UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE BIOCIÊNCIAS
DEPARTAMENTO DE BOTÂNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA VEGETAL

YENNIFER CAROLINA MATA SUCRE

ANÁLISE DA DIVERSIFICAÇÃO DA HETEROCROMATINA DO GRUPO CAESALPINIA (Leguminosae) BASEADO EM DADOS CITOMOLECULARES E GENOMICOS

Recife
2020

YENNIFER CAROLINA MATA SUCRE

ANÁLISE DA DIVERSIFICAÇÃO DA HETEROCROMATINA DO GRUPO CAESALPINIA (Leguminosae) BASEADO EM DADOS CITOMOLECULARES E GENOMICOS

Dissertação de mestrado apresentado ao Programa de Pós-Graduação em Biologia Vegetal da Universidade Federal de Pernambuco, como um dos requisitos para a obtenção do título de Mestre em Biologia Vegetal.

Área de Concentração: Sistemática e Evolução.

Orientador: Prof. Dr. Luiz Gustavo Souza

Catalogação na fonte
Elaine C Barroso
(CRB4 1728)

Mata Sucre, Yennifer Carolina
Análise da diversificação da heterocromatina do grupo *Caesalpinia* (Leguminosae) baseado em dados citomoleculares e genomicos/ Yennifer Carolina Mata Sucre – 2020.

118 f.: il., fig., tab.

Orientador: Luis Gustavo Souza

Dissertação (mestrado) – Universidade Federal de Pernambuco. Centro de Biociências. Programa de Pós-Graduação em Biologia Vegetal, 2020.
Inclui referências.

1. Caesalpinia 2. Evolução do genoma 3. Filogenética I. Souza, Luis Gustavo (orient.) II. Título

583.749 CDD (22.ed.) UFPE/CB – 2020- 120

YENNIFER CAROLINA MATA SUCRE

ANÁLISE DA DIVERSIFICAÇÃO DA HETEROCROMATINA DO GRUPO CAESALPINIA (Leguminosae) BASEADO EM DADOS CITOMOLECULARES E GENOMICOS

Dissertação de mestrado apresentado ao Programa de Pós-Graduação em Biologia Vegetal da Universidade Federal de Pernambuco, como um dos requisitos para a obtenção do título de Mestre em Biologia Vegetal.

Aprovada em: 18/ 02 / 2020

BANCA EXAMINADORA

Prof. Dr. Luiz Gustavo Souza (Orientador)

Universidade Federal de Pernambuco

Prof. Dr. Andrea Pedrosa-Harand (Examinador Interno)

Universidade Federal de Pernambuco

Prof. Dr. Rita De Cassia De Moura (Examinador Externo)

Universidade de Pernambuco

Dedico este trabajo al pilar fundamental en mi vida, ***mi familia****. Y en especial a mis hermanas* ***Yennymar y Yessika Mata Sucre***

AGRADECIMENTOS

Quero agradecer a todas as pessoas que, de forma direta ou indireta, contribuíram para a realização deste trabalho. Especialmente:

Aos meus pais, **Yennys Sucre** e **Marcos Mata**, que desde o início me deram todo o apoio e motivação para continuar com meus estudos, embora eles não entendam o que faço, eternamente agradecida.

As minhas irmãs **Yennymar** e **Yessika Mata**. O que eu faria sem vocês? Obrigada pelos minutos de conversa e motivação para não desistir na minha formação profissional.

Ao meu orientador, **Prof. Dr. Gustavo Souza**, pelo constante exemplo, motivação, incentivo e inspiração no meu caminho profissional.

A **Prof. Dr. Andrea Pedrosa-Harand**, pesquisadora, mamãe, "chefa" do laboratório e admirável mulher. Obrigada pela confiança, amizade e exemplo de vida no caminho da ciência.

A futura doutora **Mariela Sader**, mulher, pesquisadora e exemplo de vida, obrigada pelos conselhos profissionais e pessoais, por estar sempre presente, sem medir tempo nem esforço. Obrigada pela dedicação mulher!

A minha amiga e coleguinha de "rolês" aleatórios, futura doutora **Amalia Ibiapino Moura**, por todos esses momentos de escuta constante das minhas frustações, pelo conhecimento transmitido nas eternas e inúmeras conversas e pela paciência em tudo, muito obrigada mulher!

A **Dr. Mariana Báez**, por todo o carinho que você pôde dar para uma pessoa, sem conhecê-la, muito obrigada por tudo!

Aos meus amigos **Giovanna La manna**, **Ambar Vallera**, **Luis Gutierrez**, **Lederle Hernandez** e **Luis Diaz**, por estarem presentes, mesmo com a distância.

Ao **Herberth Haeusler**, pela confiança, companhia, carinho, apoio incondicional, palavras de alento e infinita força para me aguentar durante tudo este caminho percorrido. Obrigada.

Ao **Dr. Jose Imery**, por me inspirar a continuar no caminho da ciência e por me ensinar que “lo fácil no es divertido”, muito obrigada por tudo professor.

Às pessoas que fizeram ou fazem parte do Laboratório de citogenética e evolução vegetal, Amandão, Amandinha, Ana, Brena, Bruna, Cláudio, Géssica, Gustavinho, Jéssica, Lucas, Natália, Pablo, Paulo, Rayssa, Tiago Esposito, Tiago Ribeiro e Yhanndra. Muito obrigada pelo apoio infinito desde o início do meu trabalho e por todos esses momentos lindos e chatos compartilhados.

Ao programa de **Pós-Graduação em Biologia Vegetal**, em especial ao pessoal administrativo, pela ajuda, paciência e constantes esclarecimentos de dúvidas durante a minha permanência no programa.

Ao programa de **Becas Brasil PAEC-OEA-GCUB**, pela bolsa e demais recursos concedidos para a realização do presente estudo.

Finalmente, a todas as pessoas que estão distantes, mas mesmo assim não estão ausentes.

A todos. Muito **obrigada**!

RESUMO

O grupo Caesalpinia é composto por 27 gêneros e 225 espécies. Apresenta distribuição pantropical em florestas tropicais sazonalmente secas (do inglês SDTF), principalmente no continente americano. Cariotipicamente o grupo é caracterizado pela estabilidade numérica (2*n* = 24), porém apresenta uma alta diversidade na quantidade e distribuição de bandas heterocromáticas. Neste trabalho, a caracterização do grupo foi expandida para 12 novas espécies, as quais mostraram bandas pericentroméricas $CMA^{+}/DAPI^{-}$ e $CMA^{0}/DAPI^{-}$ associadas com as suas distribuições geográficas. Além disso, foi demonstrado estatisticamente uma autocorrelação da intensidade CMA/DAPI ao longo do cromossomo, tamanho do genoma e a distribuição geográfica (latitude) no grupo. Embora, o retrotransposons LTR Ty3/*Gypsy* da linhagem Tekay, seja um dos principais componentes da heterocromatina CMA^{+} pericentromerica nas espécies de Caesalpinia do Nordeste brasileiro (*Cenostigma microphyllum*, *Libidibia ferrea* e *Paubrasilia echinata*), nossos resultados demonstram que não é o elemento mais abundante em *Erythrostemon hughesii*. De fato, *E. hughesii* (uma espécie sem bandas CMA^{+} pericentromericas) tem uma baixa proporção de retrotransposons Ty3/*Gypsy*-Tekay e uma elevada abundância de sequências de DNA satélites na heterochromatina. Isso demonstra que as bandas de heterocromatina CMA^{+} são altamente polimórficas no grupo Caesalpinia e representa um *hotspot* de sequências repetitivas. Os resultados do presente trabalho corroboram a importância das análises citogenômicas no grupo Caesalpinia para a caracterização cromossômica, assim como a importância do uso de sequências repetitivas nos métodos comparativos.

Palavras-chave: Coloração CMA/DAPI. Evolução do genoma. Métodos comparativos filogenéticos. Sequências repetitivas.

ABSTRACT

The Caesalpinia group is formed by 27 genera and 225 species with pantropical distribution in seasonally dry tropical forests (SDTF), mainly in the American continent. Karyotypically the group is characterized by numerical stability (2*n* = 24), however was observed a high diversity in the amount and distribution of heterochromatin bands. In this work, the characterization of the group was expanded to 12 new species, which showed pericentromeric CMA^{+}/$DAPI^{-}$ and CMA^{0}/$DAPI^{-}$ bands, associated with their geographic distributions. In addition, was statistically demonstrated a self-correlation of CMA / DAPI intensity along the chromosome, genome size and geographical distribution (latitude) of the group. Although the retrotransposons LTR Ty3 / *Gypsy* of the Tekay lineage, being one of the main components of the heterochromatin CMA^{+} pericentromeric in the Northeast Brazil Caesalpinia species (*Cenostigma microphyllum*, *Libidibia ferrea* and *Paubrasilia echinata*), our results demonstrate that it is not the most abundant element in *Erythrostemon hughesii*. In fact, *E. hughesii* (a species without CMA^{+} pericentromeric bands) has a low proportion of Ty3 / *Gypsy*-Tekay retrotransposons and a high abundance of satellite DNA sequences in their heterochromatin. This demonstrates that the CMA^{+} heterochromatin bands are highly polymorphic in the Caesalpinia group and represent a *hotspot* of repetitive sequences. The results of the present study corroborate the importance of cytogenomic analyzes in the Caesalpinia group for chromosome characterization, as well as the importance of using repetitive sequences in comparative methods.

Keywords: CMA / DAPI double staining. Genome evolution. Heterochromatin. Phylogenetic comparative methods. Repetitive sequences.

SUMÁRIO

1 INTRODUÇÃO

O grupo Caesalpinia pertence à família Leguminosae e inclui 27 gêneros e 225 espécies com uma ampla distribuição no mundo, ocorrendo principalmente em regiões tropicais. Desde as primeiras análises citogenéticas, o grupo mostrou distintas características interessantes, como uma variação no tamanho cromossômico e, como consequência, no tamanho do genoma. Análises citomoleculares têm descrito polimorfismo no número, posição e composição das bandas heterocromáticas, sendo distribuídas preferencialmente nas regiões pericentromérica e subtelomérica dos cromossomos.

Nos últimos anos, o uso de análises comparativas para inferir a história evolutiva de diferentes caracteres aumentou, mas ainda são poucos os que integram dados citogenéticos, genômicos e ambientais. Em Caesalpinia, a integração de dados citomoleculares e geográficos revelou três padrões de heterocromatina pericentromérica, relacionados aos principais centros de diversidade das espécies neotropicais do grupo. A identificação desses padrões demonstrou a utilidade das marcas citogenéticas nas análises citogeográficas. Adicionalmente a combinação com outros dados citológicos, como o tamanho do genoma, pode revelar mais precisamente a história de evolução cromossômica do grupo.

A história evolutiva de um cariótipo é frequentemente difícil de determinar, especialmente para eventos de diferenciação profunda. Recentemente, o sequenciamento de próxima geração (NGS) de três espécies brasileiras do grupo (*Cenostigma microphyllum*, *Libidibia ferrea* e *Paubrasilia echinata*) permitiu obter uma visão geral do conteúdo genético e das sequências repetitivas desses genomas. Constatou-se que os elementos transponíveis do tipo retrotransposon integram a maior parte da heterocromatina pericentromérica do tipo CMA^{+}. Sabe-se que elementos transponíveis, devido à sua capacidade de se mover, podem gerar alterações nos genomas e ter um papel adaptativo em resposta ao meio ambiente, podendo neste caso, explicar os padrões heterocromáticos observados em Caesalpinia. Entretanto, análises citogenéticas e genômicas de mais espécies são necessárias para elucidar essas relações entre o genoma e o ambiente. Portanto, o objetivo deste trabalho foi expandir a caracterização das espécies do grupo Caesalpinia do ponto de vista citomolecular e genômico, buscando discutir a diversificação da heterocromatina no grupo.

O conteúdo desta dissertação foi divido em dois capítulos, o primeiro contém a caracterização citomolecular de diferentes espécies dos gêneros *Coulteria*, *Erythrostemon*, *Libidibia*, *Mezoneuron*, *Pomaria* e *Tara*, assim como análises comparativas e correlações dos

caracteres citogenéticos, tamanho do genoma e bandeamento CMA/DAPI com dados ecológicos (manuscrito submetido ao periódico *Plant Systematics and Evolution*). No capitulo dois são abordados os aspectos citogenômicos de uma espécie do gênero *Erythrostemon*, que possui bandas heterocromáticas CMA^+ subterminal e com ocorrência no México. Foi realizada a caraterização da fração repetitiva do genoma de *E. hughesii* e o mapeamento das sequencias repetitivas mais abundantes. Além disso, realizamos uma análise comparativa dessa espécie com as espécies do grupo que ocorrem no Brasil (*Cenostigma*, *Libidibia* e *Paubrasilia*) cujos genomas foram caracterizados previamente (manuscrito a ser submetido no periódico *Planta*).

2 REFERENCIAL TEÓRICO

2.1 GENOMA E O MEIO AMBIENTE

Os organismos evoluem e se adaptam através de mudanças em seus genomas, e essas mudanças podem favorecer certos genes ou vias moleculares e melhorar a sobrevivência e o fitness reprodutivo do organismo hospedeiro. Alguns estudos têm demonstrado adaptação do genoma ao meio ambiente, assim como relação entre polimorfismos genéticos e exposições a diferentes condições ambientais (Pluess et al. 2016; Yang et al. 2016; Sun et al. 2017). As combinações específicas de interação genótipo × ambiente oferecem múltiplos efeitos em resposta ao ambiente (Des Marais et al. 2013). Por exemplo, Deatherage et al. (2017) sequenciaram o genoma completo de 30 linhagens de *Escherichia coli* que evoluíram por 2.000 gerações em cinco ambientes que diferiam apenas nas temperaturas experimentadas. Essas colônias apresentaram mutações gênicas específicas em cada temperatura e algumas tenderam a ser benéficas, representando assinaturas genômicas de adaptação. Modificações adaptativas dentro de um conjunto comum de mutações em diferentes genes, permite a adaptação e evolução do genoma nos estágios iniciais.

A adaptação no nível genômico é um processo complexo que resulta do grande número de genes que podem potencialmente evoluir de forma convergente e ao acaso (Thomas e Hahn 2015; Zou e Zhang 2015). Uma adaptação convergente de organismos filogeneticamente distintos ao mesmo ambiente foi observada em diferentes espécies de manguezais. Lyu et al. (2018) estudaram a composição de sequências repetitivas e seus possíveis efeitos na adaptação dos genomas de diferentes espécies de manguezais e seus parentes não-manguezais que colonizam independentemente a interface entre terra e mar. Neste trabalho, todas as linhagens de verdadeiros manguezais reduziram significativamente seu conteúdo repetitivo em comparação com seus parentes não-manguezais, e como consequência, apresentaram diminuição do tamanho do genoma. A eliminação da maioria dos elementos móveis, representou uma estratégia convergente empregada pelos mangues para se adaptar a novos ambientes estressantes.

Diferentes estudos têm relatado que as sequências repetitivas, especificamente os ETs, tornam-se ativas sob condições de estresse, e essa ativação é frequentemente tomada como evidência de um papel adaptativo (Chenais et al. 2012; Casacuberta e Gonzalez 2013; Negi et

al. 2016; Rey et al. 2016; Horváth et al. 2017). Essa hipótese se baseia no fato de que a ativação de ETs levaria a um aumento na taxa de mutação, gerando variabilidade sob a qual a seleção natural pode atuar (Horváth et al. 2017). Além disso, como alguns ETs são conhecidos por conter sequências reguladoras da resposta ao estresse, se os ETs forem ativados pelo estresse, eles podem distribuir elementos de resposta por todo o genoma que podem ajudar a reprogramar essas redes de genes (Cowley et al. 2013). Atualmente, existe uma concordância relativamente ampla de que os ETs têm contribuído para várias inovações fundamentais na evolução adaptativa (Chuong et al. 2017; Esnault et al. 2019), por exemplo, servindo como origem para pequenos RNAs (Slotkin e Martienssen 2007; Berezikov 2011) ou como enriquecedores de regiões promotoras de genes (Niu et al. 2019).

A ativação e a transposição de ETs associados ao stress foram discutidas em detalhes em algumas revisões (Negi et al. 2016; Galindo-González et al. 2017; Horváth et al. 2017; Schrader e Schmitz 2019). Como unidades de transcrição, os ETs possuem suas próprias sequências reguladoras, onde diferentes fatores podem impactar algumas cópias de uma família particular de retroelementos (Lanciano e Mirouze 2018). Por exemplo, em formigas do gênero *Cardiocondyla* ilhas de ETs tem um papel fundamental na rápida adaptação a diferentes habitats. Os retroelementos das superfamílias BEL/Pao, DIRS, LOA/Loa, Ngaro, R1/R2 e RTE, bem como transposons de DNA das superfamílias Academ, Kolobok-Hydra, Maverick, Merlin e TcMar-Mariner, ocupam as ilhas de ET com números de cópias significativamente maiores do que outros elementos transponíveis. Essas ilhas de ET funcionam como poços gênicos para a diversificação genética em populações fundadoras dessa espécie invasora, atuando na diferenciação, adaptação e especiação do grupo (Schrader et al. 2014).

Em plantas, as inserções de ET fornecem um mecanismo mutagênico potente para a evolução de novos genes e suas funcionalidades. Em soja, um retrotransposon tipo Ty1/*copia*, designado SORE-1, encontra-se inserido em genes responsáveis pela sensibilidade ao fotoperíodo. Quando o gene é duplicado, a inserção do retrotransposon causa a perda da função genética por silenciamento, levando à adaptação do cultivo de soja em altas latitudes (Kanazawa et al. 2009). Adicionalmente, em girassol, três espécies de origem híbrida antiga têm eventos de proliferação de sequências repetitivas dentro das famílias de retrotransposon LTR-Ty3/*gypsy* e Ty1/*copia*. Os elementos do tipo Ty1/*copia* sofreram aumento no número de cópias após ou associados às origens dessas espécies, embora a escala de proliferação seja

menor do que a dos elementos do tipo Ty3/*gypsy*. Nesse trabalho, a similaridade das sequências de Ty1/*copia* nos genomas das três espécies híbridas e das duas espécies parentais revelam que uma única sub-linhagem desses elementos exibe características de amplificação recente e provavelmente serviu como linhagem de fonte proliferativa. A proliferação de LTR-retrotransposons indica que as condições genômicas e/ou ambientais associadas às origens desses táxons híbridos de girassol propiciaram à não repressão de pelo menos dois grupos principais de elementos transponíveis (Kawakami et al. 2010). A relação ET-estresse é complexa, embora as evidências tenham demonstrado que os ETs são ativados sob estresse na maioria dos estudos, em outros as inserções de ET demonstraram ter um efeito deletério.

2.2 ANALISES CITOGENÉTICAS X MEIO AMBIENTE

2.2.1 Evolução do tamanho genômico

Variações cromossômicas podem estar relacionadas à distribuição ecológica e geográfica das plantas, e ao mesmo tempo sua divergência nos genomas ocorre como consequência adaptativa às condições ambientais específicas (Tapia-Pastrana et al. 2012). Diferentes atributos são o resultado de uma combinação de genótipo e ambiente, por exemplo, o tamanho do genoma. Até hoje questiona-se a evolução correlacionada entre GS e fatores ecológicos, devido aos resultados contrastantes obtidos em diferentes estudos (Šmarda et al. 2007; Díez et al. 2013; Kang et al. 2014; Jordan et al. 2015). Por exemplo, nas tribos Cardueae e Anthemideae (Asteraceae) o GS está correlacionado com características cariológicas, fisiológicas e ambientais (Garnatje et al. 2004). No entanto, Jakob et al. (2004) observaram que no nível taxonômico mais alto das espécies de *Hordeum* L. (Poaceae), as correlações ambientais estavam ausentes. E em *Zea mays* L. a diversidade de GS nas linhagens de milho filogeneticamente independentes está negativamente correlacionada com a altitude (Díez et al. 2013).

Embora existam discrepâncias sobre a existência ou não de relações entre o GS e o ambiente, a grande maioria dos trabalhos apontam para uma diversidade de GS associada a adaptações ambientais ou ecológicas. Du et al. (2017) em *Lilium* L. (Liliaceae) por meio de uma análise integrativa de citogenética e filogeografia, descreveram como eram as relações entre a diversidade e evolução do GS e sua correlação com características cariológicas e ecológicas em vários táxons do grupo. Eles mostraram que existe uma correlação negativa

entre o GS e a temperatura/precipitação anual, dois fatores ambientais que são diretamente afetados pela posição ou elevação geográfica. Nesse caso, as espécies que ocorrem nos ambientes das montanhas Hengduan e Himalaia, na China, exibem um GS relativamente pequeno, e geralmente crescem acima de 3.000 m em ambientes relativamente extremos. Por outro lado, as espécies do Extremo Oriente e da América do Norte normalmente crescem em altitudes mais baixas, com ambientes relativamente menos severos, e exibem um GS maior, sugerindo que um GS pequeno evolui como uma adaptação a ambientes estressantes.

Do mesmo modo, Bilinski et al. (2018) analisaram como eram as mudanças no tamanho do genoma em linhagens domésticas e selvagens de milho que ocorriam em gradientes altitudinais na Mesoamérica e na América do Sul. Eles encontraram que as diferenças nos tempos de floração em diferentes altitudes afetam indiretamente os clines no tamanho do genoma, devido a uma relação mecanicista entre o tamanho do genoma e a produção celular e a taxa de desenvolvimento. Nessa interação, o tamanho do genoma sofre pressões paralelas e diferencial de seleção natural de acordo com a altitude. Os autores apontam uma relação da variação do genoma a importantes diferenças nos fenótipos com adaptações independentes à alta altitude.

O tamanho do genoma é uma característica adaptativa importante (Bilinski et al. 2018). Ele varia muitas vezes de magnitude entre as espécies, devido a alterações na ploidia e no conteúdo de DNA haploide (Hidalgo et al. 2017; Pellicer et al. 2018). Na ausência de poliploidia, as alterações na quantidade de DNA repetitivo são as principais responsáveis pelas diferenças de GS entre as espécies (Du et al. 2017). Conforme relatado na *Fritillaria* L., um caso extremo de expansão genômica através do acúmulo de DNA repetitivo, 80-90% do DNA dessas espécies é derivado de repetições altamente heterogêneas. A falta de eliminação e a baixa transposição de DNA repetitivo desempenham papéis importantes na evolução de genomas, principalmente naqueles extremamente grandes (Kelly et al. 2015).

Os ETs geralmente proliferam mais rapidamente do que podem ser removidos, contribuindo assim para o crescimento do tamanho do genoma (Brookfield 2005; Lisch 2013). A maioria das espécies da família Brassicaceae (Burnett) tem GS pequenos, no entanto as espécies do clado monofilético *Hesperis* L. possuem os maiores genomas do grupo (Mandáková et al. 2017). Estudo recentes mostraram que existe uma correlação positiva entre o aumento do tamanho do genoma e o conteúdo de ETs, onde a proliferação de retrotransposons LTR, iniciou no ancestral do clado *Hesperis* e, posteriormente, nos táxons

das seis tribos. Supõe-se que a predominância da obesidade genômica no grupo esteja associada à seleção para hábitos de vida bienais ou perenes, e em alguns casos a expansão do genoma foi neutralizada pela eliminação de ETs, permitindo em algumas espécies uma transição adaptativa à estratégia de vida anual (Hloušková et a. 2019).

2.2.2 Evolução do cariótipo

A diversidade cromossômica pode ter efeitos importantes na evolução de alguns grupos, pois essas alterações afetam a estrutura cromossômica e a simetria do cariótipo (Peruzzi et al. 2009; Gao et al. 2015). Nesse sentido, as análises citogenéticas contribuem frequentemente para o entendimento da evolução das plantas e para a determinação dos fatores envolvidos na diversificação dos grupos e táxons (Lan e Albert 2011; Lamo et al. 2016). As análises de cariótipo são uma ferramenta importante para revelar padrões de evolução cromossômica. As principais informações fornecidas pelas análises citogenéticas para o entendimento da evolução cromossômica são: número cromossômico, posição dos centrômeros, número e posição das regiões organizadoras nucleares (RONs), quantidade e distribuição da heterocromatina, composição do DNA repetitivo e conteúdo total do DNA (Poggio et al. 2008). Esses estudos podem gerar dados quantitativos e estatisticamente reproduzíveis que permitem estabelecer correlações com outros fatores como a morfologia, filogenia e as condições ambientais (Van-Lume et al. 2017).

Variações fenotípicas e padrões de diversidade biótica tendem a existir em gradientes ambientais, cujas causas ainda representam um enigma para os biólogos evolucionistas e biogeográficos. Certas diferenças cariotípicas terem o potencial de atuar como barreira genética e, consequentemente, podem revelar fortes evidências de estruturação populacional (Blöch et al. 2009; Amaro et al. 2012). Menezes et al. (2017) analisaram sequência de DNA multilocus, modelos de nicho climático e características cromossômicas para investigar e comparar os padrões filogeográficos de duas vespas sociais parapátricas do gênero *Synoeca*. Nesse trabalho, os autores observaram que as espécies de vespa exibiram padrões contrastantes de dispersão espaço-temporal e experimentaram alterações cromossômicas substanciais como variação no número e tamanho cromossômico, assim como padrões distintos em relação aos locais ricos em GC e AT. Essas alterações cromossômicas e a direção da dispersão populacional foram orientadas latitudinalmente ao longo da Mata Atlântica Brasileira.

A integração de análises filogeográficas e citogenéticas, baseadas em características estruturais e quantitativas do cariótipo, mostrou-se útil em estudos evolutivos e taxonômicos em vários grupos de angiospermas (Guerra 2000; Weiss-Schneeweiss e Schneeweiss 2013). Em *Nierembergia* (Solananeceae) uma combinação de dados cariotípicos dentro de uma filogenia datada foram usados para analisar como aconteceu a evolução cromossômica no grupo e inferir como os processos geológicos ou climáticos influenciaram à diversificação do gênero (Acosta et al. 2016). O uso de marcadores citogenéticos clássicos e moleculares, como coloração convencional e fluorescente, e hibridações por FISH do DNA ribossômico, permitiram demonstrar que *Nierembergia* está mais relacionada ao gênero *Bouchetia* do que *Leptoglossis*, assim como a existência de duas linhagens dentro do gênero. Essas duas linhagens se diferençam por uma diminuição no comprimento do cariótipo e no tamanho dos cromossomos. Portanto, quando as duas linhagens ancestrais de *Nierembergia* foram isoladas, ocorreram grandes divergências na evolução cromossômica e, em seguida, cada linhagem passou por especiação separadamente. Deste modo, os dados cariológicos fornecem outra fonte de marcadores para a compreensão da sistemática, padrões evolutivos e processos de divergência nas plantas (Crawford et al. 2005).

A combinação de caráteres cromossômicos com a morfologia, biogeografia e marcadores moleculares, muitas vezes auxiliam na identificação de casos de hibridação e rearranjos cromossômicos envolvidos na especiação (Weiss-Schneeweiss et al. 2008; Baltisberger e Horandl 2016; Chiarini e Gauthier 2016). No entanto, algumas vezes a morfologia cromossômica conservada gera dificuldade na identificação de marcadores espécie-específicos. No caso das espécies de Caesalpinia, o uso de outros marcadores cromossômicos, como à distribuição e quantidade de heterocromatina, pode contribuir na caracterização citomolecular do grupo assim como na diferenciação das espécies (Almeida et al. 2007; Van-Lume et al. 2017; Moreno et al. 2018; Rodrigues et al. 2018).

Nas angiospermas a heterocromatina aparentemente não apresenta uma distribuição aleatória, em vez disso, encontra-se em regiões cromossômicas preferenciais (Guerra 2000) e em alguns casos relacionadas com as condições ambientais. Em *Notolathyrus*, por exemplo, Chalup et al. (2015) evidenciaram padrões diferenciais de distribuição e conteúdo de bandas heterocromáticas $DAPI^+$, apresentando uma posição diferencial dependendo da distribuição geográfica das plantas. Nos biomas subtropicais o conteúdo de heterocromatina e o número de bandas era menor em comparação com os biomas temperados. Os dados evolutivos

cromossômicos revelaram que as espécies sul-americanas são um grupo homogêneo e monofilético da seção. Por outro lado, a variação na quantidade de heterocromatina não foi diretamente relacionada à variação no conteúdo de DNA das espécies de *Notolathyrus*, apesar disso, a correlação observada entre a quantidade de heterocromatina e algumas variáveis geográficas e bioclimáticas sugere que essa variação deve ter um valor adaptativo.

Recentemente alguns trabalhos têm mostrado diferentes padrões heterocromáticos usados como marcador citomolecular para contar relações evolutivas entre grupos (Deanna et al. 2018; Moreno et al. 2018; Wahlang et al. 2019). Por exemplo, em *Cynodon* (Poaceae) um gênero com ampla distribuição em áreas tropicais e subtropicais, diferentes eventos de ploidização causaram rearranjos estruturais, como deleções, inserção ou duplicações de sequências de DNA, permitindo variação na morfologia, tamanho e número dos cromossomos. Esses eventos resultaram em uma ampla diversidade de cariótipos, que contribuíram no isolamento reprodutivo e, consequentemente, para a especiação de *Cynodon*. As análises citogenéticas comparativas do grupo utilizando bandeamento CMA/DAPI e hibridações por FISH, mostraram que as espécies estão caracterizadas pela ausência ou presença de bandas de heterocromatina CMA ou DAPI e poucos sítios de DNA ribossômico. As análises PCM indicaram que o ancestral comum mais recente de *Cynodon* tinha poucos sítios de DNA ribossômico e diferentes números de bandas CMA/DAPI e que durante os eventos de poliploidização, houve perdas e ganhos de sequências heterocromáticas principalmente nas regiões centroméricas e dos braços curtos dos cromossomos (Chiavegatto et al. 2019). Dessa forma, análises comparativas dos números, quantidade e posições das sequências de DNA ribossômico e bandas heterocromáticas têm valor não apenas para a identificação de cromossomos e genomas, mas também para a elucidação de diferenças e relações evolutivas, ecológicas e taxonômicas entre grupos.

2.3 A CIÊNCIA UNIFICADORA: CITOGENÔMICA

A relação íntima entre a sequência de DNA e a estrutura e função cromossômica destaca a necessidade de integrar dados genômicos e citogenéticos para entender de maneira mais abrangente o papel que a arquitetura do genoma desempenha na plasticidade do genoma (Deakin et al. 2019). Embora os avanços da tecnologia NGS tornaram o sequenciamento de todo o genoma de um organismo accessível, existe muita informação sobre a composição de muitos genomas eucariotos que ainda não compreendemos. O genoma dos eucariotos está composto por diferentes tipos de sequências de DNA, que podem ser classificadas como

sequencias únicas e sequências repetitivas, sendo estas últimas divididas em duas grandes famílias, chamadas “repetições em tandem” e “repetições dispersas”. Cada uma dessas duas famílias é dividida em várias subfamílias (Figura 4) (Richard et al. 2008).

As repetições dispersas contêm geralmente todos os transposons, genes de RNA de transferência e genes parálogos, enquanto as repetições em tandem contêm sequências em tandem, o DNAr e o DNA satélite. Os mecanismos moleculares que criam e propagam as repetições dispersas e em tandem são específicos para cada classe e são utilizados para sua classificação (ver glossário de conceitos). Atualmente, sabe-se que os elementos repetitivos podem ser amplamente abundantes em alguns eucariotos, compondo mais de 50% do genoma humano (De Koning et al. 2011), mais do 70% do genoma do milho (Meyer 2001) e mais do 90% do genoma da cebola (Fu et al. 2019).

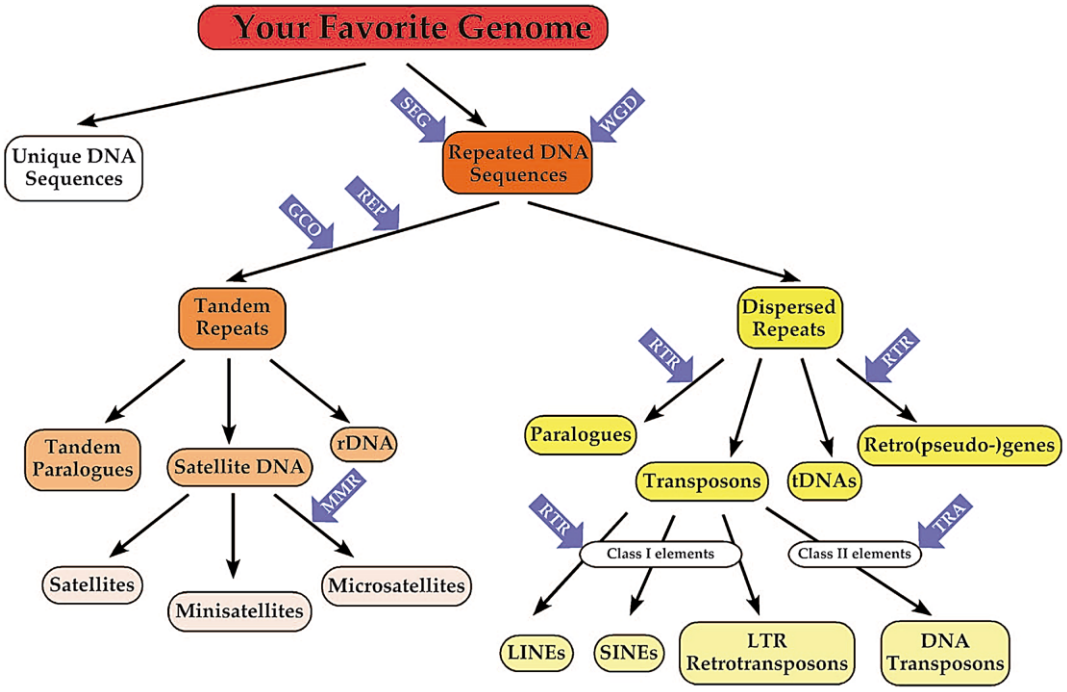


Figura 4. Sequências de DNA repetitivas em genomas eucarióticos e seus mecanismos de evolução, reproduzido de Richard et al. (2008). As duas categorias principais de elementos repetidos (repetições em tandem e repetições dispersas) são mostradas, juntamente com subcategorias, conforme descrito no texto. As setas azuis apontam para mecanismos moleculares envolvidos na propagação e evolução de sequências repetidas. REP, desvio de replicação; GCO, conversão de genes; WGD, duplicação de genoma inteiro; SEG, duplicações segmentares; RTR, transcrição reversa; TRA, transposição.

Devido a seu mecanismo de transposição "cortar e colar" e "copiar e colar”, os ET podem modificar a sequência cromossômica original, ocasionando diferentes efeitos e

rearranjos que influenciam a evolução da regulação gênica e sua função (Kazazian 2004; Liu et al. 2019). Os ETs fornecem agentes de evolução adaptativa e adaptação, criando variantes e diversidade genética. Em *Deschampsia* P.Beauv., uma Poaceae que ocorre na América do Sul e na Antártica, 73.3% do genoma total é representado por DNA repetitivo, sendo os ETs os mais abundantes e os DNA satélites os mais variáveis. Essa variação observada no padrão de DNA satélite poderia ter facilitado o isolamento reprodutivo entre as populações de *D. antarctica* devido à falha no reconhecimento homólogo dos cromossomos, e como consequência a sua diferenciação geográfica (González et al. 2017). Esse estudo destaca a importância de mudanças na organização cromossômica com base na amplificação diferencial não aleatória de unidades repetitivas de DNA no genoma.

Quanto à organização das sequências repetitivas, diferentes famílias de ET podem ocorrer em agrupadas em tandem ou apresentar-se espalhadas ao longo dos cromossomos (Gaeta et al. 2010; Macas et al. 2015; Horváth et al. 2017). A literatura indica que os retroelementos geralmente se acumulam em padrões dispersos pelos cromossomos, diferentemente dos DNAs satélites que formam clusters mais definidos (Heslop-Harrison e Schmidt 2007; Heslop-Harrison e Schwarzacher 2011; Ribeiro et al. 2017). No entanto, existem exemplos que relatam que os membros das superfamílias *Copia* e *Gypsy* têm um perfil heterogêneo, podendo ser localizados dispersos, agrupados ou uma combinação de ambos em alguns cromossomos (Belyayev et al. 2001; Gaeta et al. 2010). De Souza et al. (2018) reportaram distribuições independentes para cada família de retroelementos LTR, apresentando um acúmulo variado dessas sequências no genoma de *Eleocharis* R.Br. (Poaceae). Essa variabilidade foi maior em *E. montana*, onde as sondas Oryco, Del e CRM hibridizaram na metade dos cromossomos dessa espécie, em contraste à sonda Athila/Ogre-Tat teve sinais pontuais em todos os cromossomos. A variação na distribuição dos retroelementos indicam que cada família de LTR tem atividade independente no genoma e nos cromossomos, com diferentes histórias e destinos evolutivos.

Os novos métodos de sequenciamento de nova geração (NGS), combinado com a abordagem de *genome skimming*, tem permitido a identificação de diferentes famílias de elementos repetitivos em espécies não modelos, assim como a análise comparativa desses elementos na compreensão da composição, abundância e evolução de sequências repetitivas (Ekblom e Galindo 2011; Du et al. 2017; González et al. 2017; De Souza et al. 2018). Dentre as ferramentas em bioinformática recentes, destaca-se o RepeatExplorer que, por meio de uma

análise de *clustering*, identifica a composição e abundância de cada classe de elementos repetidos que compõem o genoma (Novák et al. 2013). As técnicas citogenéticas voltadas especificamente para as regiões de heterocromatina são complementares a essas abordagens NGS, revelando como os genomas são organizados estruturalmente em cromossomos, como eles são moldados pela interação de forças evolutivas e como as mudanças na estrutura dos cromossomos contribuem para a especiação (Dion-Côté et al. 2016). Tais abordagens integrativas em plantas começam a revelar que as alterações na estrutura cromossômica estão associadas à divergência e evolução.

Do ponto de vista dos cariótipos e da biologia cromossômica, o entendimento sobre a dinâmica e evolução das sequências repetitivas ainda é superficial, porque a diversidade da fração repetitiva de DNA depende não apenas dos genomas analisados, mas também da história evolutiva e o papel de cada um (De Souza et al. 2018). Em diferentes espécies do gênero *Paphiopedilum* Pfitzer, (Orchidaceae), sequências de DNA satélite especificas apresentaram uma distribuição cromossômica mais semelhante em espécies estreitamente relacionadas do que em espécies mais distantes filogeneticamente. Os autores sugerem que desde a origem do DNA satélite nesse grupo, houve claramente mudanças consideráveis em sua abundância e distribuição cromossômica entre as diferentes espécies, evoluindo rapidamente, e adquirindo um forte sinal filogenético (Lee et al. 2018).

Igualmente, *Arachis* L. (Fabaceae) é um gênero que constitui o pool genético secundário de amendoim, com espécies diploides e alopoliploides com diferentes genomas (A, B, D, F, G e K), e nesse grupo as sequências repetitivas têm desempenhado um papel importante para a diferenciação do genoma A e B no grupo. Nesse genoma, os retroelementos LTR e quatro DNAs satélites são os elementos mais abundantes e compõem as regiões de heterocromatina pericentromerica e/ou regiões eucromáticas. Análises comparativas revelaram que o satélite Agla_CL8sat está presente na heterocromatina dos genomas A, K e F, sugerindo que os DNA satélites estão intimamente relacionados entre os diferentes genomas e que foram amplificados diferencialmente a partir de uma biblioteca ancestral, levando a grandes mudanças nos padrões de heterocromatina (Samoluk et al. 2019).

2.4 O GRUPO CAESALPINIA

2.4.1 Caracterização do grupo

O grupo Caesalpinia é um grupo morfologicamente heterogêneo composto por árvores, arbustos, cipós ou herbáceas com uma grande quantidade de homoplasia e convergência morfológicas, que dificulta a determinação de carácteres diagnósticos únicos para cada clado (Lewis e Schrire 1995, Lewis 1998, Gagnon et al. 2013, 2016). O grupo inclui 27 gêneros e cerca de 225 espécies e, embora não tenha sinapomorfias morfológicas diagnósticas únicas para o clado como um todo, o grupo Caesalpinia pode ser reconhecido por uma combinação de características, como a presença de tricomas glandulares, espinhos, flores simétricas bilaterais com uma sépala inferior um tanto modificada e estames livres aglomerados ao redor do pistilo. As flores variam muito e podem ser fortemente modificadas, dependendo do sistema de polinização, e os frutos de cada clado são extremamente diversos, refletindo uma impressionante variação nas estratégias de dispersão de sementes (Gagnon et al. 2016).

Atributos morfológicos são utilizados para dividir o grupo em dois grandes subclados, o clado I contém todas as espécies que estão armadas com espinhos ao longo dos galhos (embora *Coulteria* careça de espinhos), e que possuem idioblastos na lâmina folhar. Por outro lado, o clado II contém apenas espécies que não possuem espinhos e quase todas as espécies que não apresentam idioblastos nas folhas (os últimos também estão ausentes em *C. mimosoides* no clado I (Lersten e Curtis 1996) e em *Haematoxylum* Gronov.). todas espécies do clado II são caracterizadas pela presença de estruturas glandulares multicelulares nas hastes, folhas e inflorescências (embora *Haematoxylum dinteri* Harms, *Caesalpinia mimosoides* Lam. e membros de *Coulteria* no clado I também possuam). No nível genérico, os frutos são altamente variáveis e do ponto de vista taxonômico são mais úteis como caracteres diagnósticos que as flores. Vários dos gêneros reconhecidos até agora podem ser diferenciados com base nas características dos frutos (Gagnon et al. 2016). No entanto, um conjunto de caraterísticas morfológicas combinados pode auxiliar na distinção de cada gênero, por exemplo, a presença de tricomas glandulares, espinhos, folhas geralmente pulvinadas, bipinadas ou pinadas definem o gênero *Tara* (Gagnon et al. 2016; LPWG, 2017). Nesse sentido, um enfoque de sistemática integrativa, utilizando outros conjuntos de caracteres (citogenéticos, bioquímicos, anatômicos, etc.) auxiliam na melhor caracterização dos gêneros no grupo.

O grupo Caesalpinia é distribuído nas regiões tropicais do mundo, com apenas pequenas incursões envolvendo apenas quatro gêneros no bioma temperado (por exemplo

Pomaria), apresentando um sinal claro de conservadorismo de nicho tropical (Gagnon et al. 2019). Muitos subclados do grupo estão restritos ao bioma suculento, mas outras espécies também ocorrem em savanas tropicais, florestas tropicais, manguezais costeiros ou outros habitats costeiros e áreas secas temperadas quentes e propensas à geada (Simpson et al. 2005, 2006). O grupo está praticamente ausente da Amazônia e pouco representado em florestas tropicais nos países neotropicais e na África, mas é mais comum nas florestas tropicais do Sudeste Asiático, onde ocorre os clado de cipós (Gagnon et al. 2019).

As espécies neotropicais crescem principalmente nos habitats sazonalmente secos e semiáridos, especialmente nas SDTF (Gagnon et al. 2016). A distribuição atual das SDTF na América do Sul é descontínua em grandes áreas desde a Caatinga no Nordeste do Brasil até o vale do Rio Uruguai (Werneck et al. 2011). Evidências fósseis e climáticas indicam que o bioma SDTF é relativamente antigo e data do Eoceno médio da América do Norte; sugerindo que a atual distribuição disjunta do SDTF representa os remanentes de um bioma contínuo (Werneck et al. 2011).

2.4.2 Estudos citogenéticos

As análises citogenéticas no grupo abrangem analises citogenéticas clássicas e moleculares. O primeiro trabalho realizado por Beltrão e Guerra (1990), revelou os números cromossômicos de diferentes espécies no grupo. Cangiano e Bernardello (2005) fizeram as primeiras descrições cromossômicas em células somáticas de três espécies de *Caesalpinia* L. (*C. gilliesii* Wall. ex Hook.,, *C. mimosifolia* Griseb. e *C. paraguariensis* (D.Parodi) Burkart) endêmicas da Argentina. Borges et al. (2012) descreveram cariotipicamente citotipos diploides e tetraploides de *Libidibia ferrea* (Benth.) L.P.Queiroz que se encontravam isolados reprodutivamente. Rodrigues et al. (2012) apresentaram o cariótipo, a morfometria cromossômica e o padrão de heterocromatina por bandeamento C em *Caesalpinia calycina* (=*Erythrostemon calycinus*), *Caesalpinia microphylla* (=*Cenostigma microphyllum*), *Caesalpinia pluviosa* var. *peltophoroides* (=*Cenostigma pluviosum*), e em *Caesalpinia ferrea* (=*Libidibia ferrea*). Adicionalmente, dois anos depois, Rodrigues et al. (2014) incluíram uma caracterização citogenética de outras 14 espécies, desta vez com ocorrência na América do Sul. López et al. (2014) relataram as descrições citogenéticas como número cromossômico e morfologia cromossômica de diferentes indivíduos de *Caesalpinia spinosa* (=*Tara spinosa*) de duas localidades peruanas. Em resumo, todos esses trabalhos descrevem um cariótipo de

2*n* = 24 cromossomos, de pequeno tamanho (~2 μm) e com uma morfologia predominantemente meta / submetacêntricos.

As primeiras abordagens citomoleculares no grupo começaram com bandeamento com fluorocromos, principalmente CMA e DAPI, realizadas por Souza e Benko-Iseppon (2004), em *Caesalpinia pulcherrima* (L.) Sw. Esse bandeamento revelou diversas bandas terminais (ricas em GC) em todos os cromossomos. Adicionalmente, Van-Lume et al. (2017) usando também bandeamento CMA/DAPI relataram em 20 espécies de Caesalpinia, um maior número de bandas em diferentes posições cromossômicas dependendo das espécies analisadas. Nesse trabalho, também foi apresentado o mapeamento cromossômico dos DNAr 5S e 35S, demonstrando pouca variação no número e na posição. Recentemente, Rodrigues et al. (2018) estudaram os locais de heterocromatina rica em GC e DNAr 35S para avaliar a diversidade do cariótipo em 10 espécies do grupo, revelando blocos de CMA/DAPI exclusivamente nas regiões terminais dos cromossomos, coincidindo com os locais de 35S DNA ribossômico em todas as espécies analisadas. Esses trabalhos relatam uma elevada diversidade citomolecular no grupo, reforçando a necessidade de continuar com as análises citogenéticas em outras espécies.

2.4.3 Caesalpinia grupo modelo

Quais atributos fariam o grupo Caesalpinia se tornar um grupo modelo para futuros estudos relacionando marcadores citomoleculares com traços ecológicos?

· A idade (55 Milhões de anos), distribuição geográfica (Fig. 5) e as taxas evolutivas constantes, sugerem estabilidade ao longo do tempo e adaptação ao Bioma Suculento em escala global.

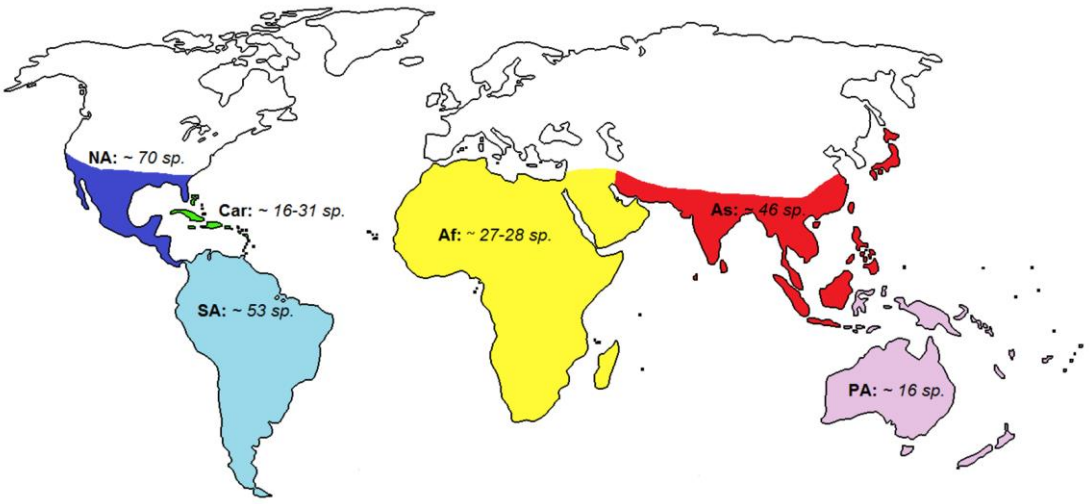


Figura 5. Áreas geográficas de ocorrências do grupo Caesalpinia, reproduzido de Gagnon et al. (2019). Azul escuro: América Central e América do Norte temperada e quente; Azul claro: América do Sul; Verde: as ilhas do Caribe, incluindo as Antilhas maiores e menores; Amarelo: África, incluindo Madagascar e a Península Arábica; Vermelho: sul da Ásia, da Índia à península indo / malaio; Lilás: região da Austrália / Pacífico, separada da Ásia pela linha de Wallace, entre Bornéu e Sunderland. É indicada uma estimativa do número de espécies do grupo que habitam cada área.

· As espécies têm um genoma principalmente diploide com 2*n* = 24 cromossomos (Beltrão e Guerra 1990; Souza e Benko-Iseppon 2004; Cangiano e Bernardello 2005; Borges et al. 2012; Rodrigues et al. 2012, 2014, 2018; Lopez et al. 2014; Van-Lume et al. 2017, 2019) com poucas espécies descritas como poliploides (Alves e Custódio 1989; Beltrão e Guerra 1990; Caponio et al. 2012). A poliploidia é disploidia são fontes de variação numérica na evolução das plantas com flores. Na sua ausência outros eventos ou processos contribuem com a diversificação e especiação em Caesalpinia.

· O grupo apresenta uma variação natural do tamanho genômico. No geral, os valores de 2C variam 7,73 vezes, de 0,92 pg/2C em *Cenostigma bracteosum* (Tul.) G.P.Lewis a 7,11 pg / 2C em *Pomaria lactea* (Schinz) B.B.Simpson & G.P.Lewis . As espécies do Clado II presentam genomas maiores (valores de 2C de 1,70 a 7,11 pg), em comparação com o Clado I, onde os valores de 2C variaram de 0,92 a 2,98 pg (Rodrigues et al. 2018; Souza et al. 2019). Adicionalmente essa variação foi relacionada com diferentes traços ecológicos. Souza et al. (2019) utilizando métodos comparativos filogenéticos para 40 espécies do grupo em um contexto espaço-temporal, observaram uma correlação positiva entre o GS e a latitude, assim como correlações adicionais com as variáveis de temperatura e precipitação. O tamanho genômico no grupo aumentou em 10% quando existia uma diminuição da temperatura de 0,39 °C. Como consequência, o GS no grupo é menor em temperaturas mais altas e latitudes mais

próximos do equador, enquanto que o GS aumenta em temperaturas menores e latitudes mais afastadas do equador (Fig. 6). Os autores sugerem que a variação do GS é o resultado do impacto contrastante do estresse de alta temperatura nas espécies que ocupam o bioma suculento em comparação com o estresse hídrico ou de temperatura menos pronunciado nas espécies das zonas temperadas.

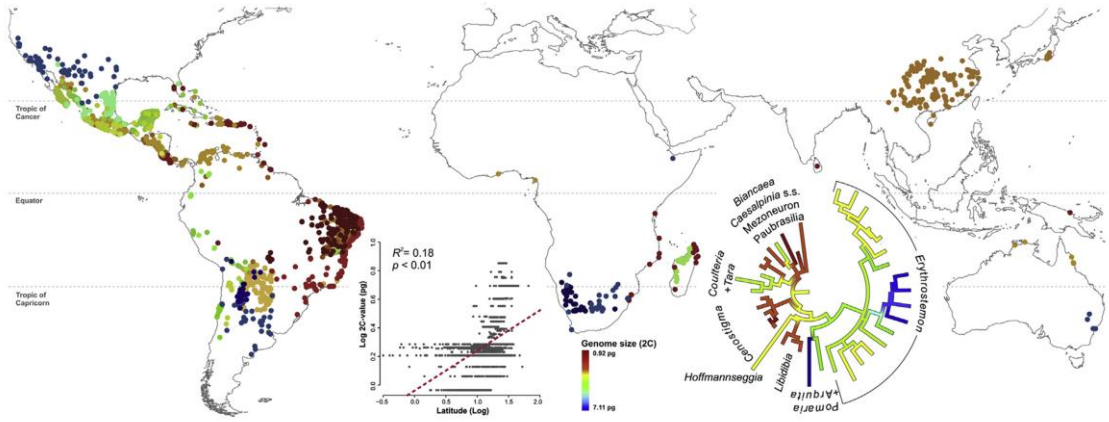


Figura 6. A distribuição geográfica da variação do GS para as espécies do grupo Caesalpinia, reproduzida de Souza et al. (2019). À direita, reconstrução ancestrais do tamanho genômico (2C pg) na filogenia. À esquerda, gráfico de latitude versus tamanho do genoma, mostrando a correlação entre essas variáveis. Cores quentes menor GS e cores frias maiores GS.

· A presença de polimorfismos citomoleculares. Van-Lume et al. (2017) ao fazer analises citomoleculares usando os fluorocromos CMA e DAPI em 20 espécies do grupo, encontrou três padrões distintos na heterocromatina proximal dos táxons estudados ($CMA^{+}/DAPI^{-}$, $CMA^{0}/DAPI^{-}$ e $CMA^{0}/DAPI^{0}$), assim como variabilidade nas quantidades de heterocromatina em cada cariótipo. Espécies dos gêneros *Guilandina*, *Paubrasilia*, *Cenostigma*, *Libidibia* e *Biancaea* apresentaram bandas proximais $CMA^{+}/DAPI^{-}$, espécies dos gêneros *Arquita*, *Balsamocarpon* e *Erythrostemon* mostraram bandas proximais $CMA^{0}/DAPI^{-}$ com menos intensidade que as outras bandas e o resto das espécies não apresentaram bandas proximais CMA^{+}, por exemplo, *Caesalpinia pulcherrima* e *Erythrostemon hughesii*, cujas bandas CMA^{+} são terminais. Adicionalmente, esses padrões mostraram uma relação com o ambiente. Ao investigar os padrões de variação filogenética, ambiental e geográfica, os autores descobriram que existe correlação entre os padrões CMA/DAPI e o nicho ecológico onde elas ocorrem (Fig. 7), permitindo caracterizar os

cariótipos de acordo com a presença das bandas CMA em três centros de diversidades geográficas: Cordilheira dos Andes, Mesoamérica (incluindo México, América Central, sul dos EUA e Caribe) e Nordeste do Brasil. Os autores sugerem que no grupo, a heterocromatina evoluiu de forma não aleatória e está correlacionada com as distribuições geográficas / nichos ecológicos das espécies, indicando que o ambiente desempenha um papel na fixação desses cariótipos, embora análises adicionais sejam necessárias para estabelecer os mecanismos causais subjacentes à conservação com o nicho observado.

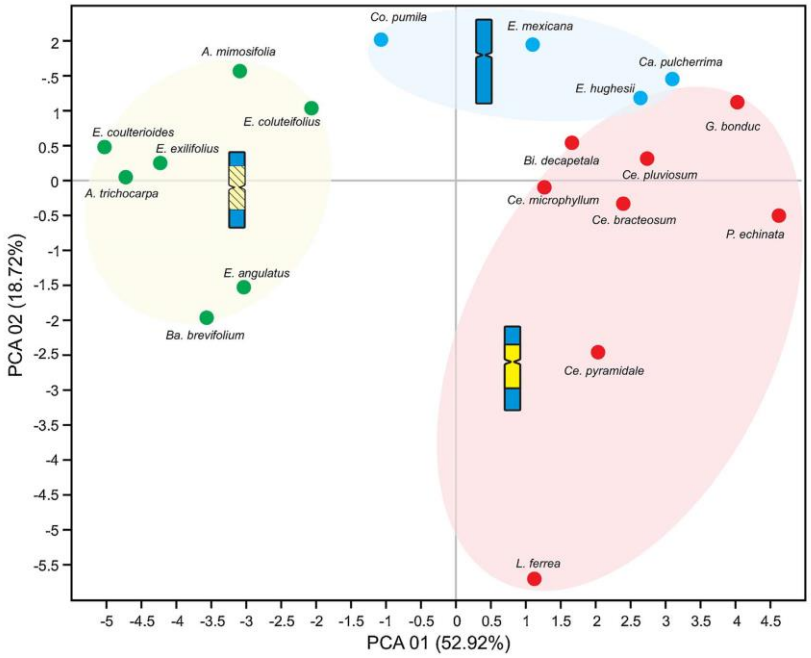


Figura 7. Gráfico de dispersão mostrando os resultados da Análise de Componentes Principais (PCA) baseada em indivíduos do grupo Caesalpinia, tomada de Van-Lume et al. (2017). Os dois primeiros eixos explicam 52,92% e 18,72% da variação entre as 19 variáveis climáticas. Os tipos heterocromáticos $CMA^{+}/DAPI^{-}$ (vermelho), $CMA^{0}/DAPI^{-}$ (verde) e $CMA^{0}/DAPI^{0}$ (azul) são rotulados nas elipses de 95% de inércia.

· Existência de homologia citogenômica das bandas heterocromáticas entre espécies filogeneticamente distantes. Van-Lume et al. (2019) por meio de analises NGS e ferramentas bioinformáticas caracterizaram a composição da heterocromatina de três espécies do grupo que ocorrem no Nordeste do Brasil, *Cenostigma microphyllum*, *Libidibia ferrea* e *Paubrasilia echinata* e testaram a homologia dessas regiões cromossômicas. Estimou-se que as frações repetitivas representam 41,70%, 38,44% e 72,51% em *C. microphyllum*, *L. ferrea* e *P. echinata*, respectivamente. Os elementos repetitivos do tipo LTR da superfamília Ty3/*Gypsy*

foram os mais abundantes nos três genomas, especificamente a linhagem Tekay, seguido da linhagem Athila e poucos DNA satélites. As hibridações por FISH revelou uma distribuição proximal para os elementos Tekay espécies-específico em todos os cromossomos das três espécies, co-localizados com as bandas CMA^+, e padrões proximal na heterocromatina de *L. ferrea* ou restritos aos cromossomos acrocêntricos de *C. microphyllum* para o elemento Athila (Fig. 8). A análise filogenética baseada nas sequências de DNA indicou que os elementos Tekay identificados nas três espécies do nordeste do Brasil formavam um grupo monofilético. Os autores sugerem uma colonização ancestral da linhagem Tekay na heterocromatina proximal, sendo, a composição atual da heterocromatina pericentromérica nessas espécies o resultado de uma combinação da manutenção de uma distribuição ancestral do Tekay com um acúmulo espécies-específico de outras repetições.

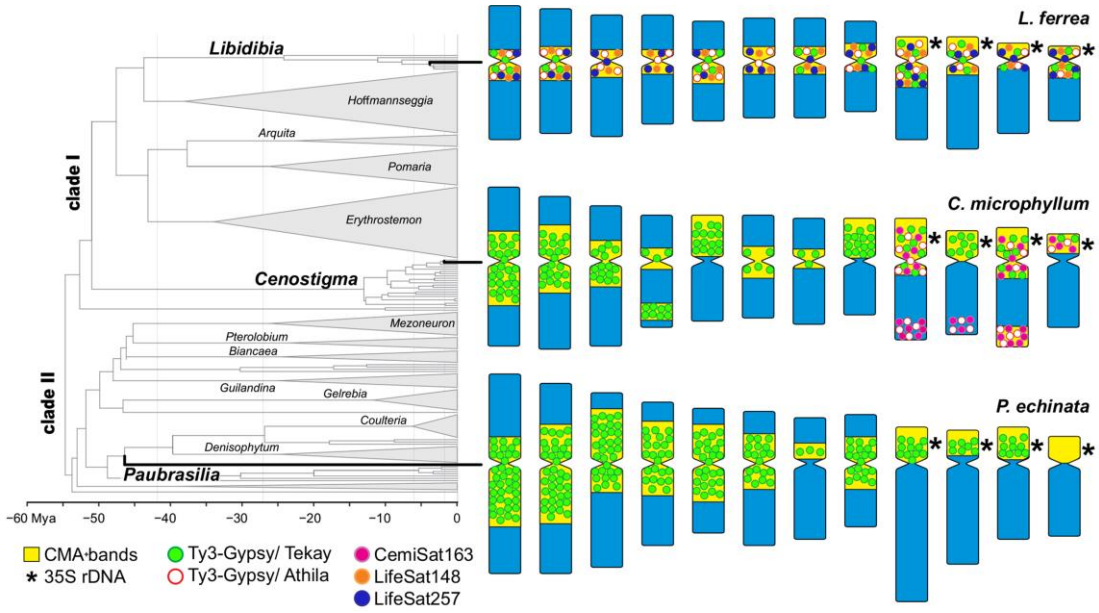


Figura 8. Relações filogenéticas (Gagnon et al. 2019) e mapeamento citogenômico comparativo de bandas heterocromáticas $CMA^+/DAPI^-$, principais retrotransposons LTR Ty3/*Gypsy* (Tekay e Athila) e DNAs satélite em três espécies pertencentes ao grupo Caesalpinia do nordeste do Brasil, reproduzido de Van-Lume et al. (2019). As espécies analisadas são destacadas no cladograma com uma linha preta grossa.

Esse histórico demonstra quão recente e pouco, porem interessante, é o conhecimento de citogenômica do grupo. O desenvolvimento do NGS, e novas ferramentas bioinformáticas cada vez mais accessíveis, espécies não-modelo tornaram-se mais passíveis de análises aprofundadas de caracterização, principalmente do componente repetitivo do DNA de seus genomas, como é o caso das espécies do grupo. No entanto, ainda há muito por estudar e analisar, pelo que consideramos o grupo Caesalpinia um grupo modelo muito conveniente

para estudos citogenômicos, biogeográficos e evolutivos que permitam fornecer respostas para as questões ainda não resolvidas. Por fim, supõe-se que esse conhecimento permita extrapolar as informações obtidas para outros organismos, considerando uma origem comum das espécies.

Tabela 1: Número de espécies do grupo Caesalpinia que possui caracterização citogenética e citomolecular disponíveis. Pg= picogramas. rDNA= DNA ribossômico.

Gênero/ espécies	2C (pg)	2*n*	Sítios de rDNA 5S/35S	Referência (2C/ 2*n*)
Arquita Gagnon, G.P.Lewis & C.E.Hughes				
A. trichocarpa (Griseb.) Gagnon, G.P.Lewis & C.E.Hughes	3.57	24	1-4	Souza et al. (2019)/ Van-Lume et al. (2017)
A. mimosifolia (Griseb.) E. Gagnon, G.P. Lewis & C.E.Hughes	-	24	2-4	/ Van-Lume et al. (2017)
Balsamocarpon				
B. brevifolium Clos.	-	24	-	/ Van-Lume et al. (2017)
Biancaea Tod.				
B. decapetala (Roth) O.Deg.	1.93	24	-	Souza et al. (2019)/ Van-Lume et al. (2017)
Caesalpinia L.				
C. pulcherrima Sw.	1.61	24	1-5	Souza et al. (2019)/ Van-Lume et al. (2017)
	3.60			Ohri et al. 2004
Cenostigma Tul.				
C. bracteosum (Tul.) Gagnon & G.P.Lewis	0.92	24, 48	1-4	Souza et al. (2019)/ Alves and Custódio (1989)
C. eriostachys (Benth.) Gagnon & G.P.Lewis	1.76	-	1-4	Souza et al. (2019)
C. microphyllum (Mart. ex G.Don) Gagnon & G.P.Lewis	1.88	24	1-4	Souza et al. (2019)/ Van-Lume et al. (2017)
C. pluviosum (DC.) Gagnon & G.P.Lewis	1.88	24	1-4	Souza et al. (2019)/ Van-Lume et al. (2017)
C. pyramidale (Tul.) Gagnon & G.P.Lewis	1.80	24	1-4	Souza et al. (2019)/ Van-Lume et al. (2017)
Coulteria Kunth				
C. mollis Kunth	1.72	-	-	Souza et al. (2019)
C. pumila (Britton & Rose) Sotuyo & G.P.Lewis	1.92	24	-	Souza et al. (2019)/ Van-Lume et al. (2017)
Erythrostemon Klotzsch				
E. acapulcensis (Standl.) Gagnon & G.P.Lewis	2.27	-	-	Souza et al. (2019)
E. angulatus (Hook. & Arn.) Gagnon & G.P.Lewis	4.03	24	-	Souza et al. (2019)/ Van-Lume et al. (2017)
E. calycinus (Benth.) L.P.Queiroz	1.54	24	-	Rodrigues et al. (2018)
E. coccineus (G.P.Lewis & J.L.Contr.) Gagnon & G.P.Lewis	2.34	-	-	Souza et al. (2019)
E. coluteifolius (Griseb.) Gagnon & G.P.Lewis	5.69	24	1-5	Souza et al. (2019)/ Van-Lume et al. (2017)
E. coulterioides (Griseb.) Gagnon & G.P.Lewis	6.18	24	1--	Souza et al. (2019)/

E. exilifolius (Griseb.) Gagnon & G.P.Lewis	5.62	24	1-7	Van-Lume et al. (2017) Souza et al. (2019)/ Van-Lume et al. (2017)
E. exostemma (Sessé & Moc. ex DC.) Gagnon & G.P.Lewis	2.29	-	-	Souza et al. (2019)
E. gilliesii (Hook.) Klotzsch	4.94	24	-	Souza et al. (2019)/ Van-Lume et al. (2017)
E. hintonii (Sandwith) Gagnon & G.P.Lewis	2.29	-	-	Souza et al. (2019)
E. hughesii (G.P.Lewis) Gagnon & G.P.Lewis	2.18	24	1-4	Souza et al. (2019)/ Van-Lume et al. (2017)
E. melanadenius (Rose) Gagnon & G.P.Lewis	2.24	-	-	Souza et al. (2019)
E. mexicanus (A.Gray) Gagnon & G.P.Lewis	4.01	24	-	Ohri et al. 2004/ Fedorov (1974)
E. nelsonii (Britton & Rose) Gagnon & G.P.Lewis	2.27	-	-	Souza et al. (2019)
E. pannosus (Brandegee) Gagnon & G.P.Lewis	2.47	-	-	Souza et al. (2019)
E. placidus (Brandegee) Gagnon & G.P.Lewis	2.59	-	-	Souza et al. (2019)
E. sousanus J.L.Contr., Sotuyo & G.P.Lewis	3.41	-	-	Souza et al. (2019)
E. yucatanensis (Greenm.) Gagnon & G.P.Lewis	2.15	-	-	Souza et al. (2019)
Guilandina L.				
G. bonduc L.	1.34	24	1-4	Souza et al. (2019)/ Van-Lume et al. (2017)
Hoffmannseggia Cav. H. doelli Phil	2.40	-	-	Souza et al. (2019)
Libidibia Schltdl.				
L. coriaria Schltdl.	1.70	24	-	Ohri, 1998/ Kumari and Bir (1989)
L. ferrea (Mart. ex Tul.) L.P.Queiroz	1.83	24, 48, 72	1/5-4/8/12	Souza et al. (2019)/ Van-Lume et al. (2017)
L. paraguariensis (D.Parodi) G.P.Lewis	1.81	24	-	Souza et al. (2019)/ Cangiano e Bernardello (2005)
L. punctata (Willd.) Britton	1.70	-	-	Souza et al. (2019)
Mezoneuron Desf.				
M. hildebrandtii Vatke	2.55	-	-	Souza et al. (2019)
Paubrasilia Gagnon, H.C.Lima & G.P.Lewis				
P. echinata (Lam.) Gagnon, H.C.Lima & G.P.Lewis	2.89	24	1-4	Rodrigues et al. (2018)/ Van-Lume et al. (2017)
Pomaria Cav.				
P. lactea (Schinz) B.B.Simpson & G.P.Lewis	7.11	-	-	Souza et al. (2019)
Tara Molina				
T. cacalaco (Bonpl.) Molinari & Sanchez Och.	2.98	24	-	Souza et al. (2019)/ Sarkar et al. (1982)
T. spinosa (Molina) Britton & Rose	2.65	24	-	Souza et al. (2019)/ Bandel (1974)
T. vesicaria (L.) Molinari, Sanchez Och. & Mayta	2.78	24	-	Souza et al. (2019)

3 RESULTADOS

3.1 ARTIGO 1- REVISITING THE CYTOMOLECULAR EVOLUTION OF THE CAESALPINIA GROUP (LEGUMINOSAE): A BROAD SAMPLING REVEALS NEW CORRELATIONS BETWEEN CYTOGENETIC AND ENVIRONMENTAL VARIABLES

Mata-Sucre Y[1], Costa L[1], Gagnon E[2], Lewis GP[3], Leitch IJ[3], Souza G[1]

Artigo submetido à revista **Plant Systematic and Evolution (Qualis A3)**

Revisiting the cytomolecular evolution of the Caesalpinia Group (Leguminosae): a broad sampling reveals new correlations between cytogenetic and environmental variables

Yennifer Mata-Sucre [1], Lucas Costa [1], Edeline Gagnon [2], Gwilym P Lewis [3], Ilia J Leitch [3], Gustavo Souza [1]

[1] Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Rua Nelson Chaves S/N, Cidade Universitária, Recife, PE. 50670-420, Brazil.

[2]Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5NZ, UK

[3] Comparative Plant and Fungal Biology Department, Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK.

Correspondence author: lgrsouza@hotmail.com

Abstract

The pantropical Caesalpinia Group includes 225 species in 27 monophyletic genera, and the group has undergone recent phylogenetic, taxonomic and biogeographic revisions. Previous works have reported a diverse pattern of heterochromatin distribution related to ecological niche/ geographic distribution and variation in genome size, also correlated with environmental variables. In order to investigate the relationship between cytogenetic and ecological traits using the Caesalpinia Group as a model, new cytomolecular data (chromosome number and morphology, CMA/DAPI staining and number and position of 5S and 35S rDNA sites) for 14 species in six genera were generated. These data were analysed by phylogenetic comparative methods. All species studied have $2n = 24$ (16M/SM + 8A) and most of them just one pair of 5S rDNA sites and two to five pairs of 35S rDNA sites. Three heterochromatic patterns were observed on the chromosomes: (i) pericentromeric $CMA^+/DAPI^-$ bands, (ii) pericentromeric $CMA^0/DAPI^-$ bands, and (iii) terminal $CMA^+/DAPI^-$ bands. The '*Coulteria* + *Tara*' and '*Arquita* + *Balsamocarpon* + *Erythrostemon* + *Pomaria*' clades (except for *E. gilliesii*, *E. hughesii* and *E. mexicanus*) independently showed $CMA^0/DAPI^-$ bands associated with larger genomes and geographic distributions at higher

latitudes. We statistically demonstrate that heterochromatin (CMA/DAPI intensity along the chromosome), genome size and latitude are autocorrelated in the Caesalpinia Group. On the other hand, we found a non-significant correlation between genome size and amount of heterochromatin. We argue that environmental factors associated with different latitude may have played a role in contributing to the diversification of the heterochromatin in Caesalpinia Group.

Key words: CMA/DAPI staining, fluorescent *in situ* hybridization, genome size, heterochromatin, phylogenetic comparative methods.

Introduction

The repetitive fraction of the eukaryotic genome is very complex and variable. The differential accumulation and elimination of specific repeats are key drivers of genome size variation in flowering plants (Bennetzen and Wang 2014; Pellicer et al. 2018). Part of this genome fraction can be cytologically revealed as heterochromatin by several techniques such as C-banding (Schwarzacher and Schweizer 1982) or differential fluorochrome staining (Schweizer 1976; Guerra 2000). Diverse studies have described polymorphisms in heterochromatin distribution and their potential biological significance (Ambrožová et al. 2011; Siljak-Yakovlev et al. 2017; Van-Lume and Souza 2018) revealing the importance of this type of study in the interpretation of plant biodiversity.

Recent cytogenetic studies have revealed that some heterochromatic polymorphisms can be related to ecological traits such as latitude (Menezes et al. 2017), altitude (Bilinski et al. 2017), historical changes in sea level (Acosta et al. 2016), or ecological niche (Van-Lume et al. 2017). In general, the heterochromatin comprises repetitive sequences that often evolve quickly and, hence, can vary in nucleotide composition, abundance and chromosomal distribution, impacting genome structure and/or genome size (Plohl et al. 2008). The Caesalpinia Group (Leguminosae) presents an interesting model to test these relationships because the group is karyotypically stable in terms of chromosome number and ploidy level and yet has ecologically-differentiated cytotypes, with species showing a correlation between the composition and distribution of heterochromatin, geographic distribution and ecological niche of the species (Van-Lume et al. 2017). In addition, a correlation between genome size and latitude/temperature has also been demonstrated in this group (Souza et al. 2019),

highlighting that species of the Caesalpinia group represent clear examples of plants showing a link between genomic and environmental variables.

The Caesalpinia Group comprises 27 genera and approximately 225 species of mainly arboreal and shrubby species (several are suffrutescent) (Gagnon et al. 2016, 2019). The group is distributed pantropically, with the Americas as the center of species diversity. In South America, the group represents an important ecological component with a higher intra/intergeneric diversity and abundance in tropical forests, especially in seasonally dry tropical forests (SDTF) within the Succulent Biome (Dryflor et al. 2016; Gagnon et al. 2019). Several species of this group are well-known and of considerable economic importance, such as *Paubrasilia echinata* (Lam.) Gagnon, H.C.Lima & G.P.Lewis known as "pau-brasil" (Brazil wood), the national tree of Brazil and the only wood used to make professional high-quality violin bows (Gasson et al. 2009). *Cenostigma pyramidale* (Tul.) Gagnon & G.P. Lewis is valued in the Caatinga region of northeast Brazil as one of several species used for firewood and charcoal production (Gasson et al. 2009). Other genera, such as *Erythrostemon* Klotzsch, *Libidibia* Schltdl., *Mezoneuron* Desf., *Tara* Molina and *Pomaria* Cav., contain species which are widely cultivated in warm climates, either as attractive ornamentals or for their medicinal properties (Castañeda et al. 2017). The group has a complex taxonomic history and has undergone recent nomenclatural alterations based mostly on molecular phylogenetic studies (Manzanilla and Bruneau 2012; Gagnon et al. 2013, 2015, 2016). Nevertheless, an integrative approach using additional sets of characters (cytogenetic, phylogenetic, biochemical, anatomical, etc.) may also help to further elucidate the systematics and evolution of the group.

Many species in the Caesalpinia group have been extensively analysed cytogenetically (e.g. Beltrão and Guerra 1990; Borges et al. 2012; Rodrigues et al. 2014, 2018; Van-Lume et al. 2017), and so far, they all present a stable karyotype ($2n = 24$), with small (~2 μm) predominantly meta/submetacentric chromosomes (Beltrão and Guerra 1990; Rodrigues et al. 2014, 2018; Van-Lume et al. 2017). Double staining with chromomycin A_3^+ (CMA) and 4',6-diamidino-2-phenylindole (DAPI) fluorochromes revealed the presence of heterochromatin on terminal or pericentromeric chromosomal regions, and high variation in the intensity of CMA and DAPI staining along the chromosomes (Van-Lume et al. 2017; Rodrigues et al. 2018). Thus, the CMA/DAPI index (CDI) was introduced to mathematically express the ratio between the intensity of the CMA and DAPI fluorescence in the pericentromeric region of the

chromosomes to highlight further diversity in the distribution and composition of heterochromatin (Van-Lume et al. 2017). The CDI discriminated three distinct patterns ($CMA^{+}/DAPI^{-}$, $CMA^{0}/DAPI^{-}$ and $CMA^{0}/DAPI^{0}$) in the pericentromeric heterochromatin of 20 species (Van-Lume et al. 2017). These three heterochromatic patterns were shown to be correlated with the main centers of diversity in the neotropical species of the Caesalpinia Group: i.e. (i) the Andes ($CMA^{0}/DAPI^{-}$), (ii) Mesoamerica ($CMA^{0}/DAPI^{0}$), and (iii) Northeastern Brazil ($CMA^{+}/DAPI^{-}$) (Van-Lume et al. 2017).

Despite these cytogenetic insights, there are still only a few molecular cytogenetic studies for the Caesalpinia Group species. For example, the distribution of ribosomal DNA (rDNA) was characterized for 20 species, and in general shows conservation of the number/position of 5S rDNA (a single pair per karyotype) and diversity in the number of 35S rDNA, ranging from three to seven pairs situated only at the terminal region of the chromosomes (Van-Lume et al. 2017; Rodrigues et al. 2018). In three species of northeast Brazil (i.e. *Cenostigma microphyllum* [Mart. ex G.Don] Gagnon & G.P.Lewis, *Libidibia ferrea* [Mart. ex Tul.] L.P.Queiroz and *Paubrasilia echinata*) the composition of heterochromatin was genomically investigated (Van-Lume et al. in press). The most abundant repetitive element in the heterochromatin in all three species were Ty3/Gypsy retrotransposons belonging to the Tekay lineage. However, Ty3/Gypsy Athila and satellite DNA were also identified in some CMA^{+} bands of *C. microphyllum* and *L. ferrea* (Van-Lume et al. in press). In species with numerically stable karyotypes, as seen in Caesalpinia Group species, the physical mapping of heterochromatin and rDNA sites have been used in classical cytotaxonomic studies (Seijo et al. 2004; Silvestri et al. 2015; Moreno et al. 2018). More recently, the use of phylogenetic comparative approaches (PCM) which integrate cytogenetic and phylogenetic data to identify karyotypic synapomorphies supporting clades identified from molecular phylogenies (Moreno et al. 2015; Chiarini et al. 2018; Lee et al. 2018; Ribeiro et al. 2018) are increasing. These approaches represent a new frontier in plant cytogenetics, being used to shed insights on genomic evolution (Puttick et al. 2015; Kolano et al. 2016; Costa et al. 2017; Van-Lume et al. 2017; Sader et al. 2019).

The objective of this study was to characterize the heterochromatin and rDNA distribution in chromosomes of Caesalpinia Group species, specifically investigating by PCM the relationships among heterochromatin, genome size and environmental variables. We provide new cytomolecular analysis for 14 species of six genera in the group based on chromosome

number and morphology, double staining with the fluorochromes CMA and DAPI and fluorescent *in situ* hybridization (FISH) for 5S and 35S ribosomal DNA (rDNA). Our sampling included species of the genera *Coulteria*, *Erythrostemon*, *Libidibia*, *Mezoneuron*, *Tara* and *Pomaria*, significantly increasing the number of species previously analysed by Van-Lume et al. (2017) and Rodrigues et al. (2018). Additionally, the molecular phylogeny of Gagnon et al. (2016) was used to interpret karyotypic evolution and the relationships between cytogenetic and ecological data. By combining the new data obtained here together with previous studies, we aimed to address three questions in the Caesalpinia Group: 1) What is the extent of heterochromatic variability in the broad taxonomic sampling? 2) Are the Caesalpinia Group clades (Gagnon et al. 2016) supported by karyotypic synapomorphies? 3) How are heterochromatin, genome size and ecological variables interrelate?

Materials and Methods

Plant material

A total of 14 species (on average six individuals per species) belonging to the genera *Coulteria*, *Erythrostemon*, *Libidibia*, *Mezoneuron*, *Pomaria* and *Tara* were analysed (Table 1). The seeds were obtained from the Millennium Seed Bank (Royal Botanic Gardens, Kew, U.K.) or from field collections (see Table 1), then germinated and grown in the experimental station of the Laboratory of Cytogenetics and Plant Evolution, in Recife, northeastern Brazil.

Chromosomal preparations and CMA/DAPI banding

Root tips obtained from germinated seeds were pre-treated with 0.002 M 8-hydroxyquinoline for 5 h at 18 °C, fixed in ethanol:acetic acid (3:1, v/v) for 2–24 h at room temperature and stored at -20 °C. For preparation of slides, fixed root tips were washed in distilled water and digested in a 2% (w/v) celullase (Onozuka)/20% (v/v) pectinase (Sigma) solution at 37 °C, for 90 min. Meristems were macerated in a drop of 45% acetic acid and spread on a hot plate following Ruban et al. (2014).

The CMA/DAPI double-staining technique was used for fluorochrome banding following Vaio et al. (2018) with few modification (stained with CMA 0.1 mg/mL for 60 min). Subsequently, the slides were aged for three days before analysis in an epifluorescence Leica DMLB microscope with filters for CMA and DAPI. Images were captured with a Cohu CCD

video camera using the Leica QFISH software. Finally, the best images were edited in Adobe Photoshop CS3 version 10.0.

Fluorescent in situ *hybridization (FISH)*

All *in situ* hybridizations followed Pedrosa et al. (2002). To localize the rDNA sites, 5S rDNA (D2) from *Lotus japonicus* labelled with Cy3-dUTP (GE) and 35S rDNA (p*Ta*71) from *Triticum aestivum* labelled with digoxigenin-11-dUTP (Roche) were used as probes. Labelling of probes was done by nick translation. The 35S rDNA probe was detected with sheep anti-digoxigenin FITC conjugate (Roche) and amplified with goat anti-sheep FITC conjugate (Serotec). The hybridization mixture contained formamide 50% (v/v), dextran sulfate 10% (w/v), 2× SSC and 5 ng μL^{-1} of each probe. The slides were denatured at 75 °C for 5 min. Stringent washes were performed, reaching a final stringency of approx. 76%. Images of the best metaphase plates were captured as indicated above.

Chromosomal measurements

Three metaphases of each species showing clear chromosome morphology were measured using Adobe Photoshop CS3 version 10.0 software. The ratio of chromosome arms (AR = the long arm length/short arm length) was used to classify the chromosomes as metacentric (AR = 1-1.49), submetacentric (AR = 1.50–2.99), or acrocentric (AR > 3.00), following Guerra (1986). Average lengths of the entire chromosome complement, homologous pairs and size of chromosome CMA bands and rDNA sites were measured and compared for the construction of idiograms in Drawid V0.26 program (Kirov et al. 2017). Intensity measurements were made at 10 points along the largest chromosomal pair in three metaphases per species using ImageJ 1.8v. The CDI (CMA/DAPI index; Van-Lume et al. 2017) was used to express the ratio between the intensity of the CMA and DAPI fluorescence in the pericentromeric region of the chromosomes. Thus, (i) CDI < 0.90 indicates CMA^- bands; (ii) CDI = 1 ± 0.1 represents neutral CMA^0 centromeres; and (iii) CDI > 1.1 indicates CMA^+ bands. Based on these results, the pericentromeric region was categorized as: (1) $CMA^+/DAPI^-$, (2) $CMA^0/DAPI^-$ or (3) $CMA^0/DAPI^0$ (Van-Lume et al. 2017).

Phylogenetic Comparative Method

In order to investigate heterochromatic patterns within the Caesalpinia Group, species occurrence data were downloaded from the Global Biodiversity Information Facility (GBIF)

website (https://www.gbif.org) and a distribution map was plotted using the software QGIS v. 2.18 (QGIS Development Team 2014). Two criteria were used to minimize the effect of erroneous GBIF distribution data: i) we carried out a taxonomic survey to filter only valid names based on the most recent Caesalpinia Group classification (Gagnon et al. 2016); and ii) only species with *vouchers* deposited in herbaria were recorded. The data were cleaned to exclude oceanic points and locations that were unlikely to be natural occurrence (e.g., occurrences in botanical gardens at more northern latitudes). For the cultivated species included in our study (*Biancaea decapetala*, *Caesalpinia pulcherrima*, *Cenostigma pluviosum*, and *Erythrostemon gilliesii*) only their neotropical distribution was considered because Africa and Asia are outside the scope of our study. From the collection sites of every species, we extracted climatic variables from the WorldClim 1.4 (5 min) generic grid format (Hijmans et al. 2005), utilizing the package "raster" 2.6–7 (Hijmans 2017) implemented in the R software 3.3.3 (R Core Team 2017).

Data analyses were performed in R software v.3.0.1 (R Core Team 2011). Climatic niches [sets of temperature and precipitation conditions where a species can occur (Bonetti and Wiens, 2014)] of karyotypes were also compared using a dimension-reducing principal component analysis (PCA) for all 19 bioclim variables using the *prcomp* function of the "stats" package in R Studio (R Core Team 2017). The PCA biplot was constructed with the R package "*factoextra*" (Kassambara and Mundt 2017) and the biplot was converted to a density plot with the package "*ggplot*" (R Core Team 2017) to improve data visualization.

For comparative analysis, were used the CMA/DAPI patterns and CDI coefficient of the 14 newly characterized species and the 18 species already published by Van-Lume et al. (2017) plotted in the molecular phylogeny of Gagnon et al. (2016). In total, we included 13 genera from the two main lineages of the Caesalpinia Clade (Gagnon et al. 2016). Clade I includes *Coulteria* + *Tara* + *Guilandina* + *Biancaea* + *Paubrasilia* + *Mezoneuron* + *Caesalpinia sensu stricto*, and Clade II includes *Erythrostemon* + *Cenostigma* + *Libidibia* + *Balsamocarpon* + *Pomaria* + *Arquita* (Fig. 4). Genome size data were obtained from Souza et al. (2019). Because the estimated heterochromatin amounts and CDI coefficient are distributed non-randomly with respect to phylogeny, PICs were obtained considering phylogenetic relationships among taxa (Felsenstein 1985; Harvey and Pagel 1991).

Correlations between genome size and the total amount of heterochromatin were investigated in two ways: (i) the amount of heterochromatin, as estimated from the extension of CMA

bands along the chromosome and (ii) intensity of CMA/DAPI staining, evaluated by calculating CDI values for all species analysed. Potential correlation were evaluated by PICs analysis implemented in the package *ape* v.4.0 (Paradis 2012). The Comparative Analysis by Independent Contrasts (CAIC) software package (Purvis and Rambaut 1995) was used to identify and calculate PICs using three different phylogenetic frameworks: (i) the entire phylogenetic tree containing all 32 species with cytological data, (ii) a tree that just included the nine species in Clade I, and (iii) a tree that just included the 20 Clade II species. As normality is not an assumption for Pearson correlation analyses and as various transformations (e.g. square root or log) did not improve skewness of the cytological data, we used untransformed data for all analyses.

To assess the variation of the CDI coefficient in relation to genome size and latitude, a multiple regression analysis was performed including all 32 species sampled. The *scatter3D* function in the *plot3D* package (Soetaert 2014) was used to create a three-dimensional scatterplot graph to visualize how the CDI depends linearly on the genome size and latitude as predictor variables. A 3D plane of tendency and standard statistical analyses were also performed using the same package, but without considering phylogenetic-statistical analyses. CorelDRAW version X7 software was used to plot graphic data and draw the tree topology.

Results

Karyotype and heterochromatin distribution in Caesalpinia Group species

Analysis of all 14 species showed stability in chromosome number ($2n = 24$) and morphology with 16 meta/submetacentric and eight acrocentric chromosomes in each karyotype (Fig. 1, Table 1, Online Resource 1). The total chromosome length ranged from 1.50 µm in *Tara spinosa* to 9.1 µm in *Pomaria lactea* (Online Resource 1). Measurements of the percentage of the genome occupied by CMA/DAPI-staining heterochromatic bands, the type of CMA/DAPI staining and the number of rDNA sites are listed in Table 1.

Double staining analyses using CMA and DAPI fluorochromes revealed $CMA^{+}/DAPI^{-}$ bands on the short arms of the acrocentric chromosomes in all species analysed. In the pericentromeric region of the chromosomes, three patterns of CMA/DAPI banding were identified, corresponding to those previously described by Van-Lume et al. (2017). Only one species (*Mezoneuron hildebrandtii*) showed no heterochromatic $CMA^{+}/DAPI^{-}$ bands in the pericentromeric region (Fig. 1n). Eleven species (*Coulteria mollis*, *C. pumila*, *C. platyloba*,

Erythrostemon pannosus, *E. angulatus*, *E. placidus*, *E. calycinus*, *Pomaria lactea*, *Tara cacalaco*, *T. spinosa* and *T. vesicaria*) presented pericentromeric $CMA^0/DAPI^-$ bands (Fig. 1) while the remaining two species (*Libidibia coriaria* and *L. punctata*) showed pericentromeric $CMA^+/DAPI^-$ bands (Fig. 1l-m). The intensity of the bands varied between the species (Online Resource 1). The percentage of heterochromatin in each karyotype, based on CMA/DAPI staining, ranged from 4.67% (*M. hildebrandtii*) to 51.07% (*E. calycinus*, Fig. 1g) of the total chromosome complement (Table 1).

Chromosomal distribution of rDNA sites

In general, FISH analyses revealed a single 5S rDNA site per karyotype (Fig. 2), except in *M. hildebrandtii* (Fig. 2j) and *P. lactea* (Fig. 2k) which had two sites, and *T. spinosa* (Fig. 2m) which had three. Most species presented a 5S rDNA site located in the interstitial or terminal region of submetacentric chromosomes. In contrast, in the three *Coulteria* species analysed and *M. hildebrandtii*, the 5S rDNA site was located on the short arm an acrocentric chromosome in synteny with the 35S rDNA (Figs. 2 a-c, j and Figs 4).

The number of 35S rDNA sites was variable between species, although typically they were located on the short arms of the acrocentric chromosomes, and co-localized with CMA^+ bands. In general, three or four sites were observed per species. However, a divergent number of 35S sites were observed in *E. calycinus* (Fig. 2e) and *T. spinosa* (Fig. 2m) with five and two sites, respectively. Interestingly, for *E. angulatus* only three of the seven CMA^+ sites overlapped with the 35S rDNA sites. Idiograms showing all the cytogenetic markers obtained for the Caesalpinia Group species are illustrated in Figs. 4.

Evolution of the heterochromatin and its correlation with environmental variables and genome size

Climatic niches of all analysed species were compared using principal component analysis (PCA) on a full set of 19 WorldClim variables (Fig. 3a). This PCA revealed that the distribution and climatic niches of *Tara*/*Coulteria* were ecologically distinct from the other two groups of CMA^0 species. Similarly, CMA^+ species also clustered into a group that was distinct from the two groups of CMA^0 species, demonstrating that each group of species with a particular type of heterochromatin has its own climatic niche (Fig. 3a).

While a relationships between CMA/DAPI banding and ecological niche/geographical distribution (Van-Lume et al. 2017), and the correlation between genome size and environment (latitude/ temperature) (Souza et al. 2019) was previously found for the Caesalpinia Group, our increased sampling of species now allows the investigation of how genome size is correlated with heterochromatin. The PICs analysis showed a non-significant relationship between genome size and the heterochromatin amount (R = 0.06, p = 0.629 and DF = 24) in species of the Caesalpinia Group. In contrast, when analysing the two major clades separately, a negative (but non-significant) correlation was found in clade I (R = -1.098, p = 0.31 and DF=7) while a positive (but non-significant) correlation was observed in clade II (R = 0.83, p = 0.41 and DF=15) (Online Resource 2). In both cases, the high p values may be a result of low sampling, as indicated by the low DF values.

We also evaluated the relationship between genome size, latitude and CMA/DAPI intensity in the heterochromatin by combining the CMA/DAPI fluorochrome intensity (CDI) index values generated in this work (see Methods and Materials, Online Resource 1) with those from Van-Lume et al. (2017), and presenting the results as a 3D graph (Fig. 3b). Overall, there was a negative relationship between the CDI index and genome size (R= -2.81, p < 0.001 and DF=26) and between the CDI and genome size and latitude (R= -1.54, p < 0.01 and DF=26) (Fig. 3b). Thus, species with smaller genome sizes and distributed at lower latitudes were shown to be associated with positive CMA bands and hence a higher GC content in the heterochromatin. In contrast, species with larger genome sizes which are typically distributed at mid latitudes were associated with more neutral CMA banding patterns and hence lower GC content, with the CDI values close or equal to one. The previously reported correlation between genome size and environmental conditions (latitude and temperature) was thus maintained with the increase in species number analysed here (R= 20.04, p < 0.0001 and DF=26).

Discussion

New sampling confirms three consistent patterns of heterochromatin and cytogenetic markers among Caesalpinia Group species

The phylogenetically-broader sampling combining our new with already available data (Van-Lume et al. 2017) represent 13 of the 27 genera recognized in the group (Gagnon et al. 2016), increased the number of species by 85%, confirming the existence of three patterns of

heterochromatin distribution ($CMA^0/DAPI^0$, $CMA^0/DAPI^-$ and $CMA^+/DAPI^-$) in the Caesalpinia Group (see Van-Lume et al. 2017). This larger sampling allowed us to demonstrate that CMA^0 bands have arisen several times independently (homoplasy) in clade II (*Arquita*, *Balsamocarpum*, *Erythrostemon* and *Pomaria*) and clade I (*Coulteria* and *Tara*) (Fig. 4). The occurrence of CMA^0 bands in species which occupy similar habitats but are in phylogenetically distinct subclades reinforces the previously identified trend showing that similarity in ecological conditions can be associated with similar karyological pattern (see Van-Lume et al. 2017). In the Caesalpinia group, the species with CMA^0 bands are distributed in higher latitudes and to have the largest genomes (Souza et al. 2019).

Phylogenetic interpretation of the cytogenetic markers revealed few putative synapomorphies, e.g. variability in the clade comprising *Biancaea* + *Mezoneuron* + *Guilandina* + *Caesalpinia sensu strictu*, with each genus is characterized by different patterns with or without pericentromeric CMA^+ bands (Fig. 4). However, most analyzed characters (i.e. type of heterochromatin, number of rDNA sites, etc.) were shown to be homoplastic. Such homoplasy makes interpretation of karyotypic markers difficult in species and genera of the Caesalpinia group (Van-Lume et al. 2017). Despite this, we did observe a common 5S and 35S rDNA synteny within Clade I, in contrast with the variation in number of sites and chromosomal positions observed in clade II (except perhaps in *Cenostigma*), suggesting that *Cenostigma* species are more closely related to clade I than to the rest of the clade II. Similar results were observed in the genus *Aristolochia* (Aristolochiaceae) (Berjano et al. 2009).

Cytofluorometric analyses of the intensity of heterochromatic bands has been used as a valid method for estimating the *in situ* base composition of DNA (Schweizer 1976; Leemann and Ruch 1982). In the Caesalpinia Group, most terminal CMA^+ bands on acrocentric chromosomes coincided with a 35S rDNA hybridization site, as previously observed in related species of this group (Van-Lume et al. 2017; Rodrigues et al. 2018) and other unrelated genera (Berjano et al. 2009; Gaeta et al. 2010). The other CMA^+ bands, observed in the pericentromeric region of sub/metacentric chromosomes, correspond to other guanine-cytosine (GC)-rich pericentromeric sequences, such as satellite DNA and LTR-retrotransposons (Van-Lume et al. 2019). In the Caesalpinia Group, the CMA^+ pericentromeric heterochromatin was previously shown to be evolving dynamically and non-randomly by undergoing expansions and contractions in association with biogeographical and ecological conditions (Van-Lume et al. 2017). Such rapid evolution of repetitive sequences,

which typically occupy the heterochromatin, can lead to changes in composition and/or abundances of a particular sequence (especially retrotransposons and satellite repeats) even between related species (Alkhimova et al. 2004; Salim and Gerton 2019; Van-Lume et al. 2019). The recent study of the identity and abundance of LTR Ty3/*Gypsy*-Tekay, Athila retrotransposons and distinct satellite DNAs in the pericentromeric heterochromatin of some Caesalpinia Group species support the observation that rapid repeat evolution is related to the cytological variability observed in heterochromatin (Van-Lume et al. 2019). Based on this, we suggest that the presence of CMA^0 bands in unrelated species (i.e. as observed in *Coulteria* + *Tara* [clade I] and *Arquita* + *Balsamocarpon* + *Erythrostemon* + *Pomaria* [clade II]) reflects convergence caused by the ecological and environmental similarity of the habitats that these species occupy (Fig. 3a).

Despite the heterochromatic variability observed across the Caesalpinia group, the genera *Cenostigma* and *Libidibia* (both with CMA^+ bands) and the *Coulteria* + *Tara* subclade (with CMA^0 bands) are noteworthy by their stable CMA/DAPI banding pattern. Biogeographical data appear to correlate with these results, since each stable karyotype group tends to share a similar biome distribution in contrast to the species belonging to genera that are more polymorphic in their heterochromatin bands (e.g. *Erythrostemon*). We suggest that these heterochromatically-stable genera arise because of their distribution within the Succulent Biome, which presents a strong phylogenetic conservatism despite its fragmented distribution (Gagnon et al. 2019). In contrast, the rest of the species analysed, which are distributed in two or more biomes (Gagnon et al. 2019), show either polymorphic or no heterochromatic bands. It is also possible that the heterochromatic stability in *Cenostigma*, *Libidibia* and *Coulteria*+*Tara* is related to their relatively recent divergence time (~ 17 Mya) compared with the Caesalpinia Group as a whole which arose in the early Eocene, 54.78 Mya (Fig. 4, Gagnon et al. 2019).

Different patterns of CMA/DAPI staining indicates complex evolution of the heterochromatin in the Caesalpinia group

In the Caesalpinia group, genome size evolution and diversity is predicted to be driven, in part, by the contrasting impact of high-temperature stress in species occupying the Succulent Biome compared with species occupying more Temperate zones (Souza et al. 2019). Recently, Gagnon et al. (2019) clearly demonstrated that the distribution of Caesalpinia group species show a strong relationship between biomes and growth forms: species in the

Succulent Biome are almost all trees or shrubs while in temperate regions they are suffrutices. In addition to a correlation with climatic variables/environment, we have shown that the CDI patterns can also be driven by plant life cycle/morphological traits (Gagnon et al. 2019). In our study, comparative analyses showed that the smallest genomes and highest CDI values are found in tree species, such as *Paubrasilia echinata* (2C = 1.15 pg and CDI=1.11), while the largest genomes and mid value CDI indices were found in suffrutices, e.g. *Pomaria lactea* (2C = 7.11 pg and CDI=0.98). Thus, the differences in genome size and CDI patterns observed in the Caesalpinia Group may be driven partly by the impact of genome size at the nuclear and cellular level which, in turn, can have a knock-on effect at the whole plant level to influence the parameters that various adaptive traits can adopt (Cavalier-Smith 2005; Souza et al. 2019) and hence the environment that a plant can occupy (Pellicer et al. 2018).

Nevertheless, this analysis revealed that the conspicuous polymorphism in amount of heterochromatin (size of CMA bands) in Caesalpinia group species did not correlate with genome size variation. Our hypothesis is that repeats detected here using CMA/DAPI fluorochrome banding (i.e. CMA^+ bands) comprise only part of the total repetitive fraction of the genome and do not play the major role in determining differences in genome size between the Caesalpinia group species. The literature indicates that retrotransposons are often dispersed across plant chromosomes, rather than being localized in discrete clusters as observed for satellite DNAs (e.g. Heslop-Harrison and Schwarzacher 2011; Ribeiro et al. 2017; Van-Lume et al. 2019). In Caesalpinia group species, for example, in *Cenostigma microphyllum*, the heterochromatin is comprised of a dispersed repetitive sequence as a conserved unit (LTR transposable elements) and few tandem sequences (satellite DNAs) (Van-Lume et al. 2019). Certainly, such observations agree with the majority of studies which have explored how repetitive DNA contributes to genome size variation. Such studies have shown than it is the dispersed transposable elements (particularly retrotransposons) which typically play the primary role in genome size variation between species, at least in species with small genomes (e.g. Macas et al. 2015).

We therefore suggest that the repetitive elements present in the heterochromatin (CMA^+ bands) do not represent a large fraction of the genome and hence do not contribute significantly to the genome size variation reported. Instead, the highlight how the repeats present in the heterochromatin may be responding to environmental factors. Subsequent comparative cytogenomic analysis by next generation sequencing (NGS) will clarify the

nature, relative abundance and diversity of these different kinds of heterochromatin (see Van-Lume et al. 2019) and help characterize how their evolution and interaction is influenced by environmental variables.

Acknowledgements

The authors wish to thank the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE - APQ- 0970- 2.03/15) for financial support and a post-doc grant to G.S. by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES - Estágio Pós-Doutoral 88881.119479/2016-01). This study was partly financed by the CAPES (finance Code 001). G.S. receives a productivity fellowship from CNPq (process number PQ - 310693/2018-7).

Authors' contributions

YMS conducted the experiments and wrote the manuscript. LC contributed in bioinformatics analyses and manuscript revisions. EG, GPL and IJL provided the plant material, contributed with critical revisions, manuscript revisions and with the correction of the English. GS conceived and designed the research, contributed with critical discussions and manuscript revisions.

Conflict of interest

The authors declare that they have no competing interests.

Table 1 List of analysed species describing their respective karyotype formula (KF), heterochromatin proportion (%H), CMA/DAPI pattern, number of 5S and 35S rDNA sites, and the type (s) of chromosome they occur on, as SM: submetacentric, AC: acrocentric, and genome size 2C in picograms (pg) together with coefficients of variation (CV).

Fig. 1 Heterochromatin banding patterns identified using CMA/DAPI fluorochrome staining. The three distinctive patterns of heterochromatin staining comprise: (i) $CMA^0/DAPI^-$ (a-k), (ii) $CMA^+/DAPI^-$ (l-m) and $CMA^0/DAPI^0$ pattern (n). CMA is shown in yellow and DAPI in blue. *Coulteria mollis* (a); *C. pumila* (b); *Erythrostemon angulatus* (c); *E. pannosus* (d); *E. placidus* (e); *Tara cacalaco* (f); *E. calycinus* (g); *Pomaria lactea* (h); *Tara vesicaria* (i); *T.*

spinosa (j); *Coulteria platyloba* (k); *Libidibia coriaria* (l); *L. punctata* (m) and *Mezoneuron hildebrandtii* (n). Scale bar = 5 μm

Fig. 2 *In situ* hybridization showing the number and location of 5S (red) and 35S (green) rDNA sites in *Coulteria mollis* (a); *C. pumila* (b); *C. platyloba* (c); *Erythrostemon angulatus* (d); *E. calycinus* (e); *E. placidus* (f); *E. pannosus* (g); *Libidibia coriaria* (h); *L. punctata* (i); *Mezoneuron hildebrandtii* (j); *Pomaria lactea* (k); *Tara cacalaco* (l) and *T. spinosa* (m). Arrows show 5S rDNA sites. Arrowheads indicate small 35S rDNA sites and empty arrows indicate CMA^+ sites. Scale bar = 5 μm

Fig. 3 Relationships between heterochromatin and environmental variables in the Caesalpinia Group. **a**, geographic distribution of species with pericentromeric CMA^+ bands (red dots), *Tara*/ *Coulteria* species with pericentromeric CMA^0 bands (blue dots), other species with CMA^0 bands (green dots); on the left, a density plot showing the results from an individual-based Principal Component Analysis (PCA) of the three groups described above. **b**, multiple linear regression showing the relationship among genome size (2C), latitude and CDI values

Fig. 4 Comparative haploid idiograms of clade I and II of Caesalpinia Group for taxa characterized in this study (bold font) and in Van-Lume et al. (2017) (gray). The idiograms show the relative chromosomes size, centromeric position, CMA^+ bands (yellow), CMA/DAPI patterns, and 5S (red) and 35S (green) rDNA site. Phylogenetic relationships are based in Gagnon et al. (2016). Bold branches represent support equal to 1. Scale bar = 5 μm

Online Resource 1 List of analysed species describing their chromosome pairs number (CP), chromosome length (CL), standard deviation (SD), the ratio of chromosome arms (AR) and the CMA/DAPI intensity index (CDI index)

Online Resource 2 Scatterplots of independent contrasts between genome size and heterochromatin proportion in clade I (a), clade II (b) and the Caesalpinia group (c). Phylogeny includes species characterized in this study (black) and those obtained from Van-Lume et al. (2017) (gray). Contrast values for genome size and heterochromatin proportion were plotted above and below branches, respectively. The two characters were statistically non-correlated

References

Acosta MC, Moscone EA, Cocucci AA (2016) Using chromosomal data in the phylogenetic and molecular dating framework: karyotype evolution and diversification in *Nierembergia* (Solanaceae) influenced by historical changes in sea level. Plant Biol 18: 514-526. https://doi.org/10.1111/plb.12430

Alkhimova OG, Mazurok NA, Potapova TA, Zakian SM, Heslop-Harrison JS, Vershinin AV (2004) Diverse patterns of the tandem repeats organization in rye chromosomes. Chromosoma 113(1): 42-52. https://doi.org/10.1007/s00412-004-0294-4

Ambrožová K, Mandáková T, Bureš P, Neumann P, Leitch IJ, Koblížková A, Lysak MA (2011) Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. Ann Bot 107: 255-268. https://doi.org/10.1093/aob/mcq235

Beltrão GTA, Guerra M (1990) Citogenética de angiospermas coletadas em Pernambuco-III. Ciência e Cultura 42(10).

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Ann Rev Plant Biol 65: 505-530. https://doi.org/10.1146/annurev-arplant-050213-035811

Berjano R, Roa F, Talavera S, Guerra M (2009) Cytotaxonomy of diploid and polyploidy (Aristolochiaceae) species based on the distribution of CMA/DAPI bands and 5S and 45S rDNA sites. Plant Syst Evol 280(3-4): 219-227. https://doi.org/10.1007/s00606-009-0184-6

Bilinski P, Albert PS, Berg JJ, Birchler J, Grote M, Lorant A, Quezada J, Swarts K, Yang J, Ross-Ibarra J (2017) Parallel altitudinal clines reveal adaptive evolution of genome size in *Zea mays*. PLOS Genetics 14(5): e1007162. https://doi.org/10.1101/134528

Borges LA, Souza LGR, Guerra M, Machado IC, Lewis GP, Lopes AV (2012) Reproductive isolation between diploid and tetraploid cytotypes of *Libidibia ferrea* (= *Caesalpinia ferrea*) (Leguminosae): ecological and taxonomic implications. Plant Syst Evol 298(7): 1371-1381. https://doi.org/10.1007/s00606-012-0643-3

Castañeda R, Gutiérrez H, Carrillo É, Sotelo A (2017) Leguminosas (Fabaceae) silvestres de uso medicinal del distrito de Lircay, provincia de Angaraes (Huancavelica, Perú). B Latinoam Caribe Pl 16(2): 136-149.

Cavalier-Smith T (2005) Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot 95(1): 147-175. https://doi.org/10.1093/aob/mci010

Chiarini F, Sazatornil F, Bernardello G (2018) Data reassessment in a phylogenetic context gives insight into chromosome evolution in the giant genus *Solanum* (Solanaceae). Syst Biodivers 16(4): 397-416. https://doi.org/10.1080/14772000.2018.1431320

Costa L, Oliveira Á, Carvalho-Sobrinho J, Souza G (2017) Comparative cytomolecular analyses reveal karyotype variability related to biogeographic and species richness patterns in Bombacoideae (Malvaceae). Plant Syst Evol 303(9): 1131-1144. https://doi.org/10.1007/s00606-017-1427-6

Dryflor et al. (2016) Plant diversity patterns in neotropical dry forests and their conservation implications. Science 6306(353): 1383-1387. https://doi.org:10.1126/science.aaf5080

Felsenstein J (1985) Phylogenies and the comparative method. Amer Nat 125(1): 1-15. https://www.jstor.org/stable/2461605

Gaeta ML, Yuyama PM, Sartori D, Fungaro MHP, Vanzela ALL (2010) Occurrence and chromosome distribution of retroelements and NUPT sequences in *Copaifera langsdorffii* Desf. (Caesalpinioideae). Chromosome Res 18(4): 515-524. https://doi.org/10.1007/s10577-010-9131-1

Gagnon E, Bruneau A, Hughes CE, de queiroz L, Lewis GP (2016) A new generic system for the pantropical Caesalpinia group (Leguminosae). PhytoKeys 71(1): 1-160. https://doi.org/10.3897/phytokeys.71.9203

Gagnon E, Ringelberg JJ, Bruneau A, Lewis GP, Hughes CE (2019) Global Succulent Biome phylogenetic conservatism across the pantropical Caesalpinia Group (Leguminosae). New Phytologist, 222(4): 1994-2008.https://doi.org/10.1111/nph.15633

Gagnon E, Lewis GP, Sotuyo JS, Hughes CE, Bruneau A (2013) A molecular phylogeny of *Caesalpinia sensu lato*: increased sampling reveals new insights and more genera than expected. S Afr J Bot 89: 111–127. https://doi.org/10.1016/j.sajb.2013.07.027

Gagnon E, Hughes CE, Lewis GP, Bruneau A (2015) A new cryptic species in a new cryptic genus in the Caesalpinia group (Leguminosae) from the seasonally dry inter Andean valleys of South America. Taxon 64: 468–490. https://doi.org/10.12705/643.6

Gasson P, Warner K, Lewis G (2009) Wood anatomy of *Caesalpinia* ss, *Coulteria*, *Erythrostemon*, *Guilandina*, *Libidibia*, *Mezoneuron*, *Poincianella*, *Pomaria* and *Tara* (Leguminosae, Caesalpinioideae, Caesalpinieae). IAWA J 30(3): 247-276. https://doi.org/10.1163/22941932-90000218

Guerra MS (1986) Reviewing the chromosome nomenclature of Levan. Brazil J Genet 9: 741-743.

Guerra M (2000) Patterns of heterochromatin distribution in plant chromosomes. Genet Mol Biol 23(4): 1029-1041. http://dx.doi.org/10.1590/S1415-47572000000400049

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Vol. 239. Oxford: Oxford University Press.

Kassambara A, Mundt F (2017) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Thierer T (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kirov I, Khrustaleva L, Van Laere K, Soloviev A, Meeus S, Romanov D, Fesenko I (2017) DRAWID: user-friendly java software for chromosome measurements and idiogram drawing. Comp Cytogenet 11(4): 747-757. https://doi.org/10.3897/CompCytogen.v11i4.20830

Kolano B, McCann J, Orzechowska M, Siwinska D, Temsch E, Weiss-Schneeweiss H (2016) Molecular and cytogenetic evidence for an allotetraploid origin of *Chenopodium quinoa* and *C. berlandieri* (Amaranthaceae). Mol Phylogenet Evol 100: 109-123. https://doi.org/10.1016/j.ympev.2016.04.009

Lee YI, Yap JW, Izan S, Leitch IJ, Fay MF, Lee YC, Leitch AR (2018) Satellite DNA in *Paphiopedilum* subgenus *Parvisepalum* as revealed by high-throughput sequencing and fluorescent in situ hybridization. BMC Genomics 19(1): 578. https://doi.org/10.1186/s12864-018-4956-7

Leemann U, Ruch F (1982) Cytofluorometric determination of DNA base content in plant nuclei and chromosomes by the fluorochromes DAPI and Cromomicyn A3. Exper Cell Res 140: 275-282. https://doi.org/10.1016/0014-4827(82)90115-X

Macas J, Novák P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fuková I, Doležel J, Kelly LJ, Leitch IJ (2015) In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PLoS One 10(11): e0143424. https://doi.org/10.1371/journal.pone.0143424

Manzanilla V, Bruneau A (2012) Phylogeny reconstruction in the Caesalpinieae grade (Leguminosae) based on duplicated copies of the sucrose synthase gene and plastid markers. Mol Phylogenet Evol 65(1): 149-162. https://doi.org/10.1016/j.ympev.2012.05.035

Menezes RS, Brady SG, Carvalho AF, Del Lama MA, Costa MA (2017) The roles of barriers, refugia, and chromosomal clines underlying diversification in Atlantic Forest social wasps. Sci Rep 7: 7689. https://doi.org/10.1038/s41598-017-07776-7

Moreno N, Amarilla L, Las Peñas M, Bernardello G (2015) Molecular cytogenetic insights into the evolution of the epiphytic genus *Lepismium* (Cactaceae) and related genera. Bot J Linn Soc 177: 263-277. https://doi.org/10.1111/boj.12242

Moreno NC, Stiefkens L, Las Peñas ML, Bartoli A, Tortosa R, Bernardello G (2018) Karyotypes, heterochromatin distribution and rDNA patterns in South American Grindelia (Asteraceae). Plant Biosys 152(1): 166-174. https://doi.org/10.1080/11263504.2016.1265611

Paradis E, Bolker B, Claude J (2012) Package ape. Analyses of phylogenetics and evolution. R package version 2012. 04-04. http://cran.r-project.org/web/packages/ape/ape.pdf

Pellicer J, Hidalgo O, Dodsworth S, Leitch I (2018) Genome size diversity and its impact on the evolution of land plants. Genes 9(2): 88. https://doi.org/10.3390/genes9020088

Plohl M, Luchetti A, Meštrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero) chromatin. Gene 409(1-2): 72-82. https://doi.org/10.1016/j.gene.2007.11.013

Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol. 25:1253–1256. https://doi.org/10.1093/molbev/msn083

Purvis A, Rambaut A (1995) Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Bioinformatics 11(3): 247-251.

Puttick MN, Clark J, Donoghue PC (2015) Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms. Proc R Soc Lond [Biol] 282(1820): 20152289. https://doi.org/10.1098/rspb.2015.2289

R Core Team (2011) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www. R-project. org.

Rambaut A (2009) FigTree v1. 3.1: Tree Figure Drawing Tool. http://tree.bio.ed.ac.uk/software/figtree.

Ribeiro T, Buddenhagen CE, Thomas WW, Souza G, Pedrosa-Harand A (2018) Are holocentrics doomed to change? Limited chromosome number variation in *Rhynchospora* Vahl (Cyperaceae). Protoplasma 255(1): 263-272. https://doi.org/10.1007/s00709-017-1154-4

Ribeiro T, Marques A, Novák P, Schubert V, Vanzela AL, Macas J, Pedrosa-Harand A (2017) Centromeric and non-centromeric satellite DNA organisation differs in holocentric *Rhynchospora* species. Chromosoma 126(2): 325-335. https://doi.org/10.1007/s00412-016-0616-3

Rodrigues PS, Souza MM, Corrêa RX (2014) Karyomorphology and karyotype asymmetry in the South American *Caesalpinia* species (Leguminosae:Caesalpinioideae). Genet Mol Biol 13: 8278-93. http://dx.doi.org/10.4238/2014.October.20.4

Rodrigues PS, Souza MM, Melo CAF, Pereira TNS, Corrêa RX (2018) Karyotype diversity and 2C DNA content in species of the Caesalpinia group. BMC genetics 19(1): 25. https://doi.org/10.1186/s12863-018-0610-2

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Ruban A, Fuchs J, Marques A, Schubert V, Soloviev A, Raskina O, Houben A (2014) B chromosomes of *Aegilops speltoides* are enriched in organelle genome-derived sequences. PLoS One 9(2): e90214. https://doi.org/10.1371/journal.pone.0090214

Sader MA, Amorim BS, Costa L, Souza G, Pedrosa-Harand A (2019) The role of chromosome changes in the diversification of *Passiflora* L. (Passifloraceae). Syst Biodivers 1:1-15. https://doi.org/10.1080/14772000.2018.1546777

Salim D, Gerton JL (2019) Ribosomal DNA instability and genome adaptability. Chromosome Res 1: 1-15. https://doi.org/10.1007/s10577-018-9599-7

Schwarzacher T, Schweizer D (1982) Karyotype analysis and heterochromatin differentiation with Giemsa C-banding and fluorescent counterstaining in *Cephalanthera* (Orchidaceae). Plant Syst Evol 141: 91-113.

Schweizer D (1976) Reverse fluorescent chromosome banding with chromomycin and DAPI. Chromosoma 58: 307-324.

Seijo JG, Lavia GI, Fernández A, Krapovickas A, Ducasse D, Moscone EA (2004) Physical mapping of the 5S and 18S–25S rRNA genes by FISH as evidence that *Arachis duranensis* and *A. ipaensis* are the wild diploid progenitors of *A. hypogaea* (Leguminosae). Amer J Bot 91(9): 1294-1303. https://doi.org/10.3732/ajb.91.9.1294

Siljak-Yakovlev S, Godelle B, Zoldos V, Vallès J, Garnatje T, Hidalgo O (2017) Evolutionary implications of heterochromatin and rDNA in chromosome number and genome size changes during dysploidy: A case study in *Reichardia* genus. PloS one 12(8): e0182318. https://doi.org/10.1371/journal.pone.0182318

Silvestri MC, Ortiz AM, Lavia GI (2015) rDNA loci and heterochromatin positions support a distinct genome type for 'x= 9 species' of section *Arachis* (*Arachis*, Leguminosae). Plant Syst Evol 301(2): 555-562. https://doi.org/10.1007/s00606-014-1092-y

Soetaert K (2014) plot3D: Tools for plotting 3-D and 2-D data. R package version, 10-2.

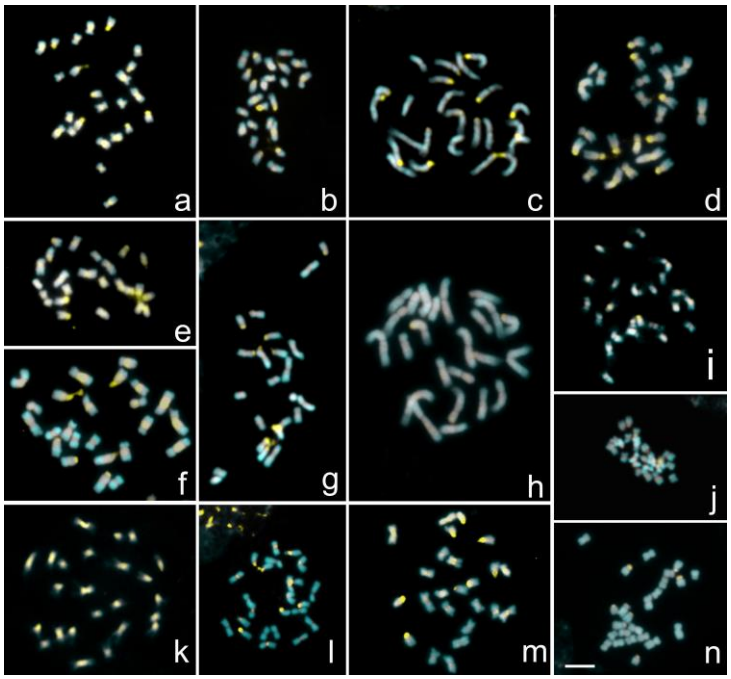
Souza G, Costa L, Guignard MS, Van-Lume B, Pellicer J, Gagnon E, Lewis GP (2019) Do tropical plants have smaller genomes? Correlation between genome size and climatic variables in the Caesalpinia Group (Caesalpinioideae, Leguminosae). Persp Plant Ecol 38:13-23. https://doi.org/10.1016/j.ppees.2019.03.002

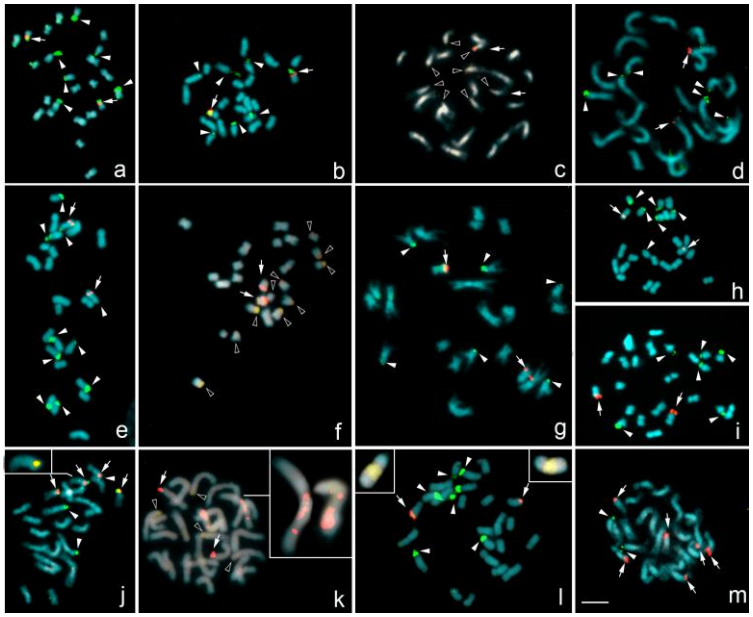
Vaio M, Nascimento J, Mendes S, Ibiapino A, Felix LP, Gardner A, Emshwiller E, Fiaschi P, Guerra M (2018). Multiple karyotype changes distinguish two closely related species of *Oxalis* (*O. psoraleoides* and *O. rhombeo-ovata*) and suggest an artificial grouping of section Polymorphae (Oxalidaceae). Bot J Linn Soc 188(3): 269-280. https://doi.org/10.1093/botlinnean/boy054

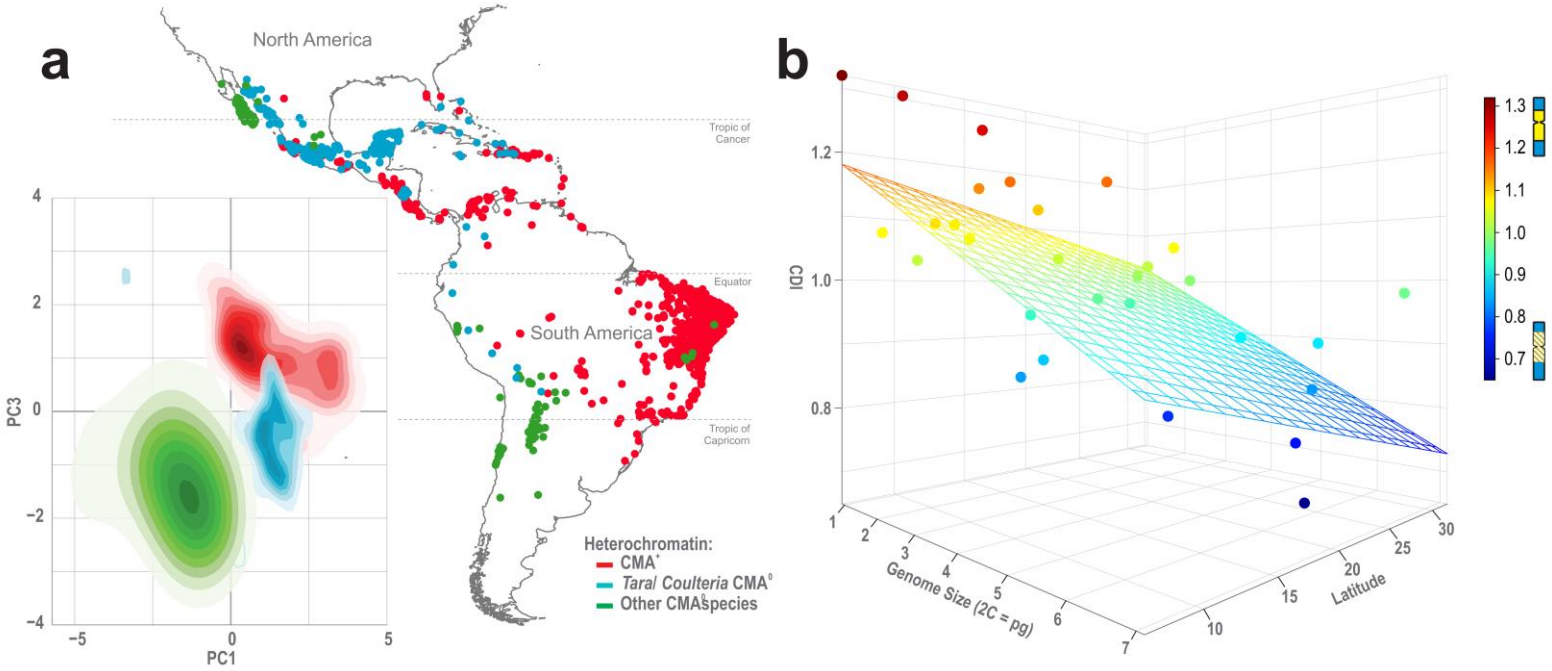
Van-Lume B, Esposito T, Diniz-filho J, Gagnon E, Lewis G, Souza G (2017) Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. Persp Plant Ecol 29: 51-63. 2017. https://doi.org/10.1016/j.ppees.2017.11.004

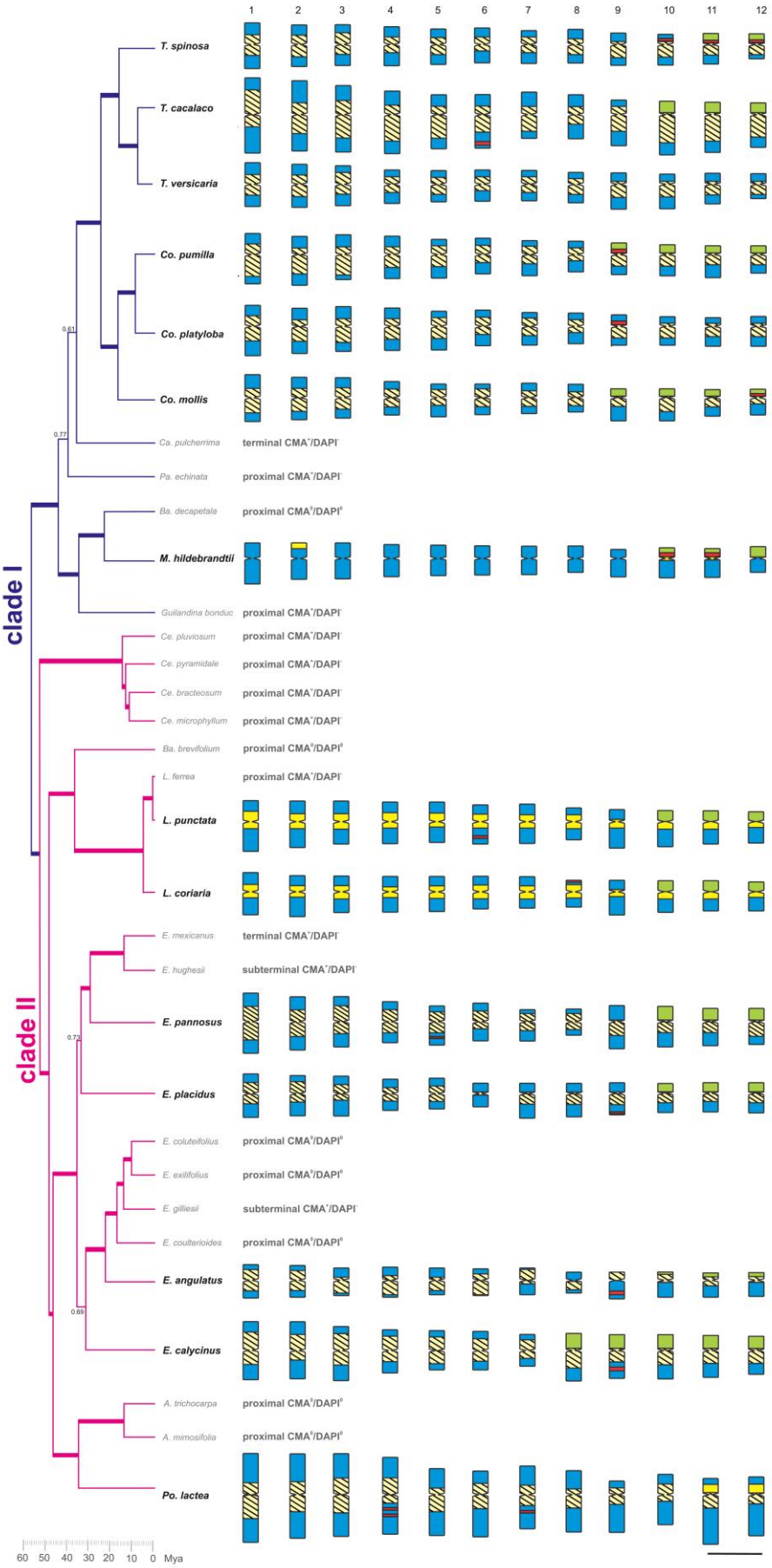
Van-Lume B, Souza G (2018) Cytomolecular analysis of species in the *Peltophorum* clade (Caesalpinioideae, Leguminosae). Brazilian J Bot 41: 385-392. https://doi.org/10.1007/s40415-018-0449-9

Van-Lume B, Mata-Sucre Y, Báez M, Ribeiro T, Huettel B, Gagnon E, Leitch IJ, Pedrosa-Harand A, Lewis GP, Souza G (2019) Evolutionary convergence or homology? Comparative cytogenomics of Caesalpinia group species (Leguminosae) reveals diversification in the pericentromeric heterochromatic composition. Planta 250: 2173-2186. https://doi.org/10.1007/s00425-019-03287-z


a
b
c
d
e
f
g
h
i
j
k
l
m
n


a
b
c
d
e
f
g
h
i
j
k
l
m


a
North America
South America
Tropic of Cancer
Equator
Tropic of Capricorn
Heterochromatin:
CMA+
Taraf Coulteria CMA0
Other CMA0 species
PC3
PC1
b
CCI
Genome Size (2C = pg)
Latitude


1
2
3
4
5
6
7
8
9
10
11
12
T. spinosa
T. cacalaco
T. versicaria
Co. pumilla
Co. platyloba
Co. mollis
Ca. pulcherrima
terminal CMA+/DAPI-
Pa. echinata
proximal CMA+/DAPI-
Ba. decapetala
proximal CMA0/DAPI0
M. hildebrandtii
Guilandina bonduc
proximal CMA+/DAPI-
Ce. pluviosum
proximal CMA+/DAPI-
Ce. pyramidale
proximal CMA+/DAPI-
Ce. bracteosum
proximal CMA+/DAPI-
Ce. microphyllum
proximal CMA+/DAPI-
Ba. brevifolium
proximal CMA0/DAPI0
L. ferrea
proximal CMA+/DAPI-
L. punctata
L. coriaria
E. mexicanus
terminal CMA+/DAPI-
E. hughesii
subterminal CMA+/DAPI-
E. pannosus
E. placidus
E. coluteifolius
proximal CMA0/DAPI0
E. exilifolius
proximal CMA0/DAPI0
E. gilliesii
subterminal CMA+/DAPI-
E. coulterioides
proximal CMA0/DAPI0
E. angulatus
E. calycinus
A. trichocarpa
proximal CMA0/DAPI0
A. mimosifolia
proximal CMA0/DAPI0
Po. lactea
clade I
clade II
0.61
0.77
0.73
0.69
60 50 40 30 20 10 0 Mya

3.2 ARTIGO 2- WHAT IS THE RELATIONSHIP BETWEEN HETEROCHROMATIN COMPOSITION AND SPECIES-RICHNESS? CYTOGENOMICS OF *Erythrostemon hughesii* GAGNON & G.P.LEWIS (CAESALPINIOIDEAE)

Yennifer C Mata-Sucre[1], Mariela Sader[1], Brena Van-Lume[1], Andrea Pedrosa-Harand[1]
Gwilym P. Lewis [2] and Gustavo Souza[1]

Artigo submetido à revista **Planta (Qualis A1)**

What is the relationship between heterochromatin composition and species-richness? Cytogenomics of *Erythrostemon hughesii* Gagnon & G.P.Lewis (Caesalpinioideae)

Yennifer C Mata-Sucre[1], Mariela Sader[1], Brena Van-Lume[1], Andrea Pedrosa-Harand[1] Gwilym P. Lewis [2] and Gustavo Souza[1]

[1] Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Rua Nelson Chaves S/N, Cidade Universitaria, Recife, PE. 50670-420, Brazil.

[2] Comparative Plant and Fungal Biology Department, Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK

Correspondence author:

lgrsouza@hotmail.com

Tel: + 55 81 2126 8846

Fax: + 55 81 2126 8348

Main Conclusion: Cytogenomic characterization of *Erythrostemon hughesii* reveals a heterogeneity of repeats in its subtelomeric heterochromatin. Comparative analyses suggest that the species-richness in the Caesalpinia group could be related to heterochromatin heterogeneity.

Abstract

In numerically stables karyotype, repetitive DNA variability is one of the main causes of genome and chromosome evolution. Species from the Caesalpinia Group are karyotypically characterized by present $2n = 24$ with small chromosomes and highly variable CMA^+ heterochromatin patterns, correlated with environmental variables. *Erythrostemon hughesii* differ from other species of the group, because of the presence

of subtelomeric CMA^+ bands, while most species in the group have pericentromeric bands. Here we perform a repeatome characterization of *E. hughesii* genome based on genome skimming and chromosomal distribution of the most abundant elements. The repetitive fraction of the *E. hughesii* represents 28.73% of the genome. The most abundant elements were satellite DNAs (7.83%) and LTR-RT *Ty1*/*copia*-Ale (1%) and *Ty3*/*gypsy* CRM (0.88%) and Athila (0.75%). The hybridization signals of four satDNAs and four LTR elements were present in most CMA^+ subtelomeric bands. The repeatome in *E. hughesii* was distinct from Northeast Brazilian species, mainly by their high heterochromatic diversity and low amount of *Ty3*/*gypsy*-Tekay. Sequence homology of Tekay element is congruent with a clade-specific origin of this repeats after the Caesalpinia Group divergence. The strong reduction in the Tekay elements abundance in *E. hughesii* genome correlated with the loss of proximal CMA^+ heterochromatin and the invasion of several other repeats in the chromosomes subtelomeric region. Repeatome variability suggest that the species-richness in Caesalpinia Group could be related to heterochromatin heterogeneity as a factor of genomic divergence, which is the major prerequisite for speciation process.

Keywords: cytogenetics, genome evolution, mobile elements, repetitive sequences, satellite DNA.

Abbreviations

CMA: chromomycin A3
DANTE: Domain-based Annotation of Transposable Elements
DAPI: 4′, 6-diamidino-2-phenylindole
FISH: Fluorescent in situ hybridization
LTR: Long terminal repeat
NGS: Next generation sequencing
rDNA: ribosome DNA
RT: retrotransposons
satDNA: satellite DNA
TAREAN: tandem repeat analyzer
TE: transposable elements

Introduction

Repetitive DNA represents a substantial fraction of many plant genomes and is frequently referred to as the ‘repeatome’ (Goubert et al. 2015; Maumus and Quesneville 2016; Jouffroy et al. 2016; Hannan 2018; Pita et al. 2019). Changes in repeat abundance and distribution can lead to genetic changes and genome evolution (Feschotte et al. 2002; Lisch 2013; Bennetzen and Wan 2014; Van't Hof et al. 2016). The repeatome comprises a diversity of different DNA sequences that can broadly be divided into two main groups comprising: (i) tandem repeats and (ii) dispersed repeats. These can be further divided into numerous families of repetitive sequences (Ma et al. 2007; Richard et al. 2008). The tandem repeat fraction comprises tandem genes such as ribosome DNA (rDNA) and satellite DNA (satDNA) (Richard et al. 2008, Garrido-Ramos 2015). The dispersed repeats which comprise the other major group of repetitive sequences are largely composed of transposable elements (TE) which can change their position within the genome by self-encoded proteins, causing a diversity of effects such as mutations, insertions, deletions and structural rearrangements that can influence genetic regulation (Kazazian 2004) and genome size (Macas et al. 2015; Hloušková et al. 2019).

Despite the diversity of TEs that have been identified, they can broadly be divided into two major classes according to their mode of transposition within the genome: (i) retrotransposons (RT, class I elements) and (ii) DNA transposons (class II elements) (Wicker et al. 2007; Bourque et al. 2018). In plant genomes, Long Terminal Repeats (LTR) RTs are a very large and diverse group of TEs comprising more than 75% of the nuclear genome in some species (e.g. Baucom et al. 2009, Wendel et al. 2016). LTR-RTs are typically classified into superfamilies that include *Ty1/copia* (also known as the family *Pseudoviridae* in the ICTV classification of viruses) and *Ty3/gypsy* (*Metaviridae*) (Wicker et al. 2007; Krupovic et al. 2018), these are further divided into a vast number of diverse elements in plants (Neumann et al. 2019). The *Ty3/gypsy* superfamily includes two major lineages, the chromovirus and non-chromovirus sequences, which differ by the presence or absence of the chromodomain protein respectively (Neumann et al. 2019). There is ample evidence demonstrating that the transposition of LTR-RTs is one of the main drivers of genome size increases in plants (e.g. Zhang et al. 2017; Wicker et al. 2018; Baniaga and Barker 2019), resulting in extremely large genomes in some species that tolerate LTR-RT accumulation (e.g. Kelly et al. 2015). Since repetitive DNA plays an important role in genome evolution,

documenting its composition and dynamics is essential to shed light on the origin and evolution of genomic diversity (Garrido-Ramos 2017, Bennetzen and Wang 2014, Wendel et al. 2016).

The development of high through-put sequencing technology has opened up the possibility of unraveling details about the repetitive fraction of different genomes and performing comparative analyses of the entire repeatome of non-model and non-cultivated species (Belyayev et al. 2019; Van-Lume et al. 2019). These advances have been accompanied by the development of novel bioinformatic approaches which enable the repeatome to be characterized. They include the RepeatExplorer pipeline (Novák et al. 2013) which uses a graph-based clustering approach for *de novo* repeat identification, classification and protein domain searches. It has been widely used to explore a diversity of plant repeatomes (e.g. Macas et al. 2015; Li et al. 2019; Liu et al. 2019; Samoluk et al. 2019; Van-Lume et al. 2019). Methods for investigating repetitive sequence structure, variation, and evolution today often combine genomic characterization with a cytogenomic approach to determine the physical location of different repetitive DNA families on the chromosomes using fluorescence *in situ* hybridization (FISH) (e.g. Lee et al. 2018; Samoluk et al. 2019; Van-Lume et al. 2019; Zhang et al. 2019; Hloušková et al. 2019).

An ideal taxonomic model for utilizing these approaches is the Caesalpinia group (Leguminosae: Caesalpinioideae). Cytogenetically, its species have been extensively studied (Beltrão and Guerra 1990; Borges et al. 2012; Rodrigues et al. 2014; 2018; Van-Lume et al. 2017; Mata-Sucre et al. 2020) and have shown a stable diploid karyotype $2n = 2x = 24$. Nevertheless, double staining with CMA (chromomycin A_3) and DAPI (4′, 6-diamidino-2-phenylindole) specific fluorochromes has revealed a large variation in the amount and distribution of heterochromatin, both in proximal and terminal chromosome regions (Van-Lume et al. 2017; Rodrigues et al. 2018; Mata-Sucre et al. 2020). Three proximal heterochromatin patterns were shown to be associated with particular environmental variables associated with the main neotropical centre of diversity: the Andean species with neutral CMA bands and negative DAPI bands ($CMA^0/DAPI^-$), species from Mesoamerica without proximal fluorochrome bands ($CMA^0/DAPI^0$), and species from northeastern Brazil with CMA^+ bands and negative DAPI bands ($CMA^+/DAPI^-$) (Van-Lume et al. 2017). Furthermore, the analysis of additional species of the Caesalpinia group showed an intensity of the CMA^+ bands (Mata-Sucre et al. 2020) and genome sizes (Souza et al. 2019) that were directly correlated with ecological

variables, highlighting that this is an ideal plant group for analysing the role of the environment in genome evolution.

Despite previous extensive cytogenetic studies, only three species with CMA^{+}/$DAPI^{-}$ bands have been analysed to determine the sequence composition and evolution of the repetitive fraction of the genome (Van-Lume et al. 2019). The most abundant type of repetitive element in the genomes of *Cenostigma microphyllum*, *Libidibia ferrea* and *Paubrasilia echinata* was the Tekay element belonging to the Chromovirus lineage of *Ty3/gypsy* LTR-RTs. Through cytogenomic analysis, the CMA^{+} heterochromatin bands were enriched with Tekay elements in all three species. However, the heterochromatin of *C. microphyllum* and *L. ferrea* also contained additional species-specific repetitive sequences (i.e. *Ty3/gypsy*-Athila and satDNAs). The authors suggested that the abundance of these TEs, modulated by ecological factors, could have led to the correlation between ecological niche and the heterochromatic banding patterns previously reported (Van-Lume et al. 2017). Nevertheless, the abundance and distribution of Tekay elements in the genomes of species without proximal heterochromatin, such as *Erythrostemon hughesii*, was not studied (Van-Lume et al. 2017).

Erythrostemon is one of the most diverse genera of the Caesalpinia group with 31 species widespread across low-elevation seasonally dry tropical forests in a disjunct geographical distribution: (i) USA, Mexico, and Central America (22 species), and the Caribbean (one species in Cuba and Hispaniola); (ii) Northeast Brazil (one species), and (iii) Argentina, Bolivia, Chile and Paraguay (seven species) (Gagnon et al. 2016, 2019). A comparative cytogenetic analysis of the Caesalpinia group showed that most of the species in the group had CMA^{0}/$DAPI^{-}$ proximal bands (Van-Lume et al. 2017; Mata-Sucre et al. 2020). However, one species, *E. hughesii* (G.P. Lewis) Gagnon & G.P. Lewis, from Oxaca state in Mexico (Lewis et al. collection number 1795, the type specimen) presented a distinctive CMA/DAPI banding pattern (Van-Lume et al. 2017). This shrub has CMA^{+} heterochromatic bands in the subterminal/terminal regions of almost all chromosomes but no proximal CMA^{0} heterochromatic bands (Van-Lume et al. 2017).

Due to its distinctive CMA^{+} bands, this study aims to identify and characterize repetitive sequences in the *Erythostemon hughesii* genome, using Next-Generation Sequencing (NGS) data, and bioinformatics analysis with RepeatExplorer and chromosome mapping using FISH. Our work poses three questions: (i) What types of

repetitive elements are found in the subtelomeric CMA^+ heterochromatin bands of *E. hughesii*? (ii) What is the abundance of *Ty3*/gypsy Tekay elements in *E. hughesii*?, the most abundant repeat identified in the genome of other Caesalpinia group species analysed to date, and (iii) How has the repetitive fraction of analysed Caesalpinia group genomes fluctuated over time when viewed within a phylogenetic context?

Materials and methods

Plant material, DNA extraction and whole-genome sequencing

Seeds of *Erythrostemon hughesii* (G.P. Lewis) Gagnon & G.P. Lewis were obtained from the Millennium Seed Bank of the Royal Botanic Gardens, Kew (ID K000264581). During the cytogenetic analyses the seedlings were grown and kept in the experimental garden of the Laboratory of Plant Cytogenetics and Evolution (UFPE, Recife-PE, Brazil).

Total genomic DNA was obtained from young leaves using the CTAB extraction protocol of Weising et al. (2005) and sequenced using the Illumina HiSeq 2000 platform in the Max Planck-Genome-Center, generating 1,984,766 reads paired-end reads of 250 bp. Since *Erythrostemon hughesii* has a genome size of 1C = 1.09 pg (1,066 Mbp) (Souza et al. 2019), the 496 Mbp of sequence data obtained is estimated to have covered 0.4x of the genome.

Identification of repetitive sequences

A graph-based clustering analysis in the pipeline RepeatExplorer2 (Novák et al. 2010, 2013) was used to identify and quantify the repetitive sequences in the 1,984,766 reads of low coverage sequencing data. The resulting clusters were classified by similarity searches against the Conserved Domain Database for the annotation of protein domains present in repeats (Marchler-Bauer et al. 2011) which is integrated into RepeatExplorer2. The genome proportion of the repetitive fraction, and for each individual repeat identified in the *E. hughesii* genome present in least 0.01% of the genome, was calculated using the number of clustered reads versus the total amount of reads used for analysis, after excluding chloroplast and mitochondrial reads which represent possible contaminants.

The TAREAN (TAndem REpeat ANalyser) pipeline (also implemented in RepeatExplorer2) was used to identify tandem repeat sequences. This tool also relies on graph-based repeat clustering, and allows the identification and characterization of

satDNAs from unassembled sequence reads (Novák et al. 2017). The BLAST tool was used to characterize, where possible, the unidentified clusters by comparisons against the public database NCBI. To confirm that the satDNA clusters obtained from TAREAN were genuine satellite sequences, the contigs classified as putative satellite sequences were aligned and their monomers were detected using a Dot-plot analysis in JDotter (Brodie et al. 2004). To investigate the degree of homology among each of the characterized satDNAs, we considered that monomeric sequences with 50 - 80% similarity belonged to different families of the same superfamily of satDNAs, sequences while those with more than 80 - 95% similarity were variants of the same family (ie. subfamily), and those showing > 95% similarity were considered to be variants of the same monomer, as proposed by Ruiz-Ruano et al. (2016).

To investigate the similarity between the satDNA EhugSat3_490 and rDNA regions (see Results), we assembled the complete 35S rDNA sequence with NOVOPlasty (Dierckxsens et al. 2016) using a random selection of all 1,984,766 read pairs. A partial ribosomal 18S RNA gene sequence from *Cuscuta campestris* (GenBank accession AY880318.1) was used as the seed sequence for the 'seed-and-extend' algorithm implemented in NOVOPlasty. The annotation of ribosomal regions was performed using Geneious ver. 7.1.9 (http://www.geneious.com, Kearse et al. 2012), then manually verified and corrected by comparison with sequences of related species in GenBank.

Phylogenetic relationships of Gypsy-Tekay elements

To explore the homology and understand the dynamic evolution of the LTR-chromovirus of the superfamily *Ty3/gypsy*-Tekay RT sequences, we analysed these elements in more detail using similarity searches against the *Ty3/gypsy*-Database (Neumann et al. 2019) and comparative approaches using data for three other species of the Caesalpinia group taken from Van-Lume et al. (2019). The reverse transcriptase protein domains of *Ty3*/*gypsy*-Tekay elements from *E. hughesii* were extracted and filtered by quality (alignment sequence identity 0.35, alignment similarity 0.45 and alignment proportion length full-length 0.8) from contigs using the DANTE (Domain-based Annotation of Transposable Elements) tool in the RepeatExplorer2 platform (Novák et al. 2013). This tool annotates and classifies protein domains based on homology comparisons with the database of available Viridiplantae protein domains (Neumann et al. 2019).

For analysis of phylogenetic relationships, the reverse transcriptase protein domains obtained from a database of retrotransposon protein domains (REXdb) (http://repeatexplorer.org/) were aligned together with a specific set of reverse transcriptases from previously published legume *Ty3/gypsy* elements (Neumann et al. 2019) using MAFFT (Katoh and Standley 2013). A total of five clusters for *E. hughesii*, two from *Cenostigma microphyllum*, two from *Libidibia ferrea* and three from *Paubrasilia echinata* (Van-Lume et al. 2019) were included in the final alignment. This alignment was used for construction of phylogenetic trees using neighbor-joining methods in Geneious Prime 2019 (http://www.geneious.com) and, for the comparative analyses, with a reverse transcriptase database of all Chromovirus lineages previously published (Neumann et al. 2019). A sequence from TatIV_Ogre family for a non-chromodomain containing *Ty3/gypsy* element was used as the outgroup.

Abundance of repeats reconstructed over time

A comparative analysis of the repetitive fraction of our *E. hughesii* data was performed by combining the data generated here with the equivalent data obtained from the three other species previously analysed from the Caesalpinia group: *Cenostigma microphyllum*, *Libidibia ferrea* and *Paubrasilia echinata* (Van-Lume et al. 2019). For this, the abundance of the five most abundant repeats in these genomes (*Ty1/copia*-Ale, *Ty3/gypsy*-Athila, *Ty3/gypsy*-CRM, *Ty3/gypsy*-Tekay and all satellite sequences identified) were compared. To explore changes in their abundance over time, these data were reconstructed using the 'character reconstruction' function of the software Mesquite 2 (Maddison and Maddison 2014) and the phylogenetic/ age relationships proposed by Gagnon et al. (2019).

Primer design, PCR amplification and probe labeling

The satDNA and TE primers were designed using Primer3 (Rozen et al. 2000) in the Geneious program from the most conserved part of the repetitive sequences identified in the clusters from RepeatExplorer2 and K-mer graphics (Online Resource 1). The sequences were amplified by PCR from 50 μL reactions containing 2 ng of product (satDNA or TEs), 0.1 mM dNTP, 1 × PCR buffer, 0.4 μM primer, 2 mM $MgCl_2$ and homemade TaqDNA polymerase. The reactions involved 30 cycles of amplification (1 min at 94 °C, 1 minute at 60 °C and 1 minute at 72 °C). The probes were labeled by

nick translation (Pedrosa et al. 2002) with Cy3-dUTP and hybridized *in situ* to *E. hughesii* metaphases, always sequentially to the CMA and DAPI banding (see below).

Chromosomal preparations and CMA / DAPI banding

Root tips obtained from germinated seeds were pretreated with 0.002 M 8-hydroxyquinoline for 5 h at 18 °C, fixed in ethanol:acetic acid (3:1 v/v) for 2–24 h at room temperature and stored at -20 °C. For preparation of slides, fixed root tips were washed in distilled water and digested in a 2% (w/v) cellulase (Onozuka)/20% (v/v) pectinase (Sigma) solution at 37 °C, for 90 min. The meristem was macerated in a drop of 45% acetic acid and spread on a hot plate following Ruban et al. (2014). The CMA/DAPI double-staining technique was used for fluorochrome banding following Mata-Sucre et al. (2020). Slides were aged for 3 days before their analysis using an epifluorescence Leica DMLB microscope. Images were captured with a Cohu CCD video camera using the Leica QFISH software. Finally, the images of the best cells were edited in Adobe Photoshop CS3 version 10.0 for brightness and contrast only.

Fluorescent in situ hybridization (FISH)

All *in situ* hybridizations were performed following Pedrosa et al. (2002). *In situ* hybridization of satDNA and TEs-Cy3 probes used a hybridization mixture containing formamide 50% (v/v), dextran sulphate 10% (w/v), 2× SSC and 50 ng/μL of each probe. The slides were denatured at 75 °C for 5 min. Stringent washes were performed, to give a final stringency of approx. 76%. To localize the rDNA sites, 5S rDNA (D2) from *Lotus japonicus* was labelled with Cy3-dUTP (GE) and 35S rDNA (pTa71) from *Triticum aestivum* labelled with digoxigenin-11-dUTP (Roche). The 35S rDNA probe was detected with sheep anti-digoxigenin FITC conjugate (Roche) and amplified with goat anti-sheep FITC conjugate (Serotec). Images of the best cells were captured as indicated above.

Results

Repeatome fraction analysis and identification

A total of 1,199,444 Illumina reads were analysed with RepeatExplorer2 (Fig 1) corresponding to a genome coverage of 0.3x (1C = 1.09 pg). Of these, 638,260 reads were grouped into 90,849 clusters while 561,184 remained unclustered (Fig 1a). Considering only those clusters with at least 0.01% genome abundance, a total of

331,829 reads were grouped into 174 clusters and identified as repetitive elements, together comprising 28.73% of the whole genome of *E. hughesii* (Fig 1b).

The satDNAs were the most abundant repetitive sequences, representing 7.83% of the *E. hughesii* genome and 27.8% of the analysed repetitive fraction (Table 1, Fig 1b). The two superfamilies of retrotransposons, *Ty3/gypsy* and *Ty1/copia*, were in similar proportions in the genome, representing 2.48% and 2.32% of the genome, respectively. Five families of *Ty3/gypsy* elements were identified, of which the CRM and Athila lineages were the most abundant, accounting for 0.88% and 0.75% respectively (Table 1). In contrast, the *Ty3/gypsy* Tekay lineage was one of the less abundant elements, representing just 0.28% in the genome. For the *Ty1/copia* superfamily, eight families were identified, among these the Ale linage was the most abundant (1.00%). Other repetitive sequences which were not LTR-RT or satellite repeats were found in lower proportions (<0.50%), this group included non-LTR-RT such as LINE elements, various DNA transposons, CACTA, Harbinger, hAT, Mutator elements and pararetroviruses (Table 1).

Chromosome distribution of LTR-RT elements

The *E. hughesii* karyotype is characterized by 11 of the 12 chromosome pairs presenting subtelomeric CMA^+ bands, the remaining chromosome pair does not have a heterochromatin band (see arrowheads in Fig 2a, c, e, g). The chromosomal locations of the most abundant identified repeats were determined by fluorescence *in situ* hybridization (FISH) using probes designed to target the Integrase domains of the Ale, Athila and CRM elements and the chromo domain of the Tekay element. Hybridization signals revealed these four LTR-RT elements were clustered within the heterochromatic subtelomeric regions (Fig 2), co-localizing with the CMA^+ bands and 35S rDNA sites (as shown, for example, in Fig 2b for the *Ty1/copia*-Ale element). Only the CRM chromovirus *Ty3/gypsy* element showed small additional signals in the centromeric region (Fig 2f).

Identification and chromosome mapping of satellite DNA sequences

The TAREAN output of RepeatExplorer2 identified five clusters that contained putative satellites, representing 27.8% of the repetitive fraction of the genome (and hence 7.83% of the whole genome) based on the characterization of 174 clusters (Fig 1b). Two

clusters (CL1 and CL2) (hereafter referred as EhugSat1_48 and EhugSat2_48), together comprised c. 7% of the whole genome (with EhugSat1_48 corresponding to 5%). Based on the sequence similarity of the 48 bp monomeric consensus sequences (89.4% similarity) and the Dot-plot comparison of the consensus sequences (Online Resource 2), EhugSat1_48 and EhugSat2 were classified as two subfamilies of the same family following Ruiz-Ruano et al. (2016). The other three clusters (CL 136, CL166 and CL167) corresponded to three distinct satDNA families (hereafter referred to as EhugSat3_490, EhugSat4_877 and EhugSat5_974). Overall they were less abundant, together comprising just 0.033% of the genome.

Chromosomal mapping of these sequences revealed distinct hybridization sites which co-localized with the CMA^+ bands (Fig 3). For EhugSat1_48 (Fig 3b) and EhugSat5_974 (Fig 3f), these sites were seen to co-localize with all the CMA^+ chromosomal bands with varying intensities, including the 35S rDNA sites (Fig 3a). EhugSat2_48 presented signals in all CMA^+ subtelomeric bands and weak signals, or absence of signals, in the 35S rDNA sites (Fig 3c). In contrast, EhugSat3_490 presented signals only in the 35S rDNA sites (Fig 3d). Comparing this satDNA sequence with the assembled 35S rDNA sequence in NovoPlasty, showed a similarity of 35%, thus we consider this cluster (i.e. EhugSat3_490) to be derived from 35S rDNA (Online Resource 3). FISH of EhugSat4_877 revealed eight hybridization sites, with strong signals in the subtelomeric regions of the chromosomes (Fig 3e).

Phylogenetic relationships between Ty3/gypsy-Tekay elements identified in Caesalpinia group species

To determine the evolutionary relationships between the LTR-RT Ty3/*gypsy*-Tekay elements in Caesalpinia group species, we inferred phylogenetic trees based on alignments of the reverse transcriptase protein domain (Fig 4). The consensus sequences from RepeatExplorer2 of five clusters of *Ty3/gypsy* elements in *E. hughesii* together with two clusters from *Cenostigma microphyllum*, two clusters from *Libidibia ferrea* and three clusters from *Paubrasilia echinata* (obtained from Van-Lume et al. 2019) were analysed. Phylogenetic analysis of these clusters indicated sequence similarity of 79.1% between the Chromovirus-Tekay elements from *E. hughesii* and the previously analysed Caesalpinia group species from Northeastern Brazil which together formed a well-supported clade (Fig 4a). Internal relationships revealed there was considerable similarity between *E. hughesii* and *L. ferrea* (83.6%), and between *P. echinata* and *C.*

microphyllum (80%). A single contig (CL307) of *E. hughesii* showed less sequence similarity (70%) with the rest of the Caesalpinia Group Tekay elements identified, instead being more similar (80%) to a Tekay sequence identified in *Glycine* genus (Fig 4).

Ancestral reconstruction of repeat abundance evolution in Caesalpinia group species

Reconstruction of the abundances of the five main repeat types identified (i.e. satDNAs and the Ale, Athila, CRM and Tekay RT) in the four Caesalpinia group species with available data showed large oscillations in their abundance over time. This was particularly striking for *Ty3/gypsy*-Tekay, which ranged from 54% in *P. echinata* to 0.28% in *E. hughesii*. Based on a previous time-calibrated phylogeny (Gagnon et al. 2019) we define *P. echinata* as the first diverging lineage of our sampling, since this species has an older origin (~ 50 Mya) than *Cenostigma* (~ 15 Mya), *Erythrostemon* (~ 35 Mya) and *Libidibia* (~ 25 Mya). Using this information, and hence assuming that the homogeneous repeat composition identified in *P. echinata* was the plesiomorphic state for these Caesalpinia group species, the ancestral reconstruction analysis points to the following possible evolution of the abundance of the five repeat types identified in the other three species. A gradual reduction in Tekay abundance was predicted to have occurred across the phylogeny for the three species, although the extent of decrease varied between the three species (Fig 5). The opposite trend was observed for the total abundance of satellites which reached their maximum abundance (8%) in *E. hughesii* (Fig 5, Online Resource 4). These predicted evolutionary trends have given rise to the contrasting heterochromatin compositions observed in the four extent species. The most homogeneous heterochromatin was observed in *P. echinata* with its CMA^+ proximal heterochromatic bands are dominated by Tekay elements. In contrast, the three other species had more complex heterochromatin, with contrasting LTR-RTs and satellites present in the heterochromatic CMA^+ bands.

Discussion

Diversity and abundance of repetitive elements within Erythrostemon hughesii *heterochromatin*

The repeat characterization of *Erythrostemon hughesii* (1.06 Gbp/1C) using RepeatExplorer2 revealed that nearly a third (i.e. 28.7%) of its genome is composed of repetitive elements. This is comparable to some other Caesalpinia group species analysed so far (e.g *Cenostigma microphyllum* (2.81 Gbp/1C) with 41% repetitive content and *Libidibia ferrea* (1.79 Gbp/1C) with 38%), but less abundant than in the *Paubrasilia echinata* genome (2.82 Gbp/1C) with 72% (Van-Lume et al. 2019). The most abundant repeats identified in *E. hughesii* were five satellite DNA (satDNA) sequences which together comprised 7.83% of the genome. This contrasts with the other Caesalpinia group species analysed so far, where satDNA sequences made up less than 1% of the genome (Van-Lume et al. 2019). However, such satDNA abundance is not a common feature of all legume genomes because different species have been reported with sequences relatively rich in satDNA and with abundances greater than 1% of their genome. (e.g. Ribeiro et al. 2017; Robledillo et al. 2018; Samoluk et al. 2019; Online Resource 4). The satellite sequences characterized here in *E. hughesii* were shown to distributed in most of the heterochromatic bands showing that these families of repeats are the major components of the CMA^+ heterochromatin bands. Other studies have shown that the DNA sequences present in satellite repeats together with their abundance and distribution can diverge rapidly between species (Koukalova et al. 2009; Plohl et al. 2012). We suggest that the high abundance, subtelomeric location and diversity of the satellite repeat sequences in *E. hughesii* are responsible for the marked difference of the *E. hughesii* karyotype compared to the other Caesalpinia group species.

Although diverse, the most abundant LTR-RT and satDNA families were clearly present co-localized in the *E. hughesii* heterochromatin CMA^+ bands. Even though RTs are typically dispersed along chromosomes, unlike satDNAs that form more defined clusters (Heslop-Harrison and Schwarzacher 2007, 2011; Ribeiro et al. 2017; Piccoli et al. 2018), there are a few examples in plants where members of the *Ty1/copia* and *Ty3/gypsy* superfamilies have been shown to be colocalized and/or intermixed, sometimes with satDNAs repeats in cytologically recognized heterochromatin, a situation that has been described as ‘grouped TEs’ (Belyayev et al. 2001; Gaeta et al. 2010). Whether this represents a lack of in-depth studies is unclear. Certainly, here we have observed a preferentially grouped pattern for the LTR-RT elements identified in the heterochromatin, as observed in other Caesalpinia group species (Van-Lume et al. 2019), and some other legumes (Albernaz et al. in prep.).

The localization of all LTR-elements and satDNA sequences together in the *E. hughesii* heterochromatin demonstrated its heterogeneous composition. How TE blocks accumulate in certain regions of the genome was still poorly understood although the importance of the epigenetic state of the repeat and the chromatin conformation of particular regions of the genome are considered to play a role (see Janssen et al. 2018). Only a few studies have revealed TE complexes, where particular TE clusters are interrupted by other TEs, by genes or by a short tandem arrangement of other TEs (e.g. Giordano et al. 2007; Paço et al. 2019), forming nests of clustered TEs as a consequence of TE integration, intra-chromosomal recombination or variant replication (Vitte et al. 2013; Gao et al. 2015). These TE complexes can be copied and amplified, resulting in many duplications of TE nests as independent fragments across the genome (Bergman et al. 2006; Coline et al. 2014; Zhang et al. 2017), and contributing to genome evolution of the species. Whether such processes are also responsible for giving rise to the diversity of repeat sequences in the subtelomeric heterochromatic CMA^+ bands in *E. hughesii* are currently unclear.

Comparative cytogenomic analysis revealed the Ty3/gypsy-Tekay evolutionary history in the Caesalpinia group

The major difference shown in the comparative analysis of the Caesalpinia group species was the significant difference in the abundance of *Ty3/gypsy*-Tekay repeat elements between species. While the three Brazilian species contained an abundance of these elements ranging from 17 to 54% of the genome (Van-Lume et al. 2019), in *E. hughesii* (Mesoamerican) Tekay elements comprised less than 1% of the genome. These observed differences suggest that the activity of these elements (i.e. amplification) and the rate at which they are deleted via recombination-based processes, clearly differ between these species, as also shown for many other species that have been studied. For example, in a comparative analysis of repeats present in sugarcane (*Saccharum officinale*) and other grasses, evolutionary analyses showed that the proliferation of LTR-RTs varied between RT-lineages in different species, because of contrasting repeat dynamics giving rise to differences in the rate at which new repeats were inserted or removed from the genome (De Setta et al. 2012).

Comparative reverse transcriptase sequence analysis showed an interesting evolutionary history of the Tekay elements in the Caesalpinia group. The Chromovirus-Tekay elements of analysed Caesalpinia group species formed a monophyletic which

also included reverse transcriptase sequences belonging to *Lotus japonicus* and *Glycine* species and which are considered to belong to an ancestral Tekay lineage. This observation reinforces the suggestion that the Tekay elements identified in the Caesalpinia group species colonized their genomes prior to the divergence of these genera (Van-Lume et al. 2019). Although our sample is still small, our analysis shows that the Tekay elements identified in *Cenostigma* are more closely related to species belonging to a different clade of the Caesalpinia group as proposed by Gagnon et al. 2016, recognizing that major clades in the group need to be revised, as previously suggested based on phylogenetic (Gagnon et al. 2016; 2019) and cytomolecular (Mata-Sucre et al. 2020) data.

RTs are an important source of genetic diversity and polymorphisms, and they have the potential to cause sequence and structural changes that lead to genetic variations within plant species (e.g. Rebollo et al. 2011; Bennetzen and Wang 2014; Nie et al. 2019). Despite the different location of the heterochromatic bands in the chromosomal sub-telomeric region, our cytogenetic analysis showed strong hybridization signals of the Tekay element in the sub-telomeric CMA^+ heterochromatic bands, similar to those observed in the Brazilian species which have proximal CMA^+ bands. Remarkably, the large reduction in abundance of Tekay elements in the *E. hughesii* genome (comprising just 0.28% of the genome, Table 1) compared to the Brazilian species (as noted above) correlated with the loss of proximal CMA^+ heterochromatin (Fig 5). This suggests that the sub-telomeric CMA^+ bands may represent an evolutionary novelty in *E. hughesii,* arising from several invasions of other repeats into the subtelomeric region of its chromosomes and/or structural rearrangements (e.g. Ren et al. 2018). What may have driven the reorganization and repeat turnover giving rise to the distinct composition and location of CMA^+ bands in *E. hughesii* is currently unclear. However, a role for the environmental impact on changes in repeat dynamics is possible, given that it has been shown in Caesalpinia group species, that proximal heterochromatic polymorphisms may arise from the shared effects of the environment, phylogeny and geography (Van-Lume et al. 2017; Mata-Sucre et al. 2020).

The observed dominance of LTR-RT in the repeat fraction of the genome has been shown to be a common feature of higher plant genomes in which retroelements represent one of the major forces driving genome evolution (e.g. Bennetzen and Wang 2014; Horváth et al. 2017). Van-Lume et al. (2019) analysing Caesalpinia group

species, suggested that Tekay and/or Athila *Ty3/Gypsy* element enrichment in the CMA^+ proximal bands may have contributed to explaining the association between heterochromatin bands/genome size and ecological traits (Van-Lume et al. 2017; Souza et al. 2019). It was hypothesized that species with heterochromatin composed predominantly of RTs (as in the Northeast Brazilian species of the Caesalpinia group), could be susceptible to genetic variation triggered by environmental changes. The data presented here suggest that this scenario may be more complex, since *E. hughesii* heterochromatin is enriched in both TEs and especially satDNA, making it difficult to disentangle the potential role of ecological factors in impacting the dynamics of these two distinctive types of repeats. The analysis of other species in the Caesalpinia group, as well as correlation tests of the abundance of specific repeats with ecological variables may help to identify more precisely how these repetitive elements influence the interaction between genomes and the environment.

The relationship between heterochromatin repeat diversity and species-richness

Our ancestral reconstruction analyses suggested that the heterogeneity of repetitive sequences present in the heterochromatin changed drastically over time (Fig 5). The species with a repeat-rich heterochromatin, i.e. *C. microphyllum*, *L. ferrea* and *E. hughesii* (in clade II of the Caesalpinia group phylogeny) belong to genera with a relatively high species-richness: *Cenostigma* (14 spp.), *Libidibia* (7 spp.) and *Erythrostemon* (31 spp.) (Gagnon et al. 2016). In contrast, *Paubrasilia echinata* (in clade I), which has its heterochromatin dominated by Tekay elements, is a monospecific genus (Gagnon et al. 2016). Although our study sample is small, these data suggest a possible association between the accumulation of diverse repeat types in the heterochromatin and species-richness in the Caesalpinia group. Interestingly, a literature search for studies analysing the cytogenomic composition of heterochromatin (looking for publications that used an approach similar to ours) revealed this trend was also apparent in other legumes, as well as in some other angiosperm families, although the r^2 values were lower (Online Resources 5 and 6). Evolutionary changes in karyotype structure have long been implicated in speciation events, and lineage-specific variations in the genome and karyotype structure can account for different levels of species diversity (Herrick and Sclavi 2019). Heterochromatin, which can be highly enriched in TEs and other repeats (as shown here), and therefore has the potential to undergo rapid

evolutionary changes, has been suggested to play a role in genome divergence and, consequently, speciation (Fig 6; Ferree and Barbash 2009; Hughes and Hawley 2009; Herrick and Sclavi 2019); certainly there are studies showing how heterochromatin can be involved in intra-genomic conflicts, greatly influencing the genomic landscape of speciation (Austin et al. 2009; Brown and O'Neill 2010; Sawamura 2012).

TE abundance, TE-derived genomic features and chromosomal rearrangements involving TE sequences are frequently lineage-specific and, therefore, suggest that TEs may have contributed to the process of speciation in some species, either as a cause, or an effect (Böhne et al. 2008; Stapley et al. 2015, 2017). The hypothesis that non-coding DNA is of structural importance cytogenetically, driving the speciation process, has been proposed for mammals, since it was observed that different amounts of non-coding DNA associated with heterochromatin account, in part, for different levels of species richness (Herrick and Sclavi 2019). In plants we suggest that species with heterochromatin composed of several diverse repeats are more likely to generate genomic polymorphisms (e.g., loss or gain of some repeats) that provide an increased potential to undergo speciation events (Fig 6; Online Resource 5 and 6). Nevertheless, to establish in more detail the evolution of heterochromatin and the sequences it contains and how this impacts speciation events and hence species diversity, it is clearly necessary to analyse a larger sample of species. For this, a survey of a broader, phylogenetically-representative sample of species in the Caesalpinia group, with its robust phylogeny (Gagnon et al. 2019) and well circumscribed genera (Gagnon et al. 2016, 2019) which range in species number from 1-225, has the potential to be an excellent model for exploring the role of heterochromatin repeats in plant speciation.

Conclusions

All repetitive sequences (four TEs and five satellites) characterized here for the *E. hughesii* genome were shown to be present in the sub-telomeric heterochromatic CMA^+ bands, revealing a high diversity of repeats in these regions. Comparative analyses showed that the main differences between *E. hughesii* and the repeat profiles of the heterochromatic CMA^+ bands in three other Caesalpinia group species were the reduction of *Ty3-gypsy*/Chromovirus RTs and an increase of satDNAs. This difference in repeat abundance may be associated with the loss/change of proximal CMA^+ bands in *E. hughesii* and the invasion of several other repeats into the sub-telomeric region of its

chromosomes. Comparative cytogenomic analyses suggest that the species-richness of genera in the Caesalpinia group could be related to the heterochromatin heterogeneity which may play a role in driving genomic divergence, which is one of the major prerequisites for speciation to occur. The results presented here contribute to understanding the complex history of genome and karyotype evolution in Caesalpinia group species.

Author's Contributions

YMS conducted the experiments and wrote the manuscript. MS contributed in bioinformatics analyses and manuscript revisions. BV provided the genome sequence and contributed with manuscript revisions. APH and EG contributed with critical revisions. IJL and GPL provided access to plant material, contributed with critical manuscript revisions, and with English corrections. GS conceived and designed the research, contributed with critical discussions and manuscript revisions.

Acknowledgments

The authors wish to thank Dra. Magdalena Vaio for her suggestions which improved the manuscript. We thank the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE - APQ- 0970- 2.03/15) for financial support. This study was partly financed by the CAPES (finance Code 001). G.S. receives a productivity fellowship from CNPq (process number PQ - 310693/2018-7).

Conflict of interest

The authors declare that they have no competing interests.

Table 1 Summary of the repetitive elements identified in the *Erythrostemon hughesii* genome estimated using RepeatExplorer2

Fig 1 Genomic clusters of *Erythrostemon hughesii*. (a) Summary of the clustering analysis. Illumina reads were clustered using RepeatExplorer2 resulting in 90,849 clusters. (b) Proportions of different repeats identified by analyzing the 174 clusters which each comprise at least 0.01% of the genome. Together, the repeats identified in these 174 clusters represent 28.7% of the genome

Fig 2 Comparative chromosome mapping of CMA^+ bands (yellow; a, c, e, and g) and different LTR-retrotransposons in *Erythrostemon hughesii* ($2n = 24$). Hybridization signals of the *Ty1/copia*-Ale element (pink; b); *Ty3/gypsy*-Athila element (blue; d); *Ty3/gypsy*-CRM element (orange; f), and *Ty3/gypsy*-Tekay element (green; h). Arrowheads indicate a chromosome pair without CMA^+ bands. Inserts in (b) represent 35S rDNA sites (light blue). Bar = 10 μm

Fig 3 Karyograms of chromosome mapping by FISH in *Erythostemon hughesii* ($2n = 24$). Chromosome CMA^+ bands (yellow) with 35S rDNA sites (blue) (a) and signals of specific satellite DNA sequences (b-f). Hybridization sites of EhugSat1_48 (b), EhugSat2_48 (c), EhugSat3_490 (d), EhugSat4_877 (e) and EhugSat5_974 (f). Bar = 10 μm

Fig 4 Phylogenetic relationships between concatenated reverse transcriptase sequences of *Ty3/gypsy* elements identified in 12 legume species using Neighbor-Joining. Branches of elements from *Erythrostemon hughesii*, *Cenostigama microphyllum*, *Libidibia ferrea* and *Paubrasilia echinata* are represented in yellow, blue, green and pink, respectively (a). Phylogenetic relationships between different Ty3/gypsy elements belonging to the Chromovirus family including the specific *Ty3/gypsy* -Tekay lineage which is shown in purple together with other legumes *Ty3/gypsy* Chromovirus lineages including CRM and Reina, shown in grey (b). Branches shown in bold represents those with >0.8 bootstrap support. For complete phylogenetic relationships between concatenated reverse transcriptase sequences of *Ty3/gypsy* elements

Fig 5 Comparative repeat landscape graphs of TE and satellite DNA content in the *Erythrostemon hughesii* genome compared with three other Caesalpinia group species (Van-Lume et al. 2019). Phylogeny with divergence times for the internal nodes for the two sequences analysed (TE and satellite DNA). Coloured numbers on the branches leading to the species show the percentages of each of the corresponding repeat classes that originate from the time interval corresponding to that branch. Repeat landscapes represent transposable elements of the four main classes of retrotransposons analysed as well as the satellite DNA sequences. The x-axis indicates the divergence time, the y-axis gives the relative percentage of each repeat class for each species. For more detail of the repeat's graphics see online resource 4

Fig 6 Two hypothetical models to explain the correlation between repeat diversity in heterochromatin and species-richness. In this model, we propose that species that possess heterochromatin composed of homogeneous repeats allow greater genetic recognition, and as a consequence, greater genomic compatibility during meiotic pairing, decreasing the rate of speciation over time due to different evolutionary pathways (eg. evolution in concert, homogenization) (left). However, species that accumulate diverse repeats in their heterochromatin are more likely to experience genomic

polymorphisms (eg, loss or gain of some repeats) that provide greater genomic incompatibility, and as a consequence of different evolutionary pathways throughout over time, different speciation events will result in contrasting repeating content and patterns and positions of heterochromatin bands (right)

Online Resource 1 List of primers used to amplify five DNA satellites in the *Erythrostemon hughesii* genome to determine their physical location using fluorescent *in situ* hybridization. Tm: melting temperature

Online Resource 2 RepeatExplorer2 analysis of NGS data for *Erythrostemon hughesii*. (a) Clusters of reads for five satDNAs show graph layouts that are typical for tandem repeats. (b) Comparisons of the five satDNA families in Dot-plot (genomic similarity search tool) where parallel lines indicate tandem repeats (the distance between the diagonals equals the lengths of the motifs)

Online Resource 3 Organization of the 35S ribosomal DNA (rDNA) sequence generated by NOVOPlasty for *Erythrostemon hughesii* (a). Comparison of the sequence similarity between the consensus sequence of the EhugSat3_490 satellite and that of 35S rDNA, showing the similar fragments within the intergeneic spacer (IGS) region (b). Blue boxes represent an amplified region of the 35S rDNA and the satellite sequences

Online Resource 4 Comparative repeats graph of transposable elements of the four main classes of retrotransposons analysed as well as the satellite DNA sequences in the *Erythrostemon hughesii* genome and other three Caesalpinia genomes obtained from Van-Lume et al. (2019)

Online Resource 5 List of species where heterochromatin has been characterized using cytogenomic approches (e.g. RepeatExplorer clustering + FISH mapping) comparing the number of repetitive sequences present in the heterochromatin and the number of species per genus. For “Repeats” the rDNA was not considered. (NRSIH = number of repeat sequences identified in heterochromatin; NSAG = number of species accepted in the genus)

Online Resource 6 Relationship between the number of repetitive sequences identified in heterochromatin (NRSIH) and the number of species accepted in the genus (NSAG). Red = Caesalpinia group; green = Papilionoides; blue = other plants (see Online Resource 4). The correlation was significant ($p<0.05$) and the r values are shown on the graph

References

Austin B, Trivers R, Burt A (2009) Genes in Conflict: the Biology of Selfish Genetic Elements. Harvard University Press

Baniaga A, Barker M (2019) Nuclear Genome Size is Positively Correlated with Median LTR-retroelement Insertion Time in Fern and Lycophyte Genomes. Am Fern J 109 (3): 248-266. https://doi.org/10.1640/0002-8444-109.3.248

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, et al (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 2009: 5:e1000732. https://doi.org/10.1371/journal.pgen.1000732

Beltrão GTA, Guerra M (1990) Citogenética de angiospermas coletadas em Pernambuco-III. Ciência e Cultura 42 (10)

Belyayev A, Josefiová J, Jandová M, Kalendar R, Krak K, Mandák B (2019) Natural history of a satellite DNA family: From the ancestral genome component to species-specific sequences, concerted and non-concerted evolution. Int J Mol Sci 20 (5): 1201. https://doi.org/10.3390/ijms20051201

Belyayev A, Raskina O, Nevo E (2001) Evolutionary dynamics and chromosomal distribution of repetitive sequences on chromosomes of *Aegilops speltoides* revealed by genomic in situ hybridization. Heredity 86 (6): 738-742. https://doi.org/10.1038/sj.hdy.6888910

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Ann Rev Plant Biol 65 (1): 505-530. https://doi.org/10.1146/annurev-arplant-050213-035811

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. Genome Biol 7: R112. https://doi.org/10.1186/gb-2006-7-11-r112

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff J N (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. Chrom Res 16 (1): 203-215. https://doi.org/10.1007/s10577-007-1202-6

Borges LA, Souza LGR, Guerra M, Machado IC, Lewis GP, Lopes AV (2012) Reproductive isolation between diploid and tetraploid cytotypes of *Libidibia ferrea* (= *Caesalpinia ferrea*) (Leguminosae): ecological and taxonomic

implications. Plant Syst Evol 298 (7): 1371-1381. https://doi.org/10.1007/s00606-012-0643-3

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS et al (2018) Ten things you should know about transposable elements. Genome Biol 19: 199. https://doi.org/10.1186/s13059-018-1577-z

Brodie R, Roper RL, Upton C (2004) JDotter: a Java interface to multiple dotplots generated by dotter. Bioinformatics 20 (2): 279-281. https://doi.org/10.1093/bioinformatics/btg406

Brown JD, O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. Ann Rev Genomics Hum Genet 11: 291-316. https://doi.org/10.1146/annurev-genom-082509-141554

Coline G, Théron E, Brasset E, Vaury C (2014) History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in *Drosophila melanogaster*. Front Genet 5: 257. https://doi.org/10.3389/fgene.2014.00257

De Setta N, Metcalfe CJ, Cruz GM, Ochoa EA, Van Sluys MA (2012) Noise or symphony: comparative evolutionary analysis of sugarcane transposable elements with other grasses. In: Grandbastien MA, Casacuberta J (eds) Plant Transposable Elements. Springer, Berlin, Heidelberg. pp 169-192. https://doi.org/10.1007/978-3-642-31842-9_10

Deng H, Xiang S, Guo Q, Jin W, Cai Z, Liang G (2019) Molecular cytogenetic analysis of genome-specific repetitive elements in *Citrus clementina* Hort. Ex Tan. and its taxonomic implications. BMC Plant Biol 19 (1): 77. https://doi.org/10.1186/s12870-019-1676-3

Dierckxsens N, Mardulyn P, Smits G (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45 (4): e18-e18. https://doi.org/10.1093/nar/gkw955

Dluhošová J, Ištvánek J, Nedělník J, Řepková J (2018) Red clover (*Trifolium pratense*) and zigzag clover (*T. medium*)–a picture of genomic similarities and differences. Front Plant Sci 9: 724. https://doi.org/10.3389/fpls.2018.00724

Ferree PM, Barbash DA (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. PLoS Biol 7 (10): e1000234. https://doi.org/10.1371/journal.pbio.1000234

Feschotte C, Zhang X, Wessler SR (2002) Miniature inverted-repeat transposable elements and their relationship to established DNA transposons. In: Craig N, Craigie R, Gellert M, Lambowitz A (ed) Mobile DNA II. ASM Press, Washington, DC. pp 1147-1158. https://doi.org/10.1128/9781555817954.ch50

Fu J, Zhang H, Guo F, Ma L, Wu J, Yue M, et al (2019) Identification and characterization of abundant repetitive sequences in *Allium cepa*. Sci Rep 9 (1): 1-7. https://doi.org/10.1038/s41598-019-52995-9

Gaeta ML, Yuyama PM, Sartori D, Fungaro MHP, Vanzela ALL (2010) Occurrence and chromosome distribution of retroelements and NUPT sequences in *Copaifera langsdorffii* Desf. (Caesalpinioideae). Chromosome Res 18 (4): 515-524. https://doi.org/10.1007/s10577-010-9131-1

Gagnon E, Bruneau A, Hughes CE, de Queiroz LP, Lewis GP (2016) A new generic system for the pantropical Caesalpinia group (Leguminosae). PhytoKeys (71): 1-160. https://doi.org/10.3897/phytokeys.71.9203

Gagnon E, Ringelberg JJ, Bruneau A, Lewis GP, Hughes CE (2019) Global succulent biome phylogenetic conservatism across the pantropical Caesalpinia group (Leguminosae). New Phytol 222 (4): 1994-2008. https://doi.org/10.1111/nph.15633

Garrido-Ramos MA (2015) Satellite DNA in plants: more than just rubbish. Cytogenet Genome Res 146 (2): 153-170. https://doi.org/10.1159/000437008

Garrido-Ramos MA (2017) Satellite DNA: An evolving topic. Genes 8 (9): 230. https://doi.org/10.3390/genes8090230

Gao D, Jiang N, Wing RA, Jiang J, Jackson SA (2015) Transposons play an important role in the evolution and diversification of centromeres among closely related species. Front Plant Sci 6: 216. https://doi.org/10.3389/fpls.2015.00216

Giordano J, Ge Y, Gelfand Y, Abrusán G, Benson G, Warburton PE (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol 3: e137. https://doi.org/10.1371/journal.pcbi.0030137

González ML, Chiapella JO, Urdampilleta JD (2018) Characterization of some satellite DNA families in *Deschampsia antarctica* (Poaceae). Polar Biol 41 (3): 457-468. https://doi.org/10.1007/s00300-017-2205-1

Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M (2015) De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). Genome Biol Evol 7 (4): 1192-1205. https://doi.org/10.1093/gbe/evv050

Hannan AJ (2018) Tandem Repeats and Repeatomes: Delving deeper into the 'Dark Matter' of genomes. EBioMedicine 31: 3-4. https://doi.org/10.1016/j.ebiom.2018.04.004

Heitkam T, Petrasch S, Zakrzewski F, Kögler A, Wenke T, Wanke S, Schmidt T (2015) Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest *Camellia japonica*. Chromosome Res 23 (4): 791-806. https://doi.org/10.1007/s10577-015-9500-x

Herrick J, Sclavi B (2019) Genome diversity and species richness in mammals. BioRxiv: 709311. https://doi.org/10.1101/709311

Heslop-Harrison JS, Schwarzacher T (2007) Domestication, genomics and the future for banana. Annals Bot-London 100 (5): 1073-1084. https://doi.org/10.1093/aob/mcm191

Heslop-Harrison JS, Schwarzacher T (2011) Organisation of the plant genome in chromosomes. Plant J 66 (1): 18-33. https://doi.org/10.1111/j.1365-313X.2011.04544.x

Horváth V, Merenciano M, González J (2017) Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends Genet 33 (11): 832-841. https://doi.org/10.1016/j.tig.2017.08.007

Hloušková P, Mandáková T, Pouch M, Trávníček P, Lysak MA (2019) The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. Annals Bot-London 124: 103-120. https://doi.org/10.1093/aob/mcz036

Hughes SE, Hawley RS (2009) Heterochromatin: a rapidly evolving species barrier. PLoS Biol 7 (10): e1000233. https://doi.org/10.1371/journal.pbio.1000233

Iwata-Otsubo A, Lin JY, Gill N, Jackson SA (2016) Highly distinct chromosomal structures in cowpea (*Vigna unguiculata*), as revealed by molecular cytogenetic analysis. Chromosome Res 24 (2): 197-216. https://doi.org/10.1007/s10577-015-9515-3

Janssen A, Colmenares SU, Karpen GH (2018) Heterochromatin: guardian of the genome. Annu Rev Cell Dev Biol 34: 265-288. https://doi.org/10.1146/annurev-cellbio-100617-062653

Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F (2016) Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genomics 17 (1): 624. https://doi.org/10.1186/s12864-016-2980-z

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30 (4): 772-780. https://doi.org/10.1093/molbev/mst010

Kazazian HH (2004) Mobile elements: drivers of genome evolution. Science 303 (5664): 1626-1632. https://doi.org/10.1126/science.1089670

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647-1649. https://doi.org/10.1093/bioinformatics/bts199

Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novák P, Neumann P et al (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol 208: 596-607. https://doi.org/10.1111/nph.13471

Kirov IV, Kiseleva AV, Van Laere K, Van Roy N, Khrustaleva LI (2017) Tandem repeats of *Allium fistulosum* associated with major chromosomal landmarks. Mol Genet Genomic 292 (2): 453-464. https://doi.org/10.1007/s00438-016-1286-9

Koukalova B, Moraes AP, Renny-Byfield S, Matyasek R, Leitch A, Kovarik A (2009) Fall and rise of satellite repeats in allopolyploids of *Nicotiana* over c. 5 million years. New Phytol 186: 148-160

Krupovic M, Blomberg J, Coffin JM, Dasgupta I, Fan H, Geering AD et al (2018) Ortervirales: new virus order unifying five families of reverse-transcribing viruses. J Virol 92:1-5. https://doi.org/10.1128/JVI.00515-18

Lee YI, Yap JW, Izan S, Leitch IJ, Fay MF, Lee YC, Leitch AR et al (2018) Satellite DNA in *Paphiopedilum* subgenus *Parvisepalum* as revealed by high-throughput sequencing and fluorescent in situ hybridization. BMC Genomics 19 (1): 578. https://doi.org/10.1186/s12864-018-4956-7

Li SF, Guo YJ, Li JR, Zhang DX, Wang BX, Li N, Gao WJ et al (2019) The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). Mobile DNA 10 (1): 3. https://doi.org/10.1186/s13100-019-0147-6

Lisch D (2013) How important are transposons for plant evolution?. Nat Rev Genet 14: 49-61. https://doi.org/10.1038/nrg3374

Liu Q, Li X, Zhou X, Li M, Zhang F, Schwarzacher T, Heslop-Harrison JS (2019) The repetitive DNA landscape in *Avena* (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. BMC Plant Biol 19 (1): 226. https://doi.org/10.1186/s12870-019-1769-z

Mata-Sucre Y, Costa L, Gagnon E, Lewis GP, Leitch IJ, Souza G (2020) Revisiting the cytomolecular evolution of the Caesalpinia group (Leguminosae): a broad sampling reveals new correlations between cytogenetic and environmental variables. Plant Syst Evol 306: 1-13. https://doi.org/10.1007/s00606-020-01674-8

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet 23 (3): 134-139. https://doi.org/10.1016/j.tig.2007.01.004

Macas J, Novák P, Pellicer J, Čížková J, Koblížková A, Neumann P, Leitch IJ et al (2015). In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PLoS One 10 (11): e0143424. https://doi.org/10.1371/journal.pone.0143424

Maddison W, Maddison D (2014) Mesquite: A modular system for evolutionary analysis, version 2.0 Available: http://mesquiteproject.org.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Gwadz M et al (2010) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39 (suppl_1): D225-D229. https://doi.org/10.1093/nar/gkq1189

Maumus F, Quesneville H (2016) Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol 30: 41-46. https://doi.org/10.1016/j.pbi.2016.01.003

Neumann P, Novák P, Hoštáková N, Macas J (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA 10 (1): 1. https://doi.org/10.1186/s13100-018-0144-1

Nie Q, Qiao G, Peng L, Wen X (2019) Transcriptional activation of long terminal repeat retrotransposons sequences in the genome of pitaya under abiotic stress. Plant Physiol Bioch 135: 460-468. https://doi.org/10.1016/j.plaphy.2018.11.014

Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res 45 (12): e111-e111. https://doi.org/10.1093/nar/gkx257

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11 (1): 378. https://doi.org/10.1186/1471-2105-11-378

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive

elements from next-generation sequence reads. Bioinformatics 29 (6): 792-793. https://doi.org/10.1093/bioinformatics/btt054

Paço A, Freitas R, Vieira-da-Silva A (2019) Conversion of DNA Sequences: From a Transposable Element to a Tandem Repeat or to a Gene. Genes 10 (12): 1014. https://doi.org/10.3390/genes10121014

Pedrosa A, Sandal N, Stougaard J, Schweizer D, Bachmair A (2002) Chromosomal map of the model legume *Lotus japonicus*. Genetics 161 (4): 1661-1672.

Peška V, Mandáková T, Ihradská V, Fajkus J (2019) Comparative dissection of three giant genomes: *Allium cepa*, *Allium sativum*, and *Allium ursinum*. Int J Mol Sci 20 (3): 733. https://doi.org/10.3390/ijms20030733

Piccoli MCA, Bardella VB, Cabral-de-Mello DC (2018) Repetitive DNAs in *Melipona scutellaris* (Hymenoptera: Apidae: Meliponidae): chromosomal distribution and test of multiple heterochromatin amplification in the genus. Apidologie 49 (4): 497-504. https://doi.org/10.1007/s13592-018-0577-z

Pita S, Díaz-Viraqué F, Iraola G, Robello C (2019) The Tritryps comparative repeatome: insights on repetitive element evolution in Trypanosomatid pathogens. Genome Biol Evol 11 (2): 546-551. https://doi.org/10.1093/gbe/evz017

Plohl M, Meštrović N, Mravinac B (2012) Satellite DNA evolution. In: Garrido-Ramos MA (ed) Repetitive DNA. Karger Publishers. pp 126-152

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Lorincz MC et al (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet 7 (9): e1002301. https://doi.org/10.1371/journal.pgen.1002301

Ren L, Huang W, Cannon EKS, Bertioli DJ, Cannon SB (2018) A mechanism for genome size reduction following genomic rearrangements. Front Genet 9: 454-454. https://doi.org/10.3389/fgene.2018.00454

Ribeiro T, dos Santos KG, Richard MM, Sévignac M, Thareau V, Geffroy V, Pedrosa-Harand A (2017) Evolutionary dynamics of satellite DNA repeats from *Phaseolus* beans. Protoplasma 254 (2): 791-801. https://doi.org/10.1007/s00709-016-0993-8

Ribeiro T, Vasconcelos E, dos Santos KG, Vaio M, Brasileiro-Vidal AC, Pedrosa-Harand A (2019) Diversity of repetitive sequences within compact genomes of *Phaseolus* L. beans and allied genera *Cajanus* L. and *Vigna* Savi. Chromosome Res: 1-15. https://doi.org/10.1007/s10577-019-09618-w

Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72 (4): 686-727. https://doi.org/10.1128/MMBR.00011-08

Robledillo LÁ, Koblížková A, Novák P, Böttinger K, Vrbová I, Neumann P, Macas J et al (2018) Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. Sci Rep-UK 8 (1): 1-11. https://doi.org/10.1038/s41598-018-24196-3

Rodrigues PS, Souza MM, Corrêa RX (2014) Karyomorphology and karyotype asymmetry in the South American Caesalpinia species (Leguminosae: Caesalpinioideae). Genet Mol Biol 13: 8278-93. http://dx.doi.org/10.4238/2014.October.20.4

Rodrigues PS, Souza MM, Melo CAF, Pereira TNS, Corrêa RX (2018) Karyotype diversity and 2C DNA content in species of the Caesalpinia group. BMC Genetics 19 (1): 25. https://doi.org/10.1186/s12863-018-0610-2

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Misener S, Krawetz SA (eds) Bioinformatics Methods and Protocols. Methods in Molecular Biology™ vol 132. Humana Press, Totowa, NJ. pp 365-386.

Ruban A, Fuchs J, Marques A, Schubert V, Soloviev A, Raskina O, Houben A (2014) B chromosomes of *Aegilops speltoides* are enriched in organelle genome-derived sequences. PLoS One 9 (2): e90214. https://doi.org/10.1371/journal.pone.0090214

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep-UK 6: 28333. https://doi.org/10.1038/srep28333

Said M, Hřibová E, Danilova TV, Karafiátová M, Čížková J, Friebe B, et al (2018) The *Agropyron cristatum* karyotype, chromosome structure and cross-genome

homoeology as revealed by fluorescence in situ hybridization with tandem repeats and wheat single-gene probes. Theor Appl Genet 131 (10): 2213-2227. https://doi.org/10.1007/s00122-018-3148-9

Samoluk SS, Chalup LM, Chavarro C, Robledo G, Bertioli DJ, Jackson SA, Seijo G (2019) Heterochromatin evolution in *Arachis* investigated through genome-wide analysis of repetitive DNA. Planta 249 (5): 1405-1415. https://doi.org/10.1007/s00425-019-03096-4

Sawamura K (2012) Chromatin evolution and molecular drive in speciation. Intern J Evol Biol 2012 (301894): 1-9. https://doi.org/10.1155/2012/301894

Schmidt T, Heitkam T, Liedtke S, Schubert V, Menzel G (2019) Adding color to a century-old enigma: multi-color chromosome identification unravels the autotriploid nature of saffron (*Crocus sativus*) as a hybrid of wild *Crocus cartwrightianus* cytotypes. New Phytol 222 (4): 1965-1980. https://doi.org/10.1111/nph.15715

Sola-Campoy PJ, Robles F, Schwarzacher T, Rejon CR, de la Herran R, Navajas-Perez R (2015) The molecular cytogenetic characterization of Pistachio (*Pistacia vera* L.) suggests the arrest of recombination in the largest heteropycnotic pair HC1. PloS one 10(12). https://doi.org/10.1371/journal.pone.0143861

Souza G, Costa L, Guignard MS, Van-Lume B, Pellicer J, Gagnon E, Lewis GP (2019) Do tropical plants have smaller genomes? Correlation between genome size and climatic variables in the Caesalpinia Group (Caesalpinioideae, Leguminosae). Persp Plant Ecol 38:13-23. https://doi.org/10.1016/j.ppees.2019.03.002

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM (2017) Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc Lond 372 (1736): 20160455. https://doi.org/10.1098/rstb.2016.0455

Stapley J, Santure AW, Dennis SR (2015) Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol 24 (9): 2241-2252. https://doi.org/10.1111/mec.13089

Van-Lume B, Esposito T, Diniz-filho J, Gagnon E, Lewis G, Souza G (2017) Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. Persp Plant Ecol 29: 51-63. 2017. https://doi.org/10.1016/j.ppees.2017.11.004

Van-Lume B, Mata-Sucre Y, Báez M, Ribeiro T, Huettel B, Gagnon E, Souza G et al (2019) Evolutionary convergence or homology? Comparative cytogenomics of Caesalpinia group species (Leguminosae) reveals diversification in the proximal heterochromatic composition. Planta 250 (6): 2173-2186. https://doi.org/10.1007/s00425-019-03287-z

Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Saccheri IJ et al. (2016) The industrial melanism mutation in British peppered moths is a transposable element. Nature 534 (7605): 102. https://doi.org/10.1038/nature17951

Vitte C, Estep MC, Leebens-Mack J, Bennetzen JL (2013) Young, intact and nested retrotransposons are abundant in the onion and asparagus genomes. Annals Bot 112 (5): 881-889. https://doi.org/10.1093/aob/mct155

Vondrak T, Robledillo LÁ, Novák P, Koblížková A, Neumann P, Macas J (2020) Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. Plant J 101 (2): 484. https://doi.org/10.1111/tpj.14546

Waminal NE, Pellerin RJ, Jang W, Kim HH, Yang TJ (2018) Characterization of chromosome-specific microsatellite repeats and telomere repeats based on low coverage whole genome sequence reads in *Panax ginseng*. Plant Breed Biotechnol 6 (1): 74-81. https://doi.org/10.9787/PBB.2018.6.1.74

Wang GX, He QY, Macas J, Novák P, Neumann P, Meng DX, et al (2017) Karyotypes and distribution of tandem repeat sequences in *Brassica nigra* determined by fluorescence in situ hybridization. Cytogenet Genome Res 152 (3): 158-165. https://doi.org/10.1159/000479179

Weising K, Nybom H, Pfenninger M, Wolff K, Kahl G (2005) DNA fingerprinting in plants: principles, methods, and applications. CRC Press, Boca Raton.

Wendel JF, Jackson SA, Meyers BC, Wing RA (2016) Evolution of plant genome architecture. Genome Biol 17 (1): 1-14. https://doi.org/10.1186/s13059-016-0908-1

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Paux E et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8 (12): 973-982. https://doi.org/10.1038/nrg2165

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, et al. (2018) Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol 19:103. https://doi.org/10.1186/s13059-018-1479-0

Zhang QJ, Gao LZ (2017) Rapid and recent evolution of LTR retrotransposon drives rice genome evolution during the speciation of AA-genome *Oryza* species. G3: Genes Genom Genet 7 (6): 1875-1885. https://doi.org/10.1534/g3.116.037572

Zhang S, Zhu M, Shang Y, Wang J, Dawadundup Zhuang L, Qi Z et al (2019) Physical organization of repetitive sequences and chromosome diversity of barley revealed by fluorescence in situ hybridization (FISH). Genome 62 (5): 329-339. https://doi.org/10.1139/gen-2018-0182

Zhou HC, Waminal NE, Kim HH (2019) In silico mining and FISH mapping of a chromosome-specific satellite DNA in *Capsicum annuum* L. Genes Genomics 41 (9): 1001-1006. https://doi.org/10.1007/s13258-019-00832-8

Table 1 Summary of the repetitive elements identified in the *Erythrostemon hughesii* genome estimated using RepeatExplorer2

Repeats		**Genome proportion [%]**
Satellite		7.83
rDNA		
	35S	4.30
	5S	0.17
Retroelements		
Solo LTR		4.96
LTR-RT		
	Ty1/*copia* (Total)	2.32
	Ale	1.00
	Tork	0.50
	Bianca	0.31
	Ikeros	0.23
	Alesia	0.14
	Ivana	0.07
	SIRE	0.04
	Tar	0.03
	Ty3/*gypsy* (Total)	2.48
	CRM	0.88
	Athila	0.75
	TatIV_ogre	0.48
	Tekay	0.28
	Reina	0.10
Other TE elements		
	LINE	0.62
	hAT	0.56
	Mutator	0.43
	PIF_Harbinger	0.39
	Pararetrovirus	0.20
	CACTA	0.17
Unclassified		4.15

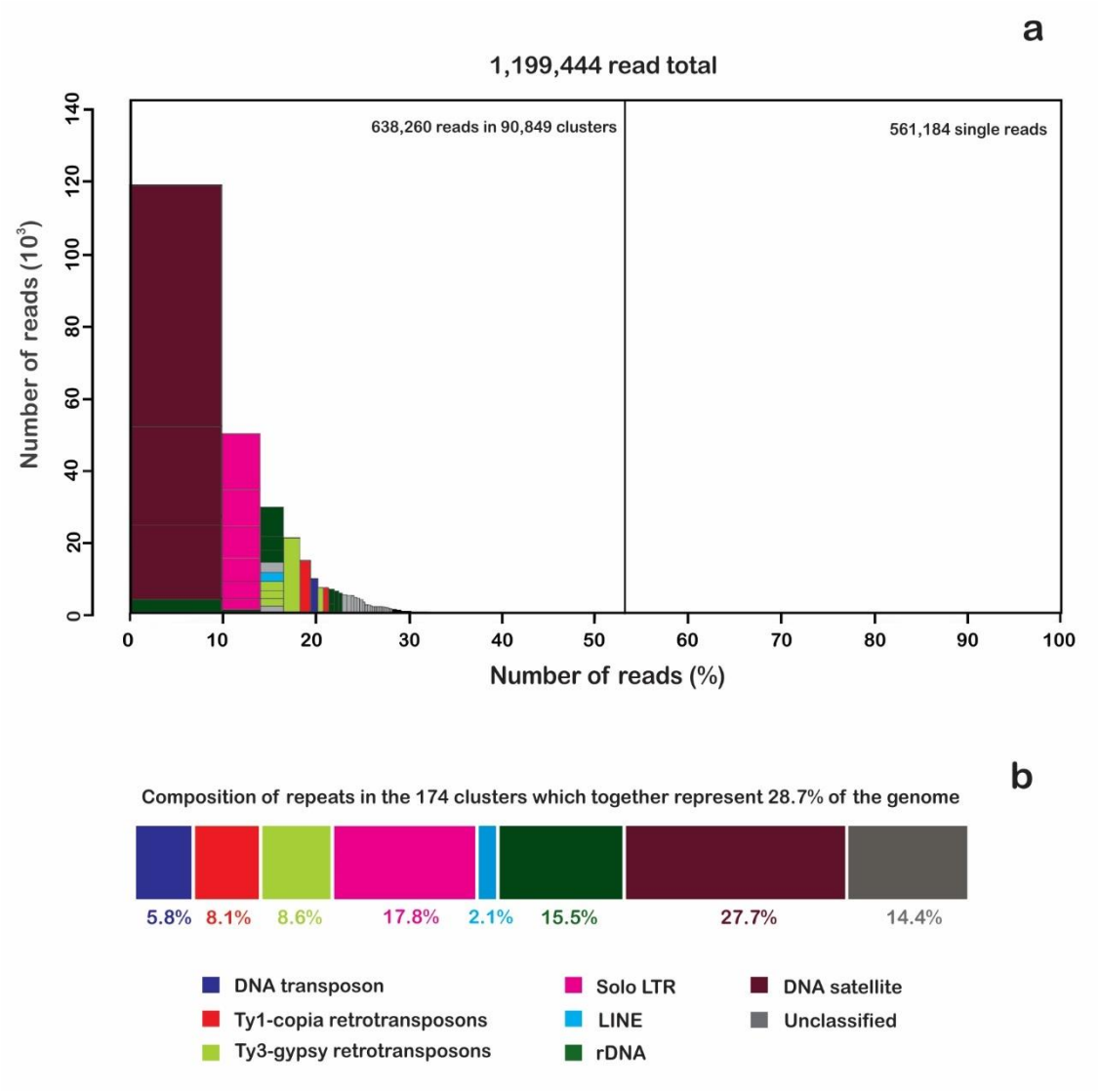


Fig 1 Genomic clusters of *Erythrostemon hughesii*. (a) Summary of the clustering analysis. Illumina reads were clustered using RepeatExplorer2 resulting in 90,849 clusters. (b) Proportions of different repeats identified by analyzing the 174 clusters which each comprise at least 0.01% of the genome. Together, the repeats identified in these 174 clusters represent 28.7% of the genome

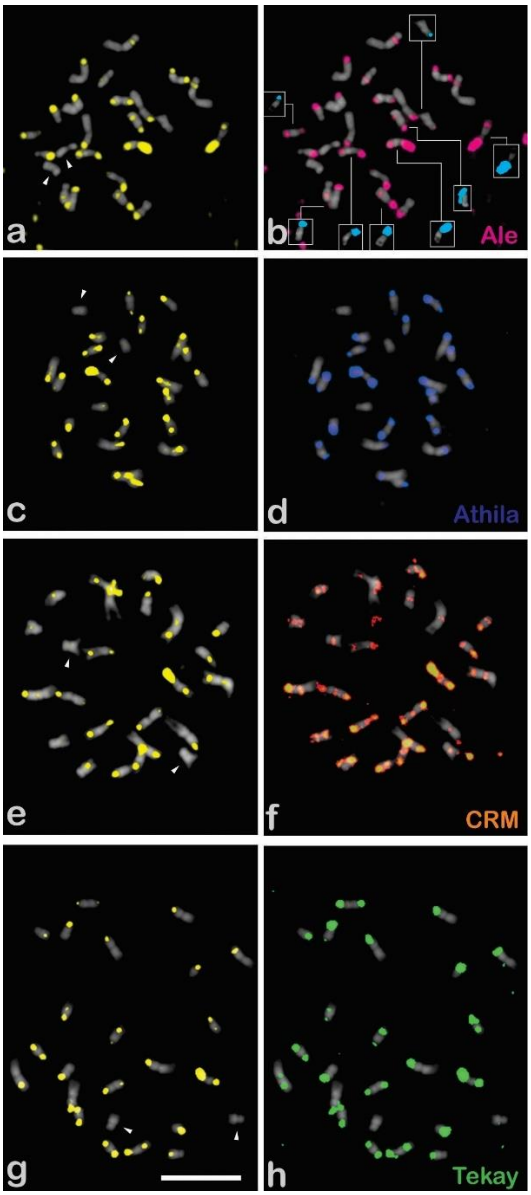


Fig 2 Comparative chromosome mapping of CMA^+ bands (yellow; a, c, e, and g) and different LTR-retrotransposons in *Erythrostemon hughesii* ($2n = 24$). Hybridization signals of the *Ty1/copia*-Ale element (pink; b); *Ty3/gypsy*-Athila element (blue; d); *Ty3/gypsy*-CRM element (orange; f), and *Ty3/gypsy*-Tekay element (green; h). Arrowheads indicate a chromosome pair without CMA^+ bands. Inserts in (b) represent 35S rDNA sites (light blue). Bar = 10 μm

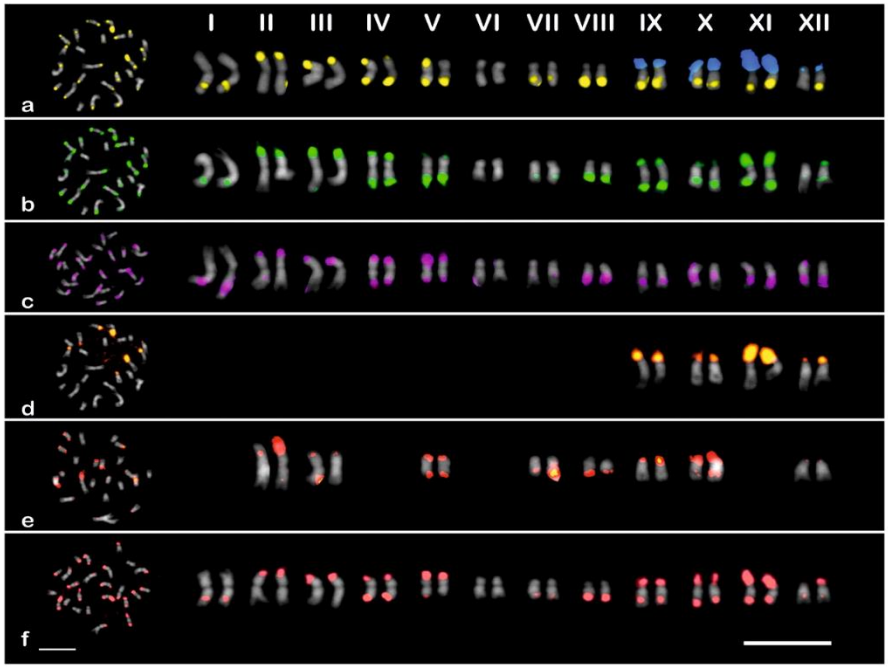


Fig 3 Karyograms of chromosome mapping by FISH in *Erythostemon hughesii* ($2n = 24$). Chromosome CMA^+ bands (yellow) with 35S rDNA sites (blue) (a) and signals of specific satellite DNA sequences (b-f). Hybridization sites of EhugSat1_48 (b), EhugSat2_48 (c), EhugSat3_490 (d), EhugSat4_877 (e) and EhugSat5_974 (f). Bar = 10 μm

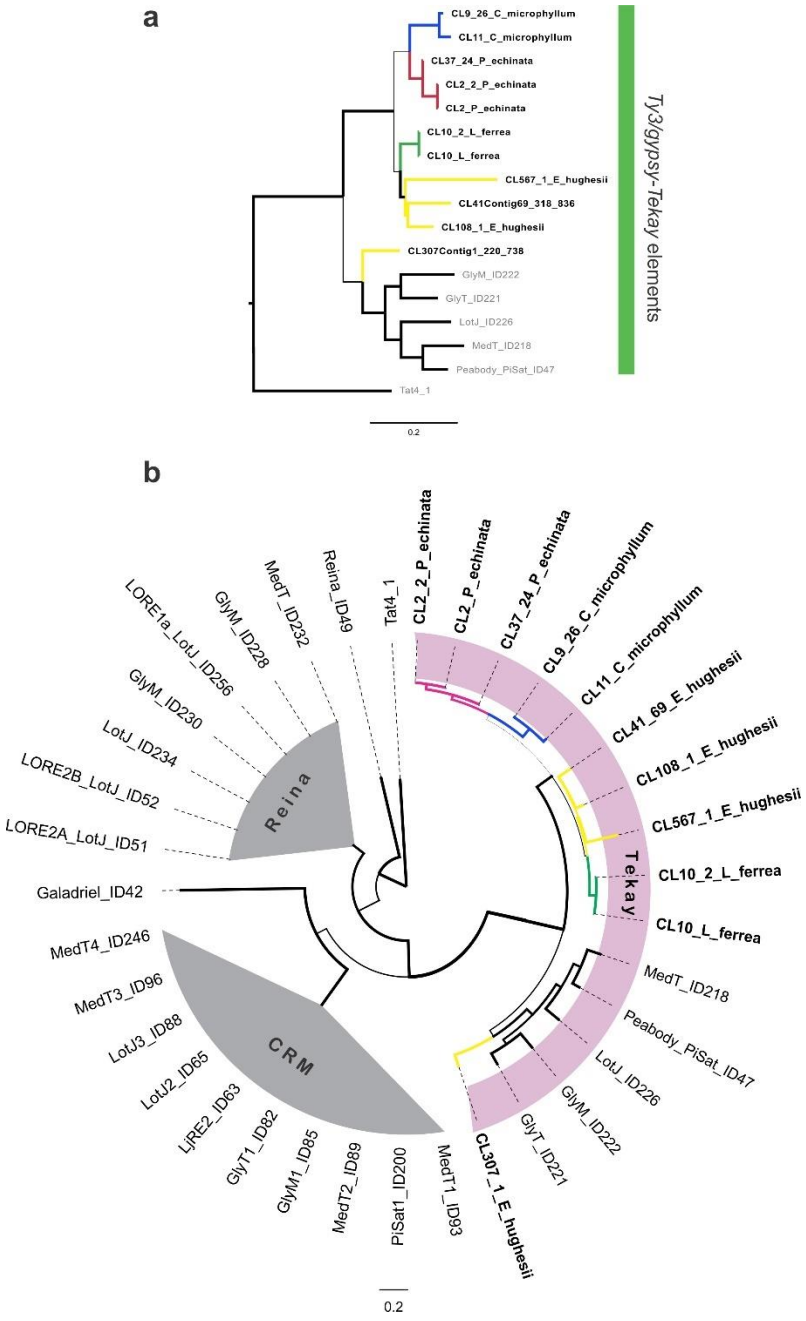


Fig 4 Phylogenetic relationships between concatenated reverse transcriptase sequences of *Ty3/gypsy* elements identified in 12 legume species using Neighbor-Joining. Branches of elements from *Erythrostemon hughesii*, *Cenostigama microphyllum*, *Libidibia ferrea* and *Paubrasilia echinata* are represented in yellow, blue, green and pink, respectively (a). Phylogenetic relationships between different Ty3/gypsy elements belonging to the Chromovirus family including the specific *Ty3/gypsy* -Tekay lineage which is shown in purple together with other legumes *Ty3/gypsy* Chromovirus lineages including CRM and Reina, shown in grey (b). Branches shown in bold represents those with >0.8 bootstrap support. For complete phylogenetic relationships between concatenated reverse transcriptase sequences of *Ty3/gypsy* elements see Online Resource 7

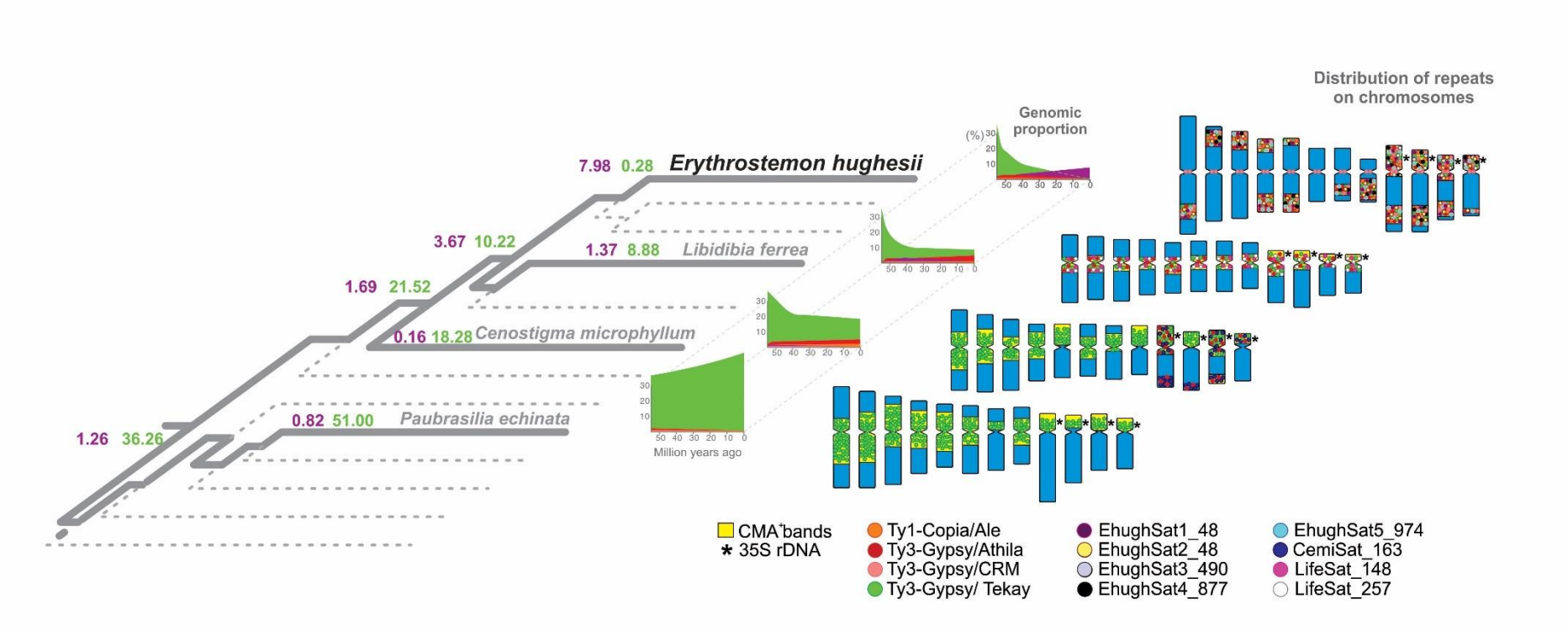


Fig 5 Comparative repeat landscape graphs of TE and satellite DNA content in the *Erythrostemon hughesii* genome compared with three other Caesalpinia group species (Van-Lume et al. 2019). Phylogeny with divergence times for the internal nodes for the two sequences analysed (TE and satellite DNA). Coloured numbers on the branches leading to the species show the percentages of each of the corresponding repeat classes that originate from the time interval corresponding to that branch. Repeat landscapes represent transposable elements of the four main classes of retrotransposons analysed as well as the satellite DNA sequences. The x-axis indicates the divergence time, the y-axis gives the relative percentage of each repeat class for each species. For more detail of the repeat's graphics see online resource 4

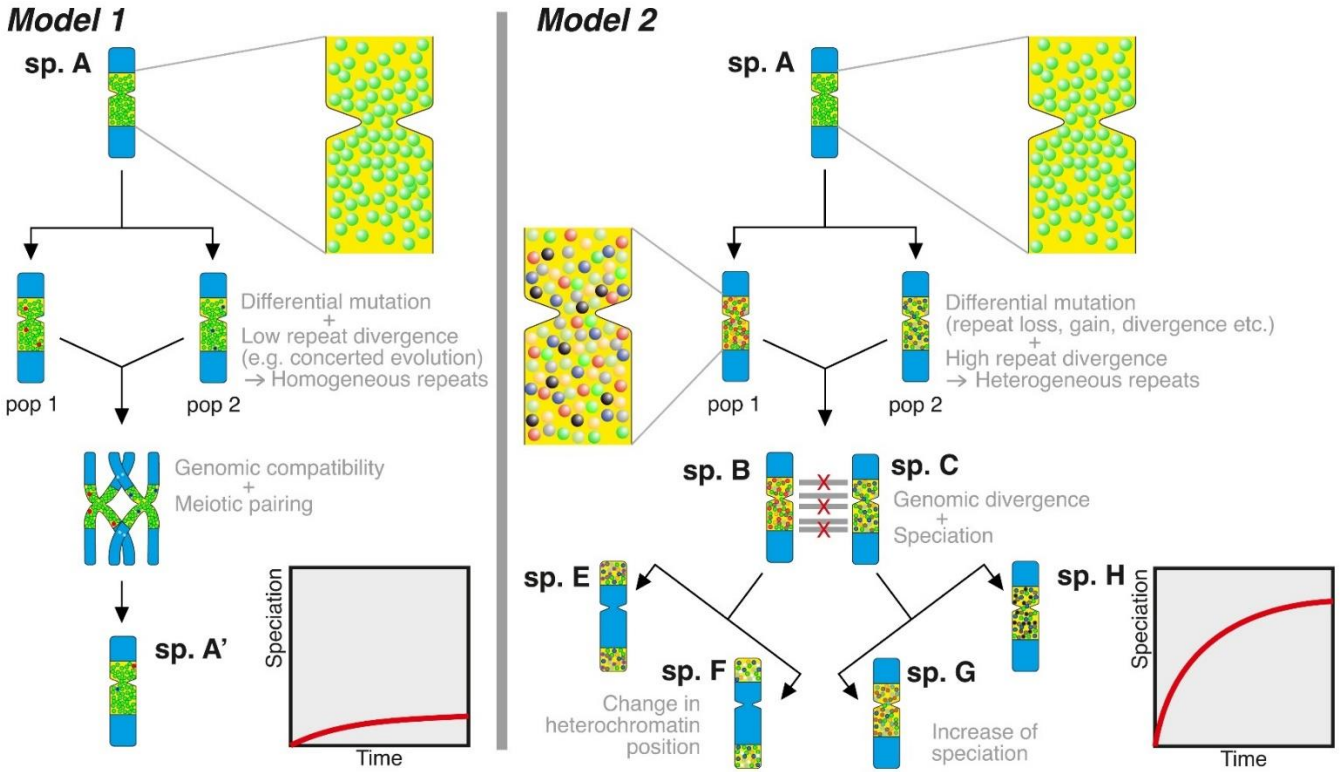
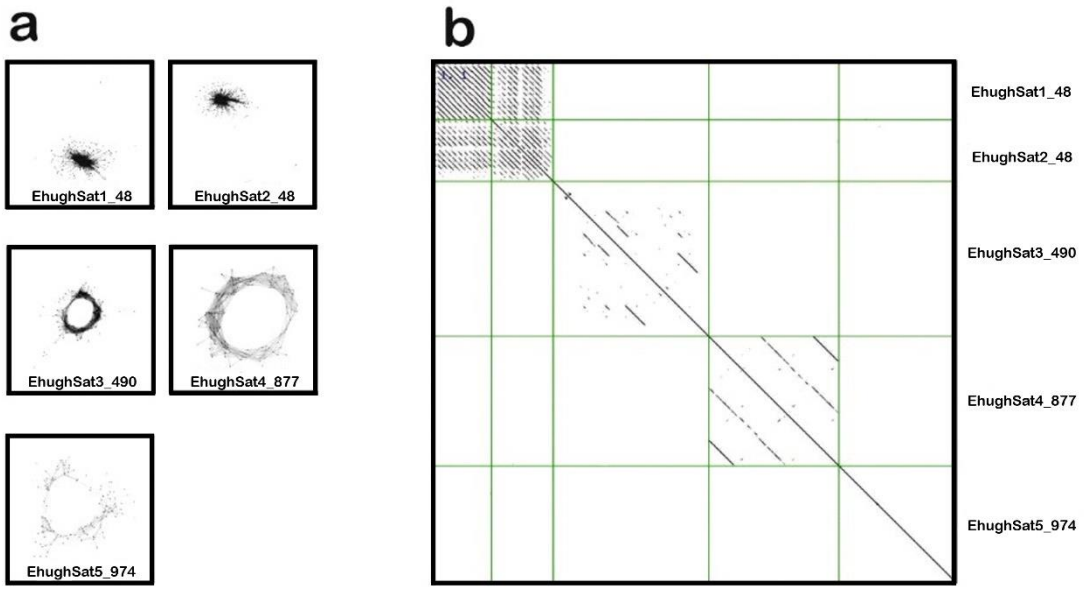


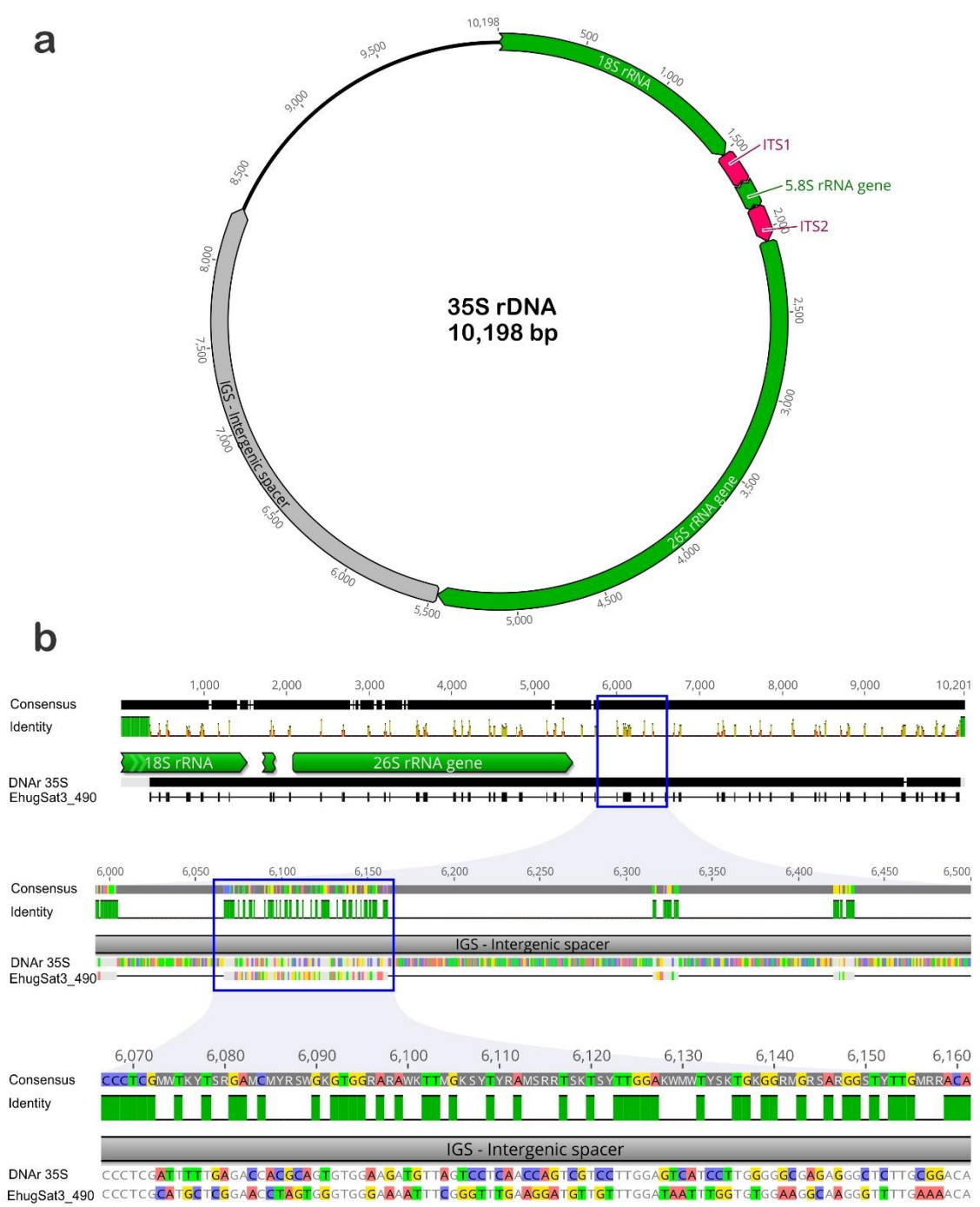
Fig 6 Two hypothetical models to explain the correlation between repeat diversity in heterochromatin and species-richness. In this model, we propose that species that possess heterochromatin composed of homogeneous repeats allow greater genetic recognition, and as a consequence, greater genomic compatibility during meiotic pairing, decreasing the rate of speciation over time due to different evolutionary pathways (eg. evolution in concert, homogenization) (left). However, species that accumulate diverse repeats in their heterochromatin are more likely to experience genomic polymorphisms (eg, loss or gain of some repeats) that provide greater genomic incompatibility, and as a consequence of different evolutionary pathways throughout over time, different speciation events will result in contrasting repeating content and patterns and positions of heterochromatin bands (right)

Online Resource 1 List of primers used to amplify five DNA satellites in the *Erythrostemon hughesii* genome to determine their physical location using fluorescent *in situ* hybridization. Tm: melting temperature

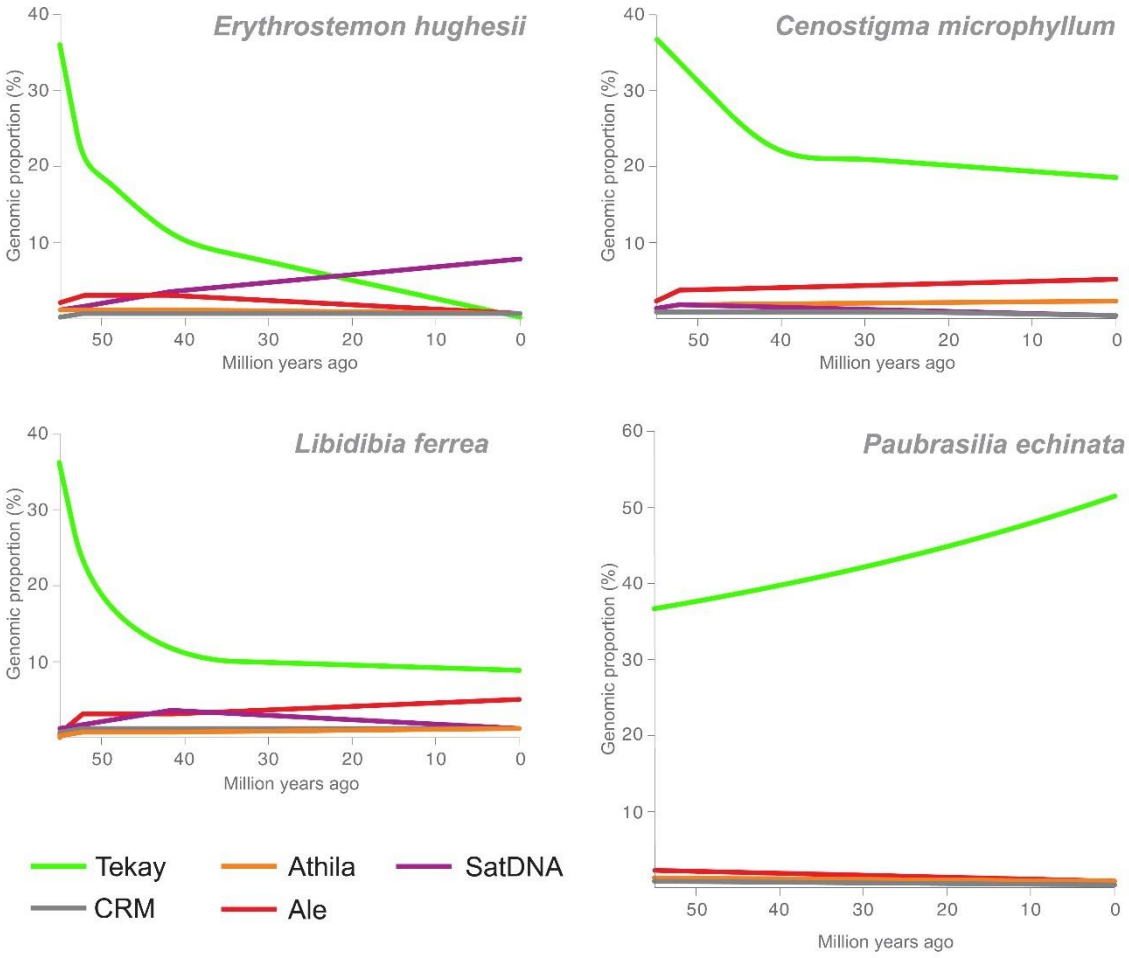
Element	Name	Sequence	Tm (°C)
EhugSat1_48	F	TGAGGTATCAGGTCTTTGGTGC	59.9
	R	ACCTTAGGCACCAAAGACCTG	
EhugSat2_48	F	GCATGAGGTCTTTGGTGCCT	60.6
	R	CCTTAGGCACCAAAGACCTCAT	
EhugSat3_490	F	ACCCCAAAACCGTTTCCCTT	58.9
	R	AGGGAAAATTTGGGTTGGGC	
EhugSat4_877	F	TGCCTAACCCGTCAATGTGT	59.6
	R	GGAATGAATTGAAGTGACGTGC	
EhugSat5_974	F	ACCCTCCTTGCCCCTCATAT	60.4
	R	CAGGGCTTGCAATTTCGGC	



Online Resource 2 RepeatExplorer2 analysis of NGS data for *Erythrostemon hughesii*. (a) Clusters of reads for five satDNAs show graph layouts that are typical for tandem repeats. (b) Comparisons of the five satDNA families in Dot-plot (genomic similarity search tool) where parallel lines indicate tandem repeats (the distance between the diagonals equals the lengths of the motifs)



Online Resource 3 Organization of the 35S ribosomal DNA (rDNA) sequence generated by NOVOPlasty for *Erythrostemon hughesii* (a). Comparison of the sequence similarity between the consensus sequence of the EhugSat3_490 satellite and that of 35S rDNA, showing the similar fragments within the intergeneic spacer (IGS) region (b). Blue boxes represent an amplified region of the 35S rDNA and the satellite sequences

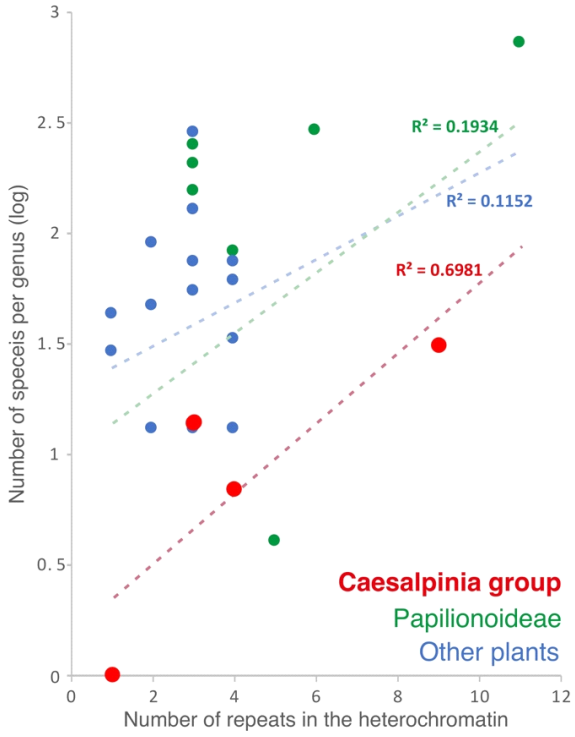


Online Resource 4 Comparative repeats graph of transposable elements of the four main classes of retrotransposons analysed as well as the satellite DNA sequences in the *Erythrostemon hughesii* genome and other three Caesalpinia genomes obtained from Van-Lume et al. (2019)

Online Resource 5 List of species where heterochromatin has been characterized using cytogenomic approches (e.g. RepeatExplorer clustering + FISH mapping) comparing the number of repetitive sequences present in the heterochromatin and the number of species per genus. For "Repeats" the rDNA was not considered. (NRSIH = number of repeat sequences identified in heterochromatin; NSAG = number of species accepted in the genus)

Family:Species	NRSIH	NSAG	Reference
Amaryllidaceae			
***Allium cepa* L.**	4	972	Fu et al. (2019)/ Peška et al. (2019)
***Allium fistulosum* L.**	2	972	Kirov et al. (2017)
Anacardiaceae			
***Pistacia vera* L.**	2	13	Sola-Campoy et al. (2015)
Araliaceae			
***Panax ginseng* C.A.Mey.**	3	13	Waminal et al. (2018)
Brassicaceae			
***Brassica nigra* (L.) Koch.**	4	73	Wang et al. (2017)
***Chorispora tenella* (Pall.) DC.**	4	13	Hloušková et al. (2019)
***Dontostemon micranthus* C.A.Mey.**	3	14	Hloušková et al. (2019)
***Hesperis sylvestris* Crantz**	4	61	Hloušková et al. (2019)
Iridaceae			
***Crocus sativus* L.**	3	127	Schmidt et al. (2019)
Leguminosae			
***Arachis glandulifera* Stalker**	4	82	Samoluk et al. (2019)
***Cenostigma michrophyllum* (Mart. ex G. Don) Gagnon & G. P. Lewis**	3	14	Van-Lume et al. (2019)
***Erythrostemon hughesii* (G. P. Lewis) E. Gagnon & G. P. Lewis**	9	31	This study
***Lathyrus sativus* L.**	3	206	Vondrak et al. (2020)
***Libidibia ferrea* (Mart. exTul.) L.P. Queiroz**	4	7	Van-Lume et al. (2019)
***Lupinus albus* L.**	11	724	Marques et al. in prep.
***Paubrasilia echinata* (Lam.) Gagnon, H. C. Lima & G.P. Lewis**	1	1	Van-Lume et al. (2019)
***Phaseolus vulgaris* L.**	2	98	Ribeiro et al. 2019
***Phaseolus coccineus* L.**	4	98	Ribeiro et al. 2019

***Pterodom pubescens* (Benth.) Benth**	5	4	Albernaz et al. in prep.
***Trifolium pratense* L.**	3	250	Dluhošová et al. (2018)
***Vicia faba* L.**	6	289	Robledillo et al. (2018)
***Vigna unguiculata* (L.) Walp**	3	153	Iwata-Otsubo et al. (2016)
Poaceae			
***Agropyron cristatum* (L.) Gaertn.**	1	29	Said et al. (2018)
***Deschampsia antarctica* É.Desv.**	2	47	González et al. (2018)
Rutaceae			
***Citrus clementina* hort.**	4	33	Deng et al. (2019)
Solanaceae			
***Capsicum annuum* L.**	1	43	Zhou et al. (2019)
Theaceae			
***Camellia japonica* L.**	3	280	Heitkam et al. (2015)



Online Resource 6 Relationship between the number of repetitive sequences identified in heterochromatin (NRSIH) and the number of species accepted in the genus (NSAG). Red = Caesalpinia group; green = Papilionoides; blue = other plants (see Online Resource 4). The correlation was significant ($p<0.05$) and the r values are shown on the graph

4 CONCLUSÕES

- No grupo Caesalpinia roram confirmados os três padrões de bandas de heterocromatina proximal: $CMA^{+}/DAPI^{-}$, $CMA^{0}/DAPI^{-}$ e $CMA^{0}/DAPI^{0}$.
- Os clados *'Coulteria + Tara'* e *'Arquita + Balsamocarpon + Erythrostemon + Pomaria'* mostraram independentemente bandas $CMA0/DAPI^{-}$ associadas a distribuições geográficas em latitudes mais altas.
- A intensidade CMA / DAPI ao longo do cromossomo, tamanho do genoma e latitude estão autocorrelacionadas no grupo Caesalpinia, sugerindo que a evolução da heterocromatina no grupo está respondendo a fatores ambientais.
- Todas as sequências repetitivas caracterizadas no genoma de *E. hughesii* mostraram sinais FISH nas bandas heterocromáticas subteloméricas CMA^{+}, mostrando uma diversidade de elementos repetitivos nessa região.
- As análises comparativas mostraram que as principais diferenças entre *E. hughesii* e outros genomas do Grupo Caesalpinia foram a redução do retrotransposon *Ty3-gypsy* /Cromovírus e o ganho de várias outras repetições em tandem na região subtelomérica de seus cromossomos.
- Análises de reconstrução de sequência integrado a análises comparativas de cariótipo sugerem que a riqueza de espécies no Grupo Caesalpinia pode estar relacionada à heterogeneidade da heterocromatina como fator de divergência genômica.

REFERÊNCIAS

ACOSTA, María Cristina; MOSCONE, Eduardo Alberto; COCUCCI, Andrea Aristides. Using chromosomal data in the phylogenetic and molecular dating framework: karyotype evolution and diversification in *Nierembergia* (Solanaceae) influenced by historical changes in sea level. **Plant Biology**, v. 18, n. 3, p. 514-526, 2016.

DE ALMEIDA-TOLEDO, L. F. et al. Chromosome evolution in fish: sex chromosome variability in *Eigenmannia virescens* (Gymnotiformes: Sternopygidae). **Cytogenetic and genome research**, v. 99, n. 1-4, p. 164-169, 2002.

ALVES, Mao; CUSTODIO, AV de C. **Citogenetica de leguminosas coletadas no Estado do Ceara (Cytogenetics of leguminosae coleted in the state of Ceara)**. 1989.

AMARO, Renata Cecília et al. Demographic processes in the montane Atlantic rainforest: molecular and cytogenetic evidence from the endemic frog Proceratophrys boiei. **Molecular Phylogenetics and Evolution**, v. 62, n. 3, p. 880-888, 2012.

BELTRÃO, G. T. A.; GUERRA, M. Citogenética de angiospermas coletadas em Pernambuco-III. **Ciência e Cultura**, v. 42, n. 10, 1990.

BELYAYEV, Alexander; RASKINA, Olga; NEVO, Eviatar. Evolutionary dynamics and chromosomal distribution of repetitive sequences on chromosomes of *Aegilops speltoides* revealed by genomic in situ hybridization. **Heredity**, v. 86, n. 6, p. 738-742, 2001.

BEREZIKOV, Eugene. Evolution of microRNA diversity and regulation in animals. **Nature Reviews Genetics**, v. 12, n. 12, p. 846-860, 2011.

BILINSKI, Paul et al. Parallel altitudinal clines reveal trends in adaptive evolution of genome size in *Zea mays*. **PLoS genetics**, v. 14, n. 5, 2018.

BLÖCH, Cordula et al. Molecular phylogenetic analyses of nuclear and plastid DNA sequences support dysploid and polyploid chromosome number changes and reticulate evolution in the diversification of *Melampodium* (Millerieae, Asteraceae). **Molecular Phylogenetics and Evolution,** v. 53, n. 1, p. 220-233, 2009.

BORGES, Laís Angélica et al. Reproductive isolation between diploid and tetraploid cytotypes of *Libidibia ferrea* (= *Caesalpinia ferrea*) (Leguminosae): ecological and taxonomic implications. **Plant Systematics and Evolution**, v. 298, n. 7, p. 1371-1381, 2012.

BROOKFIELD, J. F. Y. Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families. **Cytogenetic and genome research**, v. 110, n. 1-4, p. 383-391, 2005.

CAPONIO, Irene et al. Ploidy dimorphism and reproductive biology in *Stenodrepanum bergii* (Leguminosae), a rare South American endemism. **Genome**, v. 55, n. 1, p. 1-7, 2012.

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

CHIARINI, Franco Ezequiel; GAUTHIER, Martha J. Chromosome differentiation in three species of *Leptostemonum* (*Solanum*, Solanaceae) endemic to oceanic islands. 2016.

CHIAVEGATTO, Raquel Bezerra et al. Heterochromatin Bands and rDNA Sites Evolution in Polyploidization Events in *Cynodon* Rich. (Poaceae). **Plant Molecular Biology Reporter**, v. 37, n. 5, p. 477-487, 2019.

CANGIANO, Alejandra Maria; BERNARDELLO, Gabriel. Karyotype analysis in Argentinean species of Caesalpinia (Leguminosae). **Caryologia**, v. 58, n. 3, p. 262-268, 2005.

CASACUBERTA, Elena; GONZÁLEZ, Josefa. The impact of transposable elements in environmental adaptation. **Molecular ecology**, v. 22, n. 6, p. 1503-1517, 2013.

CHALUP, Laura et al. Karyotype characterization and evolution in South American species of *Lathyrus* (Notolathyrus, Leguminosae) evidenced by heterochromatin and rDNA mapping. **Journal of plant research**, v. 128, n. 6, p. 893-908, 2015.

CHUONG, Edward B.; ELDE, Nels C.; FESCHOTTE, Cédric. Regulatory activities of transposable elements: from conflicts to benefits. **Nature Reviews Genetics**, v. 18, n. 2, p. 71, 2017.

COLOMBO, P.; CONFALONIERI, V. Cytogeography and the evolutionary significance of B chromosomes in relation to inverted rearrangements in a grasshopper species. **Cytogenetic and genome research**, v. 106, n. 2-4, p. 351-358, 2004.

COWLEY, Michael; OAKEY, Rebecca J. Transposable elements re-wire and fine-tune the transcriptome. **PLoS genetics**, v. 9, n. 1, 2013.

DE KONING, AP Jason et al. Repetitive elements may comprise over two-thirds of the human genome. **PLoS genetics**, v. 7, n. 12, 2011.

DE QUEIROZ, Luciano Paganucci, et al. **Diversity and evolution of flowering plants of the Caatinga Domain**. En Caatinga. Springer, Cham, 2017. p. 23-63.

DE SOUZA, Thaíssa B. et al. Analysis of retrotransposon abundance, diversity and distribution in holocentric *Eleocharis* (Cyperaceae) genomes. **Annals of botany**, v. 122, n. 2, p. 279-290, 2018.

DEAKIN, Janine E. et al. Chromosomics: bridging the gap between genomes and chromosomes. **Genes**, v. 10, n. 8, p. 627, 2019.

DEANNA, Rocío et al. Patterns of chromosomal evolution in the florally diverse Andean clade Iochrominae (Solanaceae). **Perspectives in plant ecology, evolution and systematics**, v. 35, p. 31-43, 2018.

DEATHERAGE, Daniel E. et al. Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. **Proceedings of the National Academy of Sciences**, v. 114, n. 10, p. E1904-E1912, 2017.

DES MARAIS, David L.; HERNANDEZ, Kyle M.; JUENGER, Thomas E. Genotype-by-environment interaction and plasticity: exploring genomic responses of plants to the abiotic environment. **Annual Review of Ecology, Evolution, and Systematics**, v. 44, p. 5-29, 2013.

DIAZ-URIARTE, Ramon; GARLAND JR, Theodore. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. **Systematic Biology**, v. 45, n. 1, p. 27-47, 1996.

DÍEZ, Concepción M. et al. Genome size variation in wild and cultivated maize along altitudinal gradients. **New Phytologist**, v. 199, n. 1, p. 264-276, 2013.

DION-CÔTÉ, Anne-Marie, et al. Standing chromosomal variation in Lake Whitefish species pairs: the role of historical contingency and relevance for speciation. **Molecular ecology**, 2017, vol. 26, no 1, p. 178-192.

DU, Yun-peng et al. Genome size diversity in *Lilium* (Liliaceae) is correlated with karyotype and environmental traits. **Frontiers in plant science**, v. 8, p. 1303, 2017.

EKBLOM, Robert; GALINDO, Juan. Applications of next generation sequencing in molecular ecology of non-model organisms. **Heredity**, v. 107, n. 1, p. 1-15, 2011.

ESNAULT, Caroline et al. Transposable element insertions in fission yeast drive adaptation to environmental stress. **Genome research**, v. 29, n. 1, p. 85-95, 2019.

FELSENSTEIN, Joseph. Phylogenies and the comparative method. **The American Naturalist**, v. 125, n. 1, p. 1-15, 1985.

FU, Jiaping et al. Identification and characterization of abundant repetitive sequences in *Allium cepa*. **Scientific reports**, v. 9, n. 1, p. 1-7, 2019.

GAETA, Marcos Letaif et al. Occurrence and chromosome distribution of retroelements and NUPT sequences in *Copaifera langsdorffii* Desf. (Caesalpinioideae). **Chromosome Research**, v. 18, n. 4, p. 515-524, 2010.

GALINDO-GONZÁLEZ, Leonardo et al. LTR-retrotransposons in plants: engines of evolution. **Gene**, v. 626, p. 14-25, 2017.

GAGNON, E. et al. A molecular phylogeny of *Caesalpinia sensu lato*: Increased sampling reveals new insights and more genera than expected. **South African Journal of Botany**, v. 89, p. 111-127, 2013.

GAGNON, Edeline et al. A new generic system for the pantropical Caesalpinia group (Leguminosae). **PhytoKeys**, n. 71, p. 1, 2016.

GAGNON, Edeline et al. Global succulent biome phylogenetic conservatism across the pantropical Caesalpinia group (leguminosae). **New Phytologist**, v. 222, n. 4, p. 1994-2008, 2019.

GAO, Dongying et al. Transposons play an important role in the evolution and diversification of centromeres among closely related species. **Frontiers in plant science**, v. 6, p. 216, 2015.

GARNATJE, Teresa et al. Genome size in *Echinops* L. and related genera (Asteraceae, Cardueae): karyological, ecological and phylogenetic implications. **Biology of the Cell**, v. 96, n. 2, p. 117-124, 2004.

GONZÁLEZ, María Laura; CHIAPELLA, Jorge Oscar; URDAMPILLETA, Juan Domingo. Characterization of some satellite DNA families in *Deschampsia antarctica* (Poaceae). **Polar Biology**, v. 41, n. 3, p. 457-468, 2018.

GRAFEN, Alan. The phylogenetic regression. Philosophical Transactions of the Royal Society of London. B, **Biological Sciences**, v. 326, n. 1233, p. 119-157, 1989.

GREWAL, Shiv IS; JIA, Songtao. Heterochromatin revisited. **Nature Reviews Genetics**, v. 8, n. 1, p. 35-46, 2007.

GUERRA, Marcelo. Patterns of heterochromatin distribution in plant chromosomes. **Genetics and molecular biology**, v. 23, n. 4, p. 1029-1041, 2000.

HARVEY, Paul H. et al. **The comparative method in evolutionary biology**. Oxford: Oxford university press, 1991.

HEITZ, Emil. **Das heterochromatin der moose**. Bornträger, 1928.

HESLOP-HARRISON, J. S; SCHWARZACHER, Trude. Domestication, genomics and the future for banana. **Annals of botany**, v. 100, n. 5, p. 1073-1084, 2007.

HESLOP-HARRISON, J. S.; SCHWARZACHER, Trude. Organisation of the plant genome in chromosomes. **The Plant Journal**, v. 66, n. 1, p. 18-33, 2011.

HIDALGO, Oriane et al. Is there an upper limit to genome size?. **Trends in plant science**, v. 22, n. 7, p. 567-573, 2017.

HLOUŠKOVÁ, Petra et al. The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. **Annals of Botany**, v. 124, n. 1, p. 103-120, 2019.

HORVÁTH, Vivien; MERENCIANO, Miriam; GONZÁLEZ, Josefa. Revisiting the relationship between transposable elements and the eukaryotic stress response. **Trends in Genetics**, v. 33, n. 11, p. 832-841, 2017.

JAKOB, Sabine S.; MEISTER, Armin; BLATTNER, Frank R. The considerable genome size variation of *Hordeum* species (Poaceae) is linked to phylogeny, life form, ecology, and speciation rates. **Molecular Biology and Evolution**, v. 21, n. 5, p. 860-869, 2004.

JORDAN, Gregory J. et al. Environmental adaptation in stomatal size independent of the effects of genome size. **New Phytologist**, v. 205, n. 2, p. 608-617, 2015.

KANAZAWA, Akira et al. Adaptive evolution involving gene duplication and insertion of a novel Ty1/copia-like retrotransposon in soybean. **Journal of Molecular Evolution**, v. 69, n. 2, p. 164-175, 2009.

KANG, Ming et al. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. **New Phytologist**, v. 202, n. 4, p. 1371-1381, 2014.

KAWAKAMI, T. et al. Different scales of Ty1/copia-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species. **Heredity**, v. 104, n. 4, p. 341-350, 2010.

KAZAZIAN, Haig H. Mobile elements: drivers of genome evolution. **Science**, v. 303, n. 5664, p. 1626-1632, 2004.

KELLY, Laura 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, n. 2, p. 596-607, 2015.

LAMO, J. M.; WAHLANG, D. R.; RAO, S. R. Comparative analysis of heterochromatin distribution pattern in wild and cultivated species of *Curcuma* L. **Med. Aromat. Plants**, v. 3, p. 006, 2016.

LAN, Tianying; ALBERT, Victor A. Dynamic distribution patterns of ribosomal DNA and chromosomal evolution in *Paphiopedilum*, a lady's slipper orchid. **BMC Plant Biology**, v. 11, n. 1, p. 126, 2011.

LANCIANO, Sophie; MIROUZE, Marie. Transposable elements: all mobile, all different, some stress responsive, some adaptive?. **Current opinion in genetics & development**, v. 49, p. 106-114, 2018.

LEE, Yung-I. et al. Satellite DNA in *Paphiopedilum* subgenus *Parvisepalum* as revealed by high-throughput sequencing and fluorescent in situ hybridization. **BMC genomics**, v. 19, n. 1, p. 578, 2018.

LEGUME PHYLOGENY WORKING GROUP et al. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. 2017.

LERSTEN, Nels R.; CURTIS, John D. Survey of leaf anatomy, especially secretory structures, of tribe Caesalpinieae (Leguminosae, Caesalpinioideae). **Plant Systematics and Evolution**, v. 200, n. 1-2, p. 21-39, 1996.

LISCH, Damon. How important are transposons for plant evolution?. **Nature Reviews Genetics**, v. 14, n. 1, p. 49-61, 2013.

LIU, Qing et al. The repetitive DNA landscape in Avena (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. **BMC plant biology**, v. 19, n. 1, p. 226, 2019.

LÓPEZ, A. et al. Cytogenetic characterization of *Caesalpinia spinosa* from Tarma and Palca (Junín). **Revista Peruana de Biología**, v. 20, n. 3, p. 245-248, 2014.

LYU, Haomin et al. Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. **New phytologist**, v. 217, n. 1, p. 428-438, 2018.

MA, Jianxin et al. Plant centromere organization: a dynamic structure with conserved functions. **TRENDS in Genetics**, v. 23, n. 3, p. 134-139, 2007.

MACAS, Jiří; MESZAROS, Tibor; NOUZOVA, Marcela. PlantSat: a specialized database for plant satellite repeats. **Bioinformatics**, v. 18, n. 1, p. 28-35, 2002.

MACAS, Jiří 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, 2015.

MANDÁKOVÁ, Terezie et al. Monophyletic origin and evolution of the largest crucifer genomes. **Plant Physiology**, v. 174, n. 4, p. 2062-2071, 2017.

MENEZES, Rodolpho ST et al. The roles of barriers, refugia, and chromosomal clines underlying diversification in Atlantic Forest social wasps. **Scientific reports**, v. 7, n. 1, p. 1-16, 2017.

MEYERS, Blake C.; TINGEY, Scott V.; MORGANTE, Michele. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. **Genome Research**, v. 11, n. 10, p. 1660-1676, 2001.

MORENO, Natalia Cecilia et al. Karyotypes, heterochromatin distribution and rDNA patterns in South American *Grindelia* (Asteraceae). **Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology**, v. 152, n. 1, p. 166-174, 2018.

NEGI, Pooja; RAI, Archana N.; SUPRASANNA, Penna. Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. **Frontiers in plant science**, v. 7, p. 1448, 2016.

NEUMANN, Pavel et al. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. **Mobile DNA**, v. 10, n. 1, p. 1, 2019.

NIU, Xiao-Min et al. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. **Proceedings of the National Academy of Sciences**, v. 116, n. 14, p. 6908-6913, 2019.

NOVÁK, Petr et al. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. **Bioinformatics**, v. 29, n. 6, p. 792-793, 2013.

NOVÁK, Petr; NEUMANN, Pavel; MACAS, Jiří. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. **BMC bioinformatics**, v. 11, n. 1, p. 378, 2010.

NOVÁK, Petr et al. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. **Nucleic acids research**, v. 45, n. 12, p. e111-e111, 2017.

O'MEARA, Brian C. Evolutionary inferences from phylogenies: a review of methods. **Annual Review of Ecology, Evolution, and Systematics**, v. 43, p. 267-285, 2012.

PELLICER, Jaume et al. Genome size diversity and its impact on the evolution of land plants. **Genes**, v. 9, n. 2, p. 88, 2018.

PERUZZI, Lorenzo; LEITCH, I. J.; CAPARELLI, K. F. Chromosome diversity and evolution in Liliaceae. **Annals of Botany**, v. 103, n. 3, p. 459-475, 2009.

PLUESS, Andrea R. et al. Genome–environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. **New Phytologist**, v. 210, n. 2, p. 589-601, 2016.

POGGIO, Lidia; ESPERT, Shirley M.; FORTUNATO, Renée H. Citogenética evolutiva en leguminosas americanas. **Rodriguésia**, p. 423-433, 2008.

REY, Olivier et al. Adaptation to global change: a transposable element–epigenetics perspective. **Trends in ecology & evolution**, v. 31, n. 7, p. 514-526, 2016.

RIBEIRO, Tiago et al. Evolutionary dynamics of satellite DNA repeats from *Phaseolus* beans. **Protoplasma**, v. 254, n. 2, p. 791-801, 2017.

RICHARD, Guy-Franck; KERREST, Alix; DUJON, Bernard. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. **Microbiol. Mol. Biol. Rev**., v. 72, n. 4, p. 686-727, 2008.

RODRIGUES, P. S. et al. Karyomorphology and karyotype asymmetry in the South American Caesalpinia species (Leguminosae and Caesalpinioideae). **Genetics and Molecular Research**, v. 13, n. 4, p. 8278-8293, 2014.

RODRIGUES, Polliana Silva et al. Karyotype diversity and 2C DNA content in species of the Caesalpinia group. **BMC genetics**, v. 19, n. 1, p. 25, 2018.

RODRIGUES, Polliana Silva; SOUZA, Margarete Magalhães; CORRÊA, Ronan Xavier. Karyomorphology of Caesalpinia Species (Caesalpinioideae: Fabaceae) from Caatinga and Mata Atlantica Biomes of Brazil. **Journal of Plant Studies**, v. 1, n. 2, p. 82, 2012.

SAMOLUK, Sergio S. et al. Heterochromatin evolution in *Arachis* investigated through genome-wide analysis of repetitive DNA. **Planta**, v. 249, n. 5, p. 1405-1415, 2019.

SCHRADER, Lukas; SCHMITZ, Jürgen. The impact of transposable elements in adaptive evolution. **Molecular ecology**, v. 28, n. 6, p. 1537-1549, 2019.

SCHRADER, Lukas et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. **Nature communications**, v. 5, p. 5495, 2014.

SCHWEIZER, Dieter. Reverse fluorescent chromosome banding with chromomycin and DAPI. **Chromosoma**, v. 58, n. 4, p. 307-324, 1976.

SHENDURE, Jay; JI, Hanlee. Next-generation DNA sequencing. **Nature biotechnology**, v. 26, n. 10, p. 1135, 2008.

SIMPSON, Beryl et al. Phylogeny and biogeography of *Pomaria* (Caesalpinioideae: leguminosae). **Systematic Botany**, v. 31, n. 4, p. 792-804, 2006.

SIMPSON, Beryl B.; TATE, Jennifer A.; WEEKS, Andrea. The biogeography of *Hoffmannseggia* (Leguminosae, Caesalpinioideae, Caesalpinieae): a tale of many travels. **Journal of Biogeography**, v. 32, n. 1, p. 15-27, 2005.

SLOTKIN, R. Keith; MARTIENSSEN, Robert. Transposable elements and the epigenetic regulation of the genome. **Nature Reviews Genetics**, v. 8, n. 4, p. 272-285, 2007.

ŠMARDA, Petr; BUREŠ, Petr; HOROVÁ, Lucie. Random distribution pattern and non-adaptivity of genome size in a highly variable population of *Festuca pallens*. **Annals of Botany**, v. 100, n. 1, p. 141-150, 2007.

SOUZA, Maria Goreti C.; BENKO-ISEPPON, Ana M. Cytogenetics and chromosome banding patterns in Caesalpinioideae and Papilionioideae species of Pará, Amazonas, Brazil. **Botanical Journal of the Linnean Society**, v. 144, n. 2, p. 181-191, 2004.

SOUZA, Gustavo et al. Do tropical plants have smaller genomes? Correlation between genome size and climatic variables in the Caesalpinia Group (Caesalpinioideae, Leguminosae). **Perspectives in plant ecology, evolution and systematics**, v. 38, p. 13-23, 2019.

STRAUB, Shannon CK et al. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. **American Journal of Botany**, v. 99, n. 2, p. 349-364, 2012.

SUN, Jin et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. **Nature Ecology & Evolution**, v. 1, n. 5, p. 1-7, 2017.

TAPIA-PASTRANA, Fernando; MERCADO-RUARO, Pedro; GÓMEZ-ACEVEDO, Sandra. Contribución a la citogenética de *Tamarindus indica* (Leguminosae: Caesalpinioideae). **Acta botánica mexicana**, n. 98, p. 99-110, 2012.

THOMAS, Gregg WC; HAHN, Matthew W. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. **Molecular biology and evolution**, v. 32, n. 5, p. 1232-1236, 2015.

VAN-LUME, Brena 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, 2017.

VAN-LUME, Brena et al. Evolutionary convergence or homology? Comparative cytogenomics of Caesalpinia group species (Leguminosae) reveals diversification in the pericentromeric heterochromatic composition. **Planta**, v. 250, n. 6, p. 2173-2186, 2019.

WAHLANG, Daniel Regie et al. Karyo-morphological consistency and heterochromatin distribution pattern in diploid and colchitetraploids of *Vigna radiata* and V. mungo. **Meta Gene**, v. 21, p. 100569, 2019.

WEICHENHAN, D. et al. Evolution by fusion and amplification: the murine Sp100-rs gene cluster. **Cytogenetic and Genome Research**, v. 80, n. 1-4, p. 226-231, 1998.

WEISS-SCHNEEWEISS, Hanna; SCHNEEWEISS, Gerald M. Karyotype diversity and evolutionary trends in angiosperms. In: **Plant Genome Diversity Volume** 2. Springer, Vienna, 2013. p. 209-230.

WEISS-SCHNEEWEISS, Hanna et al. Karyotype diversification and evolution in diploid and polyploid South American *Hypochaeris* (Asteraceae) inferred from rDNA localization and genetic fingerprint data. **Annals of Botany**, v. 101, n. 7, p. 909-918, 2008.

WERNECK, Fernanda P. et al. Revisiting the historical distribution of Seasonally Dry Tropical Forests: new insights based on palaeodistribution modelling and palynological evidence. **Global Ecology and Biogeography**, v. 20, n. 2, p. 272-288, 2011.

YANG, Ji et al. Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. **Molecular biology and evolution**, v. 33, n. 10, p. 2576-2592, 2016.

ZOU, Xiaoxin; ZHANG, Yu. Noble metal-free hydrogen evolution catalysts for water splitting. **Chemical Society Reviews**, v. 44, n. 15, p. 5148-5180, 2015.

GLOSSÁRIO

· **Citogenômica:** área da genômica que está orientada à compreensão, expressão e interação dos genes no nível celular, empregando diferentes metodologias citogenéticas.

· **Citogeografia:** é a análise da distribuição geográfica de marcadores citológicos polimórficos, considerada uma ferramenta importante a ser aplicada no estudo do significado evolutivo da variabilidade cromossômica em uma espécie (Colombo e Confalonieri 2004). Essas informações não apenas ajudam a elucidar quais tipos de processos estão operando, a fim de produzir um padrão particular de distribuição da diversidade, mas também fornece insights adicionais sobre o significado da persistência de tais padrões nas populações naturais.

· **Clustering analysis:** ou análise de cluster é um algoritmo de agrupamento baseado em similaridade que avalia por comparações de sequência todos entre todos os reads de *whole-genome shotgun* (Novák et al. 2010). A análise de *clustering* empregado pelo RepeatExplorer representa os reads e suas semelhanças de sequências representadas como nós e arestas de conexão (Fig. 1a-b), em um gráfico virtual que identifica os clusters dos reads pela topologia do gráfico (Fig. 1c) (Novák et al. 2017). O número de reads em cada cluster é proporcional à abundância genômica da repetição correspondente, possibilitando sua quantificação.

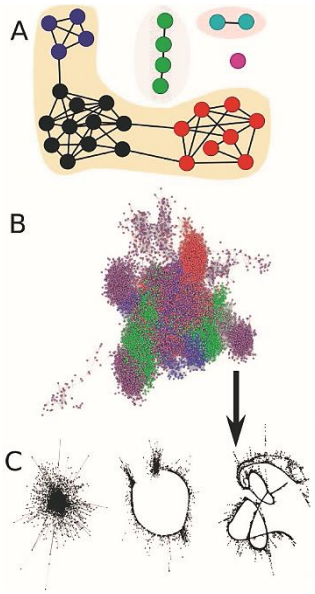


Figura 1. Reads de sequência organizadas em uma estrutura gráfica tomada de Novák et al. (2010). Reads únicos são representadas por vértices (nós) e suas sequências se sobrepõem por arestas. (A) exemplos de diferentes tipos de clusters que podem ser encontrados na estrutura do gráfico. Os nós com a mesma cor correspondem aos clusters (comunidades) identificados usando um algoritmo de aglomeração hierárquica. O nó magenta representa um *single* - um read sem similaridade com outras sequências. (B) um exemplo de um gráfico construído a partir de reads amostradas de *Pisum sativum*. (C) Gráficos calculados usando o algoritmo para três grupos com estrutura diferente.

· **Coloração CMA/DAPI:** técnica de coloração diferencial com fluorocromos que torna visíveis os blocos de heterocromatina de acordo com sua composição de pares de base específica. O 4'-6-diamidino-2-fenilindol (DAPI) colore preferencialmente as regiões de

heterocromatina ricas em AT, enquanto a cromomicina A3 (CMA) colore preferencialmente as regiões ricas em GC (Schweizer 1976). Esses mesmos fluorocromos também podem corar negativamente os blocos de heterocromatina pobres em AT ou GC, enquanto a dupla coloração com os dois fluorocromos de especificidade de bases diferentes (por exemplo, CMA / DAPI) pode destacar diferenças de coloração representadas como bandas.

· **DNA satélite**: Ou satDNA é uma classe de DNA repetitivo que se caracteriza por sua organização genômica em longas matrizes de unidades dispostas em conjunto, chamados monômeros. As sequências monoméricas são tipicamente centenas de nucleotídeos repetidos em tandem (um do lado do outro) e altamente homogeneizadas (Macas et al. 2002).

O comprimento do monômero é frequentemente usado para classificar as repetições genômicas em tandem como microssatélites (2 a 7 bp), minissatélites (dezenas de bp) ou satélites (centenas de bp). Entretanto parece que os satélites são melhor distinguidos através da formação de matrizes mais longas (dezenas de kilobases para megabases) concentradas em relativamente poucos locos genômicos, enquanto matrizes de micro e minisatélite são muito mais curtas e espalhadas pelo genoma (Novák et al. 2017).

· **DNA transposons**: ET da classe II presentes em baixa a moderadas copias no genoma vegetal e são caracterizados, em alguns casos, pelas suas repetições terminais invertidos (TIRs) de comprimento variável. São conhecidos como ET de "corte e cola" porque replicam-se através de um intermediário de DNA cortado durante a transposição (Wicker et al. 2007). Esses transposons podem aumentar seus números transpondo durante a replicação cromossômica de uma posição que já foi replicada para outra onde a forquilha de replicação ainda não passou.

· **Elementos transponíveis (ETs)**: são pequenas sequências repetidas que não contêm genes com aparente importância para a sobrevivência imediata do organismo. Em vez disso, eles contêm apenas informações genéticas suficientes para produzir cópias de si mesmas e/ou movimentar-se no genoma.

Sua classificação encontra-se baseada na presença ou ausência de um intermediário de transposição de RNA ou DNA. Transposição via RNA para retrotransposon classe I e via DNA para transposons classe II, cada classe também é subdividida em várias ordens,

superfamílias e famílias (Fig. 2). O mecanismo de transposição da classe I é comumente chamado " copiar e colar" e o da classe II, "cortar e colar" (Wicker et al. 2007).

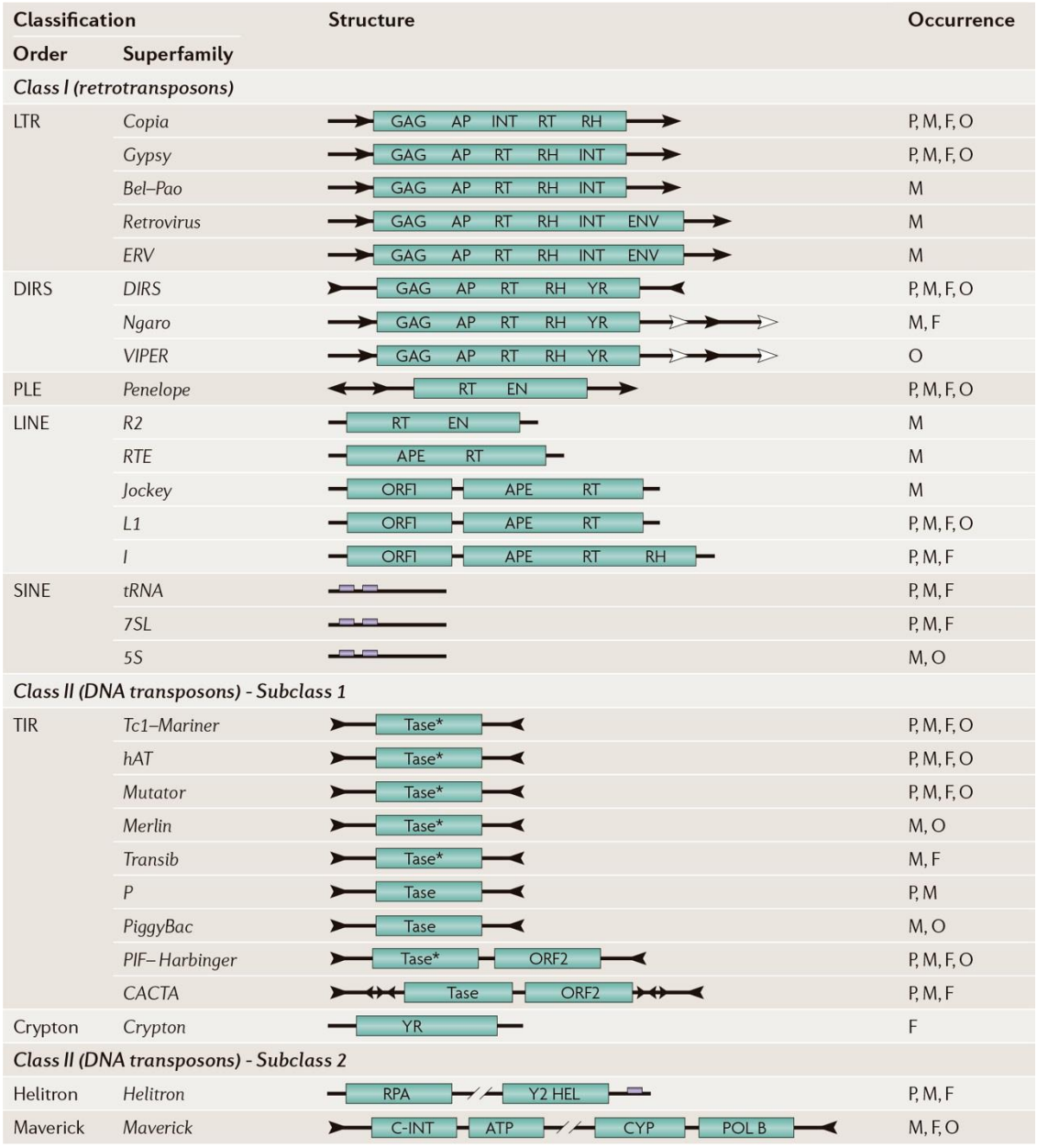


Figura 2. Sistema de classificação proposto por Wicker et al. (2007) para elementos transponíveis (ETs). A classificação é hierárquica e divide os ETs em duas classes principais com base na presença ou ausência de RNA como intermediário de transposição. Eles são subdivididos em subclasses, ordens e superfamílias. Dirs, sequência de repetição intermediária de dictyostelium; LiNE, elemento nuclear intercalado há muito tempo; LTR, repetição terminal longa; PLE, elementos do tipo Penélope; SINE, elemento nuclear intercalado curto; TIR, repetição terminal invertida. P, Plantas; M, Metazoários; F, Fungos e O, Outros.

· **FISH**: A hibridação *in situ* fluorescente (do inglês FISH) envolve uma ligação não covalente de uma determinada sequência de DNA de fita simples ou RNA, denominada sonda, com uma sequência de nucleotídeos complementar situada na célula afixado a uma lâmina (Fig. 3) (Brasileiro-Vidal e Guerra 2002). A sonda pode ser detectada direta ou indiretamente por causa de uma etiqueta fluorescente que permite a visualização com microscopia de fluorescência (Birchler e Danilova 2016).

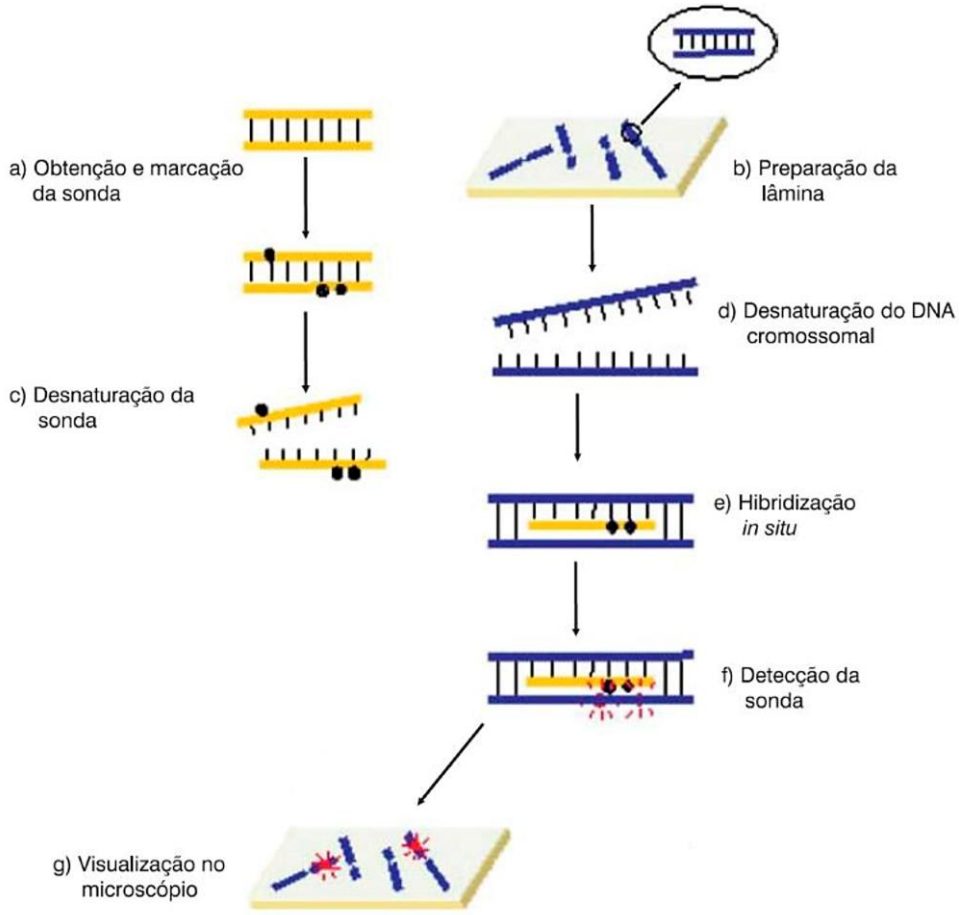


Figura 3. Esquema simplificado da técnica de hibridização *in situ* fluorescente tomada de Brasileiro-Vidal e Guerra (2002). (a) A sequência de DNA a ser usada como sonda deve ser isolada e marcada. (b) paralelamente, devem ser preparadas lâminas com cromossomos espalhados. (c, d, e) posteriormente, os DNAs da sonda e dos cromossomos devem ser desnaturados e colocados em contato para que ocorra a hibridização *in situ*. (f, g) A localização da sonda é feita por uma molécula reconhecedora ligada a um fluorocromo e observada sob microscopia de fluorescência.

· **Genome skimming**: termo empregado por Straub et al. (2012) como uma maneira de descrever o sequenciamento superficial e aleatório do DNA genômico, que resulta em uma caracterização de uma pequena porcentagem da fração de alta cópia do genoma (plastoma, mitogenoma e repeatoma).

· **Heterocromatina (HC)**: Heitz (1928) descreveu a heterocromatina como a fração genômica mais compacta, mantendo altos níveis de condensação ao longo do ciclo celular. Essas regiões densamente compactadas no núcleo são quase exclusivamente compostas por DNA repetitivo (Avramova, 2002), que também podem ser organizadas em pequenas ilhas heterocromaticas não detectáveis por análise citogenética clássica (Weichenhan et al. 1998). A HC é principalmente enriquecida nas regiões centromérica e telomérica dos cromossomos, dependendo da espécie, e compõe-se frequentemente por DNA satélite, retrotransposons e/ou sequências de DNA ribossômico (Lamo et al. 2016).

· **LTR-retrotransposons**: ET de classe I que se caracteriza por ter repetição terminal longa (do inglês LTR) nas duas extremidades do elemento, e isso pode ser usado para sua detecção e caracterização. Eles são o maior e mais diversificado grupo de ETs presentes nos eucariotos e particularmente abundantes em genomas vegetais. Este tipo de retrotransposons replicam-se através de um intermediário de RNA e uma transcriptase reversa para à transposição (mecanismo de copiar e colar), gerando novas cópias de elementos que, após a integração, aumentam o tamanho do genoma do hospedeiro (Neumann et al. 2019).

· **NGS**: sequenciamento de próxima geração (do inglês NGS), é uma metodologia de sequenciamento de DNA paralelo que permite experimentos a escala do genoma completo. As vantagens das estratégias de segunda geração ou de matriz cíclica, incluem: construção e amplificação *in vitro* de uma biblioteca de sequenciamento, sequenciamento baseado em matriz que permite um grau de paralelismo maior e o os *arrays* podem ser manipulados por enzimas por um único volume de reagente, economizando na obtenção das sequencias (Shendure e Ji 2008).

· **PCMs**: os métodos filogenéticos comparativos (do inglês PCMs) foram desenvolvidos inicialmente na década de 1980, e usa informações sobre a filogenia do grupo para testar hipóteses evolutivas em uma análise comparativa (Felsenstein 1985; Grafen 1989). Desde então, os PCMs foram estendidos para investigar os padrões e os processos evolutivos das espécies (Pennell e Harmon 2013), e incluem métodos para investigar fatores de diversificação (Maddison et al. 2007), tempo e modo da evolução das características (O'Meara 2012) e modelos de especiação e extinção.

· **PICs**: o método de contrastes independentes filogenéticos (do inglês PICs). Usa informações filogenéticas para explicar o fato de que as espécies, em uma análise comparativa, são relacionadas entre si e, portanto, podem compartilhar semelhanças entre elas, porque as herdam de seus ancestrais e não por causa da evolução independente (Felsenstein 1985; Harvey e Paggel 1991). O método possui três suposições principais: (1) que a topologia da filogenia é correta; (2) que os comprimentos dos ramos da filogenia tenham suporte; e (3) que os traços evoluem da maneira do modelo de movimento browniano, um modelo simples de evolução de traços em que a variação dos traços se acumula como uma função linear do tempo (Felsenstein 1985; Diaz-Uriarte e Garland 1996).

· **RepeatExplorer:** é um pipeline computacional projetado para identificar e caracterizar elementos repetitivos de DNA em dados de sequenciamento de última geração de genomas de plantas e animais. Emprega *clustering* baseado em gráficos (Novák et al. 2010) de leituras de sequência para identificar elementos repetitivos e vários programas adicionais que auxiliam na anotação e quantificação (Novák et al. 2013).

· **SDTF**: as florestas tropicais sazonalmente secas (do inglês SDTF). É um bioma que abrange uma variedade de vegetação tropical sazonalmente seca, desde florestas decíduas de estatura baixa ou média, até vegetação espinhosa de arbustos, cactos e matagal de baixa estatura, agrupadas como floresta tropical e bosque sazonalmente secas (De Queiroz et al. 2017).

· **Tamanho do genoma**: referido como o valor C nas espécies, representa a quantidade de DNA no genoma nuclear haploide não replicado. A quantidade de DNA é frequentemente referida como tamanho do genoma (do inglês GS), mas, estritamente falando, o tamanho do genoma em eucariotos é a quantidade de DNA em um conjunto cromossômico básico (monoploide) (x). Assim, o valor C é igual ao tamanho do genoma para espécies diploides, enquanto que para poliploides é a soma de dois ou mais genomas (Pellicer et al. 2018).