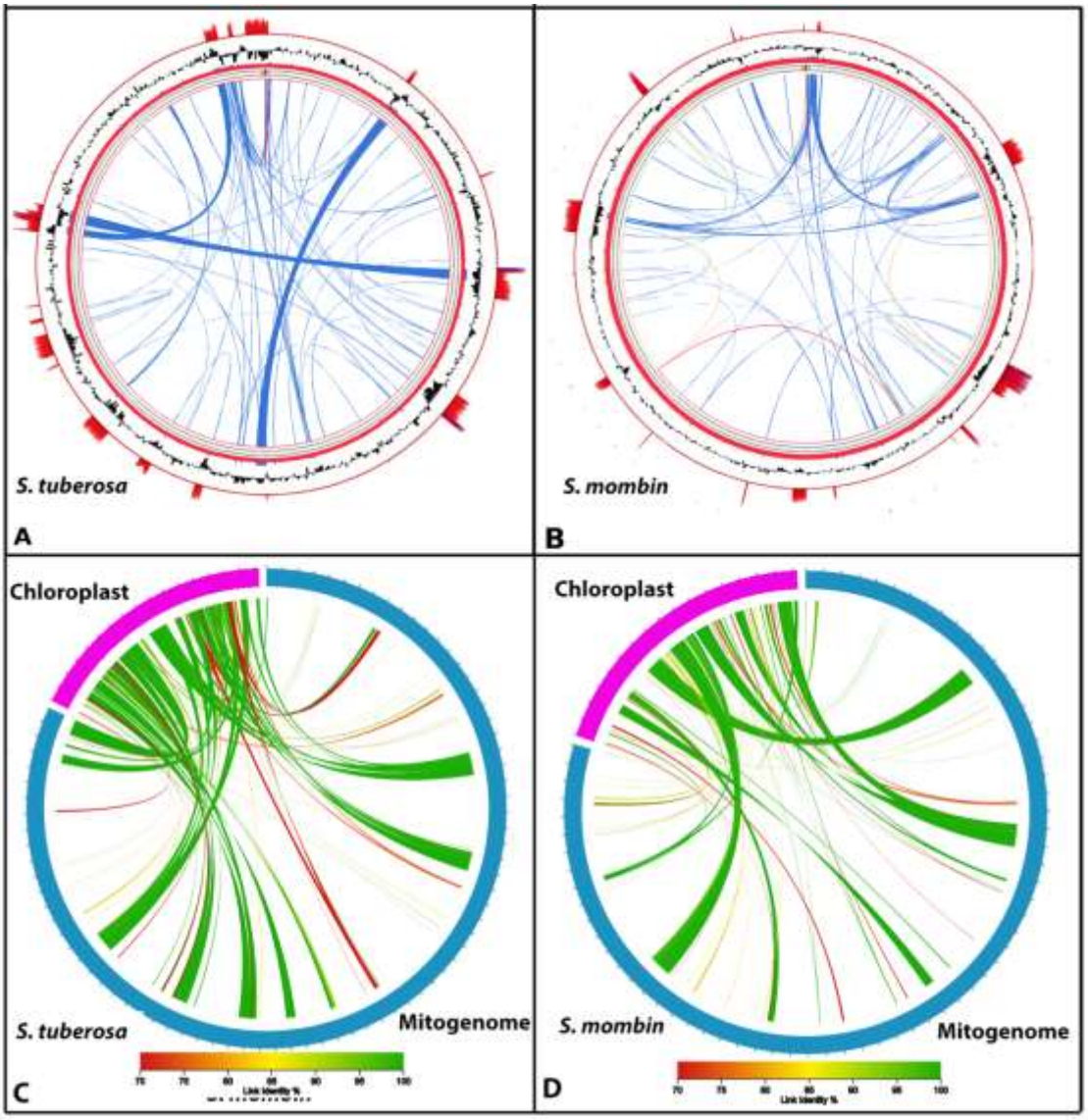


Figure 2. Comparative analyzes of mitogenomes. (A) satellite regions in the genome of *S. tuberosa* (B) satellite regions in the mitogenoma of *S. mombin* C) sharing of regions between cpDNA and mtDNA of *S. tuberosa* D) comparison between the mitogenoma *S. mombin* and cpDNA. The inner circle reveals the distribution of repeats in two mitogenomes, with curved lines and ribbons connecting pairs of repeats and proportional width to repeat size. The blue lines represent the repeats distributed in the genome (A, B). The green ribbons represent represents similarly high regions between cpDNA and mtDNA (C, D). The red lines represent regions with similarity low and intermediate colors represent regions that have undergone some change.

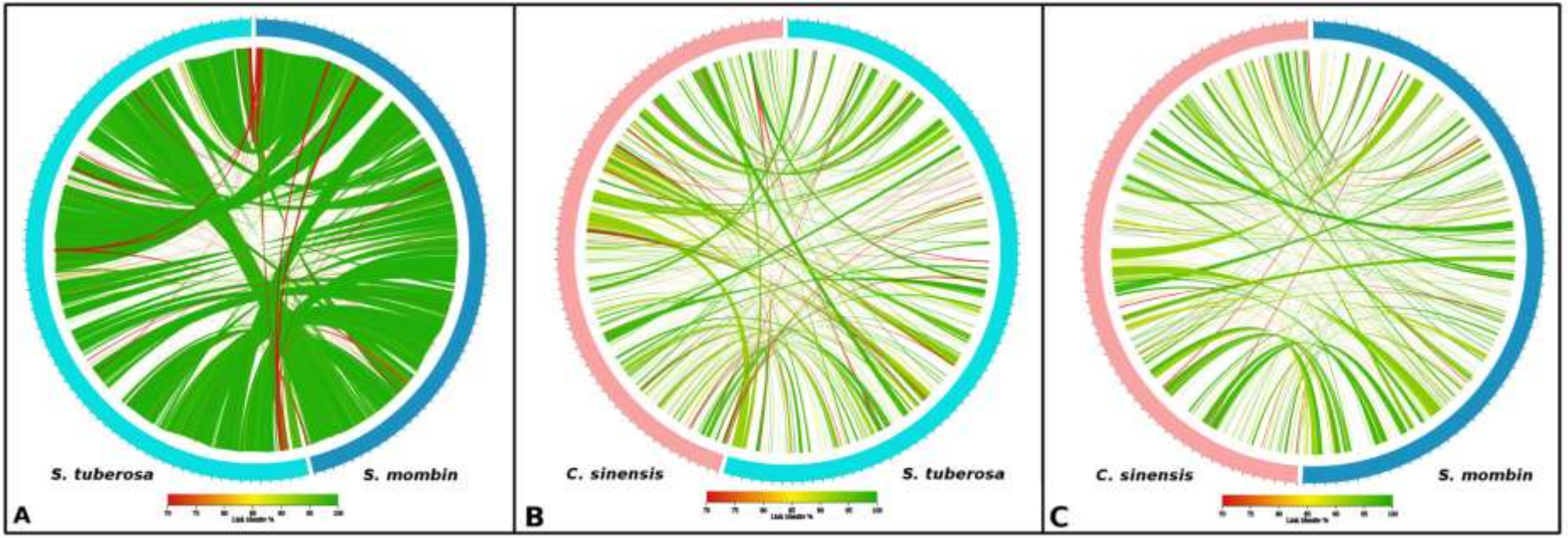


Figure 3. Dynamic of rearrangements between three species Sapindales mitochondrial genomes, with curved ribbons connecting pairs of syntenic blocks and width proportional to block size. The numbers give genome coordinates in kilobases, a) *S. tuberosa* versus *S. mombim*, b) *C. sinensis* versus S. *tuberosa* and c) *C. sinensis* versus *S.mombin*.

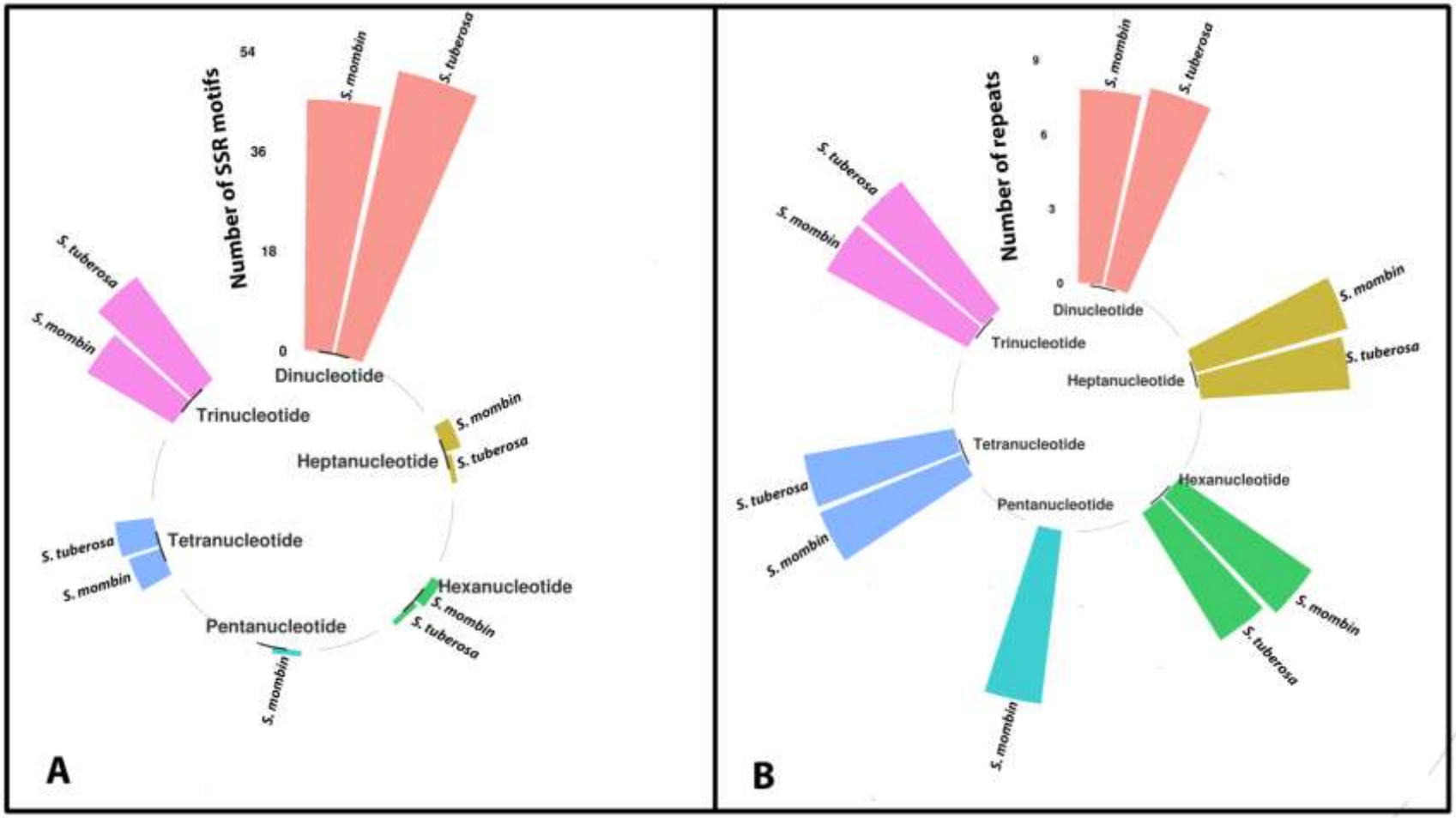


Figure 4. Distribution of the microsatellite in the mitogenoma of *S. tuberosa* and *S. mombim* based on A) number of SSR motifs B) number of repeats.

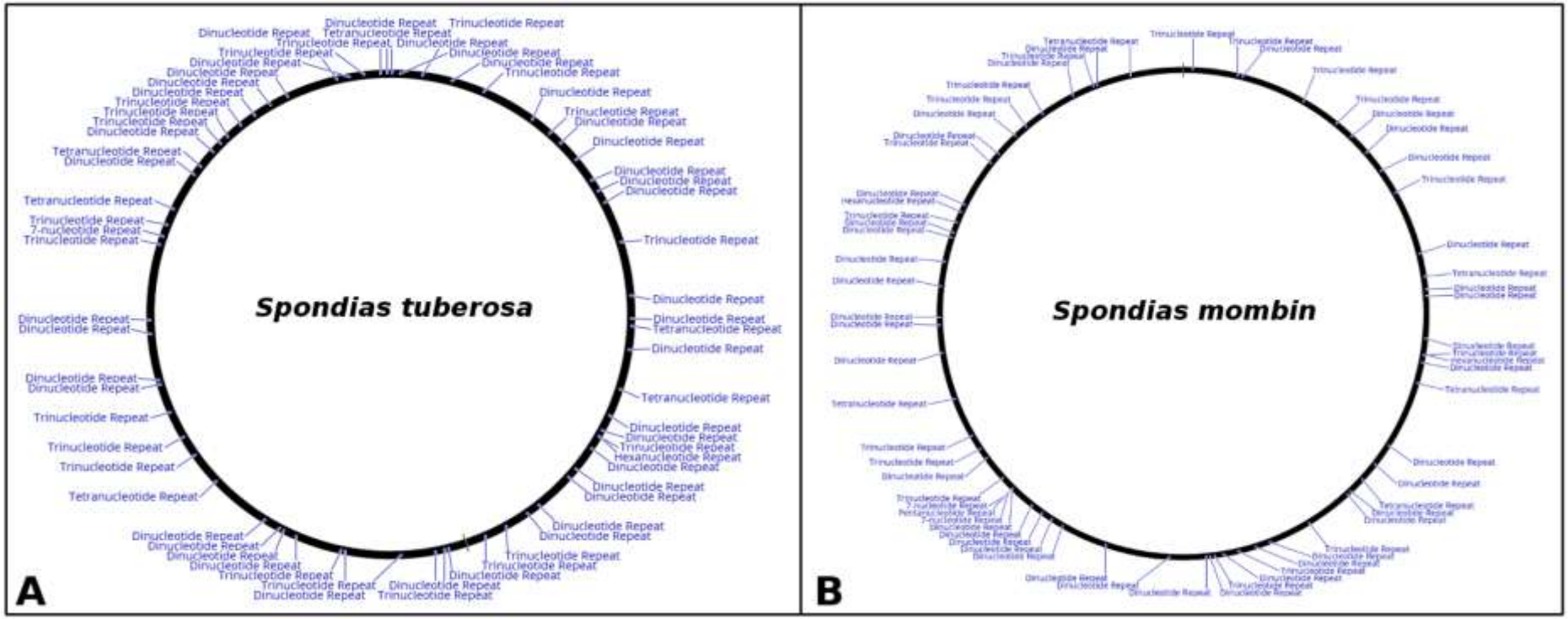


Figure S1. Microsatellites in the mitogenoma of *S. tuberosa* and *S. mombim* according to SSR motifs.

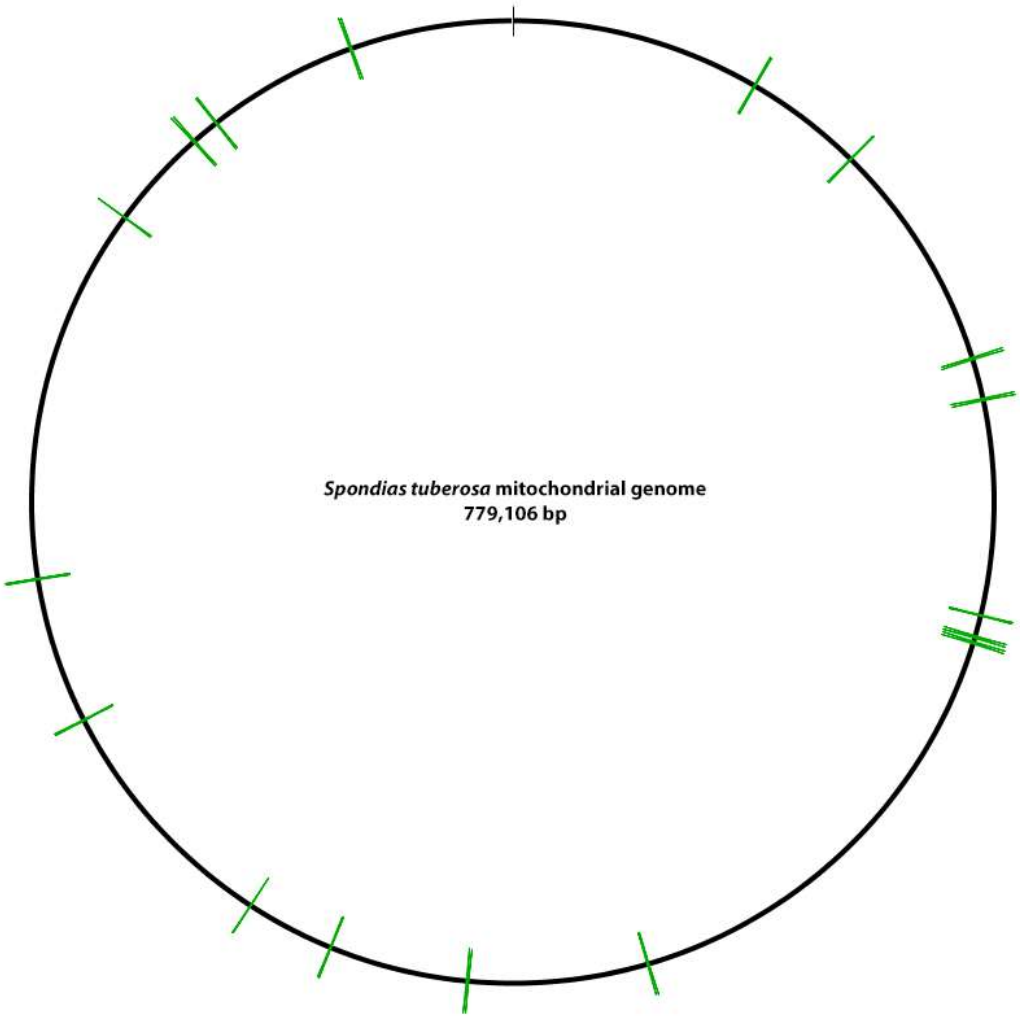


Figure S2. Distribution of 70 microsatellites in the mitogenome of *S. tuberosa*.

Table 1. Summary of Illumina sequencing and assembling

Description	*Spondias tuberosa*	*Spondias mombin*
1.Number of single-end reads	152,638,887	-
2.Number of paired-end reads	28,132,784	18,248,064
3. Number of Mate-pair (2-4kb)	19,221510	22,755,148
4. Number of Mate-pair (2-4kb)	45,962,662	95,309,416
Reads length (nt)	100	100
Total reads assembled	245,955,843	136,312,628
Total length sequenced (Gb)	24,6	13,63
Coverage	47.94x	28.57

Table S1. Primer sequence (F, forward and R, reverse), motifs, number of repetitions, allele size (bp), number of alleles (N_A) test for 17 mitochondrial microsatellites in *Spondias tuberosa*.

Loci	Primer sequence	Motifs	Allele size (bp)
tub4	F: CCGACCGAGTAGGATCCCTA R: CGATCTGCGATCCACTGGAA	(TA)19	191
tub6	F: AGAGAGCCCCTCCCTGTAAG R: AAGTGCTTGATGACTCGGCA	(TA)11	279
tub8	F: TGAAATGCGACCGAGAAGCT R: TAGGAGGGCACGGTTAAGGA	(GA)9	212
tub10	F: GAATAAGCAGCGCTCTTGGC R: TTCTCGACGCAGGAGATTCG	(CT)9	229
tub14	F: ACTTTTTGGGCCAGCTCAGT R: TGGACTGGACACGTGTTAGC	(GTT)9	273
tub18	F: CCCATATCGAGAAAGGGCCC R: GATCATGGCATGTGTCACGC	(AC)7	203
tub22	F: TGTCTTGATTGGCAGTCGCT R: TCCGGATCTCGTTGTTTGGG	(AAG)8	250
tub26	F: TCTCCCTCGACCTGCAAGTA R: TCCAATGGCCAATTCACCACT	(TAAG)15	252
tub29	F: GTACACAGTGCCCCACAACT R: TTTGATTGCTCCCGCTTCCT	(TA)11	241
tub38	F: TTTATCTTCAGGCTGCCGCA R: AGGGCCTTTAAAGAGCGTGA	(AT)15	263
tub41	F: CGCATGCATGTGTCATCACC R: TCAGACGTTTCCACTCCTGC	(AAG)7	221
tub43	F: TCAATCAATTGGTCCGCGGA R: GATTCGGATAGGGGAGCAGC	(AT)12	272
tub45	F: GAAAGCGTACCGGCAGTTTG R: TGCTCACTTCGGTAAGAGCG	(AT)6; (TAA)17; (AT)11	262
tub46	F: GTTCAGGAGCGAGCAAGACT R: ACCCACCATACCTTGCTTCG	(AT)12	244
tub47	F: TAGCCGGTTGCAAATCCACT R: TTCCTTCTTGTGGCTGGGTG	(TC)9	193
tub72	F: GCTCCGACCCAGCAAGTAAT R: ACGGACGTGGCCAAAACATA	(AAGA)6	276

Table S2. Genes in the mitochondrial genome of *Spondias*.

Gene function	**Gene name**	
	S. tuberosa	*S. mombin*
Complex I	nad1, nad2, nad4 (2x), nad4L, nad5, nad6, nad7, nad9	nad1, nad2, nad4 (2x), nad4L, nad5, nad6, nad7, nad9
Complex II	sdh4	sdh4
Complex III	cob	cob
Complex IV	cox1, cox2, cox3	cox1, cox2, cox3
ATP synthase	atp1, atp4, atp6, atp9 (2x)	atp1, atp4, atp6, atp9 (2x)
Cytochrome -c biogenesis	ccmB, ccmC, ccmFC, ccmFN	ccmB, ccmC, ccmFC, ccmFN
Ribosomal protein small (SSU)	rps3, rps4, rps7, rps10	rps3, rps4, rps10
Ribosomal protein large (LSU)	rpl2, rpl5, rpl10, rpl16	rpl2, rpl5, rpl10, rpl16
SecY-independent transport	mttb	mttb
Transfer RNAs	tRNA-Ser(4x), tRNA-Tyr, tRNA-Trp, tRNA-Pro(2x), tRNA-Phe(2x), tRNA-Met (6x), tRNA-Lys, tRNA-Ile(2x), tRNA-His, tRNA-Gly, tRNA-Glu(3x), tRNA-Cys, tRNA-Asp(3x), tRNA-Arg	tRNA-Ser(5x), tRNA-Tyr, tRNA-Trp, tRNA-Pro(2x), tRNA-Met (6x), tRNA-Lys, tRNA-Ile(2x), tRNA-His, tRNA-Gly, tRNA-Glu(3x), tRNA-Cys, tRNA-Asp(3x)
Ribossomal RNAs	rrn5, rrn18, rrn26(2x)	rrn5, rrn18, rrn26

Artigo 02

Genome size and cytomolecular analysis revealed a high similarity of satellite DNA among *Spondias* species (Anacardiaceae)

Gleica Martins[1], José Daniel Oliveira dos Santos[1], André Marques[1], Gustavo Souza[2] e Cícero Almeida[1]

[1]Laboratório de Recursos Genéticos, Arapiraca Campus, Universidade Federal de Alagoas, Brazil

[2]Laboratório de Citogenética e Evolução, Universidade Federal de Pernambuco, Brasil

Autor para correspondência:

Cicero de Almeida

Arapiraca Campus

Universidade Federal de Alagoas

Avenida Manoel Severino Barbosa s/n, Rodovia AL 115, km 6,5. Bairro Bom Sucesso

Arapiraca, AL, Brazil

Número de telefone: 55 823482-1831

E-mail: cicero@arapiraca.ufal.br

Abstract

The genus *Spondias* belongs to the Anacadiaceae family and comprises 20 species, of which 10 are native to Brazil. *Spondias* is characterized by having economically and ecologically important species. Cytogenetic studies using chromosome banding and genomic *in situ* hybridization have suggested that the genomes of these species are similar in their repetitive DNA, but cytogenomic studies are necessary for further understanding of the genomic evolution of the genus. The aim of study was the characterization of the genome of *Spondias* species by genome size estimates and repetitive DNA. The genome size estimates were obtained by flow cytometry and repetitive DNA was accessed using low coverage sequencing from Illumina single-end and paired-end reads, following the satellite DNAs being characterized by fluorescent *in situ* hybridization. The genomic size estimates showed sizes of 460 to 530 Mb (2C values ranging from 0.98 to 1.08 pg), indicating small genome sizes for the *Spondias* genus. The cluster analysis using single-end and paired-end reads revealed that the fraction of repetitive DNA ranged from 31.02% (*Spondias bahiensis*) to 49.66% (*Spondias dulcis*). Two satellite DNAs were found in all species and another two satellites were specific for some species. The first two satellites corresponded to terminal constitutive heterochromatin, and one of them, the SatSpo3 satellite, corresponded to the cromomicin A3 bands that are highly represented in the genus.

Keywords: cytogenomic, repetitive DNA, genomic evolution

Introduction

The genome size has been shown to be an important genome component; it is associated with the amount of nuclear DNA and has been correlated with the amount of repetitive DNA components. The amount of DNA in the nuclear genome has been described as a C-value paradox (1C values are measured in picograms, with 1 pg equivalent to 978 Mb) and is an important approach to evolutionary studies (Kelly and Leitch, 2011; Pellicer et al., 2018).

Repetitive DNA is an important component of eukaryotic genomes that may account for more than 90% of the genome size (Kazazian Jr., 2004) and includes diverse families of mobile elements and satellite DNA (satDNA) (Macas et al., 2015). The mobile elements are the most abundant component of plant genomes and can be found in the retrotransposons or Class I ("copy and paste") and transposons or Class II. Class I is mainly composed of long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons and Class II is characterized by a mechanism of "cut-and-paste" and non-autonomy (Rebollo et al., 2012, Wicker et al., 2007).

satDNA is characterized by long arrays of tandemly arranged units (monomers), spanning up to megabases in length and are frequently associated with blocks of heterochromatin in the nuclear genomes of plants (Macas et al., 2002). satDNA has a high evolutionary rate in sequence diversity and most satellite repeats have many families and may be species or genus-specific (Macas et al., 2002). Genomic size estimates allow access to important traits of the genome and are extensively used in evolution studies (Kelly et al., 2015; Pellicer et al., 2018) and with the advent of high throughput sequencing have been used to provide better understanding of genome

evolution, inferring the evolution of satellite DNAs, ribosomal DNA, and mobile elements (Kelly and Leitch, 2011).

The genus *Spondias* is the most important one in the Anacardiaceae family and comprises 20 species, of which 10 occur in Neotropical regions (Mitchel and Daly, 2015). Some species are characterized by being economically and ecologically important because their fruits are utilized as food by humans and animals. Among the species that occur in Brazil, *S. bahiensis* was recently described by Machado et al. (2015) and other species include are *S. mombin* L. (cajá), *S. tuberosa* Arruda Câmara (umbu), *S. dulcis* L. (cajarana), and *S. purpurea* Parkinson (siriguela). The newly described species *S. bahiensis* has been reported to be a natural hybrid and this hypothesis has been corroborated by molecular studies (Machado et al., 2015; Nobre et al., 2018). Phylogenetic studies have shown close relationships among species, with *S. tuberosa*, *S. bahiensis*, and *S. venulosa* being more derived, while *S. purpurea*, *S. dulcis*, and *S. mombin* are found in the first diverging lineage (Machado et al., 2015; Nobre et al., 2018; Silva et al., 2015).

Cytogenetic study of six *Spondias* species showed 2n = 32 chromosomes and constitutive heterochromatin rich in CG content, while genomic *in situ* hybridization across species showed high sharing of repetitive DNA (Almeida et al. 2007). In a similar way, genomic studies including genome size estimates and characterization of repetitive fractions allow for further understanding of the genomic evolution of *Spondias* genus. In the present study, we investigated the questions: (1) What are the genome size estimates for species in the genus *Spondias*? (2) What are the components and amounts of the repetitive fractions in the genome of species in the genus *Spondias*? And (3) What are the characteristics of the satDNA in *S. tuberosa*?

Materials and methods

Genome size estimation using flow cytometry

For flow cytometry, a suspension of nuclei from young leaves was prepared as described by Loureiro *et al.* (2007) using WPB buffer. The genome sizes were estimated using a CyFlow SL flow cytometer (Partec, Görlitz, Germany). The final DNA content for each accession was calculated based on at least three different measurements. As an internal control, young leaves of Glycine max cv "Polanka" (2C = 2.50 pg DNA) (Doležel and Greilhuber, 2010) were used. The software FloMax (Partec) was used for data processing.

Plant material, DNA isolation, and high-throughput DNA sequencing

Plant material from *S. tuberosa*, *S. bahiensis, S. dulcis*, *S. mombin*, and *a* hybrid (*S. tuberosa* × *S. bahieneis*) were collected in the state of Alagoas, Brazil, and the total DNA was extracted (including nuclear, chloroplast, and mitochondrial DNA) through use of approximately 2 cm^2 of the leaves following the cetyltrimethylammonium bromide (CTAB) extraction method. The quantity and quality of the extracted DNA was verified by visualization on 1% agarose gels. The DNA samples were fragmented using a mechanical mechanism into 400–600 bp to construct the sequencing paired-end library. The fragments were ligated with adapters using the "Nextera DNA Sample Preparation" (Illumina) according to the manufacturer's instructions and the 100 nt single-end and 2 × 100 bp paired-ends were sequenced on the Illumina HiSeq2500 platform. The sequencing was performed at the Central Laboratory for High Performance Technologies in Life Sciences (LacTad-*Laboratório Central de*

Tecnologias de Alto Desempenho em Ciências da Vida) at the State University of Campinas-UNICAMP, São Paulo.

Graph-based clustering of DNA reads

The reads were trimmed using the BBDuk and were used as input for comparative graph-based clustering with RepeatExplorer (Novak et al., 2010) and with Tandem Repeat Analyzer (TAREAN) software (Novák et al., 2017), implemented within the Galaxy environment (http://repeatexplorer.umbr.cas.cz/) for repeat identification. The RepeatExplorer analysis allowed us to detect the genomic proportion of the repetitive DNA while the TAREAN is a computational pipeline for identification of repeats from unassembled sequence reads. For graph-based clustering we used 495,974 reads for *S. tuberosa*, 296,743 for *S. mombim*, 476.671 for *Spondias sp*, 677.087 for *S. bahiensis*, and 252.458 for *S. dulcis* (Table 1).

Slide preparation

Root tips from *S. tuberosa* growing in pots were collected and pre-treated with 8-hydroxyquinoline for 20 h at 10°C, fixed in ethanol:acetic acid (3:1; v/v) for 2 to 24 h at room temperature and stored at –20°C. The root tips were washed in distilled water and digested with 2% cellulase (Onozuka) and 20% pectinase (Merck) at 37°C for 90 min. The apical meristems were squashed in 45% acetic acid under a coverslip. The coverslip was removed under liquid nitrogen.

Staining with CMA/DAPI and *in situ* hybridization

For the CMA/DAPI fluorochrome banding, the slides were aged for three days at room temperature and then directly stained with 0.5 mg/mL CMA (Sigma) for 1 h and 2

mg/mL of 4',6-diamidino2-phenylindole (DAPI) (Sigma) for 30 min. The slides were mounted in 1:1 (v/v) McIlvaine's buffer-glycerol with 2.5 mM $MgCl_2$ (pH 7.0) and kept in the dark at room temperature for three days before analysis (Schweizer & Ambros, 1994).

For *in situ* hybridization, the probes were obtained using the consensus sequence, for which the oligos were labeled with Cy3-dUTP and 6-carboxyfluorescein (6-FAM). *In situ* hybridization was performed according to Pedrosa et al. (2002). The hybridization mix contained 50% formamide (v/v), 10% dextran sulfate (w/v), 2 × SSC and 50 ng of each probe. The final hybridization stringency was estimated to be 76%. The slides were mounted with DAPI (4 μg mL^1)/Vectashield (Vector) 1:1 (v/v) and analyzed under an epifluorescence microscope (Leica DMLB) equipped with DAPI, FITC, and Cy3 filters. Images were recorded using a Cohu CCD camera and the software Leica QFISH before editing with the software Adobe Photoshop CS5.

Results

Genome size estimates

The results of flow cytometry showed that the genome size ranged from 460 Mb to 530 Mb (Table 1), among which the smallest genome was *S. pupurea* and the largest was *S. dulcis*. Pairwise comparison among species using Tukey's test ($p<0.05$) showed that *S. dulcis*, *S. tuberosa*, and the hybrid had a similar genome size, while *S. mombin* had an intermediary sized genome and *S. purpurea* had the smallest genome (Figure 1A). The genome size estimates showed strong correlations with phylogenetic relationships of the species, for which the phylogenetic tree using *trn*H-*psb*A, *rbc*L, and *mat*K plastid sequences showed a highly phylogenetic signal ($\lambda = 0.88$, $p < 0.05$),

suggesting that the derived clade showed the largest genome sizes, while the first diverging lineage showed smaller genome sizes (Figure 1B).

Repeat composition of *Spondias*

High-throughput DNA sequencing was performed on four species and a natural hybrid, which obtained 100 nt single-end reads for *S. bahiensis* and the hybrid, 2 × 100 nt paired-end reads for *S. tuberosa* and *S. mombin*, and 2 × 150 nt for *S. dulcis*. The sequences were analyzed using Graph-based clustering and the repetitive fraction corresponded to 49.99% in *S. dulcis*, 34.60% in the hybrid, 35.57% in *S. tuberosa*, 31.02% in *S. bahiensis*, and 33.43% in *S. mombin* (Table 1). Clusters representing repeats making up at least 0.01% of the genome were characterized, of which the results for each species separately showed 101 (15.11% of the genome) clusters in *S. tuberosa*, 106 (22.24 of the genome) in *S. mombin*, 110 (18.82% of the genome) in the hybrid, 109 (14.12% of the genome) in *S. bahiensis* and 137 (29.65% of the genome) in *S. dulcis* (Table 1).

The repeat composition of the species displayed that transportable elements (TE) were more abundant in all genomes, except for the hybrid (Table 2), among which the LTR retrotransposons, including ty1/copia and Ty3/gypsy, represented the major fraction of all of the analyzed genomes. The TE was mostly represented in *S. dulcis*, which corresponded to 5.021% for Ty1/copia and 13.145% for Ty3/gypsy (Table 2). Among the LTR retrotransposons families, Angela, chromovirus, and Tork were the most abundant in the genus. The repetitive fraction that corresponded to satellite DNA ranged from 3.46% (*S. dulcis*) to 7.47% (*S. mombin*), indicating a considerable proportion of satDNA. The rDNA showed moderately repetitive DNA, which ranged from 0.94% (*S. mombin*) to 4.256% (the hybrid) (Table 2).

High-throughput search for satDNAs

The results of graph-based clustering of unique nuclear reads showed two satDNAs in all genomes (SatSpo1 and SatSpo2) (Figure S1 and Table S1). Additionally, we found two satDNAs in *S. tuberosa* and *S. mombin* (SatSpo3 and SatSpo4) (Table S1) and one satDNA exclusively in *S. tuberosa*, interlaced on the intergenic spacer of DNAr (Figure S2). The monomers length showed a low variation, such as SatSpo2 with 145 bp (detected as a variation with 156 bp), SatSpo2 with 363 bp, and SatSpo3 and SatSpo4 with 185 pb and 213 bp, respectively (Table S2). The satDNA found on the intergenic spacer of DNAr showed a monomer length of 564 bp. Remarkably, SatSpo1 showed a CG content of 47%, while SatSpo2 displayed a CG content of 70% (Table S2). The analysis of satDNA families showed that SatSpo1 had 12 families in *S. tuberosa*, *S. mombin*, *S. bahiensis* and *S. dulcis*; however, *S. bahiensis* had the nuclear genome shared families from *S. dulcis* and *S. tuberosa* (Figure 2A). The phylogenetic relationship among the SatSpo1 families showed the families of *S. mombin* in the first clade and the families of *S. tuberosa* and *S. dulcis* in the second large clade; however, in the second clade we observed a subdivision between *S. tuberosa* and *S. dulcis* (Figure 2B). The phylogenetic tree revealed that *S. bahiensis* shared families with *S. tuberosa* and *S. dulcis* (Figure 2B). SatSpo2 was observed in only one family in the genus *Spondias*.

In situ hybridization in *S. tuberosa*

Heterochromatin that is CMA positive in a terminal location was observed in five chromosome pairs (Figure 3A and 3D), of which for two the labeling corresponded to 35s rDNA. The FISH for SatSpo1 satDNA labeled all complementary chromosomes,

although some chromosomes showed weak signals (Figure 3B and 3E) and sometimes the signals of SatSpo1 were close to the labeling of SatSpo2 (compare Figures 3B and 3C). The characterization of SatSpo2 in the chromosomes of *S. tuberosa* showed terminal strong signals in four chromosome pairs and double staining with CMA revealed SatSpo2 corresponded to the CMA positive bands (Figure 3A, 3C, and 3F). Two additional CMA positive blocks corresponded to 35S rDNA as described for *S. tuberosa* (Almeida et al., 2007).

Discussion

Genome size of *Spondias*

Genome size (GS) is an important approach to estimate the amount of DNA in the cell nucleus and is known as the C-value (Pellicer et al., 2018). The C-value has been used in some groups of angiosperms, as described for the Fabaceae family (Macas et al., 2015), *Taraxacum* sect. *Taraxacum* (Asteraceae) (Macháčková et al., 2018), and the genus *Jathropha* (Euphorbiaceae) (Marinho et al., 2018). The GS is classified using the C-value, for which an 1C-value < 3.5 pg is characterized as a small genome (Kelly and Leitch, 2011) The present study showed that the GS of *Spondias* species tends to be small and there was some minor variation across species, among which the derived species had larger GSs while the first diverging lineages had smaller GSs. Previous phylogenetic analysis of the genus showed that *S. mombin* was the first diverging lineage (Silva et al., 2015) and *S. mombin* was suggested to be an ancestral species of the genus (Mitchell and Daly 2015). These results corroborate prior molecular studies that have demonstrated low diversity among species in this genus (Silva et al., 2015, Balbino et al., 2019).

Repetitive sequences in Spondias

The repetitive DNA sequences in plant genomes make up a major fraction of their GS and studies are important for understudying their evolutionary dynamic, organization, distribution, and the relationships across congeneric species (Mehrotra and Goyal, 2014). Studies with repetitive DNA sequences expanded after the development of new sequencing approaches, because NGS can be used in clustering-based repeat identification of moderately to highly repeated sequences.

Here, we demonstrated the characterization of repetitive fractions in four *Spondias* species and a hybrid, for which the genome was found to be moderately repetitive, except for *S. dulcis* with 50% repetitive DNA. Transportable elements are the main components of the repetitive fraction in plants (Sveinsson et al., 2013) and in *Spondias* the TEs were more abundant in all species; however, no great difference was observed among species, and the main TE families showed similar proportions, suggesting low genome diversification among species. Prior studies have suggested that diversification of *Spondias* was recent (Machado et. al., 2015; Muellner-Riehl et al., 2016), resulting in low diversification in the genus.

satDNA in Spondias

For satDNA, we observed two main satellites distributed in all species. SatSpo1 showed high diversification in the genus, which was observed to be family and species-specific, but on the other hand, SatSpo2 showed low variation across the species. The low diversification in repetitive DNA was observed by genomic *in situ* hybridization, during which a probe from *S. tuberosa* hybridized to all chromosomes of *S. bahiensis*, suggesting high similarity in the sequences. Remarkable, *Spondias* has been revealed to have terminal heterochromatin that is rich in CG (Almeida et al., 2007) and in the

present study we detected a satDNA rich-CG, which FISH and CMA labeling revealed that SatSpo2 corresponded to the heterochromatin rich-CG. However, the results of SatSpo1 suggested that constitutive heterochromatin is formed by two main satellites, for which both satDNAs were detected near or interlaced in the chromosomes of *S. tuberosa*. A prior study showed that all neotropical species of *Spondias* have one site of rDNA 35S and 5S; however, for *S. tuberosa* we detected one satellite interlaced in the intergenic spacer of the rDNA. The sub repeat satellite on the intergenic spacer of the rDNA has been detected in other species, such as in the genus *Phaselous*, for which one satellite like the intergenic spacer of rDNA was located by FISH in the centromeric or pericentromeric region of the chromosomes of *P. vulgaris* while in other species in the genus it was restricting to intergenic spacers of rDNA (Almeida et al., 2012). Future studies in *Spondias* will allow us to determine if the satDNA IGS-SPO is distributed in many chromosomes or if it is restricted to the intergenic space of *S. tuberosa*.

Remarkably, the genus *Spondias* is characterized by natural hybrids, as described by Nobre et al. (2018) for *S. bahiensis*, and the results of the satDNA analysis demonstrated the hybrid origin of *S. bahiensis*, which shared the same satellites with progenitor *S. tuberosa* for SatSpo1 and IGS-SPO satDNA. We concluded that the genus *Spondias* has a recent diversification of repetitive DNA, with similar distributions of TEs and satDNA families across species.

Acknowledgments

We thank the Federal University of Alagoas for the laboratories and scientific support and the Fundação de Apoio à Pesquisa de Alagoas (FAPEAL) for funding this project.

Reference

Almeida, C. C. S., De Lemos Carvalho, P. C. & Guerra, M. Karyotype differentiation among Spondias species and the putative hybrid Umbu-cajá (Anacardiaceae). *Bot. J. Linn. Soc.* (2007). doi:10.1111/j.1095-8339.2007.00721.x

Balbino, E., Martins, G., Morais, S. & Almeida, C. Genome survey and development of 18 microsatellite markers to assess genetic diversity in Spondias tuberosa Arruda Câmara (Anacardiaceae) and cross-amplification in congeneric species. *Mol. Biol. Rep.* (2019). doi:10.1007/s11033-019-04768-w

Doležel, J. & Greilhuber, J. Nuclear genome size: Are we getting closer? *Cytometry Part A* (2010). doi:10.1002/cyto.a.20915

Kazazian, H. H. *Mobile Elements: Drivers of Genome Evolution*.

Kelly, L. J. & Leitch, I. J. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research* (2011). doi:10.1007/s10577-011-9246-z

Macas, J., Mészáros, T. & Nouzová, M. PlantSat: A specialized database for plant satellite repeats. *Bioinformatics* (2002). doi:10.1093/bioinformatics/18.1.28

Macas, J. *Et al.* In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe fabeae. *Plos One* (2015). Doi:10.1371/journal.pone.0143424

Machado, M. C., Carvalho, P. C. L. & Van Den Berg, C. *Domestication, hybridization, speciation, and the origins of an economically important tree crop of spondias (Anacardiaceae) from the brazilian Caatinga dry forest*. (2015).

Mitchell, J. & Daly, D. C. A revision of Spondias L. (Anacardiaceae) in the Neotropics. *Phytokeys* (2015). Doi:10.3897/phytokeys.55.8489

Macháčková, P. *Et al.* New chromosome counts and genome size estimates for 28 species of Taraxacum sect. Taraxacum. *Comp. Cytogenet.* (2018). Doi:10.3897/compcytogen.v12i3.27307

Marinho, A. C. T. A. *Et al.* Karyotype and genome size comparative analyses among six species of the oilseed-bearing genus Jatropha (Euphorbiaceae). *Genet. Mol.*

Biol. (2018). Doi:10.1590/1678-4685-gmb-2017-0120

Mehrotra, S. & Goyal, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics and Bioinformatics* (2014). Doi:10.1016/j.gpb.2014.07.003

Mitchell, John, and Douglas C. Daly. "A Revision of Spondias L. (Anacardiaceae) in the Neotropics." *Phytokeys*, 2015, doi:10.3897/phytokeys.55.8489.

Muellner-Riehl, A. N. *Et al.* Molecular phylogenetics and molecular clock dating of Sapindales based on plastid rbcl, atpb and trnl-trnf DNA sequences. *Taxon* (2016). Doi:10.12705/655.5

Nobre, L. L. De M., Dos Santos, J. D. O., Leite, R. & de Almeida, C. Phylogenomic and single nucleotide polymorphism analyses revealed the hybrid origin of Spondias bahiensis (family Anacardiaceae): De novo genome sequencing and comparative genomics. *Genet. Mol. Biol.* (2018). Doi:10.1590/1678-4685-gmb-2017-0256

Novák, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* (2010). Doi:10.1186/1471-2105-11-378

Novák, P. *Et al.* TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* (2017). Doi:10.1093/nar/gkx257

Pedrosa-Harand, A., dos Santos, K. G. B., Fonsêca, A., Mosiolek, M. & Almeida, C. Contrasting evolution of a satellite DNA and its ancestral IGS rdna in Phaseolus (Fabaceae) . *Genome* (2012). Doi:10.1139/g2012-059

Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I. J. Genome size diversity and its impact on the evolution of land plants. *Genes* (2018). Doi:10.3390/genes9020088

Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu. Rev. Genet.* (2012). Doi:10.1146/annurev-genet-110711-155621

Sveinsson, S., Gill, N., Kane, N. C. & Cronk, Q. Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (Theobroma cacao L.)

And related species. *BMC Genomics* (2013). Doi:10.1186/1471-2164-14-502

Silva, J. N., Bezerra da Costa, A., Silva, J. V. & Almeida, C. DNA barcoding and phylogeny in neotropical species of the genus Spondias. *Biochem. Syst. Ecol.* (2015). doi:10.1016/j.bse.2015.06.005

Wicker, Thomas *et al*. Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat. Rev. Genet.* **10,** 276–276 (2009).

Figure legends

Figure 1. Genome size estimates in *Spondias*. (A) 2C-values in four *Spondias* species and a natural hybrid. (B) Molecular phylogenetic analysis by the maximum likelihood method and signal phylogenetics for genome size estimates.

Figure 2. Diversification analyses for SatSpo1 satDNA in *Spondias*. (A) Heatmap analyses in 11 SatSpo1 families in four *Spondias* species. The colors represent the proportion of the families, in which red is more abundant and white represents absent families. (B) Molecular phylogenetic analysis by the maximum likelihood method for SatSpo1 families, with supported values estimated by bootstrap.

Figure 3. Characterization of SatSPo1 and SatSPo2 in *Spondias tuberosa*. (A and C) CMA/DAPI labeling of chromosomes. (B and D) SatSpo1 probe and (C and E) SatSpo1 probe. Arrowheads in B and D show minor signals.

Figure S1. Monomers of the SatSpo1 and SatSpo2 satellites in *Spondias*.

Figure S2. Structure of the 35S rDNA in *Spondias*, showing the IGS-SPO subrepeat.

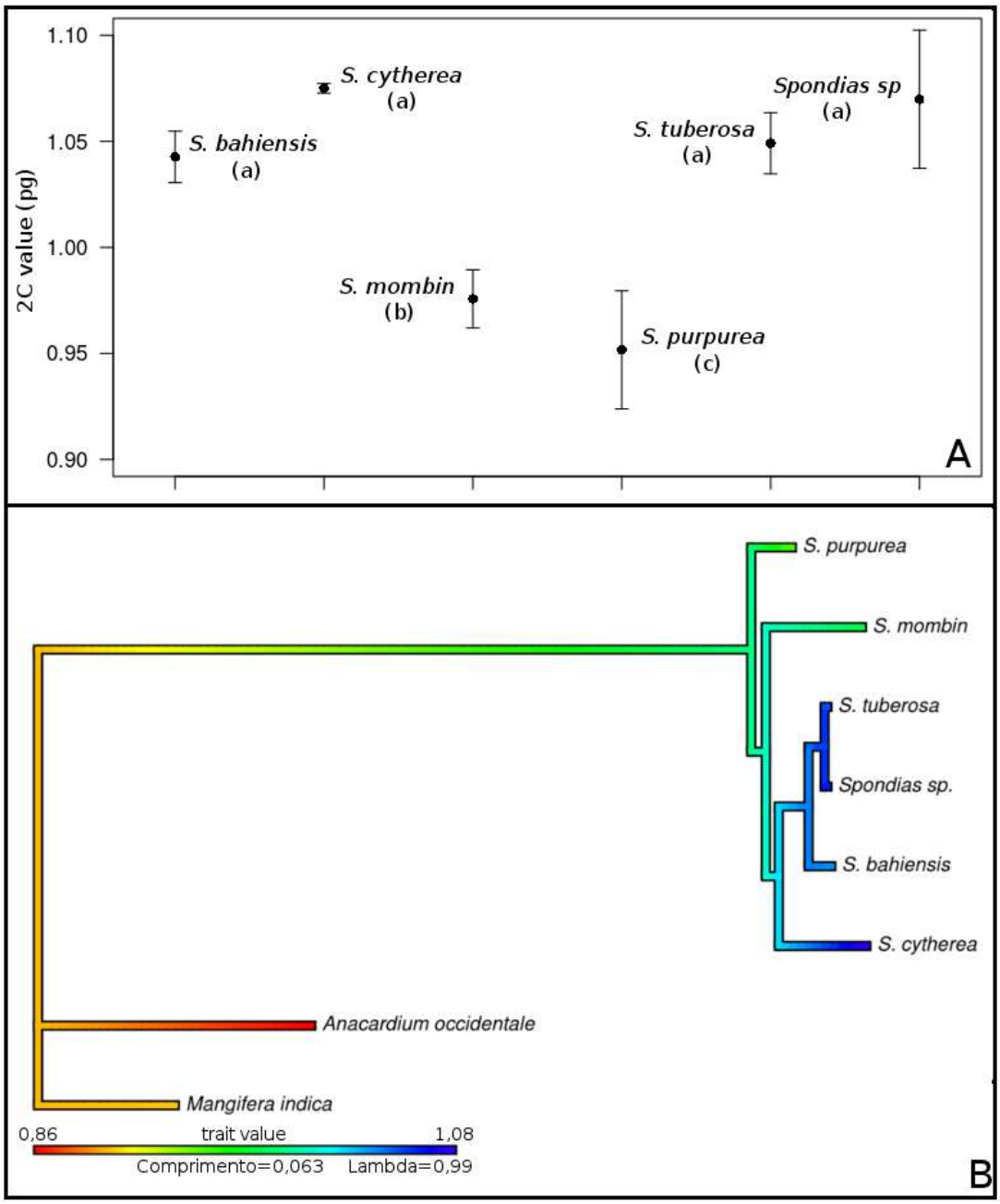


Figure 1. Genome size estimates in *Spondias*. (A) 2C-values in four *Spondias* species and a natural hybrid. (B) Molecular phylogenetic analysis by the maximum likelihood method and signal phylogenetics for genome size estimates.

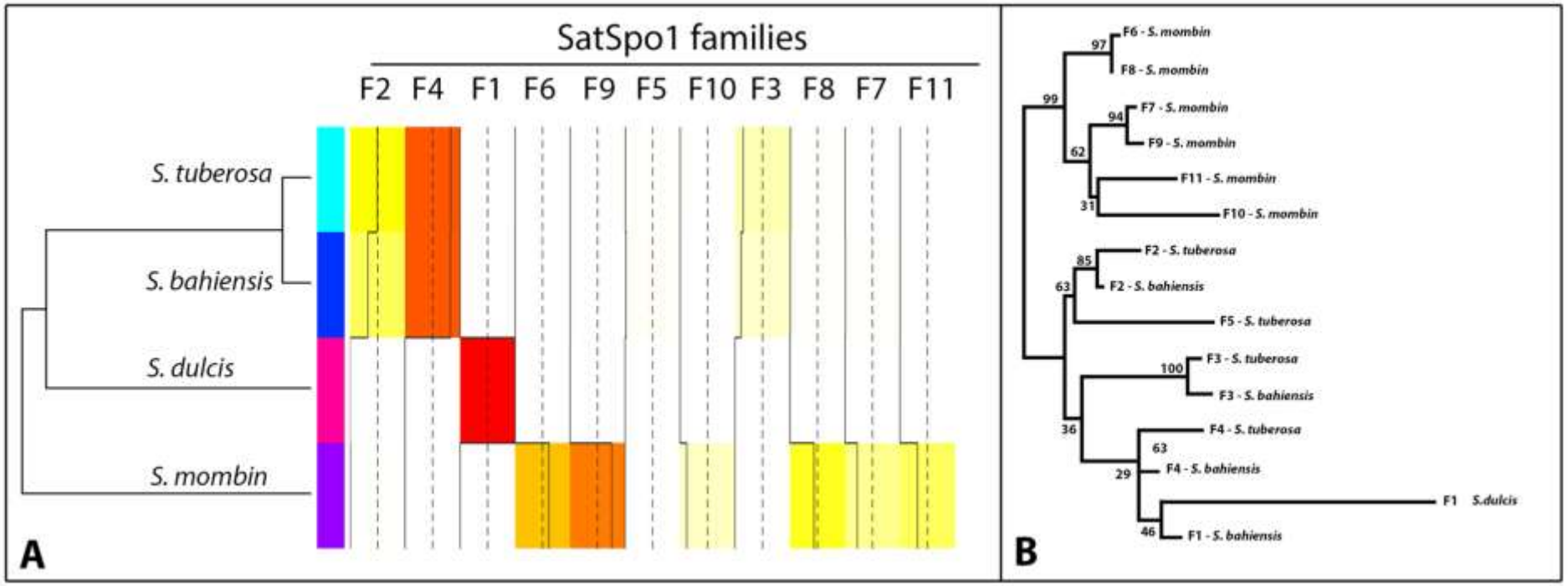


Figure 2. Diversification analyses for SatSpo1 satDNA in *Spondias*. (A) Heatmap analyses in 11 SatSpo1 families in four *Spondias* species. The colors represent the proportion of the families, in which red is more abundant and white represents absent families. (B) Molecular phylogenetic analysis by the maximum likelihood method for SatSpo1 families, with supported values estimated by bootstrap.

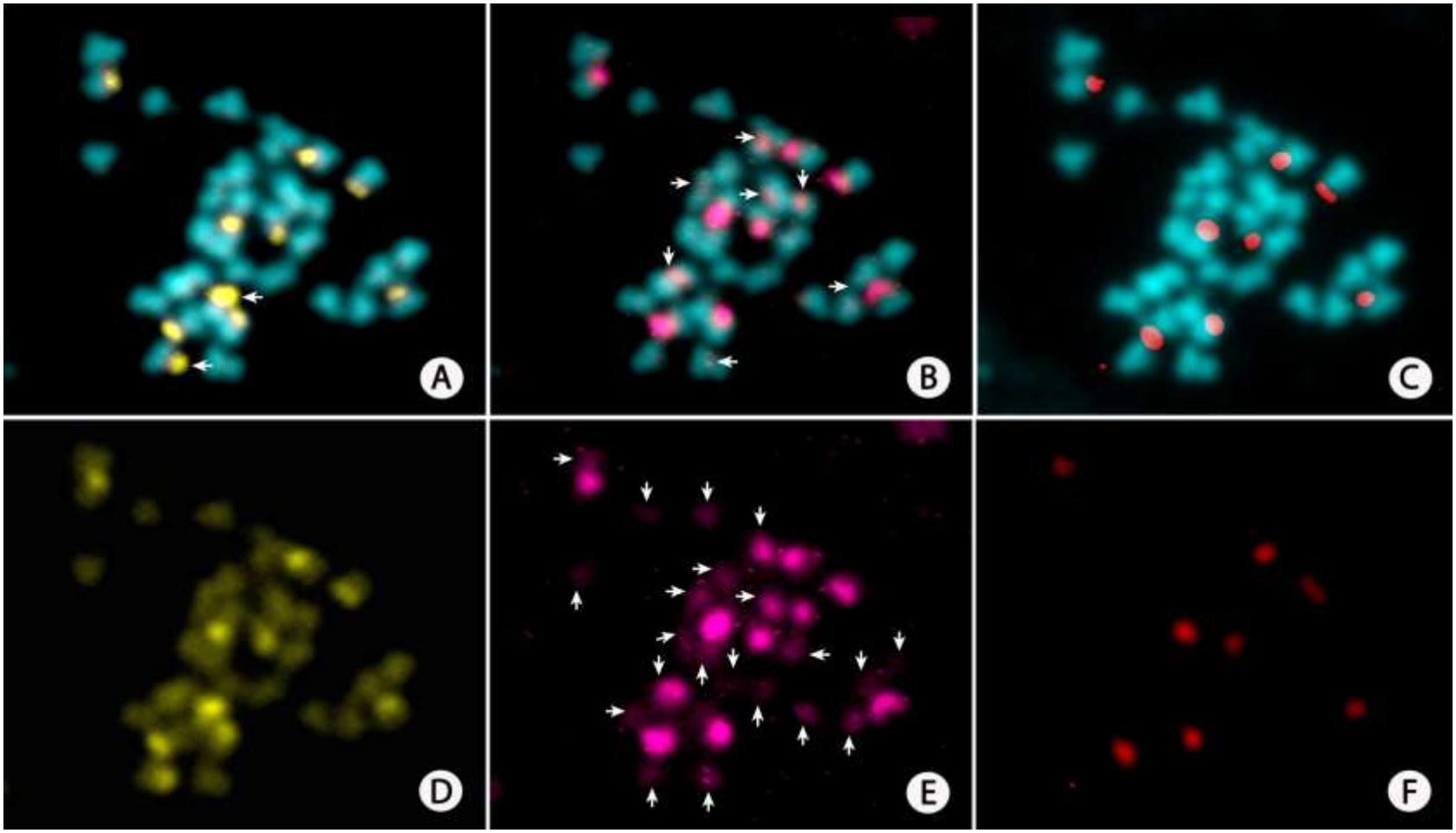


Figure 3. Characterization of SatSPo1 and SatSPo2 in *Spondias tuberosa*. (A and C) CMA/DAPI labeling of chromosomes. (B and D) SatSpo1 probe and (C and E) SatSpo1 probe. Arrowheads in B and D show minor signals.

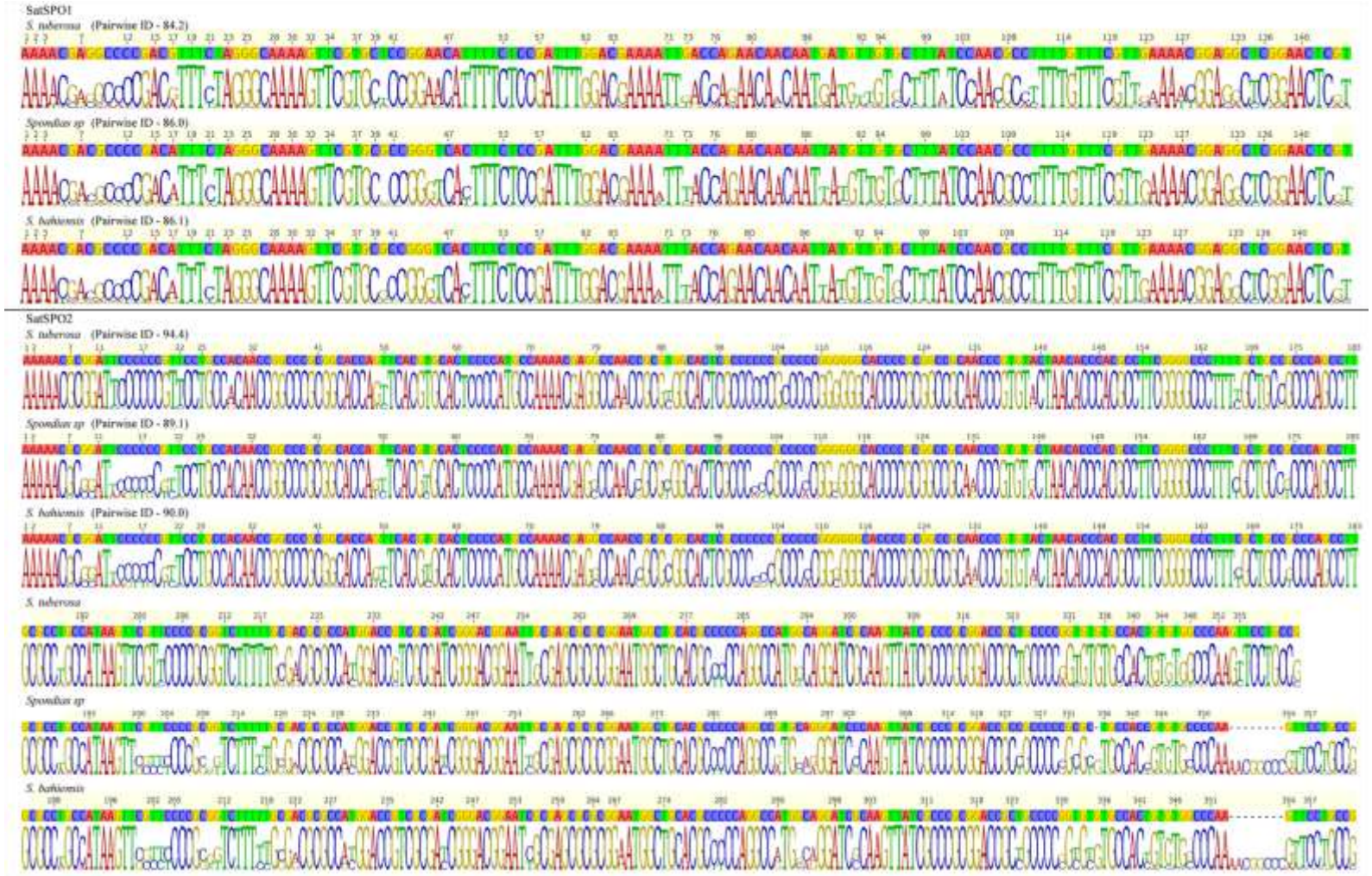


Figure S1. Monomers of the SatSpo1 and SatSpo2 satellites in *Spondias*.

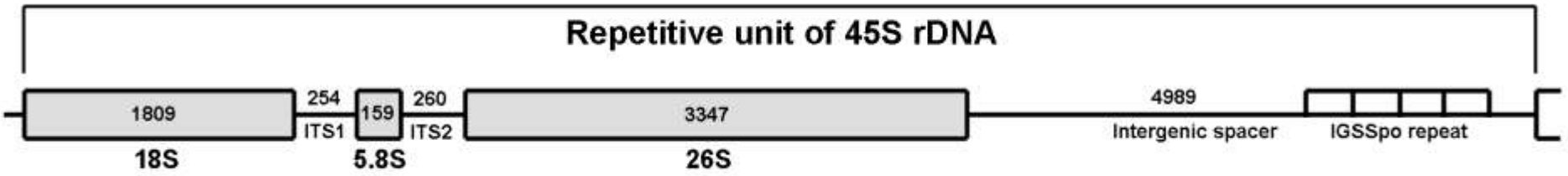


Figure S2. Structure of the 35S rDNA in *Spondias*, showing the IGS-SPO subrepeat.

Table 1. Quantification of genome size and repetitive DNA analysis in species of the genus *Spondias*.

Species	**Genome size (Mb)**	**2C-Value (pg)**	**Analyzed reads**	**Coverage**	**Repetitive fraction (%)**
S. dulcis	525.67	1.08	252.458	0.072	49.66
Spondias sp	523.16	1.07	476.671	0.091	34.60
S. tuberosa	513.01	1.05	495.974	0.145	35.57
S. bahiensis	509.86	1.04	677.087	0.133	31.02
S. mombin	477.13	0.98	296.743	0.093	33.43

Table 2. Genome proportions of repetitive sequences in the *Spondias.*

Class	Order	Superfamily	Family	Genome [%]				
				S. tuberosa	*S. mombin*	*Spondias sp.*	*S. bahiensis*	*S. dulcis*
Classified TEs-Retrotransposons	LTR	Ty1/copia	Ale	0.17	0.018	0.159	0.133	0.081
			Angela	1.47	3.28	1.686	1.223	0.918
			keros	0.40	0.242	0.344	0.325	0.037
			Bianca	-	-	-	-	0,23
			Ivana/Oryco	-	-	-	-	2.1
			SIRE	0.015	0.013	-	-	0.01
			Tork	0.26	1.499	0.035	0.182	1,645
		Total		**2.315**	**5.053**	**2.224**	**1.863**	**5.021**
		Ty3/gypsy	Athila	0.947	0.737	1.108	0.487	12.714
			Chromovirus	1.39	3.733	1.344	1.209	0.431
			Tat	0.419	-	0.23	0.192	-
		Total		**2.756**	**4.474**	**2.682**	**1.888**	**13.145**
	Non-LTR (LINE)			-	0.87	0.098	0.078	2.5
Classified TEs-Transposons		TIR	hAT	0.012	-	0.01	-	-
			Mutator	0.126	0.396	0.047	0.095	0.071
Total				**5.209**	**10.793**	**5.061**	**3.924**	**20.737**
rDNA				2.59	0.94	4.256	1.956	1.52
Classified Repeats			SAT	4.13	7.47	5.70	3.81	3.46
Total Unidentified				**3.186**	**3.031**	**3.804**	**4.423**	**3.932**
Total Repeat				**15.115**	**22.238**	**18.821**	**14.118**	**29.649**

Table S1. Satellites DNA sequence in *Spondias*

Satélites	Monômeros	Tamanho [pb]	GC [%]	Repeat [%]				
				S. tuberosa	*S. mombin*	*Spondias sp*	*S. bahiensis*	*S. dulcis*
SatPO1	AAAACGACGCCCCGACATTTCTAGGGCAAAAGTTCGTGCGCCGGGTCACTTTCTCCGATTTGGACGAAAATTGACCAGAACAACAATTATGTTGTGCTTTATCCAACGCCTTTTGTTTCGTTGAAAACGGAGGCTCGGAACTCGT	145	47	4.23	6.79		3.81	3.46
SatSPO2	AAAAACGCGGATTCCCCCCGTTCCTGCCACAACCGGCCCGCGGCACCAGTTCACGTGCACTCCCCATGCCAAAACGAGGCCAACCGCGCGGCACTCGCCCCCCGCCCCCGGGGGGCACCCCGCGGCCGCAACCCGTGTACTAACACCCACGCCTTCGGGGCCCTTTCGCTGCCGCCCAGCCTTGCGCCAGCCATAAGTTCGTCCCCCGCGGTCTTTTTGGGGCGCGCCACGGACCCTCGCGATCGGGACGGAATTGCGAGCGCGCGGAATGGCTGCACGCCCCCAGGCCATGGCAGGATCGCAAGTTATCGCCCGCGGACCGCTGCCCCGGTGTGTGCCACTGTGTGGCCCAAGTTCCTGCCG	363	70	1.20	0.83		1.47	7.5
SatSO3	TGTGTACCTATGAATTTGTGTTGCAATTTGTTTTGATAACATATAATTAATTAAATAAATATGTGTACGTATACTTAATTGAATTAATTTGTAAACTTATGAATTTGTATTGCAATTTATTTTGATGACATTCAATTAATCGAATAAATTTGTCTACTTATACATAATAATTAATTGAATTAATT	185	19	0.03	0.05	<0.01	<0.01	<0.01
SatSPO4	CCTGCAAATAAATCAACTTATGTAGTCAAATTAAAATTTTATTACTAGTAATCATAATAACTTTAACAAATAAAAGTTAAAAGGCATTTACTTATCTCCAGTGTCGATCCTTCATGCAACAAGTTACTTGATGACAGTTGAATCTAACAAACTGTCATAAATTAACATTCTTATGTAGAATACAAATAATATATTAATGCTAAAGCTTCATTT	213	25	0.026	0.0	0.0	0.0	0.0

IGS-SPO	ACACACACGACCATGCTCCTGCACACTACATAGGCACAACATAGCCACA CTTACCAGGCACCGATGGTGCCACCACCACCGCCACCGCCACCACCACC GCCACCGTTGCCGCCTTACGGTTTTAGGCCCATACAAAATGGAGAATGA CGAATTTTGAAGCCATTTCTTGCAGGGATGAAGATAAAAATGCAAGGAA TGAGTAGAAAAAAGAAAAAAAATATTTCGGGATGAAAATTAATTTTTTC GATTTTTCGAATATTAAAAATAAGAAAAATCATAAAAAATAGTAGAGA AGGAATTTTTGAGTGGGACTTTTTTCAAGGGTCTCCTAAAATCCCAAAT GAAATTATGGAATGGTTGGATGGTGTTTTGGAGAAACTTTGTGCGGGAG TTTTTTCAAAAGTTTGGGCGTTGGCACACAGTGTGGGGCAAAACAGCCC ACTGGTGCTGCACAGCATGCAGCAATGTGCAGCAAGACAGGCCACACA TGCTGCTTGCAAGCAGCAAGTCTGTGCACACATGCTGCTGCATGCATGT GGCAAGGCTGCACACCCATGCTGTTGC	564	45	0.66	0.00	0.44	0.38	0.00